2.1 Topographs

Finding Pythagorean triples is answering the question, *When is the sum of two squares equal to a square?* More generally one can ask, *Exactly which numbers are sums of two squares?* In other words, when does an equation $x^2 + y^2 = n$ have integer solutions, and how can one find these solutions? The brute force approach of simply plugging in values for $x$ and $y$ leads to the following list of all solutions for $n \leq 50$ (apart from interchanging $x$ and $y$):

$1 = 1^2 + 0^2$, $2 = 1^2 + 1^2$, $4 = 2^2 + 0^2$, $5 = 2^2 + 1^2$, $8 = 2^2 + 2^2$, $9 = 3^2 + 0^2$, $10 = 3^2 + 1^2$, $13 = 3^2 + 2^2$, $16 = 4^2 + 0^2$, $17 = 4^2 + 1^2$, $18 = 3^2 + 3^2$, $20 = 4^2 + 2^2$, $25 = 5^2 + 0^2 = 4^2 + 3^2$, $26 = 5^2 + 1^2$, $29 = 5^2 + 2^2$, $32 = 4^2 + 4^2$, $34 = 5^2 + 3^2$, $36 = 6^2 + 0^2$, $37 = 6^2 + 1^2$, $40 = 6^2 + 2^2$, $41 = 5^2 + 4^2$, $45 = 6^2 + 3^2$, $49 = 7^2 + 0^2$, $50 = 5^2 + 5^2 = 7^2 + 1^2$.

Notice that in some cases there is more than one solution for a given value of $n$. Our first goal will be to describe a more efficient way to find the integer solutions of $x^2 + y^2 = n$ and to display them graphically in a way that sheds much light on their structure. The technique for doing this will work not just for the function $x^2 + y^2$ but also for any function $Q(x, y) = ax^2 + bxy + cy^2$, where $a$, $b$, and $c$ are integer constants. Such a function $Q(x, y)$ is called a *quadratic form*, or sometimes just a *form* for short.

Solving $x^2 + y^2 = n$ amounts to representing $n$ in the form of the sum of two squares. More generally, solving $Q(x, y) = n$ is called *representing $n$ by the form* $Q(x, y)$. So the overall goal is to solve the representation problem: Which numbers $n$ are represented by a given form $Q(x, y)$, and how does one find such representations.

Before starting to describe the method for displaying the values of a quadratic form graphically let us make a preliminary observation:

If the greatest common divisor of two integers $x$ and $y$ is $d$, then $Q(x, y) = d^2 Q\left(\frac{x}{d}, \frac{y}{d}\right)$ where the greatest common divisor of $\frac{x}{d}$ and $\frac{y}{d}$ is 1. Hence it suffices to find the values of $Q$ on *primitive pairs* $(x, y)$, the pairs whose greatest common divisor is 1, and then multiply these values by arbitrary squares $d^2$.

Thus the real problem is to find the primitive representations of a number $n$ by a form $Q(x, y)$, or in other words, to find the primitive solutions of $Q(x, y) = n$.

Primitive pairs $(x, y)$ correspond almost exactly to fractions $x/y$ that are reduced to lowest terms, the only ambiguity being that both $(x, y)$ and $(-x, -y)$ correspond to the same fraction $x/y$. However, this ambiguity does not affect the value
of a quadratic form $Q(x, y) = ax^2 + bxy + cy^2$ since $Q(x, y) = Q(-x, -y)$. This means that we can regard $Q(x, y)$ as being essentially a function $f(x/y)$. Notice that we are not excluding the possibility $(x, y) = (1, 0)$ which corresponds to the “fraction” $1/0$.

We already have a nice graphical representation of the rational numbers $x/y$ and $1/0$ as the vertices in the Farey diagram. Here is a picture of the diagram with the so-called dual tree superimposed:

![Diagram of a dual tree superimposed on a Farey diagram.]

The dual tree has a vertex in the center of each triangle of the Farey diagram, and it has an edge crossing each edge of the Farey diagram. The upper half of the dual tree does actually look like a real tree, with the lower half being its reflection in still water. As with the Farey diagram, we can only draw a finite part of the dual tree. The actual dual tree has branching that repeats infinitely often, an unending bifurcation process with smaller and smaller twigs.

The tree divides the interior of the large circle into regions, each of which is adjacent to one vertex of the original diagram. We can write the value $Q(x, y)$ in the region adjacent to the vertex $x/y$. This is shown in the figure below for the quadratic form $Q(x, y) = x^2 + y^2$, where to unclutter the picture we no longer draw the triangles of the original Farey diagram.
For example the 13 in the region adjacent to the fraction 2/3 represents the value $2^2 + 3^2$, and the 29 in the region adjacent to 5/2 represents the value $5^2 + 2^2$.

For a quadratic form $Q$ this picture showing the values $Q(x, y)$ is called the topograph of $Q$. It turns out that there is a very simple method for computing the topograph from just a very small amount of initial data. This method is based on the following:

**Arithmetic Progression Rule.** If the values of $Q(x, y)$ in the four regions surrounding an edge in the tree are $p$, $q$, $r$, and $s$ as indicated in the figure, then the three numbers $p$, $q + r$, $s$ form an arithmetic progression.

We can check this in the topograph of $x^2 + y^2$ shown above. Consider for example one of the edges separating the values 1 and 2. The values in the four regions surrounding this edge are 1, 1, 2, 5 and the arithmetic progression is 1, 1 + 2, 5. For an edge separating the values 1 and 5 the arithmetic progression is 2, 1 + 5, 10. For an edge separating the values 5 and 13 the arithmetic progression is 2, 5 + 13, 34. And similarly for all the other edges.

The arithmetic progression rule implies that the values of $Q$ in the three regions surrounding a single vertex of the tree determine the values in all other regions, by starting at the vertex where the three adjacent values are known and working one’s way outward in the dual tree. The easiest place to start for a quadratic form $Q(x, y) = ax^2 + bxy + cy^2$ is with the three values $Q(1, 0) = a$, $Q(0, 1) = c$, and $Q(1, 1) = a + b + c$ for the three fractions 1/0, 0/1, and 1/1. Here are two examples:
In the first case we start with the values 1 and 2 together with the 3 just above them. These determine the value 9 above the 2 via the arithmetic progression 1, 2 + 3, 9. Similarly the 6 above the 1 is determined by the arithmetic progression 2, 1 + 3, 6. Next one can fill in the 19 next to the 9 we just computed, using the arithmetic progression 3, 2 + 9, 19, and so on for as long as one likes.

The procedure for the other form \( x^2 - 2y^2 \) is just the same, but here there are negative as well as positive values. The edges that separate positive values from negative values will be important later, so we have indicated these edges by special shading.

Perhaps the most noticeable thing in both the examples \( x^2 + 2y^2 \) and \( x^2 - 2y^2 \) is the fact that the values in the lower half of the topograph are the same as those in the upper half. We could have predicted in advance that this would happen because \( Q(x, y) = Q(-x, y) \) whenever \( Q(x, y) \) has the form \( ax^2 + cy^2 \), with no \( xy \) term. The topograph for \( x^2 + y^2 \) has even more symmetry since the values of \( x^2 + y^2 \) are unchanged when \( x \) and \( y \) are switched, so the topograph has left-right symmetry as well.

Here is a general observation: The three values around one vertex of the topograph can be specified arbitrarily. For if we are given three numbers \( a, b, c \) then the quadratic form \( ax^2 + (c - a - b)xy + by^2 \) takes these three values for \((x, y)\) equal to \((1,0), (0,1), (1,1)\).

**Proof of the Arithmetic Progression Rule:** Let the two vertices of the Farey diagram corresponding to the values \( q \) and \( r \) have labels \( x_1/y_1 \) and \( x_2/y_2 \) as in the figure below. Then by the mediant rule for labeling vertices, the labels on the \( p \) and \( s \) regions
are the fractions shown. Note that these labels are correct even when \(x_1/y_1 = 1/0\) and \(x_2/y_2 = 0/1\).

For a quadratic form \(Q(x, y) = ax^2 + bxy + cy^2\) we then have

\[
Q(x_1 + x_2, y_1 + y_2) = a(x_1 + x_2)^2 + b(x_1 + x_2)(y_1 + y_2) + c(y_1 + y_2)^2
\]

\[
= ax_1^2 + bx_1y_1 + cy_1^2 + ax_2^2 + bx_2y_2 + cy_2^2 + (\cdots)
\]

Similarly we have

\[
P = Q(x_1 - x_2, y_1 - y_2) = ax_1^2 + bx_1y_1 + cy_1^2 + ax_2^2 + bx_2y_2 + cy_2^2 - (\cdots)
\]

The terms in \((\cdots)\) are the same in both cases, namely the terms involving both subscripts 1 and 2. If we compute \(p + s\) by adding the two formulas together, the terms \((\cdots)\) will therefore cancel, leaving just \(p + s = 2(q + r)\). This equation can be rewritten as \((q + r) - p = s - (q + r)\), which just says that \(p, q + r, s\) forms an arithmetic progression. 

**Periodic Separator Lines**

For most quadratic forms that take on both positive and negative values, such as \(x^2 - 2y^2\), there is another way of drawing the topograph that reveals some hidden and unexpected properties. For the form \(x^2 - 2y^2\) there is a zigzag path of edges in the topograph separating the positive and negative values, and if we straighten this path out to be a line, called the *separator line*, what we see is the following infinitely repeated pattern:
To construct this, one can first build the separator line starting with the three values $Q(1,0) = 1$, $Q(0,1) = -2$, and $Q(1,1) = -1$. Place these as shown in part (a) of the figure below, with a horizontal line segment separating the positive from the negative values.

To extend the separator line one step farther to the right, apply the arithmetic progression rule to compute the next value 2 using the arithmetic progression $-2, 1 - 1, 2$. Since this value 2 is positive, we place it above the horizontal line and insert a vertical edge to separate this 2 from the 1 to the left of it, as in (b) of the figure. Now we repeat the process with the next arithmetic progression $1, 2 - 1, 1$ and put the new 1 above the horizontal line with a vertical edge separating it from the previous 2, as shown in (c). At the next step we compute the next value $-2$ and place it below the horizontal line since it is negative, giving (d). One more step produces (e) where we see that further repetitions will produce a pattern that repeats periodically as we move to the right. The arithmetic progression rule also implies that it repeats periodically to the left, so it is periodic in both directions:

Thus we have the periodic separator line. To get the rest of the topograph we can then work our way upward and downward from the separator line, as shown in the original
As one moves upward from the separator line, the values of $Q$ become larger and larger, approaching $+\infty$ monotonically, and as one moves downward the values approach $-\infty$ monotonically. The reason for this will become clear in the next section when we discuss something called the Monotonicity Property.

An interesting property of this form $x^2 - 2y^2$ that is evident from its topograph is that it takes on the same negative values as positive values. This would have been hard to predict from the formula $x^2 - 2y^2$. Indeed, for the similar-looking quadratic form $x^2 - 3y^2$ the negative values are quite different from the positive values, as one can see in its straightened-out topograph:

$$Q(x, y) = x^2 - 3y^2$$

Continued Fractions Again

There is a close connection between the topograph for a quadratic form $x^2 - dy^2$ and the infinite continued fraction for $\sqrt{d}$ when $d$ is a positive integer that is not a square. In fact, we will see that the topograph can be used to compute the continued fraction for $\sqrt{d}$. As an example let us look at the case $d = 2$. The relevant portion of the topograph for $x^2 - 2y^2$ is the strip along the line separating the positive and negative values:

This is a part of the dual tree of the Farey diagram. If we superimpose the triangles of the Farey diagram corresponding to this part of the dual tree we obtain an infinite strip of triangles:
Ignoring the dotted triangles to the left, the infinite strip of triangles corresponds to the infinite continued fraction \( 1 + \frac{1}{\sqrt{2}} \). We could compute the value of this continued fraction by the methods in the previous chapter, but there is an easier way using the quadratic form \( x^2 - 2y^2 \). For fractions \( \frac{x}{y} \) labeling the vertices along the infinite strip, the corresponding values \( n = x^2 - 2y^2 \) are either \( \pm 1 \) or \( \pm 2 \). We can rewrite the equation \( x^2 - 2y^2 = n \) as \( \left( \frac{x}{y} \right)^2 = 2 + \frac{n}{2y^2} \). As we go farther and farther to the right in the infinite strip, both \( x \) and \( y \) are getting larger and larger while \( n \) only varies through finitely many values, namely \( \pm 1 \) and \( \pm 2 \), so the quantity \( \frac{n}{2y^2} \) is approaching 0. The equation \( \left( \frac{x}{y} \right)^2 = 2 + \frac{n}{2y^2} \) then implies that \( \left( \frac{x}{y} \right)^2 \) is approaching 2, so we see that \( \frac{x}{y} \) is approaching \( \sqrt{2} \). Since the fractions \( \frac{x}{y} \) are also approaching the value of the infinite continued fraction \( 1 + \frac{1}{\sqrt{2}} \) that corresponds to the infinite strip, this implies that the value of the continued fraction \( 1 + \frac{1}{\sqrt{2}} \) is \( \sqrt{2} \).

Here is another example, for the quadratic form \( x^2 - 3y^2 \), showing how \( \sqrt{3} = 1 + \frac{1}{1 + \frac{1}{2}} \).

After looking at these two examples one can see that it is not really necessary to draw the strip of triangles, and one can just read off the continued fraction directly from the periodic separator line. Let us illustrate this by considering the form \( x^2 - 10y^2 \):
If one moves toward the right along the horizontal line starting at a point in the edge separating the $\frac{1}{0}$ region from the $\frac{0}{1}$ region, one first encounters 3 edges leading off to the right (downward), then 6 edges leading off to the left (upward), then 6 edges leading off to the right, and so on. This means that the continued fraction for $\sqrt{10}$ is $3 + \frac{1}{6}$.

Here is a more complicated example showing how to compute the continued fraction for $\sqrt{19}$:

$$
\begin{array}{cccccccc}
1 & 6 & 5 & 9 & 9 & 5 & 6 & 1 & 6 \\
\end{array}
$$

From this we read off that $\sqrt{19} = 4 + \frac{1}{2} + \frac{1}{3} + \frac{1}{3} + \frac{1}{2} + \frac{1}{8}$.

In section 2 of this chapter we will prove that the topograph of the form $x^2 - dy^2$ always has a periodic separator line whenever $d$ is a positive integer that is not a square. As in the examples above, this separator line always includes the edge of the dual tree separating the vertices $1/0$ and $0/1$ since the form takes the positive value +1 on $1/0$ and the negative value $-d$ on $0/1$. The periodicity then implies that the continued fraction for $\sqrt{d}$ has the form

$$
\sqrt{d} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_n}}}
$$

with the periodic part starting immediately after the initial term $a_0$. In addition to being periodic, the separator line also has mirror symmetry with respect to reflection across the vertical line corresponding to the edge connecting $1/0$ to $0/1$ in the Farey diagram. This is because the form $x^2 - dy^2$ has no $xy$ term, so replacing $x/y$ by $-x/y$ does not change the value of the form. Once the separator line has symmetry with respect to this vertical line, the periodicity forces it to have mirror symmetry with respect to an infinite sequence of vertical lines, as illustrated in the following figure for the form $x^2 - 19y^2$:
In particular, these mirror symmetries imply that the continued fraction
\[ \sqrt{d} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_n}}} \]
always has two special properties:

(a) \( a_n = 2a_0 \).

(b) The intermediate terms \( a_1, a_2, \ldots, a_{n-1} \) form a palindrome, reading the same forward as backward.

Thus in \( \sqrt{19} = 4 + \frac{1}{2 + \frac{1}{1 + \frac{1}{3 + \frac{1}{1 + \frac{1}{2 + \frac{1}{8}}}}} \) the final 8 is twice the initial 4, and the intermediate terms 2, 1, 3, 1, 2 form a palindrome. These special properties held also in the earlier examples, but were less apparent because there were fewer terms in the repeated part of the continued fraction.

In some cases there is an additional kind of symmetry along the separator line, as illustrated for the form \( x^2 - 13y^2 \):

As before there is a horizontal translation giving the periodicity and there are reflec-tional symmetries across vertical lines, but now there is an extra glide-reflection along the strip that interchanges the positive and negative values of the form. Performing this glide-reflection twice in succession gives the translational periodicity. Notice that there are also 180 degree rotational symmetries about the points marked with dots on the separator line, and these rotations account for the palindromic middle part of the continued fraction
\[ \sqrt{13} = 3 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{6}}} \} \} \]
The fact that the periodic part has odd length corresponds to the separator strip having the glide-reflection symmetry. We could rewrite the continued fraction to have a periodic part of even length by doubling the period,

$$\sqrt{13} = 3 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{6 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{6}}}}}}}$$

and this corresponds to ignoring the glide-reflection and just considering the translational periodicity.

We have been using quadratic forms \(x^2 - dy^2\) to compute the continued fractions for irrational numbers \(\sqrt{d}\), but everything works just the same for irrational numbers \(\sqrt{p/q}\) if one uses the quadratic form \(qx^2 - py^2\) in place of \(x^2 - dy^2\). Following the same reasoning as before, if the equation \(qx^2 - py^2 = n\) is rewritten as \(q(\frac{x}{y})^2 = p + \frac{n}{y^2}\) then we see that as we move out along the periodic separator line the numbers \(x\) and \(y\) approach infinity while \(n\) cycles through finitely many values, so the term \(\frac{n}{y^2}\) approaches 0 and the fractions \(\frac{x}{y}\) approach a number \(z\) satisfying \(qz^2 = p\), so \(z = \sqrt{p/q}\). This argument depends of course on the existence of a periodic separator line, and we will prove in the next section that forms \(qx^2 - py^2\) always have a periodic separator line, assuming that \(\sqrt{p/q}\) is not a rational number, i.e., that \(p\) and \(q\) are not both squares.

Here are two examples. For the first one we use the form \(3x^2 - 7y^2\) to compute the continued fraction for \(\sqrt{7/3}\).

\[
\begin{array}{cccccccccc}
3 & 5 & 12 & 17 & 20 & 21 & 12 & 5 & 3 & 5 & 12 & 17 & 20 & 21 \\
\hline
-7 & -4 & -1 & -4 & -7 & -4 & -1
\end{array}
\]

This gives \(\sqrt{7/3} = 1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{8 + \frac{1}{1 + \frac{1}{1 + \frac{1}{2}}}}}\). For the second example we use \(10x^2 - 29y^2\) to compute the continued fraction for \(\sqrt{29/10}\),

\[
\begin{array}{cccccccccc}
10 & 11 & 26 & 19 & 29 & 19 & 26 & 11 & 10 & 11 & 10 \\
\hline
\end{array}
\]

with the result that \(\sqrt{29/10} = 1 + \frac{1}{1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2}}}}}\). The period of odd length here corresponds to the existence of the glide-reflection and 180 degree rotation symmetries.
As one can see in these examples, the palindrome property and the relation \( a_n = 2a_0 \) still hold for the continued fractions for irrational numbers \( \sqrt{p/q} \) assuming that \( a_0 > 0 \), which is equivalent to the condition \( p/q > 1 \) since \( a_0 \) is the integer part of \( \sqrt{p/q} \). Fractions \( p/q \) less than 1 can easily be dealt with just by inverting them, interchanging \( p \) and \( q \). Inverting a continued fraction \( a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_n}}} \) changes it to \( \frac{1}{a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_n}}} \). For example, from the earlier computation of \( \sqrt{7/3} \) we obtain \( \sqrt{3/7} = \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{8}}}} \).

One might ask whether the irrational numbers \( \sqrt{p/q} \) are the only numbers having a continued fraction \( a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_n}}} \) or \( \frac{1}{a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots + \frac{1}{a_n}}} \) satisfying the palindrome property and the relation \( a_0 = 2a_n \). The answer is yes, as we will see later in the chapter.

**Pell's Equation**

We encountered the equation \( x^2 - dy^2 = 1 \) briefly in Chapter 0. It is traditionally called Pell’s equation, and the similar equation \( x^2 - dy^2 = -1 \) is sometimes called Pell’s equation as well. If \( d \) is a square then the equations are not very interesting since in this case \( d \) can be incorporated into the \( y^2 \) term, so one is looking at the equations \( x^2 - y^2 = 1 \) and \( x^2 - y^2 = -1 \), which have only the trivial solutions \( (x, y) = (\pm 1, 0) \) for the first equation and \( (x, y) = (0, \pm 1) \) for the second equation, since these are the only cases when the difference between two squares is \( \pm 1 \). We will therefore assume that \( d \) is not a square in what follows.

As an example let us look at the equation \( x^2 - 19y^2 = 1 \). We drew a portion of the periodic separator line for the form \( x^2 - 19y^2 \) earlier, and here it is again with some of the fractional labels \( x/y \) shown as well.

<table>
<thead>
<tr>
<th>( \frac{1}{0} )</th>
<th>( \frac{9}{2} )</th>
<th>( \frac{48}{11} )</th>
<th>( \frac{170}{39} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>0</td>
<td>( \frac{4}{1} )</td>
<td>( \frac{13}{3} )</td>
<td>( \frac{61}{14} )</td>
</tr>
</tbody>
</table>

Ignoring the label \( 741/170 \) for the moment, the other fractional labels are the first few convergents for the continued fraction for \( \sqrt{19} \) that we computed before, \( 4 + \frac{1}{\sqrt{2} + \frac{1}{\sqrt{3} + \frac{1}{\sqrt{1} + \frac{1}{\sqrt{2} + \frac{1}{\sqrt{8}}}}} \). These fractional labels are the labels on the vertices of the zigzag path in the infinite strip of triangles in the Farey diagram, which we can imagine being superimposed on the separator line in the figure. The fractional label we are most interested in is the \( 170/39 \) because this is the label on a region where
the value of the form \( x^2 - 19y^2 \) is 1. This means exactly that \((x, y) = (170, 39)\) is a solution of \( x^2 - 19y^2 = 1 \). In terms of continued fractions, the fraction \( 170/39 \) is the value of the initial portion \( 4 + \frac{1}{2} + \frac{1}{1} + \frac{1}{3} + \frac{1}{1} + \frac{1}{2} \) of the continued fraction for \( \sqrt{19} \), with the final term of the period omitted.

Since the topograph of \( x^2 - 19y^2 \) is periodic along the separator line, there are infinitely many different solutions of \( x^2 - 19y^2 = 1 \) along the separator line. Going toward the left just gives the negatives \(-x/y\) of the fractions \( x/y \) to the right, changing the signs of \( x \) or \( y \), so it suffices to see what happens toward the right. One way to do this is to use the linear fractional transformation that gives the periodicity translation toward the right. This transformation sends the edge \( \langle 1/0, 0/1 \rangle \) of the Farey diagram to the edge \( \langle 170/39, 741/170 \rangle \). Here \( 741/170 \) is the value of the continued fraction \( 4 + \frac{1}{2} + \frac{1}{1} + \frac{1}{3} + \frac{1}{1} + \frac{1}{2} + \frac{1}{4} \) obtained from the continued fraction for \( \sqrt{19} \) by replacing the final number 8 in the period by one-half of its value, 4. The figure above shows why this is the right thing to do. We get an infinite sequence of larger and larger positive solutions of \( x^2 - 19y^2 = 1 \) by applying the periodicity transformation with matrix \( \begin{pmatrix} 170 & 741 \\ 39 & 170 \end{pmatrix} \) to the vector \( \begin{pmatrix} 1 \\ 0 \end{pmatrix} \). For example,

\[
\begin{pmatrix} 170 & 741 \\ 39 & 170 \end{pmatrix} \begin{pmatrix} 170 \\ 39 \end{pmatrix} = \begin{pmatrix} 57799 \\ 13260 \end{pmatrix}
\]

so the next solution of \( x^2 - 19y^2 = 1 \) after \((170, 39)\) is \((57799, 13260)\), and we could compute more solutions if we wanted. Obviously they are getting large rather quickly.

The two 170’s in the matrix \( \begin{pmatrix} 170 & 741 \\ 39 & 170 \end{pmatrix} \) can hardly be just a coincidence. Notice also that the entry 741 factors as \( 19 \cdot 39 \) which hardly seems like it should be just a coincidence either. Let’s check that these numbers had to occur. In general, for the form \( x^2 - dy^2 \) let us suppose that we have found the first solution \((x, y) = (p, q)\) after \((1, 0)\) for Pell’s equation \( x^2 - dy^2 = 1 \), so \( p^2 - dq^2 = 1 \). Then based on the previous example we suspect that the periodicity transformation is the transformation

\[
\begin{pmatrix} p & dq \\ q & p \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} px + dqy \\ qx + py \end{pmatrix}
\]

To check that this is correct the main thing to verify is that this transformation preserves the values of the quadratic form. When we plug in \((px + dqy, qx + dy)\) for \((x, y)\) in \( x^2 - dy^2 \) we get

\[
(px + dqy)^2 - d(qx + py)^2 = p^2x^2 + 2pdqxy + d^2q^2y^2 - dq^2x^2 - 2pdqxy - dp^2y^2 = (p^2 - dq^2)x^2 - d(p^2 - dq^2)y^2 = x^2 - dy^2 \text{ since } p^2 - dq^2 = 1
\]
so the transformation \( \begin{pmatrix} p & dq \\ q & p \end{pmatrix} \) does preserve the values of the form. Also it takes 1/0 to \( p/q \), and its determinant is \( p^2 - dq^2 = 1 \), so it has to be the translation giving the periodicity along the separator line. (We haven’t actually proved yet that periodic separator lines always exist for forms \( x^2 - dy^2 \), but we will do this in the next section.)

Are there other solutions of \( x^2 - 19y^2 = 1 \) besides the ones we have just described that occur along the separator line? The answer is No because we will see in the next section that as one moves away from the separator line in the topograph, the values of the quadratic form change in a monotonic fashion, steadily increasing toward \( +\infty \) as one moves upward above the separator line, and decreasing steadily toward \( -\infty \) as one moves downward below the separator line. Thus the value 1 occurs only along the separator line itself. Also we see that the value \(-1\) never occurs, which means that the equation \( x^2 - 19y^2 = -1 \) has no integer solutions.

For an example where \( x^2 - dy^2 = -1 \) does have solutions, let us look again at the earlier example of \( x^2 - 13y^2 \). The first positive solution \((x, y) = (p, q)\) of \( x^2 - 13y^2 = -1 \) corresponds to the value \(-1\) in the middle of the figure. This is determined by the continued fraction \( p/q = 3 + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} = 18/5 \), so we have \((p, q) = (18, 5)\). The matrix \( \begin{pmatrix} p & dq \\ q & p \end{pmatrix} \) in this case is \( \begin{pmatrix} 18 & 65 \\ 5 & 18 \end{pmatrix} \) with determinant \( 18^2 - 13 \cdot 5^2 = -1 \) so this gives the glide-reflection along the periodic separator line taking 1/0 to 18/5 and 0/1 to 65/18. The smallest positive solution of \( x^2 - 13y^2 = +1 \) is obtained by applying this glide-reflection to \((18, 5)\), which gives

\[
\begin{pmatrix} 18 & 65 \\ 5 & 18 \end{pmatrix} \begin{pmatrix} 18 \\ 5 \end{pmatrix} = \begin{pmatrix} 324 + 325 \\ 90 + 90 \end{pmatrix} = \begin{pmatrix} 649 \\ 180 \end{pmatrix}
\]

Repeated applications of the glide-reflection will give solutions of \( x^2 - 13y^2 = +1 \) and \( x^2 - 13y^2 = -1 \) alternately.
2.2 The Classification of Quadratic Forms

We can divide quadratic forms $Q(x, y) = ax^2 + bxy + cy^2$ into four broad classes according to the signs of the values $Q(x, y)$ for $(x, y) \neq (0, 0)$, where as always we restrict $x$ and $y$ to integers. (We assume at least one of the coefficients $a, b, c$ is nonzero, so $Q$ is not identically zero.)

(I) $Q(x, y)$ takes on both positive and negative values but not 0. In this case we call $Q$ a hyperbolic form.

(II) $Q(x, y)$ takes on both positive and negative values and also 0. Then we call $Q$ a 0-hyperbolic form.

(III) $Q(x, y)$ takes on only positive values or only negative values. Then we call $Q$ elliptic.

(IV) $Q$ takes on the value 0 and either positive or negative values, but not both. Then $Q$ is called parabolic.

The hyperbolic-elliptic-parabolic terminology is motivated in part by what the level curves $ax^2 + bxy + cy^2 = k$ are, where we now allow $x$ and $y$ to take on all real values so that one gets actual curves. The level curves are hyperbolas in cases (I) and (II), and ellipses in case (III). In case (IV), however, the level curves are not parabolas as one might guess, but straight lines. Case (IV) will be the least interesting of the four cases.

There is an easy way to distinguish the four cases by looking at the discriminant $\Delta = b^2 - 4ac$:

(I) If $\Delta$ is positive but not a square then $Q$ is hyperbolic.

(II) If $\Delta$ is positive and a square then $Q$ is 0-hyperbolic.

(III) If $\Delta$ is negative then $Q$ is elliptic.

(IV) If $\Delta$ is zero then $Q$ is parabolic.

These statements will be proved later in this section.

We will analyze each of the four types of forms in turn, but before doing this let us make a couple preliminary general comments. In the arithmetic progression rule for labeling the four regions surrounding a given edge of the topograph, we can label the edge by the common increment $h = (q + r) - p = s - (q + r)$ as in the figure at the right. The edge can be oriented by an arrow showing the direction in which the progression increases by $h$. Changing the sign of $h$ corresponds to changing the orientation of the edge. In the special case that $h$ happens to be 0 we will not orient the edge.
The values of the increment $h$ along the boundary of a region in the topograph have the interesting property that they also form an arithmetic progression:

![Diagram of the topograph showing the arithmetic progression of edge labels along the boundary of a region.]

We will call this property the Second Arithmetic Progression Rule. To see why it is true, start with the edge labeled $h$ in the figure, with the adjacent regions labeled $p$ and $q$. The original Arithmetic Progression Rule then gives the value $p + q + h$ in the next region to the right, and another application of the rule gives the label $h + 2p$ on the next edge. Thus the edge label increases by $2p$ when we move from one edge to the next edge to the right, so by repeated applications of this fact we see that we have an arithmetic progression of edge labels all along the border of the region labeled $p$.

Another thing worth noting at this point is something that we will refer to as the Monotonicity Property: If the three labels $p$, $q$, and $h$ adjacent to an edge are all positive, then so are the three labels for the next two edges in front of this edge, and the new labels are larger than the old labels. It follows that when one continues forward out this part of the topograph, all the labels become monotonically larger the farther one goes. Similarly, when the original three labels are negative, all the labels become larger and larger negative. This is really just the same principle applied to the negative $-Q(x, y)$ of the original form $Q(x, y)$.

**Proposition.** If an edge in the topograph of $Q(x, y)$ is labeled $h$ with adjacent regions labeled $p$ and $q$, then the quantity $h^2 - 4pq$ is equal to the discriminant of $Q(x, y)$.

**Proof:** For the given form $Q(x, y) = ax^2 + bxy + cy^2$, the regions $1/0$ and $0/1$ in the topograph are labeled $a$ and $c$, and the edge in the topograph separating these two regions has $h = b$ since the $1/1$ region is labeled $a + b + c$. So the statement of the proposition is correct for this edge. For other edges we proceed by induction, moving farther and farther out the tree. For the induction step suppose we have two adjacent edges labeled $h$ and $k$ as in the figure,
and suppose inductively that the discriminant equals \( h^2 - 4pq \). Then \( r = p + q + h \), and hence \( k = (q + r) - p = q + (p + q + h) - p = h + 2q \), so we have \( k^2 - 4qr = (h + 2q)^2 - 4q(p + q + h) = h^2 + 4hq + 4q^2 - 4pq - 4q^2 - 4hq = h^2 - 4pq \), which means that the result holds for the edge labeled \( k \) as well.

\[ \square \]

Hyperbolic Forms

The most interesting of the four types of quadratic forms are the hyperbolic forms. We will show that these all have a periodic separator line as in the examples \( x^2 - dy^2 \) and \( qx^2 - py^2 \) that we looked at earlier.

**Theorem.** For a hyperbolic form \( Q(x, y) \) the edges of the topograph for which the two adjacent regions are labeled by numbers of opposite sign form a line which is infinite in both directions, and the topograph is periodic along this line.

**Proof:** Since the form is hyperbolic, all regions of the topograph have labels that are either positive or negative, never zero. There must exist two regions of opposite sign since \( Q \) is hyperbolic, and by moving along a path in the topograph joining these two regions we will somewhere encounter two adjacent regions of opposite sign. Thus there must exist edges whose two adjacent regions have opposite sign. Let us call these edges *separating edges*. If we apply the discriminant formula \( \Delta = h^2 - 4pq \) in preceding proposition to a separating edge, we see that \( \Delta \) must be positive since \( p \) and \( q \) are nonzero and have opposite sign, so \(-4pq\) is positive while \( h^2 \) is positive or zero. Thus a hyperbolic form must have positive discriminant.

At an end of a separating edge the value of \( Q \) in the next region must be either positive or negative since \( Q \) does not take the value 0:

\[
\begin{array}{c}
+ \\
- \\
+ \\
- 
\end{array}
\]

This implies that exactly one of the two edges at the end of the first separating edge is also a separating edge. Repeating this argument, we see that each separating edge is part of a line of separating edges that is infinite in both directions (and the edges that lead off from this edge are not separating edges).

As we move off this separating line the values of \( Q \) are steadily increasing on the positive side and steadily decreasing on the negative side, by the monotonicity property, so there are no other separating edges that are not on this line.
It remains to prove that the topograph is periodic along the separating line. We can assume all the edges along the line are oriented in the same direction, by changing the signs of the \( h \) values where necessary. For an edge of the line labeled \( h \) with adjacent regions labeled \( p \) and \(-q\), with \( p, q > 0 \), we know that \( h^2 + 4pq \) is equal to the discriminant \( \Delta \). From the equation \( \Delta = h^2 + 4pq \) we obtain the inequalities \( |h| < \sqrt{\Delta} \), \( p \leq \Delta/4 \), and \( q \leq \Delta/4 \). Thus there are only finitely many possible values for \( h \), \( p \), and \( q \) along the separator line. Hence there are only finitely many possible combinations of values \( h \), \( p \), and \( q \) for each edge on the separator line. It follows that there must be two edges on the line that have the same values of \( h \), \( p \), and \( q \). Since the topograph is uniquely determined by the three labels \( h \), \( p \), \( q \) at a single edge, the translation of the line along itself that takes one edge to another edge with the same three labels must preserve all the labels on the line. This shows that the separator line is periodic, including the values of \( Q \).

Conceivably there might be just a single region on one side of the separator line, but this doesn't actually happen: There must be edges leading away from the separating line on both the side where the form is positive and on the side where it is negative. For if there was just a single region on one side of the line, the second arithmetic progression rule would say that the \( h \) labels along the line formed an infinite arithmetic progression, and hence the \( h \) values would not be bounded, contradicting the fact that there are only finitely many different values for \( h \) along the separator line, as we just showed.

Here is an interesting consequence of the periodicity of the separator line:

**Corollary.** For a hyperbolic form \( Q(x, y) = ax^2 + bxy + cy^2 \), if the equation \( ax^2 + bxy + cy^2 = n \) has one integer solution then it has infinitely many integer solutions.

**Proof:** Suppose \((x, y)\) is a solution of \( Q(x, y) = n \). If \((x, y)\) is a primitive pair, then the number \( n \) appears in the topograph of \( Q \) infinitely many times, via the periodicity of the separator line, so there are infinitely many solutions in this case. If \((x, y)\) is not primitive then it is \( m \) times a primitive pair \((x', y')\) with \( Q(x', y') = n/m^2 \). This latter equation has infinitely many solutions as we just saw, so after multiplying these solutions \((x', y')\) by \( m \) we get infinitely many solutions of \( Q(x, y) = n \).

In the previous chapter we gave an argument that showed that infinite continued fractions that are eventually periodic always represent quadratic irrational numbers. This is one half of Lagrange’s Theorem, and now we can prove the other half, the converse statement:
Theorem. The continued fraction expansion of every quadratic irrational is eventually periodic.

Proof: A quadratic irrational number α has the form $A + B\sqrt{n}$ where $A$ and $B$ are rational numbers and $n$ is a positive integer that is not a square. Letting $\overline{\alpha}$ be the conjugate $A - B\sqrt{n}$ of $\alpha$, we see that $\alpha$ and $\overline{\alpha}$ are roots of the quadratic equation $(x - \alpha)(x - \overline{\alpha}) = x^2 - 2Ax + (A^2 - nB^2) = 0$ whose coefficients are rational numbers. After multiplying through by a common denominator we can replace this equation by an equation $ax^2 + bx + c = 0$ with integer coefficients having $\alpha$ and $\overline{\alpha}$ as roots. The leading coefficient $a$ is nonzero since it arose from multiplying by a common denominator.

From the quadratic equation $ax^2 + bx + c = 0$ we obtain a quadratic form $Q(x,y) = ax^2 + bxy + cy^2$ with the same coefficients $a, b, c$. We claim that this quadratic form is hyperbolic. It cannot take on the value 0 at an integer pair $(x, y) \neq (0, 0)$ since if $ax^2 + bxy + cy^2 = 0$ then we cannot have $y = 0$, otherwise the equation would become $ax^2 = 0$ with $a \neq 0$, forcing $x$ to be 0 as well. Since $y \neq 0$ we can divide the equation $ax^2 + bxy + cy^2 = 0$ by $y^2$ to get a quadratic equation $a(x/y)^2 + b(x/y) + c = 0$ with a rational root $x/y$, contrary to the assumption that the root $\alpha$, and hence also $\overline{\alpha}$, was irrational. Thus the quadratic form $Q(x,y)$ does not take on the value 0. To see that $Q(x,y)$ takes on both positive and negative values, note that $a(x/y)^2 + b(x/y) + c$ takes on both positive and negative values at rational numbers $x/y$ since the graph of the function $ax^2 + bxy + cy^2$ is a parabola crossing the $x$-axis at two distinct points $\alpha$ and $\overline{\alpha}$. Multiplying the formula $a(x/y)^2 + b(x/y) + c$ by the positive number $y^2$, it follows that $ax^2 + bxy + cy^2$ also takes on both positive and negative values at integer pairs $(x, y)$.

Since $Q$ is hyperbolic, its topograph contains a periodic line separating the positive and negative values. This corresponds to a strip in the Farey diagram which is infinite in both directions. The fractions $x_n/y_n$ labeling the vertices along this strip have both $x_n$ and $y_n$ approaching $\pm\infty$ as $n$ goes to $\pm\infty$. (The only way this could fail for a path consisting of an infinite sequence of distinct edges in the dual tree would be if all the edges from some point onward bordered the 1/0 or 0/1 region, which is not the case here since periodic separator lines have only a finite number of edges bordering a given region.) The values $Q(x_n, y_n) = ax_n^2 + bxy_n + cy_n^2 = k_n$ are bounded, ranging over a finite set along the strip. Thus $a(x_n/y_n)^2 + b(x_n/y_n) + c = k_n/y_n^2 \to 0$ as $n \to \pm\infty$, so at one end of the strip we have $x_n/y_n \to \alpha$ and at the other end we have $x_n/y_n \to \overline{\alpha}$. Joining either end of the strip to 1/0 in the Farey diagram then
gives infinite strips corresponding to infinite continued fractions for $\alpha$ and $\overline{\alpha}$ that are eventually periodic.

Let us look at an example to illustrate the procedure in the proof of this theorem. We will use a quadratic form to compute the continued fractions for the two quadratic irrationals $\frac{10 + \sqrt{2}}{14}$. The equation $(x - \alpha)(x - \overline{\alpha}) = 0$ is $x^2 - \frac{10}{7}x + \frac{1}{2} = 0$, so with integer coefficients this becomes $14x^2 - 20x + 7 = 0$. The associated quadratic form is $14x^2 - 20xy + 7y^2$. To compute the topograph we start with the three values at 1/0, 0/1, and 1/1 and work toward the separator line:

This figure lies in the upper half of the circular Farey diagram where the fractions $x/y$ are positive, so if we follow the separator line out to the right we approach the smaller of the two roots of $14x^2 - 20x + 7 = 0$, which is $\frac{10 - \sqrt{2}}{14}$, and if we follow the separator line to the left we approach the larger root, $\frac{10 + \sqrt{2}}{14}$. To get the continued fraction for the smaller root we follow the path in the figure that starts with the edge between 1/0 and 0/1, then zigzags up to the separator line, then goes out this line to the right. If we straighten this path out it looks like the following:

The continued fraction is therefore

$$\frac{10 - \sqrt{2}}{14} = \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{2}$$

It is not actually necessary to redraw the straightened-out path since in the original form of the topograph we can read off the sequence of left and right “side roads” as we go along the path, the sequence $LRLRRR$ where $L$ denotes a side road to the left and $R$ a side road to the right. This sequence determines the continued fraction. For the other root $\frac{10 + \sqrt{2}}{14}$ the straightened-out path has the following shape:
The sequence of side roads is \( LRRRLRRR \) so the continued fraction is

\[
\frac{10 + \sqrt{2}}{14} = \frac{1}{1} + \frac{1}{4} + \frac{1}{2}
\]

A natural question to ask is whether every periodic line in the dual tree of the Farey diagram is the separator line of some hyperbolic form, and the answer is yes. To find the form one first uses the periodic line to construct a continued fraction that is eventually periodic, then one computes the value of this continued fraction by finding a quadratic equation that it satisfies, and this quadratic equation gives the desired quadratic form. As an example, let us find a quadratic form whose periodic line looks like the following:

A periodic continued fraction corresponding to this strip is \( \frac{1}{1} + \frac{1}{2} + \frac{1}{3} \). The value of this continued fraction was worked out in an example in Section 1.3 by finding a quadratic equation that it satisfies, which was \( 3x^2 + 8x - 7 = 0 \) with roots \((-4 \pm \sqrt{37})/3\). The positive root \((-4 + \sqrt{37})/3\) is the value of \( \frac{1}{1} + \frac{1}{2} + \frac{1}{3} \). The corresponding quadratic form is \( 3x^2 + 8xy - 7y^2 \), and if we compute its separator line we get

This provides a realization of the given periodic line as the separator line in the toponograph of a quadratic form. Notice that the separator line is not symmetric under reflection across any vertical line, unlike all the separator lines we have seen up to this point. This is the simplest example without this bilateral symmetry property.
since the strips associated to continued fractions $\frac{1}{a_1}$ and $\frac{1}{a_2} + \frac{1}{a_3}$ obviously have bilateral symmetry, as do the strips for continued fractions $\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3}$ if two of the numbers $a_1, a_2, a_3$ are equal.

**Elliptic Forms**

An elliptic quadratic form $Q(x, y)$ takes on either all positive or all negative values at integer pairs $(x, y) \neq (0, 0)$. The two cases are equivalent since one can switch from one to the other just by putting a minus sign in front of $Q$. Thus it suffices to consider the case that $Q$ takes on only positive values, and we will assume we are in this case from now on.

Let $p$ be the minimum value taken on by $Q$, and consider a region of the topograph where $Q$ takes the value $p$. All the edges having one endpoint at this region are oriented away from the region, by the arithmetic progression rule and the assumption that $p$ is the minimum value of $Q$. The monotonicity property then implies that all edges farther away from the $p$ region are also oriented away from the region, and the values of $Q$ increase as one moves away from the region.

We know that the $h$-labels on the edges making up the border of the $p$ region form an arithmetic progression with increment $2p$. There are two possibilities for these $h$-labels:

(I) Some edge bordering the $p$ region has the label $h = 0$. The topograph then has the form shown in the first figure below. An example of such a form is $px^2 + qy^2$. We call the 0-labeled edge a source edge since all other edges are oriented away from this edge.

(II) No edge bordering the $p$ region has label $h = 0$. Since the labels on these edges form an arithmetic progression, there must be some vertex where the terms in the
progression change sign, and so all three edges meeting at this vertex will be oriented away from the vertex, as in the second figure above. We call this a source vertex since all edges in the topograph are oriented away from this vertex. The fact that the three edges leading from a source vertex all point away from the vertex is equivalent to the three triangle inequalities
\[ p < q + r \quad q < p + r \quad r < p + q \]
In the case of a source edge one of these inequalities becomes an equality \( r = p + q \).

**Proposition.** Elliptic forms have negative discriminant.

**Proof:** In the case of a source edge with the label \( h = 0 \) separating regions labeled \( p \) and \( q \), the discriminant is \( \Delta = h^2 - 4pq = -4pq \), which is negative. In the case of a source vertex with adjacent regions labeled \( p, q, r \), the edge between the \( p \) and \( q \) regions is labeled \( h = p + q - r \) so we have
\[
\Delta = h^2 - 4pq = (p + q - r)^2 - 4pq
= p^2 + q^2 + r^2 - 2pq - 2pr - 2qr
= p(p - q - r) + q(q - p - r) + r(r - p - q)
\]
In the last line the three quantities in parentheses are negative by the triangle inequalities, so \( \Delta \) is negative. \( \square \)

**Parabolic and 0-Hyperbolic Forms**

These are the forms whose topograph has at least one region labeled 0. Each edge adjacent to the 0 region has the same label \( h \), and the labels on the regions adjacent to the 0 region form an arithmetic progression. The discriminant is \( \Delta = h^2 \), a square.

A special case is \( h = 0 \). Then the topograph is as shown in the next figure, and the form is parabolic with discriminant \( \Delta = h^2 = 0 \). Notice that the topograph is periodic along the 0 region since it consists of the same tree pattern repeated infinitely often.
Chapter 2 Quadratic Forms

In the other case that \( h \) is nonzero, so that the discriminant \( \Delta = h^2 \) is a nonzero square, the arithmetic progression of values of \( Q \) adjacent to the 0 region includes both positive and negative numbers, so \( Q \) is 0-hyperbolic. If the arithmetic progression includes the value 0, this gives a second 0 region adjacent to the first one, and the topograph is as shown at the right. If the arithmetic progression does not include 0, there will be an edge separating the positive from the negative values in the progression. We can extend this separating edge to a line of separating edges as we did with hyperbolic forms, but the extension will eventually have to terminate with a second 0 region, otherwise the reasoning we used in the hyperbolic case would yield two edges along this line having the same \( h \) and the same positive and negative labels on the two adjacent regions, which would force the line to be periodic and hence extend infinitely far in both directions, which is impossible since it began at a 0 region at one end. Thus the topograph contains a finite separator line connecting two 0 regions. If one of the 0 regions is the 1/0 region and the other is the \( p/q \) region, then the separator line follows the strip of triangles in the Farey diagram corresponding to the continued fraction for \( p/q \). The quadratic form \( Q(x, y) = qx \cdot y - py^2 = (qx - py)y \) has a topograph with this behavior. For example, for \( p/q = 2/5 \) the topograph of the form \( 5xy - 2y^2 = (5x - 2y)y \) is the following:
We have seen that the discriminant of a form that takes the value 0 is a square. Here is the converse:

**Proposition.** If the discriminant of a form $Q(x, y)$ is a square, then $Q(x, y) = 0$ for some pair of integers $(x, y) \neq (0, 0)$.

**Proof:** Suppose first that the form $Q(x, y) = ax^2 + bxy + cy^2$ has $a \neq 0$. Then the equation $aX^2 + bX + c = 0$ has roots $X = (-b \pm \sqrt{b^2 - 4ac})/2a$. If $b^2 - 4ac$ is a square, this means the roots are rational. If $X = p/q$ is a rational root then $a(p/q)^2 + b(p/q) + c = 0$ and hence $ap^2 + bpq + cq^2 = 0$ so $Q$ takes the value 0 at a pair $(p, q)$ with $q \neq 0$. There remains the case that $a = 0$, so $Q(x, y) = bxy + cy^2$, which is 0 at $(x, y) = (1, 0)$.

In particular, this shows the discriminant of a hyperbolic form is not a square. Since we showed earlier that a hyperbolic form has positive discriminant, this completes the characterization of the four types of forms in terms of their discriminants.

**Equivalence of Forms**

In the pictures of topographs we have drawn, we often omit the fractional labels $x/y$ for the regions in the topograph since the more important information is often just the values $Q(x, y)$ of the form. This leads to the idea of considering two quadratic forms to be equivalent if their topographs “look the same” when the labels $x/y$ are disregarded. For a precise definition, one can say that quadratic forms $Q_1$ and $Q_2$ are equivalent if there is a vertex $v_1$ in the topograph of $Q_1$ and a vertex $v_2$ in the topograph of $Q_2$ such that the values of $Q_1$ in the three regions surrounding $v_1$ are equal to the values of $Q_2$ in the three regions surrounding $v_2$. Since the three values around a vertex determine all the other values in a topograph, this guarantees that the topographs look the same everywhere, if the labels $x/y$ are omitted.

With this definition, a topograph and its mirror image correspond to equivalent forms since the mirror image topograph has the same three labels around each vertex as in the corresponding vertex of the original topograph. For example, switching the variables $x$ and $y$ reflects the circular Farey diagram across its vertical axis and hence reflects the topograph of a form $Q(x, y)$ to the topograph of the equivalent form $Q(y, x)$. As another example, the forms $ax^2 + bxy + cy^2$ and $ax^2 - bxy + cy^2$ are equivalent since they are related by changing $(x, y)$ to $(-x, y)$, reflecting the Farey diagram across its horizontal axis, with a corresponding reflection of the topograph.
Theorem. Up to equivalence, there are just a finite number of forms with a given discriminant, except in the special case that the discriminant is zero.

This fails to hold for forms of discriminant 0, the parabolic forms, since multiplying such a form by different integers produces infinitely many inequivalent forms.

Proof: Consider first the case of forms of positive discriminant. These are either hyperbolic or 0-hyperbolic. Hyperbolic forms have a separator line. For an edge in the separator line labeled $h$ with adjacent regions labeled $p > 0$ and $-q < 0$ we have $\Delta = h^2 + 4pq$, so each of the quantities $|h|$, $p$, and $q$ is bounded in size by $\Delta$. This means that for fixed $\Delta$ there are only finitely many possibilities for $h$, $p$, and $q$ for each edge of the separator line, hence just finitely many possible combinations of $h$, $p$, and $-q$ for each edge, so there are just finitely many possibilities for the form, up to equivalence. The same reasoning applies also to 0-hyperbolic forms that have a separating edge in their topograph. The only ones that do not have a separating edge are the ones with two adjacent regions labeled 0. In this case the edge separating these two regions has $h^2 = \Delta$, so the value of $h$ on this edge is determined by $\Delta$, hence the form is determined by $\Delta$.

For forms of negative discriminant we can assume we are dealing with positive elliptic forms since a form $Q$ and its negative $-Q$ have the same discriminant. If a positive elliptic form has a source edge in its topograph, this edge has $h = 0$ so $\Delta = -4pq$ where $p$ and $q$ are the values of $Q$ in the adjacent regions. For fixed $\Delta$ there are only finitely many choices of $p$ and $q$ satisfying $\Delta = -4pq$, so there are only finitely many positive elliptic forms of discriminant $\Delta$ having a source edge. In the other case of a source vertex surrounded by values $p,q,r$ of the form, we obtained the formula $\Delta = p(p - q - r) + q(q - p - r) + r(r - p - q)$ with the three quantities in parentheses being negative, so $p + q + r \leq |\Delta|$ and hence there are only finitely many possibilities for $p$, $q$, and $r$ for each $\Delta$. □

As an example, let us determine all the quadratic forms of discriminant 60, up to equivalence. Two obvious forms of discriminant 60 are $x^2 - 15y^2$ and $3x^2 - 5y^2$, whose separator lines consist of periodic repetitions of the following two patterns:

\[
\begin{array}{c|c|c|c|c|c|c|c}
1 & 1 \\
\end{array}
\quad
\begin{array}{c|c|c|c|c|c|c}
3 & 7 & 7 & 3 \\
-5 & -2 & -5 \\
\end{array}
\]

From these topographs it is apparent that the two forms are not equivalent, and also that the negatives of these two forms, $-x^2 + 15y^2$ and $-3x^2 + 5y^2$, give two more
inequivalent forms. To see whether there are others we use the formula $\Delta = 60 = h^2 + 4pq$ relating the values $p$ and $-q$ along an edge labeled $h$ in the separator line, with $p > 0$ and $q > 0$. The various possibilities are listed in the table below. Note that the equation $60 = h^2 + 4pq$ implies that $h$ has to be even.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$pq$</th>
<th>$(p,q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15</td>
<td>(1,15), (3,5), (5,3), (15,1)</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>(1,14), (2,7), (7,2), (14,1)</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>(1,11), (11,1)</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>(1,6), (2,3), (3,2), (6,1)</td>
</tr>
</tbody>
</table>

Each combination of values for $h$, $p$, and $q$ in the table occurs at some edge along the separator line in one of the two topographs shown above, or the negatives of these topographs. Hence every form of discriminant 60 is equivalent to one of these four. If it had not been true that all the possibilities in the table occurred in the topographs of the forms we started with, we could have used these other possibilities for $h$, $p$, and $q$ to generate new topographs and hence new forms, eventually exhausting all the finitely many possibilities.

For finding all the positive elliptic quadratic forms of a given discriminant, up to equivalence, the procedure is simpler since one doesn’t have to actually draw any topographs. At a source vertex or edge in the topograph for such a form $Q$ let the smaller two of the three adjacent values of $Q$ be $a \leq c$, with the edge between them labeled $h \geq 0$, so that the third adjacent value of $Q$ is $a + c - h$. The form is then equivalent to the form $ax^2 + hxy + cy^2$. Since $a$ and $c$ are the smallest values of $Q$ we have $a \leq c \leq a + c - h$, and the latter inequality implies that $h \leq a$. Thus we have the inequalities $0 \leq h \leq a \leq c$. Note that these inequalities imply the three triangle inequalities at the source vertex or edge: $a + c - h \leq a + c$, $a < c + (a + c - h)$, and $c < a + (a + c - h)$. For the discriminant $\Delta = -D$ we have $D = 4ac - h^2$, so we are seeking solutions of

$$4ac = h^2 + D \quad \text{with} \quad 0 \leq h \leq a \leq c$$

The number $h$ must have the same parity as $D$, and we can bound the choices for $h$ by the inequalities $4h^2 \leq 4a^2 \leq 4ac = D + h^2$ which imply $3h^2 \leq D$, or $h^2 \leq D/3$. Every positive elliptic form is equivalent to one of the forms $ax^2 + hxy + cy^2$ for triples $(a,h,c)$ satisfying these conditions $4ac = h^2 + D$, $0 \leq h \leq a \leq c$, and $h^2 \leq D/3$. Different choices of $(a,h,c)$ satisfying these conditions never give forms that are
equivalent since \( a \) and \( c \) are the labels on the two regions in the topograph where the form takes its smallest values, and \( h \) is determined by \( a \), \( c \), and \( D \) by the formula \( 4ac = h^2 + D \).

As an example, when \( D = 80 \) we must have \( h \) even and \( h^2 \leq 80/3 \) so \( h \) must be 0, 2, or 4. The corresponding values of \( a \) and \( c \) that are possible can then be computed from the equation \( 4ac = 80 + h^2 \), keeping in mind that \( h \leq a \leq c \). The possibilities are shown in the following table:

<table>
<thead>
<tr>
<th>( h )</th>
<th>( ac )</th>
<th>( (a,c) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
<td>(1,20), (2,10), (4,5)</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>(3,7)</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>(4,6)</td>
</tr>
</tbody>
</table>

Thus every positive elliptic form of discriminant \(-80\) is equivalent to one of the forms \( x^2 + 20y^2 \), \( 2x^2 + 10y^2 \), \( 4x^2 + 5y^2 \), \( 3x^2 + 2xy + 7y^2 \), or \( 4x^2 + 4xy + 6y^2 \), and no two of these are equivalent to each other, as explained earlier.

We know there are only finitely many forms of a given nonzero discriminant, up to equivalence, but what about the question of whether every integer occurs as the discriminant of a form? For a form \( ax^2 + bxy + cy^2 \) we have \( \Delta = b^2 - 4ac \), and this is congruent to \( b^2 \) mod 4. A square such as \( b^2 \) is always congruent to 0 or 1 mod 4, so the discriminant of a form is always congruent to 0 or 1 mod 4. Conversely, for every integer \( \Delta \) congruent to 0 or 1 mod 4 there exists a form whose discriminant is \( \Delta \). Namely, if \( \Delta = 4k \) then the form \( x^2 - ky^2 \) has discriminant \( 4k \), and if \( \Delta = 4k + 1 \) then the form \( x^2 + xy - ky^2 \) has discriminant \( 4k+1 \). Here \( k \) can be positive, negative, or zero. The forms \( x^2 - ky^2 \) and \( x^2 + xy - ky^2 \) are called the principal quadratic forms of these discriminants.

The number of equivalence classes of quadratic forms with a given discriminant, where one only considers forms having positive values in the elliptic case, is known as the class number for the given discriminant. Of special interest are the cases when the class number is 1, so all forms of that discriminant are equivalent. There are nine negative discriminants \( \Delta = -D \) of class number 1:

\[
D = 3, 4, 7, 8, 11, 19, 43, 67, 163
\]

It was conjectured by Gauss around 1800 that this is the complete list for negative discriminants. It was shown in the 1930s that there is at most one more, and then in the 1960s the possibility of an elusive tenth such discriminant was finally ruled out, finishing the proof of the conjecture. For positive discriminant there are many more
cases where the class number is 1, but it is still unknown whether there are infinitely many such discriminants.

In the nine cases $D = 3, 4, 7, 8, 11, 19, 43, 67, 163$ it is very easy to check that all forms are equivalent. For example when $D = 163$ we must have $h$ odd with $h^2 \leq 163/3$ so the only possibilities are $h = 1, 3, 5, 7$. From the equation $4ac = 163 + h^2$ the corresponding values of $ac$ are $41, 43, 47, 53$ which all happen to be primes, and since $a \leq c$ this forces $a$ to be 1 in each case. But since $h \leq a$ this means $h$ must be 1, and we obtain the single quadratic form $ax^2 + hxy + cy^2 = x^2 + xy + 41y^2$.

The corresponding polynomial $x^2 + x + 41$ has a curious property discovered by Euler: For each $x = 0, 1, 2, \cdots, 39$ the value of $x^2 + x + 41$ is a prime number. Here are these primes:

```
41  43  47  53  61  71  83  97 113 131 151 173 197 223 251 281 313 347 383 421
461 503 547 593 641 691 743 797 853 911 971 1033 1097 1163 1231 1301
1373 1447 1523 1601
```

Notice that the successive differences between these numbers are 2, 4, 6, 8, \cdots. The next number in the sequence would be $1681 = 41^2$, not a prime. (Write $x^2 + x + 41$ as $x(x + 1) + 41$ to see why $x = 40$ must give a nonprime value.) A similar thing happens for the other values of $D$. The nontrivial cases are:

| $D$ | $x^2 + x + 2$ | $2$
|-----|---------------|---
| 7   | $x^2 + x + 3$ | 3 5
| 11  | $x^2 + x + 5$ | 5 7 11 17
| 19  | $x^2 + x + 11$| 11 13 17 23 31 41 53 67 83 101
| 43  | $x^2 + x + 17$| 17 19 23 29 37 47 59 73 89 107 127 149 173 199 227 257

It’s interesting that these lists include all primes less than 100 except for 79.
2.3 Representing Numbers by Forms

With the various things we have learned about quadratic forms so far, let us return to the basic problem of trying to determine what values a given form $Q(x, y)$ can take on, or in different terminology, determining which numbers are represented by $Q$. Remember that it suffices to restrict attention to the values in the topograph since these are the values for primitive pairs $(x, y)$, and to get all possible values one just multiplies the values in the topograph by arbitrary squares. We focus on the forms that are either elliptic or hyperbolic, as these are the most interesting cases.

As we will see through a series of examples, the type of answer one gets to the representation problem varies from quite simple to slightly complicated to quite complicated indeed.

The First Level of Complexity

As a first example let us try to find a general pattern in the values of the form $x^2 + y^2$. In view of the symmetry of the topograph for this form it suffices to look just in the first quadrant of the topograph. A piece of this quadrant is shown in the figure at the right, somewhat distorted to squeeze more numbers into the picture. What is shown is all the numbers in the topograph that are less than 100. At first glance it seems hard to find any patterns here, but the key is to look at how the numbers in the topograph factor into primes. First of all, the primes that occur in the topograph are $2, 5, 13, 17, 29, 37, 41, 53, 61, 73, 89, 97$. Apart from 2, these are just the primes congruent to 1 modulo 4. The remaining primes are congruent to 3 modulo 4, namely, $3, 7, 11, 19, 23, 31, 43, 47, 59, 71, 79, 83$, and these do not appear in the topograph. Moreover, all the numbers in the topograph that are not prime are products of primes in the topograph: $10, 25, 26, 34, 50, 58, 65, 74, 82, 85$. If we remember that the topograph only shows the values of $Q(x, y)$ for primitive pairs $(x, y)$, this means that the remaining values of $Q(x, y)$ are obtained from those in the topograph by multiplying by an arbitrary square $m^2$. Thus we are led to predict that the following result might be true:
Theorem of Fermat. The values of the quadratic form \( Q(x, y) = x^2 + y^2 \) as \( x \) and \( y \) range over all integers are exactly the numbers of the form \( m^2p_1p_2 \cdots p_k \) where \( m \) is an arbitrary integer and each \( p_i \) is either 2 or a prime congruent to 1 modulo 4.

Here we allow the possibility that the number of prime factors \( p_i \) in \( m^2p_1p_2 \cdots p_k \) is zero, so the number represented by \( Q \) is simply \( m^2 \), which is \( Q(m, 0) \). We will prove this theorem later in this section.

As a second example consider the form \( Q(x, y) = x^2 + 2y^2 \). Here is a portion of its topograph showing all values less than 100 again:

Inspecting the values here, we see that the following two statements appear to be true:

1. The prime numbers that occur as values of \( x^2 + 2y^2 \) are 2 and the primes congruent to 1 or 3 modulo 8. In the part of the topograph shown these are 3, 11, 17, 19, 41, 43, 59, 67, 73, 83, 89, 97. The remaining primes are congruent to 5 or 7 modulo 8 and these do not occur as values of \( x^2 + 2y^2 \).

2. The values of \( x^2 + 2y^2 \) are exactly the numbers that can be expressed as products \( m^2p_1p_2 \cdots p_k \) where \( m \) is an arbitrary integer and each \( p_i \) is a prime values of \( x^2 + 2y^2 \) as in (1).

These statements are in fact true and were also known to Fermat.

These two examples were elliptic forms, but the same sort of behavior can occur for hyperbolic forms, as we see in the next example, the form \( x^2 - 2y^2 \). The negative values of this form happen to be just the negatives of the positive values, so we need only show the positive values in the topograph:
Here the primes that occur are 2 and primes congruent to 1 or 7 modulo 8. We can count the negative of a prime number as a prime as well, and then the primes represented are $\pm 2$ and the primes congruent to $\pm 1$ modulo 8. The nonprime values are the products of the primes represented and squares times these numbers.

In these three examples the crucial idea was to look at prime factorizations and at primes modulo certain numbers, the numbers 4, 8, and 8 in the three cases. Notice that these numbers are just the absolute values of the discriminants $-4$, $-8$, and 8 in the three cases. Looking at primes modulo $|\Delta|$ turns out to be a key idea for all quadratic forms, as we will see.

A special feature of the discriminants $-4$, $-8$, and 8 is that all forms of each of these discriminants are equivalent, or in other words, the class numbers are 1 for these discriminants. It is a general fact that whenever the class number is 1, the representation problem has the same sort of simple answer as in the examples above.
The Second Level of Complexity

An example with slightly more complicated behavior is the form $x^2 - 10y^2$. Here is a portion of its topograph showing all the positive values less than 100:

$$Q(x, y) = x^2 - 10y^2$$

There is no need to show any more of the negative values since these will just be the negatives of the positive values. The prime values less than 100 are $31, 41, 71, 79, 89$. These are the primes congruent to $\pm 1$ or $\pm 9$ modulo 40, the discriminant. However, in contrast to what happened in the previous examples, there are many nonprime values that are not products of these prime values. In fact these nonprime values are products of the primes $2, 3, 5, 13, 37, 43$, none of which occur as a value of the form. Rather miraculously, these prime values are realized instead by another form $2x^2 - 5y^2$ having the same discriminant as $x^2 - 10y^2$. Here is the topograph of this companion form $2x^2 - 5y^2$:

$$Q(x, y) = 2x^2 - 5y^2$$

The prime values this form takes on are 2 and 5, which are the prime divisors of the discriminant 40, along with primes congruent to $\pm 3$ and $\pm 13$ modulo 40, namely $3, 13, 37, 43, 53, 67, 83$.

Apart from the primes 2 and 5 that divide the discriminant 40, the possible values of primes modulo 40 are $\pm 1, \pm 3, \pm 7, \pm 9, \pm 11, \pm 13, \pm 17, \pm 19$ since even numbers and multiples of 5 are excluded. There are 16 different congruence classes here, and
exactly half of them, 8, are realized by one or the other of the two forms \( x^2 - 10y^2 \) and \( 2x^2 - 5y^2 \), with 4 classes realized by each form. The other 8 congruence classes are not realized by any form of discriminant 40 since every form of discriminant 40 is equivalent to one of the two forms \( x^2 - 10y^2 \) or \( 2x^2 - 5y^2 \), as is easily checked by the methods from the previous section.

This is in fact a general phenomenon, valid for all discriminants: If one looks at primes that do not divide the discriminant, then the prime values of quadratic forms of that discriminant are exactly the primes in one-half of the possible congruence classes modulo the discriminant.

Let us mention in passing a famous theorem of Dirichlet, proved in the 1820s or 1830s, which says that every arithmetic progression \( a, a + d, a + 2d, a + 3d, \cdots \) contains infinitely many primes, provided that one rules out the obvious exceptions where \( a \) and \( d \) have a common divisor, which would then be a common divisor of all the numbers in the progression. For example, when we take \( d = 40 \), each of the 16 congruence classes listed above gives an arithmetic progression containing infinitely many primes, such as the progression 1, 41, 81, 121, 161, 201, \( \cdots \) or the progression 17, 57, 97, 137, 177, 217, \( \cdots \). In fact Dirichlet proved more: If one looks at primes less than some large number \( N \) such as a million, then each of the possible congruence classes contains approximately the same number of primes less than \( N \).

The analog of Fermat’s Theorem for discriminant 40 is the following pair of statements:

1. The numbers represented by one of the two quadratic forms \( Q_1 = x^2 - 10y^2 \) or \( Q_2 = 2x^2 - 5y^2 \) of discriminant 40 are exactly the numbers \( n = \pm m^2p_1p_2\cdots p_k \) where \( m \) is an arbitrary integer and each \( p_i \) is 2, 5, or a prime congruent to \( \pm 1, \pm 3, \pm 9 \), or \( \pm 13 \) modulo 40.

2. If the number of factors \( p_i \) in \( n = \pm m^2p_1p_2\cdots p_k \) that equal 2, 5, or \( \pm 3 \) or \( \pm 13 \) modulo 40 is even, then \( n \) is represented by \( Q_1 \), and if this number is odd then \( n \) is represented by \( Q_2 \). In particular, the primes represented by \( Q_1 \) are the primes congruent to \( \pm 1 \) or \( \pm 9 \) modulo 40 and the primes represented by \( Q_2 \) are 2, 5, and primes congruent to \( \pm 3 \) or \( \pm 13 \) modulo 40.

An interesting consequence of (2) is that no number \( n \) is represented by both forms \( x^2 - 10y^2 \) and \( 2x^2 - 5y^2 \), apart from \( n = 0 \) which is trivially represented by every form \( Q(x,y) \) when \( (x,y) = (0,0) \).

Another case which is similar to the preceding one is discriminant 12. Here there are two forms up to equivalence, \( x^2 - 3y^2 \) and \( 3x^2 - y^2 \), which is equivalent to
\[-x^2 + 3y^2\], the negative of the first form. Here is the topograph for \(x^2 - 3y^2\):

\[Q(x, y) = x^2 - 3y^2\]

For the form \(-x^2 + 3y^2\) we get the negatives of the numbers represented by \(x^2 - 3y^2\).

For discriminant 12 we have the following answer to the representation problem:

The numbers represented by one of the two quadratic forms \(x^2 - 3y^2\) or \(-x^2 + 3y^2\) of discriminant 12 are exactly the numbers \(n = m^2 p_1 p_2 \cdots p_k\) where \(m\) is an arbitrary integer and each \(p_i\) is \(-1, 2, 3\), or a prime congruent to \(\pm 1\) modulo 12. If the number of factors \(p_i\) equal to \(-1, 2, 3\), or congruent to \(-1\) modulo 12 is even, then \(n\) is represented by the form \(x^2 - 3y^2\), and if this number is odd then \(n\) is represented by \(-x^2 + 3y^2\).

The Third Level of Complexity

In the preceding examples it was possible to determine which numbers are represented by a given form by looking at primes and which congruence classes they fall into modulo the discriminant. A consequence of the way the answer was formulated was that no number (except 0) could by represented by two inequivalent forms of the same discriminant. Both of these nice properties fail to hold in general, however. An example is provided by forms of discriminant \(-56\). Two inequivalent forms of this discriminant are \(Q_1 = x^2 + 14y^2\) and \(Q_2 = 2x^2 + 7y^2\). The primes 23 and 79 are congruent modulo 56, and yet 23 is represented by \(Q_1\) since \(Q_1(3, 1) = 23\), while 79 is represented by \(Q_2\) since \(Q_2(6, 1) = 79\). Also, 30 is represented by both \(Q_1\) and \(Q_2\) since \(Q_1(4, 1) = 30\) and \(Q_2(1, 2) = 30\).

The class number for discriminant \(-56\) is actually 3, and a third form with this discriminant, not equivalent to either \(Q_1\) or \(Q_2\), is the form \(Q_3 = 3x^2 + 2xy + 5y^2\). Here are portions of the topographs of these three forms:
Apart from the primes 2 and 7 that divide the discriminant $-56$, all other primes belong to the following 24 congruence classes modulo 56, corresponding to odd numbers less than 56 not divisible by 7:

$$1, 3, 5, 9, 11, 13, 17, 19, 23, 25, 27, 29, 31, 33, 37, 39, 41, 43, 45, 47, 51, 53, 55$$

The six congruence classes whose prime elements are represented by $Q_1$ or $Q_2$ are indicated by underlines, and the six congruence classes whose prime elements are
represented by $Q_3$ are indicated by overlines. Primes not represented by any of the three forms are in the remaining 12 congruence classes.

As general terminology, one says that two quadratic forms of the same discriminant belong to the same genus if they cannot be distinguished by considering their values modulo the discriminant. Thus the preceding two forms $Q_1$ and $Q_2$ are of the same genus while $Q_3$ is of a different genus from $Q_1$ and $Q_2$. Equivalent forms always belong to the same genus, of course. The first two of the three levels of complexity we have described correspond to the discriminants where there is only one equivalence class in each genus. For discriminant $-56$ there are two different genera (“genera” is the plural of “genus”). In more complicated examples there can be large numbers of genera and large numbers of equivalence classes within a genus.

For negative discriminants there is a simple formula for the number of genera of forms of a given discriminant $\Delta$, namely, the number of genera is $2^{k-1}$ where $k$ is the number of distinct prime divisors of $|\Delta|$.

**A Criterion for Representability**

Suppose a number $n$ is represented primitively by some form $Q(x, y)$ of discriminant $\Delta$, so $n$ appears in the topograph of $Q$. If we look at an edge of the topograph bordering the region labeled $n$ then we obtain an equation $\Delta = h^2 - 4nk$ where $h$ is the label on the edge and $k$ is the label on the region on the opposite side of this edge. The equation $\Delta = h^2 - 4nk$ says that $\Delta$ is congruent to $h^2$ modulo $4n$. In case $n$ is negative we interpret “modulo $4n$" to mean “modulo $4|n|$", but for the sake of simplicity we will still write “modulo $4n$".

This in fact gives an exact criterion for primitive representability:

**Proposition.** Let two numbers $n$ and $\Delta$ be given. Then the following two statements are equivalent: (1) There exists a form of discriminant $\Delta$ that represents $n$ primitively. (2) $\Delta$ is congruent to a square modulo $4n$.

**Proof:** As we saw above, if we have a form of discriminant $\Delta$ representing $n$ primitively then we get an equation $\Delta = h^2 - 4nk$ for some integers $h$ and $k$, and this equation says that $\Delta$ is the square of $h$ modulo $4n$. Conversely, suppose that $\Delta$ is the square of some integer $h$ modulo $4n$. This means that $h^2 - \Delta$ is an integer times $4n$, or in other words $h^2 - \Delta = 4nk$ for some $k$. This equation can be rewritten as $\Delta = h^2 - 4nk$. The three numbers $n$, $h$, and $k$ can be used to construct a form whose topograph contains an edge with these three labels, for example the form $Q(x, y)$.
Let us see what this proposition implies for small values of \( n \). For \( n = 1 \) it says that there is a form of discriminant \( \Delta \) representing 1 if and only if \( \Delta \) is a square modulo 4. The squares modulo 4 are 0 and 1, and we already know that discriminants of forms are always congruent to 0 or 1 modulo 4. So we conclude that for every possible value of the discriminant there exists a form that represents 1. This isn’t really new information, however, since the principal form \( x^2 + dy^2 \) or \( x^2 + xy + dy^2 \) represents 1 and there is a principal form for each discriminant.

In the next case \( n = 2 \) we will get some new information. The possible values of the discriminant modulo 8 are 0, 1, 4, 5, and the squares modulo 8 are 0, 1, 4 since \( 0^2 = 0 \), \( (\pm 1)^2 = 1 \), \( (\pm 2)^2 = 4 \), \( (\pm 3)^2 \equiv 1 \), and \( (\pm 4)^2 \equiv 0 \). Thus 2 is not represented by any form of discriminant congruent to 5 modulo 8, but for all other values of the discriminant there is a form representing 2. Explicit forms are:

\[
\begin{align*}
\Delta &= 8k : \\
2x^2 - ky^2
\end{align*}
\]

\[
\begin{align*}
\Delta &= 8k + 1 : \\
2x^2 + xy - ky^2
\end{align*}
\]

\[
\begin{align*}
\Delta &= 8k + 4 : \\
2x^2 + 2xy - ky^2
\end{align*}
\]

For \( n = 3 \) the discriminants modulo 12 are 0, 1, 4, 5, 8, 9 and the squares modulo 12 are 0, 1, 4, 9 since \( 0^2 = 0 \), \( (\pm 1)^2 = 1 \), \( (\pm 2)^2 = 4 \), \( (\pm 3)^2 = 9 \), \( (\pm 4)^2 \equiv 4 \), \( (\pm 5)^2 \equiv 1 \), and \( (\pm 6)^2 \equiv 0 \). The excluded discriminants are those congruent to 5 or 8.

One could continue farther in this direction exploring which discriminants have forms representing a given number, but this is not really the question we want to answer, which is to start with a given discriminant, or even a given form, and decide which numbers it represents. The sort of answer we are looking for, based on the various examples we looked at earlier, is also a different sort of congruence condition, with congruence modulo the discriminant rather than congruence modulo \( 4n \). So there is more work to be done before we would have the sort of answer we want. Nevertheless, the representability criterion in the preceding proposition is the starting point.

**Proof of Fermat’s Theorem**

Without a huge amount of extra work we can now settle the simplest case, the form \( x^2 + y^2 \). Recall that the statement to be proved is that a number \( n \) is representable by the form \( x^2 + y^2 \) if and only if it can be written as \( n = m^2 p_1 \cdots p_k \) where each
Chapter 2 Quadratic Forms

$p_i$ is either 2 or a prime congruent to 1 modulo 4. The possibility that $n$ is simply $m^2$ is allowed. Another way of stating the condition on $n$ is to say that every prime factor $p = 4k + 3$ of $n$ occurs to an even power, so it can be absorbed into the $m^2$ factor.

As a preliminary step to proving this, let us look at the condition in the preceding proposition for a number $n$ to be primitively represented by $x^2 + y^2$, which is that the discriminant $-4$ is a square modulo $4n$. This means that we have an equation $h^2 = -4 + 4nk$ for some integers $h$ and $k$. The number $h$ must be even for this equation to hold, so $h = 2l$ and the equation becomes $4l^2 = -4 + 4nk$. Canceling the 4's, this becomes $l^2 = -1 + nk$. This just says that $-1$ is a square modulo $n$, so this is our new criterion for $n$ to be primitively represented by $x^2 + y^2$.

Now we begin the proof proper. First we show that if $n$ is represented by the form $x^2 + y^2$ then every prime factor $p = 4k + 3$ of $n$ must occur to an even power in $n$. Suppose on the contrary that we have a number $n$ represented by $x^2 + y^2$ whose prime factorization has a prime factor $p = 4k + 3$ occurring to an odd power. If this $n$ is not primitively represented, we can cancel square factors of it until we get a new $n$ represented primitively, with the same prime $p$ still occurring to an odd power. Our representability criterion then says that $-1$ is a square modulo $n$. Since the prime $p = 4k + 3$ is a factor of $n$, this implies that $-1$ is a square modulo $p$. Applying the representability criterion in the reverse direction, this implies that $p$ is represented (primitively) by $x^2 + y^2$, so $p = x^2 + y^2$ for some $x$ and $y$. Since $p = 4k + 3$, if we look at the equation $p = x^2 + y^2$ modulo 4, it says that 3 is the sum of two squares modulo 4. But this is impossible since the squares modulo 4 are 0 and 1 so adding two of them cannot give 3. This contradiction shows that primes $p = 4k + 3$ must occur in the prime factorization of $n$ to an even power whenever $n$ is represented by $x^2 + y^2$.

To finish the proof of the theorem it suffices to prove two things:

(1) The primes $p = 2$ and $p = 4k + 1$ are represented by $x^2 + y^2$. (This is obvious for 2.)

(2) If two numbers $m$ and $n$ are represented by $x^2 + y^2$ then so is their product $mn$.

The second statement is easier to prove so we do this first. Suppose $m$ and $n$ are represented as $m = a^2 + b^2$ and $n = c^2 + d^2$. Using complex numbers we can then factor $m$ and $n$ as $m = (a + bi)(a - bi)$ and $n = (c + di)(c - di)$. This gives a factorization of $mn$ as a product of four factors, and by rearranging the factors we
can obtain a representation of $mn$ as a sum of two squares:

$$mn = (a^2 + b^2)(c^2 + d^2)$$

$$= [(a + bi)(a - bi)][(c + di)(c - di)]$$

$$= [(a + bi)(c + di)][(a - bi)(c - di)]$$

$$= [(ac - bd) + (ad + bc)i][(ac - bd) - (ad + bc)i]$$

$$= (ac - bd)^2 + (ad + bc)^2$$

The result of this calculation is the identity

$$(a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2$$

which shows that the product of two numbers that are sums of two squares is again a sum of two squares. This identity can be checked directly without using complex numbers, just by multiplying both sides out, but the advantage of using complex numbers is that they show where the identity comes from.

It remains to prove the nontrivial part of the earlier statement (1), that every prime $p = 4k + 1$ is representable as the sum of two squares. Such a representation has to be primitive since $p$ is prime. An equivalent statement is then that $-1$ is a square modulo $p$, and this is what we will show by finding an explicit but rather large number $h$ such that $h^2 \equiv -1$ modulo $p$.

Let us first illustrate how the proof will go by doing a specific example, the case $p = 13$, which is of the form $4k + 1$. Each of the numbers from 1 to $p - 1 = 12$ has a multiplicative inverse modulo 13:

$$1 \cdot 1 \equiv 1 \quad 2 \cdot 7 \equiv 1 \quad 3 \cdot 9 \equiv 1 \quad 4 \cdot 10 \equiv 1 \quad 5 \cdot 8 \equiv 1 \quad 6 \cdot 11 \equiv 1 \quad 12 \cdot 12 \equiv 1$$

The last congruence could have been written $(-1) \cdot (12) \equiv 1$. The only cases when a number equals its own inverse modulo 13 are 1 and 12. Therefore if we consider the product


all the terms that are not equal to their inverse will cancel in pairs, leaving only the last term 12. Thus we have the congruence $12! \equiv 12$ modulo 13, which we can rewrite as $12! \equiv -1$. Now notice that modulo 13 we have $7 \equiv -6$, $8 \equiv -5$, $9 \equiv -4$, etc., so we have


$$\equiv (1)(2)(3)(4)(5)(6)(-6)(-5)(-4)(-3)(-2)(-1)$$

$$= (6!)^2$$
This shows that \(-1\) is a square modulo 13, namely \((6!)^2\). We will generalize this by showing that for every prime \(p = 4k + 1\) the congruence \(-1 \equiv [(2k)!]^2\) modulo \(p\) holds.

The first fact we need about congruences modulo a prime \(p\) is that each of the numbers \(a = 1, 2, \ldots, p - 1\) has a multiplicative inverse modulo \(p\). To see why this is true, notice that each such \(a\) has no common factors with \(p\), so we know from Chapter 1 that the equation \(ax + py = 1\) has an integer solution \((x, y)\). This equation can be rewritten as \(ax \equiv 1\) modulo \(p\), which says that \(x\) is an inverse for \(a\) modulo \(p\). Note that any two choices for \(x\) here are congruent modulo \(p\) since if \(ax \equiv 1\) and \(ax' \equiv 1\) then multiplying both sides of \(ax' \equiv 1\) by \(x\) gives \(xax' \equiv x\), and \(xa \equiv 1\) so we conclude that \(x \equiv x'\).

Which numbers equal their own inverse modulo \(p\)? If \(a \cdot a \equiv 1\), then we can rewrite this as \(a^2 - 1 \equiv 0\), or in other words \((a + 1)(a - 1) \equiv 0\). This is certainly a valid congruence if \(a \equiv \pm 1\), so suppose that \(a \not\equiv \pm 1\). The factor \(a + 1\) is then not congruent to 0 modulo \(p\) so it has a multiplicative inverse modulo \(p\), and if we multiply the congruence \((a + 1)(a - 1) \equiv 0\) by this inverse, we get \(a - 1 \equiv 0\) so \(a \equiv 1\), contradicting the assumption that \(a \not\equiv \pm 1\). This argument shows that the only numbers among 1, 2, \ldots, \(p - 1\) that are congruent to their inverses modulo \(p\) are 1 and \(p - 1\).

Now if we consider the product \((p - 1)! = (1)(2) \cdots (p - 1)\) modulo \(p\), then each factor other than 1 and \(p - 1\) can be paired up with its multiplicative inverse and these two terms multiply together to give 1 modulo \(p\), so the whole product simplifies to just \((1)(p - 1)\). Thus we have a fact known as Wilson’s theorem:

\((p - 1)! \equiv -1\) modulo \(p\) whenever \(p\) is prime.

Now let us assume that \(p\) is a prime of the form \(p = 4k + 1\). In the product \((p - 1)!\) there are \(p - 1 = 4k\) terms. The first 2\(k\) of these are \((2k)!\) and the last 2\(k\), in reverse order, are \(p - 1, p - 2, \ldots, p - 2k\). Modulo \(p\) the latter are equivalent to \(-1, -2, \ldots, -2k\), so we have

\((p - 1)! = (4k)! \equiv (1)(2) \cdots (2k - 1)(2k)(-2k)(-(2k - 1)) \cdots (-2)(-1)\)

The last 2\(k\) of these factors are the negatives of the first 2\(k\) factors, and 2\(k\) is even, so the signs on all the negative terms cancel out and we see that \((p - 1)!\) is congruent to \((2k)! \cdot (2k)!\) modulo \(p\). Combining this with Wilson’s theorem we get the desired result that \(-1\) is a square modulo \(p\), namely \(-1 \equiv [(2k)!]^2\) modulo \(p\). This finishes the proof of Fermat’s theorem answering the question of which numbers are representable as sums of two squares.
Quadratic Reciprocity

We have seen that the condition for a prime \( p \) to be represented by some form of discriminant \( \Delta \) is that \( \Delta \) is a square modulo \( 4p \). (For primes there is no need to add the condition that the representation is primitive since this is automatic for numbers with no square factors, in particular for primes.) What we need is a way to convert this criterion from a condition on \( \Delta \) modulo \( 4p \) to a condition on \( p \) modulo \( \Delta \). The main tool to make this conversion is something called Quadratic Reciprocity. Here is what it says:

**Quadratic Reciprocity.** Let \( p \) and \( q \) be two distinct odd primes. If \( p \) and \( q \) are not both congruent to 3 modulo 4, then \( p \) is a square modulo \( q \) if and only if \( q \) is a square modulo \( p \). In the exceptional case that \( p \) and \( q \) are both congruent to 3 modulo 4, then \( p \) is a square modulo \( q \) if and only if \( q \) is not a square modulo \( p \).

This is a surprising statement because congruences modulo \( p \) and congruences modulo \( q \) generally have nothing to do with each other when \( p \) and \( q \) have no common factors, as is the case here with distinct primes. It is only on a special question like whether or not a number is a square modulo \( p \) or \( q \) that there is a connection. The principle of quadratic reciprocity was discovered in a more rudimentary form by Euler, and the final version given above was formulated by Legendre. The first person to prove it, however, was Gauss around 1797 or 1798. This is well over 100 years after Fermat, and it seems Fermat was not aware of quadratic reciprocity. Even though Gauss’s proof was correct, he was not satisfied that he understood quadratic reciprocity completely and later produced several different proofs. Since that time, many other proofs have been found. Many of them are elementary in the sense that they can be presented in a couple pages with no more background than we are assuming here. Some of the proofs involve a little geometry, but it is still not easy to really “see” why the result is true.

Let us do an example to see how quadratic reciprocity applies to the representability problem. This will be the case of discriminant 13, a prime, which simplifies the use of quadratic reciprocity. Our general theory tells us that a prime \( p \) is represented by some form of discriminant 13 if and only if 13 is a square modulo \( 4p \). For \( p = 2 \) this says that 13 is a square modulo 8, which is not true since the squares modulo 8 are 0,1,2,4. Thus 2 is not represented by any form of discriminant 13.

From now on we assume that \( p \) is odd. If 13 is a square modulo \( 4p \), it is certainly a square modulo \( p \) since \( p \) divides \( 4p \). We apply quadratic reciprocity now. Since 13 is not congruent to 3 modulo 4, reciprocity says that 13 is a square modulo \( p \) if
and only if \( p \) is a square modulo 13. The squares modulo 13 are easily listed: 0, 1, 4, 9, 16 \( \equiv 3 \), 25 \( \equiv 12 \), and 36 \( \equiv 10 \). (There is no need to go farther since 7 \( \equiv -6 \), 8 \( \equiv -5 \), etc.) Thus if \( p \) is represented by a form of discriminant 13, \( p \) is congruent to one of 0, 1, 3, 4, 9, 10, 12 modulo 13. The only prime congruent to 0 modulo 13 is 13 itself, so we can say that the primes represented by forms of discriminant 13 must be either 13 or primes congruent to one of 1, 3, 4, 9, 10, 12 modulo 13, or in other words, \( \pm 1, \pm 3, \pm 4 \) modulo 13. Each of these six congruence classes contains infinitely many primes by Dirichlet’s theorem on primes in arithmetic progressions.

The converse is also true, that every prime satisfying these conditions is actually represented by some form of discriminant 13. All the steps in the reasoning above were reversible except for the step of going from 13 being a square modulo 4 \( p \) to 13 being a square modulo \( p \). In fact this step is reversible too, for suppose 13 is a square modulo \( p \), so there is a number \( h \) such that \( h^2 - 13 \) is divisible by \( p \). We can assume \( h \) is odd since if it is even we can replace \( h \) by \( h + p \), which is odd since \( p \) is odd, and the new \( h \) will still satisfy \( h^2 \equiv 13 \) modulo \( p \). Since \( h \) is odd, we have \( h^2 \equiv 1 \) modulo 4. Since 1 \( \equiv 13 \), this can be rephrased as saying \( h^2 \equiv 13 \) modulo 4, which means that \( h^2 - 13 \) is divisible by 4. We already knew that \( h^2 - 13 \) was divisible by \( p \), so \( h^2 - 13 \) is divisible by 4\( p \) since 4 and \( p \) have no common factors. Thus we have shown that 13 is a square modulo 4\( p \) if it is a square modulo \( p \).

This completes the characterization of the primes that are representable by some form of discriminant 13, assuming that quadratic reciprocity is known. As it happens, the class number for discriminant 13 is one, as you can easily verify, so all forms of discriminant 13 are equivalent to the principal form \( x^2 + xy - 3y^2 \) and so we have an exact criterion for which primes this form represents: \( p = 13 \) and primes congruent to \( \pm 1, \pm 3, \pm 4 \) modulo 13. One could predict this was true by drawing a large enough part of the topograph, but a full proof requires more than this since it is obviously impossible to draw the whole topograph all at once and check all the infinitely many primes that occur in it. (For one thing, there is the difficulty of knowing when a large number is a prime.)

The full answer to which numbers, not just primes, are represented by the form \( x^2 + xy - 3y^2 \) is what you would now expect: The numbers represented are the numbers \( n = m^2 p_1 \cdots p_k \) where each \( p_i \) is 13 or a prime congruent to \( \pm 1, \pm 3, \) or \( \pm 4 \) modulo 13. The rest of the proof of this follows the same pattern as in the proof we gave for Fermat’s theorem on the form \( x^2 + y^2 \). The only new ingredient needed is a formula showing that the product of two numbers represented by \( x^2 + xy - 3y^2 \) is also represented by this form. The formula is not difficult and we will derive it in
the next chapter for all principal forms.

For other values of the discriminant $\Delta$ the procedure for determining the primes that are represented by some form of discriminant $\Delta$ is similar to the case $\Delta = 13$, but when $\Delta$ is not itself a prime there are some extra steps required to go from quadratic reciprocity, which is just a statement about pairs of primes, to determining when a nonprime $\Delta$ is a square modulo $4p$. The extra steps are not as difficult as quadratic reciprocity, however. With more effort of a similar nature one can go farther and answer the question of which numbers, not necessarily primes, are represented by some form of a given discriminant. It is only when the class number is one, however, that this answers the question of which numbers a particular form represents.