0 A Preview

In this preliminary Chapter 0 we introduce by means of examples some of the main themes of Number Theory, particularly those that will be emphasized in the rest of the book.

Pythagorean Triples

Let us begin by considering right triangles whose sides all have integer lengths. The most familiar example is the $(3,4,5)$ right triangle, but there are many others as well, such as the $(5,12,13)$ right triangle. Thus we are looking for triples (a,b,c) of positive integers such that $a^2+b^2=c^2$. Such triples are called *Pythagorean triples* because of the connection with the Pythagorean Theorem. Our goal will be a formula that gives them all. The ancient Greeks knew such a formula, and even before the Greeks the ancient Babylonians must have known a lot about Pythagorean triples because one of their clay tablets from nearly 4000 years ago has been found which gives a list of 15 different Pythagorean triples, the largest of which is $(12709,13500,18541)$. (Actually, the tablet only gives the numbers a and c from each triple (a,b,c) for some unknown reason, but it is easy to compute b from a and c.)

There is an easy way to create infinitely many Pythagorean triples from a given one just by multiplying each of its three numbers by an arbitrary number n. For example, from $(3,4,5)$ we get $(6,8,10)$, $(9,12,15)$, $(12,16,20)$, and so on. This process produces right triangles that are all similar to each other, so in a sense they are not essentially different triples. In our search for Pythagorean triples there is thus no harm in restricting our attention to triples (a,b,c) whose three numbers have no common factor. Such triples are called *primitive*. The large Babylonian triple mentioned above is primitive, since the prime factorization of 13500 is $2^2 3^3 5^3$ but the other two numbers in the triple are not divisible by 2, 3, or 5.

A fact worth noting in passing is that if two of the three numbers in a Pythagorean triple (a,b,c) have a common factor n, then n is also a factor of the third number. This follows easily from the equation $a^2+b^2=c^2$, since for example if n divides a and b, then n^2 divides a^2 and b^2, so n^2 divides their sum c^2, hence n divides c.

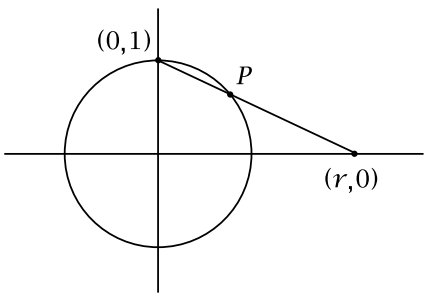
Another case is that n divides a and c. Then n^2 divides a^2 and c^2, so n^2 divides their difference $c^2 - a^2 = b^2$, hence n divides b. In the remaining case that n divides b and c the argument is similar.

A consequence of this divisibility fact is that primitive Pythagorean triples can also be characterized as the ones for which no two of the three numbers have a common factor.

If (a,b,c) is a Pythagorean triple, then we can divide the equation $a^2 + b^2 = c^2$ by c^2 to get an equivalent equation $(a/c)^2 + (b/c)^2 = 1$. This equation is saying that the point $(x,y) = (a/c, b/c)$ is on the unit circle $x^2 + y^2 = 1$ in the xy-plane. The coordinates a/c and b/c are rational numbers, so each Pythagorean triple gives a *rational point* on the circle, a point whose coordinates are both rational. Notice that multiplying each of a, b, and c by the same nonzero integer n yields the same point (x,y) on the circle. Going in the other direction, given a rational point on the circle, we can find a common denominator for its two coordinates so that it has the form $(a/c, b/c)$ and hence gives a Pythagorean triple (a,b,c). We can assume this triple is primitive by canceling any common factor of a, b, and c, and this does not change the point $(a/c, b/c)$. The two fractions a/c and b/c must then be in lowest terms since we observed earlier that if two of a, b, c have a common factor, then all three have a common factor.

From the preceding observations we can conclude that the problem of finding all Pythagorean triples is equivalent to finding all rational points on the unit circle $x^2 + y^2 = 1$. More specifically, there is an exact one-to-one correspondence between primitive Pythagorean triples and rational points on the unit circle that lie in the interior of the first quadrant (since we want all of a,b,c,x,y to be positive).

In order to find all the rational points on the circle $x^2 + y^2 = 1$ we will use a construction that starts with one rational point and creates many more rational points from this one starting point. The four obvious rational points on the circle are the intersections of the circle with the coordinate axes, which are the points $(\pm 1, 0)$ and $(0, \pm 1)$. It does not matter which one we choose as the starting point, so let us choose $(0,1)$. Now consider a line which intersects the circle in this point $(0,1)$ and some other point P, as in the figure at the right. If the line has slope m, its equation will be $y = mx + 1$. If we denote the point where the line intersects the x-axis by $(r,0)$, then $m = -1/r$ so the equation for the line can be rewritten as $y = 1 - x/r$. Here we assume r is nonzero since $r = 0$ corresponds to the slope m being infinite and the point P being $(0,-1)$, a rational point we already know about. To find the coordinates of the point P in terms of r when $r \neq 0$ we substitute $y = 1 - x/r$ into the equation



$x^2 + y^2 = 1$ and solve for x:

$$x^2 + \left(1 - \frac{x}{r}\right)^2 = 1$$

$$x^2 + 1 - \frac{2x}{r} + \frac{x^2}{r^2} = 1$$

$$\left(1 + \frac{1}{r^2}\right)x^2 - \frac{2x}{r} = 0$$

$$\left(\frac{r^2+1}{r^2}\right)x^2 = \frac{2x}{r}$$

We are assuming $P \neq (0,-1)$ so $x \neq 0$ and we can cancel an x from both sides of the last equation above and then solve for x to get $x = {}^{2r}\!/_{r^2+1}$. Plugging this into the formula $y = 1 - {}^{x}\!/_{r}$ gives $y = 1 - {}^{2}\!/_{r^2+1} = {}^{r^2-1}\!/_{r^2+1}$. Thus the coordinates (x,y) of the point P are given by:

$$(x,y) = \left(\frac{2r}{r^2+1}, \frac{r^2-1}{r^2+1}\right)$$

Note that in these formulas we no longer have to exclude the value $r = 0$, which just gives the point $(0,-1)$. Observe also that if we let r approach $\pm\infty$ then the point P approaches $(0,1)$, as we can see either from the picture or from the formulas.

If r is a rational number, then the formula for (x,y) shows that both x and y are rational, so we have a rational point on the circle. Conversely, if both coordinates x and y of the point P on the circle are rational, then the slope m of the line must be rational, hence r must also be rational since $r = -{}^{1}\!/_{m}$. We could also solve the equation $y = 1 - {}^{x}\!/_{r}$ for r to get $r = {}^{x}\!/_{1-y}$, showing again that r will be rational if x and y are rational (and y is not 1). The conclusion of all this is that, starting from the initial rational point $(0,1)$ we have found formulas that give all the other rational points on the circle.

Since there are infinitely many different choices for the rational number r, there are infinitely many rational points on the circle. But we can say something much stronger than this: every arc of the circle, no matter how small, contains infinitely many rational points. This is because every arc on the circle corresponds to an interval of r-values on the x-axis, and every interval in the x-axis contains infinitely many rational numbers. Since every arc on the circle contains infinitely many rational points, we can say that the rational points are *dense* in the circle, meaning that for every point on the circle there is an infinite sequence of rational points approaching the given point.

Now we can go back and find formulas for Pythagorean triples. If we set the rational number r equal to ${}^{p}\!/_{q}$ with p and q integers having no common factor, then the formulas for x and y become:

$$x = \frac{2({}^{p}\!/_{q})}{({}^{p}\!/_{q})^2 + 1} = \frac{2pq}{p^2+q^2} \qquad \text{and} \qquad y = \frac{({}^{p}\!/_{q})^2 - 1}{({}^{p}\!/_{q})^2 + 1} = \frac{p^2-q^2}{p^2+q^2}$$

These formulas give the ratios $x = a/c$ and $y = b/c$ for all Pythagorean triples (a,b,c), so they determine all Pythagorean triples up to multiplication by a constant. The simplest way to realize the ratios $a/c = 2pq/p^2+q^2$ and $b/c = p^2-q^2/p^2+q^2$ is just to take:

$$(a,b,c) = (2pq, p^2 - q^2, p^2 + q^2)$$

The Pythagorean triples given by this formula may not be primitive, however. For example, if x and y are both odd then $p^2 - q^2$ and $p^2 + q^2$ are both even, as is $2pq$, so the triple could be simplified by dividing by 2. The nonprimitive triples obtained in this way are the starred entries in the table below.

(p,q)	(x,y)	(a,b,c)
$(2,1)$	$(4/5, 3/5)$	$(4,3,5)$
$(3,1)^*$	$(6/10, 8/10)^*$	$(6,8,10)^* \to (3,4,5)$
$(3,2)$	$(12/13, 5/13)$	$(12,5,13)$
$(4,1)$	$(8/17, 15/17)$	$(8,15,17)$
$(4,3)$	$(24/25, 7/25)$	$(24,7,25)$
$(5,1)^*$	$(10/26, 24/26)^*$	$(10,24,26)^* \to (5,12,13)$
$(5,2)$	$(20/29, 21/29)$	$(20,21,29)$
$(5,3)^*$	$(30/34, 16/34)^*$	$(30,16,34)^* \to (15,8,17)$
$(5,4)$	$(40/41, 9/41)$	$(40,9,41)$
$(6,1)$	$(12/37, 35/37)$	$(12,35,37)$
$(6,5)$	$(60/61, 11/61)$	$(60,11,61)$
$(7,1)^*$	$(14/50, 48/50)^*$	$(14,48,50)^* \to (7,24,25)$
$(7,2)$	$(28/53, 45/53)$	$(28,45,53)$
$(7,3)^*$	$(42/58, 40/58)^*$	$(42,40,58)^* \to (21,20,29)$
$(7,4)$	$(56/65, 33/65)$	$(56,33,65)$
$(7,5)^*$	$(70/74, 24/74)^*$	$(70,24,74)^* \to (35,12,37)$
$(7,6)$	$(84/85, 13/85)$	$(84,13,85)$

Notice that the primitive versions of the starred triples occur higher in the table, but with a and b switched. This is a general phenomenon, as we will see in the course of proving the following basic result:

Proposition. *All primitive Pythagorean triples (a,b,c), after perhaps interchanging a and b, are obtained from the formula $(a,b,c) = (2pq, p^2 - q^2, p^2 + q^2)$ by letting p and q range over all positive integers with $p > q$, such that p and q have no common factor and are of opposite parity (one even and the other odd).*

Proof: We have seen that the formula $(a,b,c) = (2pq, p^2 - q^2, p^2 + q^2)$ yields all Pythagorean triples up to multiplication by a constant, so we just need to investigate when the formula gives a primitive triple and what to do when it gives a nonprimitive triple. As before we can assume that p and q have no common divisor, and we can assume that $p > q$ in order for the middle coordinate $b = p^2 - q^2$ to be positive.

Case 1: Suppose p and q have opposite parity. If all three of $2pq$, $p^2 - q^2$, and $p^2 + q^2$ have a common divisor $d > 1$ then d would have to be odd since $p^2 - q^2$ and

p^2+q^2 are odd when p and q have opposite parity. Furthermore, since d is a divisor of both p^2-q^2 and p^2+q^2 it must divide their sum $(p^2+q^2)+(p^2-q^2)=2p^2$ and their difference $(p^2+q^2)-(p^2-q^2)=2q^2$. However, since d is odd it would then have to divide p^2 and q^2, forcing p and q to have a common factor (since any prime factor of d would have to divide p and q). This contradicts the assumption that p and q have no common factors, so we conclude that $(2pq, p^2-q^2, p^2+q^2)$ is primitive if p and q have opposite parity.

Case 2: Suppose p and q have the same parity. Then their sum and difference are both even and we can write $p+q=2P$ and $p-q=2Q$ for some integers P and Q. Any common factor of P and Q would have to divide $P+Q=\frac{1}{2}(p+q)+\frac{1}{2}(p-q)=p$ and $P-Q=\frac{1}{2}(p+q)-\frac{1}{2}(p-q)=q$, so P and Q have no common factors. In terms of P and Q our Pythagorean triple becomes:

$$\begin{aligned}(a,b,c) &= (2pq, p^2-q^2, p^2+q^2)\\ &= \big(2(P+Q)(P-Q), (P+Q)^2-(P-Q)^2, (P+Q)^2+(P-Q)^2\big)\\ &= \big(2(P^2-Q^2), 4PQ, 2(P^2+Q^2)\big)\\ &= 2(P^2-Q^2, 2PQ, P^2+Q^2)\end{aligned}$$

Canceling the factor of 2 in front of this last expression gives a new Pythagorean triple $(P^2-Q^2, 2PQ, P^2+Q^2)$ of the same type $(2pq, p^2-q^2, p^2+q^2)$ that we started with but with the first two coordinates switched. This new triple is primitive by Case 1 since P and Q cannot have the same parity, otherwise $p=P+Q$ and $q=P-Q$ would both be even, which is impossible since they have no common factor.

From Cases 1 and 2 we can conclude that if we allow ourselves to switch the first two coordinates, then we get all primitive Pythagorean triples from the formula by restricting p and q to be of opposite parity and have no common factors. □

Pythagorean Triples and Quadratic Forms

There are many questions one can ask about Pythagorean triples (a,b,c). For example, we could begin by asking which numbers actually arise as the numbers a, b, or c in some Pythagorean triple. It is sufficient to answer the question just for primitive Pythagorean triples, since the remaining ones are obtained by multiplying by arbitrary positive integers. We know all primitive Pythagorean triples arise from the formula

$$(a,b,c) = (2pq, p^2-q^2, p^2+q^2)$$

where p and q have no common factor and are of opposite parity. The latter condition just amounts to saying p and q are not both odd since they cannot both be even if they have no common factor. Determining whether a given number can be expressed in one of the forms $2pq$, p^2-q^2, or p^2+q^2 is a special case of the general question

of deciding when an equation $Ap^2+Bpq+Cq^2=n$ has an integer solution p, q, for given integers A, B, C, and n. Expressions of the form $Ax^2+Bxy+Cy^2$ are called *quadratic forms*. These will be the main topic studied in Chapters 4–8, where we will develop some general theory addressing the question of what values a quadratic form takes on when all the numbers involved are integers. For now, let us just look at the special cases at hand.

First let us consider which numbers occur as a or b in primitive Pythagorean triples (a,b,c). A trivial case is the equation $0^2+1^2=1^2$ which shows that 0 and 1 can be realized by the triple $(0,1,1)$ which is primitive, so let us focus on realizing numbers bigger than 1. If we look at the earlier table of Pythagorean triples we see that all the numbers up to 15 can be realized as a or b in primitive triples except for 2, 6, 10, and 14. This might lead us to guess that the numbers realizable as a or b in primitive Pythagorean triples are the numbers not of the form $4k+2$. This is indeed true, and can be proved as follows. First note that since $2pq$ is even, p^2-q^2 must be odd, otherwise both a and b would be even, violating primitivity. Now, every odd number is expressible in the form p^2-q^2 since $2k+1=(k+1)^2-k^2$, so in fact every odd number is the difference between two consecutive squares. Taking $p=k+1$ and $q=k$ yields a primitive triple since k and $k+1$ always have opposite parity and no common factors. This takes care of realizing odd numbers. For even numbers, they would have to be expressible as $2pq$ with p and q of opposite parity, which forces pq to be even so $2pq$ is a multiple of 4 and hence cannot be of the form $4k+2$. On the other hand, if we take $p=2k$ and $q=1$ then $2pq=4k$ with p and q having opposite parity and no common factors.

To summarize, we have shown that all positive numbers $2k+1$ and $4k$ occur as a or b in primitive Pythagorean triples but none of the numbers $4k+2$ occur. To finish the story, note that a number $a=4k+2$ which cannot be realized in a primitive triple can be realized by a nonprimitive triple just by taking a triple (a,b,c) with $a=2k+1$ and doubling each of a, b, and c. Thus all numbers can be realized as a or b in Pythagorean triples (a,b,c).

Now let us ask which numbers c can occur in Pythagorean triples (a,b,c), so we are trying to find a solution of $p^2+q^2=c$ for a given number c. Pythagorean triples (p,q,r) give solutions when c is equal to a square r^2, but we are asking now about arbitrary numbers c. It suffices to figure out which numbers c occur in primitive triples (a,b,c), since by multiplying the numbers c in primitive triples by arbitrary numbers we get the numbers c in arbitrary triples. A look at the earlier table shows that the numbers c that can be realized by primitive triples (a,b,c) seem to be fairly rare: only $5,13,17,25,29,37,41,53,61,65$, and 85 occur in the table. These are all odd, and in fact they are all of the form $4k+1$. This always has to be true because p and q are of opposite parity, so one is an even number $2k$ and the other an odd number $2l+1$. Squaring, we get $(2k)^2=4k^2$ and $(2l+1)^2=4(l^2+l)+1$. Thus the

square of an even number has the form $4u$ and the square of an odd number has the form $4v+1$. Hence p^2+q^2 has the form $4(u+v)+1$, or more simply, just $4k+1$.

The argument we just gave can be expressed more concisely using congruences modulo 4. We will assume the reader has seen something about congruences before, but to recall the terminology: two numbers a and b are said to be congruent modulo a number n if their difference $a-b$ is a multiple of n. When n is negative, congruence modulo n is equivalent to congruence modulo $|n|$, so there is no loss of generality in restricting attention just to congruence modulo positive numbers. Congruence modulo 0 is the same as equality, so there is little reason to consider this case. One writes $a \equiv b \mod n$ to mean that a is congruent to b modulo n, with the word "modulo" abbreviated to "mod". One can tell whether two numbers are congruent mod n by dividing each of them by n and checking whether the remainders, which lie between 0 and $n-1$, are equal. Every number is congruent mod n to one of the numbers $0, 1, 2, \cdots, n-1$, and no two of these numbers are congruent to each other, so there are exactly n congruence classes of numbers mod n, where a congruence class means all the numbers congruent to a given number. In the preceding paragraph we were in effect dealing with congruence classes mod 4 and we saw that the square of an even number is congruent to 0 mod 4 while the square of an odd number is congruent to 1 mod 4, hence p^2+q^2 is congruent to $0+1$ or $1+0$ mod 4 when p and q have opposite parity, so $p^2+q^2 \equiv 1 \mod 4$.

Returning to the question of which numbers occur as c in primitive Pythagorean triples (a,b,c), we have seen that $c \equiv 1 \mod 4$, but looking again at the list $5, 13, 17, 25, 29, 37, 41, 53, 61, 65, 85$ we can observe the more interesting fact that most of these numbers are primes, and the ones that are not primes are products of earlier primes in the list: $25 = 5\cdot5$, $65 = 5\cdot13$, $85 = 5\cdot17$. From this somewhat slim evidence one might conjecture that the numbers c occurring in primitive Pythagorean triples are exactly the numbers that are products of primes congruent to 1 mod 4. The first prime satisfying this condition that is not on the original list is 73, and this is realized as $p^2+q^2 = 8^2+3^2$ in the triple $(48, 55, 73)$. The next two primes congruent to 1 mod 4 are $89 = 8^2+5^2$ and $97 = 9^2+4^2$, so the conjecture continues to look good. As further evidence for the conjecture, numbers congruent to 1 mod 4 that are not on the list such as $9 = 3\cdot3$, $21 = 3\cdot7$, $33 = 3\cdot11$, $45 = 3^2\cdot5$, $49 = 7\cdot7$, and $57 = 3\cdot19$ each have a prime factor that is not congruent to 1 mod 4.

More generally, if we ask which numbers can be expressed as p^2+q^2 for integers p and q having no common divisor without requiring them to have opposite parity, then we will also get the numbers c in the starred entries of the earlier table. As we saw in the proof of the proposition about Pythagorean triples, these values of c are just the doubles of the values of c in primitive Pythagorean triples. Thus one can conjecture that the numbers expressible as p^2+q^2 for positive integers p and q having no common divisor are the products of primes congruent to 1 mod 4 and the

doubles of these products. This conjecture is correct, but proving it is not easy. We will do this in Chapter 6.

After this it is easy to go the last step and ask which numbers are sums p^2+q^2 for arbitrary positive integers p and q. Now we are free to multiply p and q by the same positive integer k, which multiplies p^2+q^2 by k^2. This leads to the answer that the numbers expressible as p^2+q^2, besides 0 and 1, are all the numbers n for which each prime factor congruent to 3 mod 4 occurs to an even power in the prime factorization of n. Thus the sequence of numbers that are sums of two squares begins $0,1,2,4,5,8,9,10,13,16,17,18,20,25,26,29,32,34,36,37,40,\cdots$.

Another question one can ask about Pythagorean triples is how many there are with two of the three numbers differing only by 1. In the earlier table there are several: $(3,4,5)$, $(5,12,13)$, $(7,24,25)$, $(20,21,29)$, $(9,40,41)$, $(11,60,61)$, and $(13,84,85)$. As the pairs of numbers that differ by 1 get larger, the corresponding right triangles are either approximately 45-45-90 right triangles, as with the triple $(20,21,29)$, or long thin triangles, as with $(13,84,85)$. To analyze the possibilities, note first that if two of the numbers in a triple (a,b,c) differ by 1 then the triple has to be primitive, so we can use our formula $(a,b,c)=(2pq,p^2-q^2,p^2+q^2)$. If b and c differ by 1 then we would have $(p^2+q^2)-(p^2-q^2)=2q^2=1$ which is impossible. If a and c differ by 1 then we have $p^2+q^2-2pq=(p-q)^2=1$ so $p-q=\pm1$, and in fact $p-q=+1$ since we must have $p>q$ in order for $b=p^2-q^2$ to be positive. Thus we get the infinite sequence of solutions $(p,q)=(2,1),(3,2),(4,3),\cdots$ with corresponding triples $(4,3,5),(12,5,13),(24,7,25),\cdots$. Note that these are the same triples we obtained earlier that realize all the odd values $b=3,5,7,\cdots$.

The remaining case is that a and b differ by 1. Thus we have the equation $p^2-2pq-q^2=\pm1$. The left side does not factor using integer coefficients, so it is not so easy to find integer solutions this time. In the table there are only the two triples $(4,3,5)$ and $(20,21,29)$, with $(p,q)=(2,1)$ and $(5,2)$. After some trial and error one could find the next solution $(p,q)=(12,5)$ which gives the triple $(120,119,169)$. Is there a pattern in the solutions $(2,1),(5,2),(12,5)$? One has the numbers $1,2,5,12$, and perhaps it is not too great a leap to notice that the third number is twice the second plus the first, while the fourth number is twice the third plus the second. If this pattern continued, the next number would be $29=2\cdot12+5$, giving $(p,q)=(29,12)$, and this does indeed satisfy $p^2-2pq-q^2=1$, yielding the Pythagorean triple $(696,697,985)$. These numbers are increasing rather rapidly, and the next case $(p,q)=(70,29)$ yields an even bigger Pythagorean triple $(4060,4059,5741)$. Could there be other solutions of $p^2-2pq-q^2=\pm1$ with smaller numbers that we missed? We will develop tools in Chapters 4 and 5 to find all the integer solutions, and it will turn out that the sequence we have just discovered gives them all.

Although the quadratic form $p^2 - 2pq - q^2$ does not factor using integer coefficients, it can be simplified slightly be rewriting it as $(p-q)^2 - 2q^2$. Then if we change variables by setting $(x, y) = (p - q, q)$ we obtain the quadratic form $x^2 - 2y^2$. Finding integer solutions of $x^2 - 2y^2 = n$ is equivalent to finding integer solutions of $p^2 - 2pq - q^2 = n$ since integer values of p and q give integer values of x and y, and conversely, integer values of x and y give integer values of p and q since when we solve for p and q in terms of x and y, we again get equations with integer coefficients: $(p, q) = (x + y, y)$. Thus the quadratic forms $p^2 - 2pq - q^2$ and $x^2 - 2y^2$ are completely equivalent, and finding integer solutions of $p^2 - 2pq - q^2 = \pm 1$ is equivalent to finding integer solutions of $x^2 - 2y^2 = \pm 1$.

The equation $x^2 - 2y^2 = \pm 1$ is an instance of the equation $x^2 - Dy^2 = \pm 1$ which is known as *Pell's equation* (although sometimes this term is used only when the right side of the equation is $+1$ and the other case is called the negative Pell equation). This is a very famous equation in Number Theory which has arisen in many different contexts going back hundreds of years. We will develop techniques for finding all integer solutions of Pell's equation for arbitrary values of D in Chapters 4 and 5. It is interesting that certain fairly small values of D can force the solutions to be quite large. For example, for $D = 61$ the smallest positive integer solution of $x^2 - 61y^2 = 1$ is a rather large pair:

$$(x, y) = (1766319049, 226153980)$$

As far back as the eleventh and twelfth centuries mathematicians in India knew how to find this solution. It was rediscovered in the seventeenth century by Fermat in France, who also gave the smallest solution of $x^2 - 109y^2 = 1$, an even larger pair:

$$(x, y) = (158070671986249, 15140424455100)$$

The way that the size of the smallest solution of $x^2 - Dy^2 = 1$ depends upon D is very erratic and is still not well understood today.

Pythagorean Triples and Complex Numbers

There is another way of looking at Pythagorean triples that involves complex numbers, surprisingly enough. The starting point here is the observation that $a^2 + b^2$ can be factored as $(a+bi)(a-bi)$ where $i = \sqrt{-1}$. If we rewrite the equation $a^2 + b^2 = c^2$ as $(a + bi)(a - bi) = c^2$ then since the right side of the equation is a square, we might wonder whether each factor $a \pm bi$ on the left side would have to be a square too. For example, in the case of the triple $(3, 4, 5)$ we have $(3 + 4i)(3 - 4i) = 5^2$ with $3 + 4i = (2 + i)^2$ and $3 - 4i = (2 - i)^2$. So let us ask optimistically whether the equation $(a+bi)(a-bi) = c^2$ can be rewritten as $(p+qi)^2(p-qi)^2 = c^2$ with $a + bi = (p+qi)^2$ and $a - bi = (p - qi)^2$. We might hope also that the equation $(p + qi)^2(p - qi)^2 = c^2$

was obtained by simply squaring the equation $(p+qi)(p-qi)=c$. Let us see what happens when we multiply these various products out:

$$\begin{aligned} a+bi=(p+qi)^2&=(p^2-q^2)+(2pq)i\\ &\text{hence}\quad a=p^2-q^2\quad\text{and}\quad b=2pq\\ a-bi=(p-qi)^2&=(p^2-q^2)-(2pq)i\\ &\text{hence again}\quad a=p^2-q^2\quad\text{and}\quad b=2pq\\ c=(p+qi)(p-qi)&=p^2+q^2 \end{aligned}$$

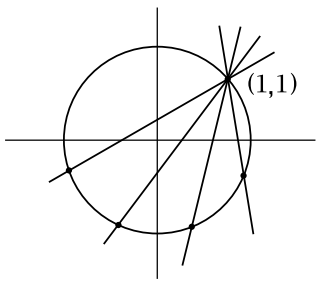
Thus we have miraculously recovered the formulas for Pythagorean triples that we obtained earlier by geometric means, with a and b switched, which does not really matter:

$$a=p^2-q^2 \qquad b=2pq \qquad c=p^2+q^2$$

Our derivation of these formulas just now depended on several assumptions that we have not justified, but it does suggest that looking at complex numbers of the form $a+bi$ where a and b are integers might be a good idea. These complex numbers $a+bi$ with a and b integers are called *Gaussian integers*, after C. F. Gauss, the first mathematician to make a thorough algebraic study of them some 200 years ago. We will develop the basic properties of Gaussian integers in Chapter 8, in particular explaining why the derivation of the formulas above is valid.

Rational Points on Quadratic Curves

The same technique we used to find the rational points on the circle $x^2+y^2=1$ can also be used to find all the rational points on other quadratic curves $Ax^2+Bxy+Cy^2+Dx+Ey=F$ with integer or rational coefficients A,B,C,D,E,F, provided that we can find a single rational point (x_0,y_0) on the curve to start the process. For example, the circle $x^2+y^2=2$ contains the rational points $(\pm1,\pm1)$ and we can use one of these as an initial point. Taking the point $(1,1)$, we would consider lines $y-1=m(x-1)$ of slope m passing through this point. Solving this equation for y and plugging into the equation $x^2+y^2=2$ would produce a quadratic equation $ax^2+bx+c=0$ whose coefficients are polynomials in the variable m, so these coefficients would be rational whenever m is rational. From the quadratic formula $x=(-b\pm\sqrt{b^2-4ac}\,)/2a$ we see that the sum of the two roots is $-b/a$, a rational number if m is rational, so if one root is rational then the other root will be rational as well. The initial point $(1,1)$ on the curve $x^2+y^2=2$ gives $x=1$ as one rational root



of the equation $ax^2+bx+c=0$, so for each rational value of m the other root x will be rational. Then the equation $y-1=m(x-1)$ implies that y will also be rational, and hence we obtain a rational point (x,y) on the curve for each rational value of m. Conversely, if x and y are both rational and $x\neq 1$ then obviously $m=\frac{y-1}{x-1}$ will be rational. Thus one obtains a dense set of rational points on the circle $x^2+y^2=2$, since the slope m can be any rational number. An exercise at the end of the chapter is to work out the formulas explicitly.

Note that the point $(1,-1)$ is a rational point on the circle which does not arise from the formulas parametrizing x and y in terms of m since it corresponds to $m=\infty$. This is analogous to the earlier case of the circle $x^2+y^2=1$ where the point $(0,-1)$ corresponded to $m=\infty$ and $r=0$. For the circle $x^2+y^2=2$ we could just as well use the parameter r instead of m, with $(r,0)$ the point where the line through $(1,1)$ intersects the x-axis. There are simple formulas relating r and m, namely $r=\frac{m-1}{m}$ and $m=\frac{1}{1-r}$. From this viewpoint the exceptional slope $m=\infty$ corresponds to $r=1$ which is not exceptional for the parametrization by r, while the exceptional value $r=\infty$ corresponds to the nonexceptional value $m=0$ when the line through $(1,1)$ is parallel to the x-axis.

If we consider the circle $x^2+y^2=3$ instead of $x^2+y^2=2$ then there are no obvious rational points. And in fact this circle contains no rational points at all. For if there were a rational point, this would yield a solution of the equation $a^2+b^2=3c^2$ by integers a, b, and c with $c\neq 0$. We can assume a, b, and c have no common factor. Then a and b cannot both be even, otherwise the left side of the equation would be even, forcing c to be even, so a, b, and c would have a common factor of 2. To complete the argument we look at the equation modulo 4. As we saw earlier, the square of an even number is 0 mod 4, while the square of an odd number is 1 mod 4. Thus, modulo 4, the left side of the equation is either $0+1$, $1+0$, or $1+1$ since a and b are not both even. So the left side is either 1 or 2 mod 4. However, the right side is either $3\cdot 0$ or $3\cdot 1$ mod 4. We conclude that there can be no integer solutions of $a^2+b^2=3c^2$ with $c\neq 0$. When $c=0$ there is of course the trivial solution $(a,b,c)=(0,0,0)$ but this is not interesting so we will generally disregard it in equations of this type.

The technique we just used to show that $a^2+b^2=3c^2$ has no nontrivial integer solutions can be used in many other situations as well. The underlying reasoning is that if an equation with integer coefficients has an integer solution, then this gives a solution modulo n for all numbers n. For solutions modulo n there are only a finite number of possibilities to check, although for large n this is a large finite number. If one can find a single value of n for which there is no solution modulo n, then the original equation has no integer solutions. However, this implication is not reversible, as it is possible for an equation to have solutions modulo n for every number n and still have no actual integer solutions. A concrete example is the equation

$2x^2 + 7y^2 = 1$. This obviously has no integer solutions, yet it does have solutions modulo n for each n, although this is certainly not obvious. Note that the ellipse $2x^2 + 7y^2 = 1$ does contain rational points such as $(1/3, 1/3)$ and $(3/5, 1/5)$. These can in fact be used to show that $2x^2 + 7y^2 = 1$ has solutions modulo n for each n, as we will show in Section 2.3 of Chapter 2 when we study congruences in more detail.

In Chapter 6 we will find a complete answer to the question of when the circle $x^2 + y^2 = n$ contains rational points by showing that there are rational points on this circle only when there are integer points on it. This reduces the problem to one we considered earlier, finding the integers n that are sums of two squares.

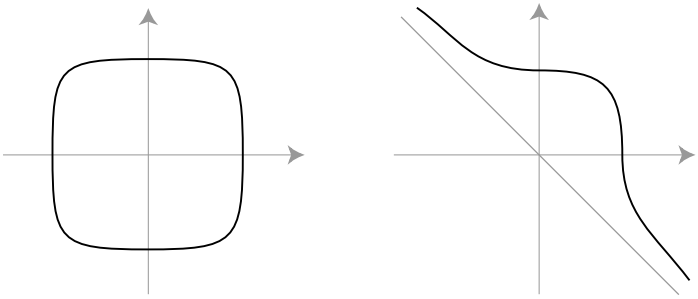
Determining when a quadratic curve contains rational points turns out to be much easier than determining when it has integer points. The general problem reduces fairly quickly to finding rational points on ellipses or hyperbolas of the special form $Ax^2+By^2 = C$ where A, B, and C are integers that are not divisible by squares greater than 1, and such that no two of A, B, and C have a common factor. A theorem of Legendre then asserts that the curve $Ax^2 + By^2 = C$ contains rational points exactly when three congruence conditions modulo A, B, and C are satisfied, namely AC must be congruent mod B to the square of some number, and likewise BC must be a square mod A and $-AB$ must be a square mod C. (There is also the obvious condition that A and B cannot both have opposite sign from C.) For example, if $C = 1$ this reduces just to saying that each of A and B is congruent to a square modulo the other one since the congruence condition mod C holds automatically when $C = 1$. For the ellipse $2x^2 + 7y^2 = 1$ this agrees with what we saw earlier since 2 is a square mod 7, namely 3^2, and 7 is a square mod 2, namely 1^2, so Legendre's theorem guarantees that the curve has a rational point. In the case of the circle $x^2 + y^2 = 3$ the congruence conditions reduce simply to -1 being a square mod 3, which it is not since every number is congruent to 0, 1, or 2 mod 3 so the squares mod 3 are just 0 and 1 since $2^2 \equiv 1 \bmod 3$.

Diophantine Equations

Equations like $x^2 + y^2 = z^2$ or $x^2 - Dy^2 = 1$ that involve polynomials with integer coefficients, and where the solutions sought are required to be integers, or perhaps just rationals, are called *Diophantine equations* after the Greek mathematician Diophantus (ca. 250 A.D.) who wrote a book about these equations that was very influential when European mathematicians started to consider this topic much later in the 1600s. Usually Diophantine equations are very hard to solve because of the restriction to integer solutions. The first really interesting case is quadratic Diophantine equations. By the year 1800 there was quite a lot known about the quadratic case, and we will be focusing on this case in this book.

Diophantine equations of higher degree than quadratic are much more challenging to understand. Probably the most famous one is $x^n + y^n = z^n$ where n is a fixed integer greater than 2. In the 1600s when the French mathematician Fermat was reading about Pythagorean triples in his copy of Diophantus' book, he made a marginal note that, in contrast with the equation $x^2 + y^2 = z^2$, the equation $x^n + y^n = z^n$ has no solutions with positive integers x, y, z when $n > 2$. This is one of many statements that he claimed were true but never wrote proofs of for public distribution, nor have proofs been found among his manuscripts. Over the next century other mathematicians discovered proofs for all his other statements, but this one was far more difficult to verify. The issue is clouded by the fact that he only wrote this statement down the one time, whereas all his other important results were stated numerous times in his correspondence with other mathematicians of the time. So perhaps he only briefly believed he had a proof. In any case, the statement has become known as Fermat's Last Theorem. It was finally proved in the 1990s by Andrew Wiles, using some very deep mathematics developed mostly over the preceding couple decades.

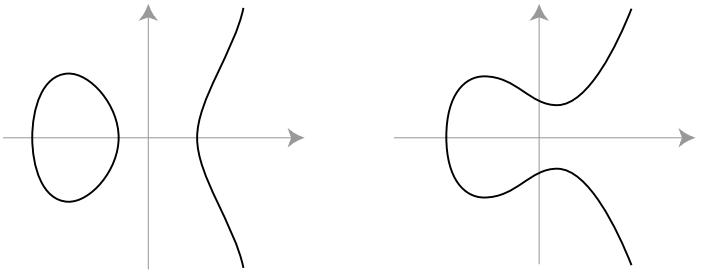
We have seen that finding integer solutions of $x^2+y^2 = z^2$ is equivalent to finding rational points on the circle $x^2+y^2 = 1$, and in the same way, finding integer solutions of $x^n + y^n = z^n$ is equivalent to finding rational points on the curve $x^n + y^n = 1$. For even values of $n > 2$ this curve looks like a flattened circle or rounded square, while for odd n it has a similar shape in the first quadrant but a rather different shape elsewhere, extending out to infinity in the second and fourth quadrants, asymptotic to the line $y = -x$:



Fermat's Last Theorem is equivalent to the statement that these curves have no rational points except their intersections with the coordinate axes, where x or y is 0. These examples show that it is possible for a curve defined by an equation of degree greater than 2 to contain only a finite number of rational points (either two points or four points here, depending on whether n is odd or even) whereas quadratic curves like $x^2 + y^2 = n$ contain either no rational points or an infinite dense set of rational points.

After quadratic curves the next case that has been studied in great depth is cubic curves such as the curves defined by equations $y^2 = x^3 + ax^2 + bx + c$. These are known as elliptic curves, not because they are ellipses but because of a connection

with the problem of computing the length of an arc of an ellipse. Depending on the values of the coefficients a, b, c elliptic curves can have either one or two connected pieces:

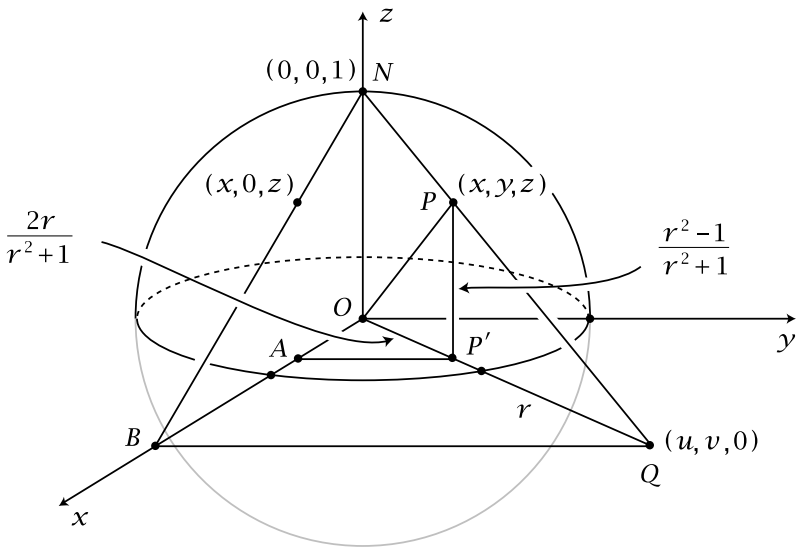


In some cases the number of rational points is finite, any number from 0 to 10 as well as 12 or 16 according to a difficult theorem of Mazur. In other cases the number of rational points is infinite and they form a dense set in the curve, or possibly just in the component that stretches to infinity when there are two components. There is no simple way known for predicting the number of rational points from the coefficients. Interestingly, elliptic curves play an important role in the proof of Fermat's Last Theorem. Their theory is much deeper than for quadratic curves, and so elliptic curves are well beyond the scope of this book.

Rational Points on a Sphere

Although we will not be discussing this later in the book, another way to generalize quadratic curves, in a different direction from considering cubic and higher degree curves, is to keep the quadratic condition but introduce more variables. After quadratic curves the next case would be quadratic surfaces, or as they are usually called, *quadric surfaces*. These are surfaces in three-dimensional space defined by an equation $Q(x, y, z) = n$ where $Q(x, y, z)$ is a quadratic function of three variables. Perhaps the simplest example is the equation $x^2 + y^2 + z^2 = 1$ which defines the sphere of radius 1 with center at the origin. Other quadric surfaces are ellipsoids, paraboloids, hyperboloids, and certain cones and cylinders.

Much of the theory of quadric surfaces parallels that for quadratic curves. To illustrate, let us consider the problem of finding all the rational points on the sphere $x^2 + y^2 + z^2 = 1$, the triples (x, y, z) of rational numbers that satisfy this equation. Some obvious rational points are the points where the sphere meets the coordinate axes such as the point $(0, 0, 1)$ on the z-axis. Following what we did for the circle $x^2 + y^2 = 1$, consider a line from $(0, 0, 1)$ to a point $(u, v, 0)$ in the xy-plane. This line intersects the sphere at some point (x, y, z), and we want to find formulas expressing x, y, and z in terms of u and v. To do this we use the following figure:



Suppose we look at the vertical plane containing the triangle ONQ. From our earlier analysis of rational points on a circle of radius 1 we know that if the segment OQ has length $|OQ| = r$, then $|OP'| = {}^{2r}/_{r^2+1}$ and $z = {}^{r^2-1}/_{r^2+1}$. From the right triangle OBQ we see that $u^2 + v^2 = r^2$. The triangle OBQ is similar to the triangle OAP' and the scaling factor to go from OBQ to OAP' is

$$\frac{|OP'|}{|OQ|} = \frac{2r/(r^2+1)}{r} = \frac{2}{r^2+1}$$

Hence

$$x = \frac{2}{r^2+1}\cdot u = \frac{2u}{u^2+v^2+1} \qquad \text{and} \qquad y = \frac{2}{r^2+1}\cdot v = \frac{2v}{u^2+v^2+1}$$

Also we have

$$z = \frac{r^2-1}{r^2+1} = \frac{u^2+v^2-1}{u^2+v^2+1}$$

Summarizing, we have expressed x, y, and z in terms of u and v by the formulas

$$x = \frac{2u}{u^2+v^2+1} \qquad y = \frac{2v}{u^2+v^2+1} \qquad z = \frac{u^2+v^2-1}{u^2+v^2+1}$$

We can also express u and v in terms of x, y, and z. The projection of the point $P = (x, y, z)$ onto the xz-plane is the point $(x, 0, z)$ which is on the line through B and N. The slope of this line is $-{}^1/_u$ so the equation for the line is $z = 1 - {}^x/_u$. Solving this for u gives $u = {}^x/_{1-z}$. Interchanging x and y corresponds to interchanging u and v so we also have $v = {}^y/_{1-z}$.

From the formulas relating (x, y, z) to (u, v) we see that x, y, and z are rational exactly when u and v are rational. Thus we have formulas for all the rational points (x, y, z) on the sphere except for the pole $(0, 0, 1)$ in terms of rational parameters u and v.

Here is a short table giving a few rational points on the sphere and the corresponding integer solutions of the equation $a^2 + b^2 + c^2 = d^2$:

(u, v)	(x, y, z)	(a, b, c, d)
$(1, 2)$	$(1/3, 2/3, 2/3)$	$(1, 2, 2, 3)$
$(2, 3)$	$(2/7, 3/7, 6/7)$	$(2, 3, 6, 7)$
$(1, 4)$	$(1/9, 4/9, 8/9)$	$(1, 4, 8, 9)$
$(2, 2)$	$(4/9, 4/9, 7/9)$	$(4, 4, 7, 9)$
$(1, 3)$	$(2/11, 6/11, 9/11)$	$(2, 6, 9, 11)$
$(3/2, 3/2)$	$(6/11, 6/11, 7/11)$	$(6, 6, 7, 11)$
$(3, 4)$	$(3/13, 4/13, 12/13)$	$(3, 4, 12, 13)$
$(2, 5)$	$(2/15, 5/15, 14/15)$	$(2, 5, 14, 15)$
$(1/2, 5/2)$	$(2/15, 10/15, 11/15)$	$(2, 10, 11, 15)$

These are in fact all the primitive positive solutions of $a^2 + b^2 + c^2 = d^2$ with $d \le 15$, up to permutations of a, b, and c.

As with rational points on the circle $x^2 + y^2 = 1$, rational points on the sphere $x^2 + y^2 + z^2 = 1$ are dense since rational points are dense in the xy-plane. Thus there are lots of rational points scattered all over the sphere. In linear algebra courses one is often called upon to create unit vectors (x, y, z) by taking a given vector and rescaling it to have length 1 by dividing it by its length. For example, the vector $(1, 1, 1)$ has length $\sqrt{3}$ so the corresponding unit vector is $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$. It is rare that this process produces unit vectors having rational coordinates, but the formulas derived above give a way to create as many rational unit vectors as we like.

The correspondence we have described between points (x, y, z) on a sphere and points (u, v) in the plane is called *stereographic projection*. One can think of the sphere and the plane as being made of clear glass, and if one looks outward and downward from the north pole of the sphere the points of the sphere are projected onto points in the plane, and vice versa. The north pole itself does not project onto any point in the plane, but points approaching the north pole project to points approaching infinity in the plane, so one can think of the north pole as corresponding to an imaginary infinitely distant "point" in the plane. This geometric viewpoint somehow makes infinity less of a mystery, as it just corresponds to a point on the sphere, and points on a sphere are not very mysterious. (Though in the early days of polar exploration the north pole may have seemed very mysterious and infinitely distant.)

One might ask also about spheres $x^2 + y^2 + z^2 = n$, following what we did for circles $x^2 + y^2 = n$. Finding an integer point on $x^2 + y^2 + z^2 = n$ is asking whether n is a sum of three squares. One can test small values of n and one finds that most numbers are sums of three squares, so it is easier to list the ones that are not: $7, 15, 23, 28, 31, 39, 47, 55, 60, 63, 71, 79, 87, 92, 95, \cdots$. The odd numbers here are just the numbers $8k + 7$, and the even numbers seem to be 4 times the earlier numbers on the list. In fact it is easy to see that numbers congruent to 7 mod 8 cannot

be expressed as sums of three squares by the following argument. The squares mod 8 are $0^2 = 0$, $(\pm1)^2 = 1$, $(\pm2)^2 = 4$, $(\pm3)^2 = 9 \equiv 1$, and $4^2 = 16 \equiv 0$, so the squares of even numbers are 0 or 4 mod 8 and the squares of odd numbers are 1 mod 8. Obviously 7 cannot be realized as a sum of three terms 0, 1, or 4, so numbers congruent to 7 mod 8 cannot be sums of three squares.

To rule out numbers $4(8k+7)$ as sums of three squares, we can work mod 4 where the squares are just 0 and 1. If we have $x^2 + y^2 + z^2 = 4n$ then $x^2 + y^2 + z^2 \equiv 0$ mod 4, and the only way to get 0 as a sum of three numbers 0 or 1 is as $0 + 0 + 0$. This means each of x, y, and z must be even, so we can cancel a 4 from both sides of the equation $x^2 + y^2 + z^2 = 4n$ to get n expressed as a sum of three squares. Thus numbers $4(8k + 7)$ are never realizable as sums of three squares since $8k + 7$ is never a sum of three squares. Repeating this argument, we see that $16(8k + 7)$ is never a sum of three squares since $4(8k + 7)$ is not a sum of three squares. Similarly $4^l(8k + 7)$ is never a sum of three squares for any larger exponent l.

The converse statement that every number not of the form $4^l(8k + 7)$ is expressible as a sum of three squares is true but is much harder to prove. This was first done by Legendre.

This answers the question of when the sphere $x^2 + y^2 + z^2 = n$ contains integer points, but could it contain rational points without containing integer points? Let us show that this cannot happen. A rational point on $x^2 + y^2 + z^2 = n$ is equivalent to an integer solution of $a^2 + b^2 + c^2 = nd^2$. It will suffice to show that if n is not a sum of three squares, then neither is nd^2 for any integer d. An equivalent statement is that if n is of the form $4^l(8k + 7)$ then so is nd^2. To prove this, let us write d as $2^p q$ with q odd and $p \geq 0$, hence $d^2 = 4^p q^2$ with $q^2 \equiv 1$ mod 8 since q is odd. Thus we have $nd^2 = 4^{l+p}(8k + 7)q^2$ where the product $(8k + 7)q^2$ is 7 mod 8 since $8k + 7$ is 7 mod 8 and q^2 is 1 mod 8. This shows what we wanted, that if n is of the form $4^l(8k + 7)$ then so is nd^2.

For a general quadric surface defined by a quadratic equation with integer coefficients there is a theorem due to Minkowski, analogous to Legendre's theorem for quadratic curves, that says that rational points exist exactly when certain congruence conditions are satisfied. In general, having rational points on a quadric surface is not equivalent to having integer points as it was for spheres, and the existence of integer points is a more delicate question.

Moving on to four variables, one could ask about integer or rational points on the spheres $x^2 + y^2 + z^2 + w^2 = n$ in four-dimensional space. Integers that could not be expressed as the sum of three squares can be realized as sums of four squares, for example $7 = 2^2 + 1^2 + 1^2 + 1^2$ and $15 = 3^2 + 2^2 + 1^2 + 1^2$, and it is a theorem of Lagrange that every positive number can be expressed as the sum of four squares. Thus the spheres $x^2 + y^2 + z^2 + w^2 = n$ always contain integer points.

Minkowski's theorem remains true for quadratic equations with integer coeffi-

cients in any number of variables, as does the fact that the existence of a single rational solution implies that rational solutions are dense.

Exercises

1. (a) Make a list of the 16 primitive Pythagorean triples (a,b,c) with $c \le 100$, regarding (a,b,c) and (b,a,c) as the same triple.
(b) How many more would there be if we allowed nonprimitive triples?
(c) How many triples (primitive or not) are there with $c = 65$?

2. (a) Find all the positive integer solutions of $x^2 - y^2 = 512$ by factoring $x^2 - y^2$ as $(x+y)(x-y)$ and considering the possible factorizations of 512.
(b) Show that the equation $x^2 - y^2 = n$ has only a finite number of integer solutions for each value of $n > 0$.
(c) Find a value of $n > 0$ for which the equation $x^2 - y^2 = n$ has at least 100 different positive integer solutions.

3. (a) Show that there are only a finite number of Pythagorean triples (a,b,c) with a equal to a given number n.
(b) Show that there are only a finite number of Pythagorean triples (a,b,c) with c equal to a given number n.

4. Find an infinite sequence of primitive Pythagorean triples where two of the numbers in each triple differ by 2.

5. Find a right triangle whose sides have integer lengths and whose acute angles are close to 30 and 60 degrees by first finding the irrational value of r that corresponds to a right triangle with acute angles exactly 30 and 60 degrees, then choosing a rational number close to this irrational value of r.

6. Find a right triangle whose sides have integer lengths and where one of the two shorter sides is approximately twice as long as the other, using a method like the one in the preceding problem. (One possible answer might be the $(8,15,17)$ triangle, or a triangle similar to this, but you should do better than this.)

7. Find a rational point on the sphere $x^2 + y^2 + z^2 = 1$ whose three coordinates are nearly equal.

8. (a) Derive formulas that give all the rational points on the circle $x^2 + y^2 = 2$ in terms of a rational parameter m, the slope of the line through the point $(1,1)$ on the circle. (The value $m = \infty$ should be allowed as well, yielding the point $(1,-1)$.) The calculations may be a little messy, but they eventually simplify to give formulas that are not too complicated:

$$x = \frac{m^2 - 2m - 1}{m^2 + 1} \qquad y = \frac{-m^2 - 2m + 1}{m^2 + 1}$$

(b) Using these formulas, find five different rational points on the circle in the first quadrant, and hence five solutions of $a^2 + b^2 = 2c^2$ with positive integers a, b, c.
(c) The equation $a^2 + b^2 = 2c^2$ can be rewritten as $c^2 = \frac{1}{2}(a^2 + b^2)$, which says that c^2 is the average of a^2 and b^2, or in other words, the squares a^2, c^2, b^2 form an arithmetic progression. One can assume $a < b$ by switching a and b if necessary. Find four such arithmetic progressions of three increasing squares where in each case the three numbers have no common divisors.

9. (a) Find formulas that give all the rational points on the upper branch of the hyperbola $y^2 - x^2 = 1$.
(b) Can you find any relationship between these rational points and Pythagorean triples?

10. (a) Show that the equation $x^2 - 2y^2 = \pm 3$ has no integer solutions by considering this equation modulo 8.
(b) Show that there are no primitive Pythagorean triples (a, b, c) with a and b differing by 3.

11. Show there are no rational points on the circle $x^2 + y^2 = 3$ using congruences modulo 3 instead of modulo 4.

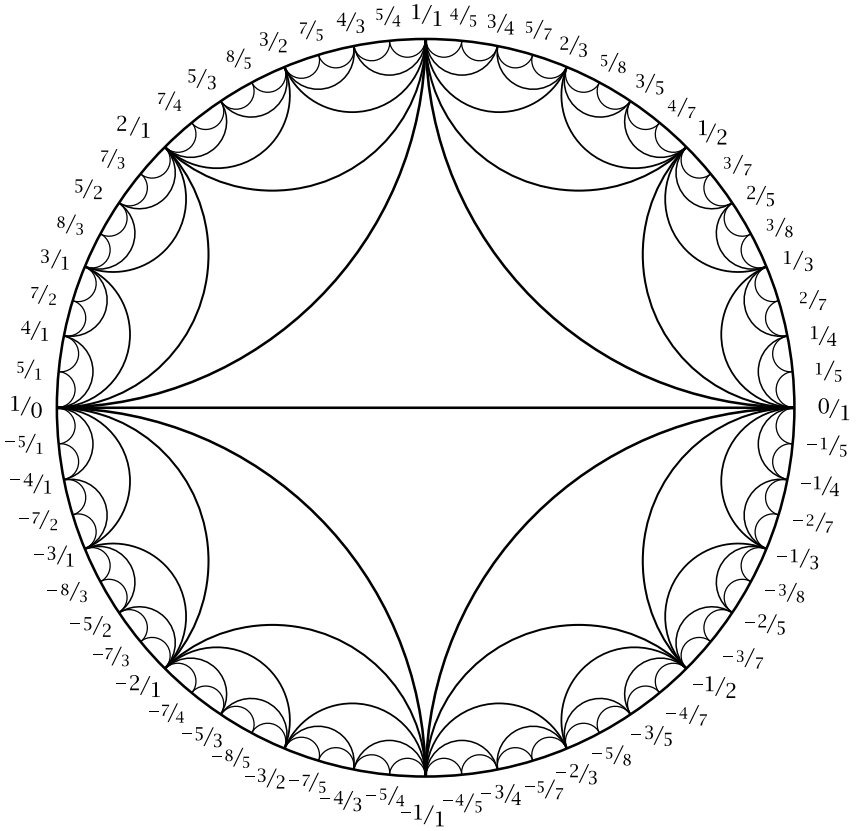
12. Show that for every Pythagorean triple (a, b, c) the product abc must be divisible by 60. (It suffices to show that abc is divisible by 3, 4, and 5.)

13. Use congruences modulo 8 to show that primitive solutions of $a^2 + b^2 + c^2 = d^2$ must have d odd and must have two of a, b, c even and the other odd.

14. Show that if the curve $x^n + y^n = 1$ has a rational point with x and y nonzero, then it has a rational point with x and y positive. *Hint*: Consider the equation $a^n + b^n = c^n$.

1 The Farey Diagram

Our goal is to use geometry to study numbers. Of the various kinds of numbers, the simplest are integers, along with their ratios, the rational numbers. Usually one thinks of rational numbers geometrically as points along a line, interspersed with irrational numbers as well. In this chapter we introduce a two-dimensional pictorial representation of rational numbers that displays certain interesting relations between them that we will be exploring. This diagram, along with several variants of it that will be introduced later, is known as the *Farey diagram*. The origin of the name will be explained when we get to one of these variants. Here is the diagram:



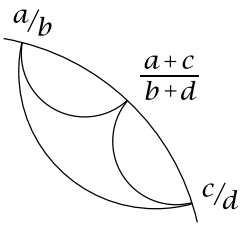
What is shown here is not the whole diagram but only a finite part of it. The actual diagram has infinitely many curvilinear triangles, getting smaller and smaller out near

the boundary circle. The diagram can be constructed by first inscribing the two big triangles in the circle, then adding the four triangles that share an edge with the two big triangles, then the eight triangles sharing an edge with these four, then sixteen more triangles, and so on forever. With a little practice one can draw the diagram without lifting one's pencil from the paper: First draw the outer circle starting at the left or right side, then the diameter, then make the two large triangles, then the four next-largest triangles, and so on.

Our first task will be to explain how the vertices of all the triangles are labeled with rational numbers. Perhaps the reader can guess what the rules are before we spell them out in detail.

1.1 The Mediant Rule

The vertices of the triangles in the Farey diagram are labeled with fractions a/b, including the fraction $1/0$ for ∞, according to the following scheme. In the upper half of the diagram, first label the vertices of the big triangle $1/0$, $0/1$, and $1/1$. Then add labels for successively smaller triangles by the rule that, if the labels at the two ends of the long edge of a triangle are a/b and c/d, then the label on the third vertex of the triangle is $\frac{a+c}{b+d}$, so the numerators and denominators are added separately, contrary to the usual way of adding fractions. The fraction $\frac{a+c}{b+d}$ is called the *mediant* of a/b and c/d.

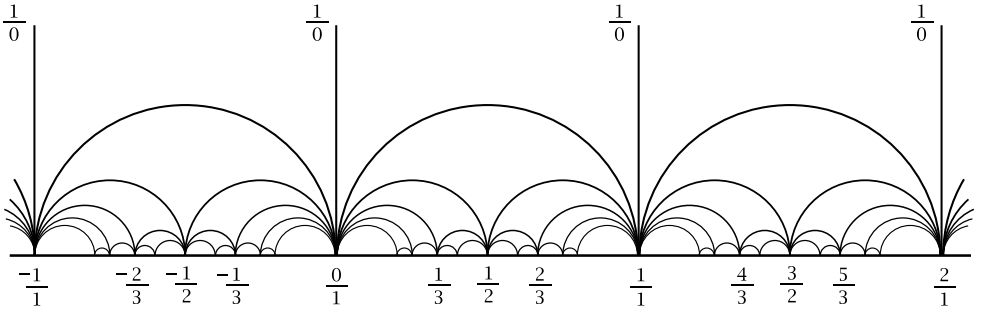


The labels in the lower half of the diagram follow the same scheme, starting with the labels $-1/0$, $0/1$, and $-1/1$ on the large triangle. Using $-1/0$ instead of $1/0$ as the label of the vertex at the far left means that we are regarding $+\infty$ and $-\infty$ as the same. The labels in the lower half of the diagram are the negatives of those in the upper half, and the labels in the left half are the reciprocals of those in the right half.

For fractions with a nonzero denominator our usual rule will be to write them with a positive denominator, so the sign of the fraction is the sign of the numerator.

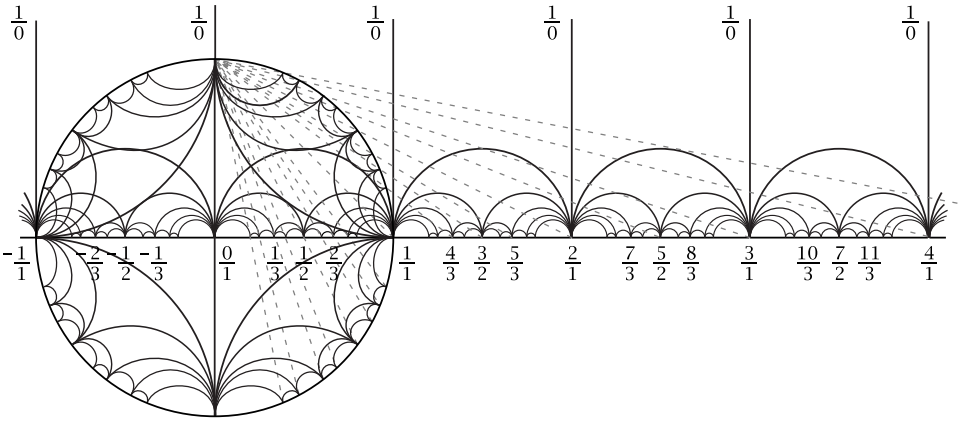
The labels generated by the mediant rule occur in their proper order around the circle, increasing from $-\infty$ to $+\infty$ as one goes around the circle in the counterclockwise direction. This is obviously true for the integer labels, and to verify it for the others it suffices to show that the mediant $\frac{a+c}{b+d}$ of a/b and c/d is always a number between a/b and c/d (hence the term "mediant"). Thus we want to show that if $a/b < c/d$ then $a/b < \frac{a+c}{b+d} < c/d$. These fractions all have positive denominators, so the inequality $a/b < c/d$ is equivalent to $ad < bc$ and $a/b < \frac{a+c}{b+d}$ is equivalent to $ab + ad < ab + bc$. Obviously $ad < bc$ implies $ab + ad < ab + bc$, so $a/b < c/d$ implies $a/b < \frac{a+c}{b+d}$. Similarly $\frac{a+c}{b+d} < c/d$ is equivalent to $ad + cd < bc + cd$ which also follows from $ad < bc$, so $a/b < c/d$ implies $\frac{a+c}{b+d} < c/d$.

There is another version of the Farey diagram with the boundary circle straightened out to a line:



Here the diagram fills up the upper half of the xy-plane, with the vertex $1/0$ of the original Farey diagram positioned "at infinity" so it is not actually shown in the new version. The edges of the diagram with one endpoint at $1/0$ are drawn as vertical lines with lower endpoints at the integer points on the x-axis. All the other edges of the diagram are semicircles with endpoints on the x-axis, and we can position these so that the vertex labeled a/b is actually the number a/b on the x-axis. This is possible since when we construct the diagram by adding more and more curvilinear triangles, we can place the new vertex of each triangle at any point between its outer two vertices, so we just choose this new vertex to be at the mediant of the outer two vertices.

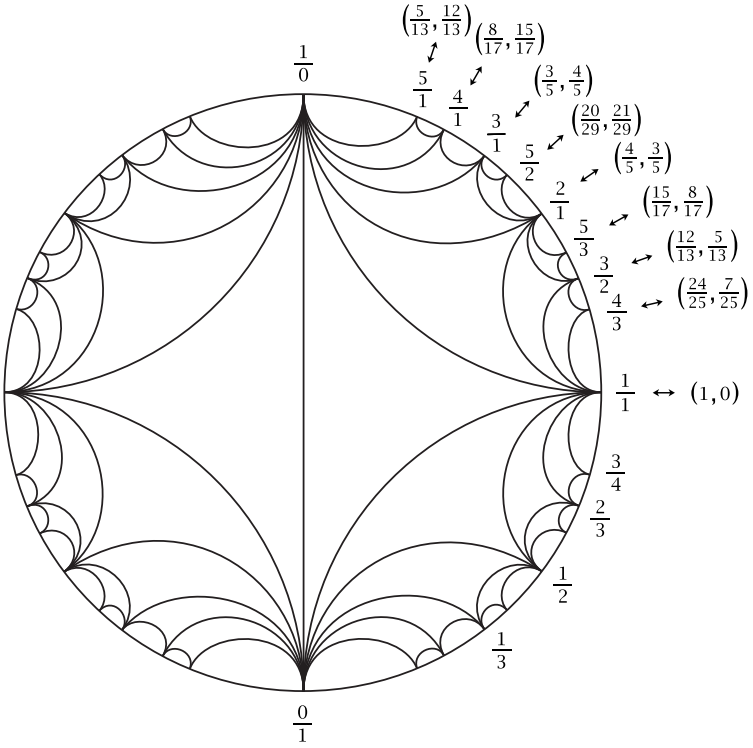
In the previous chapter we described how rational points (x, y) on the unit circle $x^2 + y^2 = 1$ correspond to rational points p/q on the x-axis by means of lines through the point $(0, 1)$ on the circle. Using this correspondence, we can label the rational points on the circle by the corresponding rational points on the x-axis and then construct a new Farey diagram in the circle by filling in triangles by the mediant rule just as before.



This gives a version of the circular Farey diagram that is rotated by 90 degrees to put $1/0$ at the top of the circle, and there are also some perturbations of the positions of the other vertices and the shapes of the triangles. For our purposes these perturbations

will not make much of a difference since it will usually be just the combinatorial pattern of the triangles that is important. We drew the circular Farey diagram the way we did at the beginning of the chapter because it looks more symmetric and is easier to draw since one does not have to figure out the exact positions of the vertices.

The next figure shows the relationship between the new circular Farey diagram and Pythagorean triples (a,b,c) using the formulas $(a,b,c) = (2pq, p^2-q^2, p^2+q^2)$ that we found in the previous chapter. The vertex with label p/q thus has coordinates $(x,y) = (a/c, b/c) = \left(\frac{2pq}{p^2+q^2}, \frac{p^2-q^2}{p^2+q^2}\right)$.

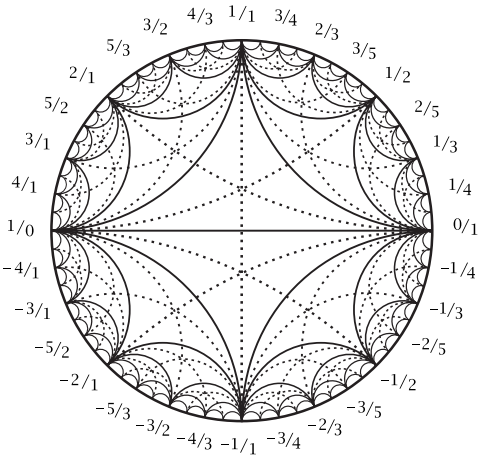


The construction we have described for the Farey diagram involves an inductive process where more and more edges and vertex labels are added in succession. With a construction like this it is not easy to tell by a simple calculation whether or not two given rational numbers a/b and c/d are joined by an edge in the diagram. Fortunately there is such a criterion:

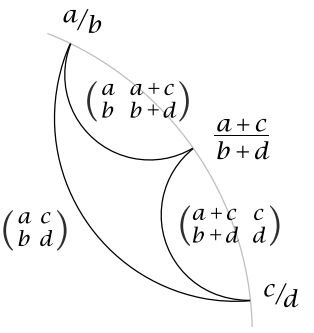
Proposition 1.1. *For each pair of fractions a/b and c/d, including $\pm 1/0$, there exists an edge in the Farey diagram with endpoints labeled a/b and c/d if and only if the determinant $ad - bc$ of the matrix $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ is equal to ± 1.*

What this means is that if one starts with the rational numbers together with $\pm 1/0$ arranged in order around a circle and one inserts circular arcs inside this circle meeting it perpendicularly and joining each pair of fractions a/b and c/d such

that $ad - bc = \pm 1$, with the circular arc replaced by a diameter in case a/b and c/d are diametrically opposite on the circle, then no two of these arcs will cross, and they will divide the interior of the circle into nonoverlapping curvilinear triangles. This is really quite remarkable when you think about it, and it does not happen for other values of the determinant besides ± 1. For example, for determinant ± 2 the edges would be the dotted arcs in the figure at the right. Here there are three arcs crossing in each triangle of the original Farey diagram, and these arcs divide each triangle of the Farey diagram into six smaller triangles.



Proof: First we show by an inductive argument that for an edge in the diagram joining two fractions a/b and c/d the associated matrix $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ has determinant ± 1. The induction starts with the edge joining $\pm 1/0$ to $0/1$ where the determinant condition obviously holds. All the other edges are added in stages, first the four edges creating the two biggest triangles, then the eight edges creating the next four triangles, and so on. Consider a triangle created at some stage by adding a new vertex labeled $\frac{a+c}{b+d}$ as the mediant of vertices a/b and c/d from an earlier stage, as in the figure at the right. We may assume by induction that $ad - bc = \pm 1$ for the long edge of the triangle which was added at an earlier stage. The determinant condition then holds also for the two shorter edges of the triangle since $a(b+d) - b(a+c) = ad - bc$ and $(a+c)d - (b+d)c = ad - bc$. Thus the determinant condition continues to hold after each stage of the construction of the diagram, so it holds for all edges.



Now we prove the converse, the statement that if $ad - bc = \pm 1$ then there is an edge in the diagram joining a/b and c/d. We may assume $b \geq 0$ and $d \geq 0$ by multiplying both numerator and denominator of either fraction by -1 if necessary, which multiplies the determinant $ad - bc$ by -1. The order of the two fractions a/b and c/d does not matter since interchanging the two columns of the matrix $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ also multiplies the determinant by -1. If b or d is 0, say $b = 0$, then the determinant condition becomes $ad = \pm 1$ so $d = 1$ and $a = \pm 1$. In this case the fractions a/b and c/d are $\pm 1/0$ and $c/1$ so they lie at the ends of an edge of the diagram, one of the vertical edges to $1/0$ in the upper halfplane version of the diagram. Thus for the rest of the proof we may assume $b > 0$ and $d > 0$.

The previous figure shows that adding a new triangle to the diagram creates two new edges corresponding to matrices obtained from $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ by replacing one of the columns by the sum of the two columns. To finish the proof we will show that for each matrix $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ of determinant ± 1 with $b > 0$ and $d > 0$ it is possible to perform a finite sequence of the inverse operations of subtracting one column from the other and end up with a matrix that we already know corresponds to an edge in the diagram. We will do this by always subtracting the column with smaller second entry from the column with larger second entry, so that these two entries remain positive. We stop the process when the two entries in the second row become equal. For example, here is how the process works for the matrix $\left(\begin{smallmatrix} 3 & 7 \\ 8 & 19 \end{smallmatrix}\right)$:

$$\begin{pmatrix} 3 & 7 \\ 8 & 19 \end{pmatrix} \to \begin{pmatrix} 3 & 4 \\ 8 & 11 \end{pmatrix} \to \begin{pmatrix} 3 & 1 \\ 8 & 3 \end{pmatrix} \to \begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix} \to \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix} \to \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

Here the last matrix corresponds to the edge joining $1/1$ and $0/1$. Reversing the steps reducing $\left(\begin{smallmatrix} 3 & 7 \\ 8 & 19 \end{smallmatrix}\right)$ to $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, we are adding one column to the other at each stage so each new matrix produced in this way corresponds to an edge of the diagram. In particular this shows that the original matrix $\left(\begin{smallmatrix} 3 & 7 \\ 8 & 19 \end{smallmatrix}\right)$ corresponds to an edge of the diagram.

For the general argument we start with a matrix $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ of determinant ± 1 with $b > 0$ and $d > 0$. If $b \neq d$ then we subtract the column with smaller second entry from the column with larger second entry, and repeat this operation until the two entries in the second row are equal. We cannot get a 0 in the second row since this would mean that the previous matrix already had equal entries in the second row. Once we get a matrix with equal entries in the second row, these entries will divide the determinant which is ± 1 so these entries must be 1. Thus the matrix is of the form $\left(\begin{smallmatrix} a & c \\ 1 & 1 \end{smallmatrix}\right)$, with determinant $a - c = \pm 1$ so a and c differ by 1. The corresponding fractions are then $n/1$ and $n+1/1$ for some integer n, and there is an edge of the diagram joining these two fractions, one of the large semicircles in the upper halfplane diagram. Hence when we reverse the sequence of column subtractions by performing a sequence of column additions, each successive matrix will correspond to an edge of the diagram and in particular $\left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ will correspond to an edge of the diagram. □

The sign of the determinant $ad - bc$ has a simple interpretation for fractions a/b and c/d with positive denominators since in this case the inequality $ad - bc > 0$ is equivalent to $a/b > c/d$ and $ad - bc < 0$ is equivalent to $a/b < c/d$. Thus the sign of the determinant tells which of a/b or c/d is larger.

Here is an interesting consequence of the preceding proposition:

Corollary 1.2. *The mediant rule for labeling the vertices in the Farey diagram always produces labels a/b that are fractions in lowest terms.*

This would follow automatically if it was always true that the mediant of two fractions in lowest terms is again in lowest terms, but this is not always the case. For

example, the mediant of $1/3$ and $2/3$ is $3/6$, and the mediant of $2/7$ and $3/8$ is $5/15$. Somehow cases like this do not occur in the Farey diagram.

Before deducing the corollary let us introduce a bit of standard terminology. For a fraction a/b to be in lowest terms means that a and b have no common factor greater than 1. This is equivalent to saying that the prime factorizations of a and b have no prime factor in common. When this is the case we say that a and b are *coprime*. An alternative terminology is to say that a and b are *relatively prime*.

Proof: From the way the Farey diagram is constructed, each labeled vertex a/b is joined to some other labeled vertex c/d by an edge of the diagram. By the easier half of Proposition 1.1 we have $ad - bc = \pm 1$. This implies that a and b are coprime since any common divisor of a and b must divide the products ad and bc, hence also the difference $ad - bc = \pm 1$, but the only divisors of ± 1 are ± 1. □

Proposition 1.1 can also be used to prove another basic fact about the Farey diagram:

Proposition 1.3. *Every fraction p/q in lowest terms occurs as the label on some vertex in the Farey diagram.*

Proof: We may assume p and q are nonzero since $0/1$ and $1/0$ certainly occur as labels in the diagram. Since the negative labels in the diagram are just the negatives of the positive labels, we can assume p and q are in fact positive. It will suffice to show that if p and q are coprime, then there is an edge in the diagram whose endpoints are labeled p/q and r/s for some integers r and s. By Proposition 1.1 this is equivalent to the existence of integers r and s such that $ps - qr = \pm 1$.

Consider a matrix $\left(\begin{smallmatrix} x & y \\ p & q \end{smallmatrix}\right)$ where the integers x and y are yet to be determined. In the proof of Proposition 1.1 there was a procedure for repeatedly subtracting the column with smaller second entry from the column with larger second entry until a matrix with equal second entries is obtained. Subtracting one column from the other does not affect coprimeness of the two second entries, so when the procedure is applied to a matrix $\left(\begin{smallmatrix} x & y \\ p & q \end{smallmatrix}\right)$ with p and q coprime, the result is a matrix whose second entries are equal and coprime, so these entries must be 1. Now let us choose a matrix of determinant ± 1 whose lower two entries are 1, say the matrix $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. If we start with this matrix and apply the reverse of the sequence of operations performed on $\left(\begin{smallmatrix} x & y \\ p & q \end{smallmatrix}\right)$ to get 1's in the second row, the resulting sequence of operations of adding one column to the other converts $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ into a matrix $\left(\begin{smallmatrix} r & s \\ p & q \end{smallmatrix}\right)$ of the same determinant ± 1. This means that we have found integers r and s such that $rq - ps = \pm 1$, or equivalently $ps - qr = \pm 1$. □

Implicit in this proof is a method for solving Diophantine equations of the form $px - qy = \pm 1$ for any two given coprime positive integers p and q. In Section 2.3 we will make this procedure explicit and streamline it to be more efficient.

Exercises

1. There is another version of the Farey diagram in which the vertex labeled p/q is placed at the point (q,p) in the plane, so p/q is the slope of the line through the origin and (q,p). The edges of this new Farey diagram are straight line segments connecting the pairs of vertices that are connected in the original Farey diagram. For example there is a triangle with vertices $(1,0)$, $(0,1)$, and $(1,1)$ corresponding to the big triangle in the upper half of the circular Farey diagram. With this model of the Farey diagram the operation of forming the mediant of two fractions just corresponds to standard vector addition $(a,b)+(c,d)=(a+c,b+d)$.

What you are asked to do in this problem is just to draw the portion of the new Farey diagram consisting of all the triangles whose vertices (q,p) satisfy $0 \le q \le 5$ and $0 \le p \le 5$. Note that since fractions p/q labeling vertices are always in lowest terms, the points (q,p) such that q and p have a common divisor greater than 1 are not vertices of the diagram.

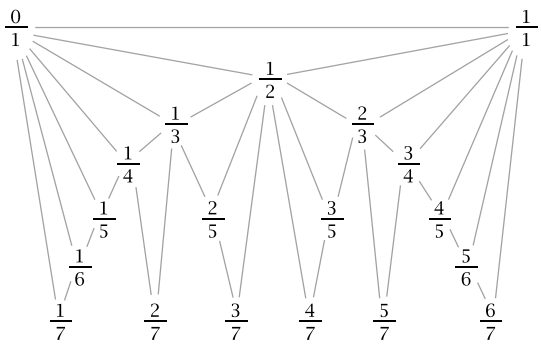
2. Consider a vertex of the Farey diagram labeled a/b with $b > 1$. Show that of all the labels on vertices connected to the a/b vertex by an edge of the diagram, exactly two have denominator smaller than b.

3. If a/b, c/d, and e/f are fractions in lowest terms such that e/f is the mediant of a/b and c/d, is it necessarily true that there is a triangle in the Farey diagram with vertices a/b, c/d, and e/f? Give either a proof or a counterexample.

4. (a) Reduce each of the matrices $\left(\begin{smallmatrix}7&3\\16&7\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}67&14\\24&5\end{smallmatrix}\right)$ to either $\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$ or $\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$ by repeatedly subtracting one column from the other as in the proof of Proposition 1.1. (b) Use Proposition 1.1 to show that this can be done for any matrix $\left(\begin{smallmatrix}a&c\\b&d\end{smallmatrix}\right)$ with non-negative entries and determinant ± 1.

1.2 Farey Series

We can build the set of rational numbers by starting with the integers and then inserting in succession the halves, thirds, fourths, fifths, sixths, and so on. Let us look at what happens if we restrict to rational numbers between 0 and 1. Starting with 0 and 1 we first insert $1/2$, then $1/3$ and $2/3$, then $1/4$ and $3/4$, skipping $2/4$ which we already have, then inserting $1/5$, $2/5$, $3/5$, and $4/5$, then $1/6$ and $5/6$, etc.



This process has an interesting property that is really quite surprising when one first sees it:

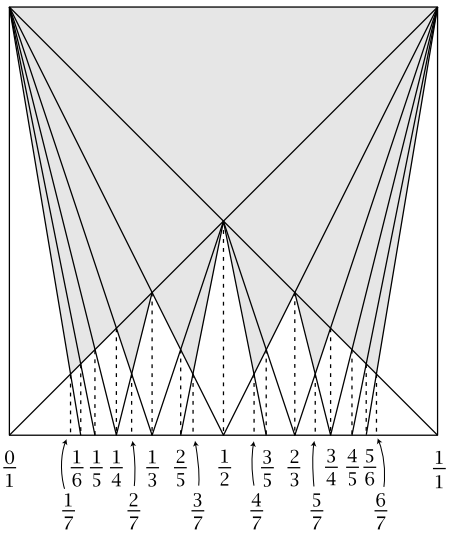
> *Each time a new number is inserted, it forms the third vertex of a triangle whose other two vertices are its two nearest neighbors among the numbers already listed, and if these two neighbors are a/b and c/d then the new vertex is exactly the mediant $\frac{a+c}{b+d}$.*

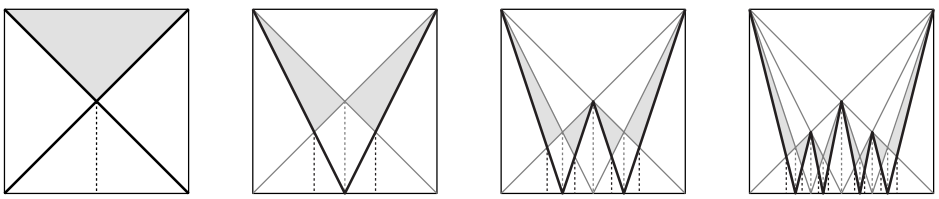
The discovery of this curious phenomenon in the early 1800s was initially attributed to a geologist and amateur mathematician named Farey, although it turned out that he was not the first person to have noticed it. In spite of this confusion, the sequence of fractions a/b between 0 and 1 with denominator less than or equal to a given number n is usually called the nth *Farey series* F_n. For example, here is F_7:

$$\frac{0}{1}\ \frac{1}{7}\ \frac{1}{6}\ \frac{1}{5}\ \frac{1}{4}\ \frac{2}{7}\ \frac{1}{3}\ \frac{2}{5}\ \frac{3}{7}\ \frac{1}{2}\ \frac{4}{7}\ \frac{3}{5}\ \frac{2}{3}\ \frac{5}{7}\ \frac{3}{4}\ \frac{4}{5}\ \frac{5}{6}\ \frac{6}{7}\ \frac{1}{1}$$

These numbers trace out the up-and-down path across the bottom of the figure above. For the next Farey series F_8 we would insert $1/8$ between $0/1$ and $1/7$, $3/8$ between $1/3$ and $2/5$, $5/8$ between $3/5$ and $2/3$, and finally $7/8$ between $6/7$ and $1/1$.

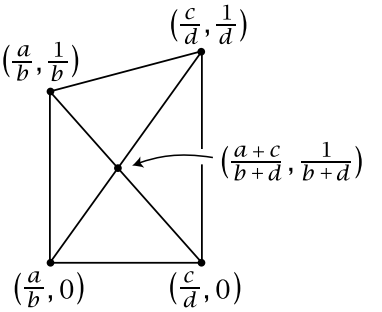
There is a cleaner way to draw the preceding diagram using straight lines in a square, as shown in the figure at the right. One can construct this diagram in stages, as indicated in the sequence of figures below. Start with a square together with its diagonals and a vertical line from their intersection point down to the bottom edge of the square. Next, connect the resulting midpoint of the lower edge of the square to the two upper corners of the square and drop vertical lines down from the two new intersection points this produces. Now add a W-shaped zigzag and drop verticals again. It should then be clear how to continue.





A nice feature of this construction is that if we start with a square whose sides have length 1 and place this square so that its bottom edge lies along the x-axis with

the lower left corner of the square at the origin, then the construction assigns labels to the vertices along the bottom edge of the square that are exactly the x-coordinates of these points. Thus the vertex labeled $1/2$ really is at the midpoint of the bottom edge of the square, and the vertices labeled $1/3$ and $2/3$ really are $1/3$ and $2/3$ of the way along this edge, and so forth. In order to verify this fact the key observation is the following: For a vertical line segment in the diagram whose lower endpoint is at the point $(a/b, 0)$ on the x-axis, the upper endpoint is at the point $(a/b, 1/b)$. This is obviously true at the first stage of the construction, and it continues to hold at each successive stage since for a quadrilateral whose four vertices have coordinates as shown in the figure at the right, the two diagonals intersect at the point $(a+c/b+d, 1/b+d)$. For example, to verify that $(a+c/b+d, 1/b+d)$ is on the upward diagonal line from $(a/b, 0)$ to $(c/d, 1/d)$ it suffices to show that the line segments from $(a/b, 0)$ to $(a+c/b+d, 1/b+d)$ and from $(a+c/b+d, 1/b+d)$ to $(c/d, 1/d)$ have the same slope. These slopes are



$$\frac{1/b+d}{a+c/b+d - a/b} = \frac{b}{b(a+c) - a(b+d)} = \frac{b}{bc - ad}$$

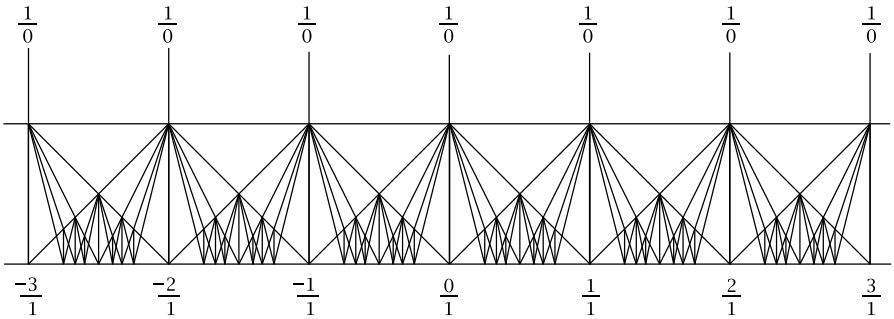
and

$$\frac{1/d - 1/b+d}{c/d - a+c/b+d} = \frac{b+d-d}{c(b+d) - d(a+c)} = \frac{b}{bc - ad}$$

so they are equal. The same argument works for the other diagonal by interchanging a/b and c/d. Note that the denominator $bc - ad$ in the slope formulas above is ± 1 since a/b and c/d are the endpoints of an edge of the Farey diagram. Thus each diagonal line in the square Farey diagram has integer slope, and this integer is, up to sign, the denominator of the rational number where the line meets the x-axis.

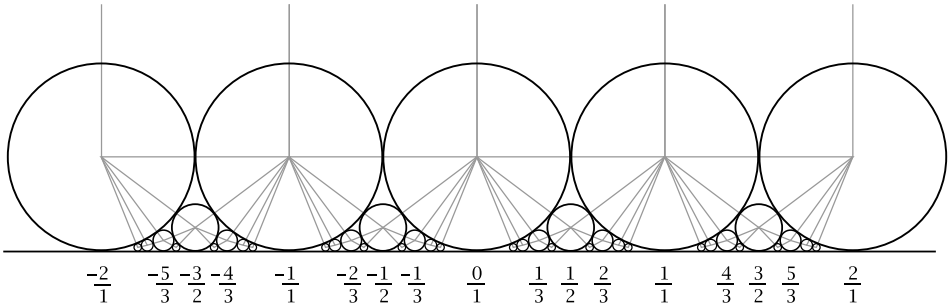
Going back to the square diagram, this fact that we have just shown implies that the successive Farey series can be obtained by taking the vertices that lie above the line $y = 1/2$, then the vertices above $y = 1/3$, then above $y = 1/4$, and so on.

We can form a linear version of the full Farey diagram by placing copies of the square side by side along the x-axis:



Here the vertical segments in the horizontal strip are not part of the resulting Farey diagram, which consists just of the triangles with nonvertical edges, along with the infinite "triangles" above the strip with a vertex at $1/0$. The original halfplane Farey diagram can be obtained from this linear Farey diagram by shrinking each vertical segment in the horizontal strip down to its lower endpoint while bending each straight edge of a triangle into a semicircle with endpoints on the x-axis.

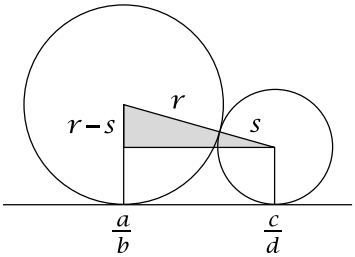
Another version of the Farey diagram can be constructed from an array of circles in the upper halfplane tangent to the x-axis and to each other as in the following figure:



This arrangement of tangent circles can be built in stages, starting with circles of diameter 1 tangent to the x-axis at the integer points. At the next stage a smaller circle is inserted in each gap between adjacent pairs of circles from the first stage. This creates new gaps, and one then puts a still smaller circle in each of these gaps. The process can then be repeated indefinitely all along the x-axis.

If we connect the centers of each pair of tangent circles by a line segment passing through the point of tangency, we obtain a pattern of triangles that is combinatorially equivalent to the pattern of triangles in the linear Farey diagram, but compressed closer to the x-axis. The vertices of these triangles are the centers of the various tangent circles, and we can label these centers by rational numbers, starting with an integer label $n/1$ at the center of the large circle tangent to the x-axis at the point n, and then labeling all the other centers by applying the mediant rule repeatedly.

The surprising thing about this construction is that the circle whose center is labeled a/b is tangent to the x-axis at exactly the point a/b on the x-axis. This can be verified as follows. For an edge of the Farey diagram with endpoints labeled a/b and c/d let us draw two circles tangent to each other and tangent to the x-axis at the points a/b and c/d. Let the radii of these two circles be r and s respectively. Note that r and s are not uniquely determined by a/b and c/d. In fact we can choose r arbitrarily and then this determines s, with s becoming small as r becomes large, and vice versa. We can find a formula



for how r and s are related by applying the Pythagorean theorem to the right triangle shown in the figure. The horizontal side of this triangle has length $|{}^{c}\!/_{d} - {}^{a}\!/_{b}|$ and the vertical side has length $|r - s|$. The condition for the two circles to be tangent is that the hypotenuse of the triangle has length $r + s$. Thus we have:

$$(r-s)^2 + \left(\frac{c}{d} - \frac{a}{b}\right)^2 = (r+s)^2$$

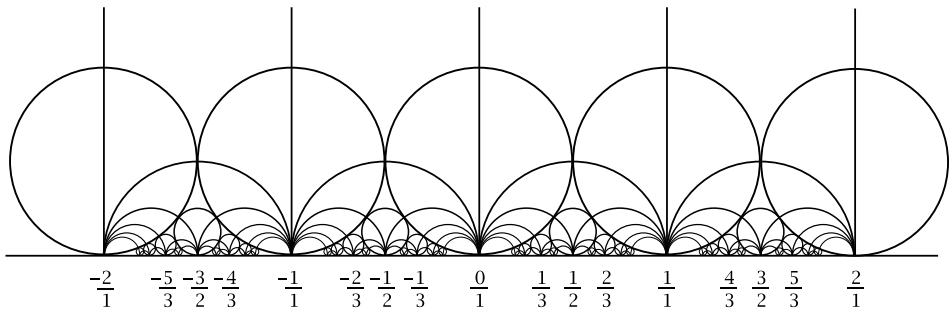
This simplifies to:

$$\left(\frac{bc-ad}{bd}\right)^2 = 4rs$$

Since we assumed the fractions ${}^{a}\!/_{b}$ and ${}^{c}\!/_{d}$ were the endpoints of an edge in the Farey diagram, we have $ad - bc = \pm 1$ so the preceding equation simplifies further to $(\frac{1}{bd})^2 = 4rs$. The easiest way to assure that this holds is to let $r = {}^{1}\!/_{2b^2}$ and $s = {}^{1}\!/_{2d^2}$, so that r depends only on ${}^{a}\!/_{b}$ and s depends only on ${}^{c}\!/_{d}$. Thus we are choosing the diameter of each circle to be the reciprocal of the square of the denominator of the fraction where the circle is tangent to the x-axis. This is consistent with how we chose the initial large circles tangent to the x-axis at integer points. Then when we build the Farey diagram inductively by adding more and more vertices labeled according to the mediant rule, each new vertex labeled ${}^{a+c}\!/_{b+d}$ between vertices labeled ${}^{a}\!/_{b}$ and ${}^{c}\!/_{d}$ is the center of a circle of diameter ${}^{1}\!/_{(b+d)^2}$ tangent to the x-axis at ${}^{a+c}\!/_{b+d}$ and tangent to each of the two circles labeled ${}^{a}\!/_{b}$ and ${}^{c}\!/_{d}$ of diameters ${}^{1}\!/_{b^2}$ and ${}^{1}\!/_{d^2}$ that are tangent to the x-axis at ${}^{a}\!/_{b}$ and ${}^{c}\!/_{d}$.

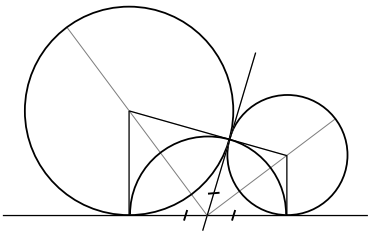
The circles tangent to the x-axis constructed in this way are called *Ford circles* after their discoverer L. R. Ford. From the formula for their diameters we see that the Ford circles whose diameter is greater than a fixed number are just the ones associated to the fractions in a Farey series, if we restrict attention to the circles tangent to the x-axis at points between 0 and 1.

Another very nice feature of Ford circles is that when we superimpose them on the upper halfplane Farey diagram, the semicircles of the Farey diagram intersect the Ford circles orthogonally at the points of tangency of the Ford circles, as shown in the following figure:



The fact that the circles and semicircles intersect orthogonally at the tangency points of the circles can be verified by considering the tangent lines to the circles at the

points where two circles are tangent. The key fact is that for any two nonparallel tangent lines to a circle, the distances from the points of tangency to the intersection point of the two tangent lines are equal. This is because reflecting across the radial line through the intersection point takes one tangent line to the other.



Exercises

1. Compute the Farey series F_{10}.

2. Draw a figure showing how Ford circles are positioned in a circular Farey diagram by the following procedure. Start with a circle C of radius 1 which will be the outer boundary of the Farey diagram. Next, draw two tangent circles of radius $1/2$ inside C and tangent to C at two opposite points of C. Label these two tangency points $1/0$ and $0/1$. Now continue drawing smaller circles inside C with the same tangency patterns as the Ford circles in the upper halfplane Farey diagram, and label the tangency points of these circles with C according to the mediant rule. After a number of these circles have been drawn, superimpose the semicircles of the Farey diagram itself.

3. In the diagram of Ford circles consider a vertical line $x = r$ for r a real number. Show that this line intersects a finite number of Ford circles if r is rational and an infinite number of Ford circles if r is irrational. Deduce that for each irrational number r there exists an infinite sequence of rational numbers p_n/q_n approaching r and approximating r in the sense that the following inequality holds for each n:

$$\left| r - \frac{p_n}{q_n} \right| < \frac{1}{2q_n^2}$$

Specifically, these are the fractions p_n/q_n labeling the circles that the line $x = r$ crosses.

4. Suppose two Ford circles tangent to the x-axis at points a/b and c/d are tangent to each other. Show that the point of tangency between the two circles is the point

$$\left(\frac{ab + cd}{b^2 + d^2}, \frac{1}{b^2 + d^2} \right)$$

so in particular the coordinates of this point are rational. *Hint*: What proportion of the way along the line segment joining the two centers is the point of tangency? This same proportion will apply to x-coordinates and y-coordinates separately.