Reduction theory and coding of geodesics on the modular surface

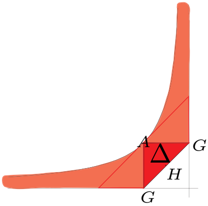
Svetlana Katok

Department of Mathematics
The Pennsylvania State University

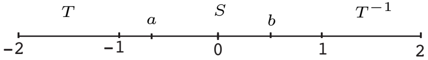
A family of maps $f_{a,b}$

suggested for consideration by Don Zagier.

Let $\mathcal{P} = \{(a,b) \,|\, a \le 0 \le b,\ b-a \ge 1,\ -ab \le 1\} \supset$
$\Delta = \{-1 \le a \le 0 \le b \le 1, b-a \ge 1\}$;



Let $f_{a,b} : \bar{\mathbb{R}} \to \bar{\mathbb{R}}$ be defined as

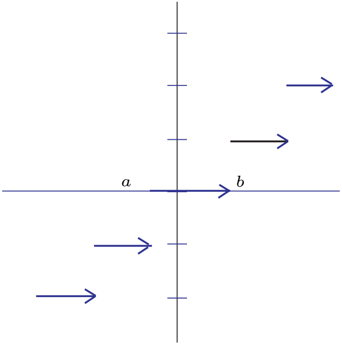


$$f_{a,b}(x) = \begin{cases} x+1 & \text{if } x < a \\ -\dfrac{1}{x} & \text{if } a \le x < b \\ x-1 & \text{if } x \ge b\,. \end{cases}$$

$Tx = x+1, Sx = -\frac{1}{x}$,
$T^{-1}x = x-1$
generators of $SL(2,\mathbb{Z})$

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

(a, b)-continued fractions: joint work with I. Ugarcovici



$(x)_{a,b}$

Theorem

If $(a, b) \in \Delta$, then any x can be expressed uniquely in the form

$$x = n_0 - \cfrac{1}{n_1 - \cfrac{1}{\ddots}} = (n_0, n_1, \cdots)_{a,b}, \;\; n_i \neq 0,$$

where $n_0 = (x)_{a,b}$, $x_1 = -\frac{1}{x - n_0}$ and $n_{i+1} = (x_{i+1})_{a,b}$, $x_{i+1} = -\frac{1}{x_i - n_i}$, i.e. $r_k = (n_0, n_1, \ldots, n_k)_{a,b} = \frac{p_k}{q_k} \to x$.

The natural extension map

Let $F_{a,b} : \bar{\mathbb{R}}^2 \to \bar{\mathbb{R}}^2$, $F_{a,b}(x,y) = \begin{cases} (x+1, y+1) & \text{if } y < a \\ (-\frac{1}{x}, -\frac{1}{y}) & \text{if } a \le y < b \\ (x-1, y-1) & \text{if } y \ge b \end{cases}$

be the (natural) extension map of $f_{a,b}$.

Numerical experiments led Zagier to conjecture that $F_{a,b}$ possesses a global attractor set $D_{a,b} = \cap_{n=0}^{\infty} F_{a,b}^n(\bar{\mathbb{R}}^2)$ with finite rectangular structure on which it is essentially bijective, and every point of the plane is mapped to $D_{a,b}$ after finitely many iterations of $F_{a,b}$.

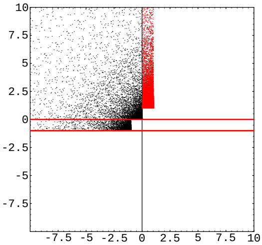
If one identifies a geodesic of the upper half-plane with a pair of real numbers $(u,w) \in \bar{\mathbb{R}}^2,\ u \ne w$ — its endpoints, then $F_{a,b}$ maps geodesics to geodesics, and the existence of an attractor for $F_{a,b}$ corresponds to a reduction of geodesics, hence we perceive $F_{a,b}$ as a reduction map.

Zagier's Reduction theory conjecture

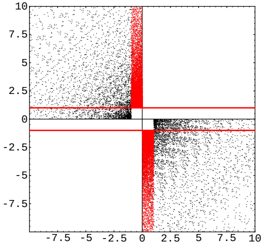
RTC: For any $(a,b) \in \mathcal{P}$ $F_{a,b}$ has an attractor $D_{a,b}$ with finite rectangular structure with the following additional property:

- for every $(u,w) \in \bar{\mathbb{R}}^2$ $\exists$ $N \geq 0$ s.t. $F_{a,b}^N(u,w) \in D_{a,b}$.

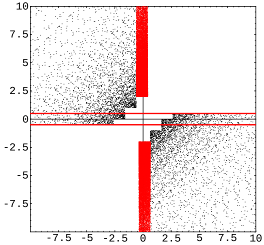
RTC holds for three classical cases:



$A_{-1,0}$
minus c.f.

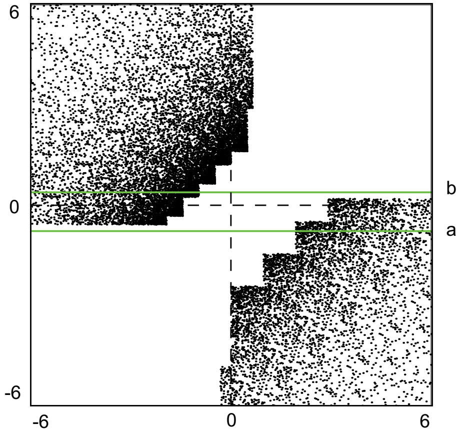


$A_{-1,1}$
alternating c.f.



$A_{-1/2,1/2}$
nearest integer c.f.

Mathematica experimentation: Stage 1



A generic example $a = -\frac{4}{5}, b = \frac{2}{5}$
attractor obtained by iterating
random points

Definition of attractor via trapping region

Definition

$\Theta_{a,b} \subset \bar{\mathbb{R}}^2$ is a trapping region for the reduction map $F_{a,b}$ if

- for every pair $(x, y) \in \bar{\mathbb{R}}^2$, $\exists N > 0$ such that $F_{a,b}^N(x, y) \in \Theta_{a,b}$;
- $F_{a,b}(\Theta_{a,b}) \subset \Theta_{a,b}$.

Definition

We define the attractor starting with the trapping region:
$D_{a,b} = \bigcap_{n=0}^{\infty} D_n$, where $D_n = \bigcap_{i=0}^{n} F_{a,b}^i(\Theta_{a,b})$.

Existence of a trapping region

Theorem

The region $\Theta_{a,b}$ is bounded by step-functions

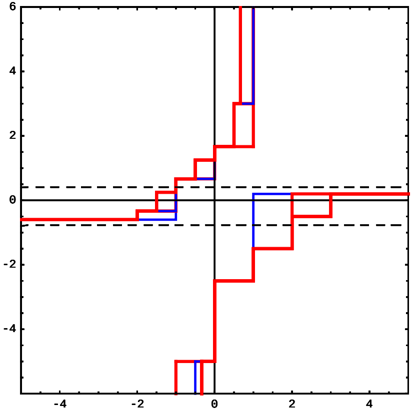
$$u(x) = \begin{cases} b-1 & \text{if } x \le -1 \\ \min(-1/(b-1)-1, -1/a) & \text{if } -1 \le x \le 0 \\ -1/(b-1) & \text{if } 0 \le x \le 1, \end{cases} \text{ and}$$

$$\ell(x) = \begin{cases} -1/(a+1) & \text{if } -1 \le x \le 0 \\ \max(-1/(a+1)+1, -1/b) & \text{if } 0 \le x \le 1 \\ a+1 & \text{if } x \ge 1 \end{cases}$$

is the trapping region for the reduction map $F_{a,b}$.

Observation: $\ell(x)$ and $u(x)$ take the initial values of the orbits of a and b.

Mathematica experimentation: Stage 2



Obtaining attractor from the trapping region

Observation: horizontal boundary levels belong to the orbits of a and b.

Orbits of a and b

The map $f_{a,b} = f$ is discontinuous for $x = a, b$, however, we can look at two maps, one on the left and one on the right of $x = a, b$, and the corresponding split orbits:

$$\mathcal{O}_\ell(a) = \{Ta, fTa, f^2Ta, \dots\},\ \mathcal{O}_u(a) = \{Sa, fSa, f^2Sa, \dots\}$$

and

$$\mathcal{O}_\ell(b) = \{Sb, fSb, f^2Sb, \dots\},\ \mathcal{O}_u(b) = \{T^{-1}b, fT^{-1}b, f^2T^{-1}b, \dots\}.$$

Observation: horizontal segments of the upper boundary of the attractor belong to $\mathcal{O}_u(a)$ and $\mathcal{O}_u(b)$, and of the lower boundary - to $\mathcal{O}_\ell(a)$ and $\mathcal{O}_\ell(b)$, hence we will refer to $\mathcal{O}_\ell(a)$ and $\mathcal{O}_\ell(b)$ as lower, and to $\mathcal{O}_u(a)$ and $\mathcal{O}_u b)$ as upper orbits of a and b.

The cycle property

Our initial experiments showed that the following patern was prevalent and generic:

Definition

- We say that a (resp., b) has the cycle property if the upper and lower orbits meet forming a cycle.
- If the product over the cycle equals the identity we say that the cycle property is strong, otherwise, the cycle property is weak.

Another pattern that we noticed was periodicity of the orbits.

A generic example: $a = -\frac{4}{5}, b = \frac{2}{5}$ - the cycle property!

$$\begin{tikzcd}
 & \boxed{5/4} \arrow[r, "T^{-1}"] & \boxed{1/4} \arrow[dr, "S"] & \\
\boxed{-4/5} \arrow[ur, "S"] \arrow[dr, "T"] & & & \boxed{-4} \\
 & \boxed{1/5} \arrow[r, "S"] & \boxed{-5} \arrow[ur, "T"] &
\end{tikzcd}$$

$$\begin{tikzcd}
 & \boxed{-3/5} \arrow[r, "S"] & \boxed{5/3} \arrow[r, "T^{-1}"] & \boxed{2/3} \arrow[r, "T^{-1}"] & \boxed{-1/3} \arrow[r, "S"] & \boxed{3} \arrow[dr, "T^{-1}"] & \\
\boxed{2/5} \arrow[ur, "T^{-1}"] \arrow[dr, "S"] & & & & & & \boxed{2} \\
 & \boxed{-5/2} \arrow[r, "T"] & \boxed{-3/2} \arrow[r, "T"] & \boxed{-1/2} \arrow[urrr, "S"] & & &
\end{tikzcd}$$

Periodic expansion: $a = \frac{1-\sqrt{5}}{2}$, $b = \frac{-1+\sqrt{5}}{2}$

$$
\begin{tikzcd}
 & \boxed{\frac{1+\sqrt{5}}{2}} \arrow[r, "T^{-1}"] & \boxed{\frac{-1+\sqrt{5}}{2}} \arrow[r, "T^{-1}"] & \boxed{\frac{-3+\sqrt{5}}{2}} \arrow[r, "S"] & \boxed{\frac{3+\sqrt{5}}{2}} \arrow[r, "T^{-1}"] & \boxed{\frac{1+\sqrt{5}}{2}} \arrow[r, "T^{-1}"] & \cdots \\
\boxed{\frac{1-\sqrt{5}}{2}} \arrow[ur, "S"] \arrow[dr, "T"] & & & & & & \\
 & \boxed{\frac{3-\sqrt{5}}{2}} \arrow[r, "S"] & \boxed{-\frac{3+\sqrt{5}}{2}} \arrow[r, "T"] & \boxed{-\frac{1+\sqrt{5}}{2}} \arrow[r, "T"] & \boxed{\frac{1-\sqrt{5}}{2}} \arrow[r, "T"] & \cdots &
\end{tikzcd}
$$

The finiteness condition

Definition

We say that (a, b) satisfies the finiteness condition if a and b either have the cycle property or their expansions are eventually periodic.

Thus the sets $\mathcal{L}_a$, $\mathcal{U}_a$, $\mathcal{L}_b$, and $\mathcal{U}_b$, called the truncated orbits, are finite.

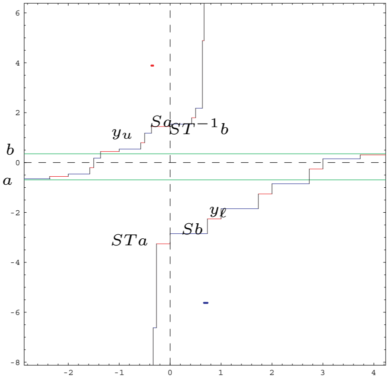
$$\mathcal{L}_a = \begin{cases} \mathcal{O}_\ell(a) & \text{if } a \text{ has periodic expansion} \\ \text{lower part of } a\text{-cycle} & \text{if } a \text{ has the cycle property} \\ \text{lower part of } a\text{-cycle} \cup \{0\} & \text{if } a \text{ has weak cycle property,} \end{cases}$$

$$\mathcal{U}_a = \begin{cases} \mathcal{O}_u(a) & \text{if } a \text{ has periodic expansion} \\ \text{upper part of } a\text{-cycle} & \text{if } a \text{ has the cycle property} \\ \text{lower part of } a\text{-cycle} \cup \{0\} & \text{if } a \text{ has weak cycle property,} \end{cases} \quad \text{and,}$$

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

Observation: Only truncated orbits appear as horizontal levels of the attractor, ends of the cycles appear only if $= 0$, i.e. if the cycle is weak.

Mathematica experimentation: Stage 3



$a = -0.693; b = 0.351$

The boundary of the attractor is computed from the data (a, b): the lower boundary consists of all levels in $\mathcal{L}_a \cup \mathcal{L}_b$ and the upper - of all levels in $\mathcal{U}_a \cup \mathcal{U}_b$. The x-levels are solutions of the (overdetermined) system of fractional-linear equations that is consistent and equivalent to the system of two equations at consecutive levels $y_u < Sa$ and $y_\ell > Sb$.

The system was solved and picture drawn by computer.

Finiteness condition implies finite rectangular structure

Definition

We say that a proper subset of $\bar{\mathbb{R}}^2$ 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 main result of our work is the following:

Theorem (FRS)

If $(a,b) \in \Delta$ satisfies the finiteness condition, the attractor set $D_{a,b} \underset{\neq}{\subset} \bar{\mathbb{R}}^2$ 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 $D_{a,b}$.

- Step 1: construction of a set $A_{a,b}$ where $F_{a,b}$ is bijective by starting with 3 connected levels $STa < Sb < y_\ell$ and proving that all levels in $\mathcal{L}_a \cup \mathcal{L}_b$ are connected, and similarly of upper levels.
- Step 2: $D_{a,b} = A_{a,b}$. $A_{a,b} \subset D_n$ for all $\forall\ n$ used for connectedness.

The set of exceptions $\mathcal{E}$ to the finiteness condition

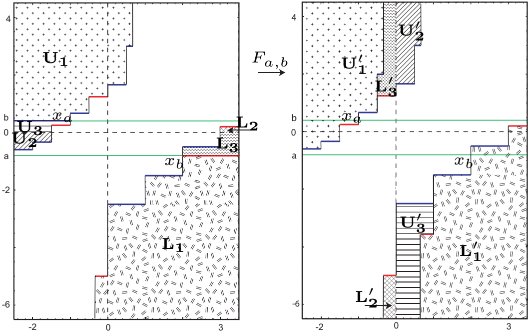
Theorem

- For all $(a,b) \in \Delta \setminus \{$ the diagonal $b = a+1\}$ the finiteness condition is satisfied, hence Theorem FRS holds.
- The exceptional set $\mathcal{E} \in \Delta$ is a uncountable Lebesgue measure 0 on the diagonal $b = a+1$.
- The finiteness condition is necessary for finite rectangular structure of the attractor, i.e. if $(a,b) \in \mathcal{E}$ s.t. $\mathcal{U}_a \cup \mathcal{U}_b \cup \mathcal{L}_a \cup \mathcal{L}_b$ is infinite, either the attractor $D_{a,b}$ is disconnected, or it consists of two connected components whose boundary functions take all values of $\mathcal{U}_a \cup \mathcal{U}_b \cup \mathcal{L}_a \cup \mathcal{L}_b$, hence are not step-functions with finitely many steps.

Reduction theory conjecture

Theorem

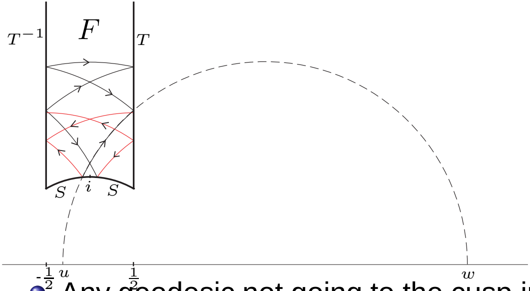
If $a, b \in \Delta$ and both have the strong cycle property, the RTC holds.

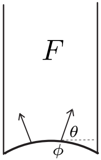


Observation: levels corresponding to the ends of the cycles are inside the attractor, hence any boundary component is mapped inside under some iteration of $F_{a,b}$.

Geometric code for the modular surface

$M = PSL(2,\mathbb{Z})\backslash\mathcal{H}$ – modular surface
$F = \{z \in \mathcal{H} \mid |z| \geq 1, \ |\mathrm{Re}\, z| \leq \frac{1}{2}\}$





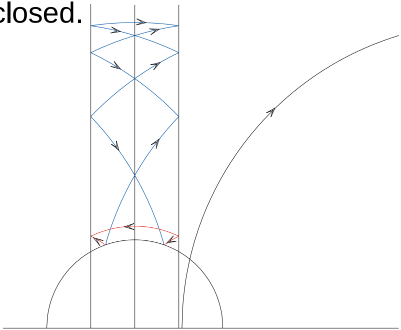
The cross-section B

$T(z) = z + 1$
$S(z) = -1/z$
$[T, T, T, T, S, T^{-1}, T^{-1}, T^{-1}, S] =$
$[4, -3] \Leftarrow$ geometric code

- Any geodesic not going to the cusp in either direction has a geometric code $[\dots n_{-2}, n_{-1}, n_0, n_1, n_2, \dots], \ n_i \neq 0$.
- Any closed geodesic passing through B - axis of $A \in SL(2,\mathbb{Z})$ - has a periodic code $[n_1, n_2, \dots, n_m]$ and $A = T^{n_1} S T^{n_2} S \cdots T^{n_m} S$.
- Left shift σ of the sequence corresponds to the first return to the cross-section B.

Which geometric codes are realized?

Denote the set of all admissible geometric codes by X.
$X \subset \mathcal{N}^{\mathbb{Z}}$, where $\mathcal{N} = \{n \in \mathbb{Z}, |n| \geq 1\}$ – alphabet, is σ–invariant and closed.



Not all sequences of non–zero integers are realized as geometric codes of geodesics on M.

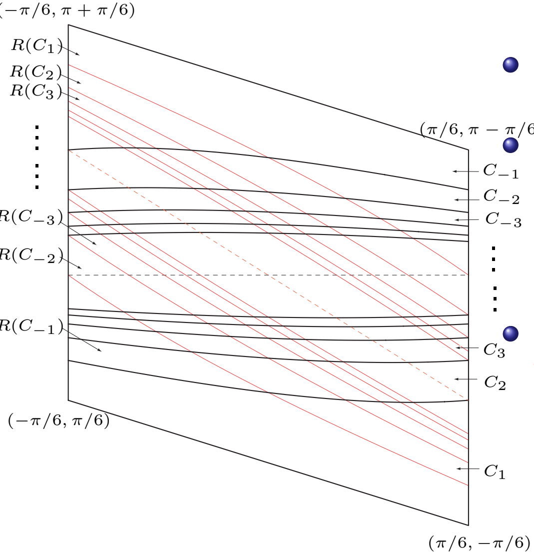
Example

$[8, 2]$ is not a geometric code since the geometric code of the axis of T^8ST^2S is $[6, -2]$ [K].

Mathematica experimentation: drawing geodesics on the modular surface.

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

Mathematica experimentation: the geometric partition



Geometric partition and its image under the return map R

- **Partition:** $B = \sqcup_{n\in\mathcal{N}} C_n$, $C_n = \{v \in B \mid n_1(v) = n\}$.
- If $n_1(v) = n$, $n_2(v) = m$ for some vector $v \in B$, then $R(C_n) \cap C_m \neq \emptyset$. Therefore, 2 cannot be followed by $1, 2, 3, 4, 5$.
- $R(C_n)$ and C_m intersect transversally if and only if $|n|, |m| \geq 2$, and $|1/n + 1/m| \leq 1/2$: geometrically Markov geodesics

A class of admissible geometric codes

Theorem ([KU])

Any bi–infinite sequence $[\ldots, n_{-1}, n_0, n_1, n_2, \ldots]$ *such that* $\left|\frac{1}{n_i} + \frac{1}{n_{i+1}}\right| \leq \frac{1}{2}$ *for* $i \in \mathbb{Z}$*, is realized as a geometric code of a geodesic on* M*.*

We denote this set by X_M *– geometrically Markov codes – it is*

- *a maximal, 1–step countable topological Markov chain in* X*;*
- *the maximal if* X_M *is symmetric (i.e. given by a symmetric transition matrix).*

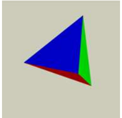
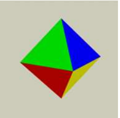
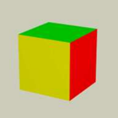
Theorem ([KU, KU1])

Complexity of the geometric code: the space X *of geometric codes is not a finite–step topological Markov chain.*

[KU] S. Katok and I. Ugarcovici, *Geometrically Markov geodesics on the modular surface*, Moscow Math. J., **1** (2005), 135-155.
[KU1] S. Katok and I. Ugarcovici, *Symbolic dynamics for the modular surface and beyond*, Bull. Am. Math. Soc., **44** (2007) 87-132

Positive geodesics

This class includes a class of codes of positive geodesics found earlier in [GK]:$[\gamma] = [\dots, n_{-1}, n_0, n_1, \dots]$, where $\frac{1}{n_i} + \frac{1}{n_{i+1}} \le \frac{1}{2}$ for all i, i.e. $[\gamma]$ does not contain 2 and $\{3,3\}, \{3,4\}, \{4,3\}, \{3,5\}, \{5,3\}$

- all segments comprising γ in F are positively (clockwise) oriented.
- the geometric code of γ coincides with its arithmetic code given by minus continued fractions $(a = -1, b = 0)$. It is obtained by juxtaposing expansions of $w = \lceil n_0, n_1, \dots \rceil$ and and $1/u = \lceil n_{-1}, n_{-2}, \dots \rceil$:
$$\lceil \gamma \rceil = \lceil \dots n_{-2}, n_{-1}, n_0, n_1, n_2, \dots \rceil,$$
where the geodesic from u to w is reduced, i.e. $0 < u < 1, \ w > 1$.

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

Coding via (a, b)-continued fractions

(a, b)-continued fraction expansion can be used for coding if

- it satisfies the RTC, and
- has a dual (or is self-dual).

Definition

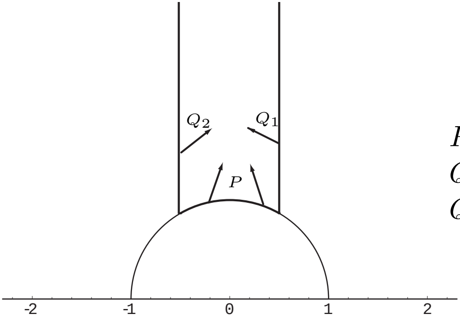
The (a, b)-expansion has a dual if the reflection of $A_{a,b}$ in the line $w = -u$ is an attractor for some (a', b')-expansion.
If $(a', b') = (a, b)$, the (a, b)-expansion is called self-dual.

Definition

A geodesic in $\mathcal{H}$ from u to w is called (a, b)-*reduced* if $(u, w) \in \Lambda_{a,b} = F_{a,b}(D_{a,b} \cap \{a \le y \le b\})$.

Coding via (a, b)-continued fractions

If $(a, b) \in \Delta$, the (a, b)- reduced geodesic from u to w intersects the unit half-circle, and let $C_{a,b} = P \cup Q_1 \cup Q_2$:



P: γ is (a, b)-reduced
Q_1: $TS(\gamma)$ is (a, b)-reduced
Q_2: $T^{-1}S(\gamma)$ is (a, b)-reduced

- $C_{a,b}$ is a cross-section, i.e. a surface inside SM that every geodesic visits infinitely many times.
- $\Lambda_{a,b}$ is a (u, w)-parametrization of $C_{a,b}$.

Coding via (a,b)-continued fractions

- Reduction theory $\Rightarrow$ every geodesic in $\mathcal{H}$ is $PSL(2,\mathbb{Z})$-equivalent to a reduced one.
- Reduced geodesic γ in $\mathcal{H}$ from u to w $\Leftrightarrow$ $(u,w)\in\Lambda_{a,b}$ $\Rightarrow (\gamma)$ - arithmetic code obtained by juxtaposing of expansion for $w=(n_0,n_1,\dots)_{a,b}$ and (dual) expansion for $1/u=(n_{-1},\dots)_{a',b'}$
$$(\gamma)=(\dots n_{-2},n_{-1};n_0,n_1,n_2,\dots)$$
- The left shift σ corresponds to the first return to the cross-section $C_{a,b}$
- $PSL(2,\mathbb{Z})$-invariance is proved via the cross-section.
- Closed geodesics have periodic coding sequences: $w=(n_1,\dots n_m)_{a,b}$, $\frac{1}{u}=(n_m,\dots,n_1)_{a',b'}$.

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