THE MODULAR SURFACE AND CONTINUED FRACTIONS

CAROLINE SERIES

Introduction

The aim of this note is to clarify the somewhat elusive connection between geodesics on the modular surface M (the quotient of the hyperbolic plane $\mathbb{H}$ by the modular group $G = \mathrm{SL}(2, \mathbb{Z})$) and continued fractions. This connection was, for example, noted by Artin [3] who, by an ingenious use of continued fractions, deduced the existence of a dense geodesic on M. Our results may be regarded as a rationale for Artin's method.

The idea that the sequence in which a geodesic γ cuts certain fixed lines on M (or their lifts to $\mathbb{H}$) is related to continued fractions is by no means new. However, when using the lines of the usual tesselation $\mathscr{T}$ of $\mathbb{H}$ by copies of the fundamental region $|\mathrm{Re}\, z| \leqslant \frac{1}{2}$, $|z| \geqslant 1$, attempts to find a precise relation between the cutting sequence of γ and the continued-fraction expansions of endpoints of suitable lifts of γ are fraught with minor discrepancies. Two possible solutions to the problem, neither entirely natural, are to be found in [11, 1].

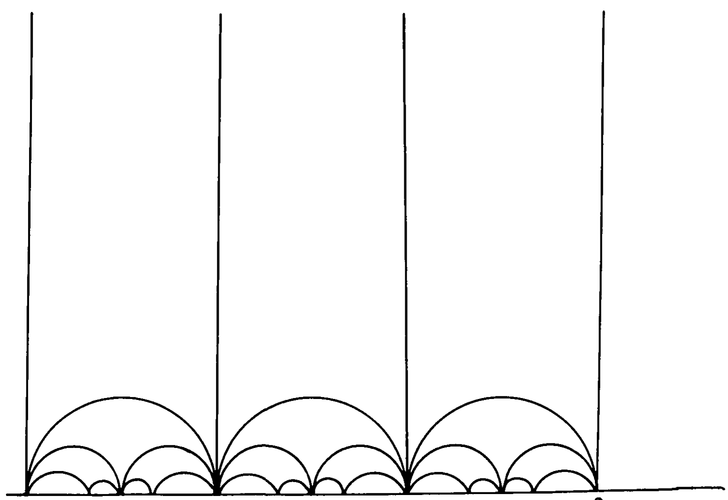

FIG. 1

In this paper all these difficulties are avoided by replacing $\mathscr{T}$ by the Farey tesselation $\mathbb{F}$, Figure 1. In this way we obtain much clearer statements of the rather remarkable facts. The Farey tesselation is a tesselation of $\mathbb{H}$ by ideal triangles, that is, triangles all of whose vertices lie on $\mathbb{R} \cup \{\infty\}$. The vertex set is precisely $\mathbb{Q} \cup \{\infty\}$. Rationals p/q, p'/q' in their lowest terms are the endpoints of a side of a triangle in

Received 13 May, 1983; revised 19 March, 1984.
1980 *Mathematics Subject Classification* 30B70.
J. London Math. Soc. (2), 31 (1985), 69–80

$\mathbb{F}$ if and only if $\begin{vmatrix} p & p' \\ q & q' \end{vmatrix} = \pm 1$. The sides of $\mathbb{F}$ turn out to be precisely the images of the imaginary axis under G.

An oriented geodesic in $\mathbb{H}$ is divided into segments as it cuts across the triangles which compose $\mathbb{F}$. In crossing such a triangle Δ a segment s cuts two sides of Δ which meet in a vertex at infinity. We label the (oriented) segment R or L according as this vertex lies to the right or left of s (Figure 2). This labelling is invariant under the action

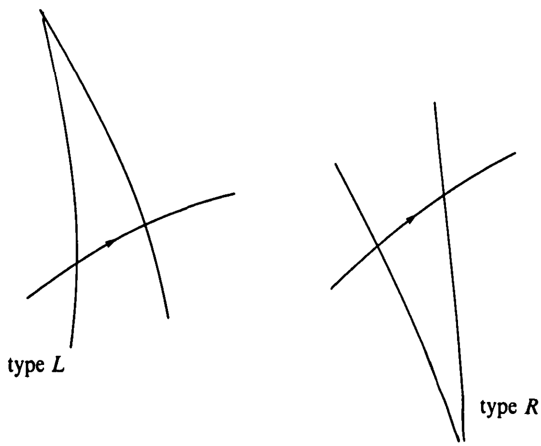

FIG. 2

of G and hence, to any geodesic $\bar{\gamma}$ on M, we may associate a cutting sequence $\ldots R^{n_0}L^{n_1}R^{n_2}\ldots$, $n_i \in \mathbb{N}$. If $x \in \bar{\gamma}$ lies at the end of a segment labelled R, we show (Theorem A) that there is a unique lift γ of $\bar{\gamma}$ such that the lift ξ of x lies on the imaginary axis and such that the positive and negative endpoints of γ on $\mathbb{R}$ are

$$\gamma_\infty = n_1 + \cfrac{1}{n_2 + \cfrac{1}{n_3 + \ldots}}, \qquad \gamma_{-\infty} = \cfrac{-1}{n_0 + \cfrac{1}{n_{-1} + \cfrac{1}{n_{-2} + \ldots}}}$$

respectively.

Motion along $\bar{\gamma}$ is obviously related to shifting the cutting sequence. This is made precise in Theorem B. The relation of the dynamics to the continued-fraction transformation $T: x \to (1/x) - [1/x]$ is explained in Theorem C. These results are described in §§1, 2.

In §3 we give some applications. We derive the relation of the hyperbolic area on M and the invariant Gauss measure for T. Theorem B allows an explicit representation of the geodesic flow as a flow over a shift [2], see also [10]. We compute the height function in 3.2. Finally, for amusement we rederive some of the well-known results about the action of G on $\mathbb{R}$ and continued fractions in 3.3.

Throughout, $\mathbb{H} = \{z \in \mathbb{C} : \operatorname{Im} z > 0]$ with the Poincaré metric $ds^2 = (dx^2 + dy^2)/y^2$. Geodesics in $\mathbb{H}$ are semicircles centred on $\mathbb{R}$ or vertical lines. For the positive and negative endpoints of an oriented geodesic γ we write $\gamma_\infty, \gamma_{-\infty}$ respectively. The modular group $G = \mathrm{SL}(2, \mathbb{Z})$ acts by isometries of $\mathbb{H}$ mapping z to $(az+b)/(cz+d)$, and $\pi: \mathbb{H} \to \mathbb{H}/\mathrm{SL}(2, \mathbb{Z}) = M$ is the projection map. Geodesics on M are exactly the images under π of geodesics in $\mathbb{H}$. We shall always be interested in *oriented* geodesics.

We write $[n_1, n_2, \ldots]$ for the continued fraction $n_1 + \cfrac{1}{n_2 + \cfrac{1}{\cdots}}$ and $[x]$ for the integer part of x, $x > 0$.

The idea of a geometric interpretation of continued fractions using $\mathbb{F}$ goes back to Humbert [7] and H. J. Smith [12]; the author was introduced to it by Moeckel's paper [9] which inspired this work.

The author would like to thank the referee for helpful suggestions regarding the presentation of this paper.

1. *Cutting sequences and continued fractions*

1.1. *The Farey tesselation*

The standard fundamental region $|\mathrm{Re}\, z| \leqslant \frac{1}{2}$, $|z| \geqslant 1$ for G is divided in half by the imaginary axis. Move the left half over using the transformation $z \to z+1$ and glue the two halves together as in Figure 3. One obtains a new fundamental region

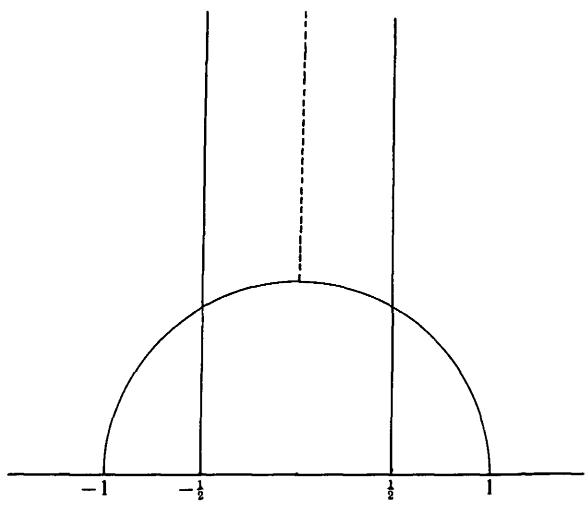

FIG. 3

for G, a quadrilateral. If S denotes the matrix $\begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \in G$ then the three images of this quadrilateral under I, S and S^2 exactly fill the ideal triangle Δ whose vertices are at $0, 1$ and ∞, (Figure 4). The images of Δ under G tesselate $\mathbb{H}$ by what we call the *Farey tesselation* $\mathbb{F}$. Notice that $\mathbb{F}$ can be regarded as the images of the imaginary axis under G.

It is not hard to see that the images of $\{0, 1, \infty\}$ under G are exactly the points $\mathbb{Q} \cup \{\infty\}$ and that two points p/q, p'/q are joined by a side of $\mathbb{F}$ if and only if $\begin{pmatrix} p & p' \\ q & q' \end{pmatrix} \in \mathrm{GL}(2, \mathbb{Z})$.

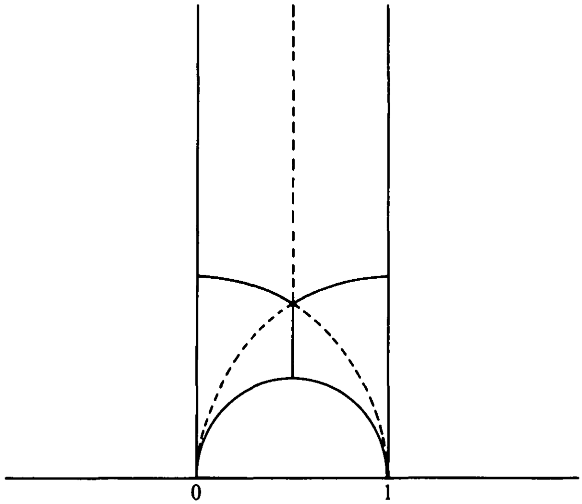

FIG. 4

There is a nice description of $\mathbb{F}$ in terms of Farey sequences. Recall that the n-th Farey sequence F_n is the set of rationals p/q with $|p|, |q| \leqslant n$ arranged in increasing order. Thus

$$F_1 \text{ is } -\infty, -1, 0, 1, \infty,$$

$$F_2 \text{ is } -\infty, -2, -1, -\tfrac{1}{2}, 0, 1, \tfrac{1}{2}, 2, \infty,$$

and so on. Rationals $p/q > p'/q'$ are adjacent in some Farey sequence if and only if $-\begin{pmatrix} p & p' \\ q & q' \end{pmatrix} \in G$ [5]. Thus $\mathbb{F}$ may be obtained by drawing the vertical line through 0 and then successively joining adjacent points in each Farey sequence, Figure 5.

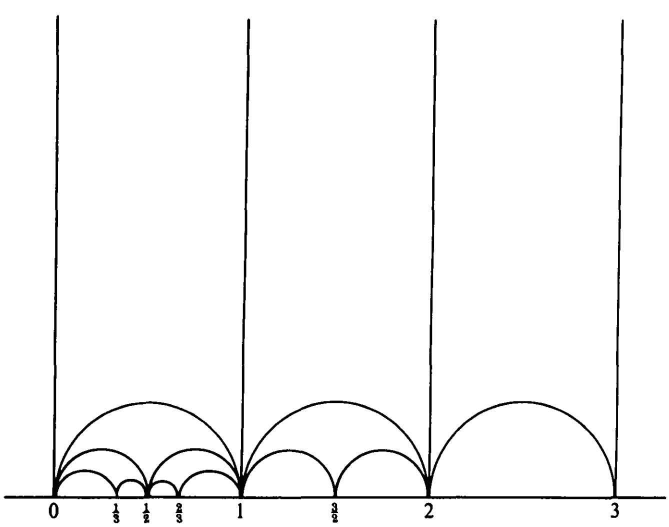

FIG. 5

1.2. *Cutting sequences*

Topologically, the modular surface M is the thrice-punctured sphere with singular points at the images of i, $\frac{1}{2}(1+i\sqrt{3})$ and ∞ respectively. The lines in the Farey tesselation project to the singular line S which runs from the cusp $\pi(\infty)$ to $\pi(i)$ and back again. If we take any geodesic $\bar{\gamma}$ on M, other than S itself, we may lift to a geodesic γ on $\mathbb{H}$ and obtain a cutting sequence $\ldots L^{n_{-1}}R^{n_0}L^{n_1}\ldots$ as described in the introduction. Since different lifts of $\bar{\gamma}$ differ by covering translations which leave $\mathbb{F}$ invariant and preserve orientation, the labels of a segment and hence the cutting sequence obtained are independent of the lift chosen.

The cutting sequence terminates if and only if γ begins or ends in the cusp $\pi(\infty)$. In this case the label of the initial or final segment may be taken to be either R or L. We call such segments initial or terminal and sometimes denote them by R_∞ or L_∞. If $x \in \bar{\gamma} \cap S$ then we may indicate the position of x in the cutting sequence by writing $\ldots R^{n_0}xL^{n_1}\ldots$. We say that the sequence *changes type* at $x \in S$ at which the segments change from R to L or vice versa, including the points immediately preceding or following initial or terminal segments.

1.3. *Statement of Theorem A*

Let A be the set of geodesics in $\mathbb{H}$ with endpoints satisfying $|\gamma_\infty| \geqslant 1$, $0 < |\gamma_{-\infty}| \leqslant 1$. Any such geodesic intersects the imaginary axis $i\mathbb{R}$ in a point ξ_γ. Notice that the cutting sequence of any such γ changes type at ξ_γ.

Since geodesics on M repeatedly cut the singular line S we have a natural cross-section X of the unit tangent bundle $T_1 M$, namely the set of unit tangent vectors with base point $x \in S$ which point along geodesics whose cutting sequences change type at x.

It is clear that if $\gamma \in A$ then the unit tangent vector u_γ to γ at ξ_γ projects to an element in X. This identification of A with X is almost a homeomorphism.

THEOREM A. *The map $i: A \to X$, $i(\gamma) = \pi(u_\gamma)$, is surjective, continuous and open. It is injective except that the two oppositely oriented geodesics joining $+1$ to -1 have the same image. Moreover, if $u_x \in X$ defines a geodesic with cutting sequence $\ldots R^{n_0}xL^{n_1}\ldots$, then $\gamma = i^{-1}(u_x)$ has endpoints given by*

$$\gamma_\infty = [n_1, n_2, \ldots], \qquad \frac{-1}{\gamma_{-\infty}} = [n_0, n_{-1}, n_{-2}, \ldots],$$

where, if the cutting sequence terminates at either end, so does the corresponding continued fraction.

If in the cutting sequence R and L are interchanged, then

$$\gamma_\infty = -[n_1, n_2, \ldots], \qquad \frac{1}{\gamma_{-\infty}} = [n_0, n_{-1}, \ldots].$$

REMARK 1.1. Notice that γ_∞ is independent of $\gamma_{-\infty}$ and the part of the cutting sequence which precedes ξ, and vice versa.

REMARK 1.2. Since $(n_r+1)+0 = n_r + \dfrac{1}{1+0}$, the terminating sequences $\ldots L^{n_r+1}$ and $\ldots L^{n_r}R$ give the same endpoint expansion, accounting for the ambiguity in labelling a segment ending in a cusp. The same remark holds for sequences beginning $LR^{n_{-r}}\ldots$ or $R^{n_{-r}+1}\ldots$.

COROLLARY. *Two geodesics with the same cutting sequence coincide.*

Proof. Fix initial points x, x' at the same division points in each sequence, and lift to points $\xi, \xi' \in i\mathbb{R}^+$ as in Theorem A. Since the endpoints at infinity of the lifts of the two geodesics have the same continued-fraction expansion, the geodesics coincide.

Proof of Theorem A. Let γ be any geodesic in $\mathbb{H}$ which intersects $i\mathbb{R}$. Since Δ is convex, $\gamma_\infty \geqslant 1$ if and only if the segment $\gamma \cap \Delta$ is of type L, and $-1 \leqslant \gamma_{-\infty} < 0$ if and only if the segment immediately preceding Δ is of type R. Similar remarks apply if $\gamma_\infty \leqslant -1$ and $0 < \gamma_{-\infty} \leqslant 1$. Since any geodesic on M cutting S at x can be lifted to a geodesic in $\mathbb{H}$ which intersects $i\mathbb{R}$ at the lift ξ of x, we see that i maps A onto X. Moreover, suppose that $\gamma, \gamma' \in A$ and that $i(\gamma) = i(\gamma')$. Since the only identification of pairs of points on $i\mathbb{R}$ under G is given by $Q: z \to -1/z$ one sees that, if $\gamma \neq \gamma'$, then $Q(\gamma) = \gamma'$. The only geodesics $\gamma \in A$ such that $Q(\gamma) \in A$ are the geodesic joining $+1$ to -1 and its inverse. This proves the first part of the statement.

Suppose now that $\gamma \in A$ and $\gamma_\infty > 1$. Let $\gamma_\infty = [n_1, n_2, \ldots]$. Let $p_1 = n_1$ if $\gamma_\infty > n_1$ and let $p_1 = n_1 - 1$ if $\gamma_\infty = n_1$. Let $\eta_\gamma = \gamma \cap (\operatorname{Re} z = p_1)$. Between ξ_γ and η_γ the tesselation $\mathbb{F}$ partitions γ into p_1 segments at the vertical lines $\operatorname{Re} z = 0, 1, \ldots, p_1$. Thus the cutting sequence of γ between ξ_γ and η_γ is L^{p_1}. Finally, if $\gamma_\infty = 1$, then the sequence starting at ξ_γ is either R_∞ or L_∞.

Apply the map $\rho_1: z \to -1/(z-p_1)$. Clearly $\rho_1 \in G$; moreover one checks easily that $\rho_1(\gamma)_\infty \leqslant -1$ and $0 < \rho_1(\gamma)_{-\infty} \leqslant 1$, and that $\rho_1(\eta_\gamma) = \xi_{\rho_1(\gamma)}$. Since $\rho_1 \in G, \rho_1(\gamma)$ is also a lift of $\pi(\gamma)$, and so the cutting sequence of γ beginning at η_γ is the same as the cutting sequence of $\rho_1(\gamma)$ beginning at $\xi_{\rho_1(\gamma)}$. If $\rho_1(\gamma_\infty) = -1$ this sequence terminates with an ambiguous segment R_∞ or L_∞; otherwise, setting $p_2 = n_2$ if $-\rho_1(\gamma_\infty) \notin \mathbb{N}$ and $p_2 = n_2 - 1$ otherwise, the cutting sequence begins with p_2 segments of type R. Applying $\rho_2: z \to -1/(z+p_2)$, the argument repeats. Exactly similar arguments apply if $\gamma_\infty \leqslant -1$.

Notice that, if $\gamma_\infty \in \mathbb{N}$, $\gamma_\infty > 1$, then the cutting sequence is ambiguously $L^{n_1}R$ or L^{n_1+1}, consistent with the ambiguity in the continued-fraction expansions.

To study the negative endpoint $\gamma_{-\infty}$ apply the map $Q: z \to -1/z$. If γ has cutting sequence $\ldots L^{n_{-1}}R^{n_0}\xi_\gamma L^{n_1}R^{n_2}\ldots$ then since $Q \in G$ the geodesic $Q(\gamma)^{-1}$ running from $Q(\gamma_\infty)$ to $Q(\gamma_{-\infty})$ has cutting sequence $\ldots L^{n_1}Q(\xi_\gamma)R^{n_0}L^{n_{-1}}\ldots$. Clearly $Q(\gamma)^{-1} \in A$ and $\xi_{Q(\gamma)^{-1}} = Q(\xi_\gamma)$. Thus $Q(\gamma_{-\infty}) = [n_0, n_{-1}, \ldots]$, which proves the result.

2. *Dynamics*

In order to set up symbolic dynamics for the geodesics on M we need to investigate the relation between shifting cutting sequences and movement along geodesics. We shall describe symbolic dynamics for the first return map P on our special cross-section X of $T_1 M$.

For $u_x \in X$, let $P(u_x)$ be the unit tangent vector where the geodesic through x in the direction u_x next enters X after u_x. Let the base point of $P(u_x)$ be $P(x)$. It is clear that $P(u_x)$ is defined unless the segment immediately following x is terminal. (Note that if the cutting sequence following x is $L^{n_1}L_\infty$ then the point preceding the terminal segment $L_\infty = R_\infty$ lies in X.) Let $X^* = \{u_x \in X$: the segment immediately following u_x is not terminal$\}$. We have defined $P: X^* \to X$, the first return map for the cross-section X of the geodesic flow.

If a geodesic on M has cutting sequence $\ldots R^{n_0} x L^{n_1} \ldots$ starting from $x \in S$, then the same geodesic reading from $P(x)$ has cutting sequence $\ldots R^{n_0} L^{n_1} P(x) R^{n_2} \ldots$. Thus the first return map corresponds to shifting cutting sequences to the left. In order to construct symbolic dynamics we introduce the space $\mathbb{N}^{\mathbb{Z}} \times \mathbb{Z}_2$, where the first coordinate will record the sequence $\ldots n_0 n_1 n_2 \ldots$ of exponents and the second will record whether the segment immediately following the base point is of type L or R. We take terminating sequences into account by adjoining points whose sequences begin or end in a row of zeros.

Thus, let

$$\Sigma = \{((n_i)_{i=N_1}^{N_2}, w): n_i \in \mathbb{N}, -\infty \leqslant N_1 \leqslant 0 < N_2 \leqslant \infty, w \in \mathbb{Z}_2\}$$

and let $\Sigma^* \subseteq \Sigma$ be the subset with $N_2 > 1$. When we shift the (n_i)-sequence we also change the type $\ldots LxR \ldots$ to $\ldots RP(x) L \ldots$ and vice versa. Thus we define

$$\hat{\sigma}: \Sigma^* \longrightarrow \Sigma, \quad \hat{\sigma}((n_i), w) = (\sigma((n_i)), w+1),$$

where $\sigma((n_i))_j = n_{j+1}$ is the left shift.

If $e = ((n_i), 0) \in \Sigma$ let $\gamma(e)$ be the geodesic in $\mathbb{H}$ whose endpoints are $\gamma_\infty(e) = [n_1, n_2, \ldots]$ and $-1/\gamma_{-\infty}(e) = [n_0, n_{-1}, \ldots]$. Likewise if $e = ((n_i), 1)$ let $\gamma(e)$ be the geodesic with endpoints $\gamma_\infty(e) = -[n_1, n_2, \ldots]$ and $1/\gamma_{-\infty}(e) = [n_0, n_1, \ldots]$. Define $D: \Sigma \to X$ so that $D(e)$ is the projection on $T_1 M$ of the unit tangent vector to $\gamma(e)$ based at $\xi_{\gamma(e)} = \gamma(e) \cap i\mathbb{R}$.

THEOREM B. *With the product topology on Σ, the map D is a continuous surjection $\Sigma \to X$ which is bijective except at points whose expansions (n_i) terminate in zeros on one or other side. The fibre of D above the image of any such point consists of the two equivalent expansions described in Remark 1.2. Moreover $D(\Sigma^*) = X^*$, and the diagram*

$$\begin{array}{ccc} \Sigma^* & \xrightarrow{\hat{\sigma}} & \Sigma \\ {\scriptstyle D}\downarrow & & \downarrow{\scriptstyle D} \\ X^* & \xrightarrow{P} & X \end{array}$$

commutes.

Proof. It is clear that the map $e \mapsto \gamma(e)$, $e \in \Sigma$, maps surjectively to A and has fibres as claimed. By Theorem A the same is true of D.

If $e = ((n_i), 0) \in \Sigma^*$ then by Theorem A one immediately knows that the cutting sequence of $\gamma(e)$ is $\ldots R^{n_0} \xi_{\gamma(e)} L^{n_1} \ldots$ and hence the geodesic on M through $D(e)$ has cutting sequence $\ldots R^{n_0} x L^{n_1}$, where $x = \pi(\xi_{\gamma(e)})$. By the discussion above, $P(D(e))$ has cutting sequence $\ldots R^{n_0} L^{n_1} P(x) R^{n_2} \ldots$. This is also the cutting sequence of $D(\hat{\sigma}(e)) = (\sigma((n_i)), 1)$. By the corollary to Theorem A, two geodesics in A with the same cutting sequence coincide. A similar argument works if $e = ((n_i), 1)$. This proves the result.

COROLLARY. *Geodesics $\gamma, \gamma' \in A$ are equivalent under G if and only if*

$$\hat{\sigma}^n(d^{-1}(\gamma)) = d^{-1}(\gamma') \quad \textit{for some } n \in \mathbb{Z},$$

where $d: \Sigma \to A$ is the map $d(e) = \gamma(e)$.

Proof. If γ, γ' are equivalent, then $\pi(\gamma) = \pi(\gamma')$ and γ, γ' have the same cutting sequence starting at different points. Thus $P^n i(\gamma) = i(\gamma')$ for some $n \in \mathbb{Z}$. The result follows from the theorem.

Conversely, if $\hat{\sigma}^n d^{-1}(\gamma) = d^{-1}(\gamma')$ for some n, then $P^n i(\gamma) = i(\gamma')$ and so $\pi(i(\gamma))$, $\pi(i(\gamma'))$ lie on the same geodesic in M, so that γ, γ' are equivalent.

This corollary is the fact on which the work in **[1]** is based.

Finally we can make precise the connection between the dynamics of the geodesic flow on M and the continued-fraction transformation $T: (0,1) \to [0,1)$, $T(\theta) = (1/\theta) - [1/\theta]$. The shift $\hat{\sigma}$ on Σ induces a map on the positive endpoints of geodesics in A. Let $W = \{t \in \mathbb{R} : |t| \geqslant 1\}$. Then the map $p^+ : \Sigma \to W$,

$$p^+((n_i), w) = \begin{cases} [n_1, n_2, \ldots], & w = 0, \\ -[n_1, n_2, \ldots], & w = 1, \end{cases}$$

can be thought of as projection onto the positive endpoint $\gamma_\infty(e)$ of the geodesic $\gamma(e)$ represented by $e \in \Sigma$.

THEOREM C. *Define* $J: W \to [0,1)$,

$$J(t) = \begin{cases} 1/|t|, & |t| > 1, \\ 0, & |t| = 1. \end{cases}$$

Then the diagram

$$\begin{tikzcd} \Sigma^* \arrow[r, "\hat{\sigma}"] \arrow[d, "J \circ p^+"'] & \Sigma \arrow[d, "J \circ p^+"] \\ (0,1) \arrow[r, "T"] & {[0,1)} \end{tikzcd}$$

commutes.

Proof. Let $u_x \in X^*$ and let $\bar{\gamma}$ be the geodesic through x in the direction u_x. Lift $\bar{\gamma}$ to a geodesic γ in $\mathbb{H}$ so that x lifts to $\xi_\gamma \in i\mathbb{R}$. With the notation of Theorem A, $P(u_x)$ lifts to the unit tangent vector to γ at the point η_γ. As in Theorem A, let $\rho_1(z) = -1/(z - p_1)$. Then $\rho_1(\gamma) \in A$ and $\rho_1(\eta_\gamma) = \xi_{\rho_1(\gamma)}$. Thus, if we now lift γ so that the lift of $P(x)$ lies on $i\mathbb{R}$, then $P(u_x)$ lifts to the unit tangent vector to $\rho_1(\gamma)$ at $\xi_{\rho_1(\gamma)}$. Combining this with Theorem B we see that $\gamma_\infty(\hat{\sigma}(e)) = \rho_1(\gamma_\infty(e))$. It is now easy to check the statement of the theorem.

REMARK 2.1. There is obviously an analogous result for the projection from Σ to negative endpoints.

REMARK 2.2. The discontinuity in J may be explained as follows. The map $p^+ \circ P$ has jump discontinuities at integer points $\mathbb{Z}^*$, while T is discontinuous at points $1/n$, $n \in \mathbb{N}$. With our definitions the jumps unfortunately occur on opposite sides of the discontinuities for the two functions. Our definition of J compensates for this. In order to have $J(x) = 1/|x|$, $|x| \geqslant 1$, one would have to lift $u \in X - X^*$ to lie on a geodesic γ with $-1 < \gamma_{-\infty} < 0$, $\gamma_\infty = \infty$ at the point where γ cuts the hyperbolic line joining -1, 0. The statement of Theorem A would then be modified for geodesics with $\gamma_\infty = \infty$. The map $p^+ \circ P$ would be defined to be continuous from the right instead of from the left at positive integer points.

Another solution, more in the spirit of the symbolic dynamics of Theorem B, would be to take both $p^+ \circ P$ and T as two-valued at the points in question.

3. *Applications*

3.1. *The Gauss measure*

It is well known that the Gauss measure

$$m(E) = \frac{1}{\log 2}\int_E \frac{dx}{1+x}$$

is invariant for the continued-fraction transformation T on $(0, 1)$ in the sense that $m(T^{-1}E) = m(E)$ for every measurable $E \subseteq (0, 1)$ [4]. It is also well known that there is a natural invariant measure μ for the geodesic flow on $T_1 M$ given by the projection of the measure $(dxdy/y^2)\, d\theta$ on $\mathbb{H} \times S^1 \approx T_1 \mathbb{H}$ to $T_1 M$. Now $u \in T_1 \mathbb{H}$ may be specified by giving the endpoints $\alpha, \beta \in \mathbb{R}$ of the geodesic $\gamma(u)$ through u, and arc length t measuring the distance of u from, say, the (Euclidean) midpoint of $\gamma(u)$. In terms of this coordinatization the above measure transforms to $(d\alpha d\beta/(\alpha-\beta)^2)\, dt$ [6].

Now, quite generally, let $\phi_t, t \in \mathbb{R}$, be a measurable flow on a Lebesgue space Y preserving a measure μ, with a cross-section $E \subseteq Y$. Let $P: E \to E$ be the first return map of ϕ_t. Then there is always a unique P-invariant measure ν on E such that μ fibres locally as $\nu \times dt$, where dt is Lebesgue measure on the flow lines of ϕ_t [2].

In our case X is a cross-section for the geodesic flow on $T_1 M$. Identifying X with A as in Theorem A, one sees that the natural P-invariant measure on X induced by $(d\alpha d\beta/(\alpha-\beta)^2)\, dt$ is $\nu = d\alpha d\beta/(\alpha-\beta)^2$.

Now let us compute the projection of ν on $(0, 1)$ under $J \circ p^+$. Since

$$\int_1^0 \frac{d\alpha}{(\alpha-\beta)^2} = \frac{1}{\beta(1+\beta)},$$

$p_*^+\nu$ is the measure $d\beta/\beta(1+\beta)$ on $[1, \infty)$. Thus $J_* p_*^+ \nu = J_*(d\beta/\beta(1+\beta)) = dy/(1+y)$. On normalizing we recover the Gauss measure for T.

3.2. *The height function*

We now compute the length of geodesic arc between successive entries of X. This will give a representation of the geodesic flow on $T_1 M$ as a flow built over $\Sigma, \hat{\sigma}$ [2].

Suppose that $\gamma \in A$ and $\gamma_\infty > 1$ and let ξ_γ, η_γ be as in the proof of Theorem A. By one of the formulae for hyperbolic distance d,

$$d(\xi_\gamma, \eta_\gamma) = \log\left|\frac{\gamma_{-\infty}-\eta_\gamma}{\gamma_{-\infty}-\xi_\gamma}.\frac{\gamma_\infty-\xi_\gamma}{\gamma_\infty-\eta_\gamma}\right|.$$

Since ξ_γ, η_γ lie on the circle radius $\frac{1}{2}(\gamma_\infty - \gamma_{-\infty})$, centre $\frac{1}{2}(\gamma_\infty + \gamma_{-\infty})$ (Figure 6), one obtains

$$\left|\frac{\gamma_\infty-\xi_\gamma}{\gamma_{-\infty}-\xi_\gamma}\right| = \frac{\operatorname{Im}\xi_\gamma}{-\gamma_{-\infty}} = \sqrt{\left|\frac{\gamma_\infty}{\gamma_{-\infty}}\right|}, \qquad \left|\frac{\gamma_\infty-\eta_\gamma}{\gamma_{-\infty}-\eta_\gamma}\right| = \frac{\gamma_\infty-[\gamma_\infty]}{\operatorname{Im}\eta_\gamma} = \sqrt{\frac{\gamma_\infty-[\gamma_\infty]}{[\gamma_\infty]-\gamma_{-\infty}}}.$$

Thus the distance between cuts $\ldots R^{n_0}\xi_\gamma L^{n_1}\ldots$ and $\ldots R^{n_0}L^{n_1}\eta_\gamma R^{n_2}\ldots$ is

(3.2.1) $$\tfrac{1}{2}\log([n_1, n_2, \ldots][n_0, n_{-1}, n_{-2}, \ldots][n_2, n_3, \ldots][n_1, n_0, n_{-1}, \ldots]).$$

It is clear by symmetry that one obtains the same formula if $\gamma_\infty \leqslant -1$.

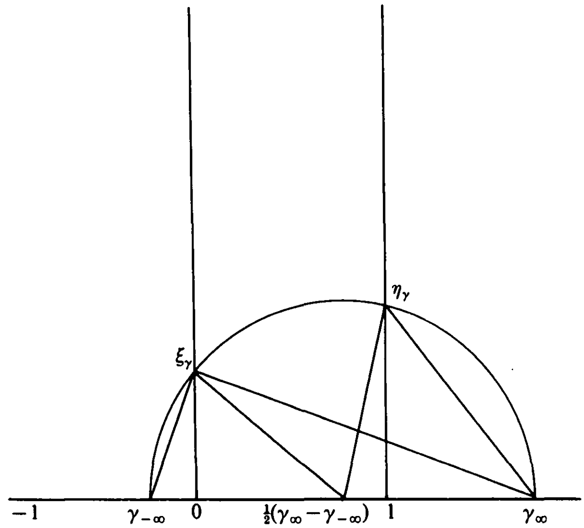

FIG. 6

With ρ_1 defined as in § 1, note that

$$\rho_1'(\gamma_\infty) = ([n_2, n_3, \ldots])^2, \qquad \rho_1'(\gamma_{-\infty}) = ([n_1, n_0, \ldots])^{-2}.$$

Thus (3.2.1) may also be written

$$\tfrac{1}{4}\log \frac{\rho_1'(\gamma_\infty)\,\rho_0'(\rho_0^{-1}\gamma_\infty)}{\rho_1'(\gamma_{-\infty})\,\rho_0'(\rho_0^{-1}\gamma_{-\infty})}.$$

In particular, if γ is periodic with period $n_1, \ldots, n_{2r}$, we obtain

$$\tfrac{1}{2}\log \frac{(\rho_{2r}\cdots\rho_1)'(\gamma_\infty)}{(\rho_{2r}\cdots\rho_1)'(\gamma_{-\infty})}$$

for the length of a period. Since $g = \rho_{2r}\ldots\rho_1 \in G$ is the primitive element fixing γ, and since $g'(\gamma_\infty) = \lambda^2$, $g'(\gamma_{-\infty}) = \lambda^{-2}$, where $\lambda > 1$ is the largest eigenvalue of g, we recover the usual formula, namely $2\log\lambda$, for the length of closed geodesics.

3.3. *Action of G on $\mathbb{R}$ and continued fractions*

We can use the results of §§ 1, 2 to recover some well-known results about continued fractions, see for example [**8**].

LEMMA 3.3.1. *Let γ, γ' be geodesics in $\mathbb{H}$ with $\gamma_\infty = \gamma'_\infty$. Then the cutting sequences of γ, γ' eventually coincide.*

Proof. We may obviously assume that the cutting sequences do not terminate, so that $\gamma_\infty \notin \mathbb{Q}$. Pick $q_1, q_2 \in \mathbb{Q}$ such that the circle joining q_1 to q_2 cuts both γ and γ'. Suppose that $q_1 \in F_n$, $q_2 \in F_m$. Then by moving closer to γ_∞ if necessary, we can find q, q' adjacent in $F_{\max(n,m)}$ such that the line C joining q to q' cuts γ, γ'. Apply $g \in G$ such that $g(q') = \infty$. It is clear that the segments of $g(\gamma), g(\gamma')$ immediately after the intersection with $g(C)$ have the same label (Figure 7). Let C_1 be the side of $\mathbb{F}$ next cut by γ, γ'. Again apply $g_1 \in G$ so that $g(C_1)$ is a vertical line, and the argument repeats.

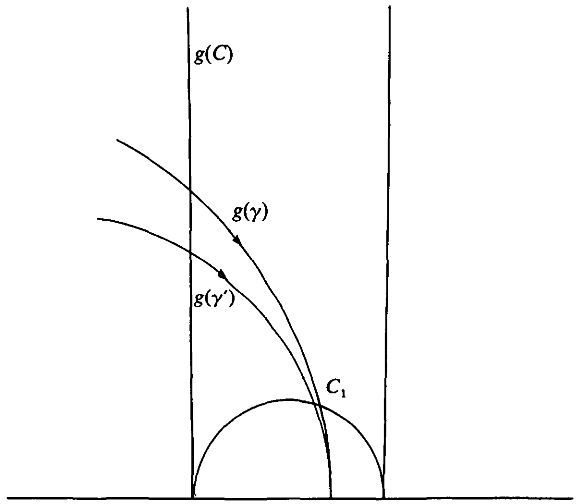

FIG. 7

DEFINITION 3.3.2. We say that two continued fractions

$$\alpha = \pm[n_1, n_2, \ldots], \qquad \beta = \pm[m_1, m_2, \ldots]$$

have the same tails mod 2 if there exist r, s such that

$$r+s \equiv \begin{cases} 0 \bmod 2 & \text{if } \alpha\beta > 0, \\ 1 \bmod 2 & \text{if } \alpha\beta < 0, \end{cases}$$

and $n_{r+k} = m_{s+k}$, $k \geqslant 0$. (We write $\pm[0, n_1, \ldots]$ if $\alpha \in (-1, 1)$.)

3.3.3. *Points $\alpha, \beta \in \mathbb{R}$ are equivalent under G if and only if they have the same tails* mod 2.

Proof. Since $-1/(\alpha - n_1) = -[n_2, n_3, \ldots]$ if $\alpha = [n_1, n_2, \ldots]$, sufficiency is clear.

Thus suppose that $\alpha, \beta \in \mathbb{R}$, $g\alpha = \beta$, $g \in G$. By applying $z \mapsto 1/z$ if necessary (note that this does not change tails mod 2), we may assume that $\alpha, \beta > 1$. Choose $\delta \in (-1, 0)$ such that the geodesics γ, γ' joining δ to α and β lie in A. Let γ, γ' have cutting sequences $\ldots \xi_\gamma L^{m_1} R^{m_2} \ldots$ and $\ldots \xi_\gamma L^{n_1} R^{n_2} \ldots$ respectively. By Lemma 3.3.1, the geodesic γ'' joining $g\delta$ to $g\alpha = \beta$ has cutting sequence $\ldots \xi_{\gamma''} \ldots R^{n_k} L^{n_{k+1}} \ldots$ for some k. Since γ'' and γ are equivalent under G, their cutting sequences, starting from equivalent initial points, coincide, and hence there exists $r \in 2\mathbb{Z}$ such that $n_{j+r} = m_j$, $j \geqslant k$. This proves the result.

3.3.4. *A number $\alpha > 1$ has a purely periodic continued-fraction expansion if and only if α is a reduced quadratic surd (that is, if the conjugate root $\bar{\alpha}$ satisfies $-1 < \bar{\alpha} < 0$). If*

$$\alpha = [\overline{n_1, n_2, \ldots, n_{2r}}]$$

then

$$-1/\bar{\alpha} = [\overline{n_{2r}, \ldots, n_1}].$$

Proof. Suppose that $\alpha = [\overline{n_1, \ldots, n_{2r}}]$. Let γ be the geodesic with endpoints $\gamma_\infty = \alpha$, $\gamma_{-\infty} = \beta$, $-1/\beta = [\overline{n_{2r}, \ldots, n_1}]$. Then $\gamma \in A$ has a periodic cutting sequence which is fixed by P^{2r}: in other words γ is fixed by $g \in G$, $g \neq I$. Thus α, β are fixed by g and hence are conjugate roots of a quadratic equation over $\mathbb{Z}$.

Conversely, let α be a reduced quadratic surd with conjugate root $\bar{\alpha}$, satisfying the equation $ax^2+bx+c=0$, where a,b,c are relatively prime and $a>0$. It is easy to show, using the inequalities $\alpha>1$, $-1<\bar{\alpha}<0$, that $|a|,|b|,|c|$ are bounded in terms of $D=b^2-4ac$. Thus there are only a finite number of reduced quadratic surds with the same discriminant D. Now consider $\rho_2\rho_1$ acting on the geodesic joining $\bar{\alpha}$ to α. Since $\rho_2\rho_1\in G$, it follows that $\rho_2\rho_1(\bar{\alpha})$, $\rho_2\rho_1(\alpha)$ are another pair of reduced quadratic surds with the same discriminant. The same holds for $\rho_{2r}\ldots\rho_1(\bar{\alpha})$, $\rho_{2r}\ldots\rho_1(\alpha)$ for any r. Thus eventually this sequence repeats so that the endpoint expansions are periodic.

3.3.5 *The tail of the expansion of $\alpha\in\mathbb{R}$ is periodic if and only if α is a quadratic surd.*

Proof. Suppose that the expansion of α has periodic tail. As in the proof of 3.3.3 we may find $g\in G$ such that $g\alpha$ is purely periodic, $g\alpha=[\overline{n_1,\ldots,n_r}]$. Applying 3.3.4, $g\alpha$ is quadratic, hence so is α.

Conversely, suppose that α is quadratic and let $\bar{\alpha}$ be the conjugate root. Let γ be the geodesic joining α to $\bar{\alpha}$. Pick a lift $g\gamma$ of $\pi(\gamma)$ with $g\gamma\in A$. By 3.3.4, $g\alpha$ has a periodic expansion and hence, by 3.3.2, the tail of α is periodic.

References

1. R. ADLER and L. FLATTO, 'Cross section maps for geodesic flows', *Ergodic theory and dynamical systems*, Progress in Mathematics 2 (ed. A. Katok, Birkhäuser, Boston, 1980).
2. W. AMBROSE, 'Representation of ergodic flows', *Ann. of Math.*, 42 (1941), 723–739.
3. E. ARTIN, 'Ein Mechanisches System mit quasi-ergodischen Bahnen', *Collected papers* (Addison Wesley, Reading, Mass., 1965), pp. 499–501.
4. P. BILLINGSLEY, *Ergodic theory and information* (Wiley, New York, 1965).
5. G. HARDY and E. M. WRIGHT, *An introduction to the theory of numbers* (University Press, Oxford, 1975).
6. E. HOPF, 'Ergodentheorie', *Abh. Sächs. Akad. Wiss. Leipzig*, 91 (1939), 261.
7. M. G. HUMBERT, 'Sur les fractions continues et les formes quadratiques binaires indéfinies', *C.R. Acad. Sci. Paris*, 162 (1916), 23–26; 67–73.
8. S. LANG, *Introduction to Diophantine approximation* (Addison Wesley, Reading Mass., 1966).
9. R. MOECKEL, Geodesics on modular surfaces and continued functions, *Ergodic Theory and Dynamical Systems*, 2 (1982), 69–84.
10. C. SERIES, 'Symbolic dynamics for geodesic flows', *Acta Math.*, 146 (1981), 103–128.
11. C. SERIES, 'On coding geodesics with continued fractions', *Enseign. Math.*, 29 (1980), 67–76.
12. H. J. SMITH, 'Mémoire sur les equations modulaires, Ac. de Lincei 1877', *Collected papers* (Chelsea, New York, 1965), pp. 224–241.

Mathematics Institute
Warwick University
Coventry CV4 7AL.