BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 44, Number 1, January 2007, Pages 87–132
S 0273-0979(06)01115-3
Article electronically published on October 2, 2006

SYMBOLIC DYNAMICS FOR THE MODULAR SURFACE AND BEYOND

SVETLANA KATOK AND ILIE UGARCOVICI

Regarding the fundamental investigations of mathematics,
there is no final ending ... no first beginning.
—Felix Klein

All new is well-forgotten old.
—A proverb

Abstract. In this expository article we describe the two main methods of representing geodesics on surfaces of constant negative curvature by symbolic sequences and their development. A geometric method stems from a 1898 work of J. Hadamard and was developed by M. Morse in the 1920s. It consists of recording the successive sides of a given fundamental region cut by the geodesic and may be applied to all finitely generated Fuchsian groups. Another method, of arithmetic nature, uses continued fraction expansions of the end points of the geodesic at infinity and is even older—it comes from the Gauss reduction theory. Introduced to dynamics by E. Artin in a 1924 paper, this method was used to exhibit dense geodesics on the modular surface. For 80 years these classical works have provided inspiration for mathematicians and a testing ground for new methods in dynamics, geometry and combinatorial group theory. We present some of the ideas, results (old and recent), and interpretations that illustrate the multiple facets of the subject.

Contents

Received by the editors February 27, 2005, and, in revised form, January 24, 2006.
2000 *Mathematics Subject Classification.* Primary 37D40, 37B40, 20H05.
Key words and phrases. Modular surface, geodesic flow, continued fractions, Markov partition.
This paper is based on the AWM Emmy Noether Lecture given by the first author at the Joint Mathematics Meetings in January 2004 in Phoenix, AZ.

1. Introduction

The origins of symbolic dynamics, according to many authors, including Birkhoff [Bh, p.184], can be traced to the 1898 work of Hadamard [Ha], where the author constructed (noncompact) surfaces in $\mathbb{R}^3$ of negative curvature and discovered that geodesics on these surfaces can be described by sequences of symbols via a certain "coding" procedure. Hadamard's idea was developed by Morse, Artin, Koebe, Nielsen and Hedlund in the 1920s and '30s, and since then symbolic dynamics has become one of the important tools in the study of systems with so-called "chaotic" behavior, of which geodesic flows on Riemannian manifolds of negative sectional curvature represent a major class of examples.

The goal of this survey is to describe from the historical perspective the development of the study of geodesic flows on surfaces of constant negative curvature by means of symbolic dynamics.

Let $\mathcal{H} = \{z = x+iy : y > 0\}$ be the upper half-plane endowed with the hyperbolic metric $ds = \frac{\sqrt{dx^2+dy^2}}{y}$. Recall that a geodesic with respect to this metric is either a vertical ray or a half-circle orthogonal to the real axis. The group of Möbius transformations

$$\left\{z \mapsto \frac{az+b}{cz+d} \mid a, b, c, d \in \mathbb{R}, ad-bc = 1\right\}$$

acting on $\mathcal{H}$ by orientation-preserving isometries can be identified with the group $PSL(2,\mathbb{R}) = SL(2,\mathbb{R})/\{\pm 1_2\}$, where 1_2 is the identity matrix. For a finitely generated Fuchsian group (i.e. a discrete subgroup $\Gamma \subset PSL(2,\mathbb{R})$), the factor space $M = \Gamma\backslash\mathcal{H}$ is a surface of constant negative curvature, possibly with some singularities (fixed points of elliptic elements) and punctures (cusps), and, in case of infinite volume, funnels. All necessary information about hyperbolic geometry and Fuchsian groups can be found in [B, K2].

Let $S\mathcal{H}$ denote the unit tangent bundle of $\mathcal{H}$. The geodesic flow $\{\tilde{\varphi}^t\}$ on $\mathcal{H}$ is defined as an $\mathbb{R}$-action on the unit tangent bundle $S\mathcal{H}$ which moves a tangent vector along the geodesic defined by this vector with unit speed. Let $v = (z,\zeta) \in S\mathcal{H}$, $z \in \mathcal{H}$, $\zeta \in \mathbb{C}, |\zeta| = \mathrm{Im}\,(z)$. Notice that $S\mathcal{H}$ can be identified with $PSL(2,\mathbb{R})$ by sending v to the unique $g \in PSL(2,\mathbb{R})$ such that $z = g(i)$, $\zeta = g'(z)(\iota)$, where ι is the unit vector at the point i to the imaginary axis pointing upwards (see Figure 1).

Under this identification the $PSL(2,\mathbb{R})$-action on $\mathcal{H}$ by Möbius transformations corresponds to left multiplications, and the geodesic flow corresponds to the right multiplication by the one-parameter subgroup

$$a_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} \text{ such that } \tilde{\varphi}^t(v) \leftrightarrow g a_t\,. \tag{1.1}$$

The orbit $\{ga_t\}$ projects to a geodesic through $g(i)$. The quotient space $\Gamma\backslash S\mathcal{H}$ can be identified with the unit tangent bundle of M, SM, although the structure of the fibered bundle is violated at elliptic fixed points and cusps (see [K2, §3.6] for details). The geodesic flow $\{\tilde{\varphi}^t\}$ on $\mathcal{H}$ descends to the *geodesic flow* $\{\varphi^t\}$ on the factor M via the projection $\pi : S\mathcal{H} \to SM$ of the unit tangent bundles (see e.g. [KH, §5.3 and §5.4] for more details).

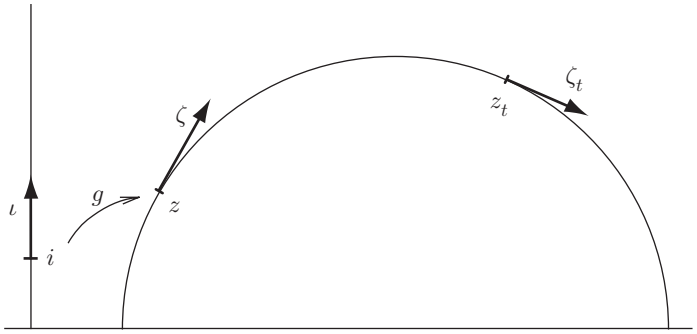

FIGURE 1. Geodesic flow on the upper half-plane $\mathcal{H}$

In all our considerations, we assume implicitly that an oriented geodesic on M is endowed with a unit tangent (direction) vector at each point and thus is an orbit of the geodesic flow $\{\varphi^t\}$ on M. For an oriented geodesic γ on M, its lift to $\mathcal{H}$ is any oriented geodesic $\tilde{\gamma}$ on $\mathcal{H}$ such that $\pi(\tilde{\gamma}) = \gamma$. In this article we mainly study the case when $\Gamma = PSL(2,\mathbb{Z}) = SL(2,\mathbb{Z})/\{\pm 1_2\}$ is the *modular group* and M is the *modular surface* which topologically is a sphere with one cusp and two singularities.

A *cross-section* C for the geodesic flow is a subset of the unit tangent bundle SM visited by (almost) every geodesic infinitely often both in the future and in the past. In other words, every $v \in C$ defines an oriented geodesic $\gamma(v)$ on M which will return to C infinitely often. The function $f : C \to \mathbb{R}$ giving the *time of the first return* to C is defined as follows: if $v \in C$ and t is the time of the first return of $\gamma(v)$ to C, then $f(v) = t$. The map $R : C \to C$ defined by $R(v) = \varphi^{f(v)}(v)$ is called the *first return map*. Thus $\{\varphi^t\}$ can be represented as the *special flow* on the space

$$C^f = \{(v,s) \mid v \in C,\ 0 \le s \le f(v)\}$$

given by the formula $\varphi^t(v,s) = (v, s+t)$ with the identification $(v, f(v)) = (R(v), 0)$.

Let $\mathcal{N}$ be a finite or countable alphabet, $\mathcal{N}^{\mathbb{Z}} = \{x = \{n_i\}_{i\in\mathbb{Z}} \mid n_i \in \mathcal{N}\}$ be the space of all bi-infinite sequences endowed with the Tikhonov (product) topology,

$$\sigma : \mathcal{N}^{\mathbb{Z}} \to \mathcal{N}^{\mathbb{Z}} \text{ defined by } \{\sigma x\}_i = n_{i+1}$$

be the left shift map, and $\Lambda \subset \mathcal{N}^{\mathbb{Z}}$ be a closed σ-invariant subset. Then (Λ, σ) is called a *symbolic dynamical system*. There are some important classes of such dynamical systems. The whole space $(\mathcal{N}^{\mathbb{Z}}, \sigma)$ is called the *Bernoulli shift*. If the space Λ is given by a set of simple transition rules which can be described with the help of a matrix consisting of zeros and ones, we say that (Λ, σ) is a *one-step topological Markov chain* or simply a *topological Markov chain* (sometimes (Λ, σ) is also called a *subshift of finite type*). Similarly, if the space Λ is determined by specifying which $(k+1)$-tuples of symbols are allowed, we say that (Λ, σ) is a *k-step topological Markov chain* (a precise definition is given in Section 4).

In order to represent the geodesic flow as a special flow over a symbolic dynamical system, one needs to choose an appropriate cross-section C and code it, i.e. to find an appropriate symbolic dynamical system (Λ, σ) and a continuous surjective map $\mathfrak{C} : \Lambda \to C$ (in some cases the actual domain of $\mathfrak{C}$ is Λ except a finite or countable

set of excluded sequences) defined such that the diagram

$$\begin{array}{ccc} \Lambda & \xrightarrow{\sigma} & \Lambda \\ {\scriptstyle \mathfrak{C}}\downarrow & & \downarrow{\scriptstyle \mathfrak{C}} \\ C & \xrightarrow{R} & C \end{array}$$

is commutative. We can then talk about *coding sequences* for geodesics defined up to a shift which corresponds to a return of the geodesic to the cross-section C. Notice that usually the coding map is not injective but only finite-to-one (see e.g. [A, §3.2 and §5]).

There are two essentially different methods of coding geodesics on surfaces of constant negative curvature. One method stems from the aforementioned work of Hadamard, which was developed by Morse [M1, M2] and Koebe [Ko]. The procedure described in Section 2 consists of recording the successive sides of a given fundamental region cut by a given geodesic. It may be applied to any finitely generated Fuchsian group Γ and assigns to the geodesic a bi-infinite sequence of generators of Γ. However, in spite of its geometric nature and seeming simplicity, this method has two major shortcomings: if the fundamental region has vertices inside $\mathcal{H}$, the geodesics passing through any of those vertices have multiple codes, and the space of all admissible codes has a complicated structure. (We believe that in general the space is not a topological Markov chain: corresponding results for the modular surface with the standard fundamental region were proved in [GL, KU1]; see Section 4.)

The second method is specific for the modular group and is of an arithmetic nature: it uses continued fraction expansions of the end points of the geodesic at infinity and a so-called *reduction theory*. This method of study and classification of indefinite binary quadratic forms goes back to 19th-century works of Gauss [Ga], Dirichlet [D], Markoff[1] [Ma] and Hurwitz [H3]. The reduction algorithm is a map on the set of all oriented geodesics in $\mathcal{H}$; it is given by a transformation of $PSL(2,\mathbb{Z})$ determined by the continued fraction expansion of the attracting end point. This map has an attractor (i.e. a set where each geodesic finds itself after finitely many iterations of the map and stays there after all further iterations). The geodesics in this attractor are called *reduced* geodesics. Based on the arithmetic of the group rather than the geometry of the fundamental region, this method produces codings of particularly simple structure—topological Markov chains. It was introduced to dynamics by Artin [Ar] in a 1924 paper, where the author used continued fractions to exhibit dense geodesics on the modular surface. If applied literally, this method gives a $GL(2,\mathbb{Z})$-invariant code, but it does not classify geodesics on the modular surface. Artin's method has been modified by Series in [S1] to eliminate this problem. Arithmetic codes for the modular group, including Artin's, and their relations to corresponding reduction theories for binary indefinite quadratic forms are discussed in Section 3, and the arithmetic coding for the congruence subgroup $\Gamma(2)$ is described in §7.5.

Considering the model of hyperbolic geometry in the unit disc $\mathcal{U}$, Nielsen [N] gave an analogue of continued fractions for representation of the points on the boundary

[1]This is the same Markov after whom *Markov chains* and *Markov processes* are named. The old-fashioned transliteration *Markoff* was used in this early publication of his Ph.D. thesis.

of $\mathcal{U}$ as infinite sequences of generators of the fundamental group Γ of a surface N_g whose fundamental region is a symmetric $4g$-sided polygon in $\mathcal{U}$.

In [He1] Hedlund represented geodesics in $\mathcal{U}$ by juxtaposing the Nielsen expansions of their end points and showed that geodesics are Γ-equivalent if and only if the corresponding sequences are shift equivalent. The author used that to prove the ergodicity of the geodesic flow on $\Gamma\backslash\mathcal{U}$ with respect to the natural Liouville measure. Artin's code was used in [He2] to obtain similar results for the modular surface. Notice that these proofs of ergodicity appeared prior to Hopf's more general analytic proof [Ho] known as "the Hopf argument". The boundary expansions method was further developed for other Fuchsian groups in [BoS, S2, S4] and is discussed in §7.1.

Ratner [R] proved the existence of a Markov partition for the geodesic flow on a compact surface of negative curvature, so that the geodesic flow is metrically isomorphic to a special flow over a topological Markov chain with a Hölder continuous ceiling function (see also [O, OW]). Ratner's construction is similar to [AW] for automorphisms of the torus based on heteroclinic connections between periodic orbits. The relationship of this construction to geometry is tenuous.

A subsequent body of work was devoted to the task of making Markov partitions geometrically explicit. In some situations the study of a first return map defined on a two-dimensional cross-section of SM, also known as a *cross-section map*, can be realized via a particular one-dimensional (non-invertible) factor-map. The latter is closely related to a map defined on the boundary of the hyperbolic plane studied by Bowen and Series [BoS], and then by Series [S2, S4]. Series showed that the geodesic flow on a surface of constant negative curvature and finite hyperbolic area is a factor of a special flow over a topological Markov chain by a continuous map which is one-to-one except for a set of the first Baire category. The symbolic dynamics derives from [BoS], and the results apply to a general class of surfaces of constant negative curvature and finite area which, however, does not include the modular surface with its standard fundamental region.

Notice that both Bowen-Series and Morse methods can be applied only to geodesics in $\mathcal{H}$ intersecting the given fundamental region $\mathcal{D}$ of Γ, which we call *$\mathcal{D}$-reduced*. Of course, any geodesic is Γ-equivalent to a $\mathcal{D}$-reduced one. In [K1] the first author developed an algorithm which $\mathcal{D}$-reduces closed geodesics on quotients by cocompact Fuchsian groups via a "reduction" map that combines two Bowen-Series-type maps on the boundary. Unfortunately, on the set of $\mathcal{D}$-reduced geodesics this map usually differs from the Bowen-Series map and the Morse map (a shift of the Morse coding sequence).

Adler and Flatto [AF1, AF2] worked on the modular surface case and obtained a representation of the geodesic flow as a special flow over a topological Markov chain by using the cross-section corresponding to the Morse code and by "linearizing" the cross-section map. In [AF4] they make a similar construction for the geodesic flow on compact surfaces of genus g with a particular $8g-4$-sided fundamental region.

This paper is organized as follows. In Section 2 we present the Morse method of coding geodesics for Fuchsian groups and its description via numerical sequences for the modular group — the *geometric code*. In §2.3 we describe the cross-section and its infinite partition for the geometric code.

In Section 3 we describe three arithmetic codes for geodesics on the modular surface obtained via generalized minus continued fractions [KU2], called the Gauss

code (G-code), the Artin code (A-code), and the Hurwitz code (H-code). All three coding procedures are actually reduction algorithms which may be considered as reduction theories for real indefinite quadratic forms translated into the matrix language. The most elegant of the three codings is the Gauss arithmetic code obtained in [K3, GuK] using minus continued fraction expansions of the end points and interpreted in [GuK] via a particular cross-section of SM. The set of such arithmetic coding sequences was identified in [GuK]: it is a symbolic Bernoulli system on the infinite alphabet $\mathcal{N}_G = \{n \in \mathbb{Z} \mid n \geq 2\}$, i.e. consists of all bi-infinite sequences constructed with symbols of the alphabet $\mathcal{N}_G$. We give similar interpretations for the Artin and the Hurwitz codes and show that the space of admissible sequences for each code is a one-step topological Markov chain with countable alphabet. We describe the corresponding symbolic representations of the geodesic flow on the modular surface as a special flow over a topological Markov chain on infinite alphabet using these arithmetic codes and give an explicit Markov partition for each code in §3.3.

In Section 4 we further analyze the geometric code for the modular group. In contrast with arithmetic codes, the set of admissible geometric coding sequences is quite complicated and, as has been proved in [KU1], is not a finite-step topological Markov chain (see Theorem 4.9). Therefore, there are geodesics whose geometric code differs from any arithmetic code. In [KU1] we identified a class of admissible geometric codes which, as well as the corresponding geodesics, we call *geometrically Markov*. We proved that geometrically Markov geodesics constitute a maximal one-step topological Markov chain in the set of all admissible geometric codes (Theorem 4.5), which is the maximal symmetric (i.e. given by a symmetric transition matrix) topological Markov chain (Theorem 4.6). It is worth noting that the H-code comes closest to the geometric code: for geometrically Markov geodesics whose codes do not contain 1's and -1's, the H-code coincides with the geometric code (Theorem 4.4). The last part of this section is devoted to a survey of the work by Grabiner and Lagarias [GL].

In Section 5 we survey other codings and interpretations for the modular group: a description of the Minkowski lattice basis reduction and connections with the geometric code and H-code, a Farey tiling interpretation of the A-code after Moeckel and Series (§5.2), a horocycle interpretation of the H-code after Fried (§5.3), and also works of Adler and Flatto (§5.4) and Arnoux (§5.5).

Section 6 is devoted to applications of arithmetic codes. In §6.2 we use the invariant Liouville measure of the geodesic flow to calculate invariant measures of one-dimensional factor-maps. In §6.3 we describe how classical results (density of closed geodesics and topological transitivity of the geodesic flow on the modular surface) can be proved using the G-code. In §6.4 we mention the work of Pollicott [P] on the asymptotic growth of the number of closed geodesics and their limit distribution proved using Artin's code. And finally, in §6.5 we explain how to obtain estimates of the topological entropy of the geodesic flow restricted to certain flow-invariant subsets of SM.

In Section 7 we describe the Bowen-Series boundary expansion for finitely generated Fuchsian groups and illustrate it with an example of the congruence subgroup $\Gamma(2)$. We develop Morse, boundary expansion, and arithmetic (via even continued fractions) codes for this group and show that in this particular case they coincide.

In this article we will consider only oriented geodesics which do not go to a cusp of M in either direction. In what follows, when we say "every oriented geodesic", we refer to every geodesic from this set. The set of excluded geodesics is insignificant from the measure-theoretic point of view, as explained in [KU2].

2. Geometric coding

2.1. **The Morse method.** We first describe the general method of coding geodesics on a surface of constant negative curvature by recording the sides of a given fundamental region cut by the geodesic. This method first appeared in a paper by Morse [M1] in 1921. However, in a 1927 paper, Koebe [Ko] mentioned an unpublished work from 1917, where the same ideas were apparently used. Starting with [S4] Series called this method *Koebe-Morse*, but since this earlier work by Koebe has not been traced, we think it is more appropriate to call this coding method the *Morse method*. We will follow [K3] in describing the Morse method for a finitely generated Fuchsian group Γ of cofinite hyperbolic area.

A Dirichlet fundamental region $\mathcal{D}$ of Γ always has an even number of sides identified by generators of Γ and their inverses; we denote this set by $\{g_i\}$. We label the sides of $\mathcal{D}$ (on the inside) by elements of the set $\{g_i\}$ as follows: if a side s is identified in $\mathcal{D}$ with the side $g_i(s)$, we label the side s by g_i. By labeling all the images of s under Γ by the same generator g_i, we obtain the labeling of the whole net $\mathcal{S} = \Gamma(\partial\mathcal{D})$ of images of sides of $\mathcal{D}$ such that each side in $\mathcal{S}$ has two labels corresponding to the two images of $\mathcal{D}$ shared by this side. We assign to an oriented geodesic in $\mathcal{H}$ a bi-infinite sequence of elements of $\{g_i\}$ which label the successive sides of $\mathcal{S}$ this geodesic crosses.

We describe the *Morse coding sequence* of a geodesic in $\mathcal{H}$ under the assumption that it does not pass through any vertex of the net $\mathcal{S}$; we call such *general position geodesics*. (Morse called the coding sequences *admissible line elements*, and some authors [S4, GL] referred to them as *cutting sequences*.) We assume that the geodesic intersects $\mathcal{D}$ and choose an initial point on it inside $\mathcal{D}$. After exiting $\mathcal{D}$, the geodesic enters a neighboring image of $\mathcal{D}$ through the side labeled, say, by g_1 (see Figure 2). Therefore this image is $g_1(\mathcal{D})$, and the first symbol in the code is g_1. If it enters the second image of $\mathcal{D}$ through the side labeled by g_2, the second image

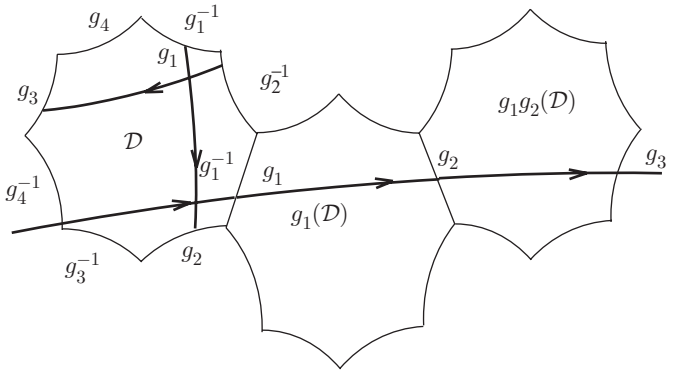

Figure 2. Morse coding

is $(g_1 g_2 g_1^{-1})(g_1(\mathcal{D})) = g_1 g_2(\mathcal{D})$, and the second symbol in the code is g_2, and so on. Thus we obtain a sequence of all images of $\mathcal{D}$ crossed by our geodesic in the direction of its orientation: $\mathcal{D}$, $g_1(\mathcal{D})$, $g_1 g_2(\mathcal{D}), \dots$, and a sequence of all images of $\mathcal{D}$ crossed by our geodesic in the opposite direction: $g_0^{-1}(\mathcal{D})$, $(g_0 g_{-1})^{-1}(\mathcal{D}), \dots$. Thus, the Morse coding sequence is

$$[\dots, g_{-1}, g_0, g_1, g_2, \dots].$$

By mapping the oriented geodesic segments between every two consecutive crossings of the net $\mathcal{S}$ back to $\mathcal{D}$ (as shown in Figure 2), we obtain a geodesic in $\mathcal{D}$. The coding sequence described above may be obtained by taking generators labeling the sides of $\mathcal{D}$ (on the outside) the geodesic hits consequently.

An element $g \in \Gamma$ is called *hyperbolic* if the associated Möbius transformation has two fixed points on the boundary of $\mathcal{H}$ (one repelling and one attracting). A geodesic on M is closed if and only if it is the projection of the axis of a hyperbolic element in Γ. For general position geodesics, a coding sequence is periodic if and only if the geodesic is closed. If a geodesic is the axis of a primitive hyperbolic element $g \in \Gamma$, i.e. a hyperbolic element which is not a power of another element in Γ, we have

$$g = g_1 g_2 \dots g_n$$

for some n. In this case the sequence is periodic with the least period $[g_1, g_2, \dots, g_n]$.

An ambiguity in assigning a Morse code occurs whenever a geodesic passes though a vertex of $\mathcal{D}$: such geodesics have more than one code, and closed geodesics have nonperiodic codes along with periodic ones (see [GL, KU1] for relevant discussions).

For free groups Γ with properly chosen fundamental regions, all reduced (here this simply means that a generator g_i does not follow or precede g_i^{-1}) bi-infinite sequences of elements from the generating set $\{g_i\}$ are realized as Morse coding sequences of geodesics on M (see [S4]), but, in general, this is not the case. Even for the classical example of $\Gamma = PSL(2, \mathbb{Z})$ with the standard fundamental region

$$F = \{z \in \mathcal{H} \mid |z| \geq 1,\ |\operatorname{Re} z| \leq 1/2\} \tag{2.1}$$

no elegant description of admissible Morse coding sequences is known and probably does not exist. Important results in this direction were obtained in [GL], where the admissible coding sequences were described in terms of forbidden blocks. The set of generating forbidden blocks found in [GL] has an intricate structure attesting the complexity of the Morse code (see §4.3 for more details).

2.2. **Geometric code for the modular surface.** Let $\Gamma = PSL(2, \mathbb{Z})$ and $M = \Gamma\backslash\mathcal{H}$ be the modular surface. Recall that the generators of $PSL(2, \mathbb{Z})$ acting on $\mathcal{H}$ are $T(z) = z + 1$ and $S(z) = -\frac{1}{z}$. The Morse code with respect to the standard fundamental region F can be assigned to any oriented geodesic γ in F (which does not go to the cusp of F in either direction) and can be described by a bi-infinite sequence of integers as follows. The boundary of F consists of four sides: left and right vertical, identified and labeled by T and T^{-1}, respectively; left and right circular, both identified and labeled by S (see Figure 3). It is clear from geometrical considerations that any oriented geodesic (not going to the cusp) returns to the circular boundary of F infinitely often. We first assume that the geodesic is in general position, i.e. does not pass through the corner $\rho = \frac{1}{2} + i\frac{\sqrt{3}}{2}$ of F (see Figure 3). We choose an initial point on the circular boundary of F and count

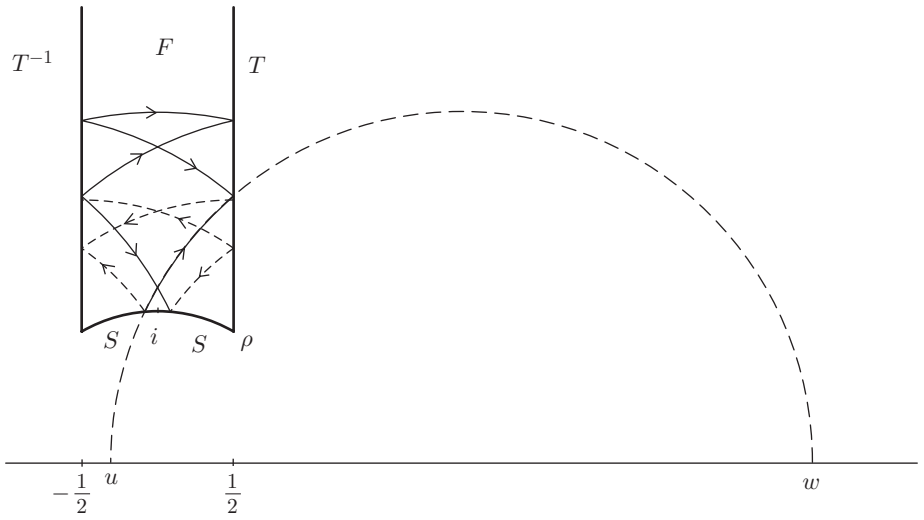

FIGURE 3. The fundamental region and a geodesic on M

the number of times it hits the vertical sides of the boundary of F moving in the direction of the geodesic. A positive integer is assigned to each block of hits of the right vertical side (or a block of T's in the Morse code), and a negative to each block of hits of the left vertical side (or a block of T^{-1}'s). Moving the initial point in the opposite direction allows us to continue the sequence backwards. Thus we obtain a bi-infinite sequence of nonzero integers

$$[\gamma] = [\dots, n_{-2}, n_{-1}, n_0, n_1, \dots],$$

uniquely defined up to a shift, which is called the *geometric code* of γ. Moving the initial point in either direction until its return to one of the circular sides of F corresponds to a shift of the geometric coding sequence $[\gamma]$. Recall that a geodesic in general position is closed if and only if the coding sequence is periodic. We refer to the least period $[n_0, n_1, \dots, n_m]$ as its geometric code. For example, the geometric code of the closed geodesic on Figure 3 is $[4, -3]$.

A geodesic with geometric code $[\gamma]$ can be lifted to the upper half-plane $\mathcal{H}$ (by choosing the initial point appropriately) so that it intersects

$$T^{\pm 1}(F), \dots, T^{n_0}(F), T^{n_0}S(F), \dots, T^{n_0}ST^{n_1}S(F), \dots,$$

in the positive direction (the sign in the first group of terms is chosen in accordance with the sign of n_0, etc.) and

$$S(F), ST^{\mp 1}(F), \dots, ST^{-n_{-1}}(F), \dots, ST^{-n_{-1}}ST^{-n_{-2}}(F), \dots,$$

in the negative direction.

The case when a geodesic passes through the corner ρ of F was described to a great extent in [GL, §7]. Such a geodesic has multiple codes obtained by approximating it by general position geodesics which pass near the corner ρ slightly higher or slightly lower. If a geodesic hits the corner only once, it has exactly two codes. If a geodesic hits the corner at least twice, it hits it infinitely many times and is closed. If it hits the corner n times in its period, it has exactly $2n + 2$

codes, i.e. shift-equivalence classes of coding sequences, some of which are not periodic. It is unknown however whether there is an upper bound on the number of shift-equivalence classes of coding sequences corresponding to closed geodesics [GL, §9].

Canonical codes considered in [K3] were obtained by the convention that a geodesic passing through ρ in the clockwise direction exits F through the right vertical side of F labeled by T (this corresponds to the approximation by geodesics which pass near the corner ρ slightly higher). According to this convention, the geometric codes of the axes of transformations $A_4 = T^4S$, $A_{3,6} = T^3ST^6S$ and $A_{6,3} = T^6ST^3S$ are $[4], [3, 6]$ and $[6, 3]$, respectively. However, all these geodesics have other codes. For example, the axis of A_4 has a code $[2, -1]$ obtained by approximation by geodesics which pass near the corner ρ slightly lower, and two nonperiodic codes for the same closed geodesic are

$$[\dots, 4, 4, 4, 4, 3, -1, 2, -1, 2, -1, 2, \dots] \text{ and } [\dots, 2, -1, 2, -1, 2, -1, 3, 4, 4, 4, \dots].$$

For more details, see [GL, KU1].

Symbolic representation of geodesics via geometric code. Let

$$\mathcal{N}^{\mathbb{Z}} = \{x = \{n_i\}_i \in \mathbb{Z} \mid n_i \in \mathcal{N}\}$$

be the set of all bi-infinite sequences on the alphabet $\mathcal{N} = \{n \in \mathbb{Z} \mid n \neq 0\}$, endowed with the Tykhonov product topology, and $\sigma : \mathcal{N}^{\mathbb{Z}} \to \mathcal{N}^{\mathbb{Z}}$ the left shift map given by $\{\sigma x\}_i = n_{i+1}$. Let X_0 be the set of admissible geometric coding sequences for general position geodesics in M, and X be its closure in the Tykhonov product topology. It was proved in [GL, Theorem 7.2] that every sequence in X is a geometric code of a unique oriented geodesic in M and every geodesic in M has at least one and at most finitely many codes (see examples above). Thus X is a closed σ-invariant subspace of $\mathcal{N}^{\mathbb{Z}}$.

2.3. **The cross-section for the geometric code.** Since every oriented geodesic which does not go to the cusp of F in either direction returns to the circular boundary of F infinitely often, the set $B \subset SM$ consisting of all unit vectors in SM with base points on the circular boundary of F and pointing inside F (see Figure 4) is a cross-section which captures the geometric code.

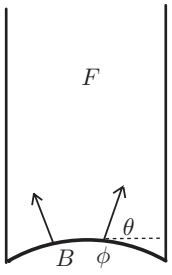

FIGURE 4. The cross-section B

The partition of the cross-section B. We parameterize the cross-section B by the coordinates (ϕ, θ), where $\phi \in [-\pi/6, \pi/6]$ parameterizes the arc and $\theta \in [-\phi, \pi - \phi]$ is the angle the unit vector makes with the positive horizontal axis in the clockwise direction. The elements of the partition of B are labeled by the symbols of the alphabet $\mathcal{N}$, $B = \cup_{n\in\mathcal{N}} C_n$ and are defined by the following condition: $C_n = \{v \in B \mid n_0(v) = n\}$; i.e. C_n consists of all tangent vectors v in B such that, for the coding sequence of the corresponding geodesic in $\mathcal{H}$, $n_0(x) = n$. Let $R : B \to B$ be the first return map. Since the first return to the cross-section exactly corresponds to the left shift of the coding sequence x associated to v, we have $n_0(R(v)) = n_1(v)$. The infinite geometric partition and its image under the return map R are sketched in Figure 5. Boundaries between the elements of the partition shown in Figure 5 correspond to geodesics going into the corner; the two vertical boundaries of the cross-section B are identified and correspond to geodesics emanating from the corner. They have more than one code. For

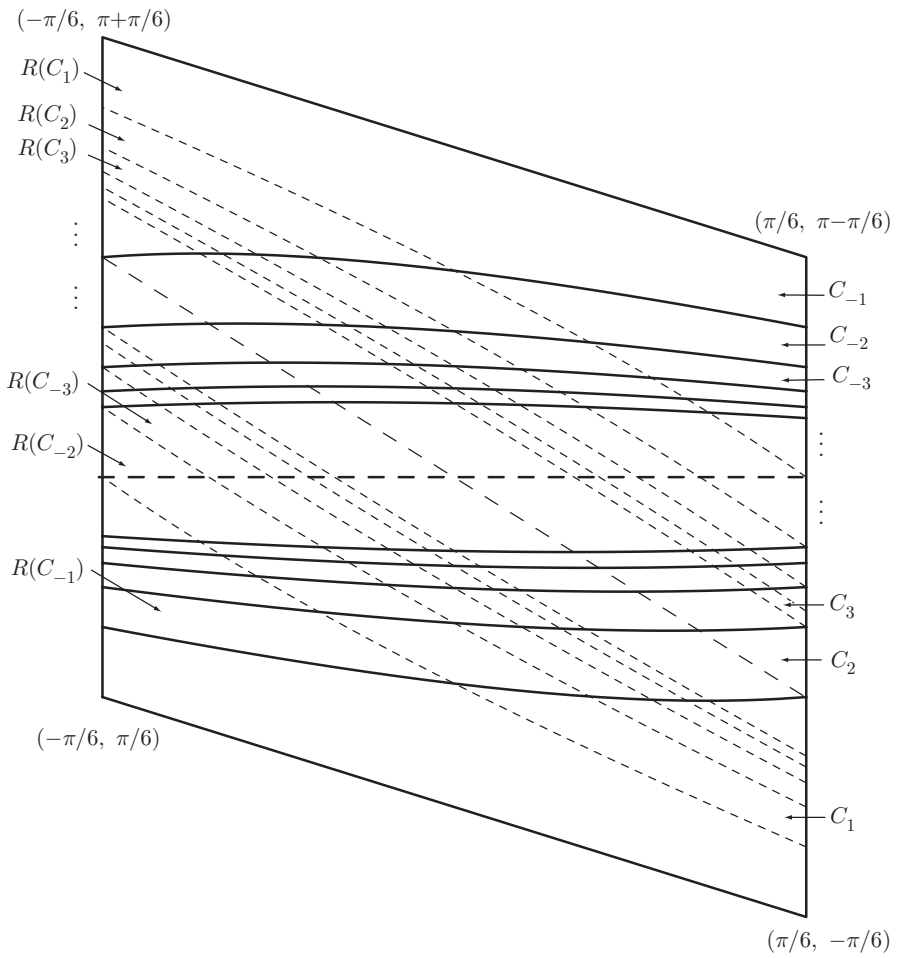

FIGURE 5. The infinite geometric partition and its image under the return map R

example, the codes [4] and $[\dots, 2, -1, 2, -1, 2, -1, 3, 4, 4, 4, \dots]$ correspond to the point on the right boundary of B between C_4 and C_3, and the codes $[2, -1]$ and $[\dots, 4, 4, 4, 4, 3, -1, 2, -1, 2, -1, 2, \dots]$ correspond to the point on the left boundary between C_2 and C_3 which are identified and are the four codes of the axis of A_4.

The coding map for the geometric code. It was proved in [GL, Lemma 7.1] that if a sequence of general position geodesics is such that the sequence of their coding sequences converges in the product topology, then the sequence of these geodesics converges to a limiting geodesic uniformly. Since the tangent vectors in the cross-section B are determined by the intersection of the corresponding geodesics with the unit circle, we conclude that the sequence of images of the coding sequences under the map $\mathfrak{C} : X \to B$ converges to the image of the limiting coding sequence. This implies that the map $\mathfrak{C}$ is continuous.

2.4. **Which geometric codes are realized?** Not all bi-infinite sequences of nonzero integers are realized as geometric codes. For instance, the periodic sequence $\{\overline{8, 2}\}$ is not a geometric code since the geometric code of the axis of T^8ST^2S is $[6, -2]$, as can be seen in Figure 6 [K3].

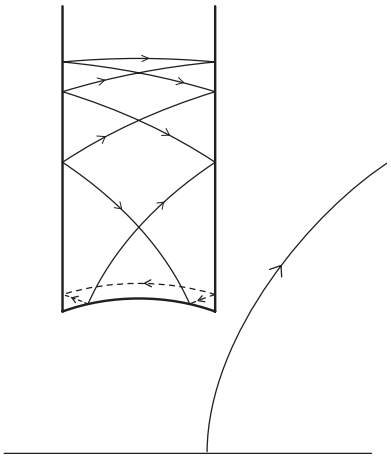

FIGURE 6. The geometric code of the axis of T^8ST^2S is $[6, -2]$

Figure 5 gives an insight into the complexity of the geometric code, where the elements C_n and their forward iterates $R(C_n)$ are shown. Each C_n is a curvilinear quadrilateral with two vertical and two "horizontal" sides, and each $R(C_n)$ is a curvilinear quadrilateral with two vertical and two "slanted" sides. The horizontal sides of C_n are mapped to vertical sides of $R(C_n)$, and the vertical sides of C_n are stretched across the parallelogram representing B and mapped to the "slanted" sides of $R(C_n)$.

If $n_0(v) = n$ and $n_1(v) = m$ for some vector $v \in B$, then $R(C_n) \cap C_m \neq \emptyset$. Therefore, as Figure 5 illustrates, the symbol 2 in a geometric code cannot be followed by 1, 2, 3, 4 and 5.

We say that C_m and $R(C_n)$ intersect "transversally" if their intersection is a curvilinear parallelogram with two "horizontal" sides belonging to the horizontal boundary of C_m and two "slanted" sides belonging to the slanted boundary of

$R(C_n)$. Notice that for each transverse intersection $R(C_n) \cap C_m$ its forward iterate under R stretches to a strip inside $R(C_m)$ between its two vertical sides. Hence, the symbol m can follow symbol n in a coding sequence.

We also observe that the elements C_m and $R(C_n)$ intersect transversally if and only if $|n| \geq 2$, $|m| \geq 2$, and

$$|1/n + 1/m| \leq 1/2.$$

This is a flow-invariant subset which constitutes the essential part of the set of geometrically Markov codes; see Theorems 4.2 and 4.5 in Section 4.

3. Arithmetic coding

3.1. **Reduction theory for indefinite quadratic forms.** Let us consider a geodesic in $\mathcal{H}$ which is a semicircle orthogonal to the real axis $\mathbb{R}$. It can be given by an equation of the form

$$A|z|^2 + B(\operatorname{Re} z) + C = 0, \tag{3.1}$$

with A, B, C real, $A \neq 0$ scaled so that $D = B^2 - 4AC = 1$. We associate to this geodesic a real quadratic form

$$Q(x, y) = Ax^2 + Bxy + Cy^2 \tag{3.2}$$

of discriminant $D = 1$. Conversely, each real quadratic form with discriminant 1 of the form (3.2) defines a geodesic in $\mathcal{H}$ given by (3.1). We denote the geodesic corresponding to the quadratic form Q by $\gamma(Q)$.

The group $SL(2, \mathbb{Z})$ acts on quadratic forms by substitutions. For $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{Z})$ we set $x = ax' + by'$, $y = cx' + dy'$, and define $Q' = g \cdot Q$ by the following equation:

$$Q'(x, y) = Q(x', y'),$$

i.e.

$$g \cdot Q = Q \circ g^{-1}.$$

Thus, the set of all real quadratic forms of discriminant 1 is decomposed into $SL(2, \mathbb{Z})$-equivalence classes. It is easy to see that this action corresponds to the action of $SL(2, \mathbb{Z})$ on geodesics by Möbius transformations: for any $g \in SL(2, \mathbb{Z})$, $\gamma(g \cdot Q) = g(\gamma(Q))$. In other words, $SL(2, \mathbb{Z})$-equivalent quadratic forms yield $SL(2, \mathbb{Z})$-equivalent geodesics in $\mathcal{H}$, hence projecting to the same geodesic in M. Therefore, we obtain a bijection between the set of geodesics in M and the set of $SL(2, \mathbb{Z})$-equivalence classes of real indefinite quadratic forms of discriminant 1. In order to classify geodesics in M we can use a version of reduction theory for binary quadratic forms.

In the most general terms, a *reduction theory* is an algorithm for finding canonical representatives in each equivalence class. Such representatives are called "reduced" elements. Each equivalence class contains a nonempty canonical set of reduced elements that form a bi-infinite sequence (which in some cases is periodic). Following the reduction algorithm one can pass from a given element to a reduced equivalent element in a finite number of steps. An application of the reduction algorithm to a reduced element yields the neighboring element on the right in the sequence.

This concept was first used by Gauss [Ga] in 1801 to classify integral binary quadratic forms of a given positive discriminant. In 1854 Dirichlet [D] described

Gauss's reduction algorithm both for $GL(2,\mathbb{Z})$- and $SL(2,\mathbb{Z})$-equivalence using regular continued fraction expansions of the roots of the corresponding quadratic equation. Dirichlet's version of Gauss's algorithm was extended by Markoff [Ma] to quadratic forms with real coefficients. Hurwitz [H1] noticed that minus (backward) continued fractions were more suited for $SL(2,\mathbb{Z})$-equivalence and expressed the reduction theory for real binary quadratic forms of positive discriminant via the closest integer minus continued fractions (see also [Fr2]).

Zagier [Z, Chapter 13] gives a complete account of the Gauss reduction theory for indefinite integral binary quadratic forms of a given discriminant $D > 0$ (which is not a perfect square) via the theory of minus continued fractions; for its translation into the matrix language see [K3]. Recall that closed geodesics on M are in one-to-one correspondence with conjugacy classes of primitive hyperbolic matrices in $PSL(2,\mathbb{Z})$ (see [K3] for details). We associate to a hyperbolic matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z})$ (which means $|a+d| > 2$) an integral quadratic form $Q_A(x,y) = cx^2 - (d-a)xy - by^2$ of discriminant $D = (a+d)^2 - 4 > 0$ (it is easy to see that D is not a perfect square). Two matrices with the same trace are $SL(2,\mathbb{Z})$-conjugate if and only if the corresponding quadratic forms are $SL(2,\mathbb{Z})$-equivalent. Conversely, to each integral quadratic form $Q(x,y)$ of discriminant $D > 0$ (which is not a perfect square) corresponds a geodesic in $\mathcal{H}$ connecting the roots of the quadratic equation $Q(z,1) = 0$. Its image in M is closed since there exists a hyperbolic matrix $A \in SL(2,\mathbb{Z})$ with the same axis (the set of integral matrices having this axis is a real quadratic field $\mathbb{Q}(\sqrt{D})$, where A corresponds to a nontrivial unit of norm 1). Two closed geodesics of the same length correspond to quadratic forms of the same discriminant; therefore the Gauss reduction theory classifies closed geodesics on M of given length.

3.2. **Continued fractions method of reduction.** In this section we describe a method of constructing arithmetic codes for geodesics on the modular surface M using expansions of the end points of their lifts to $\mathcal{H}$ in what we call *generalized minus continued fractions* [KU2]. Notice that if a geodesic does not go to the cusp of M in either direction, then the end points of all its lifts to $\mathcal{H}$ are irrational.

It was proved in [KU1, Lemma 1.1] that given a sequence of nonzero integers $\{n_i\}$, $i = 0, 1, \dots$, such that

$$|n_i| = 1 \text{ implies } n_i \cdot n_{i+1} < 0, \tag{3.3}$$

the formal minus continued fraction expression constructed out of this sequence gives a well-defined real number

$$x = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots}}} \tag{3.4}$$

(between $n_0 - 1$ and $n_0 + 1$) denoted by $(n_0, n_1, \dots)$ for short.

For a well-chosen integer-valued function $(\cdot)$, any irrational number x can be expressed uniquely in the form (3.4) where the digit n_0 is an integer equal to (x), and the digits n_i $(i \geq 1)$ are nonzero integers determined recursively by $n_i = (x_i)$, $x_{i+1} = -\frac{1}{x_i - n_i}$, starting with $x_1 = -\frac{1}{x - n_0}$. In [KU2] we described three

such functions $(\cdot)$ producing different continued fraction expansions whose digits incidentally satisfy the condition (3.3).

G-expansion. Let $\lfloor x \rfloor$ be the integer part of x (or the floor function), i.e. the largest integer less than or equal to x. The function $(x) = \lceil x \rceil = \lfloor x \rfloor + 1$ (which differs for integers from the classical ceiling function) gives the minus continued fraction expansion[2] described in [Z] and used in [K3] for coding closed geodesics. Since the coding procedure for closed geodesics is the same as the Gauss reduction theory for indefinite integral quadratic forms, we refer to this expansion as the *Gauss-* or *G-expansion* and call the corresponding code *G-code*. G-codes for oriented geodesics, not necessarily closed, were introduced in [GuK]. The digits $n_0, n_1, \dots$ of a G-expansion satisfy the condition $n_i \geq 2$ if $i \geq 1$. Conversely, any infinite sequence of integers $n_0, n_1, n_2, \dots$ with $n_i \geq 2$ for $i \geq 1$ defines a real number whose G-expansion is $\lceil n_0, n_1, n_2, \dots \rceil$.

A-expansion. The function $(x) = \lceil x \rfloor = \begin{cases} \lfloor x \rfloor & \text{if } x > 0 \\ \lceil x \rceil & \text{if } x < 0 \end{cases}$ gives an expansion which was used in [KU2] to reinterpret the classical Artin code *(A-code)*. This expansion has digits of alternating signs, and we call it the *A-expansion*. Conversely, any infinite sequence of nonzero integers with alternating signs $n_0, n_1, n_2, \dots$ defines a real number whose A-expansion is $\lceil n_0, n_1, n_2, \dots \rfloor$.

The G- and A-expansions satisfy the following properties:

(1) Two irrationals x, y are $PSL(2,\mathbb{Z})$-equivalent $\iff$ their expansions have the same tail; that is, if $x = (n_0, n_1, \dots)$ and $y = (m_0, m_1, \dots)$, then $n_{i+k} = m_{i+l}$ for some integers k, l and all $i \geq 0$.

(2) A real number x is a quadratic irrationality $\iff$ $(n_0, n_1, \dots)$ is eventually periodic.

(3) Let x and x' be conjugate quadratic irrationalities, i.e. the roots of a quadratic polynomial with integer coefficients. If $x = (\overline{n_0, n_1, \dots, n_k})$, then $\frac{1}{x'} = (\overline{n_k, \dots, n_1, n_0})$.

Let us remark that properties (2) and (3) are also valid for regular continued fraction expansions, while property (1) holds if one replaces $PSL(2,\mathbb{Z})$ by $PGL(2,\mathbb{Z})$.

H-expansion. The third expansion is obtained using the function $(x) = \langle x \rangle$ (the nearest integer to x). It was first used by Hurwitz [H1] in order to establish a reduction theory for indefinite real quadratic forms, and we call it the *Hurwitz- or H-expansion*. The digits n_i $(i \geq 1)$ of an H-expansion satisfy $|n_i| \geq 2$, and if $|n_i| = 2$, then $n_i n_{i+1} < 0$. Conversely, any infinite sequence of integers $n_0, n_1, n_2, \dots$ with the above property defines an irrational number whose H-expansion is $\langle n_0, n_1, n_2, \dots \rangle$.

The H-expansion satisfies property (2), but not (1) and (3). There is a minor exception to property (1) which was overlooked in [KU2]: it is possible for two irrationals not sharing the same tail to be $PSL(2,\mathbb{Z})$-equivalent, but this can happen if and only if one irrational has a tail of 3's in its H-expansion and the other one has a tail of -3's; i.e. the irrationals are equivalent to $r = (3 - \sqrt{5})/2$ ([H1, Fr2]). Property (3) is more serious. In order to construct a meaningful code, we need to

[2]The paper [ScSh] presents a more general form of such continued fractions used in the study of Hecke groups.

use a different expansion for $1/u$ (introduced also by Hurwitz) so that a property similar to (3) is satisfied. It uses yet another integer-valued function

$$\langle\langle x\rangle\rangle = \begin{cases} \langle x\rangle - \operatorname{sgn}(x) & \text{if } \operatorname{sgn}(x)(\langle x\rangle - x) > r = (3-\sqrt{5})/2, \\ \langle x\rangle & \text{otherwise} \end{cases}$$

and is called the *H-dual expansion*. Now if $x = \langle\overline{n_0, n_1, \dots, n_k}\rangle$, then $\frac{1}{x'}$ has a purely periodic H-dual expansion $\frac{1}{x'} = \langle\langle\overline{n_k, \dots, n_1, n_0}\rangle\rangle$. The formula for $\langle\langle\cdot\rangle\rangle$ comes from the fact that if $x = \langle n_0, n_1, \dots\rangle$, then the entries n_i satisfy the asymmetric restriction: if $|n_i| = 2$, then $n_i n_{i+1} < 0$ (for more details, see [H1, Fr2, KU2]).

Convergents. If $x = (n_0, n_1, \dots)$, then the *convergents* $r_k = (n_0, n_1, \dots, n_k)$ can be written as p_k/q_k where p_k and q_k are obtained inductively as:

$$p_{-2} = 0\,,\ p_{-1} = 1\,;\ p_k = n_k p_{k-1} - p_{k-2} \ \text{ for } k \geq 0$$
$$q_{-2} = -1\,,\ q_{-1} = 0\,;\ q_k = n_k q_{k-1} - q_{k-2} \ \text{ for } k \geq 0\,.$$

The following properties are shared by all three expansions:

- $1 = q_0 \leq |q_1| < |q_2| < \dots$;
- $p_{k-1}q_k - p_k q_{k-1} = 1$, for all $k \geq 0$.

The rates of convergence, however, are different. For the A- and H-expansions we have

$$\left| x - \frac{p_k}{q_k} \right| \leq \frac{1}{q_k^2}, \tag{3.5}$$

while for the G-expansion we have only

$$\left| x - \frac{p_k}{q_k} \right| \leq \frac{1}{q_k}. \tag{3.6}$$

A quadratic irrationality x has a purely periodic expansion if and only if x and x' satisfy certain *reduction inequalities* which give us the notion of a *reduced geodesic* for each code.

Definition 3.1. An oriented geodesic in $\mathcal{H}$ going from u to w (with u, w irrationals) is called

- *G-reduced* if $0 < u < 1$ and $w > 1$;
- *A-reduced* if $|w| > 1$ and $-1 < \operatorname{sgn}(w)u < 0$;
- *H-reduced* if $|w| > 2$ and $\operatorname{sgn}(w)u \in [r-1, r]$.

Now we can describe a reduction algorithm which works for each arithmetic code, α-code, where $\alpha = G, A, H$. For the H-code we consider only geodesics whose end points are not equivalent to r.

Reduction algorithm. Let γ be an arbitrary geodesic on $\mathcal{H}$ with end points u and w, and $w = (n_0, n_1, n_2, \dots)$. We construct the sequence of real pairs $\{(u_k, w_k)\}$ $(k \geq 0)$ defined by $u_0 = u$, $w_0 = w$ and:

$$w_{k+1} = ST^{-n_k} \dots ST^{-n_1} ST^{-n_0} w\,, \quad u_{k+1} = ST^{-n_k} \dots ST^{-n_1} ST^{-n_0} u\,.$$

Each geodesic with end points u_k and w_k is $PSL(2,\mathbb{Z})$-equivalent to γ by construction.

Theorem 3.2. *The above algorithm produces in finitely many steps an α-reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ; i.e. there exists a positive integer ℓ such that the geodesic with end points u_ℓ and w_ℓ is α-reduced.*

To an α-reduced geodesic γ we associate a bi-infinite sequence of integers

$$(\gamma) = (\dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots),$$

called its *arithmetic code*, by juxtaposing the α-expansions of $1/u = (n_{-1}, n_{-2}, \dots)$ and $w = (n_0, n_1, n_2, \dots)$ (for the H-code we need to use the dual H-expansion of $1/u$).

Remark 3.3. Any further application of the reduction algorithm to an α-reduced geodesic yields α-reduced geodesics whose codes are left shifts of the code of the initial α-reduced geodesic.

The proof of Theorem 3.2 follows the same general scheme for each code, but the notion of reduced geodesic is different in each case, and so are the properties of the corresponding expansions and estimates.

Now we associate to any oriented geodesic γ in $\mathcal{H}$ the α-code of a reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, which is obtained by the reduction algorithm described above.

Theorem 3.4. *Each geodesic γ in $\mathcal{H}$ is $PSL(2,\mathbb{Z})$-equivalent to an α-reduced geodesic ($\alpha = G, A, H$). Two reduced geodesics γ and γ' in $\mathcal{H}$ having arithmetic codes $(\gamma) = (n_i)_{i=-\infty}^{\infty}$ and $(\gamma') = (n'_i)_{i=-\infty}^{\infty}$ are $PSL(2,\mathbb{Z})$-equivalent if and only if for some integer l and all integers i one has $n'_i = n_{i+l}$.*

In [KU2] we present a geometric proof of Theorem 3.4 by constructing a cross-section C_α ($\alpha = G, A, H$) for each code directly related to the notion of α-reduced geodesics. We explain the main ideas below.

3.3. **Construction of the cross-sections for arithmetic codes.** Let $C_\alpha = P \cup Q_1 \cup Q_2$ be a subset of the unit tangent bundle SM, where P consists of all tangent vectors with base points in the circular boundary of F and pointing inward such that the corresponding geodesic is α-reduced; Q_1 consists of all tangent vectors with base points on the right vertical side of F pointing inwards such that if γ is the corresponding geodesic, then $TS(\gamma)$ is α-reduced; Q_2 consists of all tangent vectors with base points on the left vertical side of F pointing inwards such that if γ is the corresponding geodesic, then $T^{-1}S(\gamma)$ is α-reduced. If $\pi : S\mathcal{H} \to SM$ is the natural projection of the unit tangent bundles, notice that $C_\alpha = \pi(C_a)$ where C_a is the set of all unit tangent vectors with base points on the unit semi-circle $|z| = 1$ and pointing outward such that the associated geodesic on $\mathcal{H}$ is α-reduced (Figure 7). It is easy to see that for the G-code the part Q_2 is absent.

Every oriented geodesic γ on M can be represented as a bi-infinite sequence of segments σ_i between successive returns to C_α. To each segment σ_i we associate the corresponding α-reduced geodesic γ_i on $\mathcal{H}$. Thus we obtain a sequence of reduced geodesics $\{\gamma_i\}_{i=-\infty}^{\infty}$ representing the geodesic γ. If one associates to γ_i its α-code, $(\gamma_i) = (\dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots)$, then $\gamma_{i+1} = ST^{-n_0}(\gamma_i)$ and the coding sequence is shifted one symbol to the left. Thus all α-reduced geodesics γ_i in the sequence produce the same, up to a shift, bi-infinite coding sequence, which we call the *α-code* of γ and denote by (γ). The left shift of the sequence corresponds to the return of the geodesic to the cross-section C_α.

Example 3.5. Let γ be a geodesic on $\mathcal{H}$ from $u = \sqrt{5}$ to $w = -\sqrt{3}$. The G-expansions are

$$w = \lceil -1, 2, \overline{2, 3} \rceil\,, \quad 1/u = \lceil 1, \overline{2, 6, 2, 2} \rceil\,.$$

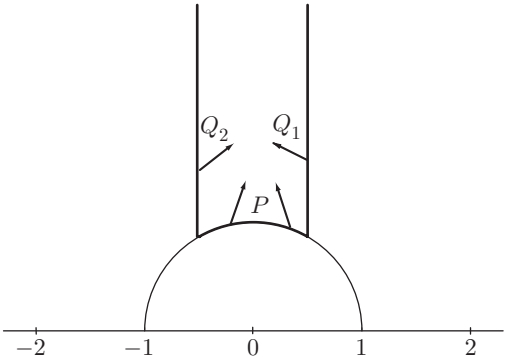

FIGURE 7. The cross-section $C_\alpha = P \cup Q_1 \cup Q_2$

First, we need to find an equivalent G-reduced geodesic. For this we use the reduction algorithm described above for G-expansions and construct the sequence (u_1, w_1), $(u_2, w_2), \dots$, until we obtain a G-reduced pair equivalent to (u, w). We have

$$w_1 = ST(w) = (1+\sqrt{3})/2, \quad u_1 = ST(u) = (1-\sqrt{5})/4\,,$$
$$w_2 = ST^{-2}(w_1) = 1 + 1/\sqrt{3},\ u_2 = ST^{-2}(u_1) = (7-\sqrt{5})/11\,,$$

and the pair (u_2, w_2) is already G-reduced. The G-expansions of $1/u_2$ and w_2 are

$$w_2 = \lceil \overline{2,3} \rceil\,, \quad 1/u_2 = \lceil 3, \overline{2,2,6,2} \rceil\,,$$

hence $\lceil \gamma \rceil = \lceil \overline{2,6,2,2}, 3, \overline{2,3} \rceil = \lceil \dots, 2,2,6,2,2,2,6,2,2,3,2,3,2,3,2,3, \dots \rceil$.

3.4. **Symbolic representation of geodesics via arithmetic codes.** Let $\mathcal{N}_G^{\mathbb{Z}}$ be the Bernoulli space on the infinite alphabet $\mathcal{N}_G = \{n \in \mathbb{Z} \mid n \geq 2\}$. We proved that each oriented geodesic which does not go to the cusp of M in either direction admits a unique G-code, $\lceil \gamma \rceil \in \mathcal{N}_G^{\mathbb{Z}}$ which does not contain a tail of 2's. Taking the closure of the set of such G-codes, we obtain the entire space $\mathcal{N}_G^{\mathbb{Z}}$. Now, each bi-infinite sequence $x \in \mathcal{N}_G^{\mathbb{Z}}$ produces a geodesic on $\mathcal{H}$ from $u(x)$ to $w(x)$, where

$$w(x) = \lceil n_0, n_1, \dots \rceil\,, \ \frac{1}{u(x)} = \lceil n_{-1}, n_{-2}, \dots \rceil\,. \tag{3.7}$$

Notice that if a sequence has a tail of 2's, then the oriented geodesic goes to the cusp. Thus the set of all oriented geodesics on M can be described symbolically as the Bernoulli space $X_G = \mathcal{N}_G^{\mathbb{Z}}$.

For the A-code, the set of all oriented geodesics (which do not go to the cusp) on M can be described symbolically as a countable one-step Markov chain $X_A \subset \mathcal{N}_A^{\mathbb{Z}}$ with the infinite alphabet $\mathcal{N}_A = \{n \in \mathbb{Z} \mid n \neq 0\}$ and transition matrix A,

$$A(n,m) = \begin{cases} 1 & \text{if } nm < 0, \\ 0 & \text{otherwise}\,. \end{cases} \tag{3.8}$$

For the H-code, recall first that the reduction algorithm and Theorem 3.4 are valid only for geodesics whose end points are not equivalent to r. Taking the closure of the set of all such H-codes, we obtain a set X_H containing also the bi-infinite sequences with a tail of 3's or -3's. These exceptional sequences are H-codes of some geodesics with one of the end points equivalent to r, but not of all such

geodesics. More precisely, each exceptional geodesic with u equivalent to r has two H-codes (see Figure 8 for the only closed such geodesic), but not all exceptional geodesics with w equivalent to r can be H-reduced (see [H1]).

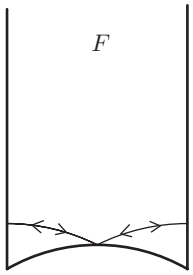

FIGURE 8. Exceptional geodesic with two H-codes, $\langle\overline{3}\rangle$ and $\langle\overline{-3}\rangle$

The set X_H is a countable one-step Markov chain $X_H \subset \mathcal{N}_H^{\mathbb{Z}}$ with infinite alphabet $\mathcal{N}_H = \{n \in \mathbb{Z} \mid |n| \geq 2\}$ and transition matrix H,

$$H(n,m) = \begin{cases} 0 & \text{if } |n| = 2 \text{ and } nm > 0, \\ 1 & \text{otherwise}. \end{cases} \tag{3.9}$$

Therefore, for $\alpha = G, A, H$, the space X_α is a closed shift-invariant subset of $\mathcal{N}_\alpha^{\mathbb{Z}}$.

Coding maps for arithmetic codes. As shown above, the coding map for each arithmetic α-code ($\alpha = G, A$), $\mathfrak{C}_\alpha : X_\alpha \to C_\alpha$ is a bijection between the cross-section C_α and the symbolic space $X_\alpha \subset \mathcal{N}_\alpha^{\mathbb{Z}}$. The map $\mathfrak{C}_H : X_H \to C_H$ is surjective and essentially one-to-one: the only exception is given by the H-codes corresponding to geodesics whose repelling end points are equivalent to r; for these exceptional H-codes the map is two-to-one.

The product topology on $\mathcal{N}_\alpha^{\mathbb{Z}}$ is induced by the distance function

$$d(x, x') = \frac{1}{m},$$

where $x = (n_i), x' = (n_i') \in \mathcal{N}_\alpha^{\mathbb{Z}}$, and $m = \min\{|i| \mid n_i \neq n_i'\} + 1$.

Proposition 3.6. *The map $\mathfrak{C}_\alpha$ is continuous.*

Proof. If $d(x, x') < \frac{1}{m}$, then the α-expansions of the attracting end points $w(x)$ and $w(x')$ of the corresponding geodesics given by (3.7) have the same first m digits. Hence the first m convergents of their α-expansions are the same, and by (3.6) and (3.5) $|w(x) - w(x')| < \frac{1}{m}$. Similarly, the first m digits of $\frac{1}{u(x)}$ and $\frac{1}{u(x')}$ are the same, and hence $|u(x) - u(x')| < \frac{u(x)u'(x)}{m} < \frac{1}{m}$. Therefore the geodesics are uniformly $\frac{1}{m}$-close. But the tangent vectors $v(x), v(x') \in C_\alpha$ are determined by the intersection of the corresponding geodesic with the unit circle. Hence, by making m large enough we can make $v(x')$ as close to $v(x)$ as we wish. □

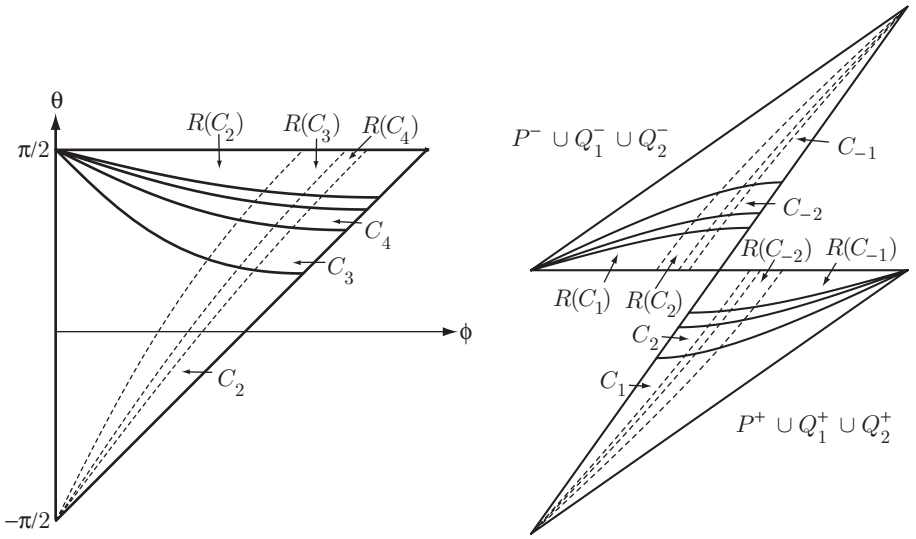

FIGURE 9. Infinite partition for the G-code (A-code, respectively) and its image under the return map R

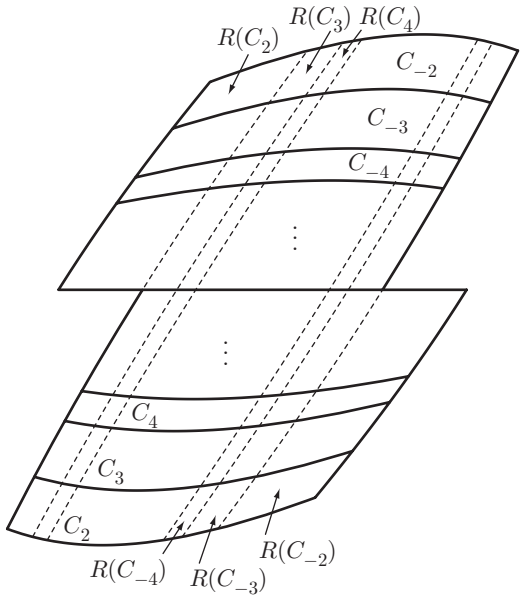

FIGURE 10. Infinite partition for the H-code and its image under the return map R

The partition of the cross-section C_α. We parameterize the lift of the cross-section C_α to $S\mathcal{H}$, C_a by the coordinates (ϕ, θ), where $\phi \in [0, \pi]$ parameterizes the unit semicircle (counterclockwise) and $\theta \in [-\pi/2, (3\pi)/2]$ is the angle the unit vector makes with the positive horizontal axis (counterclockwise). The angle θ

depends on ϕ and is determined by the condition that the corresponding geodesic is α-reduced.

The elements of the partition of C_a are labeled by the symbols of the corresponding alphabet $\mathcal{N}_\alpha$, $C_a = \bigcup_{n\in\mathcal{N}_\alpha} C_n$ and are defined by the following condition: C_n consists of all tangent vectors v in C_a such that for the coding sequence of the corresponding geodesic in $\mathcal{H}$, $n_0(x) = n$. The partitions of C_a (and therefore of C_α by projection) corresponding to the α-code ("the horizontal element") and their iteration under the first return map R to the cross-section C_a ("the vertical element") were obtained in [KU2] and are shown in Figures 9 and 10.

One can also parameterize the cross-section C_a by using the coordinates u, w and the inequalities given in Definition 3.1. In this case the pictures become even simpler: each element of the partition is a rectangle (see also §5.4). We have chosen the coordinates (ϕ, θ) to be consistent with the parametrization of the cross-section associated to the geometric code in §2.3.

Some results of this section can be illustrated geometrically since the Markov property of the partition is equivalent to the Markov property of the shift space: the symbol m follows the symbol n in the coding sequence if and only if $R(C_n)\cap C_m \neq \emptyset$, and since all intersections are transversal, according to [A, Theorem 7.9], each partition is Markov.

4. Complexity of the geometric code

Deciding which bi-infinite sequences of nonzero integers are admissible geometric codes is a nontrivial task. We present some known classes of such admissible sequences and show that the space X of all geometric codes is not a topological Markov chain.

4.1. **Classes of admissible geometric codes.**

The arithmetic codes we considered in §3.2 provide partial results: by identifying certain classes of geometric codes which coincide with arithmetic codes we obtain classes of admissible geometric codes. The first result of this kind was obtained in [GuK]:

Theorem 4.1. *A bi-infinite sequence of positive integers* $\{\dots, n_{-1}, n_0, n_1, n_2, \dots\}$ *is an admissible geometric code if and only if*

$$\frac{1}{n_i} + \frac{1}{n_{i+1}} \le \frac{1}{2} \quad \text{for all } \ i \in \mathbb{Z}\,. \tag{4.1}$$

The corresponding geodesics are exactly those for which geometric codes coincide with G-codes.

The pairs forbidden by Theorem 4.1, $\{2,p\}$, $\{q,2\}$, $\{3,3\}$, $\{3,4\}$, $\{4,3\}$, $\{3,5\}$, and $\{5,3\}$—we call them *Platonic restrictions*—are of Markov type. More precisely, the set of all bi-infinite sequences satisfying relation (4.1) can be described as a one-step countable topological Markov chain $X_P \subset \mathcal{N}_G^{\mathbb{Z}}$, with the alphabet $\mathcal{N}_G$ and transition matrix P,

$$P(n,m) = \begin{cases} 1 & \text{if } 1/n + 1/m \le 1/2\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{4.2}$$

Clearly, X_P is a shift-invariant subset of X.

The geodesics identified in Theorem 4.1 have the property that all their segments in F are positively (clockwise) oriented. Following [GuK] we call them *positive geodesics* and the corresponding class of sequences *positive coding sequences*.

A wider class of admissible coding sequences, which includes the positive ones, has been identified in [KU1]:

Theorem 4.2. *Any bi-infinite sequence of integers* $\{\dots, n_{-1}, n_0, n_1, n_2, \dots\}$ *such that*

$$\left|\frac{1}{n_i} + \frac{1}{n_{i+1}}\right| \leq \frac{1}{2} \quad \text{for } i \in \mathbb{Z} \tag{4.3}$$

is realized as a geometric code of a geodesic on M.

Remark 4.3. In order to prove that a bi-infinite sequence of nonzero integers $\{\dots, n_{-1}, n_0, n_1, n_2, \dots\}$ is a valid geometric code, it is enough to show that the geodesic from $u = 1/(n_{-1}, n_{-2}, \dots)$ to $w = (n_0, n_1, \dots)$ has the requested code. For periodic sequences we have a stronger statement; cf. [KU1, Proposition 1.3]: a periodic sequence of nonzero integers $\{\overline{n_0, n_1, n_2, \dots, n_k}\}$ satisfying (3.3) and different from $\{\overline{2}\}$ and $\{\overline{-2}\}$ is a valid geometric code if and only if the axis of the associated (hyperbolic) transformation $T^{n_0}ST^{n_1}S\dots T^{n_k}S$ going from $u = 1/(\overline{n_k, n_{k-1}, \dots, n_0})$ to $w = (\overline{n_0, n_1, \dots, n_k})$ has $[n_0, n_1, n_2, \dots, n_k]$ as a geometric code (we use the formal minus continued fractions (3.4) here). These observations were used in the proof of Theorem 4.2 above and also in the proof of Theorems 4.5 and 4.9.

The set of all bi-infinite sequences satisfying relation (4.3) can be described as a one-step countable topological Markov chain, with the alphabet $\mathcal{N} = \{n \in \mathbb{Z} \mid n \neq 0\}$ and transition matrix M,

$$M(n,m) = \begin{cases} 1 & \text{if } |1/n + 1/m| \leq 1/2\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{4.4}$$

We denote the associated one-step Markov chain by X_M. Clearly, X_M is a closed shift-invariant subset of X.

Following [KU1] we call the admissible geometric coding sequences identified in Theorem 4.2 and the corresponding geodesics *geometrically Markov*. In [KU2] we show that the H-code comes closest to the geometric code:

Theorem 4.4. *For any geometrically Markov geodesic whose geometric code does not contain* 1*'s and* -1*'s, the* H*-code coincides with the geometric code.*

The set X_M is a σ-invariant subset strictly included in X. For example, $[5, 3, -2]$ is an admissible geometric code, obtained as the code of the closed geodesic corresponding to the axis of $T^5ST^3ST^{-2}S$ (see Figure 11), but it is not geometrically Markov. Moreover, the latter is also an example of a nongeometrically Markov geodesic for which geometric and H-codes coincide. A natural question would be to characterize completely the class of geodesics for which the two codes coincide.

The following theorems were proved in [KU1]:

Theorem 4.5. *The set* X_M *is a maximal, transitive one-step countable topological Markov chain in the set of all geometric codes* X.

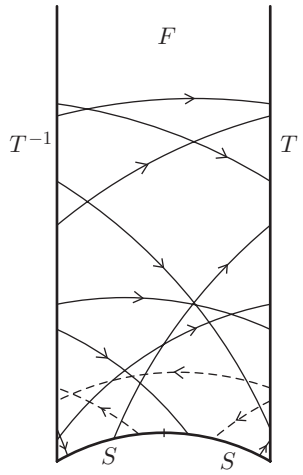

FIGURE 11. Geometric code $[5, 3, -2]$

Theorem 4.6. *The set X_M is the maximal symmetric (i.e. given by a symmetric transition matrix) one-step countable topological Markov chain in the set of all geometric codes X.*

The following result is an extension of a theorem proved in [KU2]:

Theorem 4.7. *For any geometrically Markov geodesic whose geometric code consists of symbols with alternating signs, the A-code coincides with the geometric code.*

4.2. **The space of geometric codes is not Markov.** Unlike the spaces of admissible arithmetic codes X_G, X_A, and X_H which in §3.2 were proved to form topological Markov chains, the space of admissible geometric codes X is very complicated. In order to state the complexity result proved in [KU1], we recall the notion of a k-step topological Markov chain defined on the alphabet $\mathcal{N}$ (see [KH, §1.9] for the finite alphabet definition):

Definition 4.8. Given an integer $k \geq 1$ and a map $\tau : \mathcal{N}^{k+1} \to \{0, 1\}$, the set

$$X_\tau = \{x \in \mathcal{N}^{\mathbb{Z}} \mid \tau(n_i, n_{i+1}, \ldots, n_{i+k}) = 1 \ \forall \ i \in \mathbb{Z}\}$$

with the restriction of the left-shift map σ to X_τ is called the *k-step topological Markov chain with alphabet $\mathcal{N}$ and transition map τ.*

Without loss of generality we always assume that the map τ is *essential*; i.e. $\tau(n_1, n_2, \ldots, n_{k+1}) = 1$ if and only if there exists a bi-infinite sequence in X_τ containing the $(k+1)$-block $\{n_1, n_2, \ldots, n_{k+1}\}$.

Theorem 4.9. *The space X of geometric codes is not a k-step topological Markov chain, for any integer $k \geq 1$.*

We present in what follows a sketch of the proof, correcting some minor inadvertencies that appeared in the original proof of [KU1].

Proof. Suppose that X can be represented as a k-step topological Markov chain with transition map τ. Since any k-step Markov chain is obviously $(k+1)$-step Markov, we may assume without loss of generality that k is an odd number.

In order to get a contradiction we choose two periodic sequences of integers,

$$A = [3, (4, -8)^k, 5, -2, (2, -2)^{lk}] \text{ and } B = [3, (4, -8)^k, 3, -2, (2, -2)^{lk}],$$

where l is a positive integer (to be determined later in the proof) and $(4, -8)^k$ and $(2, -2)^{lk}$ denote the fact that the pairs $\{4, -8\}$ and $\{2, -2\}$ are repeated k times and lk times, respectively, and show, using Remark 4.3, that A is a valid geometric code but B is not.

The former is checked directly: it is shown that the periodic geodesic γ_A from u_A to w_A, where

$$u_A = \frac{1}{\overline{((-2, 2)^{lk}, -2, 5, (-8, 4)^k, 3)}} \text{ and } w_A = (\overline{3, (4, -8)^k, 5, -2, (2, -2)^{lk}}),$$

is in general position and its geometric code coincides with A (here and in the rest of the proof the formal periodic minus continued fraction expansions (3.4) are used).

In order to prove the latter we consider the closed geodesic γ_B from u_B to w_B, where

$$u_B = \frac{1}{\overline{((-2, 2)^{lk}, -2, 3, (-8, 4)^k, 3)}} \text{ and } w_B = (\overline{3, (4, -8)^k, 3, -2, (2, -2)^{lk}}),$$

and compare it with the geodesic from $\bar{u}$ to $\bar{w}$, where

$$\bar{u} = \frac{1}{\overline{(-2, 2)}} = 1 - \sqrt{2}, \ \bar{w} = (3, \overline{4, -8}) = 3 - \frac{1}{\overline{(4, -8)}} = 3 - (-4 + 3\sqrt{2}) = 7 - 3\sqrt{2},$$

which passes through the left corner of $T^3(F)$. If l is large enough (depending on k) such that u_B is closer to $\bar{u}$ than w_B to $\bar{w}$, a direct computation shows that the first entry in the geometric code of γ_B will be 2 and not 3. Therefore B is not an admissible geometric code.

Since we assumed that the space X of geometric codes is k-step Markov, this implies the existence of a $(k+1)$-tuple in the infinite sequence given by B such that $\tau(n_i, n_{i+1}, \ldots, n_{i+k}) = 0$. Notice that such a $(k+1)$-tuple must contain the symbol 3 from the beginning of the sequence B. Otherwise, by using Theorem 4.2 presented above, the periodic sequence $[n_i, n_{i+1}, \ldots, n_{i+k}]$ is a valid geometric code ($k + 1$ is even), so $\tau(n_i, n_{i+1}, \ldots, n_{i+k})$ must be 1. But any $(k + 1)$-tuple containing the initial "3" appears in the sequence A, contradicting the fact that A is an admissible code. □

4.3. **Complexity results of Grabiner and Lagarias.** The main subject of [GL] is the complexity of the Morse code for the modular group and the computational complexity of conversions between different symbolic codings. They use the Morse coding sequences (referred to as "cutting sequences"), thus working with the finite alphabet of generators of $SL(2, \mathbb{Z})$: T, T^{-1}, and S. Correspondingly, instead of regular continued fraction expansions which produce a symbolic system on infinitely many symbols, the authors consider what they call *additive continued fraction expansions* by using the three symbols T, T^{-1} and S. The authors also consider what they call the *Farey tree expansion*, which is similar to our A-expansion but on a finite alphabet (see also §5.2 below).

Their main results are the following:

- The cutting sequences for irrational vertical geodesics $\{\theta + iy : y > 0\}$ (oriented downwards) are characterized in terms of forbidden blocks [GL, Theorem 6.1]. A generating set of forbidden blocks is enumerated [GL, Theorems 5.2 and 6.2]. The number of minimal forbidden blocks of length at most k grows exponentially in k [GL, Theorem 6.3]. The set of forbidden blocks for cutting sequences for all geodesics is the same as for vertical ones [GL, Theorem 7.1].
- The set of all cutting sequences is not a sofic shift—i.e. it is not a factor of a finite-step topological Markov chain (on a finite alphabet) ([GL, Theorem 7.3], related to our Theorem 4.9 on an infinite alphabet)—and it characterizes the fundamental region F of $PSL(2,\mathbb{Z})$ up to an isometry of the hyperbolic plane [GL, Theorem 8.1].
- The additive continued fraction expansion of θ can be computed from the cutting sequence expansion of $\{\theta + it : t > 0\}$ by a finite automaton [GL, Theorem 4.3] but not vice-versa.
- For real $\theta > 1$, the additive continued fraction expansion of θ can be converted into the Farey tree expansion of θ by a finite automaton, and vice versa ([GL, Theorem 3.2]; see also §5.2 below).

Without going into the technical details that can be found in [GL, §3.5], a finite automaton (finite state machine) is a finite set of relabeling rules which may involve longer and longer segments of the sequence. The key feature of a finite automaton is that it has a fixed finite amount of memory, so in order to use this notion one needs to work with sequences on a finite alphabet. Grabiner and Lagarias also show that the geometric coding encodes more information about the geodesic flow than the arithmetic codings, which retain only topological and not conformal information about the modular surface M.

5. Other codings and interpretations

5.1. **Coding geodesics and Minkowski lattice basis reduction.** We present a relationship between the coding of an oriented geodesic and the Minkowski lattice basis reduction procedure. This section is largely inspired by [GL, §3].

The definition of a Minkowski-reduced basis is of fundamental importance in the geometry of numbers [Mi, GrLe]. Here we follow the terminology of [GrLe, GL], although in dimension two the reduction algorithm described below goes back to Gauss [Ga, Article 171].

Definition 5.1. Let $L = \mathbb{Z}e_1 + \mathbb{Z}e_2$ be a lattice in $\mathbb{R}^2$. A positively oriented basis $\{m_1, m_2\}$ in L is called *Minkowski-reduced* if m_1 is the shortest vector in L and m_2 is the shortest vector, linearly independent with m_1.

It is a standard fact that a positively oriented basis $\{m_1, m_2\}$ in L is Minkowski-reduced if and only if $|m_1| \le |m_2|$ and $\frac{|<m_1,m_2>|}{|m_1|^2} \le \frac{1}{2}$. We describe now the classical algorithm of obtaining a Minkowski-reduced basis from a given basis of a lattice $L = \mathbb{Z}e_1 + \mathbb{Z}e_2$ (see also Figure 12):

[1] If $|e_2| \ge |e_1|$, go to step [2]; otherwise (if $|e_1| > |e_2|$) go to step [4].

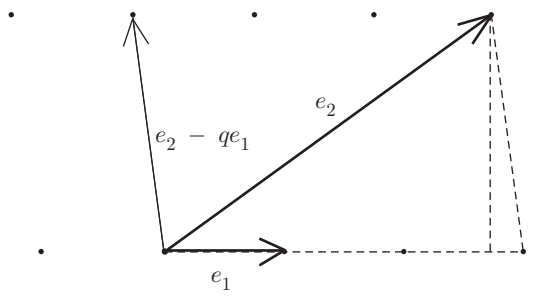

FIGURE 12. Minkowski lattice basis reduction

[2] Set $\mu \leftarrow \frac{<e_1,e_2>}{|e_1|^2}$ and $q \leftarrow \langle\mu\rangle$, where $\langle x\rangle$ is the closest integer to x. If μ is a half-integer of the form $\mu = n + \frac{1}{2}$, we make the following choice:
$$\langle\mu\rangle = \begin{cases} n \text{ if } n \geq 0 \\ n+1 \text{ if } n < 0 \end{cases}$$
SLIDE e_2 against e_1: $e_2 \leftarrow e_2 - qe_1$.

[3] If $|e_2| \geq |e_1|$, stop. (By construction $|\mu| \leq \frac{1}{2}$, hence $q = 0$, and $\{e_1, e_2\}$ is Minkowski-reduced.) Otherwise ($|e_2| < |e_1|$) go to step [4].

[4] SWAP e_1 and e_2: $e_1 \leftarrow -e_2$ and $e_2 \leftarrow e_1$, and go to step [2].

Since at step [4] the length of e_1 decreases and there are only finitely many lattice points closer to the origin than the length of the initial vector e_1, this process terminates after finitely many steps.

To any oriented geodesic on the upper half-plane $\mathcal{H}$ and a point z on it, we associate a matrix $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{R})$ which maps the positively (upwards) oriented vertical geodesic $\gamma_0 = \{iy, y > 0\}$ to this geodesic so that $g(i) = \frac{ai+b}{ci+d} = z$. We associate to $v = (z, \zeta) \in S\mathcal{H}$ (ζ is the unit vector tangent to the geodesic at z) a lattice $L = \mathbb{Z}e_1 + \mathbb{Z}e_2$ in $\mathbb{R}^2$ with the basis $\mathcal{B}_v = \{e_1 = (-c, -d),\ e_2 = (a, b)\}$. Since $\det g = 1$, the lattice is *unimodular* (i.e. the area of the fundamental parallelogram in $\mathbb{R}^2$ is equal to 1), and the basis is *positively oriented*. Conversely, any positively oriented unimodular basis $\mathcal{B} = \{e_1 = (-c, -d),\ e_2 = (a, b)\}$ in $\mathbb{R}^2$ yields an oriented geodesic in $\mathcal{H}$: the matrix corresponding to this basis is $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{R})$, and the geodesic is $g(\gamma_0)$.

Remark 5.2. Any positively oriented unimodular basis in L can be obtained from $\{e_1, e_2\}$ by using a matrix $\sigma \in SL(2,\mathbb{Z})$. This corresponds to a left multiplication of the matrix g associated to $\{e_1, e_2\}$ by $S\sigma S$. In particular, a slide of the basis corresponds to a left multiplication of g by $\begin{pmatrix} 1 & q \\ 0 & 1 \end{pmatrix} = T^q$, a swap of the basis corresponds to a left multiplication by S, and a slide followed by a swap is given by left multiplication by ST^q. Since there are nontrivial relations between the generators T and S, one can find the Minkowski-reduced basis of a lattice in many ways, not necessarily using the algorithm described above.

Lemma 5.3. *Let $v = (z, \zeta) \in S\mathcal{H}$ and $\mathcal{B}_v = \{e_1, e_2\}$ the associated basis. Then $z \in F$ if and only if B_v is a Minkowski-reduced basis.*

Proof. We have

$$z = \frac{ai+b}{ci+d} = \frac{bd+ac}{c^2+d^2} + \frac{i}{c^2+d^2}, \tag{5.1}$$

and

$$|z|^2 = \frac{a^2+b^2}{c^2+d^2}. \tag{5.2}$$

If $z \in F$, then $|z| \geq 1$ and $|\mathrm{Re}\, z| \leq \frac{1}{2}$. By (5.2) $|e_1| \leq |e_2|$. Also

$$\mu = \frac{\langle e_1, e_2 \rangle}{|e_1|^2} = -\frac{ac+bd}{a^2+b^2};$$

hence

$$|\mu| = \frac{|ac+bd|}{a^2+b^2} \leq \frac{|ac+bd|}{c^2+d^2} = |\mathrm{Re} z| \leq \frac{1}{2}.$$

Therefore the basis B_v is Minkowski-reduced. □

One can also notice that to a point z on the circular boundary of F corresponds a Minkowski-reduced basis with vectors of equal length. Based on this observation and Lemma 5.3, we explain how the geometric code is related to the Minkowski reduction procedure. The difference between the theorem presented below and the results described in [GL, §3.3] is that the authors use a finite alphabet and consider only vertical geodesics (see also §4.3 for more details of their work). Our proof is self-contained.

Let $B \subset SM$ be the cross-section for the geometric code as in §2.3, $(z, \zeta) \in B$, and γ the oriented geodesic through (z, ζ). As explained in the Introduction, this geodesic is a projection of the orbit of the geodesic flow $\tilde{\varphi}^t$ on $\mathcal{H}$, and for $t > 0$ the matrix corresponding to $v_t = (z_t, \zeta_t)$ such that $\rho(z, z_t) = t$ is

$$ga_t = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} = \begin{pmatrix} ae^{t/2} & be^{-t/2} \\ ce^{t/2} & de^{-t/2} \end{pmatrix}.$$

We denote the basis corresponding to v_t by $\mathcal{B}_{v_t} = \{e_1^{(t)}, e_2^{(t)}\}$.

Theorem 5.4. *Let $(z, \zeta) \in B$ and γ be the oriented geodesic through (z, ζ). There is an increasing sequence $0 = t_0 < t_1 < t_2 < \cdots < t_k < \cdots$ such that for any integer $k \geq 0$, $\pi(z_{t_k}, \zeta_{t_k}) \in B$, and the Minkowski-reduced basis of the lattice corresponding to (z_{t_k}, ζ_{t_k}) consists of two vectors of equal length. For $t = t_k$ the Minkowski-reduced basis can be obtained from $\{e_1^{(t)}, e_2^{(t)}\}$ by performing k slide/swap steps with $q_i = -n_i$ for $0 \leq i \leq k-1$, where $[n_0, n_1, \dots]$ is the forward part of the geometric code of γ. For $t_k < t < t_{k+1}$ the Minkowski-reduced basis can be obtained from $\{e_1^{(t)}, e_2^{(t)}\}$ by performing k slide/swap steps with $q_i = -n_i$ for $0 \leq i \leq k-1$ and an additional slide step (with appropriate sliding factor q between 0 and $-n_k$).*

Proof. By Lemma 5.3 the basis corresponding to $t_0 = 0$ is Minkowski-reduced. Suppose that the first symbol in the geometric code of γ is n_0. This means that γ intersects the vertical line $x = \frac{1}{2}$ and its $n_0 - 1$ shifts by $T(z) = z + 1$ if $n_0 > 0$, and the vertical line $x = -\frac{1}{2}$ and its $n_0 - 1$ shifts by $T^{-1}(z) = z - 1$ if $n_0 < 0$. Let t increase from $t_0 = 0$. For small t, $z_t \in F$, and by Lemma 5.3, the basis $\{e_1^{(t)}, e_2^{(t)}\}$ is already reduced. Assume that $n_0 > 0$ (the case $n_0 < 0$ is handled similarly). Then the next segment of γ corresponds to $z_t \in T(F)$. In this case T^{-1} brings z_t

to F so that the corresponding basis is Minkowski-reduced by Lemma 5.3, and it is obtained from $\{e_1^{(t)}, e_2^{(t)}\}$ by one slide step with $q_0 = -1$.

As t further increases, a similar argument shows that the Minkowski-reduced basis is obtained from $\{e_1^{(t)}, e_2^{(t)}\}$ by one slide step, where q takes consecutive values $-2, -3, \dots, -n_0$ on the consecutive segments of γ. The first critical value $t = t_1$ yields the Minkowski basis of two vectors of equal length, and a swap step assures us that $\pi(z_{t_1}, \zeta_{t_1}) \in B$. Thus, after one slide/swap step (with sliding factor $-n_0$), basically we performed a left shift in the geometric code.

If the next symbol in the geometric code is n_1, for $t_1 < t \le t_2$ the basis can be reduced by an additional slide step with an appropriate sliding factor q between 0 and $-n_1$. The general statement is proved by induction using the description of the images of F crossed by the geodesic given in §2.2. □

Remark 5.5. Let us point out that there is a natural relation between the Minkowski reduction algorithm and the Hurwitz continued fraction expansion: consider $z_t \in \gamma$ for large t, find the corresponding basis $\{e_1^{(t)}, e_2^{(t)}\}$, and reduce it using the Minkowski reduction algorithm, recording $q_0, q_1, \dots, q_k$ of the respective slides. Letting $t \to \infty$ the sequence $\{-q_i\}$ coincides with the H-code of the attracting end point of γ.

5.2. **Farey tiling interpretation of the A-code.** The *Farey sequence of order n* F_n is the set of rational numbers p/q with $(p, q) = 1$ and $|p| \le n$, $|q| \le n$ arranged in increasing order. It is convenient to include ∞ in each Farey sequence F_n. For example, the nonnegative entries of the first three sequences are

$$F_1 : 0, 1, \infty$$
$$F_2 : 0, \frac{1}{2}, 1, 2, \infty$$
$$F_3 : 0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1, \frac{3}{2}, 2, 3, \infty.$$

A basic property of the Farey sequences is the following: two rational numbers $p/q < p'/q'$ are adjacent in the Farey sequence of order $\max(|p|, |q|, |p'|, |q'|)$ if and only if $pq' - p'q = -1$.

Let $\gamma_0 = \{iy, y > 0\}$ be the standard vertical upwards oriented geodesic in $\mathcal{H}$. Its images under $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{Z})$ are geodesics in $\mathcal{H}$ with rational end points $g(0) = b/d$ and $g(\infty) = a/c$ (with the convention that $1/0 = \infty$). Since $g(0)/g(\infty) = 1 - \frac{1}{ad} \ge 0$, $g(\gamma_0)$ does not cross γ_0, and therefore the images of γ_0 under $SL(2, \mathbb{Z})$ do not cross one another. Moreover, since $ad - bc = 1$, $g(0)$ and $g(\infty)$ are adjacent in the Farey sequence of order $\max(|a|, |b|, |c|, |d|)$. They are known as *Farey edges*. The end points of Farey edges are extended rational numbers which are the images of the cusp at ∞ under $SL(2, \mathbb{Z})$; we call them *cusp points*.

Let Δ_0 denote the ideal triangle in $\mathcal{H}$ with vertices $0, 1$, and ∞. The images of Δ_0 under $SL(2, \mathbb{Z})$ are known as *Farey triangles*. If we apply to Δ_0 the transformations $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{Z})$ with $\max(|a|, |b|, |c|, |d|) \le n$, we will cover a convex subset of $\mathcal{H}$, and the images of ∞ will be exactly the Farey sequence F_n. Notice that Δ_0 is mapped onto itself by a cyclic subgroup of $SL(2, \mathbb{Z})$ of order 3; therefore each of its images will be covered three times as well. As $n \to \infty$, the images of Δ_0 will

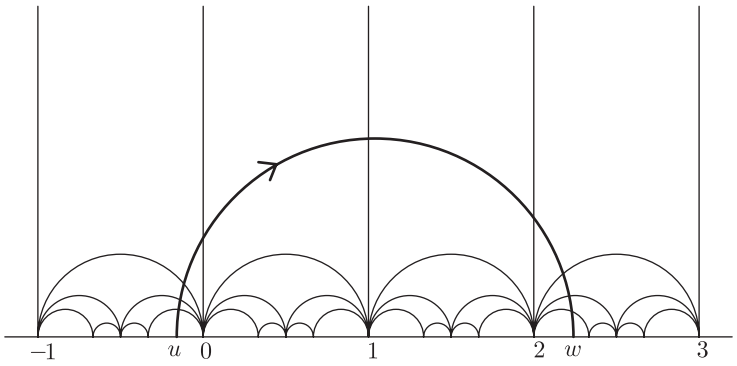

FIGURE 13. Farey tiling

cover a larger and larger part of $\mathcal{H}$, and since the rational points are dense in $\mathbb{R}$, the images of Δ_0 under $SL(2,\mathbb{Z})$ will cover $\mathcal{H}$ without overlap, forming the *Farey tiling* (see Figure 13).

Hurwitz [H3] used the Farey tiling to describe geometrically the Gauss reduction theory for $GL(2,\mathbb{Z})$. Hurwitz's elegant approach is described in [Fr2]. Series [S3] gave a similar description following an earlier work of Moeckel [Mo], where the author used the Farey tiling and its relation to continued fractions to study the asymptotic frequencies with which geodesics go into different cusps for a certain class of Fuchsian groups whose fundamental regions are made up of Farey triangles. Moeckel and Series referred to a 1916 work of Humbert [Hum], but apparently were unaware of Hurwitz's work, published in 1894.

We now use the Farey tiling to describe the A-code. Let γ be a geodesic in $\mathcal{H}$ with irrational end points u and w. Any Farey triangle Δ which meets γ must intersect it in a compact interval, and the vertices of Δ are separated by γ into a pair and a singleton. We label this interval by the singleton vertex and assign to it a "+" sign if the singleton vertex lies to the left of γ as we move in the direction from u to w, and a "−" sign if the singleton vertex lies to the right of γ. A given cusp point can label only finitely many consecutive intervals in γ, all of which have the same signs. We call a Farey edge *principal for* γ if it crosses γ and the intervals of γ on either side of it are labeled by different vertices. Notice that the signs of the intervals between principal edges for a given geodesic γ will alternate.

Suppose a geodesic γ is A-reduced, with $w = \lceil n_0, n_1, n_2, \dots \rfloor > 1$, so $n_0 \geq 1$. If we trace γ and count the number of intervals between crossings of the principal edges with the associated signs, we obtain a sequence of nonzero integers. It is easy to see that if we start tracing γ at its intersection with the vertical edge γ_0, the number of crossings between γ_0 and the second principal edge $T^{n_0}\gamma_0$ is n_0. Since the cusp at ∞ is on the left of γ, this interval is assigned a "+" sign; hence the first number in the sequence is n_0. The third principal edge is a geodesic from n_0 to $n_0 - \frac{1}{n_1}$, and the number of crossings between the second and the third principal edge is $|n_1|$. Since the cusp at n_0 is on the right of γ, this interval has a "−" sign assigned to it. Thus the second number in the sequence is $-|n_1| = n_1$, and so on. The cusp points corresponding to the singleton vertices labeling principal edges are the convergents of the A-expansion of w, $\lceil n_0, n_1, n_2, \dots, n_k \rfloor$. Thus the sequence obtained is exactly the A-expansion of w.

Since the repelling end point u of γ satisfies $-1 < u < 0$, we have $1/u = \lceil n_{-1}, n_{-2}, n_{-3}, \dots \rfloor < -1$. Therefore

$$\frac{1}{n_{-1}} < u < \frac{1}{n_{-1} - 1}\,.$$

Moving along γ from its intersection with γ_0 towards u, one observes that the principal edge preceding γ_0 is a geodesic from $1/n_{-1}$ to 0 and the number of crossings between this edge and γ_0 is exactly $|n_{-1}|$. Now, if we move from the principal edge preceding γ_0 towards γ_0, we see that the cusp at 0 is on the right of γ, so it has a "$-$" sign, and the number preceding n_0 in the sequence is n_{-1}. Thus the sequence of nonzero integers obtained by counting the number of intervals between crossings of γ with the principal edges is exactly the A-code of γ, and the changing of the original principal edge changes the sequence by a shift.

5.3. **Interpretation of the H-expansion via Ford discs.** Fried [Fr2] gives a geometric interpretation of the Hurwitz continued fraction expansion (H-expansion) in terms of Ford discs. Recall that a horodisc (i.e. a closed region bounded by a horocycle) in the hyperbolic plane $\mathcal{H}$ is either a disc tangent to the real axis or a half-plane defined by $\mathrm{Im}\,(z) \geq c$, for some $c > 0$. A *Ford disc* is a particular horodisc obtained as an image under an element of $SL(2, \mathbb{Z})$ of the standard horodisc $B(\infty)$ defined by $\mathrm{Im}\,(z) \geq 1$. Such a disc is labeled $B(r)$ if $r \in \mathbb{Q}$ is the point where it touches the real axis.

In 1917 Ford [Fo1] used horodiscs to give a geometric proof of Hurwitz's result [H2] on Diophantine approximations: if α is an arbitrary irrational number, then there are infinitely many irreducible fractions p/q (with $(p, q) = 1$) satisfying

$$\left|\frac{p}{q} - \alpha\right| < \frac{1}{\sqrt{5}\, q^2}\,,$$

and $\sqrt{5}$ is the best constant possible. (See also Ford [Fo3] for an elementary proof of this result.) Let us mention that Ford [Fo2] used horospheres in the hyperbolic 3-space to describe geometrically the properties of a sequence of complex rational fractions introduced by Hermite in order to approximate a given complex irrational.

Here are some of the basic properties of Ford discs (see [Fo3]): Ford discs do not overlap; the only Ford discs that meet $B(\infty)$ are $B(r)$ with integer r; if $B(r_1)$ and $B(r_2)$ touch, then r_1 and r_2 are consecutive Farey numbers. The H-expansion of an irrational number x can be interpreted as follows: $\langle x\rangle = \langle n_0, n_1, \dots\rangle$ if and only if the vertical oriented geodesic from ∞ to x traverses successively the Ford discs $B(p_k/q_k)$, where $p_k/q_k = \langle n_0, n_1, \dots, n_k\rangle$ are the corresponding convergents of the H-expansion.

The H-dual expansion can be given a geometric interpretation in the following way: Assign to each Ford disc $B(r)$ a *level* so that $B(\infty)$ is of level 0, $B(r)$ is of level 1 if r is an integer, $B(r')$ is of level 2 if $B(r')$ meets some Ford disc of level 1 and $B(r')$ is not of level 0 or 1, etc. In Figure 14 the large white discs are of level 1, the grey-colored discs are of level 2, and the black discs are of level 3. If the H-dual expansion of a real number x is $\langle\langle x\rangle\rangle = \langle\langle n_1, n_2, \dots\rangle\rangle$ and if $p_k/q_k = \langle\langle n_1, n_2, \dots, n_k\rangle\rangle$ are the H-dual convergents of x, then $B(p_k/q_k)$ is of level k. In other words, $\langle\langle\, \cdot\, \rangle\rangle$ is the continued fraction expansion that makes each horodisc $B(p_k/q_k)$ associated to the convergent p_k/q_k be of level k.

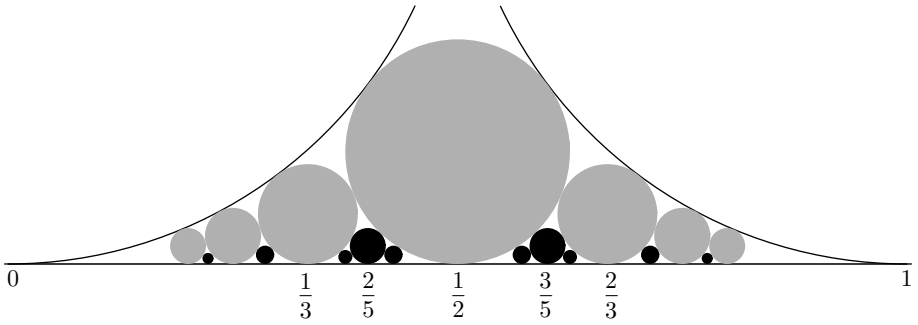

FIGURE 14. Ford discs

5.4. **Adler-Flatto's approach.** In [AF1, AF2], the following parametrization of $S\mathcal{H}$ is considered. Any $v = (z, \zeta) \in S\mathcal{H}$ defines a unique oriented geodesic in $\mathcal{H}$ with end points u and w. If s is the hyperbolic distance along this geodesic from some conveniently chosen initial point, then (u, w, s) gives another system of coordinates on $S\mathcal{H}$, in which the geodesic flow has a particularly simple form:

$$\varphi^t : (u, w, s) \mapsto (u, w, s + t). \tag{5.3}$$

The authors consider the cross-section of the geodesic flow on the modular surface given by the points on the boundary of the fundamental region F together will all unit vectors pointing to the exterior of F. This cross-section given in coordinates u, w corresponds exactly to the Morse code. However, it is not as easily expressed as the arithmetic cross-sections described in §3.2; the authors call it curvilinear. The first return map is given by fractional linear transformations. A change of coordinates is performed such that the domain becomes rectilinear. The new return map is conjugate to the original one by a one-to-one map which coincides with the identity map on much of its domain and has a simple geometric interpretation on regions where it is not the identity.

The authors use the new first return map to retrieve a version of Artin's coding which possesses a simple Markov partition. The Gauss map appears as a factor map of the new return map, from which the formula for the Gauss measure (invariant under the Gauss map) is easily obtained. Also, ergodic properties of the cross-section map and, therefore, of the geodesic flow can be derived from the ergodic properties of the Gauss map (see also 6.3).

A similar approach has been developed by the same authors in [AF4] for coding the geodesic flow on compact surfaces of constant negative curvature and genus $g \geq 2$ using a particular $8g - 4$-sided fundamental polygon.

5.5. **Arnoux's coding.** Using the algebraic definition of the geodesic flow (1.1) on the modular surface M and an explicit fundamental region, Arnoux ([Arn]) describes a coding method by regular continued fractions. The Gauss map appears as a factor map of the return map to the cross-section. More precisely, let $\{x\} = x - \lfloor x \rfloor$ be the fractional part of x, and

$$T : (0, 1) \to (0, 1), T(x) = \{1/x\}$$

be the continued fraction transformation (Gauss map). Let

$$\bar{T} : (0, 1)^2 \to (0, 1)^2 \ , \ \bar{T}(x, y) = (\{1/x\}, 1/(y + \lfloor 1/x \rfloor)).$$

The map $\bar{T}$ is an almost everywhere bijective extension of T; it is continuous on the rectangles $1/(n+1) < x < 1/n$, which are sent to $1/(n+1) < y < 1/n$. This gives a natural Markov partition and a symbolic Markov coding: to any pair (x, y) of irrational numbers in $(0, 1)$, one can associate a bi-infinite sequence of positive integers (a_n), where $x = \lfloor 0, a_1, a_2, \dots \rfloor$ and $y = \lfloor 0, a_0, a_{-1}, \dots \rfloor$ (here $\lfloor \cdot \rfloor$ denotes the regular continued fraction expansion).

Viewing the geodesic flow on the modular surface algebraically as the right action on $PSL(2, \mathbb{Z})\backslash PSL(2, \mathbb{R})$ of the group of diagonal matrices

$$a_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix},$$

Arnoux constructs a particular fundamental region Ω for this action and a cross-section $\Sigma \subset \partial\Omega$ (which can be identified under appropriate coordinates with $(0, 1)^2 \times \{0, 1\}$). He obtains an explicit formula for the first return map of the geodesic flow associated to this cross-section:

$$R : \Sigma \to \Sigma\ ,\ R(x, y, \epsilon) = (\bar{T}(x, y), 1 - \epsilon) = (\{1/x\}, 1/(y + \lfloor 1/x \rfloor), 1 - \epsilon)\,.$$

Any point in the fundamental region Ω can be written as $\varphi^t(x, y, \epsilon)$, with $(x, y, \epsilon) \in \Sigma$ and $0 \le t \le -2 \log x$. Thus, the transformation that associates to a point $\varphi^t(x, y, \epsilon)$ the point $((a_n), \epsilon, t)$ where (a_n) is the symbolic coding of the pair (x, y) conjugates the geodesic flow to a special flow $\{\psi^t\}$ defined over $\mathcal{N}^{\mathbb{Z}} \times \{0, 1\}$ with the return time $-2 \log \lfloor 0, a_1, a_2, \dots \rfloor$.

Using these computations, the author gives a short proof of Lévy's formula [Le]: for almost every real number x

$$\lim_{n\to\infty} \frac{\log q_n}{n} = \frac{\pi^2}{12 \log 2}$$

where $p_n/q_n = \lfloor 0, a_1, \dots, a_n \rfloor$ are the convergents of order n of the continued fraction expansion of x.

Arnoux's work is directly related to the modular surface being the Riemann moduli space of genus one curves. His approach is inspired by Veech's "zippered rectangles" [V] for visualizing the Teichmüller geodesic flow. This led Arnoux to the explicit fundamental region described above and the calculations presented.

6. Applications of arithmetic codes

6.1. **Geodesic flow as a special flow.** In §2.3 and §3.4 we have constructed four continuous surjective coding maps. The map $\mathfrak{C} : X \to B$ for the geometric code and the map $\mathfrak{C}_H : X_H \to C_H$ (for the H-code) are essentially one-to-one (and finite-to-one everywhere), while the maps for the other two arithmetic codes, $\mathfrak{C}_\alpha : X_\alpha \to C_\alpha$ ($\alpha = G, A$), are bijections. In all cases the first return to the cross-section corresponds to the left-shift of the coding sequence. This provides four symbolic representations of the geodesic flow $\{\varphi^t\}$ on SM as a special flow over (Λ, σ), where $\Lambda = X_G, X_A, X_H, X$, with the ceiling function f being the time of the first return to the cross-section $C = C_G, C_A, C_H, B$, i.e. four symbolic representations of the geodesic flow on the space

$$\Lambda^f = \{(x, y) : x \in \Lambda, 0 \le y \le f(x)\} \tag{6.1}$$

as explained in the Introduction.

Calculation of the return time. For $\Lambda = X_G, X_A, X_H, X$ and $C = C_G, C_A, C_H, B$, respectively, the ceiling function $f(x)$ on Λ is the time of the first return of the geodesic $\gamma(x)$ to the cross-section C. The following theorem was proved in [GuK] for the G-code, and appeared for other arithmetic codes in [KU2], and for the geometric code in [KU1]. The proof for all codes is the same. A similar formula for Artin's original code appeared earlier in [S3].

Theorem 6.1. *Let $x \in \Lambda$ and $w(x)$, $u(x)$ be the end points of the corresponding geodesic $\gamma(x)$. Then*

$$f(x) = 2\log|w(x)| + \log g(x) - \log g(\sigma x)$$

where

$$g(x) = \frac{|w(x) - u(x)|\sqrt{w(x)^2 - 1}}{w(x)^2\sqrt{1 - u(x)^2}}\,.$$

6.2. **Factor-maps associated with arithmetic codes and invariant measure on cross-sections.** Let $\alpha = G, A, H$ and $R : C_\alpha \to C_\alpha$ be the first return map. Let $p_\alpha : C_\alpha \to I_\alpha$ be a map from the cross-section to the interval I_α defined as follows: for $v \in C_\alpha$, $p_\alpha(v) = \frac{1}{w}$, where $w = (n_0, n_1, n_2, \dots)$ is the attracting end point of the α-reduced geodesic defined by v. The factor-map $f_\alpha : I_\alpha \to I_\alpha$ is such that the diagram

$$\begin{array}{ccc} C_\alpha & \xrightarrow{R} & C_\alpha \\ {\scriptstyle p_\alpha}\downarrow & & \downarrow{\scriptstyle p_\alpha} \\ I_\alpha & \xrightarrow{f_\alpha} & I_\alpha \end{array}$$

is commutative. We derive the formulas for the factor-map for all three codes. If $x = \frac{1}{w}$, then $f_\alpha(x) = \frac{1}{w'}$, where w' is the attracting point corresponding to the geodesic defined by $R(v)$. Since the first return to the cross-section corresponds to the left shift of the coding sequence, we have $w' = ST^{-n_0}w$. Hence

$$\frac{1}{w'} = n_0 - w = \left(\frac{1}{x}\right) - \frac{1}{x}\,.$$

In order to calculate the invariant measure for the map f_α we use the parametrization of $S\mathcal{H}$ by (u, w, t) considered in [AF1] and described in §5.4. The measure dm in these coordinates is given by the formula

$$dm = \frac{du\,dw\,dt}{(w-u)^2}\,, \tag{6.2}$$

and its invariance under $\{\varphi^t\}$ follows immediately from (5.3) and (6.2). The measure on the cross-section C_α invariant for the first return map is obtained by dropping dt: $dm_{C_\alpha} = \frac{du\,dw}{(w-u)^2}$, and the invariant measure on I_α is obtained by integrating dm_{C_α} with respect to du as explained in [AF1].

G-code. In this case $I_G = (0, 1)$, and

$$f_G(x) = \left\lceil \frac{1}{x} \right\rceil - \frac{1}{x} = \left\lfloor \frac{1}{x} \right\rfloor + 1 - \frac{1}{x} = 1 - \left\{ \frac{1}{x} \right\}\,.$$

In order to compute the invariant measure on I_G we first integrate dm_{C_G} over $I_G = (0,1)$:

$$\int_0^1 \frac{du\,dw}{(w-u)^2} = dw \frac{1}{w-u}\Big|_0^1 = dw\left(\frac{1}{w-1} - \frac{1}{w}\right) = \frac{dw}{w(w-1)}.$$

Going back to the variable $x = \frac{1}{w}$, we obtain the invariant measure on $I_G = (0,1)$

$$\mu_G = \frac{dx}{1-x}.$$

(See also [AF3] for a similar computation.)

***A*-code.** In this case $I_A = (-1,1)$,

$$f_A = \left\lceil \frac{1}{x} \right\rfloor - \frac{1}{x} = \begin{cases} -\{\frac{1}{x}\} \text{ if } x > 0 \\ \{-\frac{1}{x}\} \text{ if } x < 0 \end{cases}$$

and the invariant measure is $\mu_A = \dfrac{2dx}{(1+x)(1-x)}$.

***H*-code.** In this case $I_H = (-\frac{1}{2}, \frac{1}{2})$, and $f_H = \langle \frac{1}{x} \rangle - \frac{1}{x} = \frac{1}{2} - \{\frac{1}{x} + \frac{1}{2}\}$. The invariant measure is

$$\mu_H = \frac{4dx}{(2+x)(2-x)}.$$

6.3. **Classical results proved using arithmetic codes.** Artin [Ar] used regular continued fractions to prove the topological transitivity of the geodesic flow on the modular surface (i.e. the existence of a closed geodesic) and the density of closed geodesics. In fact, any arithmetic α-code ($\alpha = G, A, H$) can be used for this purpose since the Markov property allows us to list all admissible periodic coding sequences.

Proposition 6.2. *The set of closed geodesics is dense in SM, and the geodesic flow is topologically transitive.*

Proof. Recall that closed geodesics correspond to periodic α-codes. Then by Proposition 3.6 it is sufficient to find a periodic coding sequence arbitrarily close to a given coding sequence $x = (n_i)_{i\in\mathbb{Z}} \in X_\alpha$. Clearly, for any positive integer m, a periodic sequence $x_m = (\overline{n_{-m}, \dots, n_0, \dots, n_m})$ satisfies $d(x, x_m) < \frac{1}{m}$, so x_m will be arbitrarily close to x for large enough m. In order to prove topological transitivity we construct a coding sequence $\bar{x}$ which incorporates all finite codes in X_α (we can order them and write one after the other in a sequence). Then for any $\epsilon > 0$ there is $N \in \mathbb{Z}$ such that

$$d(\sigma^N \bar{x}, x) < \epsilon.$$

But this means that there is $t^* \in \mathbb{R}$ such that $\varphi^{t^*}(v(\bar{x}))$ is ϵ-close to $v(x)$, which completes the proof. □

As mentioned in the introduction, Hedlund [He2] gives a proof of the ergodicity of the geodesic flow on the modular surface by coding geodesics using regular continued fractions. Adler and Flato [AF1] prove the ergodicity of the geodesic flow using the arithmetic code described in §5.4 and reducing the problem to ergodic properties of the Gauss map, which appears as a natural one-dimensional factor map of the corresponding return map.

6.4. **Asymptotics and limit distribution of closed geodesics.** The study of the asymptotic growth of periodic orbits and their limit distribution plays an important role in the theory of dynamical systems. We restrict our discussion to the geodesic flow on surfaces of constant negative curvature, although many of the results we mention have been extended to the variable curvature case (at least for compact manifolds). Traditionally, the analysis of closed geodesics on manifolds of constant negative curvature is based on the study of the Selberg trace formula. If one denotes by $P(t)$ the number of periodic geodesics of period $\leq t$ on a surface of constant negative curvature, Huber [Hub] (for the compact case) and Sarnak [Sa] (for the general situation) proved that $P(t) \sim e^t/t$.

Concerning the limit distribution of the closed geodesics, Bowen [Bo] proved that the closed geodesics on compact manifolds of constant negative curvature are uniformly distributed with respect to the Liouville measure. To be more precise, if m_t is the probability measure formed by equidistributing Liouville measure on closed geodesics of length $\leq t$, then m_t converges to the Liouville measure (in the weak topology) as t goes to infinity. Pollicott [P] proved such a limit distribution result for the modular surface. The proof uses the arithmetic coding via regular continued fractions described by Series in [S3], an extension of the thermodynamic formalism for the continued fraction transformation obtained by Mayer [May1], and Tauberian theorems. The author also uses this result to study the distribution of quadratic irrationals.

6.5. **Estimates of the topological entropy.** Now we explain how to obtain estimates of the topological entropy of the geodesic flow restricted to certain flow-invariant subsets of SM.

We consider the following general situation presented in [GuK]. Let (Λ, σ) be a symbolic dynamical system and $L \subset \Lambda$ be a σ-invariant Borel subset of Λ. Given a Borel measurable function $g : L \to \mathbb{R}$ such that $\inf_{x \in L} g(x) > 0$, one can define a special flow $\{\psi^t\} = (L, g)$ on the space

$$L^g = \{(x, y) : x \in L,\ 0 \leq y \leq g(x)\}$$

much as the special flow defined in §6.1.

Let $\tilde{\mu}$ be an arbitrary $\{\psi^t\}$-invariant Borel probability measure on L^g and μ' its projection onto L. The sets $C_x = \{x\} \times \Delta_x,\ x \in L$, where $\Delta_x = \{y : 0 \leq y \leq g(x)\}$, constitute a measurable partition of L^g. The $\{\psi^t\}$-invariance of $\tilde{\mu}$ implies that the conditional measure on C_x induced by $\tilde{\mu}$ is the normalized Lebesgue measure on Δ_x for μ'-almost all x (here we identify C_x and Δ_x). By definition the function $x \mapsto 1/g(x)$ is bounded and hence μ'-integrable. So we can introduce a measure μ on L by

$$\mu(dx) = K \frac{\mu'(dx)}{g(x)}, \quad \text{where } K = \left[\int_L (1/g(x)) \mu'(dx) \right]^{-1}.$$

It is easy to check that $K = \int_L g d\mu$, μ is a probability measure, and $\tilde{\mu}$ is the restriction to L^g of the direct product $\mu \times \ell$, divided by K, where ℓ is the Lebesgue measure on $\mathbb{R}$. Moreover, μ is σ-invariant.

Conversely, given a σ-invariant probability measure μ on L such that $\int_L g d\mu < \infty$, one can define $\tilde{\mu}$ as above and make sure that $\tilde{\mu}$ is a $\{\psi^t\}$-invariant Borel probability measure on L^g. Thus we have a one-to-one correspondence between the set $I_g(L)$ of σ-invariant probability measures on L under which g is integrable and the set $I(L^g)$ of all $\{\psi^t\}$-invariant probability measures on L^g.

For each measure $\mu \in I_g(L)$ we denote by h_μ the measure-theoretic entropy of σ with respect to μ. The entropy of the flow $\{\psi^t\}$ with respect to the measure $\tilde{\mu}$ will be denoted by $h_{\tilde{\mu}}(\{\psi^t\})$. Recall that by definition $h_{\tilde{\mu}}(\{\psi^t\}) = h_{\tilde{\mu}}(\psi^1)$ and that by Abramov's formula [Ab] $h_{\tilde{\mu}}(\{\psi^t\}) = h_\mu / \int_L g d\mu$.

Under the definition adopted in [GuK], the topological entropy $h(\cdot)$ is the supremum of measure-theoretical entropies over the set of all flow-invariant Borel probability measures and hence is invariant with respect to a continuous conjugacy (and even a Borel measurable conjugacy) of dynamical systems.

Hence the topological entropy is defined by the formula

$$h(\{\psi^t\}) = \sup_{\mu \in I_g(L)} h_\mu \left(\int_L g d\mu \right)^{-1} \tag{6.3}$$

and has the following properties: Let $g_1 \geq g_2$ on L, and let $\{\psi_i^t\} = (L, g_i)$, $i = 1, 2$. Then by (6.3), $h(\{\psi_1^t\}) \leq h(\{\psi_2^t\})$. If two ceiling functions g_1 and g_2 are cohomologous, i.e. there exists a Borel measurable function $h : L \to \mathbb{R}$ such that $g_1(x) = g_2(x) + h(x) - h(\sigma(x))$, then the special flows (L, g_1) and (L, g_2) are conjugate [PaP] and, therefore, have the same topological entropy.

The first example of the special flow of the type described above was studied in [GuK]: the special flow over $L = X_P \subset X_G$, the space of positive coding sequences, with the ceiling function $f(x) = 2 \log w(x)$. Let Σ^+ be the subset of the unit tangent bundle SM consisting of vectors tangent to positive geodesics. Since the function f is cohomologous to the time of the first return to the cross-section C_G (Theorem 6.1), $h(\{\varphi^t_{|\Sigma^+}\}) = h(\{\phi^t\})$, where $\{\phi^t\}$ is the special flow over X_P with the ceiling function $f(x) = 2 \log w(x)$. The following two-sided estimates were obtained in [GuK]:

Theorem 6.3. $0.7771 < h(\{\varphi^t_{|\Sigma^+}\}) < 0.8161$.

Sketch of proof. The function $f(x) = 2 \log w(x)$ is well-defined on the whole space X_G, since $w(x) \geq 1$ for all $x \in X_G$. Thus, $\{\phi^t\}$ is a subflow of the special flow (X_G, f). For every $x = (n_i)_{i \in \mathbb{Z}} \in X_G$ let $n_i(x)$ denote n_i. If $x \in X_P$, then $n_i(x) \geq 3$ and it is easy to show that

$$2 \log c n_0(x) \leq f(x) \leq 2 \log n_0(x), \text{ where } c = (3 + \sqrt{5})/6 \approx 0.8726. \tag{6.4}$$

Thus the ceiling function is estimated by two functions which depend only on the first coordinate $n_0(x)$ of $w(x)$. We can now use a formula for the topological entropy developed by Polyakov [Po] based on a result of Savchenko [Sav]. The method requires the countable Markov chain to be a local perturbation of the full Bernoulli shift (i.e. the number of forbidden transitions must be finite) and the first return time function $f(x)$ to depend only on the first coordinate $n_0(x)$. For (X_P, g_δ), with $g_\delta(x) = 2 \log \delta n_0(x)$ ($\delta = 1$ and $\delta = c$), we obtain the estimates

$$h_1 = 0.7771 < h(\{\varphi^t_{|\Sigma^+}\}) < 0.8161 = h_c.$$

The estimated values h_δ are solutions of the equation $\Psi_\delta(s) = 1$, where

$$\Psi_\delta(s) = \frac{G(s)(1 + (3\delta)^{-2s} - (12\delta)^{-2s} - (15\delta)^{-2s})}{1 - (4\delta)^{-2s} - (5\delta)^{-2s}},$$

and $G(s)$ is related to the Riemann ζ-function by the formula

$$G(s) = \delta^{-2s}\left(\zeta(-2s) - \sum_{n=1}^{5} n^{-2s}\right).$$

These values were obtained with the help of the computer package Pari-GP. □

The second example was studied in [KU1]. Let Σ be the subset of the unit tangent bundle SM, consisting of vectors tangent to geometrically Markov geodesics, i.e. geodesics whose codes are in X_M (see §4.1). The set Σ is flow invariant and noncompact. Let $\{\varphi^t_{|\Sigma}\}$ be the restriction of the geodesic flow to Σ. The following theorem [KU1] gives a lower bound estimate for $h(\{\varphi^t_{|\Sigma}\})$, the topological entropy of the flow $\{\varphi^t_{|\Sigma}\}$.

Theorem 6.4. $0.8417 < h(\{\varphi^t_{|\Sigma}\})$.

The proof of this estimate follows the same scheme as in the previous theorem, but the extent of that method allows us to obtain an estimate only from below. Of course, since $h(\{\varphi^t\}) = 1$ (see [GuK]), we have a trivial estimate from above.

7. Arithmetic coding beyond the modular surface

7.1. **Boundary expansions.** In [S2, S4] Series made an explicit geometric construction of symbolic dynamics for the geodesic flow on surfaces of constant negative curvature and finite hyperbolic area. One of the main results is that the geodesic flow on a compact surface can be represented as a factor of a special flow over a topological Markov chain (with finite alphabet given by the generators of the fundamental group Γ of the surface) by a continuous map which is one-to-one except on a set of the first Baire category. The symbolic dynamics is derived from the author's earlier work with Bowen [BoS] in which the action of Γ on the boundary of the unit disc $\partial\mathcal{U}$ was shown to be orbit equivalent to a certain Markov map f_Γ used to develop the boundary expansion code geometrically. The map f_Γ is piecewise equal to the generating transformations of Γ that identify the sides of the fundamental region $\mathcal{D}$ and produces a bi-infinite sequence of generators of Γ now called the *Bowen-Series boundary expansion code*. This construction is a generalization of Nielsen's boundary expansion [N] for a surface whose fundamental region is a symmetric $4g$-sided polygon in the unit disc $\mathcal{U}$. In the presence of cusps one still obtains a Markov map, but with a countable alphabet.

Series' results apply to a general class of surfaces (which, however, does not include the modular surface with the standard fundamental domain F) and are obtained by considering specially chosen fundamental regions with *even corners* (this means that $\Gamma(\partial D)$ consists of complete geodesics in $\mathcal{U}$). A precise relation can be established in this case between the Morse code and the boundary expansion code. Series shows that cutting sequences corresponding to geodesics on a closed surface could be modified systematically to sequences which form a sofic shift (i.e. a factor of a finite-step topological Markov chain) [S4, Lemma 4.1], so that every admissible sequence corresponds to a geodesic. Thus the existence of geodesics with certain dynamical properties could be established as in earlier work of Artin and Nielsen simply by producing admissible sequences of the required kind (see also §6.3).

The arithmetic codings for the modular surface considered in Section 3 still can be viewed as boundary expansions by properly partitioning the real axis into three intervals labeled by T, T^{-1}, and S. Taking a similar approach, one can develop a reduction theory and construct arithmetic codings via continued fraction expansions for other Fuchsian groups, in particular for Hecke triangle groups, Γ_q generated by $T_q(z) = z + \lambda_q$ (where $\lambda_q = 2\cos(\pi/q)$) and $S(z) = -1/z$ for $q = 3, 4, \dots$ (work in progress). Various forms of continued fractions have been developed for such groups. To mention a few: Rosen [Ro] introduced the so-called λ_q-continued fractions to study the elements of such groups, Rosen-Schmidt [RoSc] described closed geodesics on $\Gamma_q\backslash\mathcal{H}$ and Schmidt-Sheingorn [ScSh] studied the length spectra. We refer the reader to the latter paper, which contains extensive references on the subject of Hecke groups. Let us also mention Fried's important work [Fr1], where the author develops an arithmetic coding for a larger class of nonuniform hyperbolic triangle groups. The geodesic flow on the quotients of the hyperbolic plane by such groups is symbolically coded using a generalization of Artin's continued fractions method. This allows the author to define transfer operators and to study dynamical zeta functions for such groups as done by Mayer [May2] for the modular group. Fried calculates the invariant measures for the corresponding factor-maps that generalize the Gauss measure for the continued fraction map.

In the simplest case, where the fundamental region for $\Gamma\backslash\mathcal{U}$ has no vertices in $\mathcal{U}$, the Bowen-Series and the Morse codes coincide. One such example, a three-holed sphere (the compact part of a hyperbolic surface with three infinite funnels) was studied in [S4]. In general, the codes differ if the group Γ is not free. The discrepancy is closely related to the possible different ways of representing elements of Γ as shortest words in a given set of generators.

In what follows, as an example, we describe the Bowen-Series boundary expansion code for the free group $\Gamma(2)$ with a specially chosen fundamental region.

7.2. **The congruence subgroup $\Gamma(2)$.** Consider the surface $M_2 = \Gamma(2)\backslash\mathcal{H}$ where $\Gamma(2)$ is the principal congruence subgroup of level 2,

$$\Gamma(2) = \left\{ g \in PSL(2,\mathbb{Z}) \mid g \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod 2 \right\}.$$

Notice that Γ is a subgroup of $PSL(2,\mathbb{Z})$ of index 6. Moreover it is a free group on two generators given by

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \ (A(z) = z + 2) \text{ and } B = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix} \ (B(z) = z/(-2z+1)).$$

A fundamental region F_2 for M_2 is bounded by the vertical lines $x = \pm 1$ and two semi-circles $(x \pm 1/2)^2 + y^2 = 1/4$. The identification of sides is given by the parabolic transformations A and B fixing the cusps at ∞ and 0, respectively. The $\Gamma(2)$-equivalent points -1 and 1 represent the third cusp of M_2.

7.3. **Morse coding for $\Gamma(2)$.** The Morse code with respect to the fundamental region F_2 (Figure 15) can be assigned to any geodesic γ on F_2 which does not go to any of the three cusps of F_2 in either direction. It is easy to see that the images of the cusp at ∞ under $\Gamma(2)$ are rational numbers $\frac{p}{q}$ with p odd and q even, the images of the cusp at 0 are rational numbers $\frac{p}{q}$ with p even and q odd, and the images of the cusp at 1 are rational numbers $\frac{p}{q}$ with both p and q odd. Thus we consider only geodesics whose lifts to $\mathcal{H}$ have irrational end points.

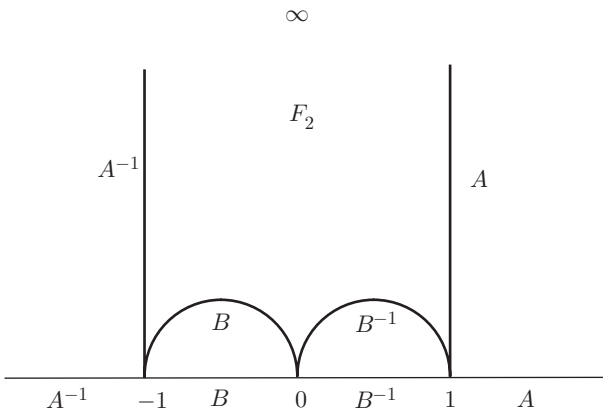

FIGURE 15. The fundamental region for M_2

We label the sides of F_2 (on the outside) as shown on Figure 15: the left vertical by A^{-1}, the right vertical by A, the left circular by B, and the right circular by B^{-1}. We consider as a starting segment of an oriented geodesic γ on M_2 a segment whose initial end point is on one of the semi-circles and whose final end point is on one of the vertical sides.

If we start coding an oriented geodesic from a segment whose initial point is on one of the semi-circles and whose final point is on one of the vertical sides, then the Morse code can be written as

$$[\gamma] = \dots A^{n_{-2}} B^{n_{-1}} A^{n_0} B^{n_1} A^{n_2} \dots$$

or equivalently, in numerical notation, as a bi-infinite sequence of nonzero integers $([\dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots], 0)$ where the additional marker symbol 0 denotes the fact that the coding sequence starts with A^{n_0}. Similarly, if we start coding a geodesic from a segment with initial point on one of the vertical sides and final point on one of the semi-circles, then the Morse code can be written as

$$[\gamma] = \dots B^{n_{-2}} A^{n_{-1}} B^{n_0} A^{n_1} B^{n_2} \dots$$

or, equivalently, as $([\dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots], 1)$ (the additional marker symbol 1 is being used to denote the fact that the coding sequence starts now with B^{n_0}).

Thus, the set of all numerical coding sequences is a subspace of the symbolic space $\mathcal{N}^{\mathbb{Z}} \times \{0, 1\}$, where $\mathcal{N} = \{n \in \mathbb{Z}, n \neq 0\}$, and the shift map σ is defined as

$$\sigma([n_i], \epsilon) = ([n_{i+1}], 1 - \epsilon).$$

We will show (Corollary 7.5) that this set is the entire symbolic space $\mathcal{N}^{\mathbb{Z}} \times \{0, 1\}$, except for the countable set of sequences having a tail of 1's or -1's. (For a geometric proof see [St].)

7.4. **Boundary expansions for $\Gamma(2)$.** The Bowen-Series boundary expansion can be easily translated to the upper half-plane model $\mathcal{H}$. We define a map

$f_{\Gamma(2)} : \mathbb{R} \cup \{\infty\} \to \mathbb{R} \cup \{\infty\}$ by

$$f_{\Gamma(2)} = \begin{cases} A(x) & \text{if } x \in [-\infty, -1] \\ B^{-1}(x) & \text{if } x \in [-1, 0] \\ B(x) & \text{if } x \in [0, 1] \\ A^{-1}(x) & \text{if } x \in [1, \infty] \end{cases}$$

and label the elements of the partition of $\mathbb{R} \cup \{\infty\}$ as shown on Figure 15: $[-\infty, -1]$ is labeled by A^{-1}, $[-1, 0]$ is labeled by B, $[0, 1]$ is labeled by B^{-1}, and $[1, \infty]$ is labeled by A. Let $\gamma(u, w)$ be a geodesic in $\mathcal{H}$ with repelling end point u and attracting end point w, intersecting F_2. The boundary expansion of w is the sequence $(w_0, w_1, w_2, \dots)$, where w_n is the label of the segment to which $f^n_{\Gamma(2)}(w)$ belongs ($w_n \in \{A, A^{-1}, B, B^{-1}\}$). The boundary expansion of u is the sequence $(u_0, u_1, u_2, \dots)$, where u_n is the label of the segment to which $f^n_{\Gamma(2)}(u)$ belongs ($u_n \in \{A, A^{-1}, B, B^{-1}\}$). Let $\bar{u}_n$ denote the inverse of u_n. Following [S4] we represent the geodesic from $\gamma(u, w)$ by a bi-infinite sequence

$$u * w = (\dots, \bar{u}_3, \bar{u}_2, \bar{u}_1, \bar{u}_0, w_0, w_1, w_3, \dots)$$

called the *Bowen-Series boundary expansion code*.

Notice that because of the particular type of the chosen fundamental region (without vertices in $\mathcal{H}$), the boundary expansion code coincides with the Morse code for any geodesic in M_2.

7.5. **Arithmetic coding for $\Gamma(2)$ via even continued fractions.** In his 1877 work, H. J. S. Smith [Sm] used the nearest even integer continued fractions to develop a reduction theory for integral indefinite binary quadratic forms with respect to the congruence subgroup $\Gamma(2)$ much as Dirichlet used regular continued fractions for $GL(2, \mathbb{Z})$- and $SL(2, \mathbb{Z})$-reduction theory 23 years earlier (see §3.1). It is interesting to remark that Smith's work came 12 years prior to Hurwitz's work [H1] on the $SL(2, \mathbb{Z})$-reduction theory using the nearest integer continued fractions. Apparently, Hurwitz was not acquainted with Smith's work.

We describe a method of constructing an arithmetic code for geodesics on $M_2 = \Gamma(2)\backslash\mathcal{H}$ based on Smith's reduction theory in a way similar to that described for $SL(2, \mathbb{Z})$-reduction in §3.2. Let us mention that a similar method was used by Kraikaamp and Lopes [KrLo] for studying the theta group (i.e. the discrete subgroup of $PSL(2, \mathbb{Z})$ generated by $z \mapsto z + 2$ and $z \mapsto -1/z$) and the geodesics on the corresponding surface.

Every irrational number x has a unique representation in the form

$$x = 2n_0 - \cfrac{1}{2n_1 - \cfrac{1}{2n_2 - \cfrac{1}{\ddots}}}$$

which we call the *even continued fraction expansion* (or *E-expansion*) and denote by $(\!(2n_0, 2n_1, \dots)\!)$ for short; $2n_0$ is the integer equal to $(\!(x)\!)$, where $(\!(x)\!)$ is the nearest even integer to x, and the nonzero integers n_i ($i \geq 1$) are determined recursively by $2n_i = (\!(x_i)\!)$, $x_{i+1} = -\frac{1}{x_i - 2n_i}$, starting with $x_1 = -\frac{1}{x - 2n_0}$. Conversely, any infinite sequence of integers $n_0, n_1, n_2, \dots$ with $n_i \neq 0$ if $i \geq 1$, and not having a tail of 1's or -1's, defines an irrational number whose E-expansion is $(\!(2n_0, 2n_1, \dots)\!)$.

The following properties are satisfied:

(1) Two irrational numbers x and y are $\Gamma(2)$-equivalent $\Longleftrightarrow$ their E-expansions have the same tail.
(2) x is a quadratic irrationality $\Longleftrightarrow (\!(2n_0, 2n_1, \ldots)\!)$ is eventually periodic.
(3) Let x and x' be conjugate quadratic irrationalities, i.e. the roots of the same quadratic polynomial with integer coefficients. For any quadratic irrationality x with purely periodic expansion $x = (\!(\overline{2n_0, 2n_1, \ldots, 2n_k})\!)$, the expansion of $\frac{1}{x'}$ is also purely periodic and $\frac{1}{x'} = (\!(\overline{2n_k, \ldots, 2n_1, 2n_0})\!)$.
(4) A quadratic irrationality x has a purely periodic E-expansion if and only if $|x| > 1$ and $|x'| < 1$, where x' is conjugate to x.

Definition 7.1. An oriented geodesic on $\mathcal{H}$ is called *E-reduced* if its repelling and attracting end points, denoted by u and w, respectively, satisfy $|w| > 1$ and $|u| < 1$ or $|w| < 1$ and $|u| > 1$.

Reduction algorithm. Let γ be an arbitrary geodesic on $\mathcal{H}$, with end points u and w, and $w = (\!(2n_0, 2n_1, 2n_2, \ldots)\!)$. We construct the following sequence of real pairs $\{(u_k, w_k)\}$ $(k \geq 0)$ defined by $u_0 = u$, $w_0 = w$ and:

$$w_{2k} = B^{-n_{2k-1}} A^{-n_{2k-2}} \ldots B^{-n_1} A^{-n_0} w\,,\ u_{2k} = B^{-n_{2k-1}} T^{-n_{2k-2}} \ldots B^{-n_1} A^{-n_0} u\,;$$
$$w_{2k+1} = A^{-n_{2k}} w_{2k}\,,\ u_{2k+1} = A^{-n_{2k}} u_{2k}\,.$$

Each geodesic with end points u_k and w_k is $\Gamma(2)$-equivalent to γ by construction.

Theorem 7.2. *The above algorithm produces in finitely many steps an E-reduced geodesic $\Gamma(2)$-equivalent to γ; i.e. there exists a positive integer ℓ such that the geodesic with end points u_ℓ and w_ℓ is reduced.*

To such a reduced geodesic γ we can associate the following code:

- if $|u| < 1$ and $|w| > 1$, then

$$(\!(\gamma)\!) := ((\ldots, n_{-2}, n_{-1}, n_0, n_1, n_2, \ldots), 0)\,,$$

  where $1/u = (\!(2n_{-1}, 2n_{-2}, \ldots)\!)$ and $w = (\!(2n_0, 2n_1, 2n_2, \ldots)\!)$;
- if $|u| > 1$ and $|w| < 1$, then

$$(\!(\gamma)\!) := ((\ldots, n_{-2}, n_{-1}, n_0, n_1, n_2, \ldots), 1)\,,$$

  where $-1/w = (\!(2n_0, 2n_1, \ldots)\!)$ and $-u = (\!(2n_{-1}, 2n_{-2}, \ldots)\!)$.

Remark 7.3. Any further application of the reduction algorithm to an E-reduced geodesic yields E-reduced geodesics whose codes are left shifts of the code of the initial E-reduced geodesic (together with an appropriate change of the marker symbol $\epsilon \in \{0, 1\}$).

The proof of Theorem 7.2 goes along the lines presented in the proof of [KU2, Theorem 1.3] for the $PSL(2, \mathbb{Z})$-reduction procedure using the G-code. In a situation similar to that described also in Section 3, we define the E-code of an oriented geodesic γ on $\mathcal{H}$ to be the E-code of a reduced geodesic $\Gamma(2)$-equivalent to γ and prove its $\Gamma(2)$-invariance by constructing a cross-section of the geodesic flow on M_2, directly related to the notion of E-reduced geodesics.

Construction of the cross-section. We describe the cross-section C_E for the geodesic flow on M_2 such that successive returns to the cross-section correspond to left-shifts in the arithmetic E-code. Let $C_E = P \cup Q$ be a subset of the unit tangent bundle SM_2, where P consists of all tangent vectors with base points in the circular sides of F_2 and pointing inward such that the corresponding geodesic is E-reduced (with $|u| < 1$ and $|w| > 1$); Q consists of all tangent vectors with base points on the vertical sides of F_2 pointing inward such that the corresponding geodesic is E-reduced (with $|u| > 1$ and $|w| < 1$).

One can show that C_E is indeed a cross-section for the geodesic flow on M_2; hence every geodesic γ can be represented as a bi-infinite sequence of segments σ_i between successive returns to C_E. To each segment σ_i is associated the corresponding E-reduced geodesic γ_i, so that $(\!(\gamma_{i+1})\!)$ differs from $(\!(\gamma_i)\!)$ by a left shift of the bi-infinite sequence and a switch between 0 and 1 symbols. Thus we associate to γ a bi-infinite coding sequence, defined up to a shift, which we call the *E-code* of γ and denote by $(\!(\gamma)\!)$. A similar argument as for the G-code shows that the E-code is $\Gamma(2)$-invariant.

Figure 16 shows the infinite partition of C_E (parameterized by $(u, 1/w)$) and its image under the first return map.

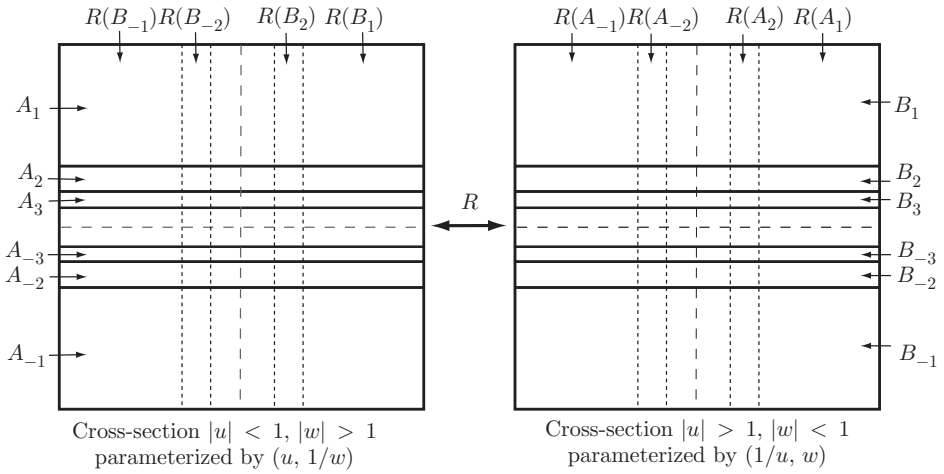

FIGURE 16. Infinite partition for the E-code and its image under the return map R

Symbolic representation of geodesics via E-code. Let $\mathcal{N}^{\mathbb{Z}}$ be the Bernoulli space on the infinite alphabet $\mathcal{N} = \{n \in \mathbb{Z}, n \neq 0\}$. We proved that each oriented geodesic which does not go to a cusp of M_2 in either direction admits a unique E-code, $(\!(\gamma)\!) \in \mathcal{N}^{\mathbb{Z}} \times \{0, 1\}$, which does not contain a tail of 1's or -1's. Taking the closure of the set of all such E-codes, we obtain the space $\mathcal{N}^{\mathbb{Z}} \times \{0, 1\}$. Each element $x = ((n_i), \epsilon) \in \mathcal{N}^{\mathbb{Z}} \times \{0, 1\}$ produces a geodesic on $\mathcal{H}$ from $u(x)$ to $w(x)$ where $w(x) = (\!(2n_0, 2n_1, \ldots)\!)$, $\frac{1}{u(x)} = (\!(2n_{-1}, 2n_{-2}, \ldots)\!)$ if $\epsilon = 0$, and $-\frac{1}{w(x)} = (\!(2n_0, 2n_1, \ldots)\!)$, $-u(x) = (\!(2n_{-1}, 2n_{-2}, \ldots)\!)$ if $\epsilon = 1$. Notice that if a sequence (n_i) has a tail of 1's or -1's, then the associated geodesic goes to the cusp at 1 in the corresponding direction. Thus, the set of oriented geodesics on M_2 which do not go to the cusps 0 and ∞ can be described by the symbolic space $X_E = \mathcal{N}^{\mathbb{Z}} \times \{0, 1\}$.

7.6. **Relation between the E-code and the geometric code.** We have already noticed that the Morse code and the boundary expansion code coincide for any geodesic γ in M_2. The following theorem establishes a similar property for the E-code and the geometric code.

Theorem 7.4. *For any geodesic in M_2 the E-code coincides with the geometric code.*

Proof. Take a geodesic γ on M_2 and suppose its E-code is $(\!(\gamma)\!) = ((n_i), 0)$; a similar argument works for $(\!(\gamma)\!) = ((n_i), 1)$. Its natural lift $\tilde{\gamma}$ to $\mathcal{H}$ has end points given by

$$w = (\!(2n_0, 2n_1, \ldots)\!)\,,\ \frac{1}{u} = (\!(2n_{-1}, 2n_{-2}, \ldots)\!)\,.$$

One needs to show that the geometric code of $\tilde{\gamma}$ (and therefore of γ) is given by $([n_i], 0)$. For that reason it is enough to see that, since the nearest even integer to w is $2n_0$, the geodesic $\tilde{\gamma}$ intersects the vertical side labeled by A and precisely the next $n_0 - 1$ consecutive images of it (in the case $n_0 > 0$), or $\tilde{\gamma}$ intersects the vertical side labeled by A^{-1} and precisely the next $|n_0| - 1$ consecutive images of it (in the case $n_0 < 0$). Therefore the first entry in the geometric code of γ is A^{n_0}.

Now we conjugate $\tilde{\gamma}$ by A^{-n_0} and look at the next coding sequence. The new geodesic $\tilde{\gamma}_1$ has end points given by $w_1 = w - 2n_0$, and $u_1 = u - 2n_0$. Notice that the mapping $S(z) = -1/z$ transfers the cusp at 0 into the cusp at ∞ and the boundary components labeled $B^{\pm 1}$ to the boundary components labeled $A^{\pm 1}$. For that reason, instead of tracing the behavior of the geodesic from u_1 to w_1, we can as well study the geodesic from $S(u_1) = -1/u_1$ to $S(w_1) = -1/w_1$. We have $-1/u_1 = 1/((2n_0, 2n_{-1}, \ldots))$ and $-1/w_1 = ((2n_1, 2n_2, \ldots))$. This brings us to the previously studied situation; hence the geodesic from $-1/u_1$ to $-1/w_1$ has its first coding sequence given by A^{n_1}. This implies that the first entry in the coding sequence of γ_1 is B^{n_1}. Continuing by induction we obtain that the geometric code of γ coincides with its E-code. □

Corollary 7.5. *The space of all geometric codes of geodesics on M_2 (not ending at a cusp) is the entire symbolic space $\mathcal{N}^{\mathbb{Z}} \times \{0, 1\}$, except for the countable set of sequences having a tail of 1's or -1's.*

Acknowledgments

We would like to thank Roy Adler, Pierre Arnoux, Boris Hasselblatt, Jeffrey Lagarias, Thomas Schmidt, Don Zagier, and an anonymous referee for useful comments and insights which helped us improve the exposition of the paper.

About the authors

Svetlana Katok is a professor at The Pennsylvania State University. She gave the 2004 AWM Emmy Noether Lecture at the Joint Mathematics Meetings in Phoenix. She is a recipient of the Eberly College of Science Alumni Society Distinguished Service Award.

Ilie Ugarcovici received a Ph.D. in mathematics from The Pennsylvania State University in 2004. He was a G. C. Evans Instructor at Rice University and currently is an assistant professor at DePaul University.

References

[Ab] L. M. Abramov, *On the entropy of a flow* (in Russian), Sov. Math. Doklady **128** (1959), no. 5, 873–875. MR0113985 (22:4816)

[A] R. Adler, *Symbolic dynamics and Markov partitions*, Bull. Amer. Math. Soc. **35** (1998), no. 1, 1–56. MR1477538 (98j:58081)

[AF1] R. Adler, L. Flatto, *Cross section maps for geodesic flows, I (The Modular surface)*, Birkhäuser, Progress in Mathematics (ed. A. Katok) (1982), 103–161. MR0670077 (84h:58113)

[AF2] R. Adler, L. Flatto, *Cross section map for geodesic flow on the modular surface*, Contemp. Math. **26** (1984), 9–23. MR0737384 (85j:58128)

[AF3] R. Adler, L. Flatto, *The backward continued fraction map and geodesic flow*, Ergod. Th. & Dynam. Sys. **4** (1984), 487–492. MR0779707 (86h:58116)

[AF4] R. Adler, L. Flatto, *Geodesic flows, interval maps, and symbolic dynamics*, Bull. Amer. Math. Soc. **25** (1991), no. 2, 229–334. MR1085823 (92b:58172)

[AW] R. Adler, B. Weiss, *Entropy, a complete metric invariant for automorphisms of the torus*, Proc. Nat. Acad. Sci. U.S.A. **57** (1967), 1573–1576. MR0212156 (35:3031)

[Arn] P. Arnoux, *Le codage du flot géodésique sur la surface modulaire*, Enseign. Math. **40** (1994), 29–48. MR1279059 (95c:58136)

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

[B] A. F. Beardon, *The Geometry of Discrete Groups*, Springer, New York, 1983. MR0698777 (85d:22026)

[BiS] J. Birman, C. Series, *Dehn's algorithm revisited, with applications to simple curves on surfaces*, in Combinatorial Group Theory and Topology (Alta, Utah, 1984), 451–478, Ann. of Math. Stud. 111, Princeton Univ. Press, 1987. MR0895628 (88k:20059)

[Bh] G. D. Birkhoff, *Nouvelles recherches sur les systèmes dynamique*, Memoriae Pont. Acad. Sci. Novi Lyncaei, s. 3, Vol. 1, 1935, pp. 85–216 (according to the collected works).

[Bo] R. Bowen, *The equidistribution of closed geodesics*, Amer. J. Math. **94** (1972), 413–423. MR0315742 (47:4291)

[BoS] R. Bowen, C. Series, *Markov maps associated with Fuchsian groups*, Inst. Hautes Études Sci. Publ. Math. No. 50 (1979), 153–170. MR0556585 (81b:58026)

[D] P. G. L. Dirichlet, *Vereinfachung der Theorie der binaren quadratischen Formen von positiver Determinante*, Abh. K. Akad. Wiss. Berlin Math. (1854), 99–115.

[Fo1] L. Ford, *A geometric proof of a theorem of Hurwitz*, Proc. Edinburgh Math. Soc. **35** (1917), 59–65.

[Fo2] L. Ford, *Rational approximations to irrational complex numbers*, Trans. Amer. Math. Soc. **19** (1918), no. 1, 1–42.

[Fo3] L. Ford, *Fractions*, Amer. Math. Monthly **45** (1938), no. 9, 586–601. MR1524411

[Fr1] D. Fried, *Symbolic dynamics for triangle groups*, Invent. Math. **125** (1996), no. 3, 487–521. MR1400315 (97g:58126)

[Fr2] D. Fried, *Reduction theory over quadratic imaginary fields*, J. Number Theory **110** (2005), no. 1, 44–74. MR2114673 (2005k:11139)

[Ga] C. F. Gauss, *Disquisitiones Arithmeticae*, 1801. English edition, Springer, New York, 1986. MR0837656 (87f:01105)

[GL] D. J. Grabiner, J. C. Lagarias, *Cutting sequences for geodesic flow on the modular surface and continued fractions*, Monatsh. Math. **133** (2001), no. 4, 295–339. MR1915877 (2003g:37051)

[GrLe] P. M. Gruber, C. G. Lekkerkerker, *Geometry of numbers*, North-Holland, 1987. MR0893813 (88j:11034)

[GuK] B. Gurevich, S. Katok, *Arithmetic coding and entropy for the positive geodesic flow on the modular surface*, Moscow Math. J. **1** (2001), no. 4, 569–582. MR1901076 (2003h:37040)

[Ha] J. Hadamard, *Les surfaces à courbures opposées et leurs lignes géodesiques*, J. Math. Pures Appl. (5) **4** (1898), 27–73.

[He1] G. A. Hedlund, *On the metrical transitivity of geodesics on closed surfaces of constant negative curvature*, Ann. Math. **35** (1934), 787–808. MR1503197

[He2] G. A. Hedlund, *A metrically transitive group defined by the modular group*, Amer. J. Math. **57** (1935), 668–678. MR1507102

[He3] G. A. Hedlund, *The dynamics of geodesic flows*, Bull. Amer. Math. Soc. **45** (1939), 241–260.

[Ho] E. Hopf, *Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung*, Ber. Verh. Sächs. Akad. Wiss. Leipzig **91** (1939), 261–304. MR0001464 (1:243a)

[Hub] H. Huber, *Zur analytischen Theorie hyperbolischen Raumformen und Bewegungsgruppen I*, Math. Ann. **138** (1959), 1–26; *II*, Math. Ann. **142** (1961), 385–398. MR0109212 (22:99)

[Hum] M. G. Humbert, *Sur les fractions continues et les formes quadratiques binaires indéfinies*, C. R. Acad. Sci. Paris **162** (1916), 23–26.

[H1] A. Hurwitz, *Über eine besondere Art der Kettenbruch-Entwicklung reeler Grossen*, Acta Math. **12** (1889) 367–405.

[H2] A. Hurwitz, *Über die angenaherte Darstellungen der Irrationalzahlen durch rationale Brüche*, Math. Ann. **39** (1891) 279–284.

[H3] A. Hurwitz, *Über die Reduktion der binaren quadratischen Formen*, Math. Ann. **45** (1894), 85–117.

[KH] A. Katok, B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, 1995. MR1326374 (96c:58055)

[K1] S. Katok, *Reduction theory for Fuchsian groups*, Math. Ann. **273** (1985), 461–470. MR0824433 (87h:11064)

[K2] S. Katok, *Fuchsian Groups*, University of Chicago Press, 1992. MR1177168 (93d:20088)

[K3] S. Katok, *Coding of closed geodesics after Gauss and Morse*, Geom. Dedicata **63** (1996), 123–145. MR1413625 (97j:20045)

[KU1] S. Katok, I. Ugarcovici, *Geometrically Markov geodesics on the modular surface*, Moscow Math. J. **5** (2005), 135–151. MR2153471

[KU2] S. Katok, I. Ugarcovici, *Arithmetic coding of geodesics on the modular surface via continued fractions*, 59–77, CWI Tract **135**, Math. Centrum, Centrum Wisk. Inform., Amsterdam, 2005. MR1901076 (2003h:37040)

[Ko] P. Koebe, *Riemannsche Mannigfaltigkeiten und nicht euklidische Raumformen*, Sitzungsberichte der Preußischen Akademie der Wissenschaften, *I* (1927), 164–196; *II, III* (1928), 345–442; *IV* (1929), 414–557; *V, VI* (1930), 304–364, 504–541; *VII* (1931), 506–534.

[KrLo] C. Kraaikamp, A. Lopes, *The theta group and the continued fraction expansion with even partial quotients*, Geom. Dedicata **59** (1996), no. 3, 293–333. MR1371228 (97g:58135)

[Le] P. Lévy, *Sur le développement en fraction continue d'un nombre choisi au hasard*, Compositio Math. **3** (1936), 286–303.

[Ma] A. A. Markoff, *Sur les formes quadratiques binaires indéfinies*, Math. Ann. **15** (1879), 381–406.

[May1] D. Mayer, *On a zeta function related to the continued fraction transformation*, Bull. Soc. Math. France **104** (1976), 195–203. MR0418168 (54:6210)

[May2] D. Mayer, *The thermodynamic formalism approach to Selberg's zeta function for* $\mathrm{PSL}(2, Z)$, Bull. Amer. Math. Soc. **25** (1991), no. 1, 55–60. MR1080004 (91j:58130)

[Mi] H. Minkowski, *Geometrie der Zahlen*, Chelsea Publishing Company, New York, 1953. MR0249269 (40:2515)

[M1] M. Morse, *A one-to-one representation of geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc. **22** (1921), 33–51.

[M2] M. Morse, *Recurrent geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc. **22** (1921), 84–100.

[Mo] R. Moeckel, *Geodesics on modular surface and continued fractions*, Ergod. Th. & Dynam. Sys. **2** (1982), 69–83. MR0684245 (84k:58176)

[N] J. Nielsen, *Untersuchungen zur Topologie der geschlossenen zweiseitigen Flächen*, Acta Math. **50** (1927), 189–358.

[O] D. Ornstein, *The isomorphism theorem for Bernoulli flows*, Advances in Math. **10** (1973), 124–142. MR0318452 (47:6999)

[OW] D. Ornstein, B. Weiss, *Geodesic flows are Bernoullian*, Israel J. Math. **14** (1973), 184–198. MR0325926 (48:4272)

[PaP] W. Parry, M. Pollicott, *Zeta functions and periodic orbit structure of hyperbolic dynamics*, Astérisque **187–188**, 1990. MR1085356 (92f:58141)

[P] M. Pollicott, *Distribution of closed geodesics on the modular surface and quadratic irrationals*, Bul. Soc. Math. France **114** (1986), 431–446. MR0882589 (88j:58102)

[Po] A. B. Polyakov, *On a measure with maximal entropy for a special flow over a local perturbation of a countable topological Bernoulli scheme*, Mat. Sb. **192** (2001), no. 7, 73–96 (Russian). English translation in Sb. Math. **192** (2001), no. 7-8, 1001–1024. MR1861374 (2002m:37024)

[R] M. E. Ratner, *Markov decomposition for the U-flow on a three-dimensional manifold* (in Russian), Mat. Zametki **6** (1969), 693–704. English translation in Math. Notes **6** (1969), 880–886. MR0260977 (41:5597)

[Ro] D. Rosen, *A class of continued fractions associated with certain properly discontinuous groups*, Duke Math. J. **21** (1954), 549–563. MR0065632 (16:458d)

[RoSc] D. Rosen, T. Schmidt, *Hecke groups and continued fractions*, Bull. Austral. Math. Soc. **46** (1992), no. 3, 459–474. MR1190349 (93i:11050)

[Sa] P. Sarnak, Prime geodesic theorems, PhD thesis, Stanford, 1980.

[Sav] S. V. Savchenko, *Special flows constructed from countable topological Markov chains*, Funktsional. Anal. i Prilozhen. **32** (1998), no. 1, 40–53 (in Russian). English translation in Funct. Anal. Appl. **32** (1998), no. 1, 32–41. MR1627271 (99m:28040)

[ScSh] T. Schmidt, M. Sheingorn, *Length spectra of the Hecke triangle groups*, Math. Z. **220** (1995), no. 3, 369–397. MR1362251 (97c:11048)

[S1] C. Series, *On coding geodesics with continued fractions*, Enseign. Math. **29** (1980), 67–76. MR0609896 (82h:30052)

[S2] C. Series, *Symbolic dynamics for geodesic flows*, Acta Math. **146** (1981), 103–128. MR0594628 (82f:58071)

[S3] C. Series, *The modular surface and continued fractions*, J. London Math. Soc. (2) **31** (1985), 69–80. MR0810563 (87c:58094)

[S4] C. Series, *Geometrical Markov coding of geodesics on surfaces of constant negative curvature*, Ergod. Th. & Dynam. Sys. **6** (1986), 601–625. MR0873435 (88k:58130)

[Sm] H. J. S. Smith, *Mémoire sur les Équations Modulaires*, Atti R. Accad. Lincei, Mem. fis. mat. (3), 1 (1877), 136–149; English transl. Coll. Math. Papers, II, 224–241.

[St] J. Stillwell, *Geometry of surfaces*, Springer, 1992. MR1171453 (94b:53001)

[V] W. Veech, *The Teichmüller geodesic flow*, Annals of Math. **124** (1986), 441–530. MR0866707 (88g:58153)

[Z] D. Zagier, *Zetafunkionen und quadratische Körper: eine Einführung in die höhere Zahlentheorie*, Springer-Verlag, 1982. MR0631688 (82m:10002)

Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania 16802

E-mail address: katok_s@math.psu.edu

Department of Mathematics, Rice University, Houston, Texas 77005

E-mail address: idu@rice.edu

Current address: Department of Mathematical Sciences, DePaul University, Chicago, Illinois 60614

E-mail address: iugarcov@depaul.edu