BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 25, Number 2, October 1991

GEODESIC FLOWS, INTERVAL MAPS, AND SYMBOLIC DYNAMICS

ROY ADLER AND LEOPOLD FLATTO

1. Introduction

Geodesic flows and interval maps are two topics in the theory of dynamical systems with a long mathematical history. The first of these seems to have originated with Jacobi who related the flows to the study of Hamiltonian systems (for a detailed description of the connection, see [CFS]). The second arises in diverse settings, such as the modelling of population genetics [Ma] and the frequency count of digits in continued fraction expansions [Bi]. In both subjects the main problem is to describe the distribution of orbits. Thus we wish to know how the geodesics spread over the manifold containing them and how iterates of points under an interval map vary over the interval. Ergodic theory provides answers to these questions, particularly the notions of ergodicity and invariant measure which will be elaborated below.

At first sight the two topics seem unrelated, geodesic flow being a continuous time action and interval map a discrete one. Nevertheless, we shall relate them and their associated symbolic dynamics when the flow takes place on a compact surface of constant negative curvature. In this case we use a graphic approach enabling us to find a series of reductions from geodesic flow to interval map. In relating the topics, we show how each sheds light on the other. We use ergodicity of interval maps to prove ergodicity of flows and conversely. Furthermore, explicit formulas for invariant measures of interval maps can be derived from invariant measures for flows. This fact is interesting as there is a paucity of explicit formulas for invariant measures of interval maps.

The steps in our reduction scheme are known to exist abstractly. However, for the dynamical systems considered here, the reductions are carried out by means of elementary geometry. Our graphic

Received by the editors January 3, 1989 and, in revised form, January 2, 1991.
1991 *Mathematics Subject Classification.* Primary 58F11, 58F17.

0273-0979/91 $1.00 + $.25 per page

approach puts into evidence certain facts not easily discernible otherwise, e.g. the existence of a Markovian partition and of factors for the cross section map which we associate with the flow. An attractive feature of our graphic approach is that it renders concrete, by means of simple geometric examples, many abstract notions of ergodic theory. However, the method seems intrinsically two-dimensional and it is not clear how to extend it to higher dimensions.

We describe in detail the main concepts employed in our work, obtaining at the same time an overview of the paper.

Ergodicity. The ergodic theorem describes the long term average behavior of certain systems evolving in time. The theorem has two versions, a discrete and a continuous one. We first describe the discrete one.

Let X be a measure space—meaning there exists a nonnegative, not identically zero, countably additive measure m assigned to certain subsets of X which are designated as measurable. We assume from now on that all sets under discussion are measurable. A transformation T of X to itself is measurable if for any set E, so is $T^{-1}E = \{x: Tx \in E\}$. If, in addition, $m(T^{-1}E) = m(E)$, then T is measure preserving. A measurable transformation T is ergodic if $T^{-1}E = E$ implies either $m(E) = 0$ or $m(X-E) = 0$. We assume X to be σ-finite, i.e. X is a countable union of sets of finite measure. Let $T^n(x)$ denote the n th iterate of $x \in X$ under T, with $T^o(x) = x$. The sequence $\{T^n x\}$ is called the T-orbit of x.

Ergodic theorem (discrete version). (i) *Let T be a measure preserving transformation of X. Then for any integrable function $f(x)$,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) \quad \textit{exists for almost all } x \in X. \tag{1.1}$$

(ii) *If, in addition, T is ergodic, then the limit in* (1.1) *equals the constant* $(1/m(X)) \int f(x)dm(x)$.

In (1.1) the phrase "for almost all" refers to the invariant measure m (this meaning is adopted throughout the paper).

The ergodic theorem asserts that if T is ergodic, then for almost all x the *time average* $\lim_{n\to\infty}(1/N)\sum_{n=0}^{N-1} f(T^n x)$ equals the *space average* $(1/m(X)) \int f(x)dm(x)$. Ergodicity of T is crucial for this fact. Indeed it is easily shown that (1.1), with right

side equaling $(1/m(X))\int f(x)dm(x)$, implies ergodicity. In the special case when f is the indicator function χ_E of a set E of finite measure and T is ergodic, (1.1) gives the following formula for the frequency of visits of T-iterates to E:

$$\lim_{N\to\infty} \frac{|\{n: T^n x \in E, 0 \le n < N\}|}{N} = \frac{m(E)}{m(X)} \quad \text{for almost all } x \in X$$

where $|\cdot|$ denotes cardinality of a set.

We generalize (1.1) to the continuous case. Let $\{T_t(x)\}$, $0 \le t < \infty$, be a one-parameter family of measure preserving transformations of X which form a semigroup. This means that $T_0(x) = x$ and $T_{s+t}(x) = T_s \circ T_t(x)$ for $s, t \ge 0$, $T_s \circ T_t$ denoting the composition product. $\{T_t\}$ is a measurable flow on X if, for any measurable function $f(x)$ on X, $f(T_t x)$ is measurable on the product space $R \times X$, R denoting the real line. If, in addition, T_t is measure preserving for all t, then $\{T_t\}$ is a measure preserving flow on X. A measurable flow $\{T_t\}$ is ergodic if $T_t^{-1}E = E$ for all t implies either $m(E) = 0$ or $m(X - E) = 0$. For given x, the set $\{T_t(x)\}$, $0 \le t < \infty$, is called the T_t-orbit or flow line through x. Both measurable transformations and flows are referred to as measurable dynamical systems.

Ergodic theorem (continuous version). (i) *Let T_t be a measure preserving flow on X. Then for any integrable function $f(x)$,*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t f(T_t x)du \quad \textit{exists for almost all } x \in X. \tag{1.2}$$

(ii) *If in addition, $\{T_t\}$ is ergodic, then the limit in* (1.2) *equals the constant* $(1/m(X))\int f(x)dm(x)$.

The ergodic theorem asserts that if $\{T_t\}$ is ergodic, then for almost all x, the time average $\lim_{t\to\infty}(1/t)\int_0^t f(T_u x)du$ equals the space average $(1/m(X))\int f(x)dm(x)$. In the special case when $f = \chi_E$, $m(E) < \infty$, and T is ergodic, (1.2) gives the following formula for the proportion of time spent by T_t-orbits in E:

$$\lim_{T\to\infty} \frac{\lambda\{u: T_u(x) \in E, 0 \le u \le t\}}{t} = \frac{m(E)}{m(X)} \quad \text{for almost all } x \in X$$

where λ denotes Lebesgue measure on the real line.

In practice, ergodicity of transformations and flows becomes difficult to check, thus limiting the applications of the ergodic theorem. It is therefore of interest to obtain examples for which ergodicity can be verified. We describe a class of flows for which this is the case.

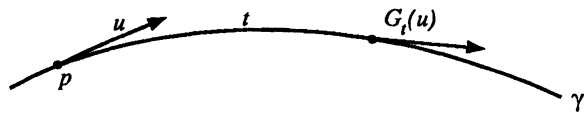


FIGURE 1.1. Geodesic flow.

Geodesic flows. Let S be a compact n-dimensional Riemannian manifold and $\mathbf{S}$ its unit tangent bundle. For a given vector $u \in \mathbf{S}$ with base point $p \in S$, let γ be the unique geodesic passing through p and tangent to u. Parametrize γ as $\gamma(t)$, t being arc length along γ measured from p. Let $G_t(u)$ be the unit tangent to γ at $\gamma(t)$, $-\infty < t < \infty$, as shown in Figure 1.1.

The set of homeomorphisms of $\mathbf{S}$: $u \to G_t(u)$, $-\infty < t < \infty$, is called the geodesic flow on $\mathbf{S}$. For given u, the curve $\tilde{\gamma} = G_t(u)$, $-\infty < t < \infty$, is called a flow line or G_t-orbit in $\mathbf{S}$ (observe that $\tilde{\gamma}$ consists of unit tangents to γ and is to be distinguished from γ). Liouville's Theorem asserts that the geodesic flow preserves the measure m on $\mathbf{S}$ induced on it by the metric of S. We refer to m as the Liouville measure and, for S of constant negative curvature, as the hyperbolic measure.

In the case where S is of negative curvature, Hedlund and Hopf have shown the geodesic flow to be ergodic [H, Ho].

In the sequel we limit ourselves to S 2-dimensional and of constant negative curvature. In this case our graphic approach enables us to reduce the subject of geodesic flows to that of interval maps.

Interval maps. This subject deals with the behavior of the iterates of a map $f(x)$ of an interval I to itself. We may assume I to be the unit interval. We also assume that $f(x)$ is noninvertible and piecewise continuous. The subject centers mainly on two problems (i) proving existence of an f-invariant measure m equivalent to Lebesgue measure λ—equivalence meaning that m and λ have the same sets of measure 0. (ii) proving ergodicity of f.

Observe that since m and λ are equivalent, the phrase "for almost all x" in (1.1) refers to either m or λ. It is a simple consequence of the ergodic theorem that any finite measure m satisfying (i) and (ii) is uniquely determined up to a multiplicative constant.

We give several examples of interval maps. Let (a) $f(x) = (2x)$, where $(\cdot)$ denotes the fractional part; (b) $f(x) = (\beta x)$ where

$\beta = (1+\sqrt{5})/2$; (c) $f(x) = (1/x)$; (d) $f(x) = (1/1-x)$. The graphs of these functions are depicted in Figure 1.2.

The above maps are among those rare cases for which an explicit formula is known for the invariant measure. We have (a) m is Lebesgue measure; (b) $dm = \beta\, dx$, $0 \le x < \beta^{-1}$, and $dm = dx$, $\beta^{-1} < x \le 1$; (c) $dm = dx/(1+x)$; (d) $dm = dx/x$.

Formulas (a), (b) are easily guessed. (c) is a famous formula due to Gauss and (d) is due to Renyi [R]. The measures (a)–(c) are finite and (d) is infinite. (c) is called the *continued fraction map* because of its relation to continued fractions (for the relation and some interesting consequences about continued fractions, consult [Bi]). The graph of (d) is obtained from that of (c) by reflecting about the line $x = \frac{1}{2}$. Hence we call it the *backward continued fraction map.*

In this paper we deal exclusively with Markovian interval maps defined below. For this case we prove in Appendix B a theorem of obscure origin, which we refer to as the *folklore theorem*, giving conditions guaranteeing that (i), (ii) hold and that m be finite.

Let X be a one dimensional space—i.e. an interval or a circle—and $\{I_i\}$ a finite partition of X into subintervals. Let $f: X \to X$

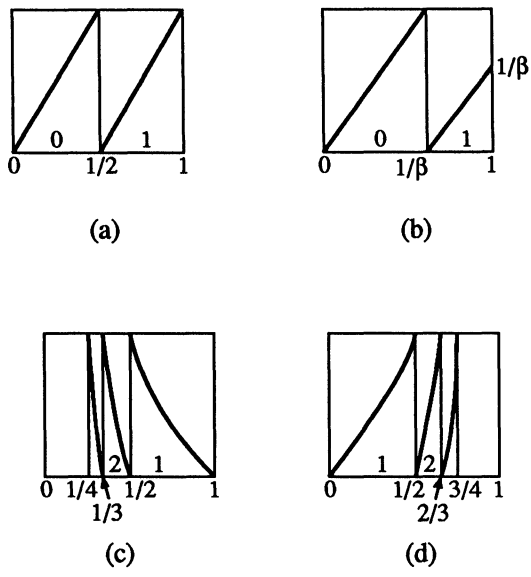


FIGURE 1.2. Interval maps.

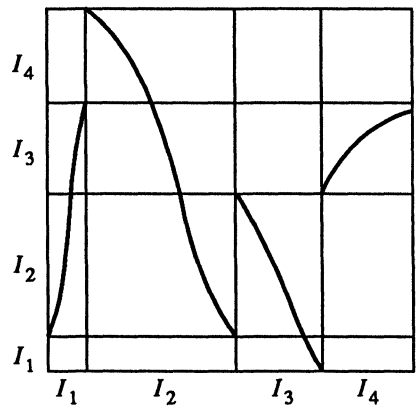


FIGURE 1.3. Markovian interval map.

satisfy:

1. piecewise smoothness—i.e. f/I_i, the restriction of f to I_i, has a C^2 extension to the closure $\overline{I}_i$ of I_i.
2. local invertibility—i.e. f/I_i is strictly monotone.
3. Markov property—i.e. $f(\overline{I}_i) =$ union of some $\overline{I}_j$'s.
4. Aperiodicity—i.e. there exists an integer p such that $f^p(\overline{I}_i) = \overline{X}$ for all i.

Figure 1.3 gives a typical example of a map satisfying (1)–(4). In this case $p = 4$.

If (1)–(3) hold, then $\{I_i\}$ is called a Markovian partition for f (or f is a Markovian map for $\{I_i\}$). Condition (4) is added to guarantee the

Folklore theorem. *Assume that* (1)–(4) *hold and that* f *is eventually expansive—i.e. for some iterate* f^n, $|df^n/dx| \geq \theta > 1$ *for all* x. *Then* f *has a finite Lebesgue-equivalent measure* m *and is ergodic. Furthermore* $dm = \rho(x)dx$, *where* $\rho(x)$ *is piecewise continuous and* $1/D < \rho < D$ *for some* $D > 0$.

Under conditions (1)–(4), the converse of the folklore theorem also holds.

For examples (a)–(d) we choose the partition $\{I_i\}$ as indicated in Figure 1.2, where I_i is labeled by i.. It is readily checked that examples (a),(b) satisfy the conditions of the folklore theorem. We conclude in these cases that f is ergodic. The folklore theorem does not cover examples (c), (d) as in these cases $\{I_i\}$ is an infinite partition. There also exists a folklore theorem for eventually

expansive Markovian maps when $\{I_i\}$ is infinite, but more conditions are required [MVP, Bo1]. These conditions cover example (c). Note that for (d), $f(x)$ is not eventually expansive because 0 is a fixed point where the derivative of the map and all its iterates is 1. This explains why the invariant measure for (d) does not satisfy the conclusions of the theorem.

Reduction. For the geodesic flows considered in this paper, the motion is continuous in time and takes place in a three dimensional manifold—the unit tangent bundle $\mathbf{S}$. On the other hand, for interval maps the motion is discrete in time—the times $1, 2, \ldots$ refer to the iterates $f(x), f^2(x), \ldots$—and is taking place in a 1-dimensional manifold—the unit interval. The connection between geodesic flow and interval map is achieved by two successive reductions, each lowering the dimension by one and reducing the motion from continuous to discrete.

The first, attributed to Poincaré, reduces the study of the flow to that of a *cross section map.* Roughly speaking, a cross section is a subset C of $\mathbf{S}$ meeting all flow lines again and again, both past and future. The measurable subsets of C are obtained by intersecting the measurable subsets of $\mathbf{S}$ with C. The correspondence between successive points of return serves to define the map T_C of C (Figure 1.4). T_C is called a cross section map for G_t. Conversely G_t is called a flow over T_C. As $\mathbf{S}$ is three-dimensional, C will be two-dimensional. The cross section map T_C preserves a measure m_C on C inherited from the hyperbolic measure m on $\mathbf{S}$ preserved by G_t. It is defined by:

$$m_C(B) = \lim_{h \to 0} \frac{1}{h} m\{G_t u : u \in B, 0 \leq t \leq h\} \tag{1.3}$$

for any measurable $B \subset C$. The set $\{G_t u : u \in B, 0 \leq t \leq h\}$ is the small cylinder depicted in Figure 1.4 on p. 236. $m_C(B)$ is the rate at which m flows through B. Ergodicity of G_t can be reduced to that of T_C.

There are many possible choices for C, but we choose one for which a second reduction is possible, namely T_C is to have a one-dimensional factor map. This means that there exists a map π, called a *projection*, from C onto the unit interval I, and a map f from I onto I, called the *factor map*, such that

$$f \circ \pi = \pi \circ T_C \tag{1.4}$$

i.e. we have the commuting diagram shown in Figure 1.5 on p. 236.

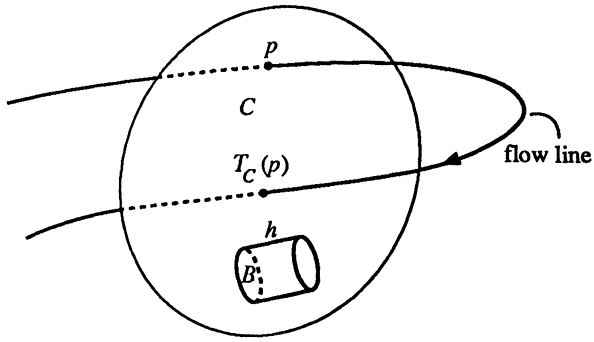


FIGURE 1.4. Cross section of geodesic flow.

$$\begin{array}{ccc} C & \xrightarrow{T_C} & C \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ I & \xrightarrow[f]{} & I \end{array}$$

FIGURE 1.5 The factor map f.

(1.4) may be worded as follows. Call the sets $\pi^{-1}x$, $x \in I$, fibers of C. T_C maps the fiber $\pi^{-1}x$ into the fiber $\pi^{-1}(f(x))$. The space of fibers is identified with I, the mappings between fibers given by $x \to f(x)$. In general f will not be invertible even though T_C is. The factor map preserves a measure m_f on I inherited from the invariant measure m_C. It is defined by

$$(1.5) \qquad m_f(E) = m_C(\pi^{-1}E) \quad \text{for any measurable } E \subset I.$$

T_C is called an extension of f. Ergodicity of T_C can be reduced to that of f.

The following example serves to illustrate the notion of factor map. Let $T(x, y)$ be the one-to-one map of the unit square $S = I \times I$ onto itself defined by $T(x, y) = ((2x), ([2x] + y)/2)$, where $[x]$, (x) denote the greatest integer less than x and the fractional part of x. T can be described geometrically. Contract the vertical lines of S by $1/2$ and magnify the horizontal ones by 2, obtaining the rectangle R: $0 \le x < 2$, $0 \le y < 1/2$. Then translate the right half of R to the top half of S. Since these actions are reminiscent of the kneading of dough, T is called the *baker transformation.* T is depicted in Figure 1.6 where 1,2 refer

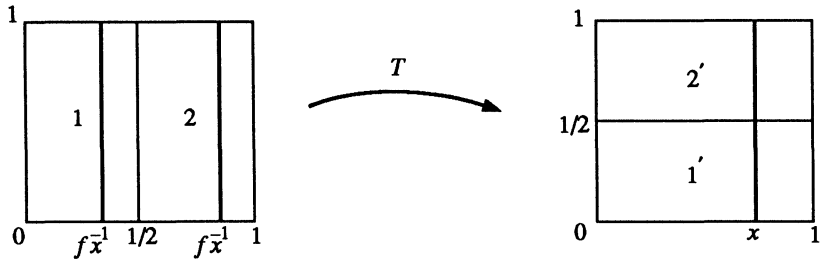


FIGURE 1.6. The Baker transformation.

to the left and right halves of S. T maps 1,2 respectively to $1'$, $2'$, the bottom and top halves of S. $T(x, y) = (2x), \cdot)$ so that it has the factor map $f(x) = (2x)$ depicted in Figure 1.2(a) (the projection map π of (1.4) is given in this case by $\pi(x, y) = x$).

The fibers are the vertical lines of S, the preimage of the vertical line over x consisting of the two verticals over the two points $f^{-1}(x)$. The inverse map T^{-1} is given by $T^{-1}(x, y) = ((x + [2y])/2), (2y))$. Hence $T^{-1}(x, y)$ also has a factor map given by $g(y) = (2y)$, the fibers now being the horizontal lines of S. T contracts vertical directions and magnifies horizontal ones and preserves the 2-dimensional Lebesgue measure $dx\ dy$.

We remark that in [AF1, 3] we showed that both the continued fraction and backward continued fraction maps arise as factors of cross section maps for the geodesic flow on the modular surface. Using this we obtain another derivation of the formulas for the invariant measures of these maps.

Cross section. With the above reductions in mind, we have found it suitable to make a special choice for C. Its description relies on some geometrical facts about S. S is assumed to be a compact surface of constant negative curvature. It is known that such a surface is orientable of genus $g \geq 2$, hence realizable as a sphere with g handles. On S there exists a system of $2g$ simple closed geodesics $\gamma_1, \ldots, \gamma_{2g}$ as exhibited in Figure 1.7 on p. 238 (we illustrate here and later with the typically sufficient case $g = 2$).

Speaking loosely $\gamma_1, \gamma_3, \ldots, \gamma_{2g-1}$ go around the "holes" and $\gamma_2, \ldots, \gamma_{2g}$ around the "waists" of S. Distinct curves γ_i, γ_j intersect if and only if $|i - j| = 1$, in which case they intersect in precisely one point. The existence of such a system of curves is intuitive. Less obvious and not so well known is that the γ_i's can be chosen as geodesics. This fact is a consequence of the

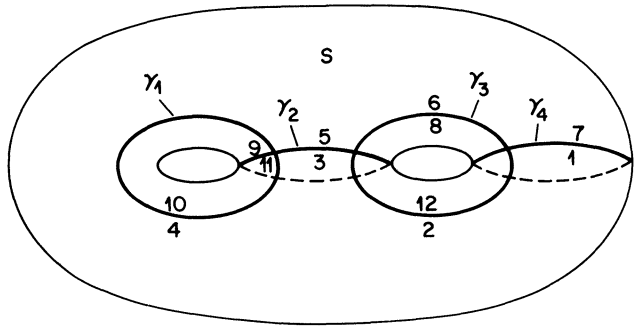


FIGURE 1.7. A system of closed geodesics on S.

curvature assumption. A proof does not seem readily accessible and so we provide one in Appendix A. The system of curves can also be shown unique; we shall not prove this, as no use of it is made in the sequel.

Let B be the union of $\gamma_1, \ldots, \gamma_{2g}$ and $\mathbf{B}$ the set of unit tangents to B. $\mathbf{B}$ consists of a finite number of closed flow lines. Hence $\mathbf{B}$ is a set of measure 0 invariant under G_t, $-\infty < t < \infty$. $\mathbf{S} - \mathbf{B}$ is also invariant under G_t. Hence the study of G_t on $\mathbf{S}$ is equivalent to the study of G_t on $\mathbf{S} - \mathbf{B}$. Choose C to be the set of unit tangents in $\mathbf{S} - \mathbf{B}$ having their base point in B. As S is compact, it is intuitive that C is a cross section for the geodesic flow on $\mathbf{S} - \mathbf{B}$.

Fundamental polygon. The factor map of T_C is obtained by coordinatizing C in a suitable manner. Our coordinatization depends on another description of S and $\mathbf{S}$. In the ensuing discussion we rely on notions of the hyperbolic plane and its associated discrete groups. These will be discussed in detail in §2.

Let $\mathbb{D}$ be the unit disk $|z| < 1$ endowed with the metric $ds = 2|dz|/(1 - |z|^2)$. $\mathbb{D}$ is the hyperbolic plane of constant negative curvature -1. The orientation preserving isometries of $\mathbb{D}$ onto itself are the Möbius transformations $\tau(z) = e^{i\beta}(z + \alpha)/(1 + \bar{\alpha}z)$, $0 \leq \beta < 2\pi$, $|\alpha| < 1$. The geodesics of $\mathbb{D}$ are the circular arcs orthogonal to $|z| = 1$. Any compact surface S of constant negative curvature -1 can be thought of as an orbit space $\mathbb{D}/\Gamma$ consisting of the Γ-orbits Γz, $z \in \mathbb{D}$, where Γ is a discrete group[1] of sense

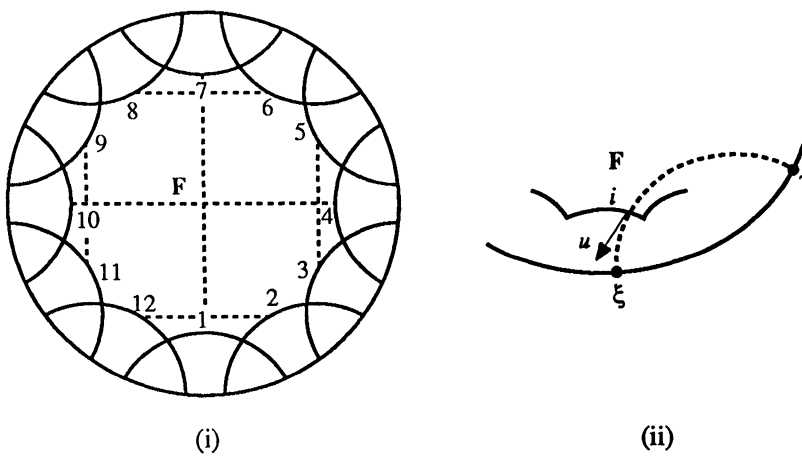
[1]There are discrete groups Γ acting freely on $\mathbb{D}$ for which $\mathbb{D}/\Gamma$ is not compact. If $\mathbb{D}/\Gamma$ is compact then Γ is called a surface group. We deal exclusively with this case.

preserving isometries of $\mathbb{D}$ acting freely on $\mathbb{D}$—i.e. the elements of Γ have no fixed points in $\mathbb{D}$. $\mathbb{D}$ is called the universal covering space of S. Γ can be identified as the fundamental group of S. Similarly the unit tangent bundle $\mathbf{S}$ can be thought to be the set of Γ-orbits Γu, $u \in \mathbb{U}$, where U is the unit tangent bundle of $\mathbb{D}$. Let $\pi(z) = \Gamma(z)$, $z \in \mathbb{D}$, and $\overline{\pi}(u) = \Gamma(u)$, $u \in \mathbb{U}$, be the respective projection maps from $\mathbb{D}$ onto S and from $\mathbb{U}$ onto $\mathbf{S}$. The existence of the system $\{\gamma_i\}$ has the following meaning for Γ.

Theorem 1.1. *Let $S = \mathbb{D}/\Gamma$ be a compact surface of genus g. There exists an $(8g-4)$ sided fundamental polygon $\mathbf{F}$ whose sides are geodesic segments. These satisfy the extension condition: that is the geodesic extensions of these segments never pierce the interior of the fundamental regions $\tau\mathbf{F}$, $\tau \in \Gamma$.*

$\mathbf{F}$ is depicted in Figure 1.8(i) where the sides are labeled successively by $1, \ldots, 8g-4$. The dotted lines indicate the pairing of sides, side i mapped into its mate by a unique group element $T_i \in \Gamma$. In general $\mathbf{F}$ need not be regular. We choose one which is as it can then be constructed by ruler and compass for $g = 2$. The projection map π takes $\mathbf{F}$ onto S and the boundary $\partial\mathbf{F}$ to $B = \bigcup \gamma_i$.

The extension condition of Theorem 1.1 is a reflection of the fact that the γ_i's are closed geodesics. As shown in §5, this condition plays a crucial role in establishing the existence of the rectilinear map T_R discussed below.

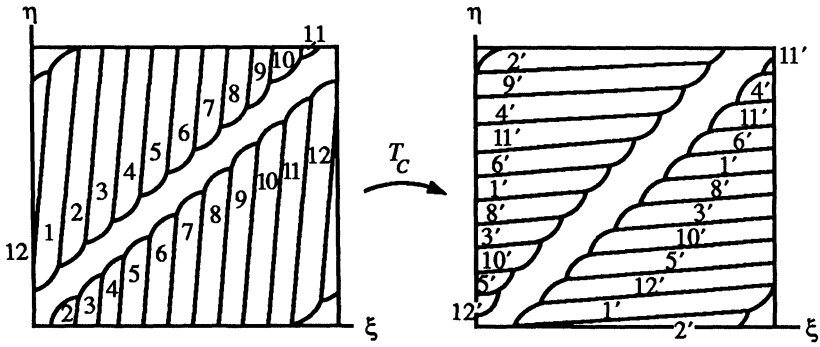


(i) (ii)

FIGURE 1.8. Fundamental polygon.

The cross section C can be thought to be the set of elements $\overline{\pi}(u)$, u varying over the unit tangent vectors in $\mathbb{U}$ with base point in $\partial\mathbf{F}$ and pointing out of $\mathbf{F}$. C partitions into $C_1, \ldots, C_{8g-4}$, C_i consisting of the elements $\overline{\pi}(u)$ with base point in side i. We coordinatize C by assigning (ξ, η) to $\overline{\pi}(u)$, where ξ and η are respectively the head and tail of the geodesic determined by u, as in Figure 1.8(ii). (ξ, η) is a point in the 2-dimensional torus $\partial\mathbb{D} \times \partial\mathbb{D}$. Identifying ξ and η with their arguments, we think of these variables as going from 0 to 2π. The *almost everywhere*[2] description of C and T_C in these coordinates is given in Figure 1.9, where C_i and $C_i' = T_C(C_i)$ are respectively labeled by i and i'. It seems difficult, if not impossible, to prove the ergodicity of T_C from its formula. To remedy the situation, we define another set R, also described in (ξ, η) coordinates, together with an invertible map T_R. The almost everywhere description of R and T_R is given in Figure 1.10 (here too, there is a more comprehensive description given in §5). R is a straightened out version of C. For this reason we call C and R respectively the curvilinear and rectilinear domains, and T_C and T_R the curvilinear and rectilinear maps.

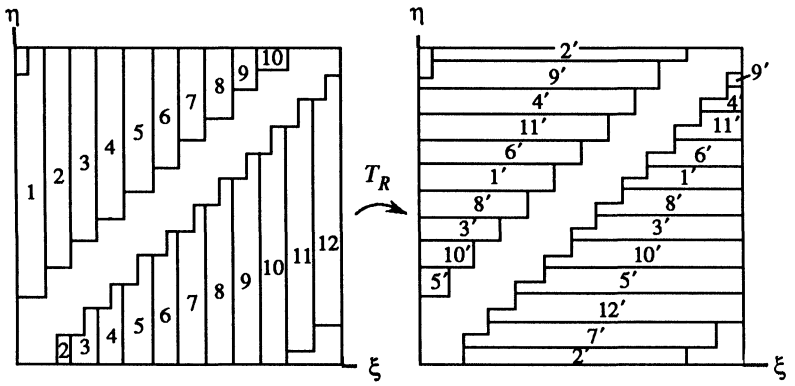
T_C and T_R are conjugate—meaning there exists a map $\Phi: C \to R$ such that $T_C = \Phi \circ T_R \circ \Phi^{-1}$—hence ergodicity of T_C is equivalent to that of T_R. The advantage of T_R over T_C is that we



$$T_C(\xi, \eta) = (T_i\xi, T_i\eta), (\xi, \eta) \in C_i$$

FIGURE 1.9. Curvilinear map.

[2]The description of T_C given by Figure 1.9 becomes ambiguous at the boundaries of the C_i's. To circumvent the ambiguity, we remove all points whose T_C-orbit hits the boundaries, thus obtaining the almost everywhere description of C and T_C. A more comprehensive description including boundaries is given in §4.



$$T_R(\xi, \eta) = (T_i\xi, T_i\eta), \qquad (\xi, \eta) \in R_i$$

FIGURE 1.10. Rectilinear map.

have the formulas

$$T_R(\xi, \eta) = (f(\xi), \cdot), \qquad T_R^{-1}(\xi, \eta) = (\cdot, g(\eta))$$

putting into evidence the factor maps $f(\xi)$, $g(\eta)$. f and g are maps of the unit circumference $|\xi| = 1$ to itself. We call these the Bowen-Series maps after Bowen and Series who studied them [BoS]. In §6 we show that these maps satisfy the hypothesis of the folklore theorem. Hence they are ergodic, which in turn implies the ergodicity of T_R and G_t, as shown in §7.

Symbolic sequences. Further insight into the cross section map T_C can be obtained via symbolic dynamics. We explain this concept. A pair (X, T) consisting of a space X and a transformation T from X to itself is called an *abstract dynamical system.* The special case (Σ, σ), where Σ is a space of sequences $\{s_n\}$ of symbols chosen from a finite alphabet $\{1, \ldots, N\}$ and Σ closed under the shift $\sigma\{s_n\} = \{s_{n+1}\}$, is called a *symbolic dynamical system.* We consider both two-sided sequences, i.e. $-\infty < n < \infty$, and one-sided ones, i.e. $0 \leq n < \infty$. We shall code abstract dynamical systems to symbolic ones, meaning that to each point $x \in X$, we assign a symbolic sequence $\phi(x) = \{s_n(x)\}$, the assignment ϕ being one-to-one and respecting the rule $\phi \circ T = \sigma \circ \phi$. The coding reduces the dynamics of T to the more transparent dynamics of the shift. For instance, the existence of periodic points p of T—i.e. $T^k p = p$ for some $k \neq 0$—reduces to the existence of sequences $\{s_n\}$ with the periodicity $s_{n+k} = s_n$. Similarly, the existence of an everywhere dense T-orbit reduces to the existence of a sequence containing every finite admissible block infinitely

often. The coding is done as follows. Let $\mathscr{P} = \{X_1, \ldots, X_N\}$ be a finite partition of X into disjoint sets X_i. For $x \in X$, define $\phi(x) = \{s_n(x)\}$ by $T^n x \in X_{s_n}$, where $-\infty < n < \infty$ for T invertible and $0 \leq n < \infty$ for T noninvertible. $\{s_n(x)\}$ is called the orbit history of x through the partition $\mathscr{P}$. The coding procedure is illustrated in Figure 1.11. Since $\{s_n(Tx)\} = \{s_{n+1}(x)\}$ we have $\phi \circ T = \sigma \circ \phi$. If ϕ is injective, i.e. distinct points have distinct orbit histories, then (X, T) is coded into (Σ, σ) where Σ is the image of X under ϕ. We call $\{s_n(x)\}$ the T-expansion of x. In the sequel we write f instead of T when the latter is noninvertible, in which case $\{s_n(x)\}$ is called an f-expansion.

Symbolic coding becomes a useful tool in the study of abstract dynamical systems only in case the resulting symbolic system is capable of a simple description. In this paper we concentrate on two such systems defined as follows.

Definition 1.1. (i) Let $\mathscr{A} = \{1, \ldots, N\}$ be a finite alphabet and $\mathscr{B}$ a set of pairs of elements (i, j) of $\mathscr{A}$. The space Σ of sequences $\{s_n\}$, $s_n \in \mathscr{A}$ and $(s_n, s_{n+1}) \in \mathscr{B}$ for all n, is called the *Markov* system defined by $\mathscr{A}$ and $\mathscr{B}$. The pairs $(i, j) \in \mathscr{B}$ are called the admissibility rules and we denote these by $i \to j$. (ii) Let Σ be a Markov system defined as in (i). Let $\phi(i)$, $1 \leq i \leq N$, be a function on $\mathscr{A}$. The space Σ' of sequences $\{s'_n\}$, $s'_n = \phi(s_n)$, is called a *sofic* system.

Thus sofic systems are obtained from Markov ones by amalgamation of the alphabet. We discuss both of these systems in further detail in Appendix C.

To code abstract dynamical systems to Markov systems, we require the notions of *Markov* partitions and *Markov* maps. We

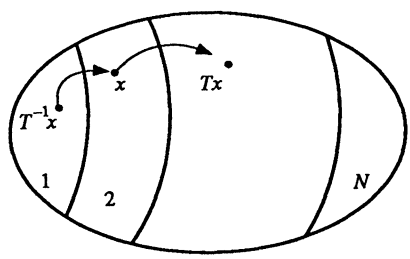


FIGURE 1.11. Obtaining symbolic sequences from partitions.

need two kinds of Markov maps: noninvertible 1-dimensional and invertible 2-dimensional ones. The 1-dimensional partitions and maps were introduced earlier in the discussion of noninvertible interval maps. We defer the definition of invertible 2-dimensional Markov partitions and maps to Appendix C, where these concepts are discussed in greater generality, and just mention here that the invertible 2-dimensional maps are typified by the baker transformation and rectilinear map. Thus 2-dimensional Markov partitions are finite unions of rectangles, each consisting of horizontal and vertical fibers. The horizontal fibers are mapped one-to-one onto a union of horizontal fibers of various rectangles, a similar statement holding for the inverse map applied to vertical fibers.

Similarly sofic systems can be derived from *sofic* partitions, which are obtained by amalgamating Markov partitions.

Cutting sequences. To code (C, T_C) into a symbolic system, we choose the earlier mentioned partition $\{C_1, \ldots, C_{8g-4}\}$. In Figure 1.7, label the sides of $B = \bigcup \gamma_i$, to conform with the labeling of sides of the fundamental polygon $\mathbf{F}$. C_i is the set of unit tangent vectors $\tilde{u}$ of $\mathbf{S}$ with base point in B and pointing away from side i. The sequences $\{s_n(\tilde{u})\}$ corresponding to the partition $\{C_1, \ldots, C_{8g-4}\}$ are called *curvilinear*. They have a geometric interpretation. Let $\gamma(\tilde{u})$ be the geodesic in S determined by $\tilde{u}$. Then $\{s_n(\tilde{u})\}$ is the sequence of labels encountered by γ as it cuts successively the system of curves B (to obtain $\{s_n(\tilde{u})\}$ we always choose the first of the two labels encountered by γ as it cuts B). For this reason curvilinear sequences are called *cutting sequences*. Observe that for γ passing through an intersection point $\gamma_r \cap \gamma_s$, the definition of cutting sequence becomes ambiguous. Assigning a cutting sequence in this case requires a special convention given in §9.

We shall find it is useful to interpret cutting sequences on the universal covering surface $\mathbb{D}$. Let $\mathbf{N}$ be the union of sets $\tau(\partial\mathbf{F})$, $\tau \in \Gamma$ (we call these Γ-translates of $\partial\mathbf{F}$). Label the sides of edges in $\mathbf{N}$ in Γ-invariant fashion—i.e. corresponding sides of $\mathbf{F}$ and $\tau\mathbf{F}$ obtain the same label. Let $u = (\xi, \eta)$ be the unit tangent vector with base point in $\partial\mathbf{F}$ and pointing out of $\mathbf{F}$, and $\gamma(\xi, \eta)$ the geodesic with end points ξ, η. The curvilinear sequence $\{s_n(\xi, \eta)\}$ is identical with the sequence of labels encountered by γ as it leaves successive Γ-translates of $\mathbf{F}$ (Figure 1.12 on p. 244).

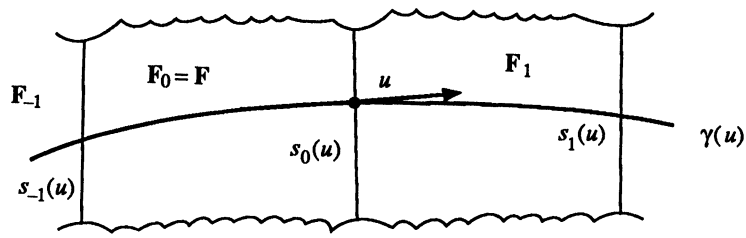


FIGURE 1.12. Cutting sequences.

Unfortunately there is no known useful description of the totality of cutting sequences, and there is good reason to believe that none exists by virtue of the comments made in §10. The difficulty disappears for the rectilinear map T_R. We refer to the partition $\{R_1, \ldots, R_{8g-4}\}$ of R, given in Figure 1.10, as the *coarse* partition. Splitting each R_i into a left and right rectangle, we obtain the *fine* partition. The sequences $\{s_n(\xi, \eta)\}$, $(\xi, \eta) \in R$, corresponding to either partition are called *rectilinear*. The fine partition is Markov. In this case, the rectilinear sequences form a Markov system, from which we remove certain sequences specified in §8 (this removal is due to an inherent difficulty in symbolic representations of dynamical systems explained in Appendix C, part IIa). The coarse partition is sofic. For this case, the rectilinear sequences form a sofic system, from which we must again remove certain sequences specified in §8. We also obtain sequences corresponding to the factor maps f and g, and show that the T_R-expansion of (ξ, η) is obtained by splicing the f-expansion of ξ to the g-expansion of η (Artin [A] used splicing to analyze geodesics on the modular surface).

The sofic rectilinear sequence $\{s_n(\bar{\xi}, \bar{\eta})\}$, where $(\bar{\xi}, \bar{\eta}) = \Phi(\xi, \eta)$ and $T_C = \Phi \circ T_R \circ \Phi^{-1}$, also has a geometric interpretation due to Morse [Mo]. It is the cutting sequence of a curve $\bar{\gamma}(\bar{\xi}, \bar{\eta})$ obtained by modifying $\gamma(\xi, \eta)$; we call it the *modified cutting sequence.* The modification is somewhat complicated and we defer its description to §9 (see in particular Figure 9.5).

We conclude with some historical background, a few further comments on our work, and a summary of the various sections.

Historical background. The earlier works on symbolic dynamics for geodesic flows deal with specific surfaces of constant negative curvature. Artin [A] codes the geodesics on the modular surface by the continued fraction expansions of the end points, deducing from this symbolism the existence of periodic and everywhere

dense geodesics. The works of Hedlund and Morse [H, Mo], relating geodesic flows to symbolic sequences, deal exclusively with surfaces of genus $g \geq 2$ defined by a regular $4g$-sided fundamental region with interior angles $\pi/2g$. Hedlund, using an idea of Nielsen, represents orbits by a Markovian system; while Morse achieves a characterization of orbits by sofic systems. Morse's method was extended to general compact surfaces by Birman and Series [BS]. It is unclear whether Hedlund's method can be likewise extended.

The reductions discussed in our paper have previously been given general formulation by various authors. Ambrose and Kakutani [AK] prove a general representation theorem for measurable flows as one built over a function and over a transformation. Their theorem is about the existence of an abstract cross section and cross section map. The concepts of factor map and natural extension (see §7) were studied by Rohlin [Ro]. The existence of a symbolism with Markovian rules was established for geodesic flows on negative curvature surfaces by Ornstein and Weiss [OW]. Thus the steps in our reduction scheme are known to exist abstractly from these works. However in our present work the reduction is made explicit and has a simple geometric interpretation.

In a series of papers Series [S1-3] also studies the problem of representing geodesic flows by symbolic systems. Comparing her approach with ours one might say that we have reached more or less the same mathematical conclusions from opposite directions. Generally speaking, her theorems become our definitions and vice versa. She first identifies the natural extension Σ of the f-expansions for the Bowen-Series map as the space of modified cutting sequences (the natural extension Σ is defined as the space of two-sided sequences obeying the same admissibility rules as the f-expansions). She then establishes that the geodesic flow can be represented as a flow over (Σ, σ), and proves that (Σ, σ) is conjugate to (Σ', σ'), where Σ' is the set of cutting sequences. In our treatment all this naturally proceeds in the opposite direction. In addition our graphic approach puts into evidence a Markovian partition leading to Markovian systems rather than to merely sofic ones, and to formulas for the invariant measure of the Bowen-Series maps.

Comments. To reiterate, the basic ideas of our paper are: we graph the curvilinear cross section map, straighten out boundaries in this

graph to obtain the nicer rectilinear map, and establish a conjugacy between the maps. In this last act we encounter a discovery which may be considered the cornerstone of our work—namely "bulges fit into corners" (see Theorem 5.1 and Figures 5.3, 5.6). All other ideas in the paper follow naturally from this basic observation.

Our treatment of geodesic flow is based on simple geometric ideas, but unfortunately, these carry along with them a quantity of forbidding detail. We are confronted with the problem of putting geometrical ideas into words ("one picture is worth a thousand words provided one uses another thousand words to justify the picture" [St, p. 225]). Another difficulty stems from the fact that we describe all orbits in the flow. Removal of a set of flow lines of measure zero—namely the flow lines of $\mathbf{S}$ passing through the intersection points of the curves in $B = \bigcup \gamma_i$—would have simplified the analysis considerably just as in [AF2]; but then we would have failed to account for all orbits.

We remark that the concepts of this paper—i.e. the introduction of the rectilinear map and its conjugacy to the curvilinear one—can be applied to any fundamental region satisfying the extension condition. We have limited our discussion to the $(8g - 4)$-sided regions of Theorem 1.1 for two reasons:

1. Universality—such a region exists for all surface groups Γ.
2. Simplicity—in this case the rectilinear domain has the simplest possible description, traceable to the fact that the number of fundamental regions meeting at a vertex equals 4, which is the smallest possible number when the extension property holds.

Summary. We summarize the contents of the various sections. Our work relies on various results which seem all but impossible to reference. To assist the reader, we include several appendices providing the proofs of these results.

In §2 we introduce some background in hyperbolic geometry with notation necessary for our discussion of geodesic flows.

In §3 we introduce the $(8g - 4)$-sided fundamental polygon $\mathbf{F}$. We use $\mathbf{F}$ to read off a set of generators for Γ and a set of group relations these satisfy. The relations are used to establish conjugacy between the curvilinear and rectilinear maps introduced in §4.5.

In §4 we define the cross section C and its associated cross section map T_C. C is coordinatized. The coordinate description of T_C, also designated by T_C, is called the curvilinear map. The

dynamics of T_C is difficult to understand. To remedy the situation we introduce in §5 a second map T_R, called rectilinear, and show that T_C and T_R are conjugate.

In §6 we show that the Bowen-Series maps—the factors of T_R and T_R^{-1}—are ergodic. We also derive formulas for the invariant measures of these maps. Then in §7 we show how the ergodicity of the Bowen-Series maps imply the ergodicity of the flow G_t.

In §8 we obtain the admissibility rules for the rectilinear sequences. The symbolic sequences for T_R are obtained by splicing the f-expansions of the factors of T_R and T_R^{-1}. The splicing assumes somewhat different forms depending as to whether we use Markovian or sofic sequences. In §9 we obtain geometric interpretations of both curvilinear and rectilinear sequences. These interpretations lead to rules for coding curvilinear sequences to rectilinear ones.

In §10 we discuss the main unresolved questions suggested by our work.

In Appendix A we prove the existence of the $(8g-4)$-sided polygon. The proof is based on two theorems which we leave unproved but for which we provide findable references: namely, a theorem of Poincaré giving conditions for a finite sided polygon to be a fundamental region for a discrete group acting on the hyperbolic plane, and the theorem of Fenchel-Nielsen on extending homeomorphisms between universal covering spaces to their boundaries. For another approach, more direct but somewhat lengthier, see Stillwell's appendix in [D, pp. 379–386].

In Appendix B we prove the folklore theorem and its converse. The converse seems to be a new addition to the subject, and we make use of it to prove ergodicity of Bowen-Series maps.

In Appendix C, we discuss symbolic systems. We give several definitions of Markovian and sofic systems, illustrating these with examples. We then develop the theory of Markov partitions in a manner more basic and elementary than found in the literature.

Finally, in Appendix D, we present an alternate proof to Theorem 9.7 of §9, which provides further insight to the interplay between geometry and the combinatorics of symbolism.

Acknowledgments. We thank Jeff Lagarias and Henry Landau for a critical reading of the introduction and for suggested improvements.

2. Geodesic flow

We begin with a brief review of two-dimensional hyperbolic geometry leading to the definition of geodesic flow. We then show how this notion carries over to surfaces defined by the hyperbolic plane modulo a discrete group (= Fuchsian group).

The hyperbolic plane may be represented by either of the following two models: the upper-half plane $\mathbb{H} = \{z: \mathcal{I}mz > 0\}$ endowed with the metric $ds = |dz|/(\mathcal{I}mz)$, or the unit disk $\mathbb{D} = \{z: |z| < 1\}$ endowed with the metric $ds = 2|dz|/(1 - |z|^2)$. The equivalence of the two models is exhibited by the map $w = (z - i)/(z + i)$ which is an isometry of $\mathbb{H}$ onto $\mathbb{D}$. The resulting surface is the hyperbolic plane of constant negative curvature -1.

In [AF 1–3] we used the model $\mathbb{H}$, but in the present paper it is more convenient to work with $\mathbb{D}$. The orientation preserving isometries of $\mathbb{D}$ onto itself are the Möbius transformations $\tau: z \to e^{i\beta}(z + \alpha)/(\bar{\alpha}z + 1)$, $0 \leq \beta < 2\pi$, $|\alpha| < 1$. We refer to such transformations simply as *motions*. The geodesics in $\mathbb{D}$ are arcs of circles orthogonal to the boundary $\partial\mathbb{D} = \{z: |z| = 1\}$. The unit tangent bundle $\mathbb{U}$ consists of unit tangent vectors u on $\mathbb{D}$ which we coordinatize by $u = u(x, y, \theta)$, where (x, y) is the base point of u and θ the angle measured in the counter-clockwise direction between the positive x-axis and u (see Figure 2.1).

We call the portion of the geodesic γ, tangent to $u \in \mathbb{U}$, beginning at its base point and going to the boundary of $\mathbb{D}$ in the direction of u, the *forward geodesic ray at* u, and denote it by $\gamma^+(u)$. We call $\gamma - \gamma^+(u)$ the *backward geodesic ray at* u and denote it by $\gamma^-(u)$.

A motion τ induces the following map on $\mathbb{U}$: $\bar{\tau}: u(z, \theta) \to u(\tau z, \theta + \arg \tau'(z))$, where we have written $u(z, \theta) = u(x + iy, \theta)$ for $u(x, y, \theta)$. The hyperbolic measures on $\mathbb{D}$ and $\mathbb{U}$ are defined respectively by $dA = 4dx\,dy/(1 - |z|^2)^2$ and $dAd\theta$. Since τ is an isometry on $\mathbb{D}$, it preserves the measure dA. Hence, the map $\bar{\tau}$, being a skew product of a transformation preserving dA and a family of translations preserving $d\theta$, must itself preserve the measure $dAd\theta$.

There is another coordinate system on $\mathbb{U}$ which will prove to be more convenient than (x, y, θ) in the ensuing discussion of geodesic flows: namely, for each $u \in \mathbb{U}$ we assign the triple (ξ, η, s) where ξ, η are complex numbers of modulus one and s is real. The pair ξ, η designate points of intersection of the

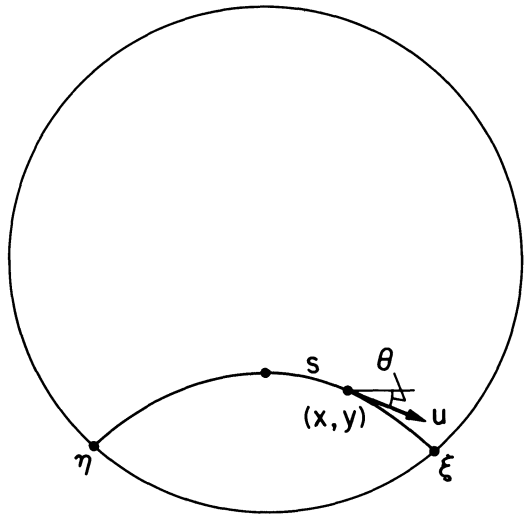


FIGURE 2.1. Hyperbolic geodesic and unit tangent vector.

geodesic, determined by u, with $\partial\mathbb{D} = \{z: |z| = 1\}$. ξ is the forward end and η the backward end as shown in Figure 2.1. The third coordinate s is the hyperbolic length parameter along the geodesic measured from its midpoint to the base point of u. If we let $\xi' = \tau\xi$ and $\eta' = \tau\eta$, where τ is a Möbius transformation, then the relation

$$\frac{d\xi' d\eta'}{(\xi' - \eta')^2} = \frac{d\xi d\eta}{(\xi - \eta)^2} \tag{2.1}$$

follows from differentiating $\tau(z) = (az + b)/(cz + d)$ (for some a, b, c, d, satisfying $ad - bc \neq 0$) and elementary algebra.

In the (ξ, η, s)-coordinate system the map $\bar{\tau}$, induced by a motion τ, becomes

$$\bar{\tau}(\xi, \eta, s) = (\tau\xi, \tau\eta, s - f(\xi, \eta)) \tag{2.2}$$

where $f(\xi, \eta)$ corrects for the fact that midpoints of geodesics are not preserved under τ. The previous remarks about skew products apply; and so from (2.1) follows that

$$dm = \frac{|d\xi|\,|d\eta| ds}{|\xi - \eta|^2} \tag{2.3}$$

is another measure preserved by $\bar{\tau}$.

Because the group of motions acts transitively on $\mathbb{U}$ and the transformation defining the change of variables from (x, y, θ)

to (ξ, η, s) is nonsingular, the two invariant measures dm and $dA d\theta$ are related by

$$dm = c\, dA\, d\theta \tag{2.4}$$

for some $c > 0$ (from a Jacobian computation which we omit one can get $c = 1/4$, a fact which is not used in this paper).

The *geodesic flow* G_t, $-\infty < t < \infty$, consists of the homeomorphisms of $\mathbb{U}$ defined by $u \to u_t$, where u and u_t are unit tangent vectors to the initial and terminal points of a geodesic segment of hyperbolic length t (see Figure 2.2).

In terms of the (ξ, η, s) coordinate system the geodesic flow has a particularly simple description: namely,

$$G_t\colon (\xi, \eta, s) \to (\xi, \eta, s + t). \tag{2.5}$$

From (2.2) and (2.5), it is clear that G_t and $\bar{\tau}$ commute, and from (2.5) that G_t preserves the measure dm.

Let Γ be a discrete subgroup of motions acting freely on $\mathbb{D}$, i.e. Γ has no elliptic elements (which is equivalent to stating that elements of Γ have no fixed points in $\mathbb{D}$). Γ acts both on $\mathbb{D}$ and $\mathbb{U}$. Let $S = \mathbb{D}/\Gamma = \{\Gamma z\colon z \in \mathbb{D}\}$ denote the set of all Γ-orbits on $\mathbb{D}$ and $\mathbf{S} = \mathbb{U}/\Gamma = \{\Gamma u\colon u \in \mathbb{U}\}$ the Γ-orbits on $\mathbb{U}$. The topologies on S and $\mathbf{S}$ are the smallest ones that render continuous the projection maps $\pi(z) = \Gamma z$ from $\mathbb{D}$ to S, and $\bar{\pi}(u) = \Gamma u$ from $\mathbb{U}$ to $\mathbf{S}$. Since Γ has no elliptic elements, π and $\bar{\pi}$ are locally

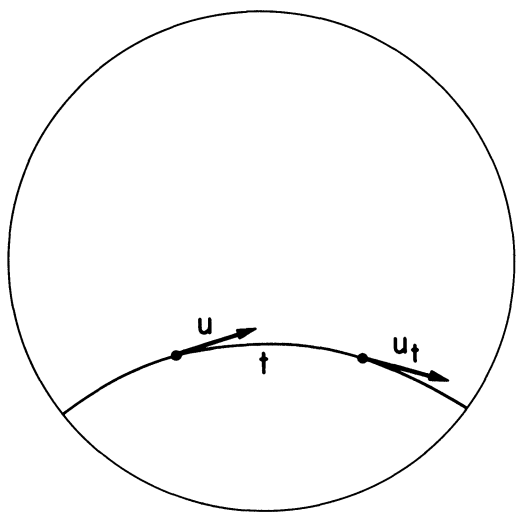


FIGURE 2.2. Geodesic flow.

one-to-one; and projections by π and $\overline{\pi}$ of neighborhoods of $\mathbb{D}$ and $\mathbb{U}$ form a basis of neighborhoods for S and $\mathbf{S}$. Locally, π and $\overline{\pi}$ are homeomorphisms, and (ξ, η), (ξ, η, s) provide local coordinates respectively for S and $\mathbf{S}$. Because the elements of Γ preserve the hyperbolic metric, formulas for ds and dm carry over to S and $\mathbf{S}$. We remark that S has constant negative curvature -1 (in fact, it is a classical result from differential geometry that all complete connected two-dimensional Riemannian manifolds of constant negative curvature -1 can be realized in this manner). We shall assume that S is compact, which implies that S is an orientable surface of genus $g \geq 2$ (see [Sp, chapter 9]).

$\mathbf{S}$ inherits the geodesic flow defined by $\overline{G}_t = \overline{\pi} \circ G_t \circ \overline{\pi}^{-1}$. (The fact that $\overline{G}_t$ is well defined—i.e. independent of the choice of $\overline{\pi}^{-1}$—is an immediate consequence of the commutativity of G_t and Γ.) $\overline{G}_t$ has an invariant measure that coincides with the m-measure of any local inverse of $\overline{\pi}$. If we call this measure $\overline{m}$, then

$$d\overline{m} = \frac{|d\xi|\,|d\eta|ds}{|\xi - \eta|^2}. \tag{2.6}$$

We conclude this section with a collection of elementary facts concerning geodesics used in this paper. We state these as a theorem and omit the proof.

Theorem 2.1. (i) *Let γ_1, γ_2 be distinct geodesics. Then γ_1, γ_2 cannot intersect twice in $\overline{\mathbb{D}} = \mathbb{D} \cup \partial\mathbb{D}$.*

(ii) *Let $\gamma_1 = \gamma_1(t)$, $\gamma_2 = \gamma_2(t)$ be two geodesics parametrized by t, and assume that the end points $\xi_1 = \lim_{t\to+\infty} \gamma_1(t)$, $\xi_2 = \lim_{t\to+\infty} \gamma_2(t)$ are distinct. Then $\lim_{t\to+\infty} d(\xi_1(t), \xi_2(t)) = +\infty$, where $d(\cdot, \cdot)$ denotes hyperbolic distance.*

(iii) *Let pqr be a triangle whose sides are geodesic segments $(p, q, r \in \overline{\mathbb{D}})$. Then pqr has finite hyperbolic area.*

3. Fundamental regions

Let Γ be a Fuchsian group acting on $\mathbb{D}$. A closed subset $\mathbf{F}$ of $\mathbb{D}$ with interior $\mathbf{F}^o$ is called a fundamental region of Γ if and only if

(i) $\tau_1\mathbf{F}^o \cap \tau_2\mathbf{F}^o = \varnothing$ for $\tau_1 \neq \tau_2$, $\tau_1, \tau_2 \in \Gamma$.

(ii) $\bigcup_{\tau\in\Gamma} \tau\mathbf{F} = \mathbb{D}$.

(iii) $\lambda(\mathbf{F} - \mathbf{F}^o) = 0$, λ denoting two-dimensional Lebesgue measure.

A fundamental region is known to exist for any Fuchsian group Γ [B]. In this paper, we concern ourselves with surface groups—i.e. Γ acts freely on $\mathbb{D}$ and $S = \mathbb{D}/\Gamma$ is compact. For such groups there exists a fundamental region with special properties described in Theorem 3.1, the proof of which is given in Appendix A.

Theorem 3.1. *Let $S = \mathbb{D}/\Gamma$ be a compact surface of genus $g \geq 2$. There exists a bounded fundamental polygon $\mathbf{F}$ whose boundary $\partial\mathbf{F}$ consists of $(8g-4)$ geodesic segments. Let $s_1, \ldots, s_{8g-4}$ be the consecutive edges of $\partial\mathbf{F}$ with counterclockwise orientation, and $\sigma(i)$ a permutation of order 2 of $1, \ldots, 8g-4$ defined by*

$$\sigma(i) = \begin{cases} 4g - i \bmod(8g-4), & i \text{ odd}, \\ 2 - i \bmod(8g-4), & i \text{ even}. \end{cases} \tag{3.1}$$

Let s_i^{-1} be the same edge as s_i but with the reverse orientation. Then

(i) *for each s_i there exists a unique element $T_i \in \Gamma$ such that $T_i(s_i) = s_{\sigma(i)}^{-1}$;*

(ii) *$T_i(s_{i-1})$, $T_i(s_{i+1})$ are contained respectively in the geodesics determined by $s_{\sigma(i)+1}$, $s_{\sigma(i)-1}$.*

By virtue of (i), $\sigma(i)$ is called a *pairing*. The content of the theorem is illustrated in Figure 3.1 where s_i is labeled by i. We have chosen here a regular polygon—(ii) then forces $\mathbf{F}$ to have interior right angles—as $\mathbf{F}$ can then be constructed by ruler and compass[3] when $g = 2$. The odd labeled edges are paired off vertically and the even ones horizontally. We define s_i for all integers i by letting $s_i = s_j$ whenever $i \equiv j \bmod(8g-4)$. Let p_i be the intersec-

[3]The ruler and compass construction of a regular $n = 8g-4$ sided hyperbolic polygon with interior right angles involves the construction of a regular Euclidean polygon which provides n equally spaced rays emanating from center 0 of $\mathbb{D}$. One draws on each ray a circle orthogonal to $\partial\mathbb{D}$ with radius y and center at a distance x from 0. In order that neighboring circles intersect at right angles we have from plane geometry the formulas $x = \sqrt{\sec\frac{2\pi}{n}}$ and $y = \sqrt{(\sec\frac{2\pi}{n}) - 1}$. It is well known [V, p. 186] that there is a ruler and compass construction for x, y, and angle $2\pi/n$ if and only if $n = 2^s p_1 \cdots p_m$ where the p's are distinct Fermat primes—i.e. primes of the form $p = 2^k + 1$. Thus, for a ruler and compass construction of a regular $n = 8g-4$ sided fundamental polygon, we require that $2g-1$ is a product of distinct Fermat primes. Since 3, 5, 17, and 257 are Fermat primes, regular $8g-4$ sided fundamental regions can be constructed by ruler and compass for $g = 2, 3, 8, 9, 26, 43, 128, 129$, *etc.*

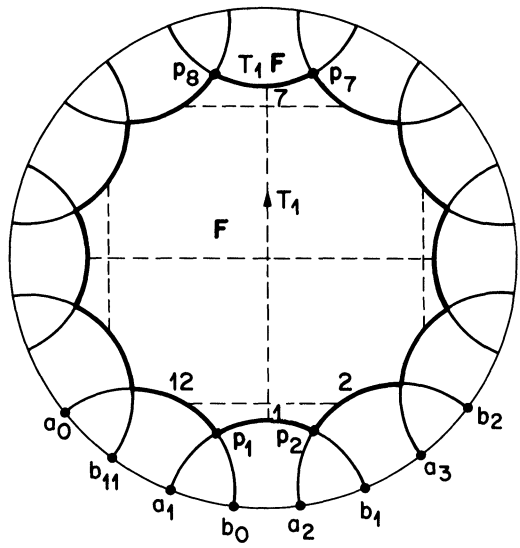


FIGURE 3.1. 12-sided fundamental polygon with pairing.

tion of s_{i-1} and s_i. Then $T_i(p_i) = p_{\sigma(i)+1}$ and $T_i(p_{i+1}) = p_{\sigma(i)}$. Since T_i is angle preserving, (ii) is equivalent to stating that the interior angles of $\mathbf{F}$ at p_i and $p_{\sigma(i)+1}$ are supplementary. It is easily shown that condition (ii) is also equivalent to demanding that the geodesics extending the edges of $\partial\mathbf{F}$ be completely in the *net* $\mathbf{N}$, which is defined as the union of all Γ-translates of $\partial\mathbf{F}$. We shall therefore refer to (ii) as the *extension condition.* As explained later in §5, the extension condition is crucial to our analysis of the cross section map associated with the geodesic flow $\overline{G}_t$.

The remaining theorems of this section hold whenever there exists a fundamental region satisfying the conditions of Theorem 3.1.

Theorem 3.2. *The T_i's satisfy the following relations*: (i) $T_{\sigma(i)}T_i = 1$ *and* (ii) $T_{\vartheta^3(i)}T_{\vartheta^2(i)}T_{\vartheta(i)}T_i = 1$, *where* $\vartheta(i) = \sigma(i) + 1$.

We remark that more is true. Namely, $T_1, \ldots, T_{8g-4}$ generate Γ and (i), (ii) are a complete set of relations for Γ [M]. We do not prove this here as the statement of Theorem 3.2 suffices for our purposes.

Proof. (i) Since $T_i(s_i) = s_{\sigma(i)}^{-1}$, $T_{\sigma(i)}(s_{\sigma(i)}^{-1}) = s_i$, we have $T_{\sigma(i)}T_i(s_i) = s_i$. A motion which fixes an oriented geodesic segment must be the identity. Hence $T_{\sigma(i)}T_i = 1$.

(ii) We label both sides of each edge in $\mathbf{N}$ as follows. Let s_i be an edge of $\partial\mathbf{F}$, so that τs_i is an edge of $\partial(\tau\mathbf{F})$, $\tau \in \Gamma$. The side of τs_i inner to $\tau\mathbf{F}$ receives the label i. It is readily shown that $\tau T_i^{-1}\mathbf{F}$ is the fundamental polygon on the other side of τs_i and that the label on that side is $\sigma(i)$. We conclude that the four fundamental regions meeting at p_i are given as in Figure 3.2.

Passing from $T_i^{-1}T_{\vartheta(i)}^{-1}T_{\vartheta^2(i)}^{-1}\mathbf{F}$ into $\mathbf{F}$ we obtain

$$T_i^{-1}T_{\vartheta(i)}^{-1}T_{\vartheta^2(i)}^{-1}T_{\vartheta^3(i)}^{-1}\mathbf{F} = \mathbf{F}. \tag{3.2}$$

Hence the composition of the above four transformations is the identity, which implies (ii).

Let $\bar{s}_i$ be the geodesic containing s_i and with the same orientation. Let $\bar{s}_i$ meet $\partial\mathbb{D}$ at a_i in the backward direction and b_i in the forward direction. We define a_i, b_i for all i by requiring $a_i = a_j$, $b_i = b_j$ wherever $i \equiv j \mod (8g-4)$.

Theorem 3.3. *The points a_i, b_i $1 \le i \le 8g-4$, are all distinct and are encountered along $\partial\mathbb{D}$ in the counterclockwise direction in the order $a_1, b_0, a_2, b_1, \ldots, a_{8g-4}, b_{8g-5}$.*

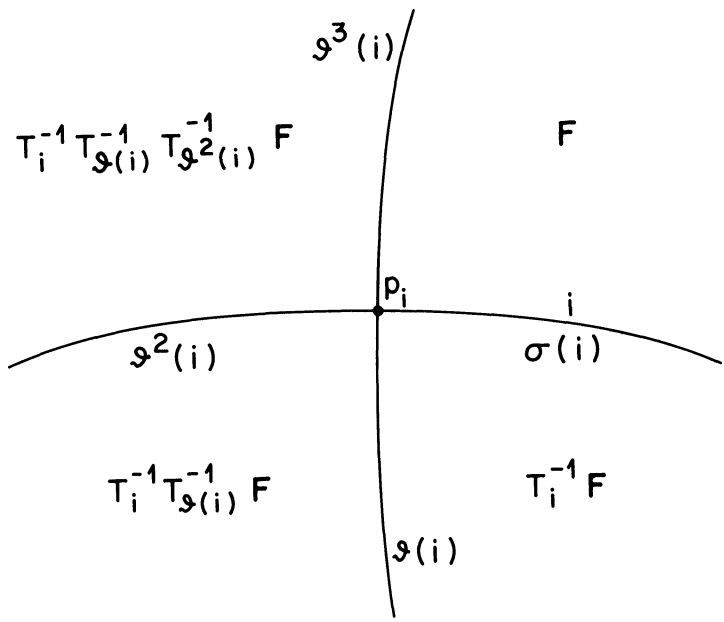


FIGURE 3.2. Fundamental regions at a vertex.

The content of Theorem 3.3 is illustrated in Figure 3.1. Another way of stating the theorem is that: if s_i, s_j are distinct nonconsecutive sides of $\mathbf{F}$, then $\overline{s}_i$, $\overline{s}_j$ don't intersect in $\overline{\mathbb{D}}$. Even though a proof of Theorem 3.3 is given in [BoS], we give one here for the sake of completeness. We use the following

Lemma 3.1. $\mathbf{F}$ *is convex.*

Proof. The interior angles at p_i and $p_{\sigma(i)+1}$ are supplementary. Thus all the interior angles are less than π, which implies that $\mathbf{F}$ is convex [B, p. 154 Theorem 7.16.1].

Proof of Theorem 3.3. Suppose that e and e' are nonconsecutive edges of $\mathbf{F}$ contained respectively in the geodesic $\overline{e}$ and $\overline{e}'$ meeting at p (which may possibly lie in $\partial\mathbb{D}$). Let q, q' be the respective end points of e and e' closest to p, and $e_1, \ldots, e_n$ the consecutive intermediate edges joining q to q' (Figure 3.3 (i)).

We show that, without loss of generality, we may assume $n = 1$. For suppose $n > 1$. Since $\mathbf{F}$ is convex, the open geodesic segments qp, $q'p$ lie outside $\mathbf{F}$. Hence the curve consisting of the segments $e_1, \ldots, e_n$, qp, $q'p$ is simple and bounds a region R. Let r be the end point of e_n distinct from q', and e_n^+ the part of $\overline{e}_n - e_n$ starting at r. e_n^+ enters R and leaves it at a point t. Since e_n^+ lies outside $\mathbf{F}$ and does not intersect $\overline{e}'$ (otherwise $\overline{e}_n$, $\overline{e}'$ would intersect twice), we conclude that t lies in the open segment qp. That is, $\overline{e}$, $\overline{e}_n$ intersect and $e_1, \ldots, e_{n-1}$ are the consecutive

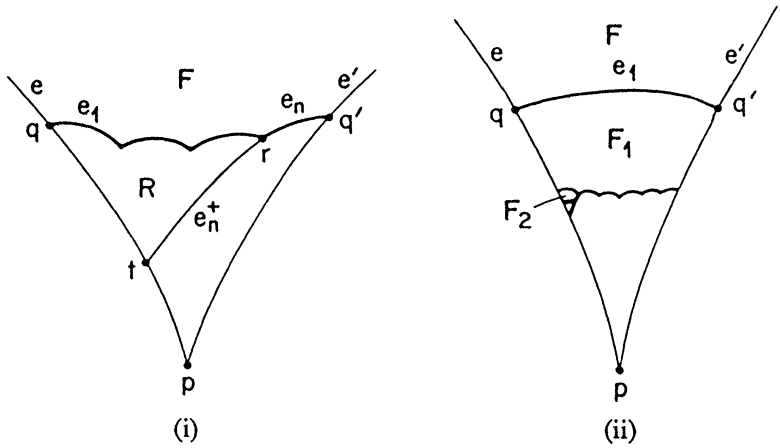


FIGURE 3.3. If edges of $\mathbf{F}$ met.

edges between e and e_n. Repetition of the argument eventually leads to the case $n = 1$.

Let e and e' be then separated by e_1 as in Figure 3.3 (ii). Let $\mathbf{F}_1$ be the fundamental region adjacent to $\mathbf{F}$ on the other side of e_1. $\mathbf{F}_1$ is inside the triangle $qq'p$. Repeating the argument of the first paragraph, we obtain a fundamental region $\mathbf{F}_2$ adjacent to $\mathbf{F}_1$, inside $qq'p$, which has a pair of nonconsecutive sides whose extensions meet within $qq'p$. Continuing in this manner, we obtain an infinite number of distinct fundamental regions inside $qq'p$, which is impossible since $qq'p$ has finite area (Theorem 2.1 (iii)).

We conclude this section with Theorem 3.4, which will be used later on.

Theorem 3.4. *T_i maps the points a_{i-1}, a_i, b_{i-1}, a_{i+1}, b_i, b_{i+1} respectively to $a_{\sigma(i)+1}$, $b_{\sigma(i)}$, $b_{\sigma(i)+1}$, $a_{\sigma(i)-1}$ $a_{\sigma(i)}$, $b_{\sigma(i)-1}$.*

Proof. The above points are illustrated in Figure 3.4. Theorem 3.4 follows from the fact that T_i maps circles to circles. For instance,

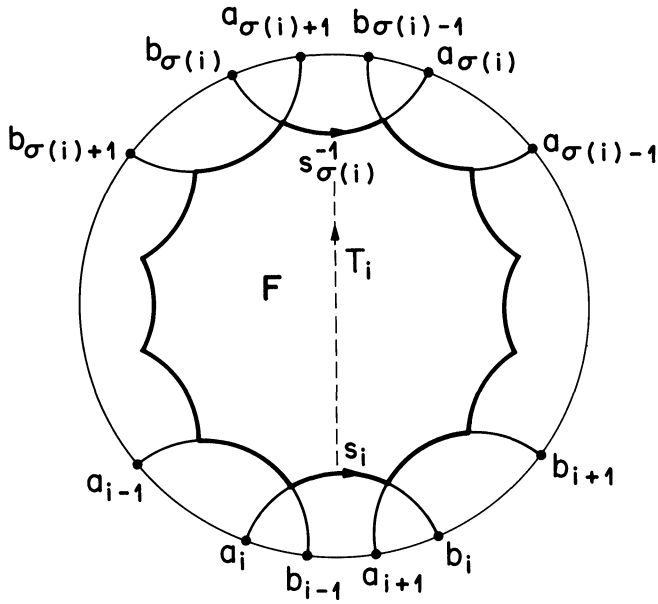


FIGURE 3.4. The points of Theorem 3.4.

T_i maps $\bar{s}_i$ to $\overline{s_{\sigma(i)}^{-1}}$. Hence T_i maps b_i, the forward end point of $\bar{s}_i$, to $a_{\sigma(i)}$, the forward end point of $\overline{s_{\sigma(i)}^{-1}}$. Similar reasoning applies to the remaining points.

4. Cross section

To study the geodesic flow $\overline{G}_t$ on $\mathbf{S}$ we introduce the notion of a cross section map. Roughly speaking, a cross section is a subset of $\mathbf{S}$ which every $\overline{G}_t$-orbit hits again and again, both in the future and in the past. The correspondence between successive points of return to the cross section serves to define the cross section map.

Let $\mathbf{B}$ consist of the $\overline{\pi}$-projections of vectors tangent to $\partial\mathbf{F}$. $\mathbf{B}$ is the union of $2g$ closed $\overline{G}_t$-orbits. $\mathbf{B}$ and $\mathbf{S}_0 = \mathbf{S} - \mathbf{B}$ are both $\overline{G}_t$-invariant, so that the study of the flow on $\mathbf{S}$ reduces to the study of the flow on $\mathbf{S}_0$. As a preliminary choice for our cross section, we choose $\mathbf{C}^p$ to consist of the $\overline{\pi}$-projections of vectors with base point in $\partial\mathbf{F}$ and pointing into the interior of one of the fundamental regions bordering on $\mathbf{F}$. Referring to Figure 1.7, $\mathbf{B}$ is the set of vectors tangent to the system of curves $\{\gamma_i\}$ and $\mathbf{C}^p$ is the set of vectors not tangent to but with base point in that system. Because of the pairing of edges of $\mathbf{F}$, every element in $\mathbf{C}^p$ is also a $\overline{\pi}$ projection of a vector based on $\partial\mathbf{F}$ and pointing into $\mathbf{F}$. The geodesic determined by the latter eventually leaves $\mathbf{F}$ and thus it is clear that $\mathbf{C}^p$ is a cross section.

In order to obtain a simple coordinate description of the cross section map, we add to $\mathbf{C}^p$ the seemingly artificial sets $\mathbf{C}_{i1}$ described below. The enlarged cross section $\mathbf{C}$ consists of the sets $\mathbf{C}_i^p$, $\mathbf{C}_{ij}$, $1 \le i \le 8g-4$ and $j = 0, 1$, defined as follows:

$\mathbf{C}_i^p$ is the set of elements $v = \overline{\pi}(u)$, u based at an interior point of the edge s_i and pointing to the exterior of $\mathbf{F}$.

$\mathbf{C}_{i0}$ is the set of elements $v = \overline{\pi}(u)$, u based at the vertex p_i and pointing to the interior of $T_i^{-1}T_{\vartheta(i)}^{-1}\mathbf{F}$, the fundamental region opposite to $\mathbf{F}$ at p_i (see Figure 4.1 (ii)).

To define $\mathbf{C}_{i1}$ we choose $\varepsilon > 0$ so that the closed disks of radius ε centered respectively at $p_1, \ldots, p_{8g-4}$ are disjoint. Let γ be a geodesic starting at p_i and going into $T_i^{-1}\mathbf{F}$. Let u be the unit tangent vector to γ at the point which is at distance ε from p_i.

$\mathbf{C}_{i1}$ is the set of elements $\overline{\pi}(u)$.

In Figures 4.1 (i)–(iii) we depict respectively vectors u for which $\overline{\pi}(u) \in \mathbf{C}_i^p$, $\mathbf{C}_{i0}$, $\mathbf{C}_{i1}$. We refer to the vectors u as representatives of $\mathbf{C}$.

The sets $\mathbf{C}_i^p$, $\mathbf{C}_{i0}$, $\mathbf{C}_{i1}$, $1 \leq i \leq 8g-4$ are mutually disjoint. $\mathbf{C}^p$ is the union of the sets $\mathbf{C}_i^p$, $\mathbf{C}_{i0}$, and $\mathbf{C}$ is obtained from $\mathbf{C}^p$ by adding the sets $\mathbf{C}_{i1}$. $\mathbf{C}$ is coordinatized as follows. For $\nu = \overline{\pi}(u)$, $u = u(\xi, \eta, \cdot)$ chosen as in Figure 4.2, assign to ν the coordinates (ξ, η).

We describe the sets $\mathbf{C}_i^p$, $\mathbf{C}_{ij}$ in (ξ, η) coordinates and henceforth dispense with bold type. For $|\alpha| < 1$, let $\eta_\alpha(z)$ be the fractional linear transformation of period two which fixes α and carries $\partial\mathbb{D}$ into itself. For $|\xi| = 1$, $\eta_\alpha(\xi)$ can also be described as the other end point of the geodesic passing through α and ξ. We write $\eta_i(\xi)$ for $\eta_{p_i}(\xi)$, $1 \leq i \leq 8g-4$.

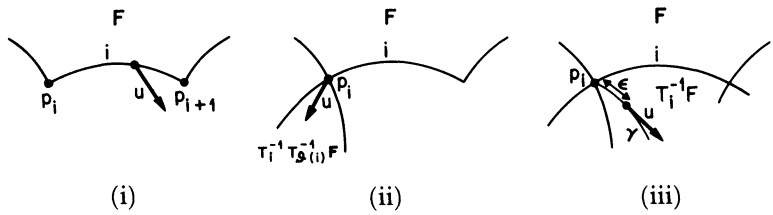


FIGURE 4.1. Representatives of the cross-section.

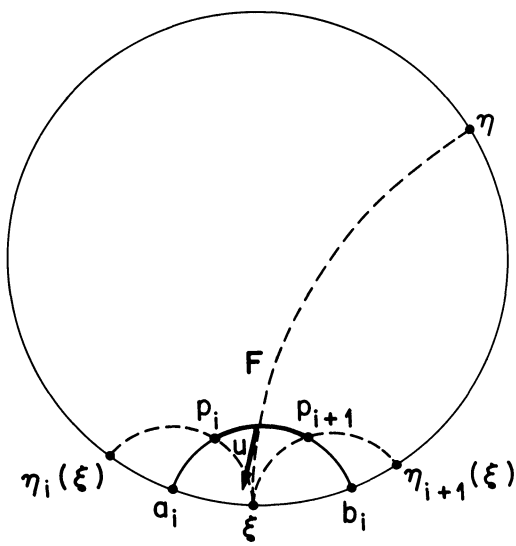


FIGURE 4.2. Coordinatization of cross section representatives.

We introduce the following notation: for a, b, two points on $\partial\mathbb{D}$, let $[a, b]$, (a, b), $[a, b)$, and $(a, b]$ be the closed, open, and half open arcs going from a to b in the counter-clockwise direction. We refer to these arcs as *intervals.*

Theorem 4.1. *In* (ξ, η) *coordinates*

$$
\begin{aligned}
C_i^p &= \{(\xi, \eta): \xi \in (a_i, b_i), \eta \in (\eta_{i+1}(\xi), \eta_i(\xi))\} \\
C_{i0} &= \{(\xi, \eta): \xi \in (a_i, b_{i-1}), \eta = \eta_i(\xi)\} \qquad (4.1) \\
C_{i1} &= \{(\xi, \eta): \xi \in (b_{i-1}, b_i), \eta = \eta_i(\xi)\}.
\end{aligned}
$$

Proof. We refer to Figure 4.2.

We show that the (ξ, η)-coordinate description of C_i^p is given by (4.1). Let $\nu = \overline{\pi}(u)$, $u = (\xi, \eta, \cdot)$ chosen as in Figure 4.2. Then $\xi \in (a_i, b_i)$. Fixing ξ and varying the base point of u over s_i, η varies over the circular arc $(\eta_{i+1}(\xi), \eta_i(\xi))$. The arguments leading to the coordinate descriptions for C_{i0}, C_{i1} are similar and we omit them.

Let $C_i = C_i^p \cup C_{i0} \cup C_{i1}$, $1 \le i \le 8g - 4$, and T_C the cross section map of $\overline{G}_t$ for the cross section C.

Theorem 4.2. *The* C_i*'s are disjoint,* $C = \bigcup_{i=1}^{8g-4} C_i$*, and* T_C *satisfies*

$$
T_C(\xi, \eta) = T_i(\xi, \eta), \quad (\xi, \eta) \in C_i \qquad (4.2)
$$

where $T_i(\xi, \eta) = (T_i(\xi), T_i(\eta))$.

Before proving the above, we first sketch C_i and $C_i' = T_C(C_i)$. The sketches are in the θ, ϕ plane where $\theta = \arg\xi$, $\phi = \arg\eta$. To avoid additional notation, we shall freely interchange ξ with $\arg\xi$ and η with $\arg\eta$. For purposes of visualization, we distort distances but preserve order among the points a_i, b_i along the $\xi - \eta$-axes. Let $l_i = (a_i, b_i)$, $u_i = (b_i, a_i)$. We have the following graph for C_i. (See Figure 4.3 on p. 260.)

The sets C_i^p, C_{i0}, C_{i1} are respectively the interior, left boundary, and upper boundary of C_i. The corner points l_i, l_{i+1}, u_{i-1}, u_i are not in C_i, nor are its right and lower boundaries. The omissions are indicated in Figure 4.3 by dots and dashed lines.

By Theorem 3.4, we find that T_i maps the four corner points of C_i: l_i, l_{i+1}, u_{i-1}, u_i respectively to $u_{\sigma(i)}$, $l_{\sigma(i)-1}$, $u_{\sigma(i)+1}$, $l_{\sigma(i)}$. Thus, we obtain the following sketch of C_i'. (See Figure 4.4 on p. 260.)

We observe from Figures 4.3, 4.4 that T_C stretches distances in the ξ-direction and contracts in the η-direction, reflecting a well-known property of the flow $\overline{G}_t$ [Ar].

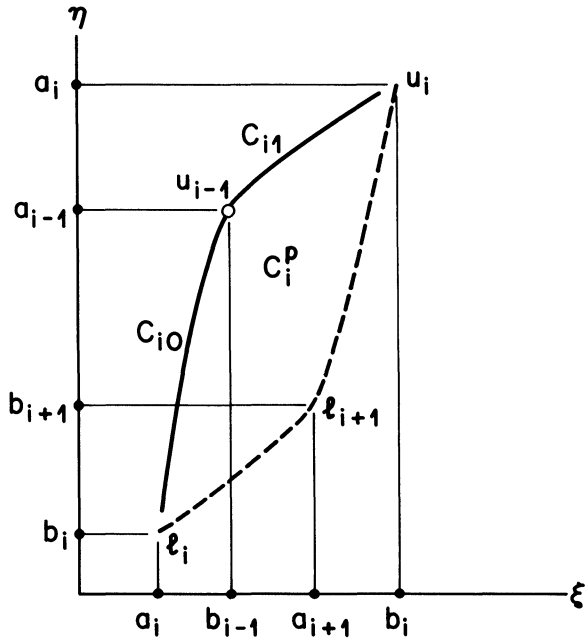


FIGURE 4.3. Graph of C_i.

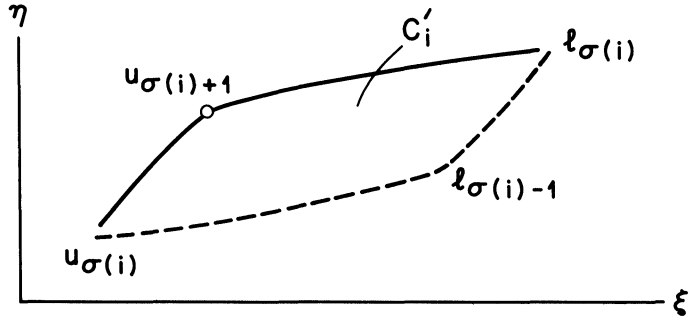


FIGURE 4.4. Image of C_i under T_i.

Proof of Theorem 4.2. We refer to Figures 4.3 and 4.4.

For $(\xi, \eta) \in C_i^p$, let $\nu = (\xi, \eta) = \overline{\pi}(u)$, $u = (\xi, \eta, \cdot)$ as in Figure 4.5 (i). The geodesic ray $\gamma^+(u)$, which begins at the base point

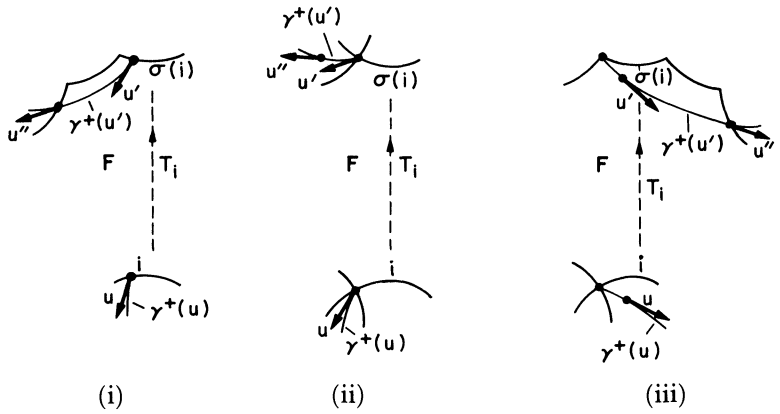


FIGURE 4.5. Action of T_C on vectors.

of u and leaves $\mathbf{F}$, is identified with the geodesic ray $\gamma^+(u') = T_i(\gamma^+(u))$ which begins at the base point of $u' = (T_i(\xi, \eta), \cdot)$ and enters $\mathbf{F}$. Let $u'' = (T_i(\xi, \eta), \cdot)$ be the unit tangent to $\gamma^+(u')$ at the point of exit from $\mathbf{F}$. Then $T_C(\nu) = \overline{\pi}(u'') = T_i(\xi, \eta)$.

The cases C_{i0}, C_{i1} are treated similarly as illustrated in Figures 4.5 (ii), (iii). In all cases $\nu = (\xi, \eta) = \overline{\pi}(u)$ and $T_C(\nu) = \overline{\pi}(u'') = T_i(\xi, \eta)$.

In Figure 4.6 we sketch $\mathbf{C}$ both as a union of C_i's and C_i''s, C_i and C_i' being labeled respectively by i and i'. To facilitate the sketching, we have treated the set of $(8g - 4)$ points a_i, b_i as if they were spaced uniformly over $\partial\mathbb{D}$. In reality this is not so, and the graphs are distortions of the real ones for C_i and C_i'.

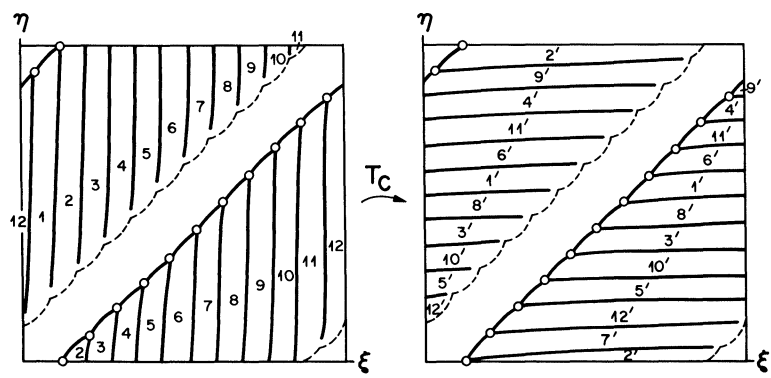


FIGURE 4.6. Action of T_C.

Observe that the set of points $\{u_i\}$ and the lower boundary of C are not in C. The first omission reflects the fact that the vectors u used to coordinatize $\mathbf{C}$ are not tangent to $\partial\mathbf{F}$. The second omission reflects the fact that the vectors u, which are the tangents to geodesics through p_{i+1} and used to coordinatize $\mathbf{C}$, do not point into $T_i^{-1}\mathbf{F}$.

We remark that the $\overline{G}_t$-invariant measure $\overline{m}$ induces a T_C-invariant measure of m_C on C. The measure is defined as

$$m_C(B) = \lim_{\Delta s \to 0} \frac{\overline{m}(\overline{G}_t(u)\colon u \in B,\ 0 \le t \le \Delta s)}{\Delta s}$$

for any measurable subset $B \subset C$. In the ξ, η coordinates dm_C is obtained by dropping ds from $d\overline{m}$, i.e.

$$dm_C = \frac{|d\xi|\,|d\eta|}{|\xi - \eta|^2}. \tag{4.3}$$

A geometric proof that m_C is T_C-invariant is given in [AF1]. The T_C-invariance also follows from formula (4.3) and the observation that, in the ξ, η coordinates, T_C piecewise transforms these variables by the same Möbius transformation.

5. Conjugacy of curvilinear and rectilinear maps

It seems difficult, if not impossible, to obtain the ergodic properties of $T_C(\xi, \eta)$ directly from formula (4.2). To do so, we define an auxiliary map $T_R(\overline{\xi}, \overline{\eta})$ which will be shown conjugate to $T_C(\xi, \eta)$.

The set R is obtained from C by replacing the boundary curves of C by polygonal lines joining successive l_i's and u_i's in the manner indicated in Figure 5.1: i.e. l_i is joined to l_{i+1}, and u_i to u_{i+1}, by moving first up and then right.

As in the case of C, R does not contain the points l_i and u_i. R contains the boundary curve joining the u_i's and does not contain the boundary curve joining the l_i's. Let R_i be the part of R for which $\xi \in [a_i, a_{i+1})$ and S_i the part for which $\eta \in (b_i, b_{i+1}]$. R_i and S_i are sketched in Figure 5.2. Observe that in drawing R_i and S_i we have used the fact that b_{i-1} lies between a_i and a_{i+1}, and a_{i+2} between b_i and b_{i+1}. This is a consequence of Theorem 3.4, which itself is a consequence of the extension condition. Thus the definition if R is intimately tied to the extension condition.

The boundaries of R_i are obtained by passing the appropriate horizontal and vertical lines through the points l_i, l_{i+1}, u_{i-1}, u_i, as indicated in Figure 5.2 (i). Similarly, the boundaries of S_i are

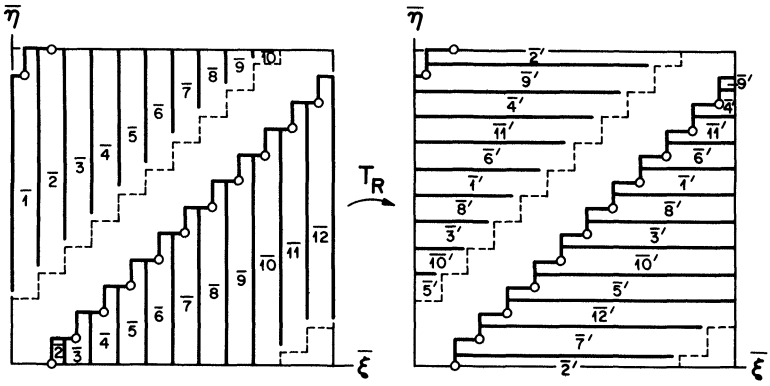


FIGURE 5.1. Action of the rectilinear map on its domain.

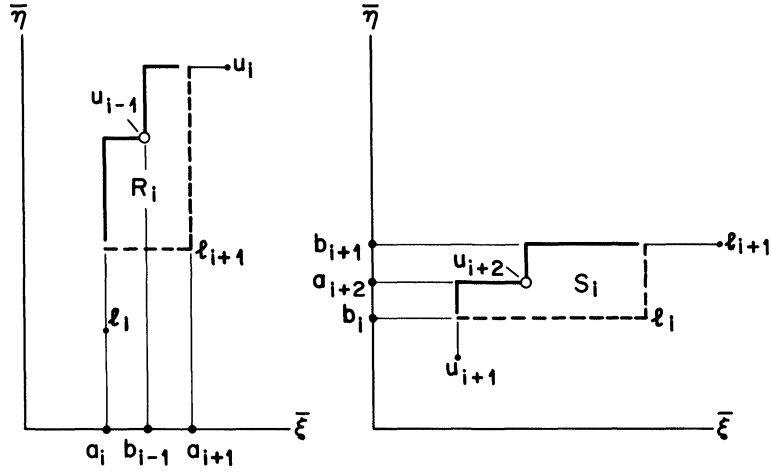


FIGURE 5.2. Portion of the rectilinear domain.

obtained with the aid of the points l_i, l_{i+1}, u_{i+1}, u_{i+2}. T_R is defined by

$$T_R(\bar{\xi}, \bar{\eta}) = T_i(\bar{\xi}, \bar{\eta}), \qquad (\bar{\xi}, \bar{\eta}) \in R_i. \tag{5.1}$$

Let $R_i' = T_R(R_i) = T_i(R_i)$. Since T_i maps l_i, l_{i+1}, u_{i-1}, u_i respectively to $u_{\sigma(i)}$, $l_{\sigma(i)-1}$, $u_{\sigma(i)+1}$, $l_{\sigma(i)}$, we get $R_i' = S_{\sigma(i)-1}$. Thus T_R is a one-to-one mapping of R onto itself. T_R is depicted in Figure 5.1, where R_i and R_i' have been replaced respectively by the labels $\bar{i}$ and $\bar{i}'$.

The R_i's are straightened out versions of the C_i's. For this reason we call C and R respectively the curvilinear and rectilinear

domains. Similarly we call $T_C(\xi, \eta)$ and (ξ, η) the *curvilinear transformation and coordinates*, and $T_R(\bar{\xi}, \bar{\eta})$ and $(\bar{\xi}, \bar{\eta})$ the *rectilinear transformation and coordinates.*

Since $T_R(\bar{\xi}, \bar{\eta})$ is defined by piecewise transforming the variables by the same linear transformation, it also preserves the measure $dm_C = |d\bar{\xi}| \cdot |d\bar{\eta}|/|\bar{\xi} - \bar{\eta}|^2$.

The reason for introducing T_R stems from the fact that it has a feature which T_C does not have. Namely, T_R maps any vertical line into part of a vertical line and T_R^{-1} maps any horizontal line into part of a horizontal line. Thus we have

(5.2)

$$T_R(\bar{\xi}, \bar{\eta}) = (f(\bar{\xi}), \cdot), \quad \text{where } f(\bar{\xi}) = T_i(\bar{\xi}),\ \bar{\xi} \in [a_i, a_{i+1})$$

$$T_R^{-1}(\bar{\xi}, \bar{\eta}) = (\cdot, g(\bar{\eta})), \quad \text{where } g(\bar{\xi}) = T_i(\bar{\xi}),\ \bar{\xi} \in (b_{i-1}, b_i].$$

The maps $f(\bar{\xi})$, $g(\bar{\xi})$ have been studied by Bowen and Series [BoS]. We refer to them as the left and right Bowen-Series maps. $f(\bar{\xi})$ is called a factor of T_R, and T_R an extension of f. Similarly, $g(\bar{\xi})$ is a factor of T_R^{-1} and T_R^{-1} an extension of g. In §7, we shall derive the ergodic properties of T_R from those of f and g.

We set up a conjugacy between T_C and T_R by constructing a one-to-one map Φ from C onto R satisfying

$$T_R \circ \Phi = \Phi \circ T_C \qquad \Phi = \text{ identity on } O = C \cap R. \tag{5.3}$$

Let

$$U_i = \text{ part of } C - R \quad \text{for which } \xi \in [a_i, a_{i+1}),$$
$$\overline{U}_i = \text{ part of } R - C \quad \text{for which } \xi \in [b_{i-1}, b_i).$$

U_i and $\overline{U}_i$ are depicted in Figure 5.3. We refer to the U_i's and $\overline{U}_i$'s as "bulges" and "corners."

Thus C is the union of O and bulges, and R the union of O and corners. We obtain a conjugacy map Φ which maps each bulge into a corner.

Let $\mathscr{P}_C = \{O, U_1, \ldots, U_{8g-4}\}$, $\mathscr{P}_R = \{O, \overline{U}_1, \ldots, \overline{U}_{8g-4}\}$. Partition C_i by $\mathscr{P}_C \vee T_C^{-1}\mathscr{P}_C$, and R_i by $\mathscr{P}_R \vee T_R^{-1}\mathscr{P}_R$. The resulting partitions for C_i, R_i and their T_C, T_R-images are sketched in Figures 5.4, 5.5, where we have used the notation $D_i' = T_C(D_i)$, etc. Each of the sets appearing in Figures 5.4, 5.5, contains its upper and left boundaries but not its lower and right boundaries.

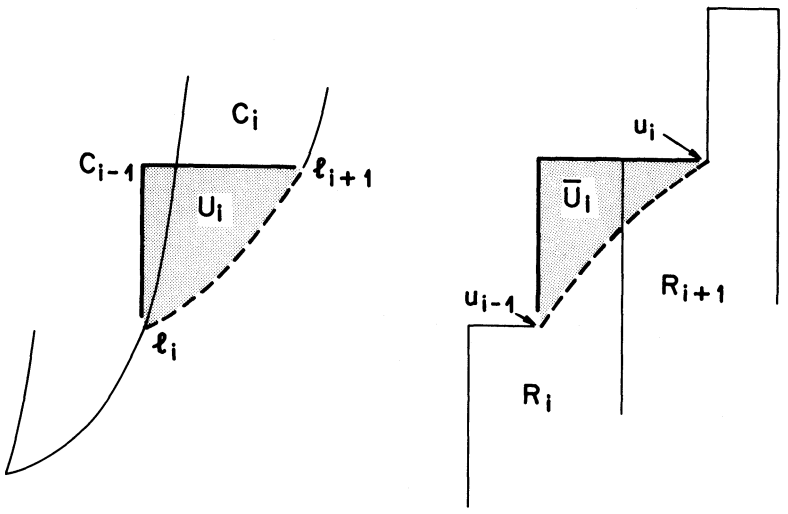


FIGURE 5.3. Bulges and corners.

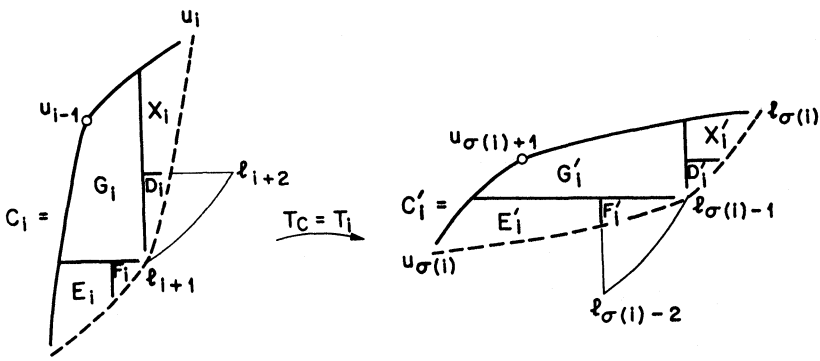


FIGURE 5.4. Action of curvilinear map on D_i, etc.

Figures 5.4, 5.5 put into evidence the following set relations. They will be used to prove Theorem 5.1.

$$
\begin{aligned}
&G_i = \overline{G}_i,\ X_i = \overline{X}_{i+1} \\
&O = \bigcap_{i=1}^{8g-4} (G_i \cup X_i) = \bigcup_{i=1}^{8g-4} (G'_i \cup E'_i) \\
&U_i = D_{i-1} \cup E_i \cup F_i, \qquad \overline{U}_i = \overline{D}_i \cup \overline{E}_i \cup \overline{F}_{i+1} \\
&D'_i \cup X'_i \subset U_{\sigma(i)-1}, \qquad F'_i \subset U_{\sigma(i)-2}.
\end{aligned} \tag{5.4}
$$

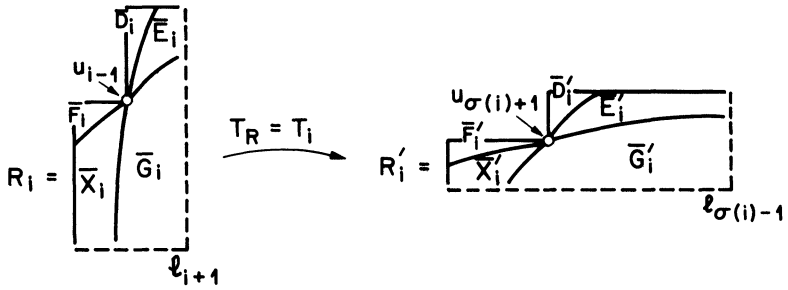


FIGURE 5.5. Action of rectilinear map on $\overline{D}_i$, etc.

We define the conjugacy map Φ by

$$\Phi = \begin{cases} I = \text{identity} & \text{on } O \\ T_{\rho(i)} \circ T_i & \text{on } U_i,\ 1 \le i \le 8g-4 \end{cases} \tag{5.5}$$

where $\rho(i) = \sigma(i) - 1$.

We prove that Φ is a one-to-one map from C onto R satisfying (5.3). We have arrived at the formula for Φ from the following considerations. Suppose that $\Phi = I$ on O and Φ satisfies the conjugacy relation of (5.3) on X_i. This forces $\Phi = T_{i+1}T_{\sigma(i)}$ on X_i', which is the same as $\Phi = T_{\rho^2(i)}T_{\rho(i)}$ on X_i' (see formula (5.11)). By (5.4) X_i' is a proper subset of $U_{\rho(i)}$, and we simply guess that $\Phi = T_{\rho^2(i)}T_{\rho(i)}$ extends to all of $U_{\rho(i)}$. This formula for Φ becomes (5.5) after replacing $\rho(i)$ by i. We remark that Φ is uniquely determined by (5.5). We omit the proof of uniqueness since it is not required later on.

Theorem 5.1. *$T_{\rho(i)} \circ T_i$ is a one-to-one map form U_i onto $\overline{U}_{i+4g-1}$, mapping D_{i-1}, E_i, F_i respectively onto $\overline{D}_{i+4g-1}$, $\overline{E}_{i+4g-1}$, $\overline{F}_{i+4g}$.*

To prove Theorem 5.1 we first collect some of the mapping properties of T_i. We recall from §3 that $\vartheta(i) = \sigma(i) + 1$.

Lemma 5.1. *Let $u_i = (b_i, a_i)$, $l_i = (a_i, b_i)$, and γ_i the curve with equation $\eta = \eta_i(\xi)$. T_i maps the points u_{i-1}, u_i, u_{i+1}, l_{i-1}, l_i, l_{i+1} respectively to $u_{\vartheta(i)}$, $l_{\sigma(i)}$, $u_{\rho(i)}$, $l_{\vartheta(i)}$, $u_{\sigma(i)}$, $l_{\rho(i)}$, and the curves γ_i, γ_{i+1} respectively to $\gamma_{\vartheta(i)}$, $\gamma_{\sigma(i)}$.*

The lemma follows directly from Theorem 3.4 and the fact that T_i maps the points (p_i, p_{i+1}) to $(p_{\vartheta(i)}, p_{\sigma(i)})$.

Proof of Theorem 5.1. We illustrate the decompositions of U_i, $\overline{U}_i$ in Figure 5.6. The pairs of regions (D_{i-1}, E_i), (E_i, F_i), $(\overline{D}_i, \overline{E}_i)$, $(\overline{E}_i, \overline{F}_{i+1})$ are separated respectively by: the curve γ_i, the vertical line $\xi = T_i^{-1}(a_{\sigma(i)-2})$, the curve $T_i^{-1}(\gamma_{\sigma(i)+2})$, the vertical line $\xi = a_{i+1}$.

We must show that $T_{\rho(i)} \circ T_i$ maps l_i, γ_{i+1}, l_{i+1} respectively to u_{i+4g-2}, γ_{i+4g-1}, u_{i+4g-1} (hence $T_{\rho(i)} \circ T_i$ maps U_i onto $\overline{U}_{i+4g-1}$), and γ_i, $T_i^{-1}(a_{\sigma(i)-2})$ to $T_{i+4g-1}^{-1}(\gamma_{\sigma(i+4g-1)+2})$, a_{i+4g}. These facts follow readily from Theorem 3.4, Lemma 5.1, and the relation

$$\sigma\rho(i) = i + 4g - 1, \tag{5.6}$$

which is a consequence of (3.1). To illustrate, we give the details for l_i. By Lemma 5.1

$$T_{\rho(i)}T_i(l_i) = T_{\rho(i)}(u_{\sigma(i)}) = T_{\rho(i)}(u_{\rho(i)+1}) = u_{\rho^2(i)}. \tag{5.7}$$

By (5.6)

$$\rho^2(i) = \sigma\rho(i) - 1 = i + 4g - 2. \tag{5.8}$$

Hence

$$T_{\rho(i)}T_i(l_i) = u_{i+4g-2}. \tag{5.9}$$

Theorem 5.2. Φ *satisfies the conjugacy relation* (5.3).

Proof. We rewrite (5.3) as

$$T_R \circ \Phi \circ T_C^{-1} \circ \Phi^{-1} = I \tag{5.10}$$

and verify it separately on each of the five parts of C_i. In each case we draw a commuting diagram and use (5.4) along with Theorem 5.1 to express the four maps on the left side of (5.10) as products

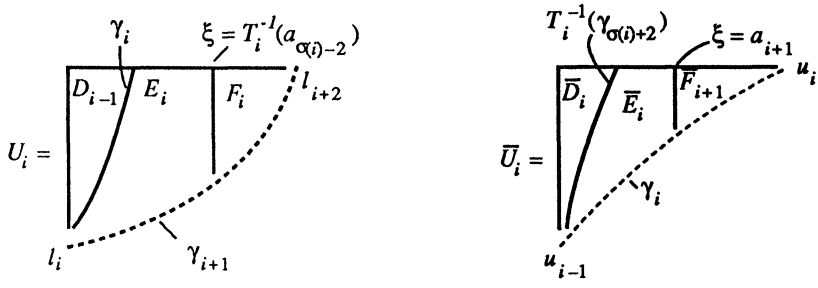


FIGURE 5.6. Partitions of bulges and corners.

of the T_i's. We work out details for X_i, the other parts being treated in similar manner. We find in each case that (5.10) is either trivial or becomes the group relation of Theorem 3.2(ii). To get (5.10) from the following commutative diagrams we start at the lower right corner and proceed counter-clockwise keeping in mind that $T_i^{-1} = T_{\sigma(i)}$.

X_i:

$$\begin{array}{ccc} X_i & \xrightarrow{T_C=T_i} & X_i' \\ {\scriptstyle \Phi=I}\downarrow & & \downarrow{\scriptstyle \Phi=T_{\rho^2(i)}T_{\rho(i)}} \\ \overline{X}_{i+1} & \xrightarrow[T_R=T_{i+1}]{} & \end{array}$$

Since $X_i \subset O$ and $X_i' \subset U_{\rho(i)}$, we have $\Phi = I$ on X_i and $\Phi = T_{\rho^2(i)}T_{\rho(i)}$ on X_i'. (5.10) becomes

$$T_{i+1} \circ T_{\sigma(i)} \circ T_{\sigma\rho(i)} \circ T_{\sigma\rho^2(i)} = I. \tag{5.11}$$

Letting $j = \sigma\rho^2(i)$, (5.11) becomes Theorem 3.2 (ii).

G_i:

$$\begin{array}{ccc} G_i & \xrightarrow{T_C=T_i} & G_i' \\ {\scriptstyle \Phi=I}\downarrow & & \downarrow{\scriptstyle \Phi=I} \\ \overline{G}_i & \xrightarrow[T_R=T_i]{} & \end{array}$$

(5.10) becomes

$$T_i \circ T_i^{-1} = I. \tag{5.12}$$

E_i:

$$\begin{array}{ccc} E_i & \xrightarrow{T_C=T_i} & E_i' \\ {\scriptstyle \Phi=T_{\rho(i)}T_i}\downarrow & & \downarrow{\scriptstyle \Phi=I} \\ \overline{E}_{i+4g-1} & \xrightarrow[T_R=T_{\rho(i)}^{-1}]{} & \end{array}$$

(5.10) becomes

$$T_{\rho(i)}^{-1} \circ T_{\rho(i)} \circ T_i \circ T_i^{-1} = I. \tag{5.13}$$

Observe that on $\overline{E}_{i+4g-1}$, $T_R = T_{i+4g-1}$. By (5.6) we rewrite $T_R = T_{\sigma\rho(i)} = T_{\rho(i)}^{-1}$ as done in the above diagram. Similar use of (5.6) is made in the following diagrams for E_i and F_i.

D_i:

$$\begin{array}{ccc} D_i & \xrightarrow{T_C=T_i} & D_i' \\ {\scriptstyle \Phi=T_{\rho(i+1)}T_{i+1}}\downarrow & & \downarrow{\scriptstyle \Phi=T_{\rho^2(i)}T_{\rho(i)}} \\ \overline{D}_{i+4g} & \xrightarrow[T_R=T_{\rho(i+1)}^{-1}]{} & \end{array}$$

(5.10) becomes

$$T_{i+1} \circ T_{\sigma(i)} \circ T_{\sigma\rho(i)} \circ T_{\sigma\rho^2(i)} = I \tag{5.14}$$

F_i:

$$\begin{array}{ccc} F_i & \xrightarrow{T_C=T_i} & F_i' \\ {\scriptstyle \Phi=T_{\rho(i)}T_i}\downarrow & & \downarrow{\scriptstyle \Phi=T_{\rho(\rho(i)-1)}T_{\rho(i)-1}} \\ \overline{F}_{i+4g} & \xrightarrow[T_R=T_{[\sigma\rho(i)+1]}]{} & \end{array}$$

(5.10) becomes

$$T_{\sigma\rho(i)+1} \circ T_{\rho(i)} \circ T_{\sigma(\rho(i)-1)} \circ T_{\sigma\rho(\rho(i)-1)} = I. \tag{5.15}$$

Letting $j = \sigma\rho^2(i)$ in (5.14), and $j = \sigma\rho(\rho(i)-1)$ in (5.15), they both become Theorem 3.2 (ii).

We conclude this section with an interpretation of both T_C and T_R as actions on subsets of $\mathbf{U}$. Although it adds nothing new to our treatment, Series [S2, S3] makes use of it to establish that T_R is a cross section map and is conjugate to T_C. The proof of Theorem 4.2 shows that we may think of T_C as an action on the set of representatives of $\mathbf{C}$ given in §4. In this interpretation $T_C(u)$ is defined to be the vector u'' appearing in the proof of Theorem 4.2. Similarly, we may think of R as the set of vectors $\Phi(u)$, u varying over the set of representatives of $\mathbf{C}$. We then define $T_R(\Phi u)$ to be $\Phi(T_C u)$. Remarkably, the vectors $\Phi(u)$, with the exception of those in $\Phi(\bigcup C_{i1})$, are all based either on $\partial\mathbf{F}$ or $\partial\mathbf{F}_i$, $1 \leq i \leq 8g-4$, the $\partial\mathbf{F}_i$'s being the fundamental regions opposite to $\mathbf{F}$ at its vertices. In all cases these vectors point out from the region on whose boundary they are based. R is obtained from C by: removing certain vectors based in $\partial\mathbf{F}$, and adding certain vectors based in $\partial\mathbf{F}_i$, $1 \leq i \leq 8g-4$. The removed vectors correspond to bulges and the added ones to corners. One

can produce a figure for T_R analogous to Figure 4.5 for T_C.

6. Bowen-Series map

In §5, we introduced the left and right Bowen-Series maps of $\partial\mathbb{D}$ onto itself. In this section, we describe properties of these maps leading to the fact that they both are expansive and ergodic.

We recall

Definition 6.1. Let

$$f(\xi) = T_i(\xi), \qquad \xi \in [a_i, a_{i+1}), \qquad 1 \le i \le 8g-4.$$
$$g(\xi) = T_i(\xi), \qquad \xi \in (b_{i-1}, b_i], \qquad 1 \le i \le 8g-4.$$

We call $f(\xi)$, $g(\xi)$ the *left and right Bowen-Series maps* of $\partial\mathbb{D}$ onto itself. Let

$$\begin{aligned} I_{i_1} &= [a_i, b_{i-1}), \quad I_{i_2} = [b_{i-1}, a_{i+1}), \\ J_{i_1} &= (a_{i+1}, b_i], \quad J_{i_2} = (b_{i-1}, a_{i+1}]. \end{aligned} \tag{6.1}$$

We depict these intervals in Figure 6.1.

Let $\mathscr{P}_I = \{I_{1_1}, I_{1_2}, \ldots, I_{(8g-4)_2}\}$, $\mathscr{P}_J = \{J_{1_1}, J_{1_2}, \ldots, J_{(8g-4)_2}\}$. $\mathscr{P}_I$, $\mathscr{P}_J$ are partitions of $\partial\mathbb{D}$.

Theorem 6.1. *$f(\xi)$ is Markovian and aperiodic with respect to $\mathscr{P}_I$. This means respectively that:*

(i) *The f-image of any interval in $\mathscr{P}_I$ is a union of intervals in $\mathscr{P}_I$.*

(ii) *There exists a positive integer n such that $\partial\mathbb{D}$ is the f^n-image of any interval in $\mathscr{P}_I$.*

Similarly, g is Markovian and aperiodic with respect to $\mathscr{P}_J$.

Proof. The proof of Theorem 6.1 is given in [BoS]. We give it here for the sake of completeness.

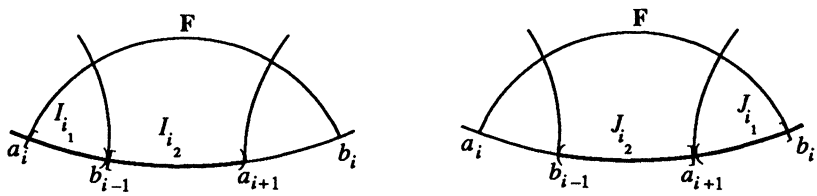


FIGURE 6.1. Intervals on $\partial\mathbb{D}$.

(i) From Theorem 3.4 we obtain

$$(6.2)\qquad \begin{cases} f(I_{i_1}) = I_{(\sigma(i)+1)_2} \cup I_{(\sigma(i)+2)_1}, \\ f(I_{i_2}) = I_{(\sigma(i)+2)_2} \cup I_{(\sigma(i)+3)_1} \cup \cdots \cup I_{(\sigma(i)-2)_1} \cup I_{(\sigma(i)-2)_2}. \end{cases}$$

and

$$(6.3)\qquad \begin{cases} g(J_{i_1}) = J_{(\sigma(i)-1)_2} \cup J_{(\sigma(i)-2)_1}, \\ g(J_{i_2}) = J_{(\sigma(i)-2)_2} \cup J_{(\sigma(i)-3)_1} \cup \cdots \cup J_{(\sigma(i)+2)_1} \cup J_{(\sigma(i)+2)_1}. \end{cases}$$

(6.2) gives (i) for f, and (6.3) gives (i) for g.

(ii) We prove (ii) for f, the proof for g being similar. We show that for any i,

$$(6.4)\qquad \partial\mathbb{D} \subset f(I_{i_2}) \cup f(I_{(i+1)_2}).$$

From (3.1) we get

$$(6.5)\qquad \sigma(i+1) \equiv \sigma(i) + 4g - 3 \bmod(8g-4).$$

From (6.1), (6.2), and (6.5) we get

$$(6.6)\qquad f(I_{i_2}) = [b_{j+2}, a_j),\ f(I_{(i+1)_2}) = [b_{j+4-1}, a_{j+4g-3})$$

where $j = \sigma(i) - 1$.

Going along $\partial\mathbb{D}$ counter-clockwise we encounter successively the points a_j, b_{j+2}, a_{j+4g-3}, b_{j+4g-1}. Hence, (6.4) follows from (6.6) (see Figure 6.2 on p. 272).

From (6.2) we get that for any i, $f(I_{i_2})$ contains two intervals I_{k_2}, $I_{(k+1)_2}$ for some k. It follows from (6.4) that $\partial\mathbb{D} \subset f^2(I_{i_2})$. From (6.2) we also get that for any i, $f(I_{i_1})$ contains an interval I_{l_2}, l depending on i_1, with a similar statement holding for $f(I_{i_2})$. We conclude that $\partial\mathbb{D} \subset f^3(I_{i_1})$, $\partial\mathbb{D} \subset f^3(I_{i_2})$.

Theorem 6.2. *For any Borel subset A of $\partial\mathbb{D}$, let $\mu(A) = \int_A h(\xi)|d\xi|$, $\nu(A) = \int_A k(\eta)|d\eta|$, where*

$$(6.7)\qquad h(\xi) = \begin{cases} \displaystyle\int_{b_{i+1}}^{a_i - 1} \frac{|d\eta|}{|\xi-\eta|^2}, & \xi \in I_{i_1}, \\ \displaystyle\int_{b_{i+1}}^{a_i} \frac{|d\eta|}{|\xi-\eta|^2}, & \xi \in I_{i_2}. \end{cases}$$

$$k(\eta) = \begin{cases} \displaystyle\int_{b_{i+1}}^{a_{i-1}} \frac{|d\eta|}{|\xi-\eta|^2}. & \eta \in J_{i_1}, \\ \displaystyle\int_{b_i}^{a_{i-1}} \frac{d\eta}{|\xi-\eta|^2}, & \eta \in J_{i_2}. \end{cases}$$

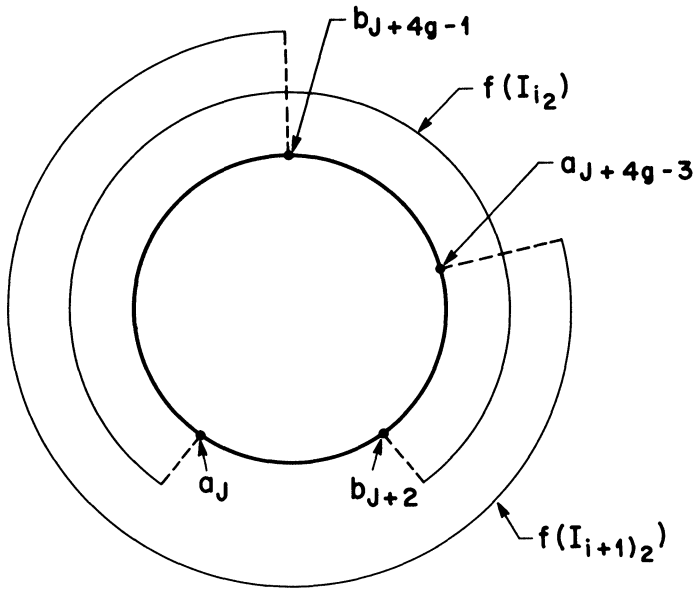


FIGURE 6.2. Images of intervals.

Then

$$\mu(f^{-1}A) = \mu(A), \qquad \nu(g^{-1}A) = \nu(A).$$

We remark that the above integrals can be evaluated in closed form. We do not write down the formulas as they are not required later.

Proof. Let $\pi_\xi(\xi, \eta) = \xi$, $\pi_\eta(\xi, \eta) = \eta$, $(\xi, \eta) \in R$. Then

$$\mu(A) = m(\pi_\xi^{-1}A), \qquad \nu(A) = m(\pi_\eta^{-1}A)$$

where $dm = |d\xi|\,|d\eta|/|\xi - \eta|^2$.

Since $f \circ \pi_\xi = \pi_\xi \circ T_R$,

$$\pi_\xi^{-1} f^{-1} = T_R^{-1}\pi_\xi^{-1}.$$

Since m is T_R-invariant,

$$\mu(f^{-1}A) = m(\pi_\xi^{-1}f^{-1}A) = m(T_R^{-1}\pi_\xi^{-1}A) = m(\pi_\xi^{-1}A) = \mu(A).$$

Similarly, we show $\nu(g^{-1}A) = \nu(A)$.

Let $f^{-n}\mathscr{P}_I = \{f^{-n}I, I \in \mathscr{P}_I\}$ and $\mathscr{P}_I^{(n)}$ be the common refinement of $\mathscr{P}_I$, $f^{-1}\mathscr{P}_I, \ldots, f^{-n}\mathscr{P}_I$. $\mathscr{P}_I^{(n)}$ partitions $\partial\mathbb{D}$ into intervals. Let $\|\mathscr{P}_I^{(n)}\| = \operatorname{Sup}|I|$, $|I|$ denoting arc length of I.

Similarly we define $\mathscr{P}_J^{(n)}$ and $||\mathscr{P}_J^{(n)}||$. The following result is needed in §§7,8. It is a consequence of Theorems 6.1, 6.2, and the converse to the folklore theorem (Theorem B2 of Appendix B).

Theorem 6.3. *f is expansive. This means that there exists an integer $n > 0$ and $\varepsilon > 0$ such that*

$$|f^{n\prime}(x)| > 1 + \varepsilon, \quad x \in \partial\mathbb{D} \text{ but } x \neq \text{ end point of interval in } \mathscr{P}^{(n)}.$$

As a consequence we get

$$||\mathscr{P}^{(n)}|| = O(\theta^{-n}) \text{ for some } 0 < \theta < 1.$$

Similar statements hold for g.

We remark that Series [S2, p. 355] proves the same result by a completely different method. Her proof has the interesting feature that it involves the interior of the disk, whereas the maps f and g are defined only on the boundary. The idea is as follows. Iterates of the map are piecewise equal to motions—i.e., Möbius transformations mapping $\mathbb{D}$ onto itself—specified by words in the generators. Upon iteration these motions map a central fundamental region to ones which are, in the Euclidean metric, exponentially small and close to the boundary of the disk. Intervals on the boundary in a basic partition refined by iteration of the map are related to fundamental regions in the interior, and this relation forces these intervals also to be exponentially small. Expansiveness is then a simple consequence of this fact.

Opposed to the proof of Series, ours involves only the domain of the maps f and g. The ingredients in our proof are Theorem B2 and formulas (6.7) for the invariant measures for f and g (although we would never have found these if not for their relation to the geodesic flow on $\mathbb{D}/\Gamma$, a concept related to the interior of $\mathbb{D}$).

As another consequence of Theorems 6.1, 6.2, and B2 we obtain

Theorem 6.4. *f and g are ergodic with respect to Lebesgue measure.*

7. Ergodic relationship of f and $\overline{G}_t$

We now employ results of [Ro] concerning *endomorphisms* and *automorphisms* of a Lebesgue space, these terms referring respectively to many-to-one a.e. and one-to-one a.e. measure preserving transformations acting on measure spaces that are measure theoretically identical with the unit interval. By virtue of the relations

$f \circ \pi_\xi = \pi_\xi \circ T_R$ and $g \circ \pi_\eta = \pi_\eta \circ T_R^{-1}$ introduced in §6, the right and left Bowen maps f and g are respectively *factor maps* of T_R and T_R^{-1}. Conversely, T_R and T_R^{-1} turn out to be the *natural extensions* of f and g. This means: (i) T_R is one-to-one onto a.e.; (ii) the Borel field generated by $\bigcup_{n=0}^{\infty} T_R^{-n}\mathscr{A}$ is $\mathbf{B}_R$, where $\mathscr{A} = \pi_\xi^{-1}\mathbf{B}_{\partial\mathbb{D}}$ and $\mathbf{B}_R$, $\mathbf{B}_{\partial\mathbb{D}}$, are respectively the Borel fields of measurable subsets of R and $\partial\mathbb{D}$. (i) is true by definition. To prove (ii) it suffices to show that the common refinement of the partitions $T_R^k(\pi_\xi^{-1}\mathscr{P}_I)$, $-n \le k \le n$, consists of rectangles, the largest of whose diameters decreases to 0 as $n \to \infty$. This follows from the fact that vertical and horizontal distances contract exponentially under the respective actions of T_R^n and T_R^{-n}, which in turn is due to the expansiveness of f and g.

An endomorphism f of a Lebesgue space is called *exact* if it has *trivial tail field*—i.e. $A \in \bigcap_{n=0}^{\infty} f^{-n}\mathbf{B}$, where $\mathbf{B}$ is the Borel field of measurable sets, implies $\mu(A) = 0$ or 1. Thus the left and right Bowen-Series maps are exact from Theorems 6.1, 6.3 and Lemma B.4. An automorphism T of a Lebesgue space is called *Kolmogoroff* if there exists a sub σ-algebra $\mathscr{A}$ of $\mathbf{B}$ such that (i) $\mathscr{A} \subset T\mathscr{A}$, (ii) the Borel field generated by $\bigcup_{n=-\infty}^{\infty} T^n\mathscr{A}$ is $\mathbf{B}$, and (iii) $\bigcap_{n=-\infty}^{\infty} T^n\mathscr{A}$ is trivial in the sense that it contains only sets of measure 0 or 1.

Theorem 7.1 [Ro]. *An automorphism is exact if and only if its natural extension is Kolmogoroff.*

Corollary 7.1. *T_R is Kolmogoroff. In particular, it is ergodic* [Ro].

Theorem 7.2. *The following are equivalent*:

(i) *$\overline{G}_t$ is ergodic.*
(ii) *T_C is ergodic.*
(iii) *T_R is ergodic.*

Proof. The equivalence of (i) and (ii) follows from elementary set considerations, and that of (ii) and (iii) from conjugacy.

8. Symbolic dynamics of the rectilinear map

Let (X, T) be an abstract dynamical system, i.e. a space X and a mapping T of X into itself. Let $\mathscr{A}$ be a finite alphabet of symbols and $\mathscr{P} = \{X_a : a \in \mathscr{A}\}$ a partition of X into disjoint subsets. We associate with each point $x \in X$ a sequence $\{s_n\}$ of symbols $s_n = s_n(x) \in \mathscr{A}$ according as $T^n x \in X_{s_n}$. The sequences

describe the history of orbits through the partition $\mathscr{P}$. They are bilateral, $-\infty < n < \infty$, or unilateral, $0 \leq n < \infty$, depending on whether T is invertible or not. We use T to indicate invertible maps and f to indicate noninvertible maps. We call respectively the sequences $\{s_n\}$ corresponding to T and f, *the T- and f-expansions of* x.

Let $\Omega = \Omega(T)$ be the space of sequences $\{s_n(x)\}$, x varying over X. We refer to this description of Ω as the orbit history description. Let ϕ denote the mapping of X onto Ω defined by $\phi(x) = \{s_n(x)\}$, and σ the shift on $\Omega(T)$ defined by $\sigma\{s_n\} = \{s_{n+1}\}$. We call the pair $(\Omega(T), \sigma)$ a symbolic dynamical system. Since $T^n(Tx) = T^{n+1}(x) \in X_{s_{n+1}}$, we have $s_n(Tx) = s_{n+1}(x)$ or $\phi(Tx) = \sigma(\phi(x))$. For invertible ϕ we obtain the conjugacy $T = \phi^{-1}\sigma\phi$.

The orbit history description of Ω is inadequate in that it does not give a procedure for deciding which sequences are in it. What is needed is another description of Ω given by a set of simple admissibility rules. Our admissibility rules are specified by forbidden blocks of symbols, the finite blocks denoted by $\mathscr{L}_0$ and the infinite ones by $\mathscr{L}_\infty$. The symbolic systems defined by the $\mathscr{L}_0$-restrictions are either Markovian or sofic. The further imposed $\mathscr{L}_\infty$-restrictions remove from these symbolic systems a *negligible* set, a concept defined in Appendix C part IIb. The sequence space defined by the admissibility rules will be denoted by $\Omega_a(T)$.

In order that the two descriptions specify the same sequence spaces, i.e. $\Omega(T) = \Omega_a(T)$, we must show that $\psi\{s_n\} = \bigcap T^{-n} X_{s_n} \neq \varnothing$ if and only if $\{s_n\} \in \Omega_a(T)$. ψ will be a one-to-one map from $\Omega_a(T)$ onto X if and only if $\psi\{s_n\}$ consists of a single point $x \in X$ for all $\{s_n\} \in \Omega_a(T)$.

Let T_R be the rectilinear map and f the factor map of T_R, i.e. the left Bowen-Series map. In the present section, we show that f and T_R are conjugate to shifts on sequence spaces described by simple admissibility rules. For the curvilinear map T_C, we have no corresponding result.

When discussing f, we consider the two partitions of $\partial\mathbb{D}$: $\{I_1, I_2, \ldots, I_{8g-4}\}$ where $I_i = [a_i, a_{i+1}]$, referred to as the *coarse* partition, and $\{I_{1_1}, I_{1_2}, \ldots, I_{(8g-4)_2}\}$ where $I_{i_1} = [a_i, b_{i-1})$, $I_{i_2} = [b_{i-1}, a_{i+1})$, referred to as the *fine* partition (see Figure 6.1). We then replace respectively $\Omega(f)$ by $\overline{\Omega}(f)$ and $\overline{\overline{\Omega}}(f)$. Similarly, when discussing T_R, consider the two partitions of R again

referred to as *coarse* and *fine*: $\{R_1, R_2, \dots, R_{8g-4}\}$ and $\{R_{1_1}, R_{1_2}, \dots, R_{(8g-4)_2}\}$, where $R_i = \pi_\xi^{-1} I_i$, $R_{i_k} = \pi_\xi^{-1} I_{i_k}$ and $i \in \{1, 2, \dots, 8g-4\}$, $i_k \in \{1_1, 1_2, \dots, (8g-4)_1, (8g-4)_2\}$. We then replace respectively $\Omega(T_R)$ by $\overline{\Omega}(T_R)$ and $\overline{\overline{\Omega}}(T_R)$. Corresponding change of notation holds for $\Omega_a(f)$ and $\Omega_a(T_R)$.

We describe the spaces $\overline{\overline{\Omega}}_a(f)$, $\overline{\overline{\Omega}}_a(T_R)$, $\overline{\Omega}_a(f)$, $\overline{\Omega}_a(T_R)$ specifying in each case $\mathscr{A}$, $\mathscr{L}_0$, $\mathscr{L}_\infty$. For $\overline{\overline{\Omega}}(f)$, $\overline{\overline{\Omega}}_a(T_R)$ the blocks of $\mathscr{L}_0$ are of length two. In these cases, $\mathscr{L}_0$ is equivalent to a set $\mathscr{T}$ of Markovian transition rules imposed on the members of $\mathscr{A}$ which require $s_n \to s_{n+1}$ for all n.

For economy of notation we drop from now on the bars from $\overline{\xi}$, $\overline{\eta}$. We let

$$\vartheta(i) = \sigma(i) + 1, \qquad \rho(i) = \sigma(i) - 1,$$
$$\alpha(i) = \sigma(i) + 2, \qquad \beta(i) = \sigma(i) - 2.$$

In each of the descriptions given below i varies from 1 to $8g-4$.

I. $\overline{\overline{\Omega}}_a(f)$: $\mathscr{A} = \{1_1, 1_2, \dots, (8g-4)_1, (8g-4)_2\}$.

$$\mathscr{T}: i_1 \to (\sigma(i)+1)_2, (\sigma(i)+2)_1,$$
$$i_2 \to (\sigma(i)+2)_2, (\sigma(i)+3)_1, \dots, (\sigma(i)-2)_1, (\sigma(i)-2)_2.$$

$$\mathscr{L}_\infty: \quad i_1, (\alpha(i))_1, (\alpha^2(i))_1, \dots,$$
$$i_2, (\beta(i))_2, (\beta^2(i))_2, \dots.$$

II. $\overline{\overline{\Omega}}_a(T_R)$: $\mathscr{A}$ and $\mathscr{T}$ same as for $\overline{\overline{\Omega}}_a(f)$. $\mathscr{L}_\infty$ consists of $\mathscr{L}_\infty$ for $\overline{\overline{\Omega}}_a(f)$ together with

$$\dots(\beta^2(i))_2, (\beta(i))_2, i_2,$$
$$\dots(\alpha^2(i))_2, (\alpha(i))_2, i_2, (\alpha(i))_2, (\alpha^2(i))_2 \dots.$$

III. $\overline{\Omega}_a(f)$: $\mathscr{A} = \{1, 2, \dots, 8g-4\}$.

$$\mathscr{L}_0: \quad i, \sigma(i); i, \rho(i);$$
$$i, \vartheta(i), \alpha\vartheta(i), \dots, \alpha^k\vartheta(i), \vartheta\alpha^k\vartheta(i); 0 \le k < \infty.$$
$$\mathscr{L}_\infty: \quad i, \beta(i), \beta^2(i), \dots.$$

IV. $\overline{\Omega}_a(T_R)$: $\mathscr{A}$ and $\mathscr{L}_0$ same as for $\overline{\Omega}_a(f)$. $\mathscr{L}_\infty$ consists of

$\mathscr{L}_\infty$ for $\overline{\Omega}_a(f)$ together with

$$\dots, \beta^2(i), \beta(i), i,$$
$$\dots, \alpha^2(i), \alpha(i), i, \alpha(i), \alpha^2(i), \dots = \dots, i, \alpha(i), i, \alpha(i), i, \dots.$$

Ignoring the $\mathscr{L}_\infty$ removals, $(\overline{\overline{\Omega}}_a(T_R), \sigma)$ is a topological Markov shift and $(\overline{\Omega}_a(T_R), \sigma)$ a sofic system. Indeed the latter is strictly sofic (see Appendix C). Although usually one prefers a topological Markov shift over a sofic system, in some respects the sofic system is superior. We shall clarify this point in the concluding paragraph of this section.

We prove a series of theorems to the effect that for each of the cases I–IV, $\Omega(T) = \Omega_a(T)$ and T is conjugate to the action of σ on $\Omega_a(T)$. This amounts to proving the following proposition:

(P)

(i) If $T^n x \in X_{s_n}$ for some $x \in X$ and all n, then $\{s_n\} \in \Omega_a(T)$.

(ii) If $\{s_n\} \in \Omega_a(T)$, then $\bigcap T^{-n} X_{s_n}$ consists of a single point $x \in X$.

All proofs of this section routinely use (P), but niggling details appear making for tedious reading. We include the proof of all cases for the sake of completeness. From here through Theorem 8.4, the sequences $\{s_n\}$ consist of symbols from the alphabet $\{1_1, 1_2, \dots, (8g-4)_2\}$.

The following lemma is the nested interval theorem modified to half-open intervals. As shown later, it explains the $\mathscr{L}_\infty$-list for $\overline{\overline{\Omega}}_a(f)$.

Lemma 8.1. *Let* $I_n = [a_n, b_n)$, $a_n < b_n$ *for* $0 \le n < \infty$; $I_1 \supset I_2 \supset \cdots \supset I_n \supset \cdots$; $\lim_{n\to\infty} |I_n| = 0$ *(where* $|I_n| = b_n - a_n$*). The* I_n*'s have a common point, which is unique, if and only if not all* b_n *agree beyond some* n*. A similar statement holds for* $I_n = (a_n, b_n]$ *with* a_n *replacing* b_n*.*

Proof. Let $I_n = [a_n, b_n)$, $\overline{I}_n = [a_n, b_n]$. $\{\overline{I}_n\}$ is a nested sequence of closed intervals with $\lim_{n\to\infty} |I_n| = 0$. By the nested interval theorem, there is a unique point $p \in \overline{I}_n$, $0 \le n < \infty$. Since $I_n \subset \overline{I}_n$, the I_n's have at most one common point, which must then be p. Suppose not all b_n agree from some n on. For each n choose $m > n$ so that $\overline{I}_m \subset I_n$. Then $p \in \overline{I}_m \subset I_n$. On the other hand, suppose $b_n = b$ for $n > n_0$. Then $b \in \overline{I}_n$,

$0 \leq n < \infty$, so that $p = b$. Since $b \notin I_n$ for $n > n_0$, the I_n's have no common point.

The proof for $I_n = (a_n, b_n]$ is identical.

Definition 8.1. A finite consecutive block $s_m, \ldots, s_n$ is called both f- and T_R-admissible if and only if $s_i \to s_{i+1}$, $m \leq i < n$, satisfies the transition rules for $\overline{\overline{\Omega}}_a(f)$ (which are identical with the transition rules for $\overline{\overline{\Omega}}_a(T_R)$). An infinite consecutive block $\{s_n\}$ is called both f- and T_R-admissible if and only if this is the case for every finite consecutive block in $\{s_n\}$.

The above notions can be stated in another way. In §6, we showed

$$(8.1) \quad \begin{cases} f(I_{i_1}) = I_{(\sigma(i)+1)_2} \cup I_{(\sigma(i)+2)_1}, \\ f(I_{i_2}) = I_{(\sigma(i)+2)_2} \cup I_{(\sigma(i)+3)_1} \cup \cdots \cup I_{(\sigma(i)-2)_1} \cup I_{(\sigma(i)-2)_2}. \end{cases}$$

From (8.1) we find that $k \to l$ satisfies the transition rules for $\overline{\overline{\Omega}}_a(f)$ if and only if $f(I_k) \cap I_l \neq \varnothing$. Since $T_R(\xi, \eta) = (f(\xi), \cdot)$, we have $T_R(R_k) \cap R_l \neq \varnothing$ if and only if $f(I_k) \cap I_l \neq \varnothing$. Thus $s_m, \ldots, s_n$ is f-admissible is equivalent to $f(I_{s_i}) \cap I_{s_{i+1}} \neq \varnothing$, $m \leq i < n$, and to $T_R(R_{s_i}) \cap R_{s_{i+1}} \neq \varnothing$, $m \leq i < n$.

The intervals in (8.1) are listed in the order in which they are encountered when $f(I_{i_k})$, $k = 1, 2$, is traversed counter-clockwise. We have chosen the order of the elements after the arrow in the transition rules for $\overline{\overline{\Omega}}_a(f)$ to conform with the order in which they appear as indices in (8.1).

We order the finite f-admissible sequence with the same s_0 lexicographically, i.e. let $s = \{s_0, \ldots, s_n, s_{n+1}\}$, $s' = \{s'_0, \ldots, s'_n, s'_{n+1}\}$ with $s_i = s'_i$, $0 \leq i \leq n$, and $s_{n+1} \neq s'_{n+1}$. We say that $s < s'$ if and only if s'_{n+1} occurs after s_{n+1} in the transition rules, i.e. we have $s_n \to \cdots s_{n+1} \cdots$, $s'_{n+1} \cdots$.

We give $\partial\mathbb{D}$ the counter-clockwise orientation, so that it is meaningful to speak of left and right end points of intervals in $\partial\mathbb{D}$.

Lemma 8.2. (i) *Let $s_0, \ldots, s_n$ be f-admissible. Then $I(s_0 \ldots s_n) = \bigcap_{k=0}^{n} f^{-k}(I_{s_k})$ is a nonempty interval closed on the left and open on the right.*

(ii) *If $\{s_n\}$, $0 \leq n < \infty$, is f-admissible, then $\lim_{n\to\infty} |I(s_0 \ldots s_n)| = 0$.*

(iii) *Let* $s=\{s_0 \dots s_n, s_{n+1}\} < s' = \{s_0 \dots s_n, s'_{n+1}\}$. $I(s_0 \dots s_{n+1})$, $I(s_0 \dots s'_{n+1})$ *are both sub-intervals of* $I(s_0 \dots s_n)$, *the first sub-interval appearing to the left of the second.*

Proof. (i), (iii) are proved by induction on n. They follow readily from the fact that f is a one-to-one, continuous, sense preserving map on each I_i, and we omit the details of the proof. (ii) follows from Theorem 6.3.

Lemma 8.2 has the

Corollary. *Let* $s_0, \dots, s_n, s_{n+1}$ *be* f*-admissible.* $I(s_0 \dots s_n)$, $I(s_0 \dots s_n s_{n+1})$ *have the same right end point if and only if* $s_n = i_1$, $s_{n+1} = (\alpha(i))_1$ *or* $s_n = i_2$, $s_{n+1} = (\beta(i))_2$ *for some* $1 \le i \le 8g-4$.

Theorem 8.1. $\overline{\overline{\Omega}}(f) = \overline{\overline{\Omega}}_a(f)$ *and* $f = \phi^{-1} \circ \sigma \circ \phi$.

Proof. We prove (P) for f. (i) Let $f^n(\xi) \in I_{s_n}$, $0 \le n < \infty$, which is the same as $\xi \in I(s_0 \dots s_n)$, $0 \le n < \infty$. Since $f^{n+1}(\xi) = f(f^n \xi) \in f(I_{s_n}) \cap I_{s_{n+1}}$, $\{s_n\}$ is f-admissible. We show that the $\mathscr{L}_\infty$-blocks of $\overline{\overline{\Omega}}(f)$ do not occur in $\{s_n\}$, so that $\{s_n\} \in \overline{\overline{\Omega}}_a(f)$. By Lemma 8.2, $\{I(s_0 \dots s_n)\}$, $0 \le n < \infty$, satisfies the hypotheses of Lemma 8.1. Since $\xi \in \bigcap_{n=0}^{\infty} I(s_0 \dots s_n)$, we conclude from Lemma 8.1 that these intervals do not have a common right end point from some n on. By the corollary to Lemma 8.2, this means that the $\mathscr{L}_\infty$-blocks of $\overline{\overline{\Omega}}_a(f)$ do not occur in $\{s_n\}$.

(ii) Let $\{s_n\} \in \overline{\overline{\Omega}}_a(f)$. Then $\{s_n\}$ is f-admissible. By Lemma 8.2, $\{I(s_0 \dots s_n)\}$, $0 \le n < \infty$, satisfies the hypotheses of Lemma 8.1. $\{s_n\}$ does not contain any of the $\mathscr{L}_\infty$-blocks of $\overline{\overline{\Omega}}_a(f)$. By the corollary to Lemma 8.2, this means that the intervals $\{I(s_0 \dots s_n)\}$ do not have a common right end point from some n on. We conclude from Lemma 8.1 that $\bigcap_{n=0}^{\infty} f^{-n}(I_{s_n}) = \bigcap_{n=0}^{\infty} I(s_0 \dots s_n)$ consists of a single point ξ.

Next, we consider $\overline{\overline{\Omega}}(T_R)$. In Figures 8.1 (i), (ii) on p. 280, we sketch R_{i_1}, R_{i_2} and their T_R-images R'_{i_1}, R'_{i_2}

Let $L_{i_k} = \pi_\eta(R'_{i_k})$, $k = 1, 2$. From Figure 8.1,

$$L_{i_1} = (b_{\sigma(i)-1}, a_{\sigma(i)+1}], \qquad L_{i_2} = (b_{\sigma(i)-1}, b_{\sigma(i)}] \tag{8.2}$$

From $R_{i_k} = T_R^{-1}(R'_{i_k})$ and $T_R^{-1}(\xi, \eta) = (\cdot, g(\eta))$ we get $g(L_{i_k}) = \pi_\eta(R_{i_k})$.

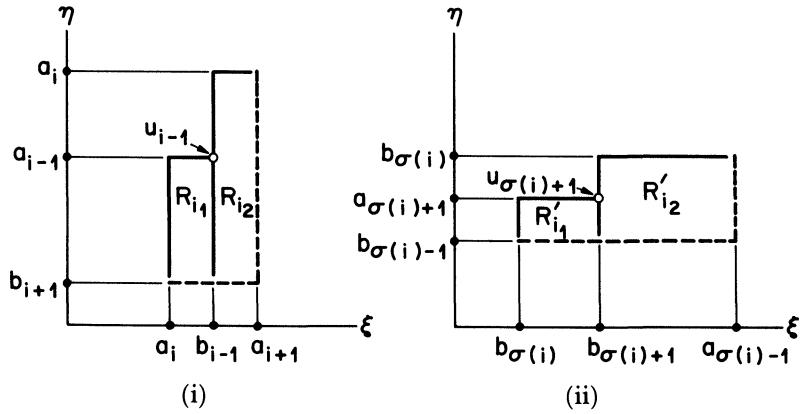


FIGURE 8.1. R_{i_1}, R_{i_2} and R'_{i_1}, R'_{i_2}.

We list the sets R'_s meeting R_t, where $s, t \in \{1_1, 1_2, \ldots, (8g-4)_2\}$. As remarked earlier, $R_t \cap R'_s \neq \varnothing$ if and only if $s \to t$ satisfies the transition rules of $\overline{\overline{\Omega}}_a(T_R)$. From these we get

R_{i_1} is met by: $R'_{(\sigma(i+2))_2} R'_{(\sigma(i+3))_2}, \cdots, R'_{(\sigma(i-3))_2}, R'_{(\sigma(i-2))_1}$

R_{i_2} is met by: $R'_{(\sigma(i+2))_2} R'_{(\sigma(i+3))_2}, \cdots, R'_{(\sigma(i-3))_2}, R'_{(\sigma(i-2))_2}, R'_{\sigma(i-1))_1}$

Applying π_η to the sets, we get

(8.3)
$$g(L_{i_1}) = L_{\sigma(i+2)_2} \cup L_{\sigma(i+3)_2} \cup \cdots \cup L_{(\sigma(i-3))_2} \cup L_{(\sigma(i-2))_1}$$
$$g(L_{i_2}) = L_{\sigma(i+2)_2} \cup L_{\sigma(i+3)_2} \cup \cdots \cup L_{(\sigma(i-3))_2} \cup L_{(\sigma(i-2))_2} \cup L_{(\sigma(i-1))_1}$$

Formulas (8.2) show that the intervals on the right side of (8.3) are disjoint and written in the order in which they are encountered when $g(L_{i_k})$ is traversed counter-clockwise. In Figure 8.2, we sketch, for fixed t, the intersections $R_t \cap R'_s$ which are labeled by s'.

Figure 8.2 puts into evidence

Lemma 8.3. *Let $s \to t$ satisfy the transition rules of $\overline{\overline{\Omega}}(T_R)$. Then*

(8.4)
$$R_t \cap R'_s = \begin{cases} I_t \times L_s - \{u_{i-1}\}, & \text{if } (t, s) = (i_2, (\alpha(i))_2) \text{ for some } i \\ I_t \times L_s & \text{otherwise,} \end{cases}$$

and

(8.5)
$$R_t \cap R'_s = R_t \cap \pi_\eta^{-1}(L_s).$$

In (8.4) we have used $\alpha(i) = \sigma(i-2)$ which is derivable from (3.1). We use later $\beta(i) = \sigma(i+2)$, also derivable from (3.1).

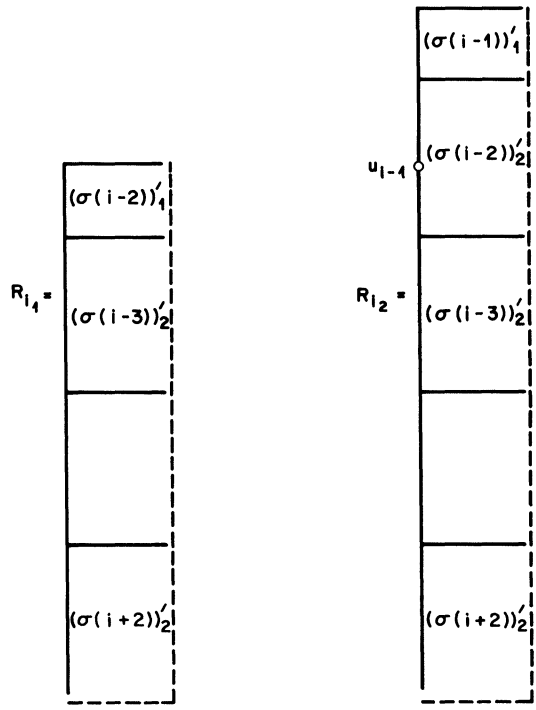


FIGURE 8.2. The sets $R_t \cup R'_s$.

To study T_R-expansions, we find it convenient to first introduce the g-expansion (recall g is the right Bowen-Series map) with respect to the class of intervals $\mathscr{L} = \{L_{1_1}, L_{1_2}, \ldots, L_{(8g-4)_2}\}$. We are faced with the difficulty that $\mathscr{L}$ is not a partition of $\partial\mathbb{D}$ as $L_{i_1} \subset L_{i_2}$. The usual definition of the g-expansion $\{s_n\}$ given by $g^n(\eta) \in L_{s_n}$, $0 \leq n < \infty$, is thus ambiguous and must be modified. In conformity with (8.3), we introduce the transition rules

(8.6)

$$\mathscr{T}^g\colon \begin{aligned} i_1 &\to (\sigma(i+2))_2, (\sigma(i+3))_2, \ldots, (\sigma(i-3))_2, (\sigma(i-2))_1 \\ i_1 &\to (\sigma(i+2))_2, (\sigma(i+3))_2, \ldots, \\ &\qquad (\sigma(i-3))_2, (\sigma(i-2))_2, (\sigma(i-1))_1. \end{aligned}$$

The transition rules $\mathscr{T}^g$ are obtained from the ones for $\overline{\overline{\Omega}}(f)$ (which are the same as the rules for $\overline{\overline{\Omega}}(T_R)$) by reversing the arrow direction.

Definition 8.2. (i) $\{s_n\}$, $0 \leq n < \infty$, is g-admissible if and only

if $s_i \to s_{i+1}$ satisfies the $\mathscr{T}^g$-rules for $0 \le n < \infty$.

(ii) Let $g^n(\eta) \in L_{s_n}$, $0 \le n < \infty$, where $\{s_n\}$ is g-admissible. $\{s_n\}$ is a *modified* g-expansion of η.

The notion of g-admissibility for finite sequences is defined in a similar manner. In view of the relation betwen $\mathscr{T}^g$ and the transition rules for $\overline{\overline{\Omega}}(f)$, $\{s_n\}$ is g-admissible if and only if the reversed sequence is f-admissible ($= T_R$-admissible). The modified g-expansion of η is unambiguously defined once s_0 is prescribed. If $\eta \in L_{i_1}$, then s_0 can be chosen as i_1 or i_2 and η has two modified g-expansions. If $\eta \in L_{i_2} - L_{i_1}$, then $s_0 = i_2$ and η has one modified g-expansion.

Examination of the proof of Theorem 8.1 shows that it remains valid for modified g-expansions. We therefore obtain

Theorem 8.2. *Let* $\overline{\overline{\Omega}}(g)$ *be the space of modified* g*-expansions.* $\overline{\overline{\Omega}}(g) = \overline{\overline{\Omega}}_a(g)$ *where* $\overline{\overline{\Omega}}_a(g)$ *is the space of sequences* $\{s_n\}$, $0 \le n < \infty$, *satisfying the* $\mathscr{T}^g$*-rules and not containing the infinite blocks*:

$\mathscr{L}_\infty : i_2, (\beta(i))_2, (\beta^2(i))_2, \ldots$.

Theorem 8.3 relates f- and modified g-expansions to T_R-expansions. For any sequence $\{s_n\}$, $-\infty < n < \infty$, let $\{s_n\}^+ = \{s_n\}$, $0 \le n < \infty$, and $\{s_n\}^- = \{s_{-n-1}\}$, $0 \le n < \infty$.

Theorem 8.3. *Let* $\xi, \eta \in \partial\mathbb{D}$. *Then* $T_R^n(\xi, \eta) \in R_{s_n}$, $-\infty < n < \infty$, *if and only if*: (i) $(\xi, \eta) \ne u_i$, $1 \le i \le 8g - 4$, (ii) $s_{-1}s_0$ *is* T_R*-admissible,* (iii) $\{s_n\}^+$ *and* $\{s_n\}^-$ *are respectively* f*- and modified* g*-expansions of* ξ *and* η.

We make the following remarks concerning the above conditions:

(1) Condition (ii) is equivalent to saying that the two expansions $\{s_n\}^+$, $\{s_n\}^-$ yield a T_R-admissible sequence $\{s_n\}$ when spliced together.

(2) Theorem 8.3 becomes false without condition (i). For let $\xi = b_{i-1} \in I_{i_2}$, $\eta = a_{i-1} \in L_{(\alpha(i))_2}$, and choose $\{s_n\}^+$, $\{s_n\}^-$ as in (iii). Then (ii) also holds as $s_{-1}s_0 = (\alpha(i))_2$, i_2 is T_R-admissible, yet $u_{i-1} = (\xi, \eta) \notin R$.

(3) Let η have two modified g-expansions. At most one of these will satisfy (ii) for given ξ. Thus the choice of g-expansion for η is unambiguous.

Proof. Suppose that $T_R^n(\xi, \eta) \in R_{s_n}$, $-\infty < n < \infty$. Since $T_R^{n+1}(\xi, \eta) \in R'_{s_n} \cap R_{s_{n+1}}$ for all n, $\{s_n\}$ is T_R-admissible. Hence (i), (ii) hold. For $n \geq 0$, $T_R^n(\xi, \eta) = (f^n(\xi), \cdot) \in R_{s_n}$. Hence $f^n(\xi) \in \pi_\xi(R_{s_n}) = I_{s_n}$, and $\{s_n\}^+$ is the f-expansion of ξ. $T_R^{-(n+1)}(\xi, \eta) \in R_{s_{-(n+1)}}$ is equivalent to $T_R^{-n}(\xi, \eta) \in R'_{s_{-(n+1)}}$. For $n \geq 0$, $T_R^{-n}(\xi, \eta) = (\cdot, g^n(\eta))$. Hence $g^n(\eta) \in \pi_\eta(R'_{s_{-(n+1)}}) = L_{s_{(n+1)}}$. Furthermore, $\{s_{-(n+1)}\}$, $0 \leq n < \infty$, is g-admissible, as $\{s_n\}$, $-\infty < n < \infty$, is T_R-admissible. Hence, $\{s_n\}^+ = \{s_{-(n+1)}\}$, $0 \leq n < \infty$, is a modified g-expansion of η.

Conversely, let (i)–(iii) hold. Then $(\xi, \eta) \neq u_i$, $1 \leq i \leq 8g-4$, $s_{-1}s_0$ is T_R-admissible, and $(\xi, \eta) \in I_{s_0} \times L_{s_{-1}}$. Hence, by (8.4), $(\xi, \eta) \in R_{s_0} \cap R'_{s_{-1}} \subset R$. For $n \geq 0$, $f^n(\xi) \in I_{s_n}$. Hence

$$T_R^n(\xi, \eta) = (f^n(\xi), \cdot) \in \pi_\xi^{-1}(I_{s_n}) = R_{s_n}, \qquad n \geq 0. \tag{8.7}$$

We prove by induction that $T_R^{-n}(\xi, \eta) \in R_{s_{-n}}$, $n \geq 0$. By (8.7) this holds for $n = 0$. Assume it holds for n. We must show that $T_R^{-(n+1)}(\xi, \eta) \in R_{s_{-(n+1)}}$, which is equivalent to $T_R^{-n}(\xi, \eta) \in R'_{s_{-(n+1)}}$. Since $g^n(\eta) \in L_{s_{-(n+1)}}$, we have

$$T_R^{-n}(\xi, \eta) = (\cdot, g^n(\eta)) \in \pi_\eta^{-1}(L_{s_{-(n+1)}}). \tag{8.8}$$

Since $s_{-(n+1)}, s_{-n}$ is T_R-admissible, we conclude from (8.8) and (8.5) that

$$T_R^{-n}(\xi, \eta) \in R_{s_{-n}} \cap \pi_\eta^{-1}(L_{s_{-(n+1)}}) = R_{s_{-n}} \cap R'_{s_{-(n+1)}} \subset R'_{s_{-(n+1)}}. \tag{8.9}$$

Theorem 8.4. *$\overline{\overline{\Omega}}(T_R) = \overline{\overline{\Omega}}_a(T_R)$ and $T_R = \phi^{-1} \circ \sigma \circ \phi$.*

Proof. Again we prove (P). (i) Let $T_R^n(\xi, \eta) \in R_{s_n}$, $-\infty < n < \infty$. Then $\{s_n\}$ is T_R-admissible. We show that the $\mathscr{L}_\infty$-blocks of $\overline{\overline{\Omega}}_a(T_R)$ do not appear in $\{s_n\}$, so that $\{s_n\} \in \overline{\overline{\Omega}}_a(T_R)$. By Theorem 8.3, $\{s_n\}^+ \in \overline{\overline{\Omega}}_a(f)$, $\{s_n\}^- \in \overline{\overline{\Omega}}_a(g)$. The $\mathscr{L}_\infty$-blocks of $\overline{\overline{\Omega}}_a(f)$, $\overline{\overline{\Omega}}_a(g)$ then show that the first three $\mathscr{L}_\infty$-blocks for $\overline{\overline{\Omega}}_a(T_R)$ do not appear in $\{s_n\}$. We show that the fourth $\mathscr{L}_\infty$-block of $\overline{\overline{\Omega}}(T_R)$ also does not appear in $\{s_n\}$. Suppose it does. Then $s_{-n} = (\alpha^n(L))_2$ for $0 \leq n < \infty$ and some i. It is readily

checked that

(8.10)

$$i_2, (\alpha(i))_2, (\alpha^2(i))_2, \cdots = f\text{-expansion of } b_{i-1},$$

$$(\alpha(i))_2, (\alpha^2(i))_2, (\alpha^3(i))_2, \ldots = \text{modified } g\text{-expansion of } a_{i-1}.$$

Since $(\alpha(i))_2$, i_2 is T_R-admissible, we conclude from Theorem 8.2 that $(\xi, \eta) = (b_{i-1}, a_{i-1}) = u_{i-1}$. But $u_{i-1} \notin R$, a contradiction. Hence the fourth $\mathscr{L}_\infty$-block of $\overline{\overline{\Omega}}_a(T_R)$ does not appear in $\{s_n\}$.

(ii) Let $\{s_n\} \in \overline{\overline{\Omega}}_a(T_R)$. Suppose $(\xi, \eta) \in \cap T_R^{-1} R_{s_n}$. By Theorems 8.1, 8.2, $\{s_n\}^+$ and $\{s_n\}^-$ are respectively the f- and modified g-expansion of ξ and η. Thus $\cap T_R^{-n} R_{s_n}$ consists of at most one point.

From the admissibility rules for $\overline{\overline{\Omega}}_a(T_R)$, $\overline{\overline{\Omega}}_a(f)$, $\overline{\overline{\Omega}}_a(g)$, we conclude that $\{s_n\}^+ \in \overline{\overline{\Omega}}_a(f)$, $\{s_n\}^- \in \overline{\overline{\Omega}}_a(g)$. Let ξ be the point with f-expansion $\{s_n\}^+$ and η the point with g-expansion $\{s_n\}^-$. We show that (ξ, η) satisfies conditions (i)–(iii) of Theorem 8.3, so that (ξ, η) is the desired point. $\{s_n\} \neq \{(\alpha^n(i))_2\}$, $1 \leq i \leq 8g-4$, as $\{s_n\} \in \overline{\overline{\Omega}}_a(T_R)$ and $\{(\alpha^n(i))_2\}$ is a forbidden block for $\overline{\overline{\Omega}}(T_R)$. We conclude from (8.10) that $(\xi, \eta) \neq u_i$, $1 \leq i \leq 8g-4$, so that (i) holds. (ii) follows from $\{s_n\}$ being T_R-admissible, and (iii) from the choice of ξ, η.

We use the description of $\overline{\overline{\Omega}}(T_R)$ to obtain that of $\overline{\Omega}(T_R)$. To distinguish between different symbolic systems, we use from now on $\{\overline{s}_n\}$, $\{\overline{\overline{s}}_n\}$ to denote sequences from the respective alphabets $\{1, 2, \ldots, 8g-4\}$, $\{1_1, 1_2, \ldots, (8g-4)_2\}$. Thus the sequences denoted earlier by $\{s_n\}$ are now denoted by $\{\overline{s}_n\}$.

For $(\xi, \eta) \in R$, let $T_R^n(\xi, \eta) \in R_{\overline{\overline{s}}_n} \subset R_{\overline{s}_m}$, $-\infty < n < \infty$. Since $R_i = R_{i_1} \cup R_{i_2}$, $\{\overline{s}_n\}$ is obtained from $\{\overline{\overline{s}}_n\}$ by dropping subscripts, i.e.

$$\overline{s}_n = i, \quad \text{if } \overline{\overline{s}}_n = i_1, i_2. \tag{8.11}$$

Conversely, we obtain $\{\overline{\overline{s}}_n\}$ from $\{\overline{s}_n\}$ by the following

Theorem 8.5.

(8.12)

$$\overline{\overline{s}}_n = \begin{cases} (\overline{s}_n)_1, & \text{if } \overline{s}_n, \ldots, \overline{s}_{n+m}, \overline{s}_{n+m+1} = i, \alpha(i), \ldots, \\ & \alpha^m(i), \vartheta\alpha^m(i) \text{ for some } m \geq 0 \text{ and } i, \\ (\overline{s}_n)_2, & \text{otherwise.} \end{cases}$$

Proof. The transition rules of $\overline{\overline{\Omega}}(T_R)$ imply:

(8.13)

if $\bar{s}_n, \bar{s}_{n+1} = i, \vartheta(i)$,	then $\bar{\bar{s}}_n = i_1$,
if $\bar{s}_n, \bar{s}_{n+1} = i, j(j \neq \vartheta(i), \alpha(i))$,	then $\bar{\bar{s}}_n = i_2$,
if $\bar{s}_n, \bar{s}_{n+1} = i, \alpha(i)$,	then $\bar{\bar{s}}_n, \bar{\bar{s}}_{n+1} = i_1, (\alpha(i))_1$ or $i_2, (\alpha(i))_2$.

Suppose that $\bar{s}_n, \ldots, \bar{s}_{n+m}, \bar{s}_{n+m+1} = i, \alpha(i), \ldots, \alpha^m(i), \vartheta\alpha^m(i)$. Repeated application of (8.13) gives $\bar{s}_{n+m} = (\alpha^m(i))_1$, $\bar{s}_{n+m-1} = (\alpha^{m-1}(i))_1, \ldots, \bar{s}_n = i_1$. The remaining part of Theorem 8.5 is broken up into two cases. If $\bar{s}_n, \ldots, \bar{s}_{n+m}, \bar{s}_{n+m-1} = i, \alpha(i), \ldots, \alpha^m(i), j(j \neq \alpha^{m+1}(i), \vartheta\alpha^m(i))$ for some $m \geq 0$ and i, then the same reasoning as before gives $\bar{s}_n = i_2$. Otherwise, we have $\bar{s}_n, \bar{s}_{n+1}, \bar{s}_{n+2}, \ldots = i, \alpha(i), \alpha^2(i), \ldots$ for some i. If $\bar{\bar{s}}_n = i_1$, then repeated application of (8.13) gives $\bar{\bar{s}}_n, \bar{\bar{s}}_{n+1}, \bar{\bar{s}}_{n+2}, \ldots = i_1, (\alpha(i))_1, (\alpha^2(i))_1, \ldots$ which is a forbidden block for $\overline{\overline{\Omega}}_a(T_R)$. Hence $\bar{\bar{s}}_n = i_2$.

Theorem 8.6. *$\overline{\Omega}(T_R) = \overline{\Omega}_a(T_R)$ and $T_R = \phi^{-1} \circ \sigma \circ \phi$.*

Proof. (i) Let $T_R^n(\xi, \eta) \in R_{\bar{\bar{s}}_n}$, $-\infty < n < \infty$. $\{\bar{s}_n\}$ is obtained from $\{\bar{\bar{s}}_n\}$ by dropping subscripts. It follows from the transition rules for $\overline{\overline{\Omega}}(T_R)$ that $i, \sigma(i)$ and $i, \rho(i)$ do not appear in $\{\bar{s}_n\}$. We show that the remaining $\mathscr{L}_0$-blocks of $\overline{\Omega}(T_R)$ do not appear in $\{\bar{s}_n\}$. Suppose to the contrary that

$$\bar{s}_{n-1}, \bar{s}_n, \bar{s}_{n+1}, \ldots, \bar{s}_{n+m}, \bar{s}_{n+m+1} = i, \vartheta(i), \alpha\vartheta(i), \ldots, \alpha^m\vartheta(i), \vartheta\alpha^m\vartheta(i) \tag{8.14}$$

for some n, m, i. By Theorem 8.5, $\bar{s}_{n-1}, \bar{s}_n = i_1, (\vartheta(i))_1$ which is a forbidden block for $\overline{\overline{\Omega}}(T_R)$. Hence (8.14) does not appear in $\{\bar{s}_n\}$.

We show next that the $\mathscr{L}_\infty$-blocks of $\overline{\Omega}(T_R)$ do not appear in $\{\bar{s}_n\}$, so that $\{\bar{s}_n\} \in \overline{\Omega}_a(T_R)$. Suppose $i, \beta(i), \ldots, \beta^m(i) \ldots$ appears in $\{\bar{s}_n\}$. By Theorem 8.5, the forbidden block $i_2, (\beta(i))_2, \ldots, (\beta^m(i))_2, \ldots$ then appears in $\{\bar{\bar{s}}_n\}$, a contradiction. The remaining $\mathscr{L}_\infty$-blocks of $\overline{\Omega}(T_R)$ are disposed of in a similar manner.

(ii) Let $\{\bar{s}_n\} \in \Omega_a(T_R)$. Define $\{\bar{\bar{s}}_n\}$ by (8.12). We show that $\{\bar{\bar{s}}_n\} \in \overline{\overline{\Omega}}_a(T_R)$. Hence $\bigcap T_R^{-n}(R_{\bar{s}_n}) = \bigcap T_R^{-n}(R_{\bar{\bar{s}}_n})$ consists of a single point.

We must prove statments A and B given below.

A. $\{\bar{\bar{s}}_n\}$ satisfies the transition rules of $\bar{\bar{\Omega}}(T_R)$. We consider several cases:

(A1) $\bar{s}_n, \bar{s}_{n+1}, \dots, \bar{s}_{n+m}, \bar{s}_{n+m+1} = i, \alpha(i), \dots, \alpha^m(i), \vartheta\alpha^m(i)$. If $m > 0$, then $\bar{\bar{s}}_n, \bar{\bar{s}}_{n+1} = i_1, (\alpha(i))_1$ which is T_R-admissible. If $m = 0$, then $\bar{\bar{s}}_n = i_1$. Since $\{\bar{s}_n\} \in \bar{\Omega}_a(T_R)$, $\bar{s}_{n+1} = \vartheta(i)$ cannot be followed by the block $\alpha\vartheta(i), \dots, \alpha^p\vartheta(i), \vartheta\alpha^p\vartheta(i)$ for some p. Hence $\bar{\bar{s}}_{n+1} = (\vartheta(i))_2$, and $i_1, (\vartheta(i))_2$ is T_R-admissible.

(A2) $\bar{s}_n, \bar{s}_{n+1}, \dots, \bar{s}_{n+m}, \bar{s}_{n+m+1} = i, \alpha(i), \dots, \alpha^m(i), j;\ j \neq \alpha^{m+1}(i), \vartheta\alpha^m(i)$. If $m > 0$, then $\bar{\bar{s}}_n, \bar{\bar{s}}_{n+1} = i_2, (\alpha(i))_2$, which is T_R-admissible. If $m = 0$, then $\bar{\bar{s}}_n, \bar{\bar{s}}_{n+1} = i_2, j_1$ or i_2, j_2, both of which are T_R-admissible.

(A3) $\bar{s}_n, \bar{s}_{n+1}, \dots, \bar{s}_{n+m}, \dots = i, \alpha(i), \dots, \alpha^m(i), \dots$. Then $\bar{\bar{s}}_n, \bar{\bar{s}}_{n+1} = i_2, (\alpha(i))_2$ which is T_R-admissible.

B. $\{\bar{\bar{s}}_n\}$ does not contain the $\mathscr{L}_\infty$-blocks of $\bar{\bar{\Omega}}(T_R)$.

Dropping indices, the last three $\mathscr{L}_\infty$-blocks of $\bar{\bar{\Omega}}(T_R)$ become the $\mathscr{L}_\infty$-blocks of $\bar{\Omega}(T_R)$. Hence these three blocks do not appear in $\{\bar{\bar{s}}_n\}$, as the corresponding ones do not appear in $\{\bar{s}_n\}$. Suppose that the first $\mathscr{L}_\infty$-block of $\bar{\bar{\Omega}}(T_R)$ appears in $\{\bar{\bar{s}}_n\}$, i.e., $\bar{\bar{s}}_{n+m} = (\alpha^m(i))_2$, $0 \leq m < \infty$, for some n, m, i. Then $\bar{s}_{n+m} = \alpha^m(i)$, $0 \leq m < \infty$. The conversion rules of (8.12) give $\bar{\bar{s}}_{n+m} = (\alpha^m(i))_2$, $0 \leq m < \infty$, a contradiction. Hence all $\mathscr{L}_\infty$-blocks of $\bar{\bar{\Omega}}(T_R)$ do not appear in $\{\bar{\bar{s}}_n\}$.

Let $\bar{\Omega}(g)$ be the space of g-expansions with respect to the partition $\{L_{1_2}, \dots, L_{(8g-4)_2}\}$. We describe the spaces $\bar{\Omega}(f)$, $\bar{\Omega}(g)$ and their relation to $\bar{\Omega}(T_R)$.

Definition 8.3. $\Omega_a(g)$ is the space of sequences $\{\bar{s}_n\}$, $0 \leq n < \infty$, with the following omitted blocks.

$$\mathscr{L}_0\colon [i, \sigma(i)]; [i, \rho^{-1}(i)];$$
$$[i, \vartheta^{-1}(i), \alpha\vartheta^{-1}(i), \dots, \alpha^k\vartheta^{-1}(i), \vartheta^{-1}\alpha^k\vartheta^{-1}(i)], 0 \leq k < \infty.$$
$$\mathscr{L}_\infty\colon [i, \beta(i), \beta^2(i), \dots].$$

Theorem 8.7. (i) $\overline{\Omega}(f) = \overline{\Omega}_a(f)$,

(ii) $\overline{\Omega}(g) = \overline{\Omega}_a(g)$,

(iii) *Let* $T_R^n(\xi, \eta) \in R_{\bar{s}_n}$, $-\infty < n < \infty$. $\{\bar{s}_n\}^+$ *is the* f*-expansion of* ξ *with respect to* $\{I_1, \ldots, I_{8g-4}\}$, *and* η *is the* g*-expansion of* η *with respect to* $\{L_{1_2}, \ldots, L_{(8g-4)_2}\}$.

We omit the proof of Theorem 8.7, which is similar to that of Theorem 8.6.

We conclude this section with some remarks comparing the above Markovian and sofic expansions. Sofic admissibility rules are more difficult than Markovian ones which have the virtue of being one-step rules. However, a price must be paid in the Markovian case. In order to get a T_R-expansion of $(\xi, \eta) \in R$ by splicing an f-expansion of ξ with a g-expansion of η, we cannot use the ordinary g-expansion and must modify the notion. The sofic case is easier in this respect as the ordinary g-expansion will do. The Markovian case has an additional difficulty over the sofic one. To get the T_R-expansion of $(\xi, \eta) \in R$, we must resolve the ambiguity concerning the modified g-expansion of η before attaching it to the f-expansion of ξ (it is removed by Theorem 8.3(ii)).

9. Geometric interpretation of symbolic sequences

Let $\{C_i\}$ and $\{R_i\}$ be the respective partitions of C and R introduced in §§4,5. For $(\xi, \eta) \in C$, let $T_C^n(\xi, \eta) \in C_{s_n}$, $-\infty < n < \infty$, and for $(\bar{\xi}, \bar{\eta}) \in R$, let $T_R^n(\bar{\xi}, \bar{\eta}) \in R_{\bar{s}_n}$, $-\infty < n < \infty$. $\{s_n\}$ and $\{\bar{s}_n\}$ are referred to respectively as *curvilinear* and *rectilinear sequences.* We obtain in the present section a geometric interpretation to these sequences.

We begin with $\{s_n\}$. Label the sides of the edges of the net $\mathbf{N}$ with the symbols $\{1, 2, \ldots, 8g-4\}$, according to the prescription of §3 (see the proof of Theorem 3.2), and assume throughout that the geodesics under discussion do not lie in $\mathbf{N}$.

Definition 9.1. (i) Let γ be a geodesic passing successively through the fundamental regions $\mathbf{F}_n = \mathbf{F}_n(\gamma)$, $-\infty < n < \infty$, with $\mathbf{F} = \mathbf{F}_0$. Assume that γ leaves $\mathbf{F}_n$ through the interior of an edge whose $\mathbf{F}_n$-side is labeled $s_n(\gamma)$. $\{s_n(\gamma)\}$ is called the *cutting sequence of* γ (Figure 9.1 (i) on p. 288).

(ii) The above definition becomes ambiguous if γ leaves a fundamental region through a vertex. In that case, let γ' be a geodesic slightly to the left of γ (Figure 9.1 (ii) on p. 288) which meets

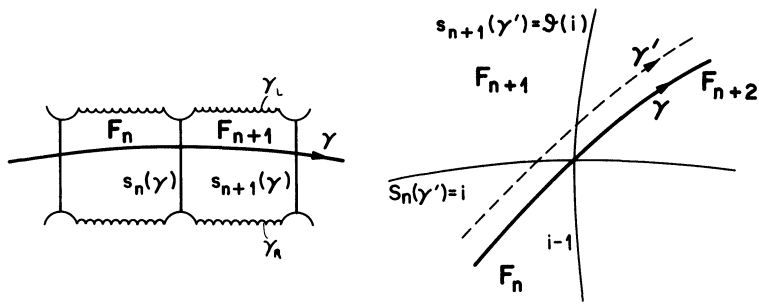


FIGURE 9.1. Cutting sequences.

successively $\mathbf{F}_n$, $\mathbf{F}_{n+1}$, $\mathbf{F}_{n+2}$. Then γ is also said to meet successively $\mathbf{F}_n$, $\mathbf{F}_{n+1}$, $\mathbf{F}_{n+2}$, and $s_n(\gamma)$, $s_{n+1}(\gamma)$ are defined to be respectively $s_n(\gamma')$, $s_{n+1}(\gamma')$.

We remark that if the vertex in (ii) is the intersection of the sides of $\mathbf{F}_n$ with labels $i-1$ and i, then $s_n(\gamma) = i$ and $s_{n+1}(\gamma) = \vartheta(i)$, as made evident by Figure 9.1 (ii).

We observe that the boundary of $\bigcup \mathbf{F}_n$ consists of two simple curves γ_L and γ_R, lying respectively to left and right of γ. We orient these curves so as to agree with the orientation of γ.

Theorem 9.1. *Let (ξ, η) be the forward and backward end points of the geodesic $\gamma = \gamma(\xi, \eta)$. Then $\{s_n(\gamma)\} = \{s_n(\xi, \eta)\}$ for $(\xi, \eta) \in C$ and $-\infty < n < \infty$.*

Theorem 9.1 is evident from the definitions of $\{s_n(\gamma)\}$ and $\{s_n(\xi, \eta)\}$ except for the case when γ passes through a vertex of some $\mathbf{F}_n$. The following is a formal proof covering all cases.

Proof. Suppose that γ passes through the interior of $\mathbf{F}_n$. Let u_n be the unit tangent vector to γ at the point where it leaves $\mathbf{F}_n$. If γ leaves the interior of $\mathbf{F}_n$ through a vertex, let u_{n+1} be the unit tangent vector to γ at the point obtained by moving distance ε away from the vertex (ε is the number appearing in the definition of $\mathbf{C}_{i1}$ in §4). We have thus chosen u_n so that $\overline{\pi}(u_n) = T_C^n(\xi, \eta)$ for $-\infty < n < \infty$.

Assume first that γ leaves the interior of $\mathbf{F}_n$ through a non-vertex point on the side labeled i. Then $s_n(\gamma) = i$. The element of Γ, which maps $\mathbf{F}_n$ to $\mathbf{F}$, maps u_n to a unit vector

based at a nonvertex point of s_i and pointing out of $\mathbf{F}$. Hence $T_C^n(\xi, \eta) = \overline{\pi}(u_n) \in C_i^p \subset C_i$, so that $s_n(\xi, \eta) = i = s_n(\gamma)$.

Assume next that γ leaves the interior of $\mathbf{F}_n$ through the vertex which is the intersection of the sides with labels $i-1, i$. Then, as remarked above, $s_n(\gamma) = i$ and $s_{n+1} = \vartheta(i)$. The elements of Γ, which map $\mathbf{F}_n$, $\mathbf{F}_{n+1}$ to $\mathbf{F}$, respectively map u_n, u_{n+1} to vectors u depicted in Figures 4.1 (ii),(iii), except that in the latter case p_i must be replaced by $p_{\vartheta(i)}$. Hence $T_C^n(\xi, \eta) = \overline{\pi}(u_n) \in C_{i0} \subset C_i$, $T_C^{n+1}(\xi, \eta) = \overline{\pi}(u_{n+1}) \in C_{\vartheta(i),1} \subset C_{\vartheta(i)}$. We conclude that $s_n(\xi, \eta) = i = s_n(\gamma)$, $s_{n+1}(\xi, \eta) = \vartheta(i) = s_{n+1}(\gamma)$.

Let $\Omega(T_C)$ be the space of sequences $\{s_n(\xi, \eta)\}$, $(\xi, \eta) \in C$. We obtain from Theorem 9.1 the

Corollary. *The mapping* $(\xi, \eta) \to \{s_n(\xi, \eta)\}$ *is one-to-one from* C *onto* $\Omega(T_C)$.

Proof. Let $\{s_n(\xi_1, \eta_1)\} = \{s_n(\xi_2, \eta_2)\}$, where $(\xi_i, \eta_i) \in C$, $i = 1, 2$. By Theorem 9.1, $\{s_n(\gamma_1)\} = \{s_n(\gamma_2)\}$, where $\gamma_i = \gamma(\xi_i, \eta_i)$. Let $d(\cdot, \cdot)$ denote hyperbolic distance. γ_1, γ_2 have identical cutting sequences, and hence $\mathbf{F}_n(\gamma_1) = \mathbf{F}_n(\gamma_2)$, $-\infty < n < \infty$. It follows that γ_1, γ_2 can be parametrized so that $\gamma_i = \gamma_i(t)$, $-\infty < t < \infty$, and $d(\gamma_1(t), \gamma_2(t))$ is uniformly bounded for $-\infty < t < \infty$. We conclude from Theorem 2.1 (ii) that γ_1, γ_2 have identical end points, i.e. $\xi_1 = \xi_2$ and $\eta_1 = \eta_2$.

Next we consider $\{\overline{s_n}\}$. To motivate the work to follow, we make a brief digression and discuss Morse's method for coding geodesics [Mo].

[Mo] considers only the special class of surfaces of genus $g \geq 2$ defined by a regular $4g$-sided fundamental region with interior angles $\pi/2g$, but the method used can be adapted readily to the fundamental regions treated in this paper. The method produces a geometric modification of the cutting sequence $\{s_n(\gamma)\}$.

Let $\{\mathbf{F}_n(\gamma)\}$, $-\infty < n < \infty$, be the sequence of fundamental regions of Definition 9.1. In general some of the interior angles of $\bigcup \mathbf{F}_n(\gamma)$ along the left boundary γ_L will be obtuse, in which case we say that γ_L is not *convex*. As illustrated in Figure 9.5, we add to $\bigcup \mathbf{F}_n(\gamma)$ fundamental regions bordering on the left side of γ_L[4] so that the left boundary of the resulting set R is (i) convex and (ii) does not contain infinitely long geodesic arcs (condition

[4]Morse chooses the right side of the right boundary γ_R; we chose the left side of γ_L to reconcile our work with his.

(ii) is inserted to guarantee that the modification procedure be injective on cutting sequences). Let $\overline{\gamma}$ be a path in R passing through the successive fundamental regions bordering on the left side of R. A portion of $\overline{\gamma}$ is illustrated in Figure 9.5, where it is denoted by $\overline{\gamma_B}$. Just as for γ, we define the cutting sequence of $\overline{\gamma}$. We call it the *modified cutting sequence* and denote it by $\{\overline{s}_n(\overline{\gamma})\}$ (a special convention is required for singling out the 0th term $\overline{s}_0(\overline{\gamma})$: this is done right after Theorem 9.8). By geometric reasoning using the convexity of the left R-boundary, Morse shows that the modified cutting sequences can be described by a list of forbidden finite and infinite blocks [Mo, p. 77]. Surprisingly to us, Morse's list of forbidden blocks, or rather its analogue as applied to our surfaces, coincides with our list for $\overline{\Omega}_a(T_R)$ given in §8. Thus the collection of sequences $\{\overline{s}_n(\overline{\xi}, \overline{\eta})\}$ is identical with the collection of sequences $\{\overline{s}_n(\overline{\gamma})\}$. This led us to speculate that even more is true: namely, $\overline{s}_n(\overline{\xi}, \overline{\eta}) = \overline{s}_n(\overline{\gamma})$, for $\gamma = \gamma(\xi, \eta)$ and $(\overline{\xi}, \overline{\eta}) = \Phi(\xi, \eta)$. This turns out to be the case, and we shall prove it as follows. (i) We reinterpret Morse's geometric modification in terms of the coding rules of Theorem 9.9 converting $s_n(\gamma)$ into $\overline{s}_n(\overline{\gamma})$. (ii) We show in Theorem 9.8 that identical coding rules convert $\{s_n(\xi, \eta)\}$ into $\{\overline{s}_n(\overline{\xi}, \overline{\eta})\}$. That $\overline{s}_n(\overline{\xi}, \overline{\eta}) = \overline{s}_n(\overline{\gamma})$ then follows from Theorems 9.1, 9.8, and 9.9.

The proof of Theorem 9.9 follows readily from the way the sides of the edges of **F** receive their labels. The proof of Theorem 9.8 is difficult, the reason for this being that simple geometric notions are replaced by combinatorial ones—particularly the notions of *component* and *block* defined later on. Indeed the conversion rules of Theorem 9.8 seem artificial and, undoubtedly, we never would have guessed them if not for the speculation $\overline{s}_n(\overline{\xi}, \overline{\eta}) = \overline{s}_n(\overline{\gamma})$. To prove Theorem 9.8, we require some preliminary results concerning admissibility rules satisfied by the sequences $\{\overline{s}_n\}$.

Let $C_i = D_i \cup E_i \cup F_i \cup G_i \cup X_i$ as in §5. We also write (D, i) and (Di) for D_i, with a similar notation for $E_i \ldots$. Let $\mathscr{P}'$ be the partition $\{(Di), (Ei) (Fi), (Gi), (Xi): 1 \leq i \leq 8g - 4\}$ of C. Define $\{s'_n\} = \{s'_n(\xi, \eta)\}$, $n \in \mathbb{Z}$, by $T_C^n(\xi, \eta) \in s'_n$, $s'_n \in \mathscr{P}'$, and denote the space of sequences $\{s'_n\}$ by $\Omega'(T_C)$. Thus $s'_n = (l_n, s_n)$, where $\{s_n\}$ is the curvilinear sequence and $\{l_n\}$ is a sequence of symbols from the alphabet $\{D, E, F, G, X\}$.

Theorem 9.2.

$$\bar{s}_n = \begin{cases} s_n, & \text{if } l_n = G, \\ s_n + 1, & \text{if } l_n = X, \\ s_n + 4g, & \text{if } l_n = D \text{ or } F, \\ s_n + 4g - 1, & \text{if } l_n = E. \end{cases} \tag{9.1}$$

Proof. By (5.4), (5.5), and Theorem 5.1, Φ maps the sets $\{G_i, X_i, D_i, E_i, F_i\}$ respectively onto the sets $\{\overline{G}_i, \overline{X}_{i+1}, \overline{D}_{i+4g}, \overline{E}_{i+4g-1}, \overline{F}_{i+4g}\}$, from which Theorem 9.2 follows.

In view of Theorem 9.2, we will have converted $\{s_n\}$ into $\{\bar{s}_n\}$ if we can determine $\{l_n\}$ from $\{s_n\}$. This will be achieved by obtaining certain restrictions on $\Omega'(T_C)$.

Theorem 9.3. *Let* $\{s'_n\} \in \Omega'(T_C)$. *For* $s'_n \to s'_{n+1}$, $-\infty < n < \infty$, *it is necessary (but not sufficient) that it be included in the following table of successor rules:*

$D_i \to D_{\sigma(i)-2}, E_{\sigma(i)-1}, F_{\sigma(i)-1}$

$X_i \to D_{\sigma(i)-2}, E_{\sigma(i)-1}, F_{\sigma(i)-1}$

$F_i \to D_{\sigma(i)-3}, E_{\sigma(i)-2}, F_{\sigma(i)-2}$

$E_i \to X_{\sigma(i)+1}, X_{\sigma(i)+2}, \ldots, X_{\sigma(i)-4}; G_{\sigma(i)+1}, G_{\sigma(i)+2}, \ldots, G_{\sigma(i)-3}$

$G_i \to X_{\sigma(i)+1}, X_{\sigma(i)+2}, \ldots, X_{\sigma(i)-4}, X_{\sigma(i)-3}; G_{\sigma(i)+1}, G_{\sigma(i)+2}, \ldots,$
$G_{\sigma(i)-3}, G_{\sigma(i)-2}$

The proof of Theorem 9.3 is made evident by Figures 5.4 and 9.2 (see p. 292), the latter exhibiting the intersection of the sets $C'_i = T_C(C_i)$ with C_j.

For instance,

$$T_C(D_i) \subset U_{\sigma(i)-1} = D_{\sigma(i)-2} \cup E_{\sigma(i)-1} \cup F_{\sigma(i)-1}$$

which gives the first of the above rules.

Due to the manner in which $\sigma(i)$ appears in the rules, it proves convenient to restate Theorem 9.3 in terms of the sequences $\{(l_n, \delta_n)\}$, where $\delta_n = \sigma(s_{n-1}) - s_n$. We give a geometrical interpretation to $\{\delta_n\}$ used later on. Let the geodesic γ enter and leave $\mathbf{F}_n$ through the edges whose $\mathbf{F}_n$-side carry the respective labels s and s'; thus $s = \sigma(s_{n-1})$, $s' = s_n$. s' is the δ_n th side of $\mathbf{F}_n$ after s in the clockwise direction (see Figure 9.3 on p. 292 where $\delta_n = 4$).

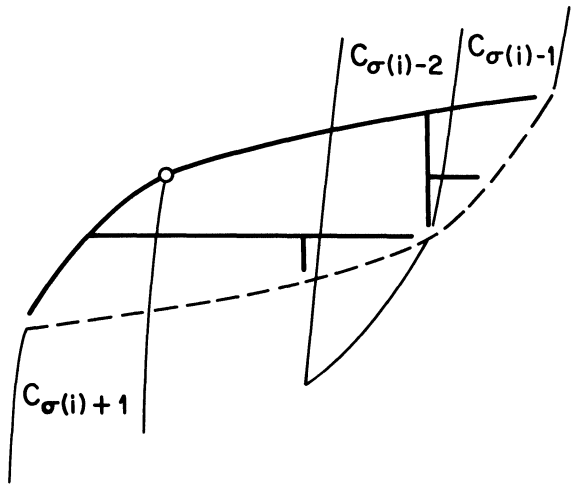


FIGURE 9.2. Intersection of C_i' with C_j.

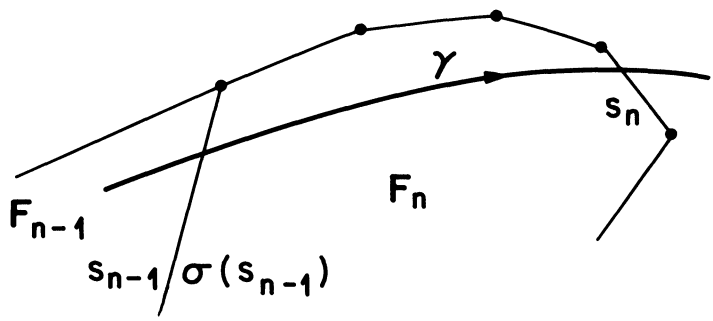


FIGURE 9.3. Geometric interpretation of δ_n.

In terms of (l_n, δ_n), Theorem 9.3 becomes

Theorem 9.4. *For* $(l_n, \delta_n) \to (l_{n+1}, \delta_{n+1})$ *it is necessary (but not sufficient) that it be included in the following table of successor rules.*

(9.2)

(i) $(D\cdot) \to (D2), (E1), (F1)$

(ii) $(X\cdot) \to (D2), (E1), (F1)$

(iii) $(F\cdot) \to (D3), (E2), (F2)$

(iv) $(E\cdot) \to (X4), \ldots, (X, 8g-5); (G3), \ldots, (G, 8g-5)$

(v) $(G\cdot) \to (X3), (X4), \ldots, (X, 8g-5); (G2), \ldots, (G, 8g-5).$

Inspection of the right side of the above rules shows that

(vi) $(l_n, \delta_n) \neq (X1), (X2), (D1), (G1), \qquad -\infty < n < \infty$

The following theorem lists certain forbidden blocks for $\{\delta_n\}$ and $\{(l_n, \delta_n)\}$.

Theorem 9.5. (i) $\{\delta_n\}$ *does not contain the blocks*

$$\begin{array}{cccccc} 1 & 2 & . & . & 2 & 1 \\ 1 & 2 & 2 & . & . & . \\ . & . & . & 2 & 2 & 1 \\ . & . & . & 2 & 2 & 2\ldots \end{array}$$

(ii) $\{(l_n, \delta_n)\}$ *does not contain the blocks*

$$\begin{array}{cc} (G2), & (G2), \ldots \\ \ldots, (G2), & (G2). \end{array}$$

Proof. (i) Translated into geometric terms, the four blocks of (i) become respectively Figures 9.4 (a–d), illustrating how the geodesic $\gamma = \gamma(\xi, \eta)$ passes successively through the sequence of regions $\{\mathbf{F}_n(\gamma)\}$. In each of the figures, the two geodesics γ and γ^* intersect twice in $\overline{\mathbb{D}}$, contradicting Theorem 2.1 (i). Hence Figures 2 (a–d) are impossible, and so $\{\delta_n\}$ does not contain any of the four blocks.

(ii) The appearance of these blocks in $\{(l_n, \delta_n)\}$ is equivalent respectively to the appearance of the following blocks in $\{(l_n, s_n)\}$:

$$(G, i), (G, \beta(i)), (G, \beta^2(i)), \ldots$$
$$\ldots, (G, \beta^2(i)), (G, \beta(i)), (G, i).$$

By Theorem 9.2, the appearance of the above implies respectively the appearance of the following blocks in $\{\bar{s}_n\}$

$$i, \beta(i), \beta^2(i), \ldots$$
$$\ldots, \beta^2(i), \beta(i), i$$

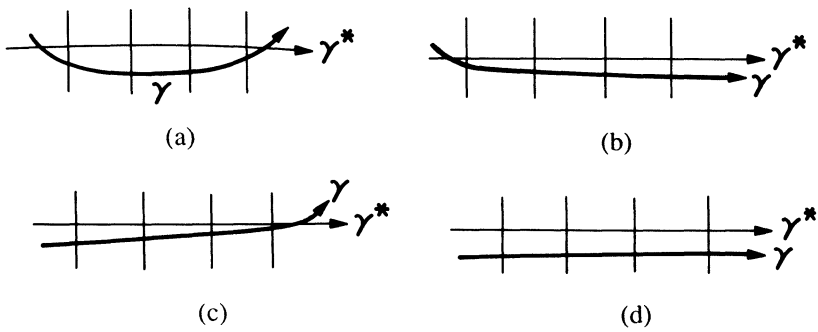


FIGURE 9.4. γ is not a geodesic.

which is impossible by Theorem 8.6. Hence the blocks do not appear in $\{(l_n, \delta_n)\}$.

We introduce some concepts seemingly artificial but necessary to obtain $\{l_n\}$ from $\{\delta_n\}$ (Theorem 9.6). In the sequel n always denotes an integer. Let $\overline{\mathbb{Z}} = \mathbb{Z} \cup \{+\infty, -\infty\}$, where $\mathbb{Z}$ is the set of all integers, and $J = \{n: \delta_n = 1\}$ augmented possibly by $+\infty$, or $-\infty$. We augment by $+\infty$ if $\delta_n = 2$ for all sufficiently large n, and by $-\infty$ if $\delta_n = 2$ for all sufficiently small n.

Definition 9.2. (i) Let $a \leq b$ be elements of J. We say that a and b are *connected* if and only if: either $a = b$, or $\delta_n = 3$ for a unique $a < n < b$ and $\delta_n = 2$ for all remaining $a < n < b$.

(ii) We call $M \subset J$ a *component* of J if and only if for any two elements $a < b$ in M, there exists a finite sequence $a = a_0 < a_1 < \cdots a_k = b$ in M such that a_i is connected to a_{i+1} for $0 \leq i < k$.

It is readily checked that distinct components M_1, M_2 do not overlap: i.e. either $M_1 < M_2$ (meaning $m_1 < m_2$ whenever $m_1 \in M_1$, $m_2 \in M_2$) or $M_2 < M_1$.

We enlarge any component M to a set $B = B(M) \subset \overline{\mathbb{Z}}$, which we call the M-*block determined by* M as follows. Let

$$M^- = \{x \in \overline{\mathbb{Z}}: x < m, m \in M, \text{ and } \delta(x) \neq 2\}$$
$$M^+ = \{x \in \overline{\mathbb{Z}}: x > m, m \in M, \text{ and } \delta(x) \neq 2\}.$$

Let $a = -\infty$ if $M^- = \varnothing$, $a = \sup M^-$ if $M^- \neq \varnothing$. Similarly $b = +\infty$ if $M^+ = \varnothing$, $b = \inf M^+$ if $M^+ \neq \varnothing$. Then $B = B(M) = \{x \in \overline{\mathbb{Z}}: a \leq x < b\}$. It is readily checked that $B(M_1) < B(M_2)$ when $M_1 < M_2$. We remark that J is the union of disjoint components; but, in general, $\overline{\mathbb{Z}}$ is not the union of disjoints M-blocks.

We illustrate the *component* and *block* concepts geometrically. Let γ be a geodesic with the cutting sequence $\{s_n\}$, and let δ_{n+1}, $\cdots$, δ_{n+11} be the sequence:

$$4, 2, 1, 2, 2, 3, 2, 1, 2, 2, 5.$$

γ passes successively through $\mathbf{F}_{n+1}, \ldots, \mathbf{F}_{n+11}$ and is depicted in Figure 9.5.

The indices $n+3$, $n+8$ are connected since $\delta_{n+3} = \delta_{n+8} = 1$ and all intermediate values of δ_k equal 2, except for δ_{n+6} which equals 3. $n+3$ is not connected to a preceding element as $\delta_{n+1} = 4$, $\delta_{n+2} = 2$. Similarly, $n+8$ is not connected to a succeeding element as $\delta_{n+9} = \delta_{n+10} = 2$, $\delta_{n+11} = 5$. Hence $\{n+3, n+8\}$ is

a component and $\{n+1, \ldots, n+10\}$ is the M-block determined by it.

We observe in Figure 9.5 that the interior angles, along the left boundary of $\bigcup \mathbf{F}_n(\gamma)$, which meet $\mathbf{F}_{n+3}$, $\mathbf{F}_{n+8}$ are bigger than π, and the one which meets $\mathbf{F}_{n+6}$ is less than π. This description holds in general. Namely, the interior angle meeting $\mathbf{F}_k$ is bigger than π if $\delta_k = 1$, and less than π if $\delta_k = 3$. (As illustrated in Figure 9.5, the interior angle lies partially in $\mathbf{F}_k$ when $\delta_k = 1$, and completely in $\mathbf{F}_k$ when $\delta_k = 3$.)

Theorem 9.6. (i) *Let* M *be a component of* J *and* $a < c$ *successive elements of* M. *Thus there is a unique* $a < b < c$, *for which* $\delta_b = 3$. *Then*

$$l_n = \begin{cases} F, & a \le n < b, \\ D, & b \le n < c. \end{cases} \tag{9.3}$$

(ii) *Let* a *be the largest element of* M *and* $a \neq +\infty$. *By Theorem* 9.5(i), *there exists an integer* b *such that* $\delta_n = 2$, $a < n < b$, *and* $\delta_b \neq 1, 2$. *Then*

$$l_n = \begin{cases} F, & a \le n < b-1, \\ E, & n = b-1. \end{cases} \tag{9.4}$$

(iii) *Let* a *be the smallest element of* M *and* $a \neq -\infty$. *By Theorem* 9.5 (i) *there exists an integer* b *such that* $\delta_n = 2$,

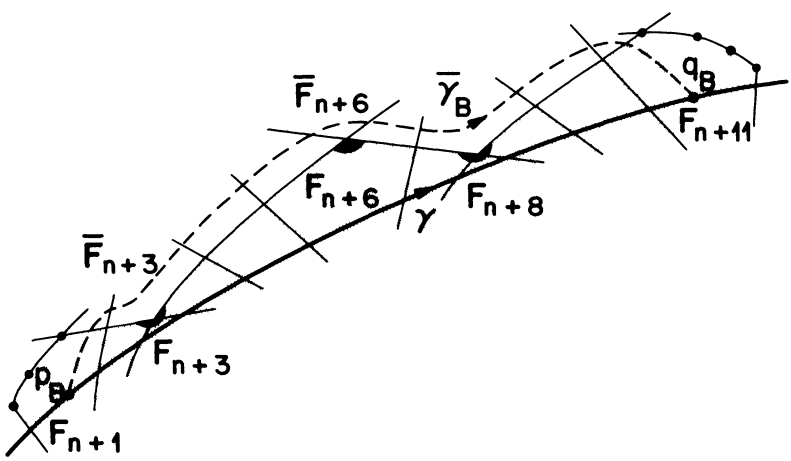


FIGURE 9.5. M-block and modified geodesic.

$b < n < a$, and $\delta_b \neq 1, 2$. Then

$$l_n = \begin{cases} D, & b < n < a, \\ X, & n = b. \end{cases} \tag{9.5}$$

(iv) *$l_n = G$, $n \notin \bigcup B$ where $\bigcup B$ is the union of all* M*-blocks.*

Proof. We show that the transition rules satisfied by $\{(l_n, \delta_n)\}$ (Theorem 9.4) and the list of omitted blocks of Theorem 9.5 force the values of l_n listed above. Though the reasoning is straightforward, it is tedious and we proved the details for (i) and (ii) only, the reasoning for (iii), (iv) being analogous.

(i) Since a has a successor c, $a \neq +\infty$. Consider first $a \neq -\infty$, so that $\delta_a = 1$. By 9.2 (vi), $l_a = E$ or F. If $\delta_{a+1} = 2$, then 9.2 (iv) forces $l_a = F$. By 9.2 (iii) $l_{a+1} = E$ or F. Repeating the argument, we get $l_n = F$, $a \leq n \leq b-2$, and $l_{b-1} = E$ or F.

Suppose $l_{b-1} = E$. Since $\delta_b = 3$, 9.2 (iv) forces $l_b = G$. Repeated use of 9.2 (v) gives $l_n = G$, $b < n < c$. If c is finite, then $\delta_c = 1$ and 9.2 (v) gives $l_{c-1} \neq G$, a contradiction. If $c = +\infty$, then $(l_n, \delta_n) = (G, 2)$, $b+1 \leq n < \infty$, which is a forbidden block in Theorem 9.5. Hence $l_{b-1} = F$. By 9.2 (iii), $l_b = D$. Repeated use of 9.2 (i) gives $l_n = D$, $b \leq n < c$.

Consider next $a = -\infty$, so that $\delta_n = 2$ for $-\infty < n < b$. We show that for $-\infty < n < b$, $\delta_n = E$ or F, the remainder of the argument proceeding as before. By 9.2 (vi) $l_n \neq X$. Suppose $\delta_n = D$. Repeated use of 9.2 (i) gives $l_k = D$, $n \leq k < b$. Since $\delta_b = 3$, 9.2 (i) gives $l_{b-1} \neq D$, a contradiction. Suppose $l_n = G$. From the transition rules of (9.2), $l_n = G$ and $\delta_n = 2$ force $l_{n-1} = G$. Repeated use of this argument gives $(l_k, \delta_k) = (G, 2)$, $-\infty < k < n$, which is a forbidden block in Theorem 9.5. Hence $\delta_n = E$ or F.

(ii) Duplicating the reasoning of (i) we get $l_n = F$, $a \leq n < b-1$ and $l_{b-1} = E$ or F. Suppose $l_{b-1} = F$. Since $\delta_b \neq 2$, 9.2 (iii) forces $l_b = D$, $\delta_b = 3$. If $\delta_{b+1} = 1$ then $a, b+1$ are connected, contradicting that a is the largest element in M. Hence $l_{b+1} \neq 1$ and 9.2 (i) gives $l_{b+1} = D$, $\delta_{b+1} = 2$. Repeating this argument we get $l_k = D$, $\delta_k = 2$, $b \leq k < \infty$. But then a, $+\infty$ are connected, again contradicting that a is the largest element in M. Hence $l_{b-1} = E$.

In the proofs of (i)–(ii) we have made use of the successor table of Theorem 9.4. To prove (iii), (iv) we also require the predecessor table derivable from it.

Let b_0, b_1 be the first and the last integer of a given M-block B, when these exist. We conclude from Theorem 9.6 that

Theorem 9.7.

$$l_n = \begin{cases} G & \text{if } n \notin \bigcup B, \\ X & \text{if } n = b_0, \\ D \text{ or } F & \text{if } n \in B - \{b_0, b_1\}, \\ E & \text{if } n = b_1. \end{cases} \tag{9.6}$$

The proof of Theorem 9.7 is based on the admissibility rules of Theorem 9.4. In Appendix D, we give another proof based on geometric considerations. Theorems 9.2, 9.7 give

Theorem 9.8.

$$\bar{s}_n = \begin{cases} s_n & \text{if } n \notin \bigcup B, \\ s_n + 1 & \text{if } n = b_0, \\ s_n + 4g & \text{if } n \in B - \{b_0, b_1\}, \\ s_n + 4g - 1 & \text{if } n = b_1. \end{cases} \tag{9.7}$$

We now obtain for $\{\bar{s}_n\}$ the geometric interpretation mentioned earlier in this section. We refer to Figure 9.5. Let γ_L be the left boundary of $\bigcup \mathbf{F}_n(\gamma)$. Suppose B is a finite M-block with smallest integer b_0 and largest integer b_1. Let $\overline{\mathbf{F}}_{b_0} = \mathbf{F}_{b_0}$, and let $\overline{\mathbf{F}}_{b_0+1}$ be the fundamental region adjacent to $\mathbf{F}_{b_0}$ and opposite to $\mathbf{F}_{b_0+1}$. $\overline{\mathbf{F}}_{b_0+2}, \ldots, \overline{\mathbf{F}}_{b_1}$ are then chosen to be the sequence of successive fundamental regions encountered along the left side of γ_L. Choose $p_B \in \gamma \bigcap \mathbf{F}_{b_0}$, $q_B \in \gamma \bigcap \mathbf{F}_{b_1+1}$ and connect p_B to q_B by a curve $\bar{\gamma}_B$ passing successively through the interiors of $\mathbf{F}_{b_0}$, $\overline{\mathbf{F}}_{b_0+1}, \ldots, \overline{\mathbf{F}}_{b_1}, \overline{\mathbf{F}}_{b_1+1} = \mathbf{F}_{b_1+1}$ (for infinite M-blocks B, the definition of $\bar{\gamma}_B$ needs some minor changes which we leave to the reader). For each M-block B, replace the portion of γ between p_B and q_B by $\bar{\gamma}_B$ and denote the resulting curve by $\bar{\gamma}$. We call $\bar{\gamma}$ the *modification* of γ or the *modified geodesic.* Just as for γ, we associate to $\bar{\gamma}$ a cutting sequence $\{\bar{s}_n(\bar{\gamma})\}$ called the *modified cutting sequence.*

Theorem 9.9.

$$\bar{s}_n(\bar{\gamma}) = \begin{cases} s_n(\gamma) & \text{if } n \notin \bigcup B, \\ s_n(\gamma) + 1 & \text{if } n = b_0, \\ s_n(\gamma) + 4g & \text{if } n \in B - \{b_0, b_1\}, \\ s_n(\gamma) + 4g - 1 & \text{if } n = b_1. \end{cases} \tag{9.8}$$

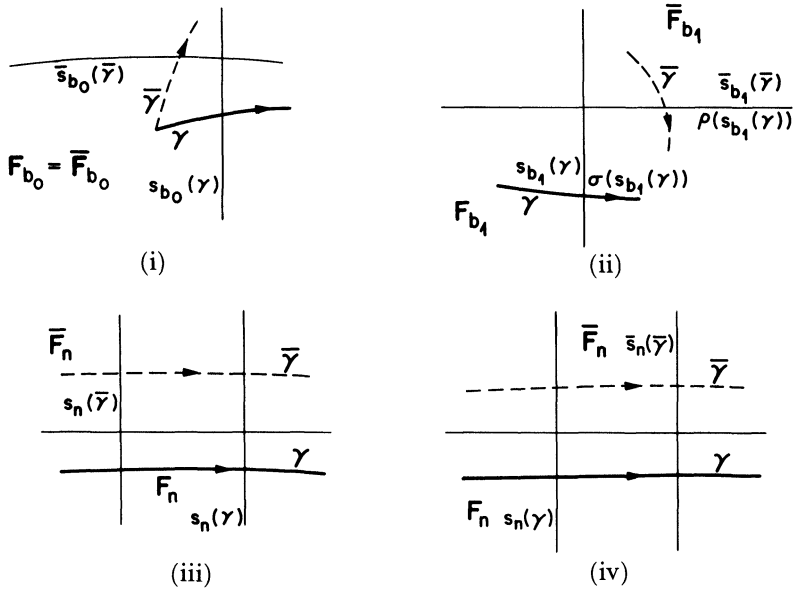


FIGURE 9.6. Cutting sequences $\{s_n(\gamma)\}$ and $\{\bar{s}_n(\bar{\gamma})\}$.

Proof. The following figures (See Figure 9.6.) depict the position of $\bar{\mathbf{F}}_n$ relative to $\mathbf{F}_n$. For $b_0 < n < b_1$, we require the two figures (iii)–(iv) since both of these will arise. Thus, referring to Figure 9.5, we observe that $\mathbf{F}_{n+2}$, $\bar{\mathbf{F}}_{n+2}$ appear as in (iii), and $\mathbf{F}_{n+4}$, $\bar{\mathbf{F}}_{n+4}$ as in (iv).

From the figures we get (9.8). For instance, from Figure 9.6 (i) we obtain $\bar{s}_{b_0}(\bar{\gamma}) = s_{b_0}(\gamma) + 1$; from Figure 9.6 (ii) and (5.6) we obtain $\bar{s}_{b_1}(\bar{\gamma}) = \sigma\rho(s_{b_1}(\gamma)) = s_{b_1}(\gamma) + 4g - 1$.

Theorem 9.10. *Let* $(\bar{\xi}, \bar{\eta}) = \Phi(\xi, \eta)$, $\gamma = \gamma(\xi, \eta)$. *Then* $\{\bar{s}_n(\bar{\xi}, \bar{\eta})\} = \{\bar{s}_n(\bar{\gamma})\}$.

Proof. By Theorem 9.1, $\{s_n(\xi, \eta)\} = \{s_n(\gamma)\}$. Theorems 9.8, 9.9 show that the rules for converting $\{s_n(\xi, \eta)\}$ to $\{\bar{s}_n(\bar{\xi}, \bar{\eta})\}$ are the same as for converting $\{s_n(\gamma)\}$ to $\{\bar{s}_n(\bar{\gamma})\}$. Hence $\{\bar{s}_n(\bar{\xi}, \bar{\eta})\} = \{\bar{s}_n(\bar{\gamma})\}$.

10. EPILOGUE

Our study has been restricted to compact surfaces because for surface groups we have the existence of a fundamental region satisfying the crucial extension condition. It might be interesting to see to what extent our treatment can be extended to the noncompact case.

Despite the fact that rectilinear sequences are in one-to-one correspondence with cutting sequences and we have a nice description of the former, the question remains: is there a satisfactory specification of the latter? It is a striking fact that,for given genus $g \geq 2$, we get the same set of rectilinear sequences, regardless of the shape of the $(8g-4)$ sided fundamental region. Furthermore, the rules for coding cutting sequences to rectilinear ones are also the same. Thus, it is impossible to get a description of cutting sequences from devising "finitary decoding" rules applied to rectilinear sequences.

Certain things are known about the set of cutting sequences. For each genus $g \geq 2$ there is a countable set of universally forbidden finite blocks—i.e., they are forbidden for all $(8g-4)$ sided fundamental regions satisfying the conditions of Theorem 3.1. These are the finite blocks containing one of the following as a sub-block:

(10.1)

$$i, \rho(i), \beta\rho(i), \ldots, \beta^k\rho(i), \rho\beta^k\rho(i); \qquad 1 \leq i \leq 8g-4 \quad \text{and} \quad 0 \leq k < \infty$$

(10.2) $$i, \vartheta(i), \alpha\vartheta(i), \ldots, \alpha^k\vartheta(i), \vartheta\alpha^k\vartheta(i);$$

The blocks (10.1) correspond to the curves depicted in Figure 9.4a and (10.2) to the reflection of these curves about γ^*. For a given fundamental region there are other forbidden finite blocks aside from the universal ones: e.g., let γ be a geodesic not passing through a vertex of $\mathbf{F}$ with $s_0(\gamma) = a$, $s_1(\gamma) = b$, and define the sequence $\{s_n'\}$, $n \in Z$, by

$$s_n' = \begin{cases} c, & \text{if } n = 0, \\ d, & \text{if } n = 1, \\ s_n(\gamma), & \text{otherwise}, \end{cases}$$

where a b, c and d are as in Figure 10.1 on p. 300.

We claim that $\{s_n'\}$, $|n| \leq N$, cannot appear in any cutting sequence for N sufficiently large. For suppose there is a geodesic γ' for which $s_n(\gamma') = s_n'$, $|n| \leq N$. Then $s_n(\gamma) = s_n(\gamma')$ for $|n| \leq N$, $n \neq 0, 1$. The above condition forces γ, γ' to be arbitrarily close for N sufficiently large. Hence, when N is sufficiently large, $s_n(\gamma') = s_n(\gamma)$ for $n = 0, 1$ as well, a contradiction. On the other hand, it is readily checked that $\{s_n'\}$, $|n| \leq N$, does not contain any of the universally forbidden blocks prescribed by

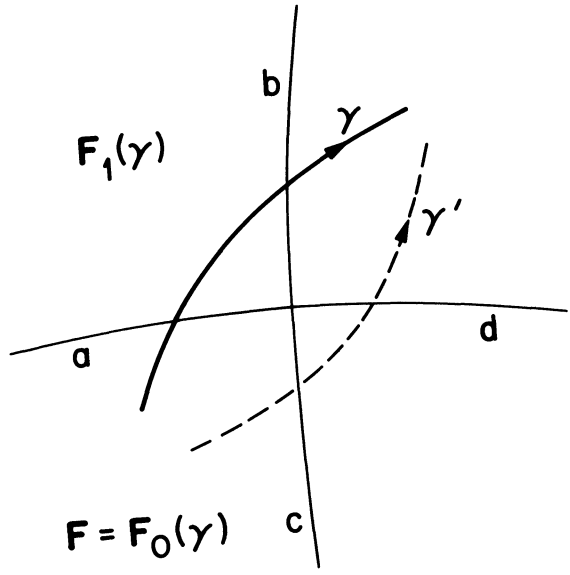


FIGURE 10.1. Suggested alternate geodesic route.

(10.1), (10.2). We finish with two questions:

1. Are the blocks described by (10.1), (10.2) the only universally forbidden ones?
2. Do the curvilinear sequences determine the fundamental region up to conjugacy within the group of motions?

APPENDIX A. EXISTENCE OF THE $(8g-4)$-SIDED FUNDAMENTAL POLYGON

In this appendix we prove the existence of the $(8g-4)$-sided fundamental polygon described in Theorem 3.1. The proof is based on classical theorems of Poincaré and Frenchel–Nielsen in the theory of Fuchsian groups. We do not state these theorems in their entirety, and quote only those parts used to prove Theorem 3.1.

Poincaré's theorem gives sufficient conditions for a polygon to be a fundamental region of a Fuchsian group. Let $\mathbf{F}$ be a bounded polygon with an even number of consecutive edges $s_1, \ldots, s_n$ oriented in the counter-clockwise direction, and let $p_1, \ldots, p_n$ be the respective initial points of these edges. Assume that s_i, $s_{\sigma(i)}$, $1 \leq i \leq n$, are of equal hyperbolic length, where $\sigma(i)$ is a given permutation of $1, \ldots, n$ of order two without fixed elements. Let

T_i be the motion carrying s_i to $s_{\sigma(i)}^{-1}$, i.e. $T_i(p_i) = p_{\vartheta(i)}$, $1 \le i \le n$. Express the permutation $i \to \vartheta(i)$ as a product of disjoint cycles. The set of vertices p_i, i varying over a cycle, is called a vertex cycle.

Poincaré's theorem. *Suppose that the sum of the interior angles of* $\mathbf{F}$ *over each vertex cycle equals* $2\pi/m_c$, *where* m_c *is a positive integer. Then the group* Γ *generated by* $T_1, \ldots, T_n$ *is discrete and* $\mathbf{F}$ *is a fundamental region for* Γ. Γ *acts freely on* $\mathbb{D}$ *if and only if* $m_c = 1$ *for each vertex cycle* c.

A proof of Poincaré's theorem can be found in [M]. As an application, consider the regular $(8g-4)$-sided polygon $\mathbf{F}'$ with interior angles $\pi/2$ centered at 0, with the pairing of Theorem 3.1 (we replace s_i, T_i respectively by s_i', T_i'). From formula (3.1) for $\sigma(i)$, one readily verifies that $i \to \vartheta(i)$ is a product of disjoint cycles, all of length four, and that the sum of the angles in each vertex cycle equals 2π. By Poincaré's theorem, the group generated by $T_1', \ldots, T_{8g-4}'$ is a Fuchsian group acting freely on $\mathbb{D}$ with $\mathbf{F}'$ for fundamental region.

Let Γ_1, Γ_2 be two surface groups. Suppose that $\mathbb{D}/\Gamma_1$ is homeomorphic to $\mathbb{D}/\Gamma_2$, which is equivalent to stating that $\mathbb{D}/\Gamma_1$, $\mathbb{D}/\Gamma_2$ have the same genus.

Fenchel-Nielsen theorem. *There exists an orientation preserving homeomorphism* h *from* $\overline{\mathbb{D}}$ *onto* $\overline{\mathbb{D}}$ *such that* $\Gamma_2 = h \circ \Gamma_1 \circ h^{-1}$.

We remark that if we replace $\overline{\mathbb{D}}$ by $\mathbb{D}$, then the above theorem follows readily from the theory of covering spaces [Sp, Chapter 4]. We just choose an orientation preserving homeomorphism φ from $\mathbb{D}/\Gamma_1$ onto $\mathbb{D}/\Gamma_2$ and lift it up to an orientation preserving homeomorphism h from $\mathbb{D}$ onto $\mathbb{D}$. The Fenchel-Nielsen theorem asserts that h can be extended to a homeomorphism from $\overline{\mathbb{D}}$ onto $\overline{\mathbb{D}}$. A proof of the theorem can be found in [T, §3; see, in particular, Proposition 3.5 and its corollary].

Proof of Theorem 3.1. Let g be the genus of $\mathbb{D}/\Gamma$. Let $\mathbf{F}'$ be the regular $(8g-4)$-sided polygon described above and Γ' the associated Fuchsian group with generators $T_1', \ldots, T_{8g-4}'$. Let ν be the number of vertex cycles and e the number of paired sides. Then $\nu = 2g-1$, $e = 4g-2$. By Euler's formula, the genus g' of $\mathbb{D}/\Gamma'$ is given by

$$2 - 2g' = \nu - e + 1 = 2 - 2g.$$

Thus $g' = g$. By the Fenchel-Nielsen theorem, there exists an orientation preserving homeomorphism h from $\overline{\mathbb{D}}$ onto $\overline{\mathbb{D}}$ such that $\Gamma = h\Gamma' h^{-1}$. The restriction of h to $\partial\mathbb{D}$ provides an orientation preserving homeomorphism from $\partial\mathbb{D}$ onto itself.

For $\mathbf{F}'$, let $\bar{s}_i'$ be the geodesic containing s_i' and with the same orientation, and let a_i', b_i' be respectively the backward and forward end points of $\bar{s}_i'$. Let $a_i = h(a_i')$, $b_i = h(b_i')$. Since h is orientation preserving on $\partial\mathbb{D}$, we conclude from Theorem 3.3 that the points a_i, b_i, $1 \le i \le 8g-4$ are all distinct and encountered along $\partial\mathbb{D}$ in the counter-clockwise direction in the order

$$a_1, b_0, a_2, b_1, \ldots, a_{8g-4}, b_{8g-5}.$$

Let $\bar{s}_i$ be the geodesic from a_i to b_i. Since a_{i-1}, b_{i-1} separate a_i, b_i on $\partial\mathbb{D}$, the geodesics $\bar{s}_{i-1}$, $\bar{s}_i$ meet in a point p_i in $\mathbb{D}$. The points p_i, $1 \le i \le 8g-4$ form the successive vertices of a polygon $\mathbf{F}$ with edges s_i, s_i being the geodesic segment from p_i to p_{i+1}. We show that $\mathbf{F}$ is the desired polygon.

Let $T_i = hT_i'h^{-1}$. The map $\tau' \to h\tau' h^{-1}$, $\tau' \in \Gamma'$, is an isomorphism from Γ' onto Γ and hence maps the generators $T_1', \ldots, T_{8g-4}'$ of Γ' respectively to the generators $T_1, \ldots, T_{8g-4}$ of Γ. If $T_i'x = y$, then $T_i(hx) = hy$. It follows from Theorem 3.4 that T_i maps the points a_{i-1}, b_{i-1}, a_i, b_i, a_{i+1} b_{i+1} respectively to $a_{\vartheta(i)}$, $b_{\vartheta(i)}$, $b_{\sigma(i)}$, $a_{\sigma(i)}$, $a_{\rho(i)}$, $b_{\rho(i)}$, i.e. T_i maps the geodesics $\bar{s}_{i-1}$, $\bar{s}_i$, $\bar{s}_{i+1}$ respectively to $\bar{s}_{\vartheta(i)}$, $\bar{s}_{\sigma(i)}^{-1}$, $\bar{s}_{\rho(i)}$. Hence $T_i(p_i) = p_{\vartheta(i)}$, $T_i(p_{i+1}) = p_{\sigma(i)}$—i.e., T_i maps s_i to $s_{\sigma(i)}^{-1}$ and the interior angle θ_i at p_i to the exterior angle at $p_{\vartheta(i)}$ (Figure A1 on p. 303).

Since T_i is angle preserving, we conclude that $\theta_i + \theta_{\vartheta(i)} = \pi$. The sum of the interior angles of a vertex cycle is given by

$$(\theta_i + \theta_{\vartheta(i)}) + (\theta_{\vartheta^2(i)} + \theta_{\vartheta^3(i)}) = \pi + \pi = 2\pi.$$

Thus $\mathbf{F}$ satisfies the conditions of Poincaré's theorem, and we conclude that $\mathbf{F}$ is a fundamental region for Γ.

Appendix B. The Folklore theorem

This appendix deals with a topic that we call the "Folklore theorem." It is not clear to us to whom it should be credited. One could justifiably attribute it to Renyi [R]. His theorem has the desired conclusions: yet his work is so heavily influenced by the classical presentation of continued fractions that not enough emphasis is placed on the fact that his theorem is really about

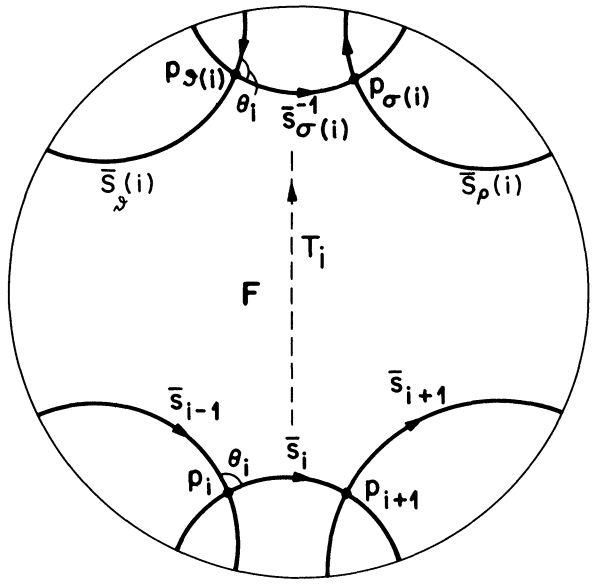


FIGURE A1. Diagram used in proof of Theorem 3.1.

iteration of a map. Consequently certain important issues such as the Markovian property are not made manifest. He recognizes a key step in the theorem—namely, the uniform bounding of ratios $|f^{n'}(x)/f^{n'}(y)|$ of derivatives of iterates of a map f; but he assumes it as hypothesis leaving one with the nontrivial task of verifying it for particular maps. Even earlier, in the proof of ergodicity of geodesic flows, Hedlund [H] had employed some of the techniques used in the folklore theorem, but he did not single out this theorem as an important one in its own right.

The folklore theorem is about Markovian interval maps. We shall recall some of the concepts discussed in the introduction. Let X be a one dimensional space, say an interval or circle of unit length, and $\mathscr{P}$ a finite partition of X into subintervals. More specifically, let $\mathscr{P} = \{I_1, \ldots, I_N\}$ where $X = \bigcup_{i=1}^{N} I_i$ and $I_i \cap I_j = \varnothing$, $i \neq j$, $N \geq 2$. We allow intervals to be open, closed, or half open without restriction.

Let f be a map of X onto itself with f^n denoting the n-fold composition of f with itself. We make the following assumptions.

(i) (smoothness) f/I_i has a C^2-extension to the closure $\bar{I}_i$ of I_i. For economy of notation we designate all these extensions again by f.

(ii) (local invertibility) f is strictly monotone on $\overline{I}_i$ and therefore determines a one-to-one mapping of $\overline{I}_i$ onto some closed subinterval $f(\overline{I}_i)$ of X.

(iii) (Markov) For each $I \in \mathscr{P}$ there is a subset $\mathscr{P}(I)$ of $\mathscr{P}$ such that $f(\overline{I}) = \bigcup\{\overline{J}: J \in \mathscr{P}(I)\}$.

(iv) (aperiodicity) There exists a positive integer q such that $f^q(\overline{I}) = \overline{X}$.

Figure 1.3 of the introduction gives a typical example of a map satisfying (i)–(iv).

We introduce some notation: $f^{-n}\mathscr{P} = \{f^{-n}I: I \in \mathscr{P}\}$, $\mathscr{P}^{(n)} =$ the common refinement of $\mathscr{P}$, $f^{-1}\mathscr{P}, \ldots, f^{-n}\mathscr{P}$, $n \geq 0$. Accordingly, we use $I^{(n)}$ to denote a generic member of $\mathscr{P}^{(n)}$. We remark that two points x, y are in the same $I^{(n)}$ iff $f^i x$, $f^i y$ lie in the same element of $\mathscr{P}$ for $0 \leq j \leq n$. From (i)–(iii) we get that $I^{(n)}$ is an interval which we call an n th *stage interval*; and most importantly f maps monotonically each $\overline{I}^{(n)}$ onto some $\overline{I}^{(n-1)}$. Thus f^n has a C^2-extension to $\overline{I}^{(n)}$, also denoted by f^n, which maps $\overline{I}^{(n)}$ monotonically onto some $\overline{I}^{(0)}$. Let $M(I^{(n)}) = \sup_{x,y \in I^{(n)}} |f^{n\prime}(x)/f^{n\prime}(y)|$ and $M_n = \sup_{I^{(n)} \in \mathscr{P}^{(n)}} M(I^{(n)})$. Finally, let $||\mathscr{P}^{(n)}|| = \sup_{I^{(n)} \in \mathscr{P}^{(n)}} \lambda(I^{(n)})$ where λ is Lebesgue measure.

We call f *expansive* if there exists a positive integer p such that $|f^{p\prime}(x)| \geq 1 + \varepsilon$ for some $\varepsilon > 0$, all $x \in I^{(p)}$, and all $I^{(p)}$. For expansive maps we conclude from the chain rule that

$$L = \inf |f^{j\prime}(x)| > 0,$$

the infimum taken as x varies over $I^{(j)}$, $I^{(j)}$ over $\mathscr{P}^{(j)}$, and $0 \leq j < p$; and

(B1)

$$\inf_{\substack{x \in I^{(n)} \\ I^{(n)} \in \mathscr{P}^{(n)}}} |f^{n\prime}(x)| \geq (1+\varepsilon)^{[n/p]} \geq L(1+\varepsilon)^{(n/p-1)}, \qquad n \geq 1.$$

Since for each $I^{(n)}$ there is an $I^{(0)}$ such that $f^n(\overline{I}^{(n)}) = \overline{I}^{(0)}$, we have from the mean value theorem that there is an $x \in \overline{I}^{(n)}$ such that $\lambda(I^{(n)}) = \lambda(I^{(0)})/|f^{n\prime}(x)|$. For expansive maps it follows from (B1) that

$$\lambda(I^{(n)}) \leq B\theta^n \tag{B2}$$

where $\theta = (1+\varepsilon)^{-1/p} < 1$ and $B = (1+\varepsilon)\max \lambda(I^{(0)})/L$: i.e.

$$||\mathscr{P}^{(n)}|| = O(\theta^n). \tag{B3}$$

Theorem B1(Folklore). *If f satisfies assumptions* (i)–(iv) *and is expansive, then it has an ergodic (hence unique) invariant probability measure μ equivalent to λ with density function $d\mu/d\lambda$ which can be chosen piecewise continuous, the discontinuities only at end points of intervals in $\mathscr{P}$, and satisfying*

$$1/D \leq \frac{d\mu}{d\lambda} \leq D \quad \text{for some } D > 0.$$

Theorem B2(Converse). *If f satisfies assumptions* (i)–(iv) *and it has an invariant measure μ with $d\mu/d\lambda$ satisfying the above conditions, then f is expansive (hence ergodic by the Folklore theorem itself).*

The proof of the theorem proceeds by a series of lemmas for which we make the blanket assumption that f is expansive and satisfies assumptions (i)–(iv). The key idea is to obtain certain estimates on $f^{n\prime}(x)$ which are independent of n.

Lemma B1. *There is a $c > 0$ such that*

$$\text{(B4)} \qquad \left|\frac{d}{dx}|f^{n\prime}(x)|^{-1}\right| < c \quad \text{for all } n \geq 0.$$

Proof. By monotonicity $f^{n\prime}(x)$ does not vanish on $\bar{I}^{(n)}$ and so $|f^{n\prime}(x)|$ is C^1 on $\bar{I}^{(n)}$. We use the chain rules $f^{k\prime}(x) = f'(x)f'(fx)\dots f'(f^{k-1}x)$ and $f^{n\prime}(x) = f^{(n-k)\prime}(f^k x)f^{k\prime}(x)$ along with logarithmic differentiation to get

$$\begin{aligned}
\left|\frac{d}{dx}|f^{n\prime}(x)|^{-1}\right| &= |f^{n\prime}(x)|^{-1}\left|\frac{d}{dx}\log|f^{n\prime}(x)|\right| \\
&= |f^{n\prime}(x)|^{-1}\left|\sum_{k=0}^{n-1}\frac{d}{dx}\log|f'(f^k x)|\right| \\
&\leq \sum_{k=0}^{n-1}\left|\frac{f''(f^k x)}{f'(f^k x)}\cdot\frac{f^{k\prime}(x)}{f^{n\prime}(x)}\right| \\
&\leq \sup|f''(x)/f'(x)|\cdot\sum_{k=0}^{n-1}|f^{(n-k)}(f^k x)|^{-1}
\end{aligned}$$

whereupon by (B1)

$$\begin{aligned}
\left|\frac{d}{dx}|f^{n\prime}(x)|^{-1}\right| &\leq \sup|f''(x)/f'(x)|\sum_{k=0}^{n-1}\frac{(1+\varepsilon)}{L}\theta^{n-k} \\
&\leq \sup|f''(x)/f'(x)|\frac{(1+\varepsilon)}{L}\sum_{j=1}^{\infty}\theta^j < \infty.
\end{aligned}$$

As a corollary we now easily obtain the following

Lemma B2. *There exists* $M > 0$ *such that*

$$M_n < M \quad \text{for all } n \geq 0.$$

Proof. For $x, y \in I^{(n)}$, we get by monotonicity of f^n on $I^{(n)}$ and (B4)

$$\begin{aligned}\log|f^{n\prime}(x)/f^{n\prime}(y)| &= \log(f^{n\prime}(x)/f^{n\prime}(y)) \\ &= \int_y^x f^{n\prime\prime}(t)/f^{n\prime}(t)dt \\ &= -\int_y^x f^{n\prime}(t)\frac{d}{dt}(f^{n\prime}(t))^{-1}dt \\ &\leq c\left|\int_y^x f^{n\prime}(t)dt\right| \leq cf^n(\bar{I}^{(n)}) \\ &= c.\end{aligned}$$

The lemma follows by taking $m = e^c$.

Lemma B3. *There exists* $D > 0$ *such that for all* $I^{(n)}$ *and all sets* E *of positive Lebesgue measure*

$$1/D \leq \frac{\lambda(f^{-(n+q)}E \cap I^{(n)})}{\lambda(E)\lambda(I^{(n)})} \leq D. \tag{B5}$$

Proof. Since $E = \bigcup_{I\in\mathscr{P}}(E \cap I)$ and λ is additive it suffices to establish (B5) for $E \subset I \in \mathscr{P}$.

Suppose $F \subset I \in \mathscr{P}$ and $f^k\bar{I}^{(k)} = \bar{I}$. From the change of variables formula for integrals

$$\lambda(F) = \int_{f^{-k}F\cap I^{(k)}} |f^{k\prime}(x)|d\lambda$$

we get

$$\frac{1}{\max|f^{k\prime}(x)|} \leq \frac{\lambda(f^{-k}F \cap I^{(k)})}{\lambda(F)} \leq \frac{1}{\min|f^{k\prime}(x)|} \tag{B6}$$

where max and min are taken over $x \in I^{(k)}$. Dividing the inequalities obtained from (B4) for $k = n + q$, $F = E \subset I$, by the one obtained for $k = n + q$, $F = I$, we get

$$\begin{aligned}1/M \leq 1/M_{n+q} &\leq \frac{\lambda(f^{-(n+q)}E \cap I^{(n+q)})}{\lambda(E)\lambda(I^{(n+q)})} \\ &\leq M_{n+q}/\lambda(I) \leq \frac{M}{\min_{I\in\mathscr{P}}\lambda(I)}\end{aligned} \tag{B7}$$

for $E \subset I$, $f^{(n+q)}\overline{I}^{(n+q)} = \overline{I}$. Since $\lambda(f^{-(n+q)}E \cap I^{(n+q)}) = 0$ when $f^{(n+q)}\overline{I}^{(n+q)} \neq \overline{I}$, the right side of (B7) remains true for $E \subset I$ and *all* $I^{(n+q)}$. Multiply the right inequality of (B7) by $\lambda(E)\lambda(I^{(n+q)})$. Since each $I^{(n)}$ is a union of $(n+q)$ th-stage intervals, we obtain the right inequality of (B5) from that of (B7) by additivity of λ. To obtain the left inequality of (B5) argue as follows.

For $I^{(n+q)} \subset I^{(n)}$, let $\overline{I}^{(0)} = f^n(\overline{I}^{(n)})$ and $\overline{I}^{(q)} = f^n(\overline{I}^{(n+q)}) \subset \overline{I}^{(0)}$. Then $f^{-n}\overline{I}^{(q)} \cap \overline{I}^n = \overline{I}^{(n+q)}$. In the left inequality of (B7) replace $n+q$, E, I by $n, \overline{I}^{(q)} \cap I^{(0)}, I^{(0)}$. We get

$$\text{(B8)} \qquad \frac{\min \lambda(I^{(q)})}{M} \leq \frac{\min \lambda(I^{(q)})}{M} \leq \frac{\lambda(I^{(n+q)})}{\lambda(I^{(n)})}$$

for $\overline{I}^{(n+q)} \subset \overline{I}^{(n)}$. By property (iv) of f,

$$\overline{I} \subset \overline{X} = f^{(n+q)}\overline{I}^{(n)} = \bigcup_{\overline{I}^{(n+q)} \subset \overline{I}^{(n)}} f^{(n+q)}\overline{I}^{(n+q)}$$

so we may choose $I^* \in \mathscr{P}^{(n+q)}$ such that $I^* \subset I^{(n)}$ and $f^{n+q}\overline{I}^* = \overline{I}$. Therefore, from (B8) and the left side of (B7),

$$\text{(B9)} \qquad \begin{aligned} \lambda(f^{-(n+q)}E \cap I^{(n)}) &\geq \lambda(f^{-(n+q)}E \cap I^*) \geq \lambda(E)\lambda(I^*)/M \\ &\geq \lambda(E)\lambda(I^{(n)}) \min \lambda(I^{(q)})/M^2 \end{aligned}$$

for $E \subset I$ and *all* $I^{(n)}$, which gives the left inequality of (B5).

Lemma B4. *Tail sets are trivial.*

Proof. Multiplying (B5) by $\lambda(I^{(n)})$ and summing over $I^{(n)}$, we get

$$\text{(B10)} \qquad 1/D \leq \lambda(f^{-(n+q)}E)/\lambda(E) \leq D,$$

for $\lambda(E) > 0$, $n \geq 0$. Now if A is a tail set, which means for each n there is an E_n such that $A = f^{-n}E_n$, then we have by (B5) and (B10)

$$\begin{aligned} \frac{\lambda(A \cap I^{(n)})}{\lambda(A)\lambda(I^{(n)})} &= \frac{\lambda(f^{-(n+q)}E_{n+q} \cap I^{(n)})}{\lambda(f^{-(n+q)}E_{n+q})\lambda(I^{(n)})} \\ &\geq \frac{1}{D}\frac{\lambda(f^{-(n+q)}E_{n+q} \cap I^{(n)})}{\lambda(E_{n+q})\lambda(I^{(n)})} \geq \frac{1}{D^2} > 0 \end{aligned}$$

so

$$D^2\lambda(A \cap I^{(n)}) \geq \lambda(A)\lambda(I^{(n)}).$$

Since $||\mathscr{P}^{(n)}|| \to 0$, we can approximate any set—in particular the complement A^c of A—by a union of disjoint n th stage intervals from which we conclude

$$0 = D^2\lambda(A \cap A^c) \geq \lambda(A)\lambda(A^c) - \varepsilon$$

for any $\varepsilon > 0$. Hence $\lambda(A) = 0$ or $\lambda(A^c) = 0$.

Proof of Theorem B1. From (B10) we observe that λf^{-n}, $n \geq q$, is equivalent to λ, and its Radon-Nikodym derivative satisfies

$$1/D \leq \frac{d(\lambda f^{-n})}{d\lambda}(y) \leq D \quad \text{a.e.}$$

From the change of variables formula, $\lambda f^{-n}(E) = \int_E h_n(y)d\lambda$ where $h_n(y) = \sum_{x \in f^{-n}(y)} |f^{n\prime}(x)|^{-1}$.

As made evident by Figure B1, $f^{-n}(y)$ is a finite set whose cardinality is constant for $y \in \operatorname{int} I$, $I \in \mathscr{P}$. The density function h_n is therefore continuous except possibly at end points of intervals of $\mathscr{P}$. The function $h_n = d(\lambda f^{-n})/d\lambda$ a.e., and hence by continuity,

$$1/D \leq h_n(y) \leq D, \tag{B11}$$

for $y \in \operatorname{int} I^{(0)}$. Also for $y \in \operatorname{int} I^{(0)}$, the derivative of h_n exists

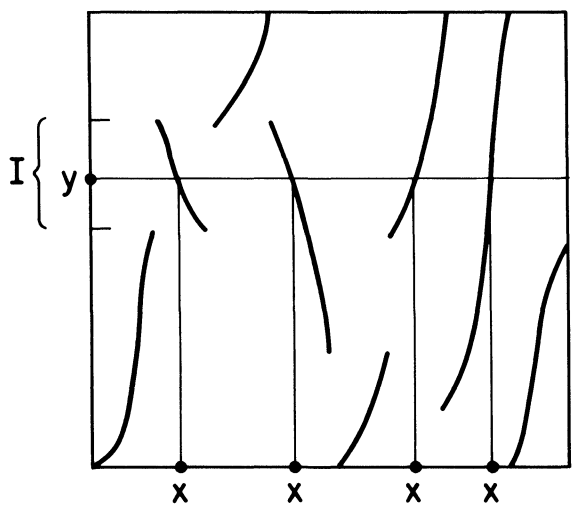


FIGURE B1. Graph of $y = f^n(x)$

and

$$\left|\frac{d}{dy}h_n(y)\right| = \left|\sum_{x\in f^{-n}(y)} |f^{n\prime}(x)|\frac{d}{dx}|f^{n\prime}(x)|\right|.$$

By (B4) and (B11)

$$\left|\frac{d}{dy}h_n(y)\right| < cD.$$

We introduce the following average,

$$S_n(y) = 1/n\sum_{k=0}^{n-1} h_k(y),$$

the advantage of which will become apparent. S_n obeys the same inequalities as h_n: namely, $1/D \le S_n(y) \le D$ and $|S^{n\prime}(y)| \le cD$ for $y \in \operatorname{int} I^{(0)}$. Therefore, $\{S_n(y)\}$ is an equicontinuous family of functions uniformly bounded in n on any $I^{(0)}$. By the Ascoli-Arzela theorem, there is a convergent subsequence $S_{n_i}(y)$ which converges uniformly to a function $h(y)$ which is continuous everywhere except possibly at endpoints of intervals $I^{(0)}$. Define $\mu(E) = \int_E h(y)d\lambda$. We have

$$\begin{aligned}\mu(E) &= \lim(1/n_i)\sum_{k=0}^{n_i-1}\lambda(f^{-k}E)\\ &= \lim(1/n_i)\sum_{k=1}^{n_i}\lambda(f^{-k}E) = \mu(f^{-1}E),\end{aligned}$$

i.e. the f-invariance of μ. The measure of μ also obeys $\lambda(E)/D \le \mu(E) \le D\mu(E)$ which implies that it is equivalent to λ. We remark that actually $\mu(E) = \lim\lambda(f^{-n}E)$ because

$$\lambda(f^{-n}E) = \int_{f^{-n}E}\frac{d\lambda}{d\mu}\,d\mu = \int_{f^{-n}E} E\left(\frac{d\lambda}{d\mu}|f^{-n}\mathbf{B}\right)d\mu$$

where $\mathbf{B}$ is the Borel field of measurable sets, and the Martingale convergence theorem and Lemma B4 imply that the integrand goes to 1 as $n \to \infty$.

Finally, μ is ergodic since invariant sets are tail sets, these have Lebesgue measure either 0 or 1, and hence μ-measure either 0 or 1.

A rate can be obtained directly for the convergence of $\lambda f^{-n}(E)$ to $\mu(E)$; but this takes much more intricate estimates than above.

These can be found in [MVP] and are derived from results of Doeblin [D] for continued fractions.

Proof of Theorem B2. Let φ be the homeomorphism of X onto itself given by $\varphi(x) = \mu([0, x]) = \int_0^x \frac{d\mu}{d\lambda} d\lambda$, and define $g(x) = \varphi f \varphi^{-1}(x)$. The function g preserves the measure $\mu\varphi^{-1}$. However, $\mu\varphi^{-1} = \lambda$ because $\lambda\varphi[0, x] = \lambda[0, \mu(0, x)] = \mu(0, x)$ which extends to $\lambda\varphi(E) = \mu(E)$ for all E.

By the fundamental theorem of calculus, φ is differentiable except at end points of intervals $I^{(0)}$, and $\varphi'(x) = d\mu/d\lambda(x)$ for $x \in \operatorname{int} I^{(0)}$. On $\operatorname{int} I^{(0)}$, φ is C^1 with $\varphi' \geq 1/D > 0$. Thus g satisfies conditions (i)–(iv) with respect to $\varphi\mathscr{P}$ except that in (i) C^2 is to be replaced by C^1.

Since g preserves Lebesgue measure so does g^q. Therefore, by the change of variables formula

$$1 = \sum_{x \in g^{-q}(y)} 1/|g^{q\prime}(x)|$$

for $y \in \operatorname{int} \varphi I^0$. From (4) we have $g^q(\varphi \overline{I}^0) = \overline{X}$ so that the cardinality of $g^{-q}(y) \geq N \geq 2$. Hence $1 \geq 1/|g^{q\prime}(x)| + 1/\sup|g^{q\prime}(x)|$; or in other words $\inf|g^{q\prime}(x)| > 1$ for $x \in \operatorname{int}\varphi I^{(q)}$.

From hypothesis

$$1/D \leq |\varphi'| \leq D$$

on $\operatorname{int} I^{(0)}$, so

$$\begin{aligned} |f^{nq\prime}(x)| &= |g^{nq\prime}(\varphi x)|\,|\varphi'(x)|/|\varphi'(f^{nq}(x))| \\ &\geq (\inf|g^{q\prime}(x)|)^n/D^2 \end{aligned}$$

for $x \in \operatorname{int} I^{(nq)}$. Therefore, we can choose n large enough so that $\inf|f^{nq\prime}(x)| > 1$ for $x \in \operatorname{int} I^{(nq)}$. We can extend this to right and left hand derivatives at end points of $I^{(nq)}$.

Appendix C. Symbolic dynamics

This appendix is split into two parts. The first describes several symbolic systems pertinent to this paper. The second applies these to the study of abstract dynamical systems.

I. **Symbolic dynamical systems.** The main aim of this section is to show that the symbolism which we use for rectilinear maps belongs to a class of symbolic systems called strictly sofic (Corollary to Theorem C7).

We begin with a description of the full N-shift. Let $\mathscr{A}$ be an *alphabet*, that is a finite set of N symbols: we label these $0, 1, \ldots, N-1$. The *full N-shift* (Σ_N, σ) consists of the space Σ_N of all bi-infinite sequences $s = (\cdots, s_{n-1}, s_n, s_{n+1}, \cdots)$, $s_n \in \mathscr{A}$, $n \in \mathbb{Z}$, upon which the *shift transformation* σ acts by shifting each coordinate one step to the left: i.e., $(\sigma(s))_n = s_{n+1}$ for $n \in \mathbb{Z}$.

We develop some of the basic theory for bi-infinite sequences, but much of it can be modified to fit the *one-sided shift system* (Σ_N^+, σ^+). Here Σ_N^+ is the space of all sequences $s^+ = (s_1, s_2, \ldots)$, $s_n \in \mathscr{A}$, acted upon by the shift $\sigma^+ : (s_1, s_2, \ldots) \to (s_2, s_3, \ldots)$.

We endow the sequence space Σ_N with a topology so that notions needed in symbolic dynamics have a dual description: namely, they can be stated either in combinatorial or topological terms. Even though combinatorial ones are usually more intuitive, the topological are often more efficient.

We define the *distance* between two distinct sequences to be $1/(|n|+1)$ where n is the coordinate of smallest modulus at which they differ. This means that the more two sequences agree around the central coordinate the closer they are. Some elementary consequences of this distance function is that Σ_N is a compact metric space (the Cantor discontinuum), the shift a homeomorphism, and *finite cylinder sets*—i.e. sets of sequences where a finite set of coordinates are specified—are closed-open sets. There are a countable number of these cylinder sets and they generate the topology. Furthermore, any closed-open subset is compact and hence the union of a finite number of finite cylinder sets.

A *subshift* (Σ, σ) is defined as a shift-invariant subspace Σ of some Σ_N together with the restriction to Σ of σ. Unless otherwise specified Σ is assumed to be closed, hence compact. The combinatorial description of Σ is as follows. The complement of Σ is also shift invariant and is a countable union of finite cylinder sets. Without loss of generality we can assume that the cylinder sets in question are specified by consecutive coordinates. We call a finite string of k consecutive symbols a *k-block*. A k-block plus a location specifies such a finite cylinder set. However, without reference to location, it specifies the union of all shifts of a finite cylinder set. Thus a countable shift-invariant union of finite cylinder sets corresponds to a countable list of finite k-blocks. We call members of this list *forbidden blocks*. So Σ consists of those

sequences that do not contain any forbidden block. This twin description of Σ—as a closed shift invariant subset of Σ_N or as specified by a list of forbidden blocks—is a sample of the duality between topology and combinatorics mentioned above.

A subshift (Σ, σ) is said to be a *subshift of finite type* (SFT) if the list of forbidden blocks is finite. Stated another way, the complement of Σ is the union of all shifts of a finite collection of finite cylinder sets. As a simple illustration of a subshift of finite type, we mention the *Fibonacci shift.* It is defined as a subshift Σ of Σ_2 consisting of the set of all sequences of 0's and 1's with no two successive 1's.

Given two subshifts $(\Sigma, \sigma), (\Sigma', \sigma')$ we call a map π of Σ onto Σ' a *homomorphism* if it is continuous and shift commuting—i.e., $\sigma'\pi = \pi\sigma$ as illustrated in Figure C1.

The second shift system is called a *factor* of the first and the first an *extension* of the second. We call π an *isomorphism* if additionally it is a homeomorphism, in which case we have the relation $\sigma' = \pi\sigma\pi^{-1}$ called *topological conjugacy.*

Again as an illustration of duality, a homomorphism π can be described combinatorially as a *sliding k-block map* by which we mean the following. Let $(\ldots, y_n, y_{n+1}, \ldots) = \pi(\ldots, x_n, x_{n+1}, \ldots)$. Because π is continuous, the inverse image of a cylinder set specified by a single coordinate is a closed-open subset, hence the union of a finite number of finite cylinder sets. This implies that there exists functions f_n of k_n variables,

$$k_n = m_n + a_n + 1, \quad \text{where } m_n, a_n \geq 0,$$

such that $y_n = f_n(x_{n-m_n}, \ldots, x_{n+a_n})$ for all $n \in Z$. Because π commutes with the shift, the functions f_n and the parameters k_n, m_n, a_n do not vary with n. Economizing on notation we abuse it slightly and write $f_n = \pi$.

The space Σ_N can be thought of as all bi-infinite walks on the complete directed graph of N vertices which are distinctly labeled. The case $N = 2$ is illustrated in Figure C2.

$$\begin{array}{ccc} \Sigma & \xrightarrow{\sigma} & \Sigma \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ \Sigma' & \xrightarrow[\sigma']{} & \Sigma' \end{array}$$

FIGURE C1. Commutativity diagram.

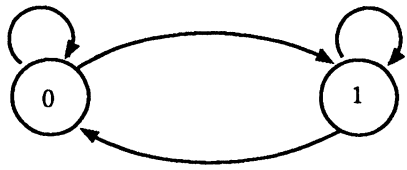


FIGURE C2. Full 2-shift (vertex labeled).

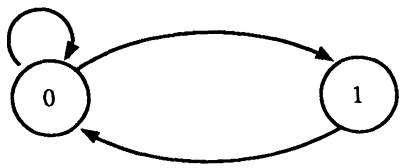


FIGURE C3. Fibonacci shift.

A *topological Markov shift* (TMS) is defined as the shift restricted to the set of bi-infinite walks on a subgraph derived from a complete directed one by possibly removing some edges. A directed subgraph is specified by a matrix $A = (a_{ij})$ where $a_{ij} = 1$ or 0 depending on whether or not the ith vertex is connected to the jth. We denote a topological Markov shift so specified by (Σ_A, σ). We write $i \to j$ if $a_{ij} = 1$ and say that Σ_A is given by the set of 1-*step admissibility rules* $i \to j$. A topological Markov shift (Σ_A, σ) is a subshift of finite type, the list of forbidden blocks being the 2-blocks $[i, j]$ where $a_{ij} = 0$.

Again the Fibonacci shift is a simple example of a topological Markov one.

As depicted in Figure C3, its graph is obtained by removing an edge from the one of Figure C2. The matrix which specifies this graph is

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix},$$

and its 1-step admissibility rules are $0 \to 0$, $0 \to 1$, and $1 \to 0$. For economy we combine the first two rules and write $0 \to 0, 1$.

As just stated, a topological Markov shift can be defined in terms of a 0, 1 matrix. In fact any nonnegative integer matrix serves to determine a topological Markov shift. This is done by giving distinct labels to edges instead of vertices in a graph. The elements a_{ij} then specify the number of edges from the ith vertex to the jth. For example, exhibited in Figure C4 on p. 314 is an edge labeled graph for the full 2-shift given by the one-by-one matrix $A = (2)$.

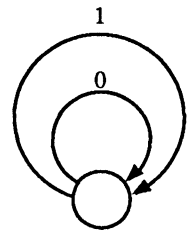


FIGURE C4. Full 2-shift (edge labeled).

It is easily checked that the symbolic system (Σ, σ) corresponding to the vertex label of a graph is conjugate to the symbolic system corresponding to the edge label of a graph. Indeed, the latter is just the higher 2-block presentation of (Σ, σ) defined below.

We give a series of theorems about the concepts SFT and TMS culminating in Theorem C5 which states that the class of SFT's is invariant under isomorphism. Theorem C5 does not hold for the class TMS. Indeed SFT is precisely the class of subshifts isomorphic to TMS, this fact following from Theorems C2 and C5. It is useful to introduce the notion of the *higher k-block presentation* $\Sigma^{(k)}$ of Σ. Informally, $\Sigma^{(k)}$ is obtained by reading k symbols of Σ at a time: i.e. we read x_n, $x_{n+1}, \ldots, x_{n+k-1}$, then we read $x_{n+1}, \ldots, x_{n+k}$, etc. We give a formal definition, seemingly pedantic but which facilitates the proofs of theorems to come. We first define the *k-block presentation* $(\Sigma_N^{(k)}, \sigma)$ *of the full N-shift.* This is the topological Markov shift with alphabet consisting of all k-blocks $[a_1, a_2, \ldots, a_k]$ of symbols a_i from the alphabet for Σ_N, and with admissiblity rules $[a_1, a_2, \ldots, a_k] \to [b_1, b_2 \ldots, b_k]$ if and only if $b_1 = a_2, \ldots, b_{k-1} = a_k$. Observe that the k-block presentation $(\Sigma_N^{(k)}, \sigma)$ of the full shift on N symbols is a proper subshift of the full N^k-shift (Σ_{N^k}, σ).

Next we define the k-block injection φ from Σ_N into Σ_N, by $\varphi(x_n, x_{n+1}, \ldots, x_{n+k-1}) = [x_n, x_{n+1}, \ldots, x_{n+k-1}]$, where the latter is the nth coordinate of $\varphi(\ldots, x_n, x_{n+1}, \ldots)$. Thus $\Sigma_N^{(k)} = \varphi(\Sigma_N)$. We call φ the *canonical injection* and observe that it has a 1-block inverse $x_n = \varphi^{-1}([x_n, x_{n+1}, \ldots, x_{n+k-1}])$ from $\Sigma_N^{(k)}$ to Σ. The *higher k-block presentation* $(\Sigma^{(k)}, \sigma)$ for a general subshift (Σ, σ) is defined by $\Sigma^{(k)} = \varphi(\Sigma)$. The following inclusions summarize what we have just said: if $\Sigma \subset \Sigma_N$, then $\varphi(\Sigma) = \Sigma^{(k)} \subset \Sigma_N^{(k)} \subset \Sigma_{N^k}$. (Σ, σ) is conjugate to $(\Sigma^{(k)}, \sigma)$ for all $k > 0$.

Higher block presentations lead to the following theorem which is immediate from their definition. It is useful because in many arguments it enables us to replace k-block maps by 1-block ones.

Theorem C1. *Every k-block map $\pi: \Sigma \to \Sigma'$ gives rise to a 1-block map $\pi^{(k)}: \Sigma^{(k)} \to \Sigma'$ given by $\pi^{(k)} = \pi\varphi^{-1}$ where φ is the canonical injection from Σ onto $\Sigma^{(k)}$.*

As mentioned earlier, TMS's are SFT's. Although SFT's are not necessarily Markov, they are conjugate to Markov as shown below.

Theorem C2. *If (Σ, σ) is a subshift of finite type, then $(\Sigma^{(k)}, \sigma)$ is a topological Markov shift for all sufficiently large k.*

Proof. Let (Σ, σ) be a SFT. Choose a fixed integer $k \geq l - 1$ where l is the length of the longest forbidden block. One merely has to observe that the k-block presentation $(\Sigma^{(k)}, \sigma)$, which is conjugate to (Σ, σ) via the canonical injection, is specified by the following 1-step admissibility rules: one k-block follows another if and only if

1. the last $k - 1$ symbols of the predecessor block coincides with first $k - 1$ of the successor,
2. the $(k+1)$-block, formed by appending the last symbol of the successor block to the predecessor k-block, does not contain a forbidden sub block.

Theorem C3. *A subshift is of finite type if and only if any of its higher block presentations is also a subshift of finite type.*

Proof. Both the injection ϕ and its inverse map finite cylinder sets to closed-open sets—i.e. to finite unions of finite cylinder sets. Thus the complement of Σ in Σ_N is generated by a finite number of finite cylinder sets if and only if the same holds for the complement of $\Sigma^{(k)}$ in $\Sigma_N^{(k)}$. In addition, since $\Sigma_N^{(k)}$ is a topological Markov shift, its complement in Σ_{N^k} is generated by a finite number of finite cylinder sets. It follows that the complement of $\Sigma^{(k)}$ in $\Sigma_N^{(k)}$ is generated by a finite number of finite cylinder sets if and only if the same holds for the complement of $\Sigma^{(k)}$ in Σ_{N^k}.

It follows from the definition that higher block presentations of Markov shifts remain Markov, but in general these shifts are not preserved under topological conjugacy. However, for 1-block

isomorphism we have

Theorem C4. *If a subshift* (Σ, σ) *is isomorphic to a topological Markov shift* (Σ_A, σ) *via a* 1*-block isomorphism* π *from* Σ *to* Σ_A, *then* (Σ, σ) *also is a topological Markov shift.*

Proof. Suppose π^{-1} is an m-block map. It follows that a symbol j follows i in a sequence of Σ if and only if

1. $\pi(i) \to \pi(j)$ is admissible in Σ_A,
2. there is an $(m+1)$-block $(s_1, \ldots, s_m, s_{m+1})$ of Σ_A such that $\pi^{-1}(s_1, \ldots, s_m) = i$ and $\pi^{-1}(s_2, \ldots, s_{m+1}) = j$. In other words Σ is specified by a 1-step rule.

Theorem C5. *If a subshift* (Σ, σ) *is topologically conjugate to a subshift of finite type* (Σ', σ), *then* (Σ, σ) *is* SFT.

Proof. By Theorem C2, there is an isomorphism from Σ' to the space Σ_A of some topological Markov shift. Combining this isomorphism with that assumed in the hypothesis of the theorem, we have an isomorphism ψ of Σ onto Σ_A, ψ being a k-block map for some k. According to Theorem C1, for ϕ the canonical injection from Σ to $\Sigma^{(k)}$, $\psi\phi^{-1}$ is a 1-block isomorphism of the k-block presentation $\Sigma^{(k)}$ onto Σ_A. Hence $(\Sigma^{(k)}, \sigma)$ is a TMS by Theorem C4, and (Σ, σ) a SFT by Theorem C3.

We can consider a larger more complicated class of dynamical systems than SFT's but one that still retains a strong finitary character. *A sofic system* is defined as a subshift that is a factor of some subshift of finite type. By Theorems C1 and C2, we can assume it to be a factor under a 1-block map of a topological Markov shift. Thus one can view a sofic system as the shift acting on the space of bi-infinite walks on a directed graph, the vertices (edges) of which are not necessarily distinctly labeled. SFT's are clearly sofic systems. But the converse is not true. A sofic system which is not a SFT is referred to as *strictly sofic.*

An example of a strictly sofic system is given in Figure C5 depicting the sub-shift that consists of the set of sequences consisting

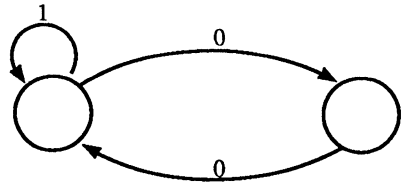


FIGURE C5. Even system.

of 0's and 1's with only even run-lengths of 0's. This system of sequences is called the *even system.* It can be shown directly or as a consequence of Theorem C7 that it is not of finite type. Yet the system retains a certain strong finitary character as revealed in the graph. Another system which in some sense has a similar finitary description is given by allowing only prime run-lengths of 0's. However as a result of Theorem C6 below, it turns out not to be sofic. These examples are to be constrasted with the system of curvilinear sequences encountered in this work which appear to defy any sort of finitary description.

Since SFT's are properly included among sofics, it becomes natural to distinguish those sofics which are SFT's from those which are not. For this a characterization of "sofic" is needed which does not depend on a graph representation. This is done with the notion of "follower set" defined as follows. Given a left infinite sequence $s^- = (\dots, s_{-2}, s_{-1})$ we call a right infinite sequence $s^+ = (s_0, s_1, \dots)$ a *follower sequence of* s^- if the bi-infinite sequence $(\dots, s_{-1}, s_0, s_1, \dots)$ is admissible—i.e. a member of the symbolic system. The *follower set* of s^- is defined to be the set $F(s^-)$ of all its follower sequences. A *sofic system* can be defined as a subshift in which there are only a finite number of different follower sets (this is not to be confused with the fact that each follower set usually contains an infinite number of elements).

Theorem C6. *The two definitions of sofic are equivalent.*

Proof. According to the first, we can assume as we have done before that a sofic system is a 1-block homomorphic image of a topological Markov shift. Each follower set in a topological Markov shift is the set of all one-sided paths from a vertex. Since there are only a finite number of vertices, there are only a finite number of follower sets in a topological Markov shift. The follower sets in the sofic system are just the images of follower sets in the topological Markov shift under the 1-block map. There are only a finite number of these as well. Therefore, the first definition implies the second.

For the converse, suppose we have a sofic system according to the second definition. We construct a directed graph as follows. Let the vertices be the pairs $(s_{-1}, F(s^-))$ where s_{-1} is the right most symbol in a left infinite sequence s^-. By virtue of the second definition the number of vertices is finite. Draw an edge from vertex $(s_{-1}, F(s^-))$ to $(t_{-1}, F(t^-))$ if $F(s^-)$ contains

a right infinite sequence whose initial symbol is t_{-1}, and $t^- = (\ldots, s_{-2}, s_{-1}, t_{-1})$. The map $\pi(s_{-1}, F(s^-)) = s_{-1}$ defines a 1-block homomorphism from a topological Markov shift onto the sofic system in question.

We apply theorem C6 to the *prime system*—namely, the subshift consisting of sequences in which the finite run-lengths of 0's are prime. There are arbitrarily long runs of composite integers as substantiated by the numbers $n! + 2, \ldots, n! + n$. Thus for every positive integer n there exists a prime number $p(n)$ such that the gap to the next prime $q(n)$ is $> n$. The follower set of a left infinite sequence which ends with a run of $p(n) + 1$ zeros consists of right infinite sequences which begin with runs of m zeros, $m \geq n$. This follower set includes sequences which begin with a run of $q(n) - p(n) - 1$ zeros. The follower set of a left infinite sequence ending in $p(q(n) - p(n)) + 1$ zeros then begins with a run of m zeros, $m \geq q(n) - p(n)$, so that it is different from the previous one. In this manner one exhibits an infinite number of different follower sets.

Theorem C7. *A sofic system is a subshift of finite type, if and only if there exists an n such that $F(s^-) = F(t^-)$ whenever $(s_{-n}, \ldots, s_{-1}) = (t_{-n}, \ldots, t_{-1})$. A strictly sofic system is, therefore, one for which no such n exists.*

Proof. If such an n exists, then a string of n symbols in the sofic system determines a unique vertex on the graph in the proof of Theorem C6. This means the homomorphism π is an isomorphism, its inverse being an n-block map. By Theorem C5, the sofic system is a subshift of finite type. We omit the proof of the converse which is easy.

We apply Theorem C7 to show that the even system is strictly sofic because no such n exists. This follows from the fact that for arbitrarily large n there exist two left infinite sequences among an infinitude of such, which end in a run of 0's of different parity and agree at the right-most n coordinates. Such a pair have different follower sets.

Corollary. *The rectilinear system $\overline{\Omega}_a(T_R)$ defined in* IV *of* §8 *is stricly sofic.*

Proof. Consider left infinite sequences ending in the block $j, \alpha\vartheta(i), \ldots, \alpha^k\vartheta(i)$. The symbol $\vartheta\alpha^k\vartheta(i)$ is allowed as the first symbol of a right follower sequence according as $j = i$ or not. Since k is

arbitrarily large, Theorem C6 implies that $\overline{\Omega}_a(T_R)$ is not a SFT.

II. Application to abstract dynamical systems

IIa. *The fundamental complication.* The pair (X, T), where T is a mapping of a space X onto itself, is called an abstract dynamical system. Additional topological and measurability structure is usually assumed. A symbolic representation of such a dynamical system is a useful tool in understanding it. By symbolic representation we mean that (X, T) is a factor of a subshift (Σ, σ) under a homomorphism π. However, representing an abstract dynamical system by a symbolic one involves a fundamental complication. We have two desires: we would like a continuous one-to-one correspondence between orbits $T^n x$ of the first and orbits $\sigma^n(\ldots, s_{-1}, s_0, s_1, \ldots)$ of the second; and we want Σ to be a closed set, preferably a subshift of finite type. Unfortunately these two desires are in conflict: constraints placed by topology must be observed. On the one hand a continuous one-to-one correspondence makes X homeomorphic to Σ. On the other hand Σ is totally disconnected while X is often a smooth or piecewise smooth manifold. By sacrificing one-to-one correspondence we can still salvage a satisfactory symbolization of orbits. This is reminiscent of the familiar situation in arithmetic: namely, certain rational numbers have two decimal expansions.

We consider two examples, chosen for their simplicity, as illustrations of a resolution of the conflict in dynamical systems. The first is continuous, the second merely piecewise so. The second, a slight variation of the first, exhibits a further complication to be resolved.

Example C1. Let (X, f) be the noninvertible dynamical system consisting of X, the unit interval with end points 0 and 1 identified, acted upon by the continuous map $f: x \to (2x)$ where (y) denotes the fractional part of y. A natural choice of π is the map from Σ_2^+ to X defined by $\pi(s_1, s_2, \ldots) = (s_1/2 + s_2/4 + \cdots)$. It is readily verified that π is: (i) commuting—i.e. $f\pi = \pi\sigma^+$, (ii) continuous, (iii) onto, (iv) no more thant n-to-1 (in this case $n = 2$), and (v) one-to-one except on a *negligible* set.[5] Thus we have a representation of the dynamical system by a topological

[5]The set in question here consists of binary expansions ending in either an infinite run of zeros or ones. An abstract notion of *negligibility* can be defined both measure theoretically and topologically. For us the notion of *negligible* will turn out to be the nondoubly transitive points defined later on.

Markov shift such that: every point has at least one symbolic representative; there is a finite upper limit to the number of representatives of any point; and every symbolic sequence represents some point.

We call a map π from one abstract dynamical system to another a *homomorphism* if it satisfies properties (i), (ii), and (iii) above. If it further satisfies (iv) and (v) we say it is *boundedly finite-to-one* and *essentially one-to-one.* An isomorphism theory for abstract dynamical systems with continuous T was developed in [AM], based on maps satisfying the above five properties. Unfortunately in the case where T is merely piecewise continuous, matters are not as simple as the next example reveals.

Example C2. Let X be the unit interval without identification of end points. Define

$$f(x) = \begin{cases} (2x), & 0 \leq x < 1 \\ 1, & x = 1. \end{cases}$$

Here f is not continuous but rather piecewise continuous, a situation shared by the Bowen-Series factor map. A natural choice for π is $\pi(s_1, s_2, \ldots) = s_1/2 + s_2/4 + \cdots$. π satisfies (ii)–(v); but we lose (i)—namely, $f(\pi(0, 1, 1, \ldots)) \neq \pi(\sigma(0, 1, 1, \ldots))$.

To restore commutativity, we remove an invariant set of troublesome sequences—namely, all those ending in an infinite run of 1's. This forces us to remove the number 1, which has no pre-image under π, from the unit interval. Hence we must also remove all pre-images of 1 under f which in turn forces us to remove all sequences ending in an infinite run of 0's.

Although the two examples above deal with noninvertible abstract dynamical systems symbolized by a one-sided shift, they also typify the general problem for the invertible case. The rectilinear map T_R of §5 is not a homeomorphism: it is only piecewise continuous. So we are in the situation of Example C2. In a manner similar to this example, in order to retain a topological Markov shift representation for T_R we must remove from the subshift an invariant continuum of points from a set which has an otherwise simple description. Although the description of the removed set is somewhat tedious, fortunately for the rectilinear maps it can be completely specified. The map T_R is piecewise continuous on rectangles, the closed side of one abutting on the open side of the next. As shown in §8 this property of T_R implies the situation is

not quite as bad as that of Example C2. Namely, even after removal of symbolic sequences, each point of X still has at least one symbolic representative. So we do not have to remove an invariant set of points from the space X as well.

IIb. *Markov partitions.* As mentioned in detail in §8, one associates a symbolic sequence with elements of a dynamical (X, T) by tracking the history of an orbit through a partition $\mathscr{P} = \{R_a : a \in \mathscr{A}\}$ of X. In order to get a topological Markov shift representation (Σ_A, σ) one must find a partition which satisfies certain properties. This type of partition is called *Markov*, and the existence of a boundedly finite-to-one, essentially one-to-one homomorphism is a consequence of its properties.

Before giving a very general version of the theory of Markov partitions, we reexamine some dynamical systems for examples of them. In example C1, we have an expansive map f and a partition $\mathscr{P} = \{R_0 = [0, 1/2],\ R_1 = [1/2, 1]\}$. The elements of this partition are closed sets equal to the closure of their interiors and which overlap only at boundary points. Observe we are deviating somewhat from the introduction and §8 where the sets of the partition were assumed to be strictly disjoint. The map π of that example has an alternate expression in terms of the partition: namely,

$$\pi(s_1, s_2, \ldots) = \bigcap_{n=0}^{\infty} \overline{R^o_{s_1} \cap f^{-1}(R^o_{s_2}) \cap \cdots \cap f^{-n}(R^o_{s_{n+1}})}$$

where $R^o_{s_i}$ denotes the interior of R_{s_i}.

The partition $\mathscr{P}$ is a prototype of a Markov one for a continuous noninvertible map. The partitions encountered in example C2 and in the hypothesis of the folklore theorem of Appendix B are samples of ones for piecewise continuous noninvertible maps; but commutativity might fail unless certain symbolic sequences and perhaps their images are removed from the discussion. The first Markov partitions for diffeomorphisms were constructed by Berg [Be]. He discovered them for hyperbolic automorphisms of the 2-torus. A simple construction of partitions for these maps can be found also in the work of Adler and Weiss [AW]. The baker's transformation on the unit square has a Markov partition—namely, the partition $\mathscr{P}$ consisting of the two sets exhibited in Figure 1.6. This map is invertible but only piecewise continuous. The rectilinear map is another example of this kind, the finer partition of

§8 being Markov. Both these cases involve the removal of certain symbolic sequences in the associated symbolic extension.

The partition for a topological Markov shift that consists of elementary cylinder sets determined by fixing the 0 th coordinate is archetypal of Markov ones. Abstracting certain properties of this partition yields a satisfactory general theorem. We treat the case of invertible dynamical systems. An analogous discussion holds for noninvertible ones.

Let (X, T) be an abstract dynamical system, where X is a compact metric space with distance function d, and T a homeomorphism.

The following property of T plays a key role in the theory of Markov partitions, and we assume it throughout except in Theorem C10.

(1) T is *expansive*: by which we mean there exists $c > 0$ such that if $d(T^n x, T^n y) < c$ for all $n \in \mathbb{Z}$ then $x = y$.

We call a family of sets $\mathscr{P} = \{R_0, R_1, \ldots, R_{N-1}\}$ a *Markov partition* for the pair (X, T) if it satisfies:

(2) $\mathscr{P}$ covers X—i.e. $X = R_0 \cup R_1 \cup, \ldots, \cup R_{N-1}$;

(3) $R_i = \overline{R_i^o}$;

(4) $R_i^o \cap R_j^o = \varnothing$, $i \neq j$;

(5) $\lim_{n\to\infty}$ diameter $(\bigcap_{-n}^{n} T^{-k} R_{s_k}^o) = 0$;

(6) $R_{s_i}^o \cap T^{-1} R_{s_{i+1}}^o \neq \varnothing$, $-n \leq i \leq n-1 \Rightarrow T^n R_{s_{-n}}^o \cap \cdots \cap T^{-n} R_{s_n}^o \neq \varnothing$, for $n > 1$.

Conditions (2), (3), and (4) are about the R_i's and (5) and (6) relate the map T to the partition. (6) is called the *Markov* property. It is crucial for getting a TMS representation of a dynamical system. Establishing this property is usually nontrivial. A useful concept for this purpose is that of *local product structure.* We give an informal discussion of this concept. Call the R_i's abstract rectangles and think of these to be striated by two families of sets, which we call "vertical" and "horizontal" fibers. Each vertical of a rectangle intersects each horizontal in a unique point of the rectangle, and each point of a rectangle is the intersection of a unique vertical and horizontal. Furthermore, each rectangle is canonically homeomorphic to the cartesian product of a vertical and horizontal fiber, hence the term "local product structure." Under the action of T the vertical fibers are contracting sets in the sense that the image of each is contained in the vertical fiber of

possibly another rectangle, and the diameter of the fiber shrinks to zero under iteration of T. Similarly, the horizontal fibers are expanding sets in the sense that the image of each is a finite union of horizontal ones in different rectangles. Their diameters shrink to zero under the iteration of T^{-1}.

Figure C6 illustrates how an image of a rectangle typically intersects another rectangle in a Markov partition. Thus, if TR_i meets R_j^o then the image of each horizontal fiber of R_i "goes across" and does not "end" in the interior of R_j. If $x = \bigcap_{n\in\mathbb{Z}} T^{-n}R_{s_n}$, then the vertical fiber through x is identified with $\bigcap_{n\geq 0} T^{-n}R_{s_n}$, and the horizontal with $\bigcap_{n\leq 0} T^{-n}R_{s_n}$. The properties of vertical and horizontal fibers under the action of T, as assumed by local product structure, lead to the Markov property (6). We remark that in §8 we got (6) without explicitly mentioning local product structure even though it occurs there implicitly.

For a given Markov partition $\{R_0, \ldots, R_{N-1}\}$, we define a symbolic system Σ_A with alphabet $\{0, \ldots, N-1\}$ and admissibility rules $i \to j$ iff $R_i^o \cap T^{-1}R_j^o \neq \varnothing$. Thus

$$\Sigma_A = \{s = (\ldots, s_{-1}, s_0, s_1, \ldots): R_{s_n}^o \cap T^{-1}R_{s_{n+1}}^o \neq \varnothing,\ n \in \mathbb{Z}\}.$$

We also define a map $\pi: \Sigma_A \to X$ by

$$\pi(\ldots, s_{-1}, s_0, s_1, \ldots) = \bigcap_{n=0}^{\infty} \overline{T^n R_{s_{-n}}^o \cap \cdots \cap T^{-n}R_{s_n}^o}.$$

That $\pi(s)$ is a well-defined point of X follows from compactness, (5), and (6).

We shall require the following notion in order to define a negligible set on which π is invertible. A *doubly transitive point* is defined as one whose forward orbit and whose backward orbit are

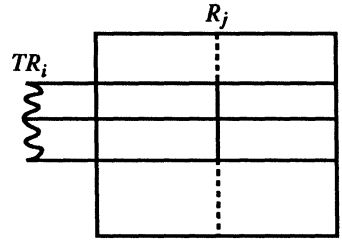


FIGURE C6. Image in a partition.

both everywhere dense. As we shall see, the set of nondoubly transitive points is appropriately negligible.

Theorem C8. *Under the assumptions* (1) *to* (6) *the map* π *is a boundedly finite-to-one homomorphism for which each doubly transitive point has a unique pre-image.*

Proof. (i) π *satisfies the commutativity property* $\pi\sigma = T\pi$. This follows immediately from the definition of π and the fact that T is a homeomorphism.

(ii) π is *continuous*: *hence uniformly continuous.* This follows from (5). In fact (5) can be strengthened to a uniform version: namely, from uniform continuity of π we get that for $\varepsilon > 0$ there exists an integer n (depending only on ε) such that diameter $\left(\bigcap_{-n}^{n} T^{-k} R^o_{s_k}\right) < \varepsilon$.

(iii) π *is onto.* Because of (3) and T being a homeomorphism, $\bigcup_{n\in\mathbf{Z}} T^n \partial\mathscr{P}$ is the countable union of closed nowhere dense sets. Therefore, by the Baire category theorem, its complement—namely, the set of points x which can be expressed by $x = \bigcap_{n\in\mathbf{Z}} T^{-n} R^o_{i_n}$—is everywhere dense. From the definition of Σ_A, each such point has the unique pre-image $\{i_n\}$ under π. Thus the range of π is dense in Σ_A. From the compactness of Σ_A and continuity of π, its range is closed, hence all of Σ_A.

We now impose a restriction which we shall later remove.

(7) diameter $(R_i) < c/2$.

We remark that under the assumption (7), (5) is automatically true and furthermore

$$\pi(s) = \bigcap_{n=0}^{\infty} \overline{T^n R^o_{s_{-n}} \cap \cdots \cap T^{-n} R^o_{s_n}} = \bigcap_{n\in\mathbf{Z}} T^{-n} R_{s_n}.$$

(iv) *There is a bound on the number of pre-images under* π. To prove this we first make the following definition. The map π is said to have a *diamond* if there exists s, $t \in \Sigma_A$ and indices $k < l < m$ such that $\pi(s) = \pi(t)$, and $s_k = t_k$, $s_l \neq t_l$, $s_m = t_m$. Observe that if the number of pre-images of a point is more than N^2, then by the "pigeon hole principle" π would have a diamond. So to establish (iv) we prove

(*) π *has no diamonds.* Without loss of generality, we may assume $k = 1$ in the definition of diamond. Assume then that $x = \pi(s) = \pi(t)$ where

$$s = (\ldots, s_{-2}, a, b_0, b_1, \ldots, b_{m-1}, d, s_{m+1}, \ldots),$$

$$t = (\dots, t_{-2}, a, c_0, c_1, \dots, c_{m-1}, d, t_{m+1}, \dots).$$

We must show that $b_l = c_l$ for $0 \le l \le m-1$. Because $[a, b_0, b_1, \dots, b_{m-1}, d]$ is admissible, $TR_a^o \cap R_{b_0}^o \cap T^{-1}R_{b_1}^o \cap \cdots \cap T^{-m+1}R_{b_{m-1}}^o \cap T^{-m}R_d^o \neq \varnothing$. Choose a point y in this open set. Because of (iii) there is a sequence $\eta = (\dots, \eta_{-2}, a, b_0, b_1, \dots, b_{m-1}, d, \eta_{m+1}, \dots) \in \Sigma_A$ such that $\pi(\eta) = y$. Also since $[a, c_0, c_1, \dots, c_{m-1}, d]$ is admissible, $\zeta = (\dots, \eta_{-2}, a, c_0, c_1, \dots, c_{m-1}, d, \eta_{m+1}, \dots) \in \Sigma_A$. Thus $\pi(\zeta) = z \in TR_a \cap R_{c_0} \cap T^{-1}R_{c_1} \cap \cdots \cap T^{-m+1}R_{c_{m-1}} \cap T^{-m}R_d$. Because of (7) and $T^l(x) \in R_{b_l} \cap R_{c_l}$ for $0 \le l \le m-1$, we conclude by the triangle inequality that $d(T^l y, T^l z) < c$. Furthermore, $d(T^n y, T^n z) < c/2$ for $n < 0$ or $n > m-1$. Therefore by (1), $y = z$. Thus $R_{b_l}^o \cap R_{c_l} \neq \varnothing$, and we conclude from (3) and (4) that $b_l = c_l$.

(v) *A doubly transitive point has a unique pre-image.* As remarked in the proof of (iii), any point whose T-orbit does not meet $\partial\mathscr{P}$ has a unique pre-image under π. We show that orbits of doubly transitive points do not meet $\partial\mathscr{P}$. The orbit of a doubly transitive point meets the interior of an element of $\mathscr{P}$ infinitely often in the past and future. Consequently, no part of its orbit can ever lie in $\partial\mathscr{P}$: for then there would be a diamond.

In order to remove the restriction imposed by (7) we form a new partition

$$\mathscr{P}^{(n)} = \{\overline{T^n R_{s_{-n}}^o \cap \cdots \cap T^{-n}R_{s_n}^o} : [s_{-n}, s_{-n+1}, \dots, s_n] \text{ a finite admissible block}\}.$$

From the fact that π is onto, we have (2)—namely, $\mathscr{P}^{(n)}$ is a cover. To get (3), (4) and (6), we need the following:

Lemma. *If* $A_i = \overline{A_i^o}$, $i = 1, \dots, N$, *then*

$$A_i^o \cap \cdots \cap A_N^o = (\overline{A_1^o \cap \cdots \cap A_N^o})^o.$$

Proof. Clearly $A_1^o \cap \cdots A_N^o \subset (\overline{A_1^o \cap \cdots \cap A_N^o})^o$. For the opposite inclusion, choose $x \in (\overline{A_1^o \cap \cdots \cap A_N^o})^o \subset A_1 \cap \cdots \cap A_N$, the inclusion being a consequence of the hypothesis. Suppose $x \notin A_1^o \cap \cdots \cap A_N^o$. Then $x \notin A_i^o$ for some i. Since $x \in A_i$, we have $x \in \partial A_i$ and every neighborhood of x intersects the complement of A_i. However, there exists a neighborhood U of x such that $U \subset (\overline{A_1^o \cap \cdots \cap A_N^o})^o \subset A_i$, a contradiction.

We get (5) from uniform convergence of diameters, since we can choose n so that diameter $(T^n R^o_{s_{-n}} \cap \cdots \cap T^{-n} R^o_{s_n}) < c/2$. Thus $\mathscr{P}^{(n)}$ satisfies (2) through (7). Then $\Sigma^{(2n+1)}$ is the set of symbolic sequences associated with $\mathscr{P}^{(n)}$. Let $\tilde{\pi}$ be the map from $\Sigma^{(2n+1)}$ to X defined by the partition $P^{(n)}$. We have $\pi = \tilde{\pi}\varphi\sigma^n$ where ϕ is the canonical injection, which implies that π has at most $N^{2(2n+1)}$ pre-images.

We can obtain a converse of Theorem D1. To do so we must introduce another concept, one that expresses irreducibility of a dynamical system. A system (X, T) is said to be *topologically transitive nonwandering* if for every two open sets U and V there exists $n \geq 0$ such that $T^n U \cap V \neq \varnothing$. For a topological Markov shift (Σ_A, σ) this is equivalent to the reachability of every node from every other in the graph of A. It is not difficult to show that in a topologically transitive non-wandering dynamical system the set of doubly transitive points is a set of the second category. Its complement is a set of the first category and is the negligible set to which we alluded earlier. In addition to being topologically negligible, it is negligible measure theoretically in the sense that, if T has a finite ergodic invariant measure which is positive on open sets then this set is also one of measure zero.

Let Σ_A be based on a finite symbol set $\mathscr{A}$ and denote by C_a the elementary cylinder set defined by $C_a = \{(\ldots, s_{-1}, s_0, s_1, \ldots): s_0 = a\}$ for $a \in \mathscr{A}$. To obtain a converse to Theorem C8, we introduce the additional property:

(vi) (Σ_A, σ) is topologically transitive nonwandering.

Theorem C9. (i), (ii), (iii), (v), *and* (vi) *imply that the partition* $\mathscr{P} = \{R_a = \pi C_a : a \in \mathscr{A}\}$ *satisfies* (2), (3), (4), (5) *and* (6).

We omit the proof. As an interesting by-product of Theorems C8 and C9 we obtain

Corollary. (1), (i), (ii), (iii), (v), *and* (vi) *imply* (iv).

The general theory of Markov partitions for dynamical systems with continuous T can be amended to apply to the expansive piecewise continuous case. Let $\mathscr{R} = \{R_0, R_1, \ldots, R_{N-1}\}$ and $\mathscr{S} = \{S_0, S_1, \ldots, S_{N-1}\}$ be two families of sets satisfying the following:

(2′) $\mathscr{R}$ and $\mathscr{S}$ cover X;

(3′) $R_i = \overline{R_i^o}$ and $S_i = \overline{S_i^o}$;

(4′) $R_i^o \cap R_j^o = S_i^o \cap S_j^o = \varnothing$, $i \neq j$.

Theorem C10. *Let T be a one-to-one mapping of X onto itself such that $T|R_i^o$ is a homeomorphism of R_i^o onto S_i^o which has a continuous extension, denoted by T_i, to R_i. Then* (1), (2′), (3′), (4′), (5), *and* (6) *imply that the map π satisfies* (i′), (ii), (iii), *and* (v′), *where* (i′) $\pi\sigma(s) = T_{s_0}\pi(s)$, *and* (v′) *there is an everywhere dense set of points for which π has a unique pre-image, the points being of the form $x = \bigcap_{n\in\mathbb{Z}} T^{-n}R_{s_n}^o$ and the unique pre-image being $\pi^{-1}(x) = \{s_n\}$.*

We omit the proof, it being a more tedious version of Theorem C9. Observe that we lost overall commutativity because $T(\overline{T^n R_{s_{-n}}^o \cap \cdots \cap T^{-n}R_{s_n}^o})$ may not be equal to $\overline{T(T^n R_{s_{-n}}^o \cap \cdots \cap T^{-n}R_{s_n}^o)}$. However, we retain that $\pi\sigma(s) = T\pi(s)$ for the everywhere dense set of points $x = \pi(s) = \bigcap_{n\in\mathbb{Z}} T^{-n}R_{s_n}^o$.

Our procedure here is at variance with §8. There it was more convenient to choose a partition whose elements were not closed (however, their closures satisfy the hypotheses Theorem C10). Fortuitously, we were able to describe precisely the sequences which had to be removed from a subshift of finite type and where commutativity fails. We got (iv) and (v) directly without recourse to the "no-diamonds" argument. It is not clear how to get (iv) and (v) in the framework of a general theorem.

Markov partitions yield topological Markov shifts, whereas partitions gotten from Markov ones by amalgamation of elements give rise to sofic systems. We call partitions gotten in this manner *sofic*. Conversely, Markov partitions are gotten from sofic ones by refinement. The coarse partition of §8 for the rectilinear map is an example of a sofic one.

We conclude this appendix with some background material concerning symbolic dynamics. Linear algebra, especially the Perron-Frobenius theory of nonnegative matrices, plays an important role in the study of topological Markov shifts. For instance, the ratio of the number of different n-blocks in sequences of Σ_A to λ^n is bounded between two positive constants, where λ is the spectral radius of A. One can prove that the spectral radius is a topological conjugacy invariant, the logarithm of which is called *channel capacity* [Sh] by engineers and *topological entropy* [AM, P] by mathematicians. It can often be used to show that two subshifts

are not isomorphic. R. Williams [Wi] showed that for topological Markov shifts topological conjugacy is equivalent to the existence of nonnegative integer solutions to certain matrix equations. Even so, the main unsolved problem in symbolic dynamics is: does there exist an algorithm to tell when two topological Markov shifts are topologically conjugate?

For an entrance into the relevant current literature on symbolic dynamics see [BKM, BMT]. For some material on Markov partitions see [B1-4, Si, Sh, Ru] and the forthcoming paper [AKS]. Theorems C8, C9 are new results of Kitchens and us, the complete details to be supplied in the future.

Appendix D. Geometric proof of Theorem 9.7

We give here a proof of Theorem 9.7, restated below, which is based on geometric considerations.

Theorem D1. *Let* $T^n(\xi, \eta) \in C_i$. *Then*

$$\begin{aligned} T_C^n(\xi, \eta) \in G_i &\Leftrightarrow n \notin \bigcup B \\ T_C^n(\xi, \eta) \in X_i &\Leftrightarrow n = b_0 \\ T_C^n(\xi, \eta) \in E_i &\Leftrightarrow n = b_1 \\ T_C^n(\xi, \eta) \in D_i \cup F_i &\Leftrightarrow n \in B - \{b_0, b_1\}. \end{aligned}$$

We remark that it suffices to prove Theorem D1 for $n = 0$. For let $(\xi^*, \eta^*) = T^n(\xi, \eta)$. Denote the blocks corresponding to (ξ^*, η^*) by B^* and let b_0^*, b_1^* be the smallest and largest integers in B^*, when these exist. Then $B^* = B - n$, $b_i^* = b_i - n (i = 0, 1)$. We can therefore rewrite Theorem D1 as $(\xi^*, \eta^*) \in G_i$ iff $0 \notin \bigcup B^*$, $(\xi^*, \eta^*) \in X_i$ iff $0 = b_0^*$, etc.

We translate the conditions of Theorem D1 on $n = 0$ into conditions on $\{\delta_n\}$. Let $\{\delta_n\}^+ = \{\delta_n\}$, $1 \leq n < \infty$, and $\{\delta_n\}^- = \{\delta_{-n}\}$, $0 \leq n < \infty$. Let m, p, r, s, t denote integers such that $0 \leq m, p$, $3 \leq r$, $s \leq 8g - 4$, $4 \leq t \leq 8g - 4$. 2^m, 2^∞ denote respectively the finite sequence $2, \ldots, 2$ (m terms) and the infinite sequence $2, 2, \ldots$.

Theorem D2.

$$0 \notin \bigcup B \Leftrightarrow \{\delta_n\}^+ = 2^m, r, \ldots$$
$$\& \{\delta_n\}^- = 2^p, s, \ldots$$
$$0 = b_0 \Leftrightarrow \{\delta_n\}^+ = 2^m, 1, \ldots \text{ or } 2^\infty$$
$$\& \{\delta_n\}^- = t, \ldots \text{ or } 3, 2^p, r, \ldots$$
$$0 = b_1 \Leftrightarrow \{\delta_n\}^- = 2^m, 1, \ldots \text{ or } 2^\infty$$
$$\& \{\delta_n\}^+ = t, \ldots \text{ or } 3, 2^p, r, \ldots$$
$$0 \in B - \{b_0, b_1\}$$
$$\Leftrightarrow \textit{either:} \quad \{\delta_n\}^+ = 2^m, 1, \ldots \text{ or } 2^\infty$$
$$\& \{\delta_n\}^- = 2, \ldots \text{ or } 3, 2^p, 1, \ldots \text{ or } 3, 2^\infty$$
$$\textit{or:} \quad \{\delta_n\}^- = 2^m, 1, \ldots \text{ or } 2^\infty$$
$$\& \{\delta_n\}^+ = 2, \ldots \text{ or } 3, 2^p, 1, \ldots \text{ or } 3, 2^\infty.$$

Proof. It is readily checked that the above list of conditions on $\{\delta_n\}$ is exhaustive. That the conditions are mutually exclusive follows from the fact that it is impossible to have both $\{\delta_n\}^+ = 2^m, 1, \ldots$ or 2^∞, and $\{\delta_n\}^- = 2^p, 1, \ldots$ or 2^∞, as stated in Theorem 9.5 (i).

Therefore, to prove Theorem D2 we must show that each of the above conditions on $\{\delta_n\}$ implies the corresponding condition on 0. This follows from the definition of M-*block* given in §9. For instance, let $\{\delta_n\}^+ = 2^m, 1, \ldots, \{\delta_n\}^- = t, \ldots$; i.e. $\{\delta_0, \delta_1, \ldots, \delta_{m+1}\} = \{t, 2, \ldots, 2, 1\}$. Since $\delta_{m+1} = 1$, $m+1$ lies in a component M. $m+1$ is not connected to a preceding element of $J = \{n: \delta_n = 1\}$ since $\delta_0 = t$ and $\delta_1 = \cdots = \delta_m = 2$. Thus $m+1$ is the smallest element in M. The integer 0 is the biggest $k < m+1$ for which $\delta_k \neq 2$. It follows that $0 = b_0$ is the smallest integer in $B = B(\mathrm{M})$.

We give a geometric interpretation of the conditions imposed on $\{\delta_n\}$ in Theorem D2. In the sequel, when we state that two curves do not intersect, we will always mean that they do not intersect in $\overline{\mathbb{D}}$.

Theorem D3. *Let s_i, s_j be the consecutive edges of* **F**. *Let γ^+ be a geodesic ray starting from a point of s_i and leaving* **F**. *Then γ^+ and $\bar{s}_j$, the geodesic containing s_j, do not intersect.*

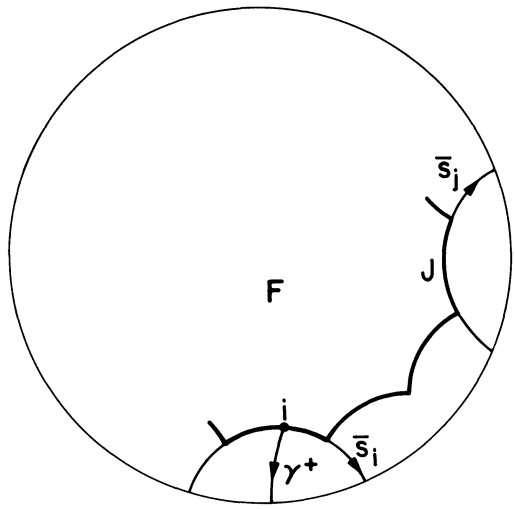


FIGURE D1. γ^+ and s_j do not intersect.

The content of Theorem D3 is illustrated in Figure D1. We remark that, by applying the group element $\tau \in \Gamma$, we obtain a similiar result for any fundamental region $\tau\mathbf{F}$.

Proof. The geodesic $\bar{s}_i$ containing s_i separates $\mathbb{D}$ into two half-planes π_1 and π_2, π_1 closed and containing γ^+, and π_2 open and containing s_j. By Theorem 3.3, $\bar{s}_i$ and $\bar{s}_j$ do not intersect. It follows that $\bar{s}_j$ is also contained in π_2, and thus does not intersect γ^+.

Let $(\xi, \eta) \in C_i$. We shall prove Theorem D1 successively for $(\xi, \eta) \in X_i$, E_i, $D_i \cup F_i$, the case $(\xi, \eta) \in G_i$ then following automatically. We need the following (ξ, η)-descriptions of the sets X_i, etc., which are obtained from Figure 5.4.

$$X_i = \{(\xi, \eta) \in C_i: \quad \xi \in [a_{i+1}, b_i), \eta \notin (b_{i+1}, b_{i+2}]\}$$
$$D_i = \{(\xi, \eta) \in C_i: \quad \xi \in [a_{i+1}, b_i), \eta \in (b_{i+1}, b_{i+2}]\}$$
$$E_i = \{(\xi, \eta) \in C_i: \quad \xi \notin [T_i^{-1}(a_{\sigma(i)-2}), a_{i+1}), \eta \in (b_i, b_{i+1}]\}$$
$$F_i = \{(\xi, \eta) \in C_i: \quad \xi \in [T_i^{-1}(a_{\sigma(i)-2}), a_{i+1}), \eta \in (b_i, b_{i+1}]\}.$$

Let $\gamma = \gamma(\xi, \eta)$ be the geodesic with end points ξ, η, where $(\xi, \eta) \in C_i$, and $u = u(\xi, \eta, \cdot)$ the outwardly pointing tangent to γ at the point p where it leaves $\mathbf{F}$. Let $\gamma^+ = \gamma^+(u)$, $\gamma^- = \gamma^-(u)$, and assume that γ passes successively through the fundamental regions $\{\mathbf{F}_n\}$, $-\infty < n < \infty$, with $\mathbf{F} = \mathbf{F}_0$.

Proof of Theorem D1. $(\xi, \eta) \in X_i$: We give a geometric interpretation to $(\xi, \eta) \in X_i$ in terms of γ^+ and γ^-. We then show that this interpretation is equivalent to the first of the $\{\delta_n\}$-conditions of Theorem D2.

As made evident by Figure D2, $\xi \in [a_{i+1}, b_i)$ iff γ^+ intersects $\bar{s}_{i+1}$, the intersection being in $\mathbb{D}$ if $\xi \neq a_{i+1}$ and in $\partial\mathbb{D}$ if $\xi = a_{i+1}$. Furthermore $(\xi, \eta) \in X_i$, i.e. $\xi \in [a_{i+1}, b_i)$ and $\eta \notin (b_{i+1}, b_{i+2}]$, iff γ^+ intersects $\bar{s}_{i+1}$ and γ^- does not intersect $\bar{s}_{i+2}$.

Suppose $\{\delta_n\}^+ = 2^m, 1, \ldots$. Then $\mathbf{F}_0, \ldots, \mathbf{F}_m$ lie on the same side of $\bar{s}_{i+1}$ and γ^+ leaves $\mathbf{F}_{m+1}$ through $\bar{s}_{m+1}$. Similarly, if $\{\delta_n\}^+ = 2^\infty$, then $\mathbf{F}_0, \ldots, \mathbf{F}_m, \ldots$ lie on the same side of $\bar{s}_{i+1}$ and γ^+ meets $\bar{s}_{i+1}$ in $\partial\mathbb{D}$. Finally, if $\{\delta_n\}^+ = 2^m, r, \ldots$, then γ^+ leaves $\mathbf{F}_{m+1}$ through an edge nonconsecutive to the edge contained in $\bar{s}_{i+1}$. It follows from Theorem D2 that γ^+ does not meet $\bar{s}_{i+1}$.

We have thus shown that: γ^+ meets $\bar{s}_{i+1}$ iff $\{\delta_n\}^+ = 2^m, 1, \ldots$ or 2^∞. Similarly, it is shown that: γ^+ meets $\bar{s}_{i+1}$ and γ^- does not meet $\bar{s}_{i+2}$ iff $\{\delta_n\}^+ = 2^m, 1, \ldots$ or 2^∞ and $\{\delta_n\}^- = t, \ldots$ or $3, 2^p, r, \ldots$. We conclude from Theorem D2 that $(\xi, \eta) \in X_i$ iff $0 = b_0$.

$(\xi, \eta) \in E_i$: As made evident by Figure D3 on p. 332, $(\xi, \eta) \in E_i$, i.e. $\xi \notin [T_i^{-1}(a_{\sigma(i)-2}), a_{i+1})$ and $\eta \in (b_i, b_{i+1}]$, iff γ^- meets $\bar{s}_{i+1}$ and γ^+ does not intersect $T_i^{-1}(\bar{s}_{\sigma(i)-2})$. We repeat the argument employed for X_i to show that $(\xi, \eta) \in E_i$ iff $0 = b_1$.

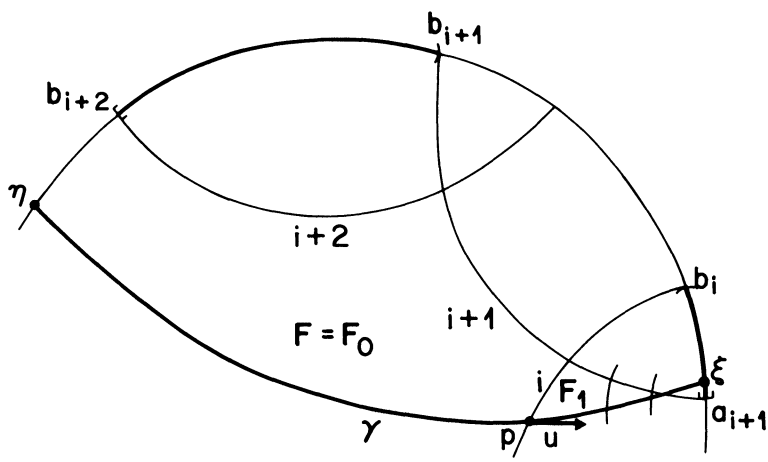


FIGURE D2. Interpretation of X_i in terms of γ.

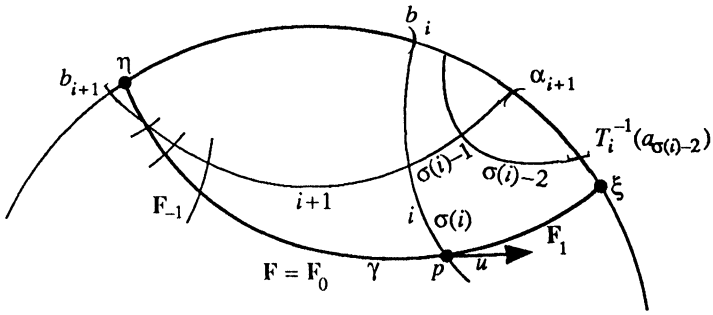


FIGURE D3. Interpretation of E_i in terms of γ.

$(\xi, \eta) \in D_i \cup F_i$: Repeating the argument employed for X_i we show that

$$(\xi, \eta) \in D_i \Leftrightarrow \{\delta_n\}^+ = 2^m, 1, \ldots \quad \text{or} \quad 2^\infty$$
$$\& \{\delta_n\}^- = 2, \ldots \quad \text{or} \quad 3, 2^p, 1, \ldots \quad \text{or} \quad 3, 2^\infty$$
$$(\xi, \eta) \in F_i \Leftrightarrow \{\delta_n\}^- = 2^m, 1, \ldots \quad \text{or} \quad 2^\infty$$
$$\& \{\delta_n\}^+ = 2, \ldots \quad \text{or} \quad 3, 2^p, 1, \ldots \quad \text{or} \quad 3, 2^\infty.$$

We conclude from Theorem D2 that $(\xi, \eta) \in D_i \cup F_i$ iff $0 \in B - \{b_0, b_1\}$.

REFERENCES

[A] E. Artin, *Ein mechanisches System mit quasiergodischen Bahnen*, Abh. Math. Sem. Univ. Hamburg **3** (1924), 170–175.

[AF1] R. L. Adler and L. Flatto, *Cross section maps for geodesic flows.* I(*the modular surface*), Ergodic Theory and Dynamical Systems, vol. 2, Proceedings Special Year Md. (1979-1980) Progress in Math., Birkhäuser, Boston, Basil, and Stuttgart, pp. 103–161.

[AF2] ——, *Cross section map for the geodesic flow on the modular surface*, Contemporary Math., vol. 26, Amer. Math. Soc., Providence, RI, 1984, pp. 9–24.

[AF3] ——, *The backward continued fraction map and the geodesic flow*, Ergodic Theory Dynamical Systems **4** (1984), 487–492.

[AM] R. L. Adler and B. Marcus, *Topological entropy and equivalence of dynamical systems*, Mem. Amer. Math. Soc., no. 219, Amer. Math. Soc., Providence, RI, 1979.

[AW] R. L. Adler and B. Weiss, *Similarity of automorphisms of the torus*, Mem. Amer. Math. Soc., no. 98, Amer. Math. Soc., Providence, RI, 1970.

[AK] W. Ambrose and S. Kakutani, *Structure and continuity of measurable flows*, Duke Math. J. **9** (1942), 25–42.

[Ar] V. I. Arnold, *Mathematical methods of classical mechanics*, Springer-Verlag, New York, Berlin, Heidelberg, 1978.

[AKS] J. Ashley, B. P. Kitchens, and M. Stafford, *Boundaries of Markov partitions*, Trans. Amer. Math. Soc. (to appear).

[B] A. F. Beardon, *The geometry of discrete groups*, Springer-Verlag, New York, Berlin, Heidelberg, 1983.

[Be] K. Berg, *On the conjugacy problem for K-systems*, Ph.D. Thesis, Univ. of Minnesota, 1967.

[Bi] P. Billingsley, *Ergodic theory and information*, John Wiley & Sons, Inc., NY, 1965.

[BS] J. S. Birman and C. Series, *Dehn's algorithm revisited, with applications to simple curves on surfaces*, Proc. of Alta, Utah Conf. on Combinatorial Group Theory (1984) (Gersten and Stallings, eds.) (to appear).

[Bo1] R. Bowen, *Markov partitions for Axiom* A *diffeomorphisms*, Amer. J. Math. **92** (1970), 725–745.

[Bo2] ——, *Markov partitions and minimal sets for Axiom* A *diffeomorphisms*, Amer. J. Math. **92** (1970), 907–918.

[Bo3] ——, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Lecture Notes in Math., vol. 470, Springer-Verlag, NY, 1975.

[Bo4] ——, *On Axiom* A *diffeomorphisms*, CBMS Regional Conf. Ser. in Math., no. 35, Amer. Math. Soc., Providence, RI, 1977.

[BoS] R. Bowen and C. Series, *Markov maps associated with Fuchsian groups*, Inst. Hautes Études Sci., Publ. Math. **50** (1979), 153–170.

[BKM] M. Boyle, B. Kitchens, and B. Marcus, *A note on minimal covers for sofic systems*, Proc. Amer. Math. Soc. **95** (1985), 403–411.

[BMT] M. Boyle, B. Marcus, and P. Trow, *Resolving maps and the dimension group for shifts of finite type*, Mem. of Amer. Math. Soc., no. 377, Amer. Math. Soc., Providence, RI, 1987.

[CP] E. M. Coven and M. E. Paul, *Sofic Systems*, Israel J. Math. **20** (1975), 165–177.

[CFS] I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinai, *Ergodic theory*, Springer-Verlag, New York, Berlin, Heidelberg, 1982.

[D] M. Dehn, *Papers on group theory and topology*, edited and translated by John Stillwell, Springer-Verlag, New York, Berlin, Heidelberg, 1987.

[Do] W. Doeblin, *Remarques sur la théorie métrique des fractions continues*, Compositio Math. (1940), 353–371.

[H] G. A. Hedlund, *On the metrical transitivity of the geodesics on closed surfaces of constant negative curvature*, Ann. of Math. (2) **35** (1934), 787–808.

[Ho] E. Hopf, *Ergodic theory and the geodesic flow on surfaces of constant negative curvature*, Bull. Amer. Math. Soc. **77** (1971), 863–877.

[KSS] A. Katok, Ya. G. Sinai, and A. M. Stepin, *Theory of dynamical systems and general transformation groups with an invariant measure*, J. Soviet Math. (1977), 974–1064. Transl. (2) **17** (1961), 277–364.

[M] B. Maskit, *On Poincaré theorem for fundamental polygons*, Adv. in Math. **7** (1971), 219–230.

[Ma] R. M. May, *Simple mathematical models with very complicated dynamics*, Nature **261** (1976), 459–467.

[Mo] M. Morse, *Symbolic dynamics, Instituted for Advanced Study Notes*, Notes by Rufus Oldenburger, Princeton, 1966 (unpublished).

[MPV] J. Moser, E. Phillips, and S. Varadhan, *Ergodic theory, a Seminar*, Lecture Notes in Math., New York Univ., NY, 1975, pp. 111–120.

[N] J. Nielsen, *Untersuchungen zur Topologie der geschlossen zweiseitigen Flächen*, Acta Math. **50** (1927), 189–358.

[O] D. S. Ornstein, *Ergodic theory, randomness, and dynamical systems*, Yale Math. Monographs 5, New Haven and London, Yale Univ. Press, 1974.

[OW] D. S. Ornstein and B. Weiss, *Geodesic flows are Bernoullian*, Israel J. Math. **14** (1973), 184–198.

[P] W. Parry, *Intrinsic Markov chains*, Trans. Amer. Math. Soc. **112** (1964), 55–66.

[R] A. Renyi, *Representation for real numbers and their ergodic properties*, Acta Math. Akad. Sci. Hungary **8** (1957), 477–493.

[Ro] V. A. Rohlin, *Exact endomorphisms of a Lebesgue space*, Izv. Acad. Nauk. SSSR Ser. Mat. **25** (1961), 499–530; Amer. Math. Soc. Transl. (2) **39** (1964), 1–37.

[Ru] D. Ruelle, *Thermodynamic formalism*, Encyclopedia of Math. and Applications **5**, Addison-Wesley, Reading, MA, 1978.

[S1] C. Series, *Symbolic dynamics for geodesic flows*, Acta Math. **146** (1981), 103–128.

[S2] ——, *The infinite word problem and limit sets in Fuchsian groups*, Ergodic Theory Dynamical Systems **1** (1981), 337–360.

[S3] ——, *Geometrical Markov coding of geodesics on surfaces of constant negative curvature*, Ergodic Theory Dynamical Systems **6** (1986), 601–625.

[Sh] M. Shub, *Global stability of dynamical systems*, Springer-Verlag, NY, 1967.

[SS] M. Shub and D. Sullivan, *Expanding maps of the circle revisited*, Ergodic Theory Dynamical Systems **5** (1985), 285–289.

[Si] Ya. G. Sinai, *Construction of Markov partitions*, Funct. Anal. Appl. **2** (1968), 70–80.

[Sm] S. Smale, *Differentiable dynamical systems*, Bull. Amer. Math. Soc. **73** (1967), 747–813.

[Sp] G. Springer, *Introduction to Riemann surfaces*, Addison-Wesley Publ. Co., Reading, MA, 1957.

[St] H. M. Stark, *An introduction to number theory*, Markham Publ. Co., Chicago, 1970.

[T] P. Tukia, *On discrete groups of the unit disk and their isomorphisms*, Ann. Acad. Sci. Fenn., Series A, I. Math. **504** (1972), 5–44.

[V] B. L. Van der Waerden, *Modern algebra*, vol. I, Frederick Ungar Publ. Co., NY, 1949.

[W] B. Weiss, *Subshifts of finite type and sofic systems*, Monatsh. Math. **77** (1973), 462–474.

IBM Research Division, T. J. Watson Research Center, Yorktown Heights, New York 10598

AT& T Bell Laboratories, Murray Hill, New Jersey 07974