Reduction Theory over Quadratic Imaginary Fields

by David Fried

In fond memory of Arnold Ross

Abstract. *Hurwitz developed a reduction theory for real binary quadratic forms of positive discriminant based on least-remainder continued fractions. For each quadratic imaginary field k, we develop a similar theory for complex binary quadratic forms of nonzero discriminant. This uses a Markov partition for the geodesic flow over the quotient of hyperbolic 3-space by the Bianchi group B_k. When k has a Euclidean algorithm, our theory is based on least-remainder continued fractions.*

This paper extends a reduction theory due to Hurwitz. To put Hurwitz's theory in the proper context, we recall an earlier and better known theory. Gauss's reduction theory for integral binary quadratic forms of fixed positive discriminant was expressed by Dirichlet using simple continued fractions. In this form it was extended by Markov to forms with real coefficients. One factors a real form with fixed discriminant $d > 0$ as

$$Q(x, y) = \sqrt{d}(x - ay)(x - zy)/(a - z), \tag{1}$$

where the *roots* z, a are assumed to be irrational. If $0 < -a < 1 < z$ we say Q is *G-reduced*. We view Q as a function on $\mathbf{R}^2$ and define an action of the group $Gl(2, \mathbf{Z})$ on forms by the rule

$$A \cdot Q = (\det A)(Q \circ A^{-1}). \tag{2}$$

For the following result, see [G, Dir, Ma] or [D1, pp. 11-16, 24-25].

Theorem 1 (Gauss, Dirichlet, Markov). *Every form Q is equivalent under $Gl(2, \mathbf{Z})$ to a G-reduced form. To a G-reduced form Q we associate the terms n_i in the simple continued fraction expansions of z and $-1/a$, namely we write*

$$z = [n_0, n_1, ...] \text{ and } -1/a = [n_{-1}, n_{-2}, ...]. \tag{3}$$

If Q' is a G-reduced form of discriminant d and n'_i are the corresponding terms then Q' is equivalent to Q under $Gl(2, \mathbf{Z})$ if and only if there is an integer l such that $n'_i = n_{i+l}$ for all integers i.

If $A \in Sl(2, \mathbf{Z})$ then Q and $A \cdot Q$ take the same values on the lattice of integral points. For this reason the $Sl(2, \mathbf{Z})$-equivalence of forms was of the greatest interest to Gauss, Dirichlet, and Markov and they found a classification much like Theorem 1. A form Q with irrational roots z, a is called *reduced* if $az < 0 < |a| < 1 < |z|$. A reduced form Q is determined by d, $\mathrm{sgn}(z)$, and the terms n_i in the simple continued fraction expansions of $|z|$ and $1/|a|$. Given such a Q, a reduced form Q' of discriminant d with roots z', a' and terms n'_i is $Sl(2, \mathbf{Z})$-equivalent to Q if and only if there is an integer l so that $n'_{i+l} = n_i$ for all i and $\mathrm{sgn}(z') = (-1)^l \mathrm{sgn}(z)$. Furthermore, every Q with distinct irrational roots is $Sl(2, \mathbf{Z})$-equivalent to a reduced form.

A related theory describes the action of $Gl(2, \mathbf{Z})$ on the space of irrational numbers, see [Se, pp. 34-37.]

Theorem 2 (Serret). *Two irrationals a and b are $Gl(2,\mathbf{Z})$-equivalent if and only if their simple continued fraction expansions have the same tail, that is $a = [a_0, a_1, ...]$ and $b = [b_0, b_1, ...]$ with $a_i = b_{i+l}$ for some integer l and all sufficiently large i.*

To show a and b have the same tail, it suffices to consider the cases $b = a + 1$, $b = -a$, and $b = 1/a$ since $PGl(2,\mathbf{Z})$ is generated by $x + 1$, $-x$, and $1/x$. If $b = a + 1$, one takes $l = 0$. For $b = -a$, one can reduce to the case $0 < a < 1/2$ and use the identities $-a = [-1, 1, a^{-1} - 1]$ and $a = [0, a^{-1}]$ to see that $l = \pm 1$ works. If $b = 1/a$ and a is negative then one can use $b = -1/(-a)$ to reduce to previously handled cases, and one finds that $l = \pm 3$ or ± 1 works. This proves Theorem 2. It follows as well that a and b are $Sl(2,\mathbf{Z})$-equivalent if and only if $a_i = b_{i+l}$ for some even l and all sufficiently large i.

In this paper we will study complex numbers and complex forms Q under the action of $Sl(2, O_k)$, where O_k is the ring of integers in a quadratic imaginary field k. Our results are analogous to those just described for $Sl(2,\mathbf{Z})$ but are more closely modeled on Hurwitz's version of reduction theory.

In section 1 we will describe Hurwitz's elegant approach to Gauss reduction via the Farey tiling. This approach generalizes to higher dimensions and a large class of discrete groups, as in the work of Vulakh [V] and the references cited in that paper. In sections 2 and 3 we will describe Hurwitz's reduction theory and relate it to hyperbolic geometry. Section 4 will present a dynamical reduction theory for compact quotients of hyperbolic n-space that motivates our later results. This reduction theory is based on a finite number of rectangles. Sections 5 and 6 will describe how Hurwitz reduction extends to a discrete group Γ of isometries of hyperbolic n-space such that the quotient space is noncompact but has finite volume. We use results of [F3], which treats a still wider class of discrete groups. Here we are interested in the case where $n = 3$ and Γ is commensurable with the Bianchi group $B_k = PSl(2, O_k)$. Sections 7 and 8 derive a reduction theory for Bianchi groups based on a finite number of rectangles, much like the theory of section 4. Section 9 is devoted to examples and a discussion of some open problems.

Our results may be viewed as extending the theory of least-remainder continued fractions to some non-Euclidean number fields. We may describe the situation by an analogy. Over any number field, the unique factorization of ideals is an algebraic substitute for the unique factorization of integers. For a quadratic imaginary field, a reduction theory based on a finite number of rectangles is a geometric substitute for the Euclidean algorithm.

We were first introduced to continued fractions and to quadratic forms in Arnold Ross's high school summer program at Ohio State University. His inspiring teaching has given many of us a lifetime of mathematical interests and we owe him an enormous debt.

Section 1. The Farey tiling and Gauss reduction

We will present here a geometric approach to Gauss reduction due to Hurwitz [H3] (see also [D1, pp. 42-44]). To begin, we will describe the projective model of hyperbolic geometry and its relation to quadratic forms.

The discriminant d of a real binary quadratic form $ux_1^2 + vx_1x_2 + wx_2^2, d = u^2 - 4vw$, defines a quadratic form in the variables u, v, and w of signature (2,1). In the corresponding real projective plane one obtains a conic C, an open disc D, and a Moebius strip M by the conditions $d = 0, d < 0$, and $d > 0$, respectively. A point of C is given by a quadratic form

with a square factor $(t_1x_1 - t_2x_2)^2$ and we identify this point with the extended real number $t_2/t_1 \in \mathbf{R} \cup \{\infty\} = \hat{\mathbf{R}}$. For a fixed $d > 0$, a point of M corresponds to a pair of forms $\pm Q(x_1, x_2)$ of discriminant d. The tangents from this point to C meet C at the roots z, a of Q. In this way we identify the forms of discriminant d with open oriented segments $\overrightarrow{az}$ in D, which for brevity we will call *rays*.

The group $Gl(2, \mathbf{R})$ acts linearly on forms as in equation (2), preserving the discriminant. Since

$$A \cdot (x_1 - tx_2)^2 = (\det A)((t\ a_{21} + a_{22})x_1 - (t\ a_{11} + a_{12})x_2)^2 \tag{4}$$

for $A = (a_{ij})$, we see that A acts on C by the usual linear fractional transformation of $\hat{\mathbf{R}}$, that is by $(t\ a_{11} + a_{12})/(t\ a_{21} + a_{22})$. It follows that the roots z, a of a quadratic form Q also transform by the usual action of A.

Let e_0 denote the open segment $\overline{0\infty}$ in D and let $\Gamma = Gl(2, \mathbf{Z})$. For $A \in \Gamma$ the segment $A(e_0)$ does not cross e_0 since $A0/A\infty = 1 - (\det A)/(a_{11}a_{22})$ is not negative. Thus the images of e_0 under Γ, known as *Farey edges*, do not cross one another and so they divide D into regions. Note that the endpoints of Farey edges are extended rational numbers, which we call *cusp points*.

Let T_0 denote the ideal triangle in D with vertices $0, 1$, and ∞. The images of T_0 under Γ, known as *Farey triangles*, can be built iteratively by a sort of crystallization process with seed T_0. Each edge of T_0 meets another Farey triangle, each of these meets just two more Farey triangles, and so on. At each stage in the construction, one has covered a convex subset of D. Since cusp points are dense in C, we see that the Farey triangles cover D without overlap, forming the *Farey tiling*.

Now fix a ray s with irrational endpoints. Any Farey triangle T that meets s must intersect s in a compact interval, since its vertices are not endpoints of s, and the three vertices of T are separated by s into a pair and a singleton . We label this interval in s by the singleton vertex. A given cusp point s can only label finitely many consecutive intervals in s. We call a Farey edge e *principal for* s if e crosses s and the intervals of s on either side of e are labeled by distinct vertices. Now trace s and count the number of intervals between crossings of principal edges. We obtain a sequence n_i of positive integers indexed by $i \in \mathbf{Z}$, unique up to an overall shift of indices by some integer l. Since the Farey tiling is invariant by Γ, it is clear that this sequence is the same, up to a shift, for s and for As, $A \in \Gamma$. We will show

Lemma 1. *Any sequence $n : \mathbf{Z} \to \mathbf{N}$ arises for some ray s. Γ-equivalence classes of s's correspond bijectively to shift-equivalence classes of n's.*

To begin, note that the stabilizer of e_0 in Γ is given by the four elements $\pm x,\ \pm 1/x$. If the edge e_0 is principal for s then we may apply exactly one of these four elements to obtain a ray s' so that the vertex label changes from 0 to ∞ as s' crosses e_0 and so that the endpoint parameters increase from one end of s' to the other. We say such a ray s' is *G-reduced*. In terms of its endpoints, $s' = \overrightarrow{az}$ is G-reduced if $0 < -a < 1 < z$. We see that for any ray s with irrational endpoints,

(*)G-reduced rays equivalent to s under Γ correspond 1-1 to principal edges of s.

Suppose $s = \overrightarrow{az}$ is G-reduced. If $n = [z]$ then the next principal edge that s crosses is $e = \overline{n\infty}$. Applying the transformation $1/(x - n)$ brings e to e_0. Let $N_G(s) = N_G(z, a)$, the

neighbor of s, be the image of s under $1/(x-n)$. Clearly $N_G(s)$ is G-reduced. By (*), the G-reduced rays equivalent to s are just the images of s under the cyclic group generated by N_G.

Moreover, in terms of endpoints we have $N_G(z,a) = (1/(z-n), 1/(a-n)),\ n = [z]$. It follows that if we expand z and $-1/a$ as simple continued fractions, as in (3), we obtain the same terms n_i as in the paragraph preceding Lemma 1. Now given n we can construct z, a by (3) and hence a G-reduced s with a given sequence n. This proves the first part of Lemma 1. Now (*) shows that the sequence n determines s up to equivalence and Lemma 1 follows. Theorem 1 is a direct consequence of Lemma 1.

To sum up, a ray with irrational endpoints is determined by its family of principal edges. Designating one principal edge as a starting point, we can define a sequence n. This sequence determines our ray up to the action of Γ. Changing the principal edge changes the sequence by a shift.

Section 2. Hurwitz reduction theory

We present Hurwitz's reduction theory based on least-remainder continued fractions [H2] (see also [D1, pp. 40-41]). One can expand any irrational number x_0 in a least-remainder continued fraction (often called a *continued fraction to the nearest integer*) as follows. Let $a_0 = \langle x_0 \rangle$ be the nearest integer to x_0 and write $x_0 = a_0 - 1/x_1$. Then $|x_1| > 2$. We take the integer $a_1 = \langle x_1 \rangle$ and write $x_1 = a_1 - 1/x_2$ with $|x_2| > 2$, etc. The resulting expansion

$$x_0 = a_0 - 1/(a_1 - 1/(a_2 - ...)) =< a_0, a_1, a_2, ... > \tag{5}$$

has the following features. The a_i are integers, $|a_i| \geq 2$ for $i \geq 1$, and if $a_i = \pm 2$ for some $i \geq 1$ then $a_i a_{i+1} < 0$. Conversely, if integers $a_i, i \geq 0$, obey these conditions then Hurwitz showed that $< a_0, a_1, a_2, ... >$ is the least-remainder continued fraction of an irrational number.

Let $\Gamma^+ = Sl(2, \mathbf{Z})$. Clearly x_0 is Γ^+-equivalent to each tail x_l for any l. So x_0 is equivalent to x_0' if $x_l = x_{l'}'$ for some l, l'. The converse fails, however. Let $r = (3-\sqrt{5})/2$. Then $x_0 = r$ and $x_0' = -r$ are equivalent since $-1/(r-1) = 2 - r$, yet their tails are all $-r^{-1}$ and r^{-1} respectivelvely. Hurwitz proved that this is the only exception, up to equivalence. That is,

Theorem 3 (Hurwitz). *If two equivalent irrationals do not share a tail then they are equivalent to $r = (3-\sqrt{5})/2$. Moreover one has r^{-1} as a tail and the other has $-r^{-1}$ as a tail.*

Thus the least-remainder analogue of Theorem 2 has a countable dense set of exceptions. One can prove Theorem 3 by showing directly that x and $-1/x$ have the same tail for x in certain intervals. One finds that the lag l becomes unbounded as x approaches $2 - r$. This contrasts with Theorem 2, where our proof shows that the lag is bounded for each element of Γ.

Hurwitz defines a reduction theory for forms in which the roots are treated in an asymmetric fashion. If a_i are integers for all $i \in \mathbf{Z}$ with $|a_i| \geq 2$ for all i and if $a_i a_{i+1} < 0$ for $|a_i| = 2$, we let

$$z =< a_0, a_1, a_2, ... > \text{ and } 1/a =< a_{-1}, a_{-2}, ... > \tag{6}$$

and we say that a form Q with roots z, a is *H-reduced*. The given continued fraction expansion for z is a least-remainder expansion whereas that for $1/a$ is a sort of dual expansion, since

the terms of absolute value 2 are treated in a backwards fashion. As we will explain in the next section, the H-reduced forms can be described by inequalities in the roots, namely

$$|z| > 2 \quad \text{and} \quad \operatorname{sgn}(z)a \in [r-1, r] \tag{7}$$

are the necessary and sufficient conditions for Q to be H-reduced. (We should mention that our condition "H-reduced" is slightly weaker than Hurwitz's condition "reduziert" in [H2]).

Theorem 4 (Hurwitz). *Every form of positive discriminant is Γ^+-equivalent to an H-reduced form. The H-reduced forms corresponding to a_i and a'_i are equivalent if for some integer l and all integers i one has $a'_i = a_{i+l}$. The converse holds unless a is equivalent to r.*

For instance, the form $Q(x,y) = 3xy - x^2 - y^2$ has $a_i = 3$ for all i and the form $Q'(x,y) = 3xy + x^2 + y^2$ has $a' = -3$ for all i, both are H-reduced forms of discriminant 5, yet $Q' = A \cdot Q$ for $A = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$.

These exceptions, minor though they may be, are perhaps the reason that Hurwitz's theory has received so little notice over the years. It will be seen, however, that this theory generalizes well to higher dimensions.

One can easily apply Hurwitz's results to classify Γ-equivalence of forms. One just takes as reduced forms those H-reduced forms with $z > 0$. We will omit the routine details.

Section 3. The Ford disc packing and Hurwitz reduction

Recall that the disc D of section 1 is the projective model of the hyperbolic plane. We will now describe the correspondence of this model with the upper halfplane model H^2 of the hyperbolic plane, construct the Ford discs, and use them to derive equation (7).

H^2 is the set of all complex numbers with positive imaginary part, thought of as a subset of the extended complex numbers $\hat{\mathbf{C}}$. The extended real numbers $\hat{\mathbf{R}}$ correspond to the boundary conic C in the manner already described. A ray $s = \overrightarrow{az}$ in D becomes an oriented hyperbolic line in H^2. If z or a is ∞ then this line is a Euclidean ray perpendicular to $\mathbf{R}$ at the finite endpoint. If both endpoints of s are finite, this line is a Euclidean semicircle perpendicular to the real line and approaching it at a and z. We measure the length of tangent vectors with the hyperbolic metric $ds^2 = (du^2 + dv^2)/v^2, w = u + iv$ on H^2. The action of $Gl(2, \mathbf{R})$ on D corresponds to the usual action by linear fractional transformations on H^2. These are isometries of the hyperbolic metric and, indeed, all the isometries of H^2 arise this way. To Γ corresponds the triangle group $(2, 3, \infty)$ generated by reflections in the 3 hyperbolic lines $\Re(w) = 0$, $\Re(w) = 1/2$, and $|w| = 1$.

Both models of the hyperbolic plane have some advantages. In D, hyperbolic lines are straight and so hyperbolic convex sets are modeled by Euclidean convex sets. In H^2, hyperbolic angle is modeled by Euclidean angle. We will work henceforth in the model H^2 and its higher-dimensional analogues because we will be using horoballs. In H^2, a horoball is either a disc tangent to $\mathbf{R}$ or a halfspace defined by an equation $\Im(w) \geq c$, for some positive c. In D, on the other hand, horoballs are awkward to describe. Furthermore, we will often be fixing a point in the boundary of the hyperbolic plane, so the upper halfspace model is ideal.

Let $B(\infty)$ be the standard horoball defined by $\Im(w) \geq 1$. The images of this horoball under elements of Γ form a family of horoballs known as *Ford discs* [C, pp. 27-33]. We label them $B(q)$ where the *base* q is the cusp point where the horoball touches $\hat{\mathbf{R}}$. Note that $B(\infty)$ lies in the union of the collection of Farey triangles with ∞ as a vertex. It follows that the same holds for all cusp points q, and hence a given Farey triangle T meets only those Ford discs $B(q)$ for which q is a vertex of T. It then follows that the only Ford discs that meet $B(\infty)$ are the $B(q)$ for integer q, and hence that the Ford discs do not overlap. They do not cover the plane but the union of the Ford discs is closed and the components of its complement are bounded. The family of Ford discs is a geometric object dual to the Farey tiling in the sense that each Ford disc corresponds to a Farey vertex, each pair of Ford discs that touch corresponds to a Farey edge, and each region in the complement of the Ford discs corresponds to a Farey triangle.

The Ford discs give rise to a geometric interpretation of Hurwitz reduction, as follows. Fix an extended real number z and a cusp point q_0 not equal to z. The ray $s_0 = \overrightarrow{q_0 z}$ exits $B(q_0)$ and must enter another Ford disc. (Reducing to the case $q_0 = \infty$, we see that the Ford discs based at integer points meet every vertical line). If the next Ford disc it enters is $B(q_1)$, we say that q_1 is the *z-successor* to q_0. If q_1 does not equal z, we may define its z-successor q_2, and so on. If $q_0 = \infty$ then $q_1 = \langle z \rangle$. If we expand z as a least-remainder continued fraction $z =< a_0, a_1, a_2, ... >$ then we see that $q_{n+1} =< a_0, a_1, ..., a_n >$ is the nth least-remainder convergent to z.

We examine the expansion of $1/a$ in (6) and some similar continued fractions. Consider a continued fraction $< b_1, b_2, ... >$, of finite or infinite length, with integer terms b_j such that for $j \geq 1$, $|b_{j+1}| \geq 2$ and $b_j b_{j+1} \leq 0$ if $|b_{j+1}| = 2$. Since such expansions were introduced by Hurwitz as the dual expansions for least-remainder continued fractions, we will call these *H-dual expansions*. One may assign to each cusp point q and to the Ford circle $B(q)$ a *level* so that ∞ is in level 0, q is in level 1 if q is an integer, q' is in level 2 if $B(q')$ meets some Ford disc $B(q)$ of level 1 but q' is not of level ≤ 1, etc. Each H-dual convergent $< b_1, b_2, ..., b_j >$ lies in level j. This is clear for $j = 1$ and it is also clear that the level is at most j. The induction step uses the exclusions $b_j \neq 0, \pm 1$ to show that one descends from level $j-1$ to level j and neither retreats to level $j-2$ nor moves to one of the two neighboring Ford discs of level $j-1$. Conversely, every cusp point of level j has a unique finite H-dual expansion consisting of j terms. The uniqueness uses the exclusion $b_{j+1} \neq 2\text{sgn}(b_j)$ to handle those level $j+1$ Ford discs that touch two Ford discs of level j.

Thus the H-dual continued fraction expansion of a rational number is unique, unlike its simple continued fraction expansion or its least-remainder continued fraction expansion.

We plot the discs $B(q)$ for $q =< b_1, ..., b_j >$ where $b_1 = 0$. One finds for $j = 2$ all the level 2 discs that meet $B(0)$ and for $j = 3$ all the level 3 discs that meet these level 2 discs, and so on.

The largest cusp point of level 2 produced this way is $1/2$, the next larger cusp point of level 2 is $2/3$, as drawn in Figure 1. The largest cusp point of level 3 produced this way is $3/5$, the next larger cusp point of level 3 is $5/8$, and so on. This sequence $1/2, 3/5, ...$ converges to $(\sqrt{5}-1)/2 = 1 - r$. Using the symmetry with respect to $-x$, it follows that H-dual expansions with $b_1 = 0$ fill out the interval $[r-1, 1-r]$. Similarly, the H-dual expansions with a fixed nonzero value $b_1 = b$ fill out the interval $[b-r, 1+b-r]$ if b is positive and the interval $[b-1+r, b+r]$ if b is negative.

The expansion of $1/a$ in (6) is an H-dual expansion with $b_i = a_{-i}$. The initial term is constrained by $a_{-1} \neq 0, \pm 1, (\text{sgn } a_0)2$. This shows that $1/a \notin (\text{sgn } z)(r-2, 3-r)$ hence $(\text{sgn } z)a \in [r-1, r]$. This establishes the condition (7) describing H-reduced forms. The converse, that (7) implies that the form is H-reduced, is proven similarly.

We have shown that the Farey tiling and the Ford discs are dual geometric configurations that underlie Gauss reduction and Hurwitz reduction, respectively.

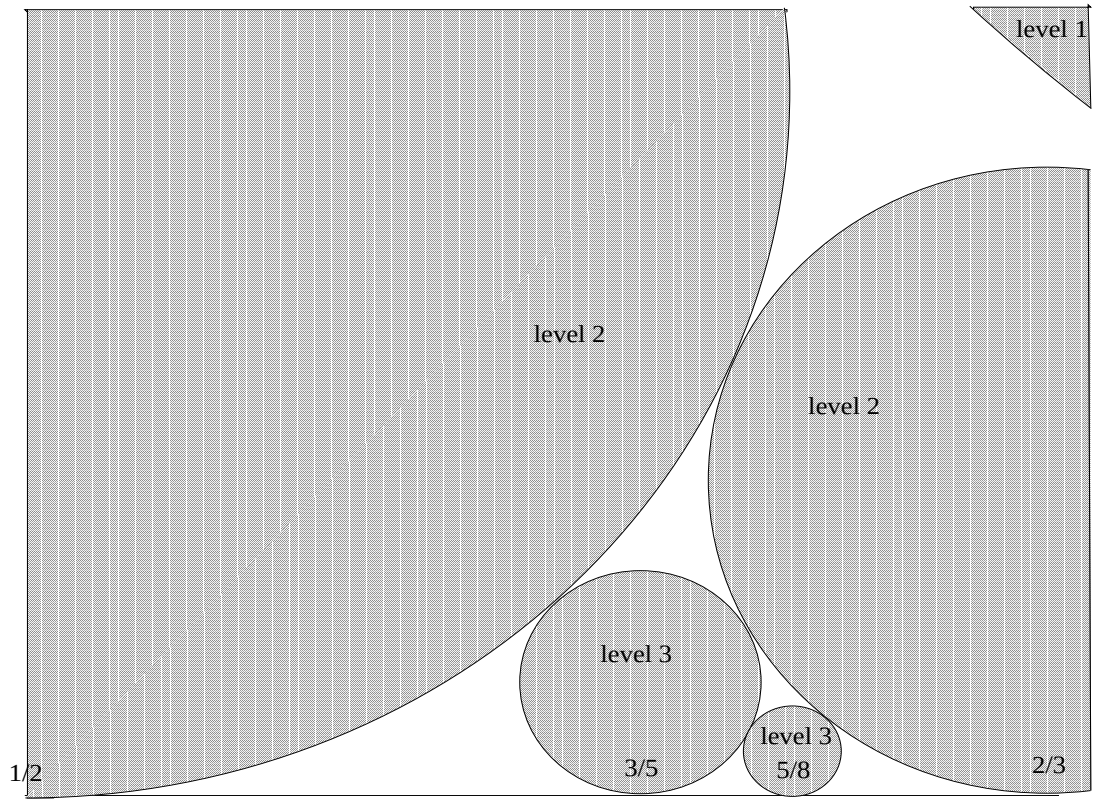

Figure 1. Ford discs for some Fibonacci quotients

Section 4. Rectangles and reduction

The model of hyperbolic n-space that we use is the n-dimensional upper halfspace H^n, consisting of the points in $x \in \mathbf{R}^n$ with $x_n > 0$. The hyperbolic metric is $(dx_1^2 + ... + dx_n^2)/x_n^2$. We regard H^n as a subset of the n-sphere $\hat{\mathbf{R}}^n = \mathbf{R}^n \cup \infty$ and identify the boundary of hyperbolic n-space with $\partial H^n = \hat{\mathbf{R}}^{n-1}$. For $n = 3$, we identify $\mathbf{R}^2$ with $\mathbf{C}$ and we identify the isometry group of H^3 with the group of conformal and anticonformal transformations of $\hat{\mathbf{C}}$, that is with the semidirect product of $PGl(2, \mathbf{C})$ by $\{z, \overline{z}\}$.

Consider a discrete group Γ of isometries of H^n such that the quotient $X = H^n/\Gamma$ is of finite volume. A reduction theory for Γ is a certain sort of description of its orbits on ordered pairs (z, a) of distinct points in the boundary ∂H^n of hyperbolic n-space. We will now describe a theory for compact X that will motivate our results for the noncompact case. This is, in fact, just a special case of a theory of Markov partitions due to Bowen and Ratner

[B, Ra], see also [F1, section 1].

Let Ω denote the open set of distinct pairs in $\partial H^n \times \partial H^n$ so that $(z,a) \in \Omega$ if $z \neq a$. A *rectangle* is just a compact subset of Ω that is a Cartesian product $Z \times A$ in the (z,a) coordinates.

Suppose X is compact and Γ acts freely on $H^n \times \partial H^n$. (The latter condition is presumably not necessary but, as the literature treats manifolds and not orbifolds, it is certainly convenient.) One finds a certain set $\mathcal{S}^*$ of *symbols* on which Γ acts freely with finitely many orbits. There is a Γ-equivariant family of rectangles $R_s = Z_s \times A_s, s \in \mathcal{S}^*$, whose union is Ω. There is a set $\mathcal{T}^*$ of *transitions*, on which Γ acts freely with finitely many orbits. There are two equivariant functions i^* and f^* (for *initial* and *final*, respectively) from $\mathcal{T}^*$ to $\mathcal{S}^*$, such that

$$A_i \subset A_f \text{ and } Z_f \subset Z_i \tag{8}$$

for $i = i^*(t), f = f^*(t), t \in \mathcal{T}^*$. On the disjoint union $\coprod R_s$ we define for each $t \in \mathcal{T}^*$ a relation $\mathcal{P}_t^*$ consisting of pairs $((z,a),(z,a))$ with $(z,a) \in (Z_f \times A_i)$. The union of these relations over all $t \in \mathcal{T}^*$ defines a relation $\mathcal{P}^*$ we call the *equivariant Poincare correspondence*. Theorem 5 will show that one can choose the rectangles so that for all $s \in \mathcal{S}^*$ and almost all $(z,a) \in R_s$,

$$\text{all instances of } (z,a) \in R_{s'},\ s' \in \mathcal{S}^*, \text{ arise by iterating } \mathcal{P}^*. \tag{9}$$

Now we factor out by Γ to get a description of geodesics on X. We select sets of orbit representatives $\mathcal{S}$ and $\mathcal{T}$ with $i^*(\mathcal{T}) \subset \mathcal{S}$. Let U be the union of the rectangles $R_s, s \in \mathcal{S}$. Since U meets every orbit of Γ on Ω, it suffices to describe the Γ-equivalence of points in U, which we will do by a *reduced Poincare correspondence* $\mathcal{P}$. We choose for each $t \in \mathcal{T}$ the elements $\gamma(t) \in \Gamma$ for which $\gamma(t) f^*(t) \in \mathcal{S}$. Let $i, f : \mathcal{T} \to \mathcal{S}, i(t) = i^*(t), f(t) = \gamma(t) f^*(t)$. We define the relation $\mathcal{P}(t) \subset R_{i(t)} \times R_{f(t)}$ to consist of the pairs $((z,a), \gamma(t)(z,a))$ for $(z,a) \in Z_{f^*(t)} \times A_{i^*(t)}$. We let $\mathcal{P}$ be the relation on $\coprod R_s,\ s \in \mathcal{S}$, that is union of these relations $\mathcal{P}(t)$ over $t \in \mathcal{T}$. Clearly, certain instances of Γ-equivalence of points in U arise from $\mathcal{P}$ and its iterates. Conversely, we obtain from equation (9) that for almost all $(z,a) \in U$,

$$\text{all points of } U \text{ } \Gamma\text{-equivalent to } (z,a) \text{ arise by iterating } \mathcal{P}.$$

This implies

Theorem 5. *Suppose X is compact and Γ acts freely on $H^n \times \partial H^n$. There are finite sets $\mathcal{S}, \mathcal{T}$ of symbols and transitions, rectangles $R_s = Z_s \times A_s \subset \Omega, s \in \mathcal{S}$, and maps $i, f : \mathcal{T} \to \mathcal{S},\ \gamma : \mathcal{T} \to \Gamma$ so that for $t \in \mathcal{T}, i = i(t), f = f(t), \gamma = \gamma(t)$ we have*

$$Z_f \subset \gamma Z_i \text{ and } \gamma A_i \subset A_f. \tag{10}$$

A relation in $R_i \times R_f$ is defined by the bijection

$$\gamma :\ \gamma^{-1} Z_f \times A_i \to Z_f \times \gamma A_i.$$

Let $\mathcal{P}$ be the relation on $U = \cup R_s$ that is the union of these relations over all $t \in \mathcal{T}$. The iterates of $\mathcal{P}$ give Γ-equivalent points in U and, conversely, for almost all $(z,a) \in U$,

$$\textit{all points of } U \text{ } \Gamma\textit{-equivalent to } (z,a) \textit{ arise by iterating } \mathcal{P}. \tag{11}$$

For each unit tangent vector v on H^n there is a unique unit-speed geodesic $x(u), u \in \mathbf{R}$ with $x'(0) = u$. We let $\phi_u^*(v) = x'(u)$. Letting $Y^* = SH^n$ be the space of unit tangent vectors to H^n, we see that ϕ_u^* is a one-parameter group of diffeomorphisms of Y^*, or a *flow* on Y^*. This is the well-known *geodesic flow* on Y^*. Each flowline $\phi_u^*(v), v \in Y^*$, traces a ray $\overrightarrow{az}$, so we may identify the space of flowlines of the geodesic flow ϕ_u^* with Ω.

Our conditions on Γ imply that $Y = Y^*/\Gamma$ is a compact manifold. Since the geodesic flow on Y^* commutes with hyperbolic isometries, there is a flow induced on Y that we call the *geodesic flow over* H^n/Γ. The product structure on Ω defines two transverse n-dimensional foliations on Y^* that descend to foliations on Y. These foliations are tangent to the geodesic flow and invariant by the geodesic flow. One of them, the *stable foliation*, is contracted for large positive u and the other, the *unstable foliation*, is contracted for large negative u, at least in directions not parallel to the flow. This foliation pair makes ϕ_u an *Anosov flow*. Note that this flow preserves a volume form, namely the Liouville measure.

Recall that any smooth flow without fixed points on a manifold corresponds in local coordinates to the standard translational flow on $\mathbf{R}^m$ given by $\tau_u(x) = x + ue_1$. A *flowbox* is a compact subset that corresponds in some such coordinate system to a set of the form $\{x \in \mathbf{R}^m : 0 \leq x_1 \leq f(x_2, ..., x_m), (x_2, ..., x_m) \in T\}$, where $T \subset \mathbf{R}^{m-1}$ is compact and $f : T \to \mathbf{R}^+$ is continuous. We call the sets corresponding to $x_1 = 0$ and $x_1 = f(x_2, ..., x_m)$ the *initial face* and *final face* of this flowbox, respectively. A *transversal* to a flow is smooth, codimension-one submanifold Z that is transverse to the flow (that is, the vector field that generates the flow is nowhere tangent to Z). Note that for our purposes, $m = \dim Y = 2n-1$.

Bowen constructs a *Markov partition* for each volume-preserving Anosov flow on a compact manifold. We will explain this result for the geodesic flow on Y. There is a transversal T such that every flowline meets T in every time interval of some fixed small length. T consists of many small connected components D_s, each of which is a smooth disc with smooth boundary. Each of these discs contains a compact set T_s which is *rectangular* in the sense that when D_s is lifted and projected to Ω, T_s maps to a rectangle. Consider a small flowbox Y_t for ϕ_u that meets T only in its initial and final face. Suppose that its initial face lies in T_i and is rectangular with the same A-factor as T_i. Suppose that its final face lies in T_f and is rectangular with the same Z-factor as T_f. Bowen's Markov partition is a finite nonoverlapping family Y_t of such flowboxes that covers Y such that the flowbox index t varies over certain ordered pairs (i, f) of disc indices, namely those for which there are short flowlines from T_i to T_f sweeping out an open set in Y. Ratner proves a uniqueness result: almost all flowlines $\phi_u(v)$ are determined by the sequence of flowboxes that they pass through. Indeed the only exception is a flowline that passes through the boundary of some flowbox for some nontrivial u-interval and the set of such flowlines has measure zero.

We lift the flowboxes Y_t to Y^* to get a Γ-equivariant family of small flowboxes which we label with an index set $\mathcal{T}^*$. These flowboxes cover SH^n with overlapping. The initial or final faces of these flowboxes lie in smooth discs that are lifts of the discs D_s. We label these lifted discs and their subsets corresponding to T_s by an index set $\mathcal{S}^*$ and project each of these subsets into Ω to get a rectangle R_s, $s \in \mathcal{S}^*$, as in Figure 2. We let i^*, f^* be the maps that assign to each flowbox the transversals containing its initial and final face, respectively. Then (8) follows from the properties of the initial and final faces of the flowboxes $Y_t \subset Y$. Ratner's uniqueness result implies (9) and Theorem 5 is a consequence of (8) and (9), as explained above.

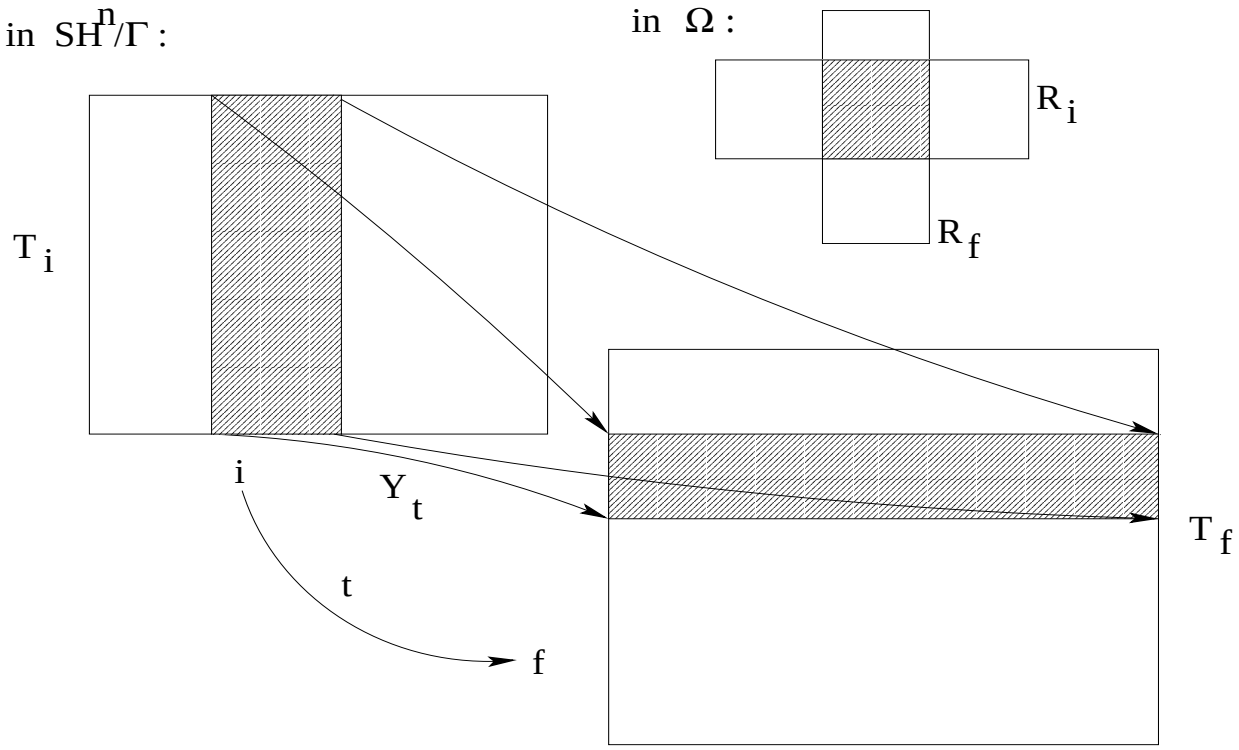

Figure 2. A transition (flowbox) between two symbols (transversals)

Note that the rectangles $R_s,\ s \in \mathcal{S}$, correspond to a finite family of rectangular regions and the transitions $t \in \mathcal{T}$ correspond to a finite family of flowboxes for the geodesic flow on the manifold Y. The maps i, f describe how flowlines pass from one transversal to another through the flowboxes. This use of transversals goes back to Poincare, hence our name for the relation $\mathcal{P}$.

Transversals can also be used to describe Gauss reduction. Referring to section 1, we form a transversal to the geodesic flow on SH^2 that consists of the unit tangent vectors to rays at points where the rays cross principal edges. For the reduced Poincare correspondence, we take $\mathcal{S}$ to consist of one point, so subscripts will be omitted. We define the rectangle $Z \times A = [1, \infty] \times [-1, 0]$ to be the completion of the space of root pairs of G-reduced forms. We take $\mathcal{T} = \mathbf{Z}^+$, the set of terms in simple continued fractions. For $n \in \mathcal{T}$ we let $\gamma_n(x) = 1/(x - n)$. Then equation (10) holds and Theorem 1 shows that (11) holds, where the set of full measure consists of pairs with z and a irrational.

We may likewise compare Theorem 5 with Hurwitz reduction. We leave the description of the relevant transversal to section 6, and simply describe the reduced Poincare correspondence. Let $\mathcal{S} = \{+, -\}, Z_\pm = \pm[2, \infty]$, and $A_\pm = \pm[r - 1, r]$. Let $\mathcal{T}$ consist of the following pairs in $Z \times \mathcal{S}$: $(\pm 2, -)$ and $(n, \pm)$ where $|n| \geq 3$. For $t = (n, \epsilon) \in \mathcal{T}$, let $\gamma_t(x) = -1/(x - n)$, $f(t) = \text{sgn}(n)$, and $i(t) = \epsilon f(t)$. Then equation (10) holds, and Theorem 4 shows that (11) holds, where the set of full measure consists of pairs with z and a irrational

and a inequivalent to r.

Thus Gauss and Hurwitz reduction theory have a structure similar to the reduction theory of Theorem 5.

Section 5. Fans, ridges, shadows, and chains

We now suppose our finite volume quotient $X = H^n/\Gamma$ is noncompact and develop a Hurwitz-style reduction theory. The main difference from Theorem 5 is that $\mathcal{T}$ must be infinite. In this section we find a certain Γ-invariant tiling of H^n. In the next section we will describe a reduction theory that includes Hurwitz's theory for $PSl(2, \mathbf{Z})$. In section 8 we will restrict to Bianchi groups and show that we can express this reduction theory using a finite system of rectangles, as in Theorem 5.

We will need many standard facts concerning Γ, such as can be found in [R].

Although the topology and singularities of a noncompact X may be complicated, X has a standard structure near ∞ that furnishes a starting point for our reduction theory. To explain this structure, we use the *trichotomy* for isometries of H^n that divides Γ into elliptic, parabolic, and hyperbolic elements.

An isometry is *elliptic* if it fixes a point of H^n. In Γ the elliptic elements have finite order and nontrivial ones produce orbifold singularities in the quotient. Nonelliptic isometries must fix a point in ∂H^n.

A nonelliptic isometry with only one such fixed point is called *parabolic* and the fixed point is called a *cusp point* for Γ. The stabilizer Γ_q of a cusp point q preserves each horoball based at q where, as for $n = 2$, a horoball based at ∞ is a closed halfspace and a horoball based at any other boundary point q is a Euclidean disc tangent to $\mathbf{R}^{n-1}$ at q [R, theorem 5.5.5]. We can find a horoball $B(q)$ based at q such that the quotient $N(q) = B(q)/\Gamma_q$ embeds in $H^n/\Gamma = X$ [R, theorem 12.5.2]. We call $N(q)$ a *cusp neighborhood* in X. It is an unbounded region bounded by a complete Euclidean $(n-1)$-orbifold (if $q = \infty$, as we may suppose, the stabilizer Γ_q acts on $\mathbf{R}^{n-1}$ by isometries and has compact quotient). $N(q)$ is isometric to a warped product of this orbifold and $(0, \infty)$ in which the metric on the orbifold fibers shrinks exponentially as one approaches ∞. If $B(q)$ is chosen smaller, the cusp neighborhood is truncated in an evident fashion.

The set Q of all cusp points for Γ consists of finitely many orbits of Γ. Selecting one cusp point $q(1), ..., q(h)$ from each such orbit and horoballs at these points that are sufficiently small, the corresponding cusp neighborhoods do not overlap. The complement of these h cusp neighborhoods has compact closure [R, corollary 4 to theorem 12.6.4]. Thus X has h ends (that is, the complement in X of a sufficiently large compact set has exactly h unbounded components) and each end is of a standard form.

An element $\gamma \in \Gamma$ of infinite order that fixes two boundary points is called *hyperbolic*. If these two points are z, a then γ stabilizes the hyperbolic line $l = \overline{az}$ (as for $n = 2$, this is represented by a ray perpendicular to $\mathbf{R}^{n-1}$ if one of a, z is ∞ and otherwise by a semicircle perpendicular to $\mathbf{R}^{n-1}$). γ acts on l by a nontrivial translation. From this it follows that γ fixes no boundary points except z and a. The elliptic elements in $Stab(l)$ form a normal subgroup and the quotient is infinite cyclic. The quotient of the ray $s = \overrightarrow{az}$ by $Stab(l)$ is a closed geodesic in X and it is prime (that is, it is not a multiple of some shorter closed geodesic). Each prime closed geodesic in X arises in this way.

To sum up, the nontrivial elements of Γ are related to orbifold singularities of X, ends of X, or closed geodesics in X according to their place in the trichotomy.

Suppose now that we choose disjoint cusp neighborhoods $N(q(i))$ as above. The inverse images of these neighborhoods in H^n form a Γ-equivariant system of nonoverlapping horoballs, or a *horoball packing*. There is a constant D such that every point of X is distance $\leq D$ from some cusp neighborhood. It follows that every point in hyperbolic space is distance $\leq D$ from some horoball in our packing, that is our packing is *D-dense*.

The packing by Ford discs is obtained for $n = 2$ and $\Gamma = PGl(2, \mathbf{Z})$ by selecting the horoball $B(\infty)$ to be the standard horoball $\Im(z) \geq 1$. Here X is an orbifold with one end and the cusp stabilizer Γ_∞ is generated by $-x$ and $1 - x$.

We now use the horoball packing $B(q), q \in Q$, to divide hyperbolic space into polyhedra akin to Dirichlet domains. We let $F(q)$ consist of all points whose distance to $B(q)$ is no more than their distance to $B(p)$ for any $p \in Q$. Then $F(q)$ is an infinite-sided convex hyperbolic polyhedron that contains $B(q)$ and the sets $F(q), q \in Q$, cover H^n without overlap. Moreover, each point of $F(q)$ is distance at most D from $B(q)$ and so $F(q)$ is contained in some horoball based at q. We call $F(q)$ a *fan* and the family $F(q), q \in Q$, the *fan tiling* associated to our horoball packing. One may regard the nonnegative function $dist(x, \cup B(q))$ as a height function on H^n and a fan $F(q)$ as a basin for a river with delta $B(q)$ and mouth q.

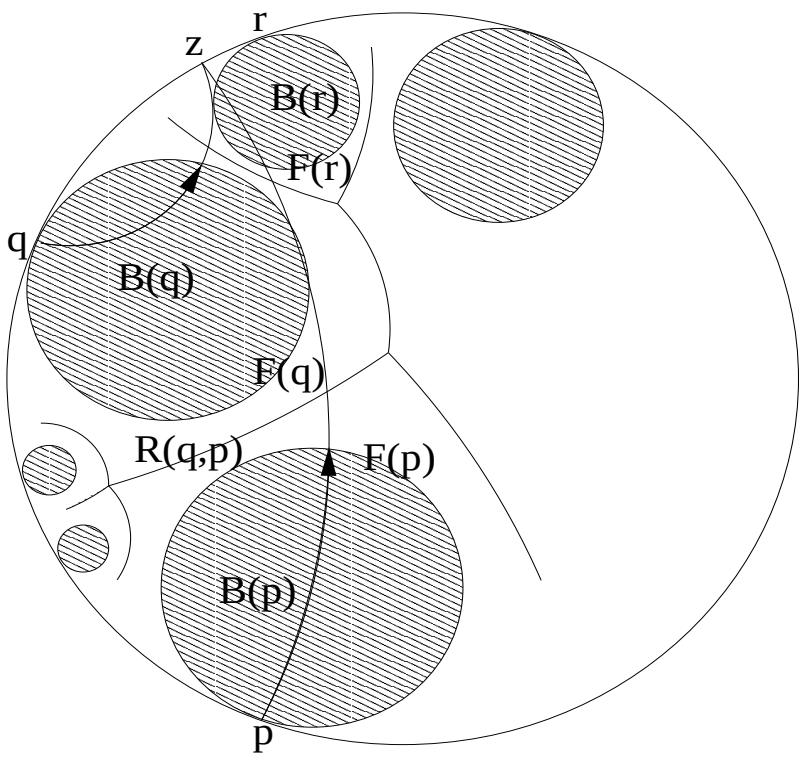

Figure 3. A point z in S(q,p) and S(r,q)

Suppose p and q are distinct cusp points. The intersection of $F(p)$ and $F(q)$ is either empty or a convex hyperbolic polyhedron of dimension at most $n-1$. If the intersection is nonempty and of dimension $n-1$ we say that p and q are *adjacent* and we define the *ridge* $R(q,p)$ to be this intersection of fans. We project $R(q,p)$ from p to ∂H^n along geodesics to define a subset $S(q,p)$ of ∂H^n that we call the *shadow* of the ridge $R(q,p)$ as lit from p. Figure 3 shows some horoballs, fans, and ridges in the Poincare disc model of the hyperbolic plane. This model is conformally equivalent to the upper halfplane model but does not have a preferred boundary point.

When we identify $\partial H^n - p$ with $\mathbf{R}^{n-1}$ (using an isometry g with $g(p)=\infty$), $S(q,p)$ is a convex Euclidean polyhedron of dimension $n-1$. We note that these polyhedra tile $\mathbf{R}^{n-1}$ without overlap, as q varies over all the cusp points adjacent to p.

We define a *chain* to be a nonempty set $S(p_l,...,p_m) = S(p_l,p_{l-1}) \cap ... \cap S(p_{m+1},p_m) \subset \partial H^n, , l > m$. Note that each shadow in the definition of a chain is a convex set in a different coordinate system on ∂H^n, so the chain may be nonconvex or even disconnected and may contain relatively open sets of different dimension.

If $p \in Q$ and $z \in \partial H^n - p$ then the ray $\overrightarrow{pz}$ meets $\partial F(p)$ at a unique point which lies in at least one ridge $R(q,p)$. Thus $z \in S(q,p)$. We call each such q (there can only be finitely many) a *z-successor to p.*

For example, consider the Ford disc packing of section 4. The fan $F(\infty)$ consists of all points in H^2 whose Euclidean distance from the set of integers is at least one. Such fans were studied by Conway [C, p.28]. The ridges form a tree dual to the Farey tiling (Conway's "topograph") that is the prototype for arboreal group theory. The term z-successor in section 4 is consistent with our present usage since for $p=\infty$ a z-successor is an integer q with $|z-q| \leq 1/2$, so after $\overrightarrow{\infty z}$ exits $B(\infty)$ it meets no Ford disc until it touches $B(q)$. The shadow $S(q,\infty)$ is just the interval $[q-1/2, q+1/2]$.

Consider in a m-sphere $S^m, m \geq 1$, a nonempty, finite family of $(m-1)$-spheres. Take a connected component of the complement of these subspheres and form its closure C. We say that C is a conformal polyhedron or *conron.* For $m=1$, a conron is just a closed interval. For $m=2$, a conron is bounded by finitely many circles and it may be called a conformal polygon or *congon.* Using stereographic projection, we may identify S^m with $\hat{\mathbf{R}}^m = \partial H^n, n = m+1$, and apply the terms congon and conron to subsets of the boundary of hyperbolic n-space. In particular, each shadow $S(p,q)$ is a conron and each chain is a finite union of conrons.

We take the *standard horoball* B_n to be the horoball based at ∞ whose Euclidean distance from $\mathbf{R}^{n-1}$ is one. If g is an isometry with $g(\infty) \neq \infty$, there is a sphere in $\mathbf{R}^n$ centered at $g^{-1}\infty$ called the *isometric sphere of* g, such that the image of this sphere by g is a sphere of the same radius [R, p. 117]. Points of H^n on the isometric sphere are just those points equidistant from B_n and $g^{-1}B_n$. The isometric sphere of g bounds an open disc in ∂H^n that we denote D_g. The derivative of g is an expansion at points of D_g.

We introduce coordinate charts in which our shadows are polyhedra. We choose isometries $g(p)$ for each cusp point p so that $g(p)B(p) = B_n$ and $g(p') \in g(p)\Gamma$ for p' equivalent to p. For adjacent cusp points p,q we let $P(p,q)$ be the polyhedron $g(q)S(p,q)$ and $g(p,q) = g(p)g(q)^{-1}$. $P(p,q)$ is the shadow of $g(q)R(p,q)$ as lit from ∞, where points of $g(q)R(p,q)$ are equidistant from $B_n = g(q)B(q)$ and $g(p,q)^{-1}B_n = g(q)B(p)$. Thus

$g(q)R(p,q)$ lies on the isometric sphere of $g(p,q)$. It follows that $g(p,q)$ is an expansion at all points of $P(p,q)$. These charts and this expansive property will play a role in our proof of Theorem 11.

Section 6. Reduction theory for finite volume quotients of hyperbolic space

We will describe each pair (z,a) of distinct points in ∂H^n using a certain family of cusp points indexed by an interval of integers I. If $f: I \to X$, with X some topological space, we define the *upper limit* $\lim^+ f(i)$ to be $f(\max\ I)$ if I is bounded above and to be $\lim_{i\to+\infty}\ f(i)$ if I is unbounded above and the limit exists. We define the *lower limit* $\lim^- f(i)$ in a similar fashion.

We say that $p_i \in Q, i \in I$, is an *aimed sequence* if I contains at least two elements and the shadows $S(p_{i+1},p_i), i \in I, i+1 \in I$, have nonempty intersection. If z lies in this intersection then p_{i+1} is a z-successor to p_i whenever $i,\ i+1 \in I$ and we say that p_i is *aimed at* z. For $p \in Q$ and $z \in \partial H^n - \{p\}$ we let $B(z,p)$ denote the unique horoball based at z that touches $B(p)$ at one point.

We can now state two theorems.

Theorem 6. Convergence: *If $p_i, i \in I$, is an aimed sequence then the limits $\lim^+ p_i = z$ and $\lim^- p_i = a$ exist, p_i is aimed at z, and $a \neq z$. I is bounded above if and only if $z \in Q$ and I is bounded below if and only if $a \in Q$.*
Continuation: *If $z \in S(p_l,...,p_m),\ m < l$, then the sequence $p_m,...,p_l$ may be extended to a sequence $p_i, i \in I$, with $\lim^+ p_i = z$.*
Existence: *Given distinct hyperbolic boundary points z, a there is an aimed sequence with these limits.*
Uniqueness: *For almost all z, a, this aimed sequence is unique up to a shift by some integer l, in which I is replaced by $I' = I + l$ and p_i is replaced by $p'_i = p_{i-l}$.*
Monotonicity: *We have $p_i \neq z$ for $i \neq \max(I)$ and $B(z,p_i)$ is contained in the interior of $B(z,p_{i'})$ for $i' < i$. The distance from $B(z,p_i)$ to $\partial B(z,p_{i'})$ is bounded below by a positive constant depending only on D.*

Theorem 6 generalizes the results of section 3 concerning the Ford disc packing of H^2. When $p_0 = \infty$, the cusp points $p_1, p_2, ...$ are the least-remainder convergents to z and the cusp points $p_{-1}, p_{-2}, ...$ are the H-dual convergents to a.

Theorem 7. *For a set of $z \in \partial H^n$ of full measure and for any two aimed sequences $p_i, i \in I$, and $q_j, j \in J$, with $\lim^+ p_i = z = \lim^+ q_j$ there is an $l \in Z$ and $j_0 \in J$ so that $q_j = p_{j+l}$ for all $j \geq j_0$.*

That is, almost every boundary point determines uniquely the tail of the sequences aimed at it.

Theorems 6 and 7 are special cases of Theorems 1 and 2 in [F3] (in the notation of that paper, take $L = \partial H^n$ and use the standard shadow family) so we will only outline the proofs here. In proving theorem 6, we may suppose that ∞ belongs to every chain defined by a finite subinterval of I. We show that the Euclidean diameter h_i of $B(i)$ satisfies $h_{i+1} \geq ch_i$ for some constant $c > 1$. This proves Monotonicity and shows that $\lim^+ p_i = \infty$ when I is not bounded above. We also bound the ratios $(p_{i+1} - p_i)/(h_{i+1} - h_i)$. This shows that $\lim^- p_i$ exists when I is not bounded below, proving Convergence. Continuation is proven

using an inductive construction. Using a diagonal argument, Continuation and the ratio bound imply Existence.

It remains to prove Uniqueness. We show that there is a Γ-equivariant family of horoballs $B'(q) \subset B(q), q \in Q$, so that if $\overline{az}$ meets $B'(q)$ then $q = p_i$ for some $i \in I$. We say z is *generic* if it does not lie in any of the countably many null sets $S(q,p) \cap S(r,p), p,q,r \in Q, q \neq r$. If z is generic then each p_i has a unique z-successor and so p_i determines p_j for $j > i$. If z is generic and $\overline{az}$ meets $B'(q)$ for cusp points q arbitrarily close to a then (z,a) is represented by a unique aimed sequence, up to a shift. This set of pairs has full measure, so Theorem 6 follows.

For Theorem 7, suppose z is generic and some ray $\overrightarrow{az}$ meets $B'(q)$ for cusp points q arbitrarily close to z. We show that such z form a set of full measure and that for such a z the tail of p_i is determined by z.

Note that Theorems 6 and 7 do not even mention Γ. In fact they are true for any D-dense horoball packing of H^n. We now take the action of Γ into account and derive a reduction theory for Γ, as follows.

Fix an adjacent pair of cusp points p, q. For each interval I of integers containing 0 and 1 and each aimed sequence $p_i,\ i \in I$, with $p_0 = q$ and $p_1 = p$ we take the limits (z,a) as in Theorem 6. These pairs (z,a) form a compact subset $K(p,q)$ of Ω. The first coordinate projection of $K(p,q)$ equals $S(p,q)$.

For the Ford disc packing of section 4, we find $K(\infty, 0)$ is the closure of the set of H-reduced pairs from equation (7), that is

$$K(\infty,0) = [2,\infty] \times [r-1, r] \cup [\infty, -2] \times [-r, 1-r]. \tag{12}$$

Let A denote the set of all ordered pairs of adjacent cusp points. A consists of finitely many Γ-orbits, and each pair has a finite stabilizer. We let $K = \coprod K(p,q)$ and $S = \coprod S(p,q)$, where the disjoint unions are indexed by A. Our group Γ acts on K and on S, permuting the indices. We let B consist of all ordered triples of cusp points (p,q,r) such that $S(p,q) \cap S(q,r)$ is nonempty. We define a relation $\mathcal{P}^*$ on K that consists of all ordered pairs $((z,a),(z,a)) \in K(p,q) \times K(q,r)$ and a relation $\mathcal{G}^*$ on S that consists of all ordered pairs $(z,z) \in S(p,q) \times S(q,r)$, where $(p,q,r) \in B$. and $(z,a) \in K(p,q,r) = K(p,q) \cap K(q,r), z \in S(p,q) \cap S(q,r)$, respectively. These relations are the *Poincare correspondence* and the *Gauss correspondence*, respectively, associated to our horoball packing. These relations are preserved by Γ and so we obtain relations on K/Γ and on S/Γ that we denote by $\mathcal{P}$ and $\mathcal{G}$, respectively. These are called the *reduced Poincare correspondence* and the *reduced Gauss correspondence* associated to Γ and the chosen system of cusp neighborhoods. (Beware that our notation in [F3] for the Gauss correspondences is slightly different than that used here.)

We briefly discuss these names. We may identify each point of $(z,a) \in K(p,q)$ with the unit tangent vector $v = v(z,p,a)$ to the ray $\overrightarrow{az}$ at the point where this ray enters the horoball $B(z,p)$. The set of such v's is a compact subset of a transversal to the geodesic flow on SH^n. If $(z,a) \in K(p,q,r)$ then the Monotonicity part of Theorem 6 implies that this ray enters $B(z,q)$ before it enters $B(z,p)$, and we get a closed interval in a flowline for the geodesic flow on SH^n bounded by $v(z,p,a)$ and $v(z,q,a)$. The union of these closed intervals over all choices of (z,a) defines a sort of flowbox for each $(p,q,r) \in B$ (beware that, unlike in section 4, this flowbox may be noncompact: this only is due to orbits that meet one face but

not the other because they diverge to a cusp). The correspondence $\mathcal{P}^*$ describes the passage of flowlines from one transversal to another, so we name it for Poincare. The term Gauss correspondence, on the other hand, comes from the *Gauss map* on the interval $[0,1]$, that sends x to the fractional part of $1/x$. This map arises as the first factor of a certain Poincare correspondence. Indeed in Theorem 1, a shift by $l = 1$ defines a Poincare correspondence (as in section 1) and the first factor of this correspondence is the graph of the Gauss map, up to a coordinate change by $1/x$.

In the charts $g(p)$ given at the end of section 5, we can identify the Gauss correspondence $\mathcal{G}^*$ with the system of maps $P(q,r) \to \mathbf{R}^{n-1}/\Gamma_q = \cup P(p,q)$ induced by $g(q,r)$.

For the Ford disc packing of section 4 with $\Gamma = PSl(2,Z)$, Γ acts simply transitively on the set of adjacent pairs of cusp points. So we may identify K/Γ with $K(\infty,0)$. We find that $\mathcal{P}$ is the completion of the relation of Theorem 4 with $l = 1$.

The following theorem, a corollary of Theorem 6, is the Γ-analogue of Theorem 4.

Theorem 8. *The sets $K(p,q)$ cover Ω. Two points $(z,a) \in K(p,q)$ and $(z',a') \in K(p',q')$ are Γ-equivalent if their equivalence classes in K/Γ are related by an iterate of the reduced Poincare correspondence. Conversely, if they are not related by such an iterate then, for almost all (z,a), they are not Γ-equivalent.*

The Γ-analogue of Theorem 3 is the following corollary of Theorem 7.

Theorem 9. *The sets $S(p,q)$ cover ∂H^n. Two points $z \in S(p,q)$ and $z' \in S(p',q')$ are Γ-equivalent if their equivalence classes in S/Γ have the same image under an iterate of the reduced Gauss correspondence. Conversely, if they do not have the same images under any iterate of the reduced Gauss correspondence then, for almost all z, they are not Γ-equivalent.*

Theorems 9 and 8 constitute a reduction theory for points and pairs of distinct points in the boundary of hyperbolic space. It is a routine matter to translate Theorem 8 into a statement about the action of Γ on quadratic forms when n is 2 or 3, using equations (1) and (2). However, such a reformulation will only be of number theoretic interest in the case of a Bianchi group or similar group.

Section 7. The chain-finite case

If for each $(p,q) \in A$ there are only finitely many chains $S(p_1,p_0,...,p_m)$, with $(p_1,p_0) = (p,q)$, we say that the horoball packing $B(q), q \in Q$, or the underlying system of cusp neighborhoods, is *chain-finite*. In this case we will derive a reduction theory based on a finite number of rectangles, as in Theorem 5, but an infinite number of transitions. In the next section we will apply these results to Bianchi groups.

Suppose each shadow $S(p,q), (p,q) \in A$, is the union of a finite family $\mathcal{C}(p,q)$ of nonoverlapping conrons $S_j(p,q), j \in J(p,q)$, and the stabilizer of each pair $(p,q) \in A$ permutes the conrons in $\mathcal{C}(p,q)$. We call $\mathcal{C}$ a Γ-*partition* of S. We will say that $\mathcal{C}$ is *consistent with the Gauss correspondence* or $\mathcal{G}$-*consistent* if for each $(p,q,r) \in B$ and each $S_{j'}(q,r)$, $j' \in J(q,r)$, that overlaps $S(p,q)$, the intersection $S_{j'}(q,r) \cap S(p,q)$ is a union of sets $S_j(p,q)$, for j in some subset of $J(p,q)$.

For motivation, consider the Ford disc packing of section 3 and let $\Gamma = PGl(2,\mathbf{Z})$. By equation (12), the shadow $S(\infty,0) = [2,-2]$ is the complement of the open interval

$(-2,2) \subset \hat{\mathbf{R}}$, and for any integer n one has $S(n,\infty) = [n-1/2, n+1/2]$. If we partition $S(\infty,0)$ into the intervals $[2,\infty]$ and $[\infty,-2]$ and partition the other shadows equivariantly, we get a Γ-partition of S. As $S(n,\infty)$ is partitioned into $[n-1/2,n]$ and $[n,n+1/2]$, we see this partition is $\mathcal{G}$-consistent. Indeed each $S(n,\infty) \cap \pm[2,\infty]$ is the union of elements of $\mathcal{C}(n,\infty)$. One can deduce from $\mathcal{G}$-consistency that the only chains $S(\infty,0,...,p_m)$ are the three intervals $[2,-2]$, $[2,\infty]$, and $[\infty,-2]$, from which it follows that the Ford disc packing is chain-finite.

Suppose we are given a chain-finite horoball packing and $(q,r) \in A$. Suppose there are N distinct chains with initial indices q,r and we enumerate them as $S(q,r,...) = C_u, u = 1,...,N$. We remove from int $S(q,r)$ the boundaries $\partial C_u, u = 1,...,N$, to get an open set $O(q,r) \subset \partial H^n$. Then $O(q,r)$ has finitely many connected components, each bounded by $(n-1)$-spheres/planes, that we denote $O_j(q,r), j \in J(q,r)$. We choose these index sets $J(q,r)$, $(q,r) \in A$, to be disjoint. $O(q,r)$ is dense in $S(q,r)$ and has full measure. We let $Z_j(q,r)$ be the closure of $O_j(q,r)$, so we get a finite family of nonoverlapping conrons that we call $\mathcal{Z}(q,r)$. The stabilizer of (q,r) permutes the chains $C_1,...,C_N$, and so it permutes these conrons. Thus in the chain-finite case, we have defined a Γ-partition of S that we call the *standard partition* $\mathcal{Z}$ associated to our chain-finite horoball packing $B(q), q \in Q$.

Lemma 2. *The standard partition is $\mathcal{G}$-consistent.*

For each component of $O_j(q,r) \cap$ int $S(p,q)$ is one of the sets $O_{j'}(p,q)$, $j' \in J(p,q)$. Thus $Z_j(q,r) \cap S(p,q)$ is the union of these $Z_{j'}(p,q)$, so Lemma 2 is proved.

For any $\mathcal{G}$-consistent Γ-partition $\mathcal{C}(p,q)$, every chain $S(p,q,...)$ is a union of congons in $\mathcal{C}(p,q)$. Thus our standard partition is the coarsest such partition in the chain-finite case.

Suppose our horoball packing is chain-finite and $(q,r) \in A$. Suppose that $p_i, i \in I$, is aimed at z, I is bounded above by 1, and $(p_1,p_0) = (q,r)$. Let $a = \lim^- p_i$. The chains $C_m = S(p_1,p_0,...,p_m), m \leq 0$, are nested and nonempty, so they stabilize at some chain C_{min}. Suppose $z' \in C_{min}$. By the Continuation part of Theorem 6, one may extend the sequence $p_m,...,p_0,p_1$ of indices in C_{min} to a sequence $p_i, i \in I'$, where I' is bounded below by m and $\lim^+ p_i = z'$. Note that the sequence $p_i, i \in I \cup I'$, is aimed at z'. We have $\lim^+ p_i = z'$ and $\lim^- p_i = a$. Thus $C_{min} \times \{a\} \subset K(q,r)$.

For a fixed chain C, each aimed sequence with $C_{min} = C$ gives a slice $C \times \{a\} \subset K(q,r)$ for some $a \in \Omega$. The closure of the union of these slices is a rectangle in $K(q,r)$ whose first factor is C. This expresses $K(q,r)$ as a finite union of rectangles whose first factor is a chain. These rectangles overlap. However $K(q,r)$ is a union of nonoverlapping rectangles of the form $K_j(q,r) = Z_j(q,r) \times A_j(q,r), j \in J(q,r)$. This follows from the section of [F3] on Markov partitions for chain-finite packings.

Suppose that $Z_{j'}(p,q) \subset Z_j(q,r)$, as in Lemma 2. Then we see that $A_j(q,r) \subset A_{j'}(p,q)$. For if $a \in A_j(q,r)$ then some aimed sequence $p_i, i \in I$, has $(p_2,p_1,p_0) = (p,q,r)$, $\lim^+ p_i = z \in Z_{j'}(p,q)$ and $\lim^- p_i = a$. But then the sequence $p'_i = p_{i+1}$ is aimed with $\lim^- p'_i = a$, $(p'_1,p'_0) = (p,q)$, and $\lim^+ p'_i = z$, so $a \in A_{j'}(p,q)$.

We let $\mathcal{S}^* = \cup J(p,q), (p,q) \in A$, be our set of symbols. For $s = j \in J(q,r)$, we let $Z_s = Z_j(q,r)$, $A_s = A_j(q,r)$, and $R_s = Z_s \times A_s$. We let $\mathcal{T}^*$ be the set of all pairs $t = (j,j')$ as in the last paragraph, for $(p,q,r) \in B$, $j \in J(q,r)$, $j' \in J(p,q)$ and $Z_{j'}(p,q) \subset Z_j(q,r)$. We let $i^*(t) = j$, $f^*(t) = j'$. Then the last paragraph shows that equation (8) holds. Moreover,

if one examines the discussion leading up to (9) and uses the Uniqueness part of Theorem 6 and the fact that the conron boundaries have measure zero, one finds that (9) holds as well.

Now we are ready to factor out by Γ to get a reduction theory for X. We assume, for simplicity, that Γ acts freely on $\mathcal{S}^*$, as will be the case in many of our examples below. Then we follow the procedure we used to derive Theorem 5 from (8) and (9). We take a finite set of representatives $\mathcal{S}$ for the action of Γ on $\mathcal{S}^*$. We let $\mathcal{T}$ consist of all $t \in \mathcal{T}^*$ with $i(t) \in \mathcal{S}$ and we choose $\gamma(t)$ so that $\gamma(t)f^*(t) \in \mathcal{S}$. We let $i, f : \mathcal{T} \to \mathcal{S},\ i(t) = i^*(t),\ f(t) = \gamma(t)f^*(t)$. We obtain the following rectangular version of reduction.

Theorem 10. *Suppose X has finite volume and is noncompact. Suppose there is a Γ-equivariant family of horoballs $B(q), q \in Q$, that is chain-finite. Then there is a finite set $\mathcal{S}$ of symbols and a countable set $\mathcal{T}$ of transitions, rectangles $R_s = Z_s \times A_s \subset \Omega, s \in \mathcal{S}$, and maps $i, f : \mathcal{T} \to \mathcal{S},\ \gamma : \mathcal{T} \to \Gamma$ so that for $t \in \mathcal{T}, i = i(t), f = f(t), \gamma = \gamma(t)$ we have*

$$Z_f \subset \gamma Z_i \text{ and } \gamma A_i \subset A_f. \tag{13}$$

A relation in $R_i \times R_f$ is defined by the bijection

$$\gamma :\ \gamma^{-1} Z_f \times A_i \to Z_f \times \gamma A_i.$$

Let $\mathcal{P}$ be the relation on $U = \cup R_s$ that is the union of these relations over all $t \in \mathcal{T}$. The iterates of $\mathcal{P}$ give Γ-equivalent points in U and, conversely, for almost all $(z, a) \in U$,

$$\text{all points of } U \text{ } \Gamma\text{-equivalent to } (z, a) \text{ arise by iterating } \mathcal{P}. \tag{14}$$

We consider the family of maps $\gamma(t)^{-1} : Z_f(t) \to Z_i(t)$. This is an infinite family of conformal or anticonformal transformations on a finite system of conrons. The union of the graphs of these maps will be called the *inverse Gauss correspondence.* This is analogous to the system of inverse branches to the Gauss map on the $[0, 1]$ given by the family of maps $(x + n)^{-1},\ n = 1, 2, \ldots$. From Theorem 7, and the fact that the boundary of any conron is a null set, we obtain

Corollary 1. *For almost all points $z \in \cup Z_s,\ s \in \mathcal{S}$, the iterates of the inverse Gauss correspondence give all points in $\cup Z_s$ that are Γ-equivalent to z.*

The advantage of this inverse Gauss correspondence over the reduced Gauss correspondence of section 6 is that its branches are defined on only finitely many sets. This makes it possible to use the inverse Gauss correspondence to calculate dynamical zeta functions for the geodesic flow. We refer to [M, F2] for a discussion of the use of the inverse Gauss map to calculate zeta functions for $PSl(2, \mathbf{Z})$ and other triangle groups. Using the results of the next section, we will have a corresponding tool for Bianchi groups.

Section 8. Rectangles for Bianchi groups

We will find an analogue of equation (12) for each Bianchi group. That is, we will choose a suitable horoball packing and show that each $K(p, q)$ is a finite union of rectangles. In this sense the reduction theory of section 6 resembles the reduction theory for compact quotients very closely when Γ is a Bianchi group.

The key point is that we can choose our horoball packing to be chain-finite. Proposition 1 below is the technical result needed to prove this. We first need a general lemma. Recall that two subgroups of some given group are *commensurable* if their intersection has finite index in each subgroup.

Lemma 3. *Let G be a Lie group, $\iota : G \to G$ an involution, and F the subgroup of fixed points of ι. For any discrete subgroup $\Gamma \subset G$ such that $\iota(\Gamma)$ is commensurable with Γ, the image of Γ in G/F is discrete.*

To see this, let $\Gamma_0 = \Gamma \cap \iota(\Gamma)$ and let $\rho : G \to G/F$ be the natural projection. The coincidence sets $C(\iota, \gamma) = \{g \in G : \iota(g) = \gamma g\}, \gamma \in \Gamma$, form a disjoint, locally finite family of closed sets in G. One finds for $\delta \in \Gamma_0$ that $\delta(C(\iota, \gamma)) = C(\iota, \gamma')$, where $\gamma' = \iota(\delta)\gamma\delta^{-1} \in \Gamma$. So these sets are permuted by the left action of Γ_0 on G. Taking γ to be the identity element $e \in G$, we see that $C(\iota, e) = F$. So $\Gamma_0 F$ is closed in G and has F as a relatively open subset. Thus $\rho(\Gamma_0)$ is discrete. As Γ_0 has finite index in Γ, we find that $\rho(\Gamma)$ is also discrete and Lemma 3 is proven.

The *commensurator* (or commensurability subgroup) $Comm_G(\Gamma)$ consists of all $g \in G$ such that $g\Gamma g^{-1}$ is commensurable with Γ. Any subgroup commensurable with Γ lies in the commensurator.

Say G is the group of isometries of H^n and suppose $R \in Comm_G(\Gamma)$ is elliptic of order two. Then we may take ι to be the inner automorphism by R and apply Lemma 3. Clearly F is the stabilizer in G of the hyperbolic subspace $Fix(R)$ of fixed points of R. Let d be the dimension of $Fix(R)$ so $0 \leq d < n$. The homogeneous space G/F may be identified with the Grassmannian of hyperbolic subspaces of dimension d, with basepoint $Fix(R)$. When $d > 0$, each of these hyperbolic subspaces is determined by its intersection with ∂H^n, which is a $(d-1)$-sphere or $(d-1)$-plane. This identifies G/F with the space of such spheres and planes in ∂H^n. Lemma 3 asserts that the orbit of the basepoint under Γ is discrete.

When H^n/Γ is noncompact and has finite volume the commensurator preserves the set Q of cusp points. We now take $n = 3$ and fix an imaginary quadratic field k. For the Bianchi group B_k we have $Q = \hat{k}$, whose stabilizer in G contains the semidirect product Σ_k of $PGl(2,k)$ with $\{z, \overline{z}\}$. Since $PGl(2,k)$ is triply transitive on $\hat{k}$, and the only elements of G that fix $0, 1$, and ∞ are $\{z, \overline{z}\}$, we see that the stabilizer of $\hat{k}$ is Σ_k. Since both $PGl(2,k)$ and $\{z, \overline{z}\}$ lie in the commensurator of B_k, we see that this commensurator is also Σ_k. It follows that $Comm_G(\Gamma) = \Sigma_k$ for any Γ commensurable with B_k.

We say a line/circle L is *k-rational* if $L = M(\mathbf{R})$ where $M \in Gl(2,k)$. For such an L, $R = M \circ \overline{z} \circ M^{-1} \in Comm_G(\Gamma)$. So Lemma 3 implies that the set of lines/circles $\gamma(L), \gamma \in \Gamma$, is discrete.

Let β be a positive constant and C a compact subset of $\mathbf{C}$. The set of lines/circles that meet C and have curvature at most β is compact. So we find

Proposition 1. *Let Γ be a group of isometries of H^3 commensurable with B_k. For any isometry g and any k-rational line/circle L, only finitely many of the images $g\gamma(L), \gamma \in \Gamma$, have bounded curvature and pass through a given bounded region in $\mathbf{C}$.*

We originally proved this proposition by a direct method that may be useful in computations. We briefly sketch this method in the case where $g = e, L = \mathbf{R}$, and Γ is given by matrices with coefficients in O_k. By commensurability, one may arrange that each of these matrices has determinant in some finite set of nonzero elements of O_k. If $\gamma(L)$ is a line, one can calculate its slope and find the point ζ on it nearest to the origin. One finds that the slope is given by a bounded element of O_k, so only finitely many slopes are possible. For a given slope, a certain multiple of ζ lies in O_k and ζ is bounded. Thus only finitely many

lines occur. The case when $\gamma(L)$ is a circle behaves similarly. One finds that the curvature takes only finitely many values. For each value of the curvature, however, the center of the circle is bounded and a certain multiple of the center lies in O_k. Thus only finitely many circles occur, and the proposition follows.

To exploit Proposition 1, we will need to arrange that various shadows and chains be bounded by k-rational lines and circles and that these circles have bounded curvature. These concerns prompt the next two lemmas.

We now define a *norm* $\nu : \Sigma_k \to \mathbf{Q}^+/(\mathbf{Q}^+)^2$. For an element g of the form $\gamma(z)$ or $\gamma(\overline{z})$ for $\gamma \in Gl(2,k)$, $\nu(g)$ is the class modulo rational squares of the norm of $\det(\gamma)$. One can easily see that ν is well-defined and a group homomorphism. If Γ is commensurable with B_k then it is finitely generated and so $\nu(\Gamma)$ is a finite abelian group of exponent 2.

We say a horoball in H^3 is *k-rational* if it is based at a point of k and its Euclidean diameter is rational or it is based at infinity and its height (i.e. its Euclidean distance to $\mathbf{C}$) is rational. It is easily checked that the image of a k-rational horoball by $g \in Comm_G(B_k)$ is k-rational if and only if $\nu(g) = 1$. One can verify this directly if the horoball and its image are based at ∞, since $az + b$ and $a\overline{z} + b$ lie in the kernel of ν if and only if $|a|$ is rational. To deduce the general case we use the transformations $\gamma(z) = (q - z)^{-1}$, $q \in k$. It is enough to check that $\nu(\gamma) = 1$ and that γ sends rational horoballs based at q to rational horoballs based at ∞.

We now suppose Γ is commensurable with B_k and $\nu(\Gamma) = 1$. Shrinking each horoball in our horoball packing slightly, we may arrange that each horoball $B(q)$ is rational.

Lemma 4. *The line or circle L containing a boundary arc $S(p,q) \cap S(p',q)$ is k-rational.*

We may reduce to the case $q = \infty$ and $B(\infty)$ of height 1, by a change of coordinates $a(q-z)^{-1}$, $a \in \mathbf{Q}$, when $q \in k$. Say that $B(p)$ has Euclidean diameter $t \leq 1$. Then $R(p,q)$ lies on the Euclidean sphere with center p and radius $\sqrt{t}$. Choosing t' the Euclidean diameter of $B(p')$ we find that these spheres meet $\mathbf{C}$ in circles $|z-p|^2 = t$ and $|z-p'|^2 = t'$. If we simplify and subtract these equations, and use that $p, p' \in k$ and $t, t' \in \mathbf{Q}$, we get an equation of the form $\Re(za) = b, a \in k, a \neq 0, b \in \mathbf{Q}$ which defines a k-rational line L. As L passes through the intersection of both spheres and $\mathbf{C}$, it is the projection from q of the intersection of the ridges $R(p,q)$ and $R(p',q)$. This proves Lemma 4.

The following lemma will be used to bound certain curvatures. We take g an isometry of H^3 that does not fix ∞ and we define the disc D_g as at the end of section 5.

Lemma 5. *Let L be a circle or line in $\mathbf{C}$ that meets D_g, and let $\kappa \geq 0$ be the curvature of L. Let $L' = g(L)$ and let κ' be its curvature. If $\kappa \leq \kappa'$ then $\kappa' \leq 2d/(r^2 - d^2)$, where r is the radius of D_g and where $d < r$ is the distance from L to the center of D_g.*

For the proof, we may compose g on both sides with Euclidean isometries to reduce to the case where g is the inversion $g(z) = r^2/\overline{z}$ and L, L' are both perpendicular to $\mathbf{R}$. By continuity, we may suppose that L and L' are circles. Say that L crosses $\mathbf{R}$ at x and y with $|x| \leq |y|$. Then $|x| = d < r$.

We simplify the inequality $\kappa = 2/|x - y| \leq \kappa' = 2/|g(x) - g(y)| = 2|xy|/(r^2|x - y|)$ to find $|xy| \geq r^2$. This implies in turn that $|xy^{-1}| \leq d^2/r^2$ and $|1 - xy^{-1}| \geq 1 - (d^2/r^2)$. But then $\kappa' = 2|x|/(r^2|1 - xy^{-1}|) \leq 2d/(r^2 - d^2)$, and Lemma 5 is proved.

We can now prove a rectangular version of reduction for Bianchi groups.

Theorem 11. *Let k be an imaginary quadratic field and O_k the ring of integers in k. Let Γ be commensurable with the Bianchi group $B_k = PSl(2, O_k)$ and suppose $\nu(\Gamma) = 1$. There is a Γ-invariant packing $B(q), q \in \hat{k}$, of H^3 by rational horoballs. Each such packing is chain-finite and gives a reduction theory for Γ based on a finite number of rectangles.*

For each cusp point p, we may choose charts $g(p)$ as in section 5 with $g(p) \in PGl(2, k)$. Since Γ has only finitely many orbits on Q, the transformations $g(p, q) = g(p)g(q)^{-1},\ p, q \in Q$, lie in a finite number of double cosets of Γ in Σ_k, hence in a finite number of left cosets of groups commensurable with B_k.

Suppose $C = S(p_1, p_0, ..., p_m),\ m \leq 0$, is a chain and define $f_0, ..., f_m \in PGl(2, k)$ by $f_i = g(p_0, p_i)$. For $i = m, ..., 0$, let $P_i = P(p_{i+1}, p_i)$. Then a fixed boundary arc of $g(p_0)C$ lies for some $i = m, ..., 0$, in the image by f_i of a line L_i that bounds P_i. We will find a bound for the curvature of this arc. Proposition 1 then shows that there are only finitely many circles of the form $f_i(L_i)$ since they have bounded curvature and pass through P_0. Hence there are only finitely many choices for $g(p_0)C$. This shows chain-finiteness and Theorem 10 gives the desired rectangles.

Let $L_{i+1}, ..., L_0$ be the lines and circles obtained from L_i by successively applying the transformations $g_j = g(p_{j+1}, p_j)$ for $j = i, ..., -1$. Suppose that the greatest curvature in this sequence occurs for $L_{j+1},\ i \leq j \leq -1$. In Lemma 5, we take $L = L_j$ and $g = g_j$. As shown in section 5, $P_j \subset D_g$. Since L meets $P_j,\ L$ crosses D_g. As $L' = L_{j+1}$, we have $\kappa \leq \kappa'$. Thus Lemma 5 gives a bound on κ', hence on the curvatures of $L_{i+1}, ..., L_0$.

Because Γ acts with only finitely many orbits on A, the transformations $g(p, q), (p, q) \in A$, take only finitely many values. Thus in Lemma 5, r takes only finitely many possible values and the ratio d/r is uniformly bounded by some constant less than one. Suppose ρ is the minimum value of r and $\mu < 1$ is the supremum of the ratios d/ρ_g. Then Lemma 5 gives the uniform bound $2/(\rho\ (\mu^{-1} - \mu))$ for the curvature of a boundary arc of $g(p_0)C$. This finishes the proof of Theorem 11.

The preceding proof relies on the fact that each transformation $g(p, q)$ in the local coordinate form of the Gauss correspondence is an expansion on the polygon $P(p, q)$. It is this expansion that bounds the curvature.

The method of proof used in this section yields the following generalization of Theorem 11.

Theorem 12. *Let G be the isometry group of H^n and Γ a discrete subgroup such that $X = H^n/\Gamma$ has finite volume but is not compact. Suppose that cusp neighborhoods for X can be chosen so that for each adjacent pair of cusp points p, q, each smooth component of $\partial S(p, q)$ is fixed by a hyperbolic reflection $R \in Comm_G(\Gamma)$. Then this system of cusp neighborhoods is chain-finite and gives a reduction theory for Γ based on a finite number of rectangles.*

For the proof, note that Lemma 5 generalizes to any $n \geq 2$. To include the case $n = 2$, we take the curvature of an interval to be twice the reciprocal of its diameter.

Section 9. Examples and open problems

Theorem 11 applies to various Γ of number-theoretic interest. Suppose J is an ideal in O_k. Consider transformations $(az+b)/(cz+d), a, d \in O_k, c \in J, b \in J^{-1}, ad-bc=1$. These form a group $PSl(O_k, J)$ commensurable to B_k. Since it is a subgroup of $PSl(2,k)$, it lies in the kernel of ν. Thus Theorem 11 applies to $PSl(O_k, J)$.

Following [EGM, chapter 7, section 4], we construct a discrete subgroup H_k of $PGl(2,k)$ and describe its properties. H_k consists of isometries $g(z) = (az+b)/(cz+d)$, where $a, b, c, d \in O_k$, and the nonzero principal ideal $< ad-bc >$ is the square of the ideal $< a, b, c, d >$. Since the norm of $ad-bc$ is a square in k, $|ad-bc|$ is rational and so $\nu(g) = 1$. H_k is maximal among the discrete subgroups of $PGl(2, \mathbf{C})$ containing B_k. Let U_k be the group of units in O_k and let I_k be the ideal class group of O_k. Let $I_k^{(2)}$ be the subgroup of I_k consisting of those ideal classes whose square is principal. Considering the ideal class of $< a, b, c, d >$, one can describe H_k/B_k as an extension of U_k/U_k^2 by $I_k^{(2)}$ and prove its order is 2^t, where t is the number of prime factors in the discriminant of k.

Let the *extended Bianchi group* E_k be the semidirect product of H_k by $\{z, \overline{z}\}$. Clearly $\nu(E_k) = 1$. Knowing that Σ_k is the commensurator of B_k it follows that E_k is maximal among discrete isometry groups of H^3 containing B_k.

We can interpret E_k as the semidirect product of B_k and the isometry group $Isom(X_k)$ of the orbifold $X_k = H^3/B_k$. For by [R, corollary 2 to Theorem 12.6.3], this isometry group has finite order. Lifting these isometries to H^3 in all possible ways gives a finite extension of B_k, hence a subgroup of E_k. Conversely, since B_k is normal in E_k we see that elements of E_k define isometries of X_k. Thus $E_k/B_k \cong Isom(X_k)$, proving our assertion. This proof also shows that E_k is just the normalizer of B_k in the isometry group of H^3.

When calculating examples, the reduction theory for H_k or E_k uses fewer symbols and transitions than that for the Bianchi group B_k itself. It is not hard, on the other hand, to recover the results for B_k from those for H_k or E_k. We will restrict ourselves to some k of small discriminant.

The cusp point a/b, $a, b \in O_k$, determines the ideal class of $< a, b >$. This gives a well-known bijection between the B_k-equivalence classes of cusp points and I_k.

Suppose k has a Euclidean algorithm. Then X_k has only one end and so all cusp neighborhoods define the same fan tiling. The cusp points adjacent to ∞ are just the elements of O_k, and all adjacent pairs are equivalent to $(0, \infty)$. The shadow $S(0, \infty)$ is just the Dirichlet domain D for the translation group O acting on $\mathbf{C}$, that is the polygon of points of $\mathbf{C}$ closer to 0 than to any other point of O_k. To calculate (or approximate) the standard partition of D for B_k, we apply to all lines through ∂D all elements $g \in B_k$ that take 0 to ∞. Some of these lines/circles cross D, and we call these curves of generation one. To each of these curves we again apply these transformations g, leading perhaps to new curves that cross D that we say are of generation two. This process may lead to new curves for a number of generations, but Theorem 10 implies that this system of curves must stabilize. These curves divide D into finitely many congons. These congons form a refinement of the standard partition (in our calculations, we typically find that they give the standard partition itself).

If k has no Euclidean algorithm then more than one polygon is needed. In this case,

one sets up the charts $g(p)$ as at the end of section 5 and uses the local coordinate version $g(p,q)$ of the Gauss correspondence to transform curves in the polygon $P(p,q)$. We may call the bounding lines curves of generation zero, their images under $\gamma g(p,q),\ \gamma \in (B_k)_\infty$, that cross one of our polygons are curves of generation one, etc. It is convenient to take $h_k = |I_k|$ distinct copies of the upper halfspace model, one for each equivalence class of cusp points, and tile the ith copy by the polygons $P(p,q(i))$, where $q(i), i = 1, ..., h_k$, are representatives for the orbits of B_k on $\hat{k}$. Then each of these polygons is equipped with an expanding mapping with values in one of the h_k copies of $\hat{\mathbf{C}}$. Applying these mappings repeatedly, one eventually finds a refinement of the standard partition of these polygons. We call the sets in this partition *tiles* for a curvilinear tiling that refines our initial tiling by polygons. The crucial property is $\mathcal{G}$-consistency: the image of a tile in $P(p,q)$ by $g(p,q)$ is, except possibly for ∞, a union of tiles.

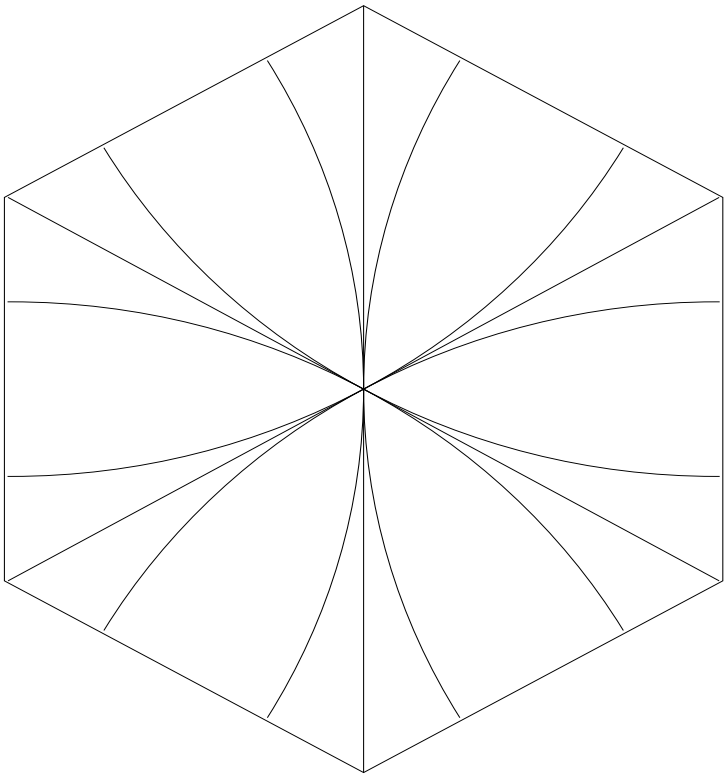

Figure 4. The standard partition for the Eisenstein field

We first take $k = \mathbf{Q}(\sqrt{-3})$, the Eisenstein field. Then D is a regular hexagon with vertices $z/2$, where $z^6 = -1$. We apply the inversion $1/\overline{z}$ to the righthand edge $\Re(z) = 1/2$ of D and obtain the circle $|z-1| = 1$. Circles equivalent to it by $U_k \subset (H_k)_\infty$ form the first generation of curves. Applying the inversion once more and translating back leads to a second generation of curves, namely the three lines through the origin satisfying $\Re(z^3) = 0$. Together these circular arcs and segments yield the standard partition of D into 18 regions, as drawn in Figure 4.

The stabilizer U_k of D in H_k acts freely on these regions with only three orbits. We may select as representative regions the three regions with $|\arg(z)| \leq \pi/6$. This leads to a reduction theory for $H_k = PGl(2,k)$ with three rectangles, or one for B_k with six rectangles. A reduction theory for E_k with two rectangles can be found from that for H_k by refining the standard partition by adding the three segments $\Im(z^3) = 0$. This yields 24 regions and we may select as representative regions the two regions for which $0 \leq \arg(z) \leq \pi/6$.

For $k = \mathbf{Q}(i)$, the Gauss field, the results are quite similar. Now D is the unit square centered at 0 and the circle $|z-1| = 1$ has 8 images that cross D, namely the unit circles centered at $\pm 1, \pm i, \pm 1 \pm i$. These 8 circles divide D into 12 congons, forming the standard partition of D, as drawn in Figure 5.

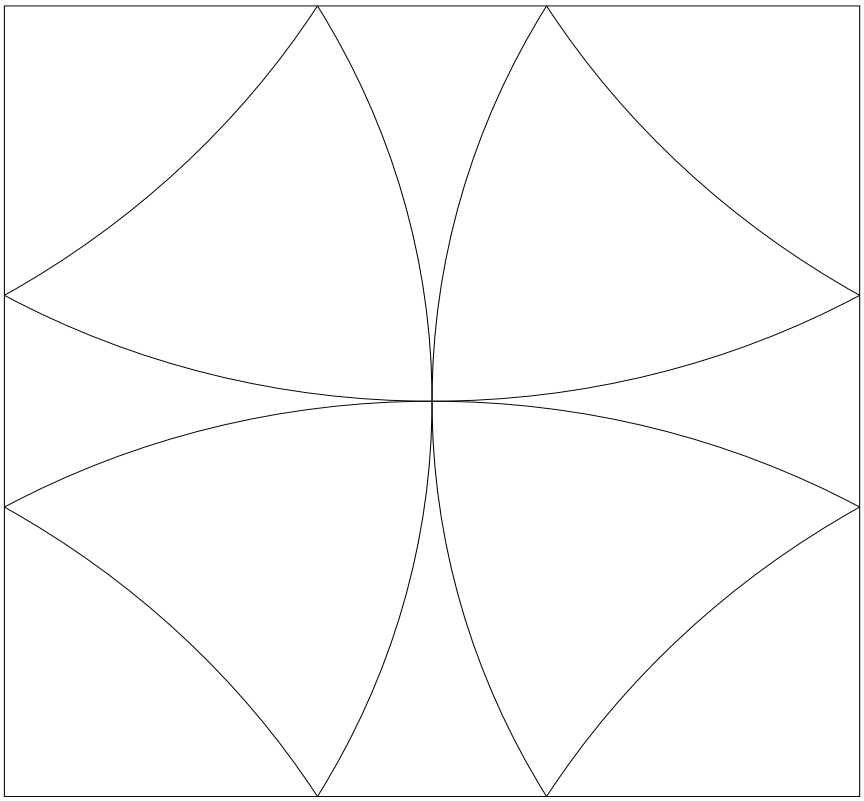

Figure 5. The standard partition for the Gauss field

The stabilizer of D in $H_k = PGl(2,k)$ is U_k, and it acts freely on these regions with only three orbits. We may take as representative regions the convex region in D defined by

$$|z-1-i| \leq 1$$

and two nonconvex regions. One of these resembles a shield

$$|z-1| \leq 1,\ |z-i| \leq 1,\ |z-1-i| \geq 1$$

and the other resembles a thorn

$$|z-1| \leq 1,\ |z-i| \geq 1,\ |z+i| \geq 1.$$

This leads to a reduction theory for H_k with three rectangles, or one for B_k with six rectangles. A reduction theory for E_k with three rectangles can be found from that for H_k by just bisecting our three regions by intersecting them with the sector $0 \leq \arg(z) \leq \pi/4$.

In [H1], Hurwitz studied least-remainder continued fractions over the Gauss and Eisenstein fields. From his calculations it is clear that the sequence of terms in a least-remainder continued fraction satisfies constraints involving arbitrarily long sequences of terms. This is unlike Hurwitz's theory over $\mathbf{Z}$ where, as noted in section 2, the only constraint involves consecutive pairs of terms. This difficulty is resolved by our use of the standard partition. We cannot change the sequences of terms, but we add a bit of information to each term, namely which congon in the standard partition of D contains the remainder. This restores $\mathcal{G}$-consistency and leads to a better description of these continued fractions.

In all cases with k Euclidean, our results can be described using least-remainder continued fraction expansions and the dual continued fraction expansions, both with terms in O_k. For other k, however, the continued fraction description breaks down since more than one remainder region must be used.

For a non-Euclidean example, let $k = \mathbf{Q}(\sqrt{-15})$ and let $s = (1 + \sqrt{-15})/4$. We have $s\overline{s} = 1, s + \overline{s} = 1/2$ or, equivalently, $s \cdot s = 1, s \cdot 1 = 1/4$. Since U_k and I_k have order two, we find that H_k/B_k is of order four. Generators can be taken to be $-z$, from the term $U_k/U_k^2 = U_k$, and the involution $R(z) = s + s/(2z - 2s)$, from the term $I_k^{(2)} = I_k$. O_k is generated by 1 and $2s$ and the stabilizer $(H_k)_\infty$ is the semidirect product of O_k and $\pm z$. The action of H_k on cusp points is transitive, so we may again use ∞ as our sole cusp point representative for this group. The set of highest points of orbits of B_k is known [S], and this set does not change when we pass to $PGl(2, O_k)$. Taking into account the action of $R(z)$ on X_k, one finds the following description of the highest points of the orbits of H_k, that is of the fan $F(\infty)$ for H_k

$$|z - m|^2 \geq 1, |z - m - s|^2 \geq 1/2, |z - m + \overline{s}|^2 \geq 1/2$$

for all $m \in O_k$. It follows that the action of H_k on adjacent pairs has three orbits, represented by $(0, \infty), (s, \infty)$, and $(-\overline{s}, \infty)$. The corresponding shadows are a hexagon $H = S(0, \infty)$ defined by

$$|\Re(z)| \leq 1/2, \ |z \cdot s| \leq 3/4, \ |z \cdot \overline{s}| \leq 3/4$$

and two parallelograms that are mirror images under reflection in the imaginary axis. One of these is $P = S(s, \infty)$, given by

$$|\Re(z - s)| \leq 1/4, \ |(z - s) \cdot 1| \leq 1.$$

Each of these pairs/shadows is stabilized in $(H_k)_\infty$ by a subgroup of order two generated by an element of the form $2c - z$, where c is the center of the shadow. Under the action of $(E_k)_\infty$, the two parallelograms become equivalent and the stabilizer of the hexagon increases to a group of order four $\{\pm z, \pm\overline{z}\}$. We may choose our charts so that $g(\infty) = id, g(s) = R$, and $g(0) = T$, where $T(z) = 1/\overline{z}$. So we use the inversion $T(z)$ on H and the involution $R(z)$ on P.

We find in this example that three generations of curves are needed. We can enumerate the curves as follows. We denote the line through a point b perpendicular to the nonzero

vector a by $a^\perp + b$. We denote the circle with center c and radius r by $c//r$. We will only write one circle/line from each orbit under $(E_k)_\infty$. Then the curves occurring in the boundary of the standard partition are the following four lines

$$1^\perp + 1/2,\ 1^\perp,\ 1^\perp + 1/4,\ s^\perp$$

and the following five circles

$$0//1,\ 0//2,\ (2s/3)//(2/3),\ s//1,\ s//(1/2).$$

For example, the line $L = s^\perp + 3$ contains the common boundary of H and P. Applying R to this line gives $s//1$. Half of this circle crosses $P-1, H-1, H$, and $H+1$. We obtain a second generation curve $0//2$ by transforming from $H+1$. Applying T to a circle crossing H equivalent to $0//2$ gives a vertical line $1^\perp + 1/4$ in the third generation.

This standard partition is permuted freely by the pair stabilizers in H_k. We have not enumerated the regions in this partition as this seems a task best left to a computer.

For an extensive number of choices of k, the fundamental domain for B_k on H^3 has been tabulated by Robert Riley, whose methods are explained in [Ri] along with several examples. With such a domain, and a knowledge of I_k, one can find the standard partition by a mechanical procedure.

We conclude with some problems that bear further inquiry. One is the nature of the dual factors A_s for the factors Z_s of a standard partition. Our calculations for $k = \mathbf{Q}(i)$ suggest that these factors have fractal boundary, much like the Koch snowflake curve. We do not know the fractal dimension of these boundaries. To those familiar with Markov partitions for Anosov flows on compact manifolds, fractal boundaries are what one expects for factors of dimension two or more [Ca] and it is the smoothness of the conron factors Z_s in Theorem 10 that is surprising. [F3] gives a recursive procedure for approximating the dual factors A_s.

It is worth exploring other reduction theories to try to avoid fractals. The thesis of J. Hurwitz [H] (see also [D, pp. 200-201]) gives a reduction theory for a finite index subgroup of the $B_k,\ k = \mathbf{Q}(i)$, based on a special class of continued fractions. In this theory, both factors Z_s and A_s are congons. We are developing a geometric method, suggested by this example, that applies to a certain geometrically defined class of finite volume, noncompact quotients of H^n. At present we know this class includes a subgroup of B_k for $k = \mathbf{Q}(\sqrt{-d}),\ d = 1, 2, 3$.

It is feasible to calculate dynamical zeta functions for Bianchi groups using the reduction theory of this paper. This uses the fact that every periodic sequence of transitions defines a closed orbit of the geodesic flow. As in the compact case [F1], it is necessary to understand the intersections of elements in the Markov partition and to introduce compensating terms to keep orbits that meet the boundary of our flowboxes from being overcounted. Although the flowbox boundaries are fractal, they can be approximated sufficiently well by small open neighborhoods that an exact calculation of the zeta functions can be made. For related work for $n = 2$, see [M, F2]. It is not so clear that the reduction theory mentioned in the preceding paragraph is adequate for zeta function computations, however, since a periodic sequence of transitions may define a parabolic element of Γ instead of a closed orbit of the geodesic flow.

The complexity of our example for $k = \mathbf{Q}(\sqrt{-15})$ suggests the problem of choosing cusp neighborhoods to minimize the number of lines and circles in the standard partition. One

could try instead some perturbed version of a fan tiling (the theory of sections 5 and 6 would still go through, as shown in [F3], Theorems 1 and 2) in hopes of obtaining as few rectangles as possible.

The results of this paper apply to many hyperbolic knot/link complements, since they are hyperbolic quotients by finite index subgroups of Bianchi groups. Indeed, by [MR, chapter 9] any noncocompact arithmetic Kleinian group is, up to conjugacy, commensurable with a Bianchi group. A large list of examples is given that includes the figure-eight knot complement, the Borromean rings complement, and the hex-link complement, corresponding respectively to the three choices of k in our examples above [MR, pp. 436-439]. It remains to examine some nonarithmetic cases to see if Theorem 12 applies or if a reduction theory with finitely many rectangles can be found by some other means.

In a more speculative vein, we might ask whether the geometric methods of this paper could be extended to include a reduction theory for real quadratic fields. It seems at present that "horoballs" in $H^2 \times H^2$ used in studying the Hilbert modular group [Hi] are not suitable for a description of pairs of boundary points along the lines of Theorem 6.

The relation between our Gauss correspondence and Diophantine approximation should be explored. [Mau,V] may be helpful here. Simple continued fractions were used to study minima of real quadratic forms in Markov's thesis, which is thoroughly developed in [D2, Chapter 7]. As a first step, one might attempt to recover this deep theory via least-remainder continued fractions.

References

[B] Bowen, R. Symbolic dynamics for hyperbolic flows. Amer. J. Math. 95 (1973) 429-460.

[Ca] Cawley, E. Smooth Markov partitions and toral automorphisms. Erg. Th. Dyn. Sys. 11 (1991) 633-651.

[C] Conway, J.H. The Sensual (Quadratic) Form. M.A.A., 1997.

[D1] Dickson, L.E. History of the Theory of Numbers, Vol. III. Quadratic and Higher Forms. Chelsea Publishing Company, New York, 1952.

[D2] Dickson, L.E. Studies in the Theory of Numbers. Chelsea Publishing Company, New York, 1957.

[Dir] Dirichlet, P.G.L. Vereinfachung der Theorie der binaren quadratischen Formen von positiver Determinante. Abh. K. Akad. Wiss. Berlin, Math. (1854) 99-115.

[EGM] Elstrodt, J., Grunewald, F. and Mennicke, J. Groups Acting on Hyperbolic Space. Springer-Verlag, Berlin, 1998.

[F1] Fried, D. The zeta functions of Ruelle and Selberg, I. Ann. Scient. Ec. Norm. Sup., 19 (1986) 491-517.

[F2] Fried, D. Symbolic dynamics for triangle groups. Invent. Math. 125 (1996) 487-521.

[F3] Fried, D. Symbolic dynamics for geometrically finite groups, in preparation.

[G] Gauss, C.F. Disquisitiones Arithmeticae, 1801. English edition, Springer-Verlag, New York, 1986.

[Hi] Hirzebruch, F. The Hilbert modular group, resolution of the singularities at the cusps and related problems. Sem. Bourbaki, 1970/71, exp. 396, L.N.M. 244, Springer-Verlag, Berlin, 1971.

[H1] Hurwitz, A. Uber die Entwicklung komplexer Grossen in Kettenbruche. Acta Math. 11 (1888) 187-200 or Mathematische Werke, Birkhauser, Basel, 1933, Band II, pp. 72-83.

[H2] Hurwitz, A. Uber eine besondere Art der Kettenbruch-Entwicklung reeller Grossen. Acta Math. 12 (1889) 367-405 or Mathematische Werke, Birkhauser, Basel, 1933, Band II, pp. 84-115.

[H3] Hurwitz, A. Uber die Reduktion der binaren quadratischen Formen. Math. Ann. 45 (1894) 85-117 or Mathematische Werke, Birkhauser, Basel, 1933, Band II, pp. 157-190.

[H] Hurwitz, J. Uber eine besondere Art der Kettenbruch-Entwicklung complexer Grossen. Acta Math. 25 (1902) 231-290.

[Mau] Maucourant, F. Sur les spectres de Lagrange et de Markoff des corps imaginaires quadratiques. Erg. Th. Dyn. Sys. 23 (2003) 193-205.

[Ma] Markoff, A. Sur les formes quadratiques binaire indefinies. Math. Ann. 15 (1879) 381-406.

[M] Mayer, D. On the thermodynamic formalism for the Gauss map. Commun. Math. Phys. 130 (1990) 311-333.

[MR] Maclachlan, C. and Reid, A. The Arithmetic of Hyperbolic 3-Manifolds. Springer-Verlag, New York, 2003.

[Ra] Ratner, M. Markov partitions for Anosov flows on n-dimensional manifolds. Israel J. Math. 15 (1973) 92-114.

[R] Ratcliffe, J. Foundations of Hyperbolic Manifolds. Springer-Verlag, New York, 1994.

[Ri] Riley, R. Applications of a computer implementation of Poincare's theorem on fundamental polyhedra. Math. Comp. 40 (1983) 607-632.

[Se] Serret, J.-A. Cours d'Algebre Superieure, 5th ed., tome 1, Gauthier-Villars 1885.

[S] Swan, R.G. Generators and relations for certain special linear groups. Adv. Math. 6,(1971) 1-77.

[V] Vulakh, L.Y. Farey polytopes and continued fractions associated with discrete hyperbolic groups. Trans. Amer. Math. Soc. 351 (1999) 2295-2323.

Department of Mathematics, Boston University, 111 Cummington St., Boston MA 02215

df@bu.edu