5 The Classification of Quadratic Forms

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

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

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

(III) If $Q(x,y)$ takes on only positive values or only negative values then we call Q an ***elliptic*** form.

(IV) If Q takes on the value 0 and either positive or negative values, but not both, then Q is called a ***parabolic*** form.

The hyperbolic-elliptic-parabolic terminology is motivated in part by what the level curves $ax^2 + bxy + cy^2 = k$ are when we 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. From the classical perspective of conic sections parabolas are the transitional case between hyperbolas and ellipses, but from another viewpoint one can pass from hyperbolas to ellipses through a transitional case of a pair of parallel lines as in the family of curves $x^2 - cy^2 = 1$ which are hyperbolas for $c > 0$, ellipses for $c < 0$, and a pair of parallel lines for $c = 0$. Parabolic forms are much simpler than the other types and we will not be spending much time on them.

As we will show later in the chapter, there is an easy way to distinguish the four types of forms $ax^2 + bxy + cy^2$ in terms of their discriminants $\Delta = b^2 - 4ac$:

(I) If Δ is positive but not a square then Q is hyperbolic.

(II) If Δ is positive and a square then Q is 0-hyperbolic.

(III) If Δ is negative then Q is elliptic.

(IV) If Δ is zero then Q is parabolic.

Discriminants play a central role in the theory of quadratic forms. A natural question to ask is whether every integer occurs as the discriminant of some form, and this is easy to answer. 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 Δ congruent to 0 or 1 mod 4 there exists a form whose discriminant is Δ. The simplest ones are:

$$x^2 - ky^2 \text{ with discriminant } \Delta = 4k$$

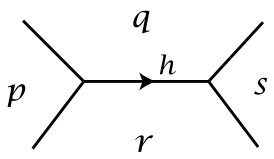
$$x^2 + xy - ky^2 \text{ with discriminant } \Delta = 4k + 1$$

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.

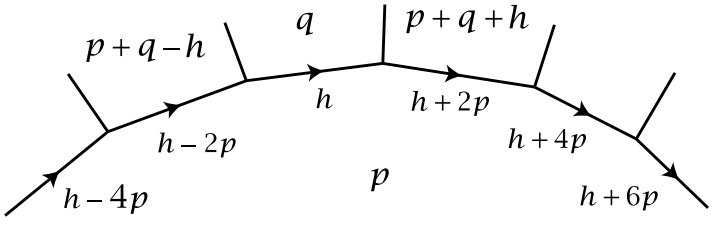
5.1 The Four Types of Forms

We will analyze each of the four types of forms in turn, but before doing this let us make a few preliminary observations that apply to all forms.

In the arithmetic progression rule controlling the labeling of the four regions surrounding an 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 the orientation of the edge is irrelevant and can be omitted.



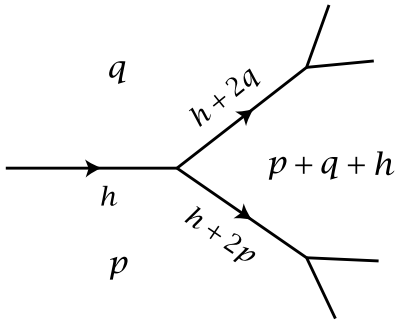
The values of the increment h along the boundary of a region in the topograph have the interesting property that they also form an arithmetic progression when all these edges are oriented in the same direction, and the amount by which h increases as we move from one edge to the next is $2p$ where p is the label on the region adjacent to all these edges:



We will call this property the ***second arithmetic progression rule***. To see why it holds, 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. From this we can deduce that the label on the edge between the regions labeled p and $p + q + h$ must be $h + 2p$ since this is the increment from q to

$p+(p+q+h)$. 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***. This says that in the figure at the right, 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 going 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 numbers.

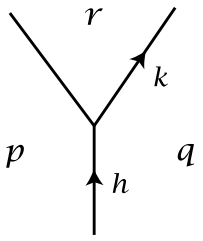


Next we have a very useful way to compute the discriminant of a form directly from its topograph:

Proposition 5.1. *If an edge in the topograph of a form $Q(x,y)$ is labeled h with adjacent regions labeled p and q, then the discriminant of $Q(x,y)$ is h^2-4pq.*

Note that the sign of h and the orientation of the edge are irrelevant here. The proposition implies that if the discriminant is known then any two of p, q, and $|h|$ determine the third.

Proof: For the given form $Q(x,y)=ax^2+bxy+cy^2$, the $1/0$ and $0/1$ regions 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. We have $r=p+q+h$, and from the second arithmetic progression rule we know that $k=h+2q$. Then 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. □

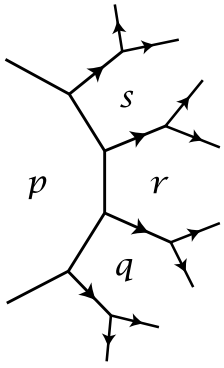


Elliptic Forms

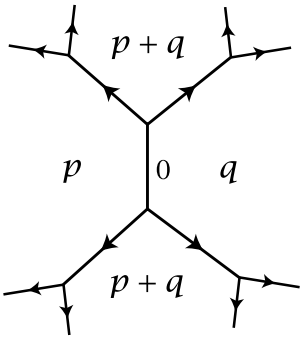
Elliptic forms have fairly simple qualitative behavior, so let us look at these forms first. Recall that we defined a form $Q(x,y)$ to be elliptic if it takes on only positive or only negative values at all integer pairs $(x,y)\neq(0,0)$. The positive and negative cases are equivalent since one can switch 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 always assume we are in this case whenever we are dealing with

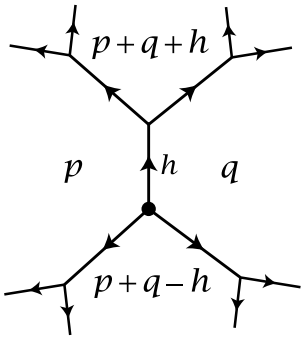
elliptic forms. We will also generally assume when we look at topographs of elliptic forms that the orientations of the edges are chosen so as to give positive h-values, unless we state otherwise.

For a positive elliptic form Q let p be the minimum positive value taken on by Q, so $Q(x,y) = p$ for some $(x,y) \neq (0,0)$. Here (x,y) must be a primitive pair otherwise Q would take on a smaller positive value than p. Thus there is a region in the topograph of Q with the label p. All the edges having one endpoint at this region must be 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 steadily as one moves away from the p region.



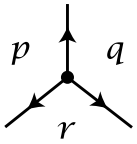
For the edges making up the border of the p region we know that the h-labels on these edges form an arithmetic progression with increment $2p$, provided that we temporarily re-orient these edges so that they all point in the same direction. If some edge bordering the p region has the label $h = 0$ then the topograph has the form shown in the first figure below, with the orientations on edges that give positive h-labels. 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.





The other possibility is that no edge bordering the p region has label $h = 0$. Then since the labels on these edges form an arithmetic progression, there must be some vertex where the terms in the progression change sign. Thus when we orient the edges to give positive h-labels, 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.

If the three regions surrounding a source vertex are labeled p, q, r then the fact that the three edges leading from this vertex all point away from the vertex is equivalent to the three inequalities $p < q + r$, $q < p + r$, and $r < p + q$. These are called triangle inequalities since they are satisfied by the lengths of the three sides of any triangle. In the case of a source edge one of the inequalities becomes an equality, for example $r = p + q$ in the earlier figure with



a source edge.

As we know, any three integers p, q, r can be realized as the three labels surrounding a vertex in the topograph of some form. If these are positive integers satisfying the triangle inequalities then this vertex is the source vertex of an elliptic form since these inequalities imply that the three edges at this vertex are oriented away from the vertex, so the monotonicity property guarantees that all values of the form are positive. The situation for source edges is simpler since any two positive integers p and q determine an elliptic form with a source edge having adjacent regions labeled p and q as in the earlier figure.

Hyperbolic Forms

The topographs of hyperbolic forms exhibit quite different behavior from the topographs of elliptic forms since they always have a ***periodic separator line*** of the sort that we saw in several of the examples in the previous chapter. Here is the general statement:

Theorem 5.2. *In the topograph of a hyperbolic form the edges 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, with other edges of the topograph leading off the line on both sides.*

Proof: For a hyperbolic form Q all regions in the topograph have labels that are either positive or negative, never zero, and there must exist two regions of opposite sign. By moving along a path in the topograph joining two such 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***.

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:

This implies that exactly one of the two edges at each 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 line are not separating edges.

The monotonicity property implies that as we move off this line of separating edges the values of Q are steadily increasing through positive integers on the positive side and steadily decreasing through negative integers on the negative side. In particular this means that there are no other separating edges that are not on the initial separator line, so there is only one separator line.

It remains to prove that the topograph is periodic along the separator line. We can assume all the edges along the separator line are oriented in the same direction by changing the signs of the h values if necessary. For an edge of the separator line labeled h with adjacent regions labeled p and $-q$ with $p > 0$ and $q > 0$, we know that $h^2 + 4pq$ is the discriminant Δ, by Proposition 5.1. The equation $\Delta = h^2 + 4pq$ with p and q positive implies that Δ is positive and furthermore that each of $|h|$, p, and q is less than Δ. Thus there are only finitely many possible values for h, p, and q along the separator line since Δ is a constant depending only on Q. It follows that there are only finitely many possible combinations of values h, p, and q at each edge on the separator line. Since the separator line is infinite, there must then 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.

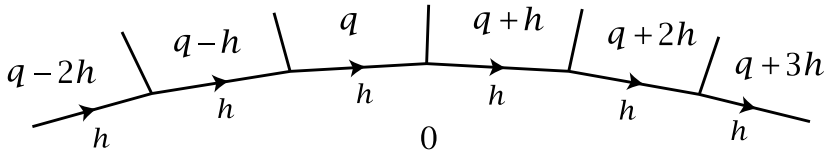
There must be edges leading away from the separator line on both the positive and the negative side, otherwise there would be just a single region on one side of the line, and then the second arithmetic progression rule would say that the h labels along the line formed an infinite arithmetic progression with nonzero increment $2p$ where p is the label on the region in question. However, this would contradict the fact that these h labels are periodic. □

The qualitative behavior of the topograph of a hyperbolic form away from the separator line fits the pattern we have seen in examples. Since the separator line is periodic the whole topograph is periodic, consisting of repeating sequences of trees leading off from the separator line on each side, with monotonically increasing positive values of the form on each tree on the positive side of the separator line and monotonically decreasing negative values on the negative side, as a consequence of the monotonicity property.

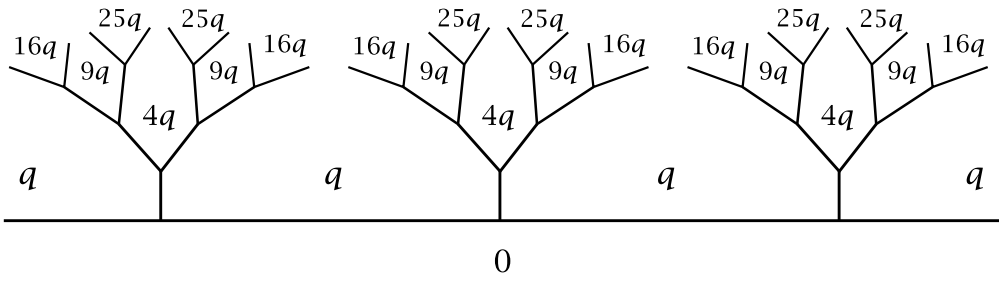
Parabolic and 0-Hyperbolic Forms

The remaining types of forms to consider are parabolic forms and 0-hyperbolic forms. These turn out to be less interesting, and they play only a minor role in the theory of quadratic forms.

Parabolic and 0-hyperbolic forms are the forms whose topograph contains at least one region labeled 0. By the second arithmetic progression rule, each edge adjacent to a 0 region has the same label h, and from this it follows that the labels on the regions adjacent to the 0 region form an arithmetic progression:

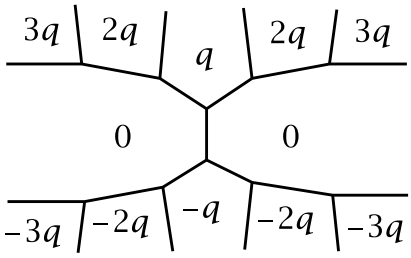


When $h = 0$ the topograph has the very simple pattern shown in the following figure:



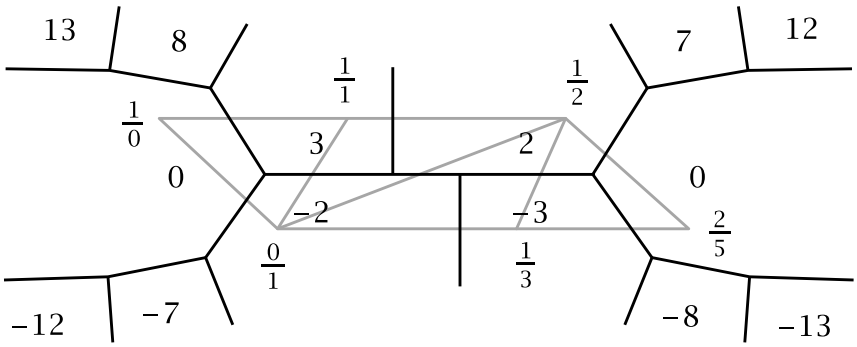
Thus the form is parabolic, taking on only positive or only negative values away from the 0 region, depending on the sign of q. We cannot have $q = 0$ since we are not allowing forms to be identically zero. An example of a form with this topograph is $Q(x, y) = qx^2$, with the 0 region at $x/y = 0/1$. The topograph is periodic along the 0 region since it consists of the same tree pattern repeated infinitely often.

The remaining case is that the label h on the edges bordering a 0 region is nonzero. The arithmetic progression of values of Q adjacent to the 0 region is then not constant, so it includes both positive and negative numbers, and hence 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. An example of a form with this topograph is $Q(x, y) = qxy$, with the two 0 regions at $x/y = 1/0$ and $0/1$.



If the arithmetic progression of values of Q adjacent to the 0 region 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. If this extension does not eventually terminate with a second 0 region, 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, forcing the line to be periodic in the direction of this extension. This in turn would force it to be periodic in both directions by the arithmetic progression rule. But this is impossible since the line began with a 0 region at one end. Thus the topograph contains a finite separator line connecting two 0 regions.

An example of such a form is $Q(x, y) = qxy - py^2 = (qx - py)y$ which has the value 0 at $x/y = 1/0$ and at $x/y = p/q$ or the reduction of p/q to lowest terms if p and q are not coprime. Here we must have $|q| > 1$ for the two 0 regions to be nonadjacent. The separator line must follow the strip of triangles in the Farey diagram corresponding to the continued fraction for p/q since the separator line is dual to a finite strip of triangles with the vertices $1/0$ and p/q at its two ends. For example, for $p/q = 2/5$ the topograph of the form $5xy - 2y^2 = (5x - 2y)y$ is shown in the following figure:



General Conclusions

Having described the topographs of the four types of forms, we can now deduce the characterization of each type in terms of the discriminant:

Proposition 5.3. *The four types of forms are distinguished by their discriminants, which are negative for elliptic forms, positive nonsquares for hyperbolic forms, positive squares for* 0*-hyperbolic forms, and zero for parabolic forms.*

Proof: Consider first an elliptic form Q, which we may assume takes on only positive values since changing Q to $-Q$ does not change the discriminant. The topograph of Q contains either a source vertex or a source edge. For a source edge with the label $h = 0$ separating regions with positive labels p and q the discriminant is $\Delta = h^2 - 4pq = -4pq$, which is negative. For a source vertex with adjacent regions having positive labels p, q, r the edge between the p and q regions is labeled $h = p + q - r$ so the discriminant can be expressed in the following way:

$$\begin{aligned}\Delta = h^2 - 4pq &= (p + q - r)^2 - 4pq \\ &= p^2 + q^2 + r^2 - 2pq - 2pr - 2qr \\ &= p(p - q - r) + q(q - p - r) + r(r - p - q)\end{aligned}$$

In the last line the three quantities in parentheses are negative by the triangle inequalities, so Δ is again negative.

For a parabolic form the topograph contains a region labeled 0 bordered by edges labeled 0, so $\Delta = h^2 - 4pq = 0$. A 0-hyperbolic form has a region labeled 0 bordered by edges all having the same nonzero label h so $\Delta = h^2$, a positive square.

For an edge in the separator line for a hyperbolic form the adjacent regions have labels p and $-q$ with p and q positive so $\Delta = h^2 + 4pq$ is positive. To see that Δ is not a square, suppose the form is $ax^2 + bxy + cy^2$. Here a must be nonzero, otherwise the form would have the value 0 at $(x, y) = (1, 0)$, which is impossible for a hyperbolic form. If the discriminant was a square then the equation $az^2 + bz + c = 0$ would have a rational root $z = x/y$ with $y \neq 0$ by the familiar quadratic formula $z = (-b \pm \sqrt{b^2 - 4ac})/2a$. Thus we would have $a(x/y)^2 + b(x/y) + c = 0$ and hence $ax^2 + bxy + cy^2 = 0$, so the form would have the value 0 at a pair (x, y) with $y \neq 0$, which is again impossible for a hyperbolic form. □

The presence or absence of periodicity in a topograph has the following consequence:

Proposition 5.4. *If an equation $Q(x,y) = n$ with $n \neq 0$ has one integer solution (x,y) then it has infinitely many integer solutions when Q is hyperbolic or parabolic, but only finitely many integer solutions when Q is elliptic or 0-hyperbolic.*

Proof: Consider first the hyperbolic and parabolic cases. Suppose (x,y) is a solution of $Q(x,y) = n$. If (x,y) is a primitive pair, then n appears in the topograph of Q so by periodicity it appears infinitely often, giving infinitely many solutions of $Q(x,y) = n$. If there is a nonprimitive solution (x,y) then it is d times a primitive pair (x',y') with $Q(x',y') = n/d^2$. The latter equation has infinitely many solutions (x',y') by what we just showed, hence $Q(x,y) = n$ has infinitely many solutions $(x,y) = (dx',dy')$.

For elliptic and 0-hyperbolic forms there is no periodicity, and the monotonicity property implies that each number appears in the topograph at most a finite number of times. Thus $Q(x,y) = n$ can have only finitely many primitive solutions. If it had infinitely many nonprimitive solutions, these would yield infinitely many primitive solutions of equations $Q(x,y) = m$ for certain divisors m of n. However, this is impossible since each equation $Q(x,y) = m$ for a fixed m can have only finitely many primitive solutions and n has only finitely many divisors since we assume it is nonzero. □

Exercises

1. (a) Find two primitive elliptic forms $ax^2 + cy^2$ that have the same discriminant but take on different sets of values. Draw enough of the topographs of the two forms to make it apparent that they do not have exactly the same sets of values. (Remember that the topograph only shows the values $Q(x,y)$ for primitive pairs (x,y).)
(b) Do the same thing with hyperbolic forms $ax^2 + cy^2$.

2. (a) Show the quadratic form $Q(x,y) = 92x^2 - 74xy + 15y^2$ is elliptic by computing its discriminant.
(b) Find the source vertex or edge in the topograph of this form.
(c) Using the topograph of this form, find all the integer solutions of $92x^2 - 74xy + 15y^2 = 60$, and explain why your list of solutions is a complete list. (There are exactly four pairs of solutions $\pm(x,y)$, three of which will be visible in the topograph.)

3. Show that if a form takes the same value on two adjacent regions of its topograph, then these regions are both adjacent to the source vertex or edge when the form is elliptic, or both lie along the separator line when the form is hyperbolic.

4. Show that the minimum value of $|h|$ for all the edges in the border of a given region in the topograph of an elliptic or hyperbolic form occurs at an edge having an

endpoint that achieves the minimum distance to the separator line or source vertex or edge of all vertices in the border of the given region.

5. (a) Show that if a quadratic form $Q(x,y) = ax^2 + bxy + cy^2$ can be factored as a product $(Ax + By)(Cx + Dy)$ with A, B, C, D integers, then Q takes the value 0 at some pair of integers $(x,y) \neq (0,0)$, hence Q must be either 0-hyperbolic or parabolic. Show also, by a direct calculation, that the discriminant of this form is a square.

(b) Find a 0-hyperbolic form $Q(x,y)$ such that $Q(1,5) = 0$ and $Q(7,2) = 0$ and draw a portion of the topograph of Q that includes the two regions where $Q(x,y) = 0$.

5.2 Equivalence of Forms

In the 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. For example if the values at v_1 are $2,2,3$ then the values at v_2 should also be $2,2,3$, in any order, but $2,3,3$ is regarded as different from $2,2,3$. Since the three values around a vertex determine all the other values in a topograph, having the same values at one vertex guarantees that the topographs look the same everywhere if the labels x/y are omitted.

An alternative definition of equivalence of forms would be to say that two forms are equivalent if there is a linear fractional transformation in $LF(\mathbb{Z})$ that takes the topograph of one form to the topograph of the other form. This is really the same as the first definition since there is a vertex of the topograph in the center of each triangle of the Farey diagram and we know that elements of $LF(\mathbb{Z})$ are determined by where they send a triangle, so if two topographs each have a vertex surrounded by the same triple of numbers, there is an element of $LF(\mathbb{Z})$ taking one topograph to the other, and conversely.

A topograph and its mirror image correspond to equivalent forms since the mirror image topograph has the same three labels around each vertex as at 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 always equivalent since they

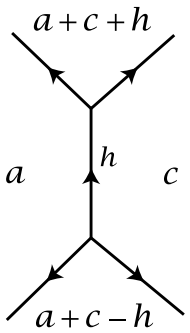
are related by changing (x,y) to $(-x,y)$, reflecting the Farey diagram across its horizontal axis, with a corresponding reflection of the topograph.

Equivalent forms have the same discriminant since the discriminant of a form is determined by the three numbers surrounding any vertex, as these three numbers determine the numbers p,q,h at each edge abutting the vertex and the discriminant is h^2-4pq for any of these edges.

Our next goal will be to see how to compute all the different equivalence classes of forms of a given discriminant. The method for doing this will depend on which of the four types of forms we are dealing with.

Reduced Elliptic Forms

Let us look at elliptic forms first to see how to determine all the different equivalence classes for a given discriminant in this case. As usual it suffices to consider only the forms with positive values. At a source vertex or edge in the topograph of a positive elliptic form Q let the smaller two of the three adjacent values of Q be a and c with $a\le c$, and let the edge between them be labeled $h\ge 0$. The third of the three smallest values of Q is then $a+c-h$. The form Q is equivalent to the form $ax^2+hxy+cy^2$ which has the values a, c, and $a+h+c$ for $(x,y)=(1,0)$, $(0,1)$, and $(1,1)$. Since a and c are the smallest values of Q we have $a\le c\le a+c-h$, and the latter inequality is equivalent to $h\le a$. Summarizing, we have the inequalities $0\le h\le a\le c$.



Thus every positive elliptic form is equivalent to a form $ax^2+hxy+cy^2$ with $0\le h\le a\le c$. An elliptic form satisfying these conditions is called ***reduced***. Two different reduced elliptic forms with the same discriminant are never 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 Δ via the formula $\Delta=h^2-4ac$ since we assume $h\ge 0$.

To avoid dealing with negative numbers let us set $\Delta=-D$ with $D>0$, so the discriminant equation becomes $D=4ac-h^2$. To find all equivalence classes of forms of discriminant $-D$ we therefore need to find all solutions of the equation

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

This equation implies that h must have the same parity as D, and we can bound the choices for h by the inequalities $4h^2\le 4a^2\le 4ac=D+h^2$ which imply $3h^2\le D$, or $h^2\le {}^{D}\!/_{3}$. This limits h to a finite number of possibilities, and for each of these values of h we just need to find all of the finitely many factorizations of h^2+D as $4ac$ with $a\le c$ and $h\le a$. In particular this shows that there are just finitely many equivalence classes of elliptic forms of a given discriminant.

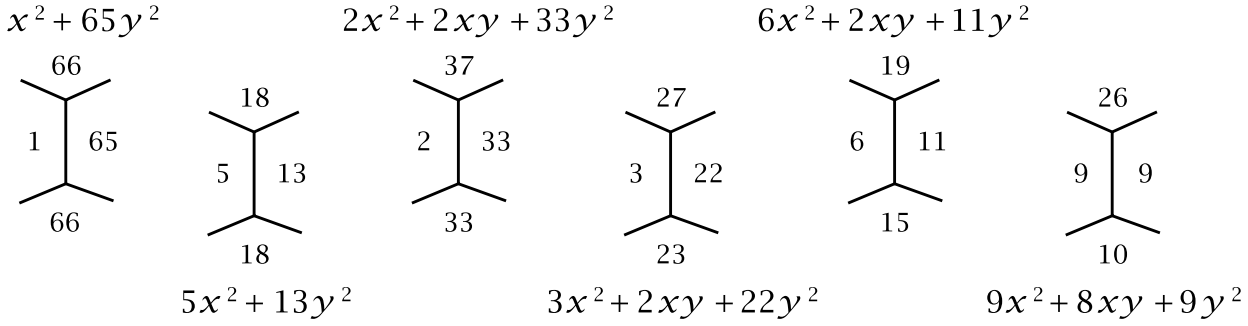
As an example consider the case $\Delta=-260$, so $D=260$. Since Δ is even, so is h, and we must have $h^2\le {}^{260}\!/_{3}$ so h must be 0, 2, 4, 6, or 8. The corresponding values

of a and c that are possible can then be computed from the equation $4ac = 260 + h^2$, always keeping in mind the requirement that $h \le a \le c$. The possibilities are shown in the following table:

h	ac	(a,c)
0	65	$(1,65),\ (5,13)$
2	66	$(2,33),\ (3,22),\ (6,11)$
4	69	—
6	74	—
8	81	$(9,9)$

As a side comment, note that the values of ac increase successively by $1, 3, 5, 7, \cdots$. This always happens when Δ is even, so the h values are $0, 2, 4, 6, \cdots$. For odd Δ the values of h are $1, 3, 5, 7, \cdots$ and the increments for ac are $2, 4, 6, 8, \cdots$. (Let it be an exercise for the reader to figure out why these statements are true.)

From the table we see that every positive elliptic form of discriminant -260 is equivalent to one of the six reduced forms $x^2 + 65y^2$, $5x^2 + 13y^2$, $2x^2 + 2xy + 33y^2$, $3x^2 + 2xy + 22y^2$, $6x^2 + 2xy + 11y^2$, or $9x^2 + 8xy + 9y^2$, and no two of these reduced forms are equivalent to each other. Here are small parts of the topographs of these forms:



In the first two topographs the central edge is a source edge, and in the last four topographs the lower vertex is a source vertex.

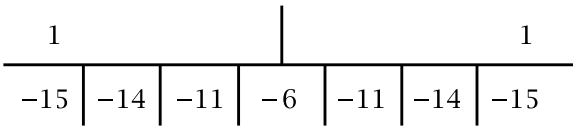
One might wonder what would happen if we continued the table with larger values of h not satisfying $h^2 \le {}^{260}/_{3}$. For example for $h = 10$ we would have $ac = 90$ so the condition $a \le c$ would force a to be 9 or less, violating the condition $h \le a$. Larger values of h would run into similar difficulties. The condition $h^2 \le {}^{D}/_{3}$ saves one the trouble of trying larger values of h.

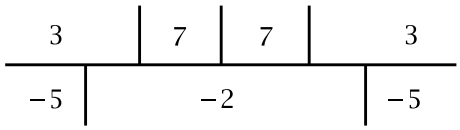
Cycles of Hyperbolic Forms

Next we consider hyperbolic forms of a given discriminant $\Delta > 0$. The topograph of a hyperbolic form has a separator line, so for each edge in the separator line we have the edge label h with the adjacent regions labeled p and $-q$ for $p > 0$ and $q > 0$. We can assume $h \ge 0$ by reorienting the edge if necessary. The discriminant equation is $\Delta = h^2 + 4pq$. Since p and q are positive this implies $h^2 < \Delta$ so there are only finitely many possibilities for h along the separator lines of forms of the given

discriminant Δ. For each h we then look at the factorizations $\Delta - h^2 = 4pq$. There can be only finitely many of these, so this means there are just finitely many possible combinations of labels $h, p, -q$ and hence only finitely many possible separator lines. Thus the number of equivalence classes of hyperbolic forms of a given discriminant is finite.

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:





From the topographs it is apparent that these 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, for a total of four equivalence classes so far. To see whether there are others we use the formula $\Delta = 60 = h^2 + 4pq$ relating the values p and $-q$ adjacent to an edge labeled h in the separator line, with $p > 0$ and $q > 0$. The various possibilities are listed in the table below. The equation $\Delta = h^2 + 4pq$ implies that h and Δ must have the same parity, just as in the elliptic case.

h	pq	(p,q)
0	15	$(1,15),\ (3,5),\ (5,3),\ (15,1)$
2	14	$(1,14),\ (2,7),\ (7,2),\ (14,1)$
4	11	$(1,11),\ (11,1)$
6	6	$(1,6),\ (2,3),\ (3,2),\ (6,1)$

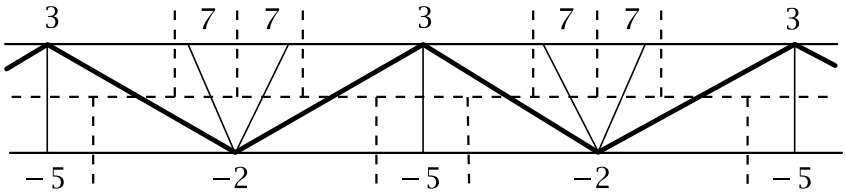
Each pair of values for (p,q) in the table occurs at some edge along the separator line in one of the two topographs shown above or the negatives of these topographs. Hence every form of discriminant 60 is equivalent to one of these four. If it had not been true that all the possibilities in the table 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 forms $px^2 + hxy - qy^2$ with new topographs, eventually exhausting all the finitely many possibilities.

The procedure in this example works for all hyperbolic forms. One makes a list of all the positive integer solutions of $\Delta = h^2 + 4pq$, then one constructs separator lines that realize all the resulting pairs (p,q). The different separator lines correspond exactly to the different equivalence classes of forms of discriminant Δ. Each solution (h,p,q) gives a form $px^2 + hxy - qy^2$. These are organized into *cycles* corresponding to the pairs $(p,-q)$ occurring along one of the periodic separator lines. Thus in the preceding example with $\Delta = 60$ the 14 pairs (p,q) in the table give rise to the four cycles along the four different separator lines.

A hyperbolic form $ax^2 + bxy + cy^2$ belongs to one of the cycles for the discriminant $\Delta = b^2 - 4ac$ exactly when $a > 0$ and $c < 0$ since a and c are the numbers p

and $-q$ lying on opposite sides of an edge of the separator line when $(x,y) = (1,0)$ and $(0,1)$.

If we superimpose the separator line of a hyperbolic form on the associated infinite strip in the Farey diagram, we see that the forms within a cycle correspond to the edges of the Farey diagram that lie in the strip and join one border of the strip to the other. For example, for the form $3x^2 - 5y^2$ we obtain the following picture, with fans of two triangles alternating with fans of three triangles:



The number of forms within a cycle can be fairly large in general. The situation can be improved somewhat by considering only the "most important" forms in the cycle, namely the forms that correspond to those edges in the strip that separate pairs of adjacent fans, indicated by heavier lines in the figure above. In terms of the topograph itself these are the edges in the separator line whose two endpoints have edges leading away from the separator line on opposite sides. The forms corresponding to these edges are traditionally called the *reduced* forms within the given equivalence class. In the example of discriminant 60 these are the forms with $(p,q) = (1,6)$, $(6,1)$, $(3,2)$, and $(2,3)$. These are the forms $x^2+6xy-6y^2$, $6x^2+6xy-y^2$, $3x^2+6xy-2y^2$, and $2x^2 + 6xy - 3y^2$. In this example there is just one reduced form for each cycle, but in more complicated examples there can be any number of reduced forms in a cycle. Note that the reduced forms do not necessarily give the simplest-looking forms, which in this example were the original forms $x^2 - 15y^2$ and $3x^2 - 5y^2$ along with their negatives $-x^2 + 15y^2$ and $-3x^2 + 5y^2$, or alternatively $15x^2 - y^2$ and $5x^2 - 3y^2$.

0-Hyperbolic and Parabolic Forms

For 0-hyperbolic forms it is rather easy to determine all the equivalence classes of forms of a fixed discriminant. As we saw in our initial discussion of 0-hyperbolic forms, their topographs contain two regions labeled 0, and the labels on the regions adjacent to each 0-region form an arithmetic progression with increment given by the label on the edges bordering the 0-region. Previously we called this edge label h but now let us change notation and call it q. We may assume q is positive by re-orienting the edges if necessary. The discriminant is $\Delta = q^2$ so both 0-regions must have the same edge label q. Either one of the two arithmetic progressions determines the form up to equivalence since two successive terms in the progression together with the 0 in the adjacent region give the three values of the form around a vertex in the topograph.

The form $qxy - py^2$ has discriminant q^2 and has $-p$ as one term of the arithmetic progression adjacent to the 0-region $x/y = 1/0$, namely in the region $x/y = 0/1$.

Thus every 0-hyperbolic form of discriminant q^2 is equivalent to one of these forms $qxy - py^2$. Arithmetic progressions with increment q can be thought of as congruence classes mod q, so only the mod q value of p affects the arithmetic progression and hence we may assume $0 \le p < q$. The number of equivalence classes of 0-hyperbolic forms of discriminant q^2 is therefore at most q, the number of congruence classes mod q. However, the number of equivalence classes could be smaller since each form has two 0 regions and hence two arithmetic progressions, which could be the same or different. Since either arithmetic progression determines the form, if the two progressions are the same then the topograph must have a mirror symmetry interchanging the two 0-regions. This always happens for example if the two 0-regions touch, which is the case $p = 0$ so the form is qxy and the mirror symmetry just interchanges x and y. If we let r denote the number of forms $qxy - py^2$ without mirror symmetry then the number of equivalence classes of 0-hyperbolic forms of discriminant q^2 is $q - r$ since each form without mirror symmetry has two different arithmetic progressions giving the same form.

For parabolic forms it is even easier to describe what all the different equivalence classes are since we have seen exactly what their topographs look like: There is a single region labeled 0 and all the regions adjacent to this have the same label q, which can be any nonzero integer, positive or negative. The integer q thus determines the equivalence class, so there is one equivalence class of parabolic forms for each nonzero integer q, with qx^2 being one form in this equivalence class. Parabolic forms all have discriminant 0, so in this case there are infinitely many different equivalence classes with the same discriminant. However, if we look only at primitive forms then there are just the two classes given by the forms $\pm x^2$.

Every parabolic form is equivalent to one of the forms qx^2 by a change of variables $T(x,y) = (sx+ty, ux+vy)$ with $sv-tu = \pm 1$, so every parabolic form factors as $q(sx+ty)^2$ for some pair of coprime integers s and t, with $q = \pm 1$ for primitive forms. Similarly, every 0-hyperbolic form is equivalent to a form $y(qx - py)$ so the form can be written as $(ux+vy)(q(sx+ty) - p(ux+vy))$ which can be simplified to a product $(Ax+By)(Cx+Dy)$ with A, B, C, D integers. Conversely, every form that factors as $(Ax+By)(Cx+Dy)$ with integer coefficients has the value 0 when $(x,y) = (-B, A)$ or $(-D, C)$ so the form must be parabolic or 0-hyperbolic. Parabolic forms are the case that the two linear factors are the same up to a constant multiple.

We have now shown how to compute all the equivalence classes of forms of a given discriminant for each of the four types of forms. In particular we have proved the following general fact:

Theorem 5.5. *There are only a finite number of equivalence classes of forms with a given nonzero discriminant.*

Exercises

1. (a) For positive elliptic forms of discriminant $\Delta = -D$, verify that the smallest value of D for which there are at least two inequivalent forms of discriminant $-D$ is $D = 12$.

(b) If we add the requirement that all forms under consideration are primitive, then what is the smallest D?

2. Determine all the equivalence classes of positive elliptic forms of discriminants -67, -104, and -347.

3. Find two elliptic forms that are not equivalent but take on the same three smallest values $a < b < c$.

4. Determine the number of equivalence classes of quadratic forms of discriminant $\Delta = 120$ and list one form from each equivalence class.

5. Do the same thing for $\Delta = 61$.

6. (a) Find the smallest positive nonsquare discriminant for which there is more than one equivalence class of forms of that discriminant. (In particular, show that all smaller discriminants have only one equivalence class.)

(b) Find the smallest positive nonsquare discriminant for which there are two inequivalent forms of that discriminant, neither of which is simply the negative of the other.

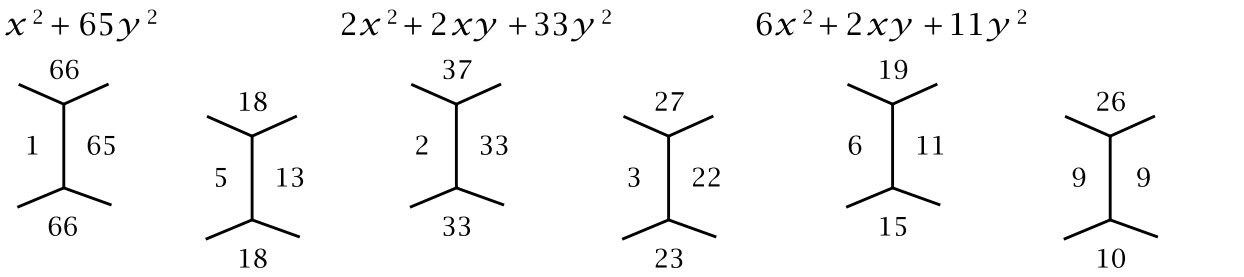
7. (a) Determine all the equivalence classes of 0-hyperbolic forms of discriminant 49.

(b) Determine which equivalence class in part (a) each of the forms $7xy - py^2$ for $p = 0, 1, 2, 3, 4, 5, 6$ belongs to.

5.3 The Class Number

When considering equivalence classes of forms of a given discriminant there are further refinements that turn out to be very useful. The first involves forms whose topographs are mirror images of each other. According to the definition we have given, two such forms are regarded as equivalent. However, there is a more refined notion of equivalence in which two forms are considered equivalent only if there is an orientation-preserving transformation in $LF(\mathbb{Z})$ taking the topograph of one form to the topograph of the other. In this case the forms are called ***properly equivalent***.

To illustrate the distinction between equivalence and proper equivalence, let us look at the earlier example of discriminant $\Delta = -260$ where we saw that there were six equivalence classes of forms:



$5x^2+13y^2$ $3x^2+2xy+22y^2$ $9x^2+8xy+9y^2$

In the first two topographs the central edge is a source edge and in the other four the lower vertex is a source vertex. Whenever there is a source edge the topograph has mirror symmetry across a line perpendicular to the source edge. When there is a source vertex there is mirror symmetry only when at least two of the three surrounding values of the form are equal, as in the third and sixth topographs above, but not the fourth or fifth topographs. Thus the mirror images of the fourth and fifth topographs correspond to two more quadratic forms which are not equivalent to them under any orientation-preserving transformation. With the more refined notion of proper equivalence there are therefore eight proper equivalence classes of forms of discriminant -260.

To obtain explicit formulas for the mirror image forms we can interchange the coefficients a and c in $ax^2+bxy+cy^2$, which corresponds to interchanging x and y, reflecting the topograph across a vertical line. Alternatively we could change the sign of b, which corresponds to changing the sign of either x or y and thus reflecting the topograph across a horizontal line.

For a general discriminant Δ each equivalence class of forms of discriminant Δ gives rise to two proper equivalence classes except when the class contains forms with mirror symmetry, in which case equivalence and proper equivalence amount to the same thing since every orientation-reversing equivalence can be converted into an orientation-preserving equivalence by composing with a mirror reflection. Here we are using the fact that the only linear fractional transformations that take a topograph to itself and reverse orientation are mirror reflections, as will be shown in Section 5.4 when we study symmetries of topographs in more detail.

Multiplying a form by an integer $d>1$ does not change its essential features in any significant way, so it is reasonable when classifying forms to restrict attention just to primitive forms, the forms that are not proper multiples of other forms. In other words, one considers only the forms $ax^2+bxy+cy^2$ for which a, b, and c have no common divisor greater than 1. The primitivity of a form is detectable just from the numbers appearing in its topograph since all the numbers in the topograph of a nonprimitive form are divisible by some number $d>1$, and conversely if all numbers in the topograph of a form $ax^2+bxy+cy^2$ are divisible by d then in particular a, c, and $a+b+c$, the values at $(1,0)$, $(0,1)$, and $(1,1)$, are divisible by d which implies

that b is also divisible by d so the whole form is divisible by d. Thus primitivity is a property of equivalence classes of forms. Multiplying a form by d multiplies its discriminant by d^2, so nonprimitive forms of discriminant Δ exist exactly when Δ is a square times another discriminant. For example, when $\Delta = -12 = 4(-3)$ one has the primitive form x^2+3y^2 as well as the nonprimitive form $2x^2+2xy+2y^2$ which is twice the form x^2+xy+y^2 of discriminant -3.

The number of proper equivalence classes of primitive forms of a given discriminant is called the ***class number*** for that discriminant, where in the case of elliptic forms one considers only the forms with positive values. The traditional notation for the class number for discriminant Δ is h_Δ. (This h has nothing to do with the h labels on edges in topographs.)

Since we have an algorithm for computing the finite set of equivalence classes of forms of a given nonzero discriminant, this leads to an algorithm for computing class numbers. When computing the table of triples (h,a,c) for elliptic forms or (h,p,q) for hyperbolic forms we omit the nonprimitive triples since these correspond to nonprimitive forms. Then we determine which of the remaining forms have mirror symmetry. For elliptic forms these are the cases when one or more of the inequalities $0 \le h \le a \le c$ is an equality, as we will see in the next section. For hyperbolic forms mirror symmetries can be detected in the separator line. Forms with mirror symmetry count once when computing the class number, and forms without mirror symmetry count twice. However, just having an algorithm to compute the class number h_Δ does not make it transparent how h_Δ depends on Δ, and indeed this is a very difficult question which is still only partially understood.

Of special interest are the discriminants for which all forms are primitive. These are called ***fundamental discriminants***. Thus a fundamental discriminant is one which is not a square times a smaller discriminant. For example, 8 is a fundamental discriminant even though it is divisible by a square, 4, since the other factor 2 is not the discriminant of any form, as it is not congruent to 0 or 1 mod 4. Technically 1 is a fundamental discriminant according to our definition, but we will exclude this trivial case. Thus fundamental discriminants are never squares, so fundamental discriminants appear only for elliptic and hyperbolic forms. With 1 excluded it is easy to check that the fundamental discriminants Δ with $|\Delta| < 40$ are 5, 8, 12, 13, 17, 20, 21, 24, 28, 29, 33, 37 and -3, -4, -7, -8, -11, -15, -19, -20, -23, -24, -31, -35, -39.

It is not hard to give a precise characterization of the discriminants Δ that are fundamental. First write $\Delta = 2^k n$ with $k \ge 0$ and n odd, possibly negative. If any odd square divides n then we can factor this out of Δ and still get a discriminant since odd squares are congruent to 1 mod 4 so multiplying by an odd square does not affect whether a number is 0 or 1 mod 4. The exponent k in 2^k can never be 1 since this would imply $\Delta \equiv 2 \bmod 4$. If $k \ge 4$ we can factor powers of 4 out of

Δ until we have k equal to 2 or 3 and still have a discriminant. If $k = 3$ we cannot factor a 4 out of Δ since this would give the excluded case $k = 1$. If $k = 2$ we can factor $4 = 2^k$ out of Δ exactly when $n \equiv 1 \bmod 4$. Finally, when $k = 0$ we have $\Delta = n$ so we must have $n \equiv 1 \bmod 4$. Thus the fundamental discriminants other than -4 and ± 8 are of three types:

- $\Delta = n$ with $|n|$ a product of distinct odd primes and $n \equiv 1 \bmod 4$.
- $\Delta = 4n$ with $|n|$ a product of distinct odd primes and $n \equiv 3 \bmod 4$.
- $\Delta = 8n$ with $|n|$ a product of distinct odd primes.

Every nonsquare discriminant can be factored uniquely as $\Delta = d^2\Delta'$ where Δ' is a fundamental discriminant and $d \geq 1$. The number d is called the ***conductor*** of Δ. Fundamental discriminants are those whose conductor is 1. Conductors will become important when we study the deeper properties of forms in later chapters. The class number h_Δ is always a multiple of $h_{\Delta'}$ and there is a not-too-complicated formula for what this multiple is, so the determination of class numbers reduces largely to the case of fundamental discriminants. However, we will not be going into more detail on the relationship between h_Δ and $h_{\Delta'}$ since this would lead us somewhat outside the scope of the book.

Discriminants of Class Number 1

The question of which discriminants have class number 1 has been much studied. This amounts to finding the discriminants for which all primitive forms are equivalent since if all primitive forms are equivalent, they are all equivalent to the principal form which has mirror symmetry so they are all properly equivalent to the principal form.

For elliptic forms the following nine fundamental discriminants have class number 1:

$$\Delta = -3,\ -4,\ -7,\ -8,\ -11,\ -19,\ -43,\ -67,\ -163$$

In addition there are four more which are not fundamental: $-12,\ -16,\ -27,\ -28$. It was conjectured by Gauss around 1800 that there are no other negative discriminants of class number 1. Over a century later in the 1930s it was shown that there is at most one more, and then in the 1950s and 1960s Gauss's conjecture was finally proved completely.

Another result from the 1930s is that for each number n there are only finitely many negative discriminants with class number n. Finding what these discriminants are is a difficult problem, however, and so far this has been done only in the range $n \leq 100$.

The situation for positive discriminants with class number 1 is not as well understood. Computations show that there are a large number of positive fundamental discriminants with class number 1, and it seems likely that there are in fact infinitely many. However, this has not been proved and remains one of the most basic unsolved problems about quadratic forms. If one allows nonfundamental discriminants then

it is known that there are infinitely many with $h_\Delta = 1$, including for example the discriminants $\Delta = 2^{2k+1}$ for $k \geq 1$ and $\Delta = 5^{2k+1}$ for $k \geq 0$.

Returning to the nine negative fundamental discriminants of class number 1, it is easy to check in each case that all forms are equivalent. For example when $\Delta = -163$ and we apply the earlier algorithm to find all reduced forms 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 prime, 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 $x^2 + xy + 41y^2$.

The corresponding polynomial $x^2 + x + 41$ has a curious property discovered by Euler: For each $x = 0, 1, 2, 3, \cdots, 39$ the value of $x^2 + x + 41$ is a prime number. Here are these forty 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 primes are $2, 4, 6, 8, 10, \cdots, 78$ since $[(x+1)^2 + (x+1) + 41] - [x^2 + x + 41] = 2(x+1)$. The next number in the sequence after 1601 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.) A similar thing happens for the other negative fundamental discriminants of class number 1. The nontrivial cases are listed in the table below, where $D = -\Delta$.

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

Satisfactory explanations are known for the occurrence of so many prime values of these quadratic polynomials but they involve fairly deep theory. It is curious that the lists of prime values account for all primes less than 100 except 79.

Suppose one asks about the next forty values of $x^2 + x + 41$ after the value 41^2 when $x = 40$. The next value, when $x = 41$, is $1763 = 41 \cdot 43$, also not a prime. After this the next two values are primes, then comes $2021 = 43 \cdot 47$, then four primes, then $2491 = 47 \cdot 53$, then six primes, then $3233 = 53 \cdot 61$, then eight primes, then $4331 = 61 \cdot 71$, then ten primes, then $5893 = 71 \cdot 83$. This last number was for $x = 76$, and the next four values are prime as well for $x = 77, 78, 79, 80$, completing the second 40 values. But then the pattern breaks down when $x = 81$ where one gets the value $6683 = 41 \cdot 163$. Thus, before the breakdown, not only were we getting sequences of 2, 4, 6, 8, 10 primes but the nonprime values were the products of two successive terms in the original sequence of prime values 41, 43, 47, 53, 61, $\cdots$.

All this seems quite surprising, even if the nice patterns do not continue forever. A partial explanation can be found in the fact that the polynomial $P(x) = x^2 + x + 41$ satisfies the identity $P(40 + n^2) = P(n-1)P(n)$ as one can easily check, so when $n = 1, 2, 3, \cdots$ we get $P(41) = P(0)P(1) = 41 \cdot 43$, $P(44) = 43 \cdot 47$, $P(49) = 47 \cdot 53$, $P(56) = 53 \cdot 61$, etc. However this does not explain why the intervening values of $P(x)$ should be prime. The polynomials in the preceding table exhibit similar behavior.

Exercises

1. Compute the class number for each of the following discriminants:
(a) -23 (b) -47 (c) -71 (d) -87 (e) -92 (f) 145 (g) 148.

2. In this extended exercise the goal will be to show that the only negative even discriminants with class number 1 are -4, -8, -12, -16, and -28. (Of these only -4 and -8 are fundamental discriminants.) The strategy will be to exhibit an explicit reduced primitive form Q different from the principal form $x^2 + dy^2$ for each discriminant $-4d$ with $d > 4$ except $d = 7$. This will be done by breaking the problem into several cases, where in each case a form Q will be given and you are to show that this form has the desired properties, namely it is of discriminant $-4d$, primitive, reduced, and different from the principal form. You should also check that the cases considered cover all possibilities.
(a) Suppose d is not a prime power. Then it can be factored as $d = ac$ where $1 < a < c$ and a and c are coprime. In this case let Q be the form $ax^2 + cy^2$.
(b) The form $ax^2 + 2xy + cy^2$ will work provided that $d + 1$ factors as $d + 1 = ac$ where a and c are coprime and $1 < a < c$. If d is odd, for example a power of an odd prime, then $d + 1$ is even so it has such a factorization $d + 1 = ac$ unless $d + 1 = 2^n$.
(c) If $d = 2^n$ the cases we need to consider are $n \geq 3$ since $d > 4$. When $n = 3$ take Q to be $3x^2 + 2xy + 3y^2$ and when $n \geq 4$ take Q to be $4x^2 + 4xy + (2^{n-2} + 1)y^2$.
(d) When $d + 1 = 2^n$ the cases of interest are $n \geq 3$. When $n = 3$ we have $d = 7$ which is one of the allowed exceptions with class number 1. When $n = 4$ we have $d = 15$ and $3x^2 + 5y^2$ works as in part (a). When $n = 5$ we have $d = 31$ and we take the form $5x^2 + 4xy + 7y^2$. When $n \geq 6$ we use the form $8x^2 + 6xy + (2^{n-3} + 1)y^2$.

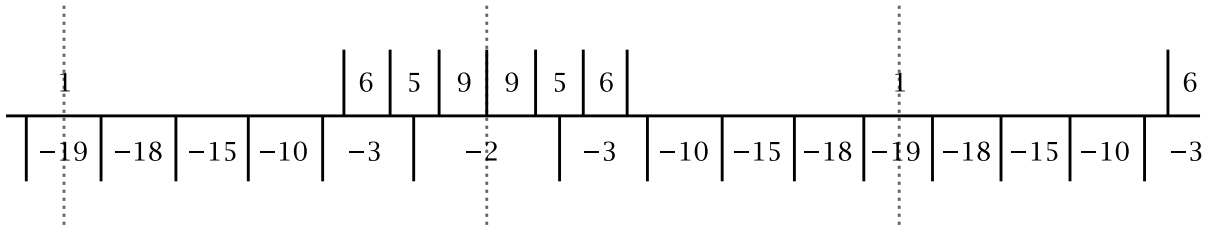
3. Show that the class number for discriminant $\Delta = q^2 > 1$ is $\varphi(q)$ where $\varphi(q)$ is the number of positive integers less than q and coprime to q.

5.4 Symmetries of Forms

We have observed that some topographs are symmetric in various ways. To give a precise meaning to this term, let us say that a ***symmetry*** of a form Q or its topograph is a transformation T in $LF(\mathbb{Z})$ that leaves all the values of Q unchanged,

so $Q(T(x,y)) = Q(x,y)$ for all pairs (x,y). For example, every hyperbolic form has a periodic separator line, which means there is a symmetry that translates the separator line along itself. If T is the symmetry translating by one period in either direction, then all the positive and negative powers of T are also translational symmetries. Strictly speaking, the identity transformation is always a symmetry but we will sometimes ignore this trivial symmetry.

Some hyperbolic forms also have mirror symmetry, where the symmetry is reflection across a line perpendicular to the separator line. This reflector line could contain one of the edges leading off the separator line, or it could be halfway between two consecutive edges leading off the separator line on the same side. Both kinds of symmetry occur along the separator line of the form $x^2 - 19y^2$, for example:

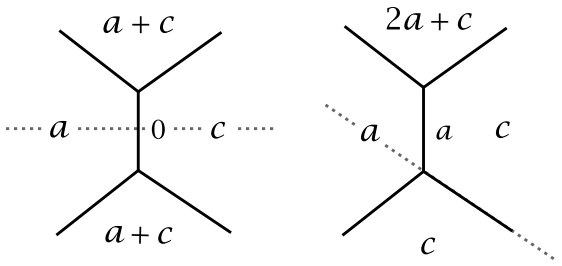


Elliptic forms can have mirror symmetries as well, as we saw in the earlier example $\Delta = -260$ where two topographs had mirror symmetry across a line perpendicular to an edge and two had symmetry across a line containing an edge.

Proposition 5.6. *A number a appears on the reflector line of a mirror symmetry of the topograph of a form Q exactly when Q is equivalent to a form $ax^2 + cy^2$ or $ax^2 + axy + cy^2$. In both cases a divides the discriminant of Q.*

In particular the principal forms $x^2 - ky^2$ and $x^2 + xy - ky^2$ have mirror symmetry, so there is at least one form with mirror symmetry in each discriminant.

Proof: The figures at the right show the two types of mirror symmetries, where the reflector line is either the perpendicular bisector of an edge of the topograph or contains an edge of the topograph. Let a and c be the labels on the left and right regions as in the figures, so the reflector line passes through the a region. If the edge between the left and right regions is labeled h then the regions above and below this edge are labeled $a + c + h$ and $a + c - h$. In the first figure the mirror symmetry forces h to be 0 so the form is equivalent to the form $ax^2 + cy^2$. In the second figure the mirror symmetry forces the lower region to be labeled c and this forces h to equal a when the edge labeled h is oriented upward. The form is then equivalent to the form $ax^2 + axy + cy^2$.



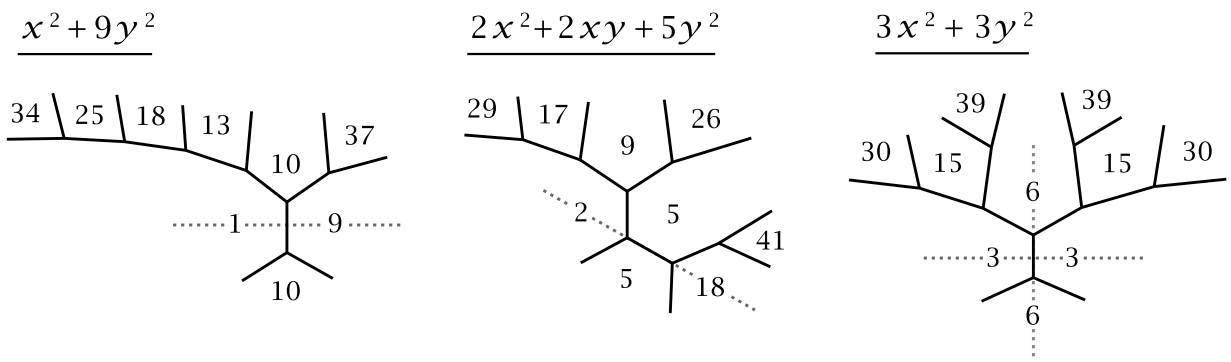
Conversely, the forms $ax^2 + cy^2$ and $ax^2 + axy + cy^2$ have topographs as shown in the figures, so these topographs have mirror symmetry with the reflector

line passing through the a region. These two forms have discriminants $-4ac$ and $a^2 - 4ac$, both divisible by a. □

As the proof showed, reflector lines crossing an edge in the topograph correspond to forms ax^2+cy^2 and reflector lines containing an edge correspond to forms $ax^2 + axy + cy^2$. For example, a form $ax^2 + bxy + ay^2$ has mirror symmetry interchanging x and y, reflecting across the vertical axis of the circular Farey diagram which contains an edge of the topograph, so this form is equivalent to a form $Ax^2+Axy+Cy^2$. The reflector line passes through regions of the topograph labeled $2a+b$ and $2a-b$ so A can be taken to be either $2a+b$ or $2a-b$, with $C = a$ since this is the value of the form at $x/y = 0/1$.

Proposition 5.7. *Let a be a divisor of the discriminant Δ that is either odd or twice an odd number. Then there exists a form ax^2+cy^2 or $ax^2+axy+cy^2$ of discriminant Δ having a in its topograph. If a is squarefree, a form of discriminant Δ with a in its topograph is unique up to equivalence, and a appears in the topograph only on a reflector line of a mirror symmetry.*

The conditions on the number a can be illuminated by looking at the case $\Delta = -36$ where there are three equivalence classes of forms:



The first two topographs have a single reflector line while the third has two reflector lines. The positive divisors of 36 are $1, 2, 3, 4, 6, 9, 12, 18$, and 36. The divisors that appear in the topographs are the ones that are odd or twice an odd number, so 4, 12, and 36 are excluded. Of the divisors that do appear, the ones that are not squarefree are 9 and 18, and these appear in more than one topograph, and off the reflector lines as well as on them.

Proof of Proposition 5.7: Suppose first that Δ is even. For the given divisor a of Δ let us first look for a form ax^2+cy^2 since this has even discriminant. Thus we want an integer c such that $\Delta = -4ac$. Since Δ is even it is divisible by 4, so if a is odd and divides Δ then $4a$ divides Δ so the desired integer c exists in this case.

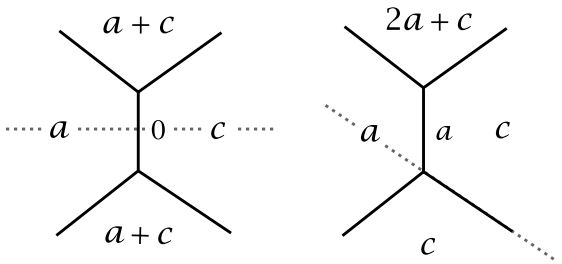
Since Δ is even it is either $8k$ or $8k+4$ for some integer k. If $\Delta = 8k$ then $\Delta = -4ac$ can again be solved for c when a is twice an odd number.

When $\Delta = 8k+4$ and a is twice an odd number the equation $\Delta = -4ac$ will not have an integer solution c since $-4ac$ is divisible by 8, so we instead look for a form $ax^2+axy+cy^2$. This has $\Delta = a(a-4c)$ and we want to find an integer c such that $\frac{\Delta}{a} = a-4c$. This is equivalent to saying $\frac{\Delta}{a} \equiv a \mod 4$. We have $a = 2(2m+1)$ so $a \equiv 2 \mod 4$. For $\frac{\Delta}{a}$, if we first divide Δ by 2 we get $4k+2$, then dividing by $2m+1$ can only change the congruence class mod 4 by a sign since odd numbers are $\pm 1 \mod 4$. Thus $\frac{\Delta}{a} \equiv 2 \mod 4$ so the congruence $\frac{\Delta}{a} \equiv a \mod 4$ is satisfied. This finished the proof of the existence of a form ax^2+cy^2 or $ax^2+axy+cy^2$ when Δ is even.

Suppose now that Δ is odd, hence also its divisor a. Since Δ is odd, we are looking for a form $ax^2+axy+cy^2$. As above, the condition for having such a form is the congruence $\frac{\Delta}{a} \equiv a \mod 4$. This is satisfied since $\Delta \equiv 1 \mod 4$ and $a \equiv \pm 1 \mod 4$.

Now we turn to the second statement in the proposition where we assume a is a squarefree divisor of Δ. Suppose that a appears in the topograph of a form of discriminant Δ. If b is one of the labels on an edge of the topograph bordering the region labeled a then we have $\Delta = b^2-4ac$ for c the label on the other region adjacent to the b edge. Since we assume a divides $\Delta = b^2-4ac$ it must also divide b^2, and if a is squarefree it will therefore divide b. Thus we have $b = ma$ for some integer m. The labels on the edges bordering the a region form an arithmetic progression with increment $2a$ so these are the numbers $b+2ka$ as k ranges over all integers. Since $b = ma$ we can factor $b+2ka$ as $(m+2k)a$. The numbers $m+2k$ for varying k form an arithmetic progression consisting of all even numbers if m is even and all odd numbers if m is odd. Thus we can choose k so that $m+2k$ is either 0 or 1, and hence the arithmetic progression $(m+2k)a$ contains either 0 or a. This means one of the edge labels on the border of the a region is either 0 or a.

The topograph near this edge has the shape shown in one of the two figures at the right. From this we see that there is a reflector line passing through the a region and the form is equivalent to either ax^2+cy^2 or $ax^2+axy+cy^2$.

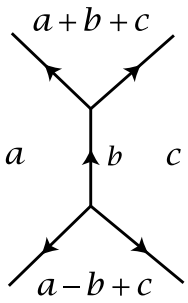


To finish the proof we only need to see that there cannot be both a form ax^2+cy^2 and a form $ax^2+axy+c'y^2$ with the same a and the same discriminant. Equating the discriminants of these two forms, we would have $-4ac = a^2-4ac'$ and therefore $a = 4(c'-c)$, but a would then be divisible by 4 and thus not squarefree. □

Symmetries of Elliptic Forms

Let us consider now what sorts of symmetries are possible in general for the various types of forms, beginning with elliptic forms. For an elliptic form each symmetry

must take the source vertex or edge to itself since this is where the smallest values of the form occur. In the case of a source edge, if a symmetry does not interchange the two ends of the source edge then the symmetry must be either the identity or a reflection across a line containing the source edge. If a symmetry does interchange the two ends of a source edge then it must either be a reflection across a line perpendicular to the edge or a 180 degree rotation of the topograph about the midpoint of the edge. Referring to the figure at the right, this rotation can only give a symmetry if $a = c$ and $a+b+c = a-b+c$ which is equivalent to having $b = 0$. Thus the form is $ax^2 + ay^2$ so if it is primitive it is just $x^2 + y^2$. Note that multiplying any form by a constant does not affect its symmetries so there is no harm in considering only primitive forms. For the form $x^2 + y^2$ note also that this form has both types of mirror symmetries, and the composition of these two mirror symmetries is the 180 degree rotational symmetry.



For a source vertex, a symmetry must take this vertex to itself. If a symmetry is orientation-preserving and not the identity then it must be a rotation about the source vertex by either one-third or two-thirds of a full turn. In either case this means that the three labels around the source vertex must be equal, so if the source vertex is the lower vertex in the figure above then the condition is $a = c = a - b + c$, which is equivalent to saying $a = b = c$. The form is then $ax^2 + axy + ay^2$ so if it is primitive it is $x^2 + xy + y^2$. The only other sort of symmetry for a source vertex is reflection across a line containing one of the three edges that meet at the source vertex. The only time there can be more than one such symmetry is when all three adjacent labels are equal so we are again in the situation of a form $ax^2 + axy + ay^2$.

For an elliptic form $ax^2 + bxy + cy^2$ that is reduced, so $0 \le b \le a \le c$, it is easy to recognize exactly when symmetries occur, namely when at least one of these three inequalities becomes an equality. Again using the figure above, when $b = 0$ one has a source edge with a mirror symmetry across the perpendicular line. When $b = a$ we have $a - b + c = c$ so there is a mirror symmetry across the lower right edge. And when $a = c$ one has mirror symmetry across the central edge. Since a and c are the two smallest labels on regions in the topograph, we see that reduced forms $ax^2 + bxy + ay^2$ occur when the smaller two of the three labels at the source vertex are equal, and reduced forms $ax^2 + axy + cy^2$ occur when the larger two labels are equal, at $0/1$ and $-1/1$.

Certain combinations of equalities in $0 \le b \le a \le c$ are also possible. If $b = 0$ and $a = c$ the form is $a(x^2 + y^2)$ with a source edge and both types of mirror symmetry as well as 180 degree rotational symmetry. Another possibility is that $b = a = c$ so the form is $a(x^2 + xy + y^2)$ with the symmetries described earlier. These are the only combinations of equalities that can occur since we must have $a > 0$ so $0 = b = a$ is impossible.

For reduced elliptic forms this exhausts all the possible symmetries since if we have strict inequalities $0 < b < a < c$ then the values of the form in the four regions shown in the figure above are all distinct. The first time this occurs is when the inequalities are $0 < 1 < 2 < 3$ so the form is $2x^2 + xy + 3y^2$ of discriminant -23.

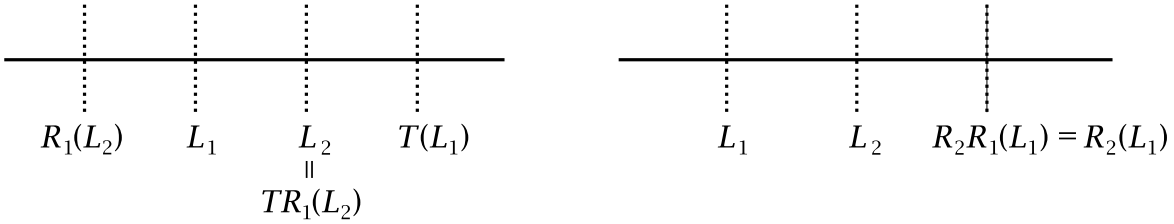
Symmetries of Hyperbolic Forms

Now consider hyperbolic forms. These all have periodic separator lines so they always have translational symmetries, and the question is what other sorts of symmetries are possible. For a hyperbolic form each symmetry must take the separator line to itself since this line consists of the edges that separate positive from negative values of the form. It is a simple geometric fact that a symmetry of a line L that is divided into a sequence of edges, say of length 1, extending to infinity in both directions, must be either a translation along L by some integer distance in either direction, or a reflection of L fixing either a vertex of L or the midpoint of an edge of L and interchanging the two halves of L on either side of the fixed point. This can be seen as follows. Symmetries of L are assumed to take vertices to vertices, so suppose the symmetry T sends a vertex v to the vertex $T(v)$. Then if T preserves the orientation of L it must be a translation along L by the distance from v to $T(v)$ as one can see by considering what T does to the two edges adjacent to v, then to the next two adjacent edges on either side, then the next two edges, and so on. If T reverses the orientation of L then either $T(v) = v$ or T fixes the midpoint of the segment from v to $T(v)$ since it sends this segment to a segment of the same length with one end at $T(v)$ but extending back toward v since T reverses orientation of L. Thus T fixes a point of L in either case, and it follows that T must reflect L across this fixed point, as one can again see by considering the edge or edges containing the fixed point, then the next two edges, and so on. If the distance from v to $T(v)$ is an even integer, the midpoint between v and $T(v)$ will be a vertex, and if it is odd, the midpoint will be a midpoint of an edge.

Returning to the situation of a symmetry T of the topograph of a hyperbolic form that takes the separator line L to itself, T must also take the side of L with positive labels to itself, so T preserves orientation of the plane exactly when it preserves orientation of L. Thus the only orientation-preserving symmetries of the topograph are translations along the separator line, and the only orientation-reversing symmetries are the two kinds of reflections across lines perpendicular to L.

If the separator line of a hyperbolic form has a mirror symmetry then because of periodicity there has to be at least one reflector line in each period, but in fact there are exactly two reflector lines in each period. To see this, let T be the translation by one period and let R_1 be a reflection across a reflector line L_1. Consider the composition TR_1, reflecting first by R_1 then translating by T, so TR_1 is an orientation-reversing symmetry. If L_2 is the line halfway between L_1 and $T(L_1)$ then $T(R_1(L_2)) = L_2$ as

we can see in the first figure below:



Thus TR_1 is an orientation-reversing symmetry that takes L_2 to itself while preserving the positive and negative sides of the separator line, so TR_1 must be a reflection R_2 across L_2. This shows that there are at least two reflector lines in each period. There cannot be more than two since if R_1 and R_2 are the reflections across two adjacent reflector lines L_1 and L_2 as in the second figure then the composition R_2R_1, first reflecting by R_1 then by R_2, is orientation-preserving and sends L_1 to $R_2(R_1(L_1)) = R_2(L_1)$ so this composition is a symmetry translating the separator line by twice the distance between L_1 and L_2. The distance between L_1 and L_2 must then be half the length of the period, otherwise if the translation R_2R_1 were some power T^n of the basic periodicity translation T with $|n| > 1$, there would be fewer than two reflector lines in a period.

For completeness let us also describe the symmetries for the remaining two types of forms besides elliptic and hyperbolic forms. For a 0-hyperbolic form, if the two regions labeled 0 in the topograph have a border edge in common then a symmetry must take this edge to itself, and it cannot interchange the ends of the edge since positive values must go to positive values. The only possibility is then a reflection across this edge, which is always a symmetry of the topograph. If the two 0-regions do not have a common border edge they are joined by a finite separator line and a symmetry must take this line to itself without interchanging the positive and negative sides. The only possibility is a reflection across a line perpendicular to the separator line and passing through its midpoint. This reflection gives a symmetry only when the finite continued fraction associated to the form is palindromic.

A parabolic form has a single 0-region in its topograph, so the bordering line for this region must be taken to itself by any symmetry. Every symmetry of this bordering line gives a symmetry of the form, either a translation along the line or a reflection across a perpendicular line.

The preceding analysis shows in particular the following fact:

Proposition 5.8. *All orientation-reversing symmetries of the topograph of a form are mirror symmetries, reflecting across a line that is either perpendicular to or contains an edge of the topograph.*

Traditionally, a form whose topograph has an orientation-reversing symmetry is called *ambiguous* although there is really nothing about the form that is ambiguous

in the usual sense of the word, unless perhaps it is the fact that such a form is indistinguishable from its mirror image.

The Symmetric Class Number

Let us define the ***symmetric class number*** h_Δ^s to be the number of equivalence classes of primitive forms of discriminant Δ with mirror symmetry. Recall that equivalence is the same as proper equivalence for forms with mirror symmetry. The ordinary class number h_Δ is thus h_Δ^s plus twice the number of equivalence classes of primitive forms without mirror symmetry. We have $h_\Delta \ge h_\Delta^s$, and in fact h_Δ is always a multiple of h_Δ^s as we will see in Proposition 7.16.

In contrast with h_Δ, the number h_Δ^s can be computed explicitly. Here is the result for elliptic and hyperbolic forms:

Theorem 5.9. *If Δ is a nonsquare discriminant and k is the number of distinct prime divisors of Δ then $h_\Delta^s = 2^{k-1}$ except in the following cases*:

(a) *If $\Delta = 4(4m+1)$ then $h_\Delta^s = 2^{k-2}$.*

(b) *If $\Delta = 32m$ then $h_\Delta^s = 2^k$.*

The exceptional cases (a) and (b) involve nonfundamental discriminants, so for fundamental discriminants we have $h_\Delta^s = 2^{k-1}$. For example, the discriminants $\Delta = 60 = 3\cdot4\cdot5$ and $\Delta = -260 = -4\cdot5\cdot13$ that we looked at in the previous section have three distinct prime divisors so the theorem says there are $2^2 = 4$ equivalence classes of mirror symmetric forms in these two cases. This agrees with what the topographs showed.

The proof of the theorem will involve considering a number of different cases. Fortunately most of the resulting complication disappears in the final answer.

Proof: By Proposition 5.6 every form with mirror symmetry is equivalent to a form $ax^2 + cy^2$ or $ax^2 + axy + cy^2$. The strategy will be to count how many of these special forms there are that are primitive with discriminant Δ, then determine which of these special forms are equivalent.

For counting the special forms ax^2+cy^2 and $ax^2+axy+cy^2$ we may assume $a > 0$ since a is the value of the form when $(x,y) = (1,0)$ and for elliptic forms we only consider those with positive values, while for hyperbolic forms we are free to change a form to its negative so it suffices to count only those with $a > 0$ and then double the result.

Case 1: Forms ax^2+cy^2. Then $\Delta = -4ac = 4\delta$ for $\delta = -ac$. Primitivity of the form is equivalent to a and c being coprime. The only way to have coprime factors a and c of $\delta = -ac$ is to take an arbitrary subset of the distinct primes dividing δ and let a be the product of these primes each raised to the same power as in δ (so $a = 1$ when we choose the empty subset). The number of such subsets is $2^{k'}$ where k' is the

number of distinct prime divisors of δ, so there are $2^{k'}$ primitive forms ax^2+cy^2 with $a>0$.

Case 2: Forms $ax^2+axy+cy^2$ with Δ odd. We have $\Delta=a^2-4ac$ so Δ and a have the same parity. From $\Delta=a(a-4c)$ we see that a divides Δ. We claim that each divisor a of Δ gives rise to a form $ax^2+axy+cy^2$ of discriminant Δ. Solving $\Delta=a^2-4ac$ for c gives $c=(a^2-\Delta)/4a$. The numerator is divisible by 4 since a and Δ are odd and hence a^2 and Δ are both 1 mod 4, making the numerator 0 mod 4. The numerator is also divisible by a if a divides Δ. Since 4 and a are coprime when a is odd, it follows that $4a$ divides the numerator so c is an integer and we get a form $ax^2+axy+cy^2$ of discriminant Δ. This form is primitive exactly when a and c are coprime. This is equivalent to saying that the two factors of $\Delta=a(a-4c)$ are coprime since any divisor of a and c must divide the two factors, and conversely any divisor of the two factors must divide a and $4c$, hence also c since this divisor of the odd number a must be odd. As in Case 1, the only way to obtain a factorization $\Delta=a(a-4c)$ with the two factors coprime is to take an arbitrary subset of the distinct primes dividing Δ and let a be the product of these primes each raised to the same power as in Δ. The number of such subsets is 2^k so this is the number of primitive forms $ax^2+axy+cy^2$ with $a>0$ when Δ is odd.

There remain the forms $ax^2+axy+cy^2$ with $\Delta=4\delta$. Again Δ and a have the same parity since $\Delta=a^2-4ac$, so a is even, say $a=2d$. From $\Delta=a^2-4ac$ we then have $\delta=d^2-2dc=d(d-2c)$.

Case 3: Forms $ax^2+axy+cy^2$ with $\Delta=4\delta$ and $a=2d$ for odd d. By primitivity c must be odd. The two factors of $\delta=d(d-2c)$ are odd and must be distinct mod 4 since c is odd. Thus one factor is 1 mod 4 and the other is 3 mod 4, so $\delta\equiv 3 \bmod 4$, say $\delta=4m+3$. We claim that when $\delta=4m+3$, each divisor d of δ gives rise to a form $ax^2+axy+cy^2$ with $a=2d$. To show this, note first that d must be odd since it divides δ which is odd. Solving $\delta=d(d-2c)$ for c gives $c=(d^2-\delta)/2d$. Since d and δ are odd, the numerator $d^2-\delta$ is even hence divisible by the 2 in the denominator. The numerator is also divisible by the d in the denominator if d divides δ. Since d is odd, this implies that $2d$ divides the numerator, so c is an integer for each divisor d of δ. In fact c is an odd integer since the numerator $d^2-\delta$ is 2 mod 4 and so $cd=(d^2-\delta)/2$ is odd, forcing c to be odd. For the form $ax^2+axy+cy^2$ to be primitive means that a and c are coprime. Since c is odd and $a=2d$ this is equivalent to c and d being coprime. This in turn is equivalent to the two factors of $\delta=d(d-2c)$ being coprime since c and d are odd. Thus when $\delta=4m+3$ we get a primitive form $ax^2+axy+cy^2$ for each choice of a subset of the distinct prime divisors of δ since this determines d as before, and d determines c and a. The number of primitive forms $ax^2+axy+cy^2$ is then $2^{k'}$ when Δ is even and $a=2d$ with d odd, where k' is the number of distinct prime divisors of δ as in Case 1.

Case 4: Forms $ax^2 + axy + cy^2$ with Δ even and $a = 2d$ for even d, say $d = 2e$. Then $\delta = d(d - 2c) = 4e(e - c)$. Since c is odd by primitivity of the form, the two factors e and $e - c$ of $\delta = 4e(e - c)$ have opposite parity, hence δ must be divisible by 8, say $\delta = 8m$. We need to determine which choices of e and c yield primitive forms $ax^2 + axy + cy^2$. Let $\delta' = \delta/4 = 2m$ so the equation $\delta = 4e(e - c)$ becomes $\delta' = e(e - c)$. Thus e must divide δ'. We have $c = e - \delta'/e$ and this will be an integer if e divides δ'. From the equation $c = e - \delta'/e$ we see that any divisor of two of the three terms c, e, and δ'/e will divide the third. In particular, c and e will be coprime exactly when e and δ'/e are coprime. Since $\delta' = e \cdot \delta'/e$ this means we want to choose e by choosing some subset of the distinct prime divisors of δ' and letting e be the product of these primes raised to the same powers as in δ'. Then e and δ'/e will be coprime and of opposite parity since they are not both even and their product δ' is even. Their difference $c = e - \delta'/e$ will then be odd. Also, c and e will be coprime so c and $a = 4e$ will be coprime, making the form $ax^2 + axy + cy^2$ primitive. The number of distinct prime divisors of δ' is the same as for $\delta = 4\delta'$ since δ' is even. Thus in Case 4 the number of primitive forms $ax^2 + axy + cy^2$ with $a > 0$ is $2^{k'}$.

Note that $k' = k$ when δ is even and $k' = k - 1$ when δ is odd. By combining the four cases above and remembering to double the number of forms when $\Delta > 0$ to account for negative coefficients of x^2, we then obtain the following table for the number of forms of either of the types $ax^2 + cy^2$ or $ax^2 + axy + cy^2$:

Δ	odd	4δ, $\delta = 4m + 1$	4δ, $\delta = 4m + 3$
Cases	(2)	(1)	(1) and (3)
$\Delta < 0$	2^k	$2^{k'} = 2^{k-1}$	$2^{k'} + 2^{k'} = 2^{k'+1} = 2^k$
$\Delta > 0$	2^{k+1}	$2^{k'+1} = 2^k$	$2^{k'+1} + 2^{k'+1} = 2^{k'+2} = 2^{k+1}$

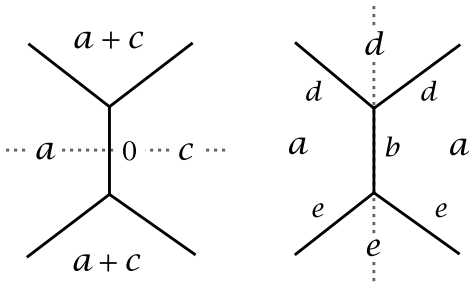
Δ	4δ, $\delta = 8m$	4δ, δ even, $\delta \neq 8m$
Cases	(1) and (4)	(1)
$\Delta < 0$	$2^{k'} + 2^{k'} = 2^{k'+1} = 2^{k+1}$	$2^{k'} = 2^k$
$\Delta > 0$	$2^{k'+1} + 2^{k'+1} = 2^{k'+2} = 2^{k+2}$	$2^{k'+1} = 2^{k+1}$

Comparing the results in the table with the statement of the theorem, we see that the proof will be finished when we show that under the relation of equivalence the special forms split up into pairs when $\Delta < 0$ and into groups of four when $\Delta > 0$.

Two easy cases that can be disposed of first are $\Delta = -3$ and $\Delta = -4$. Here all forms are equivalent and are primitive, and $k = 1$, so the theorem is true since the exceptional cases (a) and (b) in the statement of the theorem do not apply.

Our earlier analysis of symmetries of elliptic and hyperbolic forms shows that the only time that reflector lines can intersect is for elliptic forms equivalent to $ax^2 + ay^2$ or $ax^2 + axy + ay^2$, so when we restrict to primitive forms this means $\Delta = -3$ or $\Delta = -4$. Thus we may assume from now on that reflector lines do not intersect.

For a form $ax^2 + cy^2$ with a reflector line perpendicular to an edge of the topograph as in the first figure at the right we have $a \neq c$, otherwise there would be two intersecting reflector lines. Thus the reflector line corresponds to two distinct special forms, ax^2+cy^2 and cx^2+ay^2.

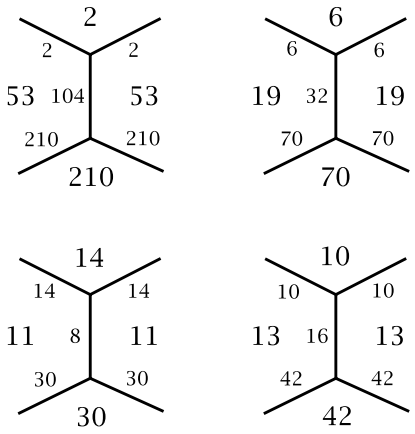


The second figure shows the case of a form with a reflector line containing an edge of the topograph. This edge corresponds to a form $ax^2 + bxy + ay^2$ and the adjacent edges correspond to two forms $dx^2 + dxy + ay^2$ and $ex^2 + exy + ay^2$ of the type $ax^2+axy+cy^2$. These two forms are distinct since if $d = e$ there would be a second reflector line intersecting the first one. Thus the reflector line accounts for two special forms $ax^2 + axy + cy^2$.

Primitive elliptic forms with mirror symmetry and $\Delta \neq -3, -4$ have just one reflector line, so each equivalence class of such forms contains exactly two special forms. For hyperbolic forms with mirror symmetry there are two reflector lines in each period, with one pair of special forms for each reflector line. These two pairs give four distinct special forms, otherwise there would be a translational symmetry taking one reflector line to the other within a single period, which is impossible. Thus each equivalence class of mirror-symmetric hyperbolic forms contains exactly four special forms, and the proof is complete. □

We illustrate the theorem with an example, the first negative discriminant with four distinct prime divisors, $\Delta = -420 = -3\cdot4\cdot5\cdot7$. In this case $\Delta = 4(4m + 3)$ so the theorem says there are $2^3 = 8$ equivalence classes of symmetric primitive forms. If we compute all the reduced forms for $\Delta = -420$ by the method in Section 5.2 we get the following table, with the letter b replacing h so we are finding solutions of $b^2 + 420 = 4ac$ with $0 \leq b \leq a \leq c$. The entries $[a,b,c]$ in the last column give the reduced forms $ax^2 + bxy + cy^2$.

b	ac	(a,c)	$[a,b,c]$
0	105	(1,105)	[1,0,105]
		(3,35)	[3,0,35]
		(5,21)	[5,0,21]
		(7,15)	[7,0,15]
2	106	(2,53)	[2,2,53]
4	109	—	
6	114	(6,19)	[6,6,19]
8	121	(11,11)	[11,8,11]
10	130	(10,13)	[10,10,13]



Thus all forms of discriminant -420 are symmetric. The first four have $b = 0$ so these arise in Case 1 in the proof of the theorem where we set $\Delta = 4\delta$, so $\delta = -3\cdot5\cdot7$ and we get a form $[a,0,c]$ for each positive divisor a of δ, the eight numbers

$1, 3, 5, 7, 15, 21, 35$, and 105. These forms $[a, 0, c]$ are the first four entries in the last column of the table along with the equivalent forms obtained by reversing a and c. The remaining four forms in the last column have b nonzero and are instances of forms $[a, a, c]$ and $[a, b, a]$. The relevant parts of the topographs of these four forms are shown in the figure to the right of the table. Each edge in the figure gives a form $[a, b, a]$, $[a, a, c]$, or $[a, c, c]$. For example the third figure gives the forms $[11, 8, 11]$, $[11, 14, 14]$, $[14, 14, 11]$, $[11, 30, 30]$, and $[30, 30, 11]$. In the proof of the theorem we were only counting the forms $[a, a, c]$, not $[a, b, a]$ or $[a, c, c]$. According to Case 3 in the proof of the theorem the numbers a in the forms $[a, a, c]$ should be twice the numbers a in the forms $[a, 0, c]$, and they are: $2 = 2\cdot 1$, $6 = 2\cdot 3$, $10 = 2\cdot 5$, $14 = 2\cdot 7$, $30 = 2\cdot 15$, $42 = 2\cdot 21$, $70 = 2\cdot 35$, and $210 = 2\cdot 105$.

Corollary 5.10. *The nonsquare discriminants Δ with $h_\Delta^s = 1$ are $\Delta = -4, \pm 8, -16, \pm p^{2k+1}$, and $\pm 4p^{2k+1}$ for odd primes p with $p \equiv 1$ mod 4 when $\Delta > 0$ and $p \equiv 3$ mod 4 when $\Delta < 0$. In particular, the only fundamental discriminants with $h_\Delta^s = 1$ are $\Delta = -4, \pm 8$, and $\pm p$ for odd primes p, with $p \equiv 1$ mod 4 when $\Delta > 0$ and $p \equiv 3$ mod 4 when $\Delta < 0$.*

Proof: Consider first the case $\Delta > 0$. If we are not in one of the exceptional cases (a) and (b) in Theorem 5.9 then Δ must have just one distinct prime divisor so it must be a power of a prime, in fact an odd power if it is not a square. Thus for p odd we have $\Delta = p^{2k+1}$ and we must have $p \equiv 1 \bmod 4$ in order to have $\Delta \equiv 1 \bmod 4$. For odd powers of $p = 2$ the only possibility is $\Delta = 8$ since Δ cannot be 2 and odd powers beyond 8 are of the form $\Delta = 32m$, the exceptional case (b) where $h_\Delta^s \geq 2$ so this is ruled out as well. In the exceptional case (a) we have $\Delta = 4(4m + 1)$ with $4m + 1$ a prime power p^{2k+1} with $p \equiv 1 \bmod 4$ since $\Delta = 4p^{2k}$ is a square.

When $\Delta < 0$ the reasoning is similar, the main difference being that $-p^{2k}$ and $-4p^{2k}$ are ruled out, not because squares are excluded, but because p^{2k} is always 1 mod 4 when p is odd, so $-p^{2k}$ is 3 mod 4. This rules out $-p^{2k}$ as a discriminant, and it rules out $-4p^{2k}$ being an exceptional case $\Delta = 4(4m + 1)$.

Requiring Δ to be a fundamental discriminant eliminates the cases $\Delta = -16$ and $\pm 4p^{2k+1}$ and restricts the exponent in $\pm p^{2k+1}$ to be 1. □

We have mentioned the fact that h_Δ is always a multiple of h_Δ^s, which will be proved in Proposition 7.17. This tells us nothing about h_Δ when $h_\Delta^s = 1$, but we will also prove that $h_\Delta^s = 1$ exactly when h_Δ is odd. Thus the preceding corollary gives a way to determine whether h_Δ is even or odd. In the examples we have looked at so far h_Δ has been either 1 or even, but odd numbers greater than 1 can also occur as class numbers. The table on the next page gives some examples for negative discriminants, so we are determining the reduced forms $ax^2 + bxy + cy^2$ by finding the solutions of $b^2 + |\Delta| = 4ac$ with $0 \leq b \leq a \leq c$. The forms other than the principal form in each discriminant lack mirror symmetry so they count twice in the class number,

making the class number odd. The discriminants in the table are all fundamental discriminants, and in each case they are the first negative discriminant with the given class number.

Δ	b	ac	(a,c)	h_Δ
-23	1	6	$(1,6),(2,3)$	3
-47	1	12	$(1,12),(2,6),(3,4)$	5
	3	14	—	
-71	1	18	$(1,18),(2,9),(3,6)$	7
	3	20	$(4,5)$	
-199	1	50	$(1,50),(2,25),(5,10)$	9
	3	52	$(4,13)$	
	5	56	$(7,8)$	
	7	62	—	
-167	1	42	$(1,42),(2,21),(3,14),(6,7)$	11
	3	44	$(4,11)$	
	5	48	$(6,8)$	
	7	54	—	
-191	1	48	$(1,48),(2,24),(3,16),(4,12),(6,8)$	13
	3	50	$(5,10)$	
	5	54	$(6,9)$	
	7	60	—	
-239	1	60	$(1,60),(2,30),(3,20),(4,15),(5,12),(6,10)$	15
	3	62	—	
	5	66	$(6,11)$	
	7	72	$(8,9)$	

Besides the cases when $h^s_\Delta = 1$, another nice situation is when $h_\Delta = h^s_\Delta$ so all primitive forms of discriminant Δ have mirror symmetry. We call such discriminants ***fully symmetric***. As we will see in later chapters, forms with fully symmetric discriminants have very special properties. A table at the end of the book lists the 101 known negative discriminants that are fully symmetric, ranging from -3 to -7392.

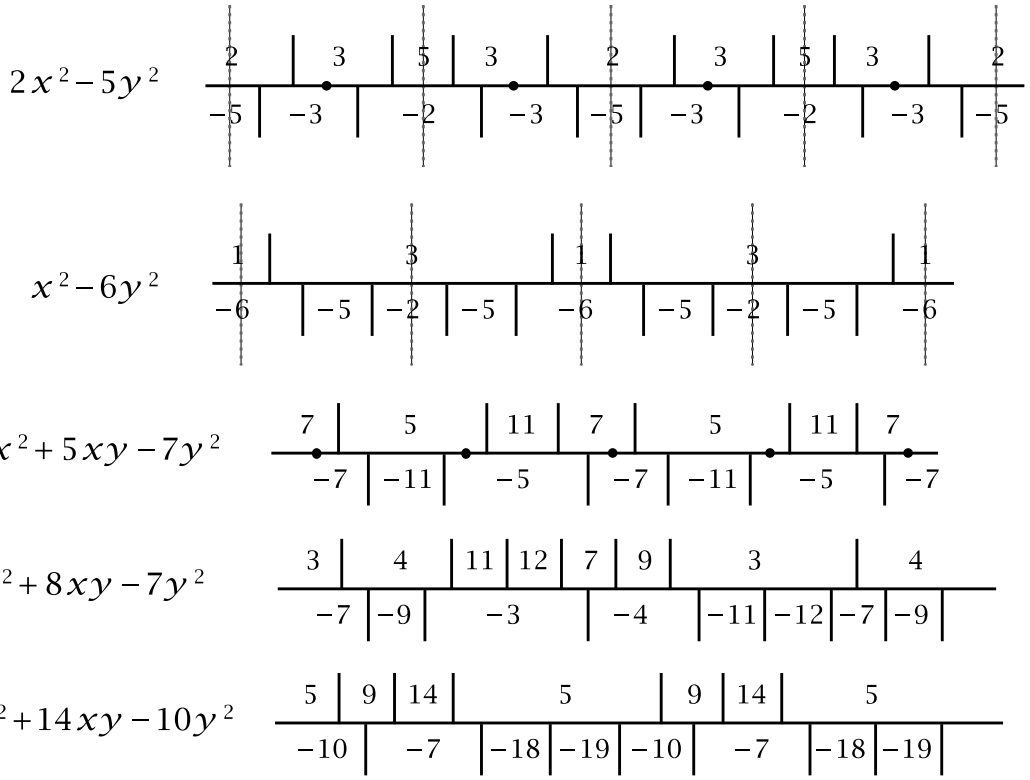
Of the 101 known fully symmetric negative discriminants, 65 are fundamental discriminants, the largest being -5460. Since 5460 factors as $3\cdot4\cdot5\cdot7\cdot13$ with five distinct prime factors, Theorem 5.9 says that $h^s_\Delta = 2^4 = 16$. This is in fact the largest value of h^s_Δ among the 101 discriminants in the list. Computer calculations have extended to much larger negative discriminants without finding any more that are fully symmetric. It has not yet been proved that no more exist, although it is known that there are at most two more. For positive discriminants there are probably infinitely many that are fully symmetric since it is likely that there are already infinitely many with $h_\Delta = 1$.

Skew Symmetries

Among the examples of hyperbolic forms we have considered there were some whose topograph had a "symmetry" which was a glide reflection along the separator line that had the effect of changing each value to its negative rather than preserving the values. These are not actual symmetries according to the definition we have given, so let us call such a transformation that takes each value of a form to its negative a ***skew symmetry***. (Compare this with skew-symmetric matrices in linear algebra which equal the negative of their transpose.)

A skew symmetry must take the separator line to itself while interchanging the two sides of the separator line, so it either translates the separator line along itself and hence is a glide reflection, or it reflects the separator line, interchanging its two ends as well as the two sides of the separator line, making it a 180 degree rotation about a point of the separator line. Examples of forms with this sort of skew symmetry occurred in Chapter 4, the forms $x^2 - 13y^2$ and $10x^2 - 29y^2$.

The figures below show forms whose separator lines have all the possible combinations of symmetries and skew symmetries.

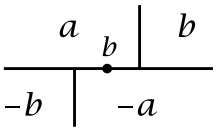


The first form has all four types: translations, mirror symmetries, glide reflections, and rotations. The next three forms have only one type of symmetry or skew symmetry besides translations, while the last form has only translational symmetries and no mirror symmetries or skew symmetries. It is not possible to have two of the three types of nontranslational symmetries and skew symmetries without having the third since the composition of two of these three types gives the third type. One can see this by considering the effect of a symmetry or skew symmetry on the orientation of

the plane and the orientation of the separator line. The four possible combinations distinguish the four types of transformations according to the following chart, where a plus sign means orientation-preserving and a minus sign means orientation-reversing.

	plane orientation	line orientation
translation	+	+
rotation	+	−
glide reflection	−	+
reflection	−	−

A rotational skew symmetry is a rotation about the midpoint of an edge of the separator line where the two adjacent regions have labels a and $-a$. If the edge separating these two regions has label b then the form associated to this edge is $ax^2 + bxy - ay^2$. Conversely, any form $ax^2 + bxy - ay^2$ whose discriminant $\Delta = b^2 + 4a^2$ is not a square (although it is the sum of two squares) will be a hyperbolic form having a rotational skew symmetry, as one can see in the figure at the right. Note that the form $ax^2 + bxy - ay^2$ will be one of the reduced forms in the equivalence class of the given form since the two edges leading off the separator line at the ends of the edge labeled b do so on opposite sides of the separator line. Thus rotational skew symmetries can be detected by looking just at the reduced forms. The same is true for mirror symmetries and glide reflection skew symmetries, but for these one must look at the arrangement of the whole cycle of reduced forms rather than just the individual reduced forms.



For rotational skew symmetries there are two rotation points along the separator line in each period, just as reflector lines occur in pairs in each period.

Exercises

1. Show that the number of symmetries of an elliptic form, including the identity transformation, is 1, 2, 4, or 6.

2. Show that the number of equivalence classes of forms of discriminant 45 with mirror symmetry is not a power of 2 if nonprimitive as well as primitive forms are allowed. (Compare this with Theorem 5.9.)

3. In the text an example was given of a hyperbolic form having only translational symmetries and no skew symmetries, the form $5x^2 + 14xy - 10y^2$. Find another example of the same sort which is not equivalent to this form or a constant times it. *Hint*: First find a separator line with the desired properties, without any labels along the line, then find a form realizing that separator line.

4. Show that a positive nonsquare number is the discriminant of some hyperbolic form whose topograph has a rotational skew symmetry if and only if the number is

the sum of two squares at least one of which is even.

5. Verify that the following discriminants are fully symmetric, so all primitive forms of that discriminant have mirror symmetry:

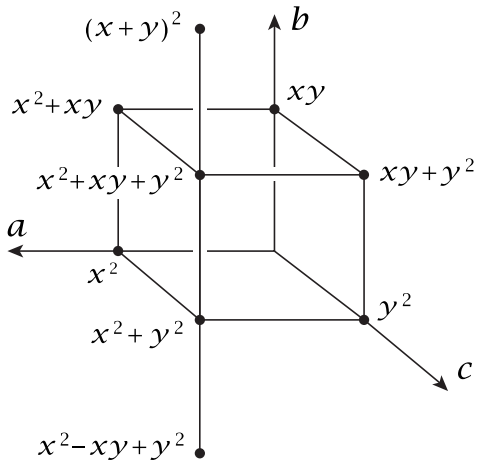
(a) -195 (b) -660 (c) 195

6. Show that the topograph of a primitive 0-hyperbolic form $qxy - py^2$ has mirror symmetry exactly when $p^2 \equiv 1 \mod q$, and has rotational skew symmetry exactly when $p^2 \equiv -1 \mod q$. (See the discussion at the end of Secion 2.1 about the relation between the continued fraction for p/q and the continued fraction obtained by reversing the order of the terms.)

5.5 Charting All Forms

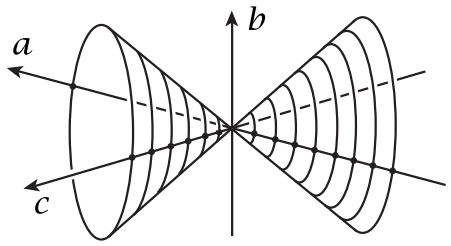
We have used the Farey diagram to study individual quadratic forms through their topographs, and in this section we will see that the Farey diagram also appears in another way when one creates a global picture mapping out all forms simultaneously. This viewpoint will not play an essential role in later chapters, however, so this section can be regarded as something of a digression from the main line of the book.

Quadratic forms are defined by formulas $ax^2+bxy+cy^2$, and our point of view will be to regard the coefficients a, b, and c as parameters that vary over all integers independently. It is natural to consider the triples (a,b,c) as points in 3-dimensional Euclidean space $\mathbb{R}^3$, and more specifically as points in the integer lattice $\mathbb{Z}^3$ consisting of points (a,b,c) whose coordinates are integers. We will exclude the origin $(0,0,0)$ since this corresponds to the trivial form that is identically zero. Instead of using the usual (x,y,z) as coordinates for $\mathbb{R}^3$ we will use (a,b,c), but since a and c play a symmetric role as the coefficients of the squared terms x^2 and y^2 we will position the a-axis and the c-axis in a horizontal plane, with the b-axis vertical, perpendicular to the ac-plane.



Along a ray starting at $(0,0,0)$ and passing through another lattice point (a,b,c) there are infinitely many lattice points (ka,kb,kc) for positive integers k. If a, b, and c have a greatest common divisor larger than 1 we can cancel this common divisor to get a primitive triple (a,b,c) corresponding to a primitive form $ax^2+bxy+cy^2$. Then all the other lattice points on the ray through (a,b,c) are the positive integer multiples (ka,kb,kc), corresponding to the nonprimitive forms $kax^2+kbxy+kcy^2$. Thus primitive forms correspond exactly to rays from the origin passing through lattice points. These are the same as rays passing through points (a,b,c) with rational

coordinates since denominators can always be eliminated by multiplying a, b, and c by a common denominator.

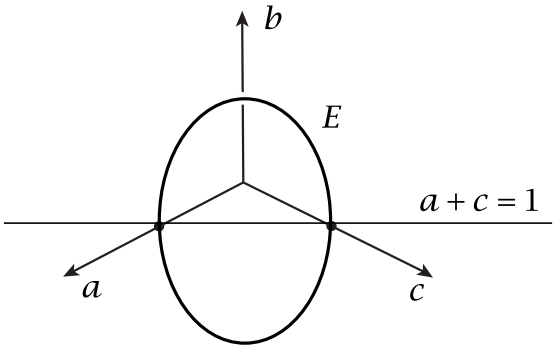


Since the discriminant $\Delta = b^2 - 4ac$ plays such an important role in the classification of forms, let us see how this fits into the picture in (a,b,c) coordinates. When $b^2 - 4ac$ is zero we have the special class of parabolic forms, and the points in $\mathbb{R}^3$ satisfying the equation $b^2 - 4ac = 0$ form a double cone with the common vertex of the two cones at the origin. The double cone intersects the ac-plane in the a-axis and the c-axis. The central axis of the double cone is the line $a = c$ in the ac-plane. Parabolic forms are the lattice points on these cones.

Elliptic and Parabolic Forms

Points (a,b,c) inside either cone have $b^2 - 4ac < 0$ so the lattice points inside the cones correspond to elliptic forms. Positive elliptic forms have $a > 0$ and $c > 0$ so they lie inside the cone projecting to the first quadrant of the ac-plane. We call this the *positive cone*. Inside the other cone are the negative elliptic forms, those with $a < 0$ and $c < 0$. Outside the cones is a single region consisting of points with $b^2 - 4ac > 0$ so the lattice points here correspond to hyperbolic forms and 0-hyperbolic forms.

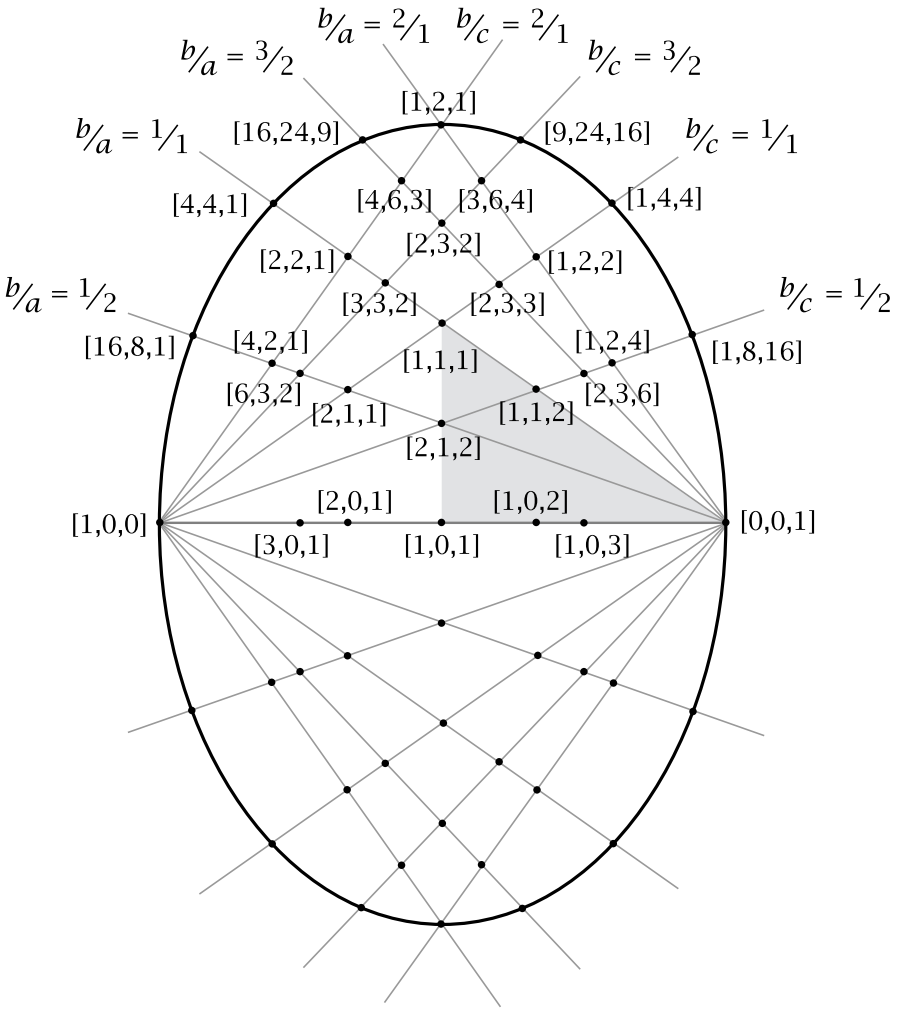
If one slices the positive cone via the vertical plane $a + c = 1$ perpendicular to the axis of the cone then the intersection of the cone with this plane is an ellipse which we denote E. The top and bottom points of E are $(a,b,c) = (\frac{1}{2}, \pm 1, \frac{1}{2})$ so its height is 2. The left and right points of E are $(1,0,0)$ and $(0,0,1)$ so its width is $\sqrt{2}$. Thus E is somewhat elongated vertically. If we wanted, we could compress the vertical coordinate to make E a circle, but there is no special advantage to doing this.



If we take a lattice point (a,b,c) corresponding to a primitive positive elliptic form and project this lattice point along the ray to the origin passing through (a,b,c), this ray intersects the plane $a + c = 1$ in the point $(\frac{a}{a+c}, \frac{b}{a+c}, \frac{c}{a+c})$ since this is the rescaling of (a,b,c) for which the sum of the first and third coordinates is 1. This point lies inside the ellipse E and has rational coordinates. Conversely, every point inside E with rational coordinates is the radial projection of a unique primitive positive elliptic form, obtained by multiplying the coordinates of the point by the least common multiple of their denominators. Thus the rational points inside E parametrize primitive positive elliptic forms. We will use the notation $[a,b,c]$ to denote both the form $ax^2 + bxy + cy^2$ and the corresponding rational point $(\frac{a}{a+c}, \frac{b}{a+c}, \frac{c}{a+c})$

inside E.

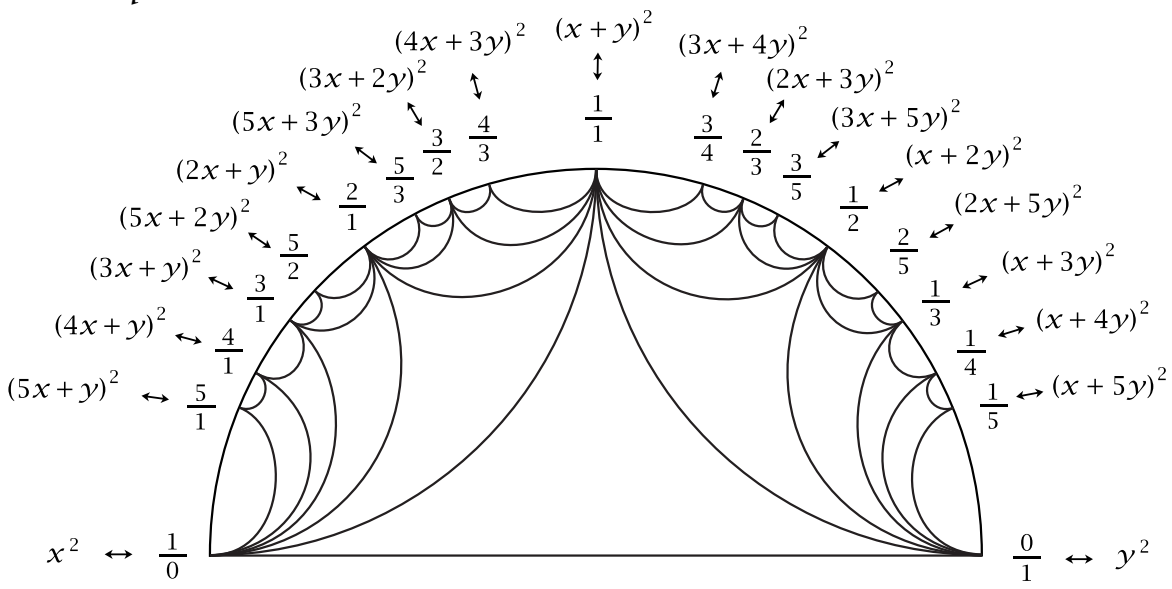
The figure below shows some examples, including a few parabolic forms on E itself. The lines radiating out from the points $[1,0,0]$ and $[0,0,1]$ consist of the points $[a,b,c]$ with a fixed ratio b/c or b/a respectively. The ratios a/c are fixed along vertical lines. For most points inside E any two out of these three ratios determine the third since $b/a \cdot a/c = b/c$. The exceptions are the points on the segment between $[1,0,0]$ and $[0,0,1]$ where b/a and b/c are both 0 but a/c can be anything.



Of special interest are the reduced primitive elliptic forms $[a,b,c]$, which are the ones satisfying $0 \le b \le a \le c$ where a, b, and c have no common divisor. These correspond to the rational points in the shaded triangle in the figure with vertices $[1,1,1]$, $[1,0,1]$, and $[0,0,1]$. The edges of the triangle correspond to one of the three inequalities $0 \le b \le a \le c$ becoming an equality, so $b = 0$ for the lower edge, $a = c$ for the vertical edge, and $a = b$ for the hypotenuse. Thus the three edges correspond to the reduced forms with mirror symmetry, the forms $[a,0,c]$ for the bottom edge, $[a,b,a]$ for the left edge, and $[a,a,c]$ for the diagonal edge. The vertices $[1,0,1]$ and $[1,1,1]$ correspond to the reduced elliptic forms with more than one mirror symmetry, and hence also rotational symmetry. Points in the interior of the triangle correspond to forms with no symmetry.

Just as rational points inside the ellipse E correspond to primitive positive elliptic forms, the rational points on E itself correspond to primitive positive parabolic forms. As we saw in Section 5.2, every parabolic form is equivalent to a form ax^2 for some nonzero integer a. For this to be primitive means that $a = \pm 1$, so every positive primitive parabolic form is equivalent to x^2. Equivalent forms can be obtained from each other by a change of variables, replacing (x, y) by $(px + qy, rx + sy)$ for integers p, q, r, s satisfying $ps - qr = \pm 1$. For the form x^2 this means that the primitive positive parabolic forms are the forms $(px+qy)^2 = p^2x^2 + 2pqxy + q^2y^2$ for coprime integers p and q. In $[a, b, c]$ notation this is $[p^2, 2pq, q^2]$, defining a point on the ellipse E.

More concisely, we could label the rational point on E corresponding to the form $(px + qy)^2$ just by the fraction p/q. Thus at the left and right sides of E we have the fractions $1/0$ and $0/1$ corresponding to the forms x^2 and y^2, while at the top and bottom of E we have $1/1$ and $-1/1$ corresponding to $(x+y)^2$ and $(x-y)^2 = (-x+y)^2$. Changing the signs of both p and q does not change the form $(px + qy)^2$ or the fraction p/q.



In the first quadrant of the ellipse the fractions p/q increase monotonically from $0/1$ to $1/1$ since the ratio b/c equals $2p/q$ and b is increasing while c is decreasing so $2p/q$ is increasing, and hence so is p/q. Similarly in the second quadrant the values of p/q increase from $1/1$ to $1/0$ since we have $b/a = 2q/p$ which decreases as b decreases and a increases. In the lower half of the ellipse we have just the negatives of the values in the upper half since the sign of b has changed from plus to minus.

This labeling of the rational points of E by fractions p/q seems very similar to the labeling of vertices in the circular Farey diagram. As we saw in Section 1.1, if the Farey diagram is drawn with $1/0$ at the top of the unit circle in the xy-plane, then the point on the unit circle labeled p/q has coordinates $(x, y) = (2pq/p^2+q^2, p^2-q^2/p^2+q^2)$. After rotating the circle to put $1/0$ on the left side by replacing (x, y) by $(-y, x)$

this becomes $(q^2-p^2/p^2+q^2, 2pq/p^2+q^2)$. Here the y-coordinate $2pq/p^2+q^2$ is the same as the b-coordinate of the point of E labeled p/q, which is the point $(a,b,c) = (p^2/p^2+q^2, 2pq/p^2+q^2, q^2/p^2+q^2)$. Since the vertical coordinates of points in either the left or right half of the unit circle or the ellipse E determine the horizontal coordinates uniquely, this means that the labeling of points of E by fractions p/q is really the same as in the circular Farey diagram.

Change of Variables

Let us return now to the general picture of how forms $ax^2 + bxy + cy^2$ are represented by points (a,b,c) in $\mathbb{R}^3$. As we know, a change of variables by a linear transformation T sends (x,y) to $T(x,y) = (px + qy, rx + sy)$, where p,q,r,s are integers with $ps - qr = \pm 1$. This change of variables transforms each form into an equivalent form. To see the effect of this change of variables on the coefficients (a,b,c) of a form $Q(x,y) = ax^2 + bxy + cy^2$ we do a simple calculation:

$$\begin{aligned} Q(px+qy, rx+sy) &= a(px+qy)^2 + b(px+qy)(rx+sy) + c(rx+sy)^2 \\ &= (ap^2 + bpr + cr^2)x^2 + (2apq + bps + bqr + 2crs)xy \\ &\qquad + (aq^2 + bqs + cs^2)y^2 \end{aligned}$$

This means that the (a,b,c) coordinates of points in $\mathbb{R}^3$ are transformed according to the following formula:

$$T^*(a,b,c) = (p^2a + prb + r^2c, 2pqa + (ps+qr)b + 2rsc, q^2a + qsb + s^2c)$$

For fixed values of p,q,r,s this T^* is a linear transformation of the variables a,b,c. Its matrix is:

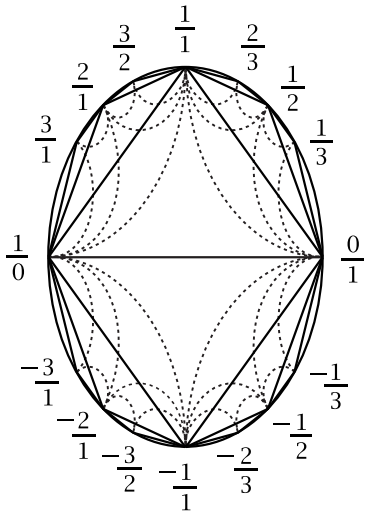
$$\begin{pmatrix} p^2 & pr & r^2 \\ 2pq & ps+qr & 2rs \\ q^2 & qs & s^2 \end{pmatrix}$$

Since T^* is a linear transformation, it takes lines to lines and planes to planes, but T^* also has another special geometric property. Since equivalent forms have the same discriminant, this means that each surface defined by an equation $b^2 - 4ac = k$ for k a constant is taken to itself by T^*. In particular, the double cone $b^2 - 4ac = 0$ is taken to itself, and in fact each of the two cones separately is taken to itself since one cone consists of positive parabolic forms and the other cone of negative parabolic forms (as one can see just by looking at the coefficients a and c), and positive parabolic forms are never equivalent to negative parabolic forms. When $k > 0$ the surface $b^2 - 4ac = k$ is a hyperboloid of one sheet and when $k < 0$ it is a hyperboloid of two sheets. In the case of two sheets the lattice points on one sheet give positive elliptic forms and the lattice points on the other sheet give negative elliptic forms.

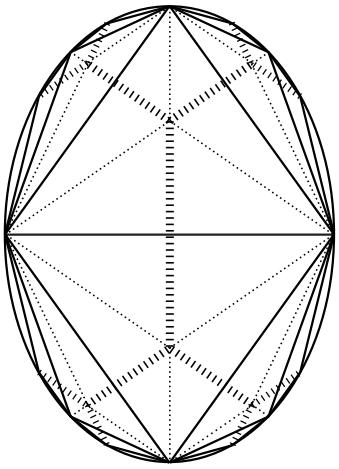
Since T^* takes lines through the origin to lines through the origin and the double cone $b^2 - 4ac = 0$ to itself, this means that T^* gives a transformation of the ellipse E

to itself, taking rational points to rational points since rational points on E correspond to lattice points on the cones. Regarding E as the boundary circle of the Farey diagram, we know that linear fractional transformations give symmetries of the Farey diagram, also taking rational points on the boundary circle to rational boundary points. And in fact, the transformation of this circle defined by T^* is exactly one of these linear fractional transformations. This is because T^* takes the parabolic form $(dx+ey)^2$ to the form $(d(px+qy)+e(rx+sy))^2 = ((dp+er)x+(dq+es)y)^2$ so in the fractional labeling of points of E this says $T^*(d/e) = (pd+re)/(qd+se)$ which is a linear fractional transformation. If we write this using the variables x and y instead of d and e it would be $T^*(x/y) = (px+ry)/(qx+sy)$. This is not quite the same as the linear fractional transformation $T(x/y) = (px+qy)/(rx+sy)$ defined by the original change of variables $T(x,y) = (px+qy, rx+sy)$, but rather T^* is obtained from T by transposing the matrix of T, interchanging the off-diagonal terms q and r.

Via radial projection, the transformation T^* determines a transformation not just of E but also of the interior of E in the plane $a+c=1$. This transformation, which we still call T^* for simplicity, takes lines inside E to lines inside E since T^* takes planes through the origin to planes through the origin. This leads us to consider a linear version of the Farey diagram in which each circular arc of the original Farey diagram is replaced by a straight line segment joining the two endpoints of the circular arc. These line segments divide the interior of E into triangles, just as the original Farey diagram divides the disk into curvilinear triangles. The transformation T^* takes each of these triangles onto another triangle, analogous to the way that linear fractional transformations provide symmetries of the original Farey diagram.



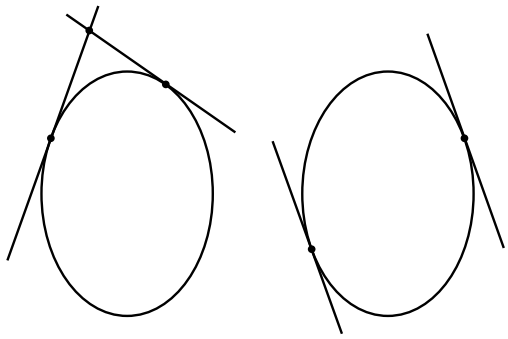
Suppose we divide each triangle of the linear Farey diagram into six smaller triangles as in the figure at the right, by adding diagonals to each quadrilateral formed by two adjacent triangles of the Farey diagram. The transformation T^* takes each of these small triangles onto another small triangle since it takes lines to lines. One of these small triangles is the triangle defined by the inequalities $0 \le b \le a \le c$ that we considered earlier. The fact that every positive primitive elliptic form is equivalent to exactly one reduced form, corresponding to a rational point in this special triangle, is now visible geometrically as the fact that there is always exactly one transformation T^* taking a given small triangle to this one special small triangle.



Elliptic forms whose topograph contains a source edge are equivalent to forms $ax^2 + cy^2$ so these are the forms corresponding to rational points on the edges of the original linear Farey diagram, before the subdivision into smaller triangles. These are the forms whose topograph has a symmetry reflecting across a line perpendicular to the source edge. (This line is just the edge in the Farey diagram containing the given form.) The other type of reflectional symmetry in the topograph of an elliptic form is reflection across an edge of the topograph. Forms with this sort of symmetry correspond to rational points in the dotted edges in the preceding figure, the edges we added to subdivide the Farey diagram into the smaller triangles. The dotted edges are of two types according to whether the two equal values of the form in the three regions surrounding the source vertex occur for the smallest value of the form (wide dotted edges) or the next-to-smallest value of the form (narrow dotted edges). The wide dotted edges form the dual tree of the Farey diagram.

Hyperbolic and 0-Hyperbolic Forms

Hyperbolic and 0-hyperbolic forms correspond to integer lattice points that lie outside the two cones. For a point (a, b, c) outside the double cone there are exactly two planes in $\mathbb{R}^3$ that are tangent to the double cone and pass through (a, b, c). Each of these planes is tangent to the double cone along a line through the origin. The two tangent planes through (a, b, c) are determined by their intersection with the plane $a + c = 1$, which consists of two lines tangent to the ellipse E. These two lines can either intersect or be parallel. The latter possibility occurs when the point (a, b, c) lies in the plane $a + c = 0$, so the two tangent planes intersect in a line in this plane. For example, if the point (a, b, c) we start with happens to lie on the b-axis, then the tangent planes are the ab-plane and the bc-plane. These intersect the plane $a + c = 1$ in the two vertical tangent lines to the ellipse E.



Our goal will be to show the following:

Proposition 5.11. *Let $Q(x, y) = ax^2+bxy+cy^2$ be a form of positive discriminant, either hyperbolic or 0-hyperbolic. Then the two points where the tangent lines to E determined by (a, b, c) touch E are the points diametrically opposite the two points that are the endpoints of the separator line in the topograph of Q in the case that Q is hyperbolic, or the two points labeling the regions in the topograph of Q where Q takes the value zero in the case that Q is 0-hyperbolic.*

Proof: We begin with a few preliminary remarks that will allow us to treat the hyperbolic and 0-hyperbolic cases in the same way. A form $Q(x, y) = ax^2 + bxy + cy^2$

of positive discriminant can always be factored as $(px+qy)(rx+sy)$ with p,q,r,s real numbers since if $a=0$ we have the factorization $y(bx+cy)$ and if $a\neq 0$ then the associated quadratic equation $ax^2+bx+c=0$ has positive discriminant so it has two distinct real roots α and β. This leads to the factorization $ax^2+bxy+cy^2 = a(x-\alpha y)(x-\beta y)$ which can be rewritten as $(px+qy)(rx+sy)$ by incorporating a into either factor. If Q is hyperbolic then the discriminant is not a square and hence the factorization $(px+qy)(rx+sy)$ will involve coefficients that are quadratic irrationals. If Q is 0-hyperbolic then the discriminant is a square so the roots α and β are rational and we obtain a factorization of Q as $(px+qy)(rx+sy)$ with rational coefficients. In fact we can take p,q,r,s to be integers in this case since we know every 0-hyperbolic form is equivalent to a form $y(bx+cy)$ so we can obtain the given form Q from $y(bx+cy)$ by replacing x and y by certain linear combinations $dx+ey$ and $fx+gy$ with integer coefficients d,e,f,g.

The points where the tangent planes touch the double cone correspond to forms of discriminant zero, with coefficients that may not be integers or even rational. A simple way to construct two such forms from a given form $Q=(px+qy)(rx+sy)$ is just to take the squares of the two linear factors, so we obtain the forms $(px+qy)^2$ and $(rx+sy)^2$, each of discriminant zero. We will show that each of these two forms lies on the line of tangency for one of the two tangent planes determined by Q.

To do this for the case of $(px+qy)^2$ we consider the line L in $\mathbb{R}^3$ passing through the two points corresponding to the forms $(px+qy)(rx+sy)$ and $(px+qy)^2$. We claim that L consists of the forms

$$Q_t(x,y) = (px+qy)\Big[(1-t)(rx+sy)+t(px+qy)\Big]$$

as t varies over all real numbers. When $t=0$ or $t=1$ we obtain the two forms $Q_0=(px+qy)(rx+sy)$ and $Q_1=(px+qy)^2$ so these forms lie on L. Also, we can see that the forms Q_t do form a straight line in $\mathbb{R}^3$ by rewriting the formula for $Q_t(x,y)$ as $ax^2+bxy+cy^2$ with the coefficients a,b,c given by:

$$(a,b,c) = \big(pr(1-t)+p^2t,\,(ps+qr)(1-t)+2pqt,\,qs(1-t)+q^2t\big)$$

This defines a line since p,q,r,s are constants, so each coordinate is a linear function of t. Since the forms Q_t factor as the product of two linear factors, they have nonnegative discriminant for all t. This means that the line L does not go into the interior of either cone. It also does not pass through the origin since if it did, it would have to be a subset of the double cone since it contains the form Q_1 which lies in the double cone. From these facts we deduce that L must be a tangent line to the double cone. Hence the plane containing L and the origin must be tangent to the double cone along the line containing the origin and Q_1. The same reasoning shows that the other tangent plane that passes through $(px+qy)(rx+sy)$ intersects the double cone along the line containing the origin and $(rx+sy)^2$.

The labels of the points of E corresponding to the two forms $(px+qy)^2$ and $(rx+sy)^2$ are p/q and r/s according to the convention we have adopted. On the other hand, when the form $(px+qy)(rx+sy)$ is hyperbolic the ends of the separator line in its topograph are at the two points where this form is zero, which occur when x/y is $-q/p$ and $-s/r$. These are the negative reciprocals of the previous two points p/q and r/s so they are the diametrically opposite points in E. Similarly, when the form $(px+qy)(rx+sy)$ is 0-hyperbolic the vertices of the Farey diagram where it is zero are at $-q/p$ and $-s/r$, again diametrically opposite p/q and r/s. □

It might have been nicer if the statement of the previous proposition did not involve passing to diametrically opposite points, but to achieve this we would have had to use a different rule for labeling the points of E, with the label p/q corresponding to the form $(qx-py)^2$ instead of $(px+qy)^2$. This 180 degree rotation of the labels would put the negative labels in the upper half of E rather than the lower half, which does not seem like a good idea.

Next let us investigate how hyperbolic and 0-hyperbolic forms are distributed over the lattice points outside the double cone $b^2-4ac=0$. This is easier to visualize if we project such points radially into the plane $a+c=1$. This only works for forms $ax^2+bxy+cy^2$ with $a+c>0$, but the forms with $a+c<0$ are just the negatives of these so they give nothing essentially new. The forms with $a+c=0$ will be covered after we deal with those with $a+c>0$.

Forms with $a+c>0$ that are hyperbolic or 0-hyperbolic correspond via radial projection to points in the plane $a+c=1$ outside the ellipse E. As we have seen, each such point determines a pair of tangent lines to E intersecting at the given point.

For a 0-hyperbolic form $(px+qy)(rx+sy)$ the points of tangency in E have rational labels p/q and r/s. We know that every 0-hyperbolic form is equivalent to a form $y(rx+sy)$ with $a=0$, so $p/q=0/1$ and one line of tangency is the vertical line tangent to E on the right side. The form $y(rx+sy)$ corresponds to the point $(0,r,s)$ in the plane $a=0$ tangent to the double cone. Projecting radially into the vertical tangent line to E, we obtain the points $(0,r/s,1)$, where r/s is an arbitrary rational number. Thus 0-hyperbolic forms are dense in this vertical tangent line to E. Choosing any rational number r/s, the other tangent line for the form $y(rx+sy)$ is tangent to E at the point labeled r/s.

An arbitrary 0-hyperbolic form $(px+qy)(rx+sy)$ is obtained from one with $p/q=0/1$ by applying a linear fractional transformation T taking $0/1$ to p/q, so the vertical tangent line to E at $0/1$ is taken to the tangent line at p/q, and the dense set of 0-hyperbolic forms in the vertical tangent line is taken to a dense set of 0-hyperbolic forms in the tangent line at p/q. Thus we see that the 0-hyperbolic forms in the plane $a+c=1$ consist of all the rational points on all the tangent lines to E at rational points p/q of E.

In the case of a hyperbolic form $ax^2+bxy+cy^2$ with $a+c>0$ the two tangent lines intersect E at a pair of conjugate quadratic irrationals, the negative reciprocals of the roots α and $\overline{\alpha}$ of the equation $ax^2+bx+c=0$. Since α determines $\overline{\alpha}$ uniquely, one tangent line determines the other uniquely, unlike the situation for 0-hyperbolic forms whose rational tangency points p/q and r/s can be varied independently. A consequence of this uniqueness for hyperbolic forms is that each of the two tangent lines contains only one rational point, the intersection point of the two lines. This is because any other rational point would correspond to another form having one of its tangent lines the same as for $ax^2+bxy+cy^2$ and the other tangent line different, contradicting the previous observation that each tangent line for a hyperbolic form determines the other. (The hypothetical second form would also be hyperbolic since the common tangency point for the two forms is not a rational point on E.)

The points in the plane $a+c=1$ that correspond to 0-hyperbolic forms are dense in the region of this plane outside E since for an arbitrary point in this region we can first take the two tangent lines to E through this point and then take a pair of nearby lines that are tangent at rational points of E since points in E with rational labels are dense in E. It is also true that points in the plane $a+c=1$ that correspond to hyperbolic forms are dense in the region outside E. To see this we can proceed in two steps. First consider the case of a point in this region whose two tangent lines to E are tangent at irrational points of E. These two irrational points are the endpoints of an infinite strip in the Farey diagram that need not be periodic. However we can approximate this strip by a periodic strip by taking a long finite segment of the infinite strip and then repeating this periodically at each end. This means that the given point in the region outside E lies arbitrarily close to points corresponding to hyperbolic forms. Finally, a completely arbitrary point in the region outside E can be approximated by points whose tangent lines to E touch E at irrational points since irrational numbers are dense in real numbers.

It remains to consider hyperbolic and 0-hyperbolic forms $(px+qy)(rx+sy)$ corresponding to points (a,b,c) in the plane $a+c=0$. Such a form determines a line through the origin in this plane, and the tangent planes to the double cone that intersect in this line intersect the plane $a+c=1$ in two parallel lines tangent to E at two diametrically opposite points p/q and $-q/p$. This means that the form is $(px+qy)(qx-py)$, up to a constant multiple. If p/q is rational this is a 0-hyperbolic form. Examples are:

- xy with vertical tangents to E at $1/0$ and $0/1$.
- $x^2-y^2=(x+y)(x-y)$ with horizontal tangents to E at $1/1$ and $-1/1$.
- $2x^2-3xy-2y^2=(2x+y)(x-2y)$ with parallel tangents at $2/1$ and $-1/2$.

If p/q and $-q/p$ are conjugate quadratic irrationals then we have a hyperbolic form $ax^2+bxy+cy^2=a(x-\alpha)(x-\overline{\alpha})$ where $\alpha\overline{\alpha}=-1$ since $c=-a$ when $a+c=0$. Thus α and $\overline{\alpha}$ are negative reciprocals of each other that are interchanged by 180

degree rotation of E. As examples we have:

$$x^2 + xy - y^2 = \left(x - \frac{-1+\sqrt{5}}{2}y\right)\left(x - \frac{-1-\sqrt{5}}{2}y\right)$$

$$2x^2 + xy - 2y^2 = 2\left(x - \frac{-1+\sqrt{17}}{4}y\right)\left(x - \frac{-1-\sqrt{17}}{4}y\right)$$

One can consider a pair of parallel tangent lines to E as the limit of a pair of intersecting tangents where the point of intersection moves farther and farther away from E in a certain direction which becomes the direction of the pair of parallel tangents.