APPENDIX A: The Geometry of the Gauss Product

Reprinted from:
R.C. Penner, The Geometry of the Gauss Product, Algebraic Geometry 4, Journal of Mathematical Science **81** (1996) (Festschrift for Yuri Manin), 2700-2718.

The Geometry of the Gauss Product

R. C. Penner

Introduction

In *Disquisitiones Arithmeticae* [1801], Gauss defined a law of composition of $PSL(2,\mathbf{Z})$ classes of suitable binary integral quadratic forms. Here we give a new geometric interpretation of this Gauss product in the case of definite forms; indeed, we shall find that the product is intimately connected with incidences of hypercycles (that is, loci equidistant to a geodesic) on the modular curve, and the product will be found to be analogous to addition on a non-singular cubic but using suitable hypercycles (instead of lines). We shall also elaborate briefly on the case of indefinite forms, which was actually our starting point.

This entire note is based on the group $\Gamma = PSL(2,\mathbf{Z})$, which is intended as a paradigm for the general case of a finite-index subgroup $\Gamma < PSL(2,\mathbf{Z})$. Throughout our discussion, though, we shall keep in mind the more general situation, say, of torsion-free finite-index subgroups $\Gamma < PSL(2,\mathbf{Z})$. Many of our constructions generalize readily as we briefly discuss at the end; however, a suitable geometric interpretation of the Gauss product should give natural analogues of the Gauss groups for each such Γ, and this we have not achieved.

Given the very classical nature of what we describe here and the activity in this realm during the period 1940-1970, we remain surprised that this picture of the Gauss product seems to be new. On the other hand, our main result is really about an algorithm for computing Gauss products rather than about the product itself. We can furthermore imagine that nobody bothered to return to the baby quadratic case of ideal class groups in the special case of definite forms armed not only with hyperbolic geometry but also with the 1968 extension of the product described by Butts-Estes-Pall in [BE] and [BP]. Indeed, recent surveys have described only special cases of this extension (and in fact, we must extend their formulation a bit further still below).

We have strived to keep this note entirely self-contained starting from scratch at least in the definite case. The only exceptions are that some routine calculations will

be suppressed, our survey below of the contemporary number-theoretic point of view on this is, after all, just a survey without proofs, and our starting point is Dirichlet's [1851] formulation of the product rather than Gauss'.

We begin by gently introducing and surveying the elementary algebra from first principles and then go on to give a self-contained proof of the existence of the Gauss product in the general (i.e., either definite or indefinite) case in Theorem 6, and our new techniques are already of value here (cf. Lemma 1 below); it is worth emphasizing that we are not saying anything new about the Gauss product at this point other than that there is a well-definedness property of an algorithm for computing it. (Experts should certainly skip the first two sections which are elementary, partly expository, and included for completeness and just glance at Lemma 1 and Theorem 6 in the third section.) We then specialize to the definite case and undertake the geometric study of fundamental roots. Our main result is Theorem 13 giving an explicit geometric formulation of the Gauss product, which is in a sense insufficient for a completely geometric description as we shall discuss. Various number-theoretic and geometric points are finally described in closing remarks, but it remains to be seen whether our formulation of the product might be of real utility in number theory; in the other direction, we can say that the existing databases of class numbers and related data can be interpreted as describing various (reasonably arcane) enumerative behaviors of hypercycles in the modular curve.

This note was originally composed as a letter from the author to Yuri Manin on the happy occasion of his birthday, and this explains the informal parenthetical remarks, most of which I have decided to leave in the text.

I am lucky to have Dennis Estes as a colleague here at USC and want to thank him for sharing his time and insights over the last months and years. Let me also thank Francis Bonahon, Bob Guralnick, Dennis Sullivan, and especially Don Zagier for helpful and stimulating questions, comments, and corrections. I finally wanted to praise [Ca] (which has been my basic reference) as well as [Za] and to acknowledge the support of the National Science Foundation.

Notation, Basics, and Context

We study here integral quadratic forms defined on the two-dimensional lattice $\mathbf{Z}^2$, i.e., we study expressions

$$f(x,y) = ax^2 + bxy + cy^2, \text{ for } a,b,c \in \mathbf{Z} \text{ and } x,y \in \mathbf{Z},$$

and we shall typically write simply $f = [a,b,c]$, referring to a,b,c respectively as the "first, middle, last" coefficient of f. We say that f is *primitive* if $gcd\{a,b,c\} = 1$ and shall also call a lattice point (x,y) *primitive* if $gcd\{x,y\} = 1$.

Of course, the symmetric bilinear form corresponding to $[a,b,c]$ is $B_f = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}$,

so $f(x,y) = (x,y) \cdot B_f \cdot (x,y)^t$, where $\cdot^t$ denotes the transpose. The *discriminant* of f is

$$D = D(f) = b^2 - 4ac = -4 \; det \; B_f,$$

where $det \; B_f$ is called the *determinant* of f. Notice that

$$b \equiv 0 (mod \; 2) \Leftrightarrow D \equiv 0 (mod \; 4)$$
$$b \equiv 1 (mod \; 2) \Leftrightarrow D \equiv 1 (mod \; 4)\,,$$

so in particular $b \equiv D (mod \; 2)$, and D is always equivalent to either 0 or 1 mod 4. We say that f is *definite* if $D(f) < 0$, that f is *indefinite* if $D(f) > 0$, and that f is *singular* if $D(f) = 0$. Of course, if f is definite, then a and c have the same sign (which we shall often take to be positive).

The natural action of $\gamma \in PSL(2,\mathbf{Z})$ (or $SL(2,\mathbf{Z})$) is by change of basis $f(\vec{x}) \mapsto f(\gamma \cdot \vec{x})$ on quadratic forms. The corresponding action on symmetric bilinear forms is $B_f \mapsto \gamma^t \cdot B_f \cdot \gamma = B_{f'}$, and we write $\gamma.f = f'$ and $f \sim f'$ in this case. The action evidently leaves invariant the discriminant and furthermore preserves primitivity since a form f is not primitive if and only if there is some prime dividing each element of $f(\mathbf{Z}^2)$ (cf. Lemma 2 below).

This action induces the natural equivalence relation on the set of forms, and if $f = [a,b,c]$ is a quadratic form, then we shall write $[f] = [[a,b,c]]$ for the class of f. In light of the previous remarks, both primitivity and the discriminant of a class are well-defined. Given a discriminant $D \in \mathbf{Z}$ (i.e., an integer equivalent to either 0 or 1 mod 4), define the *Gauss group*

$$\mathcal{G}(D) = \{[f] : f \text{ is primitive and } D(f) = D\},$$

which at this moment is to be regarded as just a set. When D is fixed, we shall write simply $\mathcal{G} = \mathcal{G}(D)$.

In fact, for each D, $\mathcal{G}(D)$ is a finite set. We shall not prove this here other than to say that one first proves the interesting fact (due to Hermite) that given a primitive f of discriminant D and determinant $d = -D/4$, there is some primitive vector $\vec{x}$ so that $f(\vec{x}) \neq 0$ and $|f(\vec{x})| \leq |d|^{\frac{1}{2}}$. (This, in turn, is proved by simply completing the square in $4af(x,y)$ and applying the Division Algorithm.)

In *Disquisitiones Arithmeticae*, Gauss defined a finite abelian group structure on each $\mathcal{G}(D)$, which is "essentially" (cf. below) the ideal class group of $\mathbf{Q}(\sqrt{D})$, and here is his original idea. Given forms $f_i, i = 1,2,3$, we shall think of each with its own copy $(x_i, y_i) \in \mathbf{Z}^2$ of the lattice, so

$$f_i(x_i y_i) = a_i x_i^2 + b_i x_i y_i + c_i y_i^2.$$

Following Gauss, we say that $[f_1][f_2] = [f_3]$ if $f_3(x_3, y_3)$ is transformed into the pointwise product $f_1(x_1, y_1) f_2(x_2, y_2)$ by a transformation

$$(x_3, y_3) = T \cdot \begin{pmatrix} x_1 x_2 \\ x_1 y_2 \\ x_2 y_1 \\ y_1 y_2 \end{pmatrix},$$

where T is a four-by-two integral matrix whose two-by-two minors generate $\mathbf{Z}$ as an ideal over $\mathbf{Z}$ (plus a further technical condition on the signs).

We have included (nearly) Gauss' original definition just in order to view the product in this natural pointwise way. In fact, our starting point is actually Dirichlet's definition to be presented below. Our basic plan in the next two sections is to recall and then generalize Cassels' version [Ca] of Dirichlet's proof that the product is well-defined on classes of forms, and then in subsequent sections to discuss the underlying geometry in the definite case. To close this section, let us give a quick description of some of the number-theoretic significance and context of this Gauss product. (It is problematic as I am sure you are more expert here than I, but I proceed nonetheless to record various facts mostly from [Ca].)

There is a further equivalence relation on each $\mathcal{G} = \mathcal{G}(D)$, where we say that two forms are in the same *genus* if for all primes p the two forms are equivalent as forms over the p-adic integers. (A specific example of two inequivalent forms in the same genus is $[1,0,82]$ and $[2,0,41]$.)

As an abelian group, $\mathcal{G}(D)$ has a unit which we shall denote $\mathbf{1} = \mathbf{1}_D \in \mathcal{G}(D)$, and the genus of the unit $\mathbf{1}_D$ is called the *principal genus*. An explicit form representing $\mathbf{1}_D$ will be given when we need it later, and we choose to write $\mathcal{G}$ multiplicatively for our notational convenience.

A celebrated calculation of Gauss (boiling down to the pigeon-hole principle!) described in [Ca] proves that the class $[f] \in \mathcal{G}$ of a form f lies in the principal genus if and only if $[f] = [g]^2$ for some $[g] \in \mathcal{G}$; that is, the principal genus is $\mathcal{G}^2$. In fact, for any finite abelian group, we may consider the kernel K and cokernel K^* of the squaring map $g \mapsto g^2$; since K and K^* are equinumerous $\mathbf{Z}/2$ vectorspaces, we find an (non-canonical) isomorphism $K \approx K^*$, so in our case, we find

$$\begin{aligned}
\mathcal{G}/\mathcal{G}^2 &= \{\text{genera}\} \\
&\approx \ker(\mathcal{G} \to \mathcal{G}^2) \\
&= \{[f] \in \mathcal{G} : [f]^2 = \mathbf{1}\} \\
&= \{\text{ambiguous classes}\},
\end{aligned}$$

where a class $[f]$ is said to be *ambiguous* if $[f]^2 = \mathbf{1}$. (The terminology is due to Gauss, and perhaps the idea is that these are the fixed points of the action of the absolute Galois group, which is by the way simply given in this quadratic context by $[a,b,c] \mapsto [a,-b,c]$.)

Let us next make precise the sense in which $\mathcal{G}(D)$ is "essentially" the ideal class group of $K = K(D) = \mathbf{Q}(\sqrt{D})$.

If D is either unity or the discriminant of a quadratic field, then it is said to be *fundamental*, so in the respective cases $D \equiv 1 (mod\ 4)$ and $D \equiv 0 (mod\ 4)$, we have equivalently that either D or $D/4$ is square-free (and in the latter case $D = 4\delta$, where δ is square free and equivalent to either 2 or 3 $mod\ 4$). Any discriminant D can be written

uniquely as $D = df^2$ where d is fundamental, and any fundamental discriminant is uniquely expressed as a product of prime discriminants, i.e., fundamental discriminants with one prime factor, the list of such being -4, ± 8, $+p$ with $p \equiv 1 (mod\ 4)$ prime, and $-q$ with $q \equiv 3 (mod\ 4)$ prime. (The fundamental discriminants are the basic ones in the sense that all class numbers are calculable in terms of those of fundamental discriminants; cf. below.)

One must introduce the finer equivalence relation of "strict" equivalence on ideals where the ideal A is identified with the ideal xA for $x \in K$ provided the norm of x is positive (this agrees with the usual notion of equivalence in the definite case), and the corresponding group of ideal classes is the "strict" ideal class group of K. This strict ideal class group surjects onto the usual ideal class group, and the kernel is of order one in the definite case and of order either one or two in the indefinite case. (Indeed for fundamental discriminants, the kernel has order 1 if and only if the ring of integers of K has a unit of norm -1.)

The strict ideal class group is isomorphic to the group $\mathcal{G}(D)$ for fundamental discriminants $D \neq 1$ in the indefinite case $D > 0$. In the definite case, one must specialize further and (following Gauss) consider only *positive* definite forms to construct a Gauss group $\mathcal{G}_+(D)$ (so $\mathcal{G} \approx \mathcal{G}_+ \times \mathbf{Z}/2\mathbf{Z}$), and then it is $\mathcal{G}_+(D)$ which is isomorphic to the strict ideal class group for fundamental discriminants $D < 0$.

In either the indefinite or definite case, though, there are vast databases of class numbers available (as well as related data). In fact, we were surprised to learn that this formulation of class numbers in terms of quadratic forms is perhaps the most tractable approach computationally, and as a practical matter, special values of Dedekind L functions are actually estimated in terms of quadratic form data (rather than the other way around!). We shall see later how to interpret this known data in the definite case in terms of the geometry of hypercycles in the modular curve.

From a contemporary point of view, then, the Gauss groups can be thought of as a sort of quadratic pre-cursor to Kummer's ideal class groups, certainly at least in the definite case to which we shall turn our attention shortly. On the other hand, Gauss' genus theory (together with the Hasse-Minkowski invariant) is the contemporary formalism for the local-to-global theory of binary quadratic forms over $\mathbf{Z}$.

Dirichlet's Definition and Its Elementary Consequences

Let us first observe that if $[a, b, c]$ is a form of discriminant D, then we may solve for $c = \frac{b^2 - D}{4a}$ to conclude that $a | \frac{b^2 - D}{4}$, where we write $u|v$ for $u, v \in \mathbf{Z}$ if u divides v.

Now suppose that f_1 and f_2 are primitive forms of the same discriminant D. We say that f_1 and f_2 are *(Dirichlet) unitable* if their classes $[f_1]$ and $[f_2]$ admit respective representatives $[a_1, b_1, c_1]$ and $[a_2, b_2, c_2]$ where

(i) $b_1 = b_2 = b$,

(ii) $gcd\{a_1, a_2\} = 1$.

We shall say that the specific forms $[a_1, b, c_1]$ and $[a_2, b, c_2]$ are themselves *(Dirichlet) united* in this case.

According to our observation above, there is an integral form $[a_1a_2, b, *]$ of discriminant D if and only $* = \frac{b^2-D}{4a_1a_2}$ is integral; meanwhile, we similarly conclude that $4a_1|b^2 - D$ and $4a_2|b^2 - D$, so the assumed relative primality of a_1 and a_2 (together with the fact that $b \equiv D(mod\ 2)$) guarantees the existence of an integral form $[a_1a_2, b, *]$ of discriminant D. At the same time, since

$$\frac{b^2 - D}{4a_1a_2} = \frac{c_1}{a_2} = \frac{c_2}{a_1},$$

one can check without pain that the form $[a_1a_2, b, *]$ is primitive if the forms f_1 and f_2 are primitive.

In fact, Dirichlet proved that two classes of primitive forms $[f_1], [f_2]$ of the same discriminant are unitable, and if $[a_1, b, c_1], [a_2, b, c_2]$ and $[a_1', b', c_1'], [a_2', b', c_2']$ are pairs of united representatives of $[f_1]$, $[f_2]$, respectively, then $[[a_1a_2, b, *]] = [[a_1'a_2', b', *']]$. Thus, the *(Gauss) product*

$$[f_1][f_2] = [[a_1a_2, b, *]]$$

is well-defined. This is Dirichlet's formulation of Gauss' product, and we shall prove (a generalization of) its well-definedness in Theorem 6 below.

Observe that the relative primality condition (ii) is not really so natural, for instance, it is not invariant under the action of $PSL(2, \mathbf{Z})$. A weaker (and in a sense weakest possible analogous) condition (in the notation above) is

(ii') $\frac{b^2-D}{4a_1a_2} \in \mathbf{Z}$, that is, $a_1a_2|\frac{b^2-D}{4}$,

A pair of forms satisfying conditions (i) and (ii') is said to be *(Cassels) concordant*, so a united pair is automatically concordant. Given a concordant pair as above, one defines a putative product using the same formula as before.

First of all, Cassels proves that the putative product is well-defined on concordance classes. We shall find that his proof generalizes handily to the more general setting of (a concordance/united type extension of) the Butts-Estes formulation to be discussed in the next section. Secondly, in the definite case, we shall prove a kind of $PSL(2, \mathbf{Z})$ invariance of concordant pairs geometrically.

Our immediate goal (in the next section) is to formulate a concordance version of the Butts-Estes-Pall product and (following Cassels) prove that this product is well-defined; only after these general considerations do we turn finally to definite forms and the modular curve.

For the remainder of this section, let us just assume temporarily that given two elements of $\mathcal{G}$, there are concordant representatives the class of whose product (as above) is well-defined, and let us investigate some of the elementary group-theoretic consequences.

It is convenient and traditional just now to call the equivalence relation generated by $PSL(2,\mathbf{Z})$ *(proper) equivalence* and denote it by $\sim$ or $\sim_p$; we shall say that two forms f, f' are *improperly equivalent* if there is a two-by-two integral matrix γ of determinant -1 so that $\gamma.f = f'$. (So improper equivalence is *not* an equivalence relation in this standard parlance.)

Here are three useful calculations and tricks:

- $[a,b,c]$ is properly equivalent to $\gamma.[a,b,c] = [c,-b,a]$ using the matrix $\gamma = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, and we call this a "flip".
- $[a,b,c]$ is properly equivalent to $\gamma.[a,b,c] = [a, b+2a\ell, a\ell^2 + b\ell + c]$ using the matrix $\gamma = \begin{pmatrix} 1 & \ell \\ 0 & 1 \end{pmatrix}$, and we call this "translation by ℓ".
- $[a,b,c]$ is improperly equivalent to $\gamma.[a,b,c] = [a,-b,c]$ using the matrix $\gamma = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$, and this is the action of Galois as was mentioned before.

The unit $\mathbf{1}_D$ of $\mathcal{G}(D)$ is evidently represented (with middle coefficient b) by $[1, b, \frac{b^2-D}{4}]$, and we may translate to arrange that $b = 0, 1$ so that

$$\mathbf{1}_D = \begin{cases} [[1, 0, -\frac{D}{4}]]; & \text{if } D \equiv 0 (mod\ 4), \\ [[1, 1, \frac{1-D}{4}]]; & \text{if } D \equiv 1 (mod\ 4). \end{cases}$$

We also have

$$[[a,b,c]]\ [[c,b,a]] = [[ac,b,1]] = [[1,-b,ac]] = \mathbf{1}_D,$$

so inversion in the group $\mathcal{G}$ is easily described in general as

$$[[a,b,c]]^{-1} = [[c,b,a]].$$

To close this section, we briefly discuss several generalities, and to begin, we claim that $[f]^2 = \mathbf{1}_D$ (that is, $[f]$ is ambiguous) if and only if any representative of $[f]$ is improperly equivalent to itself. Indeed, $[f]$ is ambiguous by definition if and only if $[f] = [f]^{-1}$, so if $[a,b,c]$ represents $[f]$, then $[a,b,c] \sim_p [c,b,a] \sim_p [a,-b,c]$, which is in turn improperly equivalent to $[a,b,c]$.

As to explicit representatives of ambiguous classes, we have the following two families of *ambiguous forms*

$$[a,0,c], \text{ for } D = -4ac,$$
$$[a,a,c], \text{ for } D = a(a-4c).$$

It is straight-forward to check that these forms represent ambiguous classes using the tricks and remarks above, but the proof that these represent all ambiguous classes requires a small further discussion of reduction theory.

For the reduction to canonical forms, one uses the tricks above and some further calculations to prove

- A definite form is equivalent to a "reduced" form $[a, b, c]$ where $|b| \leq a \leq c$, and with the exception of $[a, b, a] \sim [a, -b, a]$ and $[a, a, c] \sim [a, -a, c]$, no two reduced definite forms are equivalent.

- An indefinite form is equivalent to a "reduced" form $[a, b, c]$ with $0 < b < \sqrt{D}$ and $\sqrt{D} - b < 2|a| < \sqrt{D} + b$, and in this case, the reduced forms are partitioned into disjoint "cycles", where two reduced indefinite forms are equivalent if and only if they lie in a common cycle.

In either case, one considers the *principal root*

$$\omega = \frac{-b + \sqrt{D}}{2a},$$

so ω solves the quadric $az^2 + bz + c = 0$ and transforms as usual (by integral fractional linear transformations) under the action of $PSL(2, \mathbf{Z})$. In the definite case, a form is reduced if and only if the principal root lies in the usual fundamental domain for $PSL(2, \mathbf{Z})$. This is the classical geometric identification whose elaboration is really the main focus of this note. In the indefinite case, the corresponding identification is to consider the two real roots of $az^2 + bz + c = 0$ as the endpoints of a hyperbolic geodesic which is normalized (i.e., reduced) so that it hits the usual fundamental domain of $PSL(2, \mathbf{Z})$, and this geodesic corresponds to a cycle of equivalent reduced indefinite forms.

Returning finally to ambiguous classes in either case, the respective reduction algorithms show that every ambiguous class contains an ambiguous form. (Furthermore, among reduced definite forms, only $[a, a, a]$ and $[a, 0, a]$ admit isotropy in $PSL(2, \mathbf{Z})$, while isotropy for indefinite forms is related to integral solutions of the (positive) Pell equation $u^2 - Dv^2 = +4$.)

Concordance Extension of Butts-Estes United Forms

We shall say that two primitive classes $[f_1], [f_2]$ are *unitable* if $D(f_i) = t_i^2 d$ for some discriminant d with $t_i \in \mathbf{Z}$ for $i = 1, 2$, i.e., if $D(f_1)D(f_2)$ is an integral square. In this case, we may set

$$t_i' = t_i / gcd\{t_1, t_2\}, \text{ for } i = 1, 2.$$

Following [BE], we shall say that a pair of forms representing the classes $[f_1], [f_2]$ are *united* if there are respective representatives of the form

$$f_1 = [a_1, t_1' b, {t_1'}^2 a_2 c],$$
$$f_2 = [a_2, t_2' b, {t_2'}^2 a_1 c],$$

where $gcd\{a_1, a_2\} = 1$. We conclude from primitivity that $gcd\{a_i, t_i'\} = 1$ for $i = 1, 2$, and furthermore $gcd\{t_1', t_2'\} = 1$ always holds.

Given united forms in the notation above, we may scale the forms (destroying primitivity) to produce forms of the same discriminant $(t_2' t_1)^2 d = (t_1' t_2)^2 d$, namely,

$$t_2' f_1 = [t_2' a_1, t_1' t_2' b, {t_1'}^2 t_2' a_2 c],$$
$$t_1' f_2 = [t_1' a_2, t_1' t_2' b, t_1' {t_2'}^2 a_1 c],$$

which thus also have the same middle coefficient. Define the *product* of unitable classes to be the class of $[a_1 a_2, b, c]$, i.e.,

$$[f_1][f_2] = [[a_1 a_2, b, c]],$$

which we may think of as arising by applying the formula for the product of Dirichlet united forms here to the non-primitive forms $t_2' f_1, t_1' f_2$ and finally scaling by $(t_1' t_2')^{-1}$. One finds that $[a_1 a_2, b, c]$ is primitive and its discriminant is $(gcd\{t_1, t_2\})^2 d$.

The result from [BE] or [BP] which is relevant to us here (which follows from our Theorem 6 and is not the main result of these papers) is

Theorem [BE] *Given two unitable classes, there exist united representatives, and the class of the product is well-defined. Fixing a fundamental discriminant d and setting $\mathcal{S}(d) = \coprod_{t \geq 1} \mathcal{G}(t^2 d)$, the product above gives $\mathcal{S}(d)$ the structure of an abelian semigroup.*

Though [BP] similarly describes a product in the unitable case (using module-theoretic methods), it is the [BE] notion of united representatives that is most important for us here. By the way, [BP] remarks that Gauss already knew about extensions of the product beyond the case of primitive forms with the same discriminant, and Estes and I have checked that the full semigroup is more or less already described in *Disquisitiones Arithmeticae*. Moreover, upon further review, Estes tells me that Theorem 6 below follows from remarks of Gauss plus remarks in [BE].

To get some sense of these semigroups $\mathcal{S}(d)$ before we continue, it seems worthwhile to pause and recall what is known from order theory: If p is a prime and D is a discriminant, then all of the primitive classes with discriminant $p^2 D$ are represented by the primitive forms on the following list of forms: $[a, pb, p^2 c]$, $[ah^2 + bh + c, p(b + 2ah), p^2 a]$, where $[a, b, c]$ runs over $\mathcal{G}(D)$, and h runs through all integers from 0 to $p - 1$. Of course, it follows from primitivity that then $gcd\{a, n\} = 1$. We shall not take the time to prove this here since we shall really only need the fact that if n^2 divides the discriminant of a form f, for some integer n, then f is equivalent to a form so that n divides the middle coefficient and n^2 divides the last coefficient. (This is easily proved directly by taking a representative of the class whose first coefficient is relatively prime to n as in Lemma 2 below, completing the square, and then translating.)

In fact, there are canonical surjections $\mathcal{G}(n^2 D) \to \mathcal{G}(D)$ defined by simply taking the product with $\mathbf{1}_D$ (and the cardinalities of these kernels are known explaining why it suffices

to compute class numbers only of fundamental discriminants). Thus, the directed system of Gauss groups has a natural inverse limit, which seems to have not been studied. There is also the following amusing and immediate consequence: Given discriminants D and n^2D and any prime q, the classes of order q^k for some $k \geq 0$ in $\mathcal{G}(n^2D)$ map under the canonical surjection to classes in $\mathcal{G}(D)$ of order q^i for some $i \leq k$.

In order to give an example of the Butts-Estes-Pall product and to better explain the canonical surjections, we observe that if $[f]$ has discriminant n^2D, then by the discussion above we may find a representative of the form $[a, nb, n^2c]$ with $gcd\{a, n\} = 1$. To take the product with $\mathbf{1}_D$, we may translate (in each case of b even or odd) our standard representative $n\mathbf{1}_D$ to arrange that the middle coefficient agrees with nb and the first coefficient remains n. The product $[f]\ \mathbf{1}_D$ is thus represented by $[a, b, *]$, where we solve for $*$ so that the discriminant is D, i.e., we have $[[a, nb, n^2c]]\ \mathbf{1}_D = [[a, b, c]]$.

Here finally is the concordance extension of the Butts-Estes definition. We shall say that two unitable forms are *concordant* if they admit united representatives, but where we remove the condition that $gcd\{a_1, a_2\} = 1$. Just as in the previous case, we use the same formula

$$[a_1, t_1'b, {t_1'}^2 a_2 c]\ [a_2, t_2'b, {t_2'}^2 a_1 c] \quad = \quad [a_1a_2, b, c]$$

to define a product of concordant forms. The main result of this section is simply that this product is well-defined on the level of classes (and this subsumes all the various well-definedness-of-product results mentioned before). This extension may seem stupid until one realizes that

Lemma 1 *Suppose that $[a_i, b_i, c_i]$ for $i = 1, 2$ are primitive forms of respective discriminants D_1, D_2, where $b_1b_2 \geq 0$. Then the two forms are concordant if and only if the following two conditions hold*

- $D_1b_2^2 = D_2b_1^2$, and thus $D_i = t_i^2 d$ and $b_i = bt_i'$ for $i = 1, 2$,
- $a_1a_2 \Big| \dfrac{b^2 - d(gcd\{t_1, t_2\})^2}{4}$.

Notice that the conditions $b_1b_2 \geq 0$, $D_1b_2^2 = D_2b_1^2$ of the lemma are equivalent to the condition $b_1\sqrt{|D_2|} = b_2\sqrt{|D_1|}$. The point of Lemma 1 is that the discriminant divided by the square of the middle coefficient is an "invariant" (whose geometric significance we discover in the next section) which puts concordance into proper perspective and simplifies subsequent calculations.

Proof It is immediate that the stated conditions follow from the definition of concordance. For the converse, suppose first just that $f_1 = [a_1, b_1, c_1]$, $f_2 = [a_2, b_2, c_2]$ satisfy $D_1b_2^2 = D_2b_1^2$. It follows that D_1 and D_2 have the same square-free kernel, so we may write $D_1 = dt_1^2, D_2 = dt_2^2$ (possibly in several different ways) and set $t_i' = t_i/gcd\{t_1, t_2\}$ as before. Thus, we find that $t_2'f_1$ and $t_1'f_2$ have the same discriminant, so $D_1/b_1^2 = D_2/b_2^2$ gives $t_2'b_1 = t_1'b_2$ (and it is here that we use the hypothesis of the lemma that $b_1b_2 \geq 0$).

Since $gcd\{t_1', t_2'\} = 1$, we conclude that $b_1 = bt_1'$ and $b_2 = bt_2'$ and have proved

$$t_2' f_1 = [t_2' a_1, t_1' t_2' b, t_2' c_1],$$
$$t_1' f_2 = [t_1' a_2, t_1' t_2' b, t_1' c_2].$$

As to the integrality condition, just notice that (for $a_1 a_2 \neq 0$)

$$\frac{b^2 - d(gcd\{t_1, t_1\})^2}{4a_1 a_2} = \frac{1}{{t_i'}^2} \quad \frac{b_i^2 - dt_i^2}{4a_1 a_2}, \text{ for } i = 1, 2,$$

and so (even if $a_1 a_2 = 0$), ${a_2 t_1'}^2 | c_1$ and ${a_1 t_2'}^2 | c_2$. It follows directly from this that f_1, f_2 are as stated.

q.e.d.

In order to prove that the product of concordant forms is well-defined, we continue by recalling several standard lemmas (from [Ca]), where we must here observe that the standard hypothesis of primitivity is a red herring. For this reason, to abide by our stated goal of remaining self-contained, and because of their geometric significance, we shall also briefly recall the proofs. Here are the three lemmas:

Lemma 2 *Given a primitive form f and any integer M, there is a primitive vector (x, y) with $f(x, y)$ relatively prime to M.*

Lemma 3 *If $f_i = [a_i, b_i, c_i]$, for $i = 1, 2$ are (not necessarily primitive) forms with $gcd\{a_1, a_2\} = 1$ and $b_1 \equiv b_2 (mod\ 2)$, then there are translations $\gamma_1, \gamma_2 \in PSL(2, \mathbf{Z})$ so that $\gamma_1.f_1$ and $\gamma_2.f_2$ have the same middle coefficient.*

Lemma 4 *Suppose that $f_i = [a_i, b, c_i]$, for $i = 1, 2$ are (not necessarily primitive) forms and that there is some $\ell \in \mathbf{Z}$ so that*

$$\ell | c_1, \quad \ell | c_2, \quad and \quad gcd\{a_1, a_2, \ell\} = 1.$$

Then

$$[a_1, b, c_1] \sim [a_2, b, c_2] \quad \Rightarrow \quad [\ell a_1, b, \ell^{-1} c_1] \sim [\ell a_2, b, \ell^{-1} c_2].$$

As to the (absolutely standard) proof of Lemma 2, consider the primes p that divide M. Define x, y by taking $p|y$, $p \nmid x$ if $p \nmid a$ (and similarly taking $p|x$, $p \nmid y$ if $p \nmid c$), while if $p|a$, $p|c$, then $p \nmid b$ by primitivity and we take $p \nmid x$, $p \nmid y$. Since $gcd\{x, y\} = 1$ by construction, there is some $\gamma \in PSL(2, \mathbf{Z})$ with first column $(x, y)^t$, whence the first coefficient of $\gamma.[a, b, c]$ is relatively prime to M.

For Lemma 3, since $gcd\{a_1, a_2\} = 1$, there are integers ℓ_1, ℓ_2 with $a_1 \ell_1 - a_2 \ell_2 = 1$. Since $b_1 \equiv b_2 (mod\ 2)$, we may translate f_i by $\ell_i (b_2 - b_1)/2$ for $i = 1, 2$ to arrange that the forms have a common middle coefficient.

Lemma 4 requires a small calculation. Since we assume $f_1 \sim f_2$ (whether or not they are primitive), there is some $\gamma = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in PSL(2, \mathbf{Z})$ with $\gamma.f_1 = f_2$. Equating

coefficients (and using that f_1, f_2 have the same middle coefficient), we may eliminate r, u (this is the calculation) to get

$$a_1 s + c_2 t = 0,$$
$$a_2 s + c_1 t = 0.$$

Since $\ell | c_i$ and $gcd\{a_1, a_2, \ell\} = 1$, we conclude that $\ell | s$, and the matrix $\begin{pmatrix} r & \ell^{-1} s \\ \ell t & u \end{pmatrix}$ does the trick. Notice that the form resulting from a primitive form via Lemma 4 is not necessarily primitive (for instance, if $\ell | b$ and $\ell^2 | c$).

Proposition 5 *Unitable classes admit united representatives.*

Proof Given primitive classes $[f_1], [f_2]$ of respective discriminants $D_1 = t_1^2 d, D_2 = t_2^2 d$, we may choose (as before) representatives $f_1 = [a_1, b_1, c_1], f_2 = [a_2, b_2, c_2]$ with $gcd\{t_i, a_i\} = 1$ and $t_i' | b_i$. Let us then apply Lemma 2 to arrange that $\gcd\{a_1, a_2\} = 1$, so that then $t_1' a_2$ is relatively prime to $t_2' a_1$. Since $t_2' f_1$ has the same discriminant as $t_1' f_2$ (and discriminants and middle coefficients always have the same parity *mod* 2), we conclude that $t_2' b_1 \equiv t_1' b_2 (mod\ 2)$.

By Lemma 3, we may translate to arrange that $t_2' f_1$ and $t_1' f_2$ have the same middle coefficient. Since they also have the same discriminant, we find

$$\frac{D_1}{b_1^2} = \frac{t_2'^2 D_1}{(t_2' b_1)^2} = \frac{t_1'^2 D_2}{(t_1' b_2)^2} = \frac{D_2}{b_2^2},$$

so the first condition of Lemma 1 holds, and $b_1 b_2 \geq 0$ is automatic. One checks the integrality condition as usual using that $gcd\{a_1, a_2\} = 1$, so the forms are concordant by Lemma 1 and in fact united since $gcd\{a_1, a_2\} = 1$. *q.e.d.*

Theorem 6 *The class of a product of concordant forms is well-defined giving $\mathcal{S}(d)$ the structure of an abelian semigroup for d fundamental and $\mathcal{G}(D)$ the structure of a finite abelian group for any D.*

Proof Following [Ca], suppose that we have two concordant pairs

$$f_1' = [a_1', t_1' b', t_1'^2 a_2' c'],$$
$$f_2' = [a_2', t_2' b', t_1'^2 a_1' c']$$

and similarly f_1'', f_2'' of primitive forms representing a pair of unitable classes. Applying Lemma 2 twice (the first time to f_1' with $M = t_1' t_2' a_1' a_2' a_1'' a_2''$ and the second time to f_2' with $M = t_1' t_2' a_1' a_2' a_1'' a_2'' a_1$), we may find united representatives, say

$$f_1 = [a_1, t_1' b, t_1'^2 a_2 c],$$
$$f_2 = [a_2, t_2' b, t_2'^2 a_1 c],$$

respectively, where

$$gcd\{a_1, a_2\} = 1 = gcd\{a_1 a_2, t_1' t_2' a_1' a_2' a_1'' a_2''\}.$$

We shall show that

$$[a_1a_2, b, c] \sim [a_1'a_2', b', c'],$$

and the result then follows by symmetry (of primed and double-primed variables).

To this end, by the relative primality of a_1a_2 and $t_1't_2'a_1'a_2'$, we may apply Lemma 3 as in the proof of Proposition 5 to conclude that there are integers B, C, C' with

$$\begin{aligned} [a_1a_2, b, c] \sim [a_1a_2, B, C] = \bar{f}, \quad & [a_1'a_2', b', c'] \sim [a_1'a_2', B, C'] = \bar{f}', \\ f_1 \sim [a_1, t_1'B, t_1'^2a_2C] = \bar{f}_1, \quad & f_1' \sim [a_1', t_1'B, t_1'^2a_2'C'] = \bar{f}_1', \\ f_2 \sim [a_2, t_2'B, t_2'^2a_1C] = \bar{f}_2, \quad & f_2' \sim [a_2', t_2'B, t_2'^2a_1'C'] = \bar{f}_2'. \end{aligned}$$

Furthermore, since f_1 and f_1' have the same discriminant, we find $a_1a_1C = a_1'a_2'C'$, and by the relative primality of a_1a_2 and $a_1'a_2'$, there is some integer K with $C = a_1'a_2'K$ and $C' = a_1a_2K$. Thus, in fact

$$\begin{aligned} \bar{f} &= [a_1a_2, B, a_1'a_2'K] & \bar{f}' &= [a_1'a_2', B, a_1a_2K] \\ \bar{f}_1 &= [a_1, t_1'B, t_1'^2a_2a_1'a_2'K], & \bar{f}_1' &= [a_1', t_1'B, t_1'^2a_2'a_1a_2K], \\ \bar{f}_2 &= [a_2, t_2'B, t_2'^2a_1a_1'a_2'K], & \bar{f}_2' &= [a_2', t_2'B, t_2'^2a_1'a_1a_2K], \end{aligned}$$

and it remains only to show that $\bar{f} \sim \bar{f}'$.

Since $\bar{f}_1 \sim f_1 \sim f_1' \sim \bar{f}_1'$, we may apply Lemma 4 with $\ell = t_1'a_2'$ to $\bar{f}_1 \sim \bar{f}_1'$ to conclude that

$$[t_1'a_1a_2', t_1'B, t_1'a_2a_1'K] \sim [t_1'a_1'a_2', t_1'B, t_1'a_1a_2K],$$

and so

$$[a_1a_2', B, a_1'a_2K] \sim [a_1'a_2', B, a_1a_2K] = \bar{f}'.$$

Applying Lemma 4 in the same way to $\bar{f}_2 \sim \bar{f}_2'$ with $\ell = t_2'a_1$ gives

$$\bar{f} = [a_1a_2, B, a_1'a_2'K] \sim [a_1a_2', B, a_1'a_2K],$$

so indeed $\bar{f} = \bar{f}'$, completing the proof of well-definedness.

Associativity follows as above using Lemmas 2 and 3, units and inverses have already been discussed, and commutativity is obvious. *q.e.d.*

The Geometry of Fundamental Roots

Let $\omega(f) = \frac{-b+D^{1/2}}{2a}$ be the fundamental root of the primitive form $f = [a, b, c]$ of discriminant D. We assume in this section that f is definite, so $D < 0$. Thus,

$$\omega(f) \;=\; \frac{-b}{2a} + \sqrt{-1}\,\sqrt{\frac{-D}{4a^2}} \;=\; \frac{p}{q} + \sqrt{-1}\,\sqrt{\frac{r}{s}} \;\in\; \mathbf{Q} + \sqrt{-1}\,\sqrt{\mathbf{Q}},$$

where $p, q, r, s \in \mathbf{Z}$, and we may take $gcd\{p, q\} = 1 = gcd\{r, s\}$, and $r, s > 0$. A point in upper half space $\mathcal{U}$ with rational real and square imaginary parts will be called simply a *CM point* of $\mathcal{U}$, since (as Zagier points out) these correspond to the elliptic curves that admit a complex multiplication. In fact, one can easily check directly that each element of $PSL(2, \mathbf{Z})$ leaves invariant this CM locus, and it therefore descends to the collection of *CM points* on the modular curve $\mathcal{M} = \mathcal{U}/PSL(2, \mathbf{Z})$ itself.

A direct calculation using high-school algebra (which we omit and which is surely standard) proves

Proposition 7 *Given a CM point* $\omega = \frac{p}{q} + \sqrt{-1}\,\sqrt{\frac{r}{s}} \in \mathcal{U}$, *the positive definite primitive form* $f(\omega)$ *proportional to*

$$f'(\omega) \quad = \quad [q^2 s, -2pqs, p^2 s + q^2 r]$$

inverts the formula above for fundamental roots.

In our first proof of this (unaware of the direct calculation above starting from the fundamental root), we worked in Minkowski space, and it is worth briefly describing this for the insights it affords. Identify $\mathbf{R}^3$ with the space of all symmetric bilinear pairings as usual by $(u, v, w) \mapsto \begin{pmatrix} w - u & v \\ v & w + u \end{pmatrix}$. Pass to positive real projective classes of rational forms and identify a ray from the origin with its point of intersection with the unit hyperboloid; one sees the two disk components of projective definite forms (corresponding to first and last coefficients both positive or both negative) and the annulus of projective indefinite forms. Direct calculation shows that the function in Proposition 7 is given by radial projection of the upper sheet from (0,0,-1) followed by the usual complex fractional linear transformation mapping the Poincaré disk (the unit disk at height zero in $\mathbf{R}^3$) to $\mathcal{U}$. This establishes an isomorphism between the set of projective classes of positive definite rational quadratic forms and the set of CM points in $\mathcal{U}$.

As to the primitive form $f = f(\omega)$ in the projective class of $f' = f'(\omega)$, observe that if π is a prime dividing the coefficients of f', then $\pi | q$ or $\pi | s$ (since π divides the first coefficient), and in either case $\pi | q$ if and only if $\pi | s$ (since π divides the last coefficient and using the assumed relative primality of p, q and r, s). Thus, if π divides the coefficients of f', then $\pi | gcd\{q, s\}$.

In particular, if $gcd\{q, s\} = 1$, then $f' = f$ is itself primitive, and in this case the discriminant is $D = -4q^4 rs$. One sees that our map above is wildly discontinuous. One can go a bit further and show that $gcd\{q, s\}\ gcd\{q, s/gcd\{q, s\}\}$ actually divides the coefficients of f', but the explicit calculation of the primitive form f in the projective class of f' seems to be out of reach. We regard this overall scale as essentially non-geometric data and must stick to homogeneous rational functions of degree zero in the coefficients.

In light of the previous discussion about unitable pairs of forms, we are led to consider the level sets of $D/a^2, D/b^2, D/c^2$, for if two primitive forms lie on a common level set of one of these functions, then they are necessarily unitable. Setting $\omega = \frac{p}{q} + \sqrt{-1}\,\sqrt{\frac{r}{s}} =$

$u + \sqrt{-1}\, v \in \mathcal{U}$, one finds that

$$D/a^2 = -\alpha^2 \quad \Leftrightarrow \quad \frac{r}{s} = \frac{\alpha^2}{4} \quad \Leftrightarrow \quad v = \frac{\alpha}{2},$$

$$D/b^2 = -\beta^2 \quad \Leftrightarrow \quad \frac{r}{s} = \beta^2 \frac{p^2}{q^2} \quad \Leftrightarrow \quad v = \pm\beta u,$$

$$D/c^2 = -\gamma^2 \quad \Leftrightarrow \quad 2\gamma^{-1} = \frac{p^2}{q^2}\sqrt{\frac{s}{r}} + \sqrt{\frac{r}{s}} \quad \Leftrightarrow \quad \gamma^{-2} = u^2 + (v - \gamma^{-1})^2,$$

where $\alpha, \beta, \gamma > 0$. These respective loci are thus horizontal lines, rays from the origin, and circles tangent to $\mathbf{R}$ at zero as in Figure 1.

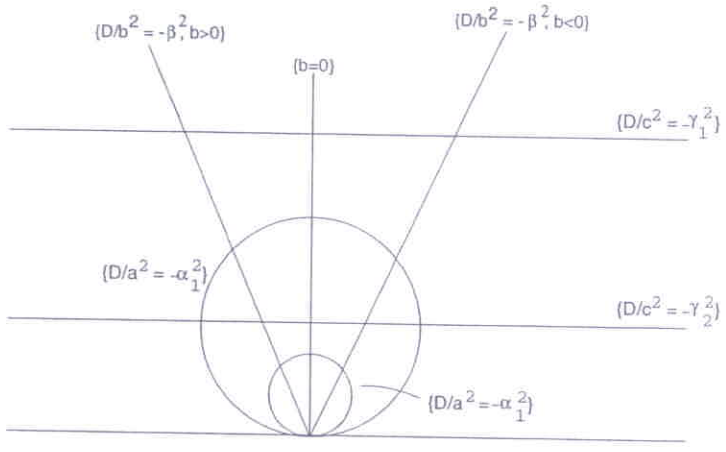

Figure 1

To understand the corresponding loci in $\mathcal{U}$ and $\mathcal{M}$, we first recall some standard hyperbolic constructions of Poincaré. The limit of a pencil of hyperbolic circles in $\mathcal{U}$ passing through a common point as the radius and center approach infinity is called a *horocycle*. In other words, a horocycle in $\mathcal{U}$ is either a horizontal line ("centered" at infinity) or a Euclidean circle in $\mathcal{U}$ tangent to $\mathbf{R}$ ("centered" at the corresponding point of tangency). An *ϵ-hypercycle* to a geodesic g in $\mathcal{U}$ (i.e., g is a vertical half-line or a Euclidean semi-circle perpendicular to $\mathbf{R}$) is a component of the locus of points at distance $\epsilon \geq 0$ from g. In particular, there are two ϵ-hypercycles for each $\epsilon > 0$, and g is itself the 0-hypercycle to g. (Put another way, these hypercycles are the loci of constant curvature which we think of as interpolating between geodesics and horocycles; as such, Bonahon has asked the reasonable question of whether the hypercyclic flow is also ergodic.) In particular, the hypercycles to the imaginary axis in $\mathcal{U}$ are simply the Euclidean rays from the origin. See Figure 1. Projections to $\mathcal{M}$ of horocycles centered at infinity or hypercycles to the imaginary axis in $\mathcal{U}$ will be called simply *horocycles* or *hypercycles* in $\mathcal{M}$. See Figure 2.

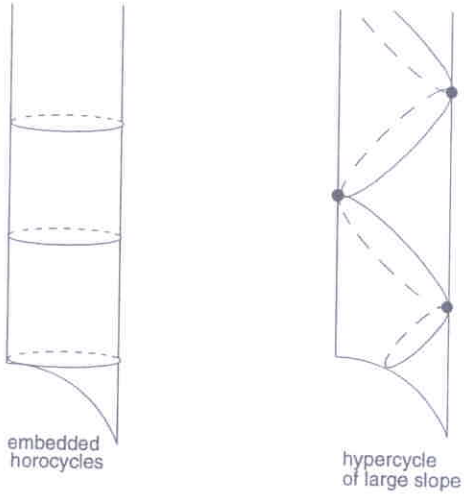

Figure 2

The horizontal foliation in $\mathcal{U}$ corresponds to the horocycles centered at infinity, and the leaves of this foliation descend to simple closed curves in $\mathcal{M}$ provided they lie above height one (i.e., provided they have hyperbolic length less than unity). We shall refer to this canonically foliated once-punctured open disk as the "cusp" of $\mathcal{M}$ and to its complement as the "body" of $\mathcal{M}$. Let us take the Euclidean heights of these embedded horocycles (or equivalently their hyperbolic lengths) as a parameter, so the "η-horocycle" in $\mathcal{M}$ is the horocycle in the cusp corresponding to the leaf at height $\eta \geq 1$ in $\mathcal{U}$.

Insofar as each hypercyclic ray is asymptotic to the puncture, there is a first and last intersection of any hypercycle with any horocycle that it meets, so each hypercycle has naturally two ends. Furthermore, an end of a hypercyclic ray is standard in the cusp, namely, it is modeled (up to rotation) on a line of constant slope in $\mathcal{U}$. Thus, an end of a hypercycle in $\mathcal{M}$ is described by one real parameter (a "slope") plus a choice of sign (so a "signed slope"), or equivalently by a signed distance (namely, along a fixed embedded horocycle to this last intersection).

In fact, the parameter ϵ above for hypercycles is useful only in that discussion, and we shall actually use a different parameter for hypercycles, namely, the σ-hypercycle to the imaginary axis in $\mathcal{U}$ is the one of Euclidean slope σ.

This is the first of several appearances of this Euclidean structure on the modular curve, and there is truly something to observe in contrast to the usual case. It is precisely because we are taking hypercycles to a bi-infinite geodesic running from puncture to puncture that this Euclidean structure is defined. (In contrast, the attempt to define

an analogous structure relative to a closed geodesic is foiled by the remaining degree of freedom given by translation along the geodesic.)

Notice that the σ-hypercycle flips to the $(-\sigma)$-hypercycle, so we shall also speak of the $|\sigma|$- hypercycle in $\mathcal{M}$. Usually when we uniformize a hypercycle on $\mathcal{M}$, we shall take the representative in $\mathcal{U}$ with positive slope.

It seems worth pausing to describe two figures. An easy calculation using Proposition 7 and our formulas for ambiguous forms shows that the fundamental roots of the ambiguous forms consist exactly of the CM points either lying on the geodesic in $\mathcal{M}$ corresponding to the imaginary axis in $\mathcal{U}$ or lying on the projection to $\mathcal{M}$ of the frontier of the usual fundamental domain for $PSL(2,\mathbf{Z})$. Furthermore, an easy verification using Proposition 7 and our formulas for units of Gauss groups, shows that

$$\omega(\mathbf{1}_D) = \begin{cases} \sqrt{\frac{\pm D}{4}}; & \text{if } D \equiv 0 (mod\ 4), \\ \frac{1}{2} + \sqrt{\frac{\pm D}{4}}; & \text{if } D \equiv 1 (mod\ 4). \end{cases}$$

Thus, all units except $\mathbf{1}_{-3}$ and $\mathbf{1}_{-4}$ have fundamental roots lying in the cusp, and their lifts to the usual fundamental domain for $PSL(2\mathbf{Z})$ in $\mathcal{U}$ alternate between real part zero and real part $\pm\frac{1}{2}$ as illustrated in Figure 3.

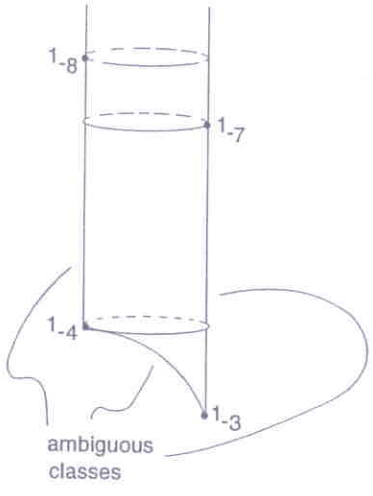

Figure 3

Our final figure to consider is standard and rather involved. First, take the Farey tesselation τ_* (i.e., the tesselation of $\mathcal{U}$ generated by reflecting the triangle spanned by $0, 1, \infty$ about its sides, and so on), so the full symmetry group of τ_* is exactly $PSL(2,\mathbf{Z})$. Furthermore, $PSL(2,\mathbf{Z})$ acts transitively on the set of oriented edges of τ_*, so any hypercycle to any geodesic in τ_* admits a lift to $\mathcal{U}$ as a ray from the origin of positive slope. As usual, the ideal points of τ_* are given their Farey enumeration by the rationals, and the

$PSL(2,\mathbf{Z})$-orbit of the horocycle at height one is a circle-packing, where at each rational number of the form p/q we take the Euclidean circle of radius $1/q^2$. See Figure 4 and imagine how hypercycles to the edges of τ_* may intersect one another and how they may intersect the horocycles in this circle-packing; these are the sorts of configurations in $\mathcal{M}$ we shall consider below.

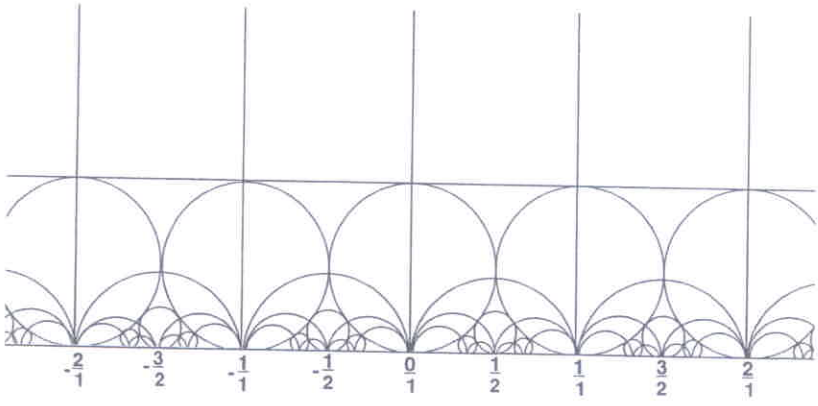

Figure 4

Armed with this discussion of hypercycles and horocycles in $\mathcal{M}$, we return to our calculations before of level sets of D/a^2, D/b^2, and D/c^2, which we find correspond respectively to horocycles centered at infinity, hypercycles, and horocycles centered at zero. It is a nice picture. We shall concentrate on the hypercyclic case and relegate the discussion of the (interesting!) horocyclic case to the next section.

It is time to reap some results from the discussion, and it is cleanest to formulate them on the level of $\mathcal{M}$. We shall simply identify not only a form with its corresponding CM point in $\mathcal{U}$ but also its class with the corresponding CM point of $\mathcal{M}$. Thus, if f is a definite form, then we think of $f = \omega(f) \in \mathcal{U}$ and $[f] = [\omega(f)] \in \mathcal{M}$. Furthermore, if h is a hypercycle in $\mathcal{U}$ to a geodesic in τ_*, then we shall let $[h]$ denote the corresponding hypercycle in $\mathcal{M}$.

Corollary 8 *$[f_1], [f_2] \in \mathcal{M}$ are unitable if and only if they lie on a common hypercycle. Furthermore, if $[f_1], [f_2]$ lie on a common horocycle, then they are unitable as well.*

The first part follows from the discussion above and Proposition 5, but notice that two CM points on a given hypercycle might not represent concordant forms, that is, we may have to choose another hypercycle to see them as concordant. Indeed, given a concordant pair f_1, f_2, there is a *corresponding* hypercycle h to the imaginary axis with $f_1, f_2 \in h$. The second part also follows from the previous discussion, and there is the amusing (and

it seems to me highly non-trivial) geometric consequence that two CM points of $\mathcal{M}$ which lie on a common horocycle also lie on a common hypercycle (and we wonder about the converse).

Corollary 9 *Suppose that f_1, f_2 are concordant, say with corresponding hypercycle h. Then $[f_1][f_2]$ also lies on the hypercycle $[h]$ in $\mathcal{M}$.*

There is truly nothing to prove in light of Theorem 6. This is a remarkable "focusing" property of hypercycles on the modular curve and explains the basic connection between geometry and the Gauss product. We shall further develop this theme and give a geometric characterization of the product in Theorem 13 below. It is worth pausing, though, and and inviting you to imagine products, relations, squares, and units in the Gauss groups from this geometric point of view.

In preparation for our next result (which is a basic compactness property of $\mathcal{M}$), we develop some generalities about the distribution of $\mathcal{G}(D)$ in $\mathcal{M}$. To this end, let us fix some discriminant D and consider the various $b \geq 0$ of the same parity as D modulo two. The points of $\mathcal{U}$ which project to $\mathcal{G}(D)$ and lie on the hypercycle of slope $\sqrt{-D/b^2}$ are in natural one-to-one correspondence with the various divisors of $(b^2 - D)/4$. In particular, the first and last such points evidently represent the unit $\mathbf{1}_D$, and we have

Proposition 10 *For any discriminant D, each point of $\mathcal{G}(D)$ lies either in the body of $\mathcal{M}$ or in the cusp of $\mathcal{M}$ below $\mathbf{1}_D$.*

It might be interesting to combine Proposition 10 with estimates (which I do not know) for the "discriminant D injectivity radius" to estimate class numbers. On the other hand, for any $[f] \in \mathcal{G}(D)$, there is some b with $0 \leq b \leq \sqrt{-D/3}$ and a representative f lying on the hypercycle of slope $\sqrt{-D/b^2}$, so one observes a seeming non-uniformity in the distribution of $\mathcal{G}(D)$ in $\mathcal{M}$ which suggests that this estimate on class numbers might not be too handy. (Zagier points out that this and "all" other such elementary hyperbolic estimates on class numbers have been tried; furthermore, much more is known about the distribution of heights of $\mathcal{G}(D)$ in $\mathcal{M}$, for instance, all other points of $\mathcal{G}(D)$ are at most half as high as $\mathbf{1}_D$, and there are at most a points whose height is $1/a$ times the height of $\mathbf{1}_D$.)

Arguing as in the proof of Theorem 6, we find that given $[f], [g] \in \mathcal{G}(D)$, there are concordant representatives lying on a common hypercycle. Furthermore, for each fixed D, we can similarly find a $b = b(D)$ so that each point of $\mathcal{G}(D)$ is represented by a point on the "saturated" hypercycle of slope $\sqrt{-D/b^2}$. Finding such a b (which I do not know how to do effectively) could also give an estimate on class numbers. At the same time, it seems like a nice combinatorics problem to try and express class numbers in terms of multiplicities on a saturated hypercycle.

Turning now to the distribution of squares of $\mathcal{G}(D)$ in $\mathcal{M}$, suppose that $[f] = [g]^2 \in \mathcal{G}(D)$, so there is a concordant pair g_1, g_2 of forms representing $[g]$ lying on the corresponding hypercycle h. In other words, $[g_1] = [g_2]$ is a double point (i.e., self-intersection) of $[h]$

in $\mathcal{M}$. A double point of the hypercycle h is said to be *concordant* if it arises as above from a concordant pair of forms.

There is a particularly simple class of double points as follows: If h is a hypercycle to the imaginary axis in $\mathcal{U}$, then a *cuspidal* double point is a point of intersection of h with some integral translate of the flip of h. Thus, the imaginary axis itself (as a hypercycle) has no cuspidal double points, and every other hypercycle has infinitely many. Notice that any cuspidal double point lying on a hypercycle of positive slope at least two necessarily lies in the cusp, and for any slope, all but a finite number of its cuspidal double points necessarily lie in the cusp, hence the terminology. (It is easy to construct non-cuspidal double points starting from Figure 4.)

Notice that hypercycles are generally dramatically far from general position. For instance, arguing as Theorem 6, we can find hypercycles with multiple points of arbitrarily high orders at any specified set of elements of the Gauss group. (The extent to which the orders of multiplicity can also be specified is an interesting question.)

There is a pleasant geometric description of concordance in the case of common discriminants (and at one point I thought just homogeneity of the formula was a small miracle), as follows.

Proposition 11 *Suppose that two forms $f_1, f_2 \in \mathcal{U}$ of the same discriminant lie on a common hypercycle. Then f_1, f_2 are concordant if and only if $|f_1||f_2| \in \mathbf{Z}$, where $|f|$ denotes the Euclidean norm of $f \in \mathcal{U}$ as a vector in $\mathbf{R}^2$.*

Proof Let us write $f_i = [a_i, b, c_i]$ with fundamental root $u_i + \sqrt{-1}\, v_i \in \mathcal{U}$ and discriminant D, for $i = 1, 2$. It follows from Proposition 7 that $b/a_i = -2p_i/q_i = -2u_i$ and $-D/a_i^2 = 4r_i/s_i = 4v_i^2$. Concordance is equivalent to integrality of

$$\begin{aligned}
\frac{b^2 - D}{4a_1 a_2} &= \frac{b}{2a_1}\frac{b}{2a_2} + \sqrt{\frac{D}{4a_1^2}\frac{D}{4a_2^2}} \\
&= u_1 u_2 + v_1 v_2 \\
&= \sqrt{u_1^2 + v_1^2}\ \sqrt{u_2^2 + v_2^2} \\
&= |f_1||f_2|,
\end{aligned}$$

where the next-to-last equality follows from the extreme case of the Cauchy-Schwarz (in)equality using that f_1 and f_2 are parallel as vectors in $\mathbf{R}^2$ since they lie on a common hypercycle. *q.e.d.*

In fact, given two forms f_1, f_2 lying on a common hypercycle (not necessarily of the same discriminant), integrality of $|f_1||f_2|$ is again a necessary condition for the forms to be concordant; a further sufficient condition for concordance is that $t_1' t_2'$ divide $|f_1||f_2|$ as one can easily check. In a sense this is a perfectly suitable answer (for given two points on a common hypercycle, we can apply the Euclidean algorithm to find their primitive representatives, hence $t_1' t_2'$, and hence verify concordance); on the other hand, we would

hope for a more intrinsic geometric characterization of concordance for two points on a common hypercycle.

Proposition 12 *Every cuspidal double point is concordant.*

Proof Let h be a hypercycle, say of slope $\beta > 0$. A general cuspidal double point on h is then given by $f_1 = n/2 + \sqrt{-1}\ \beta n/2$ for $n \in \mathbf{Z}$. This point on h translates to $-n/2 + \sqrt{-1}\ \beta n/2$ on the hypercycle of slope $-\beta$, which in turn flips to

$$f_2 = \frac{2}{n(1+\beta^2)}\ (1 + \sqrt{-1}\ \beta)$$

on h, and we find

$$|f_1||f_2| \ = \ \frac{n}{2}\frac{2}{n(1+\beta^2)} \ + \ \frac{\beta n}{2}\frac{2\beta}{n(1+\beta^2)} \ = \ 1 \ \in \ \mathbf{Z}.$$

The result then follows from Proposition 11. *q.e.d.*

Here is the promised geometric interpretation of the product (and Sullivan points out that never mind anything else, this is a theorem about the Euclidean plane).

Theorem 13 *Suppose that f_1, f_2 are concordant with corresponding hypercycle h. Then $[f_1][f_2]$ is represented by the point $f \in h$ closest to the origin with the property that whenever f_1, f_2 translate to concordant forms on the corresponding hypercycle h', then f also translates to h'.*

Proof First we show that the usual product on h has the property stated above for f, and then we show that this is actually the closest such point f to the origin.

Suppose that

$$f_1 = [a_1, t_1'b, a_2t_1'^2c], \quad \text{and } f_2 = [a_2, t_2'b, a_1t_2'^2c]$$

are primitive concordant forms which translate respectively to (primitive) concordant forms

$$f_1' = [a_1, t_1'B, a_2t_1'^2C], \quad \text{and } f_2' = [a_2, t_2'B, a_1t_2'^2C],$$

so that we have

$$B_1 = t_1'B = t_1'b + 2\ell_1a_1, \quad \text{and } B_2' = t_2'B = t_2'b + 2\ell_2a_2.$$

Solving $t_2'B_1 = t_1'B_2$, we find

$$(*) \qquad t_2'\ell_1a_1 = t_1'\ell_2a_2.$$

Of course $1 = gcd\{t_1', t_2'\}$, and also $1 = gcd\{t_1', a_1\} = gcd\{t_2', a_2\}$ by primitivity. One concludes from $(*)$ that therefore

$$(**) \qquad t_1'a_2 | \ell_1a_1;$$

furthermore, $t_1' a_2$ divides the last coefficient of f_1 by inspection, so it also divides ℓ_1 times this last coefficient. Thus, we may show that

$$(\dagger) \qquad t_1' a_2 | \ell_1 (t_1' b)$$

and use that f_1 is primitive to conclude that therefore $t_1' a_2 | \ell_1$, as required.

To establish that the product $[a_1 a_2, b, c]$ has the stated property, it therefore remains only to prove the integrality condition $(\dagger)$, i.e., we must prove $a_2 | b\ell_1$. To this end, recall that f_1, f_2 and f_1', f_2' are each supposed to be concordant pairs, so

$$a_1 a_2 | \frac{b^2 - d(gcd\{t_1 t_2\})^2}{4}, \quad \text{and } a_1 a_2 | \frac{B^2 - d(gcd\{t_1 t_2\})^2}{4},$$

using the notation of Lemma 1. Taking the difference, we find that $a_1 a_2$ divides

$$\begin{aligned} \frac{B^2 - b^2}{4} &= \frac{1}{t_i'^2} \frac{(t_i' B)^2 - (t_i' b)^2}{4} \\ &= \frac{1}{t_i'^2} \frac{(t_i' b + 2\ell_i a_i)^2 - (t_i' b)^2}{4} \\ &= \frac{1}{t_i'^2} \ell_i a_i (t_i' b + \ell_i a_i), \end{aligned}$$

for each $i = 1, 2$. Taking $i = 1$, multiplying by $t_1'^2$, and dividing through by a_1, we conclude that

$$t_1'^2 a_2 | \ell_1 (t_1' b + \ell_1 a_1).$$

On the other hand, we have that $t_1' a_2 | \ell_1 a_1$ from $(**)$, so $t_1' a_2$ divides the second term. Hence, $t_1' a_2$ also divides the first term, completing the proof of the first part (as was explained before).

We must still prove that the product is closest to the origin. Specifically, we show that if $a \in \mathbf{Z}$ satisfies

$$\left.\begin{array}{l} Bt_1' \equiv bt_1' (mod\ 2a_1) \\ Bt_2' \equiv bt_2' (mod\ 2a_2) \end{array}\right\} \quad \Rightarrow \quad B \equiv b (mod\ 2a),$$

then $a | a_1 a_2$, and the desired result then follows from the formula for the real part of the fundamental root. To this end, first notice that $Bt_i' \equiv bt_i' (mod\ 2a_i) \Leftrightarrow B \equiv b (mod\ 2a_i)$ since $gcd\{t_i, a_i\} = 1$, for $i = 1, 2$. Thus, we may take $t_1' = t_2' = 1$ in the previous inset equation.

Now, given $a \in \mathbf{Z}$, let us take $B = b + 2\ell a_1 a_2$, for $\ell \in \mathbf{Z}$, where $gcd\{\ell, a\} = 1$. Reduce modulo $2a$ to find that

$$a | \ell a_1 a_2,$$

and use $gcd\{\ell, a\} = 1$ to conclude that indeed $a | a_1 a_2$. *q.e.d.*

To the extent that our characterization of concordance is not completely geometric, as discussed above, so too is the characterization of the product in Theorem 13 deficient. (It may be that a converse of Theorem 13 holds and gives the desired entirely geometric description of concordance as well as the product.)

To close, we offer brief remarks on several disparate topics. Throughout this discussion, we shall refer to a torsion-free finite-index subgroup Γ of $PSL(2,\mathbf{Z})$ as an *arithmetic group*.

Clever Euclidean Geometry

Consider homothety $H_\lambda(z) = \lambda z$ on $z \in \mathcal{U}$, so for each $\lambda > 0$, H_λ setwise preserves each hypercycle to the imaginary axis in $\mathcal{U}$. We wish to analyze two cases of forms $[a,b,c]$, namely, in the first case: $\lambda|b, \lambda^2|c$ (and hence also $gcd\{a,\lambda\} = 1$); and in the second case: $\lambda|c$ and $gcd\{\lambda, a\} = 1$.

In the first case, the map $H_{\lambda^{-1}}$ is simply the well-defined homomorphism multiplication by $\mathbf{1}_{d/\lambda^2}$ as we calculated before. In the second case, at least $H_{\lambda^{-1}}$ is well-defined as follows from Lemma 4 above. Of course, we wonder the further extent to which homothety is well-defined on classes of points on a hypercycle, and we find it remarkable that the Euclidean geometry is somehow clever enough detect these cases so that homothety acts in its correct well-defined way on these points on a fixed hypercycle in each case.

Definite Forms in the Arithmetic Case

For any arithmetic group Γ, we can consider Γ-equivalence classes of definite forms to get a corresponding collection of CM points in the surface $\mathcal{U}/\Gamma$. It is of course tempting to define the notion of Γ-concordance and Γ-Gauss groups geometrically (perhaps by analytic continuation along hypercycles), but I am not certain of some of the details. An improvement in our geometric characterization of concordance would presumably illuminate these points and lead to a simple definition of Γ-Gauss groups. We have also worked on this algebraically for the congruence subgroups $\Gamma(N)$, and it looks promising (to simply mimic Lemmas 2-4 above with specified residues modulo N of certain coefficients).

From the point of view of the absolute Galois group and the universal Ptolemy group (cf. [P2]), the inverse limit of such Γ-Gauss groups (over the usual subgroups of $PSL(2,\mathbf{Z})$ directed by reverse-inclusion) seems a natural construction.

Multiplication on Horocycles

Turning to the case of two forms lying on a common horocycle in $\mathcal{U}$, Estes tells me that the classical point of view gives no clue on what to expect from the Gauss product in

this case, but I offer the following conjectural idea. There is a canonical involution of the divisible group $\mathbf{Q}/\mathbf{Z}$ defined as follows: Given a representative $p/q \in \mathbf{Q}$ with $gcd\{p,q\} = 1$, there is an essentially unique element $\gamma \in PSL(2,\mathbf{Z})$ mapping p/q to infinity and hence the horocycle centered at p/q of Euclidean radius $1/q^2$ to the horocycle centered at infinity of height one. Take the image of $p/q + \sqrt{-1}/q^2$ under γ to define a new point in the horocycle at height one and check that this image point is well-defined up to integral translation, so its real part is well-defined in $\mathbf{Q}/\mathbf{Z}$. More explicitly, one can compute that this involution is given by $p/q \mapsto -s/q$, where $ps - rq = 1$. I suspect that the Gauss product on embedded horocycles is related to the twisting of $\mathbf{Q}/\mathbf{Z}$ by this canonical involution.

It is worth pointing out that there are analogous twistings of $\mathbf{Q}/\mathbf{Z}$ for each arithmetic group Γ; indeed, there is one such twisting for each puncture of the corresponding surface $F = F_\Gamma = \mathcal{U}/\Gamma$. Explicitly, given a CM point ξ on a small horocycle near a specified puncture x of F, consider the geodesic through ξ asymptotic to x; the other end of this geodesic is asymptotic to another puncture y of F. If $x \neq y$, then ξ is taken to be a fixed point of the involution, whereas if $x = y$, then the involution interchanges ξ and η, where η arises as the intersection of the small horocycle about $x = y$ with this other end of the geodesic.

My original intuition about this was that it should be some sort of "jet of Gauss groups" of definite forms about a singular form. This is a reasonable geometric analogue of considering pairs (ideal class, unit in underlying ring) as is efficaciously done in number theory. Such an extension to the singular case seems to be a new idea number-theoretically however.

Indefinite Forms in the Arithmetic Case

A basic difference between the definite and indefinite cases is that $PSL(2,\mathbf{Z})$ acts discontinuously on $\mathcal{U}$ in the former case (and the quotient is the modular curve $\mathcal{M}$), but it does not act discontinuously on the one-sheeted hyperboloid $\mathcal{H}$ in Minkowski space in the latter case. (To see this, just take a minimal geodesic lamination on some surface $F = \mathcal{U}/\Gamma$, where Γ is arithmetic.)

Our basic formulas in Lemma 1 as well as many of our calculations and arguments are not sensitive to whether the forms in question are definite or indefinite, so there is definitely some indefinite version of the theory we have described in this note. It certainly seems natural to imagine $\mathcal{H}/PSL(2,\mathbf{Z})$ (the "indefinite modular curve") as a non-commutative space in the sense of Alain Connes.

A seemingly completely separate aspect in the indefinite case (and this was my starting point in quadratic forms) is the connection with "Markoff tuples" as follows. Fix an arithmetic group Γ, so the quotient $F = \mathcal{U}/\Gamma$ comes equipped with an ideal triangulation (inherited from Farey). We have described global ("lambda length") coordinates on the decorated Teichmüller space (see [P1]) relative to a specification of ideal triangulation, and we consider the decorated surface corresponding to setting all coordinates equal to unity

on this ideal triangulation. This point (the "center of the cell" in the parlance of [P1]) is uniformized by Γ; its coordinates transform under our "Ptolemy transformations" to a collection of integral coordinates, and these tuples of coordinates are called *Markoff tuples* for Γ. The reason for the terminology is that in the case of the once-punctured torus, these are exactly the classical Markoff triples as discussed in [P1]. (Estes pointed this out about my formulas ten years ago!)

It seems clear how to generalize Markoff's own argument (cf. [Di]) from $PSL(2, \mathbf{Z})$ to the case of arithmetic groups Γ, and we imagine each Γ as a group of extreme forms, in the sense of Markoff's theorem, indexed by the corresponding Markoff tuple. Kapranov tells me that Seminaire Rudakov studied a kind of generalized Markoff tuple, and it will thus be interesting to compare our constructions. As a starting point, it should be straight-forward to just check that our Markoff tuples satisfy Rudakov's generalized Markoff equations.

References

[BE] H. S. Butts and D. Estes, "Modules and binary quadratic forms over integral domains", *Linear Algebra and its Applications* **1** (1968), 153-180.

[BP] H. S. Butts and G. Pall, "Modules and binary quadratic forms", *Acta Arithmetica* **15** (1968), 23-44.

[Ca] J. W. S. Cassels, *Rational Quadratic Forms*, Academic Press (1978).

[Di] L. E. Dickson, *Studies in the Theory of Numbers*, Chelsea (1957).

[P1] R. C. Penner, "The decorated Teichmüller space of punctured surfaces", *Communications in Mathematical Physics* **113** (1987), 299-339.

[P2] —, "The universal Ptolemy group and its completions", submitted to the Proceedings of Luminy (1995); eds. P. Lochak and L. Schneps.

[Za] D. Zagier, *Zetafunctionen und quadratische Körper*, Springer-Verlag (1981).

Departments of Mathematics
and Physics/Astronomy,
University of Southern California
Los Angeles, CA 90089, USA