TRANSACTIONS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 295, Number 2, June 1986

CLOSED GEODESICS ON A RIEMANN SURFACE WITH APPLICATION TO THE MARKOV SPECTRUM

A. F. BEARDON, J. LEHNER AND M. SHEINGORN[1]

ABSTRACT. This paper determines those Riemann surfaces on which each nonsimple closed geodesic has a parabolic intersection—that is, an intersection in the form of a loop enclosing a puncture or a deleted disk. An application is made characterizing the simple closed geodesic on $H/\Gamma(3)$ in terms of the Markov spectrum.

The thrust of the situation is this: If we call loops about punctures or deleted disks *boundary curves*, then if the surface has "little" topology, each nonsimple closed geodesic must contain a boundary curve. But if there is "enough" topology, there are nonsimple closed geodesics not containing boundary curves.

1. Let R be a Riemann surface of genus g with k punctures and d disks removed; $k, d \geq 0$, $k + d > 0$. In this paper we consider the closed geodesics on R.

The method used is to represent R as a quotient, $R = H/\Gamma$, where $H = \{z = x + iy: y > 0\}$ and Γ is a fuchsian group acting on H. Let $\pi: H \to H/\Gamma$ be the projection map. We assume (H, π) is an unramified covering, so Γ has no elliptic elements. For most of this paper we shall assume $k > 0$ and $d = 0$; the remaining cases are considered in §5. Each hyperbolic axis projects to a closed geodesic in R. Each closed geodesic α in R lifts to a conjugacy class $[\alpha]$ in Γ and under known conditions $[\alpha]$ is hyperbolic.

Let $\gamma \in \Gamma$ be hyperbolic with axis A_γ. From now on we assume γ primitive; i.e., γ generates the stabilizer of A_γ. It is easily seen that

$$\pi(A_\gamma) \text{ is simple iff } A_\gamma \cap \beta A_\gamma = \varnothing \text{ for all } \beta \in \Gamma - \langle\gamma\rangle.$$

We write $A_\gamma \wedge \beta A_\gamma$ to mean that $A_\gamma \cap \beta A_\gamma = \{z\}$ for some z in H. Thus

$$\begin{aligned} &\pi(A_\gamma) \text{ is nonsimple (= self-intersecting) iff} \\ &A_\gamma \wedge \beta A_\gamma \text{ for some } \beta \in \Gamma - \langle\gamma\rangle. \end{aligned} \tag{1.1}$$

Since $\alpha A_\gamma = A_{\alpha\gamma\alpha^{-1}}$, conjugate elements are both simple or both nonsimple. We use the phrase a hyperbolic γ is simple (nonsimple) to mean $\pi(A_\gamma)$ is simple (nonsimple).

It is readily checked that if γ is nonsimple, we can choose β in (1.1) to be hyperbolic; indeed, we may replace β by $\beta\gamma^n$ for large n. However, it is not always possible to choose β to be *parabolic*, and in this connection we prove the following result.

Received by the editors Jaunary 8, 1985.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 30F35; Secondary 11J06.
[1]Third author partially supported by NSF and PSC-CUNY grants.

THEOREM 1.1. *Let R and Γ be as above and let $d = 0$. Then it is possible to select β in (1.1) to be parabolic for every nonsimple γ if and only if $g = 0$ and $k = 3$ or 4.*

The area formula for a fushsian group shows that $g = 0$ only if $k \geq 3$. For example, $\Gamma(2)$ and $\Gamma(3)$ have $g = 0$ and $k = 3$ and 4, respectively, so Theorem 1.1 applies to these groups. (Here $\Gamma(n)$ is the subgroup of the modular group consisting of all modular matrices $V \equiv \pm I \pmod{n}$.)

In §2 we develop a connection with the Markov spectrum of diophantine approximation. To introduce the Markov spectrum, let θ be a real irrational and define

$$M(\theta) = \sup_{c>0} \left\{ c: \left| \theta - \frac{p}{q} \right| < \frac{1}{cq^2} \text{ for infinitely many reduced fractions } \frac{p}{q},\ q > 0 \right\}. \tag{1.2}$$

We are interested only in the range $M(\theta) < 3$, and in this range $M(\theta)$ assumes only a countably infinite set of values M_ν. The set $\{M_\nu\}$, ordered by magnitude $(M_1 \leq M_2 \leq \cdots)$, is called the Markov spectrum and abbreviated to MS. Moreover, $M_\nu \to 3$ and, for each ν, there is a real quadratic irrational ξ_ν such that $M_\nu = M(\xi_\nu)$. The ξ_ν are known explicitly, and there is a primitive hyperbolic matrix γ_ν in $\Gamma(3)$ that fixes ξ_ν and its conjugate ξ_ν'. (See §2 for details.)

THEOREM 2.1. *Let $\gamma \in \Gamma(3)$ be hyperbolic with fixed points ξ_γ, ξ_γ'. Then $\pi(A_\gamma)$ is simple if and only if the fixed points ξ_γ, ξ_γ' of γ satisfy $M(\xi_\gamma) = M(\xi_\gamma') < 3$.*

This result, in particular, specifies the lengths of the simple closed geodesics in $H/\Gamma(3)$. Indeed, if γ is a transformation for which $M(\xi_\gamma) < 3$, then $\pi(A_\gamma)$ has length L where

$$\operatorname{trace} \gamma = 2 \cosh\left(\tfrac{1}{2} L\right). \tag{1.3}$$

Using an estimate of C. Gurwood given in [**5**], this leads to

COROLLARY 2.2. *Let $N_s(X) = \#\{$simple closed geodesics on $H/\Gamma(3)$ with length $\leq X\}$. Then $X \ll N_s(X) \ll X^2$.*

This work grew out of the characterization of the simple closed geodesics on $H/\Gamma(3)$ given by Lehner and Sheingorn in [**4**]. $H/\Gamma(3)$ is a four-times punctured sphere and the original proof of the result in [**4**] relied on the geometry of that surface. Beardon pointed out that these geometric arguments could largely be replaced by topological ones and that these new arguments applied to more general Riemann surfaces. He also saw that the new arguments had converses—that is, they pertained exactly to the surfaces with $g = 0$ and $k + d = 3$ or 4.

At about the same time, Haas was achieving his characterization of the simple geodesics on punctured tori [**3**]. He does this using binary quadratic forms studied by Cohn [**1**] and Schmidt [**6**]. For H/Γ', his characterization (in the closed case) is the same as that of [**4**]. His paper is more general than [**4**] in two ways: (i) He treats the closure of the set of simple geodesics (closed or infinite); (ii) his results apply to the signature class $(1, \infty)$, not just H/Γ'. The article of Series [**7**] is a nice

exposition of the work of Cohn, Schmidt and (decisively) Haas on simple closed geodesics on H/Γ'.

After Haas personally communicated his work to Sheingorn, the latter realized that the characterizations were the same for $H/\Gamma(3)$ and H/Γ' because they both stemmed from a characterization of the simple closed geodesics on H/Γ^3—the crucial point being that Γ^3 contains both Γ' and $\Gamma(3)$ as "large" subgroups. Again this can be generalized, à la Haas, to entire signature classes [8].

2. In this section we prove Theorem 2.1 and Corollary 2.2, which connect closed simple geodesics in $H/\Gamma(3)$ with the Markov spectrum (MS). We include a brief synopsis of known facts about MS; for a fuller account see [**2**, pp. 29–33].

In (1.2) and the following lines we defined MS to be the set of values $\{M_\nu\}$ assumed by the Markov function $M(\theta)$ in the range $M(\theta) < 3$. In order to calculate M_ν we introduce *Markov triples.* A triple of positive integers (x, y, z) is called a Markov triple if

$$x^2 + y^2 + z^2 = 3xyz, \qquad 1 \le x \le y \le z.$$

The first triples are $(1,1,1), (1,1,2), (1,2,5), \ldots$, and the rest can be recursively generated. Order the triples by the size of z, so that $1 = z_1 \le 2 = z_2 \le \cdots \le z_\nu \le \cdots$. With each triple (x_ν, y_ν, z_ν) there is associated a pair of real quadratic conjugates

$$\theta_\nu, \theta'_\nu = \frac{1}{2} + \frac{y_\nu}{x_\nu z_\nu} \pm \frac{1}{2}\left(9 - \frac{4}{z_\nu^2}\right)^{1/2}, \qquad \nu \ge 1.$$

It is a known theorem that $M(\theta_\nu) = M(\theta'_\nu) = M_\nu$, and that

$$M(\theta_\nu) = M_\nu = |\theta_\nu - \theta'_\nu| = (9 - 4/z_\nu^2)^{1/2}.$$

We have $M_1 = 5^{1/2}$, $M_2 = 8^{1/2}$, $M_3 = (221)^{1/2}/5, \ldots, \to 3$.

Next, introduce the equivalence relation:

(2.1) $\theta \sim \psi$ if $\psi = (a\theta + b)/(c\theta + d)$ with integers a, b, c, d and $ad - bc = \pm 1$.

Then

(2.2) $\theta \sim \psi \Rightarrow M(\theta) = M(\psi)$; in particular, $M(V\theta) = M(\theta)$ for $V \in \Gamma(1)$.

To understand why this is true, we adopt the usual notation for the regular continued fraction (CF)

$$\theta = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}} = [a_0, a_1, a_2, \ldots].$$

(Here $a_i \ge 1$ for $i \ge 1$.) Then it is well known that $\theta \sim \psi$ means that the CF expansions of θ and ψ agree from some point on. That is, there exist n_0 and m_0 such that $a_{n_0+k} = b_{m_0+k}$, $k = 0, 1, 2, \ldots$. (Here $\psi = [b_0, b_1, b_2, \ldots].$) Next, it can be shown that [**2**, p. 29]

$$M(\theta) = \overline{\lim_k}\{[a_{k+1}, a_{k+2}, \ldots] + [0, a_k, a_{k-1}, \ldots, a_1]\}.$$

Given this, (2.2) is clear.

We can associate the MS to hyperbolic elements of $\Gamma(3)$. For each ν there is a $\gamma_\nu \in \Gamma(3)$ whose fixed points are $\xi_{\gamma_\nu} = \theta_\nu$, $\xi'_{\gamma_\nu} = \theta'_\nu$. Namely, dropping the subscript ν, let $\varsigma = 1$ if z is odd; otherwise $\varsigma = 2$. Define

$$B = \begin{pmatrix} (N + x(2y + xz)M\varsigma^{-1})/2 & (2x^2z - 4xy + z)M\varsigma^{-1} \\ x^2zM\varsigma^{-1} & (N - x(2y + xz)M\varsigma^{-1})/2 \end{pmatrix}, \tag{2.3}$$

where $M > 0$ is the smallest integral solution of the Pell equation

$$x^4(9z^2 - 4)\varsigma^{-2}M^2 + 4 = N^2. \tag{2.4}$$

Then it can be shown that B is the $\Gamma(1)$-primitive matrix fixing ξ, ξ'. Moreover, $B \in \Gamma(3)$ if $3|M$; otherwise $B^2 \in \Gamma(3)$. But the first case never occurs, so B^2 is the $\Gamma(3)$-primitive matrix fixing ξ, ξ'. (See [**4**, **8**].)

Let $\mathcal{M} = \{\gamma \in \Gamma(3)\colon M(\xi_\gamma) < 3\}$, so Theorem 2.1 may be restated as follows: Let $\gamma \in \Gamma(3)$ be hyperbolic. Then $\pi(A_\gamma)$ is simple if and only if $\gamma \in \mathcal{M}$. Note that $\gamma \in \mathcal{M} \Rightarrow V\gamma V^{-1} \in \mathcal{M}$, $V \in \Gamma(1)$. Indeed $V\gamma V^{-1} \in \Gamma(3)$ by normality of $\Gamma(3)$ in $\Gamma(1)$. And since $\xi_{V\gamma V^{-1}} = V\xi_\gamma$, $M(\xi_{V\gamma V^{-1}}) = M(V\xi_\gamma) = M(\xi_\gamma) < 3$. That is, the conjugacy class of γ in $\Gamma(1)$ belongs to $\mathcal{M}$ if $\gamma \in \mathcal{M}$.

We shall now prove Theorem 2.1. Suppose $\pi(A_\gamma)$ is nonsimple; we wish to show that $\gamma \notin \mathcal{M}$. By Theorem 1.1 and the normality of $\Gamma(1)$, there is a $\delta \in \Gamma(3)$ conjugate to γ in $\Gamma(1)$ for which $A_\delta \wedge S^3A_\delta$; i.e., $|\xi_\delta - \xi'_\delta| > 3$. Here $S = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Denote the periodic continued fraction $[a_0, \ldots, a_{\mu-1}, b_\mu, b_{\mu+1}, \ldots, b_{\mu+k-1}, b_\mu, \ldots, b_{\mu+k-1}, \ldots]$ by $[a_0, \ldots, a_{\mu-1}, \overline{b_\mu, \ldots, b_{\mu+k-1}}]$. By translation in $\Gamma(1)$ we may assume $-1 < \xi'_\delta < 0$. Then $\xi_\delta > \xi'_\delta + 3 > 1$. It is well known that under these conditions CF ξ_δ is pure periodic [**5**, p. 75]. Let $\xi_\delta = [\overline{b_0, b_1, \ldots, b_{k-1}}]$ for a $k \geq 1$; then $-1/\xi'_\delta = [\overline{b_{k-1}, \ldots, b_1}]$ by [**5**, p. 76]. Set

$$m_\mu = [\overline{b_\mu, b_{\mu+1}, \ldots, b_{\mu+k-1}}] + [0, \overline{b_{\mu-1}, b_{\mu-2}, \ldots, b_{\mu-k}}], \qquad \mu \geq k.$$

By periodicity $m_\mu = m_{\mu+k}$, so we may restrict μ to the range $k+1 \leq \mu \leq 2k$.

Moreover,

$$M(\xi_\delta) = \overline{\lim_{\mu\to\infty}} \{[b_\mu, b_{\mu+1}, \ldots] + [0, b_{\mu-1}, \ldots, b_1]\}.$$

Now for $\mu = nk + \nu$, $0 \leq \nu < k$,

$$[0, b_{\mu-1}, \ldots, b_1] = [0, B, \ldots, B, b_{\nu-1}, \ldots, b_1],$$

where B is the block $b_{\mu-1}, \ldots, b_{\mu-k}$ and there are n B's. Hence as $\mu \to \infty$, i.e., $n \to \infty$, the right member tends to $[0, \overline{b_{\mu-1}, \ldots, b_{\mu-k}}]$, so

$$M(\xi_\delta) = \overline{\lim_{\mu\to\infty}}\, m_\mu = \max_{k+1\leq\mu\leq 2k} m_\mu.$$

Hence for n sufficiently large,

$$3 < \xi_\delta - \xi'_\delta = m_k \leq \max_\mu m_\mu = M(\xi_\delta),$$

implying $M(\xi_\gamma) = M(\xi_\delta) > 3$. That is, $\gamma \notin \mathcal{M}$, as asserted.

Conversely, assume $\gamma \notin \mathcal{M}$. Then $M(\xi_\gamma) \geq 3$; in fact $M(\xi_\gamma) > 3$, since ξ_γ is a quadratic irrational and $M(\theta)$ cannot equal 3 for any quadratic irrational θ [**2**, p. 32]. Hence

$$|\xi_\gamma - p_n/q_n| < 1/(3 + h)q_n^2, \qquad (p_n, q_n) = 1,$$

for some $h > 0$, on a sequence $q_n \to \infty$.

Write $V_n = (q_n', -p_n' : q_n, -p_n) \in \Gamma(1)$. Then with $\xi_\gamma = \xi$, $\xi_\gamma' = \xi'$,

$$\begin{aligned} |V_n\xi - V_n\xi'| &= \frac{|\xi - \xi'|}{q_n^2|\xi - p_n/q_n|\,|\xi' - p_n/q_n|} \\ &> \frac{(3+h)|\xi - \xi'|}{|\xi' - p_n/q_n|} \geq \frac{(3+h)|\xi - \xi'|}{|\xi' - \xi| + |\xi - p_n/q_n|} \\ &> \frac{3+h}{1 + 1/3q_n^2|\xi - \xi'|} > 3 \end{aligned}$$

for $n \geq n_0$. For $V = V_{n_0}$ we have

$$|V\xi_\gamma - V\xi_\gamma'| > 3. \tag{2.5}$$

Next, set $\beta = V^{-1}S^3V$, a parabolic element in $\Gamma(3)$. Since $V\xi_\gamma, V\xi_\gamma'$ are the fixed points of $VA_\gamma V^{-1}$, (2.5) gives $A_{V\gamma V^{-1}} \wedge S^3 A_{V\gamma V^{-1}}$. But $A_{V\gamma V^{-1}} = VA_\gamma$, so $A_\gamma \wedge V^{-1}S^3VA_\gamma$, i.e., $A_\gamma \wedge \beta A_\gamma$. By (1.1), $\pi(A_\gamma)$ is not simple. This completes the proof.

We turn now to the proof of Corollary 2.2. Assume that γ is a primitive hyperbolic in $\mathrm{SL}(2, \mathbf{Z})$ with positive trace and of the form (2.3) corresponding to the Markov triple (x, y, z). Using (1.3), we have

$$3z/4 \leq [x^2(9z^2 - 4)\varsigma^{-2}M^2 + 4]^{1/2}/\varsigma = \mathrm{trace}(\gamma) \leq e^{L/2} + 1,$$

so the number of γ with $L \leq X$ is not more than the number of γ with $z \leq 2e^{X/2}$. Using the estimate of this by C. Gurwood, noted in [5], we obtain $N_s(X) \ll X^2$ with an explicit constant available.

To obtain the lower bound, we simply estimate the number of γ with $L \leq X$, which, in addition, correspond to a triple $(1, y, z)$. We can generate a sequence γ_n of elements with distinct traces of the triples $(1, y_{n+1}, z_{n+1}) = (1, z_n, 3z_n - y_n)$, so $z \leq 3^n$. Moreover, we have, by (2.4),

$$(9z_n^2 - 4)(M/\varsigma)^2 + 4 = N^2,$$

which has the fundamental solution $M = \varsigma$, $N = 3z_n$. Thus $\mathrm{trace}(\gamma_n) = 3z_n/\varsigma$. If $A = (\log\varsigma)/(\log 9) - 1$, $L_n = L(\gamma_n)$, then (1.3) yields

$$\tfrac{1}{2}e^{L_n/2} \leq 3^{n+1}/\varsigma.$$

If we choose any $n < (X + \log\varsigma)/(\log 9) - 1$, then we have that $L_n < X$. These values of n, then, give distinct γ_n with $L_n < X$, and so

$$N_s(X) > X(\log 9) - 1.$$

This completes the proof of Corollary 2.2.

3. In this section we establish the positive assertion only, namely that if R $(= H/\Gamma)$ has genus zero and three or four punctures, then for every nonsimple hyperbolic γ in Γ there is a parabolic β in Γ with $\beta(A_\gamma)$ crossing A_γ. We begin by selecting a nonsimple hyperbolic γ in Γ, and there is no loss of generality in supposing that γ is primitive. Next we make some preliminary remarks about the axis A_γ and its projection $\pi(A_\gamma)$ on R. As A_γ is topologically the real line, we can define in a natural way the order relations $<$ and $\leq$ on A_γ, so, for example, if z is

on A_γ, then $z < \gamma(z)$. The meaning of $[z, w]$, (z, w), $[z, w)$ and $(z, w]$ as intervals on A_γ is self-evident.

Suppose now that the axis A_γ contains a nontrivial segment $[z, \alpha z]$ with α in Γ; thus $c = \pi([z, \alpha z])$ is a closed curve on R. If c is homotopic in R to a point of R, then, by the Monodromy Theorem, the lift from z, namely $[z, \alpha z]$, is closed. Since Γ has no elliptics, α must then be the identity, which is false. We record this result as

(3.1) no closed subarc of $\pi(A_\gamma)$ is homotopic in R to a point of R.

Next suppose that the axis A_γ contains a nontrivial segment $[z, \alpha z]$ with the closed curve $c = \pi([z, \alpha z])$ freely homotopic in R to some curve c' in a disc N (on the sphere $S = R \cup (\text{punctures})$) containing exactly one puncture. The lift of N is a maximal set of distinct Γ-equivalent open horocycles. If the hypothesis is true for some such N, then it is true for all smaller discs, and by taking N sufficiently small, these distinct horocycles are pairwise disjoint. The (connected) lift σ' of c' (from any appropriate point) therefore lies in precisely one, say Q, of these horocycles. This choice of N, however, means that if β is in Γ then $\beta(Q) = Q$ precisely when β is parabolic (or the identity) and stablizes Q. We deduce that the endpoints of σ' are, say, w and βw with β parabolic or the identity. It is well kown that as c and c' are freely homotopic, the endpoints of any lift of c must be of the form w' and $\beta' w'$, say, where β' is conjugate to β. Now (3.1) prevents β from being the identity, for if it is then so is β'; σ' is then homotopic to a point in Q, and consequently c', and hence c, is homotopic to a point of R. We deduce the following:

(3.2) if a closed subarc of $\pi(A_\gamma)$ is freely homotopic to a curve in a disc of S which contains exactly one puncture, then $\beta(A_\gamma) \cap A_\gamma \neq \varnothing$ for some parabolic β in Γ.

Observe that the conclusion of (3.2) is the result we are seeking.

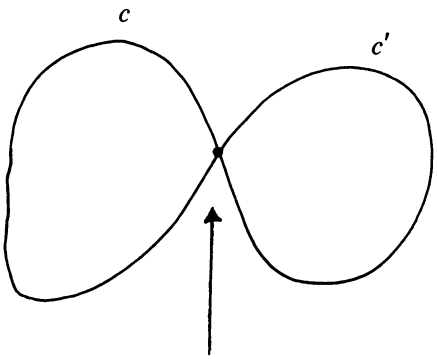
Next, suppose that, for some α and β in Γ, A_γ contains two nontrivial segments $[z, w]$ and $[\alpha z, \beta w]$ with π-images c and c' respectively such that c and c' are homotopic in R. Note that $c' = \pi([z, \alpha^{-1}\beta w])$. By the Monodromy Theorem, the lifts of c and c' from z have the same endpoint, so $\alpha = \beta$. Then α maps A_γ (the geodesic through z and w) to itself (the geodesic through αz and αw); hence, as γ is primitive, $\alpha \in \langle \gamma \rangle$. This shows that

(3.3) no two distinct nontrivial segments $[z, w]$ and $[\alpha z, \beta w]$ of a fundamental segment of A_γ have homotopic π-images in R.

Our final remark guarantees the existence of simple loops in A_γ. Consider a fundamental segment $[z, \gamma z]$ of A_γ. By assumption, its projection is a nonsimple loop, so there exist equivalent points z_1 and z_1' with $z \leq z_1 < z_1' < \gamma z$. Write $z_1' = \gamma_1 z_1$ and repeat the argument for $[z_1, \gamma_1 z_1]$. This process must terminate; else we could find a sequence of pairwise disjoint equivalent points in $[z, \gamma z]$. Using discreteness, one can show that infinitely many of these must be paired by the same element, say δ, of Γ. Thus (as in the proof of (3.3)) δ is in $\langle \gamma \rangle$, and this is clearly false. Thus we have proved

(3.4) $\pi(A_\gamma)$ contains a simple closed subarc.

CASE I



$$\pi(z) = \pi(\alpha(z)) = \pi(w) = \pi(\beta(w))$$

FIGURE 1

We can now complete the proof of Theorem 1.1. First, $\pi(A_\gamma)$ contains a simple closed subarc c. Now c is a Jordan curve on the sphere S and so has two complementary Jordan domains D and D'. By (3.1), D and D' each contain at least one puncture. If there are exactly three punctures (i.e., if $k = 3$), then one of D and D' contains exactly one puncture, and the conclusion follows from (3.2).

If $k = 4$ (which we now assume), then D and D' may each contain two punctures, and we need to refine this argument. As $\pi(A_\gamma)$ contains a simple closed subarc, we can construct a fundamental segment $[z, \gamma z]$ of A_γ containing a point αz (with α in Γ) with π 1-1 on $(z, \alpha z]$. The simple closed subarc is $\pi([z, \alpha z])$. Observe now that $(z, \alpha z] \subset (z, \gamma z)$ and that π is 1-1 on $(z, \alpha z]$ but not on $(z, \gamma z]$ (else γ is simple). It follows that there is a first point w' after αz and before γz which is equivalent to some point w in (z, w'). Write $w' = \beta w$. Thus we have

$$z < \alpha z < \beta w < \gamma z; \quad z < w < \beta w.$$

Three possibilities now arise, and we shall give a proof in each case. The easiest is

Case I: $w = \alpha z$ (Figure 1). Our assumptions imply that

$$c = \pi([z, \alpha z]) = \pi([z, w]) \quad \text{and} \quad c' = \pi([\alpha z, \beta w]) = ([w, \beta w])$$

are simple closed subarcs of $\pi(A_\gamma)$ which meet at one point only, namely the common projection of $z, \alpha z, w$ and βw. Thus $c \cup c'$ is a "figure eight" curve on S, and this has three complementary domains on S. Two of these are discs (bounded by c and by c' respectively); the third is a topological disc bounded by $c \cup c'$. If our assertion is false, then these must contain at least 2, 2 and 1 punctures respectively (see (3.1), (3.2) and (3.3)), and so $k \geq 5$. As $k = 4$, the proof for Case I is complete.

It is worth noting that the figure eight curve (or a variant of it) dominates all stages of the proof of Theorem 1, and the reader may find it helpful to bear this in mind.

CASE II

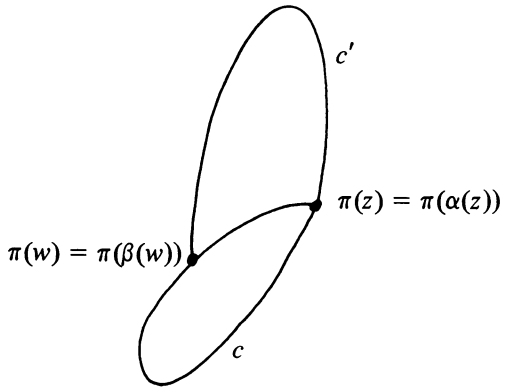


FIGURE 2

CASE III

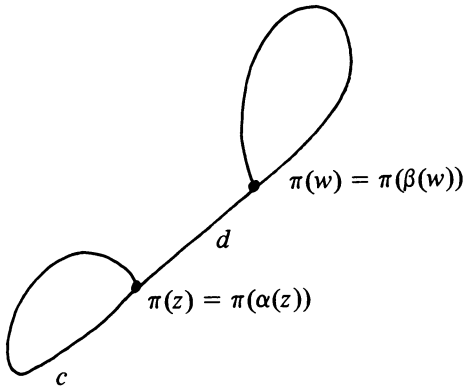


FIGURE 3

Case II: $w \in (z, \alpha z)$ (Figure 2).

The curve $c = \pi([z, \alpha z])$ is a Jordan curve which contains the two distinct points $\pi(z)$ and $\pi(w)$. The curve $c' = ([\alpha z, \beta w])$ joins these two points. Moreover, by assumption, c' is simple and does not meet c at any points other than $\pi(z)$ and $\pi(w)$. Thus the complement of c in S is the union of two discs, say V and W, and c' is a simple cross-cut of W, say. It follows that the complement of $c \cup c'$ on S is the disjoint union of three Jordan domains bounded respectively by the π-images of $[z, \alpha z], [z, w] \cup [\alpha z, \beta w], [w, \beta w]$. By (3.1) to (3.3), the result can only fail if these domains contain 2, 1, 2 punctures respectively, and this cannot be so.

Case III: $w \in (\alpha z, \beta w)$ (Figure 3).

The π-image of $[z, \beta w]$ consists of the disjoint union of the simple loop $c = \pi([z, \alpha z])$, the simple arc $d = \pi((\alpha z, w))$ and the simple loop $c' = \pi([w, \beta w])$. By construction, d joins the first (and last) point of c to that of c'. Thus the complement on S of these three curves consists of three Jordan domains, say U (bounded by c), V (bounded by c') and W (a simply connected domain bounded by the two curves $c \cup d$ and $d \cup c'$, each joining $\pi(z)$ to $\pi(w)$). Again, if the result fails, then these must contain at least 2, 2 and 1 punctures, and the proof of the "if" part of Theorem 1.1 is complete.

4. We now complete the proof of Theorem 1.1. Recall that R $(= H/\Gamma)$ is a Riemann surface of genus g with k punctures. Since Γ has no elliptic elements, Γ is isomorphic to the fundamental group Γ^* of R, and in view of this we shall regard Γ and Γ^* as being the same group; for example, we shall speak of parabolic elements of Γ^*. Likewise, any homomorphism $\theta\colon \Gamma^* \to \mathbf{Z}$ may be regarded as being defined on Γ: if x in Γ^* corresponds to γ in Γ, then $\theta(x) = \theta(\gamma)$.

In this section we verify that if $g \geq 1$, or if both $g = 0$ and $k \geq 5$, then Γ contains a primitive nonsimple hyperbolic γ such that $\beta(A_\gamma) \cap A_\gamma = \varnothing$ for every parabolic β in Γ. The proofs for $g = 0$ and for $g \geq 1$ follow a similar argument, which we now describe, but are different in their details. We construct a closed curve x on R (actually a figure eight) with certain specified geometric properties. Now x represents an element of Γ^*, and this lifts from z to γz for some γ in Γ. Because of the geometric properties of x, γ will possess the stated properties.

The proof for $g = 0$, $k \geq 5$. We assume that $k = 5$ (only a trivial change is needed if $k > 5$), and without loss of generality we may assume that $R = S - \{\infty, w_1, w_2, w_3, w_4\}$, where S is the sphere and the punctures are at ∞ and w_j. Choosing a base point w in R for the fundamental group of R, construct two simple closed curves a and b (each starting and ending at w) with the following properties (in which $N(x, v)$ denotes the winding number of x about v):

(i) $N(a, w_1) = N(a, w_2) = 1$, $N(a, w_3) = N(a, w_4) = 0$;

(ii) $N(b, w_1) = N(b, w_2) = 0$, $N(b, w_3) = N(b, w_4) = -1$;

(iii) a and b intersect only at w.

Now let $x = ab$, the curve obtained by transversing a first and then b. As winding numbers are invariant under free homotopies of curves in R, we can deduce immediately that x is not freely homotopic to any of the following curves:

(iv) a point in R;

(v) a simple closed curve;

(vi) a curve lying in any disc which contains exactly one of the punctures.

Now choose a point z in H over w and lift the curves a and x from z: this gives two curves $\tilde{a}$ and $\tilde{x}$ with terminal points $\alpha(x)$ and $\gamma(z)$ respectively (and $\tilde{a}$ is an initial segment of $\tilde{x}$). By (iv), x (and hence $\tilde{x}$) is not homotopic to a point: thus γ is not the identity. By (vi), γ is not parabolic (else $\tilde{x}$ could be moved towards the fixed point of γ and (vi) would be violated). Thus γ is hyperbolic. Next, γ is primitive, for suppose that $\gamma = \eta^m$ for some η in Γ. Then η is hyperbolic and x is freely homotopic to a curve in R transversed m times, so m divides each winding number $N(x, w_j)$. Thus $m = \pm 1$ and γ is primitive. Finally, γ is nonsimple, for otherwise $\tilde{x}$ could be deformed to a fundamental segment of A_γ and then x would violate (v). Thus we have shown that γ *is a primitive nonsimple hyperbolic element of* Γ.

Finally, we show that if $\beta(A_\gamma) \cap A_\gamma \neq \varnothing$, then β is hyperbolic or the identity I. Construct the curve $L = \bigcup_{n \in \mathbf{Z}} \gamma^n(\tilde{x})$ with the fixed points of γ adjoined (the final point of $\gamma^n(x)$ is the initial point of $\gamma^{n+1}(x)$, and because $\tilde{x}$ is compact, L converges to the fixed points of γ). If we assume that $\beta(A_\gamma)$ crosses A_γ, then $\beta(L)$ crosses L. Thus there are points u and v on $\tilde{x}$ with $\beta\gamma^r(u) = \gamma^s(v)$. Because of (iii), the only Γ-equivalent points on $\tilde{x}$ are z, $\alpha(z)$ and $\gamma(z)$: thus $\gamma^{-s}\beta\gamma^r \in \{I, \alpha, \alpha^{-1}, \gamma, \gamma^{-1}, \gamma\alpha^{-1}, \alpha^{-1}\gamma\}$. It follows that there is a conjugate β_1 of β, or of β^{-1}, either in $\langle\gamma\rangle$ or in one of the cosets $\langle\gamma\rangle\alpha$, $\langle\gamma\rangle\alpha^{-1}$. If β is parabolic, then so is

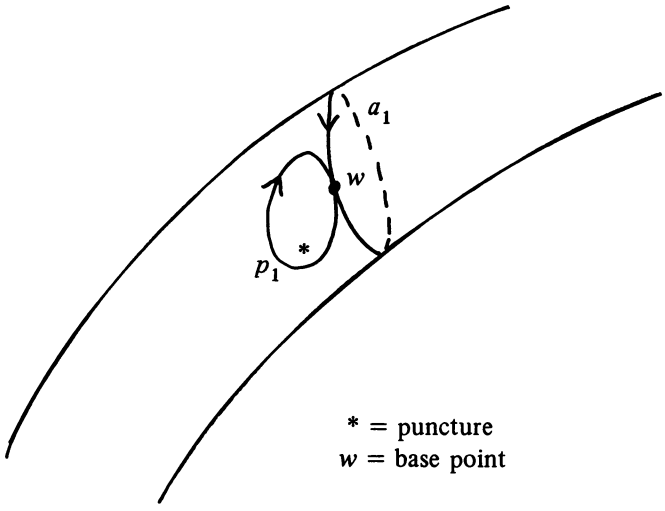


FIGURE 4

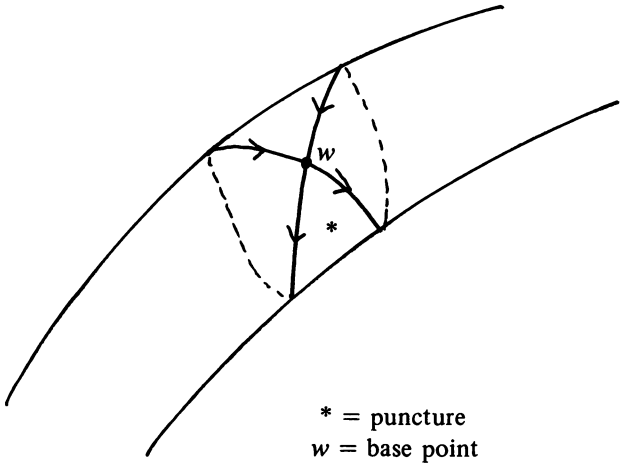


FIGURE 5

β_1, and since γ is hyperbolic we find that $\beta_1 = \gamma^m \alpha^e$ for some integer m and some choice of $e = \pm 1$. We deduce that if $y = x^m a^e$, then y is homotopic to some curve lying in a small neighborhood of one of the punctures. This is impossible, however, for

$$N(y, w_j) = \begin{cases} m + e & \text{if } j = 1, 2, \\ -m & \text{if } j = 3, 4, \end{cases}$$

and at least two of these four winding numbers must be nonzero. The proof for the case $g = 0$ is complete.

The proof for $g \geq 1$. Clearly we may assume that $k \geq 1$ (else Γ has no parabolic elements): thus we assume that $g \geq 1$, $k \geq 1$. In this part of the proof we use the standard presentation of the fundamental group Γ^* of R as the substitute for the

winding number, namely

$$\Gamma^* = \left\langle a_1, b_1, \ldots, a_g, b_g, p_1, \ldots, p_k : \prod_{i=1}^{g} [a_i, b_i] p_k \cdots p_1 = I \right\rangle \tag{4.1}$$

where $[a, b] = aba^{-1}b^{-1}$ and the p_i are parabolic. As $k \geq 1$, we can eliminate p_1 from the single relation in (4.1) and thereby realize Γ^* as a free group on the free generators

$$a_1, b_1, \ldots, a_g, b_g, p_k, \ldots, p_2. \tag{4.2}$$

Every element of Γ^* has a representation (not unique) as a word W in the generators in (4.1) and a unique representation as a reduced word W' in the generators (4.2). W' is obtained from W by replacing $(p_1)^{-1}$ by

$$\prod_{i=1}^{g} [a_i, b_i] p_k \cdots p_2 \tag{4.3}$$

wherever it occurs in W and then cancelling where possible.

Now let $\theta(u)$ be the sum of the exponents of a_1 in any word representing u. This sum is independent of the chosen word, and $\theta : \Gamma^* \to \mathbf{Z}$ is a homomorphism. Clearly, θ has the same value on conjugate elements of Γ, and, by definition, $\theta(p_i) = 0$ for $j = 1, \ldots, k$. It is well known that every parabolic element of Γ is conjugate to some p_i, so $\theta(p) = 0$ for every parabolic p in Γ.

Now define

$$x = (p_1)^{-1} a_1 a_1 = \prod_{i=1}^{g} [a_i, b_i] p_k \cdots p_2 a_1 a_1, \tag{4.4}$$

a reduced word with no p_i present if $k = 1$. We illustrate a_1 and p_1 in Figure 4 and a curve y freely homotopic to x in Figure 5, all lying on a section of the (a_1, b_1) handle of R and beginning and ending at w.

As before, select z over w and lift x to a curve $\tilde{x}$ from z to $\gamma(z)$. It is now a matter of showing that γ has the stated properties. Since $\theta(x) = 2$, γ is necessarily hyperbolic.

Next, we show that γ is primitive. If not, then we can write

$$x = (t_1 \cdots t_m)^s = (t_1 \cdots t_m) \cdots (t_1 \cdots t_m)$$

for some reduced word $t_1 \cdots t_m$ and some $s \geq 2$. Because $\theta(x) = 2$, we can only have $s = 2$ (or $s = -2$, which is essentially the same case), and then

$$[a_1, b_1] \cdots [a_g, b_g] p_k \cdots p_2 a_1 a_1 = t_1 \cdots t_m t_1 \cdots t_m. \tag{4.5}$$

The only cancellation possible on the right is $t_m t_1 = I$ (the identity), then (possibly) $t_{m-1} t_2$, and so on. In any event, the word on the right when reduced begins with t_1 and ends with t_m. Since the word on the left of (4.5) is reduced, $t_1 = a_1 = t_m$, and so no cancellation was originally possible. In view of this, both words in (4.5) are already reduced, and we have a contradiction because a_1 occurs at least four times with exponent one on the right of (4.5).

Now let β be a parabolic element of Γ. To show that $\beta(A_\gamma) \cap A_\gamma = \varnothing$, we assume the contrary and derive a contradiction. We replace x by $y = (p_1^{-1} a_1)(a_1)$ (Figure

2), constructed in the same free homotopy class and consisting of two simple loops meeting only at w. Lift y and z to obtain a curve $\tilde{y}$ from z to $\gamma(z)$ (as x and y are freely homotopic) and construct L $(= \bigcup \gamma^n(\tilde{y}))$ as in the proof for $g = 0$, using $\tilde{y}$ rather than $\tilde{x}$. Observe that the only distinct Γ-equivalent points on $\tilde{y}$ are z, $\alpha(z)$ and $\gamma(z)$. Exactly as before, we now obtain a conjugate β_1 of β with $\beta_1 = \gamma^m \alpha^e$, where m and e are integers and $e = \pm 1$. In this case we have

$$0 = \theta(\beta_1) = m\theta(\gamma) + e\theta(\alpha) = 2m \pm 1,$$

which is a contradiction as $2m \pm 1$ is an odd integer.

It remains only to show that γ is nonsimple. We shall suppose that γ is simple and again derive a contradiction. First, as $\theta(\alpha) = 1$ and $\theta(\gamma) = 2$, we see that $\alpha \notin \langle \gamma \rangle$. Thus A_γ and $\alpha(A_\gamma)$ are disjoint. This means that the endpoints of L and $\alpha(L)$ do not separate each other on the boundary of H, so as $\alpha(L)$ and L cross at $\alpha(z)$, they must also cross at some other point z'. We deduce that for some integers m and n, $\gamma^n(z') \in \tilde{y}$, $\gamma^m \alpha^{-1}(z') \in \tilde{y}$. Thus either

(i) $\gamma^n(z') = \gamma^m \alpha^{-1}(z')$ or

(ii) $\gamma^n(z'), \gamma^m \alpha^{-1}(z') \in \{z, \alpha(z), \gamma(z)\}$.

Now (i) is excluded as $\alpha \notin \langle \gamma \rangle$. From (ii) we deduce that z' lies in each of the sets

$$\{\gamma^{-n}(z), \gamma^{-n}\alpha(z), \gamma^{1-n}(z)\}, \quad \{\alpha\gamma^m(z), \alpha\gamma^{-m}\alpha(z), \alpha\gamma^{1-m}(z)\}. \tag{4.6}$$

Of the nine possible identities obtained from the different possible choices of z', five are excluded because $\alpha \notin \langle \gamma \rangle$. The remaining possibilities yield

(iii) $\alpha^{-1}\gamma^n\alpha\gamma^{-m} = I$,

(iv) $\gamma^n\alpha\gamma^{-m}\alpha = I$,

(v) $\gamma^{n-1}\alpha\gamma^{-m}\alpha = I$,

(vi) $\alpha^{-1}\gamma^n\alpha\gamma^{1-m} = I$.

Now any word in α and γ when written in terms of the free generators (4.2) is automatically reduced for α corresponds to a_1 and γ corresponds to x, which is itself reduced and which starts and ends with a_1. Thus (iv) and (v) cannot occur, and

(iii) implies that $m = n = 0$,

(iv) implies that $n = 0$ and $m = 1$.

Returning to (4.6), we find that in each case $z' = \alpha(z)$, so $\alpha(L)$ contains a loop from $\alpha(z)$ to $\alpha(z)$. Thus L contains a loop from z to z, and this cannot be because the only points on L which are equivalent to z are of the form $\gamma^r\alpha^e(z)$ with $e = 0$ or 1, and this is z only when $e = r = 0$. The proof is now complete.

5. Concluding remarks. The following modification of Theorem 1.1 is true (with essentially the same proof). Let Γ be any finitely generated Fuchsian group: then R $(= H/\Gamma)$ is a surface S of genus g with k punctures and d discs removed. Corresponding to each of these $k+d$ components of the complement of R in S, there is a conjugacy class of cyclic subgroups of Γ (corresponding to the subgroup of the fundamental group of R generated by a simple loop around the component). We call the elements in these subgroups the boundary elements of Γ (every parabolic element of Γ is of this type). Then, as suggested by Theorem 1.1, we can select β in (1.2) to be a boundary element of Γ for every nonsimple γ if and only if $g = 0$ and $k + d = 3$ or 4.

BIBLIOGRAPHY

1. H. Cohn, *Approach to Markoff's minimal forms through modular functions*, Ann. of Math. (2) **61** (1955), 1–12.
2. J. F. Koksma, *Diophantische Approximationen*, Chelsea, New York, n.d.
3. A. Haas, *Diophantine approximation on hyperbolic Riemann surfaces*, Acta Math. (Uppsala) (to appear).
4. J. Lehner and M. Sheingorn, *Simple closed geodesics on $H^+/\Gamma(3)$ arise from the Markov spectrum*, Bull. Amer. Math. Soc. (N.S.) **11** (1984), 359–362.
5. O. Perron, *Die Lehre von den Kettenbrucken*, Teubner, Stuttgart, 1954.
6. A. Schmidt, *Minimum of quadratic forms with respect to Fuchsian groups.* I, J. Reine Angew. Math. **286/7** (1976), 341–368.
7. C. Series, *The geometry of Markoff numbers*, Math. Intelligencer **7** (1985), 20–30.
8. M. Sheingorn, *Characterization of simple closed geodesics on Fricke surfaces*, Duke Math. J. **52** (1985), 535–545.
9. D. Zagier, *On the number of Markov numbers below a given bound*, Math. Comp. **39** (1982), 709–723.

DEPARTMENT OF MATHEMATICS, CAMBRIDGE UNIVERSITY, CAMBRIDGE CB2 1SB, ENGLAND (Current address of A. F. Beardon)

SCHOOL OF MATHEMATICS, INSTITUTE FOR ADVANCED STUDY, PRINCETON, NEW JERSEY 08540

DEPARTMENT OF MATHEMATICS, BARUCH COLLEGE (CUNY), NEW YORK, NEW YORK, 10010 (Current address of M. Sheingorn)

Current address (J. Lehner): 314-N Sharon Way, Jamesburg, New Jersey 08831