Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Topology 46 (2007) 469–491
TOPOLOGY
www.elsevier.com/locate/top

Homology at infinity; fractal geometry of limiting symbols for modular subgroups

M. Kesseböhmer[a,*], B.O. Stratmann[b]

[a] *Fachbereich 3 - Mathematik und Informatik, Universität Bremen, D-28359 Bremen, Germany*
[b] *Mathematical Institute, University of St Andrews, St Andrews KY16 9SS, Scotland, United Kingdom*

Received 2 November 2006; received in revised form 2 March 2007; accepted 2 March 2007

Abstract

In this paper we use fractal geometry to investigate boundary aspects of the first homology group for finite coverings of the modular surface. We obtain a complete description of algebraically invisible parts of this homology group. More precisely, we first show that for any modular subgroup the geodesic forward dynamic on the associated surface admits a canonical symbolic representation by a finitely irreducible shift space. We then use this representation to derive a complete multifractal description of the higher-dimensional level sets arising from the Manin–Marcolli limiting modular symbols.

MSC: 11F06; 20H05; 37F35

Keywords: Limiting modular symbols; Modular subgroups; Non-commutative tori; Thermodynamical formalism; Multifractal formalism; Lyapunov spectra

1. Introduction

Let $\mathcal{C}_2(G)$ refer to the space of cusp forms of weight 2 for some arbitrary modular subgroup G. That is, G is a finite index subgroup of the modular group $\Gamma := \mathrm{PSL}_2(\mathbb{Z})$. It is well known that there is a dual pairing between $\mathcal{C}_2(G)$ and the first homology group $H_1(M_G, \mathbb{R})$ of the associated compactified cusped Riemann surface M_G of genus $\mathfrak{g}$. That is, we have

$$\langle \cdot, \cdot \rangle : H_1(M_G, \mathbb{R}) \times \mathcal{C}_2(G) \to \mathbb{C}, \quad \text{where } \langle \gamma, f \rangle := \int_\gamma f(z)\, \mathrm{d}z.$$

* Corresponding author.
E-mail addresses: mhk@math.uni-bremen.de (M. Kesseböhmer), bos@maths.st-and.ac.uk (B.O. Stratmann).

doi:10.1016/j.top.2007.03.004

Each element of $H_1(M_G, \mathbb{R})$ can be represented by integrating the 1-form $f\,\mathrm{d}z$ along some smooth path between two points ξ, η in $\mathbb{H} \cup P^1(\mathbb{Q})$, and this determines the modular symbol $\{\xi, \eta\}_G \in H_1(M_G, \mathbb{R})$. A possible way to extend these symbols to the non-cuspidal boundary of hyperbolic space, and therefore to give a non-trivial homological meaning to algebraically invisible parts of $H_1(M_G, \mathbb{R})$, has been suggested by Manin and Marcolli [18]. They introduced the concept 'limiting modular symbol', which is given for $x \in \mathbb{R}$ by (whenever the limit exists)

$$\ell_G(x) := \lim_{t\to\infty} \frac{1}{t}\{i, x + i\exp(-t)\}_G \in H_1(M_G, \mathbb{R}).$$

Note that the limit in the definition of ℓ_G exists if and only if it exists for each 1-form $f\,\mathrm{d}z$ with $f \in \mathcal{C}_2(G)$, and hence it is sufficient to compute it for a fixed complex basis $\widehat{f_1}, \ldots, \widehat{f_\mathfrak{g}}$ of $\mathcal{C}_2(G)$.

The aim of this paper is to use fractal geometry in order to investigate the level sets which arise naturally from these limiting modular symbols. That is, for $\alpha \in \mathbb{R}^{2\mathfrak{g}}$ we consider

$$\mathcal{F}_\alpha := \left\{x \in \mathbb{R} : \left(\langle \ell_G(x), f_1\rangle, \ldots, \langle \ell_G(x), f_{2\mathfrak{g}}\rangle\right) = \alpha\right\},$$

where $f_{2k-1} := \mathfrak{Re}(\widehat{f_k})$ and $f_{2k} := \mathfrak{Im}(\widehat{f_k})$, for $k = 1, \ldots, \mathfrak{g}$.

A first analysis of this type of level sets was given in [18,20] for modular subgroups which satisfy the there so called 'Red-condition' (see [18]). There it was shown that for these groups $\frac{1}{t}\{i, x + i\exp(-t)\}_G$ converges weakly to zero with respect to the Lebesgue measure on the unit interval. Subsequently, this result was improved in [20] by showing that $\ell_G(x)$ is equal to zero Lebesgue—almost everywhere. Besides, these papers obtained "non-vanishing " of limiting modular symbols *only* for the end points of closed geodesics, that is for quadratic surds. In these trivial cases the limiting modular symbol turns out to be given by integrating along the closed geodesic and then normalising by the hyperbolic length of that geodesic.

The aim of this paper is to extend these results to arbitrary modular subgroups and to obtain that the limiting modular symbol is not equal to zero for a large class of perfect sets of positive Hausdorff dimension. For this we will give a detailed construction of a shift space Σ_G, referred to as the modular shift space, which is canonically associated with G. We then show that Σ_G is finitely irreducible (see Proposition 3.1 for the definition), which in particular implies that every modular subgroup satisfies the Red-condition, and hence shows that the results of [18,20] do in fact hold for arbitrary modular subgroups.

Our main results are summarised in the following theorem, where $\widehat{\beta_G} : \mathbb{R}^{2\mathfrak{g}} \to \mathbb{R}$ refers to the proper concave (negative) Legendre transform of the proper convex function $\beta_G : \mathbb{R}^{2\mathfrak{g}} \to \mathbb{R}$, given by $\widehat{\beta_G}(\alpha) := \inf_{t\in\mathbb{R}^{2\mathfrak{g}}}(\beta_G(t) - (\alpha|t))$.

Main Theorem. *For an arbitrary modular subgroup G we have that the modular shift space Σ_G is finitely irreducible.*

Moreover, for $\mathfrak{g} \geq 1$ there exists a strictly convex and differentiable function $\beta_G : \mathbb{R}^{2\mathfrak{g}} \to \mathbb{R}$ such that for each $\alpha \in \nabla\beta_G(\mathbb{R}^{2\mathfrak{g}}) \subset \mathbb{R}^{2\mathfrak{g}}$,

$$\dim_H(\mathcal{F}_\alpha) = \widehat{\beta_G}(\alpha). \tag{1.1}$$

In here, we have that $\beta_G(0) = 1$, and that β_G has a unique minimum at 0. Also, we in particular have

$$\ell_G(\mathcal{F}_\alpha) = \{h_\alpha\},$$

where $h_\alpha \in H_1(M_G, \mathbb{R})$ *is uniquely determined by* $(\langle h_\alpha, f_1\rangle, \ldots, \langle h_\alpha, f_{2\mathfrak{g}}\rangle) = \alpha$. *Furthermore, the description of the spectrum in* (1.1) *is complete in the sense that*

$$\overline{\nabla\beta_G\left(\mathbb{R}^{2\mathfrak{g}}\right)} = \left\{\alpha \in \mathbb{R}^{2\mathfrak{g}} : \mathcal{F}(\alpha) \neq \emptyset\right\}.$$

The paper is organised as follows. In Section 2 we recall some basic facts on modular symbols. This includes a brief histogram on exact dual pairings. In Section 3 we first give a reminder on a beautiful construction which allows to visualise real numbers from a modular surface perspective. Subsequently, we show how this can be generalised to arbitrary modular subgroups. Here, the main result will be to use ergodicity of the geodesic flow to obtain that the so obtained generalised modular shift spaces are always necessarily finitely irreducible. In Section 4 we introduce a limiting modular symbol naturally arising from these generalised modular shift spaces, and show how these relate to the underlying branched geometry of numbers. In Section 5 we will collect facts from the thermodynamic formalism which turn out to be crucial in the proof of our main theorem, which will then be given in Section 6.

Remark. 1. It is well known that the modular subgroup quotient $\mathbb{H}/G$ can be viewed as a complex algebraic curve permitting an arithmetic structure. Similar to the familiar picture in which $\mathbb{H}/\Gamma$ represents the moduli space of elliptic curves, $\mathbb{H}/G$ represents the moduli space $\mathcal{M}(G)$ of elliptic curves $\mathcal{E}(G)$ equipped with some finite additional structure determined by $G \setminus \Gamma$. Hence, by adding the cusp points we obtain a pre-compactification of $\mathcal{M}(G)$, given by including all possible ways in which $\mathcal{E}(G)$ degenerates to $\mathbb{C} \setminus \{0\}$. Moreover, if we further include all degenerations of $\mathcal{E}(G)$ to non-commutative tori, we derive the compactified moduli space $\mathcal{M}_0(G)$ whose boundary is given by the non-commutative space $P^1(\mathbb{R})/G$. Since the level sets $\mathcal{F}_\alpha$ are clearly G-invariant, transferring the results in this paper to the language of elliptic curve degenerations yields that the modular multifractal spectrum corresponds to a continuous family of elements of the boundary of $\mathcal{M}_0(G)$ consisting of the 'bad quotients' $\mathcal{F}_\alpha/G$. (For the relation of non-commutative geometry and modular subgroups, and for some of the literature on this, we refer to the recent survey article [4] and to [21]).

2. Let us also mention that the concept 'limiting modular symbol' could easily be extended to more general concepts of 'modular symbols at infinity'. For instance, one could consider $\ell_{G,\phi,\psi}$ given by

$$\ell_{G,\phi,\psi}(x) := \operatorname*{Lim}_{t\to\infty} \phi(x,t)\{i, x + i\psi(x,t)\}_G,$$

for functions $\phi, \psi : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+$ such that $\phi(x,t)$ and $\psi(x,t)$ tend to zero for t tending to infinity. Here, Lim represents either lim, lim sup, or lim inf. However, in this paper we concentrate on the Manin–Marcolli limiting modular symbols for which $\phi(x,t) := 1/t$, $\psi(x,t) := \exp(-t)$ and Lim := lim.

2. Preliminaries for modular symbols

For a modular subgroup G consider the space $\mathcal{C}_k(G)$ of cusp forms $f : \mathbb{H} \to \mathbb{C}$ of weight $k \in \mathbb{Z}$, given by

- f is holomorphic on $\mathbb{H}$ as well as in each cusp of G;
- $f = (g')^{k/2} \cdot (f \circ g)$ for all $g \in G$;
- f vanishes at each cusp of G.

Throughout this paper, let M_G refer to the (possibly branched) covering surface $(\mathbb{H} \cup P^1(\mathbb{Q}))/G$ of M_Γ of genus $\mathfrak{g}$, assumed to be compactified by having added the cusps.

Recall that a p-chain is a formal sum $c_p = \sum_i k_i N_i$, where the N_i are p-dimensional smooth oriented submanifolds of M_G, and the coefficients k_i are elements of some Abelian group $\mathbb{K}$. The p-homology group $H_p(M_G, \mathbb{K})$ of M_G is then given via cycles and boundaries by $H_p(M_G, \mathbb{K}) := \{c_p : \partial c_p = \emptyset\}/\{\partial c_{p+1}\}$. Obviously, we have that $H_p(M_G, \mathbb{K}) = 0$, for each $p > 2$. In particular, $H_1(M_G, \mathbb{Z})$ is obtained geometrically by taking all loops in M_G as generators and then factoring out the relation that two loops are homologous, that is they only differ by some boundary. By triangulating M_G such that the directed edges of the triangulation represent the generators of the Abelian group $H_1(M_G, \mathbb{Z})$ (modulo the relations given by zero-homologous edge cycles), each element of $H_1(M_G, \mathbb{Z})$ can be written as a $\mathbb{Z}$-linear combination of the directed edges. Hence, $H_1(M_G, \mathbb{Z})$ is a finitely generated Abelian group which is always equal to either $\mathbb{Z}^{2\mathfrak{g}}$ or to a free product of $\mathbb{Z}^{2\mathfrak{g}}$ with some torsion subgroups of the form $\mathbb{Z}_2$ and/or $\mathbb{Z}_3$. Given that $H_1(M_G, \mathbb{Z})$ is known, one can then apply the universal coefficient theorem to determine $H_1(M_G, \mathbb{K})$. Indeed, we have that $H_1(M_G, \mathbb{K}) = H_1(M_G, \mathbb{Z}) \otimes_{\mathbb{Z}} \mathbb{K}$, and hence $H_1(M_G, \mathbb{K})$ is a free $\mathbb{K}$-module of dimension $2\mathfrak{g}$, provided that $\mathbb{K}$ is a field. In particular, this shows that the space $H_1(M_G, \mathbb{R})$ is a real vector space of dimension $2\mathfrak{g}$.

In order to see how $H_1(M_G, \mathbb{R})$ fits in with $\mathcal{C}_2(G)$, note that by de Rham theory we have that the 1-cohomology group $H^1(M_G, \mathbb{R})$ is isomorphic to the de Rham cohomology $H^1_{\mathrm{dR}}(M_G, \mathbb{R}) := \{\text{closed 1-forms}\}/\{\text{exact 1-forms}\}$, and hence defines a dual pairing (see e.g. the survey article [7])

$$\langle \cdot, \cdot \rangle : H_1(M_G, \mathbb{R}) \times H^1_{\mathrm{dR}}(M_G, \mathbb{R}) \to \mathbb{R}, \quad \text{given by } \langle \gamma, \omega \rangle := \int_\gamma \omega.$$

By Hodge decomposition, there exists an isometry between $H^1_{\mathrm{dR}}(M_G, \mathbb{R})$ and the space $H^1_\Delta(M_G)$ of harmonic 1-forms. By considering real and imaginary parts of the pull-backs of these harmonic forms to $\mathbb{H} \cup P^1(\mathbb{Q})$, we finally obtain an isomorphic representation of $H^1_\Delta(M_G)$ by the holomorphic cusp forms $\mathcal{C}_2(G)$. It follows that $\mathcal{C}_2(G)$ is a $\mathfrak{g}$-dimensional $\mathbb{C}$-vector space (and hence a $2\mathfrak{g}$-dimensional $\mathbb{R}$-vector space) (cf. also [28])). Hence, summarising the above, there is an exact dual pairing of homology and cusp forms, which is given by the $\mathbb{R}$-bi-linear map

$$\langle \cdot, \cdot \rangle : H_1(M_G, \mathbb{R}) \times \mathcal{C}_2(G) \to \mathbb{C}, \qquad \langle \gamma, f \rangle := \int_\gamma f(z)\, \mathrm{d}z.$$

Also, let us recall for later use that by de Rham theory we have for a fixed basis $(\gamma_k)_{k=1}^{2\mathfrak{g}}$, consisting of cycles in $H_1(M_G, \mathbb{R})$, that the set

$$\left\{ \left(\left\langle \gamma_k, \widehat{f_1} \right\rangle, \ldots, \left\langle \gamma_k, \widehat{f_{\mathfrak{g}}} \right\rangle \right) : k = 1, \ldots, 2\mathfrak{g} \right\}$$

is a basis for $\mathbb{C}^{\mathfrak{g}}$ over $\mathbb{R}$ (cf. [5, p. 10]). For the $2\mathfrak{g} \times 2\mathfrak{g}$ real period matrix $\Pi_G := (\langle \gamma_k, f_j \rangle)_{k,j}$, this means that

$$\Pi_G \in \mathrm{GL}_{2\mathfrak{g}}(\mathbb{R}). \tag{2.1}$$

Note that the above dual pairing shows that an element of $H_1(M_G, \mathbb{R})$ can be viewed as a path in M_G, or alternatively as a path in $\mathbb{H} \cup P^1(\mathbb{Q})$. Therefore, by viewing it as a path in $\mathbb{H} \cup P^1(\mathbb{Q})$ and noting that only the end points $\xi, \eta \in \mathbb{H} \cup P^1(\mathbb{Q})$ of the path matter, this allows to represent an elements of $H_1(M_G, \mathbb{R})$ by the so called modular symbol $\{\xi, \eta\}_G \in H_1(M_G, \mathbb{R})$. In particular, for each $g \in G$, any smooth path from $\xi \in \mathbb{H} \cup P^1(\mathbb{Q})$ to $g(\xi)$ projects to a closed path in M_G, and hence corresponds to a homology class in $H_1(M_G, \mathbb{Z})$. Clearly, this class is represented by the modular symbol $\{\xi, g(\xi)\}_G$, obtained by integrating 1-forms $f\, \mathrm{d}z$, for $f \in \mathcal{C}_2(G)$, along any smooth path from ξ to $g(\xi)$. One easily verifies that $\{\xi, g(\xi)\}_G$ does not depend on ξ, and that the assignment $g \mapsto \{\xi, g(\xi)\}_G$ gives rise to a surjection of G onto $H_1(M_G, \mathbb{Z})$ (the kernel of this group homomorphism is generated by the commutators of G). For the calculus with modular symbols, the following immediate identities are useful. For each $\xi, \eta, \zeta \in \mathbb{H} \cup P^1(\mathbb{Q})$ and $g \in G$, we have

$$\{\xi, \xi\}_G = 0, \qquad \{\xi, \eta\}_G = -\{\eta, \xi\}_G, \qquad \{\xi, \eta\}_G + \{\eta, \zeta\}_G = \{\xi, \zeta\}_G,$$
$$\{\xi, \eta\}_G = \{g(\xi), g(\eta)\}_G, \quad \text{and} \quad \{\xi, g(\xi)\}_G = \{\eta, g(\eta)\}_G.$$

So far, we only considered paths between points in $\mathbb{H} \cup P^1(\mathbb{Q})$ which lie in a single G-orbit of some element of $\mathbb{H} \cup P^1(\mathbb{Q})$. In general, that is for arbitrary modular subgroups G, it is not clear how to define modular symbols between elements of $P^1(\mathbb{Q})$ which are not in a single G-orbit. However, for congruence subgroups $\Gamma_0(N)$, defined by

$$\Gamma_0(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma \,\middle|\, c \equiv 0 \bmod N \right\} \quad \text{for } N \in \mathbb{N},$$

it is well known that this can be resolved, and this is the essence of the following theorem.

Theorem (*Manin–Drinfeld [17,6]*). *For each $\xi, \eta \in P^1(\mathbb{Q})$, we have*

$$\{\xi, \eta\}_{\Gamma_0(N)} \in H_1\left(M_{\Gamma_0(N)}, \mathbb{Q}\right).$$

Finally, let us recall a few useful facts about modular subgroups and in particular also congruence subgroups $\Gamma_0(N)$.

For instance, for the index $\kappa_N := [\Gamma : \Gamma_0(N)]$ of $\Gamma_0(N)$ in Γ, we have [26]

$$\kappa_N = N \prod_{p|N} \left(1 + \frac{1}{p}\right).$$

Also, for the number N_k of $\Gamma_0(N)$-inequivalent elliptic fixed points of order $k \in \mathbb{N}$ and the number N_∞ of $\Gamma_0(N)$-inequivalent parabolic fixed points, we have [28]

$$N_2 = \begin{cases} 0 & \text{if } 4|N \\ \prod_{p|N} (1 + [(-1) : p]) & \text{otherwise,} \end{cases} \qquad N_3 = \begin{cases} 0 & \text{if } 9|N \\ \prod_{p|N} (1 + [(-3) : p]) & \text{otherwise,} \end{cases}$$

and

$$N_\infty = \sum_{d|N, d>0} \phi\left(\text{g.c.d.}\,(d, N/d)\right),$$

where ϕ is the Euler function and [:] refers to the Legendre symbol of quadratic residues.

For arbitrary modular subgroups G, the following formula for the genus $\mathfrak{g}$ of M_G is an immediate consequence of the Riemann–Hurwitz formula. With R_k the number of G-inequivalent elliptic fixed points of order $k \in \mathbb{N}$, R_∞ the number of G-inequivalent parabolic fixed points, and κ the index of G in Γ, we have [28]

$$\mathfrak{g} = 1 + \frac{\kappa}{12} - \frac{R_2}{4} - \frac{R_3}{3} - \frac{R_\infty}{2}.$$

3. Modular shift spaces

One of the main ideas of this paper is to investigate limiting modular symbols by using a shift space which generalises the shift space for the continued fraction expansion of elements of $[-1, 1]$ (see e.g. [1,2,9,27,13]). In this section we give the construction of this shift space canonically associated with the geodesic dynamic on the Riemann surface arising from a modular subgroup. Note that this construction extends the usual coding procedure for the modular surface to arbitrary modular subgroups. For ease of notation, we put $\mathcal{I} := [-1, 1] \cap \mathbb{I}$, $\mathcal{I}_{-1} := [-1, 0] \cap \mathbb{I}$ and $\mathcal{I}_{+1} := [0, 1] \cap \mathbb{I}$, where $\mathbb{I}$ denotes the set of irrational numbers.

We begin with recalling from [27] the notion of 'type-change' for an oriented geodesic in the upper half-plane $\mathbb{H}$. For this note that $\mathbb{H}$ can be tiled by the so called Farey tessellation, that is the tessellation by Γ-translates of the ideal triangle with cusp-vertices at 0, 1 and $\{i\infty\}$. Consequently, each oriented geodesic l with irrational end points is covered by infinitely many tiles of this tessellation. By travelling on l in positive direction, each of these tiles gets intersected such that there is always a single vertex of the three cusp-vertices seen either on the left or on the right of the intersection of l with the tile (the other two vertices are seen on the opposite side). In case the single vertex is seen on the left, we say that the visit is of type L, otherwise it is called of type R. If in here two successive visits are of different type, then one says that l changes type at the point where it intersects the edge at which the two involved tiles intersect.

Now, let us consider the set $\widetilde{\mathcal{L}}_\Gamma$ of oriented geodesics l in $\mathbb{H}$ with initial point l_- and end point l_+, given by

$$\widetilde{\mathcal{L}}_\Gamma := \{l = (l_-, l_+) : 0 < |l_-| \le 1 \le |l_+|, l_- \cdot l_+ < 0, \text{ and } l_-, l_+ \in \mathbb{I}\}.$$

Each element l of $\widetilde{\mathcal{L}}_\Gamma$ can then be coded by its successive 'type-changes', that is

$$l \text{ is coded by } \begin{cases} \ldots\ldots L^{n_{-2}} R^{n_{-1}} y_l L^{n_1} R^{n_2} \ldots\ldots & \text{if } l_+ \ge 1 \\ \ldots\ldots R^{n_{-2}} L^{n_{-1}} y_l R^{n_1} L^{n_2} \ldots\ldots & \text{if } l_+ \le -1, \end{cases}$$

where y_l refers to the point at which l intersects the imaginary axis. This type of coding is closely related to the continued fraction expansion of elements $y \in \mathcal{I}_{+1}$, given for $y_1, y_2, \ldots \in \mathbb{N}$ by

$$y = [y_1, y_2, \ldots] := \cfrac{1}{y_1 + \frac{1}{y_2 + \cdots}}.$$

Namely, we have that

$$l_- = -[n_{-1}, n_{-2}, \ldots] \quad \text{and} \quad l_+ = [n_1, n_2, \ldots]^{-1} \quad \text{if } l_+ \ge 1$$
$$l_- = [n_{-1}, n_{-2}, \ldots] \quad \text{and} \quad l_+ = -[n_1, n_2, \ldots]^{-1} \quad \text{if } l_+ \le -1.$$

Next, consider the subset $\widetilde{\mathcal{C}}_\Gamma$ of the unit tangent space $UT(\mathbb{H})$ consisting of all those unit tangent vectors which are based at the imaginary axis and which give rise to geodesics $l \in \widetilde{\mathcal{L}}_\Gamma$. We then have that the Poincaré section $\widetilde{\mathcal{S}}_\Gamma$ associated with $\widetilde{\mathcal{L}}_\Gamma$ is given by the canonical projection of $\widetilde{\mathcal{C}}_\Gamma$ onto $UT(M_\Gamma)$. More precisely, let $l \in \widetilde{\mathcal{L}}_\Gamma$ be given such that l is coded by $\ldots L^{n_{-2}} R^{n_{-1}} y_l L^{n_1} R^{n_2} \ldots$. With $T : z \mapsto z+1$ referring to the parabolic generator of Γ, we have that $T^{-n_1}(l)$ is a geodesic which starts in $[-(n_1+1), -n_1]$ and ends in $\mathcal{I}_{+1}$, and hence $T^{-n_1}(l)$ is not an element of $\widetilde{\mathcal{L}}_\Gamma$. However, if we additionally apply the elliptic generator $S : z \mapsto -1/z$ of Γ, then we obtain that the resulting geodesic $l' := ST^{-n_1}(l)$ is an element of $\widetilde{\mathcal{L}}_\Gamma$, and one immediately verifies that

$$l'_- = [n_1, n_{-1}, \ldots] \quad \text{and} \quad l'_+ = -[n_2, n_3, \ldots]^{-1}.$$

Hence, in this situation we have

$$ST^{-n_1} : l = (-[n_{-1}, n_{-2}, \ldots], [n_1, n_2, \ldots]) \mapsto l' = ([n_1, n_{-1}, \ldots], -[n_2, n_3, \ldots]).$$

The dynamical idea behind this coding step is as follows. Let $v_l \in \widetilde{\mathcal{C}}_\Gamma$ be given, and let v'_l be the vector in $\Gamma(\widetilde{\mathcal{C}}_\Gamma)$ obtained by sliding v_l in positive direction along l until the next type-change takes place. The significance of ST^{-n_1} is that $ST^{-n_1}(v'_l)$ is an element of $\widetilde{\mathcal{C}}_\Gamma$ such that its projection onto $UT(M_\Gamma)$ is precisely the first return to $\widetilde{\mathcal{S}}_\Gamma$ when starting from the projection of y_l onto M_Γ in the direction of v_l.

Clearly, we can proceed in a similar way if the geodesic l turns out to be coded by $\ldots R^{n_{-2}} L^{n_{-1}} y_l R^{n_1} L^{n_2} \ldots$. In this case ST^{n_1} gives rise to the assignment

$$l = ([n_{-1}, n_{-2}, \ldots], -[n_1, n_2, \ldots]^{-1}) \mapsto l' = (-[n_1, n_{-1}, \ldots], [n_2, n_3, \ldots]^{-1}).$$

This procedure is summarised by the Poincaré-map $\widetilde{\mathcal{P}}_\Gamma : \widetilde{\mathcal{L}}_\Gamma \to \widetilde{\mathcal{L}}_\Gamma$, given by

$$\widetilde{\mathcal{P}}_\Gamma(l) := \begin{cases} ST^{-n_1}(l) & \text{if } l = (-[n_{-1}, n_{-2}, \ldots], [n_1, n_2, \ldots]^{-1}) \\ ST^{n_1}(l) & \text{if } l = ([n_{-1}, n_{-2}, \ldots], -[n_1, n_2, \ldots]^{-1}). \end{cases}$$

Here it is important to remark that the restriction $\mathcal{P}_\Gamma$ of the action of $\widetilde{\mathcal{P}}_\Gamma$ to the second coordinate can also be described by the 'twisted Gauss-map'

$$\mathcal{G}_\Gamma : \mathcal{I} \to \mathcal{I}, \qquad x \mapsto S\mathcal{P}_\Gamma S(x).$$

The reason why $\mathcal{G}_\Gamma$ is called twisted Gauss-map originates from its link to the usual Gauss-map $\mathcal{G} : \mathcal{I}_{+1} \to \mathcal{I}_{+1}, x \mapsto 1/x - \lfloor 1/x \rfloor$ (where $\lfloor 1/x \rfloor$ denotes the integer part of $1/x$). Namely, one immediately verifies

$$\mathcal{G}_\Gamma (:= S\mathcal{P}_\Gamma S) : x \mapsto -\mathrm{sign}(x)\mathcal{G}(|x|).$$

Note that $\mathcal{G}_\Gamma$ can alternatively be derived directly if instead of $\widetilde{\mathcal{L}}_\Gamma$ one starts with the following set of oriented $\mathbb{H}$-geodesics

$$\mathcal{L}_\Gamma := \{l = (l_-, l_+) : 0 < |l_+| \le 1 \le |l_-|, l_- \cdot l_+ < 0, \text{ and } l_-, l_+ \in \mathbb{I}\}.$$

These geodesics can then also be coded by the type-change mechanism as explained above. Here, the relevant section $\mathcal{C}_\Gamma \subset UT(\mathbb{H})$ is the set of unit tangent vectors based at the imaginary axis, giving rise to the geodesics in $\mathcal{L}_\Gamma$. The Poincaré section arising from this will be denoted by $\mathcal{S}_\Gamma$. One then immediately verifies that $\mathcal{G}_\Gamma$ coincides with the map obtained by restricting the Poincaré map arising from this alternative approach to the second coordinate.

Also, note that for $k \in \mathbb{Z}^{\times} := \mathbb{Z} \setminus \{0\}$ the k-th inverse branch $\mathcal{P}_{\Gamma,k}^{-1}$ of $\mathcal{P}_{\Gamma}$ has the property

$$\mathcal{P}_{\Gamma,k}^{-1} : \begin{cases} (-\infty, -1] \cap \mathbb{I} \to \{y^{-1} \in [1, \infty) : y = [k, y_2, \ldots]\} & \text{if } k \in \mathbb{N} \\ [1, \infty) \cap \mathbb{I} \to \{-y^{-1} \in (-\infty, -1] : y = [|k|, y_2, \ldots]\} & \text{if } k \notin \mathbb{N}. \end{cases}$$

In particular, this shows that $\mathcal{P}_{\Gamma,k}^{-1}$ can be expressed in terms of the generators of Γ by $\mathcal{P}_{\Gamma,k}^{-1} = T^k S$. On the other hand, for the corresponding k-th inverse branch $\mathcal{G}_{\Gamma,k}^{-1}$ of the twisted Gauss-map we immediately have

$$\mathcal{G}_{\Gamma,k}^{-1} = S\mathcal{P}_{\Gamma,k}^{-1}S = ST^k SS = ST^k : \begin{cases} \mathcal{I}_{+1} \to I_{-k} & \text{if } k \in \mathbb{N} \\ \mathcal{I}_{-1} \to I_{-k} & \text{if } k \notin \mathbb{N}, \end{cases}$$

where $I_k := \{\operatorname{sign}(k)[y_1, y_2, \ldots] \in \mathcal{I} : y_1 = |k|\}$ refers to the k-th basic interval.

We can now use standard ergodic theory to obtain our actual code space via the inverse branches of $\mathcal{G}_{\Gamma}$ as follows (cf. [3,25]). One immediately verifies that the set $\alpha := \{I_k : k \in \mathbb{Z}^{\times}\}$ is a partition of $\mathcal{I}$, such that the sequence of refinements $(\bigvee_{i=0}^{n-1} \mathcal{G}_{\Gamma}^{-i}(\alpha))_{n\in\mathbb{N}}$ generates the Borel σ-algebra. Hence, in terms of inverse branches of $\mathcal{G}_{\Gamma}$ the twisted continued fraction coding of $\mathcal{I}$ is given as follows. For $n_1, n_2, \ldots \in \mathbb{N}$, we have

$$\begin{aligned} [n_1, n_2, \ldots] &= \lim_{k\to\infty} ST^{-n_1}ST^{n_2}\ldots ST^{(-1)^k n_k}\left(\mathcal{I}_{(-1)^k}\right) \\ -[n_1, n_2, \ldots] &= \lim_{k\to\infty} ST^{n_1}ST^{-n_2}\ldots ST^{(-1)^{k+1} n_k}\left(\mathcal{I}_{(-1)^{k+1}}\right). \end{aligned}$$

Therefore, by defining the shift space

$$\Sigma_* := \left\{(x_1, x_2, \ldots) \in \left(\mathbb{Z}^{\times}\right)^{\mathbb{N}} : x_i x_{i+1} < 0, \text{ for all } i \in \mathbb{N}\right\}$$

equipped with the shift map $\sigma_* : (x_1, x_2, \ldots) \mapsto (x_2, x_3, \ldots)$, one immediately verifies that

$$\rho : \Sigma_* \to \mathcal{I}, \qquad (x_1, x_2, \ldots) \mapsto -\operatorname{sign}(x_1)[|x_1|, |x_2|, \ldots]$$

is a bijection for which $\rho \circ \sigma_* = \mathcal{G}_{\Gamma} \circ \rho$.

Our next goal is to generalise this modular coding procedure to arbitrary modular subgroups G. For this, let E_G refer to a fixed set of left-coset representatives of the quotient space $G \setminus \Gamma$. In this more general setting the relevant set of oriented geodesics is given by $\mathcal{L}_G := \bigcup_{e\in E_G} e(\mathcal{L}_{\Gamma})$. Note that there is a 1-1-correspondence between $\mathcal{L}_G$ and the Poincaré section $\mathcal{S}_G$ for the geodesic flow on M_G, where $\mathcal{S}_G$ is given by the canonical projection of $\mathcal{C}_G := \bigcup_{e\in E_G} e(\mathcal{C}_{\Gamma})$ onto $UT(M_G)$. We then adopt the above modular coding procedure in order to obtain a code space also in this more general situation. For this we proceed as follows. Assume that $\overline{\Sigma}_G := \bigcup_{e\in E_G}(e(\mathcal{I}) \times \{e\})$ is equipped with the topology inherited from $\mathbb{R}$. The G-generalised twisted Gauss-map $\mathcal{G}_G : \overline{\Sigma}_G \to \overline{\Sigma}_G$ is then given by

$$\mathcal{G}_G(x, e) := \left(eS\mathcal{P}_{\Gamma}Se^{-1}(x), e\right), \text{ for } e \in E_G, x \in e(\mathcal{I}).$$

Analogous to the situation before, we now have that for $k \in \mathbb{Z}^{\times}$ and $e \in E_G$ the (k, e)-th inverse branch $\mathcal{G}_{G,(k,e)}^{-1}$ of $\mathcal{G}_G$ is given by

$$\mathcal{G}_{G,(k,e)}^{-1} : \begin{cases} e(\mathcal{I}_{+1}) \times \{e\} \to e(\mathcal{I}_{-1}) \times \{e\}, & (x, e) \mapsto \left(eST^k e^{-1}(x), e\right) \quad \text{if } k \in \mathbb{N} \\ e(\mathcal{I}_{-1}) \times \{e\} \to e(\mathcal{I}_{+1}) \times \{e\}, & (x, e) \mapsto \left(eST^k e^{-1}(x), e\right) \quad \text{if } k \notin \mathbb{N}. \end{cases}$$

Hence, we can again use standard ergodic theory to obtain our actual code space via the inverse branches of $\mathcal{G}_G$. This time the basic intervals are $I_{k,e} := e(I_k) \times \{e\}$, for $k \in \mathbb{Z}^{\times}$ and $e \in E_G$. Also, $\alpha_G := \{I_{k,e} : k \in \mathbb{Z}^{\times}, e \in E_G\}$ is a partition of $\overline{\Sigma}_G$ such that the sequence of refinements $(\bigvee_{i=0}^{n-1} \mathcal{G}_G^{-i}(\alpha))_{n \in \mathbb{N}}$ generates the Borel σ-algebra of $\overline{\Sigma}_G$. Hence, in terms of inverse branches of $\mathcal{G}_G$ the G-generalised twisted continued fraction coding of $\overline{\Sigma}_G$ is given as follows. Let $(x, e) \in \overline{\Sigma}_G$ be given. Then $x \in e(\mathcal{I})$, and we have that there exist $n_1, n_2, \ldots \in \mathbb{N}$ such that $e^{-1}(x) = \pm[n_1, n_2, \ldots]$. By the above modular coding, we then have

$$x = \begin{cases} \lim_{k \to \infty} eST^{-n_1} ST^{n_2} \ldots ST^{(-1)^k n_k} \left(\mathcal{I}_{(-1)^k}\right) & \text{if } e^{-1}(x) \in \mathcal{I}_{+1} \\ \lim_{k \to \infty} eST^{n_1} ST^{-n_2} \ldots ST^{(-1)^{k+1} n_k} \left(\mathcal{I}_{(-1)^k}\right) & \text{if } e^{-1}(x) \in \mathcal{I}_{-1}. \end{cases} \tag{3.1}$$

Clearly, the assignment $(e(\pm[n_1, n_2, \ldots]), e) \mapsto ((\mp n_1, \pm n_2, \ldots), e)$ gives rise to a bijection between $\overline{\Sigma}_G$ and $\widetilde{\Sigma}_G := \Sigma_* \times E_G$. Unfortunately, the space $\widetilde{\Sigma}_G$ is not a proper shift space. Nevertheless, such a proper shift space can be obtained by keeping track of the cosets $GeST^{\pm n_1} ST^{\mp n_2} \ldots ST^{\pm n_k}$ which are visited during the approximation of x given in (3.1). That is, we successively mark down as a second parameter the cosets in which those images of the directed imaginary axis lie on which the type-changes occur. More precisely, we define the shift space Σ_G by

$$\Sigma_G := \{((x_1, e_1), (x_2, e_2), \ldots) \in (\mathbb{Z}^{\times} \times E_G)^{\mathbb{N}} : (x_1, x_2, \ldots) \in \Sigma_*,\ e_{k+1} = \tau_{x_k}(e_k) \text{ for all } k \in \mathbb{N}\},$$

where the map $\tau_{x_k} : E_G \to E_G$ is defined by, with $\equiv_G$ referring to equivalence mod G,

$$\tau_{x_k}(e_k) : \equiv_G e_k ST^{x_k}.$$

One immediately verifies that the assignment

$$((x_1, x_2, \ldots), e) \mapsto \left((x_1, e), (x_2, \tau_{x_1}(e)), (x_3, \tau_{x_2}(\tau_{x_1}(e))), \ldots\right)$$

defines an isomorphism between $\widetilde{\Sigma}_G$ and Σ_G, and hence Σ_G is also isomorphic to $\overline{\Sigma}_G$. Of course, the advantage in using Σ_G to code the geodesic rays in M_G which arise from $\mathcal{S}_G$ is, that it becomes a proper shift space when equipped with the shift map

$$\sigma : \Sigma_G \to \Sigma_G, \qquad ((x_1, e_1), (x_2, e_2), \ldots) \mapsto ((x_2, e_2), (x_3, e_3), \ldots),$$

as well as with the canonical metric d, given for $((x_k, e_k))_k, ((x_k', e_k'))_k \in \Sigma_G$ by

$$d(((x_k, e_k))_k, ((x_k', e_k'))_k) := \sum_{i=1}^{\infty} 2^{-i} \left(1 - \delta_{(x_i, e_i), (x_i', e_i')}\right).$$

Note that the system (Σ_G, σ) relates to ordinary continued fraction expansions as follows. For $((x_k, e_k))_k \in \Sigma_G$ one immediately verifies by way of finite induction, using the matrix representation of the elements in Γ,

$$\begin{aligned} e_{k+1} = \;& \tau_{x_k}(e_k) \equiv_G e_k S T^{x_k} = \tau_{x_{k-1}}(e_{k-1}) S T^{x_k} \\ \equiv_G \;& e_{k-1} S T^{x_{k-1}} S T^{x_k} \equiv_G \cdots \\ \equiv_G \;& e_1 S T^{x_1} \dots S T^{x_k} = e_1 \begin{pmatrix} -\operatorname{sign}(x_1)\, p_{k-1}(x) & (-1)^k\, p_k(x) \\ q_{k-1}(x) & (-1)^{k+1} \operatorname{sign}(x_1)\, q_k(x) \end{pmatrix}. \end{aligned}$$

Here, $p_n(x)/q_n(x)$ refers to the n-th approximant of the ordinary continued fraction expansion of $x := [|x_1|, |x_2|, \dots]$, with the usual convention $q_0(x) := p_{-1}(x) := 1$ and $q_{-1}(x) := p_0(x) := 0$.

Remark 1. 1. At first sight it might appear that the step from $\widetilde{\Sigma}_G$ and/or $\overline{\Sigma}_G$ to Σ_G is just technical and that it achieves only little. However, this step will turn out to be crucial, since it will allow us to employ certain standard results from thermodynamic formalism, a formalism which is well elaborated for shift spaces of the type (Σ_G, σ).

2. We remark that (Σ_G, σ) can also be represented by the skew product $(\widetilde{\Sigma}_G, \widetilde{\sigma})$, where $\widetilde{\sigma} : \widetilde{\Sigma}_G \to \widetilde{\Sigma}_G$ is given by

$$\widetilde{\sigma} : ((x_1, x_2, \dots), e) \mapsto \big((x_2, x_3, \dots), \tau_{x_1}(e)\big).$$

Clearly, the assignment $\widetilde{\pi}((x_1, e_1), (x_2, e_2), \dots) := ((x_1, x_2, \dots), e_1)$ gives rise to an isomorphism $\widetilde{\pi} : \Sigma_G \to \widetilde{\Sigma}_G$, which is a dynamical conjugacy in the sense that $\widetilde{\sigma} \circ \widetilde{\pi} = \widetilde{\pi} \circ \sigma$.

Also, note that throughout we will often identify an element $((x_k, e_k))_k \in \Sigma_G$ with $((x_k)_k, e_1) \in \widetilde{\Sigma}_G$, as well as with $(e_1(-\operatorname{sign}(x_1)[|x_1|, |x_2|, \dots]), e_1) \in \overline{\Sigma}_G$.

3. Let us also mention that (Σ_G, σ) can be represented by a conformal graph directed Markov system (for an extensive discussion of these systems, we refer to [22]). Namely, define $\mathcal{V} := \{(e, \pm 1) : e \in E_G\}$ to be the finite set of vertices, $\mathcal{E} := \mathbb{Z}^{\times} \times E_G$ the countable infinite set of edges, and let the two functions $i, t : \mathcal{E} \to \mathcal{V}$ be given by $i((k, e)) := (e, -\operatorname{sign}(k))$ and $t((k, e)) := (\tau_k(e), \operatorname{sign}(k))$. Furthermore, let the edge incidence matrix $A = (A_{u,v})_{u,v \in \mathcal{E}}$ be defined by $A_{u,v} = 1$ if $t(u) = i(v)$, and $A_{u,v} = 0$ otherwise. We then have that $(\mathcal{V}, \mathcal{E}, i, t, A)$ is a directed multigraph with associated incidence matrix A, and one immediately verifies that the subshift $\Lambda_G := \{(u_k)_k \in \mathcal{E}^{\mathbb{N}} : A_{u_k, u_{k+1}} = 1, \text{ for all } k \in \mathbb{N}\}$ is isomorphic to Σ_G. In order to derive the conformal graph directed Markov system, define compact sets $\mathcal{I}_{(e,+1)} := e(\mathcal{I}_{+1})$ and $\mathcal{I}_{(e,-1)} := e(\mathcal{I}_{-1})$, for all $e \in E_G$. Also, for each $(k, e) \in \mathcal{E}$ define

$$\phi_{(k,e)} := e S T^k \left(\tau_k(e)\right)^{-1} : \mathcal{I}_{(\tau_k(e), \operatorname{sign}(k))} \to \mathcal{I}_{(e, -\operatorname{sign}(k))}.$$

(Note that here the maps $\phi_{(k,e)}$ are viewed as Möbius transformations). With these preparations, one now immediately verifies that the system

$$\Phi_G := \left\{\phi_u : \mathcal{I}_{t(u)} \to \mathcal{I}_{i(u)} \mid u \in \mathcal{E}\right\}$$

satisfies all the requirements of a conformal graph directed Markov system, apart from that a priori the maps ϕ_u are not necessarily uniformly contracting. However, similar as for Schottky groups [22, Example 5.1.5], this can be resolved by replacing the system Φ_G by a sufficiently high iterate of itself.

The final aim of this section is to show that for each modular subgroup we have that the modular shift space Σ_G satisfies a certain transitivity condition called ‘finitely irreducible’ (for the definition, see Proposition 3.1 below). Let us remark that the main results in [19] are based on the assumption that the so called ‘Red-condition’ holds (see [19] for the definition). One immediately verifies that this

Red-condition is in fact equivalent to finite irreducibility of the shift space Σ_G. However, the approach in [19] allows to verify this condition only for congruence subgroups. Indeed, the proof of [19, Proposition 1.2.1] is based on the fact that for congruence subgroups $\Gamma_0(N)$ there is an isomorphism between the set of modular symbols and the set of M-symbols (that is $P^1(\mathbb{Z}/N\mathbb{Z})$, the projective line over the ring of integers mod N (see also [5])). This then allows to verify the Red-condition algebraically in terms of elementary congruence calculations. In contrast to this, our approach is completely different. To obtain the result for *all* modular subgroups G, we combine an elementary observation for the shift space Σ_G with the ergodicity of the geodesic flow on M_G.

In the following, Σ_G^n refers to the set of admissible words of length n in the alphabet $\mathbb{Z}^\times \times E_G$, and $\Sigma_G^* := \bigcup_{n\in\mathbb{N}} \Sigma_G^n$.

Proposition 3.1. *For each modular subgroup G we have that the modular shift space (Σ_G, σ) is finitely irreducible in the sense of* [22]. *That is, there exists a finite set $W \subset \Sigma_G^*$ such that for all $a, b \in \mathbb{Z}^\times \times E_G$ there exists $w \in W$ such that $awb \in \Sigma_G^*$.*

Proof. Let $(m, e'), (n, e'') \in \mathbb{Z}^\times \times E_G$ be given. For simplicity, let us only consider the case in which $m < 0$ and $n > 0$. Clearly, the remaining cases can be dealt with in an analogous way. Now, the aim is to show that there exists a universal constant $c = c(G) \in \mathbb{N}$ and a finite set $W \subset \Sigma_G^*$ of words of length at most c such that

$$(m, e')w(n, e'') \in \Sigma_G^*, \quad \text{for some } w \in W. \tag{3.2}$$

For this, observe that if $(m, e')w(n, e'') \in \Sigma_G^*$ then $(m, e')w(s, e'') \in \Sigma_G^*$, for *all* $s \in \mathbb{N}$. This shows that the assertion in (3.2) does not dependent on n. Similarly, observe that with $\widehat{e} := \tau_m(e')$ we have $(m, e')(r, \widehat{e}) \in \Sigma_G^2$, for *all* $r \in \mathbb{N}$. This shows that when starting from the second symbol $(r, \widehat{e})$ the assertion in (3.2) does not depend on the first entry m. Combining these two observations, it follows that in order to obtain (3.2) it is sufficient to show that there exists a finite set $W' \subset \Sigma_G^*$ of words of length at most $c - 1$ with the property that in the situation above we have that there exists $w' \in W'$ such that

$$(r, \widehat{e})w'(s, e'') \in \Sigma_G^*, \quad \text{for some } r, s \in \mathbb{N}. \tag{3.3}$$

For this, note that by construction we have that $(k, e) \in \mathbb{Z}^\times \times E_G$ represents the basic interval $I_{-k,e}$. Furthermore, in terms of cross sections we have that $I_{-k,e}$ represents a certain subset $\mathcal{C}_{k,e}$ of the cross section $e(\mathcal{C}_\Gamma) \subset \mathcal{C}_G$. That is, $\mathcal{C}_{k,e}$ is the set of those unit vectors v which are based at $e(\{z \in \mathbb{H} : \mathfrak{Re}(z) = 0\})$ such that the oriented $\mathbb{H}$-geodesic given by v terminates in $e(I_{-k})$ and starts in either $e([1, \infty))$ (if k is positive) or $e((-\infty, -1])$ (if k is negative). Define $\mathcal{C}_e^- := \bigcup_{k\in\mathbb{N}} \mathcal{C}_{k,e}$ (resp. $\mathcal{C}_e^+ := \bigcup_{k\in\mathbb{N}} \mathcal{C}_{-k,e}$), and let $\mathcal{S}_e^\pm \subset \mathcal{S}_G$ be the projection of $\mathcal{C}_e^\pm$ onto $UT(M_G)$. Expressing the assertion in (3.3) in these terms, it follows that we have to show that there exists $v \in \mathcal{S}_{\widehat{e}}^-$ and $v' \in \mathcal{S}_{e''}^-$ such that $v' = \phi_t(v)$ for some $t > 0$, where $(\phi_t)_{t\in\mathbb{R}}$ refers to the geodesic flow on $UT(M_G)$. In order to see this, note that we clearly have that for each $e \in E_G$ there exists $\mathcal{V}_e^- \subset \mathcal{C}_e^-$ such that for $\delta > 0$ the flow-box $V_e^-(\delta) := \{\phi_t(v) \in UT(M_G) : v \in \mathcal{V}_e^-, t \in (-\delta, \delta)\}$ is a measurable set of positive Liouville measure. Moreover, by choosing the δ sufficiently small as well as the $\mathcal{V}_e^\pm$ such that the base points of their vectors are well separated from the elliptic fixed points and are contained in some small compact interval, we can assume without loss of generality that the $V_e^\pm(\delta)$ are pairwise disjoint and that the restriction of the universal covering map to each of these boxes is a bijection. Now, recall that the geodesic flow on $UT(M_G)$ is ergodic with respect to the Liouville measure [10,11], and that this ergodicity is equivalent to the statement that for all measurable $U, V \subset UT(M_G)$ of positive Liouville measure there exists $t = t(U, V) > 0$ such that $\phi_t(U) \cap V \neq \emptyset$ (see e.g. [23, Theorem 7.2.11]). Using this and coming back to the assertion in (3.3), we now have that there exists $t_0 = t_0(V_{\widehat{e}}^-(\delta), V_{e''}^-(\delta)) > 0$ such that $\phi_{t_0}(V_{\widehat{e}}^-(\delta)) \cap V_{e''}^-(\delta) \neq \emptyset$. Hence, there exists $v \in \mathcal{S}_{\widehat{e}}^-$ and $\epsilon \in (-2\delta, 2\delta)$ such that $\phi_{t_0+\epsilon}(v) \in \mathcal{S}_{e''}^-$. By putting $v' := \phi_{t_0+\epsilon}(v)$, the assertion in (3.3) follows. This finishes the proof. □

Remark 2. Let us remark that in the proof of Proposition 3.1 the constant $c(G)$ is obtained as follows. For each of the transitions between two different boxes from the set of flow-boxes $\{V_e^{\pm}(\delta) : e \in E_G\}$ determine the number of crossings of the Poincaré section $\mathcal{S}_G$ during the transition. The constant $c(G)$ is then obtained by adding 1 to the maximum of these numbers (note that this maximum exists, since there are only finitely many flow-boxes; also, here the adding of 1 takes the transition from (m, e) to $(r, \widehat{e})$ into account). Clearly, our approach only allows to deduce that $c(G)$ is finite, but it does not permit to determine the actual value of $c(G)$ for a particular modular subgroup. However, in [19, Proposition 1.2.1] it was shown by arithmetic means that for congruence subgroups $\varGamma_0(N)$ we in fact have that $c(\varGamma_0(N)) = 3$.

4. The limiting modular symbol for Σ_G

We already introduced the limiting modular symbol ℓ_G in the introduction, which was there defined on $\mathbb{R}$. We now define a slightly different version of such a symbol, namely the limiting modular symbol $\widetilde{\ell}_G$ defined on Σ_G.

Definition 4.1. The limiting modular symbol $\widetilde{\ell}_G : \Sigma_G \to H_1(M_G, \mathbb{R})$ is defined for arbitrary $((x_k, e_k))_k \in \Sigma_G$ by (whenever the limit exists as an element of $H_1(M_G, \mathbb{R})$)

$$\widetilde{\ell}_G\left(((x_k, e_k))_k\right) := \lim_{t\to\infty} \frac{1}{t}\{i, e_1(x + i\exp(-t))\}_G.$$

Here, we have set $x := -\mathrm{sign}(x_1)[|x_1|, |x_2|, \ldots] \in \mathcal{I}$.

Note that $\ell_G(x)$ does not depend on the starting point i of the paths along which one integrates, nor does it depend on the choice of the geodesic $\{x + i\exp(-t) : t \in \mathbb{R}\}$ (in fact, any path having x as its only accumulation point in $P^1(\mathbb{R})$ would do). Also, concerning the existence of the limit in the definition of $\widetilde{\ell}_G$ the following holds, which is of course similar to what has been remarked on ℓ_G in the introduction. Namely, since $\langle\cdot,\cdot\rangle$ is a perfect dual pairing, the existence of this limit is guaranteed if the limit

$$\lim_{t\to\infty}\left\langle \frac{1}{t}\{i, e_1(x + i\exp(-t))\}_G, f\right\rangle$$

exists either for each $f \in \mathcal{C}_2(G)$, or equivalently, for each member of a basis of $\mathcal{C}_2(G)$.

The following proposition gives the main result of this section.

Proposition 4.2. *For $((x_k, e_k))_k \in \Sigma_G$ we have*

$$\widetilde{\ell}_G\left(((x_k, e_k))_k\right) = \lim_{n\to\infty} \frac{1}{2\log q_n(|x|)} \sum_{k=1}^{n} \{e_k(i\infty), e_k(0)\}_G.$$

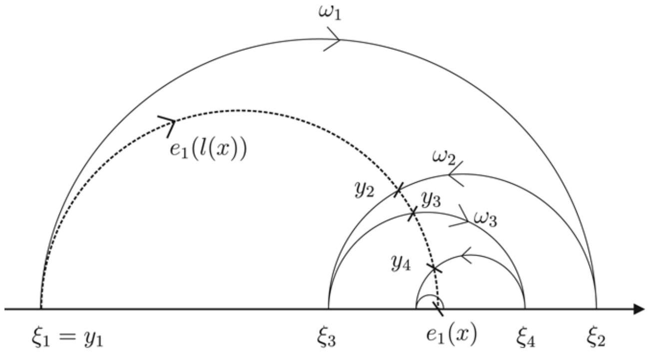

Fig. 1. Approximating path for a point $(x, e_1) \in \Sigma_G$.

Proof. Let $((x_k, e_k))_k \in \Sigma_G$ be given. Our first aim is to show that

$$\widetilde{\ell}_{G,q}\left(((x_k, e_k))_k\right) := \lim_{n\to\infty} \frac{1}{2\log q_n(|x|)} \sum_{k=1}^{n} \{e_k(i\infty), e_k(0)\}_G$$

exists if and only if $\lim_{n\to\infty} t_n^{-1}\{i, e_1(x + i\exp(-t_n))\}_G$ exists, for some sequence $(t_n)_{n\in\mathbb{N}}$ tending to infinity. More precisely, we will show that if one of these limits exists then both limits coincide, that is

$$\widetilde{\ell}_{G,q}\left(((x_k, e_k))_k\right) = \lim_{n\to\infty} \frac{1}{t_n}\{i, e_1(x + i\exp(-t_n))\}_G. \tag{4.1}$$

For this we proceed similar to [19, Proof of Theorem 0.2.1] as follows. Let $l(x)$ refer to the oriented hyperbolic geodesic from $i\infty$ to x, and define $\xi_1 = e_1(i\infty)$ and $\xi_n := e_1(-\mathrm{sign}(x_1)p_{n-2}(|x|)/q_{n-2}(|x|))$, for $n \geq 2$. Then consider the path $\omega := \omega_1\omega_2\ldots$ which runs in succession through the oriented hyperbolic geodesics ω_n which start at ξ_n and end in ξ_{n+1} (cf. Fig. 1). Clearly, when viewing ω as a path in $\mathbb{H} \cup P^1(\mathbb{Q})$ it is a connected oriented path which approximates $e_1(x)$ in its forward direction. Next, define $y_n := \omega_n \cap e_1(l(x))$ for $n \in \mathbb{N}$, and observe that the oriented geodesic path from y_n to y_{n+1} is homologous to the geodesic path which runs from y_n via ξ_{n+1} to y_{n+1}. It follows

$$\{y_n, y_{n+1}\}_G = \{y_n, \xi_{n+1}\}_G + \{\xi_{n+1}, y_{n+1}\}_G, \quad \text{for all } n \in \mathbb{N}.$$

Before we continue with this argument, first observe that we have for all $n \in \mathbb{N}$,

$$\{\xi_n, \xi_{n+1}\}_G = \{e_n(i\infty), e_n(0)\}_G \quad \text{and} \quad e_n(0) = e_{n+1}(i\infty).$$

Indeed, this can be seen as follows. Define $g_1 :=$ id., and for $n \in \mathbb{N}$ let

$$g_{n+1} := ST^{x_1}\ldots ST^{x_n} = \begin{pmatrix} -\mathrm{sign}(x_1)\, p_{n-1}(|x|) & (-1)^n\, p_n(|x|) \\ q_{n-1}(|x|) & (-1)^{n+1}\,\mathrm{sign}(x_1)\, q_n(|x|) \end{pmatrix}.$$

We then have that $e_1 g_n(i\infty) = \xi_n$ and $e_1 g_n(0) = \xi_{n+1}$. Also, since $e_n \equiv_G e_1 g_n$, there exists $\widetilde{g}_n \in G$ such that $\widetilde{g}_n e_n = e_1 g_n$. Using these facts as well as the G-invariance of the modular symbol, we obtain

$$\begin{aligned}\{e_n(i\infty), e_n(0)\}_G &= \{\widetilde{g}_n e_n(i\infty), \widetilde{g}_n e_n(0)\}_G = \{e_1 g_n(i\infty), e_1 g_n(0)\}_G \\ &= \{\xi_n, \xi_{n+1}\}_G.\end{aligned}$$

Using this observation, we proceed with the above argument as follows. For each $n \in \mathbb{N}$, we have

$$\begin{aligned}\{i, y_{n+1}\}_G &= \{i, y_2\}_G + \{y_2, y_{n+1}\}_G = \{i, y_2\}_G + \sum_{k=2}^{n} \{y_k, y_{k+1}\}_G \\ &= \{i, y_2\}_G + \sum_{k=2}^{n} \left(\{y_k, \xi_{k+1}\}_G + \{\xi_{k+1}, y_{k+1}\}_G\right) \\ &= \{i, y_2\}_G - \{\xi_2, y_2\}_G - \{y_{n+1}, \xi_{n+2}\}_G + \sum_{k=2}^{n+1} \{\xi_k, \xi_{k+1}\}_G \\ &= \{i, \xi_1\}_G - \{y_{n+1}, \xi_{n+2}\}_G + \sum_{k=1}^{n+1} \{\xi_k, \xi_{k+1}\}_G \\ &= \{i, \xi_1\}_G - \{y_{n+1}, \xi_{n+2}\}_G + \sum_{k=1}^{n+1} \{e_k(i\infty), e_k(0)\}_G .\end{aligned}$$

Now, let t_n be defined implicitly by $e_1(x + i\exp(-t_n)) := y_n$. Using elementary hyperbolic geometry in the context of for instance Ford circles (or alternatively, see e.g. [14, paragraph 3]), one immediately verifies that for all $n \in \mathbb{N}$ sufficiently large we have $\exp(t_n) \asymp (q_n(|x|))^2$.

This allows to finish the proof of (4.1) as follows.

$$\begin{aligned}\widetilde{\ell}_{G,q}\left(((x_k, e_k))_k\right) &= \lim_{n\to\infty} \frac{1}{2\log q_n(|x|)} \sum_{k=1}^{n} \{e_k\,(i\infty)\,, e_k\,(0)\}_G \\ &= \lim_{n\to\infty} \frac{1}{t_n} \left(\{i, y_n\}_G + \{y_n, \xi_{n+1}\}_G - \{i, \xi_1\}_G\right) \\ &= \lim_{n\to\infty} \frac{1}{t_n} \{i, y_n\}_G = \lim_{n\to\infty} \frac{1}{t_n}\{i, e_1(x + i\exp(-t_n))\}_G.\end{aligned}$$

In order to finish the proof of the proposition, it remains to show that the limit $\lim_{n\to\infty} t_n^{-1}\{i, e_1(x + i\exp(-t_n))\}_G$ is independent of the particular chosen sequence (t_n). That is, our final aim is to show that the existence of $\widetilde{\ell}_{G,q}(((x_k, e_k))_k)$ implies that $\widetilde{\ell}_{G,q}(((x_k, e_k))_k) = \widetilde{\ell}_G(((x_k, e_k))_k)$. In order to prove this, we argue similar as in [14, paragraph 3] as follows. Suppose that $\widetilde{\ell}_{G,q}(((x_k, e_k))_k)$ exists, and define $n_t := \sup\{n \in \mathbb{N} : 2\log q_n(|x|) \le t\}$ and $\alpha_h := \langle \widetilde{\ell}_{G,q}(((x_k, e_k))_k), h\rangle$, for arbitrary $t > 0$ and $h \in \mathcal{C}_2(G)$. We then have

$$\begin{aligned}&\limsup_{t\to\infty} \left| \frac{\left\langle \{i, e_1(x + i\exp(-t))\}_G\,, h\right\rangle}{t} - \frac{\left\langle \sum_{k=1}^{n_t} \{e_k(i\infty), e_k(0)\}_G\,, h\right\rangle}{2\log q_{n_t}(|x|)} \right| \\ &= \limsup_{t\to\infty} \left| \frac{2\log q_{n_t}(|x|)\left\langle \{i, e_1(x + i\exp(-t))\}_G\,, h\right\rangle}{2t\log q_{n_t}(|x|)} - \frac{t\left\langle \sum_{k=1}^{n_t} \{e_k(i\infty), e_k(0)\}_G\,, h\right\rangle}{2t\log q_{n_t}(|x|)} \right|\end{aligned}$$

$$\le \limsup_{t\to\infty}\left|\frac{\left\langle \{i, e_1(x+i\exp(-t))\}_G - \sum_{k=1}^{n_t}\{e_k(i\infty), e_k(0)\}_G , h\right\rangle}{t}\right|$$

$$+ \limsup_{t\to\infty}\left|\frac{2\log q_{n_t}(|x|) - t}{t}\right| \left|\frac{\left\langle \sum_{k=1}^{n_t}\{e_k(i\infty), e_k(0)\}_G , h\right\rangle}{2\log q_{n_t}(|x|)}\right|$$

$$\le \limsup_{t\to\infty}\frac{\text{const.}}{t} + \limsup_{n\to\infty}|\alpha_h|\frac{\log|x_{n+1}|}{\log q_n(|x|)} = 0 + \limsup_{n\to\infty}|\alpha_h|\frac{\log|x_{n+1}|}{\log q_n(|x|)}.$$

This shows that if $\alpha_h = 0$ holds for all $h \in \mathcal{C}_2(G)$, then $\widetilde{\ell}_G(((x_k, e_k))_k)$ exists and has to be equal to $\widetilde{\ell}_{G,q}(((x_k, e_k))_k)$. Hence, in this case the proof is finished. Therefore, we can now assume without loss of generality that there exists $h \in \mathcal{C}_2(G)$ such that $\alpha_h > 0$. By the above, in order to finish the proof of the proposition it is sufficient to show that $\limsup_{n\to\infty}\frac{\log|x_{n+1}|}{\log q_n(|x|)} = 0$. For this, observe

$$\alpha_h = \lim_{n\to\infty}\frac{\left\langle \sum_{k=1}^{n+1}\{e_k(i\infty), e_k(0)\}_G , h\right\rangle}{2\log q_{n+1}(|x|)}$$

$$= \lim_{n\to\infty}\frac{\left\langle \sum_{k=1}^{n}\{e_k(i\infty), e_k(0)\}_G + \{e_{n+1}(i\infty), e_{n+1}(0)\}_G , h\right\rangle}{2\log q_n(|x|) + 2\log|x_{n+1}|}$$

$$= \lim_{n\to\infty}\frac{\left\langle \sum_{k=1}^{n}\{e_k(i\infty), e_k(0)\}_G , h\right\rangle\left(1 + \frac{\langle\{e_{n+1}(i\infty), e_{n+1}(0)\}_G , h\rangle}{\left\langle \sum_{k=1}^{n}\{e_k(i\infty), e_k(0)\}_G , h\right\rangle}\right)}{2\log q_n(|x|)\left(1 + \frac{\log|x_{n+1}|}{\log q_n(|x|)}\right)}$$

$$= \alpha_h \lim_{n\to\infty}\frac{1 + \frac{\langle\{e_{n+1}(i\infty), e_{n+1}(0)\}_G , h\rangle}{\left\langle \sum_{k=1}^{n}\{e_k(i\infty), e_k(0)\}_G , h\right\rangle}}{1 + \frac{\log|x_{n+1}|}{\log q_n(|x|)}}.$$

Now, suppose by way of contradiction that $\limsup_{n\to\infty}\frac{\log|x_{n+1}|}{\log q_n(|x|)} > 0$. Then there exists a subsequence (n_k) such that $\lim_{k\to\infty}\frac{\log|x_{n_k+1}|}{\log q_{n_k}(|x|)} > 0$, and consequently we have $\lim_{k\to\infty}|x_{n_k+1}| = \infty$. Combining this with our assumption $\alpha_h > 0$, it follows

$$1 = \lim_{k\to\infty}\frac{\log q_{n_k}(|x|)}{\left\langle \sum_{m=1}^{n_k}\{e_m(i\infty), e_m(0)\}_G , h\right\rangle}\frac{\left\langle\left\{e_{n_k+1}(i\infty), e_{n_k+1}(0)\right\}_G , h\right\rangle}{\log|x_{n_k+1}|} = \frac{1}{\alpha_h}\cdot 0.$$

This is a contradiction, and hence it follows that $\limsup_{n\to\infty} \frac{\log|x_{n+1}|}{\log q_n(|x|)} = 0$. □

For our final result in this section, recall from the introduction that $f_1, \ldots, f_{2\mathfrak{g}}$ refers to a fixed $\mathbb{R}$-basis of $\mathcal{C}_2(G)$ given by the real and imaginary part of some complex basis of $\mathcal{C}_2(G)$. We then define for $e \in E_G$ and $\alpha \in \mathbb{R}^{2\mathfrak{g}}$,

$$\widetilde{\mathcal{F}}_\alpha(e) := \Big\{x \in \mathcal{I} : ((x_k, e_k))_k \in \Sigma_G, e_1 = e \\ \text{such that } \big(\langle \widetilde{\ell}_G(((x_k, e_k))_k), f_1\rangle, \ldots, \langle \widetilde{\ell}_G(((x_k, e_k))_k), f_{2\mathfrak{g}}\rangle\big) = \alpha\Big\},$$

where we have set $x := -\mathrm{sign}(x_1)[|x_1|, |x_2|, \ldots]$.

Lemma 4.3. *For each $e, e' \in E_G$ and $\alpha \in \mathbb{R}^{2\mathfrak{g}}$, we have*

$$\dim_H\big(\widetilde{\mathcal{F}}_\alpha(e)\big) = \dim_H\big(\widetilde{\mathcal{F}}_\alpha(e')\big).$$

Proof. Let $e, e' \in E_G$ be given. Since Σ_G is finitely irreducible, it follows that there exists $n \in \mathbb{N}$ and $((x_1, e_1), \ldots, (x_n, e_n)) \in \Sigma_G^*$ such that $e_1 = e$ and $e_n = e'$. This implies that there exists $g \in G$ such that $eST^{x_1}\ldots ST^{x_n} = ge'$. Then note that for $\widetilde{g} := e^{-1}ge' = ST^{x_1}\ldots ST^{x_n}$ one immediately verifies that $\widetilde{g}(\mathcal{I}) \subset \mathcal{I}$. Using this observation, the G-invariance of the modular symbol, and the fact that the limiting modular symbol does not depend on the starting point of the path along which one integrates, we obtain for each $y \in \widetilde{\mathcal{F}}_\alpha(e')$ and $\alpha \in \mathbb{R}^{2\mathfrak{g}}$,

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}\left\{i, e'(y + i\exp(-t))\right\}_G &= \lim_{t\to\infty}\frac{1}{t}\left\{g(i), ge'(y + i\exp(-t))\right\}_G \\ &= \lim_{t\to\infty}\frac{1}{t}\left\{i, ee^{-1}ge'(y + i\exp(-t))\right\}_G \\ &= \lim_{t\to\infty}\frac{1}{t}\{i, e\widetilde{g}(y + i\exp(-t))\}_G \\ &= \lim_{t\to\infty}\frac{1}{t}\{i, e(\widetilde{g}(y) + i\exp(-t))\}_G.\end{aligned}$$

This shows that $\widetilde{g}(\widetilde{\mathcal{F}}_\alpha(e')) \subset \widetilde{\mathcal{F}}_\alpha(e)$. Since $\widetilde{g}$ is conformal, and hence in particular bi-Lipschitz, and since $e, e' \in E_G$ were arbitrary, the lemma follows. □

5. Modular potential and pressure function

In this section we collect results from the general thermodynamic formalism which will be required in the proof of our Main Theorem.

Let $I : \Sigma_G \to \mathbb{R}$ refer to the canonical potential function associated with the Gauss-map $\mathcal{G}$, given by

$$I : ((x_k, e_k))_k \mapsto \log\left|\mathcal{G}'([|x_1|, |x_2|, \ldots])\right|.$$

Also, we require the potential function $J : \Sigma_G \to \mathbb{R}^{2\mathfrak{g}}$ given for $((x_k, e_k))_k \in \Sigma_G$ by

$$J\big(((x_k, e_k))_k\big) := \big(\langle\{e_1(i\infty), e_1(0)\}_G, f_1\rangle, \ldots, \langle\{e_1(i\infty), e_1(0)\}_G, f_{2\mathfrak{g}}\rangle\big),$$

where we will think of J as given by the vector $J =: (J_1, \ldots, J_{2\mathfrak{g}})$.

Finally, the modular pressure function $P : \mathbb{R}^{2\mathfrak{g}} \times (1/2, \infty) \to \mathbb{R}$ associated with J is then defined for $t = (t_1, \ldots, t_{2\mathfrak{g}}) \in \mathbb{R}^{2\mathfrak{g}}$ and $\beta \in (1/2, \infty)$ by (here, $[\![\ ]\!]$ refers to the cylinder set in Σ_G)

$$P(t, \beta) := \lim_{n\to\infty} \frac{1}{n} \log \sum_{\omega \in \Sigma_G^n} \exp \sup_{x \in [\![\omega]\!]} S_n \left((t|J(x)) - \beta I(x) \right).$$

Note that $(t|J) - \beta I$ is acceptable in the sense of Mauldin/Urbański [22, Def. 2.1.4], and this implies that P is well defined. Also, since J is Hölder continuous and bounded, one immediately verifies that $(t|J) - \beta I$ is summable for each $\beta > 1/2$ (for the definition of summability we refer to [22, p. 27]). In particular, this also gives that P is continuous. An argument similar to [14, paragraph 6] (see also [12,13]) then gives that $\lim_{\beta \searrow \frac{1}{2}} P(t, \beta) = \infty$ and $\lim_{\beta\to\infty} P(t, \beta) = -\infty$. Combining this with the continuity of P, it follows that there exists a function

$$\beta_G : \mathbb{R}^{2\mathfrak{g}} \to (1/2, \infty) \tag{5.1}$$

such that for each $t \in \mathbb{R}^{2\mathfrak{g}}$ we have $P(t, \beta_G(t)) = 0$.

We require the following facts from the general thermodynamic formalism, which can be found for instance in [22].

- For the potential function $(t|J) - \beta_G(t)I$ there exists a unique ergodic Gibbs measure μ_{t,β_G} which is positive on open subsets of Σ_G. In particular, we hence have that there exists a constant $Q > 1$ such that for each $\omega \in \Sigma_G^n$ and $x \in [\![\omega]\!]$ we have

$$Q^{-1} \leq \frac{\mu_{t,\beta_G}([\![\omega]\!])}{\exp\left(S_n\left((t|J(x)) - \beta_G(t)I(x)\right) - nP(t, \beta_G(t))\right)} \leq Q. \tag{5.2}$$

  For ease of notation, throughout we put $\mu_t := \mu_{t,\beta_G}$.
- By setting

$$\partial_{t_i} P(t, \beta_G(t)) := \left.\frac{\partial P(t, \beta)}{\partial t_i}\right|_{(t,\beta_G(t))} \quad \text{and} \quad \partial_\beta P(t, \beta_G(t)) := \left.\frac{\partial P(t, \beta)}{\partial \beta}\right|_{(t,\beta_G(t))},$$

  we have for all $i \in \{1, \ldots, 2\mathfrak{g}\}$,

$$\partial_{t_i} P(t, \beta_G(t)) = \int J_i \,\mathrm{d}\mu_t \quad \text{and} \quad \partial_\beta P(t, \beta_G(t)) = -\int I \,\mathrm{d}\mu_t. \tag{5.3}$$

  With $\alpha_i(t) := \partial_{t_i}\beta_G(t)$, the implicit function theorem then implies that

$$\alpha_i(t) = -\frac{\partial_{t_i} P(t, \beta_G(t))}{\partial_\beta P(t, \beta_G(t))} = \frac{\int J_i \,\mathrm{d}\mu_t}{\int I \,\mathrm{d}\mu_t}, \quad \text{for all } i \in \{1, \ldots, 2\mathfrak{g}\}. \tag{5.4}$$

Let us deduce some further results crucial for the proof of our Main Theorem. Note that parts of the proof of the following result are inspired by a similar argument given in [16].

Proposition 5.1. *The Hessian $(\nabla^2\beta_G)(t)$ is strictly positive definite for all $t \in \mathbb{R}^{2\mathfrak{g}}$. In particular, the function $\beta_G : \mathbb{R}^{2\mathfrak{g}} \to \mathbb{R}$ is strictly convex and the gradient map $\nabla\beta_G : \mathbb{R}^{2\mathfrak{g}} \to \nabla\beta_G(\mathbb{R}^{2\mathfrak{g}})$ is a diffeomorphism with a well-defined inverse $t : \nabla\beta_G(\mathbb{R}^{2\mathfrak{g}}) \to \mathbb{R}^{2\mathfrak{g}}$.*

Proof. As before, let $\mu_t := \mu_{t,\beta_G}$ denote the unique Gibbs measure for the potential function $(t|J) - \beta_G(t)I$. Also, for ease of exposition, let $J_0 := -I$, as well as $\partial_0 := \partial_\beta$ and $\partial_i := \partial_{t_i}$, for

$i=1,\dots,2\mathfrak{g}$. Since $t \mapsto P(t,\beta_G(t))$ defines a constant function, its partial derivative with respect to t_i vanishes for all $i=1,\dots,2\mathfrak{g}$. This implies

$$\partial_i P\left(t,\beta_G(t)\right) = -\partial_0 P\left(t,\beta_G(t)\right)\partial_i \beta_G(t), \quad \text{for all } i=1,\dots,2\mathfrak{g}.$$

By taking partial derivatives with respect to t_j on both sides of this equality we obtain

$$\begin{aligned}\partial_{ij}P(t,\beta_G(t))\partial_{0i}P(t,\beta_G(t))\partial_j\beta_G(t) = &-(\partial_{0j}P(t,\beta_G(t))) + \partial_{00}P(t,\beta_G(t)\partial_j\beta_G(t))\partial_i\beta_G(t)\\ &- \partial_0 P\left(t,\beta_G(t)\right)\partial_{ij}\beta_G(t).\end{aligned}$$

Hence, by defining

$$A := \left(\partial_{ij}P\left(t,\beta_G(t)\right)\right)_{i,j=0,\dots,2\mathfrak{g}} \quad \text{and} \quad C := \left(c_{ij}\right)_{i,j},$$

where

$$c_{ij} := \begin{cases}\alpha_j(t) & \text{for } i=0,\ j=1,\dots,2\mathfrak{g}\\ \delta_{ij} & \text{for } i,j=1,\dots,2\mathfrak{g},\end{cases}$$

we obtain

$$\left(-\partial_0 P\left(t,\beta_G(t)\right)\right)\left(\partial_{ij}\beta_G(t)\right)_{i,j=1,\dots,2\mathfrak{g}} =: B = C^T AC.$$

Using (5.3), it is now sufficient to show that B is strictly positive definite, or equivalently, that A is positive definite on the image $\mathrm{Im}(C)$ of C. Here,

$$\mathrm{Im}(C) := \left\{\left(\sum_{i=1}^{2\mathfrak{g}}\lambda_i\alpha_i, \lambda_1,\dots,\lambda_{2\mathfrak{g}}\right) : \left(\lambda_1,\dots,\lambda_{2\mathfrak{g}}\right) \in \mathbb{R}^{2\mathfrak{g}}\right\}.$$

For this it is sufficient to show that we have for all $y=(y_0,y_1,\dots,y_{2\mathfrak{g}}) \in \mathrm{Im}(C)\setminus\{0\}$,

$$y^T Ay > 0.$$

In order to prove this, note that by [22, Proposition 2.6.14] we have

$$\begin{aligned}\partial_{ij}P\left(t,\beta_G(t)\right) &= \sum_{k=0}^{\infty}\mu_t\left((J_i-\mu_t(J_i))(J_j\circ\sigma^k-\mu_t(J_j))\right)\\ &= \sum_{k=0}^{\infty}\mu_t\left((J_j-\mu_t(J_j))(J_i\circ\sigma^k-\mu_t(J_i))\right) =: \sigma_t^2\left(J_i,J_j\right).\end{aligned}$$

Using this, it follows

$$\begin{aligned}y^T Ay = \sum_{i,j=0}^{2\mathfrak{g}} y_i y_j \sigma_t^2\left(J_i,J_j\right) &= \sum_{i,j=0}^{2\mathfrak{g}}\sigma_t^2\left(y_iJ_i,y_jJ_j\right) = \sigma_t^2\left(\sum_{i=0}^{2\mathfrak{g}}y_iJ_i,\sum_{i=0}^{2\mathfrak{g}}y_iJ_i\right)\\ &=: \sigma_t^2\left(\sum_{i=0}^{2\mathfrak{g}}y_iJ_i\right) \geq 0.\end{aligned}$$

Since $\alpha_i = \mu_t(J_i)/\mu_t(I)$, we have for $y=(\sum\lambda_i\alpha_i,\lambda_1,\dots,\lambda_{2\mathfrak{g}}) \in \mathrm{Im}(C)$,

$$\mu_t\left(\sum_{i=1}^{2\mathfrak{g}}\lambda_iJ_i - \sum_{i=1}^{2\mathfrak{g}}\lambda_i\alpha_i I\right) = 0.$$

Let us assume by way of contradiction that $\sigma_t^2(\sum_{i=0}^{2\mathfrak{g}} y_i J_i)$ vanishes. Note that $\sigma_t^2(\sum_{i=0}^{2\mathfrak{g}} y_i J_i) = 0$ if and only if $\sum_{i=0}^{2\mathfrak{g}} y_i J_i$ is cohomologous to 0 within the class of bounded Hölder continuous functions. The latter means that there exists a bounded Hölder continuous function u on Σ_G such that (cf. [22, Lemma 4.8.8])

$$\sum_{i=0}^{2\mathfrak{g}} y_i J_i = u - u \circ \sigma. \tag{5.5}$$

Hence, it remains to show that (5.5) implies $(\lambda_1, \ldots, \lambda_{2\mathfrak{g}}) = 0$. In order to see this, we distinguish the following two cases. First, if $\sum_{i=1}^{2\mathfrak{g}} \lambda_i \alpha_i \neq 0$ then $\sum_{i=1}^{2\mathfrak{g}} \lambda_i J_i - \sum_{i=1}^{2\mathfrak{g}} \lambda_i \alpha_i I$ is an unbounded function (since I is unbounded). Since the right hand side of (5.5) is bounded, we then immediately have a contradiction. Secondly, if $\sum_{i=1}^{2\mathfrak{g}} \lambda_i \alpha_i = 0$ then consider $F := \sum_{i=1}^{2\mathfrak{g}} \lambda_i J_i$. We first investigate how F behaves on elements $\omega := ((x_k, e_k))_k \in \Sigma_G$ which are periodic in the second coordinate, that is where there exist $p \in \mathbb{N}$ such that $e_{mp+j} = e_j$ for all $m = 0, 1, \ldots$ and $j = 1, \ldots, p$. In this situation we necessarily have that $S_{p-1}F(\omega) = 0$, since otherwise we would have $\lim_{m\to\infty} |S_{mp-1}F(\omega)| = \lim_{m\to\infty} |mS_{p-1}F(\omega)| = \infty$ which contradicts (5.5). Therefore,

$$S_{p-1}F(\omega) = \sum_{j=1}^{2\mathfrak{g}} \lambda_j \left\langle \sum_{k=1}^{p} \{e_k\,(i\infty)\,, e_k(0)\}_G\,, f_j \right\rangle = 0. \tag{5.6}$$

Now let $\{\gamma_1, \ldots, \gamma_{2\mathfrak{g}}\}$ be a basis of $H_1(M_G, \mathbb{R})$ consisting of cycles. Each γ_i can be represented by an oriented closed geodesic in M_G. The forward directions of these geodesics correspond to elements $z_i = [x_1, x_2, \ldots] \in [0, 1]$ which are periodic in their continued fraction expansion, of period $2r_i$ say (if the period is odd, then replace the geodesic by twice the geodesic. Since E_G is finite, it follows that there exists $m_i \in \mathbb{N}$ and $e_{i,1} \in E_G$ such that for the element $\omega_i := ((-x_1, e_{i,1}), (x_2, e_{i,2}), (-x_3, e_{i,3}), \ldots) \in \Sigma_G$ we have $e_{i,2m_i r_i k} = e_{i,1}$, for all $k \in \mathbb{N}$ (note, in this step it is vital that the periods were chosen to be even). Hence, ω_i is of period $2m_i r_i$ and the set

$$\left\{ \sum_{k=1}^{2m_i r_i} \left\{ e_{i,k}\,(i\infty)\,, e_{i,k}(0) \right\}_G : i = 1, \ldots, 2\mathfrak{g} \right\}$$

contains a basis of $H_1(M_G, \mathbb{R})$. The assertion now follows from combining (2.1) and (5.6). □

The following immediate corollary shows that β_G and its Legendre transform $\widehat{\beta}_G$, given by

$$\widehat{\beta}_G(\alpha) := \inf_{t\in\mathbb{R}^{2\mathfrak{g}}} \{\beta_G(t) - (t|\alpha)\}\,,$$

are in fact a Legendre transform pair (cf. [24]).

Corollary 5.2. *For the Legendre transform $\widehat{\beta}_G$ of β_G, we have for each vector $\alpha \in \nabla\beta_G(\mathbb{R}^{2\mathfrak{g}})$,*

$$\widehat{\beta}_G\,(\alpha) = \beta_G(t(\alpha)) - (t(\alpha)|\alpha)\,.$$

For the final proposition of this section we require the following lemma.

Lemma 5.3. *For any measure $\mu \in \mathcal{M}(\Sigma_G, \sigma)$ we have that the first marginal of $\widetilde{\mu} := \mu \circ \widetilde{\pi}^{-1}$ is a shift invariant measure on (Σ_*, σ_*). Furthermore, if $\widetilde{\mu}$ can be written as a product measure $\nu \otimes \mathbb{P}$ on $\Sigma_* \times E_G$*

such that $\nu(U) > 0$ for all non-empty open subsets $U \subset \Sigma_*$, then $\mathbb{P}$ is equal to the equidistribution on E_G, that is $\mathbb{P}(\{e\}) = 1/\kappa$, for all $e \in E_G$.

Remark. Note that the first part of this lemma in particular shows that ν is a σ_*-invariant measure on Σ_*. It is then an immediate consequence of the ergodic theorem and the symmetry of E_G, that the limiting symbol vanishes almost surely for product measures of this type. In fact, this special situation occurs for the generalised Gauss-measure as discussed in [19], as well as for the Lyapunov spectrum arising from continued fraction expansions with bounded entries as studied in [20].

Proof of Lemma 5.3. For the first part let $A \subset \Sigma_*$ be some given Borel set. We then have

$$\widetilde{\mu}\,(A \times E_G) = \widetilde{\mu}\left(\widetilde{\sigma}^{-1}\,(A \times E_G)\right) = \widetilde{\mu}\left(\left(\widehat{\sigma}^{-1} A\right) \times E_G\right).$$

For the second part of the lemma let $p_e := \mathbb{P}(\{e\})$ for $e \in E_G$, and define $p := \max\{p_e : e \in E_G\}$. Also, let e' refer to some element of E_G such that $p_{e'} = p$, and define for $m \in \mathbb{N}$ and $e \in E_G$,

$$C^m_{e,e'} := \left\{x \in \left(\mathbb{Z}^\times\right)^{\mathbb{N}} : \tau_{x_1} \cdots \tau_{x_m}\,(e) = e'\right\}.$$

For n greater than the maximal word length of the elements in W, the $\widetilde{\sigma}$-invariance of $\widetilde{\mu}$ then gives

$$\begin{aligned} p = p_{e'} &= \frac{1}{n}\sum_{m=1}^{n} \widetilde{\mu}\left(\widetilde{\sigma}^{-m}\left(\Sigma_* \times \{e'\}\right)\right) = \frac{1}{n}\sum_{m=1}^{n} \widetilde{\mu}\left(\bigcup_{e\in E_G} C^m_{e,e'} \times \{e\}\right) \\ &= \frac{1}{n}\sum_{m=1}^{n}\sum_{e\in E_G} \widetilde{\mu}\left(C^m_{e,e'} \times \{e\}\right) = \sum_{e\in E_G}\frac{1}{n}\sum_{m=1}^{n} \nu\left(C^m_{e,e'}\right)\cdot p_e. \end{aligned}$$

Now, note that $\sum_{e\in E_G} \nu(C^m_{e,e'}) = 1$, and therefore $\sum_{e\in E_G} \frac{1}{n}\sum_{m=1}^{n} \nu(C^m_{e,e'}) = 1$. Combining this with the fact that by assumption we have $\frac{1}{n}\sum_{m=1}^{n} \nu(C^m_{e,e'}) > 0$, the calculation above implies that $p_e = p$ for each $e \in E_G$. □

Proposition 5.4. *We have $\partial_{t_i} P(0, \beta_G(0)) = 0$, for all $i \in \{1, \ldots, 2\mathfrak{g}\}$.*

Proof. Since μ_0 is the unique ergodic Gibbs measure for the potential function $-I$, it follows from Lemma 5.3 that the pull-back $\mu_0 \circ \widetilde{\pi}^{-1}$ of μ_0 to $\widetilde{\Sigma}_G$ can be written as a product measure $\nu \otimes \mathbb{P}$, where $\mathbb{P}$ refers to the equidistribution on E_G. Next note that with S referring to the elliptic generator of Γ of order 2, we have that $\{eS : e \in E_G\}$ is a set of representatives of $G \setminus \Gamma$. Using this, it follows

$$\begin{aligned} \partial_{t_i} P\,(0, \beta_G\,(0)) &= \int \left\langle \{e\,(i\infty)\,, e\,(0)\}_G\,, f_i \right\rangle \mathrm{d}\mu_0 \\ &= \frac{1}{\kappa}\sum_{e\in E_G} \left\langle \{e\,(i\infty)\,, e\,(0)\}_G\,, f_i \right\rangle \\ &= \frac{1}{\kappa}\sum_{e\in E_G} \left\langle \{eS\,(i\infty)\,, eS\,(0)\}_G\,, f_i \right\rangle \end{aligned}$$

$$= \frac{1}{\kappa} \sum_{e \in E_G} \left\langle \{e(0), e(i\infty)\}_G, f_i \right\rangle$$
$$= -\frac{1}{\kappa} \sum_{e \in E_G} \left\langle \{e(i\infty), e(0)\}_G, f_i \right\rangle = -\partial_{t_i} P(0, \beta_G(0)).$$

This implies that $\partial_{t_i} P(0, \beta_G(0)) = 0$. □

Corollary 5.5. *The function* $\beta_G : \mathbb{R}^{2\mathfrak{g}} \to \mathbb{R}$ *has a unique minimum at* $0 \in \mathbb{R}^{2\mathfrak{g}}$*, and* $\beta_G(0) = 1$.

Proof. Combining (5.4) and Proposition 5.4, it follows that $\nabla \beta_G(0) = 0$. Also, by Proposition 5.1 we have that β_G is strictly convex. Combining these two observations, it follows that β_G has a unique minimum at zero. Finally, note that $P(0, 1) = 0$ (see [14]), which immediately implies that $\beta(0) = 1$. □

6. Proof of Main Theorem

The fact that Σ_G is finitely irreducible has already been obtained in Proposition 3.1. For the remainder, let $\alpha \in \mathbb{R}^{2\mathfrak{g}}$, $e \in E_G$, and consider the set

$$\mathcal{F}_\alpha(e) := \left\{ x \in \mathcal{I} : ((x_k, e_k))_k \in \Sigma_G, e_1 = e, \lim_{n\to\infty} \frac{S_n J(((x_k, e_k))_k)}{S_n I(((x_k, e_k))_k)} = \alpha \right\},$$

where as before, we have set $x := -\text{sign}(x_1)[|x_1|, |x_2|, \ldots]$. One immediately verifies that $(q_n(|x|))^2 \asymp \exp(S_n I(((x_k, e_k))_k))$, and hence by Proposition 4.2 we have for each $((x_k, e_k))_k \in \Sigma_G$ and $\alpha \in \mathbb{R}^{2\mathfrak{g}}$,

$$\left(\langle \widetilde{\ell}_G(((x_k, e_k))_k), f_1\rangle, \ldots, \langle \widetilde{\ell}_G(((x_k, e_k))_k), f_{2\mathfrak{g}}\rangle\right) = \alpha \iff \lim_{n\to\infty} \frac{S_n J(((x_k, e_k))_k)}{S_n I(((x_k, e_k))_k)} = \alpha.$$

This shows that $\mathcal{F}_\alpha(e) = \widetilde{\mathcal{F}}_\alpha(e)$, for all $e \in E_G$. Hence, by Lemma 4.3, it is now sufficient to compute $\dim_H(\mathcal{F}_\alpha(e))$ for some fixed $e \in E_G$. For this, let $B(y, r)$ refer to the interval of length $2r$ centred at $y = [y_1, y_2, \ldots] \in \mathcal{I}_{+1}$, let $B_{y_1,\ldots,y_n} := \{[x_1, x_2, \ldots] : x_i = y_i, \text{ for } 1 \le i \le n\}$, and define

$$n_r(y) := \min\{n : B_{y_1,\ldots,y_n} \subset B(y, r)\}, \qquad m_r(y) := \max\{n : B_{y_1,\ldots,y_n} \supset B(y, r)\}.$$

Note that we clearly have that $|n_r(y) - m_r(y)|$ is uniformly bounded from above. Using this and the Gibbs property of the pull-back $\overline{\mu}_{t(\alpha)}$ of $\mu_{t(\alpha)}$ to $\overline{\Sigma}_G$ (see (5.2)), we then have for each $\alpha \in \nabla\beta_G(\mathbb{R}^{2\mathfrak{g}})$ and for $\overline{\mu}_{t(\alpha)}$–almost every $(e(x), e) \in \overline{\Sigma}_G$, where as before $x = -\text{sign}(x_1)[|x_1|, |x_2|, \ldots] \in \mathcal{F}_\alpha(e)$ and $((x_k, e_k))_k$ refers to the corresponding element in Σ_G,

$$\lim_{r\to 0} \frac{\log \overline{\mu}_{t(\alpha)}(B(e(x), r) \times \{e\})}{\log r}$$
$$= \lim_{r\to 0} \frac{\left(t(\alpha)|S_{n_r(|x|)} J\left(((x_k, e_k))_k\right)\right) - \beta(t(\alpha)) S_{n_r(|x|)} I\left(((x_k, e_k))_k\right)}{-S_{n_r(|x|)} I\left(((x_k, e_k))_k\right)}$$
$$= \beta_G(t(\alpha)) - (t(\alpha)|\alpha) = \widehat{\beta}_G(\alpha),$$

where the last equality follows from Corollary 5.2. Note that by combining (5.4) and the ergodicity of μ_t, we have that $\overline{\mu}_{t(\alpha)}(e(\mathcal{F}_\alpha(e)) \times \{e\}) / \overline{\mu}_{t(\alpha)}(e(\mathcal{I}) \times \{e\}) = 1$. Therefore, we can apply the mass distribution principle (cf. [8]) which gives

$$\dim_H(\mathcal{F}_\alpha(e)) = \widehat{\beta}_G(\alpha).$$

The remaining assertions of the Main Theorem are obtained as follows. Since $\mathcal{F}_\alpha$ can be written as a countable union of conformal images of the sets $\mathcal{F}_\alpha(e)$ for $e \in E$, we have that $\dim_H(\mathcal{F}_\alpha) = \dim_H(\mathcal{F}_\alpha(e))$. This shows that $\dim_H(\mathcal{F}_\alpha) = \widehat{\beta}_G(\alpha)$ for all $\alpha \in \nabla\beta_G(\mathbb{R}^{2\mathfrak{g}})$, and hence gives the assertion in (1.1). The facts $\beta_G(0) = 1$ and that β_G has a unique minimum at $0 \in \mathbb{R}^{2\mathfrak{g}}$ have already been obtained in Corollary 5.5. Likewise, the analytic properties of β_G stated in the Main Theorem can be deduced from Proposition 5.1. Next, for the assertion that the dimension spectrum is complete we refer to [15, Theorem 1.2] (note that the results in [15] are applicable, since β_G is strictly convex, and since by Remark 1, Section 3 we have that Σ_G can be represented by a graph directed Markov system). Finally, note that since the minimum of β_G is contained in $\nabla\beta_G(\mathbb{R}^{2\mathfrak{g}})$, an immediate application of the results in [15] gives that the dimension formula (1.1) continues to hold for all elements in the boundary of $\nabla\beta_G(\mathbb{R}^{2\mathfrak{g}})$.

References

[1] R. Adler, L. Flatto, Geodesic flows, interval maps and symbolic dynamics, Bull. Amer. Math. Soc., New Ser. 25 (1991) 229–334.

[2] E. Artin, Ein mechanisches System mit quasiergodischen Bahnen, Hamb. Math. Abh. 3 (1924) 170–177.

[3] R. Bowen, Equilibrium States and the Ergodic Theory of Anosov Diffeomorphism, in: SLNM, vol. 470, Springer-Verlag, Berlin, Heidelberg, New York, 1975.

[4] A. Connes, M. Marcolli, A walk through the noncommutative garden, 2006, Preprint, arXiv:math.QA/0601054.

[5] J.E. Cremona, Algorithms for Modular Elliptic Curves, Cambridge University Press, Cambridge, 1992.

[6] V.G. Drinfel'd, Two theorems on modular curves, Funktsional. Anal. i Prilozhen. 7 (2) (1973) 83–84.

[7] T. Eguchi, P.B. Gilkey, A.J. Hanson, Gravitation, gauge theories and differential geometry, Phys. Rep. (Review section of Phys. Lett.) 66 (6) (1980) 213–393.

[8] K. Falconer, Fractal Geometry, John Wiley & Sons Ltd., Chichester, 1990.

[9] B. Gurevich, S. Katok, Arithmetic coding and entropy for the positive geodesic flow on the modular surface, Moscow Math. J. 1 (2001) 569–582.

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

[11] E. Hopf, Fuchsian groups and ergodic theory, Trans. Amer. Math. Soc. 39 (1936) 299–314.

[12] M. Kesseböhmer, B.O. Stratmann, A multifractal formalism for growth rates and application to geometrically finite Kleinian groups, Ergodic Theory Dynam. Systems 24 (1) (2004) 141–170.

[13] M. Kesseböhmer, B.O. Stratmann, Stern–Brocot pressure and multifractal spectra in ergodic theory of numbers, Stoch. Dyn. 4 (1) (2004) 77–84.

[14] M. Kesseböhmer, B.O. Stratmann, A multifractal analysis for Stern–Brocot intervals, continued fractions and Diophantine growth rates. J. Reine Angew. Math. (605) (2007) (in press).

[15] M. Kesseböhmer, M. Urbański, Higher-dimensional multifractal value sets for infinite graph directed Markov systems. 2007, Preprint, arXiv:math.DS/0701541.

[16] S. Lalley, Distribution of periodic orbits of symbolic and Axiom flows, Adv. Appl. Math. 8 (1987) 154–193.

[17] Yu.I. Manin, Parabolic points and zeta functions of modular curves, Izv. Akad. Nauk SSSR Ser. Mat. 36 (1972) 19–66.

[18] Yu.I. Manin, M. Marcolli, Holography principle and arithmetic of algebraic curves, Adv. Theor. Math. Phys. 5 (3) (2001) 617–650.

[19] Yu.I. Manin, M. Marcolli, Continued fractions, modular symbols, and noncommutative geometry, Selecta Math. (N.S.) 8 (3) (2002) 475–521.

[20] M. Marcolli, Limiting modular symbols and the Lyapunov spectrum, J. Number Theory 98 (2) (2003) 348–376.

[21] M. Marcolli, Arithmetic Noncommutative Geometry, in: University Lecture Series, vol. 36, American Mathematical Society, Providence, RI, 2005.

[22] R.D. Mauldin, M. Urbański, Graph Directed Markov Systems, Geometry and Dynamics of Limit Sets, in: Cambridge Tracts in Mathematics, vol. 148. Cambridge, 2003.
[23] P. Nicholls, The Ergodic Theory of Discrete Groups, in: London Mathematical Society Lecture Note Series, vol. 143, Cambridge University Press, Cambridge, 1989.
[24] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, 1970.
[25] D. Ruelle, Thermodynamic Formalism, in: Encyclopedia of Mathematics and its Application, vol. 5, Addison-Wesley, 1978.
[26] B. Schoeneberg, Elliptic Modular Functions: An Introduction, Springer-Verlag, New York, 1974. Translated from the German by J.R. Smart and E. A. Schwandt, Die Grundlehren der mathematischen Wissenschaften, Band 203.
[27] C. Series, The modular surface and continued fractions, J. London Math. Soc. (2) 31 (1985) 69–80.
[28] G. Shimura, Introduction to the Arithmetic Theory of Automorphic Functions, in: Kanô Memorial Lectures, No. 1 (Publications of the Mathematical Society of Japan, No. 11), Iwanami Shoten, Publishers, Tokyo, 1971.