SYMBOLIC DYNAMICS FOR GEODESIC FLOWS

BY

CAROLINE SERIES

University of Warwick, Coventry, England

Introduction

By the classical result of Hopf [12], the geodesic flow on a surface of constant negative curvature and finite area is ergodic. In the case of a compact surface the flow has subsequently been shown to be Anosov [2], K [17], and Bernoulli [15]. By the work of Bowen and Ruelle [5] any Anosov flow on a compact manifold can be represented as a special flow over a Markov shift of finite type, with a Hölder continuous height function. Ratner [16] showed that any such special flow which is K is also Bernoulli.

In this paper we make an explicit geometrical construction of a symbolic dynamics for the geodesic flow on a surface of constant negative curvature and finite area. The construction involves the geometry of the surface and the structure of its fundamental group. The geodesic flow is shown to be a quotient of a special flow over a Markov shift, by a continuous map which is one—one except on a set of the first category. For a compact surface the height function is Hölder.

The states for the Markov shift are generators of the fundamental group Γ, and the admissible sequences are determined by the relations among the generators. If we lift the surface to its universal covering space the unit disc D, then admissible sequences correspond to geodesics in D which pass close to a fixed central fundamental region for Γ, in a sense made precise in § 3. The height function h corresponds to the time a geodesic takes to cross R, with a suitable convention if the geodesic is close to R but does not cut R.

The idea of our construction comes from three different sources. In [3] Artin obtained a representation of geodesics in the Poincaré upper half plane H (these geodesics are of course semi-circles centred on and orthogonal to the real axis) as doubly infinite sequences of positive integers, by juxtaposing the continued fraction expansions of their endpoints; two geodesics are then conjugate under the action of GL $(2, \mathbf{Z})$ on H if and only if the corresponding sequences are shift equivalent.

The second source is Hedlund's paper [11]. In [14] Nielsen gave a symbolic representation of points on S^1 as semi-infinite sequences of generators of the fundamental group Γ_1 for a surface whose fundamental region R_1 is a symmetrical $4g$-sided polygon; in [11] Hedlund represented geodesics in D by juxtaposing the Nielsen expansions of their endpoints, showed geodesics are conjugate under Γ_1 if and only if the corresponding sequences are shift equivalent, and used this to prove ergodicity of the geodesic flow on D/Γ_1. In [10] he showed that Artin's coding could be used to obtain similar results for $H/\mathrm{SL}\,(2, \mathbf{Z})$.

Finally in [13] Morse coded geodesics γ in D as sequences of generators in Γ_1 by an entirely different method: he observed that to each side of the net $\mathcal{N}_1$ of images of sides of R_1 under Γ_1 is associated a unique generator of Γ_1, and assigned to γ the sequence of generators which label the successive sides of $\mathcal{N}_1$ crossed by γ. In order to obtain a one–one correspondence between sequences with certain well-defined admissibility rules and geodesics this coding needs to be slightly modified when γ passes too near to a vertex of $\mathcal{N}_1$ and this point occupies a large part of [13]. The admissibility rules which are obtained are more or less identical with those of Hedlund.

In view of these results, and the facts about representing a general Anosov flow as a special flow over a Markov shift, it is natural to ask whether the ideas of Morse and Hedlund can be combined to give a representation of the geodesic flow as a special flow over some Markov shift whose symbols are generators of Γ and where the height function measures the time to cross the fundamental region R. This is precisely what we have done in this paper. Adler and Flatto (private communication) have obtained similar results in the $\mathrm{SL}\,(2, \mathbf{Z})$ and Γ_1 cases above.

The symbolic dynamics we use derives from the results of [6], in which the action of the fundamental group on S^1 is shown to be orbit equivalent to a certain Markov map f_Γ of finite type acting on S^1; that is, $x = gy$, $x, y \in S^1$, $g \in \Gamma \Leftrightarrow f_\Gamma^n(x) = f_\Gamma^m(y)$ for some $n, m \geqslant 0$. We copy Artin and Hedlund in representing geodesics in D by juxtaposing the f-expansions of their endpoints, and then show that these sequences have a geometrical interpretation analogous to Morse's idea of listing successive crossings of the fundamental region R. Finally we derive the representation of the geodesic flow on D/Γ as a quotient of a special flow over the natural extension of f_Γ.

To understand the constructions the reader will need to be familiar with the maps f_Γ of [6]. In [6] we first constructed f_Γ for groups Γ whose fundamental region R could be chosen to satisfy a certain symmetry condition (*), and then showed that any Γ could be deformed by a quasi-conformal deformation to a group Γ' satisfying (*). We then carried over the definition of $f_{\Gamma'}$ using the boundary homeomorphism and constructed the general f_Γ. We shall adopt the same procedure here, so that in the main part of the work, § 1–§ 4, we shall only be concerned with groups whose fundamental region satisfies (*).

In § 1 we review briefly the definition and properties of f_Γ and then determine which sequences of generators correspond to admissible f-expansions. In § 2 we describe the Γ action on S^1 in terms of sequences and show how to juxtapose sequences to represent certain pairs of points on S^1. In fact *geodesics are conjugate under* Γ *if and only if the* corresponding sequences are shift equivalent.

In § 3 we discuss the relation of this representation to the listing of successive crossings of R and in § 4 derive the symbolic representation of the flow. Finally in § 5 we show how to carry these results over to the general case using quasi-conformal maps.

We shall keep to the notation of [6]. In particular, when describing arcs on S^1, we always label in an anti-clockwise direction, so that PQ means the points lying between P and Q moving anti-clockwise from P to Q. We write (PQ), $[PQ]$, etc., to distinguish open and closed arcs on S^1.

Throughout, Γ is a finitely generated Fuchsian group of the first kind acting in the unit disc D; that is, a discrete group of linear fractional transformations $z \mapsto (az+b)/(cz+d)$, $ad-bc=1$, which map D to itself and such that there are points on S^1 with dense orbits. The corresponding surface D/Γ is a Riemann surface of constant negative curvature and finite area; we are concerned with the geodesic flow on the unit tangent bundle M of D/Γ. Γ has a fundamental region R in D which can be taken to be a polygon bounded by a finite number of circular arcs orthogonal to S^1. A vertex of R lying on S^1 is called a cusp. D/Γ is compact if and only if R has no cusps. Geodesics on D/Γ are the projections of circular arcs in D orthogonal to S^1.

If $g \in \Gamma$, $g(z) = (az+b)/(cz+d)$, then the circle $|cz+d| = 1$ is called the isometric circle of g, because $|g'(z)| > 1$ inside this circle and $|g'(z)| < 1$ outside. The isometric circle is always a circle orthogonal to S^1.

I suspect the idea that something like the ideas of this paper might work has occurred to a number of people. In particular, see the remark at the end of [10]. Certainly it had to both Adler and Moser, and I would like to thank both for the benefit of useful conversations.

§ 1. Symbolic representation of points on S^1

Let us recall briefly the constructions made in [6]. As explained in the introduction, Γ is a finitely generated Fuchsian group of the first kind acting in the unit disc D. Γ has a fundamental region R which consists of a polygon with a finite number of sides $\{s_i\}_{i=1}^{n}$; these sides extend to circular arcs $C(s_i)$ orthogonal to S^1. Each side s_i of R is identified with another side $A(s_i)$ by an element $g_i = g(s_i) \in \Gamma$; the set $\Gamma_0 = \{g_i\}_{i=1}^{n}$ is a symmetrical set of generators for Γ. The images of the sides $\{s_i\}$ under Γ form a net $\mathcal{N}$ in D. We will say R satisfies property (*) if:

(i) $C(s)$ is the isometric circle of s, and

(ii) $C(s)$ lies completely in $\mathcal{N}$.

Throughout § 1–§ 4, *we shall assume R satisfies* (*) *and moreover that R is not a triangle and does not have elliptic vertices of order* 2. (See [6].)

A typical fundamental region is shown in Fig. 1. (See also Fig. 1 of [6].)

We label the sides of R, $s_1, s_2, \ldots, s_n$ in anti-clockwise order; the vertex v_i is the intersection of s_{i-1} and s_i (with $s_0 = s_n$). $C(s_i)$ meets S^1 in P_i, Q_{i+1}, so that the order of points along $C(s_i)$ is $P_i, v_i, v_{i+1}, Q_{i+1}$.

$f = f_\Gamma$: $S^1 \to S^1$ is defined by $f_\Gamma(x) = g_i(x)$, $x \in [P_i P_{i+1})$. In [6] we showed that f_Γ has the following properties:

(a) Except for a finite number of pairs of points $x, y \in S^1$:

$$x = gy, \quad x, y \in S^1, \quad g \in \Gamma \Leftrightarrow \exists\, n, m \geqslant 0 \quad \text{such that } f^n(x) = f^m(y).$$

(b) f is Markov in the following sense:

There is a finite or countable partition of S^1 into intervals $\{I_i\}_{i=1}^\infty$ such that

(Mi) f is strictly monotonic on each I_i and extends to a C^2 function $\bar{f}_i$ on $\bar{I}_i$,

(Mii) $f(I_k) \cap I_j \neq \emptyset \Rightarrow f(I_k) \supseteq I_j$, $\forall j, k$,

(Miii) $\bigcup_{r=0}^\infty f^r(I_j) \supseteq I_k$, $\forall j, k$,

(Miv) If $\bar{I}_i = [a_i, b_i]$ then $\{\bar{f}_i(a_i), \bar{f}_i(b_i)\}_{i=1}^\infty$ is finite.

Moreover the partition $\{I_i\}$ is finite if and only if D/Γ is compact, or equivalently if R has no cusps.

(c) (Ei) If there are no cusps, then $\exists N > 0$ such that

$$\inf_{x \in S^1} |(f^N)'(x)| > \gamma > 1$$

(Eii) A cusp of R is a periodic point for f with derivative one. There is a subset $K \subseteq S^1$, consisting of a union of intervals I_i, so that if $f_K(x) = f^{n(x)}(x)$, $n(x) = \min\{n > 0: f^n(x) \in K\}$, $x \in K$, is the first return map induced on K, then $\exists N$ such that $\inf_{x \in K} |(f_K^N)'(x)| > \gamma > 1$.

To each point $x \in S^1$ we can associate a so-called f-expansion (cf. [1]). The usual way to do this is to write $x = i_0 i_1 i_2 \ldots$ if $f^n(x) \in \bar{I}_{i_n}$, $n = 0, 1, 2, \ldots$. (There is a slight ambiguity at the endpoints which we shall clarify below.) By (Mii) the rule determining which sequences $i_0 i_1 i_2 \ldots$ can occur is of finite type [8]; namely $i_r i_s$ occurs iff $f(\bar{I}_r) \supset \bar{I}_s$.

For our purposes it is better to label points using the generators Γ_0 of Γ, so we replace the partition $\{\bar{I}_i\}$ by $\{[P_i P_{i+1}] = [g_i]\}$. The rules determining which sequences are admissible is no longer of finite type. We say a sequence $e_1 e_2 \dots e_n \in \Gamma_0^n$ is admissible if $\bigcup_{r=1}^{n} f^{-r}([e_i^{-1}]) \neq \emptyset$. Let $\Sigma^+ = \{e_1 e_2 \dots \in \Gamma_0^N : e_k e_{k+1} \dots e_{k+l}$ is admissible $\forall k, l \in \mathbf{N}\}$. Define $\pi: \Sigma^+ \to S^1$ by $\pi(e_1 e_2 \dots) \to \bigcap_{r=1}^{\infty} f^{-r}([e_i^{-1}])$. The intersection is non-empty since this is true of all finite intersections and it contains at most one point because of the expanding condition (c). We discuss the topology of Σ^+ and continuity of π in § 4.

To see which sequences $e_1 e_2 \dots$ belong to Σ^+, it is enough to find those sequences $e_1 e_2 \dots e_m$ for which $\bigcap_{r=1}^{m} f^{-r}((e_r^{-1})) \neq \emptyset$, where $(e_r) = \operatorname{Int}[e_r]$.

To state the rules we need some more terminology. Starting at a vertex v_i with the side s_i and generator g_i, we get a cycle of vertices $v_i = w_1, \dots, w_p$ and corresponding generators $g_i = h_1, \dots, h_p$. ([9] Sec. 26 and [6] Lemma 2.4.) We say the anti-clockwise sequence $h_1^{-1} h_2^{-1} \dots h_p^{-1}$ is in left-hand (L) cyclic order. Similarly, starting at v_{i+1} with side s_i and generator g_i we get a cycle $v_{i+1} = z_1, z_2, \dots, z_q$ and generators $g_i = j_1, j_2, \dots, j_q$. We say the clockwise sequence $j_1^{-1} j_2^{-1} \dots$ is in right-hand (R) cyclic order. There exist integers μ, ν such that $(h_1^{-1} h_2^{-1} \dots h_p^{-1})^{\mu} = (j_1^{-1} j_2^{-1} \dots j_q^{-1})^{\nu} = 1$. $p\mu$ and $q\nu$ represent the number of sides of $\mathcal{N}$ which meet at the vertices v_i, v_{i+1} respectively, and therefore by (*), $p\mu = 2l$, $q\nu = 2k$ are even (see Fig. 1). We call L cycles of lengths $l-1$, l, $l+1$, D-(deficient), H-(half), and S-(superfluous) L cycles respectively, and similarly for R cycles of lengths $k-1$, k and $k+1$. A cycle of length $2l$ or $2k$ is called full. Notice that a full cycle is equal to the identity in Γ. If $h = g_i$, write $h^+ = g_{i+1}$ and $h^- = g_{i-1}$. If $B = b_1 \dots b_r$, $B^1 = b_1 \dots b_{r+1}$, $C = c_1 \dots c_s$ are L cycles with $c_1^{-1} = (b_{r+1}^{-1})^+$, we say B and C are adjacent or consecutive L cycles; similarly if B, B^1 and C are R cycles and $c_1^{-1} = (b_{r+1}^{-1})^-$ we say B, C are consecutive R cycles (see Fig. 2). A sequence $B_1, \dots, B_r$ of consecutive L cycles, where B_1, B_r are H-cycles and $B_2, \dots, B_{r-1}$ are D-cycles, will be called a L H-chain; such a sequence with B_1 a L D-cycle is a L D-chain. Often we represent a chain symbolically by $DD \dots DH$.

In Figs. 1 and 2 we indicate that the side s_i of R is associated to $g_i \in \Gamma_0$ by an arrow pointing into R. We write $\langle g_i^{-1} \rangle$ for the interval $[P_i P_{i+1})$ (the inverse is to make subsequent computations work properly) and write $x = g_{i_1} g_{i_2} \dots$ if $f^{n-1}(x) \in \langle g_{i_n} \rangle$, $n = 1, 2, \dots$.

Proposition 1.1. *A sequence $e_1 \dots e_p$, $e_i \in \Gamma_0$, is admissible if and only if*

(1) *gg^{-1}, $g \in \Gamma_0$, does not occur.*
(2) *No R H-cycles occur.*
(3) *No L S-cycles occur.*
(4) *No L H-chains occur.*

Proof. Referring to Fig. 1, let $P_i = C_k$, $P_{i+1} = C_1$, $Q_i = D_1$, $Q_{i+1} = D_l$. The arcs $z_1 C_1$, $z_1 C_2, \dots, z_1 C_k$ are the arcs of the net $\mathcal{N}$ emanating from z_1 and lying within the isometric

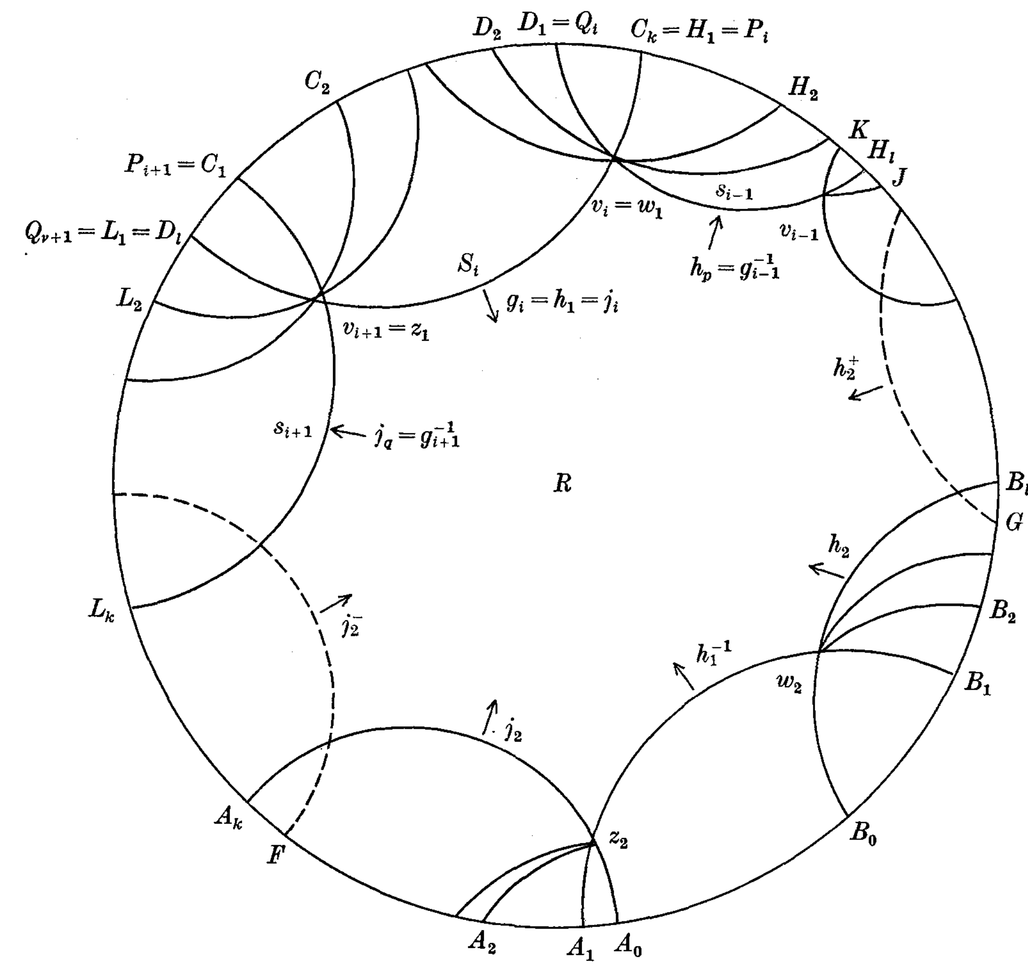

Fig. 1

circle $C(s_i)$ of g_i; similarly the arcs $w_1 D_1, \ldots, w_1 D_l$ are the arcs of $\mathcal{N}$ emanating from w_1 and lying within $C(s_i)$. By [6] Lemma 2.2, $w_1 D_{l-1}$ and $z_1 C_{k-1}$ do not intersect. $w_1, w_2, \ldots, w_p$ is the vertex cycle starting at w_1 with side s_i and $h_1, h_2, \ldots, h_p$ is the corresponding cycle of generators. Similarly $z_1, z_2, \ldots, z_q$ is the vertex cycle starting at z_1 with side s_i, with corresponding generators $j_1, j_2, \ldots, j_q$. $w_1 H_1, \ldots, w_1 H_l$; $z_1 L_1, \ldots, z_1 L_k$; $z_2 A_0, z_2 A_1, \ldots, z_2 A_k$; and $w_2 B_0, w_2 B_1, \ldots, w_2 B_l$ are all the arcs of $\mathcal{N}$ lying inside the isometric circles of h_p^{-1}, j_q^{-1}, j_2 and h_2 respectively. G, F and K are the endpoints of $C(h_2^+)$, $C(j_2^-)$, $C((h_p^{-1})^-)$ lying inside $C(h_2)$, $C(j_2)$, $C(h_p^{-1})$ respectively and J is the endpoint of the arc of $\mathcal{N}$ through v_{i-1} adjacent to but outside $C(h_p^{-1})$.

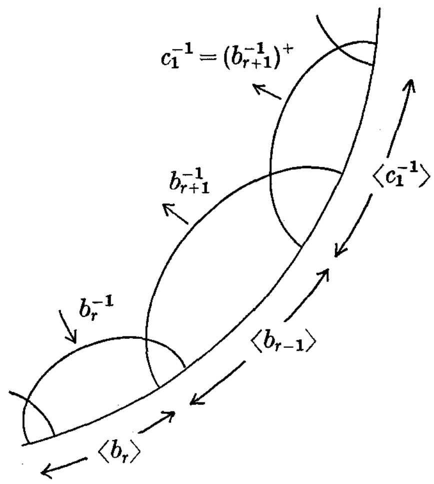

Fig. 2(*a*). Consecutive L cycles

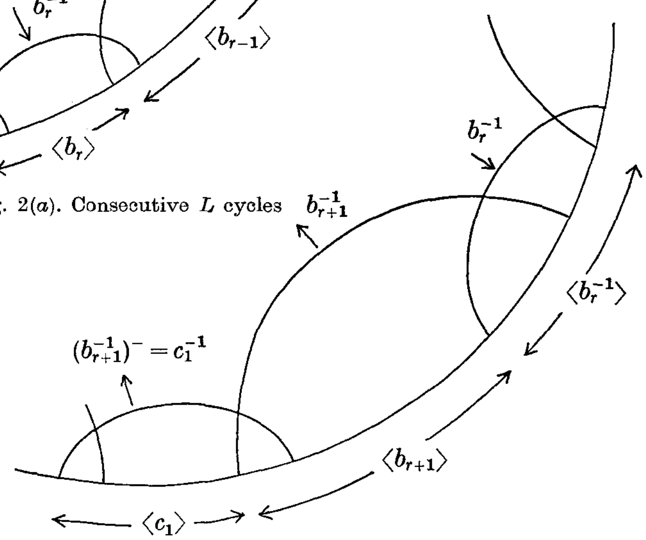

Fig. 2(*b*). Consecutive R cycles

(At a parabolic vertex, $l = \infty$ and we label points as H_∞, $H_{\infty-1}$, $H_{\infty-2}$, ... etc. and in computations treat ∞ exactly as any other integer.)

Notice that the map g_i carries D_l, z_1, w_1, C_k onto A_1, z_2, w_2, B_1 respectively; $C_1, ..., C_{k-1}$ onto $A_2, ..., A_k$ and $D_1, ..., D_{l-1}$ onto $B_2, ..., B_l$.

Now $f|_{[C_k, C_1)} = h_1 = j_1$. $f([C_k C_1))$ covers all intervals $\langle h \rangle$ except $\langle j_2^{-1} \rangle$, $\langle h_1 \rangle$ and $\langle h_2^{-1} \rangle$. Since $f(\langle h_1^{-1} \rangle) \cap \langle h_1 \rangle = \emptyset$, we get (1). $f([C_k C_r)) \cap \langle j_2^{-1} \rangle = [A_k A_{r+1})$, $1 \leqslant r \leqslant k-2$ and $f([C_k C_{k-1})) \cap \langle j_2^{-1} \rangle = \emptyset$. Moreover $f([C_k C_r)) \cap \langle h \rangle = f([C_k C_1)) \cap \langle h \rangle$ for $1 \leqslant r \leqslant k-1$ and $h \neq j_2^{-1}$. Therefore the sequence $j_1^{-1} j_2^{-1} ... j_k^{-1}$ is not admissible, but otherwise the restrictions following the symbols $j_1^{-1} ... j_r^{-1}$, $r \leqslant k-1$, are the same as those following j_r^{-1} alone. Rule (2) above follows.

Similarly we have

$$f([C_k C_1)) \cap \langle h_2^{-1} \rangle = [B_1 G),$$
$$f([D_r C_1)) \cap \langle h_2^{-1} \rangle = [B_{r+1} G), \quad 1 \leqslant r \leqslant l-2,$$
$$f([D_{l-1} C_1)) \cap \langle h_2^{-1} \rangle = \emptyset$$

and

$$f([D_r C_1]) \cap \langle h \rangle = f([C_k C_1]) \cap \langle h \rangle \quad \text{for} \quad 1 \leqslant r \leqslant l-2, \quad h \neq h_2^{-1},$$

$$f([D_{l-1} C_1]) \cap \langle h \rangle = f([C_k C_1]) \cap \langle h \rangle \quad \text{for} \quad h \neq h_2^{-1}, (h_2^+)^{-1}$$

and

$$f([D_{l-1} C_1]) \cap \langle (h_2^+)^{-1} \rangle = \langle (h_2^+)^{-1} \rangle - [GB_l).$$

Therefore the sequence $h_1^{-1} h_2^{-1} \dots h_{l+1}^{-1}$ is not admissible, which is rule (3).

The only restrictions following $h_1^{-1} \dots h_r^{-1}$, $r<l$, are the same as those following h_r^{-1} alone. Following $h_1^{-1} \dots h_l^{-1} h$, where $h \neq h_{l+1}^{-1}$, are the same restrictions as after h alone.

After $h_1^{-1} \dots h_l^{-1}(h_{l+1}^+)^{-1}$ is the same restriction as after $k^{-1}(h_{l+1}^+)^{-1}$, where k^{-1} is the element preceding $(h_{l+1}^+)^{-1}$ in the L order. Thus $(h_{l+1}^+)^{-1}$ is not the first element in a L H-cycle; also if $(h_{l+1}^+)^{-1}$ is the first element of a L D-cycle which ends in s^{-1}, followed by $(t^+)^{-1}$ where $s^{-1} t^{-1}$ are in L order, then $(t^+)^{-1}$ is not the first element of a L H-cycle.

Repetition of this argument gives rule (4), and we have examined all the possibilities for finite sequences $e_1 \dots e_p$. Σ^+ therefore consists of all sequences $e_1 e_2 \dots$ in which each finite block satisfies (1)–(4) above.

The map π: $\Sigma^+ \to S^1$ is of course not bijective. More precisely $x \in S^1$ has two representations in Σ^+ whenever $f^k(x) \in \{P_i\}_{i=1}^n$ for some $k \geqslant 0$. P_i can be written either as $DDD \dots$, an infinite string of consecutive R D-cycles, or as $HDD \dots$, an infinite string of consecutive L cycles.

Convention. In order to keep track of what is happening we shall in future adopt the following rule:

Whenever $x \in S^1$ has two symbolic expressions in Σ^+, we write $x = e_1 e_2 \dots$ where $e_1 e_2 \dots$ is the expression for x ending in L cycles.

This is equivalent to attaching P_i to the interval $(P_i P_{i+1})$ rather than $(P_{i-1} P_i)$.

Also notice $\pi\sigma(e) = f\pi(e)$, $e \in \Sigma^+$, provided e does not end in an infinite string of R D-cycles, where σ is the left shift on Σ^+.

Remark 1.2. In the case where R is a symmetric $4g$-sided polygon, our rules are identical with those of [13] p. 77 and closely related to those in [11] p. 791.

§ 2. Representation of geodesics in D

We would now like to represent a geodesic γ in D by taking the f-expansions of its endpoints P, Q, say $P = e_1 e_2 \dots$, $Q = f_1 f_2 \dots$ and writing $\gamma = \dots f_2 f_1 e_1 e_2 \dots$. Unfortunately, the sequence so obtained may not be admissible according to the rules of § 1. There are two problems: (i) Is the reversed sequence ... $f_2 f_1$ always admissible? And if so: (ii) When is ... $f_2 f_1 e_1 e_2$ ... admissible? The answer to (i) is no. It is perhaps more natural to consider the inverse sequence ... $f_2^{-1} f_1^{-1}$. This is however still in general inadmissible. To circumvent this difficulty we use the following trick:

$\bar{f}$-expansions. Recall that in defining f we made an arbitrary choice that $f|_{[P_i P_{i+1})} = g_i$. We could equally well have taken $f|_{(Q_{i-1} Q_i]} = g_i$; let us call this map $\bar{f}$. $\bar{f}$ obviously has exactly the same properties as f, and the admissibility rules are obtained by interchanging 'R' and 'L' in Proposition 1.1 above.

LEMMA 2.1. *Let $e_1 e_2$ ... be an admissible sequence for f. Then the inverse sequence ... $e_2^{-1} e_1^{-1}$ is admissible for $\bar{f}$, and vice versa.*

Proof. This follows easily from the remarks above, since an R cycle in $e_1 e_2$ ... becomes an L cycle in ... $e_2^{-1} e_1^{-1}$; and consecutive R cycles become consecutive L cycles.

Let $P, Q \in S^1$ and let $P = e_1 e_2 \ldots$, $Q = f_1 f_2 \ldots$ be the f- and $\bar{f}$-expansions of P, Q respectively. We shall call the directed geodesic γ joining Q to P admissible if $Q^{-1}.\ P = \ldots f_2^{-1} f_1^{-1} e_1 e_2 \ldots$ is admissible, and we shall also write $\gamma = \ldots f_2^{-1} f_1^{-1} e_1 e_2 \ldots$. Below in § 3 we shall see that admissible geodesics pass 'close' in a certain sense to the fundamental region R. This will deal with problem (ii) above.

Let Σ be the space of doubly infinite admissible sequences (i.e. all finite blocks satisfying (1)–(4) of Proposition 1.1) with left shift map σ.

To proceed we need to know something about the action of Γ_0 (the set of generators of Γ) on S^1 in terms of the symbolic representation of § 1.

PROPOSITION 2.2. *Let $x = e_1 e_2 \ldots \in \Sigma^+$, $g \in \Gamma_0$. Then*

(1) *$g(x) = g e_1 e_2 \ldots$ whenever $g e_1 e_2 \ldots \in \Sigma^+$ and*

(2) *$g(x) = e_2 e_3 \ldots$ if $g = e_1^{-1}$.*

Proof. We refer again to Fig. 1 with $g = h_1$.

(1) Suppose $g e_1 e_2$ ... is admissible. Then

(a) $g e_1 e_2$ ... does not begin with a R H-cycle.

(b) $g e_1 e_2$ ... does not begin with a L H-chain.

(c) $e_1 \neq g^{-1}$.

Observe that $g e_1 e_2$ ... begins with a R H-cycle iff $x = e_1 e_2 \ldots \in [H_2 H_1)$; $g e_1 e_2$ ... begins with a L H-chain iff $x \in [C_1 D_l)$. Therefore (a), (b), (c) together imply $x \notin [H_2 D_l)$.

Since $x \notin C(g)$, the isometric circle of g, $g(x) \in C(g^{-1}) \cap S^1 = \langle g \rangle \cup [B_0 B_1)$ (cf. [9] Sec. 11).

But $g(x) \notin [B_0 B_1)$ since $x \notin [H_2 H_1)$. Therefore $g(x) \in \langle g \rangle$, so $f(g(x)) = g^{-1}(g(x)) = x = e_1 e_2 \ldots$ and $g(x) = g e_1 e_2 \ldots$.

(2) Suppose $g = e_1^{-1}$. Then $x \in \langle g^{-1} \rangle$ and so $f(x) = g(x)$ and $g(x) = e_2 e_3 \ldots$.

It is possible to derive rules for the action of Γ_0 on Σ^+ in general. As this is not necessary for our development and the details become rather involved, we state without proof:

PROPOSITION 2.3. *Suppose $x \in S^1$, and $g \in \Gamma$. Let $x = e_1 e_2 \ldots$, $g(x) = f_1 f_2 \ldots$ be the f-expansions of x, $g(x)$. Then $\exists s, t > 0$ so that $g e_1 e_2 \ldots e_t = f_1 f_2 \ldots f_s$ in Γ and $e_{t+r} = f_{s+r}$, $r > 0$.*

Of course we have already proved the second part of this statement in [6], see property (a) of f_Γ in § 1.

This proposition is of interest because it enables us to prove the analogue of the results of Hedlund and Artin mentioned in the Introduction, that admissible geodesics are conjugate under Γ iff the corresponding sequences are shift equivalent. The proof is an easy consequence of Proposition 2.3:

PROPOSITION 2.4. *Let (P, Q), $(R, S) \in S^1 \times S^1$ be such that $Q^{-1} . P$, $R^{-1} . S \in \Sigma$. Then $\exists g \in \Gamma$ with $gP = R$, $gQ = S$ iff $\exists n \in S$ with $\sigma^n(Q^{-1} . P) = S^{-1} . R$.*

Proof. Let $P = e_1 e_2 \ldots$, $Q = f_1 f_2 \ldots$ be the f- and $\bar{f}$-expansions of P, Q respectively. We have $\ldots f_2^{-1} f_1^{-1} e_1 e_2 \ldots \in \Sigma$. By Proposition 2.2,

$$e_1^{-1}(P) = e_2 e_3 \ldots \quad \text{and} \quad e_1^{-1}(Q) = e_1^{-1} f_1 f_2 \ldots.$$

Hence $\sigma(Q^{-1} . P) = (e_1^{-1} Q)^{-1} . (e_1^{-1} P)$.

Conversely, suppose $P, Q \in S^1$ and $g \in \Gamma$ are such that $Q^{-1} . P$, $(gQ)^{-1} . (gP) \in \Sigma$. By Proposition 2.3, we have

$$P = e_1 \ldots e_n e_{n+1} \ldots \quad \text{and} \quad gP = u_1 \ldots u_m e_{n+1} \ldots$$

where $g e_1 \ldots e_n = u_1 \ldots u_m$.

Similarly, $Q = f_1 \ldots f_p f_{p+1} \ldots$, $gQ = v_1 \ldots v_q f_{p+1} \ldots$ and $g f_1 \ldots f_p = v_1 \ldots v_q$.

Thus $u_1 \ldots u_m e_n^{-1} \ldots e_1^{-1} = v_1 \ldots v_q f_p^{-1} \ldots f_1^{-1}$ and so

$$Q^{-1} . P = \ldots f_{p+1}^{-1} f_p^{-1} \ldots f_1^{-1} e_1 \ldots e_n e_{n+1} \ldots \quad \text{and} \quad (gQ)^{-1} . (gP) = \ldots f_{p+1}^{-1} v_q^{-1} \ldots v_1^{-1} u_1 \ldots u_m e_{n+1} \ldots$$

are shift conjugate.

This result is sufficient to show that the geodesic flow on D/Γ is ergodic, by the method used by Hedlund in [11]. Notice that the restriction to admissible geodesics with $Q^{-1} . P \in \Sigma$ corresponds to the restriction in [3] that the endpoints of geodesics lie in $(-1, 0)$ and $(0, \infty)$. For a discussion of the relevant measures, see Remark 4.4 below.

We shall instead follow the method of Morse to obtain a representation of the geodesic flow itself.

§. 3 Crossings of the fundamental region R

We now want to investigate in detail the relationship between the symbolic expansion $\gamma = \dots f_2^{-1} f_1^{-1} e_1 e_2 \dots$ of an admissible geodesic and the order in which γ cuts successive sides of the net $\mathcal{N}$. Recall that each side of R is labelled by a unique element $g \in \Gamma_0$. This label can be translated by an element of Γ to assign a unique element of Γ_0 to each (oriented) side of $\mathcal{N}$. The idea that γ should cut successively sides $\dots, f_2^{-1}, f_1^{-1}, e_1, e_2, \dots$ may unfortunately fail when γ passes too close to vertices in $\mathcal{N}$. What we shall show is

THEOREM 3.1. *For any $e \in \Sigma$, with corresponding directed geodesic γ, there is a distinguished copy $R(\gamma)$ of R such that*

(1) $\gamma \cap \overline{R(\gamma)} \neq \emptyset$

(2) $\gamma \cap \bar{R} \neq \emptyset \Rightarrow R(\gamma) = R$

(3) *γ cuts in succession $\overline{R(\gamma)}, \overline{\sigma^{-1} R(\sigma\gamma)}, \dots$ where $\sigma^{-n} = e_1 \dots e_n$ for $e = \dots f_2^{-1} f_1^{-1} e_1 e_2 \dots$.*

Throughout this section, by R we shall mean the open region bounded by the sides s_i.

Statement (3) needs a little interpretation when γ is a geodesic which goes through a vertex v of $\mathcal{N}$. Let $R_1, \dots, R_{2k}$ be the copies of R meeting at v, in anti-clockwise order round v. If γ passes from R_1 to R_{k+1} we say γ cuts $\bar{R}_1, \bar{R}_{2k}, \dots, \bar{R}_{k+1}$ in order. If γ coincides with the side of $\mathcal{N}$ between R_1 and R_2, we say γ cuts $\bar{R}_1, \bar{R}_{2k}, \dots, \bar{R}_{k+2}$ in order and if γ coincides with the side between R_1 and R_{2k}, γ cuts $\bar{R}_{2k}, \dots, \bar{R}_{k+1}$.

The idea of Theorem 3.1 is that if $\gamma \cap \bar{R} = \emptyset$, γ can be deformed by a sequence of 'small deformations' to a curve $\tilde{\gamma}$ such that $\tilde{\gamma} \cap R \neq \emptyset$ which cuts $R, \sigma^{-1} R$ in order. This sequence of deformations will determine $R(\gamma)$.

Let us make this more precise. As above, let v be a vertex of $\mathcal{N}$ where copies $R_1, \dots, R_{2k}$ of R meet, in anti-clockwise order round v. Let w_r, $1 \leqslant r \leqslant 2k$, be the vertex of $\mathcal{N}$ adjacent to v, along the side between R_r and R_{r+1} (see Fig. 3), and let A_r be the endpoint of this side on S^1.

A directed curve β will be said to pass near v if it passes from R_1 to R_{k+1} cutting the arcs $[vw_r)$, $1 \leqslant r \leqslant k$, or $[vw_r)$, $2k \geqslant r \geqslant k+1$, in order, see Fig. 3. If β cuts $[vw_r)$, $1 \leqslant r \leqslant k$, let $\tilde{\beta}$ be a curve which coincides with β everywhere except near v, where it cuts instead the arcs (vw_r), $2k \geqslant r \geqslant k+1$. $\tilde{\beta}$ is 'a small deformation of β round v'. R_{2k-r+2}, $2 \leqslant r \leqslant k$, is called the conjugate region to R_r, $R_{2k-r+2} = R_r^{*(\beta, v)}$. If β cuts $[vw_r)$, $2k \geqslant r \geqslant k+1$, we write $R_r = R_r^{*(\beta, v)}$, $2k \geqslant r \geqslant k+2$ and call R_r self-conjugate. We write $*(\beta, v) = *$ where there is no ambiguity.

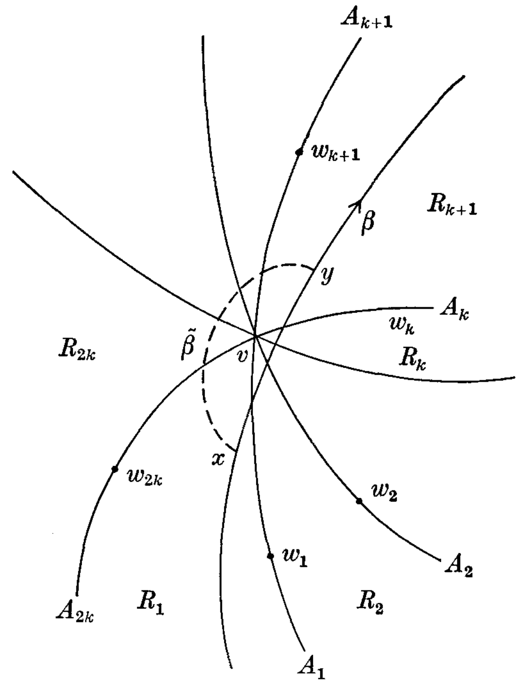

Fig. 3

We shall call a curve obtained from β by a sequence of small deformations a deformation of β. We make the same conventions about the order of regions cut by a deformed curve $\tilde{\gamma}$ through a vertex, as for geodesics γ.

Notice that the conjugate of a region S is a locally constant function of γ.

Lemma 3.2. *If the fundamental region R constructed in* [6] § 3 *has four sides, then at least eight sides meet at a vertex.*

Proof. It is straightforward to check all the cases in [6] to verify that R always has more than four sides, unless the signature of Γ is $\{1; 1; \nu_1\}$. But since $\nu_1 \geqslant 2$, and the corresponding R has interior angle $\pi/2\nu_1$, we see that in this case at least eight sides meet at a vertex.

Corollary 3.3. *There are no triangles formed by $\mathcal{N}$. If for edges of $\mathcal{N}$ form a quadrilateral, then at least eight sides meet at a vertex.*

Proof. Suppose the triangle or quadrilateral is not already a fundamental region. Then there is a vertex v of $\mathcal{N}$ on the interior of one of the sides of the region. Any other edge of

$\mathcal{N}$ through v forms a smaller triangle or quadrilateral. Proceeding in this way we eventually reach a region of minimal size which must be a copy of R.

LEMMA 3.4. *In a sequence of small deformations of a geodesic γ, a region S is associated to at most one conjugate region S^*, across a unique vertex v. Likewise S^* is the conjugate of at most one region S.*

Proof. If s is a side of S, let $B(s) \subseteq S^1$ be the arc of S^1 interior to the circle $C(s)$. Notice that if $\tilde{\gamma}$ is obtained from γ by a sequence of small deformations, and if $S^* \neq S$ is obtained by a deformation of $\tilde{\gamma}$ across the vertex v of S, and if s, s' are the sides of S meeting at v, then γ has one endpoint in $B(s) - B(s')$ and the other in $B(s') - B(s)$.

Similarly, if $\hat{\gamma}$ is a deformation of γ across a vertex w, at which meet sides t, t' of S, with conjugate region $S^{*\prime} = S$, then γ has its endpoints in $B(t) - B(t')$, $B(t') - B(t)$.

If u, u' are sides of S then since extensions of non-adjacent sides of S do not meet ([6] Lemma 2.2), we have $B(s) \cap B(t) = \emptyset$ unless $s = t$ or s, t are adjacent. After interchanging s with s' and t with t' if necessary, there are three cases:

Case 1. $s = t$, $s' = t'$. Then $v = w$ and clearly $S^* = S^{*\prime}$.

Case 2. $s = t$, $s' \neq t'$. $B(t') - B(t)$ is disjoint from $B(s) - B(s')$, so $B(t') \cap B(s') \neq \emptyset$ since it contains an endpoint of γ. Then t', s' are adjacent. But this means R has only three sides, s, t', s', which is impossible.

Case 3. $s \neq t$, $s' \neq t'$. Without loss of generality, we may suppose $(B(t) - B(t')) \cap (B(s) - B(s')) \neq \emptyset$. Then s, t are adjacent. In this case we also have $B(t') \cap B(s') \neq \emptyset$, since this set contains an endpoint of γ. Hence s', t' are adjacent. Then R has four sides, s, s', t and t'. Since non-adjacent sides of R do not meet, γ has its endpoints in sectors of the vertex star at v separated by one sector only, namely that containing S. But since by Lemma 3.2 at least eight copies of R meet at v, the endpoints of γ do not then lie in diametrically opposite sectors at v. Then γ does not pass near v, which is contrary to assumption.

The final statement is proved by exactly the same argument.

Thus we may write $S^* = S^*(\gamma)$, independent of v and the deformation $\tilde{\gamma}$.

LEMMA 3.5. *Let γ be a geodesic. Then γ cuts a region $\bar{S}$ at most once, and if $\gamma \cap \bar{S} \neq \emptyset$ and $S \neq S^*$, then $\gamma \cap S^* = \emptyset$.*

Proof. If γ cut $\bar{S}$ more than once, then $\#(\gamma \cap \partial S) > 2$. But $\#(\gamma \cap \partial S) \leqslant 2$, since S is geodesically convex. (This uses the fact that the interior angles of S are all less than π, and the formula $A = \pi(n-2) - \sum \alpha_i$ for the area of a geodesic polygon.)

Suppose γ passes near the vertex v of S and sides s, s' meet at v. If $\gamma \cap S^* \neq \emptyset$ then γ would have to cross the extensions $C(s)$, $C(s')$ of s, s' twice, which is impossible.

LEMMA 3.6. *Let $\tilde{\gamma}$ be a deformation of a geodesic γ. Suppose γ cuts in order $\bar{R}_1, ..., \bar{R}_n$ (with the above conventions if γ passes through a vertex of $\mathcal{N}$). Then $\tilde{\gamma}$ cuts in order $\bar{\tilde{R}}_1, ..., \bar{\tilde{R}}_n$ where $\tilde{R}_i$ is one of R_i, R_i^*.*

Proof. This follows easily by induction on the number of small deformations. For one deformation it is clear from the definitions.

COROLLARY 3.7. *Let $\tilde{\gamma}$ be a deformation of a geodesic γ and suppose $\tilde{\gamma} \cap S \neq \emptyset$. Then either $\gamma \cap \bar{S} \neq \emptyset$ or there is a unique region S_1 with $\gamma \cap \bar{S}_1 \neq \emptyset$ and $S = S_1^*$.*

Proof. Let $..., \bar{R}_1, \bar{R}_2, ...$ be the sequence of regions cut by γ. By Lemma 3.6, $S = R_i$ or R_i^* for some i. If $S = R_i$ we are done. If $S = R_i^*$ and $R_i = R_i^*$ then $\gamma \cap \bar{R}_i \neq \emptyset$. Suppose $\gamma \cap \bar{S} \neq \emptyset$ and there is a region $T \neq R_i$ with $\gamma \cap \bar{T} \neq \emptyset$, $T^* = S$. Then $T = R_j$ for some j and $R_i^* = R_j^*$. By Lemma 3.4, $R_i = R_j$.

LEMMA 3.8. *Let $v, R_1, ..., R_{2k}$ be as in Fig. 3. Let α be a geodesic with endpoints in $(A_{2k}A_1)$, $(A_k A_{k+1})$, cutting in order $R_2, R_3, ..., R_k$. Then there is a deformation $\tilde{\alpha}$ of α which cuts in order $R_1, R_{2k}, ..., R_{k+1}$.*

Proof. Let $x_0 = v, x_1 = w_1, x_2, ...; y_0 = v, y_1 = w_k, y_2, ...$ be the vertices of $\mathcal{N}$ along $[vA_1)$, $[vA_k)$ and suppose α cuts $[vA_1)$ on $[x_p x_{p+1})$ and $[vA_k)$ on $[y_q y_{q+1})$. Let l be any edge of $\mathcal{N}$ through $u \in \{x_i\}_0^p$, other than A_1vA_{k+1} or A_kvA_{2k}. l has an endpoint L in (A_1A_k), otherwise l, A_1vA_{k+1} and A_kvA_{2k} would form a triangle. Let z be the vertex of $\mathcal{N}$ adjacent to u on $[uL)$. Let m be a side of $\mathcal{N}$ distinct from l through z. We can suppose m has one endpoint in (LA_k), for otherwise l, m, A_kvA_{2k} and A_1vA_{k+1} form a quadrilateral. In this case pick $m^1 \neq m, l$ through z (possible since $\geqslant 8$ sides meet at z). Then either m^1, m, vA_k form a triangle, which is impossible, or m^1 has an endpoint in (LA_k). The other endpoint of m^1 lies in (A_1L), otherwise m^1, l and vA_1 form a triangle.

Then either $\alpha \cap l \in [uz)$, or m^1 cuts α twice or touches α, both of which are impossible. So $\alpha \cap l \in [uz)$.

We now see α passes near x_p. For by the above, α cuts every side of $\mathcal{N}$ through x_p between x_p and the adjacent vertex of $\mathcal{N}$ in the direction of (A_1A_k). Deforming round x_k, we see repeating the argument $\tilde{\alpha}$ passes near x_{p-1}, etc.

Similarly α can be deformed round $y_q, y_{q-1}, ...$. Let $\tilde{\alpha}$ be the curve obtained by deforming successively round $x_p, \ldots, x_1, y_q, \ldots, y_1$. Then $\bar{\alpha}$ passes near $x_0 = v$, and deforming round v we get the required result.

Let $W = \{P \in S^1 : P$ is the endpoint of a geodesic in $\mathcal{N}$ through a vertex of $R\}$.

PROPOSITION 3.9. *Suppose $\gamma = Q^{-1}_{\cdot}P \in \Sigma$. Then γ can always be deformed to a curve γ^* which cuts R, $\gamma^{-1}R$ in succession, unless possibly $P \in W$ or $Q \in W$. In this case either γ is a side of $\mathcal{N}$ and cuts $\bar{R}$, $\overline{\sigma^{-1}R}$ in succession or γ is not a side of $\mathcal{N}$ and there are geodesics $\gamma' = Q'^{-1}_{\cdot}P' \in \Sigma$ arbitrarily close to γ, with $P', Q' \notin W$.*

Proof. We refer throughout to Fig. 1. Without loss of generality we may assume $P \in [C_k C_1)$. This means $\sigma^{-1} = g_i^{-1}{}_{\cdot}\, g_i^{-1} R$ is the copy of R adjacent to R along s_i.

If Q lies outside all the circles $C(s_{i-2})$, $C(s_{i-1})$, $C(s_i)$, $C(s_{i+1})$ it is clear that $\gamma \cap R \neq \emptyset$, and that either $\gamma \cap (s_i) \neq \emptyset$, or $\gamma \cap (s_{i-1}] \neq \emptyset$. ($(s_{i-1}] = (v_{i-1} v_i]$.) In the first case γ cuts in succession $R, \sigma^{-1}R$. Otherwise $P \in [C_k D_1)$. If $P \in (C_k D_1)$, we are in the situation of Lemma 3.8 relative to v_i, so γ can be deformed to cut R, $\sigma^{-1}R$ in order.

If $P = C_k$ then $\gamma' = Q^{-1}_{\cdot}P^1$ where $P^1 \in (C_k D_1)$ is admissible. If $Q \in (C_k D_l]$ then $Q^{-1}_{\cdot}P \notin \Sigma$.

Suppose $Q \in (L_r L_{r+1}]$ $1 \leqslant r \leqslant k-1$. Then the $\mathfrak{f}$-expansion of Q begins with an L cycle of length $k-r$. Since $Q^{-1}_{\cdot}P \in \Sigma$, P begins with an R cycle of length at most $r-1$, so that $P \in [C_k C_{k-r+1})$. This means γ lies outside the circle $L_r v_{i+1} C_{k-r+1}$, so $\gamma \cap R \neq \emptyset$, and γ cuts $\sigma^{-1}R$ after R.

Suppose $Q \in (H_{s+1} H_s]$, $1 \leqslant s \leqslant l-2$, or $Q \in (K H_{l-1}]$ and $s = l-1$. The $\mathfrak{f}$-expansion of Q begins with an R cycle A_1. If A_1 is followed by consecutive R cycles $A_2, \ldots, A_n$ of lengths $D, \ldots, D, H$ respectively then A_1 has length $l-s-1$, otherwise A_1 has length $l-s$. Therefore since $Q^{-1}_{\cdot}P \in \Sigma$, if P begins with an L cycle B_1, and B_1 is followed by consecutive L cycles $B_2, \ldots, B_m$ of lengths $D, \ldots, D, H$ then B_1 has length at most $s-1$; otherwise B_1 has length at most s. This means that $P \in [D_{l-s} C_1)$.

Now if $\gamma \cap R \neq \emptyset$ the result is obvious. Otherwise unless $P = D_{l-s}$ or $Q = H_s$, or γ is a side of $\mathcal{N}$, we are in the situation of Lemma 3.8, with Q, P in the diametrically opposite sectors $(H_{s+1} H_s)$, $(D_{l-s} D_{l-s+1})$ at v. Applying Lemma 3.8 we get the required deformation. If $P = D_{l-s}$ or $Q = H_s$, and $P' \in (D_{l-s} C_1)$, $Q' \in (H_{s+1} H_s)$ then $\gamma' = Q'^{-1}_{\cdot}P' \in \Sigma$. If γ is a side of $\mathcal{N}$ γ cuts $\bar{R}$, $\overline{\sigma^{-1}R}$ in order.

If $Q \in C(S_{i-2}) - (H_l K]$, either γ already cuts $R, \sigma^{-1}R$ or γ has endpoints in the diametrically opposite sectors $(D_l H_l]$, $[H_1 D_1)$ at v_i and so can be deformed as required, or if $P = H_1$ or $Q = H_l$, replace by $P' \in (H_1 D_1)$, $Q' \in (D_l H_l)$.

Finally if $Q \in (H_l K]$ the $\mathfrak{f}$-expansion of Q begins with a sequence of consecutive R cycles of lengths $D, \ldots, D, H$ beginning with g_{i-2}^{-1}. Hence P does not begin with an L chain $DD \ldots DH$, i.e. $P \notin [C_k D_1)$. But then either γ cuts $R, \sigma^{-1}R$; or γ has endpoints in the diametrically opposite segments $(H_l H_{l-1})$, $(D_1 D_2)$ and we apply Lemma 3.8; or γ is not a side of $\mathcal{N}$ and there are curves γ' close to γ with endpoints in $(H_l H_{l-1})$, $(D_1 D_2)$; or $\gamma = H_l^{-1} \cdot D_1$ and γ cuts $\bar{R}$, $\overline{\sigma^{-1}R}$.

Now let $\gamma = Q^{-1} \cdot P \in \Sigma$ and suppose we can find a deformation γ^* with $\gamma^* \cap R \neq \emptyset$. By Corollary 3.7 either $\gamma \cap \bar{R} \neq \emptyset$ or there is a unique region R_1 with $\gamma \cap \bar{R}_1 \neq \emptyset$ and $R = R_1^*$. If $\gamma \cap \bar{R} \neq \emptyset$ set $R(\gamma) = R$; otherwise set $R(\gamma) = R_1$. It is clear from Lemma 3.4 that $R(\gamma)$ is independent of the deformation γ^*.

Suppose $\gamma = Q^{-1} \cdot P \in \Sigma$ with no deformation γ^* with $\gamma^* \cap R \neq \emptyset$, and that γ is not a geodesic in $\mathcal{N}$. By Proposition 3.9 we see there are geodesics $\gamma' = Q'^{-1} \cdot P' \in \Sigma$ arbitrarily close to γ, with $\gamma'^* \cap R \neq \emptyset$. We observed above that for any region S, S^* is a locally constant function of S. Therefore we may define $R(\gamma) = R(\gamma')$ for γ' close to γ.

If $\gamma \in \Sigma$ is a side of $\mathcal{N}$, set $R(\gamma) = R$. By Proposition 3.9, γ cuts $\bar{R}$, $\overline{\sigma^{-1}R}$ in succession. In this case $\sigma\gamma$ is also a side of $\mathcal{N}$ and so $R(\sigma\gamma) = R$. Thus γ cuts $\overline{R(\gamma)}$, $\overline{\sigma^{-1}R(\sigma\gamma)}$ in succession.

Suppose $\gamma \in \Sigma$ is not a side of $\mathcal{N}$ and let γ^* be a deformation which cuts R, $\sigma^{-1}R$ in succession. By Lemma 3.6 there are regions R_1, R_2 so that γ cuts $\bar{R}_1$, $\bar{R}_2$ in succession and $R = R_1$ or $R_1^{*(\gamma)}$, $\sigma^{-1}R = R_2$ or $R_2^{*(\gamma)}$. $R(\gamma) = R_1$ by definition.

Now $\sigma\gamma^*$ cuts R. If $\sigma\gamma \cap \bar{R} \neq \emptyset$, $R(\sigma\gamma) = R$. Then γ cuts $\overline{R(\gamma)}$, $\overline{\sigma^{-1}R(\sigma\gamma)}$ in succession.

Otherwise $\sigma\gamma \cap \bar{R} = \emptyset$ but $\sigma\gamma^* \cap R \neq \emptyset$ and $\sigma\gamma \cap \sigma\bar{R}_2 \neq \emptyset$. Thus $R \neq \sigma R_2$ and so $R = \sigma(R_2^{*(\gamma)})$. Since σ is an automorphism, $\sigma(R_2^{*(\gamma)}) = (\sigma R_2)^{*(\sigma\gamma)}$, and thus $\sigma\gamma \cap \overline{\sigma R_2} \neq \emptyset$ and $(\sigma R_2)^{*(\sigma\gamma)} = R$, which implies $R(\sigma\gamma) = \sigma R_2$. Thus γ cuts $\overline{R(\gamma)}$, $\overline{\sigma^{-1}R(\sigma\gamma)}$ in succession.

Finally suppose $\gamma \in \Sigma$ is not a side of $\mathcal{N}$ and is close to a curve γ' which cuts $\overline{R(\gamma')}$, $\overline{\sigma^{-1}R(\sigma\gamma')}$ in order. Taking γ' sufficiently close to γ we have $R(\gamma) = R(\gamma')$ and $R(\sigma\gamma') = R(\sigma\gamma')$. Moreover we may assume γ' cuts $R(\gamma')$, $\sigma^{-1}R(\sigma\gamma')$ and so γ cuts $\overline{R(\gamma)}$, $\overline{\sigma^{-1}R(\sigma\gamma)}$.

Now applying Proposition 3.9 to $\sigma^{-1}\gamma$, we may find a deformation of $\sigma^{-1}\gamma$ which cuts $\bar{R}$, $\overline{\sigma^{-1}R}$ in succession, and hence a deformation of γ which cuts σR, $\bar{R}$ in succession. Applying similar reasoning to the above, we see γ cuts $\overline{\sigma R(\sigma^{-1}\gamma)}$, $\overline{R(\gamma)}$ in succession. A simple inductive argument and repeated application of Lemma 3.5 completes the proof of Theorem 3.1.

It is obvious that, for any $\gamma \in \Sigma$, there is a unique $g \in \Sigma$ with $gR(\gamma) = R$. We shall need a converse to this:

PROPOSITION 3.10. *Let γ be any geodesic with $\gamma \cap \bar{R} \neq \emptyset$. Then there exists a unique $g \in \Gamma$ so that $g\gamma \in \Sigma$ and $R(g\gamma) = gR$.*

Proof. Suppose $g \in \Gamma$ is such that $g\gamma \in \Sigma$ and $R(g\gamma) = gR$. If $R(g\gamma) = R$, then $g = \mathrm{id}$. Otherwise, $R(g\gamma)^{*(g\gamma)} = R = g^{-1}R(g\gamma)$. Since g is an automorphism, $g^{-1}(R(g\gamma)^{*(g\gamma)}) = [g^{-1}R(g\gamma)]^{*(\gamma)}$, i.e. $R^{*(\gamma)} = g^{-1}R$. Therefore g, if it exists, is unique.

If $\gamma \in \Sigma$ then $R(\gamma)=R$ and we may take $g=\mathrm{id}$.

So suppose $\gamma=Q^{-1}_{\cdot}P \notin \Sigma$. Without loss of generality, we may assume $P\in[C_k C_1)$. If $Q^{-1}_{\cdot}P \notin \Sigma$ we must have $Q \in (H_l L_k]$ (see the proof of Proposition 3.9). Clearly $Q \notin (C_k L_1]$, for then $\gamma \cap \bar{R}=\emptyset$.

Suppose that $Q\in(L_r L_{r+1}]$, $1 \leqslant r \leqslant k-1$. Arguing as in Proposition 3.9, we see P begins with an R cycle of length at least r, so $P\in[C_{k-r+1}C_1)$. Since $\gamma \cap \bar{R} \neq \emptyset$, we must have $P \in [C_{k-r+1}C_{k-r})$, the sector at v_{i+1} diametrically opposite $(L_r L_{r+1}]$. Suppose $Q \neq L_{r+1}$, $P \neq C_{k-r+1}$. Then by Lemma 3.8 we see we can deform γ to obtain a conjugate $R^{*(\gamma)} \neq R$. Pick g so that $gR^*=R$. Now relabel the vertices so that $gP\in[C_k C_1)$. Then $g\gamma$ passes to the right of gv_{i+1} and gP, gQ are in diametrically opposite sectors at gv. Moreover gv_{i+1} is a vertex of R, and since $\gamma \cap R^*=\emptyset$, $g\gamma \cap R=\emptyset$. This forces (with the new labelling), $gv_{i+1}=v_i$, $gP\in(D_1 C_1)$ and $gQ\in(H_l H_1)$. Now as in the proof of Proposition 3.9, $(gQ)^{-1}_{\cdot}gP\in\Sigma$. Clearly $g\gamma \cap \bar{R}=\emptyset$, so as in Proposition 3.9 there is a unique region R_1 with $R_1^{*(g\gamma)}=R$ and $g\gamma \cap R_1 \neq \emptyset$, and $R_1=R(g\gamma)$. Now $R_1^{*(g\gamma)}=g((g^{-1}R_1)^{*(\gamma)})$, since g is an automorphism and thus $g^{-1}R=(g^{-1}R_1)^{*(\gamma)}$. But $g^{-1}R=R^{*(\gamma)}$, therefore by Lemma 3.4, $g^{-1}R_1=R$. Since $R_1=R(g\gamma)$, g is as required.

If either $Q=L_{r+1}$ or $P=C_{k-r+1}$ we apply the same g as for nearby γ' and use obvious continuity arguments.

Now if $Q=L_1$, $P\in[C_k C_1)$ and $\gamma \cap \bar{R} \neq \emptyset$, we must have $P=C_k$. Then we may take $g=\mathrm{id}$.

Finally suppose $Q\in(H_{s+1}H_s]$, $1\leqslant s \leqslant l-2$, or $Q\in(KH_{l-1}]$ and $s=l-1$. Since $\gamma \cap \bar{R}\neq\emptyset$ we see $P\in[D_{l-s}C_1)$. Just as in the proof of Proposition 3.9, this shows $Q^{-1}_{\cdot}P\in\Sigma$. Thus we may take $g=\mathrm{id}$.

§ 4. Symbolic representation of the geodesic flow

In this section we show that the geodesic flow on $T_1(D/\Gamma)$ can be represented as a quotient of a special flow over Σ, σ; where the height function is the time taken to cross the region $R(\gamma)$. We keep the notation and conventions of § 1–§ 3.

If γ is an admissible geodesic, let $h(\gamma)$ be the hyperbolic length of $\gamma \cap R(\gamma)$. h is infinite if an endpoint of γ is a cusp. h lifts to a function also denoted by h on Σ. Let $\Lambda=\{(e,t)\colon e\in\Sigma,\ 0\leqslant t<h(e)\}$ and let φ_τ be the special flow on Λ defined by $\varphi_\tau(e,t)=(\sigma^n e, t+\tau-S_n h(e))$ when $\tau>0$ and $0\leqslant t+\tau-S_n h(e)<h(\sigma^n e)$ with a similar definition for $\tau<0$, where $S_n h(e)=\sum_0^{n-1} h\sigma^r(e)$.

(Notice that $\sum_0^\infty h(\sigma^r\gamma)$ diverges because an arc of γ of finite length can cut only finitely many copies of R.)

Let ψ_t be the geodesic flow on the unit tangent bundle M of D/Γ, let $\tilde{M}$ be the unit tangent bundle of D and let $p\colon \tilde{M}\to M$ be projection. $\tilde{\psi}_t$ is geodesic flow on $\tilde{M}$.

For an admissible geodesic γ, let $b(\gamma) \in \tilde{M}$ be the unit tangent vector pointing along γ based at the point where γ enters $R(\gamma)$.

Define $\Pi\colon \Lambda \to M$ by

$$\Pi((e, t)) = \psi_t(pb(e)),$$

where $\pi(e)$ is the geodesic associated to e. In what follows we shall frequently identify e and $\pi(e)$.

Proposition 4.1. *Π is surjective, $\Pi\varphi_t = \psi_t \Pi$ and $\#\Pi^{-1}(\Pi(e,t)) = \#\pi^{-1}(\pi(e))$ for $e \in \Sigma$ (i.e. Π is 1–1 except on a set of the first category).*

Proof. Take $u \in M$. Lift u to $\tilde{u} \in \tilde{M}$ with the property that $\tilde{u}$ has its endpoint U in $\bar{R}$. If γ is the geodesic through U in the direction $\tilde{u}$, $\gamma \cap \bar{R} \neq \emptyset$.

By Proposition 3.10, there is a unique $g \in \Gamma$ with $g\gamma \in \Sigma$ and $R(g\gamma) = gR$. $g\tilde{u}$ is also a lifting of u, and $g\gamma \cap \overline{R(g\gamma)} \neq \emptyset$. Let τ be the hyperbolic distance along $g\gamma$ from the point V where $g\gamma$ enters $\overline{R(g\gamma)}$ to gU. Since $U \in \bar{R}$, $gU \in g\bar{R} = \overline{R(g\gamma)}$. Then $0 \leqslant \tau < h(g\gamma)$ (or $h(g\gamma) = 0$), and $g\tilde{u} = \tilde{\psi}_\tau b(g\gamma)$. Also $\Pi(g\gamma, \tau) = \psi_\tau(pb(g\gamma)) = p\tilde{\psi}_\tau b(g\gamma) = p(g\tilde{u}) = u$. Therefore Π is surjective.

Suppose also $\Pi(e, t) = u$, $e \in \Sigma$. Let $\pi(e) = \beta$. Then $u = \psi_t p(b(\beta)) = p\tilde{\psi}_t(b(\beta))$. Thus there is an $h \in \Gamma$ so that $hg\tilde{u} = \tilde{\psi}_t b(\beta)$, and so $h^{-1}b(g\gamma) = b(\beta)$. Thus $b(\beta)$ is the unit tangent vector along $h^{-1}g\gamma$ based at the point where $h^{-1}g\gamma$ enters $h^{-1}R(g\gamma)$. This means $h^{-1}g\gamma = \beta$ and $h^{-1}R(g\gamma) = R(\beta)$, i.e. $h^{-1}gR = R(\beta)$. According to Proposition 3.10, $h^{-1}g$ is unique and $h = \mathrm{id}$, $\beta = g\gamma$ certainly works. Therefore $\Pi(e, t) = u$ iff $\pi(e) = g\gamma$. Observe π is one, two or four-to-one depending on whether $g\gamma$ has neither, one or both its endpoints in $\bigcup_{r=0}^{\infty} \sigma^{-r} W$.

Suppose $(e, t) \in \Lambda$, $e = \dots f_2^{-1} f_1^{-1} e_1 e_2 \dots$, $\tau > 0$ and $S_n h(e) \leqslant t + \tau < S_{n+1} h(e)$.

Then

$$\tilde{\psi}_{h(e)} b(e) = \sigma^{-1} b(\sigma e) \tag{4.1.1}$$

by Theorem 3.1 (3).

Thus

$$\begin{aligned} &\tilde{\psi}_{S_n h(e)}(\sigma^n b(e)) \\ &= \tilde{\psi}_{S_n h(e)}(\sigma^n \tilde{\psi}_{h(e)} b(e)) \\ &= \tilde{\psi}_{S_n h(e)}(\sigma^{n-1} b(\sigma e)) \qquad \text{by (4.1.1)} \\ &= \dots = b(\sigma^n e) \end{aligned} \tag{4.1.2}$$

and

$$
\begin{aligned}
\Pi(\varphi_\tau(e,t)) &= \psi_{t+\tau-S_nh(e)}(pb(\sigma^n e))\\
&= \psi_{t+\tau}p\tilde{\psi}_{-S_nh(e)}(b(\sigma^n e))\\
&= \psi_{t+\tau}p(\sigma^n b(e)) \qquad \text{by (4.1.2)}\\
&= \psi_{t+\tau}p(b(e))\\
&= \psi_\tau(e,t).
\end{aligned}
$$

A similar computation works for $\tau < 0$.

We now want to investigate the continuity of Π and h. Put on Σ the usual product topology and metric

$$d((e_i),(e_i')) = 2^{-n}, \quad n = \sup\{m: e_i = e_i',\ |i| \leqslant m\}.$$

PROPOSITION 4.2. *$\pi: \Sigma^+ = S^1$ is continuous.*

Proof. In the no cusp case this follows easily from Property (Ei) of f in § 1, see also the last line of the proof below.

Suppose C is a cusp of R. Suppose the L cycle of generators at C is $h_1, \ldots, h_l$. Let $H = h_l \ldots h_1$. Then $H(C) = C$ and $H'(C) = 1$. By Lemma 2.8 of [6], H acting on S^1 with fixed point C is conjugate by a Möbius transformation to

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

acting on $\mathbf{R}$ with fixed point 0, with $y > 0$. Let $J(H^m) = \{P \in S^1: P = H^{-m} \ldots\}$. One sees easily $J(H^m)$ corresponds to $(\alpha(my+1)^{-1}, 0]$ for some $\alpha < 0$. Therefore $P, Q \in J(H^m) \Rightarrow |P-Q| = O(m^{-1})$ on S^1.

Now pick $P \in S^1$ and suppose P corresponds to $e = H_1^{m_1} B_1 H_2^{m_2} B_2 \ldots \in \Sigma^+$ where H_i is a cycle corresponding to a parabolic vertex and B_i is a block containing no such cycles. Suppose given $\varepsilon > 0$.

Say $\exists\, m_r$ so that $1/m_r < \varepsilon$. Let the length of the sequence $H_1^{m_1} B_1 H_2^{m_2} \ldots B_{r-1}$ be N. Then $d(e', e) < 2^{-N} \Rightarrow \sigma^N Q, \sigma^N P \in J(H_r^{m_r})$ where $Q = \pi(e')$. Also $\sigma_{e'}^r = \sigma_e^r$ for $1 \leqslant r \leqslant N$ and $|\sigma'| \geqslant 1$ on S^1. Therefore $|P-Q| < K\varepsilon$, for some K depending only on Γ.

Otherwise, $\exists L$ such that $m_r \leqslant L$, $\forall r$. Thus $P \notin J(H^L)$ for any parabolic vertex, so $\sigma^k P$ is a bounded distance away from all the parabolic vertices for each k. Since $\sigma'(x) = 1$ only at parabolic vertices, this means $\exists \lambda > 1$ such that $(\sigma_e^k)' \geqslant \lambda$ for all k. Choose N so that $\lambda^{-N} < \varepsilon$. If $d(e', e) < 2^{-N}$ then $\sigma_{e'}^k = \sigma_e^k$, $k \leqslant N$ and so $|P-Q| < \lambda^{-N}$.

COROLLARY 4.3. *π: $\Sigma \to S^1 \times S^1$ is continuous.*

Let $\Sigma^* = \{e \in \Sigma$: neither endpoint of e on S^1 is a cusp$\}$.

PROPOSITION 4.4. *h is continuous on Σ^*. In the no cusp case, h is Hölder on Σ.*

Proof. We take the no cusp case first.

Let λ be an admissible geodesic in D with endpoints $P = e^{i\theta}$, $Q = e^{i\varphi}$. Suppose C_1, C_2 are disjoint geodesics which are cut within bounded arcs by γ. The hyperbolic distance between C_1 and C_2 along γ is a smooth function of θ, φ. Hence if γ' is a geodesic with endpoints $P' = e^{i\theta'}$, $Q' = e^{i\varphi'}$, then $|d - d'| \leqslant K(|\theta - \theta'| + |\varphi - \varphi'|)$ where K depends only on C_1, C_2.

Let $\lambda > 1$ be the expansive constant for σ. Suppose $d(\gamma, \gamma') < 2^{-n}$. Then $|\theta - \theta'| \leqslant \lambda^{-n}$, $|\varphi - \varphi'| \leqslant \lambda^{-n}$.

$R(\gamma)$ always has a vertex in common with R and so is one of a finite number of regions. Thus $h(\gamma)$ is the distance along γ between a finite number of possible pairs of sides of $\mathcal{N}$. Provided γ does not pass through a vertex of $R(\gamma)$, $|h(\gamma) - h(\gamma')| \leqslant K\gamma^{-n}$ for K independent of γ.

Suppose γ enters $R(\gamma)$ across a geodesic C_1 and leaves across the intersection of C_2 and C_3. $h(\gamma')$ for γ' near γ is the distance along γ' from C_1 to one of C_2, C_3. Both these function are Hölder and their values coincide at γ. Likewise, if γ coincides with a side of $\mathcal{N}$, $R(\gamma')$ is one of a finite number of regions meeting $R(\gamma)$ and we see $h(\gamma')$ is one of a finite number of Hölder functions all of whose values agree at γ.

Now suppose R has cusps. Let K_r be the part of D outside small discs of (Euclidean) radius r round each of the cusps of R.

The above argument shows that h is continuous on geodesics γ which lie completely inside K_r. (Use continuity of the map $\Sigma \to S^1 \times S^1$ to replace the constant expansiveness of σ.) Now let $r \to 0$.

Now there is a natural topology on Λ as the suspension of Σ by h.

PROPOSITION 4.3. *Π: $\Lambda \to M$ is continuous.*

Proof. It is enough to see that $pb(e)$ varies continuously with $e \in \Sigma$, and that $\psi_t pb(\gamma) \to pb(\sigma\gamma)$ as $t \to h(\gamma)^-$.

Now $b(\gamma)$ is the unit tangent vector to γ based at the first intersection S of γ with the continuous curve $\partial R(\gamma)$. Moreover $R(\gamma)$ is locally constant as a function of γ except when γ is a side of $\mathcal{N}$. In this last case, the appropriate side of $R(\gamma')$, for γ' close to γ, is one of a finite number of continuous curves all of which pass through S.

By Corollary 4.3, the endpoints P, Q of γ vary continuously with $e \in \Sigma$ and clearly γ varies continuously with P, Q. Hence $b(\gamma)$ is a continuous function of $e \in \Sigma$.

If we lift the path $\psi_t pb(\gamma)$ to $\widetilde{\psi_t pb(\gamma)} \in \tilde{M}$ starting at $b(\gamma)$ when $t=0$, then as $t \to h(\gamma)^-$ the base point of $\widetilde{\psi_t pb(\gamma)}$ approaches the point T where γ crosses from $R(\gamma)$ to $R(\sigma\gamma)$. Therefore $\lim_{t\to h(\gamma)^-} \widetilde{\psi_t b(\gamma)} = \sigma^{-1} b(\sigma\gamma)$. Hence $\psi_t pb(\gamma) \to p(\sigma^{-1} b(\sigma\gamma)) = pb(\sigma\gamma)$ as required.

Remark 4.4. We have not said anything about measures on Λ and M. In [6] we showed there is an ergodic f_Γ-invariant measure $\bar{\mu}$ on S^1, equivalent to Lebesgue measure, finite in the no cusp case and infinite otherwise. $\bar{\mu}$ defines a unique σ-invariant measure μ on Σ which projects to μ, by

$$\mu(Z_{a_{-n}\dots a_n}) = \mu(\varrho(\sigma^{-n}(Z_{a_{-n}\dots a_n}))),$$

where $Z_{a_{-n}\dots a_n} = \{e \in \Sigma : e_r = a_r,\ |r| \leqslant n\}$ and $\varrho : \Sigma \to \Sigma^+$ is projection.

Define a measure ν on Λ by

$$\nu(E) = \int_\Sigma \int_0^{h(e)} \chi_{E_e}(t)\, dt\, d\mu(e)$$

where $E_e = \{(e, t) \in E : 0 \leqslant t < h(e)\}$.

PROPOSITION 4.5. *$\Pi_* \nu$ is the natural flow invariant measure on M.*

Proof. One verifies easily that the measure $|e^{i\theta} - e^{i\varphi}|^{-2} d\theta\, d\varphi$ on $S^1 \times S^1 -$diagonal is invariant under the natural Γ action. Since any geodesic in D is uniquely determined by its endpoints on S^1, we can identify $T_1 D$, the unit tangent bundle to D, with $(S^1 \times S^1 - \text{diag.}) \times \mathbf{R}$. The measure $\lambda = |e^{i\theta} - e^{i\varphi}|^{-2} d\theta\, d\varphi\, dt$ is invariant under Γ acting on the left and the geodesic flow on the right.

Now by Proposition 3.10, any $u \in M$ has a unique lifting $\tilde{u}$ in $T_1 D$ so that the geodesic γ defined by $\tilde{u}$ is admissible and $\tilde{u}$ has its endpoint in $\overline{R(\gamma)}$ (see Proposition 4.1). Let $A \subseteq T_1 D$ be the set of these liftings. It is clear that $\lambda|_A$ (with suitable normalisation) is the natural flow invariant measure on M. Moreover if $q : A \to S^1 \times S^1 -$diag., $q^{-1}(\gamma)$ has length $h(\gamma)$.

Π identifies $q(A) \subseteq S^1 \times S^1 -$diag. with Σ. Therefore to see $\Pi_* \nu = \lambda|_A$, it is enough to see that $w = |e^{i\theta} - e^{i\varphi}|^{-2} d\theta\, d\varphi|_{q(A)}$ and μ on Σ are the same. (We can safely ignore the sets on which Π, π are not bijective since they are null for all relevant measures.)

w is Γ invariant and hence σ invariant on $q(A)$. It is clear that w projects to a measure $\bar{w}$ equivalent to Lebesgue on $\Sigma^+ (= S^1)$, moreover $\bar{w}$ must be shift invariant on Σ^+.

Therefore $\bar{w}$ and $\bar{\mu}$ are shift invariant equivalent measures on Σ^+, and $\bar{\mu}$ is ergodic for the shift. It follows that $\bar{w} = \bar{\mu}$ (if we normalise properly), and since $\bar{w}$ determines w uniquely (just as $\bar{\mu}$ determines μ), we are done.

Notice that $\bar{\mu}$ is the Gibbs measure corresponding to the function $-\log|f'(x)|$ on S^1.

It now follows from the symbolic representation that the geodesic flow is ergodic (since the shift σ on Σ is). In the compact case we can deduce the flow is Bernoulli. One needs to know the flow is K; this is a general fact, see for example [17]. The result follows from Theorem 4.3 of [16], (a K-flow which is the special flow over a shift under a Hölder continuous function is Bernoulli). (One makes an obvious modification to deal with the fact the height function may vanish, since $\exists N$ such that $h(e)+\ldots+h(\sigma^N e)\geqslant c>0$, $\forall e\in\Sigma$.)

We hope to investigate the non-compact case elsewhere. (The flow is known to be Bernoulli in this case also, see [7].)

§ 5. Quasi-conformal deformations

Throughout § 1–§ 4, we assumed that Γ had a fundamental region R which satisfied the property (*). In [6] we showed that if Γ' is any Fuchsian group of the first kind, then there is a group Γ satisfying (*), such that there is a quasi-conformal deformation $j: \Gamma\to\Gamma'$. We now show how to use this deformation to carry over the results above to the general case.

We first summarize the facts we need about quasi-conformal maps. For details, see [4].

(1) There is an isomorphism $j: \Gamma\to\Gamma'$, and a diffeomorphism $\omega^\mu: D\to D'=D$ so that

$$j(g)=\omega^\mu g(\omega^\mu)^{-1},\quad g\in\Gamma.$$

(2) ω^μ restricts to a homeomorphism $h: S^1\to S^1$ so that $h(gx)=j(g)h(x)$, $x\in S^1, g\in\Gamma$. h is the so-called *boundary map* of ω^μ.

(3) If γ is a geodesic in D, then $\gamma'=\omega^\mu(\gamma)$ is a so-called quasi-geodesic in D'. There is a unique geodesic $\bar\gamma$ in D' with the same endpoints as $\omega^\mu(\gamma)$, $\bar\gamma$ is a bounded hyperbolic distance from $\omega^\mu(\gamma)$ (with bound depending only on ω^μ), [13].

Notice that if α, β are geodesics in D then $\alpha\cap\beta\neq\emptyset$ if and only if $\bar\alpha\cap\bar\beta\neq\emptyset$.

Let α be a geodesic in D which is an edge of $\mathcal{N}$, and let v be a vertex of $\mathcal{N}$ on α. Let $\beta_1, \ldots, \beta_r$ be the other edges of $\mathcal{N}$ through v. Then $\bar\alpha\cap\bar\beta_i\neq\emptyset$, $1\leqslant i\leqslant r$, but these intersections may all be distinct points. Let $\alpha(v)=\{\bar\alpha\cap\bar\beta_i\}_{i=1}^r$. Let w be a vertex of $\mathcal{N}$ adjacent to v along α. Then if γ is any other edge of $\mathcal{N}$ through w, $\bar\gamma\cap\bar\beta_i=\emptyset$, $1\leqslant i\leqslant r$, and so we can find disjoint closed intervals $I_\alpha(v)$, $I_\alpha(w)$ on α so that $\alpha(v)\subseteq \operatorname{Int} I_\alpha(v)$, $\alpha(w)\subseteq \operatorname{Int} I_\alpha(w)$. More generally if $\{v_i\}_{i=-\infty}^\infty$ are the vertices of $\mathcal{N}$ along α in order then there are disjoint closed intervals $\{I_\alpha(v_i)\}_{i=-\infty}^\infty$ along $\bar\alpha$ in the same order as $\{v_i\}$, $\alpha(v_i)\subseteq \operatorname{Int} I_\alpha(v_i)$.

Let $Q(v)$ be the open convex hull in D' of the set $\{I_\alpha(v)$: α is an edge of $\mathcal{N}$ through $v\}$.

Now let $t_1, \ldots, t_n$ be the sides of a copy S of R in D. Since non-adjacent sides of S do not meet, the same is true of $\bar{t}_1, \ldots, \bar{t}_n$ and thus $\bar{t}_1, \ldots, \bar{t}_n$ bound a closed polygonal region $\bar{S}$ in D'. Let $Q(S) = \bar{S} - \bigcup \{Q(v)\colon v \text{ is a vertex of } S\}$ and let $Q(D) = D' - \bigcup \{Q(v)\colon v \text{ is a vertex of } \mathcal{N}\}$.

If we collapse each of the regions $Q(v)$ to a point we obtain a net $Q(\mathcal{N})$ whose sides are the portions of the edges $\bar{\alpha}$ outside the regions $Q(v)$ and which is topologically identical with the net $\mathcal{N}$.

Now let $\bar{\gamma}$ be a geodesic in D'. We say $\bar{\gamma}$ passes across $Q(v)$ if $\bar{\gamma} \cap Q(v) \neq \emptyset$. Let the sides of $\mathcal{N}$ meeting at v be $t_1, \ldots, t_{2k}$, going in clockwise order round v. Moving clockwise round $Q(v)$ one cuts successively $\bar{t}_1, \ldots, \bar{t}_{2k}$. Let $\bar{\gamma}$ cut $\partial Q(v)$ in points P, Q. Let $\beta(v)$ be the arc of $\partial Q(v)$ joining P to Q which cuts the smaller number of sides $\bar{t}_i$. (If both arcs cut k or $k+1$ sides choose β to be the arc passing to the left of $Q(v)$.)

Now let $\hat{\gamma}$ be the curve obtained from $\bar{\gamma}$ by replacing $\bar{\gamma}$ with $[\bar{\gamma} - Q(v)] \cup \beta(v)$ in a neighbourhood of $Q(v)$, for every vertex v. In the collapsed net $Q(\mathcal{N})$, $\hat{\gamma}$ becomes a curve $Q(\gamma)$ which passes through a vertex v whenever $\bar{\gamma} \cap \overline{Q(v)} \neq \emptyset$.

THEOREM 5.1. *Let $\bar{\gamma}$ be a geodesic in D' corresponding to an admissible geodesic γ in D. We can find a distinguished region $Q(S(\gamma))$ such that*

(1) $\hat{\gamma} \cap \overline{Q(S(\gamma))} \neq \emptyset$

(2) $\hat{\gamma} \cap \overline{Q(S(\gamma))} \neq \emptyset \Rightarrow S(\gamma) = R$

(3) $\hat{\gamma}$ *cuts in succession* $\overline{Q(S(\gamma))}, \overline{\sigma^{-1}Q(S(\sigma\gamma))}, \ldots$.

Proof. The idea is obviously to imitate § 3. We define what is meant by a curve in $Q(D)$ passing near a vertex of $Q(\mathcal{N})$ just as in § 3. Lemma 3.4 depends only on the topology of $\mathcal{N}$ and the position of the endpoints of γ relative to $\mathcal{N}$; and thus carries over to $Q(\mathcal{N})$ and $\hat{\gamma}$. To prove Lemma 3.5, it is enough to see that $\bar{S}$ is geodesically convex, or equivalently that the interior angles of $\bar{S}$ are less than π. But a vertex of $\bar{S}$ is formed by the intersection of two geodesics with distinct endpoints, and therefore the angle between any adjacent pair of sides is less than π.

The proofs of Lemma 3.6 and Corollary 3.7 are unchanged. Lemma 3.8 and Proposition 3.9 again depend only on topological properties of $\mathcal{N}$ and the position of the endpoints of γ. The rest of the proof is as in § 3.

We shall say a permutation π of $\mathbf{Z}$ 'acts on finite blocks' if there are integers $\ldots < n_1 < n_2 < \ldots$ such that π maps each interval $n_i \leqslant r < n_{i+1}$ onto itself. The importance of this will be that we can keep track of a 'base point' on a sequence, by choosing the left endpoint of some fixed block to be the base point. If we require permutations to preserve a base point, the sequence $n_{\pi^{-1}(1)}, n_{\pi^{-1}(2)}, \ldots$ uniquely determines π.

PROPOSITION 5.2. *Suppose* $\ldots, l_1, l_2, l_3, \ldots$ *are geodesics in* $\mathcal{N}$ *arranged so that* $\hat{\gamma}$ *cuts* $\ldots, \bar{l}_1, \bar{l}_2, \ldots$ *in order* (*with the usual clockwise convention if* $\hat{\gamma}$ *passes through the intersection of two or more* $\bar{l}_i$). *Then* $\bar{\gamma}$ *cuts in order* $\ldots, \bar{l}_{\pi^{-1}(1)}, \bar{l}_{\pi^{-1}(2)}, \ldots$ *where* π *is a permutation of* $\mathbf{Z}$ *which acts on finite blocks.*

Proof. Define an equivalence relation on $\{l_i\}$ by $l_i \sim l_j$ iff l_i, l_j meet at a vertex v of $\mathcal{N}$ and $\hat{\gamma}$ cuts $\bar{l}_i$, $\bar{l}_j$ on $\partial Q(v)$. This is transitive since $\hat{\gamma}$ cuts each $\bar{l}_i$ exactly once and $\partial Q(v) \cap \partial Q(w) = \emptyset$ if $v \neq w$. Notice that the equivalence classes are either singletons or blocks of consecutive sides all associated to the same $Q(v)$. $\bar{\gamma}$ cuts the same sides as $\hat{\gamma}$ in the same order except possibly near $Q(v)$. If $\bar{l}_r, \ldots, \bar{l}_s$ is the block associated to $Q(v)$, then $\bar{\gamma}$ cuts in order $\bar{l}_{\pi^{-1}(r)}, \ldots, \bar{l}_{\pi^{-1}(s)}$ for some permutation π. (This means that if $\pi(1) = i$, where 1 is the base point of the sequence, $\bar{\gamma}$ cuts $\bar{l}_1$ on the ith cut after the base.)

Suppose s is the first side of $Q(S(\gamma))$ cut by $\hat{\gamma}$ and let $\bar{l}_r$ be the geodesic extending s. Define $s(\gamma) = \bar{l}_{\pi^{-1}(r)}$.

THEOREM 5.3. *The geodesic* $\bar{\gamma}$ *cuts the geodesics* $\ldots, s(\gamma), \sigma^{-1}s(\sigma\gamma), \ldots$ *in order.*

Proof. Let $\hat{\gamma}$ cut $\ldots, \bar{l}_1, \bar{l}_2, \ldots$ in order, and let $\sigma\hat{\gamma}$ cut $\ldots, \bar{m}_1, \bar{m}_2, \ldots$. By definition $s(\gamma) = \bar{l}_{\pi(\gamma)^{-1}(r)}$ and $s(\sigma\gamma) = \bar{m}_{\pi(\sigma\gamma)^{-1}(t)}$ where $\bar{l}_r$, $\bar{m}_t$ are the first sides of $Q(S(\gamma))$, $Q(S(\sigma\gamma))$ cut by $\hat{\gamma}$, $\widehat{\sigma\gamma}$ respectively. $\hat{\gamma}$ cuts $Q(S(\gamma))$, $\sigma^{-1}Q(S(\sigma\gamma))$ in order, so $\bar{l}_{r+1}$ is the first side of $\sigma^{-1}Q(S(\sigma\gamma))$ cut by $\hat{\gamma}$. Then $\sigma\bar{l}_{r+1}$ is the first side of $Q(S(\sigma\gamma))$ cut by $(\sigma\hat{\gamma}) = \widehat{\sigma\gamma}$. Therefore $\sigma\bar{l}_{r+1} = \bar{m}_t$. Since $\hat{\gamma}$ cuts $\ldots, \bar{l}_1, \bar{l}_2, \ldots$ in order, $\sigma\hat{\gamma}$ cuts $\ldots, \sigma\bar{l}_1, \sigma\bar{l}_2, \ldots, \bar{m}_j = \sigma\bar{l}_{r+1+j-t}$ for all $j \in \mathbf{Z}$. Then $\sigma\bar{\gamma}$ cuts $\ldots, \sigma\bar{l}_{\pi(\gamma)^{-1}(1)}, \sigma\bar{l}_{\pi(\gamma)^{-1}(2)}, \ldots$ in order where $\sigma\bar{l}_{\pi(\gamma)^{-1}(r+1+j-t)}$ occurs in the jth place. Thus $\bar{m}_{\pi(\sigma\gamma)^{-1}(t)} = \sigma l_{\pi(\gamma)^{-1}(r+1)}$. We have shown $s(\sigma\gamma) = \sigma l_{\pi(\gamma)^{-1}(r+1)}$. Since $\bar{\gamma}$ cuts $l_{\pi(\gamma)^{-1}(r)}, l_{\pi(\gamma)^{-1}(r+1)}$ in order, we are done.

We now want to imitate § 4, to represent the geodesic flow on $\tilde{M}$, the unit tangent bundle to D/Γ, as a special flow on a space Λ.

Let $h(\bar{\gamma})$ be the hyperbolic distance along $\bar{\gamma}$ between $s(\gamma)$ and $\sigma^{-1}s(\sigma\gamma)$. Let $\Lambda = \{(e, t): e \in \Sigma, 0 \leqslant t < h(e)\}$. Let $b(\gamma)$ be the unit tangent vector along $\bar{\gamma}$ at the point where $\bar{\gamma}$ cuts $s(\gamma)$. Define

$$\Pi: \Lambda \to \tilde{M}, \quad \Pi(e, t) = \tilde{\psi}_t p(e).$$

PROPOSITION 5.4. Π *is surjective,* $\Pi\varphi_t = \psi_t\Pi$ *and* $\#\Pi^{-1}(\Pi(e, t)) = \#\pi^{-1}(\pi(e))$ *for* $e \in \Sigma$.

Proof. Since $\bar{\gamma}$ cuts in order $s(\gamma)$, $\sigma^{-1}s(\sigma\gamma)$ the method of Proposition 4.1 shows that $\Pi\varphi_t = \psi_t\Pi$.

Using exactly the same method as in Proposition 3.10 one shows that whenever $\bar{\gamma}$ is a geodesic with $\hat{\gamma} \cap \overline{Q(R)} \neq \emptyset$, there exists a unique $g \in \Gamma$ with $g\bar{\gamma} \in \Sigma$ and $Q(S(g\bar{\gamma})) = gQ(R)$.

Take $u \in M$ and let $\tilde{u}$ be any lifting in $\tilde{M}$, with base point U. Let $\bar{\gamma}$ be the geodesic through U in the direction of $\tilde{u}$, and let $\hat{\gamma}$ be the curve obtained by deforming round $Q(v)$ for each vertex v.

Suppose, as in Proposition 5.2, that $\hat{\gamma}$ cuts geodesics ..., $\bar{l}_1$, $\bar{l}_2$, ... in order. Then $\bar{\gamma}$ cuts $\bar{l}_{\pi^{-1}(1)}$, $\bar{l}_{\pi^{-1}(2)}$, ... at points ..., M_1, M_2, ... say. Suppose $U \in [M_i M_{i+1})$. Let $Q(S)$ be the region between $\bar{l}_i$ and $\bar{l}_{i+1}$ with $\overline{Q(S)} \cap \hat{\gamma} \neq \emptyset$. (It is not hard to see there is a unique such region, because the boundary between $Q(S)$ and $Q(S')$ is either a side $\bar{l}$ of $Q(\mathcal{N})$ or a region $Q(v)$, and there are no sides of $Q(\mathcal{N})$ cutting $\hat{\gamma}$ between $\bar{l}_i$ and $\bar{l}_{i+1}$.) Applying $k \in \Gamma$ with $kS = R$, we may assume $\overline{Q(R)} \cap \hat{\gamma} \neq \emptyset$.

Now we use the analogue of Proposition 3.10 above to find $g \in \Gamma$ with $g\bar{\gamma} \in \Sigma$ and $Q(S(g\bar{\gamma})) = gQ(R)$. The first side of $Q(S(g\bar{\gamma}))$ cut by $g\hat{\gamma}$ is $g\bar{l}_i$. Therefore $s(g\gamma) = \bar{l}_{\pi^{-1}(i)}$. Hence gU lies on $g\bar{\gamma}$ between the intersection with $s(g\gamma)$ and the next side of $\overline{\mathcal{N}}$, so

$$g\hat{u} = \hat{\varphi}_\tau b(g\gamma) \quad \text{where} \quad 0 \leqslant \tau < h(g\gamma).$$

Then $\Pi(g\bar{\gamma}, \tau) = u$, as in Proposition 4.1.

Finally, it is not hard to see that Proposition 4.1 is easily modified to prove $\Pi(e, t) = u$ iff $\pi(e) = g\bar{\gamma}$.

The facts about the continuity of h and Π now follow exactly as in § 4, and we again see that in the compact case the flow is Bernoulli.

References

[1] ADLER, R., *f-expansions revisited.* Springer Lecture Notes 318, (1973).

[2] ANOSOV, D. V., Geodesic Flows on Closed Riemann manifolds with negative curvature. *Proc. Steklov. Inst. Math.*, 90 (1967).

[3] ARTIN, E., Ein Mechanisches System mit quasi-ergodischen Bahnen, *Collected Papers* pp. 499–501. Addison Wesley 1965.

[4] BERS, L., Uniformization, moduli and Kleinian groups. *Bull. London Math. Soc.*, 4 (1972), 257–300.

[5] BOWEN, R. & RUELLE, D., The ergodic theory of Axiom A flows. *Invent. Math.*, 29 (1975), 181–202.

[6] BOWEN, R. & SERIES, C., Markov maps for Fuchsian groups. *Inst. Hautes Études Sci. Publ. Math.*, 50 (1979).

[7] DANI, S., Dynamical systems on homogeneous spaces. *Bull. Amer. Math. Soc.*, 82 (1976), 950–952.

[8] DENKER, M., GRILLENBERGER, C. & SIGMUND, K., *Ergodic Theory on compact spaces.* Springer Lecture Notes 527, (1976).

[9] FORD, L. R., *Automorphic Functions.* McGraw Hill, New York, 1929.

[10] HEDLUND, G. A., A metrically transitive group defined by the modular group. *Amer. J. Math.*, 57 (1935), 668–678.

[11] —— On the metrical transitivity of the geodesics on closed surfaces of constant negative curvature. *Ann. of Math.*, 35 (1934), 787–808.

[12] HOPF, E., Ergodentheorie. *Abh. Sächs. Akad. Wiss. Leipzig*, 91 (1939), 261.

[13] MORSE, M., *Symbolic dynamics.* Institute for Advanced Study Notes, Princeton, (1966), (unpublished).

[14] NIELSEN, J., Untersuchungen zur Topologie der geschlossenen Zweiseitigen Flächen. *Acta Math.*, 50 (1927), 189–358.

[15] ORNSTEIN, D. & WEISS, B., Geodesic flows are Bernoulli. *Israel J. Math.*, 14 (1973), 184–197.

[16] RATNER, M., Anosov Flows with Gibbs measure are also Bernoulli. *Israel J. Math.*, 17 (1974), 380–391.

[17] SINAI, YA., Geodesic flows on manifolds of constant negative curvature. *Dokl. Akad. Nauk SSSR*, 131 (1960), 752–755; *Soviet Math. Dokl.*, 1 (1960), 335–339.

[18] THURSTON, W., *The Geometry and Topology of 3-manifolds*, Proposition 5.9.2. Lecture Notes, Princeton (1978).

Received February 4, 1980