Published in CWI Tracts 135 (2005), Centrum Wisk. Informatica, Amsterdam, 59–77

ARITHMETIC CODING OF GEODESICS ON THE MODULAR SURFACE VIA CONTINUED FRACTIONS

SVETLANA KATOK AND ILIE UGARCOVICI

Abstract. In this article we present three arithmetic methods for coding oriented geodesics on the modular surface using various continued fraction expansions and show that the space of admissible coding sequences for each coding is a one-step topological Markov chain with countable alphabet. We also present conditions under which these arithmetic codes coincide with the geometric code obtained by recording oriented excursions into the cusp of the modular surface.

Introduction

Let $\mathcal{H} = \{z = x+iy : y > 0\}$ be the upper half-plane endowed with the hyperbolic metric, $F = \{z \in \mathcal{H} : |z| \geq 1,\ |\mathrm{Re}\, z| \leq \frac{1}{2}\}$ be the standard fundamental region for the modular group $PSL(2,\mathbb{Z}) = SL(2,\mathbb{Z})/\{\pm I\}$, and $M = PSL(2,\mathbb{Z})\backslash\mathcal{H}$ be the modular surface which topologically is a sphere with one puncture (the cusp) and two singularities (fixed points of elliptic elements). Let $S\mathcal{H}$ denote the unit tangent bundle of $\mathcal{H}$. Then the quotient space $PSL(2,\mathbb{Z})\backslash S\mathcal{H}$ 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 [K1, §3.6] for details). Let $\pi : S\mathcal{H} \to SM$ be the projection of the unit tangent bundles. 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 therefore is an orbit of the geodesic flow $\{\varphi^t\}$ on M, which is defined as an $\mathbb{R}$-action on the unit tangent bundle SM (see e.g. [KH, §5.3, 5.4]). For an oriented geodesic γ on M, its lift to $\mathcal{H}$ is any oriented geodesic γ' on $\mathcal{H}$ such that $\pi(\gamma') = \gamma$.

In this article we will consider only oriented geodesics which do not go to the cusp of M in either direction. The corresponding geodesics in F contain no vertical segments, and both end points of all their lifts to $\mathcal{H}$ are irrational. 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, more precisely, the set of vectors tangent to the excluded geodesics $E \subset SM$ is invariant under the geodesic flow $\{\varphi^t\}$ and $\mu(E) = 0$ for any Borel probability measure μ invariant under $\{\varphi^t\}$. This can be seen from the decomposition of this set $E = E^+ \cup E^-$, so that $\varphi^t(E^+)$ (respectively, $\varphi^{-t}(E^-)$) escape to the cusp as $t \to +\infty$. For any compact $K \subset E^\pm$ there exists $T > 0$ such that $K \cap \varphi^{\pm t}(K) = \emptyset$ for any $t > T$. $\mu(E^\pm) > 0$ would then contradict the Poincaré Recurrence Theorem (see [KH, §4.1]).

Date: July 1, 2004.
2000 *Mathematics Subject Classification.* Primary 37D40, 37B40, 20H05.
Key words and phrases. Modular surface, geodesic flow, continued fractions.

Oriented geodesics on the modular surface $M = PSL(2,\mathbb{Z})\backslash\mathcal{H}$ can be symbolically coded in two different ways. The geometric code with respect to the fundamental region F is obtained by recording the successive sides of F cut by the geodesic, and can be presented by a bi-infinite sequence of non-zero integers by assigning an integer, positive or negative, depending on the orientation, to each excursion to the cusp. Another method (we call it arithmetic) is to use the boundary expansions of the end points of the geodesic at infinity and a certain "reduction theory". This method was first introduced by Artin [Ar] for the modular group who used regular continued fractions for the boundary expansions to prove the existence of a dense geodesic on M. Artin's method was used by Hedlund [Hed] to prove ergodicity of the geodesic flow on M. 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, and further developed in [BS, S2, S3] for other Fuchsian groups. Related work on coding geodesics can be also found in [AF1, AF2, AF3, Arn, GL, S4].

In this article we give a unified approach for construction of arithmetic codes for geodesics on the modular surface using *generalized minus continued fractions*. 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 $n_{i+1} = (x_{i+1})$, $x_{i+1} = -\frac{1}{x_i - n_i}$, starting with $n_0 = (x)$ and $x_1 = -\frac{1}{x-n_0}$, where $(\cdot)$ is a certain integer-valued function. The function $x \mapsto \lceil x \rceil = \lfloor x \rfloor + 1$ (where $\lfloor x \rfloor$ is the integer part of x, or the floor function, i.e. the largest integer $\leq x$) gives the minus continued fraction expansion first used for the arithmetic code in [K2, GK], although the notations in the present paper are different from [K2, GK] where only one arithmetic code was studied. (Notice that $\lceil x \rceil$ is the smallest integer greater than x, and differs at integers from the commonly used ceiling function.) This coding procedure for closed geodesics is exactly the Gauss reduction theory for indefinite integral quadratic forms translated into matrix language [K2], therefore we will refer to the above code as the *Gauss arithmetic code (G-code)*. We review it in Section 1. Using appropriate functions $(\cdot)$ we reinterpret the classical Artin code *(A-code)* in these terms in Section 2 and describe an arithmetic code based on the nearest integer continued fraction expansions of the end points in Section 3. The latter expansions were developed and used by Hurwitz [H] in order to establish a reduction theory for indefinite real quadratic forms, therefore we call the third code *Hurwitz arithmetic 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 matrix language. Although they follow the same general scheme, the notion of reduced geodesic is different in each case, and so are the estimates in Theorems 1.3, 2.4, and 3.3.

The most elegant of the three codings is the Gauss arithmetic code obtained in [K2, 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 given by a set of simple rules which can be described with the help of a transition matrix of zeros and ones, and constitutes a one-step topological Markov chain with countable alphabet. An explicit canonical Markov partition of the corresponding cross-section is presented for each arithmetic code. Symbolic representation of the geodesic flow on M as a special flow for each code is given in Section 4.

In contrast, the set of admissible geometric coding sequences is quite complicated, and, as has been proved in [KU], is not a finite-step topological Markov chain (see [KH, §1.9] for exact definitions). Therefore, there are geodesics whose geometric code differs from any arithmetic code. It is worth noting that the H-code comes closest to the geometric code: we show that for the class of geometrically Markov geodesics—identified in [KU] as the maximal one-step topological Markov chain in the set of all admissible geometric codes—the H-code coincides with the geometric code.

Acknowledgments. We thank David Fried for bringing to our attention the paper by Hurwitz [H], and the referee for helpful suggestions which improved the presentation of this article. The second author acknowledges summer support from NSF grant DMS-9704776.

1. Minus continued fraction coding (Gauss coding)

In this section we review the arithmetic coding procedure for geodesics on the modular surface, using minus (or, backward) continued fraction expansions which we call here *G-expansions*. Every real number α has a unique G-expansion $\alpha = \lceil n_0, n_1, n_2, \dots \rceil$ with $n_0 \in \mathbb{Z}$ and $n_1, n_2, \dots \geq 2$, by setting $n_0 = \lceil \alpha \rceil$ (the smallest integer greater than α), $\alpha_1 = -\frac{1}{\alpha - n_0}$, and, inductively,

$$n_i = \lceil \alpha_i \rceil \quad , \quad \alpha_{i+1} = -\frac{1}{\alpha_i - n_i} \, .$$

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 $\lceil n_0, n_1, n_2, \dots \rceil$. The following properties are satisfied (see [Z, K2], and [K3] for the proofs):

- (G1) α is rational if and only if the tail of its G-expansion consists only of 2's, i.e., there exists a positive integer l such that $n_k = 2$ for all $k \geq l$;
- (G2) α is a quadratic irrationality, i.e. a root of a quadratic polynomial with integer coefficients, if and only if its G-expansion is eventually periodic, $\alpha = \lceil n_0, n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}} \rceil$ (with the periodic part being anything but a tail of 2's);
- (G3) A quadratic irrationality α has a purely periodic G-expansion if and only if $\alpha > 1$ and $\alpha' \in (0, 1)$, where α' is conjugate to α, i.e. α' and α are roots of the same quadratic polynomial with integer coefficients;
- (G4) If $\alpha = \lceil \overline{n_1, \dots, n_k} \rceil$, then $1/\alpha' = \lceil \overline{n_k, \dots, n_1} \rceil$;

(G5) Two irrationals α, β are $PSL(2,\mathbb{Z})$-equivalent if and only if their G-expansions have the same tail, that is $\alpha = \lceil n_0, n_1, \dots \rceil$ and $\beta = \lceil m_0, m_1, \dots \rceil$ with $n_{i+k} = m_{i+l}$ for some integers k, l and all $i \geq 0$.

¿From the theory of G-expansions (see [K3]), we have that if $\alpha = \lceil n_0, n_1, \dots \rceil$, then the convergents $r_k = \lceil n_0, n_1, \dots, n_k \rceil$ can be written as p_k/q_k where p_k and q_k are obtained inductively as:

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

Proposition 1.1. *The following properties are satisfied:*

(i) $1 = q_0 < q_1 < q_2 < \dots$;

(ii) $p_{k-1}q_k - p_k q_{k-1} = 1$, *for all* $k \geq 0$;

(iii) *Let* $T(z) = z + 1$, $S(z) = -1/z$ *be the generating transformations for* $PSL(2,\mathbb{Z})$, *then for any* $z \in \mathcal{H} \cup \mathbb{R} \cup \{\infty\}$

$$T^{n_0} S T^{n_1} S \dots T^{n_k} S(z) = \lceil n_0, n_1, \dots, n_k, z \rceil = \frac{p_k z - p_{k-1}}{q_k z - q_{k-1}}\,;$$

(iv) *The sequence* $\{r_k\}$ *is monotone decreasing, converges to* α *and*

$$p_k/q_k - \alpha \leq 1/q_k\,;$$

(v) *If* α *is irrational, then there is a sequence of denominators* $\{q_{k_j}\}$ *such that* $\dfrac{q_{k_j}}{q_{k_{j-1}}} > 2$.

Proof. Proofs of the properties (i)–(iv) can be found in [K3]. We give a proof of (v) here. Since α is irrational, its G-expansion contains infinitely many entries strictly greater than 2, hence we can find a sequence k_j, such that $n_{k_j} \geq 3$. But this implies that

$$q_{k_j} = n_{k_j} q_{k_j-1} - q_{k_j-2} > 3q_{k_j-1} - q_{k_j-1} = 2q_{k_j-1}.$$

Using (i) we obtain

$$\frac{q_{k_j}}{q_{k_{j-1}}} \geq \frac{q_{k_j}}{q_{k_j-1}} > 2.$$

□

Definition 1.2. An oriented geodesic on $\mathcal{H}$ is called *G-reduced* if its repelling and attracting end points, denoted by u and w, respectively, satisfy $0 < u < 1$ and $w > 1$.

To a G-reduced geodesic γ, one associates a bi-infinite sequence of positive integers $\lceil \gamma \rceil = \lceil \dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots \rceil$, called its *G-code*, by juxtaposing the G-expansions of $1/u = \lceil n_{-1}, n_{-2}, \dots \rceil$ and $w = \lceil n_0, n_1, n_2, \dots \rceil$.

Reduction algorithm. We present the procedure of reducing any geodesic to a G-reduced one. This will help us extend the symbolic coding to all geodesics on $\mathcal{H}$.

Theorem 1.3. *Every oriented geodesic on* $\mathcal{H}$ *is* $PSL(2,\mathbb{Z})$*-equivalent to a G-reduced geodesic.*

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

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

Since w is irrational, $w_{k+1} = \lceil n_{k+1}, n_{k+2}, \dots \rceil > 1$. By Proposition 1.1 (iii),

$$w = T^{n_0} S T^{n_1} S \dots T^{n_k} S(w_{k+1}) = \frac{p_k w_{k+1} - p_{k-1}}{q_k w_{k+1} - q_{k-1}}$$
$$u = T^{n_0} S T^{n_1} S \dots T^{n_k} S(u_{k+1}) = \frac{p_k u_{k+1} - p_{k-1}}{q_k u_{k+1} - q_{k-1}},$$

hence

$$u_{k+1} = \frac{q_{k-1}u - p_{k-1}}{q_k u - p_k} = \frac{q_{k-1}}{q_k} + \frac{1}{q_k^2(p_k/q_k - u)} = \frac{q_{k-1}}{q_k} + \varepsilon_k \tag{1.1}$$

where $\varepsilon_k \to 0$. Moreover, using property (iv), we have $p_k/q_k \searrow w$, hence, for large enough k, $|p_k/q_k - u| > \frac{1}{2}|w - u|$ and

$$|\varepsilon_k| = \frac{1}{q_k^2 |p_k/q_k - u|} < \frac{2}{q_k^2 |w - u|}.$$

Property (i) and the previous relation imply that, for large enough k, $|\varepsilon_k| < 1/q_k$ and

$$0 < \frac{q_{k-1}}{q_k} - \frac{1}{q_k} < u_{k+1} = \frac{q_{k-1}}{q_k} + \varepsilon_k < \frac{q_{k-1}}{q_k} + \frac{1}{q_k} \le 1.$$

Therefore, we can find a positive integer l such that $0 < u_{l+1} < 1$. The geodesic with end points u_{l+1} and w_{l+1} is G-reduced and $PSL(2,\mathbb{Z})$-equivalent to γ. □

Remark 1.4. (i) The proof of Theorem 1.3 gives also the algorithm for G-reducing a geodesic γ: one has to construct the sequence $\{(u_k, w_k)\}$ inductively until $0 < u_k < 1$; (ii) any further application of the reduction algorithm to a reduced geodesic yields reduced geodesics whose G-codes are left shifts of the G-code of the first reduced one.

Now we associate to any oriented geodesic γ on $\mathcal{H}$ the G-code of a reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, e.g. obtained by the reduction algorithm described in the proof of Theorem 1.3. Since our goal is to define a symbolic coding for the geodesic flow on M, we need to show that this code is $PSL(2,\mathbb{Z})$-invariant, i.e. that two oriented geodesics on $\mathcal{H}$ are $PSL(2,\mathbb{Z})$-equivalent if and only if their G-codes coincide up to a shift. We are going to present a geometric proof of this fact in Corollary 1.6, by constructing a cross-section of the geodesic flow on M, directly related to the notion of G-reduced geodesics.

Construction of the cross-section. A cross-section for the geodesic flow is a subset of the unit tangent bundle SM which each geodesic (maybe with some exceptions) visits infinitely often both in the future and in the past. We construct a cross-section C_G for the geodesic flow on M, such that successive returns of a geodesic γ to C_G correspond to left-shifts in the G-code of γ. We define $C_G = P \cup Q$ to be a subset of SM, where P consists of all tangent vectors with base points in the circular side of F and pointing inward such that the corresponding geodesic on $\mathcal{H}$ is G-reduced, i.e. $0 < u < 1$ and $w > 1$ and Q consists of all tangent vectors with base points on the right vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $TS(\gamma)$ is G-reduced (Figure 1). This is a clarification of the definition given in [GK]. Notice that $C_G = \pi(C_g)$ where C_g is the set 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 G-reduced.

Theorem 1.5. *C_G is a cross-section for the geodesic flow on M.*

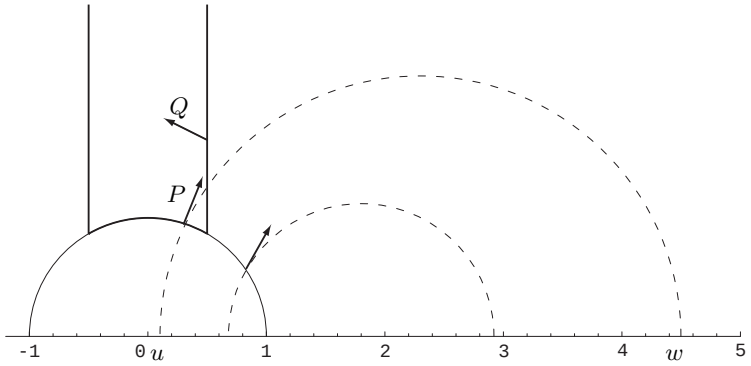

FIGURE 1. The cross-section $C_G = P \cup Q$

Proof. Let γ be an oriented geodesic on M. It is presented as a bi-infinite sequence of $PSL(2,\mathbb{Z})$-equivalent geodesic segments on F. Any segment, extended to a geodesic on $\mathcal{H}$, can be reduced according to Theorem 1.3. Thus, there exists a G-reduced geodesic γ' on $\mathcal{H}$ such that $\pi(\gamma') = \gamma$. Notice that γ' intersects the right half of the unit semicircle $|z| = 1$ in such a way that the unit tangent vector of γ' at the intersection point belongs to the set C_g. Therefore either $\gamma' \cap F$ or $ST^{-1}(\gamma') \cap F$ is one of the segments of γ on F. In either case γ intersects C_G at least once. Denote this intersection point by $\mathbf{x}_0 \in C_G \subset SM$, and let us follow the geodesic on M from this starting point. If $\mathbf{x}_0 \in P$, then the corresponding geodesic γ' on $\mathcal{H}$ from u to w is G-reduced. In order to prove that γ intersects C_G again, it is enough to notice that γ' intersects the left-half of the semi-circle $|z - n_0| = 1$, where $n_0 = \lceil w \rceil$, such that for the unit tangent vector at the intersection point $\mathbf{x}_1' \in S\mathcal{H}$, we have $ST^{-n_0}(\mathbf{x}_1') \in C_g$. Hence $\mathbf{x}_1 = \pi(\mathbf{x}_1') \in C_G$, and γ intersects C_G at $\mathbf{x}_1$. Moreover, $\mathbf{x}_1 \in C_G$ is the next intersection point of γ with C_G after $\mathbf{x}_0$. One obtains a similar property in the case where $\mathbf{x}_0 \in Q$, by studying the G-reduced geodesic γ' on $\mathcal{H}$ corresponding to $TS(\mathbf{x}_0)$. □

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

Corollary 1.6. *The G-code is $PSL(2,\mathbb{Z})$-invariant, i.e. two geodesics γ, γ' 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 $\lceil \gamma \rceil = \lceil n_i \rceil_{i=-\infty}^{\infty}$ and $\lceil \gamma' \rceil = \lceil n_i' \rceil_{i=-\infty}^{\infty}$.*

Proof. Let γ, γ' be $PSL(2,\mathbb{Z})$-equivalent. Then $\pi(\gamma) = \pi(\gamma')$ is the same oriented geodesic on M. By choosing the same starting point $\mathbf{x_0}$, one obtains the same bi-infinite sequence of segments σ_i between successive returns to C_G and hence the same G-code up to a left shift. Conversely, a left shift of a G-code corresponds

to an application of ST^{-n_0} to the end points of the geodesic, i.e., it produces a $PSL(2,\mathbb{Z})$-equivalent reduced geodesic. □

Example 1.7. Let γ be a geodesic on $\mathcal{H}$ from $u = \sqrt{5}$ to $w = -\sqrt{3}$. The G-expansions are

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

First, we need to find an equivalent G-reduced geodesic. For this we use the algorithm described in the proof of Theorem 1.3 to 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 minus continued fraction 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 $\lceil \gamma \rceil = \lceil \overline{2,6,2,2}, 3, \overline{2,3} \rceil = \lceil \dots, 2, 2, 6, 2, 2, 2, 6, 2, 2, 3, 2, 3, \dots \rceil$. This corrects a misprint in the Example of [GK].

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

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

This correspondence will extend to all oriented geodesics on M if we extend the notion of G-reduced geodesic to those with $0 < u < 1$ and $w \geq 1$, as can be easily seen from the proof of Theorem 1.3. For example, a geodesic which goes from the cusp down to the point $i \in \partial F$ and back to the cusp will be coded by the sequence $\lceil \bar{2}, 3, \bar{2} \rceil$. Thus the set of all oriented geodesics on M can be described symbolically as the Bernoulli space (minus one point) $X_G = \mathcal{N}_G^{\mathbb{Z}} \setminus \lceil \bar{2} \rceil$.

The partition of the cross-section. The infinite partition of the cross-section C_G corresponding to the G-code can be constructed as follows. We parameterize the cross-section C_g by (ϕ, θ), where $\phi \in [0, \pi/2]$ parameterizes the circle arc (counterclockwise) and $\theta \in [-\pi/2, \pi/2]$ is the angle the unit vector makes with the positive horizontal axis (counterclockwise). The angle θ depends on the position ϕ and is determined by the condition that the corresponding geodesic is G-reduced.

The partition of C_g (and that of C_G obtained by projection) corresponding to the arithmetic G-code ("the horizontal triangles") and its iteration under the first return map R to the cross-section C_g ("the vertical triangles") is shown on Figure 2. Its elements ("the horizontal triangles") are labeled by the symbols of the alphabet $\mathcal{N}_G$, $C_g = \sqcup_{n \in \mathcal{N}_G} C_n$ and are defined by the following condition: C_n consists of all tangent vectors $\mathbf{x}$ in C_g such that the corresponding geodesic in $\mathcal{H}$ goes from $0 < u < 1$ to $n - 1 < w < n$, i.e., if x is its coding sequence, then $n_0(x) = n$.

We also observe that the elements C_m and $R(C_n)$ intersect transversally for all $n, m \geq 2$, thus, according to Theorem 7.9 of [Ad], the infinite partition is Bernoulli.

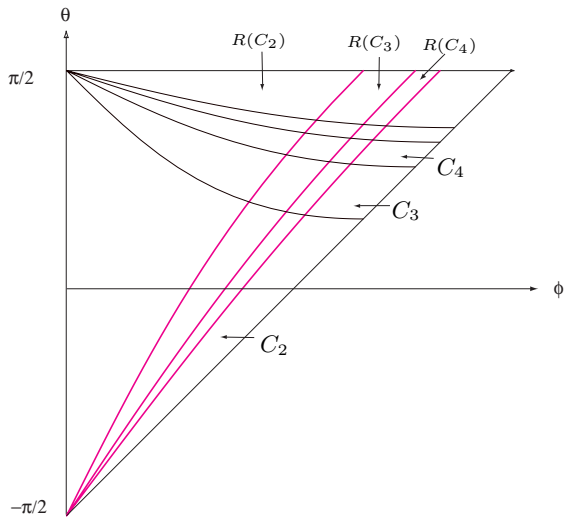

FIGURE 2. Infinite partition for the G-code and its image under the return map R

This gives an alternative geometric way to see that all arithmetic coding sequences are realized.

When does the G-code coincide with the geometric code? The relation between the geometric code and the arithmetic G-code of an oriented geodesic on M was established in [K2, GK]: *the geometric code and the arithmetic G-code of a geodesic γ on M coincide if and only if $\frac{1}{n_i} + \frac{1}{n_{i+1}} \leq \frac{1}{2}$, where $\lceil \gamma \rceil = \lceil n_i \rceil_{i=-\infty}^{\infty}$.*

2. Alternating continued fraction coding (Artin coding revisited)

In this section we describe the arithmetic coding of geodesics on the modular surface, using alternating continued fraction expansions which we call *A-expansions*. The result will be a modified Artin code, as described by Series [S1]. Every irrational number α has a unique A-expansion

$$\alpha := \lceil n_0, n_1, n_2, \dots \rfloor = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots}}}$$

with $n_0 \in \mathbb{Z}$ and $|n_i| \geq 1$, by setting $n_0 = \lceil \alpha \rfloor$, $\alpha_1 = -\frac{1}{\alpha - n_0}$, and, inductively,

$$n_i = \lceil \alpha_i \rfloor \ , \ \alpha_{i+1} = -\frac{1}{\alpha_i - n_i} \, , \qquad \text{where } \lceil \alpha \rfloor = \begin{cases} \lfloor \alpha \rfloor & \text{if } \alpha > 0 \\ \lceil \alpha \rceil & \text{if } \alpha < 0 \, . \end{cases}$$

Notice that $n_i n_{i+1} < 0$, hence the use of terminology of alternating continued fractions. 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 $\lceil n_0, n_1, n_2, \dots \rfloor$.

Remark 2.1. The properties of A-expansions can be easily established if one notices the relation with regular continued fraction expansions: if $\alpha > 0$ then

$$\alpha = \lceil n_0, n_1, n_2, \dots \rfloor = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots}}} = n_0 + \cfrac{1}{-n_1 + \cfrac{1}{n_2 + \cfrac{1}{\ddots}}} = \lfloor m_0, m_1, \dots \rfloor$$

where $m_i = (-1)^i n_i$, and $\lfloor m_0, m_1, \dots \rfloor$ is the regular continued fraction expansion of α.

The following properties are proved using the corresponding properties of the regular continued fractions, see e.g. [O]:

(A1) α is a quadratic irrationality if and only if its A-expansion is eventually periodic, $\alpha = \lceil n_0, n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}} \rfloor$;
(A2) A quadratic irrationality α has a purely periodic A-expansion if and only if $|\alpha| > 1$ and $-1 < \operatorname{sgn}(\alpha)\alpha' < 0$, where α' is conjugate to α;
(A3) If $\alpha = \lceil \overline{n_1, \dots, n_k} \rfloor$, then $1/\alpha' = \lceil \overline{n_k, \dots, n_1} \rfloor$;
(A4) Two irrationals α, β are $PSL(2, \mathbb{Z})$-equivalent if and only if their A-expansions have the same tail.

Similarly to the theory of minus continued fraction expansions, if $\alpha = \lceil n_0, n_1, \dots \rfloor$, then the partial fractions $r_k = \lceil n_0, n_1, \dots, n_k \rfloor$ can be written as p_k/q_k, where p_k and q_k are obtained inductively as:

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

Proposition 2.2. *The following properties are satisfied:*

(i) $1 = q_0 \leq |q_1| < |q_2| < \dots$;
(ii) $p_{k-1}q_k - p_k q_{k-1} = 1$, *for all* $k \geq 0$;
(iii) *Let* $T(z) = z + 1$, $S(z) = -1/z$ *be the generating transformations for* $PSL(2, \mathbb{Z})$, *then for any* $z \in \mathcal{H} \cup \mathbb{R} \cup \{\infty\}$

$$T^{n_0} S T^{n_1} S \dots T^{n_k} S(z) = \lceil n_0, n_1, \dots, n_k, z \rceil = \frac{p_k z - p_{k-1}}{q_k z - q_{k-1}}\,;$$

(iv) *The sequence* $\{r_k\}$ *converges to* α *and* $|p_k/q_k - \alpha| \leq 1/q_k^2$;
(v) *If* $\alpha > 0$, *then either* $\frac{q_{2k}}{q_{2k-1}} > \sqrt{2}$ *or* $\frac{q_{2k+1}}{q_{2k}} < -\sqrt{2}$; *if* $\alpha < 0$, *then either* $\frac{q_{2k}}{q_{2k-1}} < -\sqrt{2}$ *or* $\frac{q_{2k+1}}{q_{2k}} > \sqrt{2}$

Proof. The properties (i)–(iv) are proved similarly to those in Proposition 1.1, so we prove (v). We assume that $\alpha > 0$ (the case $\alpha < 0$ can be treated in a similar way). Then

$$1 = q_0 \leq -q_1 < -q_2 < q_3 < q_4 < -q_5 < -q_6 < q_7 < q_8 < \dots\,,$$

thus $0 < q_{2k-1}/q_{2k} < 1$, and $-1 < q_{2k}/q_{2k+1} < 0$. Taking into consideration the order of the signs and the fact that $q_k = n_k q_{k-1} - q_{k-2}$, one obtains

$$|q_k| = |n_k| \cdot |q_{k-1}| + |q_{k-2}|\,. \tag{2.1}$$

Indeed, if $k = 4m$, then $q_k > 0$, $q_{k-1} > 0$, $q_{k-2} < 0$, $n_k > 0$, and (2.1) follows. (The cases $k = 4m+1, 4m+2, 4m+3$ can be treated similarly.) ¿From (2.1) and

property (i), we get

$$|q_k| \geq |n_k| \cdot |q_{k-2}| + |q_{k-2}| = |q_{k-2}|(|n_k| + 1) \geq 2|q_{k-2}| \;\Rightarrow\; \frac{|q_{k-2}|}{|q_k|} \leq \frac{1}{2}$$

and since $\dfrac{q_{k-2}}{q_k} < 0$,

$$-\frac{1}{2} \leq \frac{q_{k-2}}{q_k} = \frac{q_{k-2}}{q_{k-1}} \cdot \frac{q_{k-1}}{q_k} < 0.$$

Therefore we either have

$$0 < \frac{q_{2k-1}}{q_{2k}} \leq \frac{1}{\sqrt{2}} \quad \text{or} \quad -\frac{1}{\sqrt{2}} \leq \frac{q_{2k}}{q_{2k+1}} < 0.$$

□

Definition 2.3. An oriented geodesic on $\mathcal{H}$ is called *A-reduced* if its repelling and attracting end points, denoted by u and w, respectively, satisfy $|w| > 1$ and $-1 < \operatorname{sgn}(w)u < 0$.

To an A-reduced geodesic γ, one associates a bi-infinite sequence of nonzero integers (with alternating signs) $\lceil\gamma\rfloor = \lceil \ldots, n_{-2}, n_{-1}, n_0, n_1, n_2, \ldots \rfloor$, called its *A-code*, by juxtaposing the A-expansions of $1/u = \lceil n_{-1}, n_{-2}, \ldots \rfloor$ and $w = \lceil n_0, n_1, n_2, \ldots \rfloor$.

Reduction algorithm. The following theorem extends this symbolic coding to all geodesics on $\mathcal{H}$.

Theorem 2.4. *Every oriented geodesic on $\mathcal{H}$ is $PSL(2,\mathbb{Z})$-equivalent to an A-reduced geodesic.*

Proof. Let γ be an arbitrary geodesic on $\mathcal{H}$, with irrational end points u and w. Let $\lceil n_0, n_1, n_2 \ldots \rfloor$ be the alternating continued fraction expansion of w. We construct the following sequence of real pairs $\{(u_k, w_k)\}$ $(k \geq 0)$ defined by $u_0 = u$, $w_0 = w$ and:

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

Notice that $w_{k+1} = \lceil n_{k+1}, n_{k+2}, \ldots \rfloor$, $|w_{k+1}| > 1$, and by Proposition 2.2 (iii),

$$w = T^{n_0} ST^{n_1} S \ldots T^{n_k} S(w_{k+1}) = \frac{p_k w_{k+1} - p_{k-1}}{q_k w_{k+1} - q_{k-1}}$$

$$u = T^{n_0} ST^{n_1} S \ldots T^{n_k} S(u_{k+1}) = \frac{p_k u_{k+1} - p_{k-1}}{q_k u_{k+1} - q_{k-1}}$$

hence

$$u_{k+1} = \frac{q_{k-1}u - p_{k-1}}{q_k u - p_k} = \frac{q_{k-1}}{q_k} \cdot \frac{u - \frac{p_{k-1}}{q_{k-1}}}{u - \frac{p_k}{q_k}} = \frac{q_{k-1}}{q_k} \cdot \delta_k \tag{2.2}$$

where $\delta_k \to 1$. If $w > 0$, we have $w_{2k} > 1$ and $w_{2k+1} < -1$. By Proposition 2.2 (v), one can find a positive integer l such that either $u_{2l} \in (-1, 0)$ or $u_{2l+1} \in (0, 1)$. Then either a geodesic from u_{2l} to w_{2l} or a geodesic from u_{2l+1} to w_{2l+1} is A-reduced and $PSL(2,\mathbb{Z})$-equivalent to γ. The case $w < 0$ is treated similarly. □

Remark 2.5. (i) The proof of Theorem 2.4 gives also the algorithm for A-reducing a geodesic γ: one has to construct inductively the sequence $\{(u_k, w_k)\}$ until $|w_k| > 1$ and $\operatorname{sgn}(w_k)u_k \in (-1, 0)$; (ii) any further application of the reduction algorithm to an A-reduced geodesic yields reduced geodesics whose A-codes are left shifts of the A-code of the first reduced one.

Similarly to the situation in Section 1, we define the A-code of an oriented geodesic γ on $\mathcal{H}$ to be the A-code of a reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, and prove its $PSL(2,\mathbb{Z})$-invariance by constructing a cross-section of the geodesic flow on M, directly related to the notion of A-reduced geodesics.

Construction of the cross-section. We describe the cross section C_A for the geodesic flow on M, such that successive returns to the cross section correspond to left-shifts in the arithmetic A-code. Let $C_A = P \cup Q_1 \cup Q_2$ be a subset of the unit tangent bundle SM, where P consists of all tangent vectors with base points in the circular side of F and pointing inward such that the corresponding geodesic is A-reduced; Q_1 consists of all tangent vectors with base points on the right vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $TS(\gamma)$ is A-reduced; Q_2 consists of all tangent vectors with base points on the left vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $T^{-1}S(\gamma)$ is A-reduced. Notice that $C_A = \pi(C_a)$ where C_a is the set 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 A-reduced (Figure 3).

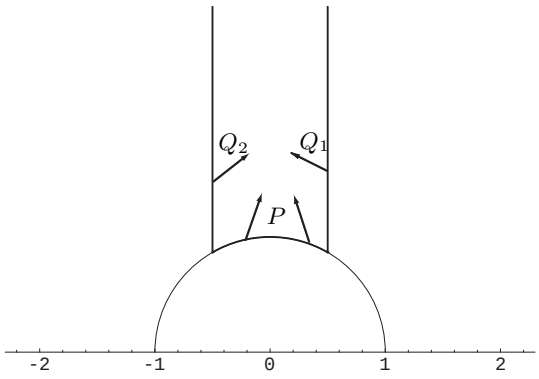

FIGURE 3. The cross-section $C_A = P \cup Q_1 \cup Q_2$

One can show similarly to the proof of Theorem 1.5 that $C_A = P \cup Q_1 \cup Q_2$ is indeed a cross-section for the geodesic flow on M, hence every geodesic γ can be represented as a bi-infinite sequence of segments σ_i between successive returns to C_A. To each segment σ_i is associated the corresponding A-reduced geodesic γ_i, so that $\lceil \gamma_{i+1} \rfloor$ differs from $\lceil \gamma_i \rfloor$ by a left shift. Thus we associate to γ a bi-infinite coding sequence, defined up to a shift, which we call the*A-code* of γ and denote by $\lceil \gamma \rfloor$. The argument of Corollary 1.6 shows that the A-code is $PSL(2,\mathbb{Z})$-invariant.

The set of all oriented geodesics on M can be described symbolically as a countable 1-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,

$$A(n,m) = \begin{cases} 1 & \text{if } nm < 0, \\ 0 & \text{otherwise}. \end{cases} \tag{2.3}$$

Each oriented geodesic γ corresponds to its A-code, $\lceil \gamma \rfloor \in X_A$ and each bi-infinite sequence of nonzero integers with alternating signs $x \in X_A$ produces a geodesic on

$\mathcal{H}$ from $u(x)$ to $w(x)$, where

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

The partition of the cross-section. The infinite partition of the cross-section C_A corresponding to the A-code can be constructed as follows. We parameterize the cross-section C_a by (ϕ, θ), where $\phi \in [0, \pi]$ parameterizes the unit semicircle (counterclockwise) and $\theta \in [-\pi/2, (3\pi)/2]$ is the angle the unit vector makes with the positive horizontal axis (counterclockwise). The angle θ depends on ϕ and is determined by the condition that the corresponding geodesic is A-reduced.

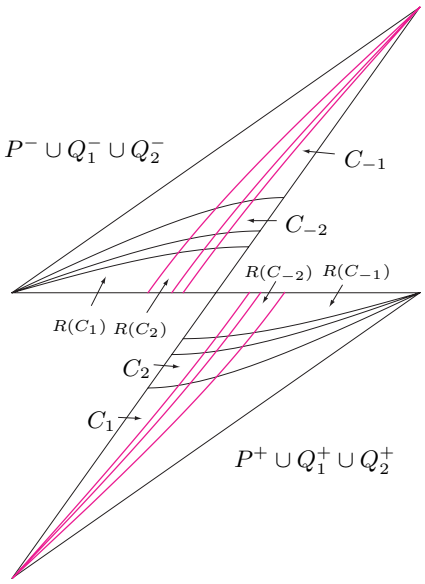

FIGURE 4. Infinite partition for the A-code and its image under the return map R

The partition of C_a (and therefore of C_A by projection) corresponding to the arithmetic A-code ("the horizontal triangles") and its iteration under the first return map R to the cross-section C_a ("the vertical triangles") is shown on Figure 4. Its elements ("the horizontal triangles") are labeled by the symbols of the alphabet $\mathcal{N}_A$, $C_a = \sqcup_{n \in \mathcal{N}_A} C_n$ and are defined by the following condition: $C_n = \{\mathbf{x} \in C_a, n_0(\mathbf{x}) = n\}$, i.e. it consists of all tangent vectors $\mathbf{x}$ in C_a such that the coding sequence $x \in X_A$ of the corresponding geodesic with this initial vector has its first symbol in the A-code $n_0(x) = n$. Thus, for $n \geq 1$, C_n consists of all tangent vectors $\mathbf{x} \in C_a$ such that the corresponding geodesic goes from $-1 < u < 0$ to $n - 1 < w < n$. We call this part of the cross section the positive part, and denote it by C_a^+ (and let $P^+ \cup Q_1^+ \cup Q_2^+$ denote its corresponding projection on C_A). For $n \leq -1$, C_n consists of all tangent vectors $\mathbf{x} \in C_a$ such that the corresponding geodesic goes from $0 < u < 1$ to $n - 1 < w < n$. We call this part of the cross section the negative part, and denote it by C_a^- (with $P^- \cup Q_1^- \cup Q_2^-$ denoting $\pi(C_a^-)$).

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. If $n_0(\mathbf{x}) = n$ and $n_1(\mathbf{x}) = m$ for some $\mathbf{x} \in C_A$, then $R(C_n) \cap C_m \neq \emptyset$. Therefore, as follows from Figure 4, the signs in the A-code must alternate, because $R(C_n) \cap C_m \neq \emptyset \Leftrightarrow nm < 0$. Moreover, all intersections are transversal, hence, according to Theorem 7.9 of [Ad], the partition is Markov.

When does the A-code coincide with the geometric code? The next theorem gives a sufficient condition for the geometric code and the arithmetic A-code of a geodesic γ on M to coincide.

Theorem 2.6. *The geometric code and the arithmetic A-code of a geodesic γ coincide if $|n_i| \geq 2$ and $n_i n_{i+1} < 0$ where $\lceil \gamma \rfloor = \lceil n_i \rfloor_{i=-\infty}^{\infty}$.*

Proof. Let $x = \{\dots, n_{-2}, n_{-1}, n_0, n_1, \dots\}$ be a sequence of integers with $|n_i| \geq 2$ and $n_i n_{i+1} < 0$. Consider the geodesic $\gamma(x)$ on $\mathcal{H}$, from

$$u(x) = \frac{1}{(n_{-1}, n_{-2}, \dots)} = \cfrac{1}{n_{-1} - \cfrac{1}{n_{-2} - \cfrac{1}{\ddots}}} \text{ to } w(x) = (n_0, n_1, \dots) = n_0 - \cfrac{1}{n_1 - \cfrac{1}{\ddots}}.$$

Since $x \in X_A$, the A-code of $\gamma(x)$ is $\lceil \gamma(x) \rfloor = \lceil n_{-2}, n_{-1}, n_0, n_1, \dots \rfloor$. We showed in [KU, Theorem 1.4], that if a sequence x satisfies $|n_i| \geq 2$ and $|\frac{1}{n_i} + \frac{1}{n_{i+1}}| \leq \frac{1}{2}$, then such geodesic $\gamma(x)$ from $u(x)$ to $w(x)$ has the geometric code $[\gamma(x)] = [\dots, n_{-1}, n_0, n_1, n_2, \dots]$. Therefore the geometric code and the A-code of $\pi(\gamma(x))$ coincide (up to a shift). □

3. Nearest integer continued fraction coding (Hurwitz coding)

In this section we describe the arithmetic coding procedure for geodesics on the modular surface, using the nearest integer continued fraction expansions, and the corresponding reduction theory for real quadratic forms with positive discriminant (indefinite real quadratic forms) developed by Hurwitz [H] (see also [F]). Every irrational number α has a unique H-expansion $\alpha = \langle n_0, n_1, n_2, \dots \rangle$ with $n_0 \in \mathbb{Z}$ and $|n_i| \geq 2$ for $i \geq 1$, by setting $n_0 = \langle \alpha \rangle$ (the nearest integer to α), $\alpha_1 = -\frac{1}{\alpha - n_0}$, and, inductively,

$$n_i = \langle \alpha_i \rangle \quad , \quad \alpha_{i+1} = -\frac{1}{\alpha_i - n_i}.$$

Notice that if $n_i = \pm 2$, then $n_i n_{i+1} < 0$. 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 $\langle n_0, n_1, n_2, \dots \rangle$. The following properties are satisfied (see [H] for more details):

- (H1) α is a quadratic irrationality, i.e. a root of a quadratic polynomial with integer coefficients, if and only if its H-expansion is eventually periodic, $\alpha = \langle n_0, n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}} \rangle$;
- (H2) A quadratic irrationality α has a purely periodic H-expansion if and only if $|\alpha| > 2$ and $\operatorname{sgn}(\alpha)\alpha' \in [r - 1, r]$, where α' is conjugate to α and $r = (3 - \sqrt{5})/2$;

(H3) Two irrationals α, β are $PSL(2,\mathbb{Z})$-equivalent if and only if their H-expansions have the same tail or one has $1/r = \langle\overline{3}\rangle$ as a tail, and the other has $-1/r = \langle\overline{-3}\rangle$ as a tail.

Definition 3.1. An oriented geodesic on $\mathcal{H}$ is called *H-reduced* if its repelling and attracting end points, denoted by u and w, respectively, satisfy $|w| > 2$ and $\mathrm{sgn}(w)u \in [r-1, r]$, where $r = (3-\sqrt{5})/2$.

We remark that the H-expansion satisfies an asymmetric restriction (if $n_i = \pm 2$, then $n_i n_{i+1} < 0$), and the statement "if $\alpha = \langle\overline{n_1, \ldots, n_k}\rangle$, then $1/\alpha' = \langle\overline{n_k, \ldots, n_1}\rangle$" is not always true. For example, if one considers the conjugate quadratic irrationalities $\alpha = (15+12\sqrt{2})/7$ and $\alpha' = (15-12\sqrt{2})/2$, then $\langle\alpha\rangle = \langle\overline{5, 2, -3}\rangle$, but $\langle\alpha'\rangle = \langle\overline{-4, -2, 4}\rangle$. For that reason, we cannot construct a meaningful symbolic sequence for an H-reduced geodesic just by juxtaposing the H-expansions of w and $1/u$. In order to associate to an H-reduced geodesic a bi-infinite sequence of integers, we use a different expansion for $1/u$ introduced by Hurwitz, and called the *H-dual expansion*. Every irrational α has a unique H-dual expansion $\alpha = \langle\!\langle n_0, n_1, n_2, \ldots\rangle\!\rangle$ with $n_0 \in \mathbb{Z}$ and $|n_i| \geq 2$ for $i \geq 1$, given by $n_0 = \langle\!\langle\alpha\rangle\!\rangle$, $\alpha_1 = -\frac{1}{\alpha - n_0}$ and, inductively,

$$n_i = \langle\!\langle\alpha_i\rangle\!\rangle \quad , \quad \alpha_{i+1} = -\frac{1}{\alpha_i - n_i}\,,$$

where

$$\langle\!\langle\alpha\rangle\!\rangle = \begin{cases} \langle\alpha\rangle - \mathrm{sgn}(\alpha) & \text{if} \quad \mathrm{sgn}(\alpha)(\langle\alpha\rangle - \alpha) > r \\ \langle\alpha\rangle & \text{otherwise} \end{cases}$$

Notice that if $n_{i+1} = \pm 2$, then $n_i n_{i+1} < 0$, and moreover if $\alpha = \langle\overline{n_1, \ldots, n_k}\rangle$, then $1/\alpha' = \langle\!\langle\overline{n_k, \ldots, n_1}\rangle\!\rangle$.

If $\alpha = \langle n_0, n_1, \ldots\rangle$, then the convergents $r_k = \langle n_0, n_1, \ldots, n_k\rangle$ 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 proof of the following Proposition is contained in [H].

Proposition 3.2. *The following properties are satisfied:*

(i) $1 = q_0 < |q_1| < |q_2| < \ldots$;
(ii) $p_{k-1}q_k - p_k q_{k-1} = 1$, *for all* $k \geq 0$;
(iii) *Let* $T(z) = z+1$, $S(z) = -1/z$ *be the generating transformations for* $PSL(2,\mathbb{Z})$, *then for any* $z \in \mathcal{H} \cup \mathbb{R} \cup \{\infty\}$

$$T^{n_0} S T^{n_1} S \ldots T^{n_k} S(z) = \lceil n_0, n_1, \ldots, n_k, z\rceil = \frac{p_k z - p_{k-1}}{q_k z - q_{k-1}}\,;$$

(iv) *The sequence* $\{r_k\}$ *converges to* α *and* $|p_k/q_k - \alpha| \leq 1/q_k^2$;
(v) $\frac{q_k}{q_{k-1}} \in [n_k - r, n_k + 1 - r]$ *if* $n_k > 0$, *and* $\frac{q_k}{q_{k-1}} \in [n_k - 1 + r, n_k + r]$ *if* $n_k < 0$.

To an H-reduced geodesic γ, one associates a bi-infinite sequence of integers $\langle\gamma\rangle = \langle\ldots n_{-1}, n_0, n_1, \ldots\rangle$, by juxtaposing the H-dual expansion of $1/u$ and the H-expansion of w. Observe that $|n_i| \geq 2$ and the only additional restriction on n_i's is that if $n_i = \pm 2$, then $n_i n_{i+1} < 0$.

Reduction algorithm. We describe the reduction procedure of any geodesic to an H-reduced one.

Theorem 3.3. *Every oriented geodesic on $\mathcal{H}$ is $PSL(2,\mathbb{Z})$-equivalent to an H-reduced geodesic.*

Proof. For the sake of completeness, we present the proof following Hurwitz [H]. Let γ be an arbitrary geodesic on $\mathcal{H}$, with irrational end points u and w, and assume that $u < w$. Let $\langle n_0, n_1, n_2, \dots \rangle$ be the H-expansion of w, and suppose that its tail is different from $\langle \bar{3} \rangle$ (the situation when the tail of w coincides with $\langle \bar{3} \rangle$ can be treated similarly). We construct the following sequence of real pairs $\{(u_k, w_k)\}$ $(k \geq 0)$ defined by $u_0 = u$, $w_0 = w$ and:

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

Notice that $w_{k+1} = \langle n_{k+1}, n_{k+2}, \dots \rangle$ and by Proposition 3.2 (iii),

$$w = \frac{p_k w_{k+1} - p_{k-1}}{q_k w_{k+1} - q_{k-1}}\,,\ u = \frac{p_k u_{k+1} - p_{k-1}}{q_k u_{k+1} - q_{k-1}}\,.$$

Hence

$$u_{k+1} = \frac{q_{k-1}u - p_{k-1}}{q_k u - p_k} = \frac{q_{k-1}}{q_k} + \frac{1}{q_k^2(p_k/q_k - u)} = \frac{q_{k-1}}{q_k} + \varepsilon_k\,,$$

where $\varepsilon_k > 0$ (for large enough k) and $\varepsilon_k \to 0$. For infinitely many k's, $n_k \neq 2, 3$, and one can find a subsequence k_j such that

$$w_{k_j+1} > 2,\ n_{k_j+1} \geq 2, \ \text{ and } n_{k_j} = -2, -3, \pm 4, \dots$$

or

$$w_{k_j+1} < -2,\ n_{k_j+1} \leq -2, \ \text{ and } n_{k_j} = -3, \pm 4, \dots\,.$$

Using Proposition 3.2 (v), one has, in the first case

$$\frac{q_{k_j}}{q_{k_j-1}} \leq -2 + r \text{ or } \frac{q_{k_j}}{q_{k_j-1}} \geq 4 - r \ \Rightarrow\ u_{k_j+1} \in \left[\frac{1}{-2+r} + \varepsilon_{k_j}, \frac{1}{4-r} + \varepsilon_{k_j}\right]$$

and, in the second case,

$$\frac{q_{k_j}}{q_{k_j-1}} \leq -3 + r \text{ or } \frac{q_{k_j}}{q_{k_j-1}} \geq 4 - r \ \Rightarrow\ u_{k_j+1} \in \left[\frac{1}{-3+r} + \varepsilon_{k_j}, \frac{1}{4-r} + \varepsilon_{k_j}\right].$$

Since $1/(-2+r) = r - 1$, $1/(-3+r) = -r$, $1/(4-r) < r$ and $0 < \varepsilon_{k_j} \to 0$, there exists an integer l such that $w_{k_l+1} > 2$ and $u_{k_l+1} \in [r-1, r]$ or $w_{k_l+1} < -2$ and $u_{k_l+1} \in [-r, 1-r]$. The geodesic with end points u_{k_l+1} and w_{k_l+1} is H-reduced and $PSL(2,\mathbb{Z})$-equivalent to γ.

We finish the proof by explaining how one can derive the reduction procedure for the case $u > w$. Let $w =< n_0, n_1, \cdots >$ and consider, as before, the sequence of real pairs $\{(u_k, w_k)\}$ $(k \geq 0)$ defined by $u_0 = u$, $w_0 = w$ and:

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

Since $-u < -w$, one can apply the reduction procedure described above to the geodesic $\tilde{\gamma}$ from $\tilde{u} = -u$ to $\tilde{w} = -w$. Notice that $-w =< -n_0, -n_1, \cdots >$ and the sequence of pairs $\{(\tilde{u}_k, \tilde{w}_k)\}$ $(k \geq 0)$ is defined by $\tilde{u}_0 = \tilde{u}$, $\tilde{w}_0 = \tilde{w}$ and:

$$\tilde{w}_{k+1} = ST^{n_k} \dots ST^{n_1} ST^{n_0} \tilde{w}\,, \quad \tilde{u}_{k+1} = ST^{n_k} \dots ST^{n_1} ST^{n_0} u\,.$$

Using the identity $ST^n(-w) = -ST^{-n}w$, one has $\tilde{w}_k = -w_k$ and $\tilde{u}_k = -\tilde{u}_k$. ¿From the proof above, there exists a positive integer k such that the geodesic with end

points $\tilde{u}_k$ and $\tilde{w}_k$ is H-reduced and $PSL(2,\mathbb{Z})$-equivalent to $\tilde{\gamma}$. Thus, the geodesic from $u_k = -\tilde{u}_k$ to $w_k = -\tilde{w}_k$ is also H-reduced and $PSL(2,\mathbb{Z})$-equivalent to γ. □

Remark 3.4. (i) The proof of Theorem 3.3 gives also the algorithm for H-reducing a geodesic γ: one has to construct inductively the sequence $\{(u_k, w_k)\}$ until $|w_k| > 2$ and $\mathrm{sgn}(w_k)u_k \in [r-1, r]$; (ii) any further application of the reduction algorithm to an H-reduced geodesic yields reduced geodesics whose H-codes are left shifts of the H-code of the first reduced one.

As in the previous sections we define the H-code of an oriented geodesic γ on $\mathcal{H}$ to be the H-code of a reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, and prove its $PSL(2,\mathbb{Z})$-invariance by constructing a cross-section of the geodesic flow on M, directly related to the notion of H-reduced geodesics.

Construction of the cross-section. We describe the construction of the cross section C_H for the geodesic flow on M, such that successive returns to the cross section correspond to left-shifts in the H-code. We define $C_H = P \cup Q_1 \cup Q_2$ to be a subset of the unit tangent bundle SM, where P consists of all tangent vectors with base points in the circular side of F and pointing inward such that the corresponding geodesic is H-reduced; Q_1 consists of all tangent vectors with base points on the right vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $TS(\gamma)$ is H-reduced; Q_2 consists of all tangent vectors with base points on the left vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $T^{-1}S(\gamma)$ is H-reduced. Notice that $C_H = \pi(C_h)$ where C_h is the set 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 H-reduced (Figure 5).

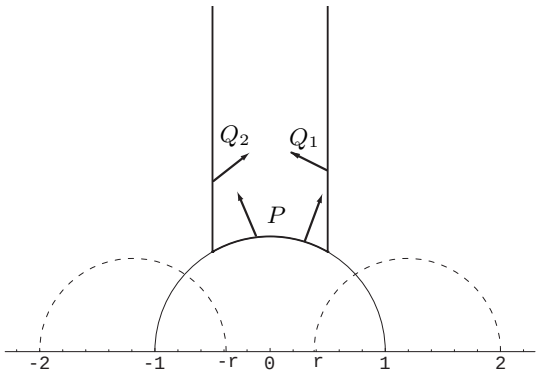

FIGURE 5. The cross section $C_H = P \cup Q_1 \cup Q_2$

One can show similarly to the proof of Theorem 1.5 that $C_H = P \cup Q_1 \cup Q_2$ is indeed a cross-section for the geodesic flow on M, hence every geodesic γ can be represented as a bi-infinite sequence of segments σ_i between successive returns to C_H. To each segment σ_i is associated the corresponding H-reduced geodesic γ_i, so that $\lceil \gamma_{i+1} \rfloor$ differs from $\lceil \gamma_i \rfloor$ by a left shift. Thus we associate to γ a bi-infinite coding sequence, defined up to a shift, which we call the *H-code* of γ and denote by $\langle \gamma \rangle$. The argument of Corollary 1.6 shows that the H-code is $PSL(2,\mathbb{Z})$-invariant.

The set of all oriented geodesics on M can be described symbolically as a countable 1-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,

$$H(n,m) = \begin{cases} 0 & \text{if } |n| = 2 \text{ and } nm > 0, \\ 1 & \text{otherwise}. \end{cases} \tag{3.1}$$

Each oriented geodesic γ corresponds to its H-code, $\langle\gamma\rangle \in X_H$ and each bi-infinite sequence of nonzero integers $x \in X_H$ produces a geodesic on $\mathcal{H}$ from $u(x)$ to $w(x)$, where

$$w(x) = \langle n_1, n_2, \ldots\rangle \quad , \quad \frac{1}{u(x)} = \langle\langle n_0, n_{-1}, \ldots\rangle\rangle .$$

The partition of the cross-section. The infinite partition of the cross-section C_H corresponding to the H-code can be constructed as follows. We parameterize the cross-section C_h by (ϕ, θ), where $\phi \in [0, \pi]$ parameterizes the unit semicircle (counterclockwise) and $\theta \in [-\pi/2, (3\pi)/2]$ is the angle the unit vector makes with the positive horizontal axis (counterclockwise). The angle θ depends on the position ϕ and is determined by the condition that the corresponding geodesic is H-reduced.

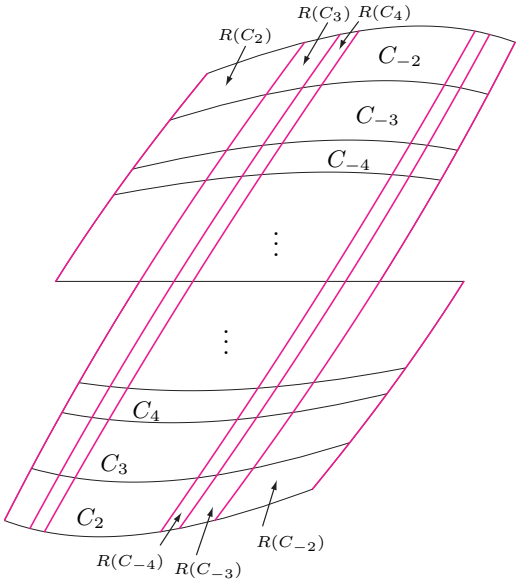

FIGURE 6. Infinite partition for the H-code and its image under the return map R

The partition of C_h (and therefore of C_H by projection) corresponding to the arithmetic H-code ("the horizontal rectangles") and its iteration under the first return map R to the cross-section C_h ("the vertical rectangles") is shown on Figure 6. Its elements ("the horizontal rectangles") are labeled by the symbols of the alphabet $\mathcal{N}_H$, $C_h = \sqcup_{n\in\mathcal{N}_H} C_n$ and are defined by the following condition: $C_n = \{\mathbf{x} \in C_h, n_0(\mathbf{x}) = n\}$, i.e. it consists of all tangent vectors $\mathbf{x}$ in C_h such that the coding sequence $x \in X$ of the corresponding geodesic with this initial vector has its first symbol in the H-code $n_0(x) = n$. Thus, for $n \geq 2$, C_n consists of all tangent

vectors $\mathbf{x} \in C_h$ such that the corresponding geodesic goes from $r-1 < u < r$ to $n-1/2 < w < n+1/2$. For $n \leq -2$, C_n consists of all tangent vectors $\mathbf{x} \in C_h$ such that the corresponding geodesic goes from $-r < u < 1-r$ to $n-1/2 < w < n+1/2$.

Some results of this section can be illustrated geometrically. If $n_0(\mathbf{x}) = n$ and $n_1(\mathbf{x}) = m$ for some vector $\mathbf{x} \in C_H$, then $R(C_n) \cap C_m \neq \emptyset$. Therefore, as follows from Figure 6, if $R(C_n) \cap C_m \neq \emptyset$ and $n = \pm 2$, then $nm < 0$. Moreover, all intersections are transversal, hence, according to Theorem 7.9 of [Ad], our partition is Markov.

When does the H-code coincide with the geometric code? The next theorem gives a sufficient condition for the geometric code and the arithmetic H-code of a geodesic γ on M to coincide.

Theorem 3.5. *The geometric code and the arithmetic H-code of γ coincide if $|n_i| \geq 2$ and*

$$\left| \frac{1}{n_i} + \frac{1}{n_{i+1}} \right| \leq \frac{1}{2}, \tag{3.2}$$

where $\langle \gamma \rangle = \langle n_i \rangle_{i=-\infty}^{\infty}$.

Proof. Let $x = \{\dots, n_{-2}, n_{-1}, n_0, n_1, \dots\}$ be a sequence of integers with $|n_i| \geq 2$ and satisfying (3.2). Consider the geodesic $\gamma(x)$ on $\mathcal{H}$, from

$$u(x) = \frac{1}{(n_{-1}, n_{-2}, \dots)} = \cfrac{1}{n_{-1} - \cfrac{1}{n_{-2} - \cfrac{1}{\ddots}}} \quad \text{to} \quad w(x) = (n_0, n_1, \dots) = n_0 - \cfrac{1}{n_1 - \cfrac{1}{\ddots}}.$$

Since $x \in X_H$, the H-code of $\gamma(x)$ is $\langle \gamma(x) \rangle = \langle n_{-2}, n_{-1}, n_0, n_1, \dots \rangle$. We showed in [KU, Theorem 1.4], that such a geodesic $\gamma(x)$ from $u(x)$ to $w(x)$ has the geometric code $[\gamma(x)] = [\dots, n_{-1}, n_0, n_1, n_2, \dots]$. Therefore the geometric code and the arithmetic code of $\pi(\gamma(x))$ coincide (up to a shift). □

Remark 3.6. There exist geodesics that are not geometrically Markov, for which the geometric code and the H-code coincide. For example, consider the closed geodesic γ given by the axis of $T^5ST^3ST^{-2}S$. Its geometric code is $[\gamma] = [5, 3, -2]$ and coincides with the H-code $\langle \gamma \rangle = \langle \overline{5, 3, -2} \rangle$. However, γ is not geometrically Markov. A natural question would be to characterize completely the class of geodesics for which the two codes coincide.

Remark 3.7. One can easily notice that the H-code and G-code of a geodesic γ coincide if $n_i \geq 3$, and, the H-code, G-code and geometric code coincide for positive geodesics.

4. Symbolic representation of the geodesic flow

Geodesic flow as a special flow. Let $C_\alpha \subset SM$ ($\alpha = G, A, H$) be a cross-section for the geodesic flow $\{\varphi^t\}$ on M constructed for one of the arithmetic codes studied in the previous sections, and X_α the corresponding set of coding sequences. Every $\mathbf{x} \in C_\alpha$ defines an oriented geodesic $\gamma(\mathbf{x})$ on M which will return to C_α infinitely often. Let $R_\alpha : C_\alpha \to C_\alpha$ be the first return map, and $f_\alpha : C_\alpha \to \mathbb{R}$ be the time of

the first return on C_α defined as follows: for $\mathbf{x} \in C_\alpha$, $R_\alpha(\mathbf{x}) = \varphi^t(\mathbf{x})$, $f_\alpha(\mathbf{x}) = t$. Then $\{\varphi^t\}$ can be represented as the special flow on the space

$$C_\alpha{}^{f_\alpha} = \{(\mathbf{x}, y) : \mathbf{x} \in C_\alpha, 0 \le y \le f_\alpha(\mathbf{x})\}$$

with the ceiling function f_α by the formula $\varphi^t(\mathbf{x}, y) = (\mathbf{x}, y + t)$ with the identification $(\mathbf{x}, f_\alpha(\mathbf{x})) = (R_\alpha(\mathbf{x}), 0)$.

In the previous sections for each arithmetic code we have established a bijective map $\mathrm{Cod}_\alpha : C_\alpha \to X_\alpha$ by $\mathrm{Cod}_\alpha : \mathbf{x} \mapsto (\gamma(\mathbf{x}))$ such that the diagram

$$\begin{array}{ccc} C_\alpha & \xrightarrow{\mathrm{Cod}_\alpha} & X_\alpha \\ {\scriptstyle \mathrm{R}_\alpha}\downarrow & & \downarrow{\scriptstyle \sigma_\alpha} \\ C_\alpha & \xrightarrow{\mathrm{Cod}_\alpha} & X_\alpha \end{array}$$

is commutative. Here σ_α is the left shift $\sigma_\alpha : X_\alpha \to X_\alpha$ defined for $x = (n_i(x))_{i=-\infty}^{\infty}$ by $(\sigma_\alpha x)_i = n_{i+1}(x)$. Thus we obtain three symbolic representations of the geodesic flow (for $\alpha = G, A, H$) on the space

$$X_\alpha{}^{f_\alpha} = \{(x, y) : x \in X_\alpha, 0 \le y \le f_\alpha(x)\}$$

given by the formula $\varphi^t(x, y) = (x, y + t)$ with the identification $(x, f_\alpha(x)) = (\sigma_\alpha x, 0)$, where $(X_\alpha, \sigma_\alpha)$ is the space of α-coding sequences, and f_α is the time of the first return to the cross-section C_α.

Calculation of the return time. Let $\alpha = G, A, H$. The ceiling function $f_\alpha(x)$ on X_α is the length of the segment between successive returns of the geodesic $\gamma(x)$ to the cross-section C_α. The following theorem was proved in [GK] for the G-code. The proof for the A- and H-code is the same.

Theorem 4.1. *Let $x \in X_\alpha$ and $w(x)$, $u(x)$ be the end points of the corresponding geodesic $\gamma(x)$. Then*

$$f_\alpha(x) = 2 \log |w(x)| + \log g(x) - \log g(\sigma_\alpha x)$$

where

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

References

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

[AF1] R. Adler and 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 and L. Flatto, *Cross section map for geodesic flow on the modular surface*, Contemp. Math. **26** (1984), 9–23.

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

[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.

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

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

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

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

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

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

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

[K1] S. Katok, *Fuchsian groups*, University of Chicago Press, 1992.

[K2] S. Katok, *Coding of closed geodesics after Gauss and Morse*, Geom. Dedicata **63** (1996), 123–145.

[K3] S. Katok, *Continued fractions, hyperbolic geometry and quadratic forms*, MASS Selecta, Amer. Math. Soc., 121–160, 2003.

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

[O] C. D. Olds, *Continued fractions*, New Mathematics Library **9**, MAA, 1992.

[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.

[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, The Pennsylvania State University, University Park, PA 16802

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