Ergod. Th. & Dynam. Sys. (1986), **6**, 601–625
Printed in Great Britain

Geometrical Markov coding of geodesics on surfaces of constant negative curvature

CAROLINE SERIES

Mathematics Institute, University of Warwick, Coventry, CV4 7AL, England and
Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104, *USA*

(*Received* 3 *May* 1985 *and revised* 5 *November* 1985)

Abstract. A natural geometrical representation of the geodesic flow on a surface M of constant negative curvature is given in which the base transformation is the shift on a (finite type) space of shortest words relative to a fixed generating set for $\pi_1(M)$ and the height function is the hyperbolic distance across a fundamental region for $\pi_1(M)$. This representation is obtained by comparing cutting sequences on M with generalised continued fraction expansions of endpoints on $\mathbb{R}$.

1. *Introduction*

The results in this paper arise from attempts over several years to understand the precise relation between various different approaches to coding geodesics on a surface M of constant negative curvature, possibly with boundary or punctures. Some of the ideas involved have already been discussed in [**15**], [**16**], [**18**]. It is of course well known by the general methods of the theory of Anosov maps that the geodesic flow on such surfaces (at least in the compact case) may be represented by a flow over a subshift of finite type [**4**]. However, work of Morse [**10**], [**12**], Artin [**2**] and Hedlund [**6**], [**7**] in the early part of this century indicates that in special cases there are other rather natural methods to obtain codings on an alphabet whose symbols are a generating set for $\pi_1(M)$, which have a very appealing relation to the underlying geometry of the surface. It is these ideas to which we return here.

There are two essentially different methods involved which we called in [**16**] the Morse and Artin methods respectively. In fact, as pointed out by Hedlund (private correspondence) what we called the Morse method in [**16**] is more properly attributed to Koebe (cf. [**8**] and our remarks below). In this paper we shall therefore refer to it as the *Koebe–Morse* method. The Artin method has the advantage of representing geodesics by sequences in a subshift of finite type, while the Koebe–Morse method has a more obvious relation to the dynamics of the geodesic flow. Briefly, the two methods are as follows. The *Koebe–Morse method* is to code a geodesic by the sequence in which it cuts a fixed set of curves on M. These fixed curves are chosen to be projections of the sides of some fundamental region for $\Gamma = \pi_1(M)$ acting in the universal cover $U \subset \mathbb{D}$ of M, where $\mathbb{D}$ is the hyperbolic disc. Since the sides of a fundamental region for Γ are naturally associated to generators of Γ, one obtains, for each geodesic γ, a doubly infinite sequence of generators called the *cutting sequence* of γ.

The *Artin method* consists of lifting a geodesic on M to U and coding the endpoints of some suitable lift at infinity. Points at infinity can be represented as semi-infinite sequences of generators of Γ using the boundary expansions of [**5**]. Thus a pair of endpoints determines a doubly infinite sequence of generators. In the special case of the modular surface $\mathbb{H}/\mathrm{SL}(2,\mathbb{Z})$, these boundary expansions reduce to the continued fraction expansions of points on $\mathbb{R}\cup\{\infty\}$. For a closed surface they may be taken to be Nielsen boundary expansions [**14**].

Our main results (theorems I and II, § 6) are that there is for quite general groups $\pi_1(M)$ a very precise relation between the Koebe–Morse and Artin codings. We consider two sets of geodesics in $\mathbb{D}$: the set $\mathscr{R}$ of geodesics which intersect some fixed fundamental region R, and the set $\mathscr{A}$ of geodesics whose doubly infinite boundary expansions satisfy a certain set of admissibility rules. There are naturally defined maps on the two sets; on $\mathscr{R}$ the first return map τ which takes a geodesic γ to the equivalent geodesic $\tau(\gamma)$ which enters R at a point equivalent to the point where γ leaves R, and on $\mathscr{A}$ the shift map σ. If we let $e(\gamma)$ be the label of the side where γ enters R, then the sequence $\ldots e(\gamma), e(\tau\gamma), e(\tau^2\gamma), \ldots$ is exactly the cutting sequence of γ.

In the simplest cases where R has no vertices in $\mathbb{D}$ the sets $\mathscr{R}$ and $\mathscr{A}$ coincide. This situation is studied in detail in § 2. More generally, $\mathscr{R}$ and $\mathscr{A}$ differ whenever $\pi_1(M)$ is not free. The discrepancy is closely related to the possible different ways of representing elements in $\pi_1(M)$ as shortest words in a given set of generators. The content of theorem I is that there is a bijection T between $\mathscr{R}$ and $\mathscr{A}$, and in theorem II we show that T conjugates the maps τ, σ. The map T is the identity on the large intersection of $\mathscr{R}$ with $\mathscr{A}$. It is moreover piecewise equal to fixed elements of $\pi_1(M)$, and the region on which T equals a fixed element has piecewise smooth and geometrically defined boundaries.

The map T may be viewed as a natural conjugacy between a cross-section of the geodesic flow and a subshift of finite type. The alphabet for this subshift consists of a set of generators of $\pi_1(M)$ and the admissible sequences are such that the finite blocks which appear run through a shortest representative for each element in $\pi_1(M)$ exactly once (theorem 4.2). From T one obtains a representation of the flow as a special flow over this subshift, where the height function is simply the hyperbolic length of the intersection of a geodesic in $\mathscr{R}$ with R.

Historically the two methods of coding seem to have arisen more or less independently. Morse in [**10**], [**11**] used a method related to the cutting sequence method (involving describing curves on M as unions of certain given geodesic segments) to study geodesics on any open surface of variable negative curvature. Contrary to the suggestion of Bott in the introduction to [**13**], Morse did not introduce cutting sequences as such until his 1938 notes [**12**]. The cutting sequence method seems to have been first exploited by Koebe [**8**]† who used exactly our cutting sequences for open surfaces (of which a special case is described in our § 2). For closed surfaces

† Added in proof: c.f. also the reference in [**8**] to the use of the same ideas in an early manuscript version (1917) of Koebe's Preisschrift (*Acta Math.* **50** (1927). We have, to date, been unable to trace this manuscript, which was deposited in the Mittag-Leffler Institute.

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

Koebe used what is now called a 'pants decomposition' of the surface and combined cutting sequences in the constituent pants with 'twist parameters' around their boundaries to obtain a one-one correspondence between geodesics (open or closed) and a certain class of symbols. In addition, Koebe dealt quite generally with both orientable and non-orientable surfaces, and allowed the possibility of both parabolic cusps and infinite connectivity.

The 1938 notes of Morse [**12**] use cutting sequences for open surfaces and the more or less equivalent idea of polygonal chains for closed surfaces. A detailed study of these chains appears in [**12**, Part II] for the symmetrical surface of genus g with the standard (a, b) generators. This analysis is closely related to the work we do to get from theorem II_0 to theorem II. The essential point was to show that polygonal chains corresponding to geodesics on a closed surface could be modified systematically to sequences in a certain subshift specified by a finite collection of rules (in modern terminology, a sofic system) so that every admissible sequence corresponded to a geodesic. Thus the existence of geodesics with certain dynamical properties could be established, as in the earlier work of Artin, Nielsen and Koebe, simply by producing admissible sequences of the required kind. The process of modification is done geometrically by replacing chains of adjacent copies of R by slightly larger geodesically convex regions. Cutting sequences are then replaced by sequences which systematically keep to one side of the enlarged region, and these sequences are shown to form a s ic system. The rules derived for this system are exactly the rules (originally found quite independently) for the boundary expansions of [**5**].

Artin in [**2**] used the continued fraction expansions of the endpoints of special lifts of geodesics to prove the existence of a geodesic dense on the modular surface. A similar method appears in Nielsen [**14**] for the symmetrical surface of genus g. In [**6**], Hedlund used Artin's ideas to prove ergodicity of the geodesic flow on this surface and later used Nielsen boundary expansions to obtain the same result on closed surfaces [**7**].

In the special case of $\mathrm{SL}(2, \mathbb{Z})$ and continued fractions, the idea of a connection between cutting sequences and boundary expansions arose in connection with number theory and goes back at least to H. J. Smith [**19**] who gave applications to the theory of reducing quadratic forms. Much more recently Adler and Flatto [**1**] have described a coding for a cross-section map of the geodesic flow on $\mathbb{H}/\mathrm{SL}(2, \mathbb{Z})$ in terms of the continued fraction transformation which is much in the spirit of our work. Related ideas appear in [**9**], and we have treated this case in detail in [**18**].

In this paper we consider all finitely generated groups Γ generated by the side pairings of fundamental regions R with *even corners*, i.e. such that $\Gamma(\partial R)$ is a union of complete geodesics in $\mathbb{H}$. This condition appears in [**8**] and was used in [**5**] to define boundary expansions geometrically. We here make heavy use of the result in [**3**] that, under this assumption, cutting sequences are shortest paths in the word metric of Γ. We also assume that R is not triangular (thus excluding the case of $\mathrm{SL}(2, \mathbb{Z})$ with the usual fundamental region). This last restriction is probably only technical (cf. remark 3.3) but the necessary geometry is somewhat different and we have not worked out the special arguments needed to treat it.

Because of the complications of the general case we begin in § 2 by describing the special case of a three-holed sphere (pair of pants). In this case the sets $\mathscr{R}$ and $\mathscr{A}$ and the maps σ, τ coincide (theorems I_0, II_0). This simple relation between cutting sequences and boundary expansions occurs because $\pi_1(M)$ is a free group and the group graph is a tree. Boundary expansions are obtained by considering points in the limit set of Γ as ends of this tree. Hedlund (private correspondence) has independently given a similar description of this example.

In § 3 we review briefly the necessary facts and results from [**3**] about geodesic cutting sequences and in § 4 collect results from [**5**] on boundary expansions. In § 5 we prove technical results which are needed for the main theorems I and II which appear in § 6.

The author would like to record her thanks to G. Hedlund for a number of enlightening comments on the history of the subject and particularly for bringing her attention to [**8**].

Since doing this work, the author has learnt that Adler and Flatto have also made use of the condition of even corners to give simple conjugacies between cross-sections of the geodesic flow and maps on the unit interval, in the same spirit as in [**1**].

Notation. Throughout this paper we use $\bar{x}$ to denote x^{-1}.

2. *The three-holed sphere*

We shall illustrate our programme by taking up the example of the three-holed sphere. This is a special case of the discussions of Morse [**10**], [**11**], [**12**] and Koebe [**8**]. We shall describe the cutting sequences and then develop a corresponding Artin-type coding by introducing suitable boundary expansions and to see that in this special case the Koebe–Morse and Artin codings coincide. As an application we show how to represent the geodesic flow on T_1M by a special flow over a shift constructed from Artin sequences.

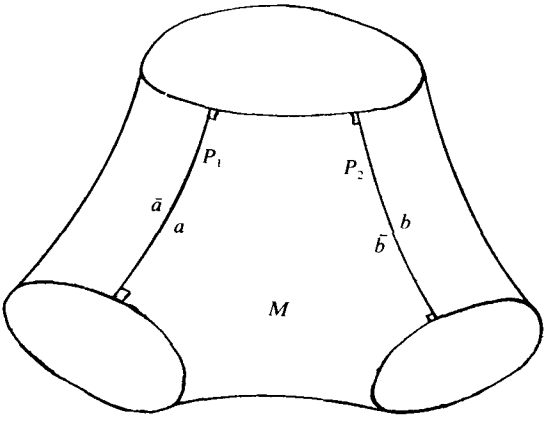

FIGURE 1(a)

Take M_∞ to be a complete hyperbolic surface with three infinite funnels, and let M be the compact part of M_∞ bounded by the unique closed geodesics which cut off the funnels. Any geodesic which cuts one of these lines goes to infinity in the

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

funnel and hence never returns to the bounded region M. Cut M along the common perpendiculars P_1, P_2 joining one of these boundary curves to the other two and lift to $\mathbb{D}$ to obtain figure 1(a). Let π denote the natural projection $\mathbb{D} \to M_\infty$.

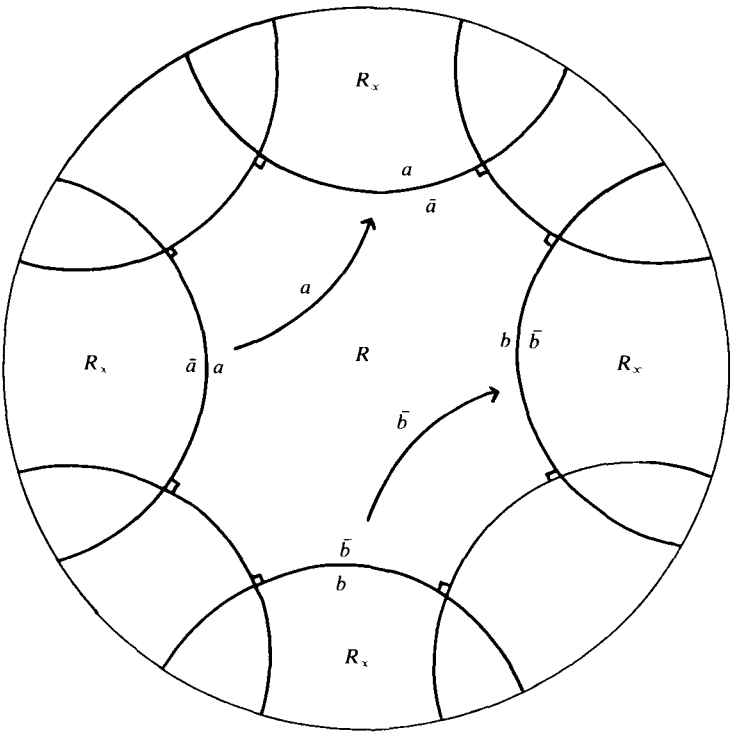

FIGURE 1(b)

The (closed) region in figure 1(b), which is a lift of M, extends to the infinite region R_∞ which projects to M_∞. Without loss of generality we may assume that $0 \in R$. The lines P_i lift to curves $\tilde{P}_i$ which form the sides of R_∞, and which are identified by isometrics a, b of $\mathbb{D}$ as shown. Then $\langle a, b |\ \rangle$ is a presentation of $\Gamma = \pi_1(M)$. The copies of R adjacent to R along sides of R are of the form eR, $e \in \Gamma_R = \{a, \bar{a}, b, \bar{b}\}$. Label the side s common to R and eR on the side of eR, by e, and on the other side by $\bar{e}$. (Equivalently, the side of s interior to R is labelled by the isometry which pairs it to some other side of R.) This labelling extends by translation under Γ to the tesselation of $\mathbb{D}$ by the images of R_∞, and induces a labelling on the oriented lines P_1, P_2 on M.

The Koebe–Morse coding. Any oriented geodesic γ on M repeatedly cuts the lines P_i. We associate to γ the sequence $\ldots e_0 e_1 e_2 \ldots$, $e_i \in \Gamma_R$, of labels on the far side of P_i in the order in which they occur, so that e_0 is the exterior label of the side of R across which γ crosses from R to $e_0 R$, and e_i is the exterior label of the side of $e_0 \ldots e_{i-1} R$ across which γ crosses from $e_0 \ldots e_{i-1} R$ to $e_0 \ldots e_i R$. Call this the *cutting sequence* of γ. Lifting to the universal cover $U \subseteq \mathbb{D}$ we can also define cutting sequences for geodesics in $\mathbb{D}$. Notice that if an arc α runs between copies R_1 and

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

R_2 of R with cutting sequence $e_1 \dots e_n$, then $R_2 = e_1 \dots e_n R_1$. The cutting sequence of γ is infinite if and only if γ is complete on M, that is, if γ never meets ∂M. Notice also that in a cutting sequence $e \in \Gamma_R$ is never immediately followed by $\bar{e}$, for this would mean that γ cut P_i twice in succession coming from opposite directions, which is impossible. Sequences with the property that e is never followed by $\bar{e}$ are called *reduced.*

Boundary Expansions. In order to discuss the Artin coding we must define the boundary expansions on $\partial\mathbb{D}$ associated to Γ. This is done for the case of the punctured torus in the example in the introduction to [17]. Let us adapt this description to the thrice punctured sphere. For $e \in \Gamma_R$ let $C(e)$ be the side of R_∞ whose exterior label is e, and let $A(e)$ be the arc cut off on $\partial\mathbb{D}$ by $C(e)$. (Notice that our convention on labelling circles differs from that used in [17].) Let $A = \bigcup \{A(e): e \in \Gamma_R\}$ and define $f: A \to \partial\mathbb{D}, f|_{A(e)}(x) = \bar{e}x$. Any point $\xi \in A$ has a finite or infinite expansion $e_0 e_1 e_2 \dots$, $e_i \in \Gamma_R$, defined by $f^n(\xi) \in A(e_n)$, $n \geq 0$, where the sequence terminates at e_n if and only if $f^n(\xi) \in A$ but $f^{n+1}(\xi) \notin A$.

Notice that $f(A(e)) \cap A(\bar{e}) = \varnothing$ for $e \in \Gamma_R$ so that a boundary expansion is necessarily reduced. Conversely any reduced sequence $e_0 e_1 \dots$ occurs as the orbit of a point in $\bigcap_{n=0}^{\infty} f^{-n}(A(e_n))$, the intersection being non-empty since $f(A(e)) \supset A(e')$ whenever $e' \neq \bar{e}$. One can show (cf. e.g. [17]) that f^N is expanding for some $N \in \mathbb{N}$, hence boundary expansions specify unique points in A.

LEMMA 2.1. *Let β be any geodesic arc joining $P \in R$ to $\xi \in A$. Then the boundary expansion of ξ is the cutting sequence of β.*

Proof. Let the cutting sequence of β be $e_0 e_1 \dots$ and let the boundary expansion of ξ be $\xi_0 \xi_1 \dots$. Clearly β leaves R across $C(\xi_0)$ so that $\xi_0 = e_0$. Suppose inductively that $\xi_i = e_i$, $i \leq n$. The $(n+2)$th region traversed by β is $e_0 \dots e_n R$. Let $g = e_0 \dots e_n$ and apply $\bar{g}$. Then $\bar{g}\beta \cap R \neq \varnothing$, and the cutting sequence of $\bar{g}\beta$ from the point where it leaves R is $e_{n+1} e_{n+2} \dots$. By definition $f^i \xi \in A(e_i)$, $i \leq n$, so that $\bar{g}\xi = f^{n+1}(\xi)$, and $f^{n+1}(\xi)$ has expansion $\xi_{n+1} \xi_{n+2} \dots$. Applying the original argument to the pair $\bar{g}\beta$, $\bar{g}\xi$ gives $e_{n+1} = \xi_{n+1}$ as required. (Notice that both sequences may terminate together, in the case when $\bar{g}\beta$ leaves R across a lift of ∂M so that $\bar{g}\beta \notin A$.)

LEMMA 2.2. *The set of points with infinite boundary expansions, $\bigcap_{n=0}^{\infty} f^{-n}A$, is equal to the limit set Λ of Γ.*

Proof. Suppose that $\xi \in \Lambda$. Since $\Lambda \subset A$, and since Λ is Γ-invariant, we have $f^n \xi \in \Gamma\Lambda = \Lambda \subset A$ for all n. Thus $\xi \in \bigcap_{n=0}^{\infty} f^{-n}A$.

Conversely, suppose that $\xi \in \bigcap_{n=0}^{\infty} f^{-n}A$. Then ξ has an infinite boundary expansion, so by lemma 2.1 the cutting sequence of the geodesic β joining 0 to ξ is also infinite. Thus $\pi(\beta)$ lies entirely within the compact part M of M_∞ so that β is within bounded distance of the orbit $\Gamma 0$ of 0. Thus β converges to a point in Λ.

COROLLARY 2.3. *There is a bijection $p^+: \sum^+ \to \Lambda$, where Σ^+ is the space of infinite reduced sequences in Γ_R.*

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

Proof. The map p^+ simply associates to $e_0e_1\ldots\in\Sigma^+$ the point $\bigcap_{n=1}^{\infty}(e_0\cdots e_n)^{-1}A(e_{n+1})$. From the lemma it follows that p^+ maps onto Λ.

Remark 2.4. (i) Notice that Σ^+ is a subshift of finite type; that is, there is a matrix $\underset{\sim}{M}=(m_{ef})$, $e,f\in\Gamma_R$, $m_{ef}\in\{0,1\}$, so that $(e_i)_{i=0}^{\infty}\in\Sigma^+\Leftrightarrow m_{e_ie_{i+1}}=1$ for $i=0,1,2,\ldots$.

(ii) Since Γ is free, shortest words in the word metric of Γ correspond exactly to reduced sequences, so that: $(e_i)_{i=0}^{\infty}\in\Sigma^+\Leftrightarrow e_k\ldots e_l$ is a shortest word in Γ for each $0\le k<l$.

Representation of geodesics. For $\xi,\eta\in\partial\mathbb{D}$, $\xi\neq\eta$, let $\gamma=\gamma(\xi,\eta)$ be the oriented geodesic joining ξ to η. We first describe those geodesics whose endpoints lie in Λ.

LEMMA 2.5. *A geodesic $\gamma(\xi,\eta)$ has endpoints $\xi,\eta\in\Lambda$ if and only if $\gamma\subset\Gamma R$. The non-wandering set for the geodesic flow on the unit tangent bundle of M corresponds exactly to unit tangent vectors directed along these geodesics.*

Proof. If $\gamma\subset\Gamma R$, then points on γ remain within a bounded distance of $\Gamma 0$ so that $\xi,\eta\in\Lambda$.

For the converse, note that since $\gamma=\bigcup_{g\in\Gamma}(\gamma\cap gR_\infty)$, and since Λ is Γ-invariant, it is enough to show that $\gamma\cap R_\infty\subset\gamma\cap R$. Now the arc on $\partial\mathbb{D}$ cut off by a lift of a boundary component of M lies entirely outside Λ. Thus if both endpoints of γ lie in Λ, we must have $\gamma\cap R_\infty\subset\gamma\cap R$.

A unit tangent vector u is in the non-wandering set of the geodesic flow ϕ_t if and only if $\phi_t(u)$ returns infinitely often within bounded distance of a fixed unit tangent vector based at $\pi(0)$. This is the case if and only if the geodesic γ in the direction of u returns infinitely often within bounded distance of $\Gamma 0$, or equivalently, if γ has both endpoints in Λ.

Now let Σ be the set of doubly infinite reduced sequences in Γ_R. For convenience we shall label such sequences $\ldots f_1f_0e_0e_1\ldots$, $e_i,f_j\in\Gamma_R$, and regard e_0 as the zero coordinate. If $\xi,\eta\in\Lambda$ have boundary expansions $\xi=\xi_0\xi_1\ldots$, $\eta=\eta_0\eta_1\ldots$, we write $\xi*\eta=\ldots\bar{\xi}_1\bar{\xi}_0\eta_0\eta_1\ldots$.

Define

$$\mathscr{A}=\{\gamma=\gamma(\xi,\eta)\colon \xi,\eta\in\Lambda \text{ and } \xi*\eta\in\Sigma\},$$

$$\mathscr{R}=\{\gamma=\gamma(\xi,\eta)\colon \xi,\eta\in\Lambda \text{ and } \gamma\cap R\neq\varnothing\}.$$

There is a bijection $p\colon\Sigma\to\mathscr{A}$ which associates to $\ldots f_1f_0e_0e_1\ldots$ the geodesic $\gamma=\gamma(p^+(\bar{f}_0\bar{f}_1\ldots),p^+(e_0e_1\ldots))$.

The left shift $\sigma\colon\Sigma\to\Sigma$ induces a natural map, also denoted σ, on $\mathscr{A}$. We have also a map $\tau\colon\mathscr{R}\to\mathscr{R}$, given by $\tau(\gamma)=\bar{e}_0\gamma$, where e_0 is the first term of the cutting sequence of γ beginning where γ leaves R. (As will be seen in the next section, τ is the first return map for the cross-section of the geodesic flow given by unit tangent vectors along geodesics where they first enter R.)

The following are the two results which we shall generalise in § 6. They can be viewed as giving the precise correspondence between the Koebe–Morse and Artin codings.

THEOREM I_0. $\mathscr{A}=\mathscr{R}$.

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

THEOREM II_0. $\sigma = \tau$.

Proof of theorem I_0. Pick $\xi, \eta \in \Lambda$, $\xi \neq \eta$. Notice that $\gamma(\xi, \eta) \in \mathscr{A}$ if and only if $\xi_0 \neq \eta_0$. If $\xi_0 = \eta_0$ then $\gamma(\xi, \eta)$ lies in the half plane bounded by $C(\xi_0)$ on the side away from R, so that $\gamma \notin \mathscr{R}$. If $\xi_0 \neq \eta_0$ then $\gamma \cap R_\infty \neq \varnothing$, so that by lemma 2.5, $\gamma \cap R \neq \varnothing$ and $\gamma \in \mathscr{R}$.

Proof of theorem II_0. Let $\xi' * \eta' = \sigma(\xi * \eta)$. Then $\eta' = \eta_1 \eta_2 \ldots$ and $\xi' = \bar{\eta}_0 \xi_0 \xi_1 \ldots$. Since $f(\eta) = \bar{\eta}_0(\eta)$ has expansion $\eta_1 \eta_2 \ldots$ and since $f(\xi') = \eta_0(\xi')$ has expansion $\xi_0 \xi_1 \ldots$, we have $\eta' = \bar{\eta}_0(\eta)$ and $\eta_0(\xi') = \xi$. Thus $\sigma(\xi * \eta)$ corresponds to the geodesic $\bar{\eta}_0(\gamma)$.

By lemma 2.1, $e_0 = \eta_0$, and so $\tau(\gamma) = \bar{e}_0 \gamma = \sigma(\gamma)$.

Application: *The geodesic flow.* Recall that the geodesic flow ϕ_t is a flow on $T_1 M$, the unit tangent bundle to M. The non-wandering set $V \subset T_1 M$ is an invariant set and contains all the interesting dynamics. Using theorems I_0, II_0 above, one can easily derive a representation of $\psi_t = \phi_t|_V$ as a flow built over the shift (Σ, σ). In fact, Σ may be identified as a cross-section to the flow.

Let W be the set of unit tangent vectors in V with base points on the lines P_i. Since vectors in the direction of P_i are not in V, the section W is transversal to the flow. There is a natural way to identify $\mathscr{R}$ and W, namely $\gamma \in \mathscr{R}$ corresponds to the projection on M of the unit tangent vector $u(\gamma)$ to γ at the point where γ first enters R. Since we have shown that $\mathscr{R} = \mathscr{A}$ and that there is a bijection of $\mathscr{A}$ with Σ, we have an identification of W with Σ. It is easy to see that this map is a homeomorphism.

This identification respects the dynamics of the situation. The first return map $P: W \to W$ lifts to a map $\hat{P}$ on $T_1 \mathbb{D}$, and we see that $\hat{P}(u(\gamma)) = v(\gamma)$ where $v(\gamma)$ is the point where $\gamma \in \mathscr{R}$ leaves R. The time taken to return to W is exactly $h(\gamma)$, the hyperbolic length of $\gamma \cap R$.

We have already defined $\tau: \mathscr{R} \to \mathscr{R}$ by $\tau(\gamma) = \bar{g}\gamma$, whenever γ leaves R across $C(g)$. Now the unit tangent vector to $\bar{g}\gamma$ at the point of entry to R is $\bar{g}v(\gamma)$. Projecting to W and identifying W and $\mathscr{R}$ we see that $P(\gamma) = \tau(\gamma)$. Since by theorem II_0, $\tau = \sigma$, we see that the system (W, P) is conjugate to the system (Σ, σ).

It is now easy to see that the flow built on (Σ, σ) under the height function h is conjugate to the geodesic flow (V, ψ_t).

This gives a very simple representation of the geodesic flow. This application will carry over without change to a much more general situation, after we have proved theorems I and II.

3. *Geodesic cutting sequences*

We shall give here a brief summary of the definitions and results we need from [3], referring the reader there for further details. Most of the definitions are also to be found in [5].

We take Γ to be a finitely generated Fuchsian group acting in the unit disc $\mathbb{D}$ by isometrics of the hyperbolic metric $ds = 2|dz|/(1-|z|^2)$. Let R be a finite sided

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

geodesic polygon which is a fundamental region for the action of Γ in $\mathbb{D}$. The sides of R are identified in pairs by elements of Γ; the set of these elements is a symmetric set of generators Γ_R for Γ. Label each oriented side of R by the corresponding generator on the interior side of R. Let N be the net of images of ∂R under Γ. Each oriented side of N is labelled by the same generator as the corresponding side of R. With this convention, if gR, hR are adjacent along side s, then the side of s interior to hR is labelled $\bar{g}h$.

Throughout we assume that R has *even corners*, in other words, that N is a union of complete geodesics in $\mathbb{D}$. This condition is not as restrictive as it appears; in fact, every surface has fundamental regions with this property [**8**], [**3**]. The curves to cut along are illustrated in figure 2. We may also assume without loss of generality that $0 \in R$.

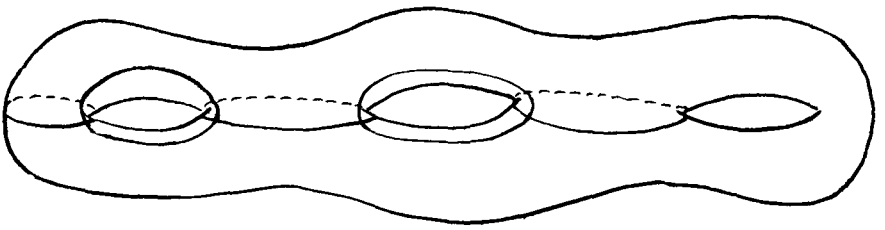

FIGURE 2

Any oriented arc γ in $\mathbb{D}$ cuts a sequence of sides $\ldots s_1 s_2 \ldots s_k \ldots$ of N and is thus associated to the corresponding sequence of labels $\ldots e_1 e_2 \ldots e_k \ldots \in \Gamma_R$, where we take the label on the far side of each s_i. (If γ passes through a vertex of N or coincides with a side of N, one modifies γ slightly to obtain an appropriate sequence as described in [**3**] (see also figure 5).) The sequence $e_1 \ldots e_k$ is called the *cutting sequence* of γ.

Conversely, to any word $w = e_1 \ldots e_k$, $e_i \in \Gamma_R$, and initial region gR, $g \in \Gamma$, we may associate an *edge path* in $\mathbb{D}$. This consists of the geodesic segments joining $g0, e_1 g0, \ldots, e_1 \ldots e_k g0$. Sometimes we replace this by the *polygonal path* consisting of the (adjacent) regions $gR, e_1 gR, \ldots, e_1 \ldots e_k gR$. The cutting sequence of this path is exactly $e_1 \ldots e_k$.

If v is any vertex of N, a small circle around N has cutting sequence $e_1 \ldots e_{2n(v)}$ where $e_1 \ldots e_{2n(v)} = 1$ is one of the defining relations of Γ. (Note that the relator has even length since R has even corners.) Any sequence of generators which appears in the order in which they occur in one of these relations we call a *cycle*; a sequence $e_1 \ldots e_{n(v)}$ we call a *half-cycle* and any cycle of greater length a *long cycle*. A cycle is *clockwise* or *anti-clockwise* depending on the sense of the corresponding edge path.

Now suppose that $v_1, \ldots, v_t$ are successive vertices of N lying along some geodesic $l \subset N$. Let α be a curve running close to and roughly parallel to l on one side possibly cutting l before v_1 and after v_t. The cutting sequence of α consists of cycles at $v_1, \ldots, v_t$ and the cycle at each intermediary vertex v_i, $l < i < t$, is of length $n(v_i) - 1$. We call such cycles *consecutive*, and the sequence of consecutive cycles we call a *chain*. (This definition differs slightly from that in [**3**], in which we allowed the cycles in a chain to have arbitrary length, but the difference is immaterial to the

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

statement of results.) We also allow infinite chains, corresponding to the case where α, l have the same endpoint at infinity. A chain is *long* if it consists of cycles of lengths $n(v_1), n(v_2)-1 \ldots, n(v_{k-1})-1, n(v_k)$. The cutting sequence of a curve α' joining the initial and final points of α but running along the other side of l is obviously also a chain, which we call *complementary* to the chain defined by α.

We measure the length of edge paths or words in Γ in the word metric defined by Γ, Γ_R. A word is *reduced* if it does not contain successive letters $x, \bar{x}, x \in \Gamma_R$. An edge path is *shortest* if the corresponding word is a shortest possible representation of the element in Γ defined by the word.

The following is the main result (theorem 2.8) of [3].

THEOREM 3.1. *Let R be a fundamental region for a group Γ, and suppose that R is not a triangle (cf. remark 3.3 below) and that R has even corners. Further, suppose that if R has four sides and if all vertices of R lie in* Int $\mathbb{D}$, *then at least three geodesics in N cross at each vertex of R. Then:*

(i) *An edge path is shortest if and only if it is reduced and contains no long cycles or long chains.*

(ii) *The cutting sequences of geodesic arcs are shortest.*

We shall also need a slight generalisation of proposition 2.7 of [3].

PROPOSITION 3.2. *Suppose that R is as in theorem* 3.1, *and let E_1, E_2 be edge paths containing no long chains with coincident initial and final points, where these endpoints may be the limits of E_1, E_2 at infinity. Then there are no copies of gR of R lying between the polygonal paths $P(E_1), P(E_2)$ defined by E_1 and E_2.*

Proof. The proof goes almost exactly as in [3]. We showed there that if R is a region lying between $P(E_1), P(E_2)$ then R has an extended side which cuts one of E_1 or E_2 twice at points in Int $\mathbb{D}$. The additional ingredient here is that if R has a cusp at infinity then an extended side of R may have one or both endpoints coincident with the endpoint(s) of E_1 at infinity, so that one cannot choose a segment cut off on E_1 of minimal length. We shall show that in fact one extended side of R does cut one of E_1, E_2 twice in Int $\mathbb{D}$, and the remainder of the proof will follow as before.

Suppose that this is not the case, so that each extended side of R either cuts each E_i at most once, or meets E_i once or twice at infinity.

Clearly R can have at most two vertices at infinity. Pick a vertex $v \in$ Int $\mathbb{D}$. The extensions of the sides s_1, s_2 of R through v each meet both E_1 and E_2, possibly at infinity. Thus a subpath of E_2, say, together with s_1, s_2, bound a region containing R. Now the extension of any side of R which intersects neither s_1 nor s_2 would, by [3, lemma 2.3], cut off a finite subpath on E_2, which we are assuming not to be the case. Thus we may reduce to the case where R has only four sides.

Let v' be the vertex of R lying on neither s_1 nor s_2. There is by assumption a side of N through v' intersecting neither s_1 nor s_2, and which therefore cuts off a finite arc on E_2, contrary to assumption.

Remark 3.3. As shown in [3], the restrictions on R in theorem 3.1 apply whenever Γ contains no elliptic elements. We have placed slightly stronger restrictions than

in [3] to ensure the validity of 3.2, which is in general false for triangular R. This is illustrated in figure 3, which refers to the group $\Gamma = \mathrm{SL}(2, \mathbb{Z})$.

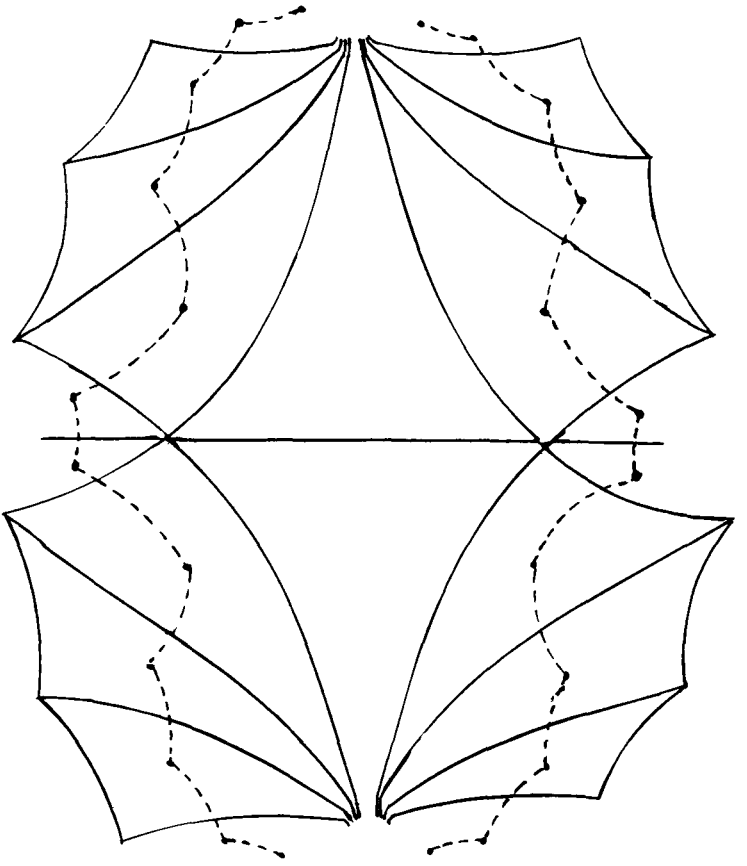

FIGURE 3

By the above proposition, whenever P, P' are shortest polygonal paths with the same endpoints, possibly at infinity, the regions forming the two paths either coincide or are adjacent. The paths differ only by forming complementary cycles round vertices v of their common boundary. These vertices we call *common vertices*, and we call such paths *adjacent*. We say that the angle on P at a common vertex v is flat, π^+, or π^- according as the number of regions in P meeting at v is $n(v)$, $n(v)+1$ or $n(v)-1$. Notice that since the maximum length of a cycle in a shortest chain is $n(v)$, and since the angles on P, P' at v together fill out the whole cycle at v, the angle on both chains at a common vertex is always one of these three types.

Now whenever the paths P, P' have common initial and final regions the two paths have equal length. Thus regions in the two paths are matched in a natural way. We want to extend this matching to paths which may meet only at infinity. For this purpose it is very important to consider only *oriented* paths. The terms 'first', 'last' below refer to order relative to this orientation.

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

Let $P=(R_i)_{i=-\infty}^{\infty}$ and $P'=(R'_i)_{i=-\infty}^{\infty}$ be shortest oriented polygonal paths with the same endpoints at infinity. We match regions in P and P' according to the following rules:

(i) If a region S is common to both paths, $S=R_i=R'_j$, then R_i is matched to R'_j.

(ii) Suppose $R_i, R_{i+1}, \ldots, R_p$ and $R'_j, R'_{j+1}, \ldots, R'_q$ are sequences of regions in P and P' which have no regions in common but which share common vertices $v_{-k}, \ldots, v_0, \ldots, v_l$, such the angle at v_i is π for $i\neq 0$, and so that the angle on P at v_0 is π^+. Suppose also that R_s, R'_t are the last regions in P, P' with $v_0\in R_s$ and $v_0\in R'_t$. Then R'_{t+r} is matched to R_{s+r+1} for $\max(i-s-1, j-t)\leq r\leq \min(q-t, p-s-1)$.

Likewise, if the angle on P at v_0 is π^- then R'_{t+r+1} is matched to R_{s+r} for $\max(i-s, j-t-1)\leq r\leq \min(p-s, q-t-1)$. Matches of this kind will be said to be *relative to the vertex* v_0.

(iii) Suppose P and P' have no common regions and the angle is flat at all common vertices. Let $R_i, \ldots, R_{i+p}$ and $R'_j, \ldots, R'_{j+p}$ be sequences of regions in P and P' which share common vertices $v_1, \ldots, v_k$. Suppose also that R_i and R'_j have a common side. Then R_{i+r} is matched to R'_{j+r} for $0\leq r\leq p$.

It is clear that the matching procedure is Γ-equivariant. The following proposition shows that it is consistent.

PROPOSITION 3.4. *If P and P' are adjacent polygonal chains, then each region in P is matched to a unique region in P', and vice versa. Further, if R_i matches R'_j and R_p matches R'_q then $p-i=q-j$.*

Proof. It is clear that the rules above match each region to at least one region in the other chain. Thus it is enough to show that if R_i and R'_j and R_p and R'_q are matched regions with $p>i$ and $q>j$, then $p-i=q-j$.

The result is clear in case (iii). For then P and P' are the opposite sides of a chain, separated by a side C of N. Since P and P' pass through the same number of regions at each common vertex, the regions match in a consistent way.

If a string of regions $R_i, \ldots, R_p$ and $R'_j, \ldots, R'_q$ are all matched by rule (i), so that $R_i=R'_j, \ldots, R_p=R'_q$, then obviously $p-i=q-j$. Thus we may assume that $R'_r\neq R_s$ for $i<r<p$ and $j<s<q$ and that P and P' are separated by a sequence of sides of N which join vertices $v_0, \ldots, v_N$ at which the angle on both paths is π, π^+ or π^-, and so that $v_0\in R_i$, $v_0\in R'_j$ and $v_N\in R_p$, $v_N\in R'_q$.

If R_i and R'_j are matched by rule (i), then the angle on P at v_0 is not flat. Suppose they are matched by rule (ii) relative to a vertex v, so that the angle at v is not flat. Between v and v_0 the paths P, P' are adjacent along a sequence of sides and vertices at which the common angles are flat. Rename the vertices in order along P as $w_0=v, w_1, \ldots$. Thus either $v_0, \ldots, v_N$ occurs as some block $w_r, \ldots, w_{r+N}$ or $w_0=v_r, w_1=v_{r+1}, \ldots, w_{N-r}=v_N$ for some $r\in\{0, \ldots, N\}$. If we rename vertices in case (i) by $v_r=w_r$, $r=0, \ldots, N$, then in both cases the angle at w_0 is not flat.

Let the regions of P, P' which have a common side joining vertices w_i, w_{i+1} be R_{i+n_i} and R'_{j+m_j}. Suppose inductively that $m_i=n_i+1$ or n_i-1 respectively according as the angle on P at the last non-flat vertex w_j preceeding w_i (i.e. with $j\leq i$), is π^-

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

or π^+. Since by assumption the angle at w_0 is not flat, the hypothesis makes sense, and clearly holds for $i=0$. Suppose it holds for $i<n<N$. If the angle at v_n on P is flat, then, since the number of regions in P and P' at v_n is equal, it holds also for $i=n+1$.

Suppose the angle at v_n on P is π^+. Then the angle on the previous non-flat vertex v_j must have been π^-, for otherwise P would contain a long chain ([**3**, lemma 2.6]). Thus by hypothesis $m_n = n_n + 1$. Continuing round v_n we see that $m_{n+1} = m_n + n(v_n) - 1$ and $n_{n+1} = n_n + n(v_n) + 1$, so that

$$m_{n+1} - n_{n+1} = m_n - n_n - 2 = -1$$

as required. A similar argument works if the angle is π^-.

This argument shows that the matching of R'_p and R_q, whether by rule (i) or (ii) at the vertex v_N, is consistent with the matching relative to w_0; in other words, $p-i=q-j$.

4. *Boundary expansions*

In this section we establish the results we need about boundary expansions to generalise the Artin-type coding described in § 2. We begin by recalling the definition of these expansions from [**5**]. As usual, we assume R is a non-triangular fundamental region for Γ with even corners.

Let the oriented sides of R be labelled by the generators Γ_R as in § 3, so that the labels on the exterior of the sides are $g_1, \ldots, g_k$ in anticlockwise order round R. As in § 2, let $C(g_i)$ be the complete geodesic in N extending the side labelled g_i, and let $A_i(g) = [P_i, Q_i]$ be the arc cut off by $C(g_i)$ on $\partial\mathbb{D}$, where P_i comes before Q_i in anticlockwise order. Depending on whether Γ is of the first or second kind, $\bigcup_{g\in\Gamma_R} A(g)$ will or will not cover $\partial\mathbb{D}$. For simplicity of exposition we shall always assume that the former is the case, so that the limit set Λ of Γ is $\partial\mathbb{D}$. The modifications needed for groups of the second kind may easily be seen by studying the example in § 2. Although the discussion in [**5**] related to groups of the first kind, this hypothesis was unnecessarily restrictive.

Define $f: \partial\mathbb{D} \to \partial\mathbb{D}$, $f|_{[P_i, P_{i+1})}(\xi) = \bar{g}_i(\xi)$. The *f-expansion* of $\xi \in \partial\mathbb{D}$ is the sequence $\xi_f = g_{i_0} g_{i_1} \ldots, g_{i_j} \in \Gamma_R$, where $f^n(\xi) \in [P_{i_n}, P_{i_n+1})$, $n \in \mathbb{N}$. Let $\Sigma^+ = \{\xi_f : \xi \in \partial\mathbb{D}\} \subset \prod_{i=0}^{\infty} \Gamma_R$.

LEMMA 4.1 ([**5**, lemma 2.3]). *The subshift Σ^+ is a sofic system. More precisely, there is an alphabet B, and a finite-to-one map $\beta: B \to \Gamma_R$, and a subshift of finite type $\Sigma_B \subset \prod_{i=1}^{\infty} B$, so that the induced map $\bar{\beta}: \Sigma_B \to \Sigma^+$ is surjective and injective except at a countable set of points where it is two-to-one, (see the remarks following the proof of* 4.2).

Proof. Partition $\partial\mathbb{D}$ into intervals whose endpoints are the set of points W where some complete geodesic in N through a vertex of R meets $\partial\mathbb{D}$. The elements of B are exactly those intervals which are bounded by adjacent points of W. Since $[P_i, P_{i+1})$ is a union of intervals in B, there is a natural map $\beta: B \to \Gamma_R$ which associates to an interval $J \in B$ the generator g_i for which $J \subset [P_i, P_{i+1})$. An easy

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

argument as in [5] shows that $f(W) \subset W$ and hence that B is a Markov partition for f.

By standard methods as in § 2 we obtain a bijection $\Sigma^+ \to \partial\mathbb{D}$. As indicated in § 2, the assumption made in [5] that $C(g_i)$ is the isometric circle of $\bar{g}_i$ is unnecessary; all we need is that f^n expands for large n.

It turns out that the finite sequences $F(\Sigma^+)$ which occur in Σ^+ run through shortest representatives of all elements in Γ, each element occurring exactly once. We use the fact that every element has a unique expression as a shortest word containing no anticlockwise half-cycles ([3, Theorem 2.8]).

THEOREM 4.2. *A word w occurs in $F(\Sigma^+)$ if and only if it is shortest and contains no anticlockwise half-cycles.*

Proof. Let R' be any image gR of R. Let $C_e(R')$ be the extended side of R' whose exterior label is e, and let $A_e(R')$ be the closed arc on $\partial\mathbb{D}$ cut off by the hyperbolic half plane $H_e(R')$ bounded by $C_e(R')$ and not containing R'. Let $A_e^*(R') = A_e(R') - A_f(R')$, where f is the exterior label of the side of R' next in anticlockwise order to e. Thus in particular $A_{g_i}(R) = [P_i, Q_i]$ and $A_{g_i}^*(R) = [P_i, P_{i+1})$, so that $\xi \in A_{g_i}^*(R)$ if and only if ξ_f begins with g_i.

More generally, let $w = e_1 \dots e_k$ be a word in Γ and let $Z(w) = \{\xi \in \partial\mathbb{D} \mid \xi_f = e_1 \dots e_k \dots\}$. We claim that

$$Z(w) = \bigcap_{i=1}^{k} A_{e_i}^*(e_1 \dots e_{i-1}R). \tag{4.2.1}$$

Suppose inductively that this is true for words of length n. Let $w = e_1 \dots e_{n+1}$. Then

$$\begin{aligned} \xi \in Z(w) &\Leftrightarrow \xi \in Z(e_1 \dots e_n) \text{ and } f^n\xi \in A_{e_{n+1}}^*(R) \\ &\Leftrightarrow \xi \in \textstyle\bigcap_{i=1}^{n} A_{e_i}^*(e_1 \dots e_{i-1}R) \text{ and } (e_1 \dots e_n)^{-1}\xi \in A_{e_{n+1}}^*(R) \\ &\Leftrightarrow \xi \in \textstyle\bigcap_{i=1}^{n} A_{e_i}^*(e_1 \dots e_{i-1}R) \text{ and } \xi \in A_{e_{n+1}}^*(e_1 \dots e_nR). \end{aligned}$$

This proves (4.2.1).

Notice that $w \in F(\Sigma^+)$ if and only if $Z(w)$ is a non-empty interval on $\partial\mathbb{D}$.

The equality (4.2.1) immediately establishes that w is reduced. For if e_ie_{i+1} are consecutive in w and $e_{i+1} = \bar{e}_i$ then $H_{e_i}(e_i \dots e_{i-1}R)$ and $H_{\bar{e}_i}(e_1 \dots e_iR)$ are the half planes bounded by $C_{e_i}(e_1 \dots e_{i-1}R)$, and hence $Z(w)$ contains at most two points.

Now suppose that w contains a cycle $e_i \dots e_{i+r}$. The half planes $H_{e_i}(e_1 \dots e_{i-1}R)$, $H_{e_{i+1}}(e_1 \dots e_iR), \dots, H_{e_{i+r}}(e_1 \dots e_{i+r-1}R)$ are bounded by sides of N through a vertex of the side s of $e_1 \dots e_{i-1}R$ with exterior label e_i. If the cycle is clockwise the half planes appear in clockwise order round the initial point of s (s is oriented to point anticlockwise round $e_1 \dots e_{i-1}R$), and if it is anticlockwise it is in anticlockwise order round the final point of s.

One sees (figure 4, for the anticlockwise case) that if either $e_i \dots e_{i+r}$ is an anticlockwise half-cycle or a long clockwise chain, then $\bigcap_{j=0}^{r} A_{e_{i+j}}^*(e_i \dots e_{i+j-1}R)$ consists of at most one point so that $w \notin F(\Sigma^+)$.

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

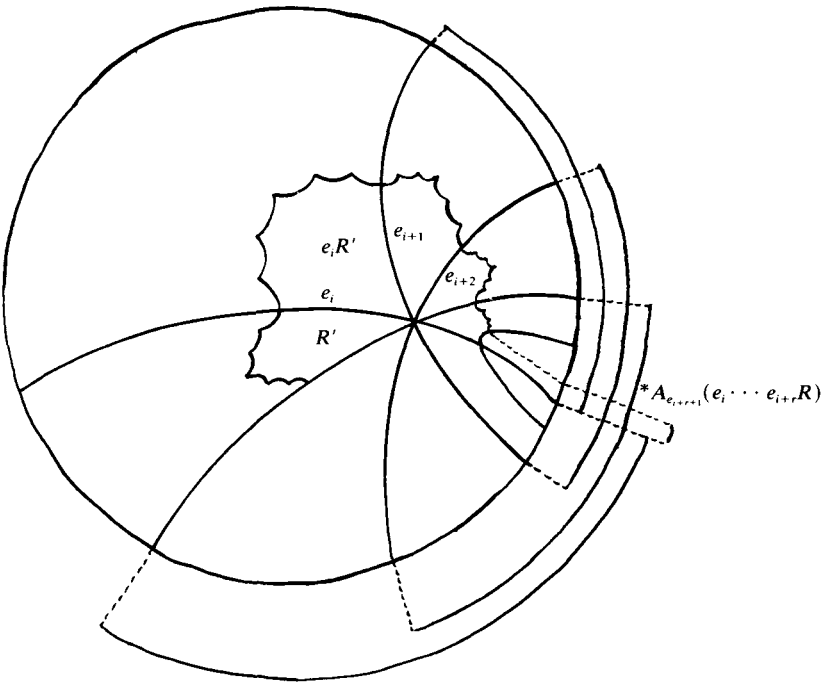

FIGURE 4

One sees also that if $e_i \dots e_{i+r+1}$ is a clockwise half-cycle followed by the first term in the consecutive cycle, then

$$\bigcap_{j=0}^{r+1} A^*_{e_{i+j}}(e_1 \dots e_{i+j-1}R) = {}^*A_{e_{i+r+1}}(e_i \dots e_{i+r}R),$$

where ${}^*A_e(R') = A^*_e(R') - A_g(R')$, where g is the next side of R' to e in clockwise order round R'. Hence, by the same arguments as above, $Z(w)$ is a point whenever w contains a long clockwise chain.

By theorem 3.1, w is shortest if and only if it is reduced and contains no long cycles or chains. Thus it only remains to show that $Z(w)$ contains an interval for all such w containing no anticlockwise cycles.

Now if e_ie_{i-1} are not successive terms in a cycle, then $H_{e_i}(e_1 \dots e_{i-1}R) \supset H_{e_{i+1}}(e_1 \dots e_iR)$. Hence one sees inductively that

$$\bigcap_{j=1}^{i} A^*_{e_j}(e_1 \dots e_{j-1}R) = A^*_{e_i}(e_1 \dots e_{i-1}R)$$

unless $e_1 \dots e_i$ terminates in a cycle or a chain. Combining this with the above observations about cycles and chains proves the result.

Theorem 4.2 gives a complete characterisation of the words occurring in $F(\Sigma^+)$. If we wish to characterise infinite words in Σ^+ there is one further constraint:

(4.3) No sequence in Σ^+ terminates in an infinite chain of anticlockwise cycles.

For if such a point existed, its image under a suitable power of f would be one of

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

the points P_{i+1}, expanded as if it belonged to the interval $A_{g_i}(R)$. But we have chosen to expand such points beginning with g_{i+1}, that is, as an infinite chain of clockwise cycles.

The points at which $\bar{\beta}: \Sigma_B \to \partial\mathbb{D}$ is two-to-one are exactly those whose expansions end in an infinite chain. Using lemma 4.1, we see that (4.3) together with theorem 4.2 gives a complete characterisation of the sequences in Σ^+.

Representation of geodesics and $\bar{f}$-expansions. In order to represent geodesics using the boundary expansions of their endpoints as in § 2, we need to ensure that such expansions lie in the natural extension Σ of Σ^+. For this to be possible we need to reverse the asymmetry in the definition of f when expanding the negative endpoint of the geodesic. Thus we introduce $\bar{f}$-expansions $\xi_{\bar{f}}$ by defining

$$\bar{f}: \partial\mathbb{D} \to \partial\mathbb{D}, \quad \bar{f}|_{(Q_i, Q_{i+1})}(\xi) = \bar{g}_i(\xi).$$

Clearly $\bar{f}$ enjoys all the properties of f, except that we interchange 'anticlockwise' and 'clockwise' throughout. In particular, $e_i \dots e_{i+r} \in F(\Sigma^+)$ if and only if $\bar{e}_{i+r} \dots \bar{e}_i \in F(\bar{\Sigma}^+)$, where $\bar{\Sigma}^+ = \{\xi_{\bar{f}}: \xi \in \partial\mathbb{D}\}$.

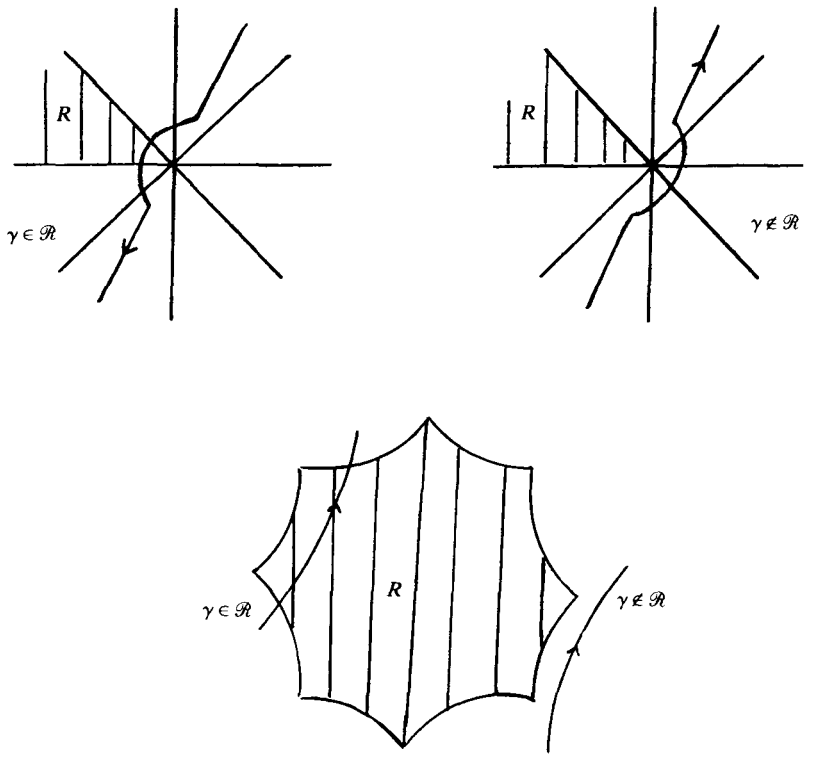

FIGURE 5

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

5. *The sets $\mathscr{A}$ and $\mathscr{R}$*

In this section we establish some preliminary results about the sets $\mathscr{A}$ and $\mathscr{R}$ described in the introduction. First, we establish some notation.

If γ is an oriented geodesic which passes through a vertex v of N then we make the convention that γ is replaced by a curve deformed *to the right* around v (see figure 5). This corresponds to our choice of right-handed boundary expansions. *From now on, we shall take as understood that all geodesic curves have been deformed, where necessary, in this way.*

Let $\xi, \eta \in \partial\mathbb{D}$, $\xi \neq \eta$, and suppose $\xi_{\bar{f}} = \xi_0\xi_{-1}\ldots$, $\eta_f = \eta_0\eta_1\ldots$. As in § 2, write $\xi * \eta = \cdots \bar{\xi}_{-1}\bar{\xi}_0\eta_0\eta_1\ldots$, and let $\gamma(\xi, \eta)$ be the oriented geodesic from ξ to η. We say $\xi * \eta$ is *shortest* if every finite block in $\xi * \eta$ is a shortest word. We write $E(\xi * \eta)$ for the edge path joining $\ldots, \xi_0\xi_10, \xi_00, 0, \eta_00, \eta_0\eta_10, \ldots$, and $E(\eta)$, $E(\xi)$ for the edge paths $0, \eta_00, \eta_0\eta_10, \ldots$ and $\ldots \xi_0\xi_10, \xi_00, 0$. The edge path of γ is denoted $E(\gamma)$.

We shall say that a sequence $(e_i)_{i=-\infty}^{\infty}$, $e_i \in \Gamma_R$, beginning or ending in an infinite chain of cycles (of lengths $n(v_1)-1, n(v_2)-1, \ldots$ at vertices $v_1, v_2, \ldots$), contains a *pseudo half cycle.* For many purposes, pseudo half cycles behave in the same way as half cycles. Notice that sequences ending in pseudo half cycles are exactly those whose endpoints lie in the set $\bigcup_{n=0}^{\infty} f^{-n}W$, where W is as in lemma 4.1.

We say the geodesic $\gamma(\xi, \eta)$ *passes near a vertex* $v \in N$ if ξ, η lie in opposite sectors defined by the net edges $N(v)$ through v, where we take the sectors to be the *closed* sets defined by the corresponding sides of v. If γ passes close to a vertex v of R and $\gamma \cap R \neq \varnothing$, then we say γ *cuts off* v *on* R if neither endpoint of γ lies in the sector at $N(v)$ containing R. Suppose ξ, η are such that $\xi_0\eta_0$ is part of a cycle or chain. Then the sides $C(\xi_0)$, $C(\eta_0)$ of R either meet in a vertex of R which we denote $v(\xi, \eta)$ or are separated by one side $s(\xi, \eta)$. Note that if in this situation γ passes near $v(\xi, \eta)$ and if $\gamma \cap R \neq \varnothing$, then γ cuts off v on R. Conversely, if γ cuts off v on R then $\bar{\xi}_0\eta_0$ lies in a cycle or chain and $v = v(\xi, \eta)$ or v is a vertex of $s(\xi, \eta)$.

Definition of the sets $\mathscr{A}$, $\mathscr{R}$. Let $\Sigma = \{(e_i)_{i=-\infty}^{\infty} | e_i \ldots e_{i+k} \in F(\Sigma^+)$, all $i < k$, and (e_i) does not begin or end with an infinite chain of anticlockwise cycles$\}$. Let

$$\mathscr{A} = \{\gamma = \gamma(\xi, \eta) \,|\, \xi * \eta \in \Sigma\}$$

and

$$\mathscr{R} = \{\gamma = \gamma(\xi, \eta) \,|\, \gamma \cap \operatorname{Int} R \neq \varnothing\}.$$

For curves γ which have been deformed because they pass through vertices of R, we have the situation illustrated in figure 5.

In contrast to the situation in § 2, it is no longer true that $\mathscr{A} = \mathscr{R}$. However, with a minor exception (lemma 5.1 below) if $\gamma \in \mathscr{R}$ then $\xi * \eta$ is shortest. The discrepancy between $\mathscr{A}$ and $\mathscr{R}$ arises from geodesics which pass close to vertices of R (lemma 5.3). These can be of four types, depending on the direction of the geodesic and the relative position of R. The possibilities are illustrated in figure 6. The cases are distinguished by the main result of this section, proposition 5.4.

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

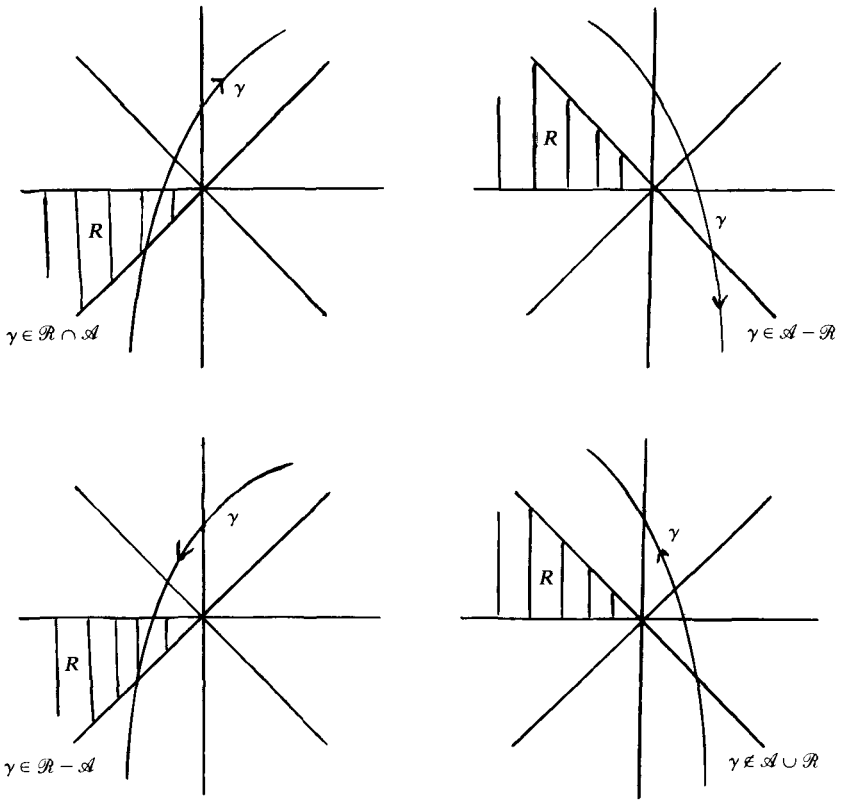

FIGURE 6

We first dispose of the exceptional case mentioned above.

LEMMA 5.1. *If $\gamma \in \mathscr{R}$ and $\xi * \eta$ is not shortest, then γ is a side of N.*

Proof. The sequence $\xi * \eta$ is not shortest because either:

(i) it reduces and $\xi_0 = \eta_0$; or

(ii) it contains a long cycle or chain which includes $\bar{\xi}_0 \eta_0$.

Case (i). Suppose $\xi_0 = \eta_0 = e_i \in \Gamma_R$. Then $\eta \in [P_i, P_{i+1})$ and $\xi \in (Q_{i-1}, Q_i]$. In particular, γ lies in the half plane $H_{e_i}(R)$ (notation as in § 4). Since $\gamma \in \mathscr{R}$, γ must coincide with $C(e_i)$.

Case (ii). Suppose $\xi * \eta$ contains a long chain. This consists of a chain of regions to one side of a side $C \subset N$, which contains the vertex $v(\xi, \eta)$ (or, where appropriate, $s(\xi, \eta)$). Since the chain is long, it cuts C at two points, one on each side of v (respectively s). Since $E(\xi)$ and $E(\eta)$ are shortest, neither of these paths can recross C. Thus γ lies in the half plane bounded by C and not containing R. Since $\gamma \in \mathscr{R}$, again, γ must be coincident with C.

In particular, the geodesics in 5.1 pass near a vertex (in fact two vertices) of R. Clearly, such geodesics lie in $\mathscr{R} - \mathscr{A}$. We now show that any other geodesic in $\mathscr{R} \triangle \mathscr{A} = \mathscr{R} - \mathscr{A} \cup \mathscr{A} - \mathscr{R}$ also passes near a vertex of R.

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

LEMMA 5.2. *Suppose that* $\gamma \in \mathscr{R} \triangle \mathscr{A}$ *and that* $\xi * \eta$ *is shortest. Then* $\bar{\xi}_0\eta_0$ *lies in a cycle or a chain and* γ *passes near* $v = v(\xi, \eta)$.

Proof. (i) Suppose $\gamma \in \mathscr{R} - \mathscr{A}$. By theorem 4.2, if $\xi * \eta$ is shortest and $\xi * \eta \notin \mathscr{A}$, then $\xi * \eta$ contains an anticlockwise half cycle containing $\xi_0\eta_0$, so that the sides $C(\xi_0)$, $C(\eta_0)$ of R meet at $v(\xi, \eta)$. The path $E(\xi * \eta)$ cuts all sides of $N(v)$ once since it contains a half cycle, and no more than once since it is shortest. Hence, ξ, η are in opposite sectors at v and we are done.

(ii) Now suppose $\gamma \in \mathscr{A} - \mathscr{R}$. Then $\xi * \eta$ is shortest. By proposition 3.2, $E(\xi * \eta)$ and $E(\gamma)$ are adjacent paths. Now $0 \in E(\xi * \eta)$ while $0 \notin E(\gamma)$ since $\gamma \notin \mathscr{R}$. Thus $\bar{\xi}_0\eta_0$ lies in a cycle or chain and $v(\xi, \eta)$ is a common vertex of the two paths. Moreover $E(\gamma)$ must cut all the sides of $N(v)$ once (possibly at infinity), otherwise $E(\xi * \eta)$ would cut some side twice, which is impossible. Thus γ has ends in opposite sectors at v.

Edge paths of geodesics passing near a vertex v always contain half cycles. More precisely:

LEMMA 5.3. *Suppose* γ *passes near* v. *Then*:

(i) *if* $\gamma \in \mathscr{R}$ *and* γ *cuts off* v *on* R, *then* $E(\gamma)$ *contains a chain beginning or ending in a half cycle or pseudo half cycle and including the cycle at* v.

(ii) *if* $\xi * \eta$ *is shortest and* $v = v(\xi, \eta)$, *then* $E(\xi * \eta)$ *has the same property as in* (i).

Proof. We shall prove only (i), case (ii) being similar. Let the sectors at v containing ξ, η be bounded by the lines l, $m \in N(v)$. Let γ cut l, m at points P, Q where possibly P or Q is on $\partial\mathbb{D}$. By theorem 3.1, $E(\gamma)$ is the edge path consisting of regions running alongside Pv, vQ on the same side as γ. If either P, $Q \in \partial\mathbb{D}$ then E is a chain beginning or ending in a pseudo half cycle, and we are done.

Otherwise, let x, y be the vertices of N along Pv, vQ closest to P, Q respectively. If $x = y = v$ then E contains a half cycle at v. Otherwise, say $x \neq v$. Then E contains a half cycle at x, which begins the chain in E which includes the cycle at v.

PROPOSITION 5.4. (see figure 6). *Suppose that* $\bar{\xi}_0\eta_0$ *lies in a cycle or chain and that* γ *passes near* $v(\xi, \eta)$. Then

$$\xi * \eta \in \mathscr{A} \Rightarrow (\gamma \text{ goes clockwise around } v \Leftrightarrow \gamma \in \mathscr{R})$$

$$\xi * \eta \notin \mathscr{A} \Rightarrow (\gamma \text{ goes anticlockwise around } v \Leftrightarrow \gamma \in \mathscr{R}).$$

Proof. (i) Suppose that $\xi * \eta$ is shortest. By 5.3, $\xi * \eta$ contains a half cycle or pseudo half cycle, with the same sense as the cycle in $\bar{\xi}_0\eta_0$. Now $\xi * \eta \in \mathscr{A}$ if and only if this cycle is clockwise. Also $\gamma \in \mathscr{R}$ if and only if $E(\xi * \eta)$ and $E(\gamma)$ agree at v, and hence have the same sense in the cycle at v. The result follows.

(ii) If $\xi * \eta$ is not shortest then certainly $\xi * \eta \notin \mathscr{A}$ and $\xi * \eta$ contains a long cycle or chain. By lemma 5.1, if $\gamma \in \mathscr{R}$ then γ is a side of N and by definition goes around v anticlockwise.

If $\gamma \notin \mathscr{R}$, then just as in the proof of 5.1, γ lies in a half plane bounded by a side C of N through v, on the side away from R. Since γ passes near v, at least one end of γ must coincide with an end α of C. Also $0 \in E = E(\xi, \eta)$ lies on the opposite

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

side of C from γ and so, since $\xi * \eta$ contains a long cycle or chain, E cuts C twice in $\operatorname{Int}\mathbb{D}$. If γ coincides with C then the conclusion holds by definition of $\mathscr{R}$. Otherwise, let P be the point where E crosses C nearest to α and let w be the next vertex along C from P towards α. Then E contains a half cycle around w, oriented in the same direction as the cycle in γ at v. This cycle is moreover not the cycle at v, since E and γ have opposite senses at vQ. Thus the cycle at w is completely contained in $E(\eta)$ or $E(\xi)$, and therefore must be clockwise. Hence, γ is oriented clockwise around v.

To define the conjugating map $\mathscr{R} \to \mathscr{A}$ in § 6 we need to define precisely the notion of *complementary path.*

Definition 5.5. Suppose we are in the situation of (5.3). Suppose that at P, the path E crosses from a region R_0 into R_1 and at Q it crosses from R_{n-1} into R_n. Let E^* be the path obtained from E by replacing $R_1, \ldots, R_{n-1}$ by the regions touching the lines Pv, vQ between R_0 and R_n and on the opposite side to E. Notice that this new path has the same length as E, for if $x, y \neq v$ the number of regions it traverses at x and y is one less than the number traversed by E while the number traversed at v is two greater. One argues similarly if x or y coincides with v. We call E^* the *complementary path to E around v.* If either P or Q lie in $\partial\mathbb{D}$ we replace E by the infinite path running along the other side of Pv or vQ.

We conclude with one further result we shall need in § 6.

LEMMA 5.6. *Let E, F be two shortest paths with the same endpoints on $\partial\mathbb{D}$, and suppose that neither E nor F contains any anticlockwise half cycles or pseudo half cycles. Then E and F coincide.*

Proof. First of all it is clear that E and F are not the chains along opposite sides of some $C \subset N$, for then one or other path would contain an anticlockwise pseudo half cycle.

Thus if E and F do not coincide, there is some common vertex v at which the angle on E, say, is π^+. Thus there is a half cycle or pseudo half cycle on E, which must be clockwise. Let $m \in N(v)$ be the side of N through v first cut by E, and let α be the endpoint at infinity such that v lies between α and the intersection with E. Since E does not cut m again, F either cuts m at some point P between v and α or the end of F coincides with α. This second case is impossible, for then F contains an anticlockwise pseudo half cycle.

In the first case let w be the last vertex of N along m between v and P. Then F contains an anticlockwise half cycle at w, which is also impossible.

6. *The conjugacy theorems*

In this section, we prove our main results, theorems I and II. As explained in § 2, these theorems allow representation of the geodesic flow as a special flow over the shift on Σ. Since the idea is essentially the same as in § 2, we shall not repeat the details here.

We begin by defining the conjugating maps $T: \mathscr{R} \to \mathscr{A}$ and $S: \mathscr{A} \to \mathscr{R}$. We keep the notation of § 5. The basic idea is that there is a symmetry between the four

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

situations illustrated in figure 6. Set:

$$T(\gamma) = S(\gamma) = \gamma, \qquad \text{if } \gamma \in \mathscr{R} \cap \mathscr{A}.$$

(6.1) *Definition of S on $\mathscr{A} - \mathscr{R}$.* Suppose $\gamma = \gamma(\xi, \eta) \in \mathscr{A} - \mathscr{R}$. The edge paths $E(\xi * \eta)$, $E(\gamma)$ meet on $\partial\mathbb{D}$ and are both shortest, so we may apply proposition 3.5 to match the paths. We have $0 \in E(\xi * \eta)$ but $0 \notin E(\gamma)$ since $\gamma \notin \mathscr{R}$. Let hR be the region matched with R and set $S(\gamma) = \bar{h}\gamma$.

(6.2) *Definition of T on $\mathscr{R} - \mathscr{A}$.* Let $\gamma = \gamma(\xi, \eta) \in \mathscr{R} - \mathscr{A}$. Suppose first that $\xi * \eta$ is not shortest. By lemma 5.1, γ is a side of N, and $0 \in E(\gamma)$ because $\gamma \in \mathscr{R}$. Let gR be the region matched to R on the opposite side of γ, and set $T(\gamma) = \bar{g}\gamma$.

Now suppose $\xi * \eta$ is shortest. By lemma 5.2, γ passes near $v(\xi, \eta)$ and the paths $E(\gamma)$, $E(\xi * \eta)$ both intersect R. Let $E^*(\xi * \eta)$ be the complementary path to $E(\xi * \eta)$ at $v(\xi, \eta)$ as in definition 5.5.

Apply proposition 3.5 to the paths $E(\gamma)$, $E^*(\xi * \eta)$ and let gR be the region matched to R. Set $T(\gamma) = \bar{g}\gamma$.

Notice that with these definitions S, T are piecewise equal to elements of Γ; moreover it is clear from figure 6 that the regions on which S and T are equal to a fixed $g \in \Gamma$ have boundaries which could easily be described geometrically.

LEMMA 6.3. *With the above definitions,*

$$S(\mathscr{A} - \mathscr{R}) \subset \mathscr{R} - \mathscr{A} \quad \textit{and} \quad T(\mathscr{R} - \mathscr{A}) \subset \mathscr{A} - \mathscr{R}.$$

Proof. By lemmas 5.1 and 5.2, γ always passes near a vertex of R.

(i) Suppose $\gamma \in \mathscr{A} - \mathscr{R}$. With the notation of (6.1), let hR be matched to R. We have $\gamma \cap \operatorname{Int}(hR) \neq \varnothing$ since $hR \in E(\gamma)$.

By proposition 5.4, γ goes anticlockwise around $v = v(\xi, \eta)$, and γ cuts off v on hR. Hence $\bar{h}\gamma \cap \operatorname{Int} R \neq \varnothing$ and $\bar{h}\gamma$ passes near $\bar{h}v(\xi, \eta)$ cutting off $\bar{h}v$ on R and going anticlockwise. Thus by (5.4), $\bar{h}\gamma \in \mathscr{R} - \mathscr{A}$.

(ii) Suppose $\gamma \in \mathscr{R} - \mathscr{A}$. With the notation of (6.2), let gR be matched to R. By (5.4), γ goes anticlockwise round $v(\xi, \eta)$ and so $\bar{g}\gamma$ passes near $\bar{g}v$ going anticlockwise. Also $\gamma \cap \operatorname{Int} gR = \varnothing$ so $\bar{g}_\gamma \cap \operatorname{Int} R = \varnothing$, hence $\bar{g}\gamma \notin \mathscr{R}$. By (5.4), $\bar{g}\gamma \in \mathscr{A} - \mathscr{R}$.

THEOREM I. *The map T is a bijection $\mathscr{R} \to \mathscr{A}$. In fact T and S are mutually inverse.*

Proof. We have only to consider the case $\gamma \in \mathscr{R} \triangle \mathscr{A}$. As in lemma 6.3, γ passes near a vertex v of R.

We shall consider only the case $\gamma \in \mathscr{A} - \mathscr{R}$; the other argument is similar. Suppose that $g0 \in E(\gamma)$ where gR is the region matched to $R \in E(\xi * \eta)$. Then $S(\gamma) = \bar{g}\gamma \in \mathscr{R} - \mathscr{A}$.

Say $E(\bar{g}\xi * \bar{g}\eta)$ is not shortest. By lemma 5.1, $\bar{g}\gamma$ and hence γ is a side of N. Clearly, gR is the region on the opposite side of γ to R matched by the rule (3.3(iii)). Thus R and $\bar{g}R$ are matched regions on opposite sides of the net side $\bar{g}\gamma$. Hence $T(\bar{g}\gamma) = g(\bar{g}\gamma) = \gamma$.

Now say $E(\bar{g}\xi * \bar{g}\eta)$ is shortest. Since $\bar{g}\gamma \notin \mathscr{A}$, there is an anticlockwise half cycle or pseudo half cycle in $\bar{g}\xi * \bar{g}\eta$ which is replaced by the corresponding clockwise cycle in $E^*(\bar{g}\xi * \bar{g}\eta)$. Thus all half cycles and pseudo half cycles in $E^*(\bar{g}\xi * \bar{g}\eta)$

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

are clockwise. The same is true of $\bar{g}E(\xi * \eta)$ since $\xi * \eta \in \mathscr{A}$. These two paths have the same endpoints on $\partial\mathbb{D}$ and so by lemma 5.6 they coincide.

Now the regions gR, R are matched in $E(\gamma), E(\xi * \eta)$ and hence $R, \bar{g}R$ are matched in $gE(\gamma) = E(\bar{g}\gamma)$ and $\bar{g}E(\xi * \eta) = E^*(\bar{g}\xi * \bar{g}\eta)$. Thus $T(\bar{g}\gamma) = g(\bar{g}\gamma) = \gamma$.

The dynamical situation is exactly as in § 2. Namely, we have the shift map σ and the first return map τ defined on $\mathscr{A}, \mathscr{R}$ respectively.

THEOREM II. *The map T conjugates the actions of σ on $\mathscr{A}$ and τ on $\mathscr{R}$.*

We first need two easy lemmas.

LEMMA 6.4. *Suppose $\xi * \eta \in \mathscr{A}$. Then $\bar{e}_0\xi * \bar{e}_0\eta \in \mathscr{A}$ and $\bar{e}_0\xi * \bar{e}_0\eta = \sigma(\xi * \eta)$.*

Proof. Let $\xi_{\bar{f}} = \bar{e}_{-1}\bar{e}_{-2}\ldots$, $\eta_f = e_0e_1e_2\ldots$. Since $\xi * \eta \in \mathscr{A}$, it follows by the characterisation of Σ^+ in § 4 that $\bar{e}_0\bar{e}_{-1}\bar{e}_{-2}\cdots \in \bar{\Sigma}^+$ and $e_1e_2\cdots \in \Sigma^+$. Define ξ', η' to be the points with these $\bar{f}$ and f expansions respectively. By definition, $\bar{f}(\xi') = e_0(\xi') = \xi$ and $f(\eta) = \bar{e}_0(\eta) = \eta'$. Thus $\xi' = \bar{e}_0\xi$, $\eta' = \bar{e}_0\eta$ as required.

LEMMA 6.5. *Let $\gamma = \gamma(\xi, \eta) \in \mathscr{R}$ and let the cutting sequence of $E(\gamma)$ be $e_0e_1\ldots$, starting at the side where γ leaves R, and let $\eta_f = \eta_0\eta_1\ldots$. Then $\eta_0 = e_0$ unless $e_0e_1\cdots$ begins with an anticlockwise chain ending in a half cycle or pseudo half cycle, in which case η_0 is the first term of the half cycle complementary to $e_0e_1\ldots$, and γ passes near a vertex of R.*

Proof. It is clear that $\eta \in C(e_0)$ and that $\eta_0 = e_0$ unless $\eta \in [P_{i+1}, Q_i]$, where $e_0 = g_i$. In this case γ passes near $v(e_0, \eta_0)$, and has anticlockwise orientation. The result now follows by the method of (5.3). It is clear that η_0 is the first term in the complementary half cycle to $e_0e_1\ldots$.

Proof of theorem II. We divide the proof into four cases, depending on whether γ and $\tau(\gamma)$ lie in $\mathscr{R} \cap \mathscr{A}$ or $\mathscr{R} - \mathscr{A}$. We always write $\gamma = \gamma(\xi, \eta)$, $\eta_f = \eta_0\eta_1\ldots$, and suppose γ leaves R across e_0, so that $\tau(\gamma) = \bar{e}_0\gamma$.

Case (i): $\gamma, \tau(\gamma) \in \mathscr{R} \cap \mathscr{A}$. We claim that $e_0 = \eta_0$. If not, by lemma 6.5, γ begins with an anticlockwise chain passing near $v(e_0, \eta_0)$. Since v is also a vertex of e_0R, we see that $\bar{e}_0\gamma$ passes anticlockwise near a vertex of R. Then by proposition 5.4, $\bar{e}_0\gamma \notin \mathscr{A}$, which is impossible.

Hence $T(\tau\gamma) = T(\bar{e}_0\gamma) = \bar{e}_0\xi * \bar{e}_0\eta$ and $\sigma(T\gamma) = \sigma(\xi * \eta)$. The result follows by lemma 6.4.

Case (ii): $\gamma \in \mathscr{R} \cap \mathscr{A}$, $\tau(\gamma) \in \mathscr{R} - \mathscr{A}$. Since $\tau(\gamma) \in \mathscr{R} - \mathscr{A}$, it follows from (5.1), (5.2) and (5.4) that $\tau(\gamma)$ passes anticlockwise round $v = v(\bar{e}_0\xi, \bar{e}_0\eta)$ and cuts off v on R.

Thus γ goes anticlockwise around e_0v and cuts off e_0v on e_0R. Now e_0v is also a vertex of R. However, γ cannot cut off e_0v on R for γ would still be going anticlockwise, which is impossible in view of (5.4) since $\gamma \in \mathscr{A} \cap \mathscr{R}$.

By lemma 5.3 the path $E(\gamma)$ contains a chain beginning or ending with a half cycle or pseudo half cycle; by the above observations this chain must begin at R. In other words the cutting sequence $e_0e_1\cdots$ of γ begins with a chain ending in a half cycle or pseudo half cycle, so by lemma 6.5, η_0 is the first term in the complementary half chain. Since the paths $E(\gamma), E(\xi * \eta)$ coincide in R, the next

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

regions e_0R, η_0R are matched in the two paths. Therefore in the paths $\bar{e}_0E(\gamma), \bar{e}_0E(\xi * \eta)$ the regions $R, \bar{e}_0\eta_0R$ are matched. Since by lemma 5.6, $\bar{e}_0E(\xi * \eta) = E(\bar{e}_0\xi * \bar{e}_0\eta)$ (both are shortest paths containing only clockwise half cycles) we have that

$$T(\bar{e}_0\gamma) = \bar{\eta}_0 e_0(\bar{e}_0\gamma) = \bar{\eta}_0\gamma.$$

Now $\sigma(T\gamma) = \sigma(\gamma) = \bar{\eta}_0\gamma$ by lemma 6.4, and we are done.

Case (iii): $\gamma \in \mathscr{R} - \mathscr{A}$, $\tau(\gamma) \in \mathscr{R} \cap \mathscr{A}$. This is illustrated in figure 7. Using arguments similar to those in case (ii) one sees that γ goes anticlockwise around a vertex v of

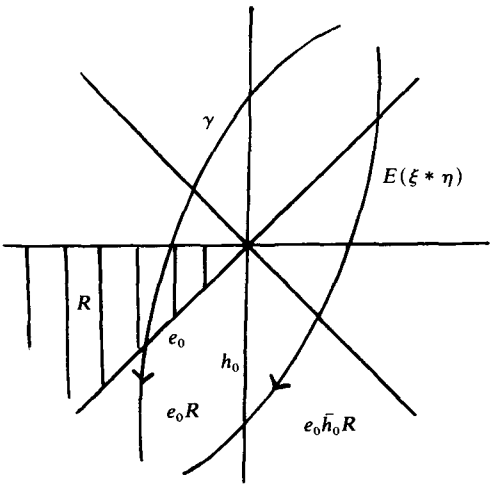

FIGURE 7

R, cutting off v on R, but that γ does not cut off the same vertex in the next region e_0R. Thus the edge sequence $E(\gamma)$ contains a chain beginning with a (pseudo) half cycle and ending with e_0. Let h_0 be the last term in the complementary chain. Since the chain in $E(\gamma)$ ends in e_0R the paths $E(\gamma), E(\xi * \eta)$ must coincide in e_0R. The preceding region in $E(\gamma), E(\xi * \eta)$ must coincide in e_0R. The preceding region in $E(\gamma)$ is R; the preceding region in $E(\xi * \eta)$ is $e_0\bar{h}_0R$. Hence $R, e_0\bar{h}_0R$ are matched so that $T(\gamma) = h_0\bar{e}_0\gamma$.

Now the cutting sequence $e_0e_1 \cdots$ does not begin with an anticlockwise cycle, or else this together with the cycle ending at e_0 would be long. Thus by lemma 6.5, $\eta_0 = e_0$. Thus $f(\eta) = \bar{e}_0\eta$ has f expansion $e_1e_2 \ldots$.

We claim that $h_0e_1e_2 \cdots \in \Sigma$. By theorem 4.2 and the remarks following it is enough to see that:

(i) $h_0 \neq \bar{e}_1$;

(ii) $h_0e_1e_2$ is not a long anticlockwise cycle or chain;

(iii) $h_0e_1 \cdots$ is not a clockwise half cycle.

Case (i) is impossible since then $\cdots e_0e_1$ would end in a long anticlockwise cycle or chain. Case (ii) is impossible for the same reason. Case (iii) is impossible for then $e_0 = \bar{e}_1$.

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

By the method of lemma 6.4, it follows that $h_0\bar{e}_0\eta$ has f expansion $h_0e_1e_2\ldots$. Thus again by (6.4),

$$\sigma(T\gamma) = \bar{h}_0(T\gamma) = \bar{e}_0\gamma = \tau(\gamma) = T(\tau\gamma).$$

Case (iv): $\gamma, \tau\gamma \in \mathscr{R} - \mathscr{A}$. In this case γ passes near a common vertex v of R and e_0R anticlockwise, cutting off the vertex in both regions. This vertex is common to the edge paths $E(\gamma), E(\xi * \eta)$. Let gR be the region matched to R and let gh_0R be matched to e_0R, so that gh_0R and gR are adjacent along the side $C_{h_0}(gR)$ of gR. Then $T(\gamma) = \bar{g}\gamma$ and

$$T(\bar{e}_0\gamma) = \bar{h}_0\bar{g}e_0(\bar{e}_0\gamma) = \bar{h}_0\bar{g}\gamma,$$

since $\bar{e}_0gh_0R$ is matched to R in the paths $E(\bar{e}_0\gamma), E(\bar{e}_0\xi * \bar{e}_0\eta) = \bar{e}_0E(\xi * \eta)$.

To compute $\sigma(T\gamma)$ we must find the first term in the f expansion of $\bar{g}\eta$. Consider the position of $\bar{g}\eta$ relative to $\bar{g}R$ and R (figure 8). Since $\bar{g}\gamma$ leaves $\bar{g}R$ across R,

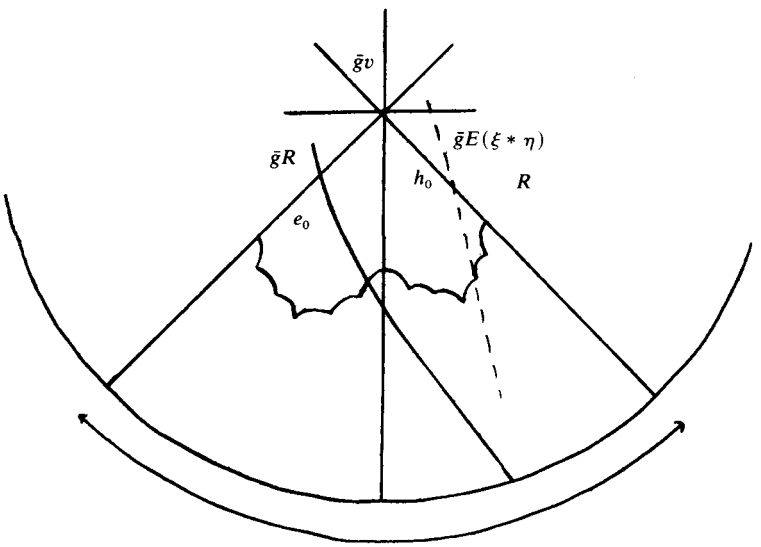

FIGURE 8

one has $\bar{g}\eta \in A_{e_0}(\bar{g}R)$. Since $\bar{g}\gamma$ cannot cross into R, we also have $\bar{g}\eta \in A_{h_0}(R)$. Since $C_{e_0}(\bar{g}R) \in N(\bar{g}v)$ and is not the same side as $C_{h_0}(R)$ (for otherwise $\bar{g}\gamma$ would enter R), we have

$$A_{e_0}(\bar{g}R) \cap A_{h_0}(R) \subset A^*_{h_0}(R).$$

Hence the f expansion of $\bar{g}\eta$ begins with h_0, so

$$\sigma(T\gamma) = \sigma(\bar{g}\gamma) = \bar{h}_0(\bar{g}\gamma) = T(\tau\gamma),$$

and we are done.

https://doi.org/10.1017/S0143385700003722 Published online by Cambridge University Press

REFERENCES

[1] R. Adler & L. Flatto. Cross section maps for the geodesic flow on the modular surface. *Contemp. Math.* **26** (1984), 9-24.
[2] E. Artin. Ein mechanisches System mit quasi ergodischen Bahnen. *Collected Papers.* Addison Wesley, 1965, 499-501.
[3] J. Birman & C. Series, Dehn's algorithm revisited, with applications to simple curves on surfaces. *Proc. Conf. on Combinatorial Group Theory, Utah*, 1984, (Ed. Gersten and Stallings).
[4] R. Bowen & D. Ruelle. The ergodic theory of Axiom A flows. *Invent. Math.* **29** (1975), 181-202.
[5] R. Bowen & C. Series. Markov maps associated to Fuchsian groups. *Publ. I.H.E.S.* **50** (1979), 153-170.
[6] G. A. Hedlund. A metrically transitive group defined by the modular group. *Amer. J. Math.* **57** (1935), 668-678.
[7] G. A. Hedlund. On the metrical transitivity of geodesics on closed surfaces of constant negative curvature. *Ann. Math.* **35** (1934), 787-808.
[8] P. Koebe. Riemannische Manigfaltigkeiten und nichteuklidische Raumformen, IV. *Sitzungberichte der Preussichen Akad. der Wissenschaften* (1929), 414-457.
[9] R. Moeckel. Geodesics on modular surfaces and continued fractions. *Ergod. Th. & Dynam. Sys.* **2** (1982), 69-84.
[10] M. Morse. A one-to-one representation of geodesics on a surface of negative curvature. *Amer. J. Math.* **XLIII** (1921), 33-51.
[11] M. Morse. Recurrent geodesics on a surface of negative curvature. *Trans. Amer. Math. Soc.* **XXII** (1921), 84-100.
[12] M. Morse. Symbolic dynamics. Institute for Advanced Study Notes, Princeton (1966) (unpublished). (First written 1938.)
[13] M. Morse. *Selected Papers*, (Ed. R. Bott). Springer-Verlag: New York (1981).
[14] J. Nielsen. Untersuchungen zur Topologie der geschlossen zweiseitige Flächen. *Act. Math.* **50** (1927), 189-358.
[15] C. Series. Symbolic dynamics for geodesic flows. *Acta Math.*, **146** (1981), 103-128.
[16] C. Series. On coding geodesics with continued fractions. *Enseignement Mathématique* **29**, Univ. de Geneve (1980), 67-76.
[17] C. Series. The infinite word problem and limit sets in Fuchsian groups. *Ergod. Th. & Dynam. Sys.* **1** (1981), 337-360.
[18] C. Series. The modular surface and continued fractions. *J. London Math. Soc.* (2), **31** (1985), 69-80.
[19] H. J. Smith. Mémoire sur les equations modulaires. *Ac. de Lincei* 1877. (*Collected Papers*, Chelsea 1965, 224-241.)