CONTINUED FRACTIONS AND GEODESICS ON THE HYPERBOLIC SURFACE

TONY WU

Abstract. In this expository paper, we develop the correspondence between continued fraction expansions and geodesics on the modular surface, following Artin, Series, and Moeckel. We show that the convergents of a real number β appear as the turning corners of a geodesic aimed at β crossing the Farey tessellation, and that projecting onto quotient surfaces translates arithmetic properties of convergents into cusp visits. Combining this with ergodicity of the geodesic flow, we prove that for almost every β, the convergents are equidistributed among congruence classes: in particular, each of the three parity types of reduced fractions appears with asymptotic frequency $1/3$.

Contents

1. Introduction

Continued fractions and the convergents of a real number β are defined by a purely algorithmic procedure. However, the dynamics of this algorithm, the growth of partial quotients, the distribution of convergents among various arithmetic classes, are chaotic and difficult to analyze directly. On a completely unrelated topic of hyperbolic geometry, many works wanted to study typical behavior and symbolic dynamics of geodesics, meaning relating the behavior of a geodesic by a sequence of its intersections with specified grid-lines. Out of those works arose the completely unexpected fact that the numbers recorded in the symbolic dynamics exactly correspond to the continued fraction expansion of a number.

The main consequence we develop is a result of Moeckel [Moe82, Proposition 2.1]: for almost every real number, the convergents are equidistributed among congruence classes, with each of the three parity types of reduced fractions appearing with asymptotic frequency $1/3$. We assume familiarity with group actions and basic linear algebra; measure-theoretic notions are introduced informally when they appear in Section 5.

Let us recall the definition of the continued fraction expansion.

Definition 1.1. Given a real number $\beta > 0$, define $a_0 = \lfloor \beta \rfloor$, and define $\beta_1 = 1/(\beta - a_0)$. We continue iteratively, defining $a_i = \lfloor \beta_i \rfloor$, and $\beta_{i+1} = 1/(\beta_i - a_i)$, thus obtaining a sequence of positive integers $a_0, a_1, a_2, \ldots$ called the partial quotients, and we write

$$\beta = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots}}} = [a_0;\, a_1, a_2, a_3, \ldots].$$

When β is irrational, this procedure never terminates and produces an infinite sequence of partial quotients. When β is rational, the algorithm terminates after finitely many steps.

Definition 1.2. The k-th convergent of β is the rational number

$$\frac{p_k}{q_k} = [a_0;\, a_1, \ldots, a_k],$$

obtained by truncating the continued fraction at the k-th step. The numerator and denominator may be found by the following recurrence

$$p_k = a_k\, p_{k-1} + p_{k-2}, \qquad q_k = a_k\, q_{k-1} + q_{k-2},$$

with $p_{-1} = 1$, $p_{-2} = 0$, $q_{-1} = 0$, $q_{-2} = 1$.

It is well known that the convergent p_k/q_k is the best rational approximation to β among all fractions with denominator at most q_k.

Artin [Art24] first observed this connection: on the hyperbolic plane, a geodesic is determined by its two endpoints α and β on the real line, and the sequence of intersections of the geodesic with the Farey tessellation encodes the continued fraction expansions of α and β. He showed that two geodesics are equivalent under the modular group if and only if their symbolic codes are shifts of each other. From this, the subject of symbolic dynamics aims to make precise the correspondence between geometric symbols and continued fraction representations, using this relation to draw out arithmetic or geometric results.

In this paper, we explore this correspondence from scratch. Section 2 introduces the hyperbolic plane and demonstrates the correspondence visually as motivation. Section 3 constructs the Farey tessellation rigorously and proves that cutting sequences encode continued fractions. Sections 4 and 5 apply the correspondence, in combination with ergodic theory, to establish Moeckel's equidistribution result.

2. The Hyperbolic Plane

To rigorously explain the phenomenon, we begin with an introduction to the hyperbolic plane. Under the Poincare half plane model, the hyperbolic plane is modeled by the upper-half plane of the complex numbers imbued with a different metric from that of the Euclidean plane. We will present the basic results on isometries and geodesics in this model and construct quotient surfaces by discrete subgroups within the group of isometries.

2.1. **The upper half-plane and geodesics.** The hyperbolic plane is represented as the upper half of the complex plane,

$$\mathbb{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\},$$

equipped with the metric

$$ds^2 = \frac{dx^2 + dy^2}{y^2}.$$

We call ds^2 the hyperbolic metric on $\mathbb{H}$; it induces a length element $ds = \frac{\sqrt{dx^2+dy^2}}{y}$, so a curve $\gamma : [a, b] \to \mathbb{H}$ has hyperbolic length $\int_a^b \frac{\sqrt{\dot{x}(t)^2+\dot{y}(t)^2}}{y(t)} dt$. The corresponding area element $dA = \frac{dxdy}{y}$ defines the hyperbolic area. Throughout this paper, we interchangeably identify the point $(x, y) \in \mathbb{R}^2$ with the complex number $x + iy$, and analogously the upper half plane in the euclidean plane is defined as $\mathbb{H} = \{(x, y) \in \mathbb{R}^2 : y > 0\}$.

Similar to how straight lines minimize the distance between two points in the euclidean plane, a geodesic minimizes the distance between two paths of a general shape.

Definition 2.1. A geodesic in (M, g) is a curve $\gamma : I \to M$ such that for any $t_0 \in I$, there exists $\varepsilon > 0$ with $\gamma|_{[t_0-\varepsilon,\, t_0+\varepsilon]}$ is the shortest path between its endpoints.

Geodesics on the hyperbolic space are easy to describe with the upper half-plane model.

Proposition 2.2 ([Kat92, §1.2]). *The geodesics in $\mathbb{H}$ are exactly the vertical lines $\{x = c\}$ and the semicircles with both endpoints on $\mathbb{R}$.*

Intuitively, the factor of $1/y^2$ in the metric means that distances shrink as y increases. A curve traveling horizontally at large y covers less hyperbolic distance per unit of Euclidean distance, so geodesics bend upward to take advantage of the cheaper metric at larger heights. This intuitively tells us why geodesics are semicircles rather than straight lines.

2.2. **A party trick.** Even without the full machinery, we can already see continued fractions hiding in the geometry of the upper half-plane. Recall that two reduced fractions $\frac{p}{q}$ and $\frac{r}{s}$ are *Farey neighbors* if $|ps - qr| = 1$ (see [HW08, Ch. 3, Thm. 28]). Connect every pair of Farey neighbors by a semicircular geodesic in $\mathbb{H}$, and add vertical geodesics from each integer up to infinity. We obtain a tiling of $\mathbb{H}$ by triangles with curved sides; this is called the Farey tessellation.

Pick an irrational number β and pick a random number $\alpha \in (-1, 0)$, and draw a semicircle in the upper half plane connecting α and β. This semi-circular path threads through the triangular tiles in the Farey tessellation. The path through each triangle T must pass through two of the three sides; the vertex between those two sides is a turning corner of the path. We define these notations formally in Section 3.

Upon repeating this procedure to each triangle that the path between α and β passes through, we obtain an infinite list of turning corners. Miraculously, the turning corners are exactly the convergents of β! We depict this procedure below using $\beta = \sqrt{3}$ and labeling the first 4 turning corners.

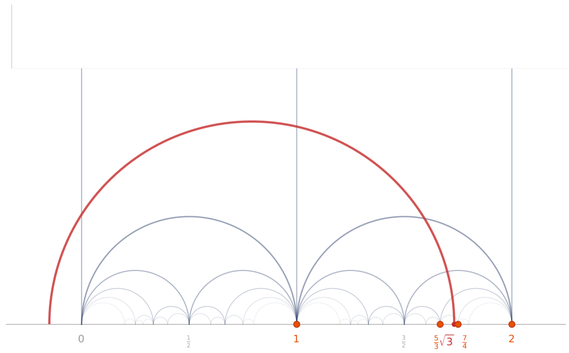

FIGURE 1. A geodesic aimed at $\beta = \sqrt{3}$ crossing the Farey tessellation. The path is marked in red, and the orange dots mark the first 4 turning corners of the path. Figure generated using Python

For $\beta = \sqrt{3}$, the turning corners come out to be

$$1, \quad 2, \quad \frac{5}{3}, \quad \frac{7}{4}, \quad \frac{19}{11}, \quad \ldots$$

Now compute the convergents of $\sqrt{3} = [1;\, 1, 2, 1, 2, \ldots]$ by hand using the recurrence from Section 1:

$$\frac{p_0}{q_0} = \frac{1}{1}, \quad \frac{p_1}{q_1} = \frac{2}{1}, \quad \frac{p_2}{q_2} = \frac{5}{3}, \quad \frac{p_3}{q_3} = \frac{7}{4}, \quad \frac{p_4}{q_4} = \frac{19}{11}, \quad \ldots$$

They are the same sequence.

This is not a special fact for $\sqrt{3}$, it works for every irrational number! The turning corners of the semicircular path aimed at β are always the convergents of β, no matter what β is. Somehow, the geometry of semicircles crossing through the Farey tessellation knows about the arithmetic of continued fractions. The sequence of turning corners does not depend on the choice of starting point $\alpha \in (-1, 0)$; a different α changes only the first few triangles the path crosses before settling into the same sequence of turning corners, so the convergents are determined by β alone.

2.3. **Isometries of the upper half-plane.** Isometries preserve the metric of a space; therefore, they also preserve the shortest paths (geodesics). To describe mappings between geodesics, we naturally want to look for isometries of $\mathbb{H}$.

Since $\mathbb{H}$ sits inside $\mathbb{C}$, we want to look for isometries represented as transformations of $\mathbb{C}$ itself. The orientation-preserving isometries of $(\mathbb{H}, ds^2)$ are the Möbius transformations

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

$\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in $\mathrm{SL}(2, \mathbb{R})$, with an additional property that composition of transformations matching matrix multiplication exactly. Since a matrix and its negative give the same Möbius transformation, the full isometry group is

$$\mathrm{PSL}(2, \mathbb{R}) = \mathrm{SL}(2, \mathbb{R}) / \{\pm I\}.$$

We require two following significant properties. First, geodesics are sent to geodesics under these transformations. Second, the group acts transitively on $\mathbb{H}$: any point can be moved to any other by some isometry. For proofs of these facts, see [Kat92, §1.3].

2.4. **The modular group.** Now we restrict to integer entries. The group

$$\mathrm{PSL}(2,\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a,b,c,d \in \mathbb{Z},\ ad-bc=1 \right\} / \{\pm I\}$$

is a discrete subgroup of $\mathrm{PSL}(2,\mathbb{R})$, called the modular group. It is generated by the two transformations

$$T : z \mapsto z+1, \qquad S : z \mapsto -\frac{1}{z},$$

corresponding to the matrices $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ respectively. The transformation T is a horizontal shift by 1, and S is an inversion and reflection through the unit circle.

2.5. **The fundamental domain and quotient surfaces.** Because $\mathrm{PSL}(2,\mathbb{Z})$ acts on $\mathbb{H}$ by isometries, it partitions $\mathbb{H}$ into congruent copies of a single region. A fundamental domain is a region that contains exactly one representative from each orbit of the group action that excludes the boundary.

PROPOSITION 2.3 ([Kat92, §3.1]). *A fundamental domain for the action of* $\mathrm{PSL}(2,\mathbb{Z})$ *on* $\mathbb{H}$ *is*

$$\mathcal{I} = \left\{ z \in \mathbb{H} : |z| \geq 1,\ |\operatorname{Re}(z)| \leq \tfrac{1}{2} \right\}.$$

In the picture below, the fundamental region and shifts of the fundamental region by compositions of T and S are labeled, each shift labeled by the transformation that takes $\mathcal{I}$ to it.

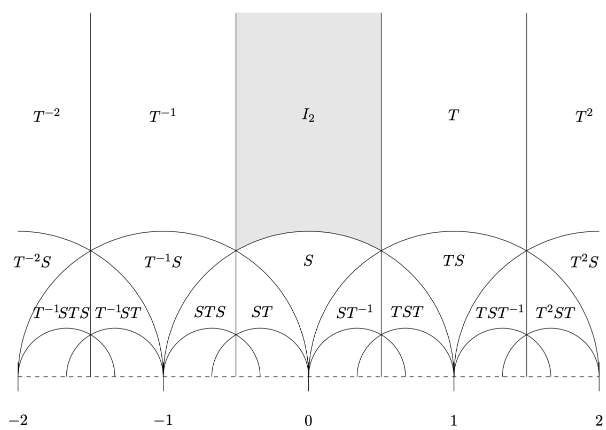

FIGURE 2. Blue section labels the fundamental domain $\mathcal{I}$ for $\mathrm{PSL}(2,\mathbb{Z})$, alongside some of its images under the group action. The left edge is mapped to the right edge under T, and the circular arc is mapped to itself under S. Figure from [Fou].

The region $\mathcal{I}$ is the shaded fundamental domain in Figure 2. Its left vertical edge $\operatorname{Re}(z) = -\frac{1}{2}$ is identified with its right edge $\operatorname{Re}(z) = \frac{1}{2}$ via T, and the circular arc $|z| = 1$ is identified with itself via S. From these boundaries, the quotient

$$\mathcal{M} = \mathrm{PSL}(2,\mathbb{Z})\backslash\mathbb{H}$$

is the modular surface. The quotient $\mathcal{M}$ has finite hyperbolic area equal to $\pi/3$. And the point at infinity captured in the region becomes a puncture, or a cusp, when we consider gluing the geometry of the shape $\mathcal{M}$. (See [Kat92, Ch. 4])

We can do the same procedure to any cofinite discrete subgroup $\Gamma \leq \mathrm{PSL}(2,\mathbb{R})$.

Definition 2.4. A discrete subgroup $\Gamma \leq \mathrm{PSL}(2,\mathbb{R})$ is cofinite if the quotient surface $\Gamma\backslash\mathbb{H}$ has finite hyperbolic area. Under this assumption, $\Gamma\backslash\mathbb{H}$ has the hyperbolic metric from $\mathbb{H}$ and is a finite-area hyperbolic surface. And it has finitely many cusps where the surface extends to points at infinity or the real line.

3. The Farey Tessellation and Cutting Sequences

We now revisit the Farey tessellation from Section 2.2 with the tools of hyperbolic geometry, and prove the correspondence rigorously.

3.1. Construction and invariance.

Recall that two reduced fractions $\frac{p}{q}$ and $\frac{r}{s}$ are Farey neighbors if $|ps - qr| = 1$.

Definition 3.1. The Farey tessellation $\mathcal{F}$ of $\mathbb{H}$ is constructed from the following.

(1) For every pair of Farey neighbors $\frac{p}{q}$ and $\frac{p'}{q'}$, we construct the semicircular geodesic with endpoints $\frac{p}{q}, \frac{p'}{q'}$.
(2) A vertical geodesic from n to ∞ for every integer n.

These arcs decompose the entire half-plane $\mathbb{H}$ into geodesic triangles whose endpoints are always on $\mathbb{R}$ or ∞. We can also imagine this tessellation being built recursively. If $\frac{p}{q}$ and $\frac{p'}{q'}$ are Farey neighbors, the next neighbor to appear between them is known to be their mediant $\frac{p+p'}{q+q'}$ [HW08, Ch. 3, Thm. 29]. So the triangle with vertices being the real numbers $\frac{p}{q}$, $\frac{p'}{q'}$ and $\frac{p+p'}{q+q'}$ is a tile in the tessellation. We can construct the whole tessellation this way, starting from the triangle $(0, 1, \infty)$, where we view ∞ as the fraction $\frac{1}{0}$, and drawing new triangles from mediants.

Proposition 3.2. *The action of* $\mathrm{PSL}(2,\mathbb{Z})$ *is transitive on the triangular tiles of the Farey tessellation* $\mathcal{F}$*.*

Proof. It suffices to check the two generators T and S. The transformation $T : z \mapsto z + 1$ shifts every fraction by 1, sending $\frac{p}{q}$ to $\frac{p+q}{q}$. If $|ps - qr| = 1$, then $|(p + q)s - q(r + s)| = |ps - qr| = 1$, so Farey neighbors map to Farey neighbors.

The transformation $S : z \mapsto -1/z$ sends $\frac{p}{q}$ to $\frac{-q}{p}$. If $|ps - qr| = 1$, then $|(-q)r - p(-s)| = |qr - ps| = 1$, so again Farey neighbors map to Farey neighbors.

Since Möbius transformations send geodesics to geodesics, the edges of $\mathcal{F}$ are permuted by T and S, and therefore by all of $\mathrm{PSL}(2,\mathbb{Z})$.

We first show that $\mathrm{PSL}(2,\mathbb{Z})$ is transitive on the triangles of the Farey tessellation, which also implies it's also transitive on the edges of the Farey tessellation. For a pair of Farey neighbors $\frac{p}{q}$ and $\frac{p'}{q'}$, we note that as $pq' - p'q = \pm 1$, the transformation $g(x) = \frac{p+p'x}{q+q'x} \in \mathrm{PSL}(2,\mathbb{Z})$ sends $(0, 1, \infty)$ to $(\frac{p}{q}, \frac{p+p'}{q+q'}, \frac{p'}{q'})$. As any Farey triangle is of form $(\frac{p}{q}, \frac{p+p'}{q+q'}, \frac{p'}{q'})$ for $|pq' - p'q| = 1$, our map is thus transitive.

To show that $\mathrm{PSL}(2,\mathbb{Z})$ is a bijection on the edges of $\mathcal{F}$, we now have to show it is also injective. Note that the set of isometries of $\mathbb{H}$, which contains the modular group, is a bijection on the upper half plane, so actions of the modular group cannot send two disjoint edges to the same image. Thus elements of the modular group are also injective on the edges of the tessellation. □

3.2. **Cutting sequences.** Cutting sequences concretely connect the geodesics to continued fractions. We will get a cutting sequence by placing a geodesic in the upper half plane and seeing what edges it hits. This identification is due to the work of Series [Ser85, Theorem A].

Definition 3.3. Let γ be an oriented geodesic in $\mathbb{H}$ from α to β with $\alpha < 0 < 1 < \beta$, both irrational. As γ crosses the edges of $\mathcal{F}$, it passes through a sequence of ideal triangles. In each triangle, γ enters through one edge and exits through another; the vertex between the two edges is uniquely determined, and it lies either to the left or to the right of γ.

We record:

- L if the opposite vertex lies to the left of γ,
- R if the opposite vertex lies to the right of γ.

This is pictured below.

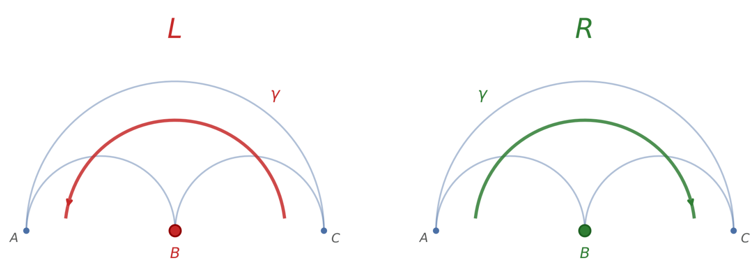

Figure 3. An ideal triangle with vertices A, B, C. The geodesic γ crosses two of the three edges, and the third vertex is labeled as B. Left: B is to the left of γ, so we record L. Right: B is to the right, so we record R. Figure generated from Tikz

Now starting from any point on the geodesic, we can obtain a sequence of left and right turns of the geodesic in each triangle, we concatenate sequences of right turns $RR\dots$ as R^n, similarly for the left turns. The resulting sequence $L^{n_0}R^{n_1}L^{n_2}\cdots$ is the cutting sequence of γ with respect to $\mathcal{F}$.

If instead of labeling the turns, we labeled the corner B and obtained a sequence of corners, these are exactly the turning corners from Section 2.2, and they are the convergents of the number β.

3.3. **The correspondence.** We can now state and prove the result observed in Section 2.2.

Theorem 3.4 (Series [Ser85, Theorem A]). *Let $\beta = [a_0; a_1, a_2, \dots] > 1$ be irrational, and let γ be the geodesic from some $\alpha \in (-1, 0)$ to β. Then:*

(1) The cutting sequence of γ is $L^{a_0}\,R^{a_1}\,L^{a_2}\,R^{a_3}\,\cdots$, where $a_0, a_1, a_2, \dots$ are the partial quotients of β.
(2) The vertex of $\mathcal{F}$ at the transition from the k-th block to the $(k+1)$-th block is the convergent p_k/q_k.

A key observation in the proof is that we can transform the hyperbolic space using the isometries of T and S to mimic the continued fractions algorithm, ensuring our geodesics continually approach the remainders of β during the continued fractions algorithm.

Proof. We induct on the blocks of the cutting sequence. For visual reference, refer back to figure 1.

Since $\alpha \in (-1, 0)$ and $\beta > 1$, the geodesic γ first crosses the vertical edge from 0 to ∞. Then, it passes through the vertical edges $x = 1, x = 2, \ldots, x = a_0 - 1$, each time going through a triangle whose labeled vertex ∞ is on the left of γ, so we record L at each crossing. At $x = a_0 = \lfloor \beta \rfloor$, since $\beta < a_0 + 1$, the geodesic no longer crosses the next vertical line and instead exits through a non-vertical edge, the labeled corner is no longer ∞, instead is a_0, which is below and to the right of γ, so we record an R. This gives the first block L^{a_0} and the next turn R, and the vertex at the transition is $a_0/1 = p_0/q_0$.

Now apply $S \circ T^{-a_0}$ to the entire picture. This is an element of $\mathrm{PSL}(2, \mathbb{Z})$, so it preserves our tessellation by Proposition 3.2. On the endpoint, it acts by

$$\beta \;\mapsto\; \beta - a_0 \;\mapsto\; \frac{-1}{\beta - a_0}.$$

Since $\beta - a_0 \in (0, 1)$, the new forward endpoint is $\beta_1 = \frac{1}{\beta - a_0} = [a_1; a_2, a_3, \ldots] > 1$, while the starting point becomes $\frac{-1}{\alpha - a_0}$, which is again between $(-1, 0)$, as $\alpha - a_0 < -1$. The transformation $S : x \mapsto -1/x$ reverses the cyclic order of points on $\mathbb{R} \cup \{\infty\}$: if $a < b < c$ on the real line, then $S(a), S(b), S(c)$ appear in the opposite order. Since each ideal triangle of the Farey tessellation has its three vertices ordered along $\mathbb{R} \cup \{\infty\}$, applying S flips every triangle, exchanging the side that lies to the left of the geodesic with the side that lies to the right. This orientation swaps all L and R turns.

The remaining cutting sequence in the new coordinates therefore begins with L^{a_1}, which reads as R^{a_1} in the original labeling. The remaining cutting sequence in the new coordinates is the cutting sequence of a geodesic aimed at β_1 with L and R swapped. By the same traversing logic as above, this begins with L^{a_1}, which reads as R^{a_1} in the original labeling. Continuing gives $L^{a_0}\, R^{a_1}\, L^{a_2}\, R^{a_3} \cdots$, proving (1).

For (2), the transition vertex in the base case is ∞. After k steps, the transition vertex in the original coordinates is the image of ∞ by reversing the transformations of (ST^{-a_i}), so it is the composed transformation $(T^{a_0}S)(T^{a_1}S)\cdots(T^{a_k}S)$. Since T has matrix $\left(\begin{smallmatrix} 1 & a_j \\ 0 & 1 \end{smallmatrix}\right)$ and S has matrix $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, their product $T^{a_j}S$ corresponds to the matrix below

$$\left(\begin{smallmatrix} 1 & a_j \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} a_j & -1 \\ 1 & 0 \end{smallmatrix}\right)$$

and composing the transformations $(T^{a_0}S)(T^{a_1}S)\cdots(T^{a_k}S)$ we get

$$\begin{pmatrix} a_0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a_1 & -1 \\ 1 & 0 \end{pmatrix}\cdots\begin{pmatrix} a_k & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} p_k & -p_{k-1} \\ q_k & -q_{k-1} \end{pmatrix}$$

by the convergent recurrence, this matrix sends $\infty = \frac{1}{0}$ to $\frac{p_k}{q_k}$. □

Remark 3.5. Since the partial quotients $a_0, a_1, a_2, \ldots$ are determined by β alone, the cutting sequence after the geodesic crosses the imaginary axis (the edge from 0 to ∞) depends only on β. The assumption $\beta > 1$ ensures that the first block of the cutting sequence is of type L, giving a clean inductive start; for general irrational β, the integer part $a_0 = \lfloor \beta \rfloor$ simply accounts for the initial L^{a_0} block before the first type change. The rationality or irrationality of α plays no role, as it only affects the geodesic's backward endpoint; we require only that α not be a vertex of $\mathcal{F}$.

At this point, we have a complete correspondence between continued fractions and geodesics: the partial quotients of β are the run-lengths of the cutting sequence, and the convergents of β are the turning corners. Next we want to use this correspondence. By projecting geodesics onto quotient surfaces $G\backslash\mathbb{H}$ with multiple cusps, we translate arithmetic properties of convergents into geometric properties of trajectories, and then use the powerful machinery of ergodic theory to answer distributional questions.

4. Arithmetic from Symbols

So far, we established a relationship between continued fractions of β, its turning sequence, and the turning corners of the geodesic. Now we want to use these relations to obtain results about the behavior of convergents themselves.

Our geodesics project onto the quotient surface $\mathcal{M} = \mathrm{PSL}(2,\mathbb{Z})\backslash\mathbb{H}$, but on this quotient all turning corners collapse into the singular cusp of $\mathcal{M}$. The resulting corner sequence is uninteresting as it would only be an infinite sequence of the same cusp.

The idea of [Moe82] is to refine the picture by taking a quotient with respect to a smaller group, so that numbers with different arithmetic properties end up becoming different corners of the quotient surface. Instead of taking all of $\mathrm{PSL}(2,\mathbb{Z})$, we choose a subgroup G that identifies fewer rationals, so that the quotient surface $G\backslash\mathbb{H}$ has multiple cusps. Each cusp corresponds to one G-orbit of rationals, and a geodesic visiting a particular cusp picks out a convergent in that orbit. If we can identify this G so that its orbits match up an arithmetic property, then using the randomness of the geodesic, we can draw conclusions about the distribution of convergents around this arithmetic property.

4.1. Cofinite subgroups and their cusps.

Recall Definition 2.4, since we consider finite-index subgroups of $\mathrm{PSL}(2,\mathbb{Z})$. They belong as cofinite subgroups of $\mathrm{PSL}(2,\mathbb{R})$ since any such subgroup inherits discreteness from $\mathrm{PSL}(2,\mathbb{Z})$ and its quotient has area equal to $[\mathrm{PSL}(2,\mathbb{Z}) : G] \cdot \pi/3$, which is finite.

Definition 4.1. Let G be a finite-index subgroup of $\mathrm{PSL}(2,\mathbb{Z})$. A G-cusp is an equivalence class of points in $\mathbb{Q} \cup \{\infty\}$ under the action of G. The set of G-cusps is finite and are the topological punctures of the quotient surface $G\backslash\mathbb{H}$.

When $G = \mathrm{PSL}(2,\mathbb{Z})$, the action on $\mathbb{Q} \cup \{\infty\}$ is transitive, so there is only one cusp. For a proper subgroup G, the rationals split into several orbits, and each orbit gives a separate cusp on the quotient.

We can try to use these G-cusps to capture arithmetic properties. An arithmetic property of rationals corresponds to a partition of $\mathbb{Q} \cup \{\infty\}$ into classes, and we want to find a subgroup G whose orbits are exactly those classes. If the above method succeeds, then the cusps of $G\backslash\mathbb{H}$ are labeled by the property, and a turning corner p_k/q_k being in a particular class is the same as the geodesic on the quotient surface visiting the corresponding cusp.

4.2. **Geodesic visits track arithmetic classes.**

Proposition 4.2. *Let G be a finite-index subgroup of* $\mathrm{PSL}(2,\mathbb{Z})$*, and let γ be a geodesic on $\mathbb{H}$. The sequence of G-cusps visited by the projected geodesic on $G\backslash\mathbb{H}$ is the sequence of G-orbits of the convergents p_k/q_k of β.*

Proof. By Theorem 3.4, the convergent corner of γ at the k-th transition is the convergent p_k/q_k, viewed as a vertex of the Farey tessellation on the boundary of $\mathbb{H}$. Under the projection $\mathbb{H} \to G\backslash\mathbb{H}$, this vertex maps to its G-equivalence class, which is by definition a G-cusp. The geodesic passing near p_k/q_k on $\mathbb{H}$ corresponds to the projected geodesic passing near that cusp on the quotient. □

So choosing a subgroup G amounts to choosing which arithmetic property of the convergents we wish to track. Different subgroups give different properties.

4.3. **Tracking parity via** $\Gamma(2)$**.** We can use this trick to capture the parity of the numerator and denominator of a convergent. Since $\gcd(p_k, q_k) = 1$, the pair $(p_k \bmod 2,\, q_k \bmod 2)$ takes one of three values: $(1,1)$, $(1,0)$, or $(0,1)$. We need a subgroup whose orbits on the rationals are exactly these three classes.

Definition 4.3. The principal congruence subgroup of level 2 is

$$\Gamma(2) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}(2,\mathbb{Z}) : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod 2 \right\}.$$

Here the congruence condition is imposed on representatives in $\mathrm{SL}(2,\mathbb{Z})$.

This is a normal subgroup of $\mathrm{PSL}(2,\mathbb{Z})$ of index 6, since $\Gamma(2)$ is the kernel of the mod 2 map $\mathrm{PSL}(2,\mathbb{Z}) \to \mathrm{PSL}(2,\mathbb{Z}/2\mathbb{Z})$ and $\mathrm{PSL}(2,\mathbb{Z}/2\mathbb{Z})$ has size 6. It is cofinite, as every finite index subgroup of a cofinite group is cofinite.

Proposition 4.4. *The $\Gamma(2)$-orbits on $\mathbb{Q}\cup\{\infty\}$ are exactly the three parity classes of reduced fractions:* (*odd*, *odd*), (*odd*, *even*), *and* (*even*, *odd*).

Proof. Let $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma(2)$ act on $\frac{p}{q}$, sending it to $\frac{ap+bq}{cp+dq}$. Since $a \equiv d \equiv 1$ and $b \equiv c \equiv 0 \pmod 2$, we have

$$ap + bq \equiv p \pmod 2, \qquad cp + dq \equiv q \pmod 2,$$

so $\Gamma(2)$ preserves parity. There are three parity classes.

To see that each parity class is a single $\Gamma(2)$-orbit, consider the class $(\text{odd}, \text{even})$, which contains $1/0 = \infty$. Any other fraction p/q with p odd and q even can be written as $g(\infty)$ for some $g \in \mathrm{PSL}(2,\mathbb{Z})$, since $\mathrm{PSL}(2,\mathbb{Z})$ acts transitively on $\mathbb{Q} \cup \{\infty\}$. Any two elements of $\mathrm{PSL}(2,\mathbb{Z})$ sending ∞ to fractions with the same parity type differ by right multiplication by an element of $\Gamma(2)$, so they give the same $\Gamma(2)$-orbit. The other two classes follow by the same argument. For full details, see [Moe82, §2]. □

By Proposition 4.2, the quotient $\Gamma(2)\backslash\mathbb{H}$ has three cusps—one for each parity type—and a geodesic aimed at β visits these cusps in the order of the parity types of the convergents of β.

This means we can translate the arithmetic question of "what is the parity sequence of the convergents of β?" to the geometric question "which cusps does the geodesic visit, and in what order?"

The following pictures depict the modular surface and the quotient surface by $\Gamma(2)$ next to each other, capturing their topological shapes but not their scale.

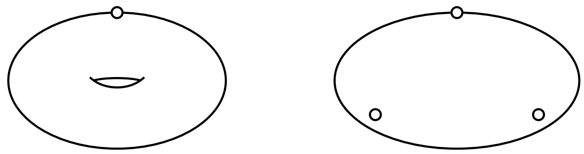

FIGURE 4. Comparison of two quotient surfaces. Left: the modular surface $\mathcal{M}$ has a single cusp because $\mathrm{PSL}(2,\mathbb{Z})$ acts transitively on $\mathbb{Q} \cup \{\infty\}$. Right: the quotient $\Gamma(2)\backslash\mathbb{H}$ has three cusps, one for each parity class of reduced fractions. A geodesic visiting a cusp corresponds to a convergent in that parity class. Figure generated from TikZ.

4.4. **Visiting $\Gamma(m)$.** The same construction applies with other subgroups. For any positive integer m, the principal congruence subgroup

$$\Gamma(m) = \ker\big(\mathrm{PSL}(2,\mathbb{Z}) \to \mathrm{PSL}(2,\mathbb{Z}/m\mathbb{Z})\big)$$

has finite index in $\mathrm{PSL}(2,\mathbb{Z})$. By the same mod-$m$ computation as above, its orbits on the rationals are determined by the residues of p and q modulo m, so the cusps of $\Gamma(m)\backslash\mathbb{H}$ are labeled by the residue classes of reduced fractions modulo m.

We have obtained a translation between the arithmetic world and the geometric. Numerical properties of the convergents of β correspond to which cusps the geodesic visits on the appropriate quotient surface. If we can say that most geodesics are random enough to visit cusps proportional to their size, then we can conclude that the distribution of the arithmetic properties is equivalent to the distribution of the sizes of cusps.

5. Ergodicity and Distribution of Convergents

To finally answer the eventual distribution of arithmetic properties, we need to calculate the random likelihood to visit any cusp associated with an arithmetic class of the property.

5.1. **The geodesic flow and its measure.** To track a geodesic, we must record both its current position and the direction it is moving. The natural space is the *unit tangent bundle*

$$T^1(G\backslash\mathbb{H}) = \{(p,v) : p \in G\backslash\mathbb{H},\ v \in T_p(G\backslash\mathbb{H}),\ \|v\|_{\mathrm{hyp}} = 1\},$$

whose points are pairs consisting of a location p on the surface and a unit tangent vector v at p specifying a direction of travel. The *geodesic flow* $\varphi_t : T^1(G\backslash\mathbb{H}) \to T^1(G\backslash\mathbb{H})$ moves a pair (p,v) forward along the geodesic through p in direction v for time t, at unit speed.

To talk about the fraction of time a trajectory spends in a region, we need a measure on $T^1(G\backslash\mathbb{H})$ that is preserved by φ_t. The hyperbolic area $dx\,dy/y^2$ is, up to scaling, the unique $\mathrm{SL}(2,\mathbb{R})$-invariant measure on $\mathbb{H}$; pairing it with the uniform measure $d\theta$ on the unit tangent circle gives a natural candidate.

DEFINITION 5.1. The *Liouville measure* on $T^1(G\backslash\mathbb{H})$ is the product of the hyperbolic area on $G\backslash\mathbb{H}$ and the uniform arclength measure on the unit tangent circle. In local coordinates (x, y, θ), where (x, y) is a point on the surface and θ parametrizes the direction of the tangent vector,

$$d\mu = \frac{dx\, dy\, d\theta}{y^2}.$$

Since the geodesic flow moves every point forward at unit speed, it preserves $d\mu$; and for the cofinite surfaces we consider, the total volume $\mu(T^1(G\backslash\mathbb{H}))$ is finite.

Definition 5.2. Let (X, μ) be a finite measure space and $\varphi_t : X \to X$ a measure-preserving flow. The flow is ergodic if every φ_t-invariant measurable set has measure either 0 or $\mu(X)$.

The following theorem of Birkhoff equates the time-average along a geodesic flow and the space average along the whole space.

Theorem 5.3 (see [Bir31]). *Let (X, μ) be a finite measure space, $\varphi_t : X \to X$ an ergodic measure-preserving flow, and $f \in L^1(X, \mu)$. Then for μ-almost every $x \in X$,*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f(\varphi_t(x))\, dt = \frac{1}{\mu(X)} \int_X f\, d\mu.$$

Theorem 5.4 (see [AA68]). *Let G be a cofinite subgroup of* $\mathrm{PSL}(2, \mathbb{R})$*. The geodesic flow on $T^1(G\backslash\mathbb{H})$ is ergodic with respect to the Liouville measure.*

So for the surfaces we have been considering, almost every geodesic spends a fraction of its time in any given region equal to the Liouville volume of that region divided by the total.

5.2. **Sizes of cusps and arithmetic classes.** To apply Theorem 5.4, we would need the size of each region near every set of cusps encoding the arithmetic class. Recall from Section 2 that the modular surface $\mathrm{PSL}(2, \mathbb{Z})\backslash\mathbb{H}$ has a single cusp at ∞. When we pass to the smaller subgroup G, if $\mathrm{PSL}(2, \mathbb{Z}) = GS_1 + GS_2 + \dots GS_n$ is the coset decomposition of $\mathrm{PSL}(2, \mathbb{Z})$ with respect to the subgroup G, then we can get a fundamental domain for G as $S_1\mathcal{I} \cup S_2\mathcal{I} \cup \dots S_n\mathcal{I}$ which is $[\mathrm{PSL}(2, \mathbb{Z}) : G]$ copies of $\mathcal{I}$. Because the fundamental region $\mathcal{G}$ is a union of translates of $\mathcal{I}$, we can always map from $\mathcal{G}$ to $\mathcal{I}$ just by mapping to the specific copy of the translate of $\mathcal{I}$. Under this map, the single cusp of $\mathcal{I}$ has many preimages in $\mathcal{G}$.

These preimages group together around cusps of $\mathcal{G}$ in bundles, and the width of a cusp C_j is defined as the number of copies in its bundle, which is the same as the number of preimages of the cusp of $\mathcal{I}$ that lie inside that cusp.

Definition 5.5. Let $C_1, \dots, C_r$ be the cusps of $G\backslash\mathbb{H}$. The width w_j of cusp C_j is the number of $\mathrm{PSL}(2, \mathbb{Z})$-translates of the standard cusp region of $\mathcal{I}$ that map to the cusp region of C_j in $G\backslash\mathbb{H}$.

By construction, the widths add up to the total number of translates:

$$\sum_{j=1}^{r} w_j = [\mathrm{PSL}(2, \mathbb{Z}) : G].$$

And since each translation of the cusp is an isometric copy of the same region of $\mathcal{I}$, the Liouville volume of the cusp neighborhood of C_j is exactly w_j times the Liouville volume of the standard cusp neighborhood.

We need one last technical restriction before stating our result. Moeckel defined an admissible subgroup $G \leq \mathrm{PSL}(2,\mathbb{Z})$ if G does not fix any point of $\mathbb{H}$. This condition ensures that the quotient $G\backslash\mathbb{H}$ is a genuine surface and that the cusps are punctures, which is what lets us integrate over disjoint cusp neighborhoods in the argument below. Now, Moeckel has noted in [Moe82, §1] that principal congruence subgroups $\Gamma(m)$ for $m \geq 2$ are all admissible.

Definition 5.6. Let $\mathfrak{I}$ denote the standard fundamental domain for $\mathrm{PSL}(2,\mathbb{Z})$. The standard cusp region of $\mathfrak{I}$ is the subset

$$\mathfrak{I}_\infty = \{z \in \mathfrak{I} : \mathrm{Im}(z) \geq 1\},$$

the region above height 1 in $\mathfrak{I}$. This forms a neighborhood of the cusp at ∞ on the modular surface.

For a finite-index subgroup $G \leq \mathrm{PSL}(2,\mathbb{Z})$ of index n, a fundamental domain for G is assembled from n copies $s_1\mathfrak{I}, \ldots, s_n\mathfrak{I}$ of the standard fundamental domain, via a coset decomposition $\mathrm{PSL}(2,\mathbb{Z}) = Gs_1 \cup \cdots \cup Gs_n$ (see [Moe82, §1]). Each copy $s_j\mathfrak{I}$ contributes its own cusp region $s_j\mathfrak{I}_\infty$, and these cusp regions group around the cusps of $G\backslash\mathbb{H}$. The *width* $w(C_j)$ of a G-cusp C_j is the number of translates $s_i\mathfrak{I}_\infty$ that map to the cusp neighborhood of C_j on the quotient surface. Since each translate is an isometric copy of $\mathfrak{I}_\infty$, the Liouville volume of the cusp neighborhood of C_j in $T^1(G\backslash\mathbb{H})$ equals $w(C_j)$ times the Liouville volume of a single standard cusp region:

$$\mathrm{Vol}(N_j) = w(C_j) \cdot \mathrm{Vol}(\mathfrak{I}_\infty \times S^1).$$

The widths satisfy $\sum_j w(C_j) = [\mathrm{PSL}(2,\mathbb{Z}) : G]$. Thus the fraction of total cusp volume belonging to C_j is $w(C_j)/[\mathrm{PSL}(2,\mathbb{Z}) : G]$, and by ergodicity this is the asymptotic frequency with which a typical geodesic visits C_j.

Theorem 5.7 (Moeckel [Moe82, Proposition 2.1]). *Let G be an admissible subgroup of $\mathrm{PSL}(2,\mathbb{Z})$, and let $C_1, \ldots, C_r$ be its cusps with widths $w_1, \ldots, w_r$. For Lebesgue-almost every $\beta \in \mathbb{R}$, the convergents p_k/q_k are distributed among the G-orbits with asymptotic frequencies*

$$\lim_{N\to\infty} \frac{1}{N}\big|\{k \leq N : p_k/q_k \in C_j\}\big| = \frac{w_j}{[\mathrm{PSL}(2,\mathbb{Z}) : G]}.$$

Proof. By Proposition 4.2, the sequence of convergent classes is exactly the sequence of cusps visited by the projected geodesic on $G\backslash\mathbb{H}$. Fix disjoint neighborhoods $N_1, \ldots, N_r$ around the cusps $C_1, \ldots, C_r$ built from the standard cusp region of $\mathfrak{I}$. By the discussion above,

$$\mathrm{Vol}(N_j) = w_j \cdot v_0$$

where v_0 is the volume of a single copy of the standard cusp region, and the total volume of $\bigsqcup_j N_j$ is $\big(\sum_j w_j\big)v_0 = [\mathrm{PSL}(2,\mathbb{Z}) : G] \cdot v_0$.

By Theorem 5.4, the geodesic flow on $T^1(G\backslash\mathbb{H})$ is ergodic with respect to Liouville measure, so by the Birkhoff ergodic theorem the asymptotic proportion of time almost every geodesic spends time in N_j equals

$$\frac{\mathrm{Vol}(N_j)}{\mathrm{Vol}\big(\bigsqcup_i N_i\big)} = \frac{w_j v_0}{[\mathrm{PSL}(2,\mathbb{Z}) : G] \cdot v_0} = \frac{w_j}{[\mathrm{PSL}(2,\mathbb{Z}) : G]}.$$

Since a cusp visit around N_j puts the convergent in the j^{th} arithmetic class, and that the above conclusion holds for almost every initial endpoint β, this ratio $\frac{w_j}{[\mathrm{PSL}(2,\mathbb{Z}):G]}$ is the asymptotic frequency of convergents in class C_j for almost all β. ☐

5.3. **Parity of convergents.** Setting $G = \Gamma(2)$, the index is $[\mathrm{PSL}(2,\mathbb{Z}) : \Gamma(2)] = 6$. Since $\Gamma(2)$ is normal in $\mathrm{PSL}(2,\mathbb{Z})$, the quotient $\mathrm{PSL}(2,\mathbb{Z})/\Gamma(2)$ acts on $\Gamma(2)\backslash\mathbb{H}$ by isometries permuting the three cusps transitively. Cusp width is preserved under such isometries, so all three cusps have the same width w, and $3w = 6$ gives $w = 2$. Theorem 5.7 then yields:

Corollary 5.8. *For Lebesgue-almost every $\beta \in \mathbb{R}$, the convergents p_k/q_k are equidistributed among the three parity types, each appearing with asymptotic frequency $\frac{1}{3}$.*

The same argument extends to congruence properties modulo any integer $m \geq 2$.

Corollary 5.9. *Fix $m \geq 2$. For Lebesgue-almost every $\beta \in \mathbb{R}$, the convergents p_k/q_k are equidistributed among the $\Gamma(m)$-orbits on $\mathbb{Q} \cup \{\infty\}$: each orbit appears with the same asymptotic frequency, equal to the reciprocal of the number of orbits.*

Proof. Since $\Gamma(m)$ is normal in $\mathrm{PSL}(2,\mathbb{Z})$, the finite quotient $\mathrm{PSL}(2,\mathbb{Z})/\Gamma(m) \cong \mathrm{PSL}(2,\mathbb{Z}/m\mathbb{Z})$ acts on $\Gamma(m)\backslash\mathbb{H}$ by isometries, and this action is transitive on the set of cusps (since the cusps correspond to $\Gamma(m)$-orbits of rationals, and $\mathrm{PSL}(2,\mathbb{Z})$ acts transitively on $\mathbb{Q} \cup \{\infty\}$). Cusp widths are preserved by isometries, so all r cusps have the same width w, and $rw = [\mathrm{PSL}(2,\mathbb{Z}) : \Gamma(m)]$ gives $w/[\mathrm{PSL}(2,\mathbb{Z}) : \Gamma(m)] = 1/r$. Theorem 5.7 then gives asymptotic frequency $1/r$ for each orbit. □

The continued fraction expansion produces a sequence of rational approximations whose arithmetic properties seem opaque. Passing to the hyperbolic plane, these approximations become turning corners of a geodesic crossing the Farey tessellation; projecting onto quotient surfaces translates arithmetic into cusp visits; and ergodicity distributes those visits in proportion to cusp widths. The "party trick" of Section 2.2 cleanly becomes an equidistribution theorem.

References

[AA68] V. I. Arnold and A. Avez. *Ergodic Problems of Classical Mechanics*. Mathematical Physics Monograph Series. W. A. Benjamin, New York, 1968.

[Art24] Emil Artin. Ein mechanisches System mit quasiergodischen Bahnen. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, 3(1):170–175, 1924.

[Bir31] George D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences of the United States of America*, 17(12):656–660, 1931.

[Fou] Charles Fougeron. Dimacs reu project page. http://reu.dimacs.rutgers.edu/~cjf151/dimacs/.

[HW08] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press, 6 edition, 2008.

[Kat92] Svetlana Katok. *Fuchsian Groups*. University of Chicago Press, Chicago, 1992.

[Moe82] R. Moeckel. Geodesics on modular surfaces and continued fractions. *Ergodic Theory and Dynamical Systems*, 2(1):69–83, 1982.

[Ser85] Caroline Series. The modular surface and continued fractions. *Journal of the London Mathematical Society. Second Series*, 31(1):69–80, 1985.

Massachusetts Institute of Technology
Email address: tonyyzwu@mit.edu