Ergod. Th. & Dynam. Sys. (1981), **1**, 107–133
Printed in Great Britain

Checking ergodicity of some geodesic flows with infinite Gibbs measure

MARY REES†
From the Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette, France

(*Received* 20 *August* 1980 *and revised* 13 *January* 1981)

Abstract. This paper concerns a problem which arose from a paper of Sullivan. Let Γ be a discrete group of isometries of hyperbolic space H^{d+1}. We study the question of when the geodesic flow on the unit tangent bundle $\mathrm{UT}(H^{d+1}/\Gamma)$ of H^{d+1}/Γ is ergodic with respect to certain natural measures. As a consequence, we study the question of when Γ is of divergence type. Ergodicity when the non-wandering set of $\mathrm{UT}(H^{d+1}/\Gamma)$ is compact is already known from the theory of symbolic dynamics, due to Bowen, or from Sullivan's work. For such a Γ, we consider a subgroup Γ_1 of Γ with $\Gamma/\Gamma_1 \cong \mathbb{Z}^v$ and prove the geodesic flow on $\mathrm{UT}(H^{d+1}/\Gamma_1)$ is ergodic (with respect to one of these natural measures) if and only if $v \leq 2$.

0. *Introduction*

The geodesic flow $\{\phi_t\}$, on the unit tangent bundle $\mathrm{UT}(M)$ of a $(d+1)$-dimensional manifold M of constant negative curvature, is a common object of study in dynamical systems and ergodic theory. Such a manifold M is of the form H^{d+1}/Γ, for Γ a discrete group of isometries of hyperbolic space H^{d+1}. In the present paper, we study the question of whether $(\mathrm{UT}(H^{d+1}/\Gamma), \{\phi_t\}, \mu)$ is ergodic, for certain groups Γ, and certain natural ϕ_t-invariant measures μ. As a consequence, we also study the question of whether Γ is *of divergence type*. These questions arose from **[10]**, as will be explained shortly.

We need to recall two classical methods of studying the geodesic flows $(\mathrm{UT}(H^{d+1}/\Gamma), \{\phi_t\})$. The first is in terms of the *limit set* of the group Γ. Recall that H^{d+1} has a natural boundary sphere S^d such that $H^{d+1} \cup S^d$ is compact, and that the action of Γ on H^{d+1} extends continuously to $H^{d+1} \cup S^d$. $H^{d+1} \cup S^d$ identifies in a natural way with the unit ball in $\mathbb{R}^{d+1}$. Γ acts smoothly on the unit sphere, and has the property that, for $\xi, \eta \in S^d$, $\|\gamma(\xi-\eta)\| = |\gamma'(\xi)|\,|\gamma'(\eta)|\,\|\xi-\eta\|$, where $\| \ \|$ denotes the Euclidean norm on $\mathbb{R}^{d+1}$, and $|\gamma'(\xi)|$ is a scalar associated to the derivative of γ (clearly $\gamma'(\xi)$ is precisely the derivative if $d=1$, when $\mathbb{R}^{d+1}$ is the complex plane). The *limit set* $L_\Gamma \subseteq S^d$ of Γ is the set of accumulation points of $\{\gamma x : \gamma \in \Gamma\}$ for any $x \in H^{d+1}$. (The definition is independent of the choice of x.) Γ leaves L_Γ invariant. $\mathrm{UT}(H^{d+1}/\Gamma)$ is the same as $(\mathrm{UT}(H^{d+1}))/\Gamma$ (where the action of Γ on $\mathrm{UT}(H^{d+1})$ is given by the derivatives of the action on H^{d+1}), and $\mathrm{UT}(H^{d+1})$ is diffeomorphic to

† Address for correspondence: Dr Mary Rees, Institute des Hautes Etudes Scientifiques, 35 route de Chartres, 91440 Bures-sur-Yvette, France.

$((S^d \times S^d)\backslash\text{diagonal}) \times \mathbb{R}$, in such a way that the lifts of $\{\phi_t\}$-orbits in $(\text{UT}(H^{d+1}))/\Gamma$ are the sets $\{(x, y)\} \times \mathbb{R}$ $(x, y \in L_\Gamma)$. The action of Γ on $\text{UT}(H^{d+1})$ transfers to an action sending the set $\{(x, y)\} \times \mathbb{R}$ to $\{(\gamma x, \gamma y)\} \times \mathbb{R}$. The non-wandering set X_Γ of the flow $(\text{UT}(H^{d+1}\backslash\Gamma), \{\phi_t\})$, when lifted to $\text{UT}(H^{d+1})$, corresponds to $((L_\Gamma \times L_\Gamma)\backslash\text{diagonal}) \times \mathbb{R}$. Thus, ϕ_t-invariant measures on X_Γ correspond to Γ-invariant measures on $(L_\Gamma \times L_\Gamma)\backslash$ diagonal. By [**5**], $(L_\Gamma \times L_\Gamma, \Gamma)$ is topologically transitive for *all* non-elementary groups Γ, so $(X_\Gamma, \{\phi_t\})$ is also topologically transitive. Questions of ergodicity are more subtle.

One class of Γ-invariant measures on $L_\Gamma \times L_\Gamma$ – which is included in those studied here – arises from the so-called 'conformal densities' studied by Sullivan [**10**] (the early work is due to Patterson [**8**]). A *Γ-invariant conformal density of dimension δ* is (abusing the notation of [**10**] slightly) a probability measure ν on L_Γ such that

$$\frac{d\gamma_*\nu}{d\nu}(\xi) = |\gamma'(\xi)|^\delta \quad \text{for all } \delta \in L_\Gamma,$$

where $\gamma_*\nu(f) = \nu(f \circ \gamma^{-1})$. If μ_ν on $L_\Gamma \times L_\Gamma$ is defined by

$$\frac{d\nu(\xi)\, d\nu(\eta)}{\|\xi - \eta\|^{2\delta}} = d\mu_\nu(\xi, \eta),$$

then μ_ν is a Γ-invariant measure on $L_\Gamma \times L_\Gamma$. Of course, if $L_\Gamma = S^d$, Lebesgue measure on S^d is a Γ-invariant conformal density of dimension d.

There is not space here for a proper review of Sullivan's results, but they include the following. Let (x, y) denote hyperbolic distance between $x, y \in H^{d+1}$. For $\alpha \in \mathbb{R}$, the Poincaré series

$$\sum_{\gamma \in \Gamma} \exp\{-\alpha(x, \gamma x)\}$$

converges or diverges independently of the choice of x. The *critical exponent* $\delta(\Gamma)$ of Γ is the supremum of the α for which the series diverges. Always, $\delta(\Gamma) \le d$. There exists a Γ-invariant conformal density ν of dimension $\delta(\Gamma) = \delta$. (This is direct imitation of [**8**], where it was proved for the case $d = 1$.) For any such ν (and Γ non-elementary) $(L_\Gamma \times L_\Gamma, \Gamma, \mu_\nu)$ is ergodic if and only if Γ is *of divergence type*, that is, the Poincaré series diverges at the critical exponent $\delta(\Gamma)$. In the case of divergence type, (L_Γ, Γ, ν) is also ergodic, for arbitrary ν, so there is only one Γ-invariant conformal density of dimension $\delta(\Gamma)$. The equivalence of ergodicity and divergence type is actually completely proved for $\delta \ge \frac{1}{2}d$ in [**10**], via a third equivalent condition, the recurrence of a certain Markov process with paths in H^{d+1}/Γ. In the classical case $\delta = d$, this process is hyperbolic Brownian motion. Aaronson and Sullivan later proved the equivalence of divergence type and ergodicity for *all* non-elementary groups Γ, by a method not using Markov processes.

If X_Γ is compact (Sullivan actually considers Γ *convex co-compact*, which is, if anything, a stronger condition, but the same proof works for X_Γ compact), then Γ is of divergence type, and ν (the conformal density) is Hausdorff measure on L_Γ, and the associated measure on X_Γ is the unique measure maximizing the entropy of $(X_\Gamma, \{\phi_t\})$. By [**8**], all finitely generated Fuchsian groups (that is, $d = 1$) are of divergence type. Classically, Γ is of divergence type if H^{d+1}/Γ has finite hyperbolic volume, in which case $\delta(\Gamma)=d$.

The divergence type condition, or equivalence conditions, have been checked by various people, for various groups Γ_1 with Γ_1 a normal subgroup of Γ, H^{d+1}/Γ finite volume, and $\Gamma/\Gamma_1 \cong \mathbb{Z}^v$. Note that a non-trivial normal subgroup Γ_1 of Γ has $L_\Gamma = L_{\Gamma_1}$, so that in these cases $X_{\Gamma_1} = \text{UT}\,(H^{d+1}/\Gamma_1)$. For Γ with H^{d+1}/Γ compact, it has been proved by Sullivan (via the non-existence of a Green's function on H^{d+1}/Γ_1) that if $\Gamma/\Gamma_1 \cong \mathbb{Z}^2$, then $\delta(\Gamma_1)=d$ and Γ_1 is of divergence type, and by Guivarc'h (using Brownian motion) that if $\Gamma/\Gamma_1 \cong \mathbb{Z}^3$ then Γ_1 is not of divergence type with $\delta(\Gamma_1)=d$. Lyons and McKean have proved [**6**] that if H^2/Γ is the thrice-punctured sphere, then the commutator subgroup $[\Gamma, \Gamma]$ (for which $\Gamma/[\Gamma,\Gamma] \cong \mathbb{Z}^2$) is *not* of divergence type, but $\delta([\Gamma,\Gamma])=1$. Their interest was in the Brownian motion result, and their proof used Brownian motion. They were also able to show, fairly easily, that if the generators of Γ are denoted a, b, and $\Gamma_2 = \{$words in a, b: sum of a-powers $=0\}$, then Γ_2 is of divergence type, and $\delta(\Gamma_2)=1$.

I propose to add to these results, and to consider the case of a normal subgroup Γ_1 of a group Γ with X_Γ compact, Γ/Γ_1 abelian, and Γ non-elementary. This includes Γ with H^{d+1}/Γ compact, and also Schottky groups, which are useful examples to bear in mind (see the beginning of § 1). Some results for 'finitely determined subabelian subgroups' of Γ will be briefly indicated in § 5. A larger class of measures than those arising from conformal densities will be considered, the so-called 'Gibbs' measures ([**3**], 1.7 of this paper, and below). Part of the motivation comes from Bowen [**4**], who proved that for some groups, Hausdorff measure on the limit set of the group is 'Gibbs'.

To explain the class of measures we consider, it is necessary to recall a second classical method of studying the geodesic flow – symbolic dynamics. If Γ is such that X_Γ is compact, then $(X_\Gamma, \{\phi_t\})$ is a hyperbolic flow in the sense of Bowen [**2**], so can be realized as the suspension of a topologically mixing subshift of finite type (Y_Γ, σ) on finitely many symbols, where σ denotes the shift. Finite-full-support-ergodic-ϕ_t-invariant measures on X_Γ are in one-to-one correspondence with finite-full-support-ergodic-σ-invariant measures on Y_Γ. So 'Gibbs' measures on X_Γ are defined to be those corresponding to 'Gibbs' measures on Y_Γ. If $\Gamma_1 \le \Gamma$ and $L_{\Gamma_1} = L_\Gamma$, 'Gibbs' measures on X_{Γ_1} are those obtained by lifting 'Gibbs' measures on X_Γ in such a way that local inverses of the natural projection are measure preserving.

The paper proceeds as follows. Suppose fixed a group Γ with X_Γ compact, and Γ_1 a subgroup of Γ with $L_{\Gamma_1} = L_\Gamma$. Denoting corresponding measures by the same symbol, we find, in § 1, a suitable subshift (Y_Γ, σ), and an equivalence relation $\sim_{\Gamma_1}$ on Y_Γ, which is a subset of the σ orbit equivalence relation, such that $(X_{\Gamma_1}, \{\phi_t\}, \mu)$ is ergodic if and only if $(Y_\Gamma, \sim_{\Gamma_1}, \mu)$ is ergodic. In § 2 it is shown that, for μ Gibbs, $(Y_\Gamma, \sim_{\Gamma_1}, \mu)$ is ergodic if and only if a certain series diverges. Specializing to the case of a Γ-invariant conformal density, it is shown this is equivalent to the divergence of:

$$\sum_{\gamma\in\Gamma} \exp\{-\delta(x, \gamma x)\}, \quad \text{for } \delta = \delta(\Gamma).$$

https://doi.org/10.1017/S0143385700001206 Published online by Cambridge University Press

In §§ 3, 4 it is shown that if Γ/Γ_1 is abelian and torsion free, $(Y_\Gamma, \sim_{\Gamma_1}, \mu)$ is ergodic if and only if rank $\Gamma/\Gamma_1 \le 2$. This result is generalized in § 5. Restricting theorem 4.7 to the conformal density case, if rank $\Gamma/\Gamma_1 = v$, and $\delta(\Gamma) = \delta$, there exist $A, B > 0$ such that

$$A/(k^{\frac{1}{2}v-1}) \le \sum_{\{\gamma\in\Gamma_1 : Ak \le (x,\gamma x) < Bk\}} \exp\{-\delta(x, \gamma x)\} \le B/(k^{\frac{1}{2}v-1})$$

for any fixed $x \in H^{d+1}$. So, in particular, $\delta(\Gamma_1) = \delta$ whenever Γ/Γ_1 is abelian and X_Γ is compact.

1. *Symbolic dynamics for the geodesic flow, and Gibbs measures*

Throughout this section, Γ is a discrete group of isometries of H^{d+1} such that $L_\Gamma \subseteq S^d$ has more than two points, and X_Γ is compact. We need to modify slightly Bowen's construction of symbolic dynamics for $(X_\Gamma, \{\phi_t\})$, associating the symbolic representation to the group Γ. Hence we obtain (for $\Gamma_1 \le \Gamma$ with $L_{\Gamma_1} = L_\Gamma$) simultaneous symbolic representations (Y_Γ, σ), (Y_{Γ_1}, σ) of $(X_\Gamma, \{\phi_t\})$, $(X_{\Gamma_1}, \{\phi_t\})$. Hence an equivalence relation $\sim_{\Gamma_1}$ is defined on (Y_Γ, σ), allowing us to reformulate the problem of the ergodicity of $(X_{\Gamma_1}, \{\phi_t\}, \mu)$, for (X_{Γ_1}, μ) a 'lift' of (X_Γ, μ) (1.3, 1.5).

(1.3) and (1.5) can be omitted if one is prepared simply to consider the case of Schottky groups: if Γ is a free group on n generators $a_1 \cdots a_n$ and has a fundamental region F obtained as the intersection in H^{d+1} of $2n$ solid 'hemispheres' with the a_iF, $a_i^{-1}F$ $(i = 1 \cdots n)$ the adjacent regions, then Y_Γ can be taken as $\{\{x_i\} \in \{a_1 \cdots a_n, a_1^{-1} \cdots a_n^{-1}\}^{\mathbb{Z}} : x_{i+1} \ne x_i^{-1}$ for any $i\}$ as in [**4**]. (For general method see [**7**] or [**9**].)

It will be helpful to bear in mind the following interpretation (in this case) of Bowen's definition of a Markov set of cross-sections for a flow [**2**]. As mentioned in the introduction, we have an identification of UT (H^{d+1}) with $(S^d \times S^d \backslash \text{diagonal}) \times \mathbb{R}$ such that $\gamma \in \Gamma$ sends $\{(x, y)\} \times \mathbb{R}$ to $\{(\gamma x, \gamma y)\} \times \mathbb{R}$, and the sets $\{(x, y)\} \times \mathbb{R}$ correspond to geodesic flow orbits.

(1.1) Note that a transverse disk C to the flow $(\text{UT}(H^{d+1}/\Gamma), \{\phi_t\})$ can be lifted (non-uniquely) to a transverse disk C' of $(\text{UT}(H^{d+1}), \{\phi_t\})$, and then all lifts are given by $\{\gamma C' : \gamma \in \Gamma\}$. The set of geodesics through C' is then identified with $D_1 \times \mathbb{R}$, for $D_1 \subseteq S^d \times S^d \backslash$diagonal. A *rectangle* is then a subset C_1 of a transverse disk C such that the set of geodesics passing through the lift $C_1' \subseteq C'$ is identified with $U \times V \times \mathbb{R}$, where $U, V \subseteq S^d$, $U \cap V = \varnothing$, $\overline{\text{interior } U} = U$, and $\overline{\text{interior } V} = V$.

$\{C_1 \cdots C_n\}$ is a *Markov set of cross-sections* for $(X_\Gamma, \{\phi_t\})$ if each C_i is a rectangle, and whenever some geodesic of X_Γ goes successively through the interiors of C_i, C_j, and nothing in between, and C_i', C_j' are lifts for which the same is true in UT (H^{d+1}), with C_i', C_j' identified with $(U_i \times V_i) \times \mathbb{R}$, $(U_j \times V_j) \times \mathbb{R}$, then $U_i \subseteq U_j$ and $V_j \subseteq V_i$. If there is such a geodesic for C_i, C_j, we say (C_i, C_j) *is admissible.*

Bowen [**2**] proves that, if $\{C_1 \cdots C_n\}$ is Markov, there is a geodesic going successively through the interiors of the cross-sections in any admissible chain $C_{i_1} \cdots C_{i_r}$. Then if $Z_\Gamma = \{\{D_j\}_{j=-\infty}^{\infty} : D_j \in \{C_1 \cdots C_n\}, D_jD_{j+1}$ admissible$\}$, there is a suspension $((Z_\Gamma \times \mathbb{R})/\mathbb{Z}, \mathbb{R})$ of (Z_Γ, σ) under a non-constant function, and a surjective homomorphism Π_Γ: $((Z_\Gamma\times\mathbb{R})/\mathbb{Z}, \mathbb{R})\to(X_\Gamma, \mathbb{R})$. Moreover, Π_Γ is one–one on a residual set whose image is residual. See [**2**] for further details. Here, σ denotes the shift $\sigma(\{D_i\})=\{D_{i+1}\}$, $\mathbb{Z}$ denotes the integers, and the $\mathbb{Z}$-action on $Z_\Gamma\times\mathbb{R}$ is that generated by $(\mathbf{z}, t)\mapsto(\sigma\mathbf{z}, t-f(\mathbf{z}))$, if f is the function we are suspending under.

(1.2) *Definition.* For discrete Γ_1, let τ: $\mathrm{UT}\,(H^{d+1}/\Gamma_1)\to\mathrm{UT}\,(H^{d+1}/\Gamma_1)$ be the map sending a unit tangent vector v to $-v$. Then $\tau X_{\Gamma_1}=X_{\Gamma_1}$. τ: $\mathrm{UT}\,(H^{d+1})=(S^d\times S^d\backslash\text{diagonal})\to\mathrm{UT}\,(H^{d+1})$ sends $\{(x, y)\}\times\mathbb{R}$ to $\{(y, x)\}\times\mathbb{R}$.

(1.3) THEOREM (modification of [**2**], § 7). *There exists a Markov set of cross-sections $\mathcal{J}_\Gamma=\{b_1\cdots b_s, \tau(b_1)\cdots\tau(b_s)\}$ for $(X_\Gamma, \{\phi_t\})$ such that the associated subshift of finite type (Z_Γ, σ) is topologically mixing. Π_Γ: $(Z_\Gamma\times\mathbb{R})/\mathbb{Z}\to X_\Gamma$ gives rise to a one–one correspondence $\mu\mapsto(\Pi_\Gamma)_*\mu$ between finite full-support invariant ergodic measures.*

Notes on proof. (1) Bowen defines hyperbolic flows only for compact manifolds, but all that is needed is that X_Γ be compact.

(2) In working through Bowen's proof in § 7 in [**2**] (and unfortunately one has to go through the whole construction making slight changes), one starts with a set of rectangles $\{B_1\cdots B_n, \tau B_1\cdots\tau B_n\}$. Note that τ interchanges stable and unstable manifolds of the flow, hence sends rectangles to rectangles.

(3) An arbitrary set of cross-sections $\mathcal{J}_\Gamma$ will not be topologically mixing. But let p be the unique strictly positive integer for which there exists $\rho:\mathcal{J}_\Gamma\to\mathbb{Z}/p\mathbb{Z}$ with $\rho(\sigma(\mathbf{z}))=\rho(\mathbf{z})+1$ for all $\mathbf{z}\in Z_\Gamma$ (if we also define $\rho: Z_\Gamma\to\mathbb{Z}/p\mathbb{Z}$ by $\rho(\{z_i\})=\rho(z_0)$), and $(\rho^{-1}(p\mathbb{Z}+r), \sigma^p)$ topologically mixing for all r. Since $\rho(\tau\mathbf{z})=-\rho(\mathbf{z})+r$ for all $\mathbf{z}\in Z_\Gamma$, some fixed r (as can be checked), there exists $C_1\in\mathcal{J}_\Gamma$ such that if $\{C_1\cdots C_n\}=\rho^{-1}\rho(C_1)$ then either $\{C_1\cdots C_n\}=\tau(\{C_1\cdots C_n\})$ or $\rho(\tau C_i)=\rho(C_1)+1$, $i=1\cdots n$. In the first case, let $\{C_1\cdots C_n\}$ be the new set $\mathcal{J}_\Gamma$. In the second case, let d_{ij} be a cross-section between C_i and $\tau(C_j)$ whenever there is a set of geodesics going successively through the interiors of C_i, $\tau(C_j)$, and nothing in between, and let d_{ij} be exactly the span of this set of geodesics in some transverse disk. Also make $\tau(d_{ji})=d_{ij}$ (this is possible). Let the new set $\mathcal{J}_\Gamma$ be the set of d_{ij} – it is topologically mixing, as required.

(4) It is not proved in [**2**] that $\mu\mapsto(\Pi_\Gamma)_*\mu$ is a one–one correspondence, but the proof is exactly analogous to that for Markov partitions for Axiom A diffeomorphisms in ([**3**] proof of theorem 4.1, page 90). □

Let $\mathcal{J}_\Gamma$ as in (1.3) be fixed.

(1.4) *Definitions.* (1) Let $\mathcal{J}$, $\mathcal{J}_{\Gamma_1}$ denote the lifted set of cross-sections in $\mathrm{UT}\,(H^{d+1})$, $\mathrm{UT}\,(H^{d+1}/\Gamma_1)$ for $\Gamma_1\le\Gamma$. Fix a 'fundamental' set of cross-sections $\mathcal{J}_1$ in $\mathcal{J}$ with $\tau\mathcal{J}_1=\mathcal{J}_1$, $\gamma\mathcal{J}_1\cap\mathcal{J}_1=\varnothing$ for $\gamma\ne1$, and $\Gamma\mathcal{J}_1=\mathcal{J}$. It is then natural to denote the cross-sections of $\mathcal{J}_{\Gamma_1}$ by $\{(C_i, \Gamma_1\gamma)$: $C_i\in\mathcal{J}_\Gamma, \gamma\in\Gamma\}$.

(2) Let $\mathcal{K}_{\Gamma_1}=\{((C_i, \Gamma_1\gamma_i), (C_j, \Gamma_1\gamma_j))$: there exists a geodesic in the cover of X_Γ in $\mathrm{UT}\,(H^{d+1}/\Gamma_1)$ going successively through the interiors of $(C_i, \Gamma_1\gamma_i)$, $(C_j, \Gamma_1\gamma_j)$ and no other cross-section in between$\}$. Define τ: $\mathcal{K}_\Gamma\to\mathcal{K}_\Gamma$ by $\tau(C_i, C_j)=(\tau C_j, \tau C_i)$. Then τ is a fixed-point-free involution of $\mathcal{K}_\Gamma$ (assuming the cross-sections are small enough, without loss of generality).

(3) Define $\phi:\mathcal{K}_\Gamma\to\Gamma$ by: $((C_i,\gamma),(C_j,\gamma\phi(C_i,C_j)))\in\mathcal{K}_{\{1\}}$ for one, hence all, $\gamma\in\Gamma$. Note $\phi(\tau a)=\phi(a)^{-1}$ for all $a\in\mathcal{K}_\Gamma$. Hence, writing $\tau a=a^{-1}$, if $\mathcal{K}_\Gamma=\{a_1\cdots a_r, a_1^{-1}\cdots a_r^{-1}\}$, ϕ can be regarded as a homomorphism $\phi:F\to\Gamma$, where F denotes the free group in $a_1\cdots a_r$.

(4) Define

$$Y_{\Gamma_1}=\{\{x_i\}: x_i\in\mathcal{K}_{\Gamma_1}\ (i\in\mathbb{Z}),\ x_i=(y_i,z_i)\text{ for } y_i, z_i\in\mathcal{J}_{\Gamma_1}\text{ and } z_i=y_{i+1}\text{ for all } i\},$$

$$\tau:\mathcal{K}_\Gamma\to\mathcal{K}_\Gamma\text{ induces }\tau: Y_\Gamma\to Y_\Gamma\text{ by }\tau(\{x_i\})=\{\tau x_{-i}\}.$$

Projection of $\mathcal{J}_{\Gamma_1}=\mathcal{J}_\Gamma\times\Gamma/\Gamma_1$ onto the first coordinate induces similar projections $\mathcal{K}_{\Gamma_1}\to\mathcal{K}_\Gamma$, and $p: Y_{\Gamma_1}\to Y_\Gamma$.

Let $\sigma: Y_{\Gamma_1}\to Y_{\Gamma_1}$ denote the shift $\sigma(\{x_i\})=\{x_{i+1}\}$. $(X_{\Gamma_1},\{\phi_t\})$ can now be represented as a factor of a suspension of the shift (Y_{Γ_1},σ) in a useful way.

In general $\mathcal{K}_{\Gamma_1}$ has infinitely many symbols. We have a commutative diagram (figure 1), where p is the natural map induced by $p: Y_{\Gamma_1}\to Y_\Gamma$, $\rho:\mathrm{UT}\,(H^{d+1}/\Gamma_1)\to\mathrm{UT}\,(H^{d+1}/\Gamma)$ is the covering map, so that $\rho^{-1}(X_\Gamma)=X_{\Gamma_1}$ if and only if $L_{\Gamma_i}=L_\Gamma$. Π_{Γ_1}, Π_Γ are both one–one on residual sets whose images are residual.

$$\begin{array}{ccc}((Y_{\Gamma_1}\times\mathbb{R})/\mathbb{Z},\{\psi_t\}) & \xrightarrow{\ \Pi_{\Gamma_1}\ } & (\rho^{-1}(X_\Gamma),\{\phi_t\})\\ \Big\downarrow p & & \Big\downarrow \rho\\ ((Y_\Gamma\times\mathbb{R})/\mathbb{Z},\{\psi_t\}) & \xrightarrow{\ \Pi_\Gamma\ } & (X_\Gamma,\{\phi_t\})\end{array}$$

FIGURE 1

(1.5) THEOREM. (1) *Let* (Y,σ) *be any subshift of type* 2 *on symbols* $\mathcal{K}=\{a_1\cdots a_r, a_1^{-1}\cdots a_r^{-1}\}$ *with involution* $\tau: Y\to Y$ *given by* $\tau(\{x_i\})=\{x_{-i}^{-1}\}$. *Let* F *be the free group on generators* $a_1\cdots a_r$. *Let* F_1 *be any subgroup of* F *and define a subshift of type* 2 (Y_{F_1},σ) *on the symbols* $\mathcal{K}\times F/F_1$ *by*: $(b_i, F_1f_i)(b_j, F_1f_j)$ *is admissible if and only if* b_ib_j *is admissible in* Y, *and* $F_1f_j=F_1f_ib_i$. *Then, if* (Y_Γ,σ), (Y_{Γ_1},σ) *are as described in* (1.4), *and* $(Y_\Gamma,\sigma)=(Y,\sigma)$, $\{(b_i,F_1f_i)\}\mapsto\{b_i,\Gamma_1\phi(f_i)\}$ *defines an isomorphism between* (Y_{Γ_1},σ) *and* (Y_{F_1},σ), *if* $F_1=\phi^{-1}(\Gamma_1)$.

(2) *If* $L_{\Gamma_1}=L_\Gamma$ *(e.g. if* $\{1\}\neq\Gamma_1\lhd\Gamma$*) then* $(Y_{\Gamma_1},\sigma)\cong(Y_{F_1},\sigma)$ *is topologically transitive (i.e. for any open* U, V *there exists* n *with* $\sigma^nU\cap V\neq\varnothing$*), and periodic points are dense.*

(3) *For* μ *an ergodic finite full-support* ϕ_t*-invariant measure on* X_Γ, *let* μ *denote also the corresponding* σ*-invariant probability measure on* Y_Γ (1.3), *and the lifts to* Y_{Γ_1}, $\rho^{-1}X_\Gamma$, *for which local inverses of* p, ρ *are measure preserving. Similarly, for* μ *a* σ*-invariant measure on any shift* (Y,σ) *as in* (1), *let* μ *also denote the lift to* (Y_{F_1},σ), *for* $F_1\le F$.

(*a*) *If* $L_{\Gamma_1}=L_\Gamma$, $(X_{\Gamma_1},\{\phi_t\},\mu)$ *is ergodic if and only if* $(Y_{\Gamma_1},\sigma,\mu)$ *is ergodic.*

(*b*) *Let* $\sim_{F_1}$ *(or* $\sim_{\Gamma_1}$ *if* $\phi^{-1}(\Gamma_1)=F_1$*) be the subset of the* σ*-orbit equivalence relation on* Y *generated by*: $\{x_i\}\sim_{F_1}\{x_{i+r}\}$ $(r>0)$ *if* $x_0\cdots x_{r-1}\in F_1$. *Suppose* (Y_{F_1},σ) *is topologi-*

cally transitive and μ has full support. Then (Y_{F_1}, σ, μ) is ergodic if and only if $(Y, \sim_{F_1}, \mu)$ is ergodic.

Proof. (2) This follows from topological transitivity of $(X_{\Gamma_1}, \{\phi_t\})$, which follows from topological transitivity of $(L_\Gamma \times L_\Gamma, \Gamma_1)$ ([**5**], 13.24).

(3) (*a*) Π_{Γ_1} is a measure isomorphism, since Π_Γ is (1.3, see also figure 1).
(*b*) $\{x_i\} \sim_{F_1} \{x_{i+r}\}$ if and only if, for $\{(x_i, F_1 f_i)\} \in Y_{F_1}$, $F_1 f_r = F_1 f_0$. 'Only if' is then clear. Ergodicity or $\sim_{F_1}$ implies:

$$\left(\left(\bigcup_{n=-\infty}^{\infty} \sigma^n\{\{(x_i, F_1 f_i)\} : \{x_i\} \in Y, F_1 f_0 = F_1 f\}\right), \sigma, \mu\right) = (A_f, \sigma, \mu) \quad (f \in F)$$

is ergodic. Topological transitivity of (Y_{F_1}, σ) implies any two A_f, $A_{f'}$ (which are open) have non-trivial intersection, hence $A_f = Y_{F_1}$ for all $f \in F_1$. □

The rest of this section concerns the characterization of 'Gibbs' measures on Y_Γ, which include conformal densities. Let (Y, σ) be any subshift of type 2 on a set of symbols $\mathcal{K}$.

(1.6) *Definition.* Let $[c_0 \cdots c_r]$ denote the following subset of Y: $\{\{d_i\}: d_i = c_i, 0 \le i \le r\}$. Let $\mathcal{A}_+$, $\mathcal{A}_{++}$, $\mathcal{A}_-$, $\mathcal{A}_{--}$ denote the σ-algebras generated by $\{\sigma^n[c]: c \in \mathcal{K}\}$ where n ranges over $\{n: n \le 0\}$, $\{n: n < 0\}$, $\{n: n \ge 0\}$, $\{n: n > 0\}$.

(1.7) *Definition.* A σ-invariant probability measure μ on Y is Gibbs if and only if:

(1) $\mu([c]) > 0$ for $c \in \mathcal{K}$.

(2) There exist constants A, $B > 0$ such that for all $[cd] \ne \varnothing$, and for all $f \in L^1(\mathcal{A}_-, \mu)$, $f \ge 0$, $(1 - \chi_{[c]})f = 0$, (for $\chi_{[c]}$ the characteristic function of $[c]$ and E_μ conditional expectation),

$$A \int f \, d\mu \, \chi_{\sigma^{-1}[d]} \le E_\mu(f \mid \mathcal{A}_{++}) \chi_{\sigma^{-1}[d]} \le B \int f \, d\mu \, \chi_{\sigma^{-1}[d]}.$$

(3) There exist constants B, $\alpha > 0$ such that, for all $f \in L^1(\mathcal{A}_-, \mu)$,

$$|E_\mu(f \mid \mathcal{A}_{++})(\mathbf{x}) - E_\mu(f \mid \mathcal{A}_{++})(\mathbf{y})| \le B(d(\mathbf{x}, \mathbf{y}))^\alpha \int |f| \, d\mu,$$

where

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \frac{d_1(x_i, y_i)}{2^i}, \quad \mathbf{x} = \{x_i\}, \quad \mathbf{y} = \{y_i\}$$

and

$$\begin{aligned} d_1(x_i, y_i) &= 1 \quad \text{if } x_i = y_i, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Note. This definition is equivalent to that in [**3**] though we do not need that here. However, the correspondence there $\phi \to \mu_\phi$ for Hölder-continuous functions, and the fact that μ maximizes $h_\mu(\sigma) + \int \phi \, d\mu$ (for h_μ denoting entropy) shows that there are many τ-invariant Gibbs measures – corresponding to τ-invariant ϕ, for example.

(1.8) Lemma. *Let ν be a conformal density on L_Γ of dimension δ. Let μ_ν denote both the corresponding Γ-invariant measure $d\mu_\nu(\xi, \eta) = c \, d\nu(\xi) \, d\nu(\eta)/|\xi - \eta|^{2\delta}$ on $L_\Gamma \times L_\Gamma$*

https://doi.org/10.1017/S0143385700001206 Published online by Cambridge University Press

(normalized so the corresponding measure on X_Γ has mass 1) and the corresponding τ- and σ-invariant probability measure on Y_Γ. Then μ_ν is Gibbs.

Proof. Let $c, d \in \mathcal{K}_\Gamma$, $c = e_0 e_1$, $d = e_1 e_2$, $e_i \in \mathcal{J}_\Gamma$. As in (1.1), let the set of geodesics through e_i be identified with $U_i \times V_i \times \mathbb{R}$, where $U_0 \subseteq U_1 \subseteq U_2$, $V_0 \supseteq V_1 \supseteq V_2$. Then

$$\mathcal{A}_- \cap [c] = \{B \cap [c] : B \in \mathcal{A}_-\} \text{ identifies with } \{U \times V_1 : U \subseteq U_0\}$$

$$\mathcal{A}_{++} \cap \sigma^{-1}[d] = \{B \cap \sigma^{-1}[d] : B \in \mathcal{A}_{++}\} \text{ identifies with } \{U_1 \times V : V \subseteq V_2\}.$$

So on $\sigma^{-1}[d] = U_1 \times V_2$, $E(f|\mathcal{A}_{++})(\xi, \eta)$ depends only on the second coordinate η, and if f is $\mathcal{A}_-$-measurable, and zero except on $U_0 \times V_1 = [c]$, f depends only on the first coordinate ξ, and

$$\chi_{U_1 \times V_2}(\xi, \eta) E_\mu(f|\mathcal{A}_{++})(\eta) = \frac{\displaystyle\int_{U_1} \frac{cf(\xi)}{|\xi - \eta|^{2\delta}}\, d\nu(\xi)}{\displaystyle\int_{U_1} \frac{c\, d\nu(\xi)}{|\xi - \eta|^{2\delta}}}.$$

Because $|\xi - \eta|$ is bounded above and below on $U_1 \times V_2$, and is C^1 in η, it is not hard to see that (2) is true, and (3) is true if the semi-metric d is replaced by the semi-metric ρ on $U_1 \times V_2$ given by

$$\rho((\xi_1, \eta_1), (\xi_2, \eta_2)) = |\eta_1 - \eta_2|$$

where $|\quad|$ denotes Euclidean metric on S^d. So we only need to show ρ and d are 'Lipshitz equivalent'. This follows from (1.9) since there exist constants A and $B > 0$ such that for any $[d_0 \cdots d_p] \neq \varnothing$, any p, $Ap \leq (x_0, \gamma x_0) \leq Bp$, if $\gamma = \phi(d_0)\phi(d_1) \cdots \phi(d_p)$, $x_0 \in H^{d+1}$ is fixed, and $(x_0, \gamma x_0)$ denotes hyperbolic distance. (These inequalities are true because any fundamental set of cross-sections $\mathcal{J}_1$ is bounded, and distance between two cross-sections is bounded below.)

(1.9) LEMMA. *Let $[d_0 \cdots d_p] \neq \varnothing$, $\phi(d_0) \cdots \phi(d_p) = \gamma$, $x_0 \in H^{d+1}$. Then there exist constants $C, D > 0$ such that*:

(1) *The ρ-diameter of $[d_0 \cdots d_p]$ is bounded above by $C \exp\{-(x_0, \gamma x_0)\}$, and $[d_0 \cdots d_p]$ contains a ball of ρ-diameter $D \exp\{-(x_0, \gamma x_0)\}$.*

(2) $C \exp\{-\delta(x_0, \gamma x_0)\} \leq \mu_\nu([d_0 \cdots d_p]) \leq D \exp\{-\delta(x_0, \gamma x_0)\}$.

Proof. Let $d_i = e_i e_{i+1}$, $e_j \in \mathcal{J}_\Gamma$. Let the cross-section lift of e_i in $\mathcal{J}_1$ (the fundamental set) correspond to $U_i \times V_i \subseteq S^d \times S^d$. So

$$\prod_{i=0}^{j-1} \phi(d_i) U_j \subseteq \prod_{i=0}^{j} \phi(d_i) U_{j+1}, \qquad \prod_{i=0}^{j-1} \phi(d_i) V_j \supseteq \prod_{i=0}^{j} \phi(d_i) V_{j+1}.$$

We need to know the Euclidean diameter, and ν-measure, of γV_p, and a lower bound on the diameter of the largest possible ball contained in γV_p. Since $U_0 \subseteq \gamma U_p$, the expanding point of γ is near U_p, hence bounded away from V_p. Thus the derivative of γ on V_p is boundedly proportional to $\exp\{-(x_0, \gamma x_0)\}$, whence the result. □

Given $e, e' \in \mathcal{J}_\Gamma$, and $\gamma \in \Gamma$, there is at most one non-empty cylinder set $[d_0 \cdots d_p]$ with $d_0 = ee_1$, and $d_p = e_p e'$ (for some $e_1, e_p \in \mathcal{J}_\Gamma$), and $\phi(d_0) \cdots \phi(d_p) = \gamma$. This follows from the Markov property (1.1), because if e identifies with $U \times V \subseteq L_\Gamma \times L_\Gamma$, and e' identifies with $U' \times V'$, where $U \subseteq \gamma U'$, $\gamma V' \subseteq V$, (1.1) implies the intervening U_i, V_i are uniquely determined. Thus, (1.9) gives:

(1.10) COROLLARY. *There exist constants* A, $B > 0$ *such that*

$$A \exp\{-\delta(x_0, \gamma x_0)\} \leq \sum_{p,[d_0 \cdots d_p]} \mu_\nu[d_0 \cdots d_p] \leq B \exp\{-\delta(x_0, \gamma x_0)\}$$

with $\phi(d_0) \cdots \phi(d_p) = \gamma$.

This will be needed in (4.7).

2. *Ergodic equivalence relations for Gibbs measures – a 'divergence type' condition*

In this section (Y, σ) is a topologically mixing subshift of type 2 on a finite set of symbols $\mathscr{K} = \{a_1 \cdots a_r, a_1^{-1} \cdots a_r^{-1}\}$, Y is invariant under τ, $\tau(\{x_i\}) = \{x_{-i}^{-1}\}$, and μ is a Gibbs measure on Y. For $F_1 \leq F$, the free group on $a_1 \cdots a_r$, $\sim_{F_1}$ is an equivalence relation on Y, as in (1.5). We find a 'divergence type' condition for the ergodicity of $\sim_{F_1}$. The proof, although it looks different, was originally based on that of ([**10**], § 7). We assume that (Y_{F_1}, σ) (as in (1.5)) is topologically transitive.

(2.1) LEMMA. (Y, σ, μ) *is strong mixing (hence ergodic).*

Proof. Define $\phi = \sum_{c \in \mathscr{K}} \chi_{[c]} \log E_\mu(\chi_{[c]} | \mathscr{A}_{++})$, with the convention $0 \log 0 = 0$. Then ϕ is Hölder-continuous with respect to the semi-metric d (1.7.3). In the notation of ([**3**], p. 13), $\mathscr{L}_\phi^* \mu = \mu$, $\mathscr{L}_\phi 1 = 1$, hence μ is Gibbs in the sense of [**3**], and strong mixing ([**3**], 1.14). □

Note. The lemma can also be proved directly, by approximating μ by Markov measures μ_m as in (3.4), and then applying a contraction mapping argument to the μ_m with a uniform contraction constant. (Part of (3.2) is needed for this.)

(2.2) LEMMA. $(Y, \sim_{F_1}, \mu)$ *is ergodic for* μ *Gibbs if and only if* $A = \{\mathbf{x}: \sigma^r \mathbf{x} \sim_{F_1} \mathbf{x}$, *some* $r > 0\}$ *has* μ*-measure* 1.

Proof. Suppose $\mu(A) < 1$. Let $B = \{\mathbf{x}: \sigma^r \mathbf{x} \sim_{F_1} \mathbf{x}$, some $r < 0\}$. We can define a μ-measure-preserving map $\psi: A \xrightarrow[\text{onto}]{} B$ by $\psi(\mathbf{x}) = \sigma^r(\mathbf{x})$, for r the least integer >0 with $\mathbf{x} \sim_{F_1} \sigma^r(\mathbf{x})$. By assumption, $0 < \mu(Y \backslash A) = \mu(Y \backslash B)$. Choose $a, b \in \mathscr{K}$ such that $\mu((Y \backslash A) \cap \{\mathbf{x}: x_0 = a\}) > 0$, $\mu((Y \backslash B) \cap \{\mathbf{x}: x_0 = b\}) > 0$. By topological transitivity, there exists an admissible sequence $a_0 \cdots a_n$ with $\pi a_i \in F_1$, $a_0 = b$, $a_n = a$. Let $C = \{\mathbf{x}$: there exist at most n integers $r_1 \cdots r_n$ with $\sigma^{r_i} \mathbf{x} \sim_{F_1} \mathbf{x}\}$. $\mu(C) < 1$ by topological transitivity of (Y_{F_1}, σ). $\mu(C) > 0$ by (1.7.2), because C contains

$$\{\mathbf{x}: x_i = a_i, 0 \leq i \leq n, x_i = y_{i-n}, i \geq n, \text{some } \mathbf{y} \in Y \backslash A, x_i = z_i, \text{some } \mathbf{z} \in Y \backslash B, i \leq 0\}.$$

C is a set of equivalence classes. So $\sim_{F_1}$ is not ergodic.

If $\mu(A) = 1$, then ψ is defined a.e. on Y. By the Martingale convergence theorem for $f \in L^1(\mathscr{A}_+, \mu)$, $\lim_{n \to \infty} E_\mu(f | \psi^{-n} \mathscr{A}_+)$ exists a.e. and equals $E\left(f \middle| \bigcap_{n=0}^{\infty} \psi^{-n} \mathscr{A}_+\right)$. But $\psi^{-n} \mathscr{A}_+ \subseteq \sigma^{-n} \mathscr{A}_+$ for $n \geq 0$, and $\bigcap_{n=0}^{\infty} \sigma^{-n} \mathscr{A}_+$ is trivial, so (Y, ψ, μ) is mixing, hence ergodic, hence $(Y, \sim_{F_1}, \mu)$ is ergodic. □

(2.3) *Definition.* Let $S_k^n = \sum\{\mu[x_0 \cdots x_{k-1}]$: there exist $i_0 = 0 < i_1 \cdots < i_n = k-1$ such that $x_{i_r+1}x_{i_r+2} \cdots x_{i_{r+1}} \in F_1$, and no such decomposition exists for larger $n\}$.

Let $S_k = \sum_n S_k^n$, $S^n = \sum_k S_k^n$.

Lemma 2.2 says $\sim_{F_1}$ is ergodic if and only if $S^1 = 1$.

(2.4) THEOREM. *$(Y, \sim_{F_1}, \mu)$ is ergodic if and only if $\sum_k S_k = \sum_n S^n = \infty$.*

Proof. If $S^1 = 1$, $S^n = 1$ for all n, and $\sum_n S^n = \infty$.

Conversely, suppose $S^1 < 1$. Let $B_k = \{\mathbf{x}: \psi^k(\mathbf{x})$ exists$\}$. Then $\mu(B_1) < 1$, by assumption. Choose $b \in \mathcal{K}$ such that $\mu((Y \backslash B_1) \cap [b]) > 0$. By topological transitivity, for each $a \in \mathcal{K}$, there exist r, $a_0 \cdots a_r$ with $a_0 = a$, $a_r = b$, and $a_0 \cdots a_{r-1} \in F_1$. Hence, by (1.7.2), $\mu([a_0 \cdots a_{r-1}] \cap \sigma^r(Y \backslash B_1)) > 0$. Hence there exist k, λ such that $\mu((Y \backslash B_k) \cap [a]) \geq \lambda > 0$ for all $a \in \mathcal{K}$.

B_n is open, hence can be represented as a disjoint union of cylinder sets. Write $B_{n,a,p}$ for the union of cylinder sets of length p which end in a.

$$\mu(B_{n,a,p} \cap \sigma^p((Y \backslash B_k) \cap [a])) \geq A\lambda\mu(B_{n,a,p}),$$

where $A < 1$ is as in (1.7.2). Hence

$$\mu(B_{n+k}) < (1 - \lambda A)\mu(B_n).$$

Hence, inductively,

$$S^{kn} < \lambda(1 - \lambda A)^{n-1}.$$

Hence

$$\sum_n S^n \leq k \sum_n S^{kn} < \infty. \qquad \square$$

We complete this section by noting that (2.4), together with the results of § 1, give part of the Aaronson–Sullivan result (see introduction).

(2.5) THEOREM. *Let Γ be a discrete group of isometries of H^{d+1} with X_Γ compact, Γ non-elementary, and ν a Γ-invariant conformal density of dimension $\delta = \delta(\Gamma)$. For $\Gamma_1 \leq \Gamma$ with $L_{\Gamma_1} = L_\Gamma$, $(L_\Gamma \times L_\Gamma, \Gamma_1, \mu_\nu)$ is ergodic if and only if $\sum_{\gamma \in \Gamma_1} \exp\{-\delta(x_0, \gamma x_0)\}$ diverges for any fixed $x_0 \in H^{d+1}$. (We are using the notation of the introduction.)*

Proof. This follows from (1.5), (1.8), (1.10) and (2.4). $\square$

3. *First stage in estimating the 'Poincaré series'*

Throughout this section, (Y, σ) is a topologically mixing subshift of type 2 on symbols $\mathcal{K} = \{a_1 \cdots a_r, a_1^{-1} \cdots a_r^{-1}\}$, and μ is a σ- and τ-invariant Gibbs measure on Y, where $\tau: \{x_i\} \mapsto \{x_{-i}^{-1}\}$ maps Y onto Y. F_1 is a fixed subgroup of the free group F on generators $a_1 \cdots a_r$ with $F/F_1 \cong \mathbb{Z}^v$, some $v > 0$. We fix a homomorphism with kernel F_1, $\boldsymbol{\theta}: F \to \langle\theta_1\rangle \oplus \cdots \oplus \langle\theta_v\rangle$, the free abelian group on generators $\theta_1 \cdots \theta_v$ (regarded as real variables). So for each $c \in \mathcal{K} \subseteq F$, $\boldsymbol{\theta}(c)$ is a linear function of the θ_i with integer coefficients. Sometimes, $\boldsymbol{\theta}$ or $\boldsymbol{\theta}(c)$ will mean evaluation at an element of $\mathbb{R}^v$ (or $(\mathbb{R}/2\pi)^v$).

We also make the assumption that (Y_{F_1}, σ) (as in (1.5)) is topologically transitive, hence with periodic points dense. (This is meant to include (Y_{Γ_1}, σ) if $Y = Y_\Gamma, \Gamma_1 \le \Gamma$ with Γ/Γ_1 abelian – see (1.5).)

In this section we begin to estimate

$$S_k = \sum \{\mu([c_0 \cdots c_{k-1}]) : [c_0 \cdots c_{k-1}] \neq \emptyset \text{ and } c_0 \cdots c_{k-1} \in F_1\}.$$

We call $\sum_{k=1}^{\infty} S_k$ the *Poincaré series for* μ, F_1 for a reason which is clear from (1.10). For a cylinder $[c_0 \cdots c_{k-1}]$, write $\boldsymbol{\theta}([c_0 \cdots c_{k-1}]) = \boldsymbol{\theta}(c_0 \cdots c_{k-1})$. If

$$S_k(\boldsymbol{\theta}) = S_k(\theta_1 \cdots \theta_v) = \sum_{c \text{ a } k\text{-cylinder}} \mu(\mathbf{c}) \exp\{i\boldsymbol{\theta}(\mathbf{c})\},$$

$$S_k(\boldsymbol{\theta}, \mathbf{x}) = \sum_{c \text{ a } k\text{-cylinder}} \chi_{\mathbf{c}}(\mathbf{x}) \exp\{i\boldsymbol{\theta}(\mathbf{c})\} \quad (\mathbf{x} \in Y),$$

then

$$S_k = \frac{1}{(2\pi)^v} \int_{[0,2\pi]^v} S_k(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \frac{1}{(2\pi)^v} \int_{[0,2\pi]^v} \int_Y S_k(\boldsymbol{\theta}, \mathbf{x})\, d\mu(\mathbf{x})\, d\boldsymbol{\theta}$$
$$= \frac{1}{(2\pi)^v} \int_{[0,2\pi]^v} \int_Y wA(\boldsymbol{\theta}, \sigma^{k-1}\mathbf{x}) A(\boldsymbol{\theta}, \sigma^{k-2}\mathbf{x}) \cdots A(\boldsymbol{\theta}, \mathbf{x})\, v(\boldsymbol{\theta}, \mathbf{x})\, d\mu(\mathbf{x})\, d\boldsymbol{\theta}.$$

Here, the rows and columns of the matrix $A(\boldsymbol{\theta}, \mathbf{x})$ and the rows of the column vector $v(\boldsymbol{\theta}, \mathbf{x})$ are indexed by $\{c : c \in \mathcal{K}\}$,

$$A(c, d)(\boldsymbol{\theta}, \mathbf{x}) = \exp\{i\boldsymbol{\theta}(c)\}\chi_{[dc]}(\mathbf{x}),$$
$$v(d)(\boldsymbol{\theta}, \mathbf{x}) = \exp\{i\boldsymbol{\theta}(d)\}\chi_{[d]}(\mathbf{x}),$$
$$w \text{ is the row vector } \underbrace{(1 \cdots 1)}_{2r}.$$

The rows and columns of a matrix $A_m(\boldsymbol{\theta})$ and the rows of the column vector $\nu_m(\boldsymbol{\theta})$, are indexed by $\{\mathbf{c} = [c_0 \cdots c_{m-1}] : \mathbf{c} \text{ is a non-empty } m\text{-cylinder}\}$:

$$A_m(\boldsymbol{\theta})(\mathbf{c}, \mathbf{d}) = \exp\{i\boldsymbol{\theta}(c_{m-1})\} \frac{\mu(\mathbf{d} \cap \sigma^{-1}\mathbf{c})}{\mu(\mathbf{d})},$$
$$v_m(\boldsymbol{\theta})(\mathbf{d}) = \exp\{i\boldsymbol{\theta}(\mathbf{d})\}\mu(\mathbf{d}),$$

w_m is the row vector of 1s with dimension equal to the number of non-empty m-cylinders.

The aim of this section is to prove:

(3.1) There exist constants $c > 0$ and $\eta < 1$ such that

$$|S_k(\boldsymbol{\theta}) - w_m A_m^{k-m}(\boldsymbol{\theta}) v_m(\boldsymbol{\theta})| < c((1 + c\eta^m)^{k-m} - 1).$$

(3.2) If $v = (v_i)$ is a vector in $\mathbb{C}^n$, let $\|v\|_1 = \sum_{i=1}^{n} |v_i|$ and for a $n \times n$ matrix $A = (a_{ij})$, let $\|A\|_1 = \sup_{\|v\|_1 = 1} \|Av\|_1 \le \sup_j \sum_i |a_{ij}|$.

(1) There exist s, B independent of m such that if $\|A_m(\boldsymbol{\theta})^{m+s} v\|_1 > 1 - \varepsilon$ for $\|v\|_1 = 1$, then either $|\boldsymbol{\theta}(c)| < B\varepsilon^{\frac{1}{8}}$ for all $c \in \mathcal{K}$ or $|\boldsymbol{\theta}(c) - \boldsymbol{\alpha}(c)| < B\varepsilon^{\frac{1}{8}}$ for all $c \in \mathcal{K}$.

If $z = A_m(\boldsymbol{\theta})^s v$, then in the first case $\|z - \exp(i\beta) v_m(\mathbf{0})\|_1 < B\varepsilon^{\frac{1}{8}}$, some $\beta \in \mathbb{R}$. In the second case, $\|z - \exp(i\beta)\Lambda_{\boldsymbol{\alpha}}^{-1} v_m(\mathbf{0})\|_1 < B\varepsilon^{\frac{1}{8}}$, some β. Here, $\boldsymbol{\alpha}$, $\Lambda_{\boldsymbol{\alpha}}$, are as in part (2).

(2) There exists at most one $\boldsymbol{\alpha}$ in {evaluations of $\boldsymbol{\theta}:\mathcal{K}\to\mathbb{R}/\langle 2\pi\rangle$} for which there is a solution γ to the equations

$$\gamma(c)+\boldsymbol{\alpha}(c)=\gamma(d)+\pi \bmod 2\pi, \quad \text{for all admissible } cd,$$
$$\boldsymbol{\alpha}(c)=0 \text{ or } \pi \bmod 2\pi, \quad \text{for each } c\in\mathcal{K},$$

and γ is unique up to addition of a constant, and we may assume $\gamma(c)=0$ or $\pi \bmod 2\pi$ for each $c\in\mathcal{K}$.

If $\Lambda_{\boldsymbol{\alpha}}$ is the diagonal matrix with rows and columns indexed by non-empty m-cylinders with

$$\Lambda_{\boldsymbol{\alpha}}(\mathbf{c},\mathbf{c})=\exp\{i\gamma(c_m)\} \quad \text{whenever} \quad \mathbf{c}=[c_0\cdots c_{m-1}] \quad \text{and} \quad c_{m-1}c_m \text{ is admissible}$$

(by the above equations, this is well-defined), then

$$A_m(\boldsymbol{\alpha})\Lambda_{\boldsymbol{\alpha}}v_m(\mathbf{0})=-\Lambda_{\boldsymbol{\alpha}}v_m(\mathbf{0}) \quad \text{and} \quad \Lambda_{\boldsymbol{\alpha}}^{-1}A_m(\boldsymbol{\alpha}+\boldsymbol{\theta})\Lambda_{\boldsymbol{\alpha}}=-A_m(\boldsymbol{\theta}).$$

This is clear from the definitions.

The motivation behind (3.1), (3.2) is to adopt a method Jon Aaronson showed me for evaluating S_k for a specific Markov measure, by approximating an arbitrary Gibbs measure function S_k by the corresponding function for approximating Markov measures (this is (3.1)), and showing the estimates for the approximating measures work, in some sense, uniformly. Part 1 of (3.2) shows that the functions $w_mA_m(\boldsymbol{\theta})^{k-m}v_m(\boldsymbol{\theta})$ tend to 0 at least as fast as $\nu^{k/m^{8t}}$ (for some $\nu<1$) outside neighbourhoods of $\mathbf{0}$, $\boldsymbol{\alpha}$ of width $O(1/m^t)$. Specifically, (3.1), (3.2) show:

(3.3) THEOREM. *For* $m^{8t+2}\le k\le m^u$

$$S_k=\frac{1}{(2\pi)^v}\int_{[-1/m^t,1/m^t]^v} w_mA_m(\boldsymbol{\theta})^{k-m}v_m(\boldsymbol{\theta}) + (-1)^{k-m}w_m\Lambda_{\boldsymbol{\alpha}}A_m(\boldsymbol{\theta})^{k-m}\Lambda_{\boldsymbol{\alpha}}^{-1}v_m(\boldsymbol{\theta}+\boldsymbol{\alpha})\,d\boldsymbol{\theta}+O(\eta^m)$$

for some $\eta<1$, *for any fixed t, u, where the second term is omitted if* $\boldsymbol{\alpha}$ *of* (3.2) *does not exist.*

(3.1) follows from (3.4), since the coefficients of the trigonometric polynomials $S_k(\boldsymbol{\theta})$ and $w_mA_m(\boldsymbol{\theta})^{k-m}v_m(\boldsymbol{\theta})$ are all positive and add to 1, if one of the coefficients of $S_k(\boldsymbol{\theta})$ is $\mu(c^1)+\cdots+\mu(c^n)$ for k-cylinders $c^1\cdots c^n$, then the corresponding coefficient of $w_mA_m(\boldsymbol{\theta})^{k-m}v_m(\boldsymbol{\theta})$ is $\mu_m(c^1)+\cdots+\mu_m(c^n)$ where μ_m is a Markov measure determined by the measure it gives to $(m+1)$-cylinders; that is, if $k\ge m$ and $[c_0\cdots c_k]\ne\varnothing$, then

$$\frac{\mu_m([c_0\cdots c_k])}{\mu_m([c_0\cdots c_{m-1}])}=\prod_{i=0}^{k-m}\frac{\mu([c_i\cdots c_{i+m}])}{\mu([c_i\cdots c_{i+m-1}])}$$

and $\mu_m([c_0\cdots c_m])=\mu([c_0\cdots c_m])$.

(3.4) LEMMA. *There exist* $c>0$, $\eta<1$ *such that for all* $k\ge m$

$$\frac{1}{(1+c\eta^m)^{k-m}}\mu[c_0\cdots c_k]\le\mu_m[c_0\cdots c_k]\le(1+c\eta^m)^{k-m}\mu[c_0\cdots c_k].$$

Proof. The statement is trivial for $k=m$ since $\mu=\mu_m$ on cylinders of length $\le m+1$. Assume the statement is true for $k-1$, $k>m$. Consider only the left-hand inequality

(the other is similar)

$$\mu([c_0\cdots c_k])=\frac{\mu([c_0\cdots c_m])}{\mu([c_1\cdots c_m])}\mu([c_1\cdots c_k])$$
$$+\left(\mu([c_0\cdots c_k])-\frac{\mu([c_0\cdots c_m])}{\mu([c_1\cdots c_m])}\mu([c_1\cdots c_k])\right).$$

By the inductive hypothesis, the first term is majorized by

$$(1+c\eta^m)^{k-1-m}\frac{\mu([c_0\cdots c_m])}{\mu([c_1\cdots c_m])}\mu_m([c_1\cdots c_k])=(1+c\eta^m)^{k-1-m}\mu_m([c_0\cdots c_k]).$$

For the second term,

$$\mu([c_0\cdots c_k])=\int_{[c_1\cdots c_k]}E(\chi_{[c_0]}\circ\sigma^{-1}|\mathscr{A}_+)\,d\mu.$$

By (1.7) 2–3, there exist $c>0$, $\eta<1$ such that

$$\left|E(\chi_{[c_0]}\circ\sigma^{-1}|\mathscr{A}_+)-\frac{\mu([c_0\cdots c_m])}{\mu([c_1\cdots c_m])}\right|<c\eta^m\frac{\mu([c_0\cdots c_m])}{\mu([c_1\cdots c_m])}$$

on $[c_1\cdots c_k]$.

So the second term is majorized by

$$c\eta^m\frac{\mu([c_0\cdots c_m])}{\mu([c_1\cdots c_m])}\mu([c_1\cdots c_k])$$

which, by the inductive hypothesis, is majorized by

$$c\eta^m(1+c\eta^m)^{k-1-m}\mu_m([c_0\cdots c_k]).$$

Adding gives the required result. □

In the proof of (3.2), the standard lemma 3.5 will be used:

(3.5) LEMMA. *Let $a_1\cdots a_n$, $b_1\cdots b_n$ be any real numbers with $b_i\geq 0$, $a_i\leq b_i$. Then for $\varepsilon>0$, if $\sum_{i=1}^n a_i>(1-\varepsilon)\sum_{i=1}^n b_i$ and $I=\{i: a_i\geq(1-\sqrt{\varepsilon})b_i\}$, then*

$$\sum_{i\notin I}b_i<\sqrt{\varepsilon}\sum_{i=1}^n b_i.$$

As an immediate corollary, using the mean value theorem:

(3.6) COROLLARY. *There exists a constant C such that, for any n, any complex numbers $a_1\cdots a_n$ with $\operatorname{Arg}(a_i)=\alpha_i$ and*

$$\operatorname{Arg}\left(\sum_{i=1}^n a_i\right)=\alpha,\quad \text{if } \left|\sum_{i=1}^n a_i\right|>(1-\varepsilon)\sum_{i=1}^n |a_i|$$

and $I=\{j: |\exp(i\alpha_j)-\exp(i\alpha)|\leq C\varepsilon^{\frac14}\}$ then

$$\sum_{i\notin I}|a_i|<\sqrt{\varepsilon}\sum_{i=1}^n|a_i|.$$

(3.7) LEMMA. *Let p be such that $\sigma^p[c]\cap[d]\neq\varnothing$ for any $c, d\in\mathscr{K}$ (p exists since (Y,σ) is topologically mixing). Then given ε small and r there exists $\alpha>0$ independent of m, ε*

such that if $\|v\|_1 \le 1$ *and* $\|A_m^{p+r}(\boldsymbol{\theta})v\|_1 > 1-\varepsilon$, *some* $\boldsymbol{\theta}$ *and* $w = (w(c)) = A_m^{p+r}(\boldsymbol{\theta})v$, *then*

$$\sum_{\substack{\mathbf{c}=[c_0\cdots c_{m-1}] \\ c_{m-r}\cdots c_{m-1}=e_0\cdots e_{r-1}}} |w(\mathbf{c})| > \alpha, \quad \text{for all non-empty } r\text{-cylinders } [e_0 \cdots e_{r-1}].$$

Proof. Write $A_m^{p+r}(\boldsymbol{\theta}) = (A(\mathbf{c}, \mathbf{d}))$. Write $\mathbf{e} = e_0 \cdots e_{r-1}$ and $\mathbf{c}_r = c_{m-r} \cdots c_{m-1}$. Then

$$\sum_{\substack{\mathbf{c}=[c_0\cdots c_{m-1}] \\ \mathbf{c}_r=\mathbf{e}}} |w(\mathbf{c})| = \sum_{\substack{\mathbf{c}=[c_0\cdots c_{m-1}] \\ \mathbf{c}_r=\mathbf{e}}} \left|\sum_{\mathbf{d}} A(\mathbf{c}, \mathbf{d})v(\mathbf{d})\right|$$

$$\ge (1-\sqrt{\varepsilon}) \sum_{\substack{\mathbf{c}\in I \\ \mathbf{c}_r=\mathbf{e}}} \sum_{\mathbf{d}} |A(\mathbf{c}, \mathbf{d})v(\mathbf{d})|,$$

where $I = \left\{\mathbf{c}: \left|\sum_{\mathbf{d}} A(\mathbf{c}, \mathbf{d})v(\mathbf{d})\right| \ge (1-\sqrt{\varepsilon})\sum_{\mathbf{d}}|A(\mathbf{c}, \mathbf{d})v(\mathbf{d})|\right\}$

$$\ge (1-\sqrt{\varepsilon}) \sum_{\mathbf{d}} |v(\mathbf{d})| \sum_{\mathbf{c}:\mathbf{c}_r=\mathbf{e}} |A(\mathbf{c}, \mathbf{d})| - \sqrt{\varepsilon} \quad \text{by (3.5)},$$

$$= (1-\sqrt{\varepsilon}) \sum_{\mathbf{d}} |v(\mathbf{d})| \sum_{\mathbf{c}:\mathbf{c}_r=\mathbf{e}} \frac{\mu_m(\sigma^{-p-r}\mathbf{c}\cap\mathbf{d})}{\mu(\mathbf{d})} - \sqrt{\varepsilon}$$

by definition of $A(\mathbf{c}, \mathbf{d})$,

$$\ge B_1(1-\sqrt{\varepsilon}) \sum_{\mathbf{d}} |v(\mathbf{d})| \sum_{\mathbf{c}:\mathbf{c}_r=\mathbf{e}} \frac{\mu(\sigma^{-p-r}\mathbf{c}\cap\mathbf{d})}{\mu(\mathbf{d})} - \sqrt{\varepsilon}$$

for B_1 independent of m by (3.4),

$$= B_1(1-\sqrt{\varepsilon}) \sum_{\mathbf{d}} |v(\mathbf{d})| \frac{\mu(\sigma^{-p-m}[\mathbf{e}]\cap\mathbf{d})}{\mu(\mathbf{d})} - \sqrt{\varepsilon}$$

$$\ge B(1-\sqrt{\varepsilon}) \sum_{\mathbf{d}} |v(\mathbf{d})| - \sqrt{\varepsilon}$$

some B, by (1.7.2) (because $\sigma^{-p-m}[\mathbf{e}]\cap\mathbf{d} \ne \emptyset$),

$$\ge B(1-\sqrt{\varepsilon})(1-\varepsilon) - \sqrt{\varepsilon} = \alpha$$

since $\|v\|_1 \ge 1-\varepsilon$, because, as is easily checked, $\|A_m(\boldsymbol{\theta})\|_1 \le 1$ for all m, $\boldsymbol{\theta}$.

(3.8) Lemma. *Again, let p be such that* $\sigma^p[c]\cap[d] \ne \emptyset$ *for any* $c, d \in \mathcal{K}$. *Then there exists D independent of m such that if* $\|A_m^{p+m}(\boldsymbol{\theta})v\|_1 > 1-\varepsilon$ *for some* $\boldsymbol{\theta}$, $\|v\|_1 \le 1$, *then there exist* $\{\gamma(e): e \in \mathcal{K}\}$ *such that*

$$\sum_{\mathbf{d}\notin I} |w(\mathbf{d})| < D\varepsilon^{\frac{1}{4}}, \quad \text{where } w(\mathbf{d}) = w = A_m^p(\boldsymbol{\theta})v,$$

$$I = \{\mathbf{d}: |w(\mathbf{d}) - \exp(i\gamma(e)|w(\mathbf{d})\|\le D|w(\mathbf{d})|\varepsilon^{\frac{1}{8}} \text{ whenever } d \in K_e\},$$

$$K_e = \{\mathbf{d} = [d_0 \cdots d_{m-1}]: d_{m-1}e \text{ admissible}\},$$

and ε *is sufficiently small independently of m.*

Proof. Write $A_m(\theta)^m = (E(\mathbf{c}, \mathbf{d}))$. Since $\sum_{\mathbf{c}} \left| \sum_{\mathbf{d}} E(\mathbf{c}, \mathbf{d}) w(\mathbf{d}) \right| > 1 - \varepsilon$, we have by (3.5),

$$\begin{aligned} \sum_{\substack{\mathbf{c}=[c_0\cdots c_{m-1}] \\ c_0=e}} \left| \sum_{\mathbf{d}} E(\mathbf{c}, \mathbf{d}) w(\mathbf{d}) \right| &\geq (1-\sqrt{\varepsilon}) \sum_{\substack{\mathbf{c}=[c_0\cdots c_{m-1}] \\ c_0=e}} \sum_{\mathbf{d}} |E(\mathbf{c}, \mathbf{d}) w(\mathbf{d})| - \sqrt{\varepsilon} \\ &\geq D(1-\sqrt{\varepsilon})\mu[e] \sum_{\substack{\mathbf{d}=[d_0\cdots d_{m-1}] \\ d_{m-1}e \text{ admissible}}} |w(\mathbf{d})| - \sqrt{\varepsilon} \\ &> D\alpha(1-\sqrt{\varepsilon})\mu[e] - \sqrt{\varepsilon} \end{aligned} \tag{1}$$

by (3.7), (1.7.2) (see (5) below).

Also by (3.5) there exists a set J of $\mathbf{c}$ such that

$$\left| \sum_{\mathbf{d}} E(\mathbf{c}, \mathbf{d}) w(\mathbf{d}) \right| > (1-\sqrt{\varepsilon}) \sum_{\mathbf{d}} |E(\mathbf{c}, \mathbf{d}) w(\mathbf{d})| \quad \text{for } \mathbf{c} \in J \tag{2}$$

and

$$\sum_{\mathbf{c} \notin J} \sum_{\mathbf{d}} |E(\mathbf{c}, \mathbf{d}) w(\mathbf{d})| < \sqrt{\varepsilon}.$$

By (1), for each $c_0 \in \mathcal{K}$, there exists $\mathbf{c} = [c_0 \cdots c_{m-1}] \in J$, if ε is sufficiently small independently of m.

Let $\gamma(\mathbf{c})$ be defined by

$$\operatorname{Arg}\left(\sum_{\mathbf{d}} E(\mathbf{c}, \mathbf{d}) w(\mathbf{d}) \right) = \boldsymbol{\theta}(\mathbf{c}) + \gamma(\mathbf{c}).$$

Then (3.6) and (2) imply there exists, for $\mathbf{c} \in J$, $L_{\mathbf{c}} \subseteq K_{c_0}$ ($\mathbf{c} = [c_0 \cdots c_{m-1}]$) such that, for $\mathbf{d} \in L_{\mathbf{c}}$,

$$|w(\mathbf{d}) - |w(\mathbf{d})| \exp\{i\gamma(\mathbf{c})\}| \leq C|w(\mathbf{d})|\varepsilon^{\frac{1}{8}} \quad \text{for } C \text{ independent of } m, \tag{3}$$

$$\sum_{\mathbf{d} \in K_{c_0} \setminus L_{\mathbf{c}}} |E(\mathbf{c}, \mathbf{d}) w(\mathbf{d})| \leq \varepsilon^{\frac{1}{4}} \sum_{\mathbf{d} \in K_{c_0}} |E(\mathbf{c}, \mathbf{d}) w(\mathbf{d})|. \tag{4}$$

(The facts that $\operatorname{Arg}(E(\mathbf{c}, \mathbf{d})) = \boldsymbol{\theta}(\mathbf{c})$ for all $\mathbf{d}$, and $E(\mathbf{c}, \mathbf{d}) \neq 0$ for $\mathbf{d} \in K_{c_0}$ have been used.)

(3.4) and (1.7.2) imply there exist A, B independent of m such that

$$A\mu([\mathbf{c}]) \leq |E(\mathbf{c}, \mathbf{d})| \leq B\mu([\mathbf{c}]) \quad \text{for } \mathbf{d} \in K_{c_0}, \text{ since } E(\mathbf{c}, \mathbf{d}) = \frac{\mu_m(\sigma^{-m}\mathbf{c} \cap \mathbf{d})}{\mu(\mathbf{d})}. \tag{5}$$

So (4) becomes:

$$\sum_{\mathbf{d} \in K_{c_0} \setminus L_{\mathbf{c}}} |w(\mathbf{d})| \leq \frac{B}{A} \varepsilon^{\frac{1}{4}} \sum_{\mathbf{d} \in K_{c_0}} |w(\mathbf{d})|. \tag{6}$$

So for ε sufficiently small independently of m, $L_{\mathbf{c}} \cap L_{\mathbf{c}'} \neq \emptyset$ if $\mathbf{c} = [c_0 \cdots c_{m-1}]$, $\mathbf{c}' = [c'_0 \cdots c'_{m-1}]$ and $c_0 = c'_0$. So (3), (6) become

$$\text{for } \mathbf{d} \in L_{c_0} |w(\mathbf{d}) - |w(\mathbf{d})| \exp\{i\gamma(c_0)\}| < 3C|w(\mathbf{d})|\varepsilon^{\frac{1}{8}} \tag{7}$$

and

$$\sum_{\mathbf{d} \in K_{c_0} \setminus L_{c_0}} |w(\mathbf{d})| < \frac{B}{A} \varepsilon^{\frac{1}{4}} \sum_{\mathbf{d} \in K_{c_0}} |w(\mathbf{d})|,$$

where $L_{c_0} = \bigcup \{L_{\mathbf{c}} : \mathbf{c} \in J, \ \mathbf{c} = [c_0 \cdots c_{m-1}]\}$ and $\gamma(c_0)$ is chosen to be $\gamma(\mathbf{c})$, some $\mathbf{c} = [c_0 \cdots c_{m-1}] \in J$.

https://doi.org/10.1017/S0143385700001206 Published online by Cambridge University Press

If $I = \bigcup_{c_0} L_{c_0}$ then

$$\sum_{\mathbf{d}\notin I} |w(\mathbf{d})| < \frac{2rB}{A} \varepsilon^{\frac{1}{4}}. \tag{8}$$

(7) and (8) give the result. □

Proof of (3.2). For some $s \geq p$ (p as in (3.7), (3.8)) yet to be chosen, we assume $\|A_m^{s+m}(\boldsymbol{\theta})v\|_1 > 1-\varepsilon$ for a v, $\|v\|_1 \leq 1$.

Let $w = w^0 = A_m^p(\boldsymbol{\theta})v$ as in (3.7), (3.8) and $w^t = A_m^{t+p}(\boldsymbol{\theta})v$, $0 \leq t \leq s-p$. Let $\{\gamma^t(e): e \in \mathcal{K}\}$ be the arguments corresponding to w^t, whose existences were proved in (3.8), i.e.

$$\text{for } \mathbf{c} \in I_t \cap K_e, \ |w^t(\mathbf{c}) - \exp\{i\gamma^t(e)\}|w^t(\mathbf{c})\| \leq D|w^t(\mathbf{c})|\varepsilon^{\frac{1}{8}} \tag{1}$$

and

$$\sum_{\mathbf{c}\notin I_t} |w^t(\mathbf{c})| < D\varepsilon^{\frac{1}{4}}.$$

However, we also have

$$\text{for } \mathbf{c} \in J_t, \ |w^t(\mathbf{c}) - \exp\{i\theta[c_{m-t}\cdots c_{m-1}] + i\gamma^0(c_{m-t})\}|w^t(\mathbf{c})\| < G\varepsilon^{\frac{1}{8}}|w^t(\mathbf{c})| \tag{2}$$

and

$$\sum_{\mathbf{c}\notin J_t} |w^t(\mathbf{c})| < G\varepsilon^{\frac{1}{8}},$$

for a G independent of m, if s is bounded independently of m. (2) follows from the fact that, if $A_m^t(\boldsymbol{\theta}) = (F_t(\mathbf{c}, \mathbf{d}))$, then $\arg F_t(\mathbf{c}, \mathbf{d}) = \boldsymbol{\theta}[c_{m-t}\cdots c_{m-1}]$ if $\mathbf{c} = [c_0 \cdots c_{m-1}]$, and then

$$\sum_{\mathbf{d}} F_t(\mathbf{c}, \mathbf{d})w^0(\mathbf{d}) = \exp\{i\boldsymbol{\theta}[c_{m-t}\cdots c_{m-1}]\} \sum_{\mathbf{d}} |F_t(\mathbf{c}, \mathbf{d})| w^0(\mathbf{d}) = w^t(\mathbf{c}).$$

Put

$$J_t = \left\{\mathbf{c}: \sum_{\mathbf{d}\notin I_0} |F_t(\mathbf{c}, \mathbf{d})\|w^0(\mathbf{d})| \leq \varepsilon^{\frac{1}{8}}|w^t(\mathbf{c})|\right\}.$$

$$\sum_{\mathbf{c}\notin J_t} |w^t(\mathbf{c})| \leq \frac{1}{\varepsilon^{\frac{1}{8}}} \sum_{\mathbf{d}\notin I_0} \sum_{\mathbf{c}} |F_t(\mathbf{c}, \mathbf{d})\|w^0(\mathbf{d})| \leq D\varepsilon^{\frac{1}{8}} \quad \text{by (1)}.$$

Combining (1) and (2) gives

$$|\exp\{i\gamma^t(d_t)\} - \exp\{i\theta[d_0\cdots d_{t-1}] + i\gamma^0(d_0)\}| < (D+G)\varepsilon^{\frac{1}{8}}, \tag{3}$$

for any non-empty cylinder $[d_0 \cdots d_t]$, any $t \leq s-p$, if ε is sufficiently small independently of m, since, by (3.7), the set of $\mathbf{c} = [c_0 \cdots c_{m-1}]$ with $c_{m-t}\cdots c_{m-1} = d_0 \cdots d_{t-1}$ is not contained in $I_t \cup J_t$.

Fix d_0, t, d_t and let $\boldsymbol{\theta}[d_1 \cdots d_{t-1}]$ vary with the restriction that $[d_0 \cdots d_t] \neq \emptyset$. For t large enough (depending only on (Y, σ), $\boldsymbol{\theta}$) the $\boldsymbol{\theta}[d_1 \cdots d_{t-1}] - \boldsymbol{\theta}[d'_1 \cdots d'_{t-1}]$ will generate a subgroup of finite index in $\langle\theta_1 \cdots \theta_v\rangle$ (since (Y_{F_1}, σ) is topologically transitive, and periodic points are dense). So there exist H, $q > 0$ (q integer) independent of m such that $\exp\{i\boldsymbol{\theta}(c)\}$ lies within $H\varepsilon^{\frac{1}{8}}$ of $\langle\exp(2\pi i/q)\rangle$ for all $c \in \mathcal{K}$. For ε sufficiently small independently of m, this uniquely defines $\boldsymbol{\theta}_0: \mathcal{K} \to \langle 2\pi/q\rangle/\langle 2\pi\rangle$ with $|\boldsymbol{\theta}(c) - \boldsymbol{\theta}_0(c)| < B\varepsilon^{\frac{1}{8}}$.

It has now been proved that $\boldsymbol{\theta}$ must lie within $O(\varepsilon^{\frac{1}{8}})$ of a finite set of points. The rest of the proof is algebraic manipulation – in the course of which we show the cardinality of the finite set is ≤ 2.

Replacing the $\gamma^t(c)$ by $\beta+\gamma_0^t(c)$ (some fixed $\beta\in\mathbb{R}$) which are $O(\varepsilon^{\frac{1}{8}})$ close, we can assume $\gamma_0^0(a)\in\langle 2\pi/q\rangle$, some $a\in\mathcal{K}$, and (3) can become

$$\gamma_0^t(d_t)=\boldsymbol{\theta}_0[d_0\cdots d_{t-1}]+\gamma_0^0(d_0)\quad \bmod 2\pi, \tag{4}$$

whenever $[d_0\cdots d_t]\neq\varnothing$. Since (Y,σ) is topologically mixing, we deduce that $\boldsymbol{\theta}_0$ and one $\gamma_0^0(a)$ determine all $\gamma_0^t(b)$ (all t, all $b\in\mathcal{K}$). In particular, all $\gamma_0^t(b)$ lie in $\langle 2\pi/q\rangle$. Since (4) is satisfied with γ_0^0, γ_0^t replaced by γ_0^1, γ_0^{t+1}, subtract the modified equation from (4), and deduce that, if t is large enough for $[a]\cap\sigma^{-t}[b]\neq\varnothing$ for all $a,b\in\mathcal{K}$,

$$\gamma_0^{t+1}(b)-\gamma_0^t(b)=\gamma_0^1(a)-\gamma_0^0(a)=\lambda_{\boldsymbol{\theta}_0} \bmod 2\pi, \tag{5}$$

for some constant $\lambda_{\boldsymbol{\theta}_0}\in\langle 2\pi/q\rangle$ for all $a,b\in\mathcal{K}$. So, putting $t=1$ in (4), we obtain

$$\gamma_0^0(b)+\lambda_{\boldsymbol{\theta}_0}=\boldsymbol{\theta}_0[a]+\gamma_0^0(a) \bmod 2\pi,\quad \text{whenever } [ab]\neq\varnothing, \tag{6}$$

where $\lambda_{\boldsymbol{\theta}_0}\in\langle 2\pi/q\rangle$ and $\gamma_0^0(a)\in\langle 2\pi/q\rangle$ for all $a\in\mathcal{K}$. $\boldsymbol{\theta}_0$ completely determines $\lambda_{\boldsymbol{\theta}_0}$, and determines the $\gamma_0^0(a)$ up to addition of a constant.

It is clear from (6) that the set of $\boldsymbol{\theta}_0$ we are considering lie in a finite group, and $\boldsymbol{\theta}_0\mapsto\lambda_{\boldsymbol{\theta}_0}$ is a group homomorphism. We shall show it is injective. So suppose $\lambda_{\boldsymbol{\theta}_0}=0$. (6) gives:

$$\gamma_0^0(b)=\gamma_0^0(a)\quad \text{whenever there exists } [d_0\cdots d_t]\neq\varnothing \text{ with } d_0=a,\ d_t=b, \tag{7}$$

and $\boldsymbol{\theta}_0[d_0\cdots d_{t-1}]=0$.

But this condition is satisfied for all $a,b\in\mathcal{K}$, since the shift $(Y_{\mathrm{Ker}\,\boldsymbol{\theta}_0},\sigma)$ (in the notation of (1.5)) is topologically transitive by assumption. Substituting in (6), we obtain $\boldsymbol{\theta}_0=0$. So $\boldsymbol{\theta}_0\mapsto\lambda_{\boldsymbol{\theta}_0}$ is injective.

Now for any $\boldsymbol{\theta}_0$, (6) implies $\gamma_0^0(d)=\gamma_0^0(e)$ if there exists c with cd, ce admissible. So if $\delta(d)=-\gamma_0^0(c^{-1})$ whenever cd is admissible, δ is well-defined.

We now use the uniqueness of γ_0^0 given $\boldsymbol{\theta}_0$. If $[cde]\neq\varnothing$,

$$\delta(e)+\lambda_{\boldsymbol{\theta}_0}=-\gamma_0^0(d^{-1})+\lambda_{\boldsymbol{\theta}_0}=-\gamma_0^0(c^{-1})+\boldsymbol{\theta}_0(d^{-1})=\delta(d)-\boldsymbol{\theta}_0(d).$$

Hence $\lambda_{\boldsymbol{\theta}_0}=\lambda_{-\boldsymbol{\theta}_0}=-\lambda_{\boldsymbol{\theta}_0}$, and $\lambda_{\boldsymbol{\theta}_0}=0$ or $\pi \bmod 2\pi$. By the injectivity of the homomorphism, there is at most one $\boldsymbol{\theta}_0$ with $\lambda_{\boldsymbol{\theta}_0}=\pi \bmod 2\pi$. Let $\boldsymbol{\alpha}$ be this $\boldsymbol{\theta}_0$ if it exists. The corresponding γ_0^0 satisfying (6) is the γ required in statement (2) of the theorem. □

4. *Second stage in estimating the 'Poincaré series'*

We continue with the notation of § 3, and the estimation of S_k. The main result is theorem 4.7. Recall that S_k depends on a σ-invariant and τ-invariant Gibbs measure μ on a subshift of finite type (Y,σ), and on a homomorphism $\boldsymbol{\theta}:F\to\mathbb{Z}^v$. Recall F is the free group on the symbols $\mathcal{K}=\{a_1\cdots a_r, a_1^{-1}\cdots a_r^{-1}\}$ of Y.

By theorem 3.3, we are reduced to estimating, for $m^{8t+2}\le k\le(m+1)^{8t+2}$ (any fixed t)

$$\frac{1}{(2\pi)^v}\int_{[-1/m^t,1/m^t]^v}[w_mA(\boldsymbol{\theta})^{k-m}v_m(\boldsymbol{\theta})+(-1)^{k-m}w_m\Lambda_{\boldsymbol{\alpha}}^{-1}A(\boldsymbol{\theta})^{k-m}\Lambda_{\boldsymbol{\alpha}}v_m(\boldsymbol{\theta}+\boldsymbol{\alpha})]\,d\boldsymbol{\theta}.$$

Here we are using the notation of (3.2). The general method is to obtain a local diagonalization of $A_m(\boldsymbol{\theta})$, hence reducing the calculation to estimating the integral of $\lambda_m(\boldsymbol{\theta})^{k-m}$, for $\lambda_m(\boldsymbol{\theta})$ the largest eigenvalue of $A_m(\boldsymbol{\theta})$. This integral is estimated by studying the second-order terms of $\lambda_m(\boldsymbol{\theta})$. As remarked before, this is a generalized version of a calculation for a specific Markov measure shown to me by Aaronson.

The main stages in the estimation are:

(4.1) LEMMA (needed for (4.2)). *$A_m(\boldsymbol{\theta})$ is conjugate to its adjoint $(A_m(\boldsymbol{\theta}))^*$, by some $C_m(\boldsymbol{\theta})$, with $C_m(\mathbf{0})$ fixing $v_m(\mathbf{0})$, and $C_m(\boldsymbol{\theta})$ continuous in $\boldsymbol{\theta}$.*

(4.2) THEOREM. *Suppose $A_m(\boldsymbol{\theta})$ is a $p\times p$ matrix. There exists a C^∞ map λ_m from $[-c/m^2, c/m^2]^v$ to $[0, 1]$, and a C^∞ map P_m from $[-c/m^2, c/m^2]^v$ to $\{P: P:\mathbb{R}^p\to\mathbb{R}^p$ is a projection with image space of dimension 1$\}$ for some $c>0$, such that λ_m, P_m have the following properties.* Ker $P_m(\boldsymbol{\theta})$, Im $P_m(\boldsymbol{\theta})$ are invariant under $A_m(\boldsymbol{\theta})$. $A_m(\boldsymbol{\theta})v=\lambda_m(\boldsymbol{\theta})v$ *for* $v\in \operatorname{Im} P_m(\boldsymbol{\theta})$. $\lambda_m(\mathbf{0})=1$, Ker $P_m(\mathbf{0})=\{(v(\mathbf{c})): \sum_{\mathbf{c}} v(\mathbf{c})=0\}$, Im $P_m(\mathbf{0})=$ sp $(\mu(\mathbf{c}))$. *There exist constants C_k, n_k independent of m such that $|D^k\lambda_m(\boldsymbol{\theta})|$, $\|D^kP_m(\boldsymbol{\theta})\|_1\le C_k m^{n_k}$. It will be useful to note that we can take $n_1=1$.*

(4.3) COROLLARY. *For $k\ge m^2$*

$$\int_{[-c/m^2,c/m^2]^v} w_mA_m(\boldsymbol{\theta})^{k-m}v_m(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$
$$=(1+O(1/m))\int_{[-c/m^2,c/m^2]^v}(\lambda_m(\boldsymbol{\theta}))^{k-m}\,d\boldsymbol{\theta}+O(\beta^m),\quad \textit{some } \beta<1.$$
$$\int_{[-c/m^2,c/m^2]^v}(-1)^{k-m}w_m\Lambda_{\boldsymbol{\alpha}}^{-1}A_m(\boldsymbol{\theta})^{k-m}\Lambda_{\boldsymbol{\alpha}}v_m(\boldsymbol{\theta}+\boldsymbol{\alpha})\,d\boldsymbol{\theta}$$
$$=(-1)^k(B_m+O(1/m))\int_{[-c/m^2,c/m^2]^v}(\lambda_m(\boldsymbol{\theta}))^{k-m}\,d\boldsymbol{\theta}+O(\beta^m),$$

for some B_m with $|B_m|\le 1$.

Proof. Write

$$v_m(\boldsymbol{\theta})=P_m(\boldsymbol{\theta})v_m(\boldsymbol{\theta})+(I-P_m(\boldsymbol{\theta}))v_m(\boldsymbol{\theta}),$$
$$w_m=w_m(P_m(\boldsymbol{\theta}))^T+w_m(I-P_m(\boldsymbol{\theta}))^T.$$

Thus, $w_mA_m(\boldsymbol{\theta})^{k-m}v_m(\boldsymbol{\theta})$ decomposes into four terms. By (3.2), $\|A_m(\mathbf{0})^{m+s}\|_1<\beta<1$ on Ker $P_m(\mathbf{0})$, for some $\beta<1$. By the given bounds on derivatives in (4.2), this estimate also holds for $\boldsymbol{\theta}$ with $|\theta_i|\le c/m^2$.

Also, $w_m(I-P_m(\mathbf{0})^T)P_m(\mathbf{0})v_m(\mathbf{0})=0$. By the bounds on derivatives, this quantity is $\le O(1/m)$ for $|\theta_i|\le c/m^2$. (Note that a bound on $\|P_m(\boldsymbol{\theta})-P_m(\mathbf{0})\|_1$ gives the same bound for $\|P_m(\boldsymbol{\theta})^T-P_m(\mathbf{0})^T\|_\infty$.) Thus, the dominating term of the four is

$$\lambda_m(\boldsymbol{\theta})^{k-m}\cdot w_mP_m(\boldsymbol{\theta})^TP_m(\boldsymbol{\theta})v_m(\boldsymbol{\theta}),$$

and, similarly, in the second part of the integral, the dominating term is

$$\lambda_m(\boldsymbol{\theta})^{k-m}\cdot(-1)^{k-m}w_m\Lambda_{\boldsymbol{\alpha}}^{-1}P_m(\boldsymbol{\theta})^TP_m(\boldsymbol{\theta})\Lambda_{\boldsymbol{\alpha}}v_m(\boldsymbol{\theta}+\boldsymbol{\alpha}).$$

The result follows, since for each of these dominating terms, the coefficient of $\lambda_m(\boldsymbol{\theta})^{k-m}$ is within $O(1/m)$ of the coefficient at $\boldsymbol{\theta}=\mathbf{0}$. □

Note. If we consider S_k+S_{k+1} instead of S_k we can just consider the integral

$$\int_{[-c/m^2,c/m^2]^v}\lambda_m(\boldsymbol{\theta})^{k-m}\,d\boldsymbol{\theta},$$

because then the second terms cancel. We have to do this, because it can happen that $S_k = 0$ for k odd in explicit examples. (For instance, the reduced word length of any element of the commutator subgroup of the free group on two generators is even.)

(4.4) THEOREM. *The first derivative of* λ_m *at* $\mathbf{0}$, $D\lambda_m(\mathbf{0})$, *is* $\mathbf{0}$. *The second derivative satisfies*

$$\begin{aligned}\sum_{i,j=1}^{v} \frac{\partial^2\lambda_m(\mathbf{0})}{\partial\theta_i\,\partial\theta_j}\theta_i\theta_j &= -\sum_{c\in\mathcal{K}} \mu(c)(\boldsymbol{\theta}(c))^2 \\ &\quad +2\sum_{c,d\in\mathcal{K}}\sum_{r=1}(\mu(\sigma^{-r}[d]\cap[c])-\mu([c])\mu([d]))\boldsymbol{\theta}(d)\boldsymbol{\theta}(c) \\ &\quad +H_m(\theta_1\cdots\theta_v) \\ &= G(\theta_1\cdots\theta_v)+H_m(\theta_1\cdots\theta_v),\end{aligned}$$

with $|H_m(\theta_1\cdots\theta_v|<A\beta^m(\theta_1^2+\cdots+\theta_v^2)$, *some* A, β, $\beta<1$, *and the expression for the quadratic polynomial* G *is convergent. Thus, the second-order terms of* λ_m *are essentially independent of m. Presumably, the same is also true of higher derivatives.*

(4.5) COROLLARY.

$$\begin{aligned}&\int_{[-1/m^{n_3+1},1/m^{n_3+1}]^v}\lambda_m(\boldsymbol{\theta})^{k-m}\,d\boldsymbol{\theta}\\ &\quad=\int_{[-1/m^{n_3+1},1/m^{n_3+1}]}\exp\{\tfrac12(k-m)(G(\boldsymbol{\theta})+O(|\boldsymbol{\theta}|^2))\}\,d\boldsymbol{\theta}.\end{aligned}$$

Proof. $x\mapsto\exp(x)$ has derivative and inverse derivative 1 at $x=0$. Because of the bound on third derivatives in (4.2), and the bound on H_m in (4.4), $|\lambda_m(\boldsymbol{\theta})-(1+\frac12G(\boldsymbol{\theta}))|\le O(|\boldsymbol{\theta}|^2/m)$ for $|\theta_i|\le 1/m^{n_3+1}$. □

(4.6) THEOREM.

$$\sum_{i,j=1}^{v}\frac{\partial^2\lambda_m(\mathbf{0})}{\partial\theta_i\,\partial\theta_j}\theta_i\theta_j\le -K(\theta_1^2+\cdots+\theta_v^2)$$

for all m, for some constant K.

The required theorem is now a corollary of this. We consider the integral in (4.5) for $k\ge m^{8n_3+10}$, and replace the variable $(\theta_1\cdots\theta_v)$ by $(k^{\frac12})(\theta_1\cdots\theta_v)$.

(4.7) THEOREM. *If* μ *is a* σ- *and* τ-*invariant Gibbs measure on* (Y,σ,τ) *and* F_1 *is a subgroup of the free group on the symbols* $\{a_1\cdots a_r, a_1^{-1}\cdots a_r^{-1}\}$ *of* Y *with* $F/F_1\cong_{\boldsymbol{\theta}}\mathbb{Z}^v$ *and* (Y_{F_1},σ) *topologically transitive, then*

$$S_k+S_{k+1}\sim\frac{2}{(2\pi)^v k^{v/2}}\int_{\mathbf{R}^v}\exp\{\tfrac12G(\theta_1\cdots\theta_v)\}\,d\theta_1\cdots d\theta_v,$$

where G *is a negative definite quadratic polynomial of rank* v, *and*

$$S_k=\sum\{\mu(\mathbf{c})\colon \mathbf{c}\text{ is a }k\text{-cylinder with }\boldsymbol{\theta}(\mathbf{c})=\mathbf{0}\}.$$

Hence, $(Y,\sim_{F_1},\mu)$ *is ergodic if and only if* $v\le 2$.

In particular, the assumption that (Y_{F_1},σ) *is topologically transitive holds if* (Y_F,σ) *and* (Y_{F_1},σ) *are simultaneous symbolic representations for the geodesic flows* $(X_\Gamma,\{\phi_t\})$ *and* $(X_{\Gamma_1},\{\phi_t\})$ *as in* § 1, *for* Γ *a discrete group of isometries with* X_Γ

https://doi.org/10.1017/S0143385700001206 Published online by Cambridge University Press

compact, $\Gamma_1 \le \Gamma$ with $\Gamma/\Gamma_1 \cong \mathbb{Z}^v$ and $F_1 = \phi^{-1}(\Gamma_1)$ for a homomorphism $\phi: F \to \Gamma$ like ϕ in (1.4). *In particular, μ can then be the measure corresponding to a Γ-invariant conformal density of dimension $\delta = \delta(\Gamma)$ on $L_\Gamma = L_{\Gamma_1}$, and the estimate* (1.10) *implies there exist constants A, $B > 0$ such that*

$$\frac{A}{k^{\frac{1}{2}v-1}} \le \sum_{\substack{Ak \le (x, \gamma x_0) \le Bk \\ \gamma \in \Gamma_1}} \exp\{-\delta(x_0, \gamma x_0)\} \le \frac{B}{k^{\frac{1}{2}v-1}}$$

for any fixed $x_0 \in H^{d+1}$, where $(x_0, \gamma x_0)$ denotes the hyperbolic distance between x_0 and γx_0.

Hence Γ_1 has the same critical exponent δ as Γ, and Γ_1 is of divergence type if and only if $v \le 2$.

We have given an outline of the proof. It remains to prove (4.2), (4.4), and (4.6). First we have to prove the lemma 4.1:

Proof of (4.1). Define $T: \mathbb{R}^p \to \mathbb{R}^p$ by $T(v(\mathbf{c})) = (w(\mathbf{c}))$ with $w(\mathbf{c}) = v(\mathbf{c}^{-1})$, where $\mathbf{c}^{-1} = [c_{m-1}^{-1} \cdots c_0^{-1}]$ if $\mathbf{c} = [c_0 \cdots c_{m-1}]$. Then if

$$A_m(\boldsymbol{\theta}) = A(\mathbf{c}, \mathbf{d})),\ TA_m(\boldsymbol{\theta})T^{-1} = (B(\mathbf{c}, \mathbf{d})),$$

where $B(\mathbf{c}, \mathbf{d}) = A(\mathbf{c}^{-1}, \mathbf{d}^{-1})$, so

$$\begin{aligned} B(\mathbf{c}, \mathbf{d}) &= \exp\{i\boldsymbol{\theta}(c_0^{-1})\} \frac{\mu[\mathbf{d}^{-1} \cap \sigma^{-1}\mathbf{c}^{-1}]}{\mu[\mathbf{d}^{-1}]} \\ &= \frac{\mu[\mathbf{c} \cap \sigma^{-1}\mathbf{d}]}{\mu[\mathbf{c}]} \cdot \exp\{-i\boldsymbol{\theta}(d_{m-1})\} \cdot \frac{\exp\{i\boldsymbol{\theta}[\mathbf{d}]\}}{\mu[\mathbf{d}]} \cdot \frac{\mu[\mathbf{c}]}{\exp\{i\boldsymbol{\theta}[\mathbf{c}]\}} \\ &= \overline{A(\mathbf{d}, \mathbf{c})} \cdot \frac{\exp\{i\boldsymbol{\theta}[\mathbf{d}]\}}{\mu[\mathbf{d}]} \cdot \frac{\mu[\mathbf{c}]}{\exp\{i\boldsymbol{\theta}[\mathbf{c}]\}}. \end{aligned}$$

So $(B(\mathbf{c}, \mathbf{d}))$ is conjugate to $(\overline{A(\mathbf{d}, \mathbf{c})}) = (A_m(\boldsymbol{\theta}))^*$ by a diagonal matrix. □

Proof of (4.2). Consider the function $F: \mathbb{R}^v \times \mathbb{C}^{p+1} \to \mathbb{C}^{p+1}$ given by

$$F(\boldsymbol{\theta}, \lambda, y_1 \cdots y_p) = \begin{pmatrix} (\lambda - A_m(\boldsymbol{\theta}))(\boldsymbol{\mu} + \mathbf{y}) \\ \sum_{i=1}^{p} y_i \end{pmatrix} \quad \text{where } \boldsymbol{\mu} = (\mu(\mathbf{c})),\ \mathbf{y} = (y_i).$$

Then $F(\mathbf{0}, 1, \mathbf{0}) = \mathbf{0}$. We want to use the implicit function theorem to solve the equation $F(\boldsymbol{\theta}, \lambda_m(\boldsymbol{\theta}), \mathbf{y}(\boldsymbol{\theta})) = \mathbf{0}$ for $\boldsymbol{\theta}$ near $\mathbf{0}$. We use the standard procedure of defining

$$\begin{pmatrix} \lambda_m^0(\boldsymbol{\theta}) \\ \mathbf{y}^0(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix}$$

$$\begin{pmatrix} \lambda_m^{r+1}(\boldsymbol{\theta}) \\ \mathbf{y}^{r+1}(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \lambda_m^r(\boldsymbol{\theta}) \\ \mathbf{y}^r(\boldsymbol{\theta}) \end{pmatrix} - (DF_{\lambda_m^r, \mathbf{y}^r})^{-1} F(\boldsymbol{\theta}, \lambda_m^r, \mathbf{y}^r),$$

choosing a suitable set of $\boldsymbol{\theta}$ for which $DF_{\lambda_m^r(\boldsymbol{\theta}), \mathbf{y}^r(\boldsymbol{\theta})}$ is invertible, and the sequence converges to a solution. Each component of F is a quadratic polynomial in λ, $\mathbf{y}$, quadratic term $\lambda \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$, and

$$DF_{\lambda, \mathbf{y}} = \begin{pmatrix} \boldsymbol{\mu} + \mathbf{y} & \lambda - A_m(\boldsymbol{\theta}) \\ 0 & 1 \cdots 1 \end{pmatrix}.$$

By (3.2), $\|(DF_{1,0})^{-1}\| \le Dm$, some constant D, since $\|A_m(\mathbf{0})^{m+s}\|_1 < \beta$, some $\beta < 1$ on $\mathrm{sp}\,\{(v(\mathbf{c})): \sum_{\mathbf{c}} v(\mathbf{c}) = 0\}$. So if $|\theta_i|$, $|\lambda^r_m - 1|\,\|\mathbf{y}^r\|_1 \le D'/m$, then $\|(DF_{\lambda^r_m,\mathbf{y}^r})^{-1}\|_1 \le 2Dm$. If $\|F(\boldsymbol{\theta}, \lambda^0_m, \mathbf{y}^0)\|_1 \le \varepsilon_0$ for $|\theta_i| \le b_m$, $i = 1 \cdots v$, then it can be proved inductively that, for such $\boldsymbol{\theta} = (\theta_1 \cdots \theta_v)$,

$$\left\|\begin{pmatrix} \lambda^r_m(\boldsymbol{\theta}) - 1 \\ \mathbf{y}^r(\boldsymbol{\theta}) \end{pmatrix}\right\|_1 \le \sum_{s=0}^{r-1} (2Dm)^{2^s} \varepsilon_0^{2^s} \quad (r \ge 1)$$

(if this is also $\le D'/m$), and

$$\|F(\boldsymbol{\theta}, \lambda^r_m(\boldsymbol{\theta}), \mathbf{y}^r(\boldsymbol{\theta}))\|_1 \le (2Dm)^{2^r - 1} \varepsilon_0^{2^r}.$$

Thus it suffices to make $|\theta_i|$, $\sum_{s=0}^{\infty} (2Dm)^{2^s} \varepsilon_0^{2^s} \le D'/m$, for which it suffices to make $\max_i |\theta_i| \le c/m^2$, some constant c.

$\begin{pmatrix} \lambda^r_m \\ \mathbf{y}^r \end{pmatrix}$ then converges to $\begin{pmatrix} \lambda_m \\ \mathbf{y} \end{pmatrix}$ satisfying

$$\frac{\partial}{\partial \theta_i}\begin{pmatrix} \lambda_m \\ \mathbf{y} \end{pmatrix} = (DF_{\lambda_m,\mathbf{y}})^{-1}\begin{pmatrix} \dfrac{\partial}{\partial \theta_i}(A_m(\boldsymbol{\theta}))(\mathbf{y} + \boldsymbol{\mu}) \\ 0 \end{pmatrix}.$$

Inductively, $\left\|D^k\begin{pmatrix} \lambda_m \\ \mathbf{y} \end{pmatrix}\right\|_1 \le E_k m^{n_k}$ for constants n_k, E_k, since $\|(DF_{\lambda,\mathbf{y}})^{-1}\| \le 2Dm$.

The bound on the $\|\;\;\|_1$ norm of $\begin{pmatrix} \boldsymbol{\mu} & I - A_m(\mathbf{0}) \\ 0 & 1 \cdots 1 \end{pmatrix}^{-1}$ gives a bound on the $\|\;\;\|_\infty$-norm of $\begin{pmatrix} 1 & I - A_m(\mathbf{0})^T \\ \vdots & \\ 1 & \\ 0 & \boldsymbol{\mu}^T \end{pmatrix}^{-1}$. Hence we can, by an exactly dual process, extend the eigenvalue 1 of $A_m(\mathbf{0})^T$ to an eigenvalue $\lambda_m(\boldsymbol{\theta})$ of $A_m(\boldsymbol{\theta})^T$, and the eigenvector $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ to an eigenvector $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + (z(\boldsymbol{\theta})(\mathbf{c}))$ (the eigenvalue is, of course, the same) with $\sum_{\mathbf{c}} \mu(\mathbf{c}) z(\boldsymbol{\theta})(\mathbf{c}) = 0$. All estimates are now in terms of the $\|\;\;\|_\infty$ norm. $P_m(\boldsymbol{\theta})$ is defined by its kernel and its image. Its image is $\mathrm{sp}\,(\boldsymbol{\mu} + \mathbf{y}(\boldsymbol{\theta}))$. Its kernel is $\mathrm{Ann}\,(\mathrm{sp}\,((1 \cdots 1) + \mathbf{z}(\boldsymbol{\theta})))$, where Ann denotes the annihilator. Using the duality of the $\|\;\;\|_1$ and $\|\;\;\|_\infty$ norms, we obtain the bounds on the $\|\;\;\|_1$ norms of $P_m(\boldsymbol{\theta})$ and its derivatives.

By lemma 4.1, $\overline{\lambda_m(\boldsymbol{\theta})}$ is also an eigenvalue of $A_m(\boldsymbol{\theta})$, and since $A_m(\boldsymbol{\theta}) - \overline{\lambda_m(\boldsymbol{\theta})}$ has kernel of dimension one for $|\theta_i| \le c/m$ (because $\|A_m(\boldsymbol{\theta})^{m+s}\|_1 < \beta < 1$ on $\mathrm{sp}\left(v: \sum_{\mathbf{c}} v(\mathbf{c}) = 0\right)\Big)$, the corresponding eigenvector extends $\boldsymbol{\mu}$ smoothly. By the uniqueness in the implicit function theorem, $\lambda_m(\boldsymbol{\theta}) = \overline{\lambda_m(\boldsymbol{\theta})}$, and so $\lambda_m(\boldsymbol{\theta})$ is real. So, since clearly $|\lambda_m| \le 1$, λ_m maps into $[0, 1]$. □

https://doi.org/10.1017/S0143385700001206 Published online by Cambridge University Press

Proof of (4.4). Let F be as in (4.2). Note that

$$(DF_{1,\mathbf{0}})^{-1}=\begin{pmatrix}1\cdots 1 & 0\\ B & \boldsymbol{\mu}\end{pmatrix}$$

where, if

$$M=\underbrace{(\boldsymbol{\mu}\cdots\boldsymbol{\mu})}_{p\text{ times}},$$

then $B(I-A_m(\mathbf{0}))=(I-A_m(\mathbf{0}))B=I-M=I-P_m(\mathbf{0})$, and $B\boldsymbol{\mu}=\mathbf{0}$. B exists since $I-A_m(\mathbf{0})$ has one-dimensional kernel by (3.2). Moreover, if $\sum_{\mathbf{c}} v(\mathbf{c})=0$, then $Bv=\sum_{r=0}^{\infty} A_m(\mathbf{0})^r v$, and by (3.2) this series converges, since on this subspace $\|A_m(\mathbf{0})^{m+s}\|_1<\beta<1$, some β. Recall from (4.2) that

$$\frac{\partial}{\partial\theta_i}\begin{pmatrix}\lambda_m\\ \mathbf{y}\end{pmatrix}=(DF_{\boldsymbol{\lambda},\mathbf{y}})^{-1}\begin{pmatrix}\dfrac{\partial}{\partial\theta_i}(A_m(\boldsymbol{\theta})(\mathbf{y}+\boldsymbol{\mu}))\\ 0\end{pmatrix}. \tag{1}$$

Differentiating this, we see that $\partial^2\lambda_m/\partial\theta_j\,\partial\theta_i$ is the first row of

$$(DF_{\boldsymbol{\lambda},\mathbf{y}})^{-1}\begin{pmatrix}\dfrac{\partial^2 A_m}{\partial\theta_j\,\partial\theta_i}(\boldsymbol{\mu}+\mathbf{y})\\ 0\end{pmatrix}-(DF_{\boldsymbol{\lambda},\mathbf{y}})^{-1}\begin{pmatrix}\dfrac{\partial\mathbf{y}}{\partial\theta_j} & \left(\dfrac{\partial\lambda_m}{\partial\theta_j}I-\dfrac{\partial A_m}{\partial\theta_j}\right)\\ 0 & 0\cdots 0\end{pmatrix}(DF_{\boldsymbol{\lambda},\mathbf{y}})^{-1}\begin{pmatrix}\dfrac{\partial A_m}{\partial\theta_i}(\boldsymbol{\mu}+\mathbf{y})\\ 0\end{pmatrix}$$

$$+(DF_{\boldsymbol{\lambda},\mathbf{y}})^{-1}\begin{pmatrix}0 & \dfrac{\partial A_m}{\partial\theta_i}\\ \vdots & \\ 0 & 0\end{pmatrix}(DF_{\boldsymbol{\lambda},\mathbf{y}})^{-1}\begin{pmatrix}\dfrac{\partial A_m}{\partial\theta_j}(\boldsymbol{\mu}+\mathbf{y})\\ 0\end{pmatrix}-(DF_{\boldsymbol{\lambda},\mathbf{y}})^{-1}\begin{pmatrix}0 & \dfrac{\partial A_m}{\partial\theta_i}\\ \vdots & \\ 0 & 0\end{pmatrix}\begin{pmatrix}\dfrac{\partial\lambda_m}{\partial\theta_j}\\ 0\end{pmatrix}. \tag{2}$$

(1), together with the fact that $\mu[c]=\mu[c^{-1}]$, and $\boldsymbol{\theta}[c]=-\boldsymbol{\theta}[c^{-1}]$, gives $\partial\lambda_m(\mathbf{0})/\partial\theta_i=0$. $\sum_{\mathbf{c}} y(\mathbf{c})=0$ gives $\sum_{\mathbf{c}}\partial y(\mathbf{c})/\partial\theta_i=0$. So

$$\frac{\partial^2\lambda_m}{\partial\theta_j\,\partial\theta_i}(\mathbf{0})=(1\cdots 1)\left(\frac{\partial^2 A_m(\mathbf{0})}{\partial\theta_j\,\partial\theta_i}+\frac{\partial A_m(\mathbf{0})}{\partial\theta_j}B\frac{\partial A_m(\mathbf{0})}{\partial\theta_i}+\frac{\partial A_m(\mathbf{0})}{\partial\theta_i}B\frac{\partial A_m(\mathbf{0})}{\partial\theta_j}\right)\boldsymbol{\mu}.$$

If $v=(v(\mathbf{c}))=(\mu(\mathbf{c})\cdot i\boldsymbol{\theta}(c_{m-1}))$, where $\mathbf{c}=[c_0\cdots c_{m-1}]$ and Δ is the diagonal matrix with $\Delta(\mathbf{c},\mathbf{c})=\mu(\mathbf{c})$, then

$$\sum_{i,j=1}^{v}\theta_i\theta_j\frac{\partial^2\lambda_m(\mathbf{0})}{\partial\theta_i\,\partial\theta_j}=v^T(\Delta^{-1}+2\Delta^{-1}A_m(\mathbf{0})B)v.$$

Letting μ_m be the Markov measure of § 3, and recalling the definition of $A_m(\boldsymbol{\theta})$, we find that

$$\sum_{i,j=1}^{v}\theta_i\theta_j\frac{\partial^2\lambda_m(\mathbf{0})}{\partial\theta_i\,\partial\theta_j}=-\sum_{c\in\mathcal{X}}\mu_m(c)(\boldsymbol{\theta}(c))^2$$
$$+2\sum_{c,d\in\mathcal{X}}\sum_{r=1}^{\infty}(\mu_m(\sigma^{-r}[d]\cap[c])-\mu_m[d]\mu_m[c])\boldsymbol{\theta}(c)\boldsymbol{\theta}(d). \tag{3}$$

(The last term is added on for convenience. It is zero since $\mu_m[c]=\mu_m[c^{-1}]$ and $\boldsymbol{\theta}(c)=-\boldsymbol{\theta}(c^{-1})$.) We claim that

$$\sum_{c,d\in\mathcal{X}}\sum_{r=m(m+s)}^{\infty}|\mu_m(\sigma^{-r}[d]\cap[c])-\mu_m([d])\mu_m([c])|\le B(m+s)\sum_{r=m}^{\infty}\beta^r \tag{4}$$

for some constant B. This is because $\mu_m(\sigma^{-r}[d]\cap[c])-\mu_m([d])\mu_m([c])$ is $w_d^T A_m(\mathbf{0})^{m+r}(v_c-\mu_m(c)\boldsymbol{\mu})$, where

$$\begin{aligned} v_c(\mathbf{e}) &= \mu_m(\mathbf{e}) \quad \text{if } \mathbf{e}=[e_0\cdots e_{m-1}] \quad \text{with } e_{m-1}=c,\\ &=0 \qquad \text{otherwise},\\ w_d(\mathbf{e}) &= 1 \qquad \text{if } e_0=d,\\ &=0 \qquad \text{otherwise}. \end{aligned}$$

By (3.2), this is majorized by $B\mu_m(c)\beta^{(m+r)/(m+s)}$, because $\|v_c-\mu_m(c)\cdot\boldsymbol{\mu}\|_1\le 2\mu_m(c)$, and the sum of the coefficients of $v_c-\mu_m(c)\cdot\boldsymbol{\mu}$ is 0.

We shall need in (4.6) (and can use now) a result from [**3**], 1.10–1.14, which cannot be deduced from § 3 here.

If μ is Gibbs, there exist constants $A, \beta, \beta<1$, such that if $\mathbf{a}, \mathbf{b}$ are any two cylinder sets with $\mathbf{a}$ length t,

$$|\mu(\mathbf{a}\cap\sigma^{-r}\mathbf{b})-\mu(\mathbf{a})\mu(\mathbf{b})|<A\beta^{r-t}\mu(\mathbf{a})\mu(\mathbf{b}). \tag{5}$$

It follows from (5) that the series

$$\sum_{r=1}^{\infty}|\mu(\sigma^{-r}[d]\cap[c])-\mu([d])\mu([c])|$$

converges. By (3.4), the earlier terms in the series (3) are approximated by the corresponding ones for μ. Thus $\sum_{i,j=1}^{v}\theta_i\theta_j\,\partial^2\lambda_m(\mathbf{0})/\partial\theta_i\,\partial\theta_j$ tends to

$$-\sum_{c\in\mathcal{K}}\mu(c)(\boldsymbol{\theta}(c))^2+2\sum_{c,d\in\mathcal{K}}\sum_{r=1}(\mu(\sigma^{-r}[d]\cap[c])-\mu([c])\mu([d]))\cdot\boldsymbol{\theta}([c])\boldsymbol{\theta}([d])$$

as $m\to\infty$, the difference being $\le O(\eta^m)$, some $\eta<1$. □

Proof of (4.6). Instead of proving $\lambda_m(\boldsymbol{\theta})$ is boundedly negative definite of rank v, we shall prove it for $(\lambda_m(\boldsymbol{\theta}))^p$, for some suitable p independent of m. This is the same, because $D\lambda_m(\mathbf{0})=0$ implies $D^2\lambda_m^p(\mathbf{0})=p\cdot D^2\lambda_m(\mathbf{0})$. We can also obtain an expression for $D^2(\lambda_m^p)$ by differentiating

$$\begin{pmatrix}(\lambda_m^p-A_m(\boldsymbol{\theta})^p)(\boldsymbol{\mu}+\mathbf{y})\\ \sum_{\mathbf{c}} y(\mathbf{c})\end{pmatrix}=\mathbf{0},$$

as we did for $p=1$ in (4.4). Then we obtain

$$\sum_{i,j=1}^{v}\theta_i\theta_j\frac{\partial^2(\lambda_m^p)(\mathbf{0})}{\partial\theta_i\,\partial\theta_j}=v_p^T(\Delta^{-1}+2\Delta^{-1}A_m(\mathbf{0})^pB_p)v_p, \tag{1}$$

where $v_p=(v_p(\mathbf{c}))$, $v_p(\mathbf{c})=\mu(\mathbf{c})\cdot i\boldsymbol{\theta}[c_{m-p}\cdots c_{m-1}]$ if $\mathbf{c}=[c_0\cdots c_{m-1}]$, and $B_p(I-A_m(\mathbf{0})^p)=(1-A_m(\mathbf{0})^p)B_p=I-M$, for M as in theorem 4.4, so that

$$B_pv_p=\sum_{r=0}^{\infty}A_m(\mathbf{0})^{rp}v_p.$$

Let

$$B_pv_p=i\Delta w_p. \tag{2}$$

So $v_p=i(I-A_m(\mathbf{0})^p)\Delta w_p$, and w_p is real. Then

$$v_p^T(\Delta^{-1}+2\Delta^{-1}A_m(\mathbf{0})^pB_p)v_p=-w_p^T(\Delta-\Delta(A_m(\mathbf{0})^p)^T\Delta^{-1}A_m(\mathbf{0})^p\Delta)w_p.$$

https://doi.org/10.1017/S0143385700001206 Published online by Cambridge University Press

Write $\Delta(A_m(\mathbf{0})^p)^T\Delta^{-1}A_m(\mathbf{0})^p\Delta=(E_p(\mathbf{c},\mathbf{d}))$. Then

$$E_p(\mathbf{c},\mathbf{d})=\sum_{\mathbf{e}\ m\text{-cylinder}}\frac{\mu_m(\sigma^{-p}\mathbf{c}\cap\mathbf{e})\mu_m(\sigma^{-p}\mathbf{d}\cap\mathbf{e})}{\mu_m(\mathbf{e})} \tag{3}$$

Note that

$$\sum_{\mathbf{c}}E_p(\mathbf{c},\mathbf{d})=\sum_{\mathbf{c}}E_p(\mathbf{d},\mathbf{c})=\mu_m(\mathbf{d}).$$

Note that, for any matrix (a_{ij}), if $\sum_j a_{ij}=\sum_j a_{ji}=a_i$, then

$$\begin{aligned}\sum_i a_ix_i^2-\sum_{i,j}a_{ij}x_ix_j&=\tfrac{1}{2}\Big(\sum_{i,j}a_{ij}x_i^2+\sum_{i,j}a_{ij}x_j^2-2\sum_{i,j}a_{ij}x_ix_j\Big)\\&=\tfrac{1}{2}\sum_{i,j}a_{ij}(x_i-x_j)^2.\end{aligned}$$

Thus, (1) becomes

$$\sum_{i,j=1}^{v}\theta_i\theta_j\frac{\partial^2\lambda_m^p(\mathbf{0})}{\partial\theta_i\,\partial\theta_j}=-\tfrac{1}{2}\sum_{\mathbf{c},\mathbf{d}}E_p(\mathbf{c},\mathbf{d})(w_p(\mathbf{c})-w_p(\mathbf{d}))^2. \tag{4}$$

From (2),

$$w_p(\mathbf{c})=(1/\mu_m(\mathbf{c}))\sum_{r=0}\ \sum_{\mathbf{d}\ m\text{-cylinder}}\mu_m(\sigma^{-pr}\mathbf{c}\cap\mathbf{d})\boldsymbol{\theta}[d_{m-p}\cdots d_{m-1}]$$

(for $\mathbf{d}=[d_0\cdots d_{m-1}]$),

$$\begin{aligned}w_p(\mathbf{c})&=(1/\mu_m(\mathbf{c}))\sum_{r=1}\sum_{d\in\mathscr{K}}\mu_m(\sigma^{m-r}\mathbf{c}\cap\mathbf{d})\boldsymbol{\theta}(\mathbf{d}),\\&=\boldsymbol{\theta}[\mathbf{c}]+(1/\mu_m(\mathbf{c}))\sum_{r=1}\sum_{d\in\mathscr{K}}\mu_m(\sigma^{-r}\mathbf{c}\cap\mathbf{d})\boldsymbol{\theta}(\mathbf{d}),\\&=\boldsymbol{\theta}[\mathbf{c}]+(1/2\mu_m(\mathbf{c}))\sum_{r=1}\sum_{d\in\mathscr{K}}\boldsymbol{\theta}(\mathbf{d})(\mu_m(\sigma^{-r}\mathbf{c}\cap\mathbf{d})-\mu_m(\sigma^{-r}\mathbf{c}\cap\mathbf{d}^{-1})).\end{aligned}$$

Hence we claim

$$|w_p(\mathbf{c})-\boldsymbol{\theta}[\mathbf{c}]|\le K_1(\theta_1^2+\cdots+\theta_v^2)^{\frac{1}{2}}. \tag{5}$$

For by (4) of (4.4), the tail of the series ($r\ge m(m+s)$) tends to 0. By (3.4), the terms $r\le m(m+s)$ can be replaced by the corresponding ones for μ. By (5) of (4.4), we can bound $|\mu(\sigma^{-r}\mathbf{c}\cap d)-\mu(\sigma^{-r}\mathbf{c}\cap d^{-1})|$ by $2A\beta^{r-1}\mu(\mathbf{c})\mu(d)$, and the claim is proved.

Now $E_p(\mathbf{c},\mathbf{d})\ne 0$ only if $[c_0\cdots c_{m-p-1}]=[d_0\cdots d_{m-p-1}]$, for $\mathbf{c}=[c_0\cdots c_{m-1}]$ and $\mathbf{d}=[d_0\cdots d_{m-1}]$, in which case, by (3) and (1.7.2), $E_p(\mathbf{c},\mathbf{d})\ge \alpha_p\cdot\max(\mu_m[\mathbf{c}],\mu_m[\mathbf{d}])$, for α_p independent of m (but not p). By topological transitivity of $(Y_{\mathrm{Ker}\,\boldsymbol{\theta}},\sigma)$, we can find p, and two p-cylinders $\mathbf{c}'$, $\mathbf{d}'$ with $c_0'=d_0'$ ($\mathbf{c}'=[c_0'\cdots c_{p-1}']$ and $\mathbf{d}'=[d_0'\cdots d_{p-1}']$) and

$$|\boldsymbol{\theta}(\mathbf{c}')-\boldsymbol{\theta}(\mathbf{d}')|\ge 3K_1(\theta_1^2+\cdots+\theta_v^2). \tag{6}$$

Then, if $\mathbf{c}=[c_0\cdots c_{m-p-1},c_0'\cdots c_{p-1}']$, $\mathbf{d}=[c_0\cdots c_{m-p-1},d_0'\cdots d_{p-1}]$, from (5), (6) we have

$$|w(\mathbf{c})-w(\mathbf{d})|\ge K_1(\theta_1^2+\cdots\theta_v^2)^{\frac{1}{2}}. \tag{7}$$

https://doi.org/10.1017/S0143385700001206 Published online by Cambridge University Press

The sum of the $E(\mathbf{c}, \mathbf{d})$ for such $\mathbf{c}$, $\mathbf{d}$ is minorized by $\alpha_p\mu[c'_0 \cdots c'_{p-1}]$, which is independent of m. So

$$\sum_{i,j} \frac{\partial^2 \lambda_m^p(\mathbf{0})}{\partial\theta_i\,\partial\theta_j}\theta_i\theta_j \leq -K_1^2\alpha_p\mu[c'_0 \cdots c'_{p-1}](\theta_1^2+\cdots+\theta_v^2),$$

and hence the expression $\sum_{i,j} \frac{\partial^2 \lambda_m(\mathbf{0})}{\partial\theta_i\,\partial\theta_j}\theta_i\theta_j$ is boundedly negative definite as required. □

5. *Finitely determined subabelian groups*

The results in this section will be rather sketchy. As in §§ 2–4, we consider a topologically mixing subshift of finite type (Y, σ) on symbols $\mathcal{K} = \{a_1 \cdots a_r, a_1^{-1} \cdots a_r^{-1}\}$ with a τ-invariant Gibbs measure μ on Y. For $G \leq F$, the free group on $a_1 \cdots a_r$, $\sim_G$ is defined as in (1.5). We find a condition for $(Y, \sim_G, \mu)$ to be ergodic, for G 'finitely determined subabelian'. No attempt will be made to translate the definition of finitely determined to a subgroup of isometries of Γ, because I am not sure of the best way to do this in general. However, for a Schottky group Γ, when we can take $F = \Gamma$, the symbolic dynamics need no interpretation.

(5.1) *Definition.* $F_r \leq F$ *is subabelian finitely determined of degree r with chain* $F_1, F_2 \cdots F_r$ if:

(1) $F = F_0 \rhd F_1 \rhd F_2 \cdots \rhd F_r$ with $F_i/F_{i-1} \cong \mathbb{Z}^{v_i}$, $v_i = v_i(F)$.

(2) There exists a set of free generators and their inverses, W, of F_1, with a *finite* $W_0^{-1} = W_0 \subseteq W$, $W_1 = W - W_0$, such that $F_i \geq \operatorname{Ker} \pi$, $i \geq 2$, where $\pi : F_1 = F_W \to F_{W_0}$ is the homomorphism obtained by deleting all symbols of W_1 in a word in F_W, and such that $F_r/\operatorname{Ker}\pi$ is subabelian finitely determined in $F_1/\operatorname{Ker}\pi$ $(\cong F_{W_0})$ of degree $r-1$ with chain

$$F_2/\operatorname{Ker}\pi, \ldots, F_r/\operatorname{Ker}\pi.$$

Thus this definition is inductive on r. We start by defining F_1 subabelian finitely determined of degree 1 if $F/F_1 \cong \mathbb{Z}^{v_1}$, some v_1. Note this condition eliminates the possibility $F \rhd F_r$ and $F/F_r \cong \mathbb{Z}^{v_1+\cdots+v_r}$ $(r>1)$. It is easy to construct examples of subabelian finitely determined subgroups.

(5.2) THEOREM. *If F_r is subabelian finitely determined of degree r in F with chain $F_1 \cdots F_r$, and $v_i = v_i(F)$, and (Y_{F_r}, σ) (as in (1.5)) is topologically transitive, then $(Y, \sim_{F_r}, \mu)$ is ergodic if and only if $v_i(F) \leq 2$, $i = 1 \cdots r$.*

Proof. §§ 2–4 show the theorem is true for $r = 1$. The proof is by induction. Suppose it is true for all subshifts and subgroups with $r-1$. Suppose $v_1 \leq 2$, so that $(Y, \sim_{F_1}, \mu)$ is ergodic. Let W, W_0, W_1 be as in (5.1). We shall construct a new shift (Y_1, μ) on symbols $\mathcal{K}_1 = \{b_1 \cdots b_s, b_1^{-1} \cdots b_s^{-1}\}$ together with a map $q: \mathcal{K}_1 \to W_0 \cup \{1\}$ with $q(c^{-1}) = q(c)^{-1}$, hence inducing an isomorphism $q: F_{\mathcal{K}_1} \to F_1/\operatorname{Ker}\pi$ (with the notation of (5.1)), and a τ-invariant Gibbs measure μ_1 on Y_1 such that $(Y_1, \sim_{G_i}, \mu_1)$ is ergodic if and only if $(Y, \sim_{F_i}, \mu)$ is, $i \geq 2$, where $G_i = q^{-1}(F_i/\operatorname{Ker}\pi)$. This is the inductive step.

Let W'_1 be the group generated by W_1. Define $w: Y \to (W_0 \cup W'_1)^{\mathbb{Z}}$ (actually only defined almost everywhere with respect to μ) as follows. For almost every $\mathbf{x} \in Y$,

$\mathbf{x}=\{x_i\}$, there is a unique way of inserting words of the form $c_1\cdots c_nc_n^{-1}\cdots c_1^{-1}$ between x_i and x_{i+1} for $i\neq -1$, so that $c_i\in\mathcal{K}$, $c_1\cdots c_n$ is a proper endpart of a word in W_0, and the augmented sequence from $\mathbf{x}$ can be decomposed into words y_i, with $y_i\in W_0\cup W_1'$ for all i, not both y_i, y_{i+1} in W_1' for any i, and x_0 part of the word $y_0=z_{-t}\cdots z_{-1}x_0\cdots z_u$, say, so that both $z_{-t}\cdots z_{-1}$ and $x_0\cdots z_u$ decompose into words of W. (Here, some of the z_i are the added symbols.) We are using here the fact that $(Y,\sim_{F_1},\mu)$ is ergodic.

Define $w_i(\mathbf{x})=y_i$. Let G denote the set of endparts of words in W_0. Define $p:Y\to((W_0\cup\{1\})\times\mathcal{K}^2\times G^2)^{\mathbf{Z}}$ by

$$p(\mathbf{x})=\{p_i(\mathbf{x})\}=\{(p_{i0}(\mathbf{x}),c_i(\mathbf{x}),d_i(\mathbf{x}),w_{i1}(\mathbf{x}),w_{i3}(\mathbf{x}))\},$$

where $p_{i0}(\mathbf{x})=w_i(\mathbf{x})$ if $w_i(\mathbf{x})\in W_0$, $=1$ if $w_i(\mathbf{x})\in W_1'$. $w_{i1}(\mathbf{x})$, $w_{i3}(\mathbf{x})$ are defined by writing $w_i(\mathbf{x})=w_{i1}(\mathbf{x})w_{i2}(\mathbf{x})w_{i3}(\mathbf{x})$, where $w_{i2}(\mathbf{x})$ is the piece of word from the original sequence $\mathbf{x}$, and $w_{i1}(\mathbf{x})$, $w_{i3}(\mathbf{x})$ are the inserted pieces. $c_i(\mathbf{x})$, $d_i(\mathbf{x})$ are the first and last elements respectively of the word $w_{i2}(\mathbf{x})$.

Let $\mathcal{K}_1$ be the set of symbols from the sequences of $p(Y)$, and $q:\mathcal{K}_1\to W_0\cup\{1\}$ projection onto the first coordinate. $p(Y)$ itself is not shift-invariant, but if Y_1 is the shift-invariant set generated by $p(Y)$, (Y_1,σ) is a subshift of finite type on $\mathcal{K}_1$, which is finite. $\tau:\mathcal{K}_1\to\mathcal{K}_1$ is defined by $\tau(w,c,d,r,s)=(w^{-1},d^{-1},c^{-1},s^{-1},r^{-1})$. There is a unique σ- and τ-invariant measure μ_1 on Y_1 with $\mu_1(A)=\mu(p^{-1}A)$ whenever $A\subseteq pY$. It can be checked that μ_1 is Gibbs.

$G_r=q^{-1}(F_r/\mathrm{Ker}\,\pi)$ is now subabelian finitely determined of degree $r-1$ in F_1, with chain $G_2\cdots G_r$. We claim that $(Y_1,\sim_{G_i},\mu_1)$ is ergodic if and only if $(Y,\sim_{F_i},\mu)$ is ergodic. Suppose $(Y_1,\sim_{G_i},\mu_1)$ is ergodic. Since words in W_0 have length at most n, say, this means that for almost all $\mathbf{x}=\{x_i\}$, the product $x_0\cdots x_p$ is in F_iz_p, for some word z_p in the symbols of $\mathcal{K}$ of length at most n, for infinitely many p. It follows from the properties of Gibbs measures that the product $x_0\cdots x_p\in F_i$ infinitely often. Hence $(Y,\sim_{F_i},\mu)$ is ergodic by lemma 2.2. The converse is immediate, once the notation is understood. □

I am greatly indebted to Jon Aaronson (see [**1**]) who first showed me how to prove the main theorem – in § 4 – for Markov measures with Schottky groups. I should like to thank him, and T. Lyons, H. McKean and D. Sullivan, for fruitful discussion concerning these related problems.

I should also like to thank St Hilda's College, Oxford for a Junior Research Fellowship during the time of this work.

REFERENCES

[1] J. Aaronson & M. Keane. Deterministic random walks and returns to zero. *J. London Math. Soc.* (in the press).
[2] R. Bowen. Symbolic dynamics for hyperbolic flows. *Amer. J. Math.* **95** (1973), 429–460.
[3] R. Bowen. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms.* Springer Lecture Notes in Math. no. 470. Springer: Berlin, 1975.
[4] R. Bowen. Hausdorff dimension of quasi-circles. *I.H.E.S. Publ. Math.* **50** (1979), 11–25.
[5] W. H. Gottschalk & G. A. Hedland. *Topological Dynamics.* Amer. Math. Soc. Coll. Publ. no. 36 (1955).

[**6**] T. Lyons & H. McKean. Winding of the plane Brownian motion. *Advances in Math.* (in the press).
[**7**] M. Morse. *Symbolic Dynamics* (notes by R. Oldenburger). I.A.S.: Princeton, 1966.
[**8**] S. J. Patterson. The limit set of a Fuchsian group. *Acta. Math.* **136** (1976), 241–273.
[**9**] C. Series. Symbolic dynamics for geodesic flows. *Acta. Math.* (in the press).
[**10**] D. Sullivan. The density at infinity of a discrete group of hyperbolic motions. *I.H.E.S. Publ. Math.* **50** (1979), 171–202.