ILLINOIS JOURNAL OF MATHEMATICS
Volume 41, Number 3, Fall 1997

RIEMANN SURFACES HAVE HALL RAYS AT EACH CUSP

THOMAS A. SCHMIDT AND MARK SHEINGORN

Overview

The main result of this paper is that every Riemann surface has a Hall ray at each cusp. By this we mean that the spectrum of maximal penetration heights of geodesics into a horocycle about the cusp fills out a real half-line. For the modular surface, this result is well known and derives from Hall's Theorem for continued fractions.

We also show that a Hall ray can exist without the presence of cusps in two settings: First, on a surface derived as a limit of cusped surfaces, whose fundamental region contains two entire horocycles. And second, with respect to a *hyperbolic continued fraction* for which the former role of a cusp is played by a simple closed geodesic.

The limiting process mentioned above also produces an infinite class of closed geodesics on the theta surface, the quotient of the upper half-plane by the usual theta group, that are pair-wise equal in length, but whose precursors in the limit process are never equal—this equality is then *accidental*. That is, there is a change in length spectrum multiplicity at the limit surface.

1. Hyperbolic geometric preliminaries

We will employ elementary hyperbolic geometry of the Poincaré upper half plane, $\mathcal{H} := \{z = x + iy \mid y > 0\}$. The geodesics of $\mathcal{H}$ are semi-circles perpendicular to the real axis, and vertical lines; these geodesics are called *h-lines*. The group of conformal hyperbolic isometries of $\mathcal{H}$ is PSL(2, $\mathbb{R}$).

Every Riemann surface is the quotient of $\mathcal{H}$ by some discontinuous subgroup of PSL(2, $\mathbb{R}$); such a subgroup is called a *Fuchsian* group. Each conformal isometry of a surface is realized by some element of the PSL(2, $\mathbb{R}$) normalizer of the corresponding Fuchsian group. The horocycles of $\mathcal{H}$ are the various PSL(2, $\mathbb{R}$) images of the sets of the form $\{z = x + iy \mid y > \alpha\}$ for some α. The natural boundary of $\mathcal{H}$ is $\mathbb{R} \cup \infty$; the action of PSL(2, $\mathbb{R}$) extends to this boundary. A Riemann surface is said to have a *cusp* at a point p if p is on this boundary and there is a non-trivial element of the corresponding Fuchsian group which fixes p. Such an element is called *parabolic*. A *hyperbolic* element of PSL(2, $\mathbb{R}$) fixes two real points and the h-line connecting them; this h-line is called the *axis* of the hyperbolic element. The real endpoints of

Received December 8, 1995.
1991 Mathematics Subject Classification. Primary 30F35, 11J70.
The first-named author thanks Widener University for granting generous leaves of absence.

an axis are called its *feet*. All these matters are beautifully presented in the text of Alan Beardon [B].

2. Statement of results

We begin by defining the concept of a Riemann surface with a Hall ray at a particular cusp. Let Σ be a Riemann surface, with a cusp p. Let $\Sigma = \Gamma\backslash\mathcal{H}$, for some Fuchsian group Γ. We make no assumptions concerning Γ apart from the fact that it has at least one parabolic conjugacy class. There is a maximal horocyclic neighborhood of the cusp—the largest horocyclic disc on Σ which is a punctured disc.

We assume, without loss of generality, that ∞ lies over p. Choose any closed geodesic τ on Σ. It is the projection to Σ from $\mathcal{H}$ of any of the axes of a hyperbolic conjugacy class in Γ. The *height* of τ is the maximum euclidean radius of these axes. The maximum exists since the geodesic is closed. This concept can be extended to arbitrary geodesics by replacing maximum by the supremum (which may be infinite) and replacing the axes of the hyperbolic conjugacy class by the Γ-orbit of an arbitrary fixed lift of τ. In general, the *naive height* of a lift to the Poincaré upper half-plane of a geodesic on a Riemann surface is the euclidean diameter of this lift. The *height* of a geodesic is the supremum of the naive heights of the lifts of the geodesic.

We are now ready for the following:

Definition. We say that Σ has a *Hall ray* at p if there exists an N depending only on p and Σ such that there is a subset of the set of all heights of geodesics on Σ which is dense in $[N, \infty)$.

In [SS2], we showed that every Hecke triangle surface (the quotient of $\mathcal{H}$ by a Hecke triangle group G_q, see below) admits a Hall ray with respect to the cusp at infinity. Here we show that possessing a Hall ray is in fact a general phenomenon—every Riemann surface admits a Hall ray with respect to each of its cusps. We also show that there is a surface (denoted $\Gamma_{inf}\backslash\mathcal{H}$)—arising as a particular limit of small covers of the Hecke triangle surfaces denoted Γ_q, defined below and studied in [S]—which has no cusps but with fundamental region containing two entire horocycles. We show that the surface admits a Hall ray with respect to each of these horocycles.

In [SS1] and [SS2], we discussed the length spectra of the Hecke triangle surfaces. In particular, we pointed out that two closed geodesics on a surface of this family can be of equal length in one of exactly two manners: either the geodesics correspond to a family of pairs of geodesics which are of equal length on all of the surfaces, or the length equality is what we called accidental. Here we show, using our limiting procedure, that there are infinitely many distinct pairs of equi-length closed geodesics on the theta surface, $\Gamma_\Theta\backslash\mathcal{H}$, which are accidentally the same length. The group Γ_Θ contains the group Γ_{inf}, and just as Γ_{inf} is the limit in the appropriate sense of Γ_q, so Γ_Θ is the limit of G_q.

The authors wish to thank the referee for many useful comments which led us to clarify and simplify the geometric arguments presented herein.

3. The existence of the Hall ray

We announce the main result of this paper.

MAIN THEOREM. *Let Γ be a Fuchsian group and $\Sigma = \Gamma\backslash\mathcal{H}$ be a surface with a puncture at p. Then Σ has a Hall ray with respect to p.*

We note that this means that if the surface has more than one puncture, then there is a Hall ray with respect to each. We will employ open geodesics to prove this theorem. Also, we remark that the techniques of [SS2] for the infinite volume Hecke triangle surfaces go through for surfaces arising as quotients of the upper half-plane by Fuchsian groups of the second kind. The proof below, however, does not use any properties of Γ beyond the facts that it is Fuchsian and has at least one cusp. Indeed, the key idea to the proof is simply that one can *slide* any sufficiently high h-line by Euclidean translation so that it intersects the horocycles of $\mathcal{H}$ which correspond to lifts of the cusp of Σ far from the cusp at the base of these horocycles.

Proof. To begin, we construct the continued fraction expansion for Γ as given in [LS]. To each lift to $\mathcal{H}$ of a geodesic of Σ, we will associate a doubly infinite sequence. The sequences of different lifts will differ only by a shift. We will show that by sliding an h-line on $\mathcal{H}$, we can force the entries of the associated double sequence to be small without change of depth of penetration into the horocycle of p.

By conjugation in $SL(2, \mathbb{R})$, we assume that p lifts to ∞. The fundamental horocycle H at ∞ is

$$H = \{z \in \mathcal{H} \mid z = x + iy,\ y > m\}$$

where m is the smallest value such that H projects to a punctured disc on Σ. Now striate H into fundamental (vertical) "strips", i.e. maximal strips containing no equivalent points. Again by $SL(2, \mathbb{R})$ conjugation, we assume $m = 1$. By Shimizu's Lemma ([Sh]), this forces all non-zero c-entries in the matrices of Γ to have $|c| \geq 1$. Isometric circle considerations now ensure that the translation length at ∞ cannot be smaller than 1.

Consider the orbit of H under Γ. This is a set of disjoint horocycles. An h-line passes through at most a countable number of these horocycles. It may pass through none, or a finite number—the latter happens if it terminates in both directions in the cusp under discussion.

In general, we get a doubly infinite sequence of horocycles H_i, $-\infty < i < \infty$. For each H_i we keep track of the strips encountered by the h-line in the excursion through H_i. The number of such strips encountered is denoted b_i. We attach a sign,

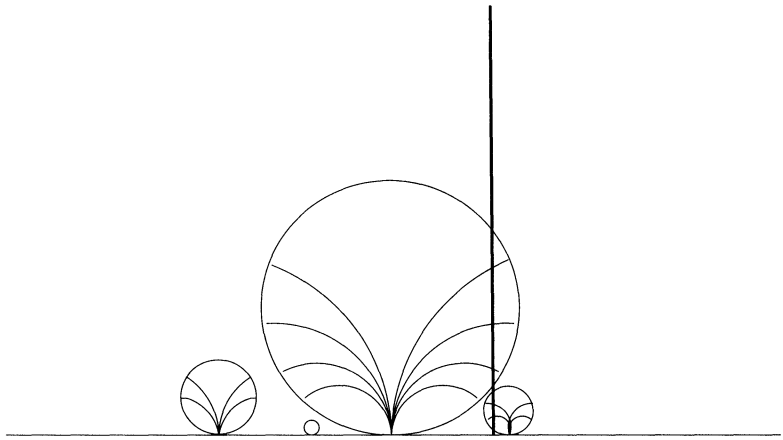

Figure 1. Defining continued fractions.

+ to denote exit to the right of the cusp, − denotes exit to the left. Our continued fraction expansion is then the double sequence (H_i, b_i). (See Figure 1.) ■

One more piece of notation: the cusp at the base of the horocycle H_i will be denoted a_i/c_i, where $\begin{pmatrix} a_i & * \\ c_i & * \end{pmatrix}$ takes ∞ to this cusp. It is easily seen that this is well defined. We now choose some h-line, which has euclidean radius greater than 100 and consider the corresponding H_i and b_i. The following observations give our proof.

A. The euclidean radii of H_i go monotonically to zero as $i \to \pm\infty$. This follows from the following two facts. (a) Our h-line enters a horocycle (other than the horocycle at ∞), which has euclidean radius at most 1, in its (euclidean) top half, exiting through the bottom half. This is because the h-line is nearly vertical as far as all the horocycles are concerned. (b) If two consecutive horocycles have nearly equal euclidean radii, the succeeding horocycle then has radius at most half as big, for it is trapped (in its entirety) in the near-triangle under the two consecutive large ones.

B. For a geodesic with a lift of large naive height and small (other) $|b_i|$, the naive height is the actual height. Indeed, the naive height can be bounded in terms of $|b_0|$, where H_0 is the horocycle at ∞. Having chosen a naive height greater than 100, small here may be taken to be 10.

For example, if $b_{15} = 100$ while all other $|b_i| \leq 10$, then no other penetration exits an H_i as close to the cusp as the exit of H_{15}. It is not necessary to refine this; indeed, difficulty arises if the large $|b_i|$ and small $|b_j|$ are close in size.

C. A simple computation shows that the euclidean radius of H_i is $1/2c_i^2$.

D. A small $|b_i|$ indicates that the h-line passes through H_i near the edge of the horocycle, rather than near its cusp. This is because the striations accumulate at

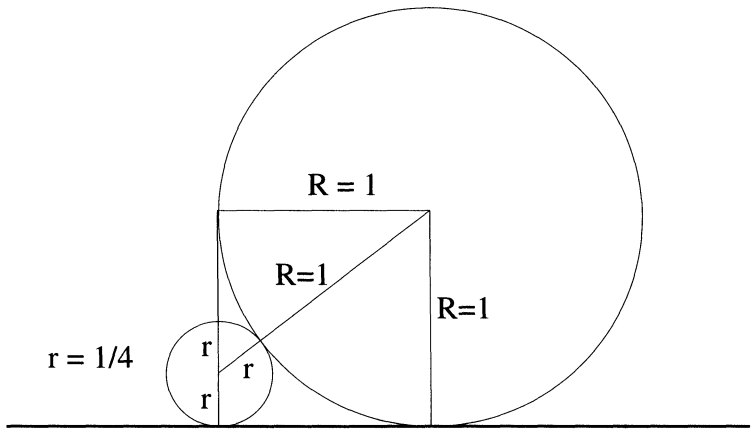

Figure 2. A Euclidean tangency problem (special case).

its cusp. Any near vertical h-line exiting close to a cusp must pass through many striations.

More precisely (with the normalization we have chosen), the y-coordinate of the exit point lies between $1/(c_i^2[(b_i - 1)^2 + 1])$ and $1/(c_i^2[b_i^2 + 1])$. Note that the dependence on c_i factors out; for all horocycles, the b_i determine comparable 'fractional division points' independent of the euclidean size of a horocycle. The top of the horocycle corresponds to $b_i = 0$. (See Figure 3.)

E. Consider consecutive pairs (H_i, b_i); (H_{i+1}, b_{i+1}). Say $|b_i| \leq 10$ but that $|b_{i+1}| > 10$. First of all, this forces $1/2c_{i+1}^2 \leq 1/6c_i^2$. (This restates the simple observations: (a) that because $|b_{i+1}|$ is large, we enter H_{i+1} near the top; (b) that at worst we are tangent to H_i (as H_i appears in the continued fraction; and (c) that, for r defined by Figure 2, $r = 1/4$. (This is the extremal case of tangency in (b), entering through the top in (a), and also the geodesic being vertical. The constant 10 above assures that replacing $1/4$ by $1/3$ leaves the inequality valid.)

By A, we have $1/2c_{i+1}^2 \leq 1/6c_j^2$, for $1 \leq j \leq i$ also. Now slide the h-line along the real axis, with a euclidean translation, to obtain an h-line of the same height but with $|b_{i+1}| \leq 10$. What is the effect on previous b_i?

This is another euclidean tangency problem; it is not difficult to solve. We have an exit from H_i at height between 0 and 9, measured by b_i. The worst we could then do is to slide an entire radius of H_{i+1}.

Since the fractional division points are invariant, we will consider H_i having diameter 1, as in Figure 3. We thus consider the case when the center of H_i is $(0, 1)$, and the cusp is thus $(0, 0)$. The parameter b is the exit-height (y-coordinate) as the h-line exits H_i. The x-coordinate is then $\sqrt{2b - b^2}$, a quantity we shall refer to as m_b. The maximal decrease in H_i exit-height stemming from sliding across the horocycle H_{i+1} occurs when the cusp of H_{i+1} is at $(m_b, 0)$, and the horocycle is tangent to H_i.

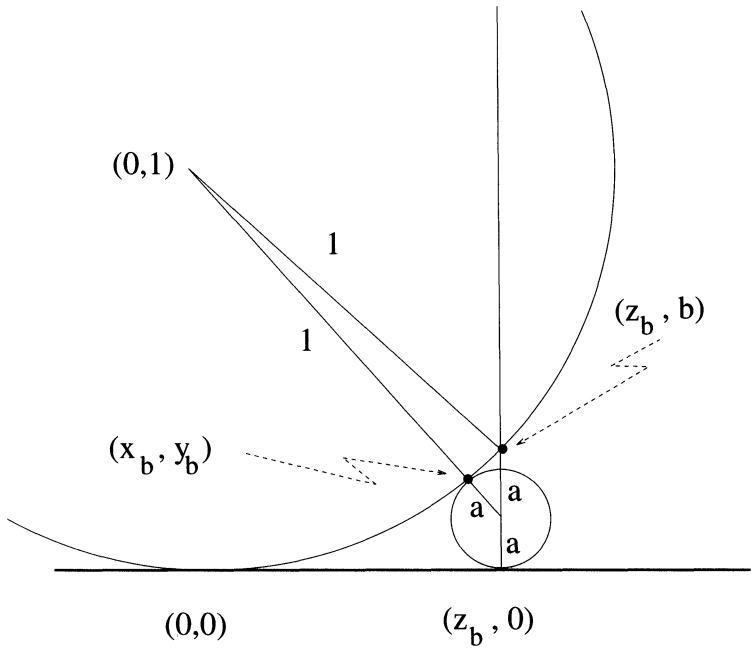

Figure 3. A Euclidean tangency problem (general case). Here $z_b = \sqrt{2b - b^2}$.

This will occur when the radius of H_{i+1} is $m_b/2$.

(In the following paragraph, quantities are evaluated as if the h-line with large naive height were in fact a vertical line. This is justified, since the hyperbolic distance between this vertical and the given h-line (of naive height ≥ 100) with the same foot goes to zero as we approach the real axis, but hyperbolic distance between the first 10 fractional division points remains constant amongst all our horocycles.)

The h-line will be slid such that its foot near $(m_b, 0)$ is moved to nearly $(m_b - m_b^2/4, 0)$. Call this new foot x_b. The new exit-height is then $y_b := 1 - \sqrt{1 - x_b^2}$. It is easy to see that y_b is minimized when b is as small as possible. With our parametrization ($c = 1/\sqrt{2}$, $|b_i| = 9$) gives the minimum value of $b = 1/41 = .024390244\ldots$. In turn, this yields $y_b = .021757727\ldots$. In terms of the fractional division points, this is an exit at $b_i = 9.535269614\ldots$. (Not a possible value, of course, as b_i is an integer.) We have shown that sliding all the way across H_{i+1} increments b_i by at most 1. This argument applies, *mutatis mutandis* to previous b_j, $1 \leq j < i$.

F. We can slide an h-line recursively, with smaller and smaller euclidean moves, generating a sequence with $|b_i| \leq 10$ for $i \neq 0$, and identical naive height to the original h-line. We do this for positive and negative i in decreasing order of the euclidean radii of the corresponding H_i.

G. We address the effect a slide at H_i for, say, a positive i has on a horocycle incursion at H_j for a negative j. Our previous analysis applies directly if the exit from H_j is higher than the entrance into H_i.

But this need not be so, the worst case being when the horocycles are of equal size, and sliding across H_i takes our geodesic exit very near the cusp of H_j. This potentially fatal situation is salvaged as follows.

First, we can slide across the H_i in the other direction—not both directions, left and right, can result in exiting near its cusp of a single H_j, because in one direction we are sliding away from that cusp.

The remaining difficulty is that sliding in the 'other' direction may enter a new horocycle, again of equal height to i, and exit near the cusp of the new horocycle. But this is not possible.

For, we would then have a configuration of three horocycles of nearly the same euclidean radius: two tangent, (H_j and another), and also H_i. The original geodesic exits near the cusp of H_i. If sliding in one direction forces an exit near the cusp of H_j, sliding a nearly euclideanly equal amount in the opposite direction cannot force an exit near the cusp of the horocycle tangent to H_j.

H. Since $1/|c_i|^2 \to 0$, this sliding process converges to an h-line of naive height equal to the original and with feet having no $|b_i| > 11$, $i \neq 0$. This means that from our original geodesic of naive height of 100, the naive height becomes the actual height of the new geodesic and we have demonstrated that $[100, \infty)$ is in the Hall ray.

It is worth discussing the prospects for terminating this process. In the algorithm as given, if the geodesic were in fact closed, we would not know it. We would simply never encounter a further need to slide. This is endemic—our tracking is not exact enough to detect closure. This is no different from ordinary continued fractions; one cannot look at a finite sequence of partial quotients and conclude that the continued fraction expansion is periodic.

There is however, a mechanism for selecting closed geodesics if Σ is of finite volume with elliptic fixed points of even order. This consists of terminating the process as follows: when we are quite near the real axis on each side—at height y_ϵ say, replace the geodesic with one connecting the two elliptic fixed points of even order nearest the points on the original geodesic at height y_ϵ. It can be shown, under the hypotheses above, first that the continued fraction of the new geodesic (up to this point on the geodesic) differs little from that of the old, and second that the naive heights are also very close (both of these require y_ϵ to be small). Moreover, the new geodesic is closed, being fixed by the hyperbolic that is the product of the two elliptics of order two fixing the elliptic fixed points. With the assumption that the elliptic fixed points are in fact computable, this fact would greatly aid calculation.

We close this section with some remarks on the effectiveness of this continued fraction. It is not hard to see that our expansion can be computed if and only we

can effectively produce the set of cusps $a/c \in [0, 1)$, in order of size of $|c|$. This is a notoriously difficult problem, even for the Hecke triangle groups (where there has been recent progress [RS]). Presumably, it is an easier task than testing membership in Γ of matrices in $SL(2, \mathbb{R})$.

4. Intermezzo: Accidental multiplicity on $\Gamma_\Theta\backslash\mathcal{H}$

In this section, we focus upon the surfaces which arise as quotients of the Poincaré upper half-plane by the Hecke triangle groups of the first kind. Each of these groups contains the element

$$T = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

For each integer $q \geq 3$, let $\lambda_q = 2\cos \pi/q$ and

$$S = \begin{pmatrix} 1 & \lambda_q \\ 0 & 1 \end{pmatrix}.$$

The Hecke group G_q is then defined as $\langle S, T\rangle$. The theta group, traditionally denoted Γ_Θ, is also generated by such a pair, but is the limit as $q \to \infty$; thus its translation length at infinity is 2. For each of these groups, we will refer to $S^m T S^{-m}$ as T_m. We define Γ_q to be $\langle T_0, T_1, \ldots, T_{q-1}\rangle$.

As the parameter q increases, the $G_q\backslash\mathcal{H}$ form a discrete deformation family. That is, by way of the underlying groups, a closed geodesic has an avatar on each surface of bigger q, see [SS2]. We say that two equi-length geodesics on a particular surface are accidentally of the same length if the corresponding geodesics on each other surface of the family are unequal in length. In general, such accidental pairs will correspond to a change in the length multiplicity of the surface from that of its nearest neighbors.

THEOREM. *There exist infinitely many distinct pairs of equi-length closed geodesics on $\Gamma_\Theta\backslash\mathcal{H}$, each of which is the limit of pairs of distinct length $G_q\backslash\mathcal{H}$ geodesics.*

Following Eichler, we denote by (z, w) the hyperbolic length of the geodesic connecting z and w in $\mathcal{H}$. We note the following formula, a standard computation:

$$(i, a + bi) = \log \frac{(a^2 + b^2 + 1) - \sqrt{(a^2 + b^2)^2 + 2(a^2 - b^2) + 1}}{2b}$$

Next we note some geometric facts about $\Gamma_\Theta\backslash\mathcal{H}$. As indicated above, Γ_Θ is generated by S: $z \mapsto z + 2$ and T: $z \mapsto -1/z$. This is a level 2 congruence subgroup with signature $(0; 2, \infty, \infty)$. Figure 4 gives a useful fundamental region for this group.

Note that the surface admits two reflections: L: $z \mapsto -\bar{z}$ and M: $z \mapsto (\bar{z} + 1)/(\bar{z} - 1)$.

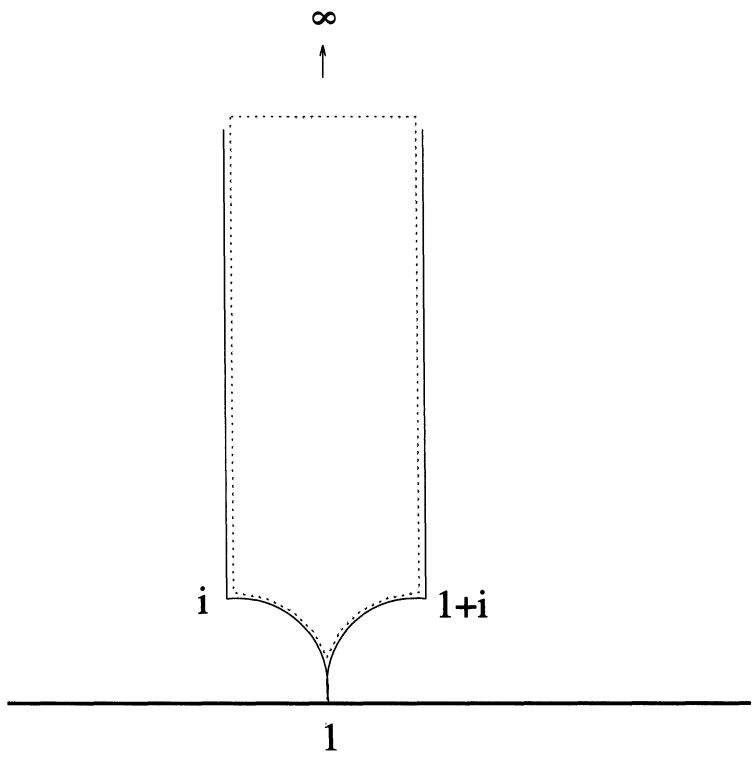

Figure 4. 'Standard' fundamental region for $\Gamma_\Theta \backslash \mathcal{H}$.

The reflection L fixes (point-wise) two 'lines' on the surface: one connects ∞ and 1—this passes through $i+1$, which is an ordinary point on $\Gamma_\Theta \backslash \mathcal{H}$; and one connecting ∞ and i, an elliptic fixed point of order 2 here, then from i to 1 along the unit circle. We think of the first of these lines as running down the back of the surface, and the second as running down the front, where the elliptic fixed point of order 2 is in view. See Figure 5.

The reflection M has the single fixed h-line connecting i and $i+2$, both of these points are lifts of the same elliptic fixed point of order 2. This h-line meets the real axis at $1 \pm \sqrt{2}$. The corresponding line separates $\Gamma_\Theta \backslash \mathcal{H}$ into two isometric punctured hemispheres, one containing ∞—which we will call northern, the other containing 1, called southern. Note LM is an Atkin-Lehner involution at $1 + \sqrt{2}i$. It is also worth remarking that, for Γ_Θ, the *cusp-width* of ∞ is 2, while that of (the other cusp) 1 is 1. Since we have shown these cusps are isometric, we see that *cusp-width is not an intrinsic geometric property.* We illustrate L and M in Figure 5.

We are now ready to define two sets of geodesics. Let $m \geq 2$ be an integer. The collection $\mathcal{N}$ consists of the geodesics whose lifts connect i and $i + 2m$. These are

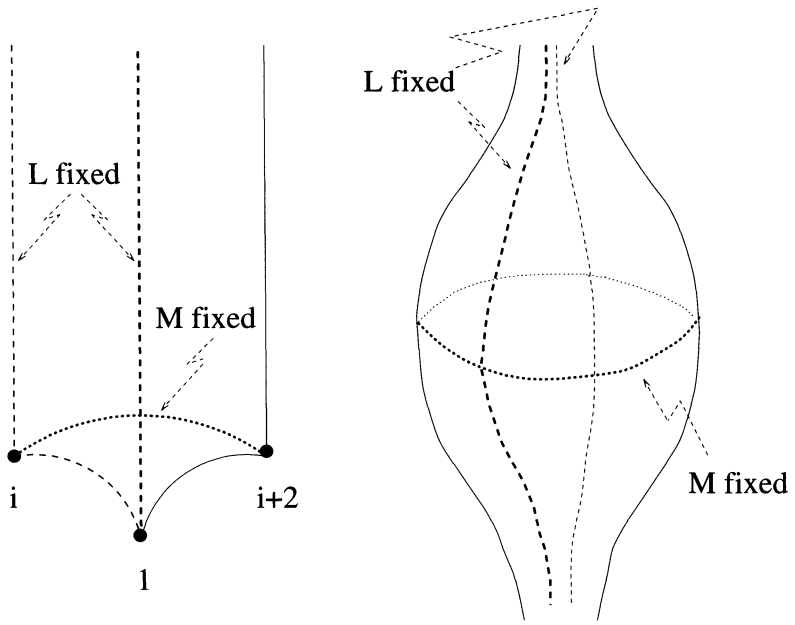

Figure 5. Isometries of $\Gamma_\Theta \backslash \mathcal{H}$.

closed and a comparison with the reflection line of M in the fundamental region shows that they lie entirely in the northern hemisphere. The images of these geodesics under M constitute $\mathcal{S}$. Because M defines an isometry, corresponding geodesics have the same length. Also, a geodesic in $\mathcal{N}$ or $\mathcal{S}$ begins and ends at the elliptic fixed point of order 2.

The geodesics in $\mathcal{N}$ are fixed by the Γ_Θ transformations $T_0 T_m$.

(*Aside*: Direct computation shows that the lengths of these geodesics are $\sim \log m$ where $m = 2, 3, \ldots$ Thus we have a set of geodesics on $G_q \backslash \mathcal{H}$ which can be used to show that $\#\{A \in G_q \mid \operatorname{tr} A \leq X\} \geq cX$, where c may be computed explicitly. Note that no discussion of quadratic forms nor of sophisticated analytic techniques is needed, by virtue of the elliptic fixed point of order 2.)

In [S], the surfaces $G_q \backslash \mathcal{H}$ are discussed and the notation $T_0 \uparrow T_m \downarrow T_1$ is defined as $T_0 T_1 \ldots T_{m-1} T_m T_{m-1} \ldots T_2 T_1$. The precise path of the geodesic fixed by $T_0 \uparrow T_m \downarrow T_1$ is determined in [S, §3, p. 552, paragraph 2]. This geodesic is in $\mathcal{S}$; in particular, it is simple. The elements $T_0 \uparrow T_m \downarrow T_1$ and $T_0 T_m$ cannot be Γ_Θ conjugate, let alone identical, as their geodesics are different—they lie in separate hemispheres. Indeed, for $m < q$, they cannot be even be Γ_q conjugate, by the same arguments. We will show that their traces are not the same, thus ruling out conjugacy in general.

It remains to show that these pairs of geodesics are not the 'limits' of equi-length pairs on the $G_q \backslash \mathcal{H}$—i.e., that the equality of length in Γ_Θ is indeed accidental.

There are at least two ways to do this. Fix m and consider the corresponding pair of geodesics. Most directly, we show that for each q, the G_q primitive hyperbolics fixing each of the two $G_q\backslash\mathcal{H}$ avatars of these geodesics have differing traces. Alternatively, we invoke the cone metrics of C. Judge [J] to show that the G_q lengths differ.

Direct Trace Computations. The following calculation takes place in G_q, not Γ_Θ. Note that now T_m fixes $i + m\lambda_q$. The $\operatorname{tr} T_0 T_m$ is $m^2\lambda_q^2 + 2$. We next compute $\operatorname{tr}\sigma_m$ where $\sigma_m := T_0 \uparrow T_m \downarrow T_1$.

Let $\tau_m := T\sigma_m T$. We have $\tau_m = (ST_0)^m(S^{-1}T_0)^m$. Now let $ST_0 =: U$, and so

$$U^m = \begin{pmatrix} a_m & -a_{m-1} \\ a_{m-1} & -a_{m-2} \end{pmatrix},$$

where $a_m := a_m(\lambda_q)$ is monic of degree m in λ_q. For $V := S^{-1}T_0$, we find

$$V^m = (-1)^m \begin{pmatrix} a_m & a_{m-1} \\ -a_{m-1} & -a_{m-2} \end{pmatrix}.$$

From these, simple computation shows

$$|\operatorname{tr}\sigma_m| = a_m^2 + 2a_{m-1}^2 + a_{m-2}^2.$$

For the limit case of $\lambda = 2$, one easily shows that $a_{m-1} = m$. Thus the traces of $T_0 T_m$ and of σ_m are the same. We show that this is not the case for any $\lambda = \lambda_q$. Now, $a_{m-1} = (\sin m\pi/q)/(\sin \pi/q)$, see say [LS]. Thus, $|\operatorname{tr}\sigma_m| = a_{m-1}^2\lambda_q^2 + 2$. But, a_m increases with $\lambda = \lambda_q$. This last can be shown by induction: Since $0 \le U^n(\infty) \le \infty$ and $U^{-n}(\infty) = \lambda_q - U^n(\infty)$, for n in $\{1, \ldots, q-1\}$; we find that $U^m(\infty)$ increases if $m \le \lfloor q/2 \rfloor$. But, a_m/a_{m-1} is $U^m(\infty)$. Hence, a_{m-1} increases to m. That is, the difference in lengths between the geodesics goes to zero as $q \to \infty$.

We now give bounds on the lengths of the geodesics on $G_q\backslash\mathcal{H}$ corresponding to σ_m. If $m = q/2$, then $a_{m-1} = 1/(\sin \pi/q)$. Thus, $a_{m-1} \ge q/\pi$, or $a_{m-1} \ge 2m/\pi$. Therefore, for $q \ge 2m$, $|\operatorname{tr}\sigma_m| \ge (4m/\pi)^2\lambda_q^2 + 2$. Thus, a lower bound for the length of such a geodesic is $2\log(m\lambda_q/2)$.

Cone Metrics. First, the quadrant $\mathcal{Q} := \{x + iy \mid 0 \le x < \infty,\ y > 0\}$ may be mapped to $\mathcal{H}$. A sequence of maps that does this (and takes ∞ to ∞) is:

$$z \mapsto i\pi z \mapsto e^{i\pi z} \mapsto \left(\frac{e^{i\pi z} + 1}{e^{i\pi z} - 1}\right)^2$$

Next, note that the 1-dimensional cone metric with angle α on $\mathcal{Q}$ is given by Judge as

$$d\tau := \frac{|\alpha\, ds|}{|\sinh \alpha y|}$$

where ds is euclidean metric. Expanding the sinh in a Taylor series, we find:

$$d\tau = \frac{|ds|}{y(1 + \frac{\alpha^2 y^2}{3!} + \frac{\alpha^4 y^4}{5!} + \ldots)}$$

It is clear from this that the length with respect to these metrics of any curve in $\mathcal{Q}$ increases as $\alpha \downarrow 0$. (The Hecke groups in Judge's normalization have $\alpha = \pi^2/q$.) This will also be true then, if we push each metric forward to $\mathcal{H}$. This means that if we measure the geodesic between i and $i + m\lambda_q$ for $\alpha = 0$ and then $\alpha = \pi^2/q$, the latter will be shorter. The first of these is just the northern hemisphere length of the geodesic between these points on $\Gamma_q\backslash\mathcal{H}$. It is also the length of this geodesic on $G_q\backslash\mathcal{H}$, since no elliptic fixed point of order 2 intervenes.

Now the second measurement gives the length of the $G_q\backslash\mathcal{H}$ geodesic whose lift is fixed by σ_m. This is because, again, we know there is no intervening elliptic fixed point of order 2, and also the fact that this geodesic has an intermediate lift contained entirely in the southern hemisphere on $\Gamma_q\backslash\mathcal{H}$ is equivalent to it lying entirely in the π^2/q cone on $G_q\backslash\mathcal{H}$. We have shown that the geodesics are of unequal length until we reach $\Gamma_\Theta\backslash\mathcal{H}$.

This technique may be profitably applied to the $\Gamma_q\backslash\mathcal{H}$ as well. Not only do we find that the northern and southern routes between i and $i + m\lambda_q$ are longer and shorter, we may take certain routes 'woven' between these two points and lying in the disc bounded by them. Such curves are characterized by a choice for each of the $m - 2$ intervening elliptic fixed points of order 2 lying within said disc of whether we pass by to the north or the south of the elliptic fixed point of order 2. There are thus 2^{m-2} distinct such geodesics.

Further their lengths are all longer than the southern route and shorter than the (piece-wise geodesic) path from i to $i + m\lambda_q$ gotten by connecting successive elliptic fixed points of order 2 by (the projections of) shortest northern arcs. (This last path has length $\approx m \log \lambda_q$, which is longer than the northern route, which has length $\approx 2(\log m + \log \lambda_q)$. The piecewise-geodesic path would be in the homotopy class of the northern route, if we adjusted the path so that it passed through $i(1 + \epsilon) + s\lambda_q$, rather than $i + s\lambda_q$ (the elliptic fixed point of order 2), for $2 \leq s \leq m - 1$.)

The length inequality holds because, for a given geodesic, if we flip all northern (resp. southern) passages to the south (resp. north), we (a) shorten (resp. lengthen) the curve, by Judge's formula, and (b) obtain a curve in the homotopy class of the southern (resp. northern) route. We have thus produced $2^{m-2} + 2$ (simple) closed geodesics on $\Gamma_q\backslash\mathcal{H}$, each having length less than $m \log \lambda_q$ and greater than $2 \log(m\lambda_q/2)$.

Example. ($q \geq 13$). Intermediate in absolute trace to the piece-wise connection and southern route elements ($= T_0 \uparrow T_5 \downarrow T_1$), we have $A := T_0 \cdot T_1 \cdot T_4 \cdot T_5 \cdot T_4 \cdot T_1$. The geodesic corresponding to A bounces between i and $i + 5\lambda_q$, passing to the north of $i + 2\lambda_q$ and $i + 3\lambda_q$ and to the south of $i + \lambda_q$ and $i + 4\lambda_q$. Now $\operatorname{tr} A = 9\lambda_q^4 + 3\lambda_q^2 + 1$;

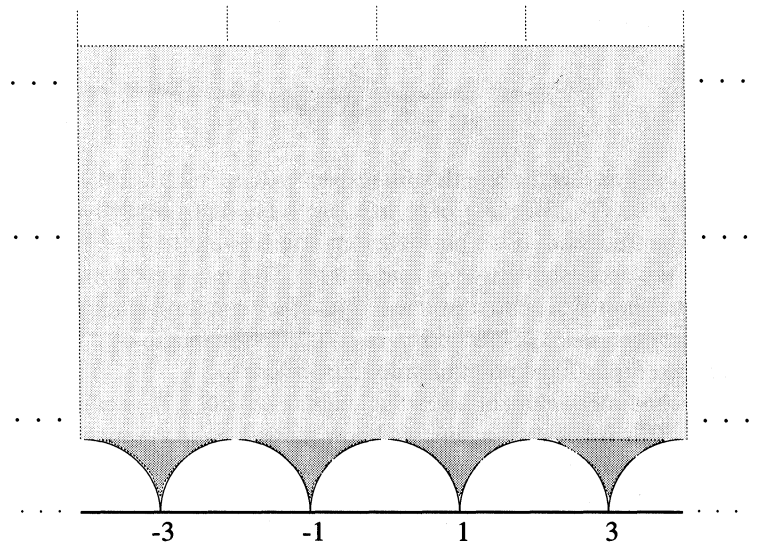

Figure 6. $\mathcal{F}_{\Gamma_{inf}}$, a fundamental region for Γ_{inf}.

whereas the trace for the southern route is $10\lambda_q^{10} - 6\lambda_q^8 + 11\lambda_q^6 - 6\lambda_q^4 + \lambda_q^2 + 2$. A simple calculation shows this trace is smaller than that of A in the indicated range. As to the upper bound, $4\log\lambda_q + \log 9 < \ell A < 4\log\lambda_q + \log 10$. The piece-wise route length is easily seen to be greater than $10\log\lambda_q$.

5. The geometry of Γ_{inf}

This subgroup of Γ_Θ is easy to define: $\Gamma_{inf} :=< T_m \mid m \in \mathbb{Z} >$. We will show that Γ_{inf} is infinitely generated, that it is normal in Γ_Θ, and that it has no parabolic elements. It is of the first kind, but with infinite volume, and has two very special limit points on the boundary of the most illuminating fundamental region which we shall construct.

5.1. *Fundamental regions.* First of all, Γ_{inf} is obviously Fuchsian, being a subgroup of the discrete group Γ_Θ. From the indicated generators, it is clear that a fundamental region is contained in $\cup_{m\in\mathbb{Z}} \mathcal{F}_{\Gamma_\Theta}$, where $\mathcal{F}_{\Gamma_\Theta}$ is the standard fundamental region for Γ_Θ. See Figure 6.

That this is a fundamental region follows easily. Indeed, if two points herein were equivalent, Γ_Θ considerations force their equivalence under a power of S^2: $z \mapsto z+2$. But no power of S^2 lies in Γ_{inf}, as this would give a new relation in Γ_Θ. It is now clear that ∞ is not a cusp of Γ_{inf}, as the fundamental region contains an entire horocycle at ∞. Next, note that all real vertices of this fundamental region are equivalent under

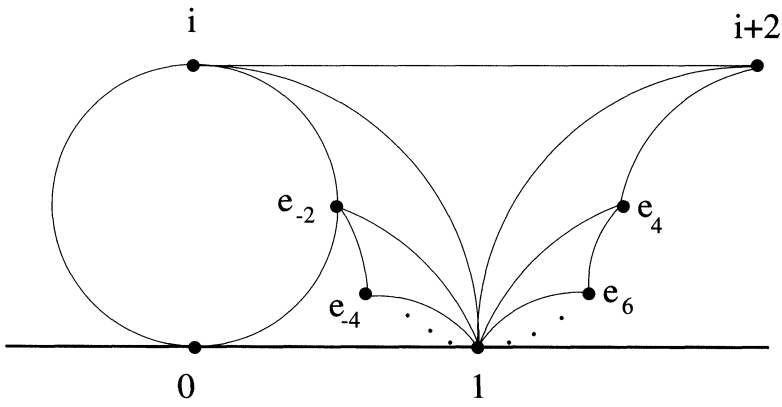

Figure 7. $\mathcal{F}^*_{\Gamma_{inf}}$, a better (?) fundamental region for Γ_{inf}.

sequential products of the various T_m. Thus we will know there is no cusp when we show that 1 is not a cusp.

Clearly, $T_0 S^{-2}$ generates the stabilizer of 1 in Γ_Θ. It is not hard to show that no power of this element is in Γ_{inf}. For, if the kth power of this element lies in Γ_{inf}, then so does $T_0 S^{-2k}$, by a simple conjugation argument. But this implies $S^{-2k} \in \Gamma_{inf}$, and we have seen that this is false. So the surface $\Gamma_{inf}\backslash\mathcal{H}$ has no cusps.

(We note in passing that $S^2\Gamma_{inf}S^{-2} = \Gamma_{inf}$, which shows that $\Gamma_{inf} \lhd \Gamma_\Theta$. This is a start towards identifying the conformal isometries of $\Gamma_{inf}\backslash\mathcal{H}$.)

We are going to transform $\mathcal{F}_{\Gamma_{inf}}$ so that the geometry at 1 and ∞ is clear. This will result in a new fundamental region, which is neither Dirichlet nor Ford, nor even convex for that matter. Indeed it will have some horocyclic sides.

Basically, we use T_0 to map the fundamental horocycle to 0, and then map all the triangles (not geodesic, the tops are horocyclic) anchored at $1 \pm 2m$ to images anchored at 1. First of all, T_0 maps the fundamental horocycle to the (euclidean) disc centered at $i/2$. The images (on smaller horocycles) of the elliptic fixed points of order 2 on $y = 1$ will be denoted $e_{\pm 2m}$, in an obvious abuse of notation. The triangle based at -3 is mapped to one based at $1/3$. Applying the elliptic of order 2 fixing e_{-2} sends this triangle to one anchored at 1. Its top is on a new horocycle, running between e_{-2} and e_{-4} (the latter on the new horocycle). An image of the triangle based at -5 lies just below the image of the one at -3, which we have just moved to 1. An application of the elliptic of order 2 fixing e_{-4} (the one just above), moves this to 1. Continuing in this way gives this fundamental region, see Figure 7. Note that every neighborhood of 1 contains infinitely many sides of this fundamental region, further demonstration that 1 is not a cusp.

We will describe an anti-conformal involution, M, of Γ_{inf} which maps ∞ to 1, showing that these points are geometrically the same. Consider $\mathcal{F}_{\Gamma_{inf}}$. Connect

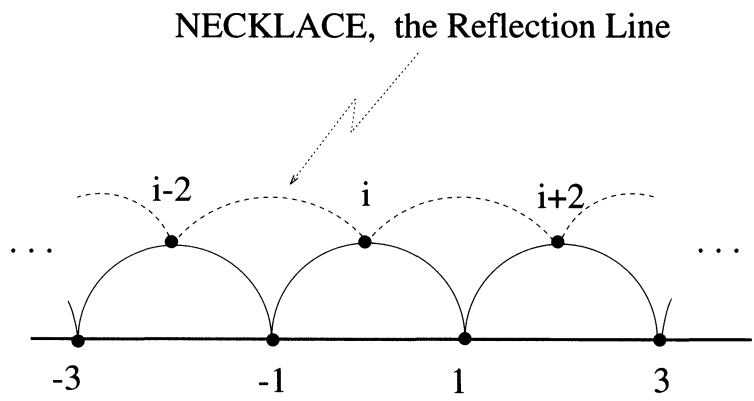

Figure 8. The necklace.

successive elliptic fixed points of order 2 in this fundamental region with h-lines $(i + 2m, i + 2m + 2)$. (This h-arc divides the Γ_Θ fundamental region anchored at $2m + 1$ in two.) Reflect the points of $\mathcal{F}_{\Gamma_{inf}}$ whose real part lie in $[2m, 2m + 2)$ in this h-line. This interchanges the two aforementioned pieces of the Γ_Θ fundamental region anchored at $2m + 1$. Do this for all m. It is not hard to check that this gives a well-defined anti-conformal involution of Γ_{inf}. (The action at $m = 0$ is indeed that given by M of the previous section.) It is also easy to check that the ∞-horocycle $y \geq 1$ is mapped to the 1-horocycle of euclidean center $i + 1$. Thus the geometry near these two points (of the limit set) is the same. Last, the necklace of h-arcs strung between successive elliptic fixed points of order 2 is point-wise fixed by M; this necklace separates the simply connected $\Gamma_{inf}\backslash\mathcal{H}$ into two simply connected pieces. Note that $\Gamma_{inf}\backslash\mathcal{H}$ cannot be isometrically compactified.

5.2. *The Hall Ray on* $\Gamma_{inf}\backslash\mathcal{H}$. In this section we observe that the method of §1 shows that there is Hall ray at ∞ (and therefore 1, by the isometry M) on $\Gamma_{inf}\backslash\mathcal{H}$. The reader may have noted that nothing in that section required the existence of a group element fixing the cusp of each horocycle.

THEOREM. *Let p be either of* 1 *or* ∞, *then* $\Gamma_{inf}\backslash\mathcal{H}$ *has a Hall ray with respect to* p.

This follows in a straight-forward manner from the fact that $\Gamma_{inf} < \Gamma_\Theta$. The only technicality is that each of 1 and ∞ on $\Gamma_{inf}\backslash\mathcal{H}$ projects to ∞ on $\Gamma_\Theta\backslash\mathcal{H}$.

Alternatively, we could note that the sizes, geodesic incidence patterns, and striation arguments for cusped horocycles go over unaffected, word for word. Indeed the argument of section 2 is simplified, because we cannot have infinitely many tangent

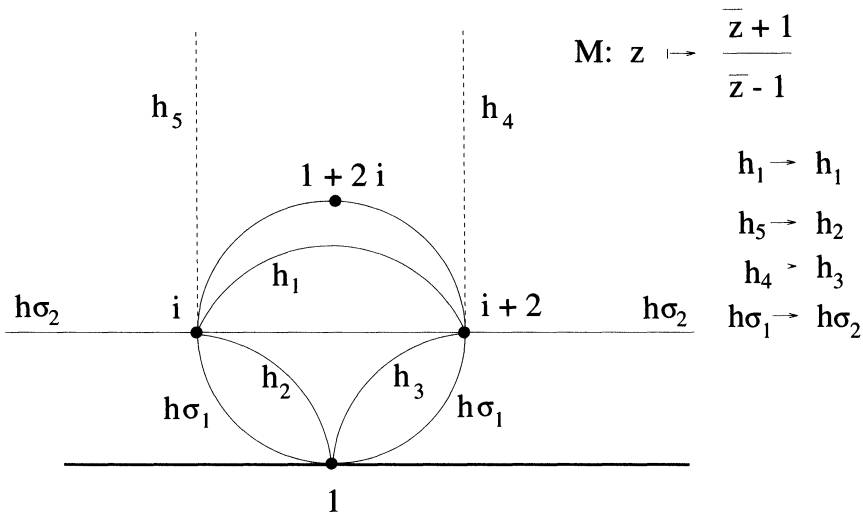

Figure 9. The action of M at $m = 0$.

horocycles all (with 'cusp') having $|c_\gamma| = 1$. Thus Hall rays and punctures are not so intimately related as one might have thought.

The difference geometrically is that on $\Gamma\backslash\mathcal{H}$, in the presence of a parabolic element, a geodesic with large height wraps around the puncture (many times) when it achieves that height; this forces a large number of self-intersections (so-called *parabolic* intersections—see [BLS, H, LS2]) in that part of the geodesic. But the analogous portion of a geodesic on $\Gamma_{inf}\backslash\mathcal{H}$ has no self-intersections, as the entire horocycle is in the fundamental region.

6. Hyperbolic continued fractions

In this section we introduce a hyperbolic analog of the continued fraction algorithm of §1. Whereas the parabolic continued fractions track the depth of penetration of geodesics into horocycles and thus proximity to a distinguished cusp, our hyperbolic continued fractions track proximity (in an appropriate sense) of geodesics to a distinguished simple closed geodesic. They admit a Hall ray phenomenon, as we demonstrate. We intend to study this algorithm in further work on the symbolic dynamics of geodesics. Also of interest is their relationship to Fenchel-Nielsen coordinates in Teichmüller space and to hyperbolic Poincaré series.

Our idea is straightforward. Given a simple closed geodesic γ on a Riemann

surface $\Sigma = \Gamma\backslash\mathcal{H}$, by $SL(2,\mathbb{R})$ conjugation, γ may be considered to lift to the imaginary axis, denoted $\mathcal{I}$. By quasi-conformal deformation, the primitive hyperbolic fixing the imaginary axis is $z \mapsto 2z$. Next, consider a *lens* $\mathcal{L}$ about $\mathcal{I}$—a neighborhood lying between $y = \pm Nx$, for some (large) N. We observe that N may be chosen so that the intersection of $\mathcal{L}$ with the annulus centered at the origin with inner and outer radii 1 and 2 respectively lies in a single fundamental region for the deformed Γ.

The lens $\mathcal{L}$ is the analog of the fundamental horocycle. Its striations are given by intersecting $\mathcal{L}$ with origin-centered annuli, each of radii 2^n and 2^{n+1}, for $n \in \mathbb{Z}$. The images of this lens under Γ are the analogues of the other horocycles. Each is centered at a lift of γ in $\mathcal{H}$; no two intersect. Finding them effectively amounts to solving the conjugacy problem in the deformed group for $\begin{pmatrix} \sqrt{2} & 0 \\ 0 & -1/\sqrt{2} \end{pmatrix}$. These lifts are the simplest of Thurston's laminations. We call the euclidean radius of the h-line in the lens the *radius* of said lens.

Given a geodesic ν of Σ, we choose an h-line lift and form the continued fraction with respect to γ: $(\mathcal{L}_i, f_i)$, where f_i is the number of striations of $\mathcal{L}_i$ which the lift of ν meets.

The appropriate notion of naive and actual height captures the proximity of the trajectory of ν to that of γ; it is then more than a matter of intersection, or near intersection.

Definition. Let z on ν and w on γ be the closest points of the respective geodesics in $\mathcal{L}$. Let ϑ_z and φ_w be the angle (to the horizontal in $\mathcal{H}$, say) made by the geodesics at z and w respectively. Then the *naive height* of ν in $\mathcal{L}$ is $(|z - w| + |\vartheta_z - \varphi_w|)^{-1}$. The *actual height* is the supremum of the naive heights of all lenses encountered by ν.

An h-line with large naive height at the lens about $\mathcal{I}$ has a large diameter and a foot quite near the origin—just as in as the parabolic case. A geodesic with such a lift would have a pair of lift feet close to (i.e. well approximated by) those of (any deformed Γ translate of) $\mathcal{I}$—i.e., the geodesic would seem to run alongside γ on Σ for many 'windings', and this is the same as having a large f for this lens.

We remark that some geodesics will have no continued fraction expansion—they never come close to γ. However, the fact that the geodesic flow is mixing makes such geodesics highly non-generic.

Sliding of h-lines is now done by means of transformations $H_a : z \mapsto az$, thus preserving the naive height in the lens about $\mathcal{I}$. In one respect, the lenses are easier to work with: Whereas the intervals consisting of 'shadows' cast on the real axis by our horocycles can overlap in annoying ways, this is not so for the intervals between the feet of our lenses—this by simplicity of the original geodesic.

We gather some useful calculations:

LEMMA. *Let* $V = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *and* N *as above, then*

$$V(x + iNx) = \frac{(ax+b)(cx+d) + acN^2x^2}{(cx+d)^2 + c^2N^2x^2} + \frac{iNx}{(cx+d)^2 + c^2N^2x^2}.$$

The point as above whose image under V *has largest imaginary part has* $x = \pm d/c\sqrt{N^2+1}$; *this maximal height is* $\approx 1/2cd(1 + N^{-1} + N^{-2})$. *The imaginary part of* $V(iy)$ *is* $|c/(c^2 + d^2y^2)|$, *which is maximized at* $|1/2cd|$ *when* $y = c/d$; *the height is* $1/2cd$. *The imaginary part of* $V(i2^nc/d)$ *is* $1/2cd(2^{n-1} + 2^{-n-1})$. *The feet of the lens* $V(\mathcal{L})$ *are at* a/c *and* b/d.

This allows us to construct vertical h-lines with small 'partial denominators,' f_i. Indeed, temporarily ignoring the initial infinite segment of the vertical in the lens at $\mathcal{I}$, given $T > 100$, we construct a vertical which enters the first lens for $f_1 \geq T$ striations, but whose subsequent f_i are all less than 10. (Note that sliding such a line by H_a produces another vertical h-line.)

This is similar to the parabolic case, thus we simply sketch the argument. Begin with a vertical line that encounters a (first) lens upon leaving the lens at $\mathcal{I}$ with $f_1 = 100$ and crossing through this lens. (That is, the vertical hits, and so crosses, a lift of γ.) Next, note that the lemma guarantees that the next lens encountered will have radius $1/2c_2d_2 \leq \delta_N(1/2c_1d_1)$, for some $\delta_N < 1$. (This amounts to N being a constant depending only on Γ and γ.) We would like to slide toward the peak (euclidean) of this next lens, if necessary, to make $f_2 \leq 10$. This poses a problem only if the second lens is almost as large as the first. Should this be so, simply slide toward the nearest foot of the first lens (increasing T) and avoid the second lens altogether! Subsequent lenses encountered will now have small radii and sliding toward their centers cannot affect $f_1 \geq T$. We remark that all sliding takes place within the first lens encountered, and as we remarked, no new lens above this one may be entered by such a slide.

In the case of $\Gamma^3 \subset SL(2, \mathbb{Z})$ (the group generated by cubes of elements of the modular group, see [S2] for the associated geometry), such a vertical h-line would terminate in a real number having a single excellent approximant by certain quadratic irrationalities—the analogue of a continued fraction with a single large partial denominator. To be specific:

Example. The shortest geodesic on $\Gamma^3\backslash\mathcal{H}$ is fixed by $\left(\begin{smallmatrix} 1 & -1 \\ -1 & 2 \end{smallmatrix}\right)$ and is perforce simple. A particular lift runs between the Golden Means $\eta, \eta' := (1 \pm \sqrt{5})/2$.) By choosing vertical lines terminating near η, and then sliding these lines, can obtain real numbers α with a single excellent approximant, of the form $(a\eta + b)/(c\eta + d)$, where $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma^3$. If we wish to expand the set of approximants to all $\Gamma(1)$ images of η—i.e., to ask which real numbers can be have an excellent approximant from the modular group, we answer that one of $\alpha, \alpha + 1, \alpha + 2$ has an excellent approximant in $\Gamma^3(\eta)$. (There are no simple geodesics on $\Gamma(1)\backslash\mathcal{H}$, of course.)

In fact, the lemma computations make it possible to construct a hyperbolic Hall ray. This is done in exact analogy with the parabolic case. As we remarked above, lenses are more tractable than horocycles.

THEOREM. *Let γ be a simple closed geodesic on a Riemann surface Σ. The set of heights with respect to γ of the geodesics of Σ contains a real half-line.*

Proof (sketch). Without loss of generality, assume γ lifts to $\mathcal{I}$. Begin with an h-line of naive height greater than 100 for the $\mathcal{I}$ lens. We will show that this h-line can be slid so as to have all (other) partial denominators $f_i \leq 10$, $i \neq 0$. As we noted, this amounts to an upper bound (depending only on γ and Γ, but not ν) for the naive heights of ν in $\mathcal{L}_i$. Thus the actual height of the limit geodesic will be the original naive height, left unchanged by the slides, and we will have our Hall ray.

All quantitative estimates necessary for the following are in the lemma above. Steps A through C are exactly as in the parabolic case. In D, the radius is $1/2c_id_i$. In E, the striations are located at heights $1/2c_id_i(2^{b_i-1} + 2^{-b_i-1})$. Thus the dependence on $1/2c_id_i$ factors out. So again, the fractional division points on the lens do not depend on the radius. This is true for the lens width as well. Also then, by increasing N at the outset, this width can be ignored in continued fraction computations; we can work with the lift of γ itself.

Justified by the factoring out mentioned in the previous paragraph, we normalize with $1/2c_id_i = 1$ and a lens having feet at the points 0 and 2 to simplify and clarify calculation. A simple estimate shows that $|f_i| \leq 10$ forces our geodesic to terminate within .98 of 1. (In other words, only 2% of termini require any slide.)

As $\mathcal{L}_{i+1}$ has radius smaller than $\mathcal{L}_i$, we can clearly slide to the peak of $\mathcal{L}_{i+1}$, unless that peak is too close to a foot of $\mathcal{L}_i$. In that case the radius of $\mathcal{L}_{i+1}$ is tiny compared to that of $\mathcal{L}_i$ and we slide towards 1, avoiding the $\mathcal{L}_{i+1}$ lens altogether, while sustaining $f_i \leq 10$. The worst case is that of two lenses of radii about $1/2$ located beneath $\mathcal{L}_i$; and here we can indeed slide to the peak of $\mathcal{L}_{i+1}$. The issue of the sliding at one foot of our geodesic harming the continued fraction expansion at the other foot is handled exactly as in the parabolic case.

One final remark on hyperbolic continued fractions. There is no reason not to use them *in tandem* with the parabolic fractions. Judicious choice of the sizes of horocycles and lenses can ensure that neither intersects the other. On surfaces like $\Gamma'\backslash\mathcal{H}$, it is reasonable to suppose that the horocycle and the lens 'almost' fill out the surface—meaning that the tracking of geodesics by the combined mechanism would be quite exact. We intend to return to this in a subsequent paper.

REFERENCES

[B] A. Beardon, *The Geometry of discrete groups*, Graduate Texts in Mathematics, no. 91, Springer-Verlag, New York 1983.

[BLS] A. Beardon, J. Lehner and M. Sheingorn, *Closed geodesics on a Riemann surface with application to the Markoff spectrum*, TAMS **295** (1986), 635–647.

[H] A. Haas, "Geometric Markoff theory and a theorem of Millington" in *Number theory with an emphasis on the Markoff spectrum*, A. Pollington and W. Moran, eds. Dekker, New York, 1993, pp. 107–112.

[J] C. Judge, *On the angular moduli of constant curvature surfaces with conic singularities*, Indiana University, 1993, preprint.

[L] J. Lehner, *Discontinuous groups and automorphic functions*, Amer. Math. Soc., Providence, RI, 1964.

[LS] J. Lehner and M. Sheingorn, *A symbolic dynamics for geodesics on punctured Riemann surfaces*, Math. Ann. **268** (1984), 425–448.

[LS2] J. Lehner and M. Sheingorn, *Simple closed geodesics on* $\Gamma(3)$ *arise from the Markoff spectrum*, BAMS **11** (1984), 359–362.

[RS] D. Rosen and T. A. Schmidt, *Hecke groups and continued fractions*, *Bull. Austral. Math. Soc.* **46** (1992) 459–475.

[SS1] T. Schmidt and M. Sheingorn, *Length spectra for Hecke triangle surfaces*, Math. Z. **220** (1995), 369–397.

[SS2] T. Schmidt and M. Sheingorn, *On the infinite volume Hecke surfaces*, Compositio Math. **95** (1995), 247–262.

[S] M. Sheingorn, "Low height Hecke triangle group geodesics" in *A tribute to Emil Grosswald*, a volume of contributed papers edited by Marvin Knopp and Mark Sheingorn, Contemp. Math., no. 143, AMS, Providence, RI, 1993, pp. 545–560.

[Sh] H. Shimizu, *On discontinuous groups operating on the product of upper half planes*, Ann. of Math. **77** (1963), 33–71.

Thomas A. Schmidt, Oregon State University, Corvallis, OR 97331
toms@math.orst.edu

Mark Sheingorn, CUNY - Baruch College, New York, NY 10010
marksh@panix.com