1

A Marriage for the Ages

1.1 Mathematics and the Hawaiian Islands

We begin our journey with a surprising analogy between mathematics and the Hawaiian Islands. Each can be thought of as an archipelago. In mathematics, as in the Hawaiian Islands, some islands are large — areas such as classical algebra, geometry, topology, number theory, and so on. There are smaller ones, too, say Martingale methods in statistics. In both math and the Hawaiian Islands, the islands are not actually separate but have connections that are not always obvious. From the viewpoint of someone in an airplane looking down, Hawaii looks like a collection of land masses separate from each other. But in reality the airplane passenger is just seeing the tops of a mostly submerged mountain range, and the whole mountain range isn't seen because the water obscures the view. If we could pull a plug and let the water drain away, we would gradually see the whole mountain range come into view, and we'd discover that the islands are in fact connected. The mathematical analog of water is ignorance, as it clouds our ability to see the whole. Pulling the plug in this case corresponds to decreasing the level of ignorance, and as that happens, mathematicians discover connections — often unsuspected — between areas of math that were previously assumed to be separate.

Discovering connections between seemingly unrelated ideas has often been a basic feature of significant mathematics. For example, old-fashioned geometry and old-fashioned algebra each developed separately

for centuries, but with the birth of analytic geometry and its coordinate systems, two big islands got connected: Points became ordered pairs, and lines, circles, and conics became equations. Geometric theorems could be translated into algebra and proved using algebra's great power. The connection was a two-way road, meaning that algebra could be translated into geometry. For example, fancy polynomial equations eventually became translated into algebraic curves, giving rise to the mathematical branch of algebraic geometry. This marriage between geometry and algebra was destined to be one for the ages.

This book is about another marriage for the ages. Let's identify the actors.

Definition 1.1.1.

> A *plane algebraic curve* is the set of solutions in the real plane $\mathbb{R}^2$ of a polynomial equation $p(x, y) = 0$. If the coefficients of $p(x, y)$ are rational — or equally well, integers — the curve is called *rational*. In this book we assume any curve is rational unless stated otherwise.

Algebraic curves slink and curve around in the plane. Rational curves are intimately connected to discrete points forming the essence of number theory, and in this book we attempt to lower the water level to reveal basic connections between number theory (which is discrete), rational curves (which usually slink around), as well as some topology and even statistics.

1.2 A Model

A great idea in mathematics often has a model containing the basic core of that idea. For us, our model is one of the most famous equations in mathematics: $a^n + b^n = c^n$, where we look for solutions in positive integers a, b, c, and n. It turns out that there is one simple operation that has defined much of the recent progress in number theory. It has proved to be one of the most important steps in bringing together the two big fields of algebraic curves and number theory. Applied to our model, here's that one simple operation:

Divide both sides of $a^n + b^n = c^n$ by c^n to get $x^n + y^n = 1$, so that x and y are rational.

What happens to these pictures when we let all nonrational points drain away? That is, when we plot $x^n + y^n = 1$ in $\mathbb{Q}^2$, where $\mathbb{Q}$ stands for the set of all rational numbers? It turns out that for $n = 2$, the unit circle still looks like a unit circle. For all other n, the transformation is entirely different. Instead of a plot in the real plane looking like a curve, when n is even we see in $\mathbb{Q}^2$ only the four points $(\pm 1, 0)$ and $(0, \pm 1)$. Things get even more sparse for odd n, because then only $(1, 0)$ and $(0, 1)$ survive, a consequence of Fermat's Last Theorem:

Theorem 1.2.1 (Fermat's Last Theorem).

For integers $n > 2$, the equation $a^n + b^n = c^n$ cannot be solved using only nonzero integers a, b, and c.

Comment 1.2.2. Fermat stated his famous "Last Theorem" in 1637. He claimed he'd proved it but that the margins of his notebook were too small to contain his proof, a claim that virtually no mathematician believes today. Fermat's conjecture stubbornly resisted seemingly countless attempts at proof and became arguably the most famous unsolved problem in mathematics. Finally, after seven years of intense work, Andrew Wiles cracked it in 1994, 357 years after its original statement. His proof today stands as one of the outstanding success stories in mathematics. ◇

1.3 Appearance of the Rational Unit Circle, Part I

Everyone knows what the circle $x^2 + y^2 = 1$ looks like in the real plane $\mathbb{R}^2$. Remarkably, when we eliminate nearly all the original points and look at the solution set in only the rational plane $\mathbb{Q}^2$, what we see still looks like a circle. But what does $x^2 + y^2 = 1$ *actually* look like in $\mathbb{Q}^2$? For example, suppose we magnify this set a billion or trillion times. Will we find missing arcs of positive length — "gaps" — in it? Or, no matter how much we magnify, will such gaps never appear? We find the answer in this section.

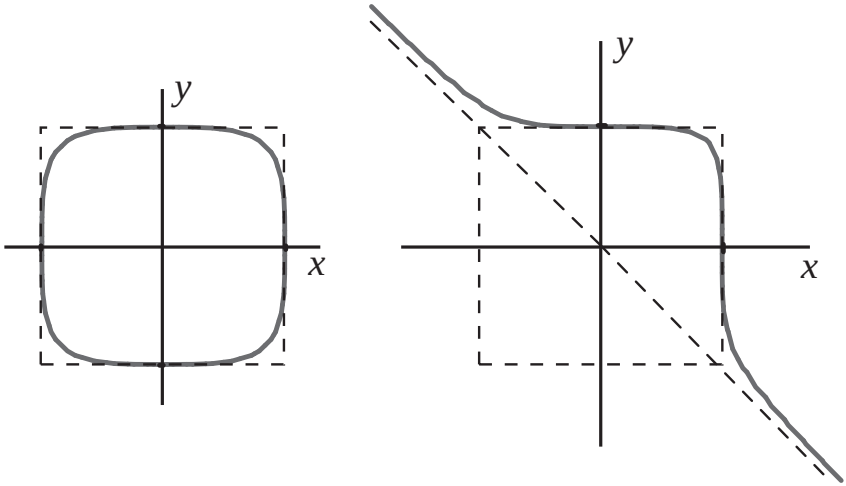

Figure 1.1. Plots in $\mathbb{R}^2$ of "Fermat curves" $x^n + y^n = 1$ illustrating their appearance for even versus odd n. The curve on the left is of $n = 4$, and on the right, $n = 5$. As n increases, each curve hugs the dashed lines more closely. When $n = 100$, for example, the curve on the left would look just like a square. When magnified enough, however, we'd see that the corners are actually rounded.

Let's begin by remembering that $x^2 + y^2 = 1$ came from the Pythagorean Theorem $a^2 + b^2 = c^2$ applied to positive integers a, b, and c, where we've divided each side by c^2 to get

$$\left(\frac{a}{c}\right)^2 + \left(\frac{b}{c}\right)^2 = 1.$$

As before, x and y in $x^2+y^2 = 1$ are the rational numbers $\frac{a}{c}$ and $\frac{b}{c}$. Some of these (a, b, c) right triangles are famous, like $(3, 4, 5)$, $(6, 8, 10)$, $(5, 12, 13)$, and $(8, 15, 17)$. There are also not-so-famous ones, like $(11, 60, 61)$, and ones that hardly anybody knows or cares about, like (4,691, 6,480, 8,161). The point here is that when we divide any of these triples by the integer c, we get rational numbers x and y satisfying $x^2 + y^2 = 1$. That is, we get points on the circle $x^2 + y^2 = 1$ in $\mathbb{Q}^2$. But under tremendous magnification, does what we see through our microscope still look like a very tiny circular arc greatly magnified?

First steps toward an answer go back at least 3,800 years to the Old Babylonian Empire in Mesopotamia. Evidence is strong that even then people knew some form of the Pythagorean Theorem. Construction of right angles was important for laying out buildings, plots of land, and countless other projects from large to small.

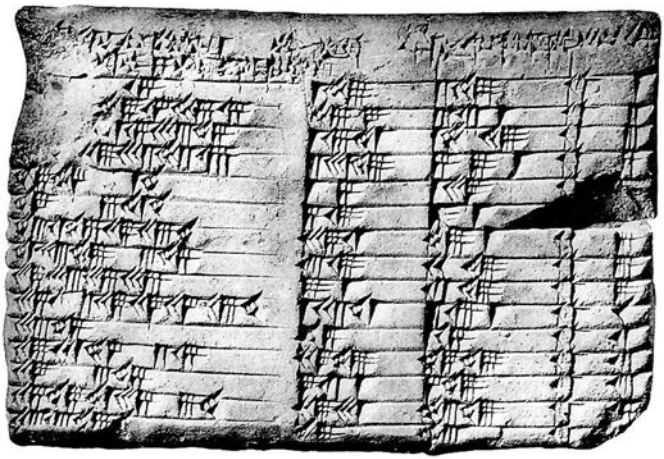

Figure 1.2. Plimpton 322. Source: Wikimedia Commons.

In a major stroke of luck, in the early 1900s, a clay cuneiform tablet was discovered in southern Iraq, and around 1922 a publisher from New York — George Plimpton — bought the tablet from an archaeological dealer for what was even then a very reasonable $10. This tablet turned out to open a rare window on some of the mathematics those ancient Mesopotamians knew. Among each of the tablet's 15 rows are integers corresponding to a, b, c in an (a, b, c) right triangle. What's so remarkable is that the tablet contains triples far removed from simple ones like $(3, 4, 5)$, $(6, 8, 10)$, $(5, 12, 13)$, or $(8, 15, 17)$ which could have been discovered by trial alone. Here, for example, are four triplets from four rows in the tablet:

- (319, 360, 481),
- (799, 960, 1,249),
- (2,291, 2,700, 3,541),
- (12,709, 13,500, 18,541).

The remarkable size of these triples has lead many historians to believe those ancients must have had some sort of formula or algorithm for generating them, perhaps even that they had figured out the gist of Euclid's algorithm. Here's one form of his algorithm:

Choose two positive integers p and q. Then, as in Figure 1.3, these integers generate the Pythagorean triple (a, b, c), with

$$\mathbf{a = |p^2 - q^2|}\,,$$

$$\mathbf{b = 2pq}\,,$$

$$\mathbf{c = p^2 + q^2}\,.$$

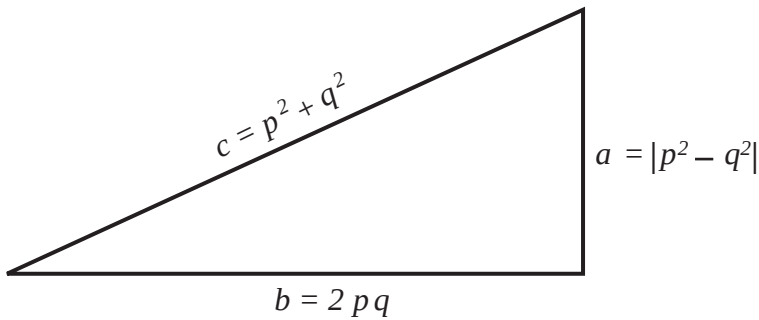

Figure 1.3. Euclid's formula.

These equations are sometimes called Euclid's formula since Euclid includes a proof in his *Elements*. From Figure 1.3 we see that

$$a^2 + b^2 = (p^4 - 2p^2q^2 + q^4) + (4p^2q^2) = p^4 + 2p^2q^2 + q^4 = c^2\,.$$

Comment 1.3.1. Much of ancient Greek mathematics was based on lengths, areas, and volumes. Negative numbers were not meaningfully addressed until the Indian mathematician Brahmagupta used them in the 7th century A.D. to represent debts. For us, however, p, q in Figure 1.3 can be chosen from the set of all integers $\mathbb{Z}$ rather than the positive ones — the natural numbers $\mathbb{N}$ — meaning the absolute sign in $|p^2 - q^2|$ is unnecessary. ◇

With the abundance of integer pairs, it's easy to generate a huge number of points $(\frac{a}{c}, \frac{b}{c})$ on the unit circle. However, this alone isn't enough to insure that in the rational plane $\mathbb{Q}^2$, the unit circle centered at the origin has no missing arcs when viewed at arbitrarily high magnifications. As an example, consider the line $y = \pi x$. If x is rational, then y isn't, so although in $\mathbb{R}^2$ the line intersects this unit circle in a point P, P isn't in $\mathbb{Q}^2$. To establish that there is no gap around P in this rational circle, we need to find a rational number m arbitrarily close to π so that $y = mx$ intersects the rational circle.

Here's the gist of a general argument. From Figure 1.3 we have

$$m = \frac{p^2 - q^2}{2pq} = \frac{1}{2}\left[\frac{p}{q} - \frac{q}{p}\right].$$

For a moment, consider $\frac{p}{q}$ to be any real number, writing $\frac{1}{2}\left[\frac{p}{q} - \frac{q}{p}\right]$ as $\frac{1}{2}\left[\frac{r}{1} - \frac{1}{r}\right]$. Now for a little magic: Let r be $\pi + \sqrt{\pi^2 + 1}$. A bit of algebra shows that $\frac{1}{2}\left[\frac{r}{1} - \frac{1}{r}\right]$ simplifies to exactly π. We can use this and continuity as a way to get our desired approximation by choosing $\frac{p}{q}$ to be rational and close to $\pi + \sqrt{\pi^2 + 1}$. A little thought shows that $\frac{1}{2}\left[\frac{p}{q} - \frac{q}{p}\right]$ will then be rational and close to π.

The following example shows the above argument in action. Begin by approximating

$$\pi + \sqrt{\pi^2 + 1} \in \mathbb{R}$$

to some number of decimal places — say,

$$3.14159 + \sqrt{3.14159 + 1} = 6.43850 = \frac{643{,}850}{100{,}000}.$$

Let p be the numerator 643,850 and q, the denominator 100,000. The slope

$$m = \frac{p^2 - q^2}{2pq} = \frac{1}{2}\left[\frac{p}{q} - \frac{q}{p}\right] = \frac{1}{2}\left[\frac{643{,}850}{100{,}000} - \frac{100{,}000}{643{,}850}\right]$$

computes to $m = 3.14159 \in \mathbb{Q}$, correct to six figures. The more places of the approximation to $\pi + \sqrt{\pi^2 + 1}$ we take, the closer our computed m will be to π.

Notice that in all this, any irrational number can be used in place of π, and we can go a bit further and say that P above can be any point of the circle not in $\mathbb{Q}^2$. One can then argue that all such P form a set in the circle with no gaps — that is, there's no missing arc of positive length in the set. In official language, this no-gaps locus in $\mathbb{Q}^2$ is *everywhere dense* in the locus in $\mathbb{R}^2$ of $x^2 + y^2 = 1$. So what we've shown above says that the rational points in the circle are everywhere dense there.

Exercise 1.3.2. Determine the fewest number of places $\pi + \sqrt{\pi^2 + 1}$ requires to get π accurate to 20 places.

1.4 Appearance of the Rational Unit Circle, Part II

The method we just met uses Euclid's formula whose roots go back to the Babylonians, some 1,500 years before Euclid. We now look at a more modern and powerful approach where, instead of a picture of the circle $x^2+y^2 = 1$ and lines passing through the origin, we pick a rational point P on the circle and consider lines passing through P. This seemingly small change turns out to pay huge dividends. Let's see how it works.

First, pick a point P on our circle. For convenience, let P be $(-1, 0)$. Then a line L with slope m through P has equation $y = m(x + 1)$. We take m to be rational, an assumption central to our new approach. The set of points on our rational circle is tremendously thinned out compared to the entire real circle, so what are the chances that L ends up intersecting the circle in some rational point Q? For an answer, plug L's equation $y = m(x + 1)$ into $x^2 + y^2 = 1$ and see what happens. The substitution gives

$$x^2 + m^2(x + 1)^2 = 1\,,$$

which can be expanded and rewritten as

$$(m^2 + 1)x^2 + 2m^2x + (m^2 - 1) = 0\,. \tag{1.1}$$

This is a quadratic equation with rational coefficients, and it's easy to check that indeed, $x = -1$ is one root of this equation, no matter what m is. We could solve for the other root, but actually we don't need to. Instead, divide the quadratic in (1.1) by the leading coefficient to make the polynomial monic. Now in any monic quadratic, the coefficient of x is the negative sum of the two roots:

$$x^2 + ax + b = (x - r)(x - s) = x^2 + (-r - s)x + rs\,.$$

In our case this coefficient is rational, as is the root $x = -1$. So the other root, whatever it may be, must be rational — that is, the x-coordinate of Q is rational. But the line L's equation is $y = m(x + 1)$, so the y-coordinate of Q must be rational, too. Therefore L does in fact intersect our thin set of rational points on the circle.

We know even more: Because both coordinates of any rational point of the circle are rational, substituting those coordinates into $y = m(x+1)$ means that the slope m is rational. We can even take $m = \infty$, corresponding to a vertical line through $(-1, 0)$. This intersects the rational circle in a double point consisting of the original point $(-1, 0)$ counted twice. We have now established that $\mathbb{Q} \cup \{\infty\}$, looked at as slopes of lines through $(-1, 0)$, naturally parameterizes via intersection all rational points of our circle, and the circle has no gaps in it.

Although we took our base point P to be $(-1, 0)$, any other rational point on our circle works, too. The algebra may be more complicated, but the overall logical arguments are the same.

Exercise 1.4.1. Redo the above argument using $P = (0, 1)$ instead of $(-1, 0)$.

1.5 What About Other Rational Circles?

We have shown in two different ways that the locus of $x^2 + y^2 = 1$ in $\mathbb{Q}^2$ actually does look like a circle — there are no missing arcs in the set, even of arbitrarily small positive length. But what about the loci in $\mathbb{Q}^2$ of $x^2 + y^2 = 2$ or $x^2 + y^2 = 3$ or $x^2 + y^2 = 4$, and so on? Are *they* circles in this same sense? Starting with any positive integer n and a rational point P on $x^2 + y^2 = n$, we can follow through the steps in the last section, and it appears that the locus of $x^2 + y^2 = n$ in $\mathbb{Q}^2$ does look like a circle with no missing arcs. But concrete examples are often great teachers, so let's take a look!

Example 1.5.1. For the locus of $x^2+y^2 = 2$ we can take $(1, 1)$ as a rational base point P, since $1^2 + 1^2 = 2$. The steps in the section just above lead to a locus in $\mathbb{Q}^2$ that is dense in the full circle — that is, it has no missing arcs of positive length in it. ◇

Example 1.5.2. What about $x^2 + y^2 = 3$? In order to follow the steps in the last section to ensure that the locus of $x^2 + y^2 = 3$ in $\mathbb{Q}^2$ looks like

a circle with no gaps, we need to first find a rational point P in its locus. It turns out we can try, try, and try some more, yet our desired rational point P remains elusive. One can run through many different candidates for a few hours, days, or years, even using a high-speed computer search program, yet the crucial rational point P continues to elude us. We may conclude that there simply isn't such a P, but of course all such testing doesn't constitute a proof. The next exercise pins down that there really is no rational point on the circle $x^2 + y^2 = 3$. It's essentially a parity (even versus odd) argument. ◇

Exercise 1.5.3. To show there's no rational point on the circle $x^2+y^2 = 3$, we begin by assuming that the nonzero integers a, b, and c have no common factor — if they do, just divide each of a, b, c by that factor. Now $a^2 + b^2 = 3c^2$ involves sums of squares. If c is even, say $2n$, then its square is $4n^2$. Therefore c^2 is 0 (mod 4).

(a) Show, similarly, that if c is odd, then its square is 1 (mod 4).

(b) When c is even, $3c^2$ is $3 \cdot 0 = 0$ (mod 4). Show that in this case, if there's a solution, a^2 and b^2 must both be odd or both even.

(c) Suppose in (b) that a^2, b^2, and c are all even. Then each of a, b, c has 2 as a factor. Why does this contradict our original assumption?

(d) Therefore suppose in (b) that a^2 and b^2 are both odd (and c is even). Show this means a^2 and b^2 are both 1 (mod 4), implying $a^2 + b^2$ is 2 (mod 4). Why is this impossible?

(e) Therefore if there's a solution, c must be odd. Show that means $3c^2$ is 3 (mod 4). For the left-hand side to be odd as well, show that $a^2 + b^2$ is 1 (mod 4). Therefore c can't be odd because if there's a solution, both sides must be the same (mod 4).

Conclusion: c is either even or odd, but in both cases we find that assuming a solution leads to a contradiction.

The next theorem is far more powerful and general than the parity argument above. It provides a reliable way of deciding whether or not there's a rational point P not only on the circle $x^2 + y^2 = 3$, but on any curve $Ax^2 + By^2 + C = 0$, where A, B, and C integers. It's due to the French mathematician Adrien-Marie Legendre (1752–1833).

Theorem 1.5.4 (Legendre's Criterion).

There is a rational point on the locus of $Ax^2 + By^2 + C = 0$ exactly when we know these things about A, B, and C:

- Not all of A, B, and C have the same sign. (When all of them are positive, or all are negative, the locus is empty.)
- ABC is squarefree. (A number is squarefree if there are no repeated factors in the number's factorization into primes.) As an example, $2 \cdot 3 \cdot 5$ is squarefree, but $3 \cdot 3 \cdot 5$ is not.
- $-AB$ is a square (mod C),
- $-AC$ is a square (mod B),
- $-BC$ is a square (mod A).

For a proof, see [Ireland, Propositions 17.3.1 and 17.3.2].

Let's use this powerful tool on $x^2 + y^2 = 3$. Here, $A = 1$, $B = 1$, and $C = -3$. Not all of these have the same sign, so the first hurdle is passed. Also, $ABC = -3$ which is squarefree, so the second condition is satisfied. Next, is $-AB$ a square (mod C)? We see that $-AB = -1$ and $C = -3$. Is -1 a square in $\mathbb{Z}_{-3} = \mathbb{Z}_3$? ($\mathbb{Z}_3$ means the integers (mod 3), which is the same as the 3-hour-clock with hours 0, 1, 2.) In $\mathbb{Z}_3$, -1 is 2 and the squares of $\mathbb{Z}_3$ are $0 \cdot 0 = 0$, $1 \cdot 1 = 1$, and $2 \cdot 2 = 4 = 1$. Well, well! 2 is not on the list, so Legendre's Criterion has already dealt a fatal blow — not all the requirements on $A = 1$, $B = 1$, and $C = -3$ are satisfied, so there is *no* rational point P in the locus of $x^2 + y^2 = 3$. Our suspicions are now solidly confirmed: $x^2 + y^2 = 3$ is the empty set in the rational plane! When we talked about choosing a rational point P in the locus and then all the logic used for $x^2 + y^2 = 1$ would apply, meaning that the locus would look like a circle with no gaps, that assumed there *is a rational point P to begin with.*

From this, we see that

The locus of $x^2 + y^2 = n$ (n a positive integer) in $\mathbb{Q}^2$ is either a set looking like a circle or we see nothing at all. There is no in-between — it's all or nothing.

Exercise 1.5.5. In plotting the locus of $x^2 + y^2 = n$ in $\mathbb{Q}^2$, determine whether we see a circle or the empty set for each of $n = 4, 5, 6, 7$.

Exercise 1.5.6. For which of $n = 3, 4, 5, 6$ does the locus of $x^2 + 2y^2 = n$ in $\mathbb{Q}^2$ look like an ellipse?

Exercise 1.5.7. For which of $n = 3, 4, 5, 6$ does the locus of $x^2 - y^2 = n$ in $\mathbb{Q}^2$ look like a hyperbola?

1.6 A Parametric Formula for the Unit Circle in $\mathbb{Q}^2$

We have just shown that the lines of rational slope m through $P = (-1, 0)$ parameterize the rational points of the unit circle $x^2 + y^2 = 1$. We can derive an explicit formula for the parametrization, and this will lead to a solution (a, b, c) to $a^2 + b^2 = c^2$ corresponding to each $m \in \mathbb{Q} \cup \{\infty\}$.

To get the parametrization, argue as we did just after (1.1) on p. 8: First divide the quadratic in (1.1) by the leading coefficient to make the polynomial monic. As we noted then, in any monic quadratic the coefficient of x is the negative sum of the two roots:

$$x^2 + ax + b = (x - r)(x - s) = x^2 + (-r - s)x + rs\,.$$

We know one root $x_1 = -1$, so the other root must satisfy

$$\frac{2m^2}{m^2 + 1} = -(x_2 - 1)\,.$$

Solving for x_2 gives

$$\frac{1 - m^2}{1 + m^2}\,.$$

Of course, one could alternatively apply the usual formula for the roots of a quadratic equation to the one in equation (1.1) and get the same answer.

To get the y-coordinate of our parametrization, substitute $\frac{1-m^2}{1+m^2}$ for x in $y = m(x + 1)$, which is the equation of the line with slope m through $P = (-1, 0)$; see Figure 1.4. Doing this and simplifying gives our desired parametrization:

$$m \to Q = \left(\frac{1 - m^2}{1 + m^2}\,,\ \frac{2m}{1 + m^2}\right). \tag{1.2}$$

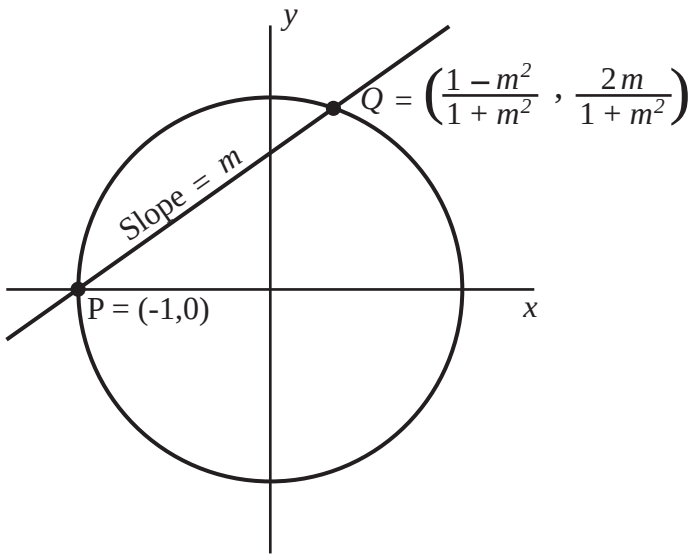

Figure 1.4. This figure shows the line $y = m(x + 1)$ through $(-1, 0)$. As $m \in \mathbb{Q}$ increases from 0 to 1, Q moves counterclockwise, covering the part of the circle in the first quadrant. As m increases from 1 to ∞, Q covers the part in the second quadrant. As m symmetrically decreases from 0 to -1, and from -1 to $-\infty$, Q moves clockwise, covering the fourth and third quadrants.

Our parametrization (1.2) leads to a solution in integers this way: Write m as a fraction $\frac{p}{q}$ of integers and substitute this directly into (1.2) to obtain explicit formulas for a, b, and c in terms of p and q. This substitution gives

$$\left(\frac{1 - \frac{p^2}{q^2}}{1 + \frac{p^2}{q^2}}, \frac{\frac{2p}{q}}{1 + \frac{p^2}{q^2}} \right),$$

which simplifies to

$$\left(\frac{-p^2 + q^2}{p^2 + q^2}, \frac{2pq}{p^2 + q^2} \right).$$

Multiplying through by p^2+q^2 gives the following solution to $a^2+b^2=c^2$ corresponding to each $m=\frac{p}{q}\in\mathbb{Q}$:

$$\begin{aligned}\mathbf{a}&=\mathbf{|p^2-q^2|}\,,\\ \mathbf{b}&=\mathbf{2pq}\,,\\ \mathbf{c}&=\mathbf{p^2+q^2}\,.\end{aligned}\tag{1.3}$$

Dividing this (a,b,c) through by the greatest common divisor of a, b, and c — written $\gcd(a,b,c)$ — yields a primitive solution, meaning a solution having no common factors other than ±1. Any solution to our problem is then obtained by multiplying each coordinate of this primitive solution by some integer n, whose sign may vary from coordinate to coordinate.

Slopes of simple fractions such as $\frac{1}{2}$, $\frac{1}{3}$, $\frac{2}{3}$, and so on yield famous right triangles, while more complicated fractions yield less famous or barely known right triangles.

Comment 1.6.1. Look again at the formulas in (1.3). Now turn to p. 6 and look at the boldfaced formulas there! These absolutely identical formulas were discovered at least 3,800 years apart by completely different methods — a remarkable testament to the timelessness of mathematics.

Example 1.6.2. With $m=\frac{p}{q}=\frac{1}{2}$, plugging $p=1$ and $q=2$ into (1.3) gives $(a,b,c)=(3,4,5)$, which is primitive. If $\frac{p}{q}$ is not in lowest terms — say, numerator and denominator share an integer factor γ — then each of a, b, and c is multiplied by γ^2. So, for example, $\frac{3}{6}$ produces $(a,b,c)=(27,36,45)$. ◇

Here are a few other examples.

Example 1.6.3. $m=\frac{1}{3}$ yields the triangle $(a,b,c)=(8,6,10)$. Since $\gcd(8,6,10)=2$, a primitive solution is $(4,3,5)$. ◇

Example 1.6.4. $m=\frac{2}{3}$ produces the triple $(a,b,c)=(5,12,13)$, which is primitive. ◇

Example 1.6.5. $m = \frac{1}{4}$ gives the solution $(a, b, c) = (15, 8, 17)$, and this is primitive. ◇

Example 1.6.6. $m = \frac{3}{4}$ yields the solution $(a, b, c) = (7, 24, 25)$, which is primitive. ◇

Example 1.6.7. $m = \frac{1}{5}$ produces $(a, b, c) = (24, 10, 26)$, with primitive solution $(12, 5, 13)$. ◇

Example 1.6.8. $m = \frac{2}{5}$ yields $(a, b, c) = (21, 20, 29)$. ◇

Here's a slightly more unusual choice:

Example 1.6.9. $m = \frac{7}{5}$ gives $(a, b, c) = (24, 70, 74)$. Since $\gcd(24, 70, 74) = 2$, a primitive solution is $(12, 25, 37)$. ◇

Exercise 1.6.10. Verify the results stated in Examples 1.6.3 through 1.6.9.

Exercise 1.6.11. Assume $m = \frac{p}{q}$ is written in lowest terms. Do finite rational slopes m and $\frac{1}{m}$ always yield the same ordered triple? Why or why not?

Exercise 1.6.12. Notice that choosing $p = 1,\ q = 2$ above yields the same shape of triangle as $p = 1,\ q = 3$. Similarly, $p = 2,\ q = 3$ yields the same shape of triangle as $p = 1,\ q = 5$. Are there similar pairs for the other examples above?

1.7 More General Quadratic Problems

We now outline a method for solving a huge swath of quadratic number theory problems. The three formulas in (1.3) lead to all integer solutions of $a^2 + b^2 = c^2$. Impressive though this is, in the world of quadratic number theory problems, $a^2+b^2 = c^2$ is pretty tame and is only one of infinitely many possibilities. What about finding formulas for all integer solutions a, b, and c to

$$a^2 + b^2 = 2c^2\,?$$

Or upping the ante and asking for formulas for

$$2a^2 + 3b^2 = 5c^2\,?$$

Or getting much wilder and asking for ones giving all integer solutions to

$$5a^2 + 8b^2 + 3ab + 11bc - 5c^2 = 0\,?$$

These seem to head toward some really serious challenges. A computer, by running through integer triples (a, b, c), may well find many specific solutions provided they exist, but what about finding *all* solutions? Or definitely establishing that there exist *no* solutions? That's quite a different matter. Notice that all these problems involve three integer variables a, b, and c, and every term has degree 2. For these problems the basic approach used above works. It's straightforward, potent, and elegant. (What more can a mathematician ask for?)

Let's make an official definition.

Definition 1.7.1. A polynomial p is *homogeneous of degree n* provided every term of the polynomial has degree n. The associated equation $p = 0$ is likewise called *homogeneous.* ◇

Now that we've had some experience in solving problems, we sharpen a bit what we said in the Preface:

> This book is about using algebraic curves over $\mathbb{Q}$ to solve homogeneous Diophantine equations of degree ≤ 3.

We've seen that when the degree is 2, then if we can find one solution, that will lead to all the rest of the solutions. To recap the method:

> Convert the homogeneous quadratic equation in integers a, b, and nonzero c to an equation in $x = \frac{a}{c}$ and $y = \frac{b}{c}$. If this equation defines a nondegenerate conic — that is, an ellipse, parabola, or hyperbola — and if there's a rational point on this conic, then any line through that point intersects the conic in a second (not necessarily distinct) point. When the line's slope is rational, the second point is also rational, and we can then go backwards from this point to an integer solution to the original equation. This leads to all its integer solutions.

If there is no rational point on the curve, then the locus in $\mathbb{Q}^2$ is empty and the original number theory problem has no integer solutions. If there's even one rational point, then there are infinitely many solutions $(x, y) \in \mathbb{Q}^2$ and therefore infinitely many integer solutions to the original number theory problem. In that case, we can work out parametrizations of x and y in terms of the rational slope m, and by writing $m = \frac{p}{q}$, we can use these parametrizations to get explicit formulas in terms of p and q for a solution (a, b, c) corresponding to m.

A good way to appreciate the power of the above method is to try finding on your own a couple of solutions to the problem below. After doing that, the solution we get will take on added significance. So using our method, let's solve this problem:

> Find all integer solutions to $2a^2 + 3b^2 = 5c^2$.

Here are the steps in the solution:

• First, transform $2a^2 + 3b^2 = 5c^2$ to an equation in x and y to get an algebraic curve. Dividing through by c^2 gives $2x^2 + 3y^2 = 5$ — that is, $\frac{2}{5}x^2 + \frac{3}{5}y^2 = 1$. This defines the ellipse depicted in Figure 1.5.

• Second, find a rational point on the ellipse; $(x, y) = (1, 1)$ does it.

• Now, between the equations of line and ellipse, eliminate y, which we can do by writing the line's equation $y-1 = m(x-1)$ as $y = mx-m+1$ and substituting this y for y in the ellipse equation $2x^2 + 3y^2 = 5$. We get

$$2x^2 + 3(mx - m + 1)^2 = 5\,.$$

By expanding and collecting terms, this can be put into the standard form $Ax^2 + Bx + C = 0$:

$$(3m^2 + 2)x^2 + 6m(1 - m)x + (3m^2 - 6m - 2) = 0\,.$$

Our challenge is to factor this to get its roots, which will then lead to the desired parametrization of (x, y) in terms of m. Since the line goes through $P = (1, 1)$, we already know that one factor is $(x - 1)$. To get the other factor, long divide the quadratic by $(x - 1)$. This division turns out to be straightforward and gives

$$(3m^2 + 2)x + (-3m^2 + 6m + 2)\,.$$

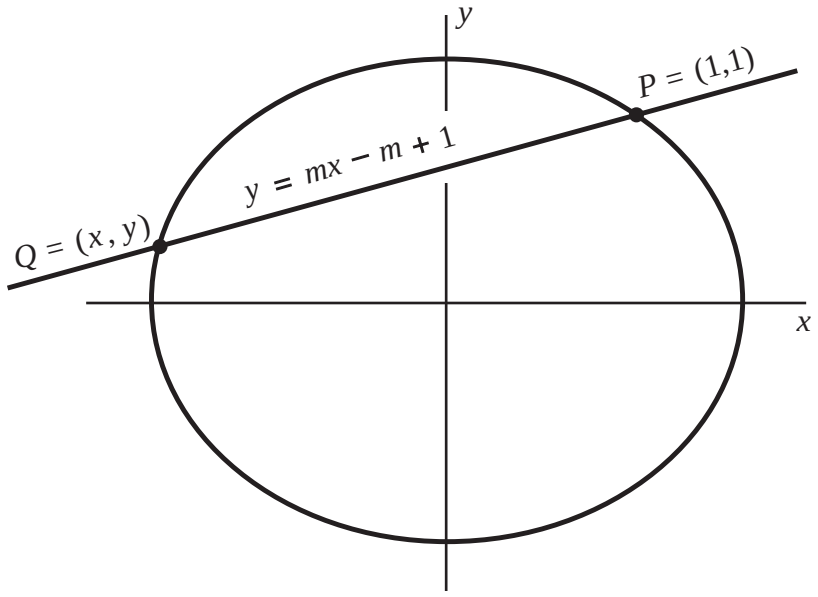

Figure 1.5. This sketch of the ellipse $2a^2 + 3b^2 = 5c^2$ shows the line $y = mx - m + 1$ through the rational point $P = (1, 1)$. As m runs through $\mathbb{Q} \cup \{\infty\}$, all the infinitely many points of the ellipse in $\mathbb{Q}^2$ are covered. When $m = \infty$, the vertical line through P intersects the ellipse in the other point $(1, -1)$. The ellipse itself has semi-major and semi-minor axes $\sqrt{\frac{5}{2}}$ and $\sqrt{\frac{5}{3}}$.

Setting this equal to zero and solving for x gives the first coordinate of our parametrization:

$$x = \frac{3m^2 - 6m - 2}{3m^2 + 2}. \tag{1.4}$$

Substituting this x into the line's equation gives the second coordinate:

$$y = \frac{-3m^2 - 4m + 2}{3m^2 + 2}. \tag{1.5}$$

We have come a long way! Equations (1.4) and (1.5) parameterize all the rational points on the ellipse, and they in turn can be used to get all integer solutions to $2a^2 + 3b^2 = 5c^2$. We can arrive at the general solution as follows.

Write m as $\frac{p}{q}$ with p, q integers, and substitute this $m = \frac{p}{q}$ into (1.4) and (1.5). After some simplification, we get for (x, y),

$$P = (x, y) = \left(\frac{3p^2 - 6pq - 2q^2}{3p^2 + 2q^2}, \frac{-3p^2 - 4pq + 2q^2}{3p^2 + 2q^2} \right).$$

Since the integers in the problem $2a^2 + 3b^2 = 5c^2$ appear only as squares, we write a solution to it using $\pm$. Our solution is, for any $m = \frac{p}{q}$,

$$\begin{aligned} \mathbf{a} &= \pm(\mathbf{3p^2 - 6pq - 2q^2}), \\ \mathbf{b} &= \pm(\mathbf{3p^2 + 4pq - 2q^2}), \\ \mathbf{c} &= \pm(\mathbf{3p^2 + 2q^2}). \end{aligned} \tag{1.6}$$

As before, divide this solution (a, b, c) by its gcd (greatest common divisor) to get a primitive solution. Then any solution corresponding to $m = \frac{p}{q}$ is obtained by multiplying the primitive one by an appropriate integer whose sign may change from coordinate to coordinate.

It's now easy to find some specific solutions corresponding to choices of m. (We choose our solutions to be positive by taking absolute values.)

Example 1.7.2. $m = \frac{1}{2}$. Substituting $p = 1$ and $q = 2$ into (1.6) gives

$a = |3 - 12 - 8| = 17,$

$b = |-3 - 8 + 8| = 3,$

$c = 3 + 8 = 11,$

so $(a, b, c) = (17, 3, 11)$, which is primitive. As a check using $(a, b, c) = (17, 3, 11)$, the left-hand side of $2a^2 + 3b^2 = 5c^2$ is $578 + 27 = 605$, and the right-hand side is $5 \cdot 121 = 605$. Agreement! Even a slope as simple as $\frac{1}{2}$ gives a nontrivial result. ◇

Example 1.7.3. $m = \frac{1}{3}$. Substituting $p = 1$ and $q = 3$ into (1.6) gives

$a = |3 - 18 - 18| = 33,$

$b = |-3 - 12 + 18| = 3,$

$c = 3 + 18 = 21.$

Since $\gcd(33, 3, 21) = 3$, $(11n, n, 7n)$ $(n \in \mathbb{Z})$ represents (up to signs) all solutions for this choice of m. As a check, substitute $(11n, n, 7n)$ into

$2a^2 + 3b^2 = 5c^2$. The left side becomes $2 \cdot 121n^2 + 3n^2 = 245n^2$, and the right-hand side becomes $5 \cdot 49n^2 = 245n^2$. Agreement. ◇

We are now witnessing just the beginning of the tremendous power of our method, and it bears repeating:

> Move from a problem in number theory to a nondegenerate conic. Check to be sure a rational solution exists, and if it does, use some algebra to parameterize the conic's points. Finally, move back to the number theory setting to get specific, easily computable results.

Let's now test-drive our creation, finding an insanely unlikely solution to $2a^2 + 3b^2 = 5c^2$.

Example 1.7.4. $m = \frac{127}{283}$. Substituting $p = 127$ and $q = 283$ into (1.6) gives

$$a = |48{,}387 - 215{,}646 - 160{,}178| = 327{,}437,$$

$$b = |-48{,}387 - 143{,}764 + 160{,}178| = 31{,}973,$$

$$c = 48{,}387 + 160{,}178 = 208{,}565.$$

This (a, b, c) happens to be primitive. Let's give our result the acid test. Write (a, b, c) as

$$(327{,}437,\ 31{,}973,\ 208{,}565)\,.$$

The left-hand side of $2a^2 + 3b^2 = 5c^2$ is

$$\begin{aligned} 2 \cdot 327{,}437^2 + 3 \cdot 31{,}973^2 &= 214{,}427{,}977{,}938 + 3{,}066{,}818{,}187 \\ &= 217{,}496{,}796{,}125\,, \end{aligned}$$

and the right-hand side is

$$5 \cdot 208{,}565^2 = 5 \cdot 43{,}499{,}359{,}225 = 217{,}496{,}796{,}125\,.$$

Those two huge computed numbers agree, so $(327{,}437,\ 31{,}973,\ 208{,}565)$ is indeed a primitive solution. All solutions corresponding to the slope $\frac{127}{283}$ are therefore $(327437n, 31973n, 208565n)$ where the sign of n may vary from coordinate to coordinate. ◇

Without formula (1.6), getting this solution would truly test human capability. We could find relatively prime integer values for p and q so large that, without (1.6), the problem would bring any computer to its knees. From just the simple case of our ellipse, algebraic curves have already shown their awesome power in number theory.

Exercise 1.7.5. Find a formula for an integer solution to $3a^2 + 4b^2 = 7c^2$ and test your solution for a few rational slopes $m = \frac{p}{q}$.

Exercise 1.7.6. Repeat the above exercise for $8a^2 - 5b^2 = 3c^2$.

Exercise 1.7.7. Repeat the above exercise for $a^2 = 6b^2 + c^2$.

1.8 Conics

In this chapter the curves we're looking at all have degree at most 2. The general second-degree equation in x and y is

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0\,, \tag{1.7}$$

where not all of A, B, and C are zero. This equation defines the general conic, and as we know, the conics we're interested in are nondegenerate — ellipses, parabolas, and hyperbolas. We now make a basic distinction.

> Until Chapter 6, the coefficients in (1.7) are assumed to be real and, most often, rational. When real, then (1.7) and the associated general conic are called *real*. If the coefficients can be chosen to be rational, then (1.7) and the conic can be called *rational*. The conics encountered in a number theory problem are rational since they arise from dividing a number theory equation by some integer, and as a consequence (1.7) ends up being rational.

Example 1.8.1. The circle $x^2 + y^2 = \pi$ is real but not rational, while the circle $\pi x^2 + \pi y^2 = \pi$ is not only real but rational since dividing through by π gives the rational form $x^2 + y^2 = 1$. ◇

In the number theory problems we've solved up to this point, the coefficients B, D, and E have all been zero, and that means our conics are

somewhat special — they're not only all "central" (that is, the center of symmetry is the origin), but the principal axes are the coordinate x- and y-axes. These are exactly the ellipses and hyperbolas plotted from equations in "standard form." Nonzero coefficients B, D, and E change all that! An xy-term always rotates the conic, and linear x- and y-terms translate it.

> The method in the last section continues to work for all quadratic problems corresponding to nondegenerate conics over the rationals. This means that even when such a homogeneous quadratic in a, b, c leads to a rational quadratic in x and y containing linear and/or mixed terms, we can still apply the approach and solve the problem.

We now illustrate our approach with two of these more general conics. The first example has tilt but no translation, and the second has both tilt and translation. In each case, simple algebra leads to the powerful solution.

Example 1.8.2. Find all integer solutions to

$$ab = c^2 .$$

Dividing through by c^2 gives $\frac{a}{c} \cdot \frac{b}{c} = 1$, or $xy = 1$. This is a (rational) hyperbola whose principal axes are tilted 45° from the (x, y)-axes. A rational point on it is $P = (-1, -1)$, so there are infinitely many solutions.

To get all solutions, begin with the line of slope m through P with equation $y + 1 = m(x + 1)$. Since y is $m(x + 1) - 1$, the x-coordinates of the two points in Figure 1.6 are the solutions of $x(mx + m - 1) = 1$ — that is, the roots of $mx^2 + (m - 1)x - 1 = 0$; the quadratic formula shows that these are $x = -1$ and $x = \frac{1}{m}$. The associated y-coordinates are -1 and m, so the points are $(-1, -1)$ and $Q = (\frac{1}{m}, m)$. Write $m = \frac{p}{q}$, so that Q becomes $\left(\frac{q}{p}, \frac{p}{q}\right)$. To move from the x, y world to integers a, b, and c, write Q over a common denominator: $\left(\frac{q^2}{pq}, \frac{p^2}{pq}\right)$. Therefore any line through $(-1, -1)$ of rational slope $m = \frac{p}{q}$ intersects the hyperbola in just one other point,

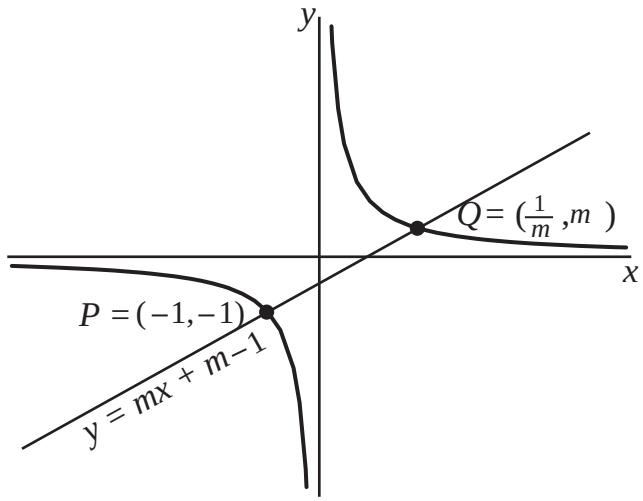

Figure 1.6. This sketch of $xy = 1$ shows the line $y = mx + m - 1$ through the rational point $(-1, -1)$. As m runs through $\mathbb{Q}$, all the points of the hyperbola in $\mathbb{Q}^2$ are covered. We'll see later on that adding points at infinity to the real plane to make it "projective," the vertical line through $(-1, -1)$ meets the hyperbola "at infinity in the projective plane."

$(x, y) = (\frac{q^2}{pq}, \frac{p^2}{pq})$. We see from Figure 1.6 that m can assume negative values, meaning that p and q can have the same or opposite signs. A little thought leads to solutions

$$\mathbf{a = q^2},$$
$$\mathbf{b = p^2},$$
$$\mathbf{c = pq}$$

and

$$\mathbf{a = -q^2},$$
$$\mathbf{b = -p^2},$$
$$\mathbf{c = pq}.$$

Dividing each solution (a, b, c) by $\gcd(a, b, c)$ gives a primitive solution for each m, with any other solution being an appropriate integer multiple of a primitive one. ◇

Example 1.8.3. Take $p = 3$ and $q = 5$. Then, for example,

$$a = q^2 = 25\,,$$

$$b = p^2 = 9\,,$$

$$c = pq = 15\,,$$

which is a primitive solution. Then $ab = c^2$ becomes $25 \cdot 9 = 15^2$, so both sides are 225. ◇

Example 1.8.4. Let $p = 7$ and $q = 11$. Then

$$a = q^2 = 121\,,$$

$$b = p^2 = 49\,,$$

$$c = pq = 77\,,$$

a primitive solution. In this case $ab = c^2$ is $121 \cdot 49 = 77^2$, so both sides are 5,929. ◇

Let's now showcase our method using an example with both tilt and translation.

Example 1.8.5. Find all integer solutions to

$$a^2 - 2ab + 3b^2 + 2ac - 2bc = 0\,.$$

Dividing through by c^2 gives $x^2 - 2xy + 3y^2 + 2x - 2y = 0$. This is a tilted ellipse going through $(0, 0)$, meaning there are infinitely many solutions (and that the algebra is simpler). This ellipse and a general line through $(0, 0)$ are depicted in Figure 1.7.

The algebra is especially simple because the general line through $(0, 0)$ has equation $y = mx$. Substituting this into $x^2 - 2xy + 3y^2 + 2x - 2y = 0$ gives

$$x^2 - 2mx^2 + 3m^2x^2 + 2x - 2mx = 0$$

which, after factoring the left-hand side, becomes

$$x[(3m^2 - 2m + 1)x - 2(m - 1)] = 0\,.$$

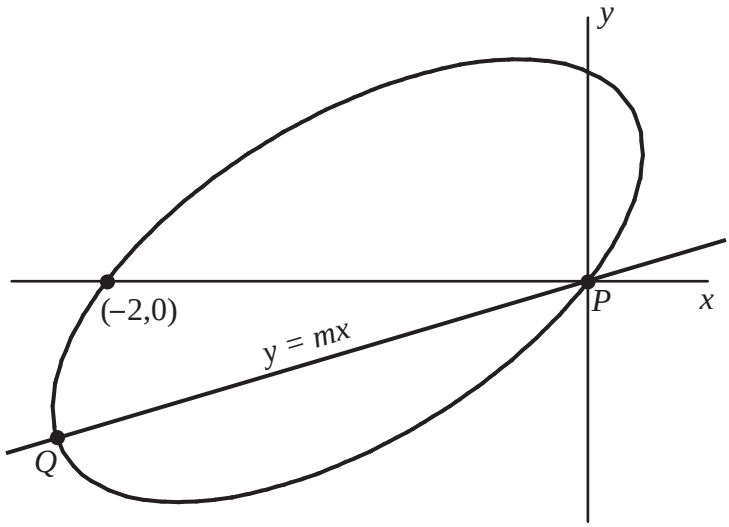

Figure 1.7. The line $y = mx$ through the origin intersects all other points Q of the rational ellipse as m runs through $\mathbb{Q} \cup \{\infty\}$. When $m = \infty$, the vertical line through the origin intersects the ellipse at $Q = (0, \frac{2}{3})$.

Set the factor in square brackets equal to zero and solve for x. This is x in the parametrization (x, y). Since y is mx, we get this for the parametrization of the ellipse:

$$(x, y) = \left(\frac{2(m-1)}{3m^2 - 2m + 1}, \frac{2m(m-1)}{3m^2 - 2m + 1} \right).$$

Writing $m = \frac{p}{q}$ and simplifying gives

$$(x, y) = \left(\frac{2(pq - q^2)}{3p^2 - 2pq + q^2}, \frac{2(p^2 - pq)}{3p^2 - 2pq + q^2} \right).$$

Corresponding to $m = \frac{p}{q}$, a solution to

$$a^2 - 2ab + 3b^2 + 2ac - 2bc = 0$$

is therefore

$$\begin{aligned} \mathbf{a} &= \mathbf{2(pq - q^2)}, \\ \mathbf{b} &= \mathbf{2(p^2 - pq)}, \\ \mathbf{c} &= \mathbf{3p^2 - 2pq + q^2}. \end{aligned} \qquad \diamond$$

Example 1.8.6. Choose $p = 1$ and $q = 5$. This gives $(a, b, c) = (-40, -8, 18)$, so

$$a^2 - 2ab + 3b^2 + 2ac - 2bc$$

becomes

$$\begin{aligned} &40^2 - 2 \cdot 40 \cdot 8 + 3 \cdot 8^2 + 2 \cdot (-40) \cdot 18 - 2 \cdot (-8) \cdot 18 \\ &\quad = 1{,}600 - 640 + 192 - 1{,}440 + 288, \end{aligned}$$

which sums to zero, as it should. A primitive solution is $(20, 4, -9)$. $\diamond$

Exercise 1.8.7. As above, find nontrivial integer solutions to

$$ab = 5c^2$$

in terms of p and q.

Exercise 1.8.8. Repeat the above exercise for

$$a^2 - 2ab + 3b^2 - 2ac + 2bc = 0.$$

1.9 Our Method Also Works in Reverse

One of the beautiful things about our method of converting a number theory problem into a geometric problem about an algebraic curve is that it works in reverse, too. That is, if you have a particular algebraic curve in mind, then as long as its defining polynomial has rational coefficients, you can see what number theory problem that curve corresponds to. (As usual, we avoid degeneracy, so a curve defined by a polynomial of degree 2 is assumed to be an ellipse, parabola, or hyperbola. In the next chapter, we will see that nondegeneracy is also important in degree 3.)

Example 1.9.1. Suppose we have a parabola such as $y = x^2$ but rotated about the origin by, say, 45°, and we'd like to find what Diophantine problem that rotated parabola corresponds to. We first need the parabola's

equation. To get it, observe that we can rotate around the origin any algebraic curve by θ if we feed the coordinates (x, y) of its defining polynomial $p(x, y)$ into the rotation matrix

$$R(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

In concrete terms, replace $p(x, y)$ by $p(X, Y)$ where $(X, Y) = (x, y)R(\theta)$. Then $p(X, Y)$ defines the same object as $p(x, y)$, but rotated about the origin by θ. Of course a pure magnification about the origin by a factor r doesn't change the amount θ of rotation, and such magnifying can sometimes keep both X and Y rational. That's just what we need of our new polynomial. It's straightforward to verify that when θ is 45° and $r = \sqrt{2}$, then

$$rR(\theta) = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

Thus $(X, Y) = (x, y)rR(\theta)$ is $(X, Y) = (x - y, x + y)$, so $p(x, y) = y - x^2$ gets replaced by

$$p(X, Y) = x + y - (x - y)^2.$$

Expanding this and setting it equal to zero gives

$$x^2 - 2xy + y^2 - x - y = 0,$$

so it's in the general form of equation (1.7) on p. 21 and we see the mixed term xy which tilts the conic. Also notice that interchanging x and y in $x^2 - 2xy + y^2 - x - y$ leaves the equation unchanged, meaning that the new parabola is symmetric with respect to the line $y = x$. This parabola is depicted in Figure 1.8.

What number theory problem does this parabola correspond to? Replacing x by $\frac{a}{c}$ and y by $\frac{b}{c}$ in $x^2 - 2xy + y^2 - x - y = 0$ and clearing denominators yields

$$a^2 - 2ab + b^2 = ac + bc.$$

Let's now solve this Diophantine equation generated by the tilted parabola. Substituting $y = mx$ into $x^2 - 2xy + y^2 - x - y = 0$ gives

$$x^2 - 2mx^2 + m^2x^2 - x - mx = 0$$

which factors into $x[(m^2 - 2m + 1)x - (m + 1)]$. Setting the second factor equal to 0 and solving for x gives $x = \frac{m+1}{m^2-2m+1}$ so, with $y = mx$, the

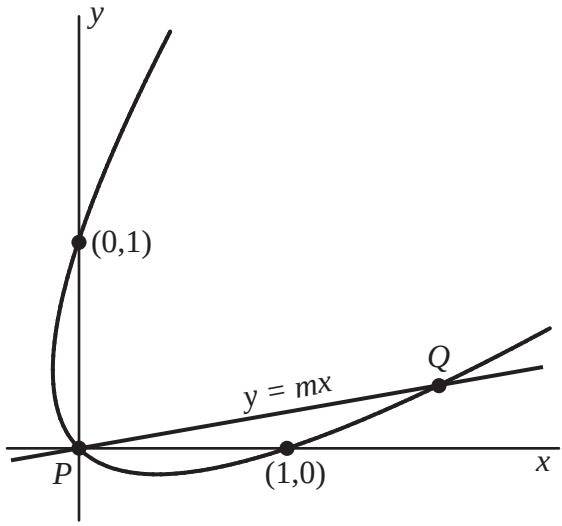

Figure 1.8. The line $y = mx$ through the origin intersects all other rational points of the parabola as m runs through $\mathbb{Q} \cup \{\infty\}$. When $m = \infty$, the vertical line through the origin intersects the parabola at $(0, 1)$.

parametrization is

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

Now substitute $m = \frac{p}{q}$ and simplify to arrive at this solution to the number theory problem arising from the tilted parabola:

$$\begin{aligned} \mathbf{a} &= \mathbf{q^2 + pq}, \\ \mathbf{b} &= \mathbf{p^2 + pq}, \\ \mathbf{c} &= \mathbf{(p - q)^2}. \end{aligned} \qquad \diamond$$

Exercise 1.9.2. In the Example 1.9.1, find a primitive solution corresponding to the slope $\frac{3}{4}$.

Exercise 1.9.3. In this same example, find a primitive solution corresponding to the slope $\frac{2}{5}$.

Exercise 1.9.4. Find the slope producing the solution $(a, b, c) = (6, 3, 1)$.

Exercise 1.9.5. Find the slope producing the solution $(a, b, c) = (15, 10, 1)$.

Exercise 1.9.6. In a way similar to what was used to tilt the parabola above, a rotation/expansion can also "untilt" a curve. This can be done to the Fermat curve $x^3 + y^3 = 1$ (which looks much like the right picture in Figure 1.1 on p. 4) to make the asymptote vertical. Find two matrices that rotate by $-45°$ and $135°$ and appropriately expand so one curve has rational points at $(0.5, \pm 0.5)$ and the other at $(-0.5, \pm 0.5)$.

1.10 Surveying All Solutions

Suppose a homogeneous quadratic in a, b, and c defines a nondegenerate conic, and suppose the conic has a rational point P. We've seen that any other rational point on the conic comes from selecting some line of slope $m = \frac{p}{q} \in \mathbb{Q} \cup \{\infty\}$ through P. Formulas for the two coordinates of Q in terms of p and q then lead to an integer-valued quadratic formula for each of a, b, c.

In a sense, this gives only a local view of the full answer to finding all integer solutions to the Diophantine problem. That is, in our examples so far, we've seen solutions corresponding to only one slope at a time. It's fair to ask for a broader perspective, a perspective giving an idea of what the full set of solutions looks like.

To do this, let's start by remembering that $\mathbb{Q} \cup \{\infty\}$ is countable, meaning we can list all fractions (written in lowest terms, say) using the natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$. For each member of this list — each slope — we create an integer triple (a, b, c) satisfying the Diophantine equation. Now as we've seen in examples above, the triple may have a greatest common divisor (gcd) other than ± 1, and by dividing the triple by its gcd, we arrive at a *primitive* integer triple. Now form two columns, the first being the slopes — essentially the m_n, with $n \in \mathbb{N}$ — and the second consisting of the corresponding primitive triples. Now create a third, fourth, fifth, ... column by successively multiplying each primitive triple by 2, 3, 4, ... We end up with a big $\mathbb{N} \times \mathbb{N}$ table whose entries beyond the first column are integer triples. Finally, we may further augment each triple by including triples with different signs of a, b, c, as appropriate.

This table then gives a view of the entire countable set of integer solutions to a quadratic Diophantine equation. Ignoring sign changes, the

solutions $(a,\ b,\ c)$, where each of a, b, and c is a quadratic in p and q, mean that the big table's typical entry is

$$\frac{(na, nb, nc)}{\gcd(a, b, c)}, \quad \text{where } n \in \mathbb{N}.$$

In this book, our primary goal is finding rational points on a plane rational algebraic curve coming from the given homogeneous Diophantine problem. Getting these rational points (x, y) constitutes the real meat of the problem; the transition from there to the corresponding primitive integer triples is, as we've seen in examples so far, reliably straightforward. For this reason:

> For the remainder of this book, any Diophantine problem will be posed as finding the rational points on a plane rational algebraic curve.

1.11 The Discriminant

Suppose the two-variable real quadratic equation

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

defines an ellipse, parabola, or hyperbola. For an easy way to tell which is which, look at the discriminant $B^2 - 4AC$. (See [Kendig 1, Chapter 9].)

- If $B^2 - 4AC > 0$, then the conic is a hyperbola.
- If $B^2 - 4AC = 0$, then the conic is a parabola.
- If $B^2 - 4AC < 0$, then the conic is an ellipse.

If we vary any or all of A, B, C, both the curve and the discriminant will likely vary in response, and if the varying is done continously and the discriminant changes sign, its value will pass through zero and we'll see the conic become a parabola or degenerate as the curves morph between ellipse and hyperbola.

The next two figures illustrate what this can look like for the parabola $x^2 - 2xy + y^2 - x - y = 0$ considered in Section 1.9. Let the coefficient of the mixed term vary — that is, let $B \in \mathbb{R}$ vary in

$$x^2 - Bxy + y^2 - x - y = 0.$$

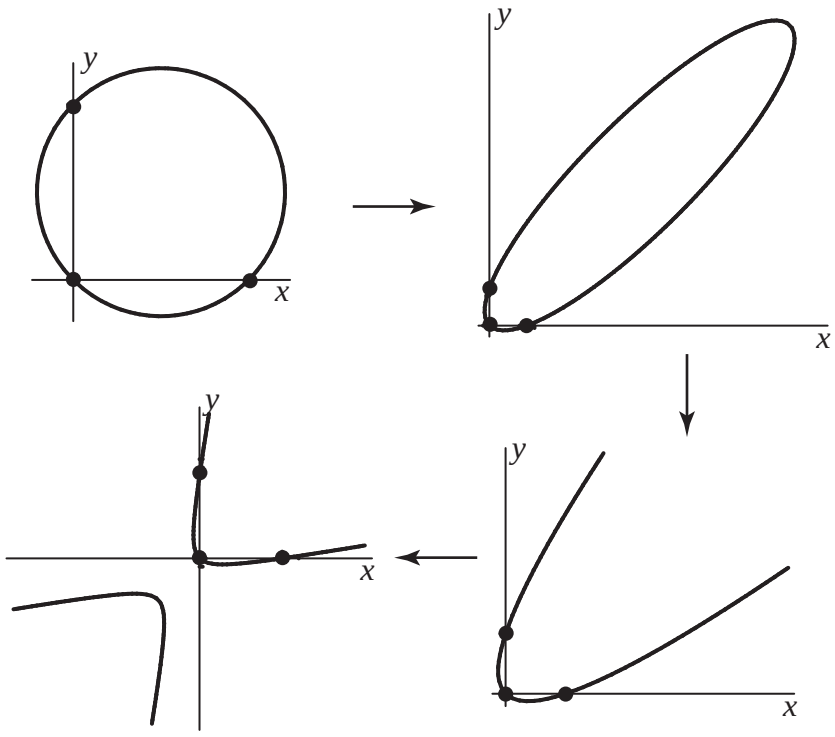

Figure 1.9. Reading clockwise from the top left sketch, this figure suggests how increasing B from zero in $x^2 - Bxy + y^2 - x - y = 0$ can lead from a circle to ellipses, then to a parabola, and on to hyperbolas.

The discriminant is $B^2 - 4$, and Figure 1.9 shows the effect of increasing B from 0, which stretches the conic. The top left sketch is of a circle with $B = 0$ and no stretching. The discriminant is -4, and because there's no xy-term, there's no tilting. As soon as B becomes nonzero, a tilt appears, as we see in the stretched-out ellipse. Encountering the sketches clockwise, we next see in the bottom right the conic corresponding to $B = 2$, which is our parabola with $B^2 - 4AC = 0$. In the next chapter we'll learn about points and the "line at infinity," and this will shed light on just how intimately ellipses, parabolas, and hyperbolas are related. For example, the parabola can be looked at as an ellipse stretched out so much that

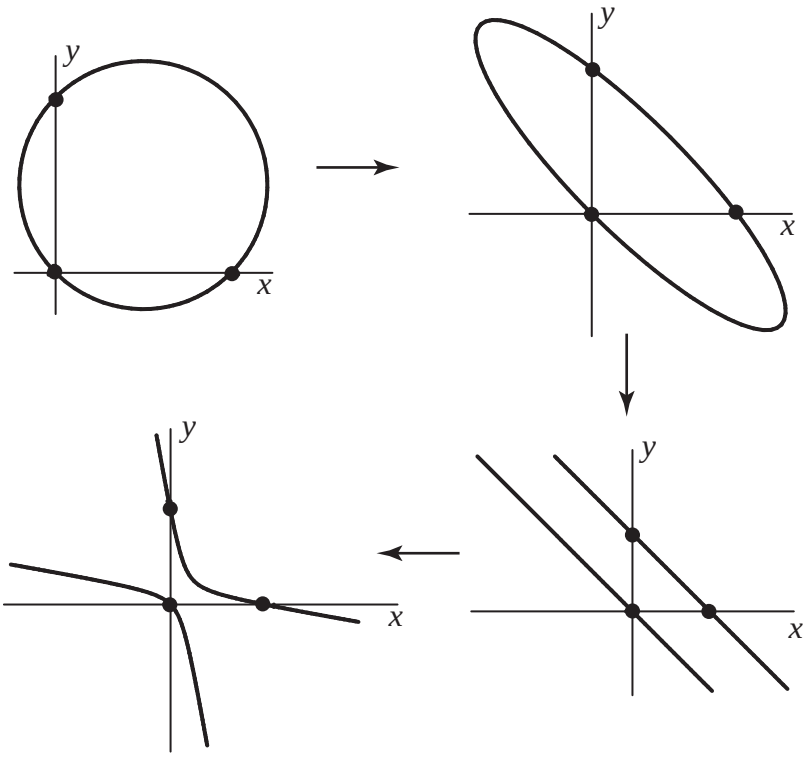

Figure 1.10. Reading clockwise from the top left sketch, this figure shows that decreasing B from zero in $x^2 - Bxy + y^2 - x - y = 0$ can in a different way lead from a circle to ellipses, to a degenerate conic of two parallel lines, and then to hyperbolas.

a point on it has actually moved onto the line at infinity. In the bottom left sketch, $B > 0$, the discriminant is positive, and the conic is a hyperbola. This value of B is large enough to have forced the ellipse over the line at infinity, and the part that crossed over appears as the branch of the hyperbola seen in the third quadrant. Notice that in all this morphing, the conic always passes through the same three points: $(0, 0)$, $(1, 0)$, and $(0, 1)$. Figure 1.10 similarly shows the effect of decreasing B from $B = 0$. When $B = -2$, the discriminant is once again zero, and the corresponding curve is degenerate — two parallel lines. Notice that again, during all

this morphing, the conic continues to pass through the three points $(0, 0)$, $(1, 0)$, and $(0, 1)$.

Comment 1.11.1. Although Figures 1.9 and 1.10 are informative, continuously animating the morphing is better than a print version. This can be done using GeoGebra, Maple, or Mathematica. For information and coding for this, see Appendix C. ◇

Exercise 1.11.2. Animate the sequence of conics $Ax^2 - 2xy + Ay^2 = 1$ as A varies through nonnegative real numbers.

1.12 Finding a Rational Point

So far, it's not been a problem to find a rational point on those conics that have one. In fact, for convenience we've sometimes chosen the conic to pass through the origin. But what if we are given a more complicated equation like, say, $-9a^2 + 7b^2 + ac = 41c^2$? To use our potent method, we need to actually find a rational point on the curve. We might try different rational values for x, but after solving for y we could well find that the answer involves square roots. Such an approach could lead to a lot of hard work with only a limited chance of success. There must be a better way.

As it turns out, there is a better way. In Appendix C we present code that searches for rational points, and this works well provided the coefficients of $p(a, b, c)$ aren't too large, say between -30 and 30. This code creates a list of all integer triples (a, b, c) in the range $-10 \le a, b, c \le 10$ satisfying $p(a, b, c) = 0$.

As an interesting test case, enter $a^2 + b^2 = 3c^2$ into the program and choose k to be 1,000, say. After returning from a leisurely lunch, we see that the computer has printed out nothing new. This is as it should be, because in Section 1.5 we found two different ways of assuring us that the corresponding circle $x^2+y^2 = 3$ has no rational points. Our computer run strongly suggests this, and we could effortlessly run a multitude of polynomials having relatively small coefficients and (likely) determine whether or not the curves contain rational points.

For larger coefficients, a method lurks in Legendre's proof of his criterion and a modern-day improvement in Legendre's approach works for really large coefficients, say 50 or 100 digits long. For integer coefficients like this, the important, much more modern paper of Cremona and Rusin

(see [Cremona]) greatly improves on Legendre's approach. This method rests on Legendre's "method of infinite descent" which Legendre himself used to prove his criterion stated on p. 11. In addition to that approach, there's a concrete algorithm for finding a rational point on a rational ellipse, parabola, or hyperbola. This is the Hasse-Minkowski Theorem, which reduces the problem to one of finding a point over $\mathbb{R}$ and over "m-hour clocks" (the integers $0, \ldots, m-1$ with addition as on an ordinary clock, but having m hours), for all positive integers m. It is beyond the scope of this book to explore these approaches, but for those with sufficient background, [Aitken] gives a good account of the first method, and J. P. Serre covers the Hasse-Minkowski Theorem in [Serre, Chapter IV].

1.13 What About Degree 1?

Our journey so far has focused on finding all integer solutions to second-degree homogeneous polynomial equations over $\mathbb{Z}$, and what we've done in this book has been a tale of success. *But what about the degree-one analog* — finding all integer solutions to degree-one homogeneous equations over $\mathbb{Z}$? The good news is that our general roadmap continues to hold. The problem is simpler, but nonetheless not trivial. There are many analogies between degree-two and degree-one problems. The basic degree-one problem takes this form:

> A degree-one equation homogeneous in the integer variables a, b, and c can be written $Aa + Bb = Cc$, with given integer coefficients A, B, C, and a, b, c being the integers we're solving for.

An example is finding all integer triples (a, b, c) satisfying $3a + 6b = 11c$. Now in the degree-two case, the first thing we did was restrict the associated curve to be nondegenerate, so that it is either an ellipse, parabola, or hyperbola, but not, say, two lines or a single point. In the degree-one case, we can write the associated line as $Ax + By = C$ (assuming c in $Aa + Bb = Cc$ isn't zero). Any real line is nondegenerate, so we needn't worry about degeneracy. Here, not both A and B are zero, so let's say $B \neq 0$. The method in the degree-two case applied to the degree-one setting tells us to find all rational points on this line. Now $Ax + By = C$ can

be solved for y:

$$By = C - Ax \Longrightarrow y = \frac{C}{B} - \frac{A}{B}x\,,$$

so a general point on the line is $\left(x, \frac{C}{B} - \frac{A}{B}x\right)$. This point is rational if and only if x is rational, so x is serving as a parameter for the rational points on the line. In this respect, x plays a role analogous to $m = \frac{p}{q} \in \mathbb{Q}$, so let's correspondingly write $x = \frac{p}{q} \in \mathbb{Q}$. We can now move from the geometric to the number-theoretic setting by substituting $x = \frac{p}{q}$ into

$$\left(x, \frac{C}{B} - \frac{A}{B}x\right)$$

to get

$$\left(\frac{p}{q}, \frac{C}{B} - \frac{A}{B}\frac{p}{q}\right)$$

and then multiply through to make the entries integers. Here, Bq does the job, leading to

$$a = Bp \quad \text{and} \quad b = Cq - Ap\,.$$

Then $Aa + Bb = Cc$ means that

$$c = \frac{Aa + Bb}{C} = \frac{ABp + BCq - ABp}{C} = Bq\,.$$

Assuming $B \neq 0$, we have

$$\begin{aligned} \mathbf{a} &= \mathbf{Bp}\,, \\ \mathbf{b} &= \mathbf{Cq} - \mathbf{Ap}\,, \\ \mathbf{c} &= \mathbf{Bq}\,. \end{aligned} \tag{1.8}$$

Example 1.13.1. Choose a degree-one Diophantine equation, say the one above, $3a + 6b = 11c$. Out of the infinitely many slopes $\frac{p}{q}$, randomly choose one, say $\frac{3}{5}$. Let's verify that the solution in (1.8) applied to this choice actually works. We have $A = 3$, $B = 6$, $C = 11$, $p = 3$, and $q = 5$. Then our general solution yields

$$a = 6 \cdot 3 = 18, \quad b = 11 \cdot 5 - 3 \cdot 3 = 55 - 9 = 46, \quad c = 6 \cdot 5 = 30\,.$$

So the triple (a, b, c) is $(18, 46, 30)$. Its gcd is 2, so $(9, 23, 15)$ is a primitive solution. Is it a solution to $3a + 6b = 11c$? Let's see:

$$3 \cdot 9 + 6 \cdot 23 = 27 + 138 = 165 = 11 \cdot 15\,.$$

So our primitive solution is verified in this case of $m = \frac{3}{5}$. The general solution, up to coordinatewise signs, is for this m

$$(9n, 23n, 15n), \quad n \in \mathbb{Z}. \qquad \diamond$$

Exercise 1.13.2. Verify that when $B \neq 0$, the solution in (1.8) in fact satisfies the general problem. What's the story when either one of p or q is zero?

Exercise 1.13.3. Derive a companion solution to (1.8), assuming $A \neq 0$.

Exercise 1.13.4. Use appropriate search code to find an integral solution to $3a + 6b = 11c$ for which $abc \neq 0$ and $a + b + c = 3$. (See Appendix C.)

Exercise 1.13.5. Find an integral solution to $3a+6b = 11c$ so that $abc \neq 0$ and $|a| + |b| + |c|$ is as small as possible.