ELECTRONIC RESEARCH ANNOUNCEMENTS
IN MATHEMATICAL SCIENCES
Volume 17, Pages 20–33 (April 21, 2010)
S 1935-9179 AIMS (2010)

doi:10.3934/era.2010.17.20

THEORY OF (a, b)-CONTINUED FRACTION TRANSFORMATIONS AND APPLICATIONS

SVETLANA KATOK AND ILIE UGARCOVICI

(Communicated by Boris Hasselblatt)

Abstract. We study a two-parameter family of one-dimensional maps and the related (a, b)-continued fractions suggested for consideration by Don Zagier and announce the following results and outline their proofs: (i) the associated natural extension maps have attractors with finite rectangular structure for the entire parameter set except for a Cantor-like set of one-dimensional zero measure that we completely describe; (ii) for a dense open set of parameters the Reduction theory conjecture holds, i.e. every point is mapped to the attractor after finitely many iterations. We also give an application of this theory to coding geodesics on the modular surface and outline the computation of the smooth invariant measures associated with these transformations.

1. Introduction

The standard generators $T(x) = x + 1$, $S(x) = -1/x$ of the modular group $SL(2, \mathbb{Z})$ were used classically to define piecewise continuous maps acting on the extended real line $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ that led to well-known continued fraction algorithms. Don Zagier suggested considering a two-parameter family of such maps $f_{a,b} : \bar{\mathbb{R}} \to \bar{\mathbb{R}}$ defined by

$$f_{a,b}(x) = \begin{cases} x + 1 & \text{if } x < a \\ -\dfrac{1}{x} & \text{if } a \le x < b \\ x - 1 & \text{if } x \ge b\,. \end{cases} \tag{1.1}$$

In order for these maps to induce continued fraction algorithms the orbit of any (irrational) point should return to the interval $[a, b)$ infinitely often, and consist of blocks of T's and T^{-1}'s separated by S's, i.e. the parameters (a, b) must belong to the set

$$\mathcal{P} = \{(a, b) \,|\, a \le 0 \le b,\, b - a \ge 1,\, -ab \le 1\}\,.$$

In all our considerations we will assume that $(a, b) \in \mathcal{P}$.

Using the first return map of $f_{a,b}$ to the interval $[a, b)$, denoted by $\hat{f}_{a,b}$, we introduce a two-parameter family of continued fraction algorithms. Let us mention

Received by the editors November 21, 2009 and, in revised form, February 23, 2010.
2000 *Mathematics Subject Classification.* Primary 37D40, 37B40; Secondary 11A55, 20H05.
Key words and phrases. Continued fractions, attractors, modular surface, invariant measure.
We are grateful to Don Zagier for helpful discussions and the Max Plank Institute for Mathematics in Bonn for its hospitality and support. The second author is partially supported by the NSF grant DMS-0703421.

here three classical examples: the case $a=-1$, $b=0$ described in [19, 6] gives the "minus" (backward) continued fractions, the case $a=-1/2$, $b=1/2$ gives the "nearest-integer" continued fractions considered first by Hurwitz in [4], and the case $a=-1$, $b=1$ was presented in [17, 7] in connection with a method of symbolically coding the geodesic flow on the modular surface following Artin's pioneering work [3] and corresponds to the regular "plus" continued fractions with alternating signs of the digits. Also, in the case $b-a=1$, the class of one-parameter maps $f_{b-1,b}$ with $b\in[0,1]$ is conceptually similar to the "α-transformations" introduced by Nakada in [14] and studied subsequently in [12, 13, 15, 16, 18].

The main object of our study is a two-dimensional realization of the *natural extension map* of $f_{a,b}$, $F_{a,b}:\bar{\mathbb{R}}^2\setminus\Delta\to\bar{\mathbb{R}}^2\setminus\Delta$, $\Delta=\{(x,y)\in\bar{\mathbb{R}}^2|\,x=y\}$, defined by

$$F_{a,b}(x,y)=\begin{cases}(x+1,y+1) & \text{if } y<a\\ \left(-\dfrac{1}{x},-\dfrac{1}{y}\right) & \text{if } a\le y<b\\ (x-1,y-1) & \text{if } y\ge b\,.\end{cases}\tag{1.2}$$

Numerical experiments led Don Zagier to conjecture that such a map $F_{a,b}$ has several interesting properties for all parameter pairs $(a,b)\in\mathcal{P}$ that we list under the **Reduction theory conjecture.**

(1) The map $F_{a,b}$ possesses a global attractor set $D_{a,b}=\cap_{n=0}^{\infty}F^n(\bar{\mathbb{R}}^2\setminus\Delta)$ on which $F_{a,b}$ is essentially bijective.
(2) The set $D_{a,b}$ consists of two (or one, in degenerate cases) connected components each having *finite rectangular structure*, i.e. bounded by non-decreasing step-functions with a finite number of steps.
(3) Every point (x,y) of the plane $(x\neq y)$ is mapped to $D_{a,b}$ after finitely many iterations of $F_{a,b}$.

Figure 1 shows the computer picture of such a the set $D_{a,b}$ with $a=-4/5$, $b=2/5$.

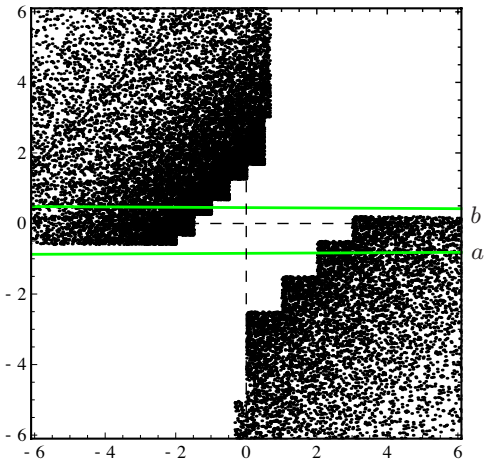

FIGURE 1. A typical attractor $D_{a,b}$ $(a=-\frac{4}{5}, b=\frac{2}{5})$

Besides the classical cases mentioned above, this conjecture has been proved in [9] for an open dense subset of parameter pairs $(a,b)\in\mathcal{P}$. Here is the main result:

Theorem 1.1. *There exists an explicit one-dimensional Lebesgue measure zero, uncountable set $\mathcal{E}$ that lies on the diagonal boundary $b = a + 1$ of $\mathcal{P}$ such that:*

(a) *for all $(a, b) \in \mathcal{P} \setminus \mathcal{E}$ the map $F_{a,b}$ has an attractor $D_{a,b}$ satisfying properties (1) and (2) above;*
(b) *for an open and dense set in $\mathcal{P} \setminus \mathcal{E}$ property (3), and hence the Reduction theory conjecture, holds. For the rest of $\mathcal{P} \setminus \mathcal{E}$ property (3) holds for almost every point of the plane.*

We point out that this approach gives explicit conditions for the set $D_{a,b}$ to have finite rectangular structure that are satisfied, in particular, for all pairs (a, b) in the interior of the maximal parameter set $\mathcal{P}$. At the same time, it provides an effective algorithm for finding $D_{a,b}$, independent of the complexity of its boundary (i.e., number of horizontal segments). The simultaneous properties satisfied by $D_{a,b}$, attracting set and bijectivity domain for $F_{a,b}$, is an essential feature that has not been exploited in earlier works. This approach makes the notions of reduced geodesic and dual expansion natural and transparent, with a potential for generalization to other Fuchsian groups. We remark that for "α-transformations" [14, 12], explicit descriptions of the domain of the natural extension maps have been obtained only for a subset of the parameter interval $[0, 1]$ (where the boundary has low complexity).

If one identifies a geodesic on the hyperbolic upper half-plane with a pair of real numbers $(x, y) \in \mathbb{R}^2$, $x \neq y$, its endpoints, then $F_{a,b}$ maps a geodesic from x to y to a geodesic $PSL(2, \mathbb{Z})$-equivalent to it, and hence can be perceived as a *reduction map*.

In this paper we announce and sketch the proof of the above theorem [9], and describe its applications to coding of geodesics [10] and the computation of invariant measures associated with these transformations [9, 11].

2. (a, b)-continued fractions

The map $f_{a,b}$ defines what we call (a, b)-*continued fractions* using a generalized integral part function:

$$\lfloor x \rceil_{a,b} = \begin{cases} \lfloor x - a \rfloor & \text{if } x < a \\ 0 & \text{if } a \leq x < b \\ \lceil x - b \rceil & \text{if } x \geq b\,, \end{cases} \tag{2.1}$$

where $\lfloor x \rfloor$ denotes the integer part of x and $\lceil x \rceil = \lfloor x \rfloor + 1$.

Theorem 2.1. *If $(a, b) \in \mathcal{P}$, then any irrational number x can be expressed uniquely as an infinite continued fraction of the form*

$$x = n_0 - \cfrac{1}{n_1 - \cfrac{1}{n_2 - \cfrac{1}{\ddots}}} = \lfloor n_0, n_1, n_2, \dots \rceil_{a,b}, \ (n_k \neq 0 \text{ for } k \geq 1),$$

where $n_0 = \lfloor x \rceil_{a,b}$, $x_1 = -\frac{1}{x - n_0}$, and $n_k = \lfloor x_k \rceil_{a,b}$, $x_{k+1} = -\frac{1}{x_k - n_k}$, i.e. the sequence of partial fractions $r_k = \lfloor n_0, n_1, \dots, n_k \rceil_{a,b}$ converges to x.

The proof follows the lines of the proof presented in [6] for the case of minus continued fractions (where $a = -1$, $b = 0$, and $n_k \geq 2$ if $k \geq 1$). We define inductively two sequences of integers $\{p_k\}$ and $\{q_k\}$ for $k \geq -2$:

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

and prove that $\frac{p_k}{q_k} = r_k$ converges to x. Here we use the important fact that $n_k \cdot n_{k+1} < 0$ if the entry $n_{k+1} = \pm 1$.

Remark 2.2. One can construct (a,b)-continued fraction expansions for rational numbers, too. However, such expansions will terminate after finitely many steps if $b \neq 0$. If $b = 0$, the expansions of rational numbers end with a tail of 2's.

In what follows, we use (in some situations) the simplified notations $\lfloor \cdot \rceil$, f, $\hat{f}$ and F for $\lfloor \cdot \rceil_{a,b}$, $f_{a,b}$, $\hat{f}_{a,b}$ and $F_{a,b}$, respectively, assuming implicitly their dependence on parameters a, b. We use also the notation f^n (or $\hat{f}^n$) for the n-fold composition operation of f (or $\hat{f}$). Also, for a given point $x \in [a,b)$ the notation $\hat{f}^{(k)}$ means the transformation of type $T^i S$ (i is an integer) such that

$$\hat{f}^k(x) = \hat{f}^{(k)}\hat{f}^{(k-1)} \cdots \hat{f}^{(2)}\hat{f}^{(1)}(x),$$

where $\hat{f}^{(1)}(x) = \hat{f}(x)$.

3. Cycle property

The structure of the attractor $D_{a,b}$ is actually "computed" from the data (a,b) as follows. We associate to the points of discontinuity of the map f, a and b, two forward orbits: to a, the *upper orbit* $\mathcal{O}_u(a)$ (i.e. the orbit of Sa) and the *lower orbit* $\mathcal{O}_\ell(a)$ (i.e. the orbit of Ta), and to b, the *upper orbit* $\mathcal{O}_u(b)$ (i.e. the orbit of $T^{-1}b$) and the *lower orbit* $\mathcal{O}_\ell(b)$ (i.e. the orbit of Sb). Now we explore the patterns in the above orbits.

The following property plays an essential role in studying the map f.

Definition 3.1. We say that a (resp., b) has the *cycle property* if the upper and lower orbits meet forming a cycle, i.e. if for some $k_1, m_1, k_2, m_2 \geq 0$ s.t.

$$f^{m_1}(Sa) = f^{k_1}(Ta) = c_a, \text{ (resp., } f^{m_2}(T^{-1}b) = f^{k_2}(Sb) = c_b).$$

We refer to the sets

$$\{Ta, fTa, f^2Ta, \ldots f^{k_1-1}Ta\} \text{ and } \{Sb, fSb, f^2Sb, \ldots, f^{k_2-1}Sb\}$$

as *lower*, and the sets

$$\{Sa, fSa, f^2Sa, \ldots, f^{m_1-1}Sa\} \text{ and } \{T^{-1}b, fT^{-1}b, \ldots, f^{m_2-1}T^{-1}b\}$$

as *upper* sides of the corresponding cycles, and the numbers c_a and c_b as the *ends of the cycles*.

If the product over the a-cycle (resp., b-cycle) equals the identity transformation

$$T^{-1}f^{-k_1}f^{m_1}S = \mathrm{Id}, \text{ (resp., } Tf^{-m_2}f^{k_2}S = \mathrm{Id}),$$

we say that a (resp., b) has the *strong cycle property*; otherwise, we say that a (resp., b) has the *weak cycle property*.

The structure of the set of points in $\mathcal{P}$ for which parameter b has the cycle property follows from the following theorem and the symmetry of the parameter set $\mathcal{P}$ with respect to the line $b = -a$, $(a, b) \mapsto (-b, -a)$. The case $a \leq -1$ is simple and can be analyzed separately. A similar result holds for parameter a.

Theorem 3.2. *Let $(a, b) \in \mathcal{P}$, $0 < b \leq -a < 1$ and $m \geq 1$ such that $a \leq T^m Sb < a + 1$.*

(1) *Suppose that there exists $n \geq 0$ such that*

$$\hat{f}^k T^m Sb \in \left(\frac{b}{b+1}, a+1\right) \text{ for } k < n, \text{ and } \hat{f}^n T^m Sb \in \left[a, \frac{b}{b+1}\right].$$

   (i) *If $\hat{f}^n T^m Sb \in (a, \frac{b}{b+1})$, then b has the cycle property; the cycle property is strong if and only if $\hat{f}^n T^m Sb \neq 0$.*
   (ii) *If $\hat{f}^n T^m Sb = a$, then b has the cycle property if and only if a has the cycle property.*
   (iii) *$\hat{f}^n T^m Sb = b/(b+1)$, then b does not have the cycle property, but the (a, b)-expansions of Sb and $T^{-1}b$ are eventually periodic.*

(2) *If $\hat{f}^k T^m Sb \in (\frac{b}{b+1}, a+1)$ for all $k \geq 0$, then b does not have the cycle property.*

We remark that the cases $m = 1, 2$ can be explicitly analyzed and the cycle relations are simple (and short); the situation $m \geq 3$ is more intricate and the following property is essential for the proof of Theorem 3.2: if

$$\frac{b}{b+1} < \hat{f}^k T^m Sb < a + 1 \text{ for all } k < n\,,$$

then

- the lower orbit of b satisfies $\hat{f}^{(k)} = T^m S$ or $T^{m+1} S$, and the upper orbit of b satisfies $\hat{f}^{(k)} = T^{-i} S$ with $i = 2$ or 3;
- there exists $q > 1$ such that $(STS)\hat{f}^n T^m S = (T^{-2} S)\hat{f}^q T^{-1}$.

The proof is by induction on n. In order to determine the upper side of the b-cycle, we use the following relation in the group $SL(2, \mathbb{Z})$

$$(STS)T^i S = (T^{-2} S)^{i-1} T^{-1} \quad (i \geq 1),$$

obtained from the "standard" relation $STS = T^{-1} S T^{-1}$.

4. Structure of the attractor

In order to state the condition under which the natural extension map $F_{a,b}$ has an attractor with finite rectangular structure mentioned in the Introduction, we follow the split orbits of a and b, and define the *truncated orbits* $\mathcal{L}_a$ and $\mathcal{U}_a$ by

$$\mathcal{L}_a = \begin{cases} \mathcal{O}_\ell(a) & \text{if } a \text{ has no cycle property} \\ \text{lower part of } a\text{-cycle} & \text{if } a \text{ has strong cycle property} \\ \text{lower part of } a\text{-cycle} \cup \{0\} & \text{if } a \text{ has weak cycle property,} \end{cases}$$

$$\mathcal{U}_a = \begin{cases} \mathcal{O}_u(a) & \text{if } a \text{ has no cycle property} \\ \text{upper part of } a\text{-cycle} & \text{if } a \text{ has strong cycle property} \\ \text{lower part of } a\text{-cycle} \cup \{0\} & \text{if } a \text{ has weak cycle property,} \end{cases}$$

and, similarly, $\mathcal{L}_b$ and $\mathcal{U}_b$.

Definition 4.1. We say that (a,b) satisfies the *finiteness condition* if the sets of values in the truncated orbits $\mathcal{L}_a$, $\mathcal{U}_a$, $\mathcal{L}_b$, and $\mathcal{U}_b$ are finite.

The following proposition follows from Theorems 3.2.

Proposition 4.2. *Suppose that the set $\mathcal{L}_b$ is finite. Then*

(1) *either b has the cycle property or the upper and lower orbits of b are eventually periodic.*
(2) *The finiteness of $\mathcal{L}_b$ implies the finiteness of $\mathcal{U}_b$.*

Similar statements hold for the sets $\mathcal{L}_a$, $\mathcal{U}_a$ and $\mathcal{U}_b$ as well.

Definition 4.3. We say that a proper subset of $\mathbb{R}^2$ has *finite rectangular structure* if it consists of two (or one, in degenerate cases) connected components bounded by non-decreasing step-functions with finitely many steps.

The following theorem is proved in [9]:

Theorem 4.4. *If $(a,b) \in \mathcal{P}$ satisfies the finiteness condition, then the attractor set $D_{a,b} \subset \bar{\mathbb{R}}^2 \setminus \Delta$ has finite rectangular structure, and $F_{a,b} : D_{a,b} \to D_{a,b}$ is a bijection except for some images of the boundary of $D_{a,b}$.*

The proof consists of the following steps:

Step 1: Construction of a set $A_{a,b}$ with finite rectangular structure where the map $F_{a,b}$ is a bijection except for some images of its boundary. We prove that there exists a unique set $A_{a,b}$ whose upper connected component is bounded by a step-function with values in the set $\mathcal{U}_{a,b} = \mathcal{U}_a \cup \mathcal{U}_b$ that we refer to as *upper levels*, and whose lower connected component is bounded by a step-function with values in the set $\mathcal{L}_{a,b} = \mathcal{L}_a \cup \mathcal{L}_b$ that we refer to as *lower levels*. Notice that each level in $\mathcal{U}_a$ and $\mathcal{U}_b$ appears exactly once, but if the same level appears in both sets, we have to count it twice in $\mathcal{U}_{a,b}$. The same remark applies to the lower levels.

Our goal is to prove that all levels of $\mathcal{L}_{a,b}$ are *connected by a vertical segment* (we will refer to this as *connected*), i.e. that the right end of a segment at a certain level is equal to the left end of the segment on the next level.

First we notice that $STa \in \mathcal{L}_a$ and $Sb \in \mathcal{L}_b$ are two consecutive levels of $\mathcal{L}_{a,b}$, and the levels $Sa \in \mathcal{U}_a$ and $ST^{-1}b \in \mathcal{U}_b$ are two consecutive levels of $\mathcal{U}_{a,b}$. Since the x-coordinate of the right end of the segment at the level STa and the x-coordinate of the left end of the segment at the level Sb are equal to 0, the levels STa and Sb are connected. Similarly, the levels Sa and $ST^{-1}b$ are connected. Let $y_\ell \in \mathcal{L}_{a,b}$ be the closest y-level to Sb with $y_\ell \geq Sb$, and $y_u \in \mathcal{U}_{a,b}$ be the closest y-level to a with $y_u \leq Sa$. Since each level in $\mathcal{U}_a$ and in $\mathcal{L}_b$ appears only once, if $y_u = Sa$, y_u can only belong to $\mathcal{U}_b$, and if $y_\ell = Sb$, y_ℓ can only belong to $\mathcal{L}_a$. We look at the rays $[-\infty, x_b]$ and $[x_a, \infty]$, where x_a and x_b are unknowns, and "transport" them (using the special form of the natural extension map $F_{a,b}$) along the truncated orbits $\mathcal{L}_a$, $\mathcal{L}_b$, $\mathcal{U}_a$ and $\mathcal{U}_b$ until we reach the levels y_u and y_ℓ. Then we set-up a system of two fractional linear equations by equating the right end of the segment at the level Sb with the left end of the segment at the level y_ℓ, and, similarly, the left end of the segment at the level Sa and the right end of the level y_u, and prove that this system has a unique solution (x_a, x_b). Therefore the levels Sb and y_ℓ, and the levels y_u and Sa are connected, so three consecutive levels $STa \leq Sb \leq y_\ell$, and $y_u \leq Sa \leq ST^{-1}b$ are connected. Moreover, their images under the same transformations in $SL(2,\mathbb{Z})$ remain connected. The main technical difficulty of the

proof is to follow the connected triples over the truncated orbits and to show that they create longer and longer chains of connected segments until all upper and all lower segments are connected.

The following proposition is instrumental to the proof. In the statement we write $f_{a,b}(x) = \rho_{a,b}(x)x$ using the following map $\rho_{a,b} : \bar{\mathbb{R}} \to \{T, S, T^{-1}\}$

$$\rho_{a,b}(x) = \begin{cases} T & \text{if } x < a \\ S & \text{if } a \le x < b \\ T^{-1} & \text{if } x \ge b. \end{cases} \tag{4.1}$$

Proposition 4.5. *Suppose that the set $\mathcal{L}_{a,b}$ is finite and $y \in \mathcal{L}_{a,b}$ with $y > STa$.*

1. *If $y \in \mathcal{L}_a$, then there exists $n_0 > 0$ such that $\rho(f^n y) = \rho(f^n STa)$ for all $0 < n < n_0$ and $\rho(f^{n_0} y) \neq \rho(f^{n_0} STa)$, or $f^{n_0} y = 0$;*
2. *If $y \in \mathcal{L}_b$, then $y > Sb$, and there exists $n_0 > 0$ such that $\rho(f^n y) = \rho(f^n Sb)$ for all $n < n_0$ and $\rho(f^{n_0} y) \neq \rho(f^{n_0} Sb)$, or $f^{n_0} y = 0$.*

A similar statement holds for the set $\mathcal{U}_{a,b}$ as well.

Bijectivity is proved by partitioning the upper and lower components of $A_{a,b}$ into 6 pieces and making sure that their images fit together without overlapping (see Figure 2). The cycle or periodic structure are used in the proof.

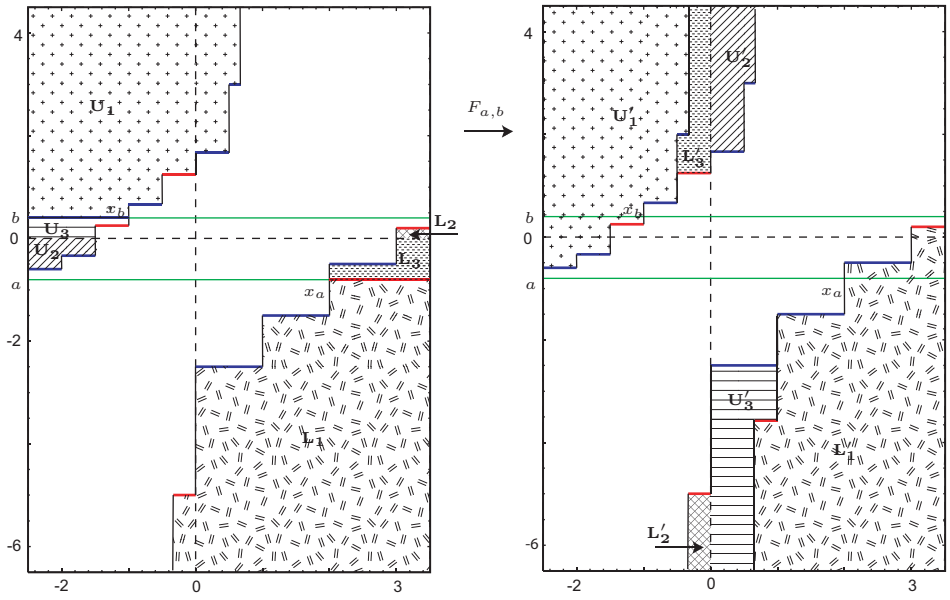

FIGURE 2. Bijectivity of the map $F_{a,b}$

Corollary 4.6. *If both a and b satisfy the strong cycle property, then for any boundary component h of $A_{a,b}$ (vertical or horizontal) there exists $N > 0$ such that $F_{a,b}^N(h)$ is in the interior of $A_{a,b}$.*

This follows from the fact that the "locking segments" at the levels corresponding to the ends of the cycles c_a and c_b are in the interior of the upper or lower connected components of $A_{a,b}$.

Step 2: Proof that the attractor $D_{a,b}$ coincides with $A_{a,b}$. The attractor of the map F is constructed by starting with a *trapping region*, i.e. a set $\Theta_{a,b} \subset \bar{\mathbb{R}}^2 \setminus \Delta$ with the following properties:

(i) for every pair $(x,y) \in \bar{\mathbb{R}}^2 \setminus \Delta$, there exists a positive integer N such that $F_{a,b}^N(x,y) \in \Theta_{a,b}$;
(ii) $F_{a,b}(\Theta_{a,b}) \subset \Theta_{a,b}$.

The precise description of $\Theta_{a,b}$ is given in [9]. To prove the "trapping" property for any initial pair $(x,y) \in \mathbb{R}^2 \setminus \Delta$, one uses the (a,b)-continued fraction expansion of $y = \lfloor n_0, n_1, \dots \rceil_{a,b}$ to show that there exists a positive integer $N > 0$ depending on (x,y) such that $F_{a,b}^N(x,y) = ST^{-n_k} \dots ST^{-n_1} ST^{-n_0}(x,y) \in \Theta_{a,b}$. Using $\Theta_{a,b}$, one defines the attractor $D_{a,b}$ by $D_{a,b} = \bigcap_{n=0}^{\infty} D_n$, where $D_n = \bigcap_{i=0}^{n} F^i(\Theta_{a,b})$.

We prove that under the finiteness condition each region D_n has a finite rectangular structure. In order to show connectedness of the upper and lower components we use the fact that $A_{a,b} \subset D_n$ for all n. Then we show that connectedness implies that all levels of $\mathcal{U}_{a,b}$ and $\mathcal{L}_{a,b}$ appear in the boundary of D_n for some n and all horizontal levels of the boundaries belong to $\mathcal{U}_{a,b} \cup \mathcal{L}_{a,b}$. Using these facts and surjectivity of $F_{a,b}$, which follows from the nesting property of the sets D_n, we conclude that the "jumps" of the step-functions between the lower levels Sb and y_ℓ and between the upper levels y_u and Sa defining the boundary of $D_{a,b}$ satisfy the same equations as the corresponding "jumps" of the boundary of $A_{a,b}$, hence the boundaries coincide and $D_{a,b} = A_{a,b}$.

Notice that generically (almost surely) the finiteness condition comes from the strong cycle property, and in this case, using Corollary 4.6, we obtain a stronger result that establishes the Reduction theory conjecture proposed by Don Zagier:

Theorem 4.7. *If both a and b have the strong cycle property, then for every point $(x,y) \in \bar{\mathbb{R}}^2 \setminus \Delta$ there exists $N > 0$ such that $F^N(x,y) \in D_{a,b}$.*

Remark 4.8. The strong cycle property is not necessary for the Reduction theory conjecture to hold. For example it holds for the two classical expansions $(-1,0)$ and $(-1,1)$ that satisfy only a weak cycle property. Moreover, if $(a,b) \in \mathcal{P}$ satisfies the finiteness condition but not the strong cycle property, then the above result remains true for almost every point (x,y) of the plane. It can be used to describe a "reduction" procedure for (almost) every geodesic on the upper half-plane, and, ultimately, a symbolic coding of the geodesic flow on the modular surface if the (a,b)-expansion admits a so-called "dual" expansion (see Section 6).

5. Exceptional set

The structure of the exceptional set $\mathcal{E} \subset \mathcal{P}$ where the finiteness condition does not hold can be explicitly described. Let us write $\mathcal{E} = \mathcal{E}_{\mathfrak{a}} \cup \mathcal{E}_{\mathfrak{b}}$, where the set $\mathcal{E}_{\mathfrak{a}}$ consists of all points $(a,b) \in \mathcal{P}$ for which a does not satisfy the finiteness condition (i.e. either the set $\mathcal{U}_a$ or $\mathcal{L}_a$ is infinite), and $\mathcal{E}_{\mathfrak{b}}$ consists of all points $(a,b) \in \mathcal{P}$ for which b does not satisfy the finiteness condition (i.e. either the set $\mathcal{U}_b$ or $\mathcal{L}_b$ is infinite). Let $\mathcal{E}_{\mathfrak{b}}^m$ denote the subset of $\mathcal{E}_{\mathfrak{b}}$ such that $a \le T^m Sb \le a+1$. We describe the recursive construction of the exceptional set $\mathcal{E}_{\mathfrak{b}}^m$: one starts with the set

$$\mathcal{T}_{\mathfrak{b}}^m = \{(a,b) \in \mathcal{P} : \frac{b}{b+1} \le T^m Sb \le a+1\}$$

and looks for all values b whose future iterations under the map $\hat{f}$ belong to the interval $[b/(b+1), a+1]$. The following two regions can be obtained at the next stage:

$$\mathcal{T}_{\mathfrak{b}}^{m,m} = \{(a,b) \in \mathcal{T}_{\mathfrak{b}}^{m} : \frac{b}{b+1} \leq T^m S T^m S b \leq a+1\}$$

$$\mathcal{T}_{\mathfrak{b}}^{m,m+1} = \{(a,b) \in \mathcal{T}_{\mathfrak{b}}^{m} : \frac{b}{b+1} \leq T^{m+1} S T^m S b \leq a+1\}\,.$$

Recursively, if $\mathcal{T}_{\mathfrak{b}}^{n_1,n_2,\ldots,n_k}$ is one of the regions obtained after k steps of this construction (where $n_1 = m$ and $n_i \in \{m, m+1\}$ for $2 \leq i \leq k$), then at the next step we get two new sets (possible empty):

$$\mathcal{T}_{b}^{n_1,n_2,\ldots,n_k,m} = \{(a,b) \in \mathcal{T}_{b}^{n_1,n_2,\ldots,n_k} : \frac{b}{b+1} \leq T^m S T^{n_k} S \ldots T^{n_1} S b \leq a+1\}$$

$$\mathcal{T}_{b}^{n_1,n_2,\ldots,n_k,m+1} = \{(a,b) \in \mathcal{T}_{b}^{n_1,n_2,\ldots,n_k} : \frac{b}{b+1} \leq T^{m+1} S T^{n_k} S \ldots T^{n_1} S b \leq a+1\}\,.$$

Now, the exceptional set $\mathcal{E}_{\mathfrak{b}}^{m}$ is obtained as the union of all sets of type

$$\mathcal{E}_{\mathfrak{b}}^{(n_i)} = \bigcap_{k=1}^{\infty} \mathcal{T}_{\mathfrak{b}}^{n_1,n_2,\ldots,n_k}$$

where $n_1 = m$, $n_i \in \{m, m+1\}$ if $i \geq 2$, and the sequence (n_i) is not eventually periodic. If such a set $\mathcal{E}_{\mathfrak{b}}^{(n_i)}$ is non-empty and (a,b) belongs to it, then b is uniquely determined by the expansion $-1/b = \lfloor -n_1, -n_2, \ldots \rceil$. Moreover, one can prove that for every $n \geq 0$, there exist integers $\mathfrak{l}_{A^{(n)}} \geq 2$, $\mathfrak{l}_{B^{(n)}} \geq 1$ such that the sequence (n_i) can be written as a concatenation of blocks

$$A^{(n+1)} = (\underbrace{A^{(n)}, \ldots, A^{(n)}}_{\mathfrak{l}_{A^{(n)}}}, B^{(n)})\,, \quad B^{(n+1)} = (\underbrace{A^{(n)}, \ldots, A^{(n)}}_{\mathfrak{l}_{A^{(n)}}-1}, B^{(n)}) \tag{5.1}$$

or

$$A^{(n+1)} = (A^{(n)}, \underbrace{B^{(n)}, \ldots, B^{(n)}}_{\mathfrak{l}_{B^{(n)}}})\,, \quad B^{(n+1)} = (A^{(n)}, \underbrace{B^{(n)}, \ldots, B^{(n)}}_{\mathfrak{l}_{B^{(n)}}+1})\,, \tag{5.2}$$

starting with $A^{(0)} = m$ and $B^{(0)} = m+1$. It turns out that the two recursive conditions are also sufficient for the set $\mathcal{E}_{\mathfrak{b}}^{(n_i)}$ to be nonempty. If in addition (n_i) is an aperiodic sequence, then $\mathcal{E}_{\mathfrak{b}}^{(n_i)}$ consists of a single point that belongs to the line segment $b - a = 1$ of $\mathcal{P}$. More precisely,

Theorem 5.1. *For any $(a,b) \in \mathcal{P}$, $b \neq a+1$, the finiteness condition holds. The set of exceptions $\mathcal{E}$ to the finiteness condition is an uncountable set of one-dimensional Lebesgue measure zero that lies on the diagonal boundary $b = a+1$ of $\mathcal{P}$.*

This is the last ingredient in the proof of Theorem 1.1.

6. Reduction theory and coding of geodesics

Let $\mathcal{H} = \{z = x + iy : y > 0\}$ be the upper half-plane endowed with the hyperbolic metric, $\mathcal{F} = \{z \in \mathcal{H} : |z| \geq 1,\ |\text{Re } z| \leq \frac{1}{2}\}$ 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. 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 [5, §3.6] for details). The geodesic flow on M is defined as an $\mathbb{R}$-action on SM, $\{\varphi^t\} : SM \to SM$.

The coding procedure for the geodesic flow on the modular surface via continued fraction expansions was presented for the three classical cases in [7]; for a survey on symbolic dynamics of the geodesic flow see also [8]. Here we describe how (a, b)-continued fractions can be used for coding purposes. This is the subject of one of our papers in preparation [10].

We will explain how Theorem 1.1 can be used to describe a reduction procedure for (almost) every geodesic in $\mathcal{H}$. In what follows we will denote the end points of geodesics by u and w, and whenever we refer to geodesics, we use (u, w) as coordinates on $D_{a,b}$.

First, we notice that the orbit of any point in $D_{a,b}$ returns to the subset $\Lambda_{a,b} = F_{a,b}(D_{a,b} \cap \{a \le w \le b\})$ infinitely often.

Definition 6.1. A geodesic in $\mathcal{H}$ from u to w is called (a, b)-*reduced* if $(u, w) \in \Lambda_{a,b}$.

In order to use (a, b)-expansions for coding geodesics we need the notion of a *dual expansion*.

Definition 6.2. The (a, b)-expansion has a *dual* expansion if the reflection of $D_{a,b}$ about the line $y = -x$ is the attractor set of some (a', b')-expansion. If $(a', b') = (a, b)$, then the (a, b)-expansion is called *self-dual*.

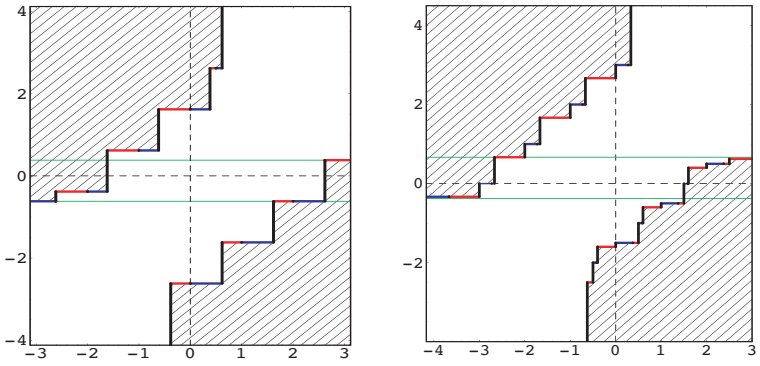

FIGURE 3. Domains of self-dual expansions

In order to determine whether an (a, b)-expansion has a dual, one notices that the parameters (a', b') of the dual must be obtained from the boundary of $D_{a,b}$ as follows: the right vertical boundary of the upper part of $D_{a,b}$ is a ray $x = 1 - b'$, and the left vertical boundary of the lower part of $D_{a,b}$ is a ray of $x = -1 - a'$. Furthermore, the boundary of the lower part of the set $D_{a,b}$ does not have y-levels with $a < y < 0$, and the boundary of the upper part of the set $D_{a,b}$ does not have y-levels with $0 < y < b$, while the boundary of the lower part of the set $D_{a',b'}$ does not have y-levels with $a' < y < 0$, and the boundary of the upper part of the set $D_{a',b'}$ does not have y-levels with $0 < y < b'$. Therefore, we have the following result:

Proposition 6.3. *If the (a, b)-expansion admits a dual expansion, then (a, b) does not satisfy the strong cycle property.*

Thus the parameter pairs $(a,b) \in \mathcal{P} \setminus \mathcal{E}$ that admit dual expansions form a discrete set in $\mathcal{D} \setminus \mathcal{E}$, where

$$\mathcal{D} = \{(a,b) \,|\, -1 \le a \le 0 \le b \le 1, b-a \ge 1\} \subset \mathcal{P},$$

and there are no parameter pairs (a,b) that admit dual expansions in the set $\mathcal{P} \setminus \mathcal{D}$. Their expansions either satisfy a weak cycle property or are periodic. The classical situations of $(-1,0)$- and $(-1,1)$-expansions are self-dual; these expansions satisfy a weak cycle property. Two more sophisticated examples are shown below: $(\frac{1-\sqrt{5}}{2}, \frac{3-\sqrt{5}}{2})$ is periodic and $(-\frac{3}{8}, \frac{2}{3})$ satisfies a weak cycle property. The expansions $(-\frac{1}{n}, 1-\frac{1}{n})$, $n \ge 1$, satisfy a weak cycle property and have dual expansions that are periodic. A classical example in this series is the Hurwitz case $(-\frac{1}{2}, \frac{1}{2})$ whose dual is $(\frac{1-\sqrt{5}}{2}, \frac{-1+\sqrt{5}}{2})$ (see [4, 7]).

In what follows we assume that $(a,b) \in \mathcal{D} \setminus \mathcal{E}$. Then every (a,b)-reduced geodesic from u to w intersects the unit half-circle. Let $C_{a,b} = P \cup Q_1 \cup Q_2$, where P consists of the unit vectors based on the circular boundary of the fundamental region $\mathcal{F}$ pointing inward such that the corresponding geodesic γ on the upper half-plane $\mathcal{H}$ is (a,b)-reduced, Q_1 consists of the unit vectors based on the right vertical boundary of F pointing inward such that $TS(\gamma)$ is (a,b)-reduced, and Q_2 consists of the unit vectors based on the left vertical boundary of $\mathcal{F}$ pointing inward such that $T^{-1}S(\gamma)$ is (a,b)-reduced (see Figure 4). Then a.e. orbit of $\{\varphi^t\}$ returns to $C_{a,b}$, i.e. $C_{a,b}$ is a *cross-section* for $\{\varphi^t\}$, and $\Lambda_{a,b}$ is a parametrization of $C_{a,b}$.

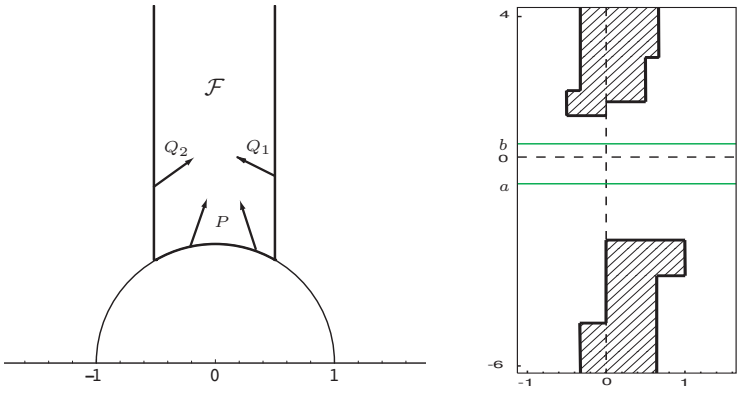

FIGURE 4. The cross-section (left) and its $\Lambda_{a,b}$ parametrization (right)

Let γ be an arbitrary geodesic on $\mathcal{H}$, from u and w, and $w = \lfloor n_0, n_1, \dots \rceil_{a,b}$. We construct the sequence of real pairs $\{(u_k, w_k)\}$ $(k \ge 0)$ defined by $u_0 = u$, $w_0 = w$ and $w_{k+1} = ST^{-n_k} w_k\,, \quad u_{k+1} = ST^{-n_k} u_k\,$. Each geodesic with end points u_k and w_k is $PSL(2,\mathbb{Z})$-equivalent to γ by construction.

According to Remark 4.8, for (almost) every geodesic in $\mathcal{H}$, the above algorithm produces, in finitely many steps, an (a,b)-reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, i.e. there exists a positive integer ℓ such that the geodesic with end points u_ℓ and w_ℓ is (a,b)-reduced. To an (a,b)-reduced geodesic γ, we associate a bi-infinite sequence of integers

$$\lfloor \gamma \rceil = \lfloor \dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots \rceil,$$

its *coding sequence*, by juxtaposing the (a,b)-expansion of $w = \lfloor n_0, n_1, n_2, \dots \rceil_{a,b}$ and the dual (a',b')-expansion of $1/u = \lfloor n_{-1}, n_{-2}, \dots \rceil_{a',b'}$.

The following theorem provides the basis for coding geodesics on the modular surface using (a,b)-coding sequences. If γ is a geodesic on $\mathcal{H}$, we denote by $\bar{\gamma}$ the canonical projection of γ on M.

Theorem 6.4. *If γ is an (a,b)-reduced geodesic, then the first return of $\bar{\gamma}$ to the cross-section $C_{a,b}$ corresponds to a left shift of the coding sequence of γ.*

The geodesic $\bar{\gamma}$ on M can be represented as a bi-infinite sequence of geodesic segments between successive returns to the cross-section $C_{a,b}$. To each segment one can associate the corresponding (a,b)-reduced geodesic γ_i on $\mathcal{H}$. Thus we obtain a sequence of reduced geodesics $\{\gamma_i\}_{i=-\infty}^{\infty}$ representing the geodesic $\bar{\gamma}$. If one associates to γ_i (with end points u, w) its coding sequence $\lfloor \gamma_i \rceil = \lfloor \dots, n_{-1}, n_0, n_1, \dots \rceil$, then $\gamma_{i+1} = ST^{-n_0}(\gamma_i)$, because the map ST^{-n_0} gives the first return to the cross-section $C_{a,b}$. For, notice that $ST^{-n_0}w = \lfloor n_1, n_2, \dots \rceil_{a,b}$ and $1/ST^{-n_0}u = -u + n_0$. Since the (a',b')-expansion is dual to the (a,b)-expansion, $(u,w) \in \Lambda_{a,b}$ implies that $-b' < u \le 1-b' \le -a'$, hence $a' \le -u < b'$, i.e. $1/ST^{-n_0}u = \lfloor n_0, n_{-1}, n_{-2}, \dots \rceil_{a',b'}$ is a legitimate dual expansion, and the left shift of the coding sequence corresponds to the first return to the cross-section. Thus all (a,b)-reduced geodesics γ_i produce, up to a shift, a bi-infinite coding sequence, which we call the (a,b)-code of $\bar{\gamma}$, and denote by $\lfloor \bar{\gamma} \rceil$. We remark that if $\bar{\gamma}$ is a closed geodesic on M then its coding sequence is periodic $w = \lfloor \overline{n_0, n_1, \dots, n_m} \rceil_{a,b}$, $1/u = \lfloor \overline{n_m, \dots, n_1, n_0} \rceil_{a',b'}$.

In conclusion, the geodesic flow becomes a special flow over a symbolic dynamical system $(X_{a,b} \subset \mathcal{N}^{\mathbb{Z}}, \sigma)$, on the infinite alphabet $\mathcal{N} = \mathbb{Z} \setminus \{0\}$, where $X_{a,b}$ is the closure of the set of admissible sequences and σ is the left shift map. The coding map $\mathrm{Cod} : X_{a,b} \to C_{a,b}$

$$\mathrm{Cod}(\lfloor \dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots \rceil) = (1/\lfloor n_{-1}, n_{-2}, \dots \rceil_{a',b'}, \lfloor n_0, n_1, n_2, \dots \rceil_{a,b})$$

is continuous, surjective, and essentially one-to-one.

7. Invariant measures and ergodic properties

Based on the finite rectangular geometric structure of the domain $D_{a,b}$ one can study the measure-theoretic properties of the Gauss-type map $\hat{f}_{a,b} : [a,b) \to [a,b)$,

$$\hat{f}_{a,b}(x) = -\frac{1}{x} - \left\lfloor -\frac{1}{x} \right\rfloor_{a,b}, \quad \hat{f}_{a,b}(0) = 0 \tag{7.1}$$

and its associated natural extension map $\hat{F}_{a,b} : \hat{D}_{a,b} \to \hat{D}_{a,b}$

$$\hat{F}_{a,b}(x,y) = \left(\hat{f}_{a,b}(x), -\frac{1}{y - \lfloor -1/x \rceil_{a,b}} \right). \tag{7.2}$$

We remark that the map $\hat{F}_{a,b}$ is obtained from the map $F_{a,b}$ induced on the set $D_{a,b} \cap \{(x,y) \mid a \le y < b\}$ by a change of coordinates $x' = y$, $y' = -1/x$. Therefore the domain $\hat{D}_{a,b}$ is easily identified knowing $D_{a,b}$ and may be considered to be its "compactification".

We present the simple case when $1 \le -\frac{1}{a} \le b+1$ and $a-1 \le -\frac{1}{b} \le -1$ described in Section 9 of [9]. The general theory is the subject our paper in preparation [11].

The truncated orbits of a and b are

$$\mathcal{L}_a = \left\{ a+1, -\frac{1}{a+1} \right\}, \quad \mathcal{U}_a = \left\{ -\frac{1}{a}, -\frac{a+1}{a} \right\}$$
$$\mathcal{L}_b = \left\{ -\frac{1}{b}, \frac{b-1}{b} \right\}, \quad \mathcal{U}_b = \left\{ b-1, -\frac{1}{b-1} \right\}$$

and the end points of the cycles are $c_a = \frac{a}{a+1}$, $c_b = \frac{b}{1-b}$.

Theorem 7.1. *If $1 \leq -\frac{1}{a} \leq b+1$ and $a-1 \leq -\frac{1}{b} \leq -1$, then the domain $\hat{D}_{a,b}$ of $\hat{F}_{a,b}$ is given by*

$$\hat{D}_{a,b} = [a, -\frac{1}{b}+1] \times [-1, 0] \cup [-\frac{1}{b}+1, a+1] \times [-1/2, 0]$$
$$\cup [b-1, -\frac{1}{a}-1] \times [0, 1/2] \cup [-\frac{1}{a}-1, b] \times [0, 1]$$

and $\hat{F}_{a,b}$ preserves the Lebesgue equivalent probability measure

$$d\nu_{a,b} = \frac{1}{\log[(1+b)(1-a)]} \frac{dxdy}{(1+xy)^2}. \tag{7.3}$$

The description of $\hat{D}_{a,b}$ follows directly from the cycle relations and the finite rectangular structure.

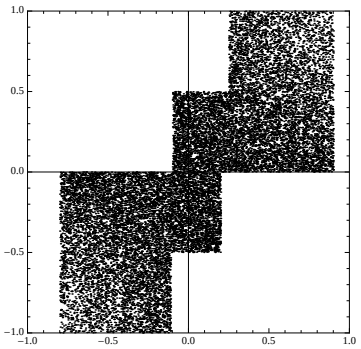

FIGURE 5. Typical domain $\hat{D}_{a,b}$ for the case studied

It is a standard computation that the measure $\frac{dxdy}{(1+xy)^2}$ is preserved by $\hat{F}_{a,b}$. Moreover, the density $\frac{1}{(1+xy)^2}$ is bounded away from zero on $\hat{D}_{a,b}$ and

$$\int_{\hat{D}_{a,b}} \frac{dxdy}{(1+xy)^2} = \log[(b+1)(1-a)] < \infty$$

hence the last part of the theorem is true.

The Gauss-type map $\hat{f}_{a,b}$ is a factor of $\hat{F}_{a,b}$ (projecting on the x-coordinate) so one can obtain its smooth invariant measure $d\mu_{a,b}$ by integrating $d\nu_{a,b}$ over $\hat{D}_{a,b}$ with respect to the y-coordinate as explained in [2]. The measure $d\mu_{a,b}$ is ergodic and the measure-theoretic entropy of $\hat{f}_{a,b}$ can be computed explicitly using Rokhlin's formula.

Theorem 7.2. *The map $\hat{f}_{a,b} : [a,b) \to [a,b)$ is ergodic with respect to the Lebesgue equivalent invariant probability measure*

$$d\mu_{a,b} = \frac{1}{C_{a,b}} \left(\frac{\chi_{(a,-\frac{1}{b}+1)}}{1-x} + \frac{\chi_{(-\frac{1}{b}+1,a+1)}}{2-x} + \frac{\chi_{(b-1,-\frac{1}{a}-1)}}{x+2} + \frac{\chi_{(-\frac{1}{a}-1,b)}}{x+1} \right) dx \tag{7.4}$$

where $C_{a,b} = \log[(1+b)(1-a)]$. The measure-theoretic entropy of $\hat{f}_{a,b}$ is given by

$$h_{\mu_{a,b}}(\hat{f}_{a,b}) = \frac{\pi^2}{3\log[(1-a)(1+b)]}\,. \tag{7.5}$$

References

[1] R. Adler and L. Flatto, *The backward continued fraction map and geodesic flow*, Ergod. Th. & Dynam. Sys., **4** (1984), 487–492. MR 0779707

[2] R. Adler and L. Flatto, *Geodesic flows, interval maps, and symbolic dynamics*, Bull. Amer. Math. Soc. (N.S.), **25** (1991), 229–334. MR 1085823

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

[4] A. Hurwitz, *Über eine besondere Art der Kettenbruch-Entwicklung reeler Grossen*, Acta Math., **12** (1889), 367–405. MR 1554778

[5] S. Katok, "Fuchsian Groups," Chicago Lectures in Mathematics, University of Chicago Press, Chicago, 1992. MR 1177168

[6] S. Katok, *Coding of closed geodesics after Gauss and Morse*, Geom. Dedicata, **63** (1996), 123–145. MR 1413625

[7] S. Katok and I. Ugarcovici, *Arithmetic coding of geodesics on the modular surface via continued fractions*, European women in mathematics—Marseille 2003, 59–77, CWI Tract, 135, Centrum Wisk. Inform., Amsterdam, 2005. MR 2223106

[8] S. Katok and I. Ugarcovici, *Symbolic dynamics for the modular surface and beyond*, Bull. Amer. Math. Soc., **44** (2007), 87–132. MR 2265011

[9] S. Katok and I. Ugarcovici, *Structure of attractors for (a,b)-continued fraction transformations*, preprint.

[10] S. Katok and I. Ugarcovici, *Coding geodesics on the modular surface and (a,b)-continued fractions*, in preparation.

[11] S. Katok and I. Ugarcovici, *Measure-theoretic properties of (a,b)-continued fractions transformations*, in preparation.

[12] L. Luzzi and S. Marmi, *On the entropy of Japanese continued fractions*, Discrete Cont. Dyn. Syst., **20** (2008), 673–711. MR 2373210

[13] P. Moussa, A. Cassa and S. Marmi, *Continued fractions and Brjuno functions*, J. Comput. Appl. Math., **105** (1999), 403–415. MR 1690607

[14] H. Nakada, *Metrical theory for a class of continued fraction transformations and their natural extensions*, Tokyo J. Math., **4** (1981), 399–426. MR 0646050

[15] H. Nakada and R. Natsui, *Some metric properties of α-continued fractions*, Journal of Number Theory, **97** (2002), 287–300. MR 1942961

[16] H. Nakada and R. Natsui, *The non-monotonicity of the entropy of α-continued fraction transformations*, Nonlinearity, **21** (2008), 1207–1225. MR 2422375

[17] C. Series, *On coding geodesics with continued fractions*, Monograph. Enseign. Math., **29** (1981), 67–76. MR 0609896

[18] F. Schweiger, "Ergodic Theory of Fibred Systems and Metric Number Theory," Oxford University Press, New York, 1995. MR 1419320

[19] D. Zagier, "Zetafunkionen und Quadratische Körper: Eine Einführung in die Höhere Zahlentheorie," Springer-Verlag, Berlin-New York, 1981. MR 0631688

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

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

Department of Mathematical Sciences, DePaul University, Chicago, IL 60614, USA

E-mail address: iugarcov@depaul.edu