2 Continued Fractions

Continued fractions are expressions of the following sort:

$$\frac{7}{16} = \cfrac{1}{2 + \cfrac{1}{3 + \cfrac{1}{2}}} \qquad \frac{67}{24} = 2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{4}}}}$$

The numerators in these two examples are all 1, and we will only be considering continued fractions of this type, although there are situations outside the scope of this book where other numerators are allowed.

To compute the value of a continued fraction one starts in the lower right corner and works one's way upward. For example, in the continued fraction for $7/16$ one starts with $3 + 1/2 = 7/2$, then taking 1 over this gives $2/7$, and adding 2 to this gives $16/7$, and finally 1 over this gives $7/16$. In the case of the continued fraction for $67/24$ the fractions arising by this process are $5/4$, $4/5$, $19/5$, $5/19$, $24/19$, $19/24$, and finally $67/24$. As we will see, there is a fairly simple way to express every rational number as a continued fraction.

The main theme of this chapter will be the close relationship between continued fractions and the Farey diagram. For example, the fact that all rational numbers occur as labels on vertices in the Farey diagram is a reflection of the fact that every rational number has an expression as a continued fraction. In fact the continued fraction for a rational number p/q will tell how to locate the vertex labeled p/q in the diagram, and conversely, from the location of the vertex p/q one can read off the continued fraction for p/q.

We will also consider continued fractions with infinitely many terms extending downward to the right. These will give expressions for irrational numbers, somewhat like expressing irrational numbers as infinite decimals. Continued fractions have the advantage that rational numbers are expressible as finite continued fractions whereas the decimal representations for rational numbers are not generally finite but are instead just eventually periodic. Infinite continued fractions that are eventually periodic correspond to a special class of irrational numbers, those that are roots of quadratic equations with integer coefficients, like $\sqrt{2}$. Thus continued fractions are better than decimals in some ways, but on the other hand simple operations like addition and multiplication of rational numbers do not have nice descriptions in terms of contin-

ued fractions. In spite of these limitations continued fractions are quite useful in Number Theory. Among other things, they can be used to solve certain Diophantine equations including linear Diophantine equations, as we will see in Section 2.3.

2.1 Finite Continued Fractions

The continued fractions we will be considering have the form shown at the right. The numbers a_i are assumed to be positive integers except for a_0 which can be any integer, possibly negative or zero. When a_0 is zero it can be omitted from the formula. To write a continued fraction in more compact form on a single line, we will often write it as $p/q = a_0 + 1/a_1 + 1/a_2 + \cdots + 1/a_n$ with diagonal arrows to indicate the extended horizontal bars in the previous notation, for example $7/16 = 1/2 + 1/3 + 1/2$ and $67/24 = 2 + 1/1 + 1/3 + 1/1 + 1/4$. An even more concise notation that is sometimes used is $[a_0; a_1, a_2, \cdots, a_n]$, or just $[a_1, a_2, \cdots, a_n]$ when there is no a_0 term. However, we will use the more suggestive arrow notation in this book.

$$\frac{p}{q} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_n}}}$$

To compute the continued fraction for a given rational number, one starts in the upper left corner and works one's way downward, as the following example shows:

$$\frac{67}{24} = 2 + \frac{19}{24} = 2 + \frac{1}{24/19} = 2 + \frac{1}{1 + 5/19} = 2 + \cfrac{1}{1 + \cfrac{1}{19/5}}$$

$$= 2 + \cfrac{1}{1 + \cfrac{1}{3 + 4/5}} = 2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{5/4}}} = 2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{4}}}}$$

The key steps are the equations $67/24 = 2 + 19/24$, $24/19 = 1 + 5/19$, $19/5 = 3 + 4/5$, and $5/4 = 1 + 1/4$. If we clear fractions in each of these equations we obtain the first four of the five equations at the right which show a sequence of repeated divisions starting with a given pair of positive integers, 67 and 24 in this case. One first divides the smaller number into the larger to obtain a quotient and a remainder which is smaller than the divisor. Then at each successive step one divides the previous remainder into the previous divisor. The process stops when one obtains a remainder of zero. This process is known as the ***Euclidean algorithm***. The numbers in the shaded box are the quotients of the successive divisions and are sometimes called the *partial quotients*. These are the numbers a_i in the continued fraction for $67/24$.

$$\begin{aligned} 67 &= 2 \cdot 24 + 19 \\ 24 &= 1 \cdot 19 + 5 \\ 19 &= 3 \cdot 5 + 4 \\ 5 &= 1 \cdot 4 + 1 \\ 4 &= 4 \cdot 1 + 0 \end{aligned}$$

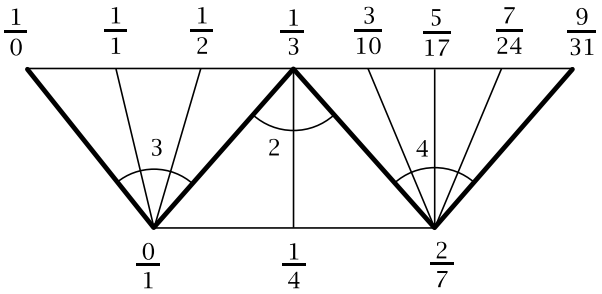
One of the classical uses for the Euclidean algorithm is to find the greatest common divisor of two given numbers. If one applies the algorithm to two numbers p and q, dividing the smaller into the larger, then the remainder into the first divisor, and so on, then the greatest common divisor of p and q turns out to be the last nonzero remainder. For example, starting with $p = 72$ and $q = 201$ the calculation is shown at the right, and the last nonzero remainder is 3, which is the greatest common divisor of 72 and 201. (In fact the fraction $^{201}/_{72}$ equals $^{67}/_{24}$, which explains why the successive quotients for this example are the same as in the preceding example.) It is easy to see from the displayed equations why 3 has to be the greatest common divisor of 72 and 201, since from the first equation it follows that any divisor of 72 and 201 must also divide 57, then the second equation shows it must divide 15, the third equation then shows it must divide 12, and the fourth equation shows it must divide 3, the last nonzero remainder. Conversely, if a number divides the last nonzero remainder 3, then the last equation shows it must also divide 12, and the next-to-last equation then shows it must divide 15, and so on until we conclude that it divides all the numbers not in the shaded rectangle, including the original two numbers 72 and 201. The same reasoning applies in general.

$$\begin{aligned} 201 &= 2\cdot 72 + 57 \\ 72 &= 1\cdot 57 + 15 \\ 57 &= 3\cdot 15 + 12 \\ 15 &= 1\cdot 12 + 3 \\ 12 &= 4\cdot 3 + 0 \end{aligned}$$

A more obvious way to try to compute the greatest common divisor of two numbers would be to factor each of them into a product of primes, then look to see which primes occurred as factors of both, and to what power. But to factor a large number into its prime factors is a very laborious and time-consuming process. For example, even a large computer would have a hard time factoring a number with a hundred or more digits into primes, so it would not be feasible to find the greatest common divisor of a pair of numbers of this size in this way. However, the computer would have no trouble applying the Euclidean algorithm to find their greatest common divisor.

Having seen what continued fractions are, let us now see what they have to do with the Farey diagram. Some examples will illustrate this best, so let us first look at the continued fraction for $^{9}/_{31}$ which is $^{1}/_{3} + {}^{1}/_{2} + {}^{1}/_{4}$. This has $3, 2, 4$ as its sequence of partial quotients, and we use these three numbers to build a strip of $3 + 2 + 4$ triangles grouped into "fans" of 3, 2, and 4 triangles:

$$\frac{9}{31} = \cfrac{1}{3 + \cfrac{1}{2 + \cfrac{1}{4}}}$$

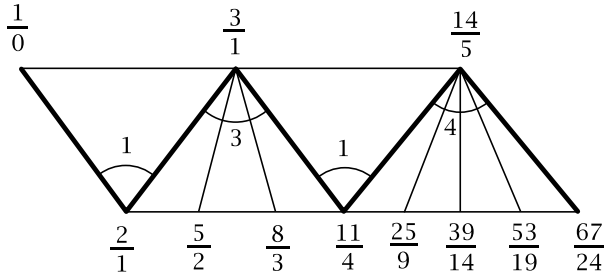


Now we begin labeling the vertices of this strip. On the left edge we start with the labels $^{1}/_{0}$ and $^{0}/_{1}$. Then we use the mediant rule for computing the third label of each

triangle in succession as we move from left to right in the strip. Thus we insert, in order, the labels $\frac{1}{1}$, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{2}{7}$, $\frac{3}{10}$, $\frac{5}{17}$, $\frac{7}{24}$, and finally $\frac{9}{31}$.

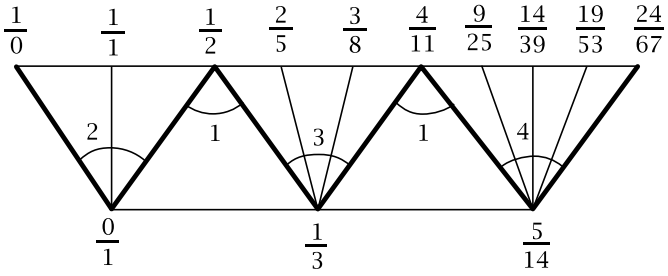
It may seem like just an accident that the final label is the fraction $\frac{9}{31}$ that we started with since the continued fraction for $\frac{9}{31}$ is computed from the equations $\frac{31}{9} = 3 + \frac{4}{9}$ and $\frac{9}{4} = 2 + \frac{1}{4}$ and these numbers have nothing to do with the fractions labeling the vertices along the strip before the final label $\frac{9}{31}$ miraculously appears. Nevertheless, we will see in Theorem 2.1 that what happened in this example always happens, at least for fractions $\frac{p}{q}$ between 0 and 1. For fractions outside this interval the procedure works if we modify it by replacing the numerator 0 of the label $\frac{0}{1}$ with a_0, the initial integer in the continued fraction $\frac{p}{q} = a_0 + \cfrac{1}{a_1 + \cdots + \cfrac{1}{a_n}}$. Thus $\frac{0}{1}$ is replaced by $\frac{a_0}{1}$. This is illustrated by the $\frac{67}{24}$ example:

$$\frac{67}{24} = 2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{4}}}}$$

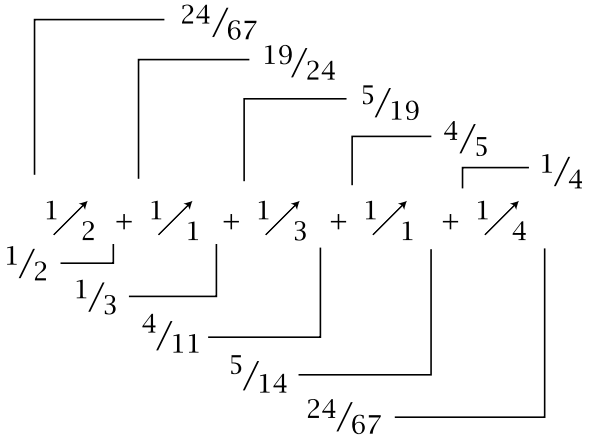


For comparison, here is the corresponding strip for the reciprocal, $\frac{24}{67}$:

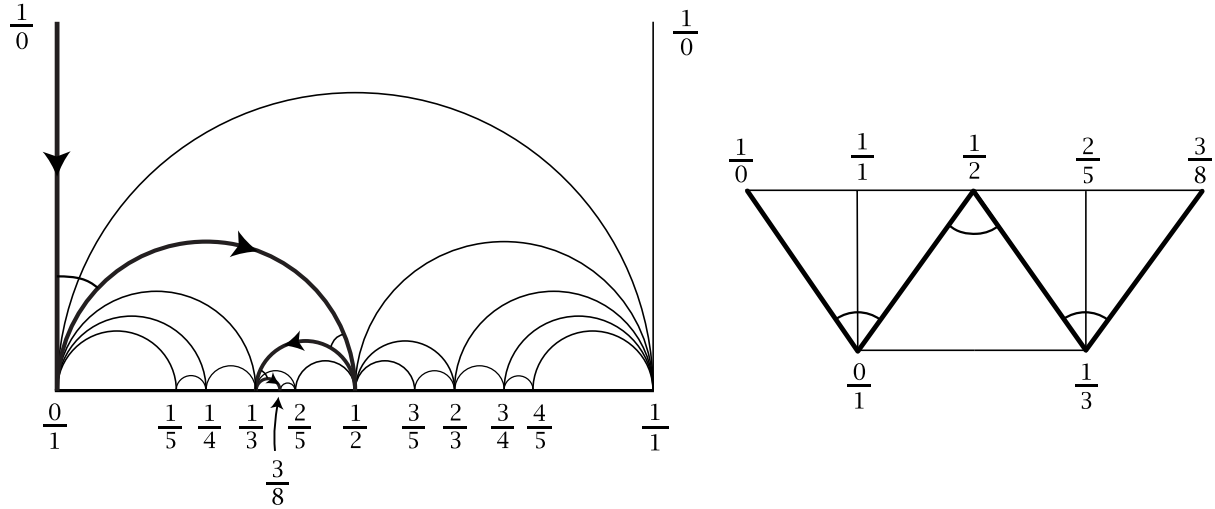
$$\frac{24}{67} = \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{4}}}}}$$



In the strip of triangles for a fraction $\frac{p}{q}$ there is a zigzag path from $\frac{1}{0}$ to $\frac{p}{q}$ that we have indicated by the heavily shaded edges. The labels on the vertices that this zigzag path passes through are the fractions that occur as the values of successively longer initial segments of the continued fraction, the continued fractions formed by the terms to the left of each plus sign in $a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots + \cfrac{1}{a_n}}}$. This is illustrated at the right for the example of $\frac{24}{67}$. These fractions are called the ***convergents*** for the given fraction. Thus the convergents for $\frac{24}{67}$ are $\frac{1}{2}$, $\frac{1}{3}$, $\frac{4}{11}$, $\frac{5}{14}$, and $\frac{24}{67}$ itself. The figure also shows the values of the terminal segments, the terms to the right of each plus sign. These are the fractions one computes in order to find the value of the continued fraction.



It is interesting to see what the zigzag paths corresponding to continued fractions look like in the upper halfplane Farey diagram. The next figure shows the simple example of the continued fraction for $3/8$. We can see here that the five triangles of the strip correspond to the four curvilinear triangles lying directly above $3/8$ in the Farey diagram, plus the fifth "triangle" extending upward to infinity, bounded on the left and right by the vertical lines above $0/1$ and $1/1$, and bounded below by the semicircle from $0/1$ to $1/1$.



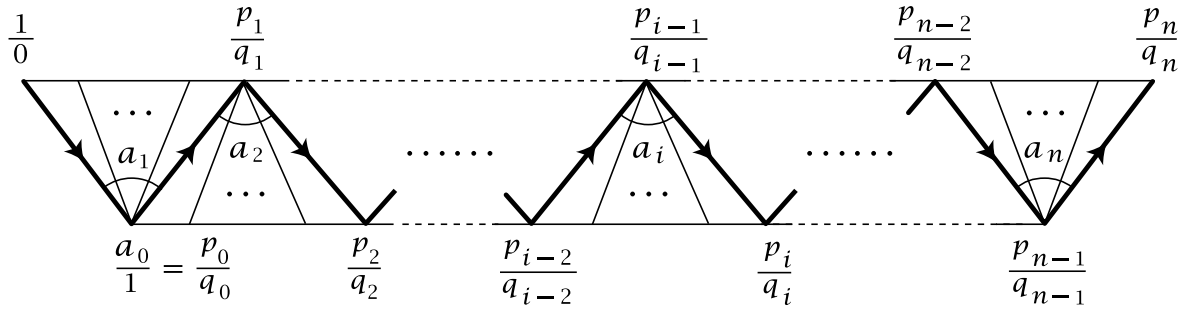
This example is typical of the general case, where the zigzag path for a continued fraction $p/q = a_0 + 1/a_1 + \cdots + 1/a_n$ becomes a "pinball path" in the Farey diagram, starting down the vertical line from $1/0$ to $a_0/1$, then turning left across a_1 triangles, then right across a_2 triangles, then left across a_3 triangles, continuing to alternate left and right turns until reaching the final vertex p/q. Two consequences of this are:

- The convergents are alternately smaller than and greater than p/q. The convergents to the left of p/q are getting successively closer to p/q from the left and the convergents to the right of p/q are getting successively closer to p/q from the right. We will see later in this section that in fact each convergent is closer to p/q than the previous one on the opposite side of p/q.
- The triangles that form the strip of triangles for p/q are exactly the triangles in the Farey diagram that lie directly above the point p/q on the x-axis. In other words, the strip of triangles for p/q consists of the triangles that the vertical line through the vertex p/q crosses.

Here is a general statement describing the relationship between continued fractions and the Farey diagram that we have observed in the preceding examples:

Theorem 2.1. *The convergents for a continued fraction $p/q = a_0 + 1/a_1 + \cdots + 1/a_n$ are the vertices along a zigzag path consisting of a finite sequence of edges in the Farey diagram, starting at $1/0$ and ending at p/q. The path starts along the edge from $1/0$ to $a_0/1$, then turns left across a fan of a_1 triangles, then right across a fan of a_2 triangles, etc., alternating left and right turns and finally ending at p/q.*

Proof: The continued fraction $p/q = a_0 + 1/a_1 + \cdots + 1/a_n$ determines a strip of triangles:



We will show that the label p_n/q_n on the final vertex in this strip is equal to p/q, the value of the continued fraction. Replacing n by i, we conclude that this holds also for each initial seqment $a_0 + 1/a_1 + \cdots + 1/a_i$ of the continued fraction. This is just saying that the vertices p_i/q_i along the strip are the convergents to p/q, which is what the theorem claims.

Each successive vertex label p_i/q_i along the zigzag path for the continued fraction $p/q = a_0 + 1/a_1 + \cdots + 1/a_n$ is computed in terms of the two preceding vertex labels according to the following formula:

$$\frac{p_i}{q_i} = \frac{a_i p_{i-1} + p_{i-2}}{a_i q_{i-1} + q_{i-2}}$$

This is because, as one can see in the figure above, the mediant rule is being applied a_i times, "adding" p_{i-1}/q_{i-1} to the previously obtained fraction each time until the next label p_i/q_i is obtained.

To prove that $p_n/q_n = p/q$ we will use 2×2 matrices. Consider the following product:

$$P = \begin{pmatrix} 1 & a_0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a_1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a_2 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & a_n \end{pmatrix}$$

We can multiply this product out starting either from the left or from the right. Suppose first that we multiply starting at the left. The two columns of the first matrix give the two fractions $1/0$ and $a_0/1$ labeling the left edge of the strip of triangles. Multiplying the first matrix by the second matrix gives:

$$\begin{pmatrix} 1 & a_0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a_1 \end{pmatrix} = \begin{pmatrix} a_0 & 1 + a_0 a_1 \\ 1 & a_1 \end{pmatrix} = \begin{pmatrix} p_0 & p_1 \\ q_0 & q_1 \end{pmatrix}$$

The two columns here give the fractions at the ends of the second edge of the zigzag path. The same thing happens for subsequent matrix multiplications, as multiplying by the next matrix in the product takes the matrix corresponding to one edge of the zigzag path to the matrix corresponding to the next edge:

$$\begin{pmatrix} p_{i-2} & p_{i-1} \\ q_{i-2} & q_{i-1} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a_i \end{pmatrix} = \begin{pmatrix} p_{i-1} & p_{i-2} + a_i p_{i-1} \\ q_{i-1} & q_{i-2} + a_i q_{i-1} \end{pmatrix} = \begin{pmatrix} p_{i-1} & p_i \\ q_{i-1} & q_i \end{pmatrix}$$

In the end, when all the matrices have been multiplied, we obtain the matrix corresponding to the last edge in the strip from p_{n-1}/q_{n-1} to p_n/q_n. Thus the second

column of the product P is $\binom{p_n}{q_n}$, and what remains to show is that this equals $\binom{p}{q}$ where p/q is the value of the continued fraction $a_0 + 1/a_1 + \cdots + 1/a_n$.

The value of the continued fraction $a_0 + 1/a_1 + \cdots + 1/a_n$ is computed by working from right to left. If we let r_i/s_i be the value of the tail $1/a_i + 1/a_{i+1} + \cdots + 1/a_n$ of the continued fraction, then we have:

$$\frac{r_n}{s_n} = \frac{1}{a_n}, \qquad \frac{r_i}{s_i} = \frac{1}{a_i + \dfrac{r_{i+1}}{s_{i+1}}} = \frac{s_{i+1}}{a_i s_{i+1} + r_{i+1}}, \qquad \text{and} \qquad \frac{p}{q} = a_0 + \frac{r_1}{s_1} = \frac{a_0 s_1 + r_1}{s_1}$$

Expressed in terms of matrices these equations become:

$$\begin{pmatrix} r_n \\ s_n \end{pmatrix} = \begin{pmatrix} 1 \\ a_n \end{pmatrix}, \qquad \begin{pmatrix} 0 & 1 \\ 1 & a_i \end{pmatrix} \begin{pmatrix} r_{i+1} \\ s_{i+1} \end{pmatrix} = \begin{pmatrix} s_{i+1} \\ r_{i+1} + a_i s_{i+1} \end{pmatrix} = \begin{pmatrix} r_i \\ s_i \end{pmatrix}$$

$$\text{and} \quad \begin{pmatrix} 1 & a_0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ s_1 \end{pmatrix} = \begin{pmatrix} r_1 + a_0 s_1 \\ s_1 \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix}$$

This means that when we multiply out the product P starting from the right, the second columns will be successively $\binom{r_n}{s_n}$, $\binom{r_{n-1}}{s_{n-1}}$, $\cdots$, $\binom{r_1}{s_1}$, and finally $\binom{p}{q}$. We have already shown that the second column of P is $\binom{p_n}{q_n}$, so $p/q = p_n/q_n$ and the proof is complete. □

An interesting fact that can be deduced from the preceding proof is that for a continued fraction $1/a_1 + \cdots + 1/a_n$ with no initial integer a_0, if we reverse the order of the numbers a_i, this leaves the denominator unchanged. For example:

$$1/2 + 1/3 + 1/4 = \frac{13}{30} \qquad \text{and} \qquad 1/4 + 1/3 + 1/2 = \frac{7}{30}$$

To see why this must always be true we use the operation of transposing a matrix to interchange its rows and columns. For a 2×2 matrix this just amounts to interchanging the upper-right and lower-left entries, so the transpose of a matrix $A = \binom{a\ b}{c\ d}$ is $A^T = \binom{a\ c}{b\ d}$. Transposing a product of matrices reverses the order of the factors, so one has $(AB)^T = B^T A^T$ as the reader can check by direct calculation. In the product

$$\begin{pmatrix} 0 & 1 \\ 1 & a_1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & a_2 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & a_n \end{pmatrix} = \begin{pmatrix} p_{n-1} & p_n \\ q_{n-1} & q_n \end{pmatrix}$$

the individual matrices on the left side of the equation are symmetric with respect to transposition, so the transpose of the product is obtained by just reversing the order of the factors:

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

Thus we see that reversing the order of the terms $a_1, \cdots, a_n$ leaves the denominator q_n unchanged, as claimed.

There is also a fairly simple relationship between the numerators. In the example of $13/30$ and $7/30$ we see that the product of the numerators, 91, is congruent to 1 modulo the denominator. In the general case the product of the numerators is

$p_n q_{n-1}$ and this is congruent to $(-1)^{n+1}$ modulo the denominator q_n. To verify this, we note that the determinant of each factor $\left(\begin{smallmatrix} 0 & 1 \\ 1 & a_i \end{smallmatrix}\right)$ is -1 so since the determinant of a product is the product of the determinants, we have $p_{n-1}q_n - p_n q_{n-1} = (-1)^n$, which implies that $p_n q_{n-1}$ is congruent to $(-1)^{n+1}$ modulo q_n.

Exercises

1. (a) Compute the values of the continued fractions $1/1 + 1/3 + 1/5 + 1/7$ and $1/1 + 1/1 + 1/1 + 1/1 + 1/1 + 1/2$.

(b) Compute the continued fraction expansions of $19/44$ and $101/1020$.

(c) Draw the strips of triangles corresponding to the continued fractions in parts (a) and (b).

2. (a) Compute the continued fraction for $38/83$ and display the steps of the Euclidean algorithm for 38 and 83 as a sequence of equations involving only integers.

(b) For the same number $38/83$ compute the associated strip of triangles grouped into fans, including the labeling of the vertices of all the triangles.

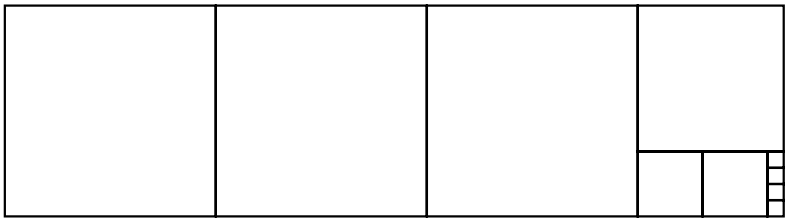
(c) Take the continued fraction $1/a_1 + 1/a_2 + \cdots + 1/a_n$ you got in part (a) and reverse the order of the numbers a_i to get a continued fraction $1/a_n + 1/a_{n-1} + \cdots + 1/a_1$. Compute the value p/q of this continued fraction, and also compute the strip of triangles for this fraction p/q. What is the relationship between p/q and $38/83$?

3. Let p_n/q_n be the value of the continued fraction $1/a_1 + 1/a_2 + \cdots + 1/a_n$ where each of the n terms a_i is equal to 2. Thus $p_1/q_1 = 1/2$, $p_2/q_2 = 1/2 + 1/2 = 2/5$, etc.

(a) Find equations expressing p_n and q_n in terms of p_{n-1} and q_{n-1}, and use these to write down the values of p_n/q_n for $n = 1, 2, 3, 4, 5, 6, 7$.

(b) Compute the strip of triangles for p_7/q_7.

4. (a) A rectangle with sides of length 13 and 48 can be partitioned into squares in the way shown in the figure at the right. Determine the lengths of the sides of all the squares, and relate the numbers of squares of each size to the continued fraction for $13/48$.



(b) Draw the analogous figure decomposing a rectangle of sides 19 and 42 into squares, and relate this to the continued fraction for $19/42$.

5. This exercise is intended to illustrate the proof of Theorem 2.1 in the concrete case of the continued fraction $1/2 + 1/3 + 1/4 + 1/5$.

(a) Write down the product $A_1A_2A_3A_4 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & a_1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 1 \\ 1 & a_2 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 1 \\ 1 & a_3 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 1 \\ 1 & a_4 \end{smallmatrix}\right)$ associated to $1/2 + 1/3 + 1/4 + 1/5$.

(b) Compute the four matrices A_1, A_1A_2, $A_1A_2A_3$, $A_1A_2A_3A_4$ and relate these to the edges of the zigzag path in the strip of triangles for $1/2 + 1/3 + 1/4 + 1/5$.

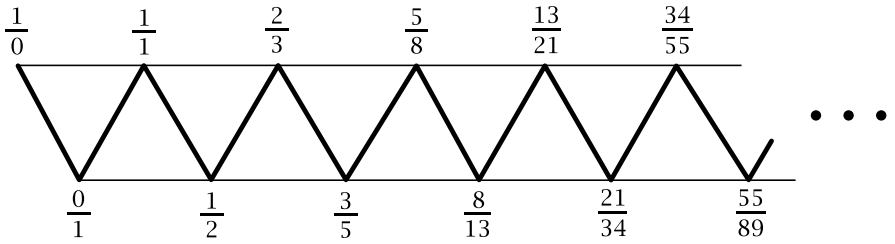
(c) Compute the four matrices A_4, A_3A_4, $A_2A_3A_4$, $A_1A_2A_3A_4$ and relate these to the successive fractions that one gets when one computes the value of $1/2+1/3+1/4+1/5$, namely $1/5$, $1/4 + 1/5$, $1/3 + 1/4 + 1/5$, and $1/2 + 1/3 + 1/4 + 1/5$.

6. Compute the strip of triangles corresponding to the continued fraction for $7/19$ and compare this with the sequence of matrices reducing $\left(\begin{smallmatrix} 3 & 7 \\ 8 & 19 \end{smallmatrix}\right)$ to $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ by a sequence of operations subtracting one column from the other. (See the proof of Proposition 1.1.)

7. Show that the continued fraction for a rational number is unique except for replacing a final term $1/a_n$ by $1/a_n-1 + 1/1$ when $a_n > 1$. For example $1/3 + 1/5 = 1/3 + 1/4 + 1/1$.

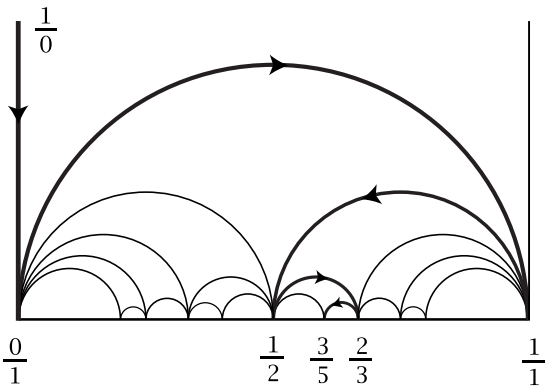
2.2 Infinite Continued Fractions

We have seen that all rational numbers can be expressed as continued fractions $a_0 + 1/a_1 + 1/a_2 + \cdots + 1/a_n$. To complete the picture we will see that irrational numbers can be represented as continued fractions with an infinite number of terms, of the form $a_0 + 1/a_1 + 1/a_2 + 1/a_3 + \cdots$. A simple example is $1/1 + 1/1 + 1/1 + \cdots$. This corresponds to an infinite strip of triangles in the Farey diagram:



Here the vertex labels along the zigzag path after the initial $1/0$ are the ratios of successive terms of the famous Fibonacci sequence $0, 1, 1, 2, 3, 5, 8, 13, 21, \cdots$ where each number after the initial 0 and 1 is the sum of its two predecessors.

The way the zigzag path looks in the upper halfplane Farey diagram is shown in the figure at the right. After the initial vertical edge from $1/0$ to $0/1$ this path consists of an infinite sequence of semicircles, each one shorter than the preceding one and sharing a common endpoint. The left endpoints of the semicircles form an increasing sequence of numbers which have to be approaching a



certain limiting value x. We know x has to be finite since it is certainly less than each of the right-hand endpoints of the semicircles, the convergents $1/1, 2/3, 5/8, \cdots$. Similarly, the right endpoints of the semicircles form a decreasing sequence of numbers approaching a limiting value y greater than each of the left-hand endpoints $0/1, 1/2, 3/5, \cdots$. Obviously $x \le y$. Is it possible that x is not equal to y? If this happened, the infinite sequence of semicircles would be approaching the semicircle from x to y. Above this semicircle there would then be an infinite number of semicircles, all the semicircles in the infinite sequence. Between x and y there would have to be a rational number p/q since there is always a rational number between any two real numbers, so above p/q there would be an infinite number of semicircles, hence an infinite number of triangles in the Farey diagram. But we know that there are only finitely many triangles above any rational number p/q, namely the triangles that appear in the strip for the continued fraction for p/q. This contradiction shows that x has to be equal to y. Thus the sequence of convergents along the edges of the infinite strip of triangles converges to a unique real number x.

This argument works for arbitrary infinite continued fractions, so we have shown the following general result:

Proposition 2.2. *For every infinite continued fraction* $a_0 + 1/a_1 + 1/a_2 + 1/a_3 + \cdots$ *the convergents converge to a unique limit.*

This limit is by definition the value of the infinite continued fraction. This is similar to the situation for infinite decimals, where the value of an infinite decimal is the limit of the values of its finite initial segments.

As a complement to the preceding proposition we have:

Proposition 2.3. *Every irrational number has an expression as an infinite continued fraction, and this continued fraction is unique.*

Proof: In the upper halfplane Farey diagram consider the vertical line L going upward from a given irrational number x on the x-axis. The lower endpoint of L is not a vertex of the Farey diagram since x is irrational. Thus as we move downward along L we cross a sequence of triangles, entering each triangle by crossing its upper edge and leaving the triangle by crossing one of its two lower edges at a point between the two endpoints of this edge. When we exit one triangle, we are entering another triangle so the sequence of triangles and edges we cross must be infinite. The left and right endpoints of the edges in the sequence must be approaching the single point x by the argument we gave earlier, so the edges themselves are approaching x. It cannot happen that an infinite number of successive edges in the sequence have a common vertex since these edges would then be approaching this vertex, which would mean that x was rational. Thus the triangles crossed by the line L form an infinite strip consisting of an infinite sequence of fans with their pivot vertices on alternate sides of the strip. The zigzag path along this strip then gives a continued fraction for x.

For the uniqueness, we have seen that an infinite continued fraction for x corresponds to a zigzag path in the infinite strip of triangles lying above x. This set of triangles is unique so the strip is unique, and there is only one path in this strip that starts at $1/0$ and then does left and right turns alternately, starting with a left turn. The initial turn must be to the left because the first two convergents are a_0 and $a_0 + 1/a_1$, with $a_0 + 1/a_1 > a_0$ since $a_1 > 0$. After the path traverses the initial edge from $1/0$ to $a_0/1$ no subsequent edge of the path can be in the border of the strip since this would entail two successive left turns or two successive right turns. □

From the preceding arguments we can see fairly explicitly why the triangles in the upper halfplane Farey diagram completely cover the upper halfplane, so every point (x, y) with $y > 0$ lies either in the interior of some triangle or on the common edge between two triangles. To see this, consider the vertical line L in the upper halfplane through the given point (x, y). If x is an integer then (x, y) is on one of the vertical edges of the diagram, so we can assume x is not an integer and hence L is not one of the vertical edges of the diagram. The line L will then be contained in the strip of triangles corresponding to the continued fraction for x. This is a finite strip if x is rational and an infinite strip if x is irrational. In either case the point (x, y), being in L, will be in one of the triangles of the strip or on an edge separating two triangles in the strip.

Periodic and Eventually Periodic Continued Fractions

Now that we have an exact correspondence between infinite continued fractions and irrational numbers, there are two natural questions that come to mind: Given an infinite continued fraction, how can one compute its value, and conversely, how can one find the infinite continued fraction for a given irrational number? These questions have very nice answers for a special class of irrational numbers, the numbers whose continued fractions have a pattern that repeats periodically from some point onward, as for example:

$$1/2 + 1/4 + 1/3 + 1/5 + 1/7 + 1/3 + 1/5 + 1/7 + 1/3 + 1/5 + 1/7 + \cdots$$

This includes the case that the whole continued fraction is periodic, for example:

$$1/3 + 1/5 + 1/7 + 1/3 + 1/5 + 1/7 + 1/3 + 1/5 + 1/7 + \cdots$$

A more concise notation is to write a bar over a block of terms in a continued fraction that repeats infinitely often. Thus the two continued fractions above can be written as:

$$1/2 + 1/4 + \overline{1/3 + 1/5 + 1/7} \qquad \text{and} \qquad \overline{1/3 + 1/5 + 1/7}$$

The value of a periodic or eventually periodic continued fraction can be computed by simple algebraic manipulations, as we illustrate now by finding the value of the

continued fraction $\overline{1/1} = 1/1 + 1/1 + 1/1 + \cdots$ involving Fibonacci numbers that we looked at earlier. Suppose we set $x = 1/1 + 1/1 + 1/1 + \cdots$. Then if we take the reciprocals of both sides of this equation we get $1/x = 1 + 1/1 + 1/1 + 1/1 + \cdots$. The right side of this equation is just $1 + x$, so we can easily solve for x:

$$\begin{aligned} \frac{1}{x} &= 1 + x \\ x^2 + x - 1 &= 0 \\ x &= (-1 \pm \sqrt{5})/2 \end{aligned}$$

We know x is positive, so this rules out the negative root and we are left with the final value $x = (-1 + \sqrt{5})/2$. The reciprocal $1/x = 1 + x = (1 + \sqrt{5})/2 \approx 1.618$ is known as the golden ratio because of its many interesting and beautiful properties.

As another example let us find the value of $1/3 + \overline{1/1 + 1/2}$. To do this we first find the value of the periodic part, so we set:

$$x = \overline{1/1 + 1/2} = 1/1 + 1/2 + 1/1 + 1/2 + 1/1 + 1/2 + \cdots$$

Taking reciprocals, we get:

$$\frac{1}{x} = 1 + 1/2 + 1/1 + 1/2 + 1/1 + 1/2 + \cdots$$

Subtracting 1 from both sides gives:

$$\frac{1}{x} - 1 = 1/2 + 1/1 + 1/2 + 1/1 + 1/2 + \cdots$$

The next step will be to take reciprocals of both sides, so before doing this we rewrite the left side as $(1-x)/x$. Then taking reciprocals gives:

$$\begin{aligned} \frac{x}{1-x} &= 2 + 1/1 + 1/2 + 1/1 + 1/2 + \cdots \\ &= 2 + x \end{aligned}$$

Thus we have $x/(1-x) = 2 + x$ which simplifies to the quadratic equation $x^2 + 2x - 2 = 0$ with roots $x = -1 \pm \sqrt{3}$. Again the negative root is discarded and we get $x = -1 + \sqrt{3}$. From this we can determine the value of the original continued fraction $1/3 + \overline{1/1 + 1/2}$ which is $1/(3 + x) = 1/(2 + \sqrt{3}) = 2 - \sqrt{3}$.

Let us consider now the complementary question of how the continued fraction for a given irrational number can be computed. Recall first how the continued fraction $a_0 + 1/a_1 + 1/a_2 + \cdots + 1/a_n$ for a rational number is computed, as in the example of $67/24 = 2 + 1/1 + 1/3 + 1/1 + 1/4$ earlier in the chapter. We first write $67/24 = 2 + 19/24$ which gives $a_0 = 2$, then we write $24/19 = 1 + 5/19$ so $a_1 = 1$, then $19/5 = 3 + 4/5$ so $a_2 = 3$, then $5/4 = 1 + 1/4$ so $a_3 = 1$ and finally $4/1 = 4 + 0$ so $a_4 = 4$. This finishes the process and we have $67/24 = a_0 + 1/a_1 + 1/a_2 + 1/a_3 + 1/a_4 = 2 + 1/1 + 1/3 + 1/1 + 1/4$.

In summary, the steps are:

(1) Write the given number x as $x = a_0 + r_1$ where a_0 is an integer and $0 \le r_1 < 1$.

(2) Write $1/r_1$ as $1/r_1 = a_1 + r_2$ where a_1 is an integer and $0 \le r_2 < 1$.
(3) Write $1/r_2$ as $1/r_2 = a_2 + r_3$ where a_2 is an integer and $0 \le r_3 < 1$.

And so on, repeatedly.

If x is a rational number, the "remainders" r_i are rational numbers with decreasing denominators until we reach a remainder r_n which is zero and the process stops after finitely many steps. We can apply the same procedure if x is irrational, but in this case the equations defining the remainders r_i show that each successive r_i must be irrational and in particular nonzero. Thus the process goes on forever, yielding an infinite continued fraction.

One can see this is the continued fraction for x by the following argument. Suppose the continued fraction for x is $a_0 + 1/a_1 + 1/a_2 + \cdots$. We can write this continued fraction as $a_0 + r_1$ for $r_1 = 1/a_1 + 1/a_2 + \cdots$. This r_1 is a number strictly between 0 and 1 since the convergents for r_1 all lie between 0 and 1 and r_1 lies between any two of its successive convergents. Thus we have $x = a_0 + r_1$ with $0 < r_1 < 1$ so a_0 is the largest integer less than x. Inverting $r_1 = 1/a_1 + 1/a_2 + \cdots$ gives $1/r_1 = a_1 + 1/a_2 + 1/a_3 + \cdots$. The preceding argument can now be repeated with $1/r_1$ in place of x to get $1/r_1 = a_1 + r_2$ with $r_2 = 1/a_2 + 1/a_3 + \cdots$ and $0 < r_2 < 1$. Then one repeats with $1/r_2$ in place of $1/r_1$, and so on.

However, there are a couple subtle points in this argument that are somewhat hidden by the notation. (These subtle points were also lurking in the background in the earlier calculations of the values of the continued fractions $\overline{1/1}$ and $1/3 + \overline{1/1 + 1/2}$.) First, we defined x and r_1 to be the infinite continued fractions $a_0 + 1/a_1 + 1/a_2 + \cdots$ and $1/a_1 + 1/a_2 + \cdots$ and then said that $x = a_0 + r_1$. For finite continued fractions this is true because they are evaluated from right to left, so the last step in evaluating $a_0 + 1/a_1 + 1/a_2 + \cdots + 1/a_n$ is to add a_0 to $1/a_1 + 1/a_2 + \cdots + 1/a_n$. Infinite continued fractions cannot be evaluated from right to left since there is no right end to start the evaluation. Instead they are evaluated from left to right as the limit of the sequence of convergents. The convergents are the values of finite continued fractions, and for these the desired result holds so the convergents for $a_0 + 1/a_1 + 1/a_2 + \cdots$ are obtained by adding a_0 to the convergents for $1/a_1 + 1/a_2 + \cdots$. Adding a fixed number a_0 to each term of a convergent sequence of numbers adds a_0 to the limit of the sequence, so the result holds for infinite continued fractions as well as finite continued fractions.

A similar issue arises when we said that the continued fraction for the reciprocal $1/r_1$ of $r_1 = 1/a_1 + 1/a_2 + \cdots$ is $a_1 + 1/a_2 + \cdots$. Again this is correct for finite continued fractions since they are evaluated from right to left, so if one stops the evaluation of $1/a_1 + 1/a_2 + \cdots + 1/a_n$ before the last step of inverting $a_1 + 1/a_2 + \cdots + 1/a_n$ one has the reciprocal of $1/a_1 + 1/a_2 + \cdots + 1/a_n$. Thus the convergents for the infinite continued fraction $1/a_1 + 1/a_2 + \cdots$ are the reciprocals of the convergents for $a_1 + 1/a_2 + \cdots$ so the limits of the convergents for the two infinite continued

fractions will also be reciprocals of each other.

Here is how the procedure works for computing the continued fraction for $\sqrt{2}$:

(1) $\sqrt{2} = 1 + (\sqrt{2} - 1)$ where $a_0 = 1$ since $\sqrt{2}$ is between 1 and 2. Thus $r_1 = \sqrt{2} - 1$.

(2) $1/r_1 = 1/(\sqrt{2}-1) = 1/(\sqrt{2}-1) \cdot (\sqrt{2}+1)/(\sqrt{2}+1) = \sqrt{2}+1$ which is between 2 and 3 so we have $1/r_1 = 2 + (\sqrt{2} - 1)$ with $a_1 = 2$ and $r_2 = \sqrt{2} - 1$.

Notice that something unexpected has happened: The remainder $r_2 = \sqrt{2}-1$ is exactly the same as the previous remainder r_1. There is then no need to do the calculation of $1/r_2$ since we know it will have to be $\sqrt{2} + 1$. This means that when we continue with step (3), this will be exactly the same as step (2), and the same will be true for all subsequent steps. Thus we can immediately write down the continued fraction for $\sqrt{2}$:

$$\sqrt{2} = 1 + \cfrac{1}{2} + \cfrac{1}{2} + \cfrac{1}{2} + \cdots$$

We can check this calculation by finding the value of the continued fraction in the same way that we did earlier for $\cfrac{1}{1} + \cfrac{1}{1} + \cfrac{1}{1} + \cdots$. It suffices to compute the value of $\cfrac{1}{2} + \cfrac{1}{2} + \cfrac{1}{2} + \cdots$ and then add 1. We set $x = \cfrac{1}{2} + \cfrac{1}{2} + \cfrac{1}{2} + \cdots$ and then take reciprocals to get $1/x = 2 + \cfrac{1}{2} + \cfrac{1}{2} + \cfrac{1}{2} + \cdots = 2 + x$. From $1/x = 2 + x$ we get the quadratic equation $x^2 + 2x - 1 = 0$ with roots $x = -1 \pm \sqrt{2}$. Since x is positive we can discard the negative root. Thus we have $-1 + \sqrt{2} = \cfrac{1}{2} + \cfrac{1}{2} + \cfrac{1}{2} + \cdots$. Adding 1 to both sides of this equation gives the continued fraction for $\sqrt{2}$.

We can compute the continued fraction for $\sqrt{3}$ by the same method as for $\sqrt{2}$, but something slightly different happens:

(1) $\sqrt{3} = 1 + (\sqrt{3} - 1)$ with $a_0 = 1$ since $\sqrt{3}$ is between 1 and 2. Thus $r_1 = \sqrt{3} - 1$.

(2) $1/r_1 = 1/(\sqrt{3}-1) = 1/(\sqrt{3}-1) \cdot (\sqrt{3}+1)/(\sqrt{3}+1) = (\sqrt{3}+1)/2$. This is between 1 and 2 since its numerator $\sqrt{3} + 1$ is between 2 and 3. Thus $a_1 = 1$ and $(\sqrt{3}+1)/2 = 1 + (\sqrt{3}-1)/2$ with $r_2 = (\sqrt{3}-1)/2$.

(3) $1/r_2 = 2/(\sqrt{3}-1) = 2/(\sqrt{3}-1) \cdot (\sqrt{3}+1)/(\sqrt{3}+1) = \sqrt{3} + 1 = 2 + (\sqrt{3} - 1)$ with $a_2 = 2$ and $r_3 = \sqrt{3} - 1$.

Now the remainder $r_3 = \sqrt{3} - 1$ is the same as r_1, so instead of the same step being repeated infinitely often as happened for $\sqrt{2}$, the same two steps will repeat infinitely often. Thus we have computed the continued fraction for $\sqrt{3}$:

$$\sqrt{3} = 1 + \cfrac{1}{1} + \cfrac{1}{2} + \cfrac{1}{1} + \cfrac{1}{2} + \cfrac{1}{1} + \cfrac{1}{2} + \cdots$$

This agrees with our earlier calculation of the value of $\overline{\cfrac{1}{1} + \cfrac{1}{2}}$ to be $-1 + \sqrt{3}$.

It is true in general that for every positive integer n that is not a square, the continued fraction for $\sqrt{n}$ has the form $a_0 + \overline{\cfrac{1}{a_1} + \cfrac{1}{a_2} + \cdots + \cfrac{1}{a_k}}$. The length of the period (the repeating block) can be large even for fairly small values of n, for example:

$$\sqrt{46} = 6 + \overline{\cfrac{1}{1} + \cfrac{1}{3} + \cfrac{1}{1} + \cfrac{1}{1} + \cfrac{1}{2} + \cfrac{1}{6} + \cfrac{1}{2} + \cfrac{1}{1} + \cfrac{1}{1} + \cfrac{1}{3} + \cfrac{1}{1} + \cfrac{1}{12}}$$

This example illustrates two other curious facts about the continued fraction for an irrational number $\sqrt{n}$:

- The last term of the period (12 in the example above) is always twice the first term a_0 (the initial 6).
- If the last term of the period is omitted, the preceding terms in the period form a palindrome, reading the same backwards as forwards.

We will see in Section 4.2 how these two properties follow from certain symmetry properties of the infinite strip of triangles in the Farey diagram associated to the continued fraction for $\sqrt{n}$.

It is natural to ask exactly which irrational numbers have continued fractions that are periodic or eventually periodic. The answer is given by:

Lagrange's Theorem. *The irrational numbers whose continued fractions are eventually periodic are exactly the numbers of the form $a + b\sqrt{n}$ where a and b are rational numbers, $b \neq 0$, and n is a positive integer that is not a square.*

These numbers $a + b\sqrt{n}$ are called *quadratic irrationals* because they are roots of quadratic equations with integer coefficients. The easier half of the theorem is the statement that the value of an eventually periodic infinite continued fraction is always a quadratic irrational. This can be proved by showing that the method we used for finding a quadratic equation satisfied by an eventually periodic continued fraction works in general. Rather than following this purely algebraic approach, however, we will develop a more geometric version of the procedure in the next chapter, so we will wait until then to give the argument that proves this half of Lagrange's Theorem, in Proposition 3.4. The more difficult half of the theorem is the assertion that the continued fraction expansion of every quadratic irrational is eventually periodic. It is not at all apparent from the examples of $\sqrt{2}$ and $\sqrt{3}$ why this should be true in general, but in Chapters 4 and 5 we will develop some theory that will make it clear, with the actual proof being given in Proposition 4.1 and Theorem 5.2. Along the way we will also develop more efficient methods for computing the continued fraction for a quadratic irrational and for computing the value of an eventually periodic infinite continued fraction.

What can be said about the continued fraction expansions of irrational numbers that are not quadratic, such as $\sqrt[3]{2}$, π, or e, the base for natural logarithms? It happens that e has a continued fraction whose terms have a very nice pattern, even though they are not periodic or eventually periodic:

$$e = 2 + \underbrace{1/1 + 1/2 + 1/1}_{} + \underbrace{1/1 + 1/4 + 1/1}_{} + \underbrace{1/1 + 1/6 + 1/1}_{} + \cdots$$

Thus the terms are grouped by threes with successive even numbers as middle denominators. Even simpler are the continued fractions for certain numbers built from

e that have arithmetic progressions for their denominators:

$$\frac{e-1}{e+1} = 1/2 + 1/6 + 1/10 + 1/14 + \cdots$$
$$\frac{e^2-1}{e^2+1} = 1/1 + 1/3 + 1/5 + 1/7 + \cdots$$

The continued fractions for e and $(e-1)/(e+1)$ were discovered by Euler in 1737 while the formula for $(e^2-1)/(e^2+1)$ was found by Lambert in 1766 as a special case of a slightly more complicated formula for $(e^x-1)/(e^x+1)$.

For $\sqrt[3]{2}$ and π, however, the continued fractions have no known pattern. For π the continued fraction begins:

$$\pi = 3 + 1/7 + 1/15 + 1/1 + 1/292 + \cdots$$

Here the first four convergents are 3, $22/7$, $333/106$, and $355/113$. We recognize $22/7$ as the familiar approximation $3\frac{1}{7}$ to π. The convergent $355/113$ is a particularly good approximation to π since its decimal expansion begins 3.14159282 whereas $\pi = 3.14159265\cdots$. It is no accident that the convergent $355/113$ obtained by truncating the continued fraction just before the 292 term gives a good approximation to π since it is a general fact that a convergent immediately preceding a large term in the continued fraction always gives an especially good approximation. This is because the next edge in the zigzag path will be rather small when viewed in the upper halfplane Farey diagram since it is the lower edge of a fan with a large number of triangles, and the value of the continued fraction lies somewhere between the two ends of this small edge.

It is easy to calculate an initial string of terms in the continued fraction for π using a reasonably capable calculator if one knows the decimal expansion of π with enough accuracy. One just repeats the two steps of subtracting the integer part and inverting, preferably using the calculator's $1/x$ button. For example, starting with 3.1415926535 one can get the initial segment $3 + 1/7 + 1/15 + 1/1 + 1/292$ this way. People who like computational challenges have used computers to find large numbers of terms of the continued fraction for π, more than a billion terms in fact.

There are nice continued fractions for π if one allows numerators larger than 1, as in the following formula discovered by Euler:

$$\pi = 3 + 1^2/6 + 3^2/6 + 5^2/6 + 7^2/6 + \cdots$$

However, it is the continued fractions with numerator 1 that have the best properties, so we will not consider the more general sort in this book.

Convergents as Rational Approximations

Let us explore in a little more detail how the convergents in the continued fraction for an irrational number x give good rational approximations to x.

As an example, consider the case $x = \sqrt{2} = 1 + \overline{1/2}$. It is a little easier to compute the convergents for $2 + \overline{1/2} = 1 + \sqrt{2}$ and then subtract 1 from each of these. For $2 + 1/2 + 1/2 + 1/2 + \cdots$ the convergents are:

$$\frac{2}{1} \quad \frac{5}{2} \quad \frac{12}{5} \quad \frac{29}{12} \quad \frac{70}{29} \quad \frac{169}{70} \quad \frac{408}{169} \quad \frac{985}{408} \quad \cdots$$

The sequence of numbers $1, 2, 5, 12, 29, 70, 169, \cdots$ generating these fractions can be constructed in a way somewhat analogous to the Fibonacci sequence, except that each number is *twice* the preceding number plus the number before that. This is because each convergent is obtained from the previous one by inverting the fraction and adding 2, and therefore the next convergent after a/b is $2 + b/a = \frac{2a+b}{a}$.

Subtracting 1 from each of the fractions in the list displayed above, we obtain the convergents for $\sqrt{2}$ as shown at the right. Notice that once an initial string of digits in the decimal expansions of the convergents occurs twice in succession, then this string is unchanged from then on. This is because for any two successive convergents, all subsequent convergents lie between these two since the convergents occur along a zigzag path in the Farey diagram. This is true generally for all infinite continued fractions.

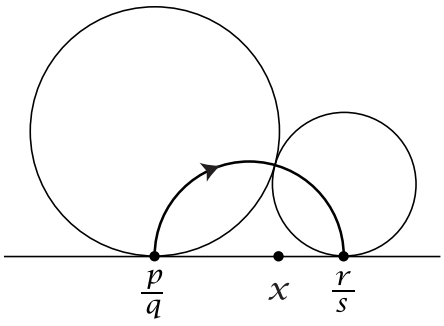
$$\begin{aligned}
\sqrt{2} &= 1.41421356\cdots \\
1/1 &= 1.00000000\cdots \\
3/2 &= 1.50000000\cdots \\
7/5 &= 1.40000000\cdots \\
17/12 &= 1.41666666\cdots \\
41/29 &= 1.41379310\cdots \\
99/70 &= 1.41428571\cdots \\
239/169 &= 1.41420118\cdots \\
577/408 &= 1.41421568\cdots
\end{aligned}$$

Information about how well an irrational number is approximated by the convergents in its continued fraction can be deduced from the geometry of Ford circles, which were introduced at the end of Chapter 1. Here is one general statement that can be made:

- Each convergent p/q in the continued fraction for an irrational number x is within $1/q^2$ of x, so $|x - p/q| < 1/q^2$.

For example, if a convergent has a denominator of 100 or greater then the convergent approximates x to within $1/10000$. Thus the approximation $239/169$ to $\sqrt{2}$ must be accurate to four decimal places.

To justify the $1/q^2$ estimate, suppose the convergent p/q is connected to the next convergent r/s by an edge of the zigzag path. We then have $s \geq q$ so the Ford circle $C_{p/q}$ at p/q, which has diameter $1/q^2$, is at least as large as the Ford circle $C_{r/s}$. The number x lies between p/q and r/s, so its distance to p/q is less than twice the radius $1/2q^2$ of $C_{p/q}$. In other words this distance is less than $1/q^2$, as claimed.



For many convergents the estimate $1/q^2$ can be improved by a factor of 2:

- At least one convergent p/q out of every two successive convergents to x is within $1/2q^2$ of x.

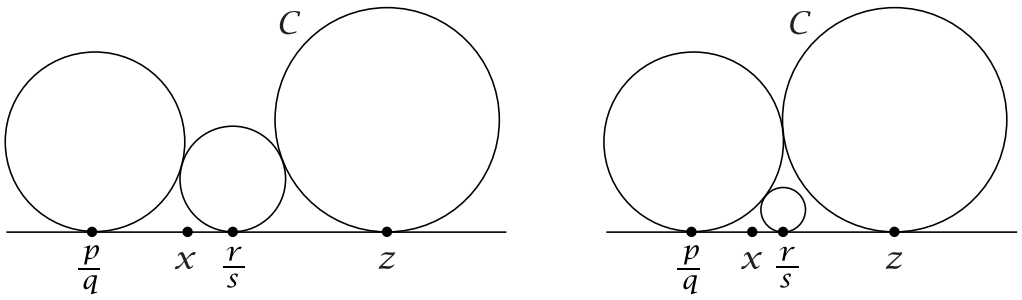
We can see this from the previous figure. For x to be within $1/2q^2$ of p/q means that x is a point in the projection of the interior of $C_{p/q}$ to the x-axis, and similarly for r/s and $C_{r/s}$. Since x lies between p/q and r/s, it must be in at least one of these two projections, except possibly in the case that $q = s$ (which can only happen when q and s are 1) when the midpoint of the interval between p/q and r/s is not in the projection of the interior of either $C_{p/q}$ or $C_{r/s}$. But this midpoint is a rational number so it cannot be x.

Next we have a very strong optimality statement about the convergents to an irrational number x:

- If p/q is a convergent in the continued fraction for x then no rational number with denominator less than or equal to q is closer to x than p/q is.

To see why this is true consider two consecutive convergents p/q and r/s as before, and let t/u be any rational number with $u \le q$, so $C_{t/u}$ is at least as large as $C_{p/q}$. The circle $C_{t/u}$ is either disjoint from or tangent to $C_{p/q}$ and $C_{r/s}$. Clearly t/u cannot be between p/q and r/s since there is no room to fit the large circle $C_{t/u}$ there. If t/u is on the opposite side of p/q from r/s then t/u would be farther from x than p/q is, so t/u would not be a closer approximation to x than p/q is.

The remaining possibility is that t/u is on the opposite side of r/s from p/q. Let C be any circle in the upper halfplane with the same geometric properties as $C_{t/u}$, namely, C is tangent to the x-axis at a point z on the opposite side of r/s from p/q, C is either disjoint from or tangent to $C_{p/q}$ and $C_{r/s}$, and C is at least as large as $C_{p/q}$.

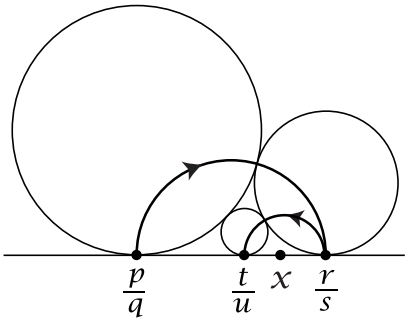


We wish to show that z is farther from x than p/q is. Sliding C along the x-axis farther from r/s moves z farther from x so we may assume C touches either $C_{p/q}$ or $C_{r/s}$. Then increasing the size of C while keeping it tangent to $C_{p/q}$ or $C_{r/s}$ also moves z farther from x so we may assume C has the same size as $C_{p/q}$. It is then evident that z is farther from x than p/q is, finishing the argument.

The last fact we will deduce from the diagram of Ford circles is that the convergents converge monotonically. We know that the convergents to the left of x are getting steadily closer to x, and the same is true for the convergents to the right. And in fact:

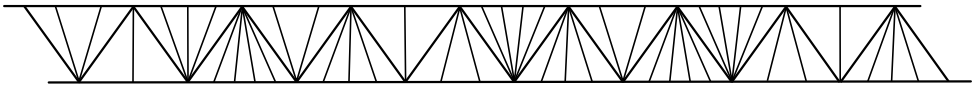
- Each convergent in the continued fraction for x is closer to x than the previous convergent.

To verify this, suppose that p/q and r/s are two consecutive convergents to x, so we wish to show that x is closer to r/s than to p/q. Consider the next convergent t/u after r/s. The circle $C_{r/s}$ is tangent to both $C_{p/q}$ and $C_{t/u}$, while $C_{t/u}$ is either tangent to $C_{p/q}$ or $C_{t/u}$ is one of the other Ford circles tangent to $C_{r/s}$ farther from $C_{p/q}$. The point x lies between r/s and t/u so to show that x is closer to r/s than to p/q it will suffice to consider just the case that $C_{t/u}$ is tangent to $C_{p/q}$. Then t/u is the mediant of p/q and r/s so as we saw in Section 1.2, t/u is closer to r/s than to p/q since $s \geq q$, or possibly t/u is equidistant from r/s and p/q if $s = q$. In either case, since x lies between t/u and r/s, it must then be closer to r/s than to p/q, which is what we wanted to show.



Doubly Infinite Strips

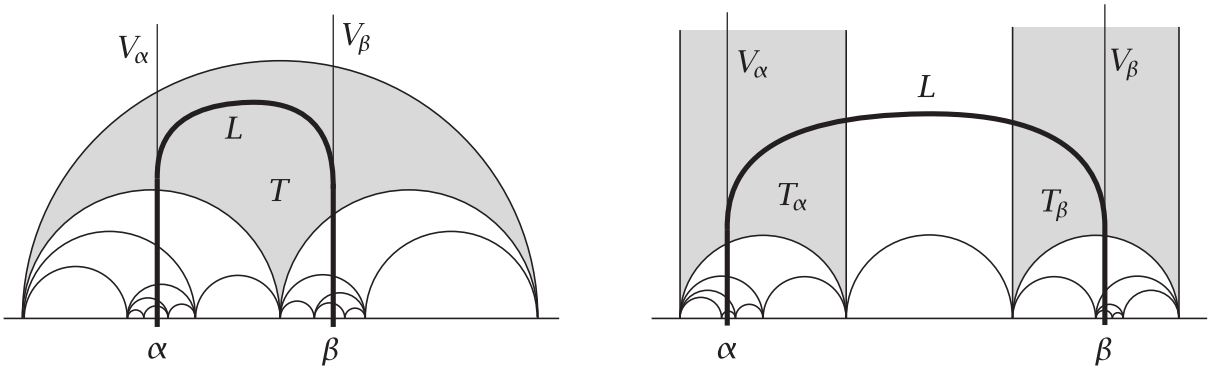
We have been considering strips of triangles in the Farey diagram consisting of fans, each fan having a finite number of triangles and each fan intersecting the next along an edge of a zigzag path in the strip. For finite continued fractions the strip has finitely many fans, while for infinite continued fractions the strip has an infinite sequence of fans at one end. In later chapters we will often be considering strips that extend infinitely far at both ends. We can think of these strips as being "doubly infinite" since they are infinite in both directions.



To see how such a doubly infinite strip lies in the upper halfplane model of the Farey diagram, let L be a line running down the middle of the strip from end to end. Viewing L as a path in the upper halfplane model of the Farey diagram, L cannot cross only vertical edges, the edges with one end at $1/0$, otherwise the strip would consist of a single infinite fan, which is not allowed as an infinite strip. Thus L must cross some semicircular edges. As we move along L crossing such a semicircular edge in the downward direction into the adjacent triangle, the next edge that L crosses will be one of the other two shorter semicircular edges of this triangle, moving downward again. All subsequent crossings will then be downward as well. The semicircles crossed are becoming smaller and smaller with diameters approaching zero, as we saw in our

initial discussion of infinite continued fractions, and there is a unique limiting point α on the x-axis for this end of the strip of triangles. This is the unique point that lies between the two endpoints of each semicircular edge crossed by L on its downward path.

Consider the vertical line V_α going upward from α. Near its lower end V_α will pass through triangles of the strip. If the whole line V_α does not stay entirely within the strip as we move upward, it will eventually leave the strip by crossing the upper semicircular edge of a triangle T of the strip as in the figure on the left below.



In this case the line L, which passes through the same upward sequence of triangles as V_α until reaching T, must exit T by turning and crossing the other smaller semicircular edge of T in the downward direction. After crossing this edge, L will then continue downward forever, passing through all the triangles of the other end of the strip and limiting on an irrational number β. The vertical line V_β going upward from β will pass through the same set of triangles until reaching the triangle T where it will also exit the strip by crossing the upper edge of T. We can then deform L so that it consists of the parts of V_α and V_β below T joined by a bending arc within T. Notice that the vertex $1/0$ is not a vertex of the strip in this case.

The other possibility is that V_α stays in the strip forever as we move upward, so eventually it lies in a triangle T_α of the strip having $1/0$ as a vertex as in the figure on the right above. One end of the line L runs parallel to V_α until it reaches T_α, then it turns right or left to cross a finite number of other triangles having $1/0$ as a vertex before turning downward to cross the lower edge of one of these triangles T_β. After this it will travel monotonically downward, limiting on an irrational number β in the x-axis. We can deform L to consist of parts of V_α and the vertical line V_β through β, joined by an arc crossing from T_α to T_β.

One conclusion we can draw from this analysis of the infinite strip is that its endpoints α and β cannot be the same number. This can be seen from the two figures above where in the first figure α and β lie below the two different lower edges of the triangle T, and in the second figure α and β lie below the two different triangles T_α and T_β with a vertex at $1/0$.

Another consequence is that the labels x/y on the vertices along the infinite strip must have denominators y approaching infinity at the ends of the strip and numer-

ators x approaching either $+\infty$ or $-\infty$ depending on the sign of the endpoint α or β being approached. This is because the labels are given by repeated applications of the mediant rule as we move vertically down either end of L toward α or β so $|x|$ and y always increase as each new triangle is added to the strip. (Near the ends of the strip the labels x/y are approaching α or β so neither x nor y is 0.)

We can also deduce that for each pair of distinct irrationals α and β there is a unique infinite strip in the Farey diagram whose ends converge to α and β. This is because α and β determine the vertical lines V_α and V_β in the figures, and these determine the triangles T or T_α and T_β since in the case that α and β lie in the same interval in the x-axis between consecutive integers, T is the smallest triangle of the Farey diagram whose projection to the x-axis contains both α and β, while in the case that α and β lie in different intervals between consecutive integers, the triangles T_α and T_β are the triangles with vertex $1/0$ that project to these two intervals.

A nice way to construct an infinite strip joining any two irrationals α and β is to take all the triangles in the Farey diagram that meet the semicircle in the upper halfplane with endpoints α and β. This semicircle can cross an edge of the Farey diagram only once since if two semicircles in the upper halfplane with endpoints on the x-axis intersect in more than one point, they must coincide. Nor can two semicircles with endpoints on the x-axis be tangent unless the point of tangency is one of the endpoints, but this does not happen here since α and β are irrational while the endpoints of edges of the Farey diagram are rational. From these observations we see that if the semicircle from α to β intersects a triangle of the Farey diagram, then it crosses this triangle from one edge to another edge. The semicircle cannot cross an infinite number of triangles having a common vertex, otherwise the semicircle would contain points arbitrarily close to the common vertex, which is impossible since the common vertex cannot be either of the irrational numbers α and β. Thus the union of all the triangles crossed by the semicircle is an infinite strip.

We have seen that an infinite strip is uniquely determined by its endpoints, so this implies that the semicircle from α to β crosses exactly the same triangles as the line we constructed earlier consisting of two vertical segments joined at the top by a 180 degree bend. This may seem odd at first glance, but what it means is that the height of the vertical segments cannot be too large compared to the distance between them.

The construction of a strip connecting two irrational numbers α and β via the semicircle with endpoints α and β works equally well when α or β is rational, but in this case the strip has only a finite number of triangles at a rational end. A very special case is when α and β are the endpoints of an edge of the Farey diagram, when the strip degenerates to just this edge.

The doubly infinite strips we will be most interested in are the ones that are periodic along their whole length. As we will see, the irrational numbers α and β

at the ends of such a strip will be the two roots of a quadratic equation with integer coefficients.

Exercises

1. Compute the values of the following infinite continued fractions:

(a) $\overline{1/4}$

(b) $\overline{1/n}$ for an arbitrary positive integer n

(c) $\overline{1/2 + 1/3}$ and $1/1 + \overline{1/2 + 1/3}$

(d) $\overline{1/1 + 1/2 + 1/1 + 1/6}$ and $1/1 + 1/4 + \overline{1/1 + 1/2 + 1/1 + 1/6}$

(e) $\overline{1/2 + 1/3 + 1/5}$

2. (a) Compute the continued fractions for $\sqrt{5}$ and $\sqrt{23}$.

(b) Using the continued fraction for $\sqrt{5}$, find the first convergent which gives a rational approximation to $\sqrt{5}$ accurate to four decimal places.

3. Compute the continued fractions for $\sqrt{n^2+1}$ and $\sqrt{n^2+n}$ where n is an arbitrary positive integer.

2.3 Linear Diophantine Equations

As an application of continued fractions let us see how they can be used to solve linear Diophantine equations $ax + by = n$, where a, b, and n are integers and the solutions are required to be integers as well. We can assume a, b, and n are nonzero, otherwise the equation is rather trivial. Changing the signs of x or y if necessary, we can rewrite the equation in the form $ax - by = n$ where a and b are both positive. Solving this equation means finding multiples of a and b that differ by n.

If a and b have greatest common divisor $d > 1$, then since d divides a and b it must divide $ax - by$, so d must divide n if the equation $ax - by = n$ is to have any solutions at all. If d does divide n we can divide both sides of the equation by d to get a new equation having the same solutions but with the new coefficients a and b coprime. For example, the equation $6x - 15y = 21$ reduces in this way to the equation $2x - 5y = 7$. Thus we can assume from now on that a and b are coprime. We will show that solutions always exist in this case, in fact infinitely many solutions, and we will see how to compute them.

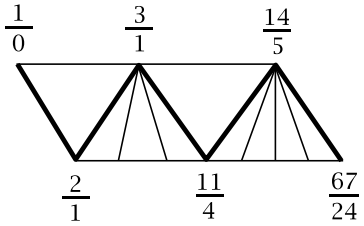
To find a solution of $ax - by = n$ it suffices to do the case $n = 1$ since if we have a solution of $ax - by = 1$, we can multiply x and y by n to get a solution of $ax - by = n$. For example, for the equation $2x - 5y = 1$ the smallest multiple of 2

that is one greater than a multiple of 5 is $2\cdot3 > 5\cdot1$, so a solution of $2x - 5y = 1$ is $(x, y) = (3, 1)$. A solution of $2x - 5y = 7$ is then $(x, y) = (21, 7)$.

The idea for solving $ax - by = 1$ when a and b are coprime is to utilize the criterion from Proposition 1.1 that the Farey diagram contains an edge joining a/b to c/d exactly when $ad - bc = \pm1$. In the case that $ad - bc = +1$ a solution of $ax - by = 1$ is then $(x, y) = (d, c)$, and when $ad - bc = -1$ a solution of $ax - by = 1$ is $(x, y) = (-d, -c)$.

For a given coprime pair of positive integers a and b we can compute the continued fraction for a/b and the corresponding strip of triangles in the Farey diagram from $1/0$ to a/b. The last edge in the zigzag path in this strip connects a fraction c/d to a/b, so we have $ad - bc = \pm1$. Since c/d is the next to last vertex along the zigzag path, the continued fraction for c/d is obtained from the continued fraction for a/b by omitting the last term. From this truncated continued fraction we can then compute c/d and hence a solution of $ax - by = 1$.

As an example, let us solve $67x - 24y = 1$. The continued fraction for $67/24$ is $2 + 1/1 + 1/3 + 1/1 + 1/4$. Omitting the last term gives $2 + 1/1 + 1/3 + 1/1$ which equals $14/5$. Thus we have $67\cdot5 - 24\cdot14 = \pm1$. The sign can be determined by observing that $67/24$ lies to the right of $14/5$ in the circular Farey diagram so $67/24 < 14/5$, hence $67\cdot5 < 24\cdot14$ and therefore $67\cdot5 - 24\cdot14 = -1$. Thus we obtain the solution $(x, y) = (-5, -14)$.



The fact that $67/24$ lies to the right of $14/5$ in the Farey diagram is a consequence of the strip of triangles having an even number of fans. With an odd number of fans the situation would be reversed. The number of fans is the number of terms in the continued fraction after the initial integer, so we see that it is not really necessary to draw the strip of triangles to figure out the correct sign.

Another way to determine the sign without using the diagram is by computing $67\cdot5 - 24\cdot14$ mod 10 to see whether we get $+1$ or -1 mod 10. Computing mod 10 means ignoring all but the last digit, so we get $7\cdot5 - 4\cdot4 = 19 \equiv -1$ mod 10 and hence the sign is negative.

We can get other solutions to $67x - 24y = 1$ by using other edges of the Farey diagram with endpoint $67/24$ instead of the edge from $14/5$. For example we could use the edge to $67/24$ in the lower border of the strip of triangles. By the mediant rule this edge joins $53/19$ to $67/24$, so we have $67\cdot19 - 24\cdot53 = \pm1$ and this time the plus sign is correct, giving the solution $(x, y) = (19, 53)$. All the other edges connected to $67/24$ are obtained by repeatedly "adding" $67/24$ either to $14/5$ for edges above $67/24$, or to $53/19$ for edges below $67/24$. In the former case these are the edges leading to the fractions $\frac{14+67k}{5+24k}$ for positive integers k, and in the latter case they are the edges to $\frac{53+67k}{19+24k}$ for positive integers k. Notice that if we let k be negative in one of these formulas, we get the fractions given by the other formula. For

example in $^{53+67k}/_{19+24k}$ the values $k = -1, -2, \cdots$ give the fractions $^{-14}/_{-5} = {}^{14}/_{5}$, $^{-81}/_{-29} = {}^{81}/_{29}, \cdots$ which are the values of $^{14+67k}/_{5+24k}$ for $k = 0, 1, \cdots$. This means that the general solution of $67x - 24y = 1$ is $(x, y) = (19 + 24k, 53 + 67k)$ for arbitrary integers k. Alternatively, we could write the general solution as $(x, y) = (-5 - 24k, -14 - 67k)$ or as $(x, y) = (-5 + 24k, -14 + 67k)$ since k can be replaced by $-k$.

This example illustrates a general fact:

Proposition 2.4. *For coprime integers a and b, if one solution of $ax - by = n$ is $(x, y) = (p, q)$ then the general solution is $(x, y) = (p + bk, q + ak)$ for k an arbitrary integer.*

Here we do not need to assume a and b are positive, so by changing the sign of b we can write the equation as $ax + by = n$ with general solution $(p - bk, q + ak)$, or alternatively as $(p + bk, q - ak)$.

Proof: One solution $(x, y) = (p, q)$ of $ax - by = n$ is given. For an arbitrary solution (x, y) we look at the difference $(x - p, y - q)$ which we denote as (x_0, y_0). This satisfies $ax_0 - by_0 = 0$, or in other words, $ax_0 = by_0$. Since a and b are coprime, the equation $ax_0 = by_0$ must have the form $a(bk) = b(ak)$ for some integer k, with $x_0 = bk$ and $y_0 = ak$. Hence every solution of $ax - by = n$ has the form $(x, y) = (p + x_0, q + y_0) = (p + bk, q + ak)$. It is easy to check that these formulas for x and y give solutions to $ax - by = n$ for all values of k. □

The Diophantine equation $ax - by = n$ can be interpreted as a congruence condition by rewriting it as $ax - n = by$ which implies that $ax \equiv n \mod b$. Conversely, if $ax \equiv n \mod b$ then this means that $ax - n = by$ for some integer y, so $ax - by = n$. Thus a solution (x, y) of $ax - by = n$ gives a solution x of $ax \equiv n \mod b$, and a solution x of $ax \equiv n \mod b$ gives a solution (x, y) of $ax - by = n$ since this equation allows y to be computed from a, b, n, and x if b is nonzero.

The special case $ax - by = 1$ is equivalent to $ax \equiv 1 \mod b$ which says that x is a multiplicative inverse to $a \mod b$. We know that $ax - by = 1$ has a solution exactly when a and b are coprime, so this means that a has a multiplicative inverse mod b exactly when a is coprime to b. For example the congruence classes mod 15 that are coprime to 15 are $1, 2, 4, 7, 8, 11, 13, 14$ and we can find multiplicative inverses for each of these by observing that the products $1 \cdot 1$, $2 \cdot 8$, $4 \cdot 4$, $7 \cdot 13$, $11 \cdot 11$, and $14 \cdot 14$ are each congruent to 1 mod 15. Thus the numbers $1, 4, 11$, and 14 are their own inverses mod 15 while the other inverses occur in pairs, the pair $2, 8$ and the pair $7, 13$. We could shorten these calculations by noting that if $ax \equiv 1 \mod b$ then $(-a)(-x) \equiv 1 \mod b$, so for example $2 \cdot 8 \equiv 1 \mod 15$ implies $(-2)(-8) \equiv 1 \mod 15$ or in other words $13 \cdot 7 \equiv 1 \mod 15$. Similarly $4 \cdot 4 \equiv 1 \mod 15$ implies $11 \cdot 11 \equiv 1 \mod 15$.

The function which assigns to each positive integer n the number of congruence classes mod n of numbers coprime to n is called the ***Euler phi function*** $\varphi(n)$. Thus in the preceding example of multiplicative inverses mod 15 we have $\varphi(15) = 8$ from the eight numbers $1, 2, 4, 7, 8, 11, 13, 14$. Later in this section we will obtain a formula for $\varphi(n)$.

Linear Diophantine equations with more than two variables can be solved by reduction to the case of two variables. Consider for example a three-variable equation $ax + by + cz = n$. Any number that divides all three coefficients a, b, c must also divide n if a solution is to exist, and in this case we can simplify the equation by dividing it by the greatest common divisor of a, b, and c, so we may as well assume that the greatest common divisor of a, b, and c is 1.

As an example that is typical of the general case for three variables, consider the equation $6x + 10y + 15z = 7$. Here the greatest common divisor of 6, 10, and 15 is 1, although when taken two at a time they have larger common divisors: 2 for 6 and 10, 3 for 6 and 15, and 5 for 10 and 15.

The idea for solving $6x+10y+15z = 7$ is to write it first as $2(3x+5y)+15z = 7$ and then to rewrite this as the two equations $3x + 5y = w$ and $2w + 15z = 7$. The first equation $3x + 5y = w$ has solutions for every w since 3 and 5 are coprime, and we can find the solutions by first solving $3x + 5y = 1$ and then multiplying these solutions by w. Since the coefficients 3 and 5 are so small, we can find a solution of $3x + 5y = 1$ by inspection rather than computing continued fractions, and we see that $(x, y) = (2, -1)$ is a solution. Then $(x, y) = (2w, -w)$ is a solution of $3x + 5y = w$. Applying Proposition 2.4, the general solution of $3x + 5y = w$ can therefore be written as $(x, y) = (2w + 5s, -w - 3s)$ for s an arbitrary integer.

Next we solve $2w + 15z = 7$. A solution of $2w + 15z = 1$ is $(w, z) = (8, -1)$ so a solution of $2w + 15z = 7$ is $(w, z) = (56, -7)$. The general solution of $2w + 15z = 7$ is then $(w, z) = (56 + 15t, -7 - 2t)$ for arbitrary integers t. Alternatively, we could notice that $2w + 15z = 7$ has the simpler solution $(w, z) = (-4, 1)$, obtained either by inspection or by letting $t = -4$ in the pair $(56 + 15t, -7 - 2t)$. Hence the general solution of $2w + 15z = 7$ can also be written as $(w, z) = (-4 + 15t, 1 - 2t)$.

Using $(w, z) = (-4 + 15t, 1 - 2t)$ we now substitute $w = -4 + 15t$ into the earlier formula $(x, y) = (2w + 5s, -w - 3s)$ to obtain the final answer in terms of the arbitrary intgers s and t:

$$\begin{aligned}(x, y, z) &= (2(-4 + 15t) + 5s, -(-4 + 15t) - 3s, 1 - 2t) \\ &= (-8 + 5s + 30t, 4 - 3s - 15t, 1 - 2t)\end{aligned}$$

In the spirit of Proposition 2.4 we can say that a particular solution of $6x+10y+15z = 7$ is $(-8, 4, 1)$, obtained by setting $s = t = 0$, and the general solution is obtained by adding this particular solution to $(5s + 30t, -3s - 15t, -2t)$ which is the general solution of the associated equation $6x + 10y + 15z = 0$ with right side zero.

The situation for equations with more variables is similar to what happened in this example, with an equation in n variables breaking up into $n-1$ equations in two variables. Each of these has solutions depending on an integer parameter, so the solutions of the n-variable equation depend on $n-1$ independent parameters.

We can apply what we have learned about linear Diophantine equations to derive a general fact about congruences often referred to as the ***Chinese Remainder Theorem*** since it was used in ancient Chinese manuscripts to solve mathematical puzzles of a certain type.

Proposition 2.5. *A collection of congruence conditions*

$$\begin{aligned} x &\equiv a_1 \mod m_1 \\ x &\equiv a_2 \mod m_2 \\ &\cdots \\ x &\equiv a_k \mod m_k \end{aligned}$$

always has a simultaneous solution provided that no two of the moduli m_i have a common divisor greater than 1, and in this case the collection of all solutions forms a single congruence class modulo the product $m_1 \cdots m_k$.

Without the hypothesis that the various moduli m_i are coprime there may not be a common solution. For example the two congruences $x \equiv 5 \mod 6$ and $x \equiv 7 \mod 15$ have no common solution since the first congruence implies $x \equiv 2 \mod 3$ while the second congruence implies $x \equiv 1 \mod 3$. Here we are using the following general fact about congruences that will be used often:

If a congruence $a \equiv b$ holds mod n then it holds mod d for each divisor d of n.

This is true because if n divides $a-b$ then so does d for each divisor d of n.

Proof of Proposition 2.5: Let us first prove the existence of a common solution x when there are just two congruences $x \equiv a_1 \mod m_1$ and $x \equiv a_2 \mod m_2$. In this case the desired number x will have the form $x = a_1 + x_1 m_1 = a_2 + x_2 m_2$ for some pair of yet-to-be-determined numbers x_1 and x_2. We can rewrite the equation $a_1 + x_1 m_1 = a_2 + x_2 m_2$ as $m_1 x_1 - m_2 x_2 = a_2 - a_1$. We know that this equation has a solution (x_1, x_2) with integers x_1 and x_2 whenever m_1 and m_2 are coprime. This is obtained by first finding integers n_1 and n_2 such that $m_1 n_1 + m_2 n_2 = 1$ and then multiplying this equation by $a_2 - a_1$ to get $(a_2-a_1)m_1 n_1 + (a_2-a_1)m_2 n_2 = a_2 - a_1$. Then in the equation $m_1 x_1 - m_2 x_2 = a_2 - a_1$ we may choose $x_1 = (a_2 - a_1)n_1$ and $x_2 = (a_2 - a_1)(-n_2)$. Thus we have:

$$\begin{aligned} x &= a_1 + x_1 m_1 \\ &= a_1 + m_1 (a_2 - a_1) n_1 \\ &= a_1 (1 - m_1 n_1) + a_2 m_1 n_1 \\ &= a_1 m_2 n_2 + a_2 m_1 n_1 \quad \text{since} \quad 1 - m_1 n_1 = m_2 n_2 \end{aligned}$$

Summarizing, we have the solution $x = a_1 m_2 n_2 + a_2 m_1 n_1$ where n_1 and n_2 satisfy $m_1 n_1 + m_2 n_2 = 1$.

For a system of more than two congruences we may suppose by induction on the number of congruences that we have a number $x = a$ satisfying all but the last congruence $x \equiv a_k \bmod m_k$. From the preceding paragraph we know that a number x exists satisfying the two congruences $x \equiv a \bmod m_1 \cdots m_{k-1}$ and $x \equiv a_k \bmod m_k$ since $m_1 \cdots m_{k-1}$ and m_k are coprime. This gives the desired solution to all k congruences $x \equiv a_i \bmod m_i$ since $x \equiv a \bmod m_1 \cdots m_{k-1}$ implies $x \equiv a \bmod m_i$ for each $i < k$, and $a \equiv a_i \bmod m_i$ for each $i < k$ by the inductive hypothesis.

Now we show that all the different solutions of the given set of congruences form a single congruence class mod $m_1 \cdots m_k$. If x and y are two solutions then the difference $x - y$ is congruent to 0 mod each of the numbers $m_1, \cdots, m_k$, which means that it is divisible by each m_i and hence by their product since they have no common factors. Thus $x \equiv y \bmod m_1 \cdots m_k$, which shows that all the solutions lie in a single congruence class mod $m_1 \cdots m_k$. Moreover every number in this congruence class is a solution since if x is one solution and $y \equiv x \bmod m_1 \cdots m_k$ then $y \equiv x \bmod m_i$ for each i, so $x \equiv a_i \bmod m_i$ implies $y \equiv a_i \bmod m_i$. □

As an illustration of the method in this proof let us find all numbers that are congruent to 7 mod 9 and to 8 mod 11. First we find a solution of $9n_1 + 11n_2 = 1$ by the earlier methods. One such solution is $(n_1, n_2) = (5, -4)$. The formula $x = a_1 m_2 n_2 + a_2 m_1 n_1$ then gives $x = -7 \cdot 11 \cdot 4 + 8 \cdot 9 \cdot 5 = -308 + 360 = 52$. We are free to change this by adding any multiple of $9 \cdot 11$, so the general solution is $52 + 99t$ for arbitrary integers t. If we were to modify the problem by adding a third congruence condition such as $x \equiv 4 \bmod 7$ then we would just be solving the two congruences $x \equiv 52 \bmod 99$ and $x \equiv 4 \bmod 7$ by the same method.

There is a geometric picture that gives a way of visualizing what the Chinese Remainder Theorem is saying. Consider the case of two simultaneous congruences $x \equiv a \bmod m$ and $x \equiv b \bmod n$ where m and n are coprime. We can then label the mn unit squares in an $m \times n$ rectangle by the numbers $1, 2, 3, \cdots$ starting in the lower left corner and continuing upward to the right at a 45 degree angle as shown in the following figure for the case of a 9×4 rectangle:

28	20	12	4	32	24	16	8	36
19	11	3	31	23	15	7	35	27
10	2	30	22	14	6	34	26	18
1	29	21	13	5	33	25	17	9

Whenever we run over the top edge, we jump back to the bottom in order to continue, and when we reach the right edge, we jump back to the left edge. This amounts to taking congruence classes mod m horizontally and mod n vertically. What the Chinese Remainder Theorem says is that when m and n are coprime, each unit square in the $m \times n$ rectangle is labeled exactly once by a number from 1 to mn. (Without the coprimeness some squares would have no labels while others would have multiple labels.) The figure thus illustrates that specifying a congruence class mod mn is equivalent to specifying a pair of congruence classes mod m and mod n via the projections onto the two axes.

For the case of three simultaneous congruences there is an analogous picture with a three-dimensional rectangular box partitioned into unit cubes. More generally, for k congruences one would be dealing with a k-dimensional box.

A common situation for applying the Chinese Remainder Theorem is to start with a number n factored as $n = p_1^{r_1} \cdots p_k^{r_k}$ for distinct primes $p_1, \cdots, p_k$, so that a congruence $x \equiv a \mod n$ is equivalent to a set of k congruences $x \equiv a_i \mod p_i^{r_i}$. If we add the condition that each a_i is not divisible by the corresponding prime p_i then a simultaneous solution $x = a$ for all k congruences must be coprime to n since $a \equiv a_i \mod p_i^{r_i}$ implies $a \equiv a_i \mod p_i$ and we assume a_i is nonzero mod p_i so a is also nonzero mod p_i. Conversely, if a is coprime to n and satisfies a set of congruences $a \equiv a_i \mod p_i^{r_i}$ and hence $a \equiv a_i \mod p_i$, then a_i must be nonzero mod p_i since a is. Thus congruence classes mod n of numbers a coprime to n are equivalent to congruence classes mod $p_i^{r_i}$ of numbers a_i coprime to p_i, one for each i.

In the geometric picture for the case $k = 2$ with a rectangular array of unit squares, if we require a_1 to be coprime to p_1 then we are omitting the numbers in certain vertical columns of squares, the columns whose horizontal coordinate is a multiple of p_1. Similarly, when we require a_2 to be coprime to p_2 we omit the numbers in the horizontal rows whose vertical coordinate is a multiple of p_2. The numbers in the boxes that are not omitted are then the numbers coprime to $n = p_1^{r_1} p_2^{r_2}$. Here is the picture for the case $n = 3^2 \cdot 2^2$:

28	20	12	4	32	24	16	8	36
19	11	3	31	23	15	7	35	27
10	2	30	22	14	6	34	26	18
1	29	21	13	5	33	25	17	9

Here the 12 unshaded squares are what is left after columns 3, 6, and 9 are excluded

along with rows 2 and 4. In other words we delete multiples of 2 and 3, leaving the numbers $1, 5, 7, 11, 13, 17, 19, 23, 25, 29, 31, 35$ as the numbers less than 36 that are coprime to 36.

In the corresponding three-dimensional picture for $k = 3$ we would be omitting the cubes in certain slices parallel to the three coordinate planes, and similarly for $k > 3$.

We can now obtain a formula for the Euler phi function $\varphi(n)$ which counts the number of congruence classes mod n of integers coprime to n. The arguments above show that $\varphi(n) = \varphi(p_1^{r_1}) \cdots \varphi(p_k^{r_k})$ when $n = p_1^{r_1} \cdots p_k^{r_k}$ for distinct primes p_i. For a prime p we have $\varphi(p^r) = p^r - p^{r-1} = p^{r-1}(p-1)$ since we are counting how many numbers remain from $1, 2, 3, \cdots, p^r$ after we delete $p, 2p, 3p, \cdots, (p^{r-1})p = p^r$. Thus we have a formula for $\varphi(n)$:

$$\begin{aligned}\varphi(n) &= p_1^{r_1-1}(p_1-1)p_2^{r_2-1}(p_2-1)\cdots p_k^{r_k-1}(p_k-1)\\ &= n\left(\frac{p_1-1}{p_1}\right)\left(\frac{p_2-1}{p_2}\right)\cdots\left(\frac{p_k-1}{p_k}\right)\end{aligned}$$

If we omit the factor n from this last product, the remaining product of the terms $(p_i-1)/p_i$ tells what proportion of the numbers less than n are coprime to n. Notice that this does not depend on the exponents r_i. For example $\varphi(36) = \varphi(4)\varphi(9) = 2 \cdot 6 = 12$, which is $1/2 \cdot 2/3 = 1/3$ times 36, in agreement with the preceding figure.

The way that $\varphi(n)$ varies with n is rather erratic since the prime factorizations of adjacent numbers are not related. For example we have $\varphi(1000) = \varphi(2^3 5^3) = 2^2(2-1)5^2(5-1) = 400$, in agreement with the fact that the numbers coprime to 2 and 5 are the numbers with last digit 1, 3, 7, or 9, which means four out of every ten numbers or 400 out of the first 1000 numbers. For the adjacent numbers 999 and 1001 we have $\varphi(999) = \varphi(3^3 \cdot 37) = 18 \cdot 36 = 648$ and $\varphi(1001) = \varphi(7 \cdot 11 \cdot 13) = 6 \cdot 10 \cdot 12 = 720$.

The Chinese Remainder Theorem can be applied to give an example of a Diophantine equation that has a solution mod n for each positive integer n but does not have an actual integer solution. The example is the equation $2x^2 + 7y^2 = 1$. This obviously has no integer solutions, although it does have rational solutions such as $(x, y) = (1/3, 1/3)$ and $(3/5, 1/5)$. We can use either of these rational solutions to get a solution mod n for certain values of n in the following way. Let us take the solution $(3/5, 1/5)$ for example. This rational solution will give an integer solution mod n provided that 5 has a multiplicative inverse "$1/5$" mod n. For example for $n = 14$ a multiplicative inverse for 5 is 3 since $5 \cdot 3 \equiv 1 \mod 14$. If we multiply the equation $2(3/5)^2 + 7(1/5)^2 = 1$ by 5^2 to get $2 \cdot 3^2 + 7 \cdot 1^2 = 5^2$ and then multiply by 3^2, the inverse of 5^2 mod 14, we get $2 \cdot 9^2 + 7 \cdot 3^2 \equiv 1 \mod 14$.

This argument gives a solution of $2x^2 + 7y^2 \equiv 1 \mod n$ whenever 5 has a multiplicative inverse mod n. As we saw earlier in this section, this happens whenever

5 is coprime to n, which means that 5 does not divide n. Similarly, using the other rational solution $(1/3, 1/3)$ we can solve $2x^2 + 7y^2 = 1 \bmod n$ whenever 3 does not divide n by finding a multiplicative inverse for 3 mod n.

There remains the possibility that n is divisible by both 3 and 5, and this is where the Chinese Remainder Theorem will be used. Consider for example the case $n = 30$. We can factor this as $5 \cdot 6$ where one factor is not divisible by 3 and the other is not divisible by 5. By the method above we can obtain a solution of $2x^2 + 7y^2 \equiv 1 \bmod 5$ from $(1/3, 1/3)$ using $3 \cdot 2 \equiv 1 \bmod 5$ so $(1/3, 1/3)$ becomes $(2, 2)$. For $2x^2 + 7y^2 \equiv 1 \bmod 6$ we use $(3/5, 1/5)$ and the fact that $5 \cdot 5 \equiv 1 \bmod 6$ so $(3/5, 1/5)$ becomes $(3 \cdot 5, 5) \equiv (3, 5) \bmod 6$. Thus we want to find (x, y) with $(x, y) \equiv (2, 2) \bmod 5$ and $(x, y) \equiv (3, 5) \bmod 6$. This we do by two applications of the Chinese Remainder Theorem, once for x and once for y. We use the earlier formula $a_1 m_2 n_2 + a_2 m_1 n_1$ where $5n_1 + 6n_2 = 1$ so $n_1 = -1$ and $n_2 = 1$. This yields $x = 2 \cdot 6 \cdot 1 - 3 \cdot 5 \cdot 1 = -3$ and $y = 2 \cdot 6 \cdot 1 - 5 \cdot 5 \cdot 1 = -13$. Thus $2(-3)^2 + 7(-13)^2 \equiv 1 \bmod 5$ and mod 6. This implies the congruence also holds mod 30 since the difference $2(-3)^2 + 7(-13)^2 - 1$ is divisible by 5 and by 6, hence by 30 since 5 and 6 are coprime. This method for the case $n = 30$ works for any n divisible by 3 and 5 since any such n can be factored as $n = kl$ where k is not divisible by 3 and l is not divisible by 5.

One might ask how rational solutions of $2x^2 + 7y^2 = 1$ such as $(1/3, 1/3)$ and $(3/5, 1/5)$ can be found. Rational solutions of $2x^2 + 7y^2 = 1$ are equivalent to integer solutions of $2x^2 + 7y^2 = z^2$, so we are looking for integers x and y such that $2x^2 + 7y^2$ is a square. This is a special case of the general problem of solving quadratic Diophantine equations $ax^2 + bxy + cy^2 = n$ which will be a central theme of the book starting in Chapter 4.

A Digression on Rational Points on Quadratic Curves

A key point in the preceding example was the existence of rational solutions of $2x^2 + 7y^2 = 1$, which correspond to rational points on the curve $2x^2 + 7y^2 = 1$, so let us consider now the general problem of determining when a quadratic curve $ax^2 + bxy + cy^2 = d$ contains rational points. Here a, b, c, and d are rational numbers but there is no loss of generality in assuming they are integers since we can multiply the equation by a common denominator for a, b, c, and d if they are not all integers.

The first step is to reduce to the case that $b = 0$. If $a \neq 0$ we can write:

$$ax^2 + bxy + cy^2 = a\Big(x + \frac{b}{2a}y\Big)^2 + \Big(c - \frac{b^2}{4a}\Big)y^2$$

Then if we change variables to $X = x + b/2a\, y$ and $Y = y$ this converts the equation $ax^2 + bxy + cy^2 = d$ into the equation $aX^2 + c'Y^2 = d$ for $c' = c - b^2/4a$. Rational values of x and y give rational values for X and Y, and conversely rational values for X and Y give rational values for x and y since the change of variables is reversible,

with $x = X - {}^{b}\!/_{2a}\, Y$ and $y = Y$. If $a = 0$ and $c \neq 0$ we can change variables as above but with a and c reversed. If both a and c are 0 the equation is $bxy = d$ which always has rational solutions if $b \neq 0$.

Thus it suffices to determine whether curves $ax^2 + by^2 = c$ have rational points. Again we can multiply through by a common denominator to make a, b, and c integers. We assume a, b, and c are nonzero to avoid trivial cases. To have solutions we obviously need to assume that a and b do not have one sign and c the opposite sign.

If rational numbers x and y satisfy $ax^2 + by^2 = c$ we can put them over a common denominator and write them as quotients X/Z and Y/Z for integers X, Y, Z, and then the equation becomes $aX^2 + bY^2 = cZ^2$ for which we are seeking integer solutions (X, Y, Z). With three variables instead of two it may appear that we have made the problem more complicated, but an advantage of the new equation is that it is *homogeneous* in the sense that all three terms have the same degree, namely 2. This means that if (X, Y, Z) is a solution, then so is (kX, kY, kZ) for any constant k. In particular, rational solutions can always be converted to integer solutions. The homogeneous equation has the trivial solution $(0, 0, 0)$ but this is not very interesting so we will always exclude this trivial solution. In fact we will need solutions with $Z \neq 0$ to get actual points $(x, y) = (X/Z, Y/Z)$ on the curve $ax^2 + by^2 = c$.

Thus we are asking when an equation $ax^2 + by^2 = cz^2$ has an integer or rational solution $(x, y, z) \neq (0, 0, 0)$. There are a few preliminary simplifications in the coefficients a, b, c that can be made. Suppose first that a factors as $a'd^2$ for some integers a' and $d > 1$. The equation can then be written as $a'(dx)^2 + by^2 = cz^2$, and finding rational solutions of $ax^2 + by^2 = cz^2$ is equivalent to finding rational solutions of $a'x^2 + by^2 = cz^2$. Square factors of b and c can be absorbed into y^2 and z^2 in the same way. Thus there is no loss of generality in assuming that each of the coefficients a, b, c in $ax^2 + by^2 = cz^2$ is ***squarefree***, that is, has no square factors greater than 1.

If all three coefficients a, b, c have a common prime factor p we can of course divide the equation by p to get a simpler equation. Repeating this step, we may assume no prime p divides all three coefficients. If p divides two of the coefficients, say $a = pa'$ and $b = pb'$, we can still simplify the equation by multiplying it by p to get $a'(px)^2 + b'(py)^2 = pcz^2$ which can be written as $a'x^2 + b'y^2 = pcz^2$ by absorbing p into x and y, and this is a simpler equation in that $|abc|$ has decreased by a factor of p. The new equation still has squarefree coefficients since we could assume that the divisor p of a and b was not also a divisor of c. By the same reasoning we can arrange also that a and c are coprime and b and c are coprime, with all three coefficients still squarefree.

Now we have Legendre's Theorem as described in Chapter 0:

Theorem 2.6. *An equation $ax^2 + by^2 = cz^2$ with a, b, and c squarefree coprime*

nonzero integers has an integer solution $(x,y,z) \neq (0,0,0)$ *exactly when the following conditions are satisfied*: ac *is a square mod* b, bc *is a square mod* a, $-ab$ *is a square mod* c, *and* a *and* b *do not both have the opposite sign from* c.

A more symmetric statement could be obtained by changing the sign of c and writing the equation as $ax^2+by^2+cz^2=0$. Then the conditions would be that $-ac$ is a square mod b, $-bc$ is a square mod a, $-ab$ is a square mod c, and the three coefficients a,b,c do not all have the same sign.

Proof: First we show that these congruence conditions must be satisfied if a solution exists. Suppose that we have a solution $(x,y,z) \neq (0,0,0)$ of $ax^2+by^2=cz^2$. We can assume each pair of x,y,z is coprime since for example if a prime p divides x and y then p^2 divides ax^2+by^2 hence it divides cz^2, which implies p divides z since c is squarefree. Then the solution (x,y,z) could be simplified by dividing by p.

The equation $ax^2+by^2=cz^2$ implies that $ax^2 \equiv cz^2 \mod b$. After multiplying this congruence by c we get $acx^2 \equiv c^2z^2 \mod b$. Now, x and b are coprime since any prime dividing both would divide $ax^2+by^2=cz^2$ and so would divide c or z, neither of which is possible since b and c are coprime and x and z are coprime. Since x is coprime to b it has a multiplicative inverse mod b. Multiplying the congruence $acx^2 \equiv c^2z^2 \mod b$ by the square of this inverse, we conclude that ac is a square mod b. In the same way we see that bc is a square mod a and $-ab$ is a square mod c.

The converse is considerably harder to prove, so let us first outline what the strategy will be. We will use the more symmetric equation $ax^2+by^2+cz^2=0$. If the left side of this equation could be factored as

$$ax^2+by^2+cz^2=(a_1x+b_1y+c_1z)(a_2x+b_2y+c_2z)$$

with all coefficients integers, then finding a solution of $ax^2+by^2+cz^2=0$ would be rather easy since we would just have to solve the linear Diophantine equation obtained by setting either factor equal to 0. However, factorizations like this rarely exist. Instead we will show that the congruence conditions in the theorem guarantee that there is a factorization modulo a suitable number n, namely $n=abc$. What this means concretely is that if one multiplies out the product of the two linear factors on the right in the displayed equation above, then the coefficients of the x^2, y^2, and z^2 terms will be congruent to a, b, and c mod n and the coefficients of the xy, yz, and xz terms will be 0 mod n. A solution of either congruence $a_ix+b_iy+c_iz \equiv 0 \mod abc$, say $a_1x+b_1y+c_1z \equiv 0 \mod abc$, will then give a solution of the congruence $ax^2+by^2+cz^2 \equiv 0 \mod abc$.

The next step in the proof will be to show that a solution (x,y,z) of the congruence $a_1x+b_1y+c_1z \equiv 0 \mod abc$ can be chosen so that the value of $ax^2+by^2+cz^2$ is a fairly small multiple of abc, in fact either 0 or $\pm abc$. The last step in the proof

will then be a rather subtle trick to convert a solution of $ax^2 + by^2 + cz^2 = \pm abc$ into a solution of $ax^2 + by^2 + cz^2 = 0$.

Now we begin to fill in details. To factor $ax^2 + by^2 + cz^2$ mod abc we first factor it mod a, b, and c separately. To factor it mod a we just need to factor $by^2 + cz^2$ mod a. Multiplying $by^2 + cz^2$ by b gives $b^2y^2 + bcz^2$. We are assuming that $-bc$ is a square mod a so we have $-bc \equiv r^2$ mod a for some integer r. Then $b^2y^2 + bcz^2 \equiv b^2y^2 - r^2z^2$ mod a with $b^2y^2 - r^2z^2$ factoring as $(by + rz)(by - rz)$. Since b is coprime to a it has an inverse b^{-1} mod a so after multiplying the congruence $b^2y^2 + bcz^2 \equiv (by + rz)(by - rz)$ mod a by b^{-1} we have the desired factorization $by^2 + cz^2 \equiv (y + b^{-1}rz)(by - rz)$ mod a. Thus there is a factorization mod a of $ax^2 + by^2 + cz^2$ as a product $(a_1x + b_1y + c_1z)(a_2x + b_2y + c_2z)$ where the coefficients a_1 and a_2 happen to be 0, but this will not be significant for the rest of the argument.

In the same way there are similar factorizations of $ax^2 + by^2 + cz^2$ mod b and mod c, with possibly different coefficients $a_1, b_1, c_1, a_2, b_2, c_2$ of the linear factors. The Chinese Remainder Theorem, applied once for each of the six coefficients, implies that there is a single choice for the coefficients that works mod a, b, and c simultaneously. Since a, b, and c are coprime, the factorization then holds mod abc.

We will be interested in triples (x, y, z) of integers satisfying three inequalities

$$0 \le x < \alpha \qquad 0 \le y < \beta \qquad 0 \le z < \gamma \qquad (*)$$

for positive real numbers α, β, and γ that are not necessarily integers. To count how many triples (x, y, z) satisfy $(*)$ let $\lambda(\alpha)$ be the number of integers x with $0 \le x < \alpha$, so $\lambda(\alpha) = \alpha$ if α is an integer and $\lambda(\alpha) = 1 + \lfloor \alpha \rfloor$ if α is not an integer, where $\lfloor \alpha \rfloor$ is the largest integer less than or equal to α. Thus $\lambda(\alpha) > \alpha$ if α is not an integer. The number of triples (x, y, z) satisfying $(*)$ is then $\lambda(\alpha)\lambda(\beta)\lambda(\gamma)$.

If $\lambda(\alpha)\lambda(\beta)\lambda(\gamma) > |abc|$ there must exist two different triples (x', y', z') and (x'', y'', z'') satisfying $(*)$ such that $a_1x' + b_1y' + c_1z' \equiv a_1x'' + b_1y'' + c_1z''$ mod abc. The triple $(x, y, z) = (x' - x'', y' - y'', z' - z'') \neq (0, 0, 0)$ will then satisfy $a_1x + b_1y + c_1z \equiv 0$ mod abc. The triple $(|x|, |y|, |z|)$ will also satisfy $(*)$ so $x^2 < \alpha^2$, $y^2 < \beta^2$, and $z^2 < \gamma^2$.

For the triple (x, y, z) we have $ax^2 + by^2 + cz^2 \equiv 0$ mod abc from the factorization of $ax^2 + by^2 + cz^2$ mod abc. Since a, b, and c do not all have the same sign, we can assume two are positive and one is negative by multiplying the equation by -1 if necessary. After a possible permutation of the coefficients we can assume that $a > 0$, $b > 0$, and $c < 0$. Since $x^2 < \alpha^2$, $y^2 < \beta^2$, and $z^2 < \gamma^2$ we then have:

$$cy^2 < cz^2 \le ax^2 + by^2 + cz^2 \le ax^2 + by^2 < a\alpha^2 + b\beta^2$$

If we choose $\alpha = \sqrt{|bc|}$, $\beta = \sqrt{|ac|}$, and $\gamma = \sqrt{|ab|}$ then these inequalities give the inequalities $-|abc| < ax^2 + by^2 + cz^2 < 2|abc|$. Since $ax^2 + by^2 + cz^2 \equiv 0$ mod $|abc|$

we must therefore have either $ax^2 + by^2 + cz^2 = 0$ or $ax^2 + by^2 + cz^2 = |abc|$. The chosen values for α, β, and γ also give $\alpha\beta\gamma = |abc|$ so the earlier hypothesis $\lambda(\alpha)\lambda(\beta)\lambda(\gamma) > |abc|$ becomes $\lambda(\alpha)\lambda(\beta)\lambda(\gamma) > \alpha\beta\gamma$ which is satisfied unless α, β, and γ are all integers. Since a, b, and c are coprime and squarefree, α, β, and γ are all integers only when a, b, and c are ± 1, but in this case the equation $ax^2 + by^2 + cz^2 = 0$ is just $x^2 + y^2 - z^2 = 0$ which has obvious integer solutions.

All that remains is to deal with the possibility $ax^2 + by^2 + cz^2 = |abc|$, so $ax^2 + by^2 + cz^2 = -abc$. Rewriting this equation as $ax^2 + by^2 + c(z^2 + ab) = 0$, we would like to convert it into an equation of the form $aX^2 + bY^2 + cZ^2 = 0$. This suggests that we multiply the equation by $z^2 + ab$ to get a term $cZ^2 = c(z^2 + ab)^2$. Multiplying $ax^2 + by^2$ by $z^2 + ab$, we have:

$$\begin{aligned}(ax^2 + by^2)(z^2 + ab) &= ax^2z^2 + a^2bx^2 + by^2z^2 + ab^2y^2 \\ &= a(xz + by)^2 + b(yz - ax)^2\end{aligned}$$

Thus we have a solution of $aX^2 + bY^2 + cZ^2 = 0$, and this is not the trivial solution $(0, 0, 0)$ since $Z = z^2 + ab > 0$. □

To apply Legendre's Theorem one needs to be able to determine which numbers are squares modulo a given number n. The brute force approach is just to compute all the possible squares. For example for $n = 15$ the numbers mod 15 are $0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6$, and ± 7 so the squares mod 15 are obtained by squaring these to get $0, 1, 4, 9, 16 \equiv 1, 25 \equiv 10, 36 \equiv 6$, and $49 \equiv 4$. Thus only six of the fifteen congruence classes mod 15 are squares mod 15, namely $0, 1, 4, 6, 9$, and 10. This approach becomes tedious for large values of n, but in Section 6.2 we will develop more efficient methods for determining whether a number m is a square mod n, which turns out to be quite a subtle question.

Exercises

1. (a) Find all integer solutions of the equations $40x + 89y = 1$ and $40x + 89y = 5$. (b) Find another equation $ax + by = 1$ with integer coefficients a and b that has an integer solution in common with $40x + 89y = 1$. *Hint*: Use the Farey diagram.

2. Find all integers x satisfying the congruence $31x \equiv 1 \bmod 71$, and then do the same for the congruence $31x \equiv 10 \bmod 71$. Are the solutions unique mod 71, i.e., unique up to adding multiples of 71?

3. Find all integer solutions of the equation $9x + 12y + 20z = 4$, and do this more generally for $9x + 12y + 20z = n$.

4. Find all solutions of the simultaneous congruences $x \equiv 6 \bmod 13$ and $x \equiv 7 \bmod 18$.

5. Show that for the Euler phi function the values $\varphi(n)$ approach infinity as n approaches infinity. In other words, show that for each number $N > 0$ there are only finitely many numbers n with $\varphi(n) < N$.

6. For each $n \leq 10$ determine which numbers are squares mod n by direct calculation.

7. Determine which curves $ax^2 + by^2 = c$ contain rational points for each triple of coprime integers a, b, c chosen from the numbers $1, 2, 3, 5$. When rational points exist, find a specific one.