# 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 by symbolic sequences and their development. A geometric method stems from a 1898 work of J. Hadamard and was developed by M. Morse and G. Hedlund in the 1920's and 30's. 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. This method was introduced to dynamics by a 1924 paper of E. Artin, where he used it 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 the ideas, results (old and recent), and interpretations that illustrate the multiple facets of the subject.

Contents



*Date*: February 27, 2005.
2000 *Mathematics Subject Classification.* Primary 37D40, 37B40, 20H05.
*Key words and phrases.* Modular surface, geodesic flow, continued fractions, Markov partition.
This survey is based on the AWM Emmy Noether Lecture given by Svetlana Katok at the annual AMS meeting 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 he 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 and Hedlund in the 1920's and 30's, and since then symbolic dynamics became one of the important tools in the study of systems with so-called "chaotic" behavior of which geodesics flows on Riemannian manifolds of negative sectional curvature represent a major class of examples.

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

Let $\mathcal{H} = \{z = x + iy : y > 0\}$ be the upper half-plane endowed with the hyperbolic metric. The group of Möbius transformations

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

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 $\Gamma \subset PSL(2,\mathbb{R})$, $M = \Gamma\backslash\mathcal{H}$ 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, K3].

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 $v = (z, \zeta) \in S\mathcal{H}$, $z \in \mathcal{H}$, $\zeta \in \mathbb{C}, |\zeta| = \mathrm{Im}\,(z)$. $S\mathcal{H}$ 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 $\iota$ is the unit vector at the point $i$ to the imaginary axis and pointing upwards.

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

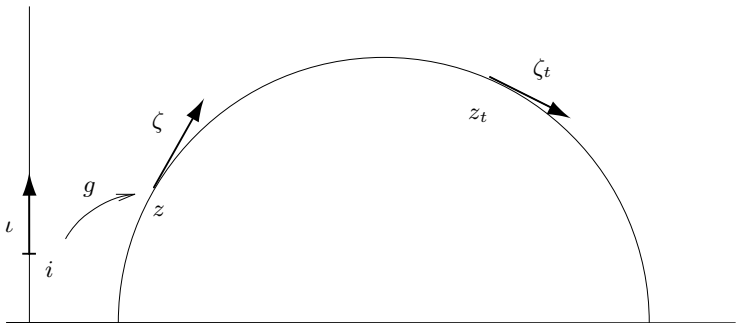


Figure 1. Geodesic flow on the upper-half plane $\mathcal{H}$

multiplication by the one-parameter subgroup

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

so that 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 has singularities at the elliptic fixed points (see [K3, §3.6] for details). The geodesic flow $\{\tilde{\varphi}^t\}$ on $\mathcal{H}$ descents to the *geodesic flow* $\{\varphi^t\}$ on the factor $\Gamma\backslash\mathcal{H} = M$ via the projection $\pi : S\mathcal{H} \to SM$ of the unit tangent bundles (see e.g. [KH, §5.3, 5.4] for more details).

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 $\gamma$ on $M$, its lift to $\mathcal{H}$ is any oriented geodesic $\tilde{\gamma}$ on $\mathcal{H}$ (a ray or a semicircle orthogonal to the real axis $\mathbb{R}$) such that $\pi(\tilde{\gamma}) = \gamma$. The main object of this article is 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 $SM$ which each geodesic (maybe with some minor exceptions) visits infinitely often both in the future and in the past. Every $v \in C$ defines an oriented geodesic $\gamma(v)$ on $M$ which will return to $C$ infinitely often. Let $R : C \to C$ be the *first return map*. The time of the first return to $C$, $f : C \to \mathbb{R}$ is defined as follows: if $v \in C$ and $R(v) = \varphi^t(v)$, then $f(v) = t$. Thus $\{\varphi^t\}$ can be represented as the *special flow* on the space

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

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 with 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 $\sigma$-invariant subset. Then $(\Lambda, \sigma)$ is called a *symbolic dynamical system*. There are some important classes of symbolic dynamical systems. The whole space $(\mathcal{N}^{\mathbb{Z}}, \sigma)$ is called the *Bernoulli shift*. If the space $\Lambda$ is given by a set of simple rules which can be described with the help of a transition matrix of zeros and ones, we say that $(\Lambda, \sigma)$ is a *one-step topological Markov chain* or simply a *topological Markov chain*. Similarly, if the space $\Lambda$ is determined by which $(k+1)$-tuples of symbols are allowed, we say that $(\Lambda, \sigma)$ is a *$k$-step topological Markov chain* (a precise definition is given in Section 5).

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 the appropriate symbolic dynamical system $(\Lambda, \sigma)$ and a continuous surjective map $\mathrm{Cod} : \Lambda \to C$ such that the diagram

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

is commutative. Thus we can 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. [Ad, §3.2 and §5]).

There are two essentially different methods of coding geodesics on surfaces of constant negative curvature. One method stems from the above mentioned work of Hadamard. It was developed by Morse [M1], Hedlund [He1, He2, He3], and Koebe [Ko]. The procedure is described in Section 2. It consists of recording the successive sides of a given fundamental region cut by the geodesic, and it may be applied to all finitely generated Fuchsian groups. 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 in this case 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 5).

The second method is of 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 integral indefinite binary quadratic forms goes back to XIX century works of Gauss [G], Dirichlet [D], Markoff[1] [Ma] and Hurwitz [H1]. 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 the 1924 paper of Artin [Ar] where he 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 the reduction theory for binary indefinite quadratic forms are discussed in Section 3, and the arithmetic coding for a congruence subgroup $\Gamma(2)$ is described in §7.5. For these arithmetic codes the map Cod is a bijection.

In [K4] the first author developed a reduction theory for closed geodesics on cocompact Fuchsian groups which serves the same purpose for the Morse code as the Gauss reduction theory for $SL(2,\mathbb{Z})$ based on continued fractions.

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

In [He1] Hedlund represented geodesics in $\mathcal{U}$ by juxtaposing the Nielsen expansions of their end points. He showed that geodesics are $\Gamma$-equivalent if and only if the corresponding sequences are shift equivalent, and used that to prove ergodicity of the geodesic flow on $\Gamma\backslash\mathcal{U}$. In [He2] he used Artin's code 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 "Hopf argument". The boundary expansions method was further developed for other Fuchsian groups in [BoS, S2, S3] and is discussed in Section 7.1.

[1]This is the Markov particularly remembered for his study of *Markov chains* and *Markov processes*. An old-fashioned spelling *Markoff* of his last name was used in this early publication of his Ph.D. thesis.

By the above mentioned classical result of Hopf, the geodesic flow on a surface of constant negative curvature and finite area is ergodic. Ratner [R] proved the existence of a Markov partition for the geodesic flow on a compact surface, i.e. 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 connection between periodic orbits. Relation of this construction with geometry is tenuous.

A subsequent body of work was devoted to the task of making Markov partitions geometrically explicit with the goal to represent the geodesic flow on a Riemann surface as a special flow over a topological Markov chain. 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 particular one-dimensional (non-invertible) factor-maps. The latter is closely related to a map $f_\Gamma$ on the boundary of the unit disc $\partial\mathcal{U}$ studied by Bowen and Series [BoS] and then by Series [S2, S4] in which she 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 of a set of the first category. The symbolic dynamics she uses derives from the results of [BoS]. Her results apply to a general class of surfaces of constant negative curvature and finite area which, however, does not include the modular surface.

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 [AF3] they make a similar construction for the geodesic flow on compact surfaces of genus $g$ with a particular $8g-4$-sided fundamental region. Arnoux [Arn] uses another method to represent the geodesic flow on the modular surface $M$ as a special flow over a symbolic Bernoulli system on infinite alphabet. He takes special coordinates to describe an appropriate fundamental region for the action of $PSL(2,\mathbb{Z})$ on $S\mathcal{H} \approx PSL(2,\mathbb{R})$ different from the unit tangent bundle of the standard fundamental region of the action of $PSL(2,\mathbb{Z})$ on $\mathcal{H}$, and, by explicit calculation, obtains an expression for the first return map associated to the boundary of this region (a cross-section) in terms of the natural (bijective) extension of the continued fractions (Gauss) map of the interval.

The paper is organized as follows. In Section 2 we describe 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 §2.5 we give a description of the geometric code through Minkowski lattice basis reduction following [GL].

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 generalized 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 [K1, GK] using minus continued fraction expansions of the end points, and interpreted in [GK] via a particular cross-section of $SM$. The set of such arithmetic coding sequences was identified in [GK]: it is a symbolic Bernoulli

system on the infinite alphabet $\mathcal{N} = \{n \in \mathbb{Z}, n \geq 2\}$, i.e. it consists of all bi-infinite sequences constructed with symbols of the alphabet $\mathcal{N}$. 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 our versions of symbolic representation 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 survey other arithmetic codings and interpretations for the modular group: a Farey tiling interpretation of the $A$-code after Moekel and Series (§4.1), a horocycle interpretation of the $H$-code after Fried (§4.2), and also works of Adler and Flatto (§4.3) and Arnoux (§4.4).

In Section 5 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 5.2). 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 5.6), which is the maximal symmetric (i.e. given by a symmetric transition matrix) topological Markov chain (Theorem 5.7). 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 5.5). §5.2 is devoted to a survey of the work by Grabiner and Lagarias [GL].

Section 6 is devoted to applications of arithmetic codes. In §6.2 we use the natural invariant Lebesgue 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. And finally, in §6.4 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 Fuchsian groups of the first kind 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, there was an unpublished work of Koebe of 1917 mentioned in his paper [Ko] of 1927, 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 it the *Morse method*. We will follow [K1] in describing the Morse method for a finitely generated Fuchsian group $\Gamma$ of the first kind. A Dirichlet fundamental region $\mathcal{D}$ for $\Gamma$ always has an even number of sides identified by generators of $\Gamma$ and their inverses; we denote this set by $\{\gamma_i\}$. We label the sides of $\mathcal{D}$ (on the inside) by elements of the set $\{\gamma_i\}$ as follows: if a side $s$ is identified in $\mathcal{D}$ with the side $\gamma_j(s)$, we label the side $s$ by $\gamma_j$. By labeling all the images of $s$ under $\Gamma$ by the same generator $\gamma_j$ 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 $\{\gamma_i\}$ which label the successive sides of $\mathcal{S}$ this geodesic crosses; at each crossing we choose the label corresponding to the image the geodesic *enters*. We first describe the *coding sequence* of a geodesic under the assumption that it does not pass through the vertices 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 may 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 $\gamma_1$ (see Figure 2). Therefore this image is $\gamma_1(\mathcal{D})$, and the first symbol in the code is $\gamma_1$. If it enters the second image of $\mathcal{D}$

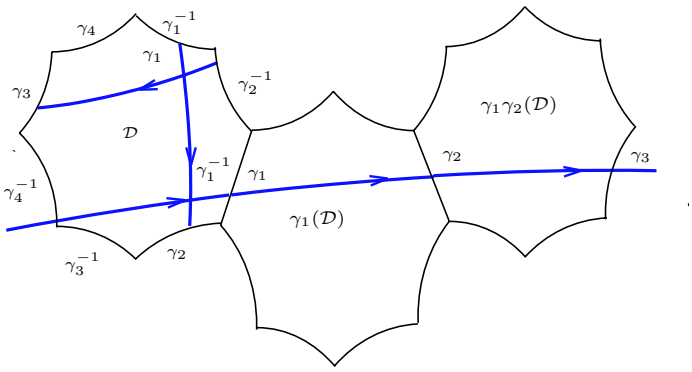


FIGURE 2. Morse coding

through the side labeled by $\gamma_2$, the second image is $(\gamma_1\gamma_2\gamma_1^{-1})(\gamma_1(\mathcal{D})) = \gamma_1\gamma_2(\mathcal{D})$, and the second symbol in the code is $\gamma_2$, and so on. Thus we obtain a sequence of all images of $\mathcal{D}$ crossed by our geodesic in the direction of its orientation: $\mathcal{D}$, $\gamma_1(\mathcal{D})$, $\gamma_1\gamma_2(\mathcal{D}), \ldots$, and a sequence all images of $\mathcal{D}$ crossed by our geodesics in the opposite direction: $\gamma_0^{-1}(\mathcal{D})$, $(\gamma_0\gamma_{-1})^{-1}(\mathcal{D})\ldots$. Thus, the Morse coding sequence is

$$\ldots\gamma_{-1}, \gamma_0, \gamma_1, \gamma_2, \ldots.$$

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 $\gamma \in \Gamma$, i.e. an element which is not a power of another element in $\Gamma$, we have

$$\gamma = \gamma_1\gamma_2\ldots\gamma_n$$

for some $n$. In this case the sequence is periodic with the least period $[\gamma_1, \gamma_2, \ldots, \gamma_n]$.

By mapping the oriented geodesic segments between every two consecutive crossings of the net $\mathcal{S}$ back to $\mathcal{D}$ (as shown on 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 ambiguity in assigning a Morse code occurs if a geodesic passes though a vertex of $\mathcal{D}$: such geodesics have more than one code, and closed geodesics have non-periodic codes along with periodic ones (see [BiS, GL] for relevant discussion).

For free groups $\Gamma$ with properly chosen fundamental regions, all reduced (here this simply means that a generator $\gamma_j$ does not follow or precede $\gamma_j^{-1}$) bi-infinite sequences of elements from the generating set $\{\gamma_i\}$ are realized as Morse coding sequences of geodesics on $M$ (see [S3]), but, in general, this is not the case. Even for the classical example of $\Gamma = PSL(2,\mathbb{Z})$ with the standard fundamental region

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

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 are described in terms of forbidden blocks. The set of generating forbidden blocks found by the authors has an intricate structure attesting the complexity of the Morse code (see §5.2 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. The Morse code with respect to the standard fundamental region $F$ can be assigned to any oriented geodesic $\gamma$ 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 $T$ ($T(z) = z+1$) and $T^{-1}$, respectively, and left and right circular identified and labeled by $S$ ($S(z) = -\frac{1}{z}$) (see Figure 3).

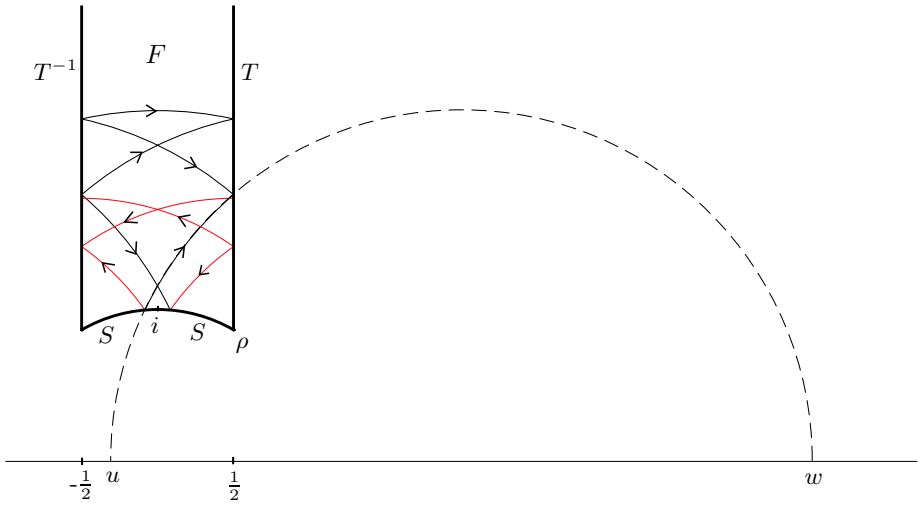


FIGURE 3. The fundamental region and a geodesic on $M$

It is clear from geometrical considerations that any oriented geodesic 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 moving in the direction of the geodesic count the number of times it hits the vertical sides of the boundary of $F$, so that a positive integer is assigned to each block of hits of the right vertical side (or block of $T$'s in the Morse code), and a negative, to each block of hits of the left vertical side (or 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 non-zero integers

$$[\gamma] = [\dots, n_{-1}, n_0, n_1, n_2, \dots],$$

uniquely defined up to a shift, which is called the *geometric code* of $\gamma$. 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 $[\gamma]$. A geodesic is closed if and only if the coding sequence is periodic with the least period $[n_1, n_1, \dots, n_m]$ which we will refer to 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

$$T^{\pm 1}(F), \dots, T^{m_1}(F), T^{m_1}S(F), \dots, T^{m_1}ST^{m_2}S(F), \dots$$

in the positive direction (the sign in the first group of terms is chosen in accordance with the sign of $m_1$, etc.) and

$$S(F), ST^{\mp 1}(F), \dots, ST^{-m_0}(F), \dots ST^{-m_0}ST^{-m_{-1}}(F), \dots,$$

in the negative direction.

The case when a geodesic passes through the corner $\rho$ of $F$ was described to a great extend in [GL, §7]. Such a geodesic has more than one code which are obtained by approximating it by general position geodesics which pass near the corner $\rho$ 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-equivalent classes of coding sequences, some of which are not periodic, but it is unknown 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 [K2] were obtained by the convention that a geodesic passing through $\rho$ 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 $\rho$ 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 $\rho$ slightly lower, and two non-periodic 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 for 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 $\mathcal{N} = \{n \in \mathbb{Z}, |n| \geq 1\}$, endowed with 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 some examples above). Thus $X$ is a closed $\sigma$-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.

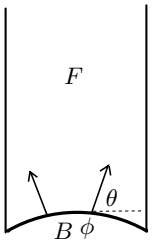


FIGURE 4. The cross-section $B$

**The partition of the cross-section** $B$**.** We parameterize the cross-section $B$ by the coordinates $(\phi, \theta)$, 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: $C_n = \{v \in B \mid n_1(v) = n\}$, i.e. it 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 $n_1(R(v)) = n_2(v)$. The infinite geometric partition and its image under the return map $R$ are sketched on Figure 5. Boundaries between the elements of the partition shown on 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 example, the codes $[4]$ and $[\ldots, 2, -1, 2, -1, 2, -1, 3, 4, 4, 4, \ldots]$ correspond to the point on the right boundary of $B$ between $C_4$ and $C_3$, and the codes $[2, -1]$ and $[\ldots, 4, 4, 4, 4, 3, -1, 2, -1, 2, -1, 2, \ldots]$ 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 Cod 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

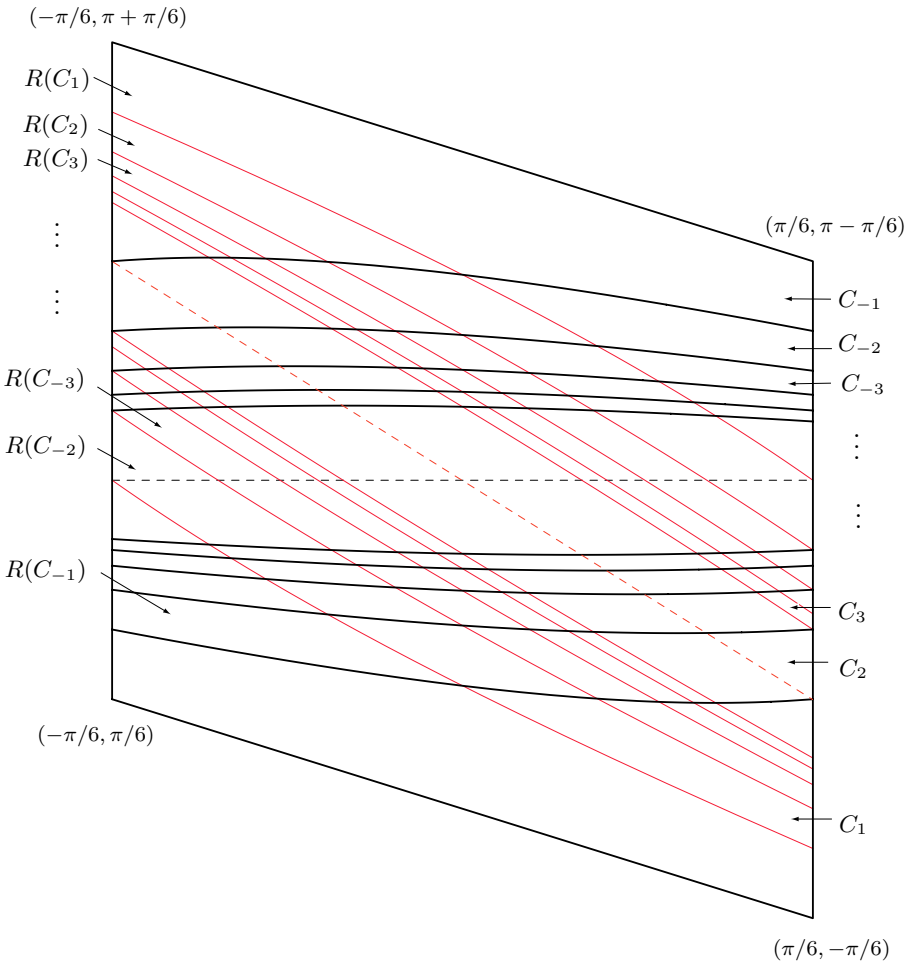


FIGURE 5. The infinite geometric partition and its image under the return map $R$

unit circle, we conclude that the sequence of images of the coding sequences under the map Cod : $X \to B$ converges to the image of the limiting coding sequence. This implies that the map Cod is continuous.

2.4. **Which geometric codes are realized?** Not all bi-infinite sequences of integers are realized as geometric codes. For instance, the periodic sequence $\{8, 2\}$ is not a geometric code since the geometric code of the axis of $T^8ST^2S$ is $[6, -2]$, as can be seen on Figure 6 [K1].

Figure 5 gives an insight into the complexity of the geometric code. The elements $C_n$ and their forward iterates $R(C_n)$ are shown on Figure 5. 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 $R(C_n)$.

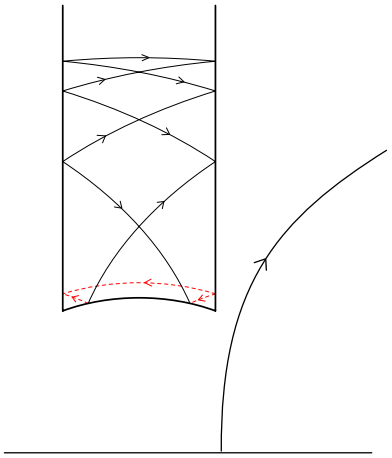

FIGURE 6. The geometric code of the axis of $T^8ST^2S$ is $[6, -2]$

If $n_1(v) = n$ and $n_2(v) = m$ 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.

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 of geometrically Markov admissible geometric codes; see Theorems 5.4 and 5.6 in Section 5.

2.5. **Geometric code via Minkowski lattice basis reduction.** A natural algorithm of providing the geometric coding sequence for a given oriented geodesic by tracing its behavior on the fundamental region $F$ is quite long and involved, but can be programmed using any standard package, like Mathematica. Here we present an algorithm based on Minkowski lattice basis reduction which computes the geometric code very efficiently. This section was largely inspired by [GL, §3].

For any oriented geodesic on the upper half-plane $\mathcal{H}$ and a point $z$ on it, there exists a unique matrix $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R})$ which maps the positively (upwards) oriented imaginary axis $I$ to this geodesic so that $g(i) = \frac{ai+b}{ci+d} = z$. We associate to $v = (z, \zeta) \in S\mathcal{H}$ ($\zeta$ is the unit vector tangent to the geodesic at $z$) a lattice $L = \mathbb{Z}e_1 + \mathbb{Z}e_2$ in $\mathbb{C} \approx \mathbb{R}^2$ with the basis $\mathcal{B}_v = \{e_1 = ai + b,\, e_2 = ci + d\}$. 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 in $\mathbb{R}^2$ yields an oriented geodesic in $\mathcal{H}$: the matrix corresponding to this basis is $g \in SL(2, \mathbb{R})$, and the geodesic is $g(I)$.

The definition of a Minkowski reduced basis is of fundamental importance in the geometry of numbers [Mi, GrL]. Here we follow the terminology of [GrL, GL], although in dimension 2 the reduction algorithm described below goes back to Gauss [G, Article 171].

**Definition 2.1.** Let $L = \mathbb{Z}e_1 + \mathbb{Z}e_2$ be a lattice in $\mathbb{C}$. 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 $m_1$.

We describe the classical algorithm of obtaining a Minkowski basis from a given basis of a lattice $L = \mathbb{Z}e_1 + \mathbb{Z}e_2$:

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

[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 $\mu$ 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].

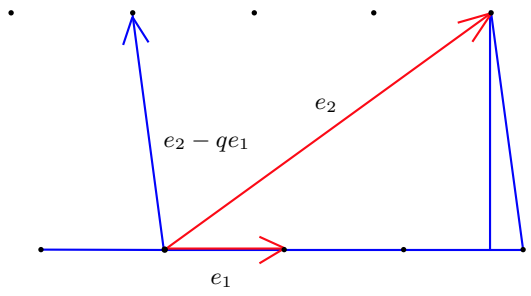


FIGURE 7. Minkowski lattice basis reduction

Since on each step the length of $e_1$ decreases, and there are only finitely many lattice points smaller than the initial $e_1$, this process terminates after finitely many steps.

*Remark* 2.2. Notice that any positively oriented unimodular basis in $L$ is obtained from $\{e_1, e_2\}$ by a matrix $\sigma \in SL(2, \mathbb{Z})$, which corresponds to a left multiplication of the corresponding $g$ by $\sigma$. In particular, a slide of the basis is given by left multiplication by $\begin{pmatrix} 1 & 0 \\ -q & 1 \end{pmatrix} = ST^qS$, a swap of the basis corresponds to a left multiplication by $S$, and a swap followed by a slide is given by left multiplication by $ST^q$.

**Lemma 2.3.** *Let $v = (z, \zeta) \in S\mathcal{H}$ with $z \in F$. If $z$ belongs to the circular boundary of $F$, then the basis $\mathcal{B}_v = \{e_1, e_2\}$ is Minkowski reduced and its vectors have equal length. Otherwise the basis reduces after one swap. Equivalently, any $v = (z, \zeta) \in S\mathcal{H}$ with $z \in S(F)$ corresponds to a Minkowski reduced basis.*

*Proof.* We have

$$z = \frac{ai+b}{ci+d} = \frac{bd+ac}{c^2+d^2} + \frac{i}{c^2+d^2}, \tag{2.2}$$

and

$$|z|^2 = \frac{a^2+b^2}{c^2+d^2}. \tag{2.3}$$

If $z \in F$, then $|z| \geq 1$ and $|\operatorname{Re} z| \leq \frac{1}{2}$. By (2.3) $|e_1| \geq |e_2|$. If $|z| = 1$, $|e_1| = |e_2|$, and

$$\mu = \frac{\langle e_1, e_2 \rangle}{|e_1|^2} = \frac{ac+bd}{a^2+b^2} = \frac{ac+bd}{c^2+d^2} = \operatorname{Re} z.$$

Hence $|\mu| \leq \frac{1}{2}$, and the basis is Minkowski reduced.

If $|z| > 1$, $|e_1| > |e_2|$, we go to step [4], and after the swap we have $e_1 = -ci - d$, $e_2 = ai + b$ with $|e_2| > |e_1|$. The next step is [2]. We have $|\mu| = |\frac{ac+bd}{c^2+d^2}| \leq \frac{1}{2}$, hence $q = 0$ and the algorithm stops. Thus if $z \in F$ is not on the circular boundary of $F$ the basis reduces after one swap. Using Remark 2.2 we see that any $v = (z, \zeta) \in S\mathcal{H}$ with $z \in S(F)$ corresponds to a Minkowski reduced basis. □

Let $B \subset SM$ be the cross-section for the geometric code as above, $(z, \zeta) \in B$, $\gamma$ be an oriented geodesic through $(z, \zeta)$. As was 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 $\mathcal{B}_{v_t} = \{e_1^{(t)}, e_2^{(t)}\}$.

The following theorem demonstrates how the geometric code of a geodesic can be obtained using the Minkowski lattice basis reduction procedure described above which is related to Minkowski geodesic continued fraction expansion of [GL]. The difference between this and Theorem 3.3 of [GL] is that they use finite alphabet and consider only vertical geodesics (see also §5.2 for more details of their work). Our proof is self-contained.

**Theorem 2.4.** *Let $(z, \zeta) \in B$ and $\gamma$ be the oriented geodesic through $(z, \zeta)$. 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 basis of the lattice corresponding to $(z_{t_k}, \zeta_{t_k})$ consists of two vectors of equal length. For $t_k < t < t_{k+1}$ the Minkowski basis is obtained from $\{e_1^{(t)}, e_2^{(t)}\}$ by $k+1$ swap/slide procedures with $q_i = -m_i$ for $1 \leq i \leq k$, where $[m_1, m_2, \ldots, m_k, \ldots]$ is the forward part of the geometric code of $\gamma$.*

*Proof.* By Lemma 2.3 the basis corresponding to $t_0 = 0$ is Minkowski reduced.

Suppose the first symbol in the geometric code of $\gamma$ is $m_1$. This means that $\gamma$ intersects the vertical line $x = \frac{1}{2}$ and its $m_1 - 1$ shifts by $T(z) = z + 1$ if $m_1 > 0$, and the vertical line $x = -\frac{1}{2}$ and its $m_1 - 1$ shifts by $T^{-1}(z) = z - 1$ if $m_1 < 0$. Let $t$ increase from $t_0 = 0$. For small $t$ $z_t \in F$, and by Lemma 2.3 the basis $\{e_1^{(t)}, e_2^{(t)}\}$ can be reduced by a single swap. Assume that $m_1 > 0$ (the case $m_1 < 0$ is handled similarly). Then the next segment of $\gamma$ corresponds to $z_t \in T(F)$. In this case $ST^{-1}$ brings $z_t$ to $S(F)$ so that the corresponding basis is Minkowski reduced by

Lemma 2.3, and it is obtained from $\{e_1^{(t)}, e_2^{(t)}\}$ by one swap/slide procedure with $q_1 = -1$. As $t$ further increases, a similar argument shows that the Minkowski reduced basis is obtained from $\{e_1^{(t)}, e_2^{(t)}\}$ by one swap/slide procedure, where $q$ takes consecutive values $-2, -3, \dots, -m_1$ on the consecutive segments of $\gamma$. The first critical value $t = t_1$ yields the Minkowski basis of two vectors of equal length so that $\pi(z_{t_1}\zeta_{t_1}) \in B$.

If the next symbol in the geometric code is $m_2$, for $t_1 < t \le t_2$ the basis will reduce in two swap/slide procedures with $q_1 = -m_1$ and $q_2$ having the opposite sign to that of $m_2$ and increasing by absolute value up to $|m_2|$. The general statement is proved by induction using the description of the images of $F$ crossed by the geodesic given in §2.2. □

Theorem 2.4 gives us a convenient algorithm for finding the geometric code of a geodesic: take $z_t \in \gamma$ for large $t$, find the corresponding basis $\{e_1^{(t)}, e_2^{(t)}\}$, and reduce it using the Minkowski reduction algorithm, recording $q_1, q_2, \dots, q_k$ of the respective slides. The geometric code will be $m_1 = -q_1, m_2 = -q_2, m_3 = -q_3, \dots, m_k = -q_k$. Increasing $t$ will add additional symbols to the code.

## 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 $\mathcal{H}$ (3.1). We denote a geodesic corresponding to the quadratic form $Q$ by $\gamma(Q)$.

The group $SL(2, \mathbb{Z})$ acts on quadratic forms by substitutions. For $g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in SL(2, \mathbb{Z})$ we set $x = \alpha x' + \beta y'$, $y = \gamma x' + \delta y'$, 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 $\mathcal{H}$, i.e. the same geodesic in $M$. Thus 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, and 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 non-empty canonical set of reduced

elements which form a bi-infinite sequence (which in some cases is periodic), and following the reduction algorithm one can pass from a given element within its equivalence class to a reduced one in a finite number of steps. An application of the reduction algorithm to a reduced element yields the right neighboring element in the sequence.

This concept was first used by Gauss [G] in 1801 to classify integral binary quadratic forms of a given positive discriminant. In 1854 Dirichlet [D] described Gauss's reduction algorithm using expansions of the roots of the quadratic form in simple continued fractions both for $GL(2,\mathbb{Z})$- and $SL(2,\mathbb{Z})$-equivalence. Dirichlet's version of Gauss's algorithm was extended by Markov [Ma] to quadratic forms with real coefficients. Hurwitz [H1] noticed that minus (backward) continued fractions are 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. A complete account of the Gauss reduction theory for indefinite integral binary quadratic forms via the theory of minus continued fractions is given in Zagier's book [Z, Chapter 13], and for its translation into the matrix language see [K1].

Closed geodesics on $M$ are in one-to-one correspondence with conjugacy classes of primitive hyperbolic matrices in $SL(2,\mathbb{Z})$ (see [K2] for details). We associate to a hyperbolic matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z})$ an integral quadratic form $Q_A(x,y) = cx^2 - (d-a)xy - by^2$ of discriminant $D = (a+d)^2 - 4 > 0$. 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 corresponding quadratic equation $Q(z,1) = 0$. Its image in $SL(2,\mathbb{Z})\backslash\mathcal{H}$ 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 non-trivial 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 will 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 $F$ in either direction then the end points of all its lifts to $\mathcal{H}$ are irrational.

Any irrational number $x$ can be expressed uniquely in the form

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

which we will denote by $x = (n_0, n_1, \cdots)$ for short. The "digits" $n_i$ are non-zero integers determined recursively by $n_i = (x_i)$, $x_{i+1} = -\frac{1}{x_i - n_i}$, starting with $n_0 = (x)$ and $x_1 = -\frac{1}{x - n_0}$, where $(\cdot)$ is a certain integer-valued function.

In [KU2] we described three such functions $(\cdot)$ producing different expansions.

**$G$-expansion.** Let $\lfloor x \rfloor$ be the integer part of $x$ (or the floor function), i.e. the largest integer $\leq x$. The function $(x) = \lceil x \rceil = \lfloor x \rfloor + 1$ gives the minus continued fraction expansion first used for the arithmetic code of closed geodesics in [K1]. 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 $\lceil n_0, n_1, n_2, \dots \rceil$.

Since the coding procedure for closed geodesic is exactly the same as Gauss reduction theory for indefinite integral quadratic forms, we refer to this expansion as *Gauss- or $G$-expansion* and corresponding code as *$G$-code*. $G$-codes for oriented geodesics were introduced in [GK].

**$A$-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 we used in [KU2] to reinterpret the classical Artin code *($A$-code)*. This expansion has digits of alternating signes, and we call it *$A$-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) $\alpha$ and $\beta$ are $PSL(2,\mathbb{Z})$-equivalent $\iff$ their expansions have the same tail.
(2) $\alpha$ is a quadratic irrationality $\iff (n_0, n_1, \cdots)$ is eventually periodic,
(3) Let $\alpha$ and $\alpha'$ be conjugate quadratic irrationalities, i.e. the roots of the same quadratic polynomial with integer coefficients. For any quadratic irrationality $\alpha$ with purely periodic expansion $\alpha = (\overline{n_1, \dots, n_k})$, the expansion of $\frac{1}{\alpha'}$ is also purely periodic and $\frac{1}{\alpha'} = (\overline{n_k, \dots, n_1})$.
(4) Two irrationals $\alpha, \beta$ are $PSL(2,\mathbb{Z})$-equivalent if and only if their expansions have the same tail, that is $\alpha = (n_0, n_1, \dots)$ and $\beta = (m_0, m_1, \dots)$ with $n_{i+k} = m_{i+l}$ for some integers $k, l$ and all $i \geq 0$.

**$H$-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 *Hurwitz- or $H$-expansion*. Conversely, any infinite sequence of integers $n_0, n_1, n_2, \dots$ with $|n_i| \geq 2$ for $i \geq 1$ which does not contain the pairs $\{2, p\}$ and $\{-2, -p\}$ for $p \geq 2$ defines an irrational number whose $H$-expansion is $\langle n_0, n_1, n_2, \dots \rangle$.

$H$-expansion satisfies properties (1), (2) and (4), but not (3), so in order to construct a meaningful code, we will need to use a different expansion for $1/u$ introduced by Hurwitz so that (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*.

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 an asymmetric restriction: if $n_i = \pm 2$, then $n_i n_{i+1} < 0$. Moreover a quadratic irrationality $x$ has a purely periodic $H$-expansion if and only if $|x| > 2$ and $\operatorname{sgn}(x)x' \in [r-1, r]$, where $x'$ is conjugate to $x$ (for more details, see [H1, Fr, KU2]).

**Convergents.** If $\alpha = (n_0, n_1, \dots)$, 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 $A$- and $H$-expansions we have

$$|\alpha - \frac{p_k}{q_k}| \leq \frac{1}{q_k^2}, \tag{3.3}$$

while for $G$-expansion we only have

$$|\alpha - \frac{p_k}{q_k}| \leq \frac{1}{q_k}. \tag{3.4}$$

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

**Definition 3.1.** An oriented geodesic in $\mathcal{H}$ with repelling point $u$ and attracting point $w$ 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, $\alpha$-code, where $\alpha = G, A, H$.

**Reduction algorithm.** Let $\gamma$ 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 following 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 geodesics with end poins $u_k$ and $w_k$ is $SL(2,\mathbb{Z})$-equivalent to $\gamma$ by construction.

**Theorem 3.2.** *The above algorithm produces a reduced geodesic $SL(2,\mathbb{Z})$-equivalent to $\gamma$ in finitely many steps, i.e. there exists a positive integer $\ell$ such that the geodesic with end points $u_\ell$ and $w_\ell$ is reduced.*

To a reduced geodesic $\gamma$, 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 expansions of $1/u = (n_{-1}, n_{-2}, \dots)$ and $w = (n_0, n_1, n_2, \dots)$ (for $H$-code we need to use the dual $H$-expansion for $1/u$). Any further application of the reduction algorithm to a reduced geodesic yields reduced geodesics whose codes are left shifts of the code of the first reduced one.

The proofs for each code follow the same general scheme, but the notion of reduced geodesic is different in each case, and so are the properties of the expansions and the estimates.

Now we associate to any oriented geodesic $\gamma$ on $\mathcal{H}$ the $\alpha$-code of a reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to $\gamma$, e.g. obtained by the reduction algorithm described

above. The following theorem shows that $\alpha$-code is well-defined for geodesics on $M$.

**Theorem 3.3.** *The $\alpha$-code is $PSL(2,\mathbb{Z})$-invariant, i.e. two geodesics $\gamma, \gamma'$ on $\mathcal{H}$ are $PSL(2,\mathbb{Z})$-equivalent if and only if for some integer $l$ and all integers $i$ one has $n'_i = n_{i+l}$, where $(\gamma) = (n_i)_{i=-\infty}^{\infty}$ and $(\gamma') = (n'_i)_{i=-\infty}^{\infty}$.*

In [KU2] we present a geometric proof of this fact by constructing a cross-section $C_\alpha$ ($\alpha = G, A, H$) for each code directly related to the notion of $\alpha$-reduced geodesics.

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 $\alpha$-reduced; $Q_1$ consists of all tangent vectors with base points on the right vertical side of $F$ pointing inwards, such that if $\gamma$ is the corresponding geodesic, then $TS(\gamma)$ is $\alpha$-reduced; $Q_2$ consists of all tangent vectors with base points on the left vertical side of $F$ pointing inwards, such that if $\gamma$ is the corresponding geodesic, then $T^{-1}S(\gamma)$ is $\alpha$-reduced. 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 $\alpha$-reduced (Figure 8). It is easy to see that for $G$-code the part $Q_2$ is absent.

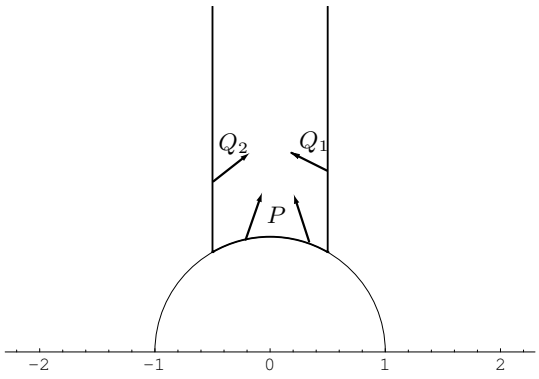


FIGURE 8. The cross-section $C_\alpha = P \cup Q_1 \cup Q_2$

Every oriented geodesic $\gamma$ on $M$ can be represented as a bi-infinite sequence of segments $\sigma_i$ between successive returns to $C_\alpha$. To each segment $\sigma_i$ we associate the corresponding $\alpha$-reduced geodesic $\gamma_i$ on $\mathcal{H}$. Thus we obtain a sequence of reduced geodesics $\{\gamma_i\}_{i=-\infty}^{\infty}$ representing the geodesic $\gamma$. If one associates to $\gamma_i$ its $\alpha$-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 $\alpha$-reduced geodesics $\gamma_i$ in the sequence produce the same, up to a shift, bi-infinite coding sequence, which we call the *$\alpha$-code* of $\gamma$ and denote by $(\gamma)$. The left shift of the sequence corresponds to the return of the geodesic to the cross-section $C_\alpha$.

**Example 3.4.** Let $\gamma$ be a geodesic on $\mathcal{H}$ from $u = \sqrt{5}$ to $w = -\sqrt{3}$. The $G$-expansions are

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

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

**Symbolic representation of geodesics for arithmetic codes.** Let $\mathcal{N}_G^{\mathbb{Z}}$ be the Bernoulli space on the infinite alphabet $\mathcal{N}_G = \{n \in \mathbb{Z}, n \geq 2\}$. We proved that each oriented geodesic which does not go to the cusp of $M$ in either direction corresponds to its $G$-code, $\lceil \gamma \rceil \in \mathcal{N}_G^{\mathbb{Z}}$. Conversely, each bi-infinite sequence $x \in \mathcal{N}_G^{\mathbb{Z}}$ which does not have an infinite tail of 2's in either direction produces a geodesic on $\mathcal{H}$ from $u(x)$ to $w(x)$ (irrational end points), where

$$w(x) = \lceil n_0, n_1, \dots \rceil \quad , \quad \frac{1}{u(x)} = \lceil n_{-1}, n_{-2}, \dots \rceil\,. \tag{3.5}$$

As was explained in [KU2], this correspondence extends to all oriented geodesics on $M$ if we extend the notion of $G$-reduced geodesic to those with $0 < u < 1$ and $w \geq 1$. Thus the set of all oriented geodesics on $M$ can be described symbolically as the Bernoulli space (minus one point) $X_G = \mathcal{N}_G^{\mathbb{Z}} \setminus \lceil \bar{2} \rceil$.

Similarly, using the $A$-code, the set of all oriented geodesics 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 $\mathcal{N}_A = \{n \in \mathbb{Z}, n \neq 0\}$ and transition matrix $A$,

$$A(n, m) = \begin{cases} 1 & \text{if } nm < 0, \\ 0 & \text{otherwise}\,, \end{cases} \tag{3.6}$$

and using the $H$-code, we can describe the set of all oriented geodesics on $M$ symbolically as a countable one-step Markov chain $X_H \subset \mathcal{N}_H^{\mathbb{Z}}$ with infinite alphabet $\mathcal{N}_H = \{n \in \mathbb{Z}, |n| \geq 2\}$ and transition matrix $H$,

$$H(n, m) = \begin{cases} 0 & \text{if } |n| = 2 \text{ and } nm > 0, \\ 1 & \text{otherwise}\,. \end{cases} \tag{3.7}$$

**The Cod maps for arithmetic codes.** Since $X_\alpha \subset \mathcal{N}_\alpha^{\mathbb{Z}}$ is a topological Markov chain, it is clearly a closed shift-invariant subset of $\mathcal{N}_\alpha^{\mathbb{Z}}$. As it was shown above, the coding map for each arithmetic $\alpha$-code, $\mathrm{Cod}_\alpha : X_\alpha \to C_\alpha$ is a bijection between the cross-section $C_\alpha$ and the symbolic space $X_\alpha \subset \mathcal{N}_\alpha^{\mathbb{Z}}$.

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 $m = \max\{k \mid n_i = n_i' \text{ for } |i| \leq k\}$.

**Proposition 3.5.** *The map* $\mathrm{Cod}_\alpha$ *is continuous.*

*Proof.* If $d(x,x') < \frac{1}{m}$, then the $\alpha$-expansions of the attracting end points $w(x)$ and $w(x')$ of the corresponding geodesics given by (3.5) have the same first $m$ digits, therefore the first $m$ convergents of their $\alpha$-expansions are the same, and by (3.4) and (3.3) $|w(x) - w(x')| < \frac{1}{m}$. 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. □

**The partition of the cross-section $C_\alpha$.** We parameterize the lift of the cross-section $C_\alpha$ to $S\mathcal{H}$, $C_a$ by the coordinates $(\phi, \theta)$, 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 $\theta$ depends on $\phi$ and is determined by the condition that the corresponding geodesic is $\alpha$-reduced.

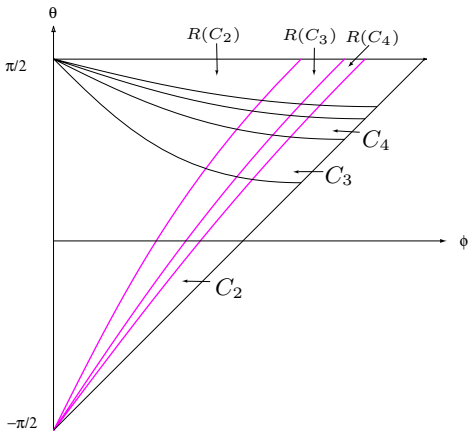


FIGURE 9. Infinite partition for the $G$-code and its image under the return map $R$

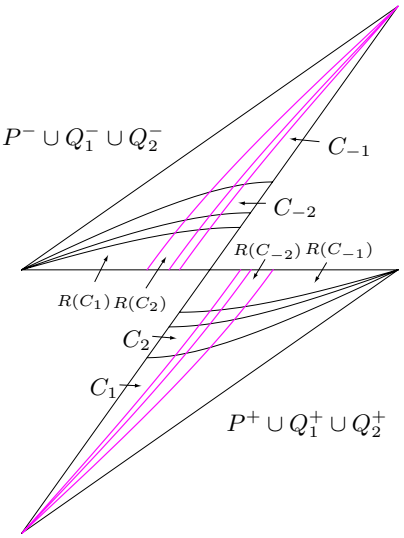


FIGURE 10. Infinite partition for the $A$-code and its image under the return map $R$

The elements of the partition of $C_a$ are labeled by the symbols of the corresponding alphabet $\mathcal{N}_\alpha$, $C_a = \cup_{n \in \mathcal{N}_G} C_n$ 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 $C_\alpha$ by projection) corresponding to the $\alpha$-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 on Figures 9, 10, and 11.

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 $R(C_n) \cap C_m \neq \emptyset$,

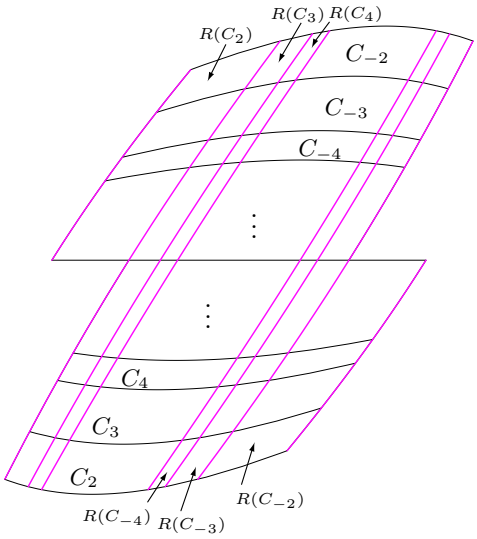


FIGURE 11. Infinite partition for the $H$-code and its image under the return map $R$

and since all intersections are transversal, according to [Ad, Theorem 7.9], each partition is Markov.

## 4. Other arithmetic codings and interpretations

4.1. **Farey tiling interpretation of the $A$-code.** The *Farey sequence of order $n$* $F_n$ is the set of irreducible rational numbers $p/q$ with $(p,q)=1$ and $|p|\le n$, $|q|\le n$ arranged in the increasing order. It is convenient to include $\infty$ in each Farey sequence $F_n$. For example, the non-negative 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 $p/q$ and $p'/q'$ are adjacent in the Farey sequence of order $\max(|p|,|q|,|p'|,|q'|)$ if and only if $pq'-p'q=-1$.

Let $e_0=\{iy, y>0\}$ be the standard vertical geodesic on $\mathcal{H}$. 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 $g(\infty)=a/c$. Since $g(0)/g(\infty)=1-\frac{1}{ad}\ge 0$, $g(e_0)$ does not cross $e_0$, and therefore the images of $e_0$ 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 $\infty$ under $SL(2,\mathbb{Z})$; we call them *cusp points*.

Let $\Delta_0$ denote the ideal triangle in $\mathcal{H}$ with vertices $0, 1$, and $\infty$. The images of $\Delta_0$ under $SL(2,\mathbb{Z})$ are known as *Farey triangles*. If we apply to $\Delta_0$ transformations

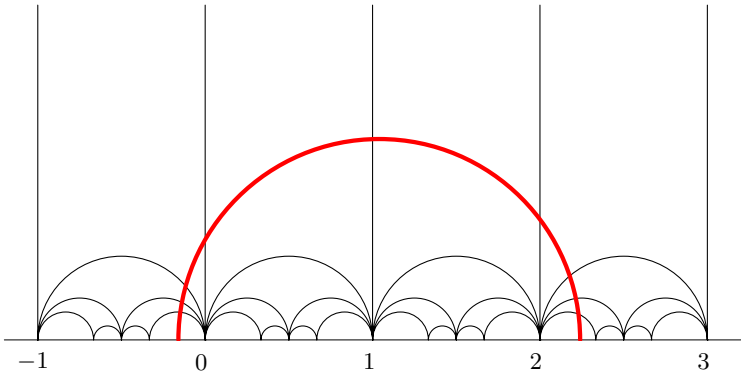


FIGURE 12. Farey tessellation

$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z})$ with $\max(|a|,|b|,|c|,|d|) \leq n$, we will cover a convex subset of $\mathcal{H}$, and the images of $\infty$ will be exactly the Farey sequence $F_n$. Notice that $\Delta_0$ 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 $\Delta_0$ will cover larger and larger part of $\mathcal{H}$, and since the rational points are dense in $\mathbb{R}$, the images of $\Delta_0$ under $SL(2,\mathbb{Z})$ will cover $\mathcal{H}$ without overlap, forming the *Farey tiling*.

Hurwitz [H2] used the Farey tiling to describe the Gauss reduction theory for $GL(2,\mathbb{Z})$ geometrically. Hurwitz's elegant approach is described in [Fr]. Series [S3] gave a similar description following an earlier work of Moeckel [Mo] where he used Farey tiling and its relation to continued fractions to study asymptotic frequencies with which geodesics go into different cusps for a certain class of subgroups whose fundamental regions are made up of Farey triangles. Moeckel and Series referred to a 1916 work of Humbert [Hu], but did not know about Hurwitz's work which has been published 22 years earlier. We will use the Farey tiling here to describe the $A$-code.

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

Suppose a geodesic $\gamma$ is $A$-reduced, with $w = \lfloor n_0, n_1, n_2, \dots \rceil > 1$, so $n_0 \geq 1$. If we trace $\gamma$, and count the number of intervals between crossings of the principal edges with the associated signs, we obtain a sequence of non-zero integers. It is easy to see that, if we start tracing $\gamma$ at its intersection with the vertical edge $e_0$, the number of crossings between $e_0$ and the second principal edge $T^{n_0}e_0$ is $n_0$. Since the cusp at $\infty$ is on the left of $\gamma$, 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 $\gamma$ 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 $w$, $\lfloor n_0, n_1, n_2, \dots n_k \rceil$. Thus the obtained sequence is exactly equal to the $A$-expansion of $w$.

Since the repelling end point $u$ of $\gamma$ satisfies $-1 < u < 0$, $1/u < -1$, we have $1/u = \lfloor n_{-1}, n_{-2}, n_{-3}, \dots \rceil$, and hence

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

Moving along $\gamma$ from its intersection with $e_0$ towards $u$ one observes that the principal edge preceding $e_0$ is a geodesic from $-1/n_{-1}$ to 0, and the number of crossings between this edge and $e_0$ is exactly $|n_{-1}|$. Now, if we move from the principal edge preceding $e_0$ towards $e_0$, we see that the cusp at 0 is on the right of $\gamma$, so it has a "-" sign, and the number preceding $n_0$ in the sequence is $n_1$. Thus the sequence of non-zero integers obtained by counting the number of intervals between crossings of $\gamma$ with the principal edges is exactly the $A$-code of $\gamma$, and the changing of the original principal edge changes the sequence by a shift.

4.2. **Interpretation of the $H$-code via Ford discs.** D. Fried in [Fr] 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) on 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

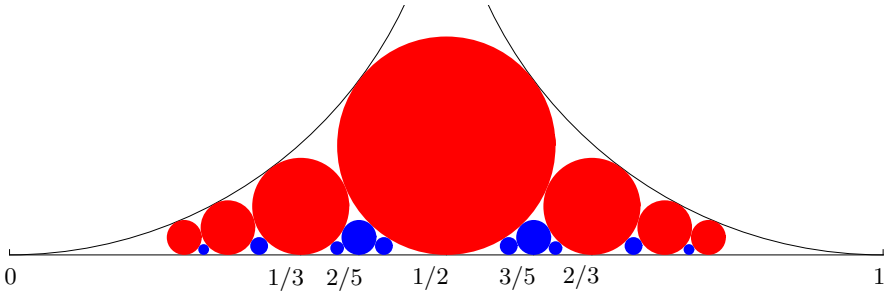


FIGURE 13. Ford discs

obtained as an image under an element of $SL(2,\mathbb{Z})$ of the standard horodisc $B(\infty)$ defined by $\mathrm{Im}\,(z) \geq 1$ (a Ford disc is labeled $B(r)$ if $r \in \mathbb{Q}$ is the point where it touches the real axis). In 1918 Ford [F1] 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. Later, Ford [F2] gave an elementary description using horodiscs to illustrate some of the properties of simple continued fractions and Farey sequences.

Here are some of the basic properties of Ford discs (see [F2]]): 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 $\infty$ 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 but $B(r')$ is not of level $\leq 1$, etc. 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 $p_k/q_k$ to be of level $k$.

4.3. **Adler-Flatto's work.** Let $v = (z = x + iy, \zeta) \in S\mathcal{H}$, where $z \in \mathcal{H}$, $\zeta \in \mathbb{C}$, $|\zeta| = y$. Thus $S\mathcal{H}$ can be coordinatized by $v = v(x, y, \theta)$, where $\theta = \arg(\zeta)$. In [AF1, AF2], the authors consider a different parametrization of $S\mathcal{H}$. Any $v = (z, \zeta) \in S\mathcal{H}$ defines a unique oriented geodesic in $\mathcal{H}$ with the end points $u$ and $w$. If $s$ is the hyperbolic distance along this geodesic from some conveniently chosen origin, then $(u, w, s)$ gives another system of coordinates on $S\mathcal{H}$, in which the geodesic flow has a particularly simple form:

$$\varphi^t : (u, w, s) \mapsto (u, w, s + t). \tag{4.1}$$

The authors consider a cross-section of the geodesic flow on the modular surface given by the points of the boundary of the standard fundamental region $F$ together will all direction vectors pointing to the exterior of $F$. The point is that in the above parametrization the first return map may be simply expressed. Having achieved this, the authors compute the return time function. They construct a linearized version of the cross-section map (the so-called "rectilinear" map) which is conjugate to the initial cross-section map by a one-to-one map which is the identity on much of its domain, and it has a simple geometric interpretation on regions where it is not the identity. From the original cross-section map the authors retrieve the Morse coding, and using the linearized version they get a version of Artin's coding which possesses a simple Markov partition. The Gauss map appears as a factor map of the rectilinear map, from where 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.

A similar approach has been developed by the same authors in [AF3] for the geodesic flow on compact surfaces of genus $g \geq 2$ and of constant negative curvature, using a particular $8g - 4$-sided fundamental polygon.

4.4. **Arnoux's work.** Using the algebraic definition of the geodesic flow 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 fractions 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 $y = \lfloor 0, a_{-1}, a_{-2}, \dots \rfloor$ (here $\lfloor \cdot \rfloor$ denotes the simple 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 $\Omega$ 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 $\Omega$ 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 return time $-2 \log \lfloor 0, a_1, a_2, \dots \rfloor$.

Using these computations, the author gives a short proof of the fact that for almost every real number

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

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

## 5. Complexity of the geometric code

Unlike the spaces of admissible arithmetic codes $X_G, X_A$, and $X_H$ which were proved to form topological Markov chains in §3.2, the space of admissible geometric codes $X$ is very complicated. In order to state a complexity result proved in [KU1] we recall the notion of a $k$-step topological Markov chain defined on the alphabet $\mathcal{N}$ (for finite alphabet see for example [KH, §1.9]):

**Definition 5.1.** Given an integer $k \ge 1$ and a map $\tau : \mathcal{N}^{k+1} \to \{0, 1\}$, the set

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

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

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

**Theorem 5.2.** *The space $X$ of geometric codes is not a $k$-step topological Markov chain, for any integer $k \ge 1$.*

Deciding which bi-infinite sequences of non-zero integers are admissible geometric codes is a nontrivial task. We describe what is known in what follows.

5.1. **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 [GK]:

**Theorem 5.3.** *A bi-infinite sequences of positive integers* $\{\dots, n_{-1}, n_0, n_1, n_2, \dots\}$ *is an admissible geometric code if and only if*

$$\frac{1}{n_i} + \frac{1}{n_{i+1}} \le \frac{1}{2} \quad \text{for all } i \in \mathbb{Z}. \tag{5.1}$$

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

The pairs forbidden by Theorem 5.3, $\{2, p\}$, $\{q, 2\}$, $\{3, 3\}$, $\{3, 4\}$, $\{4, 3\}$, $\{3, 5\}$, and $\{5, 3\}$—we call them *Platonic restrictions*—are of Markov type. More precisely, the set of all bi-infinite sequences satisfying relation (5.1) of Theorem 5.3 can be described as a one-step countable topological Markov chain, with the alphabet $\mathcal{N}_G$ and transition map $P$,

$$P(n, m) = \begin{cases} 1 & \text{if } 1/n + 1/m \le 1/2\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{5.2}$$

We will refer to $P$ as a transition matrix in this case, and denote the associated one-step Markov chain by $X_P$. Clearly, $X_P$ is a shift-invariant subset of $X$.

Geodesics identified in Theorem 5.3 have the property that all segments comprising them in $F$ are positively (clockwise) oriented. Following [GK] 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 5.4.** *Any bi-infinite sequence of integers* $\{\dots, n_{-1}, n_0, n_1, n_2, \dots\}$ *such that*

$$\left| \frac{1}{n_i} + \frac{1}{n_{i+1}} \right| \le \frac{1}{2} \quad \text{for } i \in \mathbb{Z} \tag{5.3}$$

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

The set of all bi-infinite sequences satisfying relation (5.3) of Theorem 5.4 can be described as a one-step countable topological Markov chain, with the alphabet $\mathcal{N}$ and transition matrix $M$,

$$M(n, m) = \begin{cases} 1 & \text{if } |1/n + 1/m| \le 1/2\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{5.4}$$

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 5.4 and corresponding geodesics, *geometrically Markov.* In [KU2] we show that the $H$-code comes closest to the geometric code:

**Theorem 5.5.** *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 $\sigma$-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 14), but it is not geometrically Markov. Moreover, the latter is also an example of a non-geometrically 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.

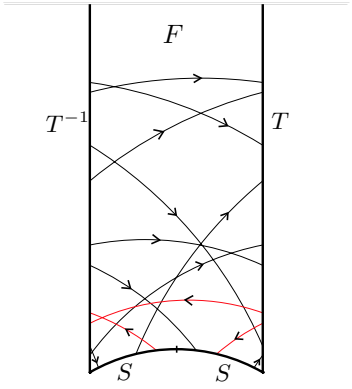


FIGURE 14. Geometric code $[5, 3, -2]$

The following theorems were proved in [KU1]:

**Theorem 5.6.** *The set $X_M$ is a maximal, transitive one-step countable topological Markov chain in the set of all geometric codes $X$.*

**Theorem 5.7.** *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 5.8.** *For any geometrically Markov geodesic whose geometric code has alternating sign symbols, the $A$-code coincides with the geometric code.*

5.2. **Grabiner-Lagarias results.** The main subject of [GL] is complexity of the Morse code for the modular group and computational complexity of conversion between different symbolic codings. They use *cutting sequences* instead of our numerical geometric coding 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 is a one-sided shift on infinitely many symbols, they consider what they call *additive continued fraction expansions* using three symbols $T$, $T^{-1}$ and $S$, and *Minkowsi geodesic continued fraction expansions* obtained by Minkowski lattice basis reduction which was already described in §2.5, also on a finite alphabet. They also consider what they call the *Farey tree expansion*, which is exactly our $A$-expansion on the finite alphabet (see §4.1 above).

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. is not a factor of a finite-step topological Markov chain (on a finite alphabet) ([GL, Theorem 7.3]; it is related to our Theorem 5.2 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 cutting sequence expansion of $\{\theta + it : t > 0\}$ and the Minkowski continued fraction expansion of $\theta$ can each be computed from the other by a finite automation ([GL, Theorem 3.3]; it is related to our Theorem 2.4). The additive continued fraction expansion of $\theta$ can be computed from either of these expansions by a finite automation [GL, Theorem 4.3].
- The Minkowski continued fraction expansion of $\theta$ cannot be computed from the additive continued fraction expansion of $\theta$ by a finite automation [GL, Theorem 6.4].
- For real $\theta > 1$, the additive continued fraction expansion of $\theta$ can be converted into the Farey tree expansion of $\theta$ by a finite automation, and vice versa ([GL, Theorem 3.2]; see also §4.1 above).

Without going into technical details which can be found in [GL, §3.5], a finite automation (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 automation is that it has a fixed finite amount of memory, and in order to use this notion one needs to work with sequences on finite alphabet. Grabiner and Lagarias show that geometric coding sequences encode more information about the geodesic flow than arithmetic ones which are topological Markov chains, but only retain topological and not conformal information about the Riemann surface $M$.

## 6. Applications of arithmetic codes

6.1. **Geodesic flow as a special flow.** In §§2.3 and 3.3 we have constructed four continuous surjective coding maps. The map $\mathrm{Cod} : X \to B$ for the geometric is not injective, while the maps for three three arithmetic codes, $\mathrm{Cod}_\alpha : X_\alpha \to C_\alpha$ ($\alpha = G, A, H$) are bijections, and 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 $(\Lambda, \sigma)$, 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 $\Lambda$ is the time of the first return of the geodesic $\gamma(x)$ to the cross-section $C$. The following theorem was proved in [GK] 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 the original Artin's code has 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, $R : C_\alpha \to C_\alpha$ be the first return map, and $p_\alpha : C_\alpha \to I_\alpha$ be a map from the cross-section to the interval $I_\alpha$ 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 $\alpha$-reduced geodesic defined by $v$. The factor-map $f_\alpha : I_\alpha \to I_\alpha$ is a map such that the following diagram

$$\begin{array}{ccc} C_\alpha & \xrightarrow{\mathrm{R}} & C_\alpha \\ {\scriptstyle \mathrm{p}_\alpha}\downarrow & & \downarrow{\scriptstyle \mathrm{p}_\alpha} \\ I_\alpha & \xrightarrow{\mathrm{f}_\alpha} & I_\alpha \end{array}$$

is commutative. We will derive the formulas for the factor-map for all three codes. Let $x = \frac{1}{w}$. 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 $\frac{1}{w'} = n_0 - w = (\frac{1}{x}) - \frac{1}{x}$.

In order to calculate the invariant measure for the map $f_\alpha$ we use the parametrization of $S\mathcal{H}$ by $(u, w, t)$ obtained in [AF1] and described in §4.3, The measure $dm$ in these coordinates is given by the formula

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

and its invariance under $\{\varphi^t\}$ follows immediately from (4.1) and (6.2). The measure on the cross-section $C_\alpha$ invariant for the first return map is obtained by dropping $dt$: $dm_{C_\alpha} = \frac{dudw}{(w-u)^2}$, and the invariant measure on $I_\alpha$ is obtained by integrating $dm_{C_\alpha}$ with respect to $du$ as explained in [AF1].

***G*-code.** In this case $I_G = (0, 1)$, and $f_G(x) = \lceil \frac{1}{x} \rceil - \frac{1}{x} = \lfloor \frac{1}{x} \rfloor + 1 - \frac{1}{x} = 1 - \{\frac{1}{x}\}$. In order to compute the invariant measure on $I_G$ we first integrate $dm_{C_G}$ over $I_G = (0, 1)$:

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

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

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

***A*-code.** In this case $I_A = (-1, 1)$, and $f_A = \lceil \frac{1}{x} \rfloor - \frac{1}{x} = \begin{cases} -\{\frac{1}{x}\} \text{ if } x > 0 \\ \{-\frac{1}{x}\} \text{ if } x < 0 \end{cases}$ and the invariant measure is $\mu_A = \frac{2dx}{(1+x)(1-x)}$.

**$H$-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 with the use of arithmetic codes.** In [Ar] Artin used continued fractions to prove topological transitivity of the geodesic flow on the modular surface and density of closed geodesics. Moreover, any arithmetic $\alpha$-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 closed geodesics are dense in $SM$, and the geodesic flow is topologically transitive.*

*Proof.* Recall that closed geodesics correspond to periodic $\alpha$-codes. Then by Proposition 3.5 it is sufficient to find a periodic coding sequence arbitrary 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 arbitrary 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 $X_\alpha$ (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 $\epsilon$-close to $v(x)$, which completes the proof. □

6.4. **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. Let $L \subset \Lambda$ be a $\sigma$-invariant Borel subset of $\Lambda$. 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

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

much as the special flow was defined in §6.1.

Let $\tilde{\mu}$ be an arbitrary $\{\psi^t\}$-invariant Borel probability measure on $L^g$ and $\mu'$ its projection onto $L$. The sets $C_x = \{x\} \times \Delta_x,\ x \in L$, where $\Delta_x = \{y : 0 \le y \le g(x)\}$, 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 $\Delta_x$ for $\mu'$-almost all $x$ (here we identify $C_x$ and $\Delta_x$). By definition the function $x \mapsto 1/g(x)$ is bounded and hence $\mu'$-integrable. So we can introduce a measure $\mu$ on $L$

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

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

Conversely, given a $\sigma$-invariant probability measure $\mu$ 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 $\sigma$-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 $h_\mu$ the measure-theoretic entropy of $\sigma$ with respect to $\mu$. 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 [GK], 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

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

Topological entropy has the following properties. Let $g_1 \geq g_2$ on $L$, and let $\{\psi_2^t\} = (L, g_2)$. Then by (6.3) $h(\{\psi^t\}) \leq h(\{\psi_2^t\})$. 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 $g_1(x) = g_2(x) + h(x) - h(\sigma(x))$, the special flows $(L, g_1)$ and $(L, g_2)$ are conjugate [PP] and therefore have the same topological entropy.

The first example of the special flow of the type described above was studied in [GK]: the special flow over $L = X_P \subset X$, the space of positive coding sequences, with the ceiling function $f(x) = 2 \log w(x)$. Let $\Sigma^+$ 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 $B$ (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 [GK]:

**Theorem 6.3.** $0.7771 < h(\{\varphi_t^+\}) < 0.8161$.

The function $f(x) = 2 \log w(x)$ can be extended to the whole space $X$ since the formulae (3.5) make sense for any $x \in X$. For every $x = (n_i)_{i \in \mathbb{Z}} \in X$ we write $n_i = n_i(x)$. It is easy to show that since for $x \in X_P$ $n_i(x) \geq 3$, we have

$$2 \log c n_1(x) \leq f(x) \leq 2 \log n_1(x), \text{ where } c = (3 + \sqrt{5})/6 \approx 0.8726. \tag{6.4}$$

Thus the ceiling function is estimated by two functions which depend only on the first coordinate $n_1(x)$. We can now use a formula for the topological entropy developed by Polyakov [Po] based on a result of Savchenko [Sa]. 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 $n_1(x)$. For $(X_A, g_\delta)$, with $g_\delta(x) = 2 \log \delta n_1(x)$, for $\delta = 1$ and $\delta = c$ we obtain the estimates

$$h_1 = 0.7771 < h(\{\varphi_t^+\}) < 0.8161 = h_c.$$

The estimated values $h_\delta$ 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 with the Riemann $\zeta$-function by the formula

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

They were obtained with the help of the computer package Pari-GP.

The second example was studied in [KU1]. Let $\Sigma$ 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 §5.1). The set $\Sigma$ is flow invariant and noncompact. Let $\{\varphi^t_{|\,\Sigma}\}$ the restriction of the geodesic flow to $\Sigma$. 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.** $0.8417 < h(\{g^t_{|\,\Sigma}\})$.

The proof of this estimate follows the same scheme as in the previous theorem, but the extend of that method only allows us to obtain an estimate from below. Of course, since $h(\{\varphi^t\}) = 1$ (see [GK]), 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 of the geodesic flow on surfaces of constant negative curvature and finite hyperbolic area. Her main result was that the geodesic flow on a compact surface could be represented as a factor of a special flow over a topological Markov chain on finite alphabet of generators of the fundamental group $\Gamma$ of the surface by a continuous map which is one-to-one except on a set of the first category. The symbolic dynamics she uses was derived from the results of her earlier work with Bowen [BoS] in which the action of $\Gamma$ on the boundary of the unit disc $\partial\mathcal{U}$ was shown to be orbit equivalent to a certain Markov map $f_\Gamma$ that they used to develop boundary expansion code geometrically. The map $f_\Gamma$ is piecewise equal to the generating transformations of $\Gamma$ that identify the sides of the fundamental region $D$, and produces a bi-infinite sequence of generators of $\Gamma$ that we call 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 symmetrical $4g$ polygon the unit disc $\mathcal{U}$. In the presence of cusps one still obtains a Markov map, but on countable alphabet.

Series' results apply to a general class of surfaces (which, however does not include the modular surface) and are obtained by considering specially chosen fundamental regions with *even corners* (this means that $\Gamma(\partial D)$ consists of complete geodesics in $\mathcal{U}$) for which she establish precise relation 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 dynamics properties could be establishes as in earlier work of Artin, Nielsen, and Koebe, simply by producing admissible sequences of required kind (see also §6.3 below).

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]. The codes differ generally if the group $\Gamma$ is not free. The discrepancy is closely related to the possible different ways of representing elements of $\Gamma$ as shortest words in a given set of generators.

In what follows we will describe the Bowen-Series boundary expansion code for an example of a 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,

$$\Gamma(2) = \{\gamma \in PSL(2,\mathbb{Z}) \,|\, \gamma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod 2\}\,.$$

Notice that $\Gamma$ 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 are given by the transformations $A$, $B$, parabolic transformations fixing the cusps at $\infty$ and 0, respectively. Two equivalent points $-1$ and 1 represent the third cusp of $M_2$.

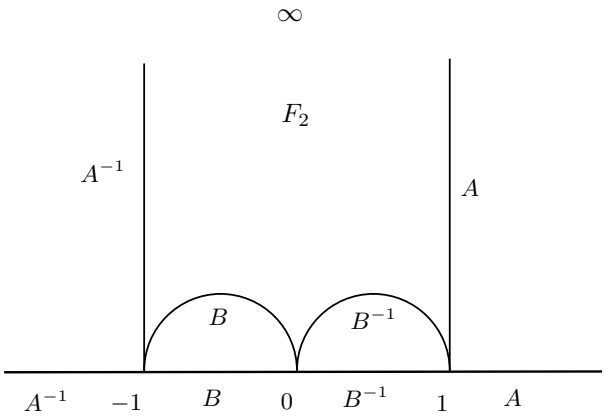


FIGURE 15. The fundamental region for $M_2$

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 $\gamma$ on $F_2$ which does not go to any of three cusps of $F_2$ in either direction. It is easy to see that the images of the cusp at $\infty$ 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 will consider only geodesics whose lifts to $\mathcal{H}$ have irrational end points.

We consider as a starting segment of an oriented geodesic $\gamma$ on $M_2$ a segment with the initial end point on one of the semi-circles and the final end point on one of the vertical sides. We label the sides of $F_2$ (on the outside) as shown on Figure 15: the left vertical by $A$, the right vertical by $A^{-1}$, the left circular by $B^{-1}$, and the right circular by $B$. We begin the coding procedure by writing down the labels $A$, $A^{-1}$, $B$, $B^{-1}$ of the sides hit by the geodesic. In this way, 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 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 the initial end point on one of the vertical sides and the final point on one of the semi-circles, 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 $B^{n_0}$).

In conclusion, the set of all Morse sequences can be modelled by a subset 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 $\sigma$ defined as

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

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 $\mathbb{R} \cup \{\infty\}$ by the inverse elements, 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 $w = w_0, w_1, w_2, \dots$, where $w_n$ is the label of the segment where $f^n_{\Gamma(2)}(w)$ belongs; $(w_n \in \{A, A^{-1}, B, B^{-1}\})$. The boundary expansion of $u$ is the sequence $u = u_0, u_1, u_2, \dots$, where $u_n$ is the label of the segment where $f^n_{\Gamma(2)}(u)$ belongs $(u_n \in \{A, A^{-1}, B, B^{-1}\})$. Let $\bar{u}_n$ denote the inverse to the label 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 coincide 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 simple continued fractions for $GL(2, \mathbb{Z})$- and $SL(2, \mathbb{Z})$-reduction theory 23 years earlier (see Section 3.1). It is interesting to notice 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 will 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 Section 3.2.

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 will call even continued fractions expansion (or *E-expansion*) and denote by $x = (\!(2n_0, 2n_1, \cdots)\!)$ for short. The non-zero integers $n_i$ are determined recursively by $2n_i = (\!(x_i)\!)$, $x_{i+1} = -\frac{1}{x_i - 2n_i}$, starting with $2n_0 = (\!(x)\!)$ and $x_1 = -\frac{1}{x-2n_0}$, where $(\!(x)\!)$ is the nearest even integer to $x$. The following properties are satisfied:

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

**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 $\gamma$ 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 geodesics with end points $u_k$ and $w_k$ is $\Gamma(2)$-equivalent to $\gamma$ by construction.

**Theorem 7.2.** *The above algorithm produces a reduced geodesic $\Gamma(2)$-equivalent to $\gamma$ in finitely many steps, i.e. there exists a positive integer $\ell$ such that the geodesic with end points $u_\ell$ and $w_\ell$ is reduced.*

To such a reduced geodesic $\gamma$ 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 $-u = (\!(2n_{-1}, 2n_{-2}, \ldots)\!)$.

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

The proof of this theorem goes along the lines presented in the proof of [KU2, Theorem 1.3.] for the $SL(2, \mathbb{Z})$-reduction procedure using the $G$-code. Similarly

to that situation described also in Section 3, we define the $E$-code of an oriented geodesic $\gamma$ on $\mathcal{H}$ to be the $E$-code of a reduced geodesic $\Gamma(2)$-equivalent to $\gamma$, 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 $\gamma$ can be represented as a bi-infinite sequence of segments $\sigma_i$ between successive returns to $C_E$. To each segment $\sigma_i$ is associated the corresponding $E$-reduced geodesic $\gamma_i$, 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 $\gamma$ a bi-infinite coding sequence, defined up to a shift, which we call the*$E$-code* of $\gamma$ and denote by $(\!(\gamma)\!)$. A similar argument as for the $G$-code shows that the $E$-code is $\Gamma(2)$-invariant.

**Symbolic representation of geodesics for $E$-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 corresponds to its $E$-code, $(\!(\gamma)\!) \in \mathcal{N}^{\mathbb{Z}} \times \{0,1\}$. Conversely, 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 2'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 into cusps can be described essentially 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 $\gamma$ in $M_2$. The following theorem establishes similar property for the $E$-code and the geometric code.

**Theorem 7.3.** *For any geodesic in $M_2$ the $E$-code coincides with the geometric code.*

*Proof.* Take a geodesic $\gamma$ 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 $\gamma$) 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 $\gamma$ is

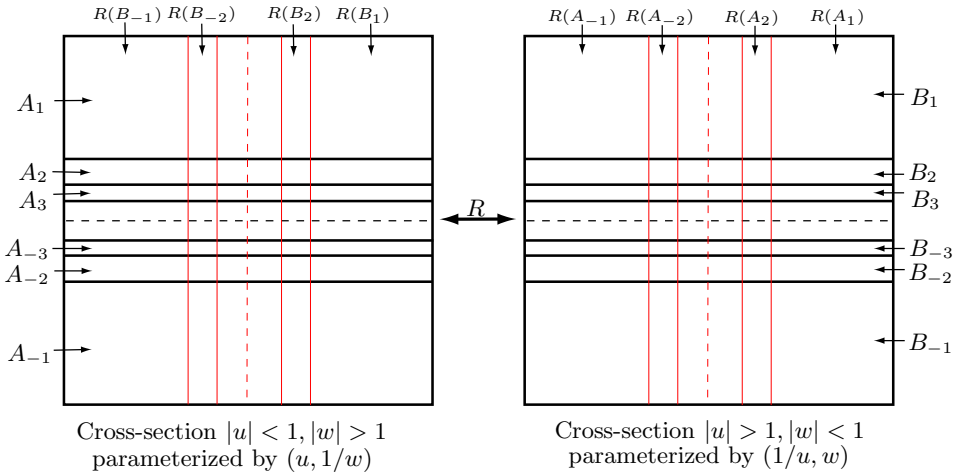


FIGURE 16. Infinite partition for the $E$-code and its image under the return map $R$

$A^{n_0}$. 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 $\infty$, 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 $-1/u_1 = 1/((2n_0, 2n_{-1}, \dots))$ and $-1/w_1 = ((2n_1, 2n_2, \dots))$. This brings us to the previous 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 coding sequence of $\gamma_1$ is $B^{n_1}$. Continuing by induction we obtain that the geometric code of $\gamma$ coincide with its $E$-code. □

This theorem shows that the space of admissible geometric codes of the geodesics on $M_2$ is (essentially) the entire symbolic space $\mathcal{N}^{\mathbb{Z}} \times \{0, 1\}$.

## REFERENCES

[Ab] L. M. Abramov, *On the entropy of a flow* (in Russian), Sov. Math. Doklady. 128, no. 5 (1959).

[Ad] R. Adler, *Symbolic dynamics and Markov partitions*, Bull. Amer. Math. Soc. **35** (1998), no. 1, 1–56.

[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.

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

[AF3] R. Adler, L. Flatto, *Geodesic flows, interval maps, and symbolic dynamics*, Bull. Amer. Math. Soc. **25** (1991), no. 2, 229–334.

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

[Arn] P. Arnoux, *Le codage des flot géodésique sur la surface modulaire*, Enseign. Math. **40** (1994), 29–48.

[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.

[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, Princeton, NJ, 1987.

[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).

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

[D] P. G. L. Dirichlet, *Vereinfachung der Theorie der binaren quadratischen Formen von positiver Determinante*, Abh. K. Akad. Wiss. Berlin, Math. (1854), 99–115; French transl., Jour. de Math., (2), 2, 1857, 353 (with additions); Werke, II, 139–158, 159–181. Zahlentheorie, §§72-82, 1863; English transl., American Mathematical Society, Providence, 1999

[F1] L. Ford, *Rational approximations to irrational complex numbers*, Trans. Amer. Math. Soc. **19** (1918), no. 1, 1–42.

[F2] L. Ford, *Fractions*, Amer. Math. Monthly **45** (1938), no. 9, 586–601.

[Fr] D. Fried, *Reduction theory over quadratic imaginary fields*, preprint.

[G] C. F. Gauss, *Disquisitiones Arithmeticae*, 1801; Werke, I, 1863; German transl. by H. Mase, 1889; French transl. by A.C.M. Poullet-Delisne, 1807; English transl. Springer-Verlag; Reissue edition, 1986.

[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.

[GrL] P.M. Gruber, C.G. Lekkerkerker, *Geometry of numbers*, North-Holland, 1987

[GK] B. Gurevich, S. Katok, *Arithmetic coding and entropy for the positive geodesic flow on the modular surface*, Moscow Mathematical Journal **1** (2001), no. 4, 569-582.

[Ha] J. Hadamard, *Les surfaces à courbures opposées et leurs lignes géodesiques*, Journal de Mathématiques pures et appliquées, (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.

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

[He3] G. A. Hedlund, *The dynamics of geodesic flows*, Bulletin of the American Mathematical Society **45** (1939) 241–260.

[Ho] E. Hopf, *Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung*, Berichte über die Verhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig, Mathematisch-Physikalische Klasse **91**, (1939), 261–304.

[Hu] M. G. Humbert, *Sur les fractions continues et les formes qudratiques 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 Reduktion der binaren quadratischen Formen*, Math. Ann **45** (1894), 85–117; Mathematische Werke, Birkhauser, Basel, 1933, Band II, 157–190

[KH] A. Katok, B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, 1995.

[K1] S. Katok, *Coding of closed geodesics after Gauss and Morse*, Geom. Dedicata **63** (1996), 123–145.

[K2] S. Katok, *Continued fractions, hyperbolic geometry and quadratic forms*, MASS Selecta, Amer. Math. Soc., 121–160, 2003.

[K3] S. Katok, *Fuchsian Groups*, University of Chicago Press, 1992.

[K4] S. Katok, *Reduction theory for Fuchsian groups*, Math. Ann. **273** (1985), 461–470.

[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.

[KU1] S. Katok, I. Ugarcovici, *Geometrically Markov geodesics on the modular surface*, to appear Moscow Mathematical Journal.

[KU2] S. Katok, I. Ugarcovici, *Arithmetic coding of geodesics on the modular surface via continued fractions*, to appear, CWI Tracts, Centrum voor Wiskunde en Informatica, Amsterdam, 2005.

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

[Mi] H. Minkowski, *Geomtrie der Zahlen*, Chelsea Publishing Company,New York, 1953

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

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

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

[O] D. Ornstein, *The isomorphism theorems for Bernoulli flows*, Advances in Math, **10** (1973), 124–142.

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

[PP] W. Parry, M. Pollicott, *Zeta functions and periodic orbit structure of hyperbolic dynamics*, Astérisque, **187–188**, 1990.

[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.

[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: Math. Notes, **6** (1969), 880–886

[Sa] S. V. Savchenko, *Special flows constructed from countable topological Markov chains*, Funktsional. Anal. i Prilozhen. **32** (1998), no. 1, 40–53, 96 (Russian). English translation in Funct. Anal. Appl. **32** (1998), no. 1, 32–41.

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

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

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

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

[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

[Z] D. Zagier, *Zetafunkionen und quadratische Körper: eine Einführung in die höhere Zahlentheorie*, Springer-Verlag, 1982.

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

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

Department of Mathematics, Rice University, Houston, TX 77005

*E-mail address*: idu@rice.edu