The geometry of continued fractions as analysed by considering Möbius transformations acting on the hyperbolic plane [1]

Richard van Rensburg

Under the supervision of Dr Meira Hockman
School of Mathematics
University of the Witwatersrand
Private Bag X3, PO WITS 2050, South Africa

October 2011

[1] Submitted to the University of the Witwatersrand, Johannesburg, in fulfillment of the requirements for the degree of Master of Science.

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

____

Richard van Rensburg

This ____________ day of October 2011, at Johannesburg, South Africa.

Abstract

Continued fractions have been extensively studied in number-theoretic ways. In this text, we will illuminate some of the geometric properties of continued fractions by considering them as compositions of Möbius transformations which act as isometries of the hyperbolic plane $\mathbb{H}^2$. In particular, we examine the geometry of *simple* continued fractions by considering the action of the extended modular group on $\mathbb{H}^2$. Using these geometric techniques, we prove very important and well-known results about the convergence of simple continued fractions. Further, we use the Farey tessellation $\mathcal{F}$ and the method of cutting sequences to illustrate the geometry of simple continued fractions as the action of the extended modular group on $\mathbb{H}^2$. We also show that $\mathcal{F}$ can be interpreted as a graph, and that the simple continued fraction expansion of any real number can be can be found by tracing a unique path on this graph. We also illustrate the relationship between Ford circles and the action of the extended modular group on $\mathbb{H}^2$. Finally, our work will culminate in the use of these geometric techniques to prove well-known results about the relationship between periodic simple continued fractions and quadratic irrationals.

Acknowledgements

I would like to thank my supervisor Dr Meira Hockman, for her unwaivering support and enthusiasm, and my parents, for allowing me the privilege of pursuing full-time postgraduate study.

Contents

Chapter 1

Hyperbolic Geometry

1.1 Historical Background

Definition 1.1 A *continued fraction* is a fraction of the form

$$b_0 + \cfrac{a_0}{b_1 + \cfrac{a_1}{b_2 + \cfrac{a_2}{b_3 + \ddots}}} \tag{1.1}$$

where the a_i and b_i, for $i = 0, 1, 2, ...$, may be real or complex numbers.

In this text, we will concern ourselves mainly with continued fractions in which $a_i = 1$ for all i and $b_i \in \mathbb{Z}^+$ for all $i \geq 1$, and $b_0 \in \mathbb{Z}$. Such continued fractions are called *simple* continued fractions and have been extensively studied by number-theorists ([1],[2],[3]). The geometric properties of continued fractions are apparent from the fact that a continued fraction is a composition of a finite or infinite sequence of Möbius maps of the form $z \mapsto b_i + \frac{a_i}{z}$ for $i = 0, 1, 2, ...$, evaluated at a suitable z. In particular, a simple continued fraction can be expressed as a composition of a sequence of Möbius maps of the form $z \mapsto b_i + \frac{1}{z}$ for $i = 0, 1, 2, \ldots$, evaluated at $z = \infty$. In 1938, L.R. Ford ([4]) discussed the geometry of continued fractions by representing each rational number as a circle in the upper half plane. In 1942, J.F. Paydon and H.S. Wall ([6]) discussed continued fractions as sequences of Möbius maps. A subsequent, more comprehensive text by Jones and Thron ([5]) also discussed the treatment of continued fractions as compositions of sequences of Möbius maps. Indeed, continued fraction theory has been revolutionized by considering continued fractions from this geometric point of view.

In recent years, mathematicians such as Alan Beardon, Caroline Series, Svetlana Katok and Ian Short have contributed to the theory of continued fractions by considering the action of particular groups of Möbius transformations on the boundaries of $\mathbb{H}^2$ and $\mathbb{H}^3$ ([7],[8],[9],[12]). The aim of this text is to explore some of the developments that have arisen from studying simple continued fractions in this way.

1.2 Theoretical Background

1.2.1 Definition of a Continued Fraction

We call (1.1) a *simple* continued fraction if $a_i = 1$ for all i, and $b_i \in \mathbb{Z}^+$ for all $i \geq 1$, and $b_0 \in \mathbb{Z}$. If $a_i = 1$ for all i and $b_0 \in \mathbb{Z}$ and $b_i \in \mathbb{R}^+$ for all $i \geq 1$ then we call (1.1) a *positive* continued fraction. If $a_i = 1$ and $b_i \in \mathbb{Z}$ for all i, then we call (1.1) an *integer* continued fraction.

Definition 1.2 The quantities $b_0, b_1, b_2, \ldots$ of the simple continued fraction given by

$$b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \ddots + \cfrac{1}{b_n}}} \tag{1.2}$$

are called the *partial quotients*.

A continued fraction may be finite or infinite, and we may denote a simple continued fraction by its sequence of partial quotients. That is, if (1.2) is a finite simple continued fraction then it is represented by $[b_0, b_1, b_2, ..., b_n]$, where $n \in \mathbb{N}$. If (1.2) is an infinite simple continued fraction, then it is represented by $[b_0, b_1, b_2, ...]$.

1.2.2 Möbius Transformations

The results stated in this subsection are well-known and can be found in the literature on Möbius transformations. We state these results for completeness as we will rely on them throughout this text. The proofs of many of the theorems are omitted, but can be found in texts such as [10], [11] and [14].

We will represent the *extended complex plane* $\mathbb{C} \cup \{\infty\}$ by $\mathbb{C}_\infty$. Similarly, we represent the *extended real line* $\mathbb{R} \cup \{\infty\}$ by $\mathbb{R}_\infty$.

Definition 1.3 A Möbius transformation (or Möbius map) is a map of the form $z \mapsto \frac{az+b}{cz+d}$ from $\mathbb{C}_\infty$ to $\mathbb{C}_\infty$ where $a, b, c, d \in \mathbb{C}$ and $ad - bc \neq 0$. We denote the group of all Möbius maps by $\mathcal{M}$.

Definition 1.4 The *General Linear Group*, $GL(2, \mathbb{C})$, is the group of 2×2 complex matrixes $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where $ad - bc \neq 0$, and I denotes the 2×2 identity matrix.

Since the mapping $z \mapsto \frac{az+b}{cz+d}$ is equivalent to the mapping $z \mapsto \frac{\lambda az+\lambda b}{\lambda cz+\lambda d}$ for $\lambda \neq 0$, we identify the matrices

$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $\begin{pmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{pmatrix}$ for $\lambda \neq 0$. Thus any matrix $A \in GL(2, \mathbb{C})$ can be multiplied by a suitable λ so that $\det(\lambda A) = 1$, and we say that A has been *normalized.*

We note that a circle C in $\mathbb{C}_\infty$ can assume one of two possible forms:

1. C is a Euclidean circle;
2. C is a Euclidean line with ∞ attached.

Hence we refer to inversions in circles and lines as *reflections.* If $S(a,r)$ is a circle with center $a \in \mathbb{C}$ and radius $r > 0$, then reflection in $S(a,r)$ is given as $\phi_{S(a,r)}(z) = a + r^2 \frac{(z-a)}{|z-a|^2}$. Reflection in the line

$$L(a,t) = \{z \in \mathbb{C} : (z \cdot a) = t\} \cup \{\infty\},$$

where $t \in \mathbb{R}$, is given by $\phi_{L(a,t)}(z) = z - 2[(z \cdot a) - t]\frac{a}{|a|^2}$.

Example : Let $S(0,1) = \{z \in \mathbb{C} : |z| = 1\}$ be the unit circle centred at 0; $\Re(z) = a$ be the line through $a \in \mathbb{R}$ orthogonal to the real axis and let $\Im(z) = 0$ be the real axis in $\mathbb{C}_\infty$. Reflection of a point z in the circle $S(0,1)$ is given by $J_1(z) = \frac{z}{|z|^2} = \frac{1}{\bar{z}}$, with $J_1(0) = \infty$ and $J_1(\infty) = 0$. Reflection in $\Im(z) = 0$ is given by $J_2(z) = z - 2i\Im(z) = \bar{z}$, with $J_2(\infty) = \infty$. Reflection in the line $\Re(z) = a$ is given by $J_a(z) = z - 2(\Re(z) - a)$ with $J_a(\infty) = \infty$. We note that the mappings $\psi(z) = \frac{1}{z}$ and $\tau(z) = z + 1$ can be expressed as $\psi = J_1 J_2 = J_2 J_1$ and $\tau = J_{a+\frac{1}{2}} J_a$ for any $a \in \mathbb{R}$.

Definition 1.5 The General Möbius Group, denoted $GM(\mathbb{R}^n_\infty)$, is the group consisting of finite compositions of reflections in spheres or planes in $\mathbb{R}^n_\infty$. The

Möbius Group is the subgroup of $GM(\mathbb{R}^n_\infty)$ consisting of compositions of an even number of reflections in lines or circles in $\mathbb{R}^n_\infty$.

Theorem 1.6 When $n = 2$ the Möbius Group is $\mathcal{M}$. Thus a Möbius transformation acting in $\mathbb{R}^n_\infty$ is a finite composition of an even number of reflections in spheres or planes. [14]

In fact, these transformations can be written as compositions of rotations, dilations, translations and the complex inversion $z \mapsto \frac{1}{z}$ in $\mathbb{C}_\infty$.

Theorem 1.7 $\mathcal{M} \cong GL(2, \mathbb{C})/\{\lambda I : \lambda \neq 0\}$

Definition 1.8 For $f \in \mathcal{M}$ with $f(z) = \frac{az+b}{cz+d}$ and $ad - bc = 1$, we define the *norm* of f as $||f|| = \sqrt{|a|^2 + |b|^2 + |c|^2 + |d|^2}$.

Definition 1.9 The *fixed points* of $f \in \mathcal{M}$ are all the values of z such that $f(z) = z$.

Theorem 1.10 If $f \in \mathcal{M} \setminus \{1_\mathcal{M}\}$ then f has one or two fixed points.

Theorem 1.11 If $\{z_1, z_2, z_3\}$ and $\{w_1, w_2, w_3\}$ are triples of distinct points in $\mathbb{C}_\infty$, then there is a unique Möbius map f such that $f(z_j) = w_j$ for $j = 1, 2, 3$.

Theorem 1.12 For any three distinct points $z_1, z_2, z_3 \in \mathbb{C}_\infty$, there is a unique circle which passes through z_1, z_2 and z_3. Further, if $f \in \mathcal{M}$ and C is a circle in $\mathbb{C}_\infty$, then $f(C)$ is also a circle in $\mathbb{C}_\infty$. In particular, any circle in $\mathbb{C}_\infty$ is the image under some $f \in \mathcal{M}$ of the real axis.

Definition 1.13 The *cross-ratio* of four distinct points $z_1, z_2, z_3, z_4 \in \mathbb{C}_\infty$ is defined as

$$(z_1, z_2; z_3, z_4) = \frac{(z_1 - z_3)(z_2 - z_4)}{(z_1 - z_2)(z_3 - z_4)}.$$

Theorem 1.14 Cross-ratios are invariant under Möbius transformations. That is, for any $f \in \mathcal{M}$ we have

$$(f(z_1), f(z_2); f(z_3), f(z_4)) = (z_1, z_2; z_3, z_4).$$

Definition 1.15 The *trace* of a Möbius map $f(z) = \frac{az+b}{cz+d}$, where $ad - bc = 1$, is the quantity $a + d$ denoted $tr(f)$. Further, $tr^2(f)$, the square of the trace of a Möbius map $f \neq 1_{\mathcal{M}}$, is used to classify f as follows:

1. f is called *parabolic* if and only if $tr^2(f) = 4$
2. f is called *elliptic* if and only if $tr^2(f) \in [0, 4)$
3. f is called *loxodromic* if and only if $tr^2(f) \notin [0, 4]$
4. f is called *strictly loxodromic* if and only if $tr^2(f) < 0$ or $tr^2(f) \notin \mathbb{R}$.

Example : Let $f(z) = e^{i\theta}z$ where $e^{i\theta} \neq 1$. This is a Euclidean rotation through θ radians. Since $e^{i\theta} \neq 1$, we must have $\theta \neq 2k\pi$, $k \in \mathbb{Z}$. Then

f is associated with the matrix $\begin{pmatrix} e^{i\theta} & 0 \\ 0 & 1 \end{pmatrix}$ which can be normalized to the matrix $\begin{pmatrix} e^{\frac{i\theta}{2}} & 0 \\ 0 & e^{-\frac{i\theta}{2}} \end{pmatrix}$. That is:

$$f(z) = \frac{e^{i\theta}z + 0}{0z + 1} = \frac{e^{\frac{i\theta}{2}}z + 0}{0z + e^{\frac{-i\theta}{2}}}.$$

We thus have

$$tr^2(f) = (e^{\frac{i\theta}{2}} + e^{\frac{-i\theta}{2}})^2 = \left(2\cos\left(\frac{\theta}{2}\right)\right)^2 = 4\cos^2\left(\frac{\theta}{2}\right).$$

Since $0 \leq \cos^2\left(\frac{\theta}{2}\right) < 1$ with $\theta \neq 2k\pi$ for all $k \in \mathbb{Z}$, we have $0 \leq 4\cos^2(\frac{\theta}{2}) < 4$. That is, $0 \leq tr^2(f) < 4$ and so f is elliptic.

Example : Let $f(z) = \lambda z$ where $\lambda \in \mathbb{C}$ and $|\lambda| \neq 0, 1$. Here we have f associated with the matrix $\begin{pmatrix} \lambda & 0 \\ 0 & 1 \end{pmatrix}$ which can be normalized to the matrix $\pm\begin{pmatrix} \sqrt{\lambda} & 0 \\ 0 & \frac{1}{\sqrt{\lambda}} \end{pmatrix}$. Then $tr^2(f) = (\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}})^2 = \lambda + \frac{1}{\lambda} + 2$. If $\lambda \in \mathbb{R}^+$ then $\sqrt{\lambda} \in \mathbb{R}$ and $\lambda + \frac{1}{\lambda} > 2$ and hence $tr^2(f) > 4$ and so f is loxodromic. This is easily seen since if $\lambda + \frac{1}{\lambda} \leq 2$ then $\lambda + \frac{1}{\lambda} - 2 \leq 0$ and so $\lambda^2 - 2\lambda + 1 = (\lambda - 1)^2 \leq 0$. This is a contradiction if $\lambda \in \mathbb{R}^+$ and $\lambda \neq 1$. If $\lambda \notin \mathbb{R}^+$ then $tr^2(f) \notin \mathbb{R}$ and so f is strictly loxodromic.

Definition 1.16 Two Möbius maps g and h are *conjugate* in $\mathcal{M}$ if there exists a map $f \in \mathcal{M}$ such that $h = fgf^{-1}$.

Theorem 1.17 The trace of a Möbius map is invariant under conjugation.

The conjugacy classes of a Möbius map can also be used to classify Möbius maps into the same categories as in Definition 1.15, and we thus have the following theorem ([12]).

Theorem 1.18 Let $g \in \mathcal{M}$ and $g \neq 1_{\mathcal{M}}$.

The following are equivalent:

1. (a) g is parabolic;

(b) g is conjugate in $\mathcal{M}$ to the translation $z \mapsto z+1$;

(c) g has exactly one fixed fixed point ζ and $g^n \to \zeta$ pointwise on $\mathbb{C}_\infty$.

The following are equivalent:

2. (a) g is elliptic;

   (b) g is conjugate in $\mathcal{M}$ to a Euclidean rotation $z \mapsto e^{i\theta}z$, where $e^{i\theta} \neq 1$;

   (c) g has two fixed points, and $g^n(z)$ converges if and only if z is a fixed point of g.

The following are equivalent:

3. (a) g is loxodromic or strictly loxodromic;

   (b) g is conjugate in $\mathcal{M}$ to the map $z \mapsto \lambda z$, where $|\lambda| \neq 0, 1$;

   (c) g has two fixed points u and v which can be chosen so that if $z \neq v$ then $g^n(z) \to u$ as $n \to \infty$.

Proof: Definition 1.15 implies that if $g \neq 1_{\mathcal{M}}$ then exactly one of 1(a), 2(a) and 3(a) is true. Since the trace of a Möbius map is invariant under conjugation (Theorem 1.17), we have that 1(b) implies 1(a), 2(b) implies 2(a) and 3(b) implies 3(a).

We know that from Theorem 1.10 that g has either one or two fixed points, so if g has exactly one fixed point, then g is conjugate to a map which has ∞ as its only fixed point. That is, g is conjugate to a translation. If g has two fixed points then g is conjugate to a map which fixes 0 and ∞. That is, g is conjugate to a map of the form $z \mapsto \lambda z$, $\lambda \neq 0, 1$. Hence at least one of 1(b), 2(b) and 3(b) is true, and so 1(a) is equivalent to 1(b), or 2(a) is equivalent to 2(b) or 3(a) is equivalent to 3(b).

Suppose that 1(b) holds. Then for some $h \in \mathcal{M}$ we have $g = hfh^{-1}$, where $f(z) = z+1$. We have that ∞ is the only fixed point of f. Further, for all z we have $f^n(z) = z+n \to \infty$ as $n \to \infty$, so that $g^n(z) = hf^n(h^{-1}(z)) \to h(\infty)$ as $n \to \infty$. Thus g has exactly one fixed point, namely $h(\infty) = \zeta$, and $g^n \to \zeta$ pointwise on $\mathbb{C}_\infty$. This shows that 1(b) implies 1(c).

Now suppose that 2(b) holds. Then for some $h \in \mathcal{M}$ we have $g = hfh^{-1}$, where $f(z) = e^{i\theta}z$. Note that $f(0) = 0$ and $f(\infty) = \infty$, so g has two fixed points. Since $e^{i\theta} \neq 1$, we have $f^n(z) = e^{in\theta}z$ and $f^n(z)$ has no limit for $z \neq 0$ and $z \neq \infty$. Thus g^n converges only at its limit points. This shows that 2(b) implies 2(c).

Now suppose that 3(b) holds. Then for some $h \in \mathcal{M}$ we have $g = hfh^{-1}$, where $f(z) = \lambda z$, $|\lambda| \neq 0, 1$. Then $f(z)$ fixes 0 and ∞, and we have $f^n(z) = \lambda^n z$. If $|\lambda| < 1$, then $\lambda^n z \to 0$ as $n \to \infty$ for all $z \neq \infty$. If $|\lambda| > 1$ then $\lambda^n z \to \infty$ as $n \to \infty$ for all $z \neq 0$. Thus g has two fixed points u and v which can be chosen so that if $z \neq v$ then $g^n(z) \to u$ as $n \to \infty$. Hence 3(b) implies 3(c).

We know that at most one of 1(a), 2(a) and 3(a) is true. Since we have that 1(b) implies 1(a), 2(b) implies 2(a), and 3(b) implies 3(a), we have that at most one of 1(b), 2(b) and 3(b) can hold. Now we also have that 1(b) implies 1(c), 2(b) implies 2(c), and 3(b) implies 3(c), so at least one of 1(c), 2(c) and 3(c) must hold. But at most one of 1(c), 2(c) and 3(c) can hold. Therefore exactly one of 1(c), 2(c) and 3(c) holds. Therefore 1(b) is equivalent to 1(c) and hence to 1(a); 2(b) is equivalent to 2(c) and hence to 2(a), and 3(b) is equivalent to 3(c) and hence to 3(a). ■

Definition 1.19 With the labeling that we have given in the above theorem, we call u the *attracting fixed point* and v the *repelling fixed point* of the loxodromic map g. That is, $g^n(z) \to u$ as $n \to \infty$ for all $z \neq v$, while $g^n(v) = v$.

1.2.3 The Action of Möbius Maps on $\mathbb{H}^2$

In this subsection, we introduce the upper half plane $\mathbb{H}^2$ and the group of Möbius maps which preserves $\mathbb{H}^2$. We introduce the hyperbolic plane metric which, together with $\mathbb{H}^2$, become a model of the hyperbolic plane.

Definition 1.20 The *upper-half plane* $\mathbb{H}^2$ is defined as

$$\mathbb{H}^2 = \{z \in \mathbb{C} : \Im(z) > 0\}.$$

We are interested in Möbius maps that leave $\mathbb{H}^2$ invariant. Some of the results in this subsection are stated without proof. These proofs can be found in texts that include an introduction to hyperbolic geometry, such as [9], [10], [13] and [14].

Definition 1.21 $PSL(2, \mathbb{R}) = \{z \mapsto \frac{az+b}{cz+d} : a, d, c, d \in \mathbb{R}, ad - bc = 1\}$.

Definition 1.22 We define *hyperbolic length* in $\mathbb{H}^2$ by the formula

$$ds^2 = \frac{dx^2 + dy^2}{y^2} = \frac{|dz|^2}{y^2},$$

where $z = x + iy$. In particular, if $f : [a, b] \to \mathbb{H}^2$ is a piecewise differentiable path with $f(t) = x(t) + iy(t)$ then its hyperbolic length $\ell(f)$ is given by

$$\ell(f) = \int_a^b \frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt}{y} = \int_a^b \frac{\left|\frac{dz}{dt}\right| dt}{y} = \int_a^b \frac{|dz|}{y} = \int_a^b \frac{|dz|}{\Im(z)}.$$

Theorem 1.23 Hyperbolic length is invariant under the elements of $PSL(2,\mathbb{R})$.

Example : Suppose $0 < a < b$, where $a, b \in \mathbb{R}$, and consider the piecewise C^1 path $f : [a,b] \longrightarrow \mathbb{H}^2$ given by $f(t) = it$. Then $f([a,b])$ is the segment of $\mathbb{I}$, the positive imaginary axis, between ia and ib. Further, we have $\Im(f(t)) = t$ and $|f'(t)| = 1$, so

$$\ell(f) = \int_f \frac{1}{\Im(z)}|dz| = \int_a^b \frac{1}{t}dt = [\ln(t)]_a^b = \ln(b) - \ln(a) = \ln\left(\frac{b}{a}\right).$$

Hyperbolic lines (also called $\mathbb{H}$-lines) are the paths of shortest hyperbolic length between two points in $\mathbb{H}^2$. That is, $\mathbb{H}$-lines are hyperbolic *geodesics*. The next theorem establishes that there is a unique path of shortest hyperbolic length between any two points in $\mathbb{H}^2$. This uniqueness leads to an understanding of the nature of $\mathbb{H}$-lines and $\mathbb{H}$-line segments.

Theorem 1.24 There is a unique $\mathbb{H}$-line segment (geodesic) joining two distinct points in $\mathbb{H}^2$. The $\mathbb{H}$-line segments are arcs of circles with centre on the real axis, or segments of Euclidean lines perpendicular to the real axis.

Definition 1.25 Let G be a group and X a set. G *acts* on X if

1. $g(x) \in X$ for $g \in G$, $x \in X$
2. $g_1 g_2(x) = g_1(g_2(x))$ for $g_1, g_2 \in G$
3. $1_G(x) = x$, where 1_G is the identity element in G.

G acts *transitively* on X if for all $x_1, x_2 \in X$, there exists a $g \in G$ such that $x_2 = g(x_1)$. G acts *doubly transitively* on X if, whenever (x_1, x_2) and (y_1, y_2) are pairs of distinct elements of X, there exists some $g \in G$ such that $g(x_i) = y_i$ for $i = 1, 2$.

As in Euclidean geometry, there is a unique $\mathbb{H}$-line between any two distinct points in $\mathbb{H}^2$. Two $\mathbb{H}$-lines are *parallel* in $\mathbb{H}^2$ if they are disjoint in $\mathbb{H}^2$. The positive imaginary axis $\mathbb{I}$ is an $\mathbb{H}$-line and plays a pivotal role in the development of the ideas that follow.

Theorem 1.26

1. $PSL(2,\mathbb{R})$ acts transitively on $\mathbb{H}^2$.
2. $PSL(2,\mathbb{R})$ acts doubly transitively on $\mathbb{R}_\infty$.
3. $PSL(2,\mathbb{R})$ acts transitively on the set of all $\mathbb{H}$-lines.

Definition 1.27 The *hyperbolic distance* $\rho(z_1, z_2)$ between two points $z_1, z_2 \in \mathbb{H}^2$ is the hyperbolic length of the unique $\mathbb{H}$-line segment joining z_1 to z_2. That is, $\rho(z_1, z_2) = \inf\{\ell(f)\}$, where the infimum is taken over all piecewise C^1 paths $f : [a, b] \to \mathbb{H}^2$ such that $f(a) = z_1$ and $f(b) = z_2$.

Theorem 1.28 $\mathbb{H}^2$, together with the metric ρ, is a metric space.

The elements of $PSL(2,\mathbb{R})$ are *isometries* of $\mathbb{H}^2$. However, there are isometries of $\mathbb{H}^2$, such as $z \mapsto -\bar{z}$, that are not in $PSL(2,\mathbb{R})$. In particular, we have the following result:

Theorem 1.29 Isom$\mathbb{H}^2$, the set of all isometries on $\mathbb{H}^2$, is generated by Möbius transformations from $PSL(2,\mathbb{R})$ together with the map $z \mapsto -\bar{z}$. The group $PSL(2,\mathbb{R})$ is a normal subgroup of index 2 of Isom$\mathbb{H}^2$.

We have seen that if ia and ib, with $b > a$, are two points on $\mathbb{I}$ then we have that $\rho(ia, ib) = \ln(\frac{b}{a})$. Using this fact together with Theorem 1.23, we have the following result:

Lemma 1.30 Let $z, w \in \mathbb{H}^2$ with $z \neq w$, and let the $\mathbb{H}$-line joining z and w have endpoints $z^*, w^* \in \mathbb{R}_\infty$, chosen in such a way that z lies between z^* and w. Then there exists a unique $g \in PSL(2, \mathbb{R})$ such that $g(z^*) = 0$, $g(w^*) = \infty$ and $g(z) = i$. We also have $g(w) = ri$, where $r > 1$, and $\rho(z, w) = \ln(r)$.

Theorem 1.31 Let $z, w \in \mathbb{H}^2$, then we have the following:

1. $\tanh \frac{1}{2}\rho(z, w) = \left|\frac{z-w}{z-\bar{w}}\right|$
2. $\sinh \frac{1}{2}\rho(z, w) = \frac{|z-w|}{2\sqrt{\Im(z)\Im(w)}}$

Theorem 1.32 If $f \in PSL(2, \mathbb{R})$ with $f(z) = \frac{az+b}{cz+d}$ then
$||f||^2 = 2\cosh\rho(i, f(i)) = 2 + 4\sinh^2\rho(i, f(i))$
where $||f||^2 = |a|^2 + |b|^2 + |c|^2 + |d|^2$.

Proof: First observe that $f(i) = \frac{ai+b}{ci+d} = \frac{(ac+bd)+i}{|c|^2+|d|^2}$.
Also observe that

$$\begin{aligned} (ac+bd)^2 + 1 &= (ac+bd)^2 + (ad-bc)^2 \\ &= |a|^2|c|^2 + |b|^2|d|^2 + |a|^2|d|^2 + |b|^2|c|^2 \\ &= (|a|^2 + |b|^2)(|c|^2 + |d|^2) \end{aligned}$$

From Theorem 1.31, we have that
$\sinh \frac{1}{2}\rho(z, w) = \frac{|z-w|}{2(\Im(z)\Im(w))^{\frac{1}{2}}}$, which gives us
$\cosh^2 \frac{1}{2}\rho(z, w) = 1 + \sinh^2 \frac{1}{2}\rho(z, w) = 1 + \frac{|z-w|^2}{4\Im(z)\Im(w)}$.
From the hyperbolic trigonometric identities, we have for all A that
$1 + \cosh 2A = 2\cosh^2 A$ and $\cosh^2 A - \sinh^2 A = 1$.
Thus $1 + \cosh 2(\frac{1}{2}\rho(z, w)) = 2\cosh^2 \frac{1}{2}\rho(z, w)$ implies
$1 + \cosh\rho(z, w) = 2\left(1 + \frac{|z-w|^2}{4\Im(z)\Im(w)}\right)$ which implies

$\cosh\rho(z,w) = 1 + \frac{|z-w|^2}{2\Im(z)\Im(w)}$ which implies

$2\cosh\rho(z,w) = 2 + 4\sinh^2\frac{1}{2}\rho(z,w)$.

So, setting $z = i$ and $w = f(i)$, we obtain

$2\cosh\rho(i,f(i)) = 2 + 4\sinh^2\frac{1}{2}\rho(i,f(i))$. This gives us

$$\begin{aligned}
\cosh\rho(i,f(i)) &= 1 + 2\sinh^2\frac{1}{2}\rho(i,f(i)) \\
&= 1 + \frac{|i-f(i)|^2}{2\Im(i)\Im(f(i))} \\
&= 1 + \frac{\left|i - \frac{ai+b}{ci+d}\right|}{2\left(\frac{1}{|c|^2+|d|^2}\right)} \\
&= 1 + \frac{\left|i - \frac{(ac+bd)+i}{|c|^2+|d|^2}\right|}{\frac{2}{|c|^2+|d|^2}} \\
&= 1 + \frac{1}{2}\left(\frac{(-1+|c|^2+|d|^2)^2}{|c|^2+|d|^2} + \frac{(ac+bd)^2}{|c|^2+|d|^2}\right) \\
&= 1 + \frac{1}{2}\left(\frac{(-1+|c|^2+|d|^2)^2}{|c|^2+|d|^2} + \frac{(|a|^2+|b|^2)(|c|^2+|d|^2)-1}{|c|^2+|d|^2}\right) \\
&= \frac{1}{2}(|a|^2+|b|^2+|c|^2+|d|^2).
\end{aligned}$$

Hence $2\cosh\rho(i,f(i)) = |a|^2+|b|^2+|c|^2+|d|^2$.

Hence $2\cosh\rho(i,f(i)) = 2 + 4\sinh^2\rho(i,f(i)) = |a|^2+|b|^2+|c|^2+|d|^2$ as required. ■

1.3 The Modular Group and the Extended Modular Group

We now introduce a subgroup of $PSL(2,\mathbb{R})$, namely the Modular group $PSL(2,\mathbb{Z})$, denoted Γ. We also introduce the Extended Modular group, denoted $\tilde{\Gamma}$. The Modular and Extended Modular groups are useful in the study of *simple* continued fractions, as any simple continued fraction can be expressed as a composition of Extended Modular maps.

Definition 1.33 $\Gamma = \{z \mapsto \frac{az+b}{cz+d} : a, b, c, d \in \mathbb{Z}, ad - bc = 1\} = PSL(2, \mathbb{Z})$

Definition 1.34 $\tilde{\Gamma} = \{z \mapsto \frac{az+b}{cz+d} : a, b, c, d \in \mathbb{Z}, |ad - bc| = 1\}$

Elements of Γ are called *Modular* maps (or Modular transformations) while elements of $\tilde{\Gamma}$ are called *Extended Modular* maps (or Extended Modular transformations).

Definition 1.35 $SL(2, \mathbb{Z})$, the *Special Linear Group*, is the group of non singular 2×2 integer matrices with determinant 1.

Theorem 1.36 Γ is a group under composition of maps where each element $z \mapsto \frac{az+b}{cz+d}$ in Γ can be associated with a non-singular matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ in $SL(2, \mathbb{Z})$. ([9],[12])

Theorem 1.37 Γ is generated by the mappings τ and φ where $\tau(z) = z + 1$ and $\varphi(z) = -\frac{1}{z}$. We write $\Gamma = \langle \tau, \varphi \rangle$. ([9],[12])

Theorem 1.38 $\tilde{\Gamma}$ is a group under the composition of maps and $\tilde{\Gamma} = \langle \tau, \psi \rangle$ where $\tau(z) = z + 1$ and $\psi(z) = \frac{1}{z}$. Γ is a normal subgroup of $\tilde{\Gamma}$. In particular $\varphi = \tau\psi\tau^{-1}\psi\tau$ and $\tilde{\Gamma} = \Gamma \cup \Gamma\psi$. ([12])

Consider the maps of the form $s_{b_i}(z) = b_i + \frac{1}{z} = \tau^{b_i}\psi(z)$ for $b_i \in \mathbb{Z}$. We note that while $\tau \in \Gamma \cap \tilde{\Gamma}$ the mapping $\psi \in \tilde{\Gamma} \setminus \Gamma$. When studying simple continued fractions, it is thus more natural to consider the action of the *Extended* Modular group on $\mathbb{H}^2$.

We denote the composition of a sequence of n such maps by

$$S_{[b_n]}(z) = s_{b_0}s_{b_1}s_{b_2}\cdots s_{b_n}(z) = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \ddots + \cfrac{1}{b_n + \cfrac{1}{z}}}} \tag{1.3}$$

Note that the sequence of exponents of the τ's gives the sequence of partial quotients of the simple continued fraction of ω where $\omega = S_{[b_n]}(\infty)$ and

$$S_{[b_n]}(\infty) = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \ddots + \cfrac{1}{b_n}}} \tag{1.4}$$

We note that the map $\psi \in \tilde{\Gamma}$ does not preserve $\mathbb{H}^2$. This is a complicating feature in the development of the geometry of simple continued fractions in $\mathbb{H}^2$, but we can find a novel way of changing the geometry in order to avoid this problem.

1.3.1 The Vertical Plane in Hyperbolic Space

The vertical plane $\mathbb{H}^\perp$ is the intersection of hyperbolic 3-space $\mathbb{H}^3$ with the vertical Euclidean plane through $\mathbb{R}_\infty$. Let $\mathbf{j} = (0,0,1) \in \mathbb{H}^\perp$. This plane is an isometric image of $\mathbb{H}^2$ where we map $x+iy$ in $\mathbb{H}^2$ to $x+\mathbf{j}y$ in $\mathbb{H}^\perp$. We will see that $\mathbb{H}^\perp$ is left invariant by an extension of ψ and is thus a more appropriate region in which to study simple continued fractions.

Definition 1.39 Hyperbolic space, denoted $\mathbb{H}^3$, is defined as

$$\mathbb{H}^3 = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 > 0\}$$

together with the metric ρ' of hyperbolic space given by $\frac{|dx|}{x_3}$ where $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. ([12])

Definition 1.40 The vertical hyperbolic plane $\mathbb{H}^\perp$ in $\mathbb{H}^3$ is given by

$$\mathbb{H}^\perp = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 = 0, x_3 > 0\}.$$

A *quarternion* is a quantity of the form $w = x_1 + x_2\mathbf{i} + x_3\mathbf{j} + x_4\mathbf{k}$ where $x_1, x_2, x_3, x_4 \in \mathbb{R}$ and we set $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$. We denote the set of all quarternions by Q. If $x_3 = x_4 = 0$ then $w = x_1 + ix_2 \in \mathbb{C}$, so $\mathbb{C}$ can be considered as a subset of Q. By considering the complex number as quarternions, Theorem 1.32 can be generalized to $\mathbb{H}^3$ as follows ([12]):

Theorem 1.41 Suppose $g(z) = \frac{az+b}{cz+d}$ where $a, b, c, d \in \mathbb{C}$, $ad - bc = 1$ and $z \in \mathbb{H}^3$. Then $||g||^2 = 2\cosh \rho'(\mathbf{j}, g(\mathbf{j}))$ where $||g|| = |a|^2 + |b|^2 + |c|^2 + |d|^2$.

Theorem 1.42 $\mathbb{C}_\infty$ is the boundary of $\mathbb{H}^3$, and the action of a Möbius map on $\mathbb{C}_\infty$ is in fact the action of a conformal isometry of $\mathbb{H}^3$ on the boundary of $\mathbb{H}^3$. [11]

1.3.2 The Poincaré Extension

Henri Poincaré (1854 - 1912) observed that each Möbius map g acting on $\mathbb{C}_\infty$ has a natural extension to a Möbius map $\tilde{g}$ acting on $\mathbb{R}^3_\infty$. We have already noted that a Möbius map can be expressed as the composition of finitely many reflections in circles or generalized circles. Poincaré showed that for each reflection in $\mathbb{R}^2$ we define a reflection in $\mathbb{R}^3$ that leaves the plane $x_3 = 0$, as well as each of the upper and lower half spaces, invariant. In this way each Möbius map acting on $\mathbb{C}_\infty$ can be regarded as a composition of a finite number of reflections in $\mathbb{R}^3$. This extension depends on the embedding $x \mapsto \tilde{x}$ of $\mathbb{R}^2$ into $\mathbb{R}^3$, where $x = (x_1, x_2)$ and $\tilde{x} = (x_1, x_2, 0)$. [14]

Definition 1.43 For each reflection g acting in $\mathbb{C}_\infty$, we define a reflection $\tilde{g}$ acting in $\mathbb{R}^3_\infty$ as follows: if g is a reflection in the circle $S(a, r)$ of radius r centred at $a = (a_1, a_2)$, then $\tilde{g}$ is a reflection in the sphere $S(\tilde{a}, r)$ in $\mathbb{R}^3$, where $\tilde{a} = (a_1, a_2, 0)$. If g is a reflection in a line

$$L(a, t) = \{x \in \mathbb{R}^2 : (x \cdot a) = t\} \cup \{\infty\}$$

then $\tilde{g}$ is a reflection in the plane

$$P(\tilde{a}, t) = \{x \in \mathbb{R}^3 : (x \cdot \tilde{a}) = t\} \cup \{\infty\},$$

where $t \in \mathbb{R}$. The extensions $\tilde{g}$ are called the *Poincaré Extensions.*

It is seen that if $x = (x_1, x_2) \in \mathbb{R}^2$ and $(y_1, y_2) = y = g(x)$ for some reflection g in $\mathbb{C}_\infty$, then the Poincaré extension $\tilde{g}$ of g satisfies

$$\tilde{g}(x_1, x_2, 0) = (y_1, y_2, 0) = \widetilde{g(x)}.$$

These extensions leave the complex plane $x_3 = 0$ and each of the half planes $x_3 > 0$ and $x_3 < 0$ invariant. This invariance proves that a Poincaré extension of a composition of reflections exists and is unique. [14]

Further we observe that if g_i and f_j are reflections in $\mathbb{C}_\infty$, then $\widetilde{g_i f_j} = \widetilde{g_i}\widetilde{f_j}$. Thus if g and f are in $\mathcal{M}$ with $g = g_1 g_2 \cdots g_n$ and $f = f_1 f_2 \cdots f_m$ where the g_i and f_j are reflections in $\mathbb{C}_\infty$ for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m$, then

$$\widetilde{gf} = (\tilde{g}_1 \tilde{g}_2 \cdots \tilde{g}_n \tilde{f}_1 \tilde{f}_2 \cdots \tilde{f}_m) = \widetilde{g}\widetilde{f}$$

Theorem 1.44 If $\tilde{g}$ is reflection in the sphere $S(\tilde{a}, r)$ with radius r and centre $\tilde{a}$, where $a \in \mathbb{R}^2$, then

$$\frac{|\tilde{g}(y) - \tilde{g}(x)|}{|y-x|} = r^2 \Big(\frac{1}{|y-a|^2} - \frac{2(x-a)\cdot(y-a)}{|x-a|^2|y-a|^2} + \frac{1}{|x-a|^2} \Big)^{\frac{1}{2}} = \frac{r^2}{|x-\tilde{a}||y-\tilde{a}|}.$$

[14]

Example : We have noted that the complex inversion $\psi(z) = \frac{1}{z}$ is the composition of the reflections $J_1(z) = \frac{z}{|z|^2} = \frac{1}{\bar{z}}$ and $J_2(z) = z - 2i\Im(z) = \bar{z}$. Thus $\tilde{\psi} = \widetilde{J_1 J_2} = \widetilde{J_1}\widetilde{J_2}$. Hence for $w = (x_1, x_2, x_3)$ we have

$$\begin{aligned} \widetilde{\psi}(w) &= \widetilde{\psi}(x_1, x_2, x_3) \\ &= \widetilde{J_1}\widetilde{J_2}(x_1, x_2, x_3) \\ &= \widetilde{J_1}(x_1, -x_2, x_3) \\ &= \left(\frac{x_1}{x_1^2 + x_2^2 + x_3^2}, \frac{-x_2}{x_1^2 + x_2^2 + x_3^2}, \frac{x_3}{x_1^2 + x_2^2 + x_3^2} \right). \end{aligned}$$

In particular, when $x_2 = 0$ we obtain

$$\tilde{\psi}(w) = \tilde{\psi}(x_1, 0, x_3) = \left(\frac{x_1}{x_1^2 + x_3^2}, 0, \frac{x_3}{x_1^2 + x_3^2} \right).$$

This is precisely the reflection of the point $(x_1, 0, x_3)$ in the unit hemisphere centred at the origin. Thus the Poincaré extension $\tilde{\psi}$ of the mapping ψ, when restricted to $\mathbb{H}^\perp$, is equivalent to inversion in the unit semi-circle centred at the origin, and therefore leaves the plane $\mathbb{H}^\perp$ invariant.

It can be similarly established that the Poincaré extension of the mapping $\tau = J_{a+\frac{1}{2}} J_a$ for any $a \in \mathbb{R}$ is the mapping $\tilde{\tau}(x_1, x_2, x_3) = (x_1 + 1, x_2, x_3)$ and acts like the translation $(x_1, 0, x_3) \mapsto (x_1 + 1, 0, x_3)$ and leaves $\mathbb{H}^\perp$ invariant.

Thus we have that $\tilde{\psi}$ and $\tilde{\tau}$ preserve $\mathbb{H}^\perp$, which is an identical copy of $\mathbb{H}^2$ in $\mathbb{H}^3$. In what follows we will identify $\tilde{\tau}$ with τ; $\tilde{\psi}$ with ψ; $\mathbb{H}^\perp$ with $\mathbb{H}^2$, and the boundary of $\mathbb{H}^{\perp}$ in $\mathbb{H}^3$ with $\mathbb{R}_{\infty}$. Further, we identify $\mathbb{I}$ with the vertical axis in $\mathbb{H}^{\perp}$. The tessellation described in Chapter 3, as well as the simple continued fraction expansions of all the real numbers, will be considered in this way. That is, as existing in this identical copy of the hyperbolic plane.

Since $\tilde{\Gamma}$ is generated by τ and ψ, we can now regard $\widetilde{\Gamma}$ as a group of hyperbolic isometries acting on hyperbolic space $\mathbb{H}^3$, with the extended real line and the plane $\mathbb{H}^{\perp}$ being left invariant by the Poincaré extensions of elements in $\widetilde{\Gamma}$.

1.3.3 A tessellation of $\mathbb{H}^2$

Definition 1.45 Let G be a group of homeomorphisms of a topological space X. We say that two points x and y in X are *equivalent* if $y = g(x)$ for some g in G. An open subset D of X is a *fundamental domain* for G if every point of X is equivalent to at most one point in D, and to at least one point in $\bar{D}$, the closure of D in X. ([12])

The following lemma provides a test to determine whether a set is a fundamental domain for the action of a group G on X.

Lemma 1.46 Let G be a group of homeomorphisms of a topological space X onto itself, and let D be an open subset of X. Then D is a fundamental domain for G if the following two conditions are satisfied [12]:

1. $g \in G$ and $g \neq 1_G$ implies that $g(D) \cap D = \Phi$
2. $$X = \bigcup_{g \in G} g(\bar{D})$$

Proof: Assume that D satisfies (1) and (2). Suppose that x is equivalent to the points y_1 and y_2 in D. Then y_1 and y_2 are equivalent, so there is some g in G such that $y_2 = g(y_1)$. Now (1) implies that $g = 1_G$, and so $y_1 = y_2$, and this shows that every point of X is equivalent to at most one point in D. It follows from property (2) that every point of X is equivalent to at least one point in $\bar{D}$. ■

Definition 1.47 If an open subset D of X is a fundamental domain for G, then we say that the collection of sets $\{g(\bar{D}) : g \in G\}$ is a *tessellation* of X.

Theorem 1.48 A fundamental domain for Γ is given by the set

$$\mathcal{D} = \left\{ z \in \mathbb{H}^2 : |z| > 1, |\Re(z)| < \frac{1}{2} \right\}.$$

([9],[10],[15],[16],[17])

Proof: We will show that $\mathcal{D}$ satisfies the two conditions of Lemma 1.46.

1. Suppose $g(\mathcal{D}) \cap \mathcal{D} \neq \Phi$, where $g \in \Gamma$ with $g(z) = \frac{az+b}{cz+d}$. That is, suppose that there is a $w \in \mathcal{D}$ such that $g(w) \in \mathcal{D}$.

   Without loss of generality, we may assume that $\Im(w) \leq \Im(g(w))$. If necessary, we can consider g^{-1} where $g^{-1}(g(w)) = w \in \mathcal{D} \cap g^{-1}(\mathcal{D})$.

   Let $w = u + iv$.

$$\begin{aligned} \Im(g(w)) &= \Im\left(\frac{(aw+b)(\bar{c}\bar{w}+\bar{d})}{(cw+d)(\bar{c}\bar{w}+\bar{d})}\right) \\ &= \Im\left(\frac{ac|w|^2 + bd + u + iv}{|cw+d|^2}\right) \\ &= \frac{v}{|cw+d|^2} \end{aligned}$$

$$= \frac{\Im(w)}{|cw+d|^2}.$$

Hence $\Im(w) \leq \Im(g(w))$ implies that $|cw+d| \leq 1$.

Since $w \in \mathcal{D}$ we have $|w| > 1$ and $|u| < \frac{1}{2}$. Assume $c \neq 0$. Then

$$\begin{aligned} |cw+d|^2 &= |cu+d+icv|^2 \\ &= (cu+d)^2+(cv)^2 \\ &= c^2(u^2+v^2)+d^2+2cdu \\ &= c^2|w|^2+d^2+2cdu \\ &> c^2+d^2+2cdu \\ &\geq c^2+d^2-|cd| \\ &= (|c|-|d|)^2+|cd| \\ &\geq 1. \end{aligned}$$

But $|cw+d|^2 \leq 1$. Hence there is a contradiction. But if $d=0$ and $c \neq 0$ then $|cw|^2 = c^2|w|^2 > c^2 \geq 1$. This is also a contradiction. Thus we must have $c=0$ and $d \neq 0$. Thus $g(z) = z+m$ for some $m \in \mathbb{Z}$.

Thus if $w \in \mathcal{D}$ and $g(w) \in \mathcal{D}$, then $|w-g(w)| < 1$.

But $g(w)-w = m \geq 1$ unless $m=0$. Therefore $m=0$ and $g = 1_{\mathcal{M}}$.

2. Let $z_0 \in \mathbb{H}^2$ with $z_0 = x_0 + y_0$. We must show that z_0 is Γ-equivalent to some point in $\bar{\mathcal{D}}$. Consider $K > 0$ and any $c' \in \mathbb{Z}$. Consider $\frac{K}{|c'|}$. Then the circle with centre z_0 and radius $\frac{K}{|c'|}$ can contain only a finite number of integers and hence at most a finite number of rationals with denominator c'. Thus there are only a finite number of integers c' and d' such that $\left|z_0 - \left(-\frac{d'}{c'}\right)\right| < \frac{K}{c'}$, or $|c'z_0+d'| < K$. Thus the set of numbers $|c'z_0+d'|$, taken over all coprime pairs (c',d'), attains a positive minimum. Let this minimum be attained when $c=c'$ and $d=d'$. Then

we can find integers a and b such that $ad - bc = 1$. Let $g(z) = \frac{az+b}{cz+d}$. Then $g \in \Gamma$, and since $\Im(g(w)) = \frac{\Im(w)}{|cw+d|^2}$, we have that $g(z_0)$ has the largest imaginary part among all Γ-images of z_0. By composing g with a suitable translation $\tau^m(z) = z + m$, we may assume that $|\Re(g(z_0))| \leq \frac{1}{2}$. Now let $g(z_0) = x_1 + iy_1$. Since $\Im(\varphi g(z_0)) \leq y_1$ we have that $\frac{y_1}{|g(z_0)|^2} \leq y_1$ so that $g(z_0) \geq 1$. Finally, since $|\Re(g(z_0))| \leq \frac{1}{2}$ and $|g(z_0)| \geq 1$, we have that $g(z_0) \in \bar{\mathcal{D}}$.

■

Chapter 2

Continued Fractions and Möbius Maps

2.1 Convergents, Convergence and Tails of Simple Continued Fractions

In this chapter, we give a formal and detailed account of a simple continued fraction as a composition of a sequence of Möbius maps. This will enable us to explore many well-known properties of continued fractions in the context of the action of Möbius maps on $\mathbb{H}^2$.

Since Möbius maps are compositions of inversions in hyperspheres, they can be defined in all dimensions, so results about the convergence of compositions of Möbius maps are likely to be true in all dimensions. We will investigate simple continued fractions from a geometric point of view, as opposed to the traditional number-theoretic point of view.

We are concerned with the convergence of simple continued fractions, which are continued fractions of the form given by $b_0 + K(1|b_n)$, where

$$K(1|b_n) = \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \ddots + \cfrac{1}{b_n + \ddots}}}} \tag{2.1}$$

where $b_0 \in \mathbb{Z}$ and the $b_i \in \mathbb{Z}^+$ for all $i \geq 1$. This simple continued fraction can be described in terms of Möbius transformations of the form $t_{b_i}(z) = \frac{1}{b_i+z}$ for $i \geq 1$. Compositions of these maps in the form $T_{[b_n]}(z) = t_{b_1}t_{b_2}\cdots t_{b_n}(z)$ will provide a way of describing the convergence of simple continued fractions in terms of Möbius maps. Analogously we may describe the simple continued fractions in terms of Möbius maps of the form $s_{b_i}(z) = b_i + \frac{1}{z}$ for $i \geq 0$. In this case we may compose the maps to form the composition of $n+1$ such maps as $S_{[b_n]}(z) = s_{b_0}s_{b_1}s_{b_2}\cdots s_{b_n}(z)$.

We note that for $b \in \mathbb{Z}$ we have $t_b(0) = \frac{1}{b}$ while $t_b(\infty) = 0$. Thus we have that $T_{[b_n]}(0) = T_{[b_n]}t_b(\infty) = T_{[b_{n+1}]}(\infty)$ for any $b \in \mathbb{Z}^+$ where $T_{[b_{n+1}]} = t_{b_1}t_{b_2}\cdots t_{b_n}t_b$ and $b_i, b \in \mathbb{Z}^+$. Similarly, $s_b(0) = \infty$ while $s_b(\infty) = b$ and so $S_{[b_n]}(\infty) = S_{[b_n]}s_b(0) = S_{[b_{n+1}]}(0)$ for any $b \in \mathbb{Z}^+$ where $S_{[b_{n+1}]} = s_{b_0}s_{b_1}\cdots s_{b_n}s_b$ and $b_i, b \in \mathbb{Z}^+$ for $i \geq 1$ and $b_0 \in \mathbb{Z}$.

In this text, we will use Möbius transformations and hyperbolic geometry to prove some of the well-known results about the convergence of simple continued fractions.

2.1.1 Definition of Convergence of a Continued Fraction

Definition 2.1 For the continued fraction $b_0 + K(1|b_n)$, the quantities

$$b_0, \quad b_0 + \frac{1}{b_1}, \quad b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2}}, \cdots \tag{2.2}$$

are called the *convergents* or *approximants* of the continued fraction.

Using the above compositions of maps, the sequence of convergents may be expressed as $S_{[b_0]}(\infty), S_{[b_1]}(\infty), S_{[b_2]}(\infty)....$

Equivalently, we can express the sequence of convergents as $b_0 + T_{[b_1]}(0), b_0 + T_{[b_2]}(0), b_0 + T_{[b_3]}(0), ...$

Definition 2.2 The continued fraction $b_0 + K(1|b_n)$ converges classically to x if and only if the limit

$$\lim_{n\to\infty} S_{[b_n]}(0)$$

exists and is equal to x.

Note that

$$\lim_{n\to\infty} S_{[b_n]}(0) = x$$

implies that

$$\lim_{n\to\infty} S_{[b_n]}(\infty) = x.$$

Alternatively, the continued fraction $b_0 + K(1|b_n)$ converges classically to x if and only if

$$b_0 + \lim_{n\to\infty} T_{[b_n]}(0) = b_0 + \lim_{n\to\infty} T_{[b_n]}(\infty) = x$$

and it is sufficient for one of these limits to equal x.

That is, the convergence of a simple continued fraction is described in terms of its convergents and a simple continued fraction converges classically to a real number x if the sequence of its convergents converges to x. ([12])

Definition 2.3 Let $x \in \mathbb{R}$ have the simple continued fraction expansion

$$x = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \ddots}} \tag{2.3}$$

Then

$$x = \lim_{k \to \infty} S_{[b_k]}(\infty)$$

where $S_{[b_k]}(z) = s_{b_0} s_{b_1} \cdots s_{b_k}(z)$. Then $S_{[b_k]}(\infty) = s_{b_0} s_{b_1} \cdots s_{b_k}(\infty)$ is called the *k-th convergent to x*. Analogously, $b_0 + T_{[b_k]}(0) = b_0 + t_{b_1} t_{b_2} \cdots t_{b_k}(0)$ is called the *k-th convergent to x*.

Since we are working in $\mathbb{C}_\infty$, ∞ is an admissible value, so a continued fraction that approaches ∞ is said to *converge* to ∞.

We note that the removal of a finite number of partial quotients at the beginning of a continued fraction will not affect its convergence. It is useful to describe the convergence and convergents of x in terms of *tails* of the simple continued fractions in the following way.

Definition 2.4 Let

$$x = \lim_{n \to \infty} S_{[b_n]}(\infty).$$

Then

$$x_{b_k} = b_k + \cfrac{1}{b_{k+1} + \cfrac{1}{b_{k+2} \ddots}} \tag{2.4}$$

is called the *k-th tail* or *k-th complete quotient* of x.

We note immediately that $x = x_{b_0}$ and thus $S^{-1}_{[b_{k-1}]}(x) = x_{b_k}$ and $x = S_{b_k}(x_{b_{k+1}})$.

2.1.2 Convergence of Positive Continued Fractions

In this section, we will show that all positive continued fractions converge. In particular, we examine the geometry of the Möbius maps of the form $t_{b_j}(z) = \frac{1}{b_j+z} = \psi\tau^{b_j}(z)$, where $\psi(z) = \frac{1}{z}$ and $\tau(z) = z+1$ and $b_j \in \mathbb{Z}^+$ for all j.

We note that $t_b = \psi\tau^b$ for $b \geq 1$ and $s_b = \tau^b\psi$ for all $b \geq 0$.

Note that $\mathbb{I}$ is the $\mathbb{H}$-line with endpoints 0 and ∞, and is mapped by t_{b_j} to the $\mathbb{H}$-line with endpoints 0 and $\frac{1}{b_j}$, for $j \geq 1$. We will consider the right half plane given by $\mathbb{K} = \{z \in \mathbb{C} : \Re(z) > 0\}$, the interior of the circle in $\mathbb{C}_\infty$ which has the imaginary axis as its circumference. This circle, and its interior $\mathbb{K}$, are mapped by the transformations t_{b_j} to the circle centered on the real axis and passing through the points 0 and $\frac{1}{b_j}$. We note further that since these transformations are all conformal, the transformed circle remains orthogonal to the real axis.

These facts lead to the following important theorem about the *convergents* of a positive real continued fraction.

Theorem 2.5 Let $K(1|b_n)$ be a positive continued fraction. Then

$$0 < T_{[b_2]}(0) < T_{[b_4]}(0) < T_{[b_6]}(0) < \cdots < T_{[b_5]}(0) < T_{[b_3]}(0) < T_{[b_1]}(0).$$

In particular, the two series

$$\sum_{n=4,6,8,...} T_{[b_n]}(0) - T_{[b_{n-2}]}(0), \sum_{n=3,5,7,...} T_{[b_{n-2}]}(0) - T_{[b_n]}(0)$$

and the two sequences $\{T_{[b_1]}(0), T_{[b_3]}(0), T_{[b_5]}(0), ...\}$, $\{T_{[b_2]}(0), T_{[b_4]}(0), T_{[b_6]}(0), ...\}$ all converge.([12])

Proof: Let $\mathbb{K} = \{z \in \mathbb{C} : \Re(z) > 0\}$ and let $t_{b_j}(z) = \frac{1}{b_j+z}$ for $j \geq 1$. Let $z \in \mathbb{K}$, with $z = x + iy$ where $x > 0$.

Now, $t_{b_j}(\infty) = \frac{1}{b_j+\infty} = 0$ and $t_{b_j}(0) = \frac{1}{b_j+0} = \frac{1}{b_j}$.

Hence $t_{b_j}(\mathbb{I}) = \ell_{b_j}$, where $\mathbb{I}$ is the positive imaginary axis and ℓ_{b_j} is the $\mathbb{H}$-line through 0 and $t_{b_j}(0)$. Since each t_{b_j} maps circles to circles, we have that $\mathbb{K}$ is mapped to $t_{b_j}(\mathbb{K})$. Each disc $T_{[b_n]}(\mathbb{K})$ is symmetric about the real axis, and as $T_{[b_{n+1}]}(\infty) = T_{[b_n]}(0)$, the real diameter of $T_{[b_n]}(\mathbb{K})$ has endpoints $T_{[b_n]}(\infty)$ and $T_{[b_{n+1}]}(\infty) = T_{[b_n]}(0)$. Thus $T_{[b_{2n}]}(\mathbb{K})$ is tangent to $T_{[b_{2n-1}]}(\mathbb{K})$ at its extreme right-hand point $T_{[b_{2n-1}]}(\infty)$, and is tangent to $T_{[b_{2n+1}]}(\mathbb{K})$ at its extreme left-hand point $T_{[b_{2n}]}(\infty)$. Thus,

$$0 < T_{[b_2]}(0) < T_{[b_4]}(0) < \cdots < T_{[b_5]}(0) < T_{[b_3]}(0) < T_{[b_1]}(0).$$

Now, the real diameter of $t_{b_n}(\mathbb{K})$ is the real closed interval $[t_{b_n}(\infty), t_{b_n}(0)]$, so the discs $T_{[b_1]}(\mathbb{K}), T_{[b_2]}(\mathbb{K}), T_{[b_3]}(\mathbb{K}), ...$ must be nested as illustrated in Figure 1 below. That is, t_{b_j} maps $\mathbb{K}$ onto a Euclidean disc $t_{b_j}(\mathbb{K})$ which lies in $\mathbb{K}$ and which is tangent to $\mathbb{I}$ at 0, and we have $\mathbb{K} \supseteq T_{[b_1]}(\mathbb{K}) \supseteq T_{[b_2]}(\mathbb{K}) \supseteq$ ... This completes the proof.

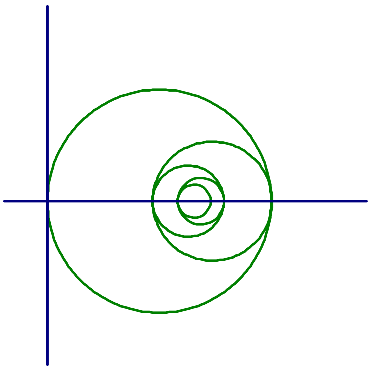

Figure 1: Nested Discs

■

A number-theoretic proof of this result can be found in [3].

Corollary 2.6 If $b_0 + K(1|b_n)$ is a simple continued fraction, then we have that the sequences $\{S_{[b_{2k+1}]}(\infty)\}_{k\geq 0}$ and $\{S_{[b_{2k}]}(\infty)\}_{k\geq 0}$ converge to α' and β' respectively, with $\beta' \leq \alpha'$.

Proof: If $K(1|b_n)$ is a simple continued fraction, then

$$0 < T_{[b_2]}(0) < T_{[b_4]}(0) < \cdots < T_{[b_5]}(0) < T_{[b_3]}(0) < T_{[b_1]}(0)$$

and the sequence $\{T_{[b_{2k+1}]}(0)\}_{k\geq 0}$ converges to α while the sequence $\{T_{[b_{2k}]}(\infty)\}_{k\geq 1}$ converges to β, where $\beta \leq \alpha$. Consider

$$\begin{aligned} T_{[b_{2k}]}(z) &= t_{b_1} t_{b_2} \cdots t_{b_{2k}}(z) \\ &= \psi\tau^{b_1}\psi\tau^{b_2}\cdots\psi\tau^{b_{2k}}(z) \end{aligned}$$

$$\begin{aligned} &= \psi\tau^{b_1}\psi\tau^{b_2}\cdots\psi\tau^{b_{2k}}\psi\psi(z) \\ &= \psi(s_{b_1}s_{b_2}\cdots s_{b_{2k}})\psi(z). \end{aligned}$$

Thus $T_{[b_{2k}]}(0) = \psi(s_{b_1}s_{b_2}\cdots s_{b_{2k}})\psi(0) = \psi(s_{b_1}s_{b_2}\cdots s_{b_{2k}})(\infty)$. Hence if

$$\lim_{k\to\infty} T_{[b_{2k}]}(0) = \beta$$

then it follows from the completeness axiom that

$$\lim_{k\to\infty} S_{[b_{2k}]}(\infty)$$

converges to some β'. Similarly, $T_{[b_{2k+1}]}(0) = t_{b_1}t_{b_2}\cdots t_{b_{2k+1}}(0) = \psi(s_{b_1}\cdots s_{b_{2k+1}})(\infty)$ and it follows from the completeness axiom that

$$\lim_{k\to\infty} S_{[b_{2k+1}]}(\infty)$$

converges to some α'. Since $\beta \leq \alpha$, we have $\beta' \leq \alpha'$. ■

We can now state and prove the following theorem about the *convergence* of positive infinite continued fractions.

Theorem 2.7 Suppose each $b_k > 0$. Then $K(1|b_n)$ converges if and only if

$$\sum_{k=1}^{\infty} b_k$$

diverges. (The Seidel-Stern Theorem [12])

Simple continued fractions are positive continued fractions, so this theorem states that every infinite simple continued fraction converges.

Proof: We prove the contrapositive of this statement. That is, we prove that $K(1|b_n)$ diverges if and only if

$$\sum_{k=1}^{\infty} b_k$$

converges.

Let $t_{b_k}(z) = \frac{1}{b_k+z}$ and let $T_{[b_n]} = t_{b_1} \cdots t_{b_n}$.

Then $t_{b_k}(z) = \frac{1}{b_k+z} = \frac{i}{ib_k+iz}$, so let t_{b_k} be represented by the matrix

$$A_k = \begin{pmatrix} 0 & i \\ i & ib_k \end{pmatrix}$$

Observe that $\det(A_k) = 1$, so A_k is *unimodular* in

$$PSL(2,\mathbb{C}) = \{z \mapsto \frac{az+b}{cz+d} : a,b,c,d \in \mathbb{C}, ad-bc = 1\}.$$

Now $||t_{b_k}||^2 = (|0|^2 + |i|^2 + |i|^2 + |ib_k|^2) = 2 + |b_k|^2$. Hence by Theorem 1.41

$$\begin{aligned} |b_k|^2 &= ||t_{b_k}||^2 - 2 \\ &= 2\cosh \rho'(\mathbf{j}, t_{b_k}(\mathbf{j})) - 2 \\ &= 2(\cosh \rho'(\mathbf{j}, t_{b_k}(\mathbf{j})) - 1) \\ &= 2(1 + 2\sinh^2 \frac{1}{2}\rho'(\mathbf{j}, t_{b_k}(\mathbf{j})) - 1) \\ &= 2(2\sinh^2 \frac{1}{2}\rho'(\mathbf{j}, t_{b_k}(\mathbf{j}))) \\ &= 4\sinh^2 \frac{1}{2}\rho'(\mathbf{j}, t_{b_k}(\mathbf{j})) \end{aligned}$$

Hence $|b_k| = 2\sinh \frac{1}{2}\rho'(\mathbf{j}, t_{b_k}(\mathbf{j}))$.

Now, let us first assume that

$$\sum_{k=1}^{\infty} b_k = \sum_{k=1}^{\infty} |b_k|$$

converges. This implies that

$$\sum_{k=1}^{\infty} 2\sinh \frac{1}{2}\rho'(\mathbf{j}, t_{b_k}(\mathbf{j}))$$

converges, since we showed above that $|b_k| = 2\sinh \frac{1}{2}\rho'(\mathbf{j}, t_{b_k}(\mathbf{j}))$. Further,

$$\sum_{k=1}^{\infty} |b_k|$$

is convergent implies that $|b_k| \to 0$ as $k \to \infty$.

Thus $2\sinh\frac{1}{2}\rho'(\mathbf{j}, t_{b_k}(\mathbf{j})) \to 0$ as $k \to \infty$.

Let $c_k = \rho'(\mathbf{j}, t_{b_k}(\mathbf{j}))$. Then $\sinh\frac{1}{2}c_k = \frac{e^{\frac{1}{2}c_k} - e^{-\frac{1}{2}c_k}}{2} \to 0$ as $k \to \infty$. Hence $c_k \to 0$ as $k \to \infty$ since the hyperbolic sine function is a homeomorphic and increasing function over its entire domain. That is $\rho'(\mathbf{j}, t_{b_k}(\mathbf{j})) \to 0$ as $k \to \infty$.

Thus, if

$$\sum_{k=1}^{\infty} \sinh\frac{1}{2}\rho'(\mathbf{j}, t_{b_k}(\mathbf{j}))$$

converges, then by the Limit Comparison Test for the convergence of positive series, we have that

$$\sum_{k=1}^{\infty} \rho'(\mathbf{j}, t_{b_k}(\mathbf{j}))$$

converges.

Since ρ' is a metric, we deduce from the triangle inequality that

$$\begin{aligned}
\rho'(\mathbf{j}, T_{[b_n]}(\mathbf{j})) &= \rho'(\mathbf{j}, t_{b_1}\cdots t_{b_n}(\mathbf{j})) \\
&\leq \rho'(\mathbf{j}, t_{b_1}\cdots t_{b_{n-1}}(\mathbf{j})) + \rho'(t_{b_1}\cdots t_{b_{n-1}}(\mathbf{j}), t_{b_1}\cdots t_{b_n}(\mathbf{j})) \\
&\leq \rho'(\mathbf{j}, t_{b_1}\cdots t_{b_{n-1}}(\mathbf{j})) + \rho'(j, t_{b_n}(\mathbf{j})) \\
&\vdots \\
&\leq \rho'(\mathbf{j}, t_{b_1}(\mathbf{j})) + \cdots + \rho'(\mathbf{j}, t_{b_n}(\mathbf{j}))
\end{aligned}$$

But

$$\rho'(\mathbf{j}, t_1(\mathbf{j})) + \cdots + \rho'(\mathbf{j}, t_n(\mathbf{j})) = \sum_{k=1}^{n} \rho'(\mathbf{j}, t_{b_k}(\mathbf{j})).$$

Hence

$$\rho'(\mathbf{j}, T_{[b_n]}(\mathbf{j})) \leq \sum_{k=1}^{n} \rho'(\mathbf{j}, t_{b_k}(\mathbf{j})).$$

But

$$\sum_{k=1}^{\infty} \rho'(\mathbf{j}, t_{b_k}(\mathbf{j}))$$

converges to some m_0, as shown above. So $\rho'(\mathbf{j}, T_{[b_n]}(\mathbf{j}))$ is bounded above by m_0, and so we may deduce that the points $T_{[b_k]}(\mathbf{j})$, $k = 1, 2, ..., n$ lie in a compact part of $\mathbb{H}^3$. That is, $\rho'(\mathbf{j}, T_{[b_n]}(\mathbf{j})) \leq m_0$. Consider the hyperbolic geodesic that has endpoints $T_{[b_k]}(0)$ and $T_{[b_{k-1}]}(0)$. We have that $T_{[b_k]}(\mathbf{j})$ lies on this geodesic. Figure 2 illustrates the closest the geodesic endpoints can get to each other when $T_{[b_k]}(\mathbf{j})$ is furthest from $\mathbf{j}$. Since we are interested in $|T_{[b_n]}(0) - T_{[b_{n-1}]}(0)|$, we note that this distance is at its minimum when the geodesic is situated as illustrated in Figure 2, when $|T_{[b_n]}(0) - T_{[b_{n-1}]}(0)| = 2T_{[b_n]}(\mathbf{j})$.

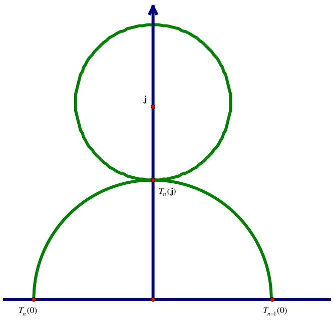

Figure 2: Minimum Distance of Geodesics

That is, the hyperbolic geodesic in $\mathbb{H}^3$ that has endpoints $T_{[b_k]}(0)$ and $T_{[b_k]}(\infty)$ contains $T_{[b_k]}(\mathbf{j})$, and so we have that $|T_{[b_k]}(0) - T_{[b_k]}(\infty)|$ or, equivalently,

$|T_{[b_k]}(0) - T_{[b_{k-1}]}(0)|$ must be bounded below by a positive number, say M_1. Hence the series

$$\sum_{n=4,6,8,\ldots} T_{[b_n]}(0) - T_{[b_{n-2}]}(0)$$

and

$$\sum_{n=3,5,7,\ldots} T_{[b_{n-2}]}(0) - T_{[b_n]}(0)$$

do not converge to the same limit, so $K(1|b_n)$ must diverge, by Theorem 2.5.

Now we prove the converse. That is, if $K(1|b_n)$ diverges, then

$$\sum_{k=1}^{\infty} b_k$$

converges.

Since cross-ratios are invariant under Möbius maps we have that

$$\begin{aligned}
(t_{b_k}(0), \infty; 0, T^{-1}_{[b_{k-1}]}(\infty)) &= (t^{-1}_{b_k}(t_{b_k}(0)), t^{-1}_{b_k}(\infty); t^{-1}_{b_k}(0), t^{-1}_{b_k}(T^{-1}_{[b_{k-1}]}(\infty))) \\
&= (0, t^{-1}_{b_k}(\infty); t^{-1}_{b_k}(0), (T_{[b_k]}t_{b_k})^{-1}(\infty)) \\
&= (0, t^{-1}_{b_k}(\infty); t^{-1}_{b_k}(0), T^{-1}_{[b_k]}(\infty)) \\
&= (T_{[b_k]}(0), T_{[b_k]}(t^{-1}_{b_k}(\infty)); T_{[b_k]}(t^{-1}_{b_k}(0)), T_{[b_k]}(T^{-1}_{[b_k]}(\infty))) \\
&= (T_{[b_k]}(0), T_{[b_{k-1}]}(\infty); T_{[b_{k-1}]}(0), \infty) \\
&= (T_{[b_k]}(0), T_{[b_{k-2}]}(0); T_{[b_{k-1}]}(0), \infty)
\end{aligned}$$

Now

$$(T_{[b_k]}(0), T_{[b_{k-2}]}(0); T_{[b_{k-1}]}(0), \infty) = \lim_{M\to\infty} \frac{(T_{[b_k]}(0) - T_{[b_{k-2}]}(0))(T_{[b_{k-1}]}(0) - M)}{(T_{[b_k]}(0) - T_{[b_{k-1}]}(0))(T_{[b_{k-2}]}(0) - M)}$$

but

$$\lim_{M\to\infty} \frac{(T_{[b_k]}(0) - T_{[b_{k-2}]}(0))(T_{[b_{k-1}]}(0) - M)}{(T_{[b_k]}(0) - T_{[b_{k-1}]}(0))(T_{[b_{k-2}]}(0) - M)} = \frac{T_{[b_k]}(0) - T_{[b_{k-2}]}(0)}{T_{[b_k]}(0) - T_{[b_{k-1}]}(0)} = \frac{T^{-1}_{[b_{k-1}]}(\infty)}{\frac{1}{b_k}}$$

So

$$\frac{T_{[b_k]}(0) - T_{[b_{k-2}]}(0)}{T_{[b_k]}(0) - T_{[b_{k-1}]}(0)} = b_k T^{-1}_{[b_{k-1}]}(\infty) \tag{2.5}$$

Further, $T^{-1}_{[b_{k-1}]}(\infty) = t_{b_k}(T^{-1}_{[b_k]}(\infty)) = \frac{1}{b_k + T^{-1}_{[b_k]}(\infty)}$

which implies $b_k T^{-1}_{[b_{k-1}]}(\infty) + T^{-1}_{[b_{k-1}]}T^{-1}_{[b_k]}(\infty) = 1$

and so

$$b_k T^{-1}_{[b_{k-1}]}(\infty) = 1 - T^{-1}_{[b_{k-1}]}(\infty)T^{-1}_{[b_k]}(\infty) \tag{2.6}$$

Now let $w_{[b_k]} = T^{-1}_{[b_k]}(\infty)$ and let $v_{[b_k]} = T_{[b_k]}(0) - T_{[b_{k-1}]}(0)$.

Then $w_{[b_k]}w_{[b_{k-1}]} = 1 - b_k w_{[b_{k-1}]} = 1 - \frac{v_{[b_k]}+v_{[b_{k-1}]}}{v_{[b_k]}} = \frac{-v_{[b_{k-1}]}}{v_{[b_k]}}$ from (2.5) and (2.6).

Hence we have

$$w_{[b_k]}w_{[b_{k-1}]} = \frac{-v_{[b_{k-1}]}}{v_{[b_k]}} \tag{2.7}$$

which implies $\frac{w_{[b_k]}w_{[b_{k-1}]}}{v_{[b_k]}} = \frac{-v_{[b_{k-1}]}}{v^2_{[b_k]}}$. This implies $v_{[b_{k-1}]}\frac{w_{[b_k]}w_{[b_{k-1}]}}{v_{[b_k]}} = \frac{-v^2_{[b_{k-1}]}}{v^2_{[b_k]}}$ which implies $\frac{w_{[b_k]}}{v_{[b_k]}} = \left(\frac{-1}{w_{[b_{k-1}]}}v_{[b_{k-1}]}\right)\left(\frac{v^2_{[b_{k-1}]}}{v^2_{[b_k]}}\right)$. Since (2.7) implies that

$$w_{[b_{k-1}]}v_{[b_{k-1}]} = \frac{-v_{[b_{k-2}]}}{w_{[b_{k-2}]}},$$

we have $\left|\frac{w_{[b_k]}}{v_{[b_k]}}\right| = \left|\frac{1}{w_{[b_{k-1}]}}v_{[b_{k-1}]}\right|\frac{v^2_{[b_{k-1}]}}{v^2_{[b_k]}}$ implies $\left|\frac{w_{[b_k]}}{v_{[b_k]}}\right| = \left|\frac{w_{[b_{k-2}]}}{v_{[b_{k-2}]}}\right|\frac{v^2_{[b_{k-1}]}}{v^2_{[b_k]}}$, which implies $\left|\frac{w_{[b_k]}}{v_{[b_k]}}\right| \geq \left|\frac{w_{[b_{k-2}]}}{v_{[b_{k-2}]}}\right|$ since $\{|v_{[b_k]}|\}$ is a decreasing sequence and $|v_{[b_k]}| \geq a > 0$ for all k, since we assumed that $K(1|b_n)$ diverges. Now we can deduce that $\left\{\left|\frac{w_{[b_k]}}{v_{[b_k]}}\right|\right\}$, $k = 1, 2, ..., n$ is bounded below by some positive number r. Hence we may conclude that the sequence $\{|w_{[b_k]}|\}$ has a positive lower bound, say R, since $\left|\frac{w_{[b_k]}}{v_{[b_k]}}\right| > r$ implies $|w_{[b_k]}| > r|v_{[b_k]}| > ar$.

Now, from (2.5), we obtain $rRb_k \leq T_{[b_k]}(0) - T_{[b_{k-2}]}(0)$ and by Theorem 2.5, the mutually disjoint open intervals $(T_{[b_2]}(0), T_{[b_4]}(0))$, $(T_{[b_4]}(0), T_{[b_6]}(0))$,..., $(T_{[b_5]}(0), T_{[b_3]}(0))$, $(T_{[b_3]}(0), T_{[b_1]}(0))$ lie in the interval $(T_{[b_2]}(0), T_{[b_1]}(0))$, and so if

$$\sum_{k=1}^{\infty} |T_{[b_k]}(0) - T_{[b_{k-2}]}(0)|$$

converges, then

$$\sum_{k=1}^{\infty} rRb_k$$

converges, and hence

$$\sum_{k=1}^{\infty} b_k$$

converges, and this completes the proof. ■

A number-theoretic proof of this result can be found in [2]. We noted in Chapter 1.2.2 that $g \in \mathcal{M}$ can be expressed as an even number of reflections in lines or circles. More generally, a Möbius transformation acting in $\mathbb{R}^n_\infty$ can be expressed as a composition of an even number of reflections in hyperplanes or hyperspheres. Since our continued fractions can be expressed as compositions of reflections, the concept of continued fractions in higher dimensions exists. This geometric interpretation of continued fractions in higher dimensions complements the algebraic methods of exploring continued fractions in higher dimensions.

Since we have shown that every simple continued fraction converges, it is natural to ask whether every real number can be expressed as a simple continued fraction. The answer to this question is "yes". In particular, every rational number can be expressed as a finite simple continued fraction in two distinct ways, while every irrational number can be expressed as an infinite simple continued fraction. We adopt the convention that the last partial quotient of the simple continued fraction of a rational must always be 1. Using this convention we will show that the simple continued fraction of any real number will be unique.

Let us first note that the simple continued fraction expansion of any rational number can be found by using Euclid's algorithm. For example, consider $\frac{151}{46} \in \mathbb{Q}$:

$151 = (3 \times 46) + 13$

$46 = (3 \times 13) + 7$

$13 = (1 \times 7) + 6$

$7 = (1 \times 6) + 1$

$6 = (1 \times 5) + 1$

Hence the simple continued fraction expansion of $\frac{151}{46}$ is given by

$$\frac{151}{46} = 3 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{5 + \cfrac{1}{1}}}}} \tag{2.8}$$

where we have adopted the convention that the final partial quotient of a finite simple continued fraction expansion must always be 1. We note that $S_{[b_n]}(\infty) = s_{b_0}s_{b_1}\cdots s_{b_n-1}s_1(\infty)$ if $b_n \geq 2$ and so this convention can always be adopted.

Theorem 2.8 A real number x has a finite simple continued fraction expansion if and only if x is rational. Further, the simple continued fraction expansion is unique if we assume that the last partial quotient must be 1. ([12])

Proof: We have seen that the finite simple continued fraction of a rational number can be written in such a way that the last partial quotient is 1, so if $x \in \mathbb{R}$ has a unique finite simple continued fraction expansion then clearly x is rational and the last partial quotient is 1.

Conversely, let $x = \frac{p}{q} \in \mathbb{Q}$ where p and q are co-prime and $q > 0$. It follows

immediately from Euclid's algorithm that the simple continued fraction expansion of $x = \frac{p}{q}$ is finite with last partial quotient 1. We need only establish the uniqueness. Suppose that
$\frac{p}{q} = s_{b_0}s_{b_1}\cdots s_{b_n}(\infty) = s_{c_0}s_{c_1}\cdots s_{c_m}(\infty)$ where $b_n = 1$ and $c_m = 1$. Assume that $m < n$.
We have $s_{b_0}(x_{b_1}) = s_{c_0}(x_{c_1})$, where c_0 and b_0 are the integer parts of x. Thus $b_0 = c_0$ and $s_{b_0} = s_{c_0}$.

Hence, applying $s_{b_0}^{-1}$ to both sides, we have $\frac{p'}{q'} = x_{b_1} = x_{c_1}$.

Since b_1 and c_1 are the integer parts of $\frac{p'}{q'}$, we have $b_1 = c_1$ and $s_{b_1} = s_{c_1}$. Hence, applying $s_{b_1}^{-1}$ to both sides of $s_{b_1}s_{b_2}\cdots s_{b_n}(\infty) = s_{c_1}s_{c_2}\cdots s_{c_m}(\infty)$, we obtain $s_{b_2}s_{b_3}\cdots s_{b_n}(\infty) = s_{c_2}s_{c_3}\cdots s_{c_m}(\infty)$. Continuing in this way, we obtain that $b_i = c_i$ for all $i \leq m$. Then $b_m = 1$ and $s_{b_{m+1}}s_{b_{m+2}}\cdots s_{b_n}(\infty) = \infty$. But this is impossible since $b_{m+i} \geq 1$ for all i, so $s_{b_{m+1}}s_{b_{m+2}}\cdots s_{b_n}(\infty)$ is finite. Hence $m \geq n$. Similarly, we can show that $n \geq m$, and so $m = n$.

The uniqueness of the simple continued fraction for $\frac{p}{q}$ is thus established if we choose the last coefficient to be 1.

■

Theorem 2.9 Every irrational number x can be expressed uniquely as an infinite simple continued fraction.([12])

Proof: Let $x \in \mathbb{R} \setminus \mathbb{Q}$ and let $[x] = q_0$ denote the integer part of x. Thus $0 < x - [x] < 1$. Let $x_{q_1} = s_{q_0}^{-1}(x) = \frac{1}{x-q_0}$ and $q_1 = [x_{q_1}]$. Then $x_{q_1} > 1$ and $x_{q_1} \in \mathbb{R} \setminus \mathbb{Q}$.
Now, $x = s_{q_0}(x_{q_1})$. Hence

$$x_{q_1} = [x_{q_1}] + x_{q_1} - [x_{q_1}] = q_1 + (x_{q_1} - q_1) = q_1 + \frac{1}{\frac{1}{x_{q_1}-q_1}} = q_1 + \frac{1}{x_{q_2}} = s_{q_1}(x_{q_2})$$

where $x_2 = \frac{1}{x_1 - q_1} > 1$ and $x_2 \in \mathbb{R} \setminus \mathbb{Q}$. Continuing in this way, we see that $x = s_{q_0} s_{q_1} \cdots s_{q_k}(x_{q_{k+1}}) = S_{[q_k]}(x_{q_{k+1}})$ where $q_0 \in \mathbb{Z}$, $q_i \in \mathbb{Z}^+$ for $i \geq 1$ and $x_{q_{k+1}} > 1$ and $x_{q_{k+1}} \in \mathbb{R} \setminus \mathbb{Q}$.

Since $q_i \in \mathbb{Z}^+$ for $i \geq 1$, we know that the simple continued fraction $S_{[q_k]}(\infty) = s_{q_0} s_{q_1} \cdots s_{q_k}(\infty)$ converges to some real X, by Theorem 2.7. We show that $x = X$. Letting $s_{q_k}(z) = q_k + \frac{1}{z}$ and $S_{[q_k]}(z) = s_{q_0} s_{q_1} \cdots s_{q_k}$, we note that for each k we have $x = S_{[q_k]}(x_{q_{k+1}})$. Note that if $a > 0$ then the mapping $x \mapsto a + \frac{1}{x}$, where $x > 0$, is a decreasing map of the interval $(0, \infty)$ into itself, and $a + \frac{1}{x}$ decreases as x increases. Thus the composition of such maps, as we have with $S_{[q_k]} = s_{q_0} s_{q_1} ... s_{q_k}$, is an increasing or decreasing map of $(0, \infty)$ into itself, depending on whether k is even or odd, by Corollary 2.6. That is, $S_{[q_0]}(\infty), S_{[q_2]}(\infty), ..., S_{[q_{2k}]}(\infty)$ is increasing while $S_{[q_1]}(\infty), S_{[q_3]}(\infty), ..., S_{[q_{2k+1}]}(\infty)$ is decreasing, for $k = 0, 1, 2, ...$ Thus

$$S_{[q_{2k+1}]}(\infty) \leq S_{[q_{2k+1}]}(x_{2k+2}) = x = S_{[q_{2k}]}(x_{2k+1}) \leq S_{[q_{2k}]}(\infty).$$

Letting $k \to \infty$, we see that $X = x$.

To establish uniqueness, suppose that

$$x = \lim_{m \to \infty} s_{a_0} s_{a_1} \cdots s_{a_m}(\infty) = \lim_{n \to \infty} s_{b_0} s_{b_1} \cdots s_{b_n}(\infty)$$

where $a_i, b_i \in \mathbb{Z}^+$ if $i \geq 1$. Equating the integer parts, we obtain $a_0 = b_0$. Now suppose that

$$x_{a_k} = \lim_{m \to \infty} s_{a_k} s_{a_{k+1}} \cdots s_{a_m}(\infty) = \lim_{n \to \infty} s_{b_k} s_{b_{k+1}} \cdots s_{b_n}(\infty) = x_{b_k}.$$

Then we have that $b_k = a_k$ and so $s_{b_k} = s_{a_k}$. Acting $s_{b_k}^{-1} = s_{a_k}^{-1}$ on both sides, we obtain

$$x_{a_{k+1}} = \lim_{m \to \infty} s_{a_{k+1}} s_{a_{k+2}} \cdots s_{a_m}(\infty) = \lim_{n \to \infty} s_{b_{k+1}} s_{b_{k+2}} \cdots s_{b_n}(\infty) = x_{b_{k+1}}$$

which gives us $b_{k+1} = a_{k+1}$. Hence by induction on i we have that $a_i = b_i$ for all $i = 0, 1, 2, ...$ ■

Since we have established the uniqueness of the simple continued fraction expansion for any real, we may simplify the notation for the convergents $S_{[b_n]}$ to just S_n and for the tails from x_{b_n} to just x_n when there is no ambiguity.

The number-theoretic proofs of Theorems 2.8 and 2.9 can be found in [3].

2.2 Equivalence of Continued Fractions

Definition 2.10 Two real numbers x and y are said to be $\sim$ *equivalent* ,denoted $x \sim y$, if we can find a transformation $g \in \widetilde{\Gamma}$ with $g(y) = x$ so that x and y are in the same orbit under $\tilde{\Gamma}$.

This $\sim$ relation is clearly an equivalence relation, and any rational number is equivalent to zero and thus any two rational numbers are equivalent to each other, by transitivity.

Definition 2.11 Let $x, y \in \mathbb{R} \setminus \mathbb{Q}$. We say x and y have the same tail, or $x \approx y$, if there exist $p, q \in \mathbb{Z}^+$ such that $x_{a_{p+n}} = y_{b_{q+n}}$ for $n = 0, 1, 2, ...$, where

$$x = \lim_{n \to \infty} S_{[a_n]}(\infty),$$

where $S_{[a_n]}(z) = s_{a_0} s_{a_1} \cdots s_{a_n}(z)$, and

$$y = \lim_{m \to \infty} S_{[b_m]}(\infty),$$

where $S_{[b_m]}(z) = s_{b_0} s_{b_1} \cdots s_{b_m}(z)$, are the simple continued fraction expansions of x and y respectively. Without loss of generality we may write $x_{p+n} = y_{q+n}$ for $n = 0, 1, 2, ...$ instead of $x_{a_{p+n}} = y_{b_{q+n}}$ for $n = 0, 1, 2, ...$

Lemma 2.12 $\approx$ is an equivalence relation on $\mathbb{R} \setminus \mathbb{Q}$.

Proof: Let $x, y, z \in \mathbb{R} \setminus \mathbb{Q}$:

1. $x \approx x$ since the simple continued fraction expansion of x is unique.
2. If $x \approx y$ then $y \approx x$ since the relationship is clearly symmetric.
3. Suppose $x \approx y$ and $y \approx z$. Then $x_{p+n} = y_{q+n}$ and $y_{t+m} = z_{s+m}$ for $p, q, t, s \in \mathbb{Z}^+$ and $m = 0, 1, 2, ...,$ $n = 0, 1, 2, ...$ This implies that $x_{r+n} = z_{v+n}$ for some $r, v \in \mathbb{Z}^+$ and $n = 0, 1, 2, ...$ Hence $x \approx z$

■

We will show that two *irrational* numbers x and y are $\sim$ equivalent (under $\widetilde{\Gamma}$) if and only if their simple continued fraction expansions have the same tail. This is stated formally in Theorem 2.13.

Theorem 2.13 Suppose $x, y \in \mathbb{R} \setminus \mathbb{Q}$. Then x and y are in the same orbit under $\tilde{\Gamma}$ if and only if x and y have the same tail, or $x \sim y$ if and only if $x \approx y$. ([12])

Proof: We know that $\tilde{\Gamma} = \langle \tau, \psi \rangle$, where $\tau(z) = z + 1$ and $\psi(z) = \frac{1}{z}$. Let $\omega(z) = -z$, where $\omega \in \tilde{\Gamma}$. We will first show the following:

1. $\tau(z) \approx z$ for all z
2. $\omega(z) \approx z$ for all z
3. $\psi(z) \approx z$ for all z

Let

$$z = \lim_{k \to \infty} S_{[a_k]}(\infty) = \lim_{k \to \infty} s_{a_0} s_{a_1} \cdots s_{a_k}(\infty) = S_{[a_k]}(z_{k+1}),$$

be the simple continued fraction expansion of z. Note that $\omega(s_{a_k}(z)) = \omega\tau^{a_k}\psi(z) = s_{-a_k}(\omega(z))$. That is, $-s_{a_k}(z) = s_{-a_k}(-z)$. Further, note that $s_a s_1 s_b(z) = s_{a+1}s_{-(b+1)}(-z)$, for $a, b \in \mathbb{Z}$. That is,

$$s_a s_1 s_b(z) = s_{a+1}s_{-(b+1)}\omega(z). \tag{2.9}$$

1. Clearly $\tau(z) \approx z$ since z and

   $$1 + z = \lim_{k\to\infty} s_{a_0+1}s_{a_1}\cdots s_{a_k}(\infty) = \tau S_{[a_k]}(z_{k+1})$$

   have the same tails.

2. $-z = \omega(z) = \omega(S_{[a_k]}(z_{k+1})) = S_{[-a_k]}\omega(z_{k+1})$

   Consider the following two cases. Case (a): $z < 0$, and Case (b): $z > 0$.

   Case (a): If $z < 0$, then $a_0 \leq -1$ and $a_i \geq 1$ for all $i \geq 1$.

   If $a_1 \neq 1$ then

   $$-z = \omega(z) = \omega(S_{[a_k]}(z_{k+1})) = S_{[-a_k]}\omega(z_{k+1}) = \lim_{k\to\infty} s_{-a_0-1}s_1 s_{a_1-1}s_{-a_2}\cdots s_{-a_k}(\infty)$$

   by (2.9), with $-a_0 = a + 1$ and $-a_1 = -b - 1$.

   If $a_1 = 1$ then

   $$-z = \lim_{k\to\infty} s_{-a_0}s_{-a_1}s_{-a_2}\cdots s_{-a_k}(\infty) = \lim_{k\to\infty} s_{-a_0-1}s_{1+a_2}s_{a_3}\cdots s_{a_k}(\infty)$$

   by (2.9), with $a = a_0$ and $b = a_2$. Hence $\omega(z) \approx z$ if $z < 0$.

   Case (b): We have $-z < 0$, so by case (a) we have $\omega(-z) \approx -z$. But $\omega(-z) = -(-z) = z$, so $z \approx -z = \omega(z)$. Hence $\omega(z) \approx z$ for all z.

3. Consider the following two cases. Case (a): $z > 0$, and Case (b): $z < 0$.

   Case (a): If $0 < z < 1$ then

   $$z = \lim_{k\to\infty} s_{a_0}s_{a_1}s_{a_2}\cdots s_{a_k}(\infty) = S_{[a_k]}(z_{k+1})$$

where $a_0 = 0$ and $a_i \in \mathbb{Z}^+$ for all $i \geq 1$. Then

$$\psi(z) = \frac{1}{z} = \psi(S_{[a_k]}(z_{k+1})) = s_{a_0}^{-1} S_{[a_k]}(z_{k+1})$$

where $a_i \in \mathbb{Z}^+$ for all i. Hence $\psi(z) = \frac{1}{z} \approx z$.

If $z > 1$ then set $v = \psi(z) = \frac{1}{z}$ so that $0 < v < 1$. By above, $v \approx \frac{1}{v} = \psi(v)$. Thus $\psi(\psi(z)) \approx \psi(z)$ and so $z \approx \psi(z)$. Hence $\psi(z) \approx z$ for all $z > 0$.

Case (b): $z < 0$

From (2) above, we have $\omega(z) \approx z$ and $\omega(z) > 0$ if $z < 0$.

From 3(a) above, we have $\psi(\omega(z)) \approx \omega(z)$ and $\omega(z) \approx z$

so $\psi(\omega(z)) \approx z$ by transitivity of $\approx$.

But $\psi(\omega(z)) = \psi(-z) = \frac{1}{-z} = -(\frac{1}{z}) = \omega(\psi(z))$.

Therefore $\omega(\psi(z)) \approx z$ and $\omega(\psi(z)) \approx \psi(z)$.

Therefore $z \approx \psi(z)$, by transitivity of $\approx$.

Hence $z \approx \psi(z)$ for $z < 0$ and thus $z \approx \psi(z)$ for all $z \in \mathbb{R} \setminus \mathbb{Q}$.

Now we can conclude that if $x \sim y$, then $x \approx y$, for all $x, y \in \mathbb{R} \setminus \mathbb{Q}$.

Conversely, suppose $x \approx y$, with $x = s_{c_0} s_{c_1} ... s_{c_t}(x_{t+1}) = S_{[c_t]}(x_{t+1})$ and $y = s_{b_0} s_{b_1} ... s_{b_r}(y_{r+1}) = S_{[b_r]}(y_{r+1})$, where $b_0, c_0 \in \mathbb{Z}$ and $b_i, c_i \in \mathbb{Z}^+$ for $i \geq 1$ and $S_{[b_r]}, S_{[c_t]} \in \tilde{\Gamma}$

Then $\beta = x_{t+1} = y_{r+1}$ for some $r, t \in \mathbb{Z}^+$. Therefore $x = S_{[c_t]}(\beta)$ and $\beta = S_{[c_t]}^{-1}(x)$ and so $y = S_{[b_r]}(\beta) = S_{[b_r]} S_{[c_t]}^{-1}(x) = S(x)$, where $S = S_{[b_r]} S_{[c_t]}^{-1} \in \tilde{\Gamma}$. Thus $x \sim y$.

■

Chapter 3

Geometry of Simple Continued Fractions

3.1 Introduction

In this chapter, we investigate the tessellation of $\mathbb{H}^2$ by Farey triangles. We develop this tessellation by considering the orbit of $\mathbb{I}$ under Γ or $\tilde{\Gamma}$.

In this section we identify $\mathbb{H}^2$ with $\mathbb{H}^\perp$ and refer to $\mathbb{H}^\perp$ exclusively. Recall that in order to deal with $\tilde{\Gamma}$ acting on $\mathbb{H}^2$, we identify $\psi : z \mapsto \frac{1}{z}$ with its Poincaré extension $\tilde{\psi}$ which preserves $\mathbb{H}^\perp$. We note that ψ acting on $\mathbb{C}_\infty$ is the composition of two reflections, namely inversion in the unit sphere and reflection in the plane through $\mathbb{R}_\infty$.

The tessellation that we develop gives a rich geometric description of Γ and $\tilde{\Gamma}$. It is this description of Γ (or $\tilde{\Gamma}$) acting on $\mathbb{H}^2$ that enables us to consider simple continued fractions in a geometric way by considering the cutting sequences across the tessellations. Further, we show how this tessellation can be interpreted as a graph, and that the simple continued fraction expansion of any real number is related to a unique path on this graph.

3.2 Farey Geodesics

Recall that geodesics in $\mathbb{H}^2$ can be vertical line segments or semicircles orthogonal to $\mathbb{R}_\infty$. If the endpoints α and β of a geodesic γ lie on $\partial\mathbb{H}^2 = \mathbb{R}_\infty$ then we denote the geodesic γ by $[\alpha : \beta]$, where $[\alpha : \beta] = [\beta : \alpha]$. The positive imaginary axis $\mathbb{I} = [0 : \infty]$ is the geodesic that has endpoints 0 and ∞ and is called the *fundamental geodesic*. Recall too that all geodesics are segments of circles or generalized circles in $\mathbb{C}_\infty$ and that $g \in \Gamma$ maps circles in $\mathbb{C}_\infty$ to circles in $\mathbb{C}_\infty$.

Definition 3.1

1. A *Farey geodesic* is the image of $\mathbb{I}$ under some $g \in \Gamma$. We denote by F the set of all Farey geodesics. That is,
   $F = \{g(\mathbb{I}) : g \in \Gamma\}$.

2. If $\gamma = g(\mathbb{I})$ is a Farey geodesic, then its endpoints $g(0)$ and $g(\infty)$ on $\mathbb{R}_\infty$ are called *Farey neighbours.*

The relationship between $\mathbb{Q}_\infty$, Farey geodesics, Farey neighbours, Γ and $\tilde{\Gamma}$ is made explicit in the following theorem.

Theorem 3.2 Let $\gamma = [\alpha : \beta]$ be a hyperbolic geodesic with $\alpha, \beta \in \mathbb{R}_\infty$.

1. If γ is a Farey geodesic then $\alpha, \beta \in \mathbb{Q}_\infty$.

2. $F = \{g(\mathbb{I}) : g \in \tilde{\Gamma}\}$.

3. $\gamma = \left[\frac{p}{q} : \frac{r}{s}\right]$ is a Farey geodesic if and only if $|ps - qr| = 1$. We call this the *Farey neighbourhood condition.*

4. $[x : \infty] \in F$ if and only if $x \in \mathbb{Z}$.

5. Every element in $\mathbb{Q}_\infty$ is the end point of infinitely many Farey geodesics.

6. If x and y are Farey neighbours then $g(x)$ and $g(y)$ are Farey neighbours for all $g \in \tilde{\Gamma}$. That is, extended modular transformations map Farey geodesics to Farey geodesics.

Proof:

1. $\gamma \in F$ implies that $\gamma = g(\mathbb{I})$ where $g \in \Gamma$ with $g(z) = \frac{az+b}{cz+d}$. Since $g(0) = \frac{b}{d}$ and $g(\infty) = \frac{a}{c}$ are in $\mathbb{R}_\infty$, they are the unique endpoints of γ and so $\gamma = \left[\frac{b}{d} : \frac{a}{c}\right]$ and $\frac{a}{c}, \frac{b}{d} \in \mathbb{Q}_\infty$.

2. We know from Theorem 1.38 that $\tilde{\Gamma} = \Gamma \cup \Gamma\psi$ and so Γ is a normal subgroup of $\tilde{\Gamma}$. Thus $F \subseteq \{g(\mathbb{I}) : g \in \tilde{\Gamma}\}$. Let $g \in \tilde{\Gamma} \setminus \Gamma$, then $g = h\psi$ where $h \in \Gamma$. We note that $\psi(\mathbb{I}) = \mathbb{I}$, so $g(\mathbb{I}) = h\psi(\mathbb{I}) = h(\mathbb{I}) \in F$. Hence $\{g(\mathbb{I}) : g \in \tilde{\Gamma}\} \subseteq F$ and thus $\{g(\mathbb{I}) : g \in \tilde{\Gamma}\} = F$.

3. If γ is a Farey geodesic then $\gamma = \left[\frac{p}{q} : \frac{r}{s}\right] = g(\mathbb{I})$ where $g(z) = \frac{pz+r}{qz+s}$ or $\frac{rz+p}{sz+q}$ and $g(0) = \frac{r}{s}$ or $\frac{p}{q}$, and $g(\infty) = \frac{p}{q}$ or $\frac{r}{s}$ and $g \in \tilde{\Gamma}$. Thus $|ps - qr| = 1$.

   Conversely, if $\frac{p}{q}$ and $\frac{r}{s}$ in $\mathbb{Q}_\infty$ are such that $|ps - qr| = 1$, then $g(z) = \frac{pz+r}{qz+s}$ and $g \in \tilde{\Gamma}$ and $g(\mathbb{I}) \in F$. Thus $\left[\frac{p}{q} : \frac{r}{s}\right] \in F$.

4. If $[x:\infty] \in F$ then there is a $g \in \tilde{\Gamma}$ such that $g(z) = \frac{az+b}{cz+d}$, $|ad-bc| = 1$ and $a, b, c, d \in \mathbb{Z}$, with $g(\infty) = \frac{a}{c}$ and $g(0) = \frac{b}{d}$ and $[x:\infty] = g(\mathbb{I})$. Either $g(0) = \infty$ or $g(\infty) = \infty$. If $g(\infty) = \infty$ and $g(0) = x$, then $\frac{a}{c} = \infty$ and $c = 0$ and $ad = \pm 1$. Since $a, d \in \mathbb{Z}$ we have $d = \pm 1$. Thus $x = g(0) = \frac{b}{d} = \pm b \in \mathbb{Z}$. Alternately, $g(0) = \infty$ and $g(\infty) = x$ implies that $\frac{b}{d} = \infty$ and $d = 0$ and $-bc = \pm 1$. As above, we obtain $c = \pm 1$ and so $x = \frac{a}{c} = \pm a \in \mathbb{Z}$. Conversely, if $x \in \mathbb{Z}$ then x and ∞ are Farey neighbours since $|x \cdot 0 - 1 \cdot 1| = 1$, and so $[x:\infty] \in F$.

5. For each integer m we know that $[m:\infty]$ is a Farey geodesic, by part 4 above. Further, for any $\frac{p}{q} \in \mathbb{Q}$ with p and q co-prime, there are infinitely many integers m and n such that $pn - qm = 1$. Let $g(z) = \frac{pz+m}{qz+n}$ where $g(\infty) = \frac{p}{q}$. Hence there are infinitely many Farey geodesics with end point $\frac{p}{q}$.

6. If x and y are Farey neighbours, then $[x:y] \in F$ and so there exists an $f \in \tilde{\Gamma}$ such that $\gamma = [x:y] = f(\mathbb{I})$. Thus $g(\gamma) = [g(x):g(y)] = gf(\mathbb{I}) \in F$ since $gf \in \tilde{\Gamma}$. That is, $g(x), g(y) \in \mathbb{Q}_\infty$ and satisfy the Farey neighbour condition. Therefore $g(x)$ and $g(y)$ are Farey neighbours for all $g \in \tilde{\Gamma}$.

■

Recall that if a group G acts on a set X then the *stabilizer* of an element $x \in X$ under G is given by $G_x = \{g \in G : g(x) = x\}$. We will consider Γ and $\tilde{\Gamma}$ acting on F. For $g \in \tilde{\Gamma}$ we denote $\tau^g = g\tau g^{-1}$ and so $\langle \tau^g \rangle = \langle g\tau g^{-1} \rangle = g\langle \tau \rangle g^{-1}$, and $(g\tau g^{-1})^n = g\tau^n g^{-1} = \langle \tau \rangle^g$. Recall that we use 1_G to denote the identity element of a group G.

Theorem 3.3 Recall that $\mathbb{I} = [0 : \infty]$, $\varphi(z) = \frac{-1}{z}$, $\psi(z) = \frac{1}{z}$ and $\omega(z) = -z$. Then Γ and $\tilde{\Gamma}$ act on F in the following way:

1. $\Gamma_{\mathbb{I}} = \{1_{\mathcal{M}}, \varphi\}$, $\tilde{\Gamma}_{\mathbb{I}} = \{1_{\mathcal{M}}, \varphi, \psi, \omega\}$.

2. $\Gamma_{g(\mathbb{I})} = \{1_{\mathcal{M}}, g\varphi g^{-1}\}$ for $g \in \Gamma$ and $\tilde{\Gamma}_{g(\mathbb{I})} = \{1_{\mathcal{M}}, g\varphi g^{-1}, g\psi g^{-1}, g\omega g^{-1}\}$ for $g \in \tilde{\Gamma}$.

3. Two Farey geodesics are either equal or disjoint. That is, two Farey geodesics will never intersect each other in $\mathbb{H}^2$.

4. If $\frac{a}{c}, \frac{b}{d} \in \mathbb{Q}$ and $\left[\frac{a}{c} : \frac{b}{d}\right] \in F$ then $\frac{a}{c}$ and $\frac{b}{d}$ lie in the same unit interval.

Proof:

1. $\Gamma_{\mathbb{I}} = \{g \in \Gamma : g(\mathbb{I}) = \mathbb{I}\}$, where $g(z) = \frac{az+b}{cz+d}$ with $ad - bc = 1$ and $a, b, c, d \in \mathbb{Z}$. Suppose $g \in \Gamma_{\mathbb{I}}$. Then $g(0) = \infty$ and $g(\infty) = 0$, or $g(0) = 0$ and $g(\infty) = \infty$. That is, $d = 0$ and $a = 0$, or $b = 0$ and $c = 0$. Since $ad - bc = 1$ and $a, b, c, d \in \mathbb{Z}$, we have $g(z) = \frac{b}{cz} = \frac{-1}{z}$ or $g(z) = z$. Thus $g = \varphi$ or $g = 1_{\mathcal{M}}$. Hence $\Gamma_{\mathbb{I}} \subseteq \{\psi, 1_{\mathcal{M}}\}$. Now suppose $g \in \{\psi, 1_{\mathcal{M}}\}$. Since $\varphi(\mathbb{I}) = \mathbb{I}$ and $1_{\mathcal{M}}(\mathbb{I}) = \mathbb{I}$, we must have that $\{\psi, 1_{\mathcal{M}}\} \subseteq \Gamma_{\mathbb{I}}$ and hence $\Gamma_{\mathbb{I}} = \{1_{\mathcal{M}}, \varphi\}$.

   Recall that we consider $\tilde{\Gamma}$ acting on $\mathbb{H}^{\perp}$ rather than $\mathbb{H}^2$ and thus $\psi(\mathbb{I}) = \tilde{\psi}(\mathbb{I}) = \mathbb{I}$.

   $\tilde{\Gamma}_{\mathbb{I}} = \{g \in \tilde{\Gamma} : g(\mathbb{I}) = \mathbb{I}\}$. If $g \in \tilde{\Gamma}_{\mathbb{I}}$ then $g(z) = \frac{az+b}{cz+d}$ with $ad - bc = \pm 1$ and $a, b, c, d \in \mathbb{Z}$. As above, we obtain that $g = 1_{\mathcal{M}}$ or $g = \varphi$ but we can also have $bc = 1$ and $ad = -1$, in which case $g = \psi$ or $g = \omega$. Thus $\tilde{\Gamma}_{\mathbb{I}} = \{1_{\mathcal{M}}, \varphi, \psi, \omega\}$, where $\varphi^2 = \psi^2 = 1_{\mathcal{M}}$, $\varphi\psi = \omega = \psi\varphi$ and $\psi\omega = \varphi = \omega\psi$ and $\varphi\omega = \psi = \omega\varphi$.

2. Let $g \in \Gamma$.

$$\begin{aligned}
\Gamma_{g(\mathbb{I})} &= \{f \in \Gamma : f(g(\mathbb{I})) = g(\mathbb{I})\} \\
&= \{f \in \Gamma : g^{-1}fg(\mathbb{I}) = \mathbb{I}\} \\
&= \{f \in \Gamma : g^{-1}fg \in \Gamma_{\mathbb{I}}\} \\
&= \{f \in \Gamma : f \in g\Gamma_{\mathbb{I}}g^{-1}\} \\
&= \{1_{\mathcal{M}}, g\varphi g^{-1}\}.
\end{aligned}$$

Similarly, if we let $f, g \in \tilde{\Gamma}$ then we have

$$\begin{aligned}
\tilde{\Gamma}_{g(\mathbb{I})} &= \{f \in \tilde{\Gamma} : f(g(\mathbb{I})) = g(\mathbb{I})\} \\
&= \{f \in \tilde{\Gamma} : g^{-1}fg(\mathbb{I}) = \mathbb{I}\} \\
&= \{f \in \tilde{\Gamma} : g^{-1}fg \in \tilde{\Gamma}_{\mathbb{I}}\} \\
&= \{f \in \tilde{\Gamma} : f \in g\tilde{\Gamma}_{\mathbb{I}}g^{-1}\} \\
&= \{1_{\mathcal{M}}, g\varphi g^{-1}, g\psi g^{-1}, g\omega g^{-1}\}.
\end{aligned}$$

3. Let $\gamma_1 = f(\mathbb{I})$, $\gamma_2 = g(\mathbb{I})$ where $f, g \in \Gamma$. Then $\gamma_1 = \gamma_2$ implies that $f(\mathbb{I}) = g(\mathbb{I})$ and thus $g^{-1}f(\mathbb{I}) = \mathbb{I}$. Thus $g^{-1}f = 1_{\mathcal{M}}$ or $g^{-1}f = \varphi$. Hence we have $f = g$ or $f = g\varphi$.

   Let $g \in \Gamma$ and $g \neq 1_{\mathcal{M}}, \varphi$. Then $g(\mathbb{I}) \neq \mathbb{I}$. Suppose $g(\mathbb{I}) \cap \mathbb{I} \neq \Phi$. If there exists a distinct pair z and w in $g(\mathbb{I}) \cap \mathbb{I}$ then by definition of a hyperbolic geodesic we have that $g(\mathbb{I})$ and $\mathbb{I}$ are identical as the unique $\mathbb{H}$-line through z and w. But this contradicts $g(\mathbb{I}) \neq \mathbb{I}$, so if $g(\mathbb{I}) \cap \mathbb{I} \neq \Phi$ then it contains at most one element. That is, $g(\mathbb{I})$ cuts $\mathbb{I}$. Thus the endpoints $g(0)$ and $g(\infty)$ of $g(\mathbb{I})$ must satisfy $g(0) < 0 < g(\infty)$ or $g(\infty) < 0 < g(0)$. That is,

$$\frac{b}{d} < 0 < \frac{a}{c} \tag{3.1}$$

   or

$$\frac{a}{c} < 0 < \frac{b}{d} \tag{3.2}$$

where $g(z) = \frac{az+b}{cz+d}$ with $ad - bc = 1$ and $a, b, c, d \in \mathbb{Z}$. We can assume that $d > 0$ and $c > 0$ and so we have that either $bc < 0 < ad$ or $ad < 0 < bc$.

Thus we have $\left(\frac{a}{c}\right)\left(\frac{b}{d}\right) < 0$, from (3.1) and (3.2).

Thus $bcad = (bc)^2(\frac{ad}{bc}) < 0$ and $bcad = bc(1 + bc)$ since $ad - bc = 1$.

Thus $bc(1 + bc) < 0$. But if $bc < 0 < ad$ and $bc(1 + bc) < 0$ then $1 + bc > 0$ and so $-1 < bc < 0$. But $b, c \in \mathbb{Z}$, so this is impossible.

If $ad < 0 < bc$ then $1 + bc < 0$ and $bc < -1$ and so $0 < bc < -1$, which is also impossible.

Thus there does not exist a singleton $z \in g(\mathbb{I}) \cap \mathbb{I}$ and so $g(\mathbb{I}) \cap \mathbb{I} = \Phi$.

4. Let $\frac{a}{c}, \frac{b}{d} \in \mathbb{Q}$ with $c \neq 0$, $d \neq 0$. Let $g \in \Gamma$ such that $g(\mathbb{I}) = \left[\frac{a}{c} : \frac{b}{d}\right] \in F$. By part 3 above, we have that $g(\mathbb{I}) = \mathbb{I}$ or $g(\mathbb{I}) \cap \mathbb{I} = \Phi$.

   Suppose that there exists a $m \in \mathbb{Z}$ such that $\frac{a}{c} < m < \frac{b}{d}$. Let $\tau^{-m}(z) = z - m$. Then $\tau^{-m} \in \Gamma$ and $\tau^{-m}g(\mathbb{I})$ will cut $\mathbb{I}$ with

$$\tau^{-m}\left(\frac{a}{c}\right) < 0 < \tau^{-m}\left(\frac{b}{d}\right).$$

   That is,

$$\frac{a}{c} - m < 0 < \frac{b}{d} - m.$$

   This is a contradiction, so the endpoints $\frac{a}{c}$ and $\frac{b}{d}$ must lie in a closed interval $[m, m + 1]$ where we may have $c = d = 1$ with $a = m$ and $b = m + 1$.

■

3.3 The Farey Tessellation

A hyperbolic n-gon is a polygon with n hyperbolic line segments as its sides. The line segments intersect in pairs at points, and we call these points the *cusps* of the n-gon, and the line segments do not intersect other than at these cusps. Further, these cusps will be used to represent the n-gons in $\mathbb{H}^{\perp}$.

In particular, if $n = 3$ then the hyperbolic n-gon is a hyperbolic triangle. That is, an open set bounded by three hyperbolic geodesics and denoted by $\mathbb{T} = \{z_1; z_2; z_3\}$ where z_1, z_2 and z_3 are the cusps of $\mathbb{T}$. If z_1, z_2 and z_3 all lie on $\mathbb{R}_\infty$ then $\mathbb{T}$ is called an *ideal triangle.*

The ideal triangle $\mathbb{T}_0 = \{0; 1; \infty\} = \left\{\frac{0}{1}; \frac{1}{1}; \frac{1}{0}\right\}$ (Figure 3) is called the *fundamental triangle.* That is, $\mathbb{T}_0$ is the hyperbolic triangle bounded by the geodesics $[0 : \infty]$, $[0 : 1]$ and $[1 : \infty]$, and plays a central role in the development of the Farey tessellation of $\mathbb{H}^2$.

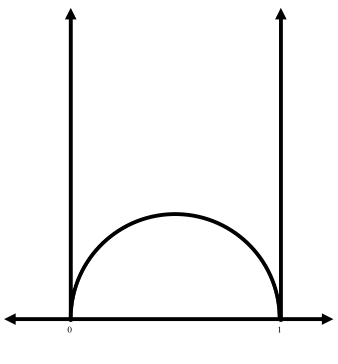

Figure 3: The Fundamental Triangle

If $\mathbb{T}$ is a hyperbolic triangle then we denote $\bar{\mathbb{T}}$ as the closure of $\mathbb{T}$ in $\mathbb{H}^{\perp}$. Thus $\bar{\mathbb{T}}_0 = \mathbb{T}_0 \cup [0 : \infty] \cup [1 : \infty] \cup [1 : 0]$.

We note that in $\mathbb{T}_0$ the cusp at $1 = \frac{1}{1}$ can be written as the median of $\frac{0}{1}$ and $\frac{1}{0}$. That is, $\frac{0}{1} \oplus \frac{1}{0} = \frac{0+1}{1+0} = \frac{1}{1}$. This property plays an important role in the Farey tessellation of $\mathbb{H}^2$. We formally define the median (or Farey sum) as follows:

Definition 3.4 Let $\frac{a}{c}$ and $\frac{b}{d}$ be reduced rationals. Without loss of generality, we may assume that $c \geq 0$ and $d \geq 0$. Then we say that $\frac{a}{c} \oplus \frac{b}{d} = \frac{a+b}{c+d}$ is the *Farey sum* of $\frac{a}{c}$ and $\frac{b}{d}$.

Definition 3.5 A *Farey triangle* is the image of $\mathbb{T}_0$ under an element of Γ. The set of all Farey triangles will be denoted by $\mathbb{F} = \{g(\mathbb{T}_0) : g \in \Gamma\}$.

Note that $\psi(0) = \infty$, $\psi(\infty) = 0$ and $\psi(1) = 1$, so $\psi(z) = \frac{1}{z}$ leaves $\mathbb{T}_0$ invariant, and hence we also have that $\mathbb{F} = \{g(\mathbb{T}_0) : g \in \tilde{\Gamma}\}$. If $g(z) = \frac{az+b}{cz+d}$ and $|ad - bc| = 1$ and $a, b, c, d \in \mathbb{Z}$, then $\mathbb{T} = g(\mathbb{T}_0)$ has cusps $g(\infty) = \frac{a}{c}$, $g(0) = \frac{b}{d}$ and $g(1) = \frac{a+b}{c+d}$. That is, the cusps of $g(\mathbb{T}_0)$ are the images under g of the the cusps of $\mathbb{T}_0$. More generally, we have the following result.

Theorem 3.6 Extended modular transformations map Farey triangles to Farey triangles.

Proof: Let $\mathbb{T}$ be any Farey triangle. Then $\mathbb{T} = g(\mathbb{T}_0)$ for some $g \in \tilde{\Gamma}$ and so $\mathbb{T}_0 = g^{-1}(\mathbb{T})$. Now let $f \in \tilde{\Gamma}$. Then $fg \in \tilde{\Gamma}$ and hence $fg(\mathbb{T}_0) = f(\mathbb{T})$ is also a Farey triangle. ■

Theorem 3.7 Let $\mathbb{T} = \left\{\frac{a}{c}; \frac{e}{f}; \frac{b}{d}\right\}$ be any Farey triangle such that $\frac{e}{f}$ lies between $\frac{a}{c}$ and $\frac{b}{d}$. Then we have the following:

1. $\mathbb{T} = g(\mathbb{T}_0)$ where $g(z) = \frac{az+b}{cz+d}$, $g \in \tilde{\Gamma}$ and $\frac{e}{f} = \frac{a}{c} \oplus \frac{b}{d}$. Hence, if $\frac{a}{c} < \frac{b}{d}$ then $\frac{a}{c} < \frac{a+b}{c+d} < \frac{b}{d}$. If $\frac{b}{d} < \frac{a}{c}$ then $\frac{b}{d} < \frac{a+b}{c+d} < \frac{a}{c}$.

2. An ideal triangle is a Farey triangle if and only if it is bounded by three Farey geodesics.

3. Two Farey triangles are either equal or disjoint.

Proof:

1. Let $g(z) = \frac{az+b}{cz+d}$ such that $g \in \tilde{\Gamma}$. Then $g(0) = \frac{b}{d}$, $g(\infty) = \frac{a}{c}$ and $g(1) = \frac{a+b}{c+d}$, so $\frac{e}{f} = \frac{a+b}{c+d} = \frac{a}{c} \oplus \frac{b}{d}$.
   Let $f \in \tilde{\Gamma}$. Then $fg \in \tilde{\Gamma}$ and $fg(\mathbb{T}_0)$ is a Farey triangle.
   $fg(0) = f\left(\frac{b}{d}\right)$, $fg(\infty) = f\left(\frac{a}{c}\right)$ and $fg(1) = f\left(\frac{a+b}{c+d}\right) = f\left(\frac{a}{c} \oplus \frac{b}{d}\right)$.
   Thus $fg(1) = fg(0) \oplus fg(\infty)$, or $f\left(\frac{a}{c} \oplus \frac{b}{d}\right) = f\left(\frac{a}{c}\right) \oplus f\left(\frac{b}{d}\right)$.
   That is, if $\mathbb{T} = g(\mathbb{T}_0)$, then $g\left(\frac{0}{1}\right) \oplus g\left(\frac{1}{0}\right) = g\left(\frac{1}{1}\right)$. Further, if $f \in \tilde{\Gamma}$ we have $f\left(\frac{a}{c} \oplus \frac{b}{d}\right) = f\left(\frac{a}{c}\right) \oplus f\left(\frac{b}{d}\right)$.

2. If $\mathbb{T}$ is a Farey triangle then, by definition of a Farey triangle, $\mathbb{T}_0$ must be bounded by three Farey triangles. Conversely, let $\mathbb{T} \neq \mathbb{T}_0$ be an ideal triangle bounded by three Farey geodesics. Let one of these geodesics be $g(\mathbb{I})$ for some $g \in \tilde{\Gamma}$. Then $g^{-1}(\mathbb{T})$ is an ideal triangle with all of its sides being Farey geodesics, one of which is $\mathbb{I}$. But $\mathbb{T}_0$ and $\{-1, 0, \infty\} = \tau^{-1}(\mathbb{T}_0)$ are the only ideal triangles which are bounded by $\mathbb{I}$. Hence $\mathbb{T} = g(\mathbb{T}_0)$ or $\mathbb{T} = g\tau^{-1}(\mathbb{T}_0)$ and so $\mathbb{T}$ is an image of $\mathbb{T}_0$ under an extended modular map, and so $\mathbb{T}$ is a Farey triangle.

3. We have shown that Farey geodesics cannot cross each other. Hence a Farey triangle cannot meet any Farey geodesic. Let $\mathbb{T}_1$ and $\mathbb{T}_2$ be two distinct Farey triangles. Thus $\mathbb{T}_2$ cannot meet any of the three

geodesics that bound $\mathbb{T}_1$. The complement, in $\mathbb{H}^{\perp}$, of the three geodesics that bound $\mathbb{T}_1$ has four components, one of which is $\mathbb{T}_1$. Since $\mathbb{T}_2$ is connected, we have that either $\mathbb{T}_1 \cap \mathbb{T}_2 = \Phi$ or $\mathbb{T}_2 \subset \mathbb{T}_1$. Similarly, since $\mathbb{T}_1$ is connected we must have that $\mathbb{T}_2 \cap \mathbb{T}_1 = \Phi$ or $\mathbb{T}_1 \subset \mathbb{T}_2$. Since $\mathbb{T}_1$ and $\mathbb{T}_2$ are distinct, we have $\mathbb{T}_1 \cap \mathbb{T}_2 = \Phi$.

■

While each Farey triangle $\mathbb{T}$ can be written as $g(\mathbb{T}_0)$ for some $g \in \tilde{\Gamma}$, there are elements of $\tilde{\Gamma}$ that leave $\mathbb{T}_0$ invariant. It is useful to find the elements of $\tilde{\Gamma}$ that leave $\mathbb{T}_0$ fixed, as this leads us to the stabilizer of any Farey triangle $\mathbb{T}$ under Γ and $\tilde{\Gamma}$.

Theorem 3.8 Γ and $\tilde{\Gamma}$ act on $\mathbb{T}_0$ in the following way:

1. $\Gamma_{\mathbb{T}_0} = \{g \in \Gamma : g(\mathbb{T}_0) = \mathbb{T}_0\} = \langle h \rangle$, where $h(z) = \varphi\tau^{-1}(z)$ and $h^3 = 1_{\mathcal{M}}$.
2. $\tilde{\Gamma}_{\mathbb{T}_0} = \{g \in \tilde{\Gamma} : g(\mathbb{T}_0) = \mathbb{T}_0\} = \langle h, \psi \rangle$, where $h(z) = \varphi\tau^{-1}(z)$ and $\psi(z) = \frac{1}{z}$.
3. If $H = \langle h \rangle$, then
$$\Gamma = \bigcup_{i=0}^{\infty} h_i H = h_0 H \cup h_1 H \cup h_2 H ...$$
is the coset decomposition of Γ with respect to H, where $h_0 = 1_{\mathcal{M}}$ and $h_i \in \Gamma \setminus H$ for $i \geq 1$, and where $h_i H \cap h_j H = \Phi$ for $i \neq j$.

Proof:

1. $h(z) = \varphi\tau^{-1}(z) = \varphi(z-1) = \frac{-1}{z-1} = \frac{1}{1-z}$. Thus $h(0) = 1$, $h(1) = \infty$ and $h(\infty) = 0$. Thus h leaves the set of cusps of $\mathbb{T}_0$ invariant. So

$h(\mathbb{T}_0) = \mathbb{T}_0$. Further, $h^2 = h^{-1}$ and $h^3 = 1_{\mathcal{M}}$ so that $H = \langle h \rangle$ is a cyclic subgroup of order 3 in Γ. Hence $\langle h \rangle \subseteq \Gamma_{\mathbb{T}_0}$.

Now suppose $g(\mathbb{T}_0) = \mathbb{T}_0$ for some $g \in \Gamma$. If $g \notin H$ then g must interchange two cusps, leaving one fixed. Say $g(0) = \infty$, $g(\infty) = 0$ and $g(1) = 1$. So we must have that $g = \psi$, but $\psi \notin \Gamma$, and so we must have that $g \in H$. Hence $\Gamma_{\mathbb{T}_0} \subseteq H$ and thus $\Gamma_{\mathbb{T}_0} = H = \langle h \rangle$. Similarly, if $g(1) = \infty$, $g(\infty) = 1$ and $g(0) = 0$ then we obtain that $g = \psi h^{-1} = \psi h^2$. If $g(0) = 1$, $g(1) = 0$ and $g(\infty) = \infty$ then we obtain $g = \psi h$.

2. $\Gamma \subset \tilde{\Gamma}$ so $\Gamma_{\mathbb{T}_0} \subseteq \tilde{\Gamma}_{\mathbb{T}_0}$. From part (1) above we have that if $g(\mathbb{T}_0) = \mathbb{T}_0$ and $g \in \tilde{\Gamma} \setminus H$, then $g = \psi$, so ψh and ψh^2 also leave $\mathbb{T}_0$ fixed. Hence $\tilde{\Gamma}_{\mathbb{T}_0} = \langle \psi, h \rangle$.

3. H is a proper subgroup of Γ, so we can partition Γ into disjoint cosets modulo H. That is, we can find $h_0, h_1, h_2, ...$, where $h_0 = 1_{\mathcal{M}}$ and $h_i \notin H$ for $i \geq 1$, such that
$$\Gamma = \bigcup_{i=0}^{\infty} h_i H = h_0 H \cup h_1 H \cup h_2 H ...$$

■

Recall that if an open subset D of X is a fundamental domain for a group G, then we say that the collection of sets $\{g(\bar{D}) : g \in G\}$ is a *tessellation* of X.

Theorem 3.9 $\mathbb{H}^{\perp}$ is tessellated by the sets
$$\left\{ h_i(\bar{\mathbb{T}}_0) : \Gamma = \bigcup_{i=0}^{\infty} h_i H \right\}$$
.

Proof: The coset decomposition of Γ by H (where H is as defined in Theorem 3.8) is into distinct, disjoint cosets modulo H. So if $i \neq j$ then $h_i H \cap h_j H = \Phi$.

Suppose $h_i(\mathbb{T}_0) \cap h_j(\mathbb{T}_0) \neq \Phi$, where $h_i, h_j \in \Gamma$ and $i \neq j$. Then $h_i(\mathbb{T}_0)$ and $h_j(\mathbb{T}_0)$ are Farey triangles and are therefore equal or disjoint. So $h_i(\mathbb{T}_0) = h_j(\mathbb{T}_0)$ and so $h_i^{-1}h_j(\mathbb{T}_0) = \mathbb{T}_0$. Thus $h_i^{-1}h_j \in \langle h \rangle = H$ and $h_iH = h_jH$, which is a contradiction. Hence $h_i(\mathbb{T}_0) \cap h_j(\mathbb{T}_0) = \Phi$ for $i \neq j$. Hence

$$\left\{ h_i(\mathbb{T}_0) : \Gamma = \bigcup_{i=0}^{\infty} h_iH \right\}$$

is a collection of disjoint subsets of $\mathbb{H}^{\perp}$.

It remains to show that the sets $h_j(\bar{\mathbb{T}}_0)$, $j = 0, 1, 2...$, cover $\mathbb{H}^{\perp}$. Let

$$\Sigma = \bar{\mathbb{T}}_0 \cup \tau^{-1}(\bar{\mathbb{T}}_0).$$

Recall from Theorem 1.48 that the fundamental region $\mathcal{D}$ of Γ is given by

$$\mathcal{D} = \left\{ z \in \mathbb{H}^{\perp} : |\Re(z)| < \frac{1}{2}, |z| > 1 \right\}.$$

Thus $\bar{\mathcal{D}} \subset \Sigma$. Also recall that $\{g(\bar{\mathcal{D}}) : g \in \Gamma\}$ tessellates $\mathbb{H}^{\perp}$. Hence every $z \in \mathbb{H}^{\perp}$ lies in some Γ-image of $\bar{\mathcal{D}}$ and hence in some Γ-image of $\bar{\mathbb{T}}_0$, since $\tau^{-1} \in \Gamma$ and $\bar{\mathcal{D}} \subset \bar{\mathbb{T}}_0 \cup \tau^{-1}(\bar{\mathbb{T}}_0)$. But the collection of Γ-images of $\bar{\mathcal{D}}$ is the same as the collection $\{h_j(\bar{\mathbb{T}}_0) : j = 0, 1, 2, ...\}$. That is, $z \in \mathbb{H}^{\perp}$ implies that $z \in g(\bar{\mathcal{D}})$ which implies $z \in g(\Sigma)$ which implies $z \in g'(\bar{\mathbb{T}}_0)$ for some $g' \in \Gamma$. Now $g' \in \Gamma$ so $g' \in h_iH$ for some $i = 0, 1, 2, ...$ That is, $g' = h_ih^r$ for some $r \in \mathbb{Z}$. Therefore $z \in h_ih^r(\bar{\mathbb{T}}_0) = h_i(\bar{\mathbb{T}}_0)$. Thus

$$\mathbb{H}^{\perp} \subseteq \bigcup_{i=0}^{\infty} h_i(\bar{\mathbb{T}}_0) \subseteq \mathbb{H}^{\perp}$$

and this completes the proof. ■

Definition 3.10 The tessellation of $\mathbb{H}^{\perp}$ by Farey triangles, as described above, is called the *Farey tessellation* (Figure 4) of $\mathbb{H}^{\perp}$ and is denoted by

$$\mathcal{F} = \left\{ h_i(\bar{\mathbb{T}}_0) : \Gamma = \bigcup_{i=0}^{\infty} h_iH \right\}.$$

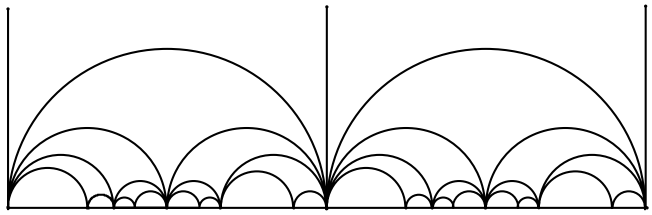

Figure 4: Part of the Farey Tessellation

Theorem 3.11 $\mathcal{F}$ is invariant under Γ.

Proof: We need to show that

$$\Gamma = \bigcup_{i=0}^{\infty} gh_iH$$

for all $g \in \Gamma$ and thus that

$$\mathcal{F} = \left\{ gh_j(\bar{\mathbb{T}}_0) : \Gamma = \bigcup_{i=0}^{\infty} gh_iH \right\}.$$

That is, we need to show that Γ is the disjoint union of the cosets gh_iH, $i = 0, 1, 2, ...$ Suppose $gh_iH \cap gh_jH \neq \Phi$ where $i \neq j$. Then for $t, p \in [0, 2]$ we have $gh_ih^t = gh_jh^p$ where $h \in H$. This implies that $h_ih^t = h_jh^p$, which implies that $h^th^{-p} = h_i^{-1}h_j$. But this implies that $h_i^{-1}h_j \in H$ and so $h_iH = h_jH$, which is impossible. Hence we must have $gh_iH \cap gh_jH = \Phi$. Since

$gh_i \in \Gamma$, we have $gh_i \in h_k H$ for some k, and so $gh_i H \subseteq h_k H$. Further, we have $|gh_i H| = |h_k H| = 3$ and so $gh_i H = h_k H$. Thus

$$\mathcal{F} = \left\{ gh_i(\bar{\mathbb{T}}_0) : \Gamma = \bigcup_{i=0}^{\infty} gh_i H \right\}.$$

■

3.4 Cutting Sequences and Simple Continued Fractions

3.4.1 Introduction

Following Series ([8]), we consider the Farey tessellation $\mathcal{F}$ that is cut by a directed geodesic σ giving rise to a sequence of cut Farey triangles.

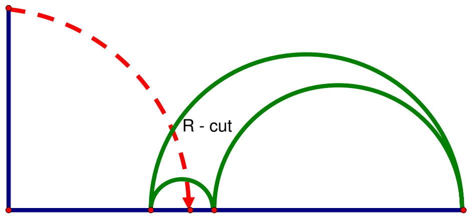

Figure 5: A Right Cut

We define the different cuts of the Farey triangles as follows: if σ cuts $\mathbb{T}_j$ so that one cusp of $\mathbb{T}_j$ lies to the right of the direction of σ, then we label the *segment* of σ that lies in $\mathbb{T}_j$ as a "right cut" or "R" of $\mathbb{T}_j$ (Figure 5).

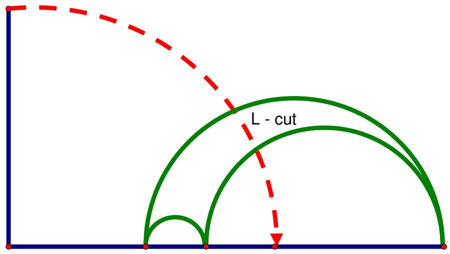

Figure 6: A Left Cut

If σ cuts $\mathbb{T}_j$ so that one cusp lies to the left of the direction of σ, then we label the segment of σ that lies in $\mathbb{T}_j$ as a "left cut" or "L" of $\mathbb{T}_j$ (Figure 6).

Accumulating the L and R labels, we obtain a sequence $\cdots R^{n_{-1}}L^{n_0}R^{n_1}L^{n_2}\cdots$ where the exponent n_i denotes a succession of n_i left or right cuts, and we assign n_0 to the succession of cuts that begins with $\mathbb{T}_0$. Series calls this sequence the *cutting sequence* of σ. She then shows that if σ cuts $\mathbb{R}$ at x' and x, where $|x| \geq 1$ and $0 < |x'| \leq 1$, then the simple continued fraction expansions of x' and x satisfy $x = [n_1, n_2, \cdots]$ and $\varphi(x') = \frac{-1}{x'} = [n_0, n_{-1}, n_{-2}, \cdots]$.

If $x \in \mathbb{Q}^+$ then x is a cusp of a Farey triangle, and so σ cuts through a finite number of Farey triangles to produce a finite cutting sequence. For a finite cutting sequence, we adopt the convention that the last cut (that is, the segment of σ that ends at x) is given the opposite label to second-last cut. This is consistent with our convention that the final partial quotient of a finite simple continued fraction must be 1. If $x \in \mathbb{R}^+ \setminus \mathbb{Q}$ then x is not a cusp of any Farey triangle, and so σ must cut through an infinite number of Farey triangles before cutting the real line at x, thereby producing a semi-infinite cutting sequence.

In this section, we show how the Farey tessellation $\mathcal{F}$ can be interpreted as a graph $\mathcal{G}$ and that the cutting sequence can be interpreted as a path on $\mathcal{G}$. In particular, we show that $\mathcal{G}$ is a tree and that these paths are thus unique. In what follows, we will show that each rational number will give rise to a unique finite path on $\mathcal{G}$ while each irrational number will give rise to a unique semi-infinite path on $\mathcal{G}$. We first introduce some basic concepts in graph theory, taken from [18] and [19].

Formally, a *graph* $G = (V, E)$ is comprised of a non-empty set V of elements called *vertices*, together with a set E of *edges*, where each edge is an unordered pair of elements from V. A graph is *infinite* if the vertex set V is infinite or if the set of edges E is infinite.

Two vertices that are joined by an edge are said to be *adjacent*, as are two edges that meet at a vertex. An edge between vertices u and v is said to have u (or v) as an *end vertex*, and the edge is said to be incident with v (or u). The number of edges incident with a vertex v in a graph G is called the *degree* of v and is denoted by *degv* or *deg*$_G v$. A graph in which every vertex has degree r is called an *r-regular* graph.

A *path* in a graph G is an ordered set $\{v_0, v_1, v_2, ...\}$ of elements of V such that v_i is adjacent to v_{i+1} for $i = 0, 1, 2, ...$ The length of a path is the number of edges in the path. A path that starts at $u \in V$ and ends at $v \in V$ is called a $(u, v) - path$ in G. A graph G is called *connected* if for each $u, v \in V$ there is a $(u, v) - path$ in G. A *closed* path is a path which begins and ends at the same vertex. A *cycle* is a closed path that contains no repeated vertices other than the beginning and end vertex. A connected graph that contains no cycles is called a *tree*. In particular, a graph G is a *tree* if and only if every two distinct vertices of G are joined by a unique path.

3.4.2 The $\rho - \tau$ Farey Tree

Recall that $\mathbb{T}_0$ is the Farey triangle bounded by $\mathbb{I}$, $\left[\frac{1}{1} : \frac{1}{0}\right]$ and $\left[\frac{0}{1} : \frac{1}{1}\right]$ with cusps at 0,1 and ∞. Also recall that every Farey triangle is in the orbit of $\mathbb{T}_0$ under $\tilde{\Gamma} = \langle \tau, \psi \rangle$, where $\tilde{\Gamma}$ acts on $\mathbb{H}^{\perp}$.

Definition 3.12 Two Farey triangles are said to be *adjacent* if they are bounded by a common Farey geodesic, where $\mathbb{T}_i \bowtie \mathbb{T}_j$ denotes $\mathbb{T}_i$ is adjacent to $\mathbb{T}_j$.

Theorem 3.13 Every Farey triangle is adjacent to exactly three other Farey triangles.

Proof: Let $\mathbb{T}_j$ be any Farey triangle. Then $\mathbb{T}_j = g(\mathbb{T}_0)$ for some $g \in \tilde{\Gamma}$ and so $\mathbb{T}_0 = g^{-1}(\mathbb{T}_j)$. Clearly $\mathbb{T}_0$ is adjacent to exactly three other Farey triangles, namely $\{-1; 0; \infty\}$, $\{1; 2; \infty\}$ and $\{0; \frac{1}{2}; 1\}$, or $\tau^{-1}(\mathbb{T}_0)$=$\rho^{-1}(\mathbb{T}_0)$, $\tau(\mathbb{T}_0)$ and $\rho(\mathbb{T}_0) = \psi\tau\psi(\mathbb{T}_0)$, because $\mathbb{T}_0$ is adjacent to these three other Farey triangles and cannot be adjacent to any others because it only has three sides. Thus $g(\mathbb{T}_0)$ is adjacent to $g\tau(\mathbb{T}_0)$, $g\tau^{-1}(\mathbb{T}_0)$ and $g\rho(\mathbb{T}_0)$. Hence every Farey triangle is adjacent to exactly three other Farey triangles. ■

We now construct an infinite graph $\mathcal{G}$ of Farey triangles. Let the set $\mathcal{V}$ of vertices be the set of all Farey triangles, and let the set $\mathcal{E}$ of edges be the set of all pairs of adjacent Farey triangles. That is, $\mathcal{V} = \{g(\mathbb{T}_0) : g \in \tilde{\Gamma}\}$ and $\mathcal{E} = \{(\mathbb{T}_i, \mathbb{T}_j) : \mathbb{T}_i \bowtie \mathbb{T}_j\}$.

In Theorem 3.13 we showed that every Farey triangle is adjacent to exactly three Farey triangles, so every vertex in $\mathcal{G}$ has degree 3. A *path* on $\mathcal{G}$ is thus a chain $\mathbb{T}_i, \mathbb{T}_{i+1}, ...$ of adjacent Farey triangles where $\mathbb{T}_i$ is adjacent to $\mathbb{T}_{i+1}$ for all i. In what follows, we will show that $\mathcal{G}$ is a tree. That is, we will show that $\mathcal{G}$ is connected and contains no cycles. We will first prove that $\mathcal{G}$ is connected.

Theorem 3.14 $\mathcal{G}$ is connected and every vertex $\mathbb{T}$ of $\mathcal{G}$ can be expressed as $\mathbb{T} = \tau^{b_0}\rho^{b_1}\tau^{b_2}\rho^{b_3}\cdots\tau^{b_n}(\mathbb{T}_0)$ where $b_0 \in \mathbb{Z}$ and $b_i \in \mathbb{Z}^+$ for all $i \geq 1$, except b_n which may be zero. Further, since $\psi(\mathbb{T}_0) = \mathbb{T}_0$ and $s_{b_i}(z) = \tau^{b_i}\psi(z)$, we have that each vertex can be written as $s_{b_0}s_{b_1}\cdots s_{b_n}(\mathbb{T}_0)$.

Proof: We have shown in Theorem 3.9 that the Farey triangles tessellate $\mathbb{H}^2$. Thus each Farey triangle is connected to every other Farey triangle by a chain of adjacent Farey triangles and $\mathcal{G}$ is thus connected. In particular, $\mathbb{T}_0$ is connected to each vertex in $\mathcal{G}$. Consider the sequence of adjacent vertices $\mathbb{T}_0, \mathbb{T}_1, ..., \mathbb{T}_j$ in $\mathcal{G}$, connecting $\mathbb{T}_0$ to $\mathbb{T}_j$.

If $\mathbb{T}_j = \mathbb{T}_1$, then $\mathbb{T}_j$ is adjacent to $\mathbb{T}_0$. Thus $\mathbb{T}_j = \tau^{-1}(\mathbb{T}_0)$, $\mathbb{T}_j = \tau(\mathbb{T}_0)$ or $\mathbb{T}_j = \rho(\mathbb{T}_0)$. Thus $b_0 = -1$ or 1 or 0 and the result holds.

Assume the result holds for $m < k$. That is, assume $\mathbb{T}_m = \tau^{b_0}\rho^{b_1}\cdots\tau^{b_k}(\mathbb{T}_0) = g(\mathbb{T}_0)$ where $b_0 \in \mathbb{Z}$ and $b_i \in \mathbb{Z}^+$ for $i \geq 1$, except b_k which may be zero. Since

$\mathbb{T}_{m+1}$ is adjacent to $\mathbb{T}_m$, we have that $\mathbb{T}_{m+1} = g\tau(\mathbb{T}_0)$ or $g\rho(\mathbb{T}_0)$ or $g\tau^{-1}(\mathbb{T}_0)$, by Theorem 3.13. Thus the result holds for $\mathbb{T}_{m+1}$ and hence for all k.

We note that since $\rho(z) = \psi\tau\psi(z)$, we can write each vertex $\mathbb{T}_m \in \mathcal{V}$ as $\mathbb{T}_m = \tau^{b_0}\psi\tau^{b_1}\psi\cdots\tau^{b_k}(\mathbb{T}_0)$, or $\tau^{b_0}\psi\tau^{b_1}\psi\cdots\tau^{b_k}\psi(\mathbb{T}_0)$ since $\psi(\mathbb{T}_0) = \mathbb{T}_0$. Thus we have $\mathbb{T}_m = s_{b_0}s_{b_1}\cdots s_{b_k}(\mathbb{T}_0)$. ■

From now on, we may refer to $\mathcal{G}$ as the $\rho - \tau$ Farey graph.

Theorem 3.15 $\mathcal{G}$ contains no cycles.

Proof: If we remove $\mathbb{T}_0$ then the graph is disconnected. Since the graph is homogeneous, if we remove any vertex then the graph is disconnected. But if there is a cycle, then removing a single vertex will not disconnect the graph. Hence $\mathcal{G}$ contains no cycles. ■

From Theorems 3.14 and 3.15, we conclude that $\mathcal{G}$ is a tree (Figure 7).

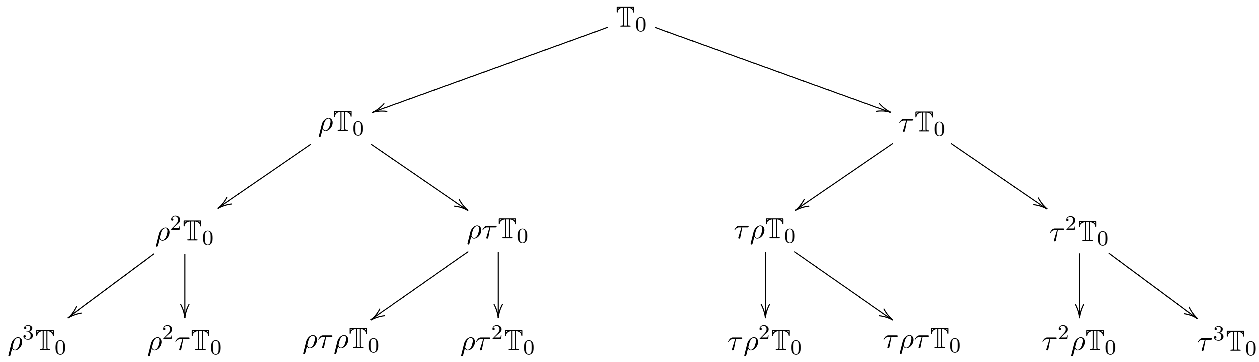

Figure 7: Part of the $\rho - \tau$ Farey Tree

Definition 3.16 On any given path on $\mathcal{G}$, we distinguish the vertices at which the direction of this path changes, and we call these vertices the *nodes* of the path.

The nodes are the vertices at which a succession of τ's ends and a succession of ρ's begins, or vice versa. For example, if $\mathbb{T}_j = \tau^{b_0}\rho^{b_1}\cdots\tau^{b_k}(\mathbb{T}_0)$ is a vertex on a path on which the next vertex is $\tau^{b_0}\rho^{b_1}\cdots\tau^{b_k}\rho(\mathbb{T}_0)$, then $\mathbb{T}_j$ is a node of this path.

Definition 3.17 We define a $\mathbb{T}_0$-path on $\mathcal{G}$ as a path that begins at the vertex $\mathbb{T}_0$.

The nodes of a $\mathbb{T}_0$-path, the j-th node of which is given by
$\mathbb{T}_j = \tau^{b_0}\rho^{b_1}\cdots\tau^{b_k}(\mathbb{T}_0) = S_{[b_k]}(\mathbb{T}_0)$, are thus given by
$S_{[b_0]}(\mathbb{T}_0), S_{[b_1]}(\mathbb{T}_0), S_{[b_2]}(\mathbb{T}_0), ..., S_{[b_k]}(\mathbb{T}_0), ...$

The cusps of the node $S_{[n_k]}(\mathbb{T}_0)$ lie in the interval $[S_{[n_k]}(0), S_{[n_k]}(\infty)]$. A $\mathbb{T}_0$-path may be finite or semi-infinite. We note that there appears to be a direct relationship between simple continued fraction expansions and the chains of nodes on a $\mathbb{T}_0$-path on $\mathcal{G}$. This relationship will be explored in the sections that follow.

3.4.3 Cutting Sequences as $\mathbb{T}_0$-paths on $\mathcal{G}$

The geodesic segment σ, starting at a point inside $\mathbb{T}_0$ and cutting the real line at a point x, is divided into subsegments as it cuts across $\mathcal{F}$. Hence we can regard $\sigma = \sigma_0 \cup \sigma_1 \cup \sigma_2 \cdots$, where the subsegment σ_i can be identified as the intersection of σ with the Farey triangle $\mathbb{T}_i$ which is the i-th triangle cut by σ. Since the path of σ links adjacent triangles, the vertices $\{V_i\}_{i=0,1,2,\cdots}$ of $\mathcal{G}$, where $V_i = \mathbb{T}_i$, form a $\mathbb{T}_0$-path and hence a chain of nodes on $\mathcal{G}$.

Thus σ produces a unique ordered path $\{\mathbb{T}_i\}_{i=0,1,2\cdots}$ on $\mathcal{G}$. We define the *cutting sequence* of σ as this unique $\mathbb{T}_0$-path, which may be finite or semi-infinite. If σ ends at a rational x then the $\mathbb{T}_0$-path is finite. If x is irrational then the $\mathbb{T}_0$-path is semi-infinite.

3.4.4 Finite $\mathbb{T}_0$-paths on $\mathcal{G}$

Consider the geodesic segment σ starting at a point in $\mathbb{T}_0$ and ending at a rational point x on $\mathbb{R}$. Assume that the simple continued fraction expansion of x is given by $S_{[a_t]}s_1(\infty) = s_{a_0}s_{a_1}\cdots s_{a_t}s_1(\infty)$ with $a_0 \in \mathbb{Z}, a_i \in \mathbb{Z}^+$ for $i \geq 1$ and where the convergents of x are

$$S_{[a_0]}(\infty), S_{[a_1]}(\infty), S_{[a_2]}(\infty), \cdots, S_{[a_t]}(\infty), S_{[a_t]}s_1(\infty) = x.$$

We note that $x \in [a_0, a_0+1]$ where a_0 is the integer part of x. Recall that we have adopted the convention that the last partial quotient of the simple continued fraction expansion of a rational must be 1. Thus $a_{t+1} = 1$ and $x = S_{[a_t]}(1) = S_{[a_t]}s_1(\infty) = S_{[a_{t+1}]}(\infty)$.

The segment σ will cross a finite number of Farey triangles in $\mathcal{F}$ and will thus describe a finite $\mathbb{T}_0$-path. The nodes on the path will be of the form $S_{[b_0]}(\mathbb{T}_0)$, $S_{[b_1]}(\mathbb{T}_0)$, $S_{[b_2]}(\mathbb{T}_0)$, $\cdots$, $S_{[b_k]}(\mathbb{T}_0)$, where $b_0 \in \mathbb{Z}$ and $b_i \in \mathbb{Z}^+$ for $i \geq 1$. Note that $S_{[b_k]}(\mathbb{T}_0)$ is the last vertex on this $\mathbb{T}_0$-path and that $S_{[b_k]}(1) = x$ is a cusp of this vertex. Thus $x = S_{[a_t]}(1) = S_{[b_k]}(1)$.

Since we know that the simple continued fraction expansion of a rational is unique if the final partial quotient is 1, we have $t = k$ and $a_i = b_i$ for all $i = 0, 1, 2, \cdots, k$. Hence the finite $\mathbb{T}_0$-path (corresponding to σ) on $\mathcal{G}$ has nodes $S_{[b_0]}(\mathbb{T}_0), S_{[b_1]}(\mathbb{T}_0), S_{[b_2]}(\mathbb{T}_0), \cdots, S_{[b_k]}(\mathbb{T}_0)$, where the convergents of x are $S_{[b_0]}(\infty), S_{[b_1]}(\infty), S_{[b_2]}(\infty), \cdots, S_{[b_k]}(\infty)$ and $S_{[b_k]}(1) = x$.

Thus we have the following result:

Theorem 3.18 The $\mathbb{T}_0$-path of a geodesic segment σ, which starts in $\mathbb{T}_0$ and ends at a rational point x, has nodes that correspond to the convergents of the simple continued fraction expansion of x. That is, $S_{[b_k]}(\mathbb{T}_0)$ is a node of the $\mathbb{T}_0$-path of the geodesic segment σ that ends at x if and only if $S_{[b_k]}(\infty)$ is a convergent of the simple continued fraction expansion of x.

3.4.5 Convergence of semi-infinite $\mathbb{T}_0$-paths on $\mathcal{G}$

Definition 3.19 If the nodes of a semi-infinite $\mathbb{T}_0$-path are given by $S_{[b_0]}(\mathbb{T}_0), S_{[b_1]}(\mathbb{T}_0), S_{[b_2]}(\mathbb{T}_0), ..., S_{[b_k]}(\mathbb{T}_0), ...$ and if the path ends at the point x on $\mathbb{R}$ where

$$x = \lim_{k\to\infty} S_{[b_k]}(\infty) = \lim_{k\to\infty} S_{[b_k]}(0) = \lim_{k\to\infty} S_{[b_k]}(1)$$

then we say that the $\mathbb{T}_0$-path *converges* to x. Here $S_{[b_k]}(\mathbb{T}_0)$ is the k-th node of the $\mathbb{T}_0$-path.

Let x be any irrational number with simple continued fraction expansion given by the convergents $S_{[a_0]}(\infty), S_{[a_1]}(\infty), ...$ where $S_{[a_t]}(\infty) \to x$ as $t \to \infty$. Thus $x \in [a_0, a_0 + 1]$. Let σ be a geodesic segment that starts in $\mathbb{T}_0$ and ends at $x \in \mathbb{R} \setminus \mathbb{Q}$. This segment will trace out a semi-infinite $\mathbb{T}_0$-path on $\mathcal{G}$. Let the nodes of this $\mathbb{T}_0$-path be $S_{[b_0]}(\mathbb{T}_0), S_{[b_1]}(\mathbb{T}_0), S_{[b_2]}(\mathbb{T}_0), ...$ where $b_0 \in \mathbb{Z}$ and $b_i \in$

$\mathbb{Z}^+$ for $i \geq 1$. For any k, the node $S_{[b_k]}(\mathbb{T}_0)$ has cusps $S_{[b_k]}(0), S_{[b_k]}(\infty), S_{[b_k]}(1)$ and x lies in the interval spanned by the cusps $S_{[b_k]}(0)$ and $S_{[b_k]}(\infty)$. Since

$$\lim_{k\to\infty} S_{[b_k]}(\infty) = \lim_{k\to\infty} S_{[b_k]}(0) = \lim_{k\to\infty} S_{[b_k]}(1)$$

we have that

$$\lim_{k\to\infty} S_{[b_k]}(\infty) = x.$$

Thus $b_i = a_i$ for all i since a real irrational number has a unique simple continued fraction expansion. Thus we have the following result:

Theorem 3.20 The $\mathbb{T}_0$-path of geodesic segment σ, which starts in $\mathbb{T}_0$ and ends at an irrational point x, has nodes that correspond to the convergents of the simple continued fraction expansion of x. That is, $S_{[b_k]}(\mathbb{T}_0)$ is a node of the $\mathbb{T}_0$-path of the geodesic segment σ that ends at x if and only if $S_{[b_k]}(\infty)$ is a convergent of the simple continued fraction expansion of x.

3.5 Ford Circles

We include this section on Ford circles because we need a result from Ford circles to prove a result in the next chapter of this text. The proofs of some of the results in this section are omitted because they can be found in [4].

In what follows, we will assume that $\frac{a}{c}$ is reduced and we will represent ∞ as $\frac{1}{0}$.

Definition 3.21 The *Ford circle* at $\frac{a}{c}$, denoted by $\mathcal{C}_{\frac{a}{c}}$, is the circle which has a radius of $\frac{1}{2c^2}$, lies above the real axis and is tangent to the real axis at $\frac{a}{c}$. The Ford circle $\mathcal{C}_\infty = \mathcal{C}_{\frac{1}{0}}$ is the line $\{z \in \mathbb{C} : \Im(z) = 1\} \cup \{\infty\}$ and is called the *fundamental Ford circle.* That is, $\mathcal{C}_\infty$ is the Ford circle that is tangential to $\mathbb{R}_\infty$ at ∞. We denote the set of all Ford circles by $\mathfrak{F}$.

It follows that the integers are represented by circles of radius $\frac{1}{2}$, and that any interval of the real axis contains infinitely many points of tangency of Ford circles.

Theorem 3.22 Two distinct Ford circles are either tangent to each other or completely external to each other. ([4])

Definition 3.23 Two fractions $\frac{a}{c}$ and $\frac{b}{d}$ are *adjacent* if their representative Ford circles are tangent to each other.

In [4], Ford describes the parts of $\mathbb{H}^\perp$ that are exterior to the Ford circles as consisting of an infinite number of circular arc triangles which he calls *mesh triangles.* Any two sides of a mesh triangle lie on circles corresponding to adjacent fractions. Consider a vertical geodesic L in $\mathbb{H}^\perp$ given by $\Re(z) = \omega$ and consider the fractions whose circles are passed through in succession by L. Each circle $\mathcal{C}_{\frac{a}{c}}$ is surrounded by mesh triangles. If L passes from $\mathcal{C}_{\frac{a}{c}}$ into one of these mesh triangles and if L does not return to $\mathcal{C}_{\frac{a}{c}}$, then it will on leaving the mesh triangle pass in general into a circle tangent to $\mathcal{C}_{\frac{a}{c}}$ because a mesh triangle is the space between three adjacent Ford circles. We adopt the convention that if L touches two circles at their point of tangency without entering either circle, then we will only consider one of these circles as crossed by L. We thus have from [4] the following principle.

Principle

If two circles in the system of Ford circles are penetrated in succession by a vertical geodesic L, then the two corresponding fractions are adjacent.

Theorem 3.24

1. Each fraction $\frac{a}{c}$ has an adjacent fraction. ([4])

2. $\mathfrak{F} = \{g(\mathcal{C}_\infty) : g \in \tilde{\Gamma}\}$, and either $\mathcal{C}_{\frac{a}{c}} \cap \mathcal{C}_{\frac{b}{d}} = \Phi$ or $\mathcal{C}_{\frac{a}{c}} \cap \mathcal{C}_{\frac{b}{d}} = \{g(i)\}$ where $g(z) = \frac{az+b}{cz+d}$ and $g \in \tilde{\Gamma}$. Further, $\Re(g(i)) \in \mathbb{Q}$.

3. If $\mathcal{C}_{\frac{a}{c}}$ is tangential to $\mathcal{C}_{\frac{b}{d}}$ then $\mathcal{C}_{\frac{b_n}{d_n}}$ is tangential to $\mathcal{C}_{\frac{a}{c}}$ if and only if $\frac{b_n}{d_n} = \frac{b+na}{d+nc}$ for $n \in \mathbb{Z}$. ([4])

4. Of the family of circles tangential to $\mathcal{C}_{\frac{a}{c}}$, exactly two correspond to fractions that have denominators numerically smaller than c ([4]). A vertical line $\Re(z) = \omega$ will cut a finite number of Ford circles if and only if $\omega \in \mathbb{Q}$. The vertical line $\Re(z) = \omega$, where $\omega \in \mathbb{R} \setminus \mathbb{Q}$, will cut infinitely many Ford circles.

5. Let ω be irrational. Apart from $\mathcal{C}_\infty$, all the Ford circles cut by the vertical line $\Re(z) = \omega$ have points of tangency (with the x axis) in the same unit interval as ω.

Proof:

1. Since a and c are coprime, we have that there exist integers b and d such that $|ad - bc| = 1$ ([4]).

2. Let $g \in \tilde{\Gamma}$ where $g(z) = \frac{az+b}{cz+d}$. Consider the three distinct points ∞, i and $1+i$ on the Ford circle $\mathcal{C}_\infty$. It is easily shown that $g(\infty)$, $g(i)$ and $g(i+1)$ all lie on $\mathcal{C}_{\frac{a}{c}}$, the Ford circle at $\frac{a}{c}$, given by the equation

$\left|z - \left(\frac{a}{c} + \frac{i}{2c^2}\right)\right| = \frac{1}{2c^2}$. But Möbius maps map circles to circles, and three points determine a circle uniquely, so we must have that $g(z) = \frac{az+b}{cz+d}$ maps $\mathcal{C}_\infty$ to $\mathcal{C}_{\frac{a}{c}}$. Hence $\{g(\mathcal{C}_\infty) : g \in \tilde{\Gamma}\} \subseteq \mathfrak{F}$.

Let $\mathcal{C}_{\frac{a}{c}} \in \mathfrak{F}$. By part 1 above, we can find $\frac{b}{d} \in \mathbb{Q}$ such that $|ad - bc| = 1$. Let $g(z) = \frac{az+b}{cz+d}$. Thus $g \in \tilde{\Gamma}$. Further, we have $g(\infty) = \frac{a}{c}$ and $g(\mathcal{C}_\infty) = \mathcal{C}_{\frac{a}{c}}$. Thus $\mathfrak{F} \subseteq \{g(\mathcal{C}_\infty) : g \in \tilde{\Gamma}\}$ and hence $\mathfrak{F} = \{g(C_\infty) : g \in \tilde{\Gamma}\}$.

Further, since $\frac{a}{c} = g(\infty)$ we can state that $g(\mathcal{C}_\infty) = \mathcal{C}_{g(\infty)}$. It follows that if $f, g \in \tilde{\Gamma}$, then $gf(\mathcal{C}_\infty) = \mathcal{C}_{gf(\infty)}$ since $gf \in \tilde{\Gamma}$. Suppose $\mathcal{C}_{\frac{a}{c}}$ is tangent to $\mathcal{C}_{\frac{b}{d}}$. Since g maps circles to circles and preserves tangency, we have that $g\left(\mathcal{C}_{\frac{a}{c}}\right) = \mathcal{C}_{g\left(\frac{a}{c}\right)}$ is tangent to $g\left(\mathcal{C}_{\frac{b}{d}}\right) = \mathcal{C}_{g\left(\frac{b}{d}\right)}$ and these circles are Ford circles. In particular, $\mathcal{C}_{\frac{1}{0}}$ and $\mathcal{C}_{\frac{0}{1}}$ touch at i, which implies that $g(\mathcal{C}_\infty)$=$\mathcal{C}_{g(\infty)}$ and $g(\mathcal{C}_0)$=$\mathcal{C}_{g(0)}$ touch at $g(i)$. Further, we have $\Re(g(i)) = \Re\left(\frac{ai+b}{ci+d}\right) = \Re\left(\left(\frac{ai+b}{ci+d}\right)\left(\frac{-ci+d}{-ci+d}\right)\right) = \frac{ac+bd}{c^2+d^2} \in \mathbb{Q}$.

3. $|(b+na)c - a(d+nc)| = |bc - ad| = 1$ and it is thus easily verified that the fractions $\frac{b_n}{d_n}$ are adjacent to $\frac{a}{c}$.

   $|(b+na)(d+(n+1)c) - (b+(n+1)a)(d+nc)| = |bc - ad| = 1$ and it is thus easily verified that $\frac{b_n}{d_n}$ is adjacent to $\frac{b_{n+1}}{d_{n+1}}$.

$$\frac{b_n}{d_n} = \frac{a}{c} + \frac{bc-ad}{c(d+nc)} = \frac{a}{c} \pm \frac{1}{c^2\left(n+\frac{d}{c}\right)} \tag{3.3}$$

As $n \to \infty$, $\frac{b_n}{d_n}$ approaches $\frac{a}{c}$ from one side. As $n \to -\infty$, $\frac{b_n}{d_n}$ approaches $\frac{a}{c}$ from the other side. Hence the circles $C_{\frac{b_n}{d_n}}$ form a ring around $C_{\frac{a}{c}}$ and are all tangent to $C_{\frac{a}{c}}$, with $C_{\frac{b_n}{d_n}}$ tangent to $C_{\frac{b_{n+1}}{d_{n+1}}}$ and $C_{\frac{b_{n-1}}{d_{n-1}}}$. It is not possible to draw a circle, lying in $\mathbb{H}^\perp$, which is tangent to both the real line and $C_{\frac{a}{c}}$ but does not intersect one of the circles of the form $C_{\frac{b_n}{d_n}}$. Thus there are no further fractions adjacent to $\frac{a}{c}$.

4. We have that $|d+nc| < |c|$ or, equivalently, $|n+\frac{d}{c}| < 1$ for exactly two values of n, namely the integers between which $\frac{-d}{c}$ lies. For one of these two values of n, $n+\frac{d}{c}$ is positive while the other is negative. We have from (3.3) that one of the these fractions is greater than $\frac{a}{c}$ while the other is less than $\frac{a}{c}$. Hence a vertical geodesic $\Re(z) = \omega$ can only cut a finite number of Ford circles. Conversely, if L cuts through a finite number of Ford circles, then the last of these Ford circles must be cut by L at its point of tangency with the real axis. Thus w is rational. It follows that if $\omega \notin \mathbb{Q}$ then the vertical geodesic $\Re(z) = \omega$ will cut infinitely many Ford circles.

5. Since $\omega \in \mathbb{R} \setminus \mathbb{Q}$, there exists an integer n such that $n < \omega < n+1$. But the Ford circles $\mathcal{C}_n$ and $\mathcal{C}_{n+1}$ are tangent to each other and to the real axis, so, with the exception of $\mathcal{C}_\infty$, all the Ford circles cut by L will lie in the mesh triangle in between $\mathcal{C}_n$ and $\mathcal{C}_{n+1}$ and will therefore have points of tangency (with the real line) in between n and $n+1$.

■

The relationship between Ford circles and the extended modular group can be investigated further, but this investigation would be beyond the scope of this text because we are only considering simple continued fractions.

Chapter 4

Periodic Simple Continued Fractions

4.1 Introduction

In this chapter, we pay special attention to the geometric properties of *periodic* simple continued fractions. In particular, we examine the relationship between periodic simple continued fractions and the fixed points of loxodromic modular transformations.

Definition 4.1 The simple continued fraction $b_0 + K(1|b_n)$ is *periodic* if the sequence $b_0, b_1, ...$ is periodic.

Definition 4.2 The simple continued fraction $b_0 + K(1|b_n)$ is *pre-periodic* if the sequence $b_0, b_1, ...$ is periodic after a finite number of initial terms have been deleted.

The *period* p of the simple continued fraction $b_0 + K(1|b_n)$ is the smallest positive integer that is a period of the sequence $b_0, b_1, ...$

Suppose the sequence $s_{b_0}, s_{b_1}, ...$ is periodic with period p, where $s_{b_n}(z) = b_n + \frac{1}{z}$. Let $S_{[b_p]} = s_{b_0} s_{b_1} \cdots s_{b_p}$. We call $S_{[b_p]}$ the *generator* of the corresponding simple continued fraction.

Definition 4.3 A real number x is a *quadratic irrational* if it is of the form $x = \frac{a+b\sqrt{D}}{c}$ where $a, b, c \in \mathbb{Z}$, $c \neq 0$ and D is a square-free positive integer.

Definition 4.4 If $x = \frac{a+b\sqrt{D}}{c}$ is a quadratic irrational and $x^* = \frac{a-b\sqrt{D}}{c}$, then x^* is also a quadratic irrational and is a solution of the same quadratic equation of which x is a solution. We call x^* the *algebraic conjugate* of x.

Lemma 4.5 The image of a quadratic irrational under an extended modular transformation is again a quadratic irrational.

Proof: Let $f \in \tilde{\Gamma}$ with $f(z) = \frac{pz+q}{rz+s}$, and let $x = \frac{a+b\sqrt{D}}{c}$ be a quadratic irrational. Then

$$\begin{aligned} f(x) &= \frac{p(\frac{a+b\sqrt{D}}{c}) + q}{r(\frac{a+b\sqrt{D}}{c}) + s} \\ &= \frac{(pa+qc) + pb\sqrt{D}}{(ra+sc) + rb\sqrt{D}} \\ &= \frac{(pq+c)((ra+sc) - rb\sqrt{D})}{((ra+sc) + rb\sqrt{D})((ra+sc) - rb\sqrt{D})} + \frac{pb\sqrt{D}((ra+sc) - rb\sqrt{D})}{((ra+sc) + rb\sqrt{D})((ra+sc) - rb\sqrt{D})} \\ &= \frac{-(sc+ra)(p+qc) - prb^2D \pm \sqrt{D}(\pm bc + pbr(a-1)^2)}{(ra+sc)^2 - (rb)^2 D} \end{aligned}$$

■

It is well-known to number-theorists that an irrational number x has a pre-periodic simple continued fraction expansion if and only if x is a quadratic irrational ([1],[2],[3]). In our text, we give a geometric proof that an irrational number has a pre-periodic simple continued fraction expansion if and only if it is the fixed point of some loxodromic modular transformation.

Recall that a loxodromic modular map is a map of the form $f(z) = \frac{az+b}{cz+d}$, with $a, b, c, d \in \mathbb{Z}$ and $ad - bc = 1$ and $tr^2(f) = (a + d)^2 > 4$ and f has two distinct fixed points in $\mathbb{R}_\infty$. Let us call these fixed points α and β, where α is the attracting fixed point. Recall that $f^n(z) \to \alpha$ as $n \to \infty$ for all $z \neq \beta$, and $f^n(\beta) = \beta$ for all n. Let $A(f)$ be the unique geodesic that has α and β as its endpoints. Then $A(f)$ is fixed by f and is called the *axis* of f.

Theorem 4.6 Let f be a loxodromic modular map and let α and β be the fixed points of f, with α the attracting fixed point and $\beta > \alpha$. Then $A(f)$ is the only geodesic that is fixed by f and f^{-1}, but all Euclidean circles in $\mathbb{C}_\infty$ that pass through α and β are also fixed by f and f^{-1}.

Proof: We know that $A(f)$ is fixed by f and f^{-1} since the endpoints of $A(f)$ are the fixed points of f and f^{-1}, and f and f^{-1} are isometries of $\mathbb{H}^\perp$ by Theorem 1.23. Further, we know that $A(f)$ is the only geodesic that is fixed by f and f^{-1} since f and f^{-1} are loxodromic maps and therefore have only two distinct fixed points in $\mathbb{R}_\infty$. Now consider the segment of any Euclidean circle in $\mathbb{H}^\perp$ that passes through α and β, and consider the map $g(z) = \frac{z-\beta}{z-\alpha}$. Then g maps $A(f)$ to $\mathbb{I}$ and maps the segment of the Euclidean circle (which passes through α and β) in $\mathbb{H}^\perp$ to an infinite ray through the origin. Thus ∞ is the attracting fixed point of $h = gfg^{-1}$. Thus $h(z) = \lambda z$, where $\lambda > 1$. That is, $h = gfg^{-1}$ fixes $\mathbb{I}$ and the system of infinite rays passing through 0. Hence all circles, including $A(f)$, that pass through α and β are fixed by f.

Similarly, they are fixed by f^{-1}. ■

4.2 Pell's Equation

Before we explore the relationship between quadratic irrationals and the fixed points of loxodromic modular maps, let us take a closer look at the properties of quadratic irrationals by examining the integer solutions of the equations

$$X^2 - DY^2 = 1 \tag{4.1}$$

and

$$X^2 - DY^2 = 4 \tag{4.2}$$

where D is a square-free positive integer. This discussion is taken from [12].

It is easy to show that if (X_i, Y_i) is a solution of (4.1), called *Pell's Equation*, then $(2X_i, 2Y_i)$ is a solution of (4.2). Hence the solutions of Pell's equation will naturally lead us to solutions of (4.2).

A solution of either (4.1) or (4.2) is called a *trivial* solution if $Y = 0$. We will show that (4.1) and (4.2) have non-trivial solutions.

Lemma 4.7 If D is a square-free positive integer and $a, b, c, d \in \mathbb{Q}$, then $a + b\sqrt{D} = c + d\sqrt{D}$ if and only if $a = c$ and $b = d$.

Proof: If $a = c$ and $b = d$ then clearly $a + b\sqrt{D} = c + d\sqrt{D}$.
Conversely, if $a + b\sqrt{D} = c + d\sqrt{D}$ then $(b - d)\sqrt{D} = c - a$.
But $(b - d) \in \mathbb{Q}$ and $\sqrt{D} \in \mathbb{R} \setminus \mathbb{Q}$, so $(b - d)\sqrt{D} \in \mathbb{R} \setminus \mathbb{Q}$.
But $(c - a) \in \mathbb{Q}$. So $(b - d)\sqrt{D} = c - a$ is only possible if $b - d = 0$. That is, if $b = d$. Hence $c - a = 0$ and so $a = c$. ■

Theorem 4.8 Suppose D is a square-free positive integer. Then there exists a solution (X, Y) of Pell's equation such that $X, Y \in \mathbb{Z}^+$. ([12])

Proof: We know from Theorem 3.24 that a vertical line $\Re(z) = \sqrt{D}$ will cut infinitely many distinct Ford circles $\mathcal{C}_{\frac{p_i}{q_i}}$ so that $\left|\sqrt{D} - \frac{p_i}{q_i}\right| \leq \frac{1}{2q_i^2}$. Hence we have an infinite sequence $\{\frac{p_i}{q_i}\}_{i\in I}$, where I is an index set, satisfying $\left|\sqrt{D} - \frac{p_i}{q_i}\right| \leq \frac{1}{2q_i^2}$. Then

$$|p_i^2 - Dq_i^2| = \left(q_i^2\left|\sqrt{D} - \frac{p_i}{q_i}\right|\right)\left(\sqrt{D} + \frac{p_i}{q_i}\right) \leq \left(\frac{1}{2q_i^2}q_i^2\right)\left(\sqrt{D} + \frac{p_i}{q_i}\right) = \frac{1}{2}\left(\sqrt{D} + \frac{p_i}{q_i}\right).$$

We know from Theorem 3.24 that if $\mathcal{C}_{\frac{p_1}{q_1}}$ and $\mathcal{C}_{\frac{p_2}{q_2}}$ are adjacent Ford circles intersected by the vertical line $\Re(z) = \sqrt{D}$, then $|p_1q_2 - q_1p_2| = 1$ and $\frac{p_1}{q_1}$, $\frac{p_2}{q_2}$ are Farey neighbours. By Theorem 3.24, we have some integer n such that

$$n \leq \frac{p_1}{q_1} < \sqrt{D} < \frac{p_2}{q_2} \leq n + 1$$

or

$$n \leq \frac{p_2}{q_2} < \sqrt{D} < \frac{p_1}{q_1} \leq n + 1.$$

That is, $\sqrt{D}$ and $\frac{p_i}{q_i}$, $i \in I$, lie in the same unit interval, since $\mathcal{C}_{\frac{p_i}{q_i}}$ and $\mathcal{C}_{\frac{p_{i+1}}{q_{i+1}}}$ are adjacent Ford circles. That is, we have

$$n < \sqrt{D} < n + 1$$

and

$$n < \frac{p_i}{q_i} < n + 1.$$

So

$$2n < \sqrt{D} + \frac{p_i}{q_i} < 2n + 2.$$

This implies that

$$n < \frac{1}{2}\left(\sqrt{D} + \frac{p_i}{q_i}\right) < n + 1.$$

Further, we have

$$n+1<\sqrt{D}+1<n+2.$$

Hence $n<\frac{1}{2}\left(\sqrt{D}+\frac{p_i}{q_i}\right)<n+1<\sqrt{D}+1$ and so $\frac{1}{2}\left(\sqrt{D}+\frac{p_i}{q_i}\right)<\sqrt{D}+1$. Thus $|p_i^2-Dq_i^2|<\sqrt{D}+1$, or $-(\sqrt{D}+1)<p_i^2-Dq_i^2<\sqrt{D}+1$. Thus $\{p_i^2-Dq_i^2\}_{i\in I}$ is a bounded sequence of integers. Since there are only a finite set of integers between $-(\sqrt{D}+1)$ and $\sqrt{D}+1$, we can find infinitely many $j\in I$ such that $p_j^2-Dq_j^2=K$, where K is some integer between $-\sqrt{D}-1$ and $\sqrt{D}+1$. Hence we can pass to this infinite subsequence of $\{p_i^2-Dq_i^2\}_{i\in I}$ in which $p_j^2\equiv Dq_j^2 \pmod K$ for all j. Further, since the sequence $\{p_i\}_{i\in I}$ is infinite, we can find infinitely many members such that $p_i\equiv p_j \pmod K$. Similarly, we can find infinitely many members of the infinite sequence $\{q_i\}_{i\in I}$ such that $q_i\equiv q_j \pmod K$. Let us now pass to this further subsequence of $\{p_i^2-Dq_i^2\}_{i\in I}$ in which $p_i^2\equiv Dq_i^2 \pmod K$, $p_i\equiv p_j \pmod K$ and $q_i\equiv q_j$ $\pmod K$. Note that

$$(p_1-\sqrt{D}q_1)(p_2+\sqrt{D}q_2)=p_1p_2-Dq_1q_2+\sqrt{D}(p_1q_2-p_2q_1)$$

where p_i,p_j,q_i,q_j satisfy $p_i\equiv p_j \pmod K$ and $q_i\equiv q_j \pmod K$.

Now let $u=p_1p_2-Dq_1q_2$ and $v=p_1q_2-p_2q_1$.

Then $u\equiv(p_1^2-Dq_1^2) \pmod K\equiv K \pmod K\equiv 0 \pmod K$

and $v\equiv(p_1q_1-p_1q_1) \pmod K\equiv 0 \pmod K$.

Thus we can write $u=KX$ and $v=KY$ for some $X,Y\in\mathbb{Z}$.

Further, we note the following:

$u^2-Dv^2=(p_1p_2-Dq_1q_2)^2-D(p_1q_2-p_2q_1)^2=(p_2^2-Dq_2^2)(p_1^2-Dq_1^2)=K^2$.

But $u^2-Dv^2=K^2X^2-DK^2Y^2=K^2(X^2-DY^2)$.

Thus $K^2=K^2(X^2-DY^2)$ and hence $X^2-DY^2=1$.

Note that $Y\neq 0$, because if $Y=0$ then $v=0$ and then $v=p_1q_2-p_2q_1=0$ and so $\frac{p_1}{q_1}=\frac{p_2}{q_2}$, which is not possible since we assumed that the p_j and q_j are coprime and that each $\frac{p_i}{q_i}$ is distinct. Hence $|Y|\geq 1$. Further, we note that

$X^2 - DY^2 = 1$ implies that $Y^2 = \frac{X^2-1}{D}$. Hence $Y \neq 0$ implies that $|X| > 1$. If $X < -1$ then let $X' = -X$. Then $(X')^2 = X^2$ and so $(X')^2 - DY^2 = 1$. So, if necessary, we can replace X by $-X$ to obtain $X > 1$. Similarly, we can replace Y by $-Y$ so that we have a solution (X, Y) of (4.1) with $X > 0$ and $Y > 0$. ■

Theorem 4.9 The map θ defined by $\theta(X + Y\sqrt{D}) = (X, Y)$ is an isomorphism between the additive groups $\mathbb{Z}(\sqrt{D}) = \{X + Y\sqrt{D} : X, Y \in \mathbb{Z}\}$ and $\mathbb{Z} \times \mathbb{Z} = \{(X, Y) : X, Y \in \mathbb{Z}\}$.

Proof: Suppose $X_1 + Y_1\sqrt{D} = X_2 + Y_2\sqrt{D}$. Then $(X_1, Y_1) = (X_2, Y_2)$, by Lemma 4.7, and so $\theta(X_1 + Y_1\sqrt{D}) = (X_1, Y_1) = (X_2, Y_2) = \theta(X_2 + Y_2\sqrt{D})$. Thus θ is well-defined. The fact that θ is one-to-one follows immediately. It is also clear that θ is onto. Now we need only show that θ is a homomorphism.

$$\begin{aligned}
\theta((X_1 + Y_1\sqrt{D}) + (X_2 + Y_2\sqrt{D})) &= \theta((X_1 + X_2) + (Y_1 + Y_2)\sqrt{D}) \\
&= (X_1 + X_2, Y_1 + Y_2) \\
&= (X_1, Y_1) + (X_2, Y_2) \\
&= \theta(X_1 + Y_1\sqrt{D}) + \theta(X_2 + Y_2\sqrt{D})
\end{aligned}$$

■

We thus identify the number $X + Y\sqrt{D}$ with the ordered pair (X, Y). This relationship is explored in the following lemma.

Lemma 4.10 Let (X, Y) be a solution of (4.2) and let $U = X + Y\sqrt{D}$ and let $V = X - Y\sqrt{D}$. Then $U \neq 0$; $U = 2$ if and only if $X = 2$ and $Y = 0$; $U = -2$ if and only if $X = -2$ and $Y = 0$. Further, we have the following:

1. $X > 2$ and $Y > 0$ if and only if $U > 2$.

2. $X > 2$ and $Y < 0$ if and only if $0 < U < 2$.

3. $X < -2$ and $Y > 0$ if and only if $-2 < U < 0$.

4. $X < -2$ and $Y < 0$ if and only if $U < -2$.

Proof: The hyperbola $X^2 - DY^2 = 4$ (represented in Figure 8) is transformed to the hyperbola $UV = 4$ (represented in Figure 9) by the transformation equations $U = X + Y\sqrt{D}$ and $V = X - Y\sqrt{D}$.

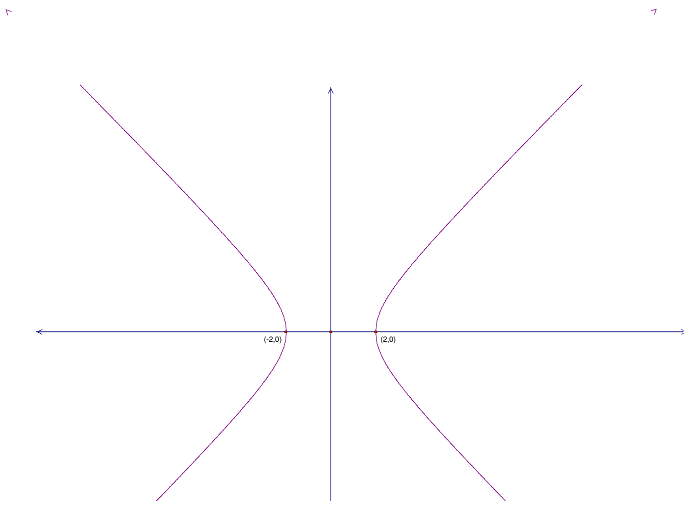

Figure 8: $X^2 - DY^2 = 4$

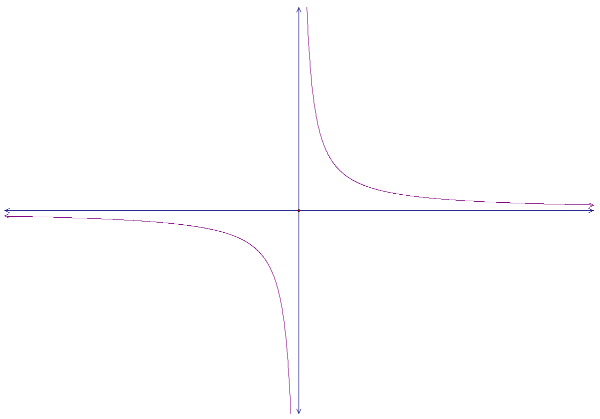

Figure 9: $UV = 4$

Note that $U = X + Y\sqrt{D}$ and $V = X - Y\sqrt{D}$ give us $X = \frac{U+V}{2}$ and $Y = \frac{U-V}{2\sqrt{D}}$. We note that the lines $X = \pm\sqrt{D}Y$ are asymptotic to $X^2 - DY^2 = 4$, while the axes $U = 0$ and $V = 0$ are asymptotic to $UV = 4$.

Note that $U = 2$ if and only if $V = 2$. Since $\frac{U+V}{2} = X$ and $\frac{U-V}{2\sqrt{D}} = Y$, we have that $U = 2$, $V = 2$ if and only if $X = 2$ and $Y = 0$.

Similarly, we see that $U = -2$ if and only if $V = -2$, and this holds if and only if $X = -2$ and $Y = 0$.

The remaining relationships can be read off the graphs. ■

Remark 4.11 From Theorem 4.8 and Lemma 4.10 we deduce that there is at least one solution (X, Y) of (4.2) such that $U = X + Y\sqrt{D}$ lies in the interval $(2, \infty)$, with $U = X + Y\sqrt{D} > |X| + |Y|$. Since the values of U do not accumulate anywhere in the interval $(2, \infty)$, there must be a solution (X_0, Y_0) which yields the smallest value of $X + Y\sqrt{D}$ taken over all (X, Y)

such that $X + Y\sqrt{D} > 2$. We call (X_0, Y_0) the *fundamental solution* of (4.2). Note that $X_0 + Y_0\sqrt{D} = X_0 + \sqrt{X_0^2 - 4}$, so X_0 is the smallest positive integer X satisfying (4.2) for some integer Y.

Lemma 4.12 If G is a non-trivial subgroup of $(\mathbb{R}, +)$, then G is either dense in $\mathbb{R}$ or cyclic. Further, if G' is a non-trivial subgroup of $(\mathbb{R}^+, \cdot)$ then G' is either dense in $\mathbb{R}^+$ or cyclic. ([20], [21])

Proof: Suppose G is a subgroup of $(\mathbb{R}, +)$ and is not dense in $\mathbb{R}$. Then G cannot contain arbitrarily small positive real numbers, and so there exists a real $\epsilon > 0$ such that the open interval $(0, \epsilon)$ is disjoint from G. We claim that G contains a least positive real number. Suppose G does not contain a least positive real number. Then there exist positive real numbers $a_1 > a_2 > \cdots$ that are all in G. But each of the positive real numbers $a_i - a_{i+1}$ is in G, and all but finitely many of them must be less than ϵ, which is impossible. Hence there is a least positive real number $a \in G$. Any $b \in \mathbb{R}$ may be expressed as $b = na + c$ where $n \in \mathbb{Z}$ and $0 \leq c < a$. If $b \in G$ then we also have that $c \in G$. But then we must have that $c = 0$ since a is the least positive real number in G. Hence $b = na$ and so $G = \langle a \rangle$.

Let $\chi : (\mathbb{R}, +) \to (\mathbb{R}^+, \cdot)$ be defined by $\chi(x) = e^x$, then $G \leq \mathbb{R}$ is mapped to $\chi(G) = G'$. Exponential functions are one-to-one and onto, so χ is a one-to-one and onto mapping. Further note that $\chi(x+y) = e^{x+y} = e^x e^y = \chi(x)\chi(y)$, and so χ is a group homomorphism. Hence $G' \cong \chi(G) \leq \mathbb{R}^+$. Suppose G is dense in $\mathbb{R}$. Since the exponential map is locally a homeomorphism, we have that χ preserves denseness and so G' is dense in $\mathbb{R}^+$. If G is not dense in $\mathbb{R}$ then G is cyclic and so $G = \langle a \rangle$ where a is the least positive real element in G. Hence $G' = \langle e^a \rangle$ since the image, under an isomorphism, of a cyclic group is a cyclic group. Thus, if G' is a non-trivial subgroup of $(\mathbb{R}^+, \cdot)$ then G' is either dense in $\mathbb{R}^+$ or cyclic. ■

Theorem 4.13 Suppose D is a square-free positive integer, and let

$$\Pi(D) = \left\{ \frac{1}{2}(X + Y\sqrt{D}) : X, Y \in \mathbb{Z}, X^2 - DY^2 = 4, X + Y\sqrt{D} > 0 \right\}.$$

Then $\Pi(D)$ is a cyclic multiplicative group generated by $\frac{1}{2}(X_0+Y_0\sqrt{D})$, where (X_0, Y_0) is the fundamental solution of (4.2). [12]

Proof: Let $\Pi(D) = \left\{ \frac{U}{2} : U > 0 \right\}$, where $U = X + Y\sqrt{D} > 0$ and $X^2 - DY^2 = 4$, $X, Y \in \mathbb{Z}$. If $X = 2$ and $Y = 0$ then $\frac{U}{2} = 1 \in \Pi(D)$. Thus $\Pi(D)$ is a non-empty set of positive numbers.

Note that if $\frac{U}{2} \in \Pi(D)$ then $\frac{V}{2} = \frac{X - Y\sqrt{D}}{2} \in \Pi(D)$, where $UV = 4$. Thus, if $\frac{U}{2} \in \Pi(D)$ then $\frac{2}{U} = \frac{V}{2} \in \Pi(D)$ so that $\Pi(D)$ is closed under taking multiplicative inverses.

We need to show that $\Pi(D)$ is closed under multiplication.

Let $\frac{1}{2}(X_i + Y_i\sqrt{D}) \in \Pi(D)$ for $i = 1, 2$.

Let $\frac{A+B\sqrt{D}}{2} = \left(\frac{X_1+Y_1\sqrt{D}}{2}\right)\left(\frac{X_2+Y_2\sqrt{D}}{2}\right)$ then $\frac{A}{2} = \frac{X_1X_2+DY_1Y_2}{4}$, $\frac{B}{2} = \frac{X_1Y_2+Y_1X_2}{4}$, or $A = \frac{X_1X_2+Y_1Y_2D}{2}$, $B = \frac{X_1Y_2+Y_1X_2}{2}$.

To show that $A, B \in \mathbb{Z}$, we must show that $X_1X_2 + DY_1Y_2$ and $X_1Y_2 + Y_1Y_2$ are even. We also need to show that $A^2 - DB^2 = 4$ and that $A + B\sqrt{D} > 0$.

Since $X_i + Y_i\sqrt{D} > 0$, $i = 1, 2$, we have that $A + B\sqrt{D} > 0$.

Further, $A^2 - DB^2 = \frac{1}{4}(X_1X_2 + DY_1Y_2)^2 - \frac{D}{4}(X_1Y_2 + Y_1X_2)^2 = 4$.

We need to establish that both $X_1X_2 + DY_1Y_2$ and $X_1Y_2 + Y_1X_2$ are even.

If Y_1 is even then $X_1^2 - DY_1^2 = 4$ implies that $X_1^2 = 4 + DY_1^2 = 4m$ for some $m \in \mathbb{Z}$, so X_1 is even. Thus $X_1X_2 + DY_1Y_2$ is even and $X_1Y_2 + Y_1X_2$ is even. Thus $A, B \in \mathbb{Z}$. Similarly, if Y_2 is even then X_2 is even and so $A, B \in \mathbb{Z}$.

Now assume that both Y_1 and Y_2 are odd.

Then $Y_1^2 \equiv Y_2^2 \equiv Y_1Y_2 \equiv 1 \pmod 4$. Since $X_i^2 - DY_i^2 = 4$ for $i = 1, 2$, we have that $X_i^2 \equiv DY_i^2 \pmod 4$ and so $X_1^2 \equiv X_2^2 \equiv D \pmod 4$ since $Y_i^2 \equiv 1 \pmod 4$ for $i = 1, 2$. So X_1 and X_2 have the same parity (both even or both odd) and so the parity of X_1X_2 will be the same as that of X_1 and X_2. Further, if X_i is even (or odd), then $X_i^2 \equiv D \pmod 4$ implies that D is even (or odd). Thus X_1, X_2 and D are of the same parity (all even or all odd) and so the parity of X_1X_2 will be the same as that of X_1, X_2 and D. So if Y_1, Y_2 are odd, then $2A = X_1X_2 + DY_1Y_2 \equiv X_1X_2 + D \equiv 0 \pmod 2$ since X_1, X_2 and D all have the same parity, and the sum of two odd integers is an even integer. Similarly, $2B = X_1Y_2 + Y_1X_2 \equiv 0 \pmod 2$. Thus $A, B \in \mathbb{Z}$ in all cases. Hence $\Pi(D)$ is a multiplicative group of positive integers. Further, $\Pi(D)$ is both non-trivial and non-dense in the interval $(1, \infty)$ (Remark 4.11). Hence by Lemma 4.12 we have that the group $\Pi(D)$ is an infinite cyclic multiplicative group generated by $\frac{X_0+Y_0\sqrt{D}}{2}$. ■

Definition 4.14 We will call a solution (X, Y) of (4.2) a *positive* solution if $X + Y\sqrt{D} > 0$.

Remark 4.15 In $\Pi(D)$ we consider only the positive solutions of (4.2). In general, if (X, Y) is a solution of (4.2) then so is $(-X, -Y)$. That is, (X, Y) is a solution of (4.2) if and only if we have $\frac{X+Y\sqrt{D}}{2} = \pm\left(\frac{X_0+Y_0\sqrt{D}}{2}\right)^n$ for some $n \in \mathbb{Z}$. [12]

Theorem 4.16 Suppose D is a square-free positive integer. Let $\mathbb{G}$ be the set of positive solutions of $X^2 - DY^2 = 4$. That is,

$$\mathbb{G} = \{(X, Y) : X, Y \in \mathbb{Z}, X^2 - DY^2 = 4, X + Y\sqrt{D} > 0\}.$$

Then $\mathbb{G}$ is an infinite cyclic group generated by (X_0, Y_0) where $\frac{X_0+Y_0\sqrt{D}}{2}$ is the generator of $\Pi(D)$. [12]

This theorem essentially states that $\mathbb{G}$ is isomorphic to $\Pi(D)$.

Proof: On $\mathbb{G}$ define the binary operation $\star$ by

$$(X_1, Y_1) \star (X_2, Y_2) = (X, Y)$$

where $X = \frac{X_1X_2+Y_1Y_2D}{2}$ and $Y = \frac{X_1Y_2+Y_1X_2}{2}$ and $X, Y \in \mathbb{Z}$. We show that $\star$ is a well-defined binary operation on $\mathbb{G}$:
$(X_1, Y_1) = (U_1, V_1)$ and $(X_2, Y_2) = (U_2, V_2)$ implies that $\frac{X_1+Y_1\sqrt{D}}{2} = \frac{U_1+V_1\sqrt{D}}{2}$ and $\frac{X_2+Y_2\sqrt{D}}{2} = \frac{U_2+V_2\sqrt{D}}{2}$.
Thus $\left(\frac{X_1+Y_1\sqrt{D}}{2}\right)\left(\frac{X_2+Y_2\sqrt{D}}{2}\right) = \left(\frac{U_1+V_1\sqrt{D}}{2}\right)\left(\frac{U_2+V_2\sqrt{D}}{2}\right)$ as $\Pi(D)$ is a group under a well-defined operation.
Thus $X_1X_2 + Y_1Y_2D = U_1U_2 + V_1V_2D$ and $X_1Y_2 + Y_1X_2 = U_1V_2 + V_1U_2$.
Hence $(X_1, Y_1) \star (X_2, Y_2) = (U_1, V_1) \star (U_2, V_2)$.

Let $\Omega : \Pi(D) \to \mathbb{G}$ be defined by $\Omega\left(\frac{X+Y\sqrt{D}}{2}\right) = (X, Y)$. Then Ω is well-defined and bijective. Further, observe that

$$\begin{aligned}
\Omega\left(\left(\frac{X_1+Y_1\sqrt{D}}{2}\right)\left(\frac{X_2+Y_2\sqrt{D}}{2}\right)\right) &= \Omega\left(\frac{A+B\sqrt{D}}{2}\right) \\
&= (A, B) \\
&= \left(\frac{X_1X_2+Y_1Y_2D}{2}, \frac{X_1Y_2+Y_1X_2}{2}\right) \\
&= (X_1, Y_1) \star (X_2, Y_2) \\
&= \Omega\left(\frac{X_1X_2+Y_1Y_2D}{2}\right) \star \Omega\left(\frac{X_1Y_2+Y_1X_2}{2}\right).
\end{aligned}$$

Hence Ω is a group homomorphism and so $\mathbb{G}$ is an infinite cyclic group generated by (X_0, Y_0). ■

4.3 Quadratic Irrationals and Loxodromic Modular Maps

Recall that the *discriminant* D of the quadratic polynomial $Az^2 + Bz + C$, where $A, B, C \in \mathbb{Z}$, $A \neq 0$, is given by $D = B^2 - 4AC$ and that this quantity D gives us information about the nature of the roots of the quadratic equation

$$Az^2 + Bz + C = 0 \tag{4.3}$$

In particular, recall that the two roots of (4.3) are distinct irrationals if and only if D is a square-free positive integer.

In this section, we show that any pair α and α^* of algebraically conjugate quadratic irrationals are precisely the pair of fixed points of some loxodromic modular map. That is, we will illuminate the geometric properties of quadratic irrationals by analysing the action of loxodromic modular maps on $\mathbb{H}^2$.

Theorem 4.17 The fixed points of a loxodromic modular map are algebraically conjugate quadratic irrationals. Further, if α and α^* are algebraically conjugate quadratic irrationals, then α and α^* are the fixed points of a loxodromic modular map ([12]).

Proof: Let g be a loxodromic modular map. Then we may write $g(z) = \frac{az+b}{cz+d}$ for $a, b, c, d \in \mathbb{Z}$ such that $ad - bc = 1$ and $(a + d)^2 > 4$. To find the fixed points of g, we solve the quadratic equation $cz^2 + (d - a)z - b = 0$. We note that if $c = 0$ then $ad - bc = 1$ implies that $ad = 1$, which implies $a, d = \pm 1$. This contradicts $(a + d)^2 > 4$, so $c \neq 0$. The discriminant of this equation is given by

$D = (d-a)^2+4bc = d^2-2ad+a^2+4(ad-1) = (a+d)^2-4$, since $ad-bc = 1$. We have that $(a+d)^2 > 4$, so $(a+d)^2-4 > 0$. Suppose that $\sqrt{D} \in \mathbb{Q}$. Then since $D \in \mathbb{Z}$, we must have that $\sqrt{D} \in \mathbb{Z}$, say $\sqrt{D} = M$, where $M > 0$ so that $D = (a+d)^2-4 = M^2$. But this means that $4 = (a+d)^2-M^2$, which implies that $(|a+d|-M)(|a+d|+M) = 4$. But this implies that $|a+d|-M = 1$ and $|a+d|+M = 4$, since $(a+d)^2 > 4$ implies that $|a+d| > 2$. Hence $|a+d|-M+|a+d|+M = 5$ and so $|a+d| = \frac{5}{2}$. But this is a contradiction, because $|a+d| = \frac{5}{2}$ is not possible since we assumed that $a, d \in \mathbb{Z}$. Thus we can conclude that the fixed points of g are algebraically conjugate quadratic irrationals, with $D = (a+d)^2-4$ having an irrational square root.

Conversely, suppose that α and α^* are the solutions of (4.3), where $D = B^2-4AC > 0$ and $\sqrt{D} \in \mathbb{R} \setminus \mathbb{Q}$. By Theorem 4.8, we have that there exist $X, Y \in \mathbb{Z}^+$ such that $X^2-DY^2 = 4$. Note that
$(X-YB)(X+YB)-(-2CY)(2AY) = X^2-Y^2B^2+4ACY^2$
$= X^2-Y^2(B^2-4AC) = X^2-DY^2 = 4$. Hence the matrix

$$\begin{pmatrix} X-YB & -2CY \\ 2AY & X+YB \end{pmatrix}$$

has determinant 4. Now $4ACY^2$ is even, and we have that
$(X-YB)(X+YB)+4ACY^2 = 4$, so $X+YB$ or $X-YB$ must be even, or $X+YB$ and $X-YB$ are both even. But if $X-YB$ is even then $X-YB = 2V$ for some $V \in \mathbb{Z}$, which implies that $X+YB = 2V+2YB$, which implies $X+YB$ is also even. Similarly, $X+YB$ is even implies that $X-YB$ is also even. Thus we may write $X-YB = 2V$ and $X+YB = 2U$ for some $U, V \in \mathbb{Z}$. Hence the map $g(z) = \frac{Vz-CY}{AYz+U}$ is a modular map since $U, CY, AY, V \in \mathbb{Z}$ and

$$\begin{aligned} UV - AYCY &= \frac{(X-YB)(X+YB)}{4} - \frac{4ACY^2}{4} \\ &= \frac{(X-YB)(X+YB)-4ACY^2}{4} \end{aligned}$$

$$\begin{aligned} &= \frac{4}{4} \\ &= 1. \end{aligned}$$

The fixed points of g satisfy $Vz - CY = z(AYz + U)$. This implies that $AYz^2 + z(U - V) + CY = 0$, which implies that $AYz^2 + zYB + CY = 0$. Hence $Az^2 + Bz + C = 0$, since $Y \neq 0$. Further, observe that

$$(U + V)^2 = X^2 = 4 + DY^2 > 4.$$

Hence the fixed points of g satisfy (4.3), and we have $tr^2(g) = (U + V)^2 > 4$. Hence the fixed points of g are α and α^*, and g is a loxodromic modular map.

■

Lemma 4.18 Let α be a quadratic irrational. Then a loxodromic modular map g fixes α if and only if g fixes α^*. Moreover Γ_α, the stabiliser of α in Γ, is cyclic. [12]

Proof: Suppose $g(z) = \frac{az+b}{cz+d}$ is a loxodromic modular map.
Then the fixed points of g satisfy $cz^2 + (d - a)z - b = 0$, so the fixed points are algebraic conjugates of each other. Let us denote them by α and α^*. Thus $\alpha = \frac{a-d+\sqrt{(d-a)^2+4bc}}{2c}$ and $\alpha^* = \frac{a-d-\sqrt{(d-a)^2+4bc}}{2c}$, where $c \neq 0$, and $g(\alpha) = \alpha$, $g(\alpha^*) = \alpha^*$ and $(\alpha^*)^* = \alpha$.

The stabiliser of α in Γ is given by $\Gamma_\alpha = \{g \in \Gamma : g(\alpha) = \alpha\}$. Thus $g \in \Gamma_\alpha$ if and only if $g(\alpha) = \alpha$ and $g(\alpha^*) = \alpha^*$, by Theorem 4.17.

Assume that $\alpha > \alpha^*$ and set $f(z) = \frac{z-\alpha}{z-\alpha^*}$. Then $f \in PSL(2, \mathbb{R})$ and $f(\alpha) = 0$ and $f(\alpha^*) = \infty$. We note that for each $g \in \Gamma_\alpha$ we have $fgf^{-1}(0) = fg(\alpha) = f(\alpha) = 0$, and $fgf^{-1}(\infty) = fg(\alpha^*) = f(\alpha^*) = \infty$.

Let $\Lambda_0 = f\Gamma_\alpha f^{-1}$. Thus each element in Λ_0 fixes both 0 and ∞ and is in

$PSL(2,\mathbb{R})$. Hence the geodesic $\mathbb{I}$ is fixed. Hence the elements of Λ_0 must be of the form $z \mapsto kz$, $k \in \mathbb{R}^+$.

Hence $\Lambda_0 = \{g_k \in PSL(2,\mathbb{R}) : g_k(z) = kz, k \in \mathbb{R}^+\}$ is isomorphic to $(\mathbb{R}^+, \cdot)$, the multiplicative group of positive real numbers.

Since Γ_α is a subgroup of Γ, which is a discrete subgroup of $PSL(2,\mathbb{R})$ ([10],[16]), the action of Γ_α on the hyperbolic geodesic joining α and α^* is non-trivial and discrete. Thus the action of $\Lambda_0 = f\Gamma_\alpha f^{-1}$ on $\mathbb{I}$ is also non-trivial and discrete. Since we have that Λ_0 is isomorphic to $(\mathbb{R}^+, \cdot)$ we have Λ_0 is cyclic and hence Γ_α is cyclic, by Lemma 4.12. ■

The next result establishes the generator of the cyclic group Γ_α.

Theorem 4.19 Let α and α^* be the solutions of (4.3), with
$D = B^2 - 4AC > 0$, $\sqrt{D} \in \mathbb{R} \setminus \mathbb{Q}$ and where

$$\mathbb{G} = \{(X,Y) : X^2 - DY^2 = 4, X + Y\sqrt{D} > 0\}.$$

Then the map $\Psi : \mathbb{G} \to \Gamma_\alpha$ defined by $\Psi(X,Y) = g$ where

$$g(z) = \frac{(X - YB)z - 2CY}{2AYz + (X + YB)},$$

is an isomorphism. Furthermore, $\Gamma_\alpha = \{g \in \Gamma : g(\alpha) = \alpha\}$ is generated by $\Psi(X_0, Y_0)$, where (X_0, Y_0) is the fundamental solution of (4.2). [12]

Proof: By Theorem 4.16, we can identify $\mathbb{G}$ with $\Pi(D)$. By Theorem 4.17, we have that the solutions α and α^* of (4.3) are the fixed points of $g(z) = \frac{Vz-CY}{AYz+U}$ where $U = \frac{X+BY}{2}$, $V = \frac{X-BY}{2}$ and $g(\alpha) = \alpha$ and $g(\alpha^*) = \alpha^*$.
We first show that $\Psi : (\mathbb{G}, \star) \to \Gamma_\alpha$ is a group homomorphism.
$\Psi((X_1, Y_1) \star (X_2, Y_2)) = \Psi\left(\frac{X_1X_2+DY_1Y_2}{2}, \frac{X_1Y_2+Y_1X_2}{2}\right) = g$

where

$$\begin{aligned} g(z) &= \frac{\left(\left(\frac{X_1X_2+DY_1Y_2}{2}\right) - B\left(\frac{X_1Y_2+Y_1X_2}{2}\right)\right)z - 2C\left(\frac{X_1Y_2+Y_1X_2}{2}\right)}{2A\left(\frac{X_1Y_2+Y_1X_2}{2}\right)z + \left(\frac{X_1X_2+DY_1Y_2}{2} + B\left(\frac{X_1Y_2+Y_1X_2}{2}\right)\right)} \\ &= \frac{((X_1X_2+DY_1Y_2) - B(X_1Y_2+Y_1X_2))z - 2C(X_1Y_2+Y_1X_2)}{2A(X_1Y_2+Y_1X_2)z + X_1X_2 + DY_1Y_2 + B(X_1Y_2+Y_1X_2)} \\ &= g_1g_2(z) \end{aligned}$$

with $g_i(z) = \frac{(X_i - BY_i)z - 2CY_i}{2AY_iz + X_i + BY_i}$ for $i = 1, 2$.

Now, the composition g_1g_2 is associated with the matrix product

$$\begin{pmatrix} X_1 - BY_1 & -2CY_1 \\ 2AY_1 & X_1 + BY_1 \end{pmatrix} \begin{pmatrix} X_2 - BY_2 & -2CY_2 \\ 2AY_2 & X_2 + BY_2 \end{pmatrix}$$

$=$

$$\begin{pmatrix} (X_1 - BY_1)(X_2 - BY_2) - 4ACY_1Y_2 & -2CY_2(X_1 - BY_1) - 2CY_1(X_2 + BY_2) \\ 2AY_1(X_2 - BY_2) + 2AY_2(X_1 + BY_1) & -4ACY_1Y_2 + (X_1 + BY_1)(X_2 + BY_2) \end{pmatrix}$$

$=$

$$\left(\begin{matrix} X_1X_2 + Y_1Y_2(B^2 - 4AC) - B(Y_1X_2 + X_1Y_2) & -2C(Y_2X_1 + Y_1X_2) \\ 2A(Y_1X_2 + Y_2X_1) & X_1X_2 + Y_1Y_2(B^2 - 4AC) + B(Y_1X_2 + X_1Y_2) \end{matrix} \right.$$

$=$

$$\begin{pmatrix} X_1X_2 + DY_1Y_2 - B(Y_1X_2 + X_1Y_2) & -2C(X_1Y_2 + Y_1X_2) \\ 2A(Y_1X_2 + Y_2X_1) & X_1X_2 + DY_1Y_2 + B(Y_1X_2 + X_1Y_2) \end{pmatrix}$$

Hence $\Psi((X_1, Y_1)\star(X_2, Y_2)) = g_1g_2 = \Psi(X_1, Y_1)\Psi(X_2, Y_2)$ and so Ψ is a group homomorphism.

Now we show that Ψ is one-to-one.

Suppose that $\Psi(X_1, Y_1) = \Psi(X_2, Y_2)$. Then $g_1 = g_2$, where $g_i(z) = \frac{(X_i - BY_i)z - 2CY_i}{2AY_iz + X_i + BY_i}$ for $i = 1, 2$.

We have $g_1(0) = g_2(0)$ and $g_1(\infty) = g_2(\infty)$, so

$$\frac{-2CY_1}{X_1 + Y_1B} = \frac{-2CY_2}{X_2 + Y_2B} \tag{4.4}$$

and

$$\frac{X_1 - Y_1B}{2AY_1} = \frac{X_2 - Y_2B}{2AY_2} \tag{4.5}$$

Hence we have

$$Y_1(X_2 + Y_2B) = Y_2(X_1 + Y_1B) \tag{4.6}$$

from (4.4). From (4.5) we have

$$Y_2(X_1 - Y_1B) = Y_1(X_2 - Y_2B) \tag{4.7}$$

Subtracting (4.6) from (4.7), we obtain

$$\frac{X_1}{X_2} = \frac{Y_1}{Y_2} \tag{4.8}$$

Since $X_i, Y_i \in \mathbb{Z}^+$ for $i = 1, 2$ and $X_i^2 - DY_i^2 = 4$, it is seen that

$$\frac{Y_1^2X_2^2}{Y_2^2} - DY_2^2 = 4$$

and thus

$$X_2^2 - DY_2^2 = 4\frac{Y_2^2}{Y_1^2} = 4.$$

Hence $\frac{Y_2^2}{Y_1^2} = 1$ and $Y_1 = Y_2$ and so $X_1 = X_2$. Thus $(X_1, Y_1) = (X_2, Y_2)$ and so Ψ is one-to-one.

Now we show that Ψ is onto.
Let $f \in \Gamma_\alpha$ with $f(z) = \frac{az+b}{cz+d}$. Then $f(\alpha) = \alpha$, $f(\alpha^*) = \alpha^*$ and $(a+d)^2 > 4$. Without loss of generality (by changing the sign of the coefficients of f if necessary), we may assume $a + d > 0$. We note that if w is a fixed point of f then $f(w) = w$, and we see that w satisfies a quadratic equation $Aw^2+Bw+C = 0$ and hence in fact $Y(Aw^2+Bw+Cw) = 0$ for $Y \neq 0$, where one can thus assume that A, B, C are co-prime. Further, since $\frac{aw+b}{cw+d} = w$ we have that $YA = c$, $YB = d - a$, $YC = -b$ with $Y \neq 0$. That is

$Y(Aw^2 + Bw + C) = cw^2 + (d - a)w - b$. Let $D = B^2 - 4AC$, and put $X = a + d > 0$. Then

$$\begin{aligned} X^2 - DY^2 &= X^2 - (B^2 - 4AC)Y^2 \\ &= X^2 - B^2Y^2 + 4ACY^2 \\ &= X^2 - (YB)^2 + 4(YA)(YC) \\ &= (a + d)^2 - (d - a)^2 + 4(c)(-b) \\ &= 4ad - 4bc \\ &= 4(ad - bc) \\ &= 4 \end{aligned}$$

Thus $f(w) = \dfrac{2aw + 2b}{2cw + 2d} = \dfrac{(X - YB)w - 2YC}{2(YA)w + (X + YB)}$. Hence $f = \Psi(X, Y)$ and so Ψ is onto Γ_α, where α is a fixed point of f.

Finally, since $\Psi : \mathbb{G} \to \Gamma_\alpha$ and $\Omega : \Pi(D) \to \mathbb{G}$ are both group isomorphisms, we have that $\Psi\Omega : \Pi(D) \to \Gamma_\alpha$ is a group isomorphism. We have shown in Theorem 4.13 that $\Pi(D)$ is an infinite multiplicative cyclic group generated by $\frac{X_0+Y_0\sqrt{D}}{2}$ where (X_0, Y_0) is the fundamental solution of $X^2 - DY^2 = 4$. Thus, since $\Psi : \mathbb{G} \to \Gamma_\alpha$ is an isomorphism, we have that Γ_α is generated by $\Psi(X_0, Y_0) = g_0$, where $g_0(z) = \dfrac{(X_0 - Y_0B)z - 2CY_0}{2AY_0z + (X_0 + Y_0B)}$ and α and α^* satisfy (4.3). ■

4.4 Simple Periodic Continued Fractions and Quadratic Irrationals

In this section, we prove well-known theorems about pre-periodic simple continued fractions. Specifically, that a real number x has a periodic simple continued fraction expansion if and only if x is a quadratic irrational. The number-theoretic proofs of these theorems are well-documented ([1],[2],[3]). Following Series [8], we will prove the results using Möbius maps acting on $\mathbb{H}^{\perp}$.

Recall that a Möbius map f is loxodromic if and only if $tr^2(f) \notin [0,4]$.

Lemma 4.20 Any composition of maps of the form $z \mapsto b + \frac{1}{z}$, where $b \geq 1$, is loxodromic. [12]

Proof: Let $s_b(z) = b + \frac{1}{z} = \frac{bz+1}{z}$, where $b \geq 1$. Then $s_b(1) = b + 1 > 1$ and $s_b(\infty) = b$, so s_b maps the interval $[1, \infty) \cup \{\infty\}$ into itself.
Note that $s_b^{-1}(z) = \frac{1}{z-b}$ so $s_b^{-1}(-1) = \frac{1}{-1-b} > -1$, and $s_b^{-1}(0) = \frac{1}{-b} < 0$, so s_b^{-1} maps the interval [-1,0] into itself. The same is therefore true of a composition g of such maps. Thus g has a fixed point in $[1, \infty)$ and g^{-1} has a fixed point in $[-1, 0]$. So g has two distinct real fixed points and so g must be loxodromic. ■

Theorem 4.21 If the infinite simple continued fraction expansion of $x \in \mathbb{R} \setminus \mathbb{Q}$ is pre-periodic, then x is a quadratic irrational. [12]

Proof: Let $x \in \mathbb{R} \setminus \mathbb{Q}$, say

$$x = \lim_{n\to\infty} S_{[b_n]}(0) = \lim_{n\to\infty} S_{[b_n]}(1) = \lim_{n\to\infty} S_{[b_n]}(\infty)$$

where $S_{[b_n]}(z) = s_{b_0}s_{b_1}s_{b_2}\cdots s_{b_n}(z)$ where $s_{b_i}(z) = b_i + \frac{1}{z}$ and $b_i \geq 1$ for all $i \geq 1$ and $b_0 \in \mathbb{Z}$. Then the sequence of convergents of the simple continued fraction expansion of x is given by $\{S_{[b_0]}(\infty), S_{[b_1]}(\infty), S_{[b_2]}(\infty), ...\}$. Now assume that the simple continued fraction expansion of x is pre-periodic.

Then the sequence of maps $\{s_{b_0}, s_{b_1}, s_{b_2}, ...\}$ is also pre-periodic and so it has a subsequence of the form $\{g, gf, gf^2, gf^3, ...\}$ where f is loxodromic and g must contain s_{b_0} in its composition if $b_0 < 1$. Further, we have that

$$\lim_{n\to\infty} gf^n(0) = x$$

and therefore

$$\lim_{n\to\infty} f^n(0) = g^{-1}(x).$$

But if f is loxodromic then $f^n(0)$ converges, and we denote this limit by y. That is,

$$\lim_{n\to\infty} f^n(0) = y.$$

Thus y is a fixed point of f and $x = g(y)$. Now, f is composed of maps of the form $s_{b_i}(z) = b_i + \frac{1}{z}$ with $b_i \geq 1$ for all i and $s_{b_i} \in \tilde{\Gamma}$. So $f^n \in \tilde{\Gamma}$ for all $n \in \mathbb{Z}$ and f^n also fixes y. While f may be in $\tilde{\Gamma}$, we have that $f^2 \in \Gamma$ and so y is a quadratic irrational by Theorem 4.17. But $x = g(y)$ and $g \in \tilde{\Gamma}$, and the image of a quadratic irrational under an extended modular map is again a quadratic irrational, by Lemma 4.5. Hence x is a quadratic irrational. ■

We now prove the converse of Theorem 4.21. In this proof we use the concept of a cutting sequence as a $\mathbb{T}_0$-path on the $\rho - \tau$ Farey tree. We recall that each vertex on the path can be written as

$$\mathbb{T} = \tau^{b_0}\rho^{b_1}\tau^{b_2}\rho^{b_3}\cdots\tau^{b_t}(\mathbb{T}_0),$$

where $b_0 \in \mathbb{Z}$ and $b_j \geq 1$ for all $j \geq 1$ and where b_t may be zero.

It is noted that if $-1 < \alpha^* < 0$ and $0 < \alpha < 1$, we can consider $\psi(\alpha^*) < -1$ and $\psi(\alpha) > 1$ where $\psi(z) = \frac{1}{z}$. Let $[\psi(\alpha^*)] = -n$ be the integer part of $\psi(\alpha^*)$ where $n > 1$. Then $-1 < \tau^{n-1}\psi(\alpha^*) < 0$ and $1 < \tau^{n-1}\psi(\alpha)$ since $n - 1 \geq 1$.

Thus there is a $g \in \tilde{\Gamma}$ such that $-1 < g(\alpha^*) < 0 < 1 < g(\alpha)$. In this case we say that $g(\alpha^*)$ and $g(\alpha)$ are in *standard normal form.* Thus if α and α^* are not in standard normal form they can be brought to this form by the action of $g \in \tilde{\Gamma}$.

Theorem 4.22 If α is a quadratic irrational, then the simple continued fraction expansion of α is pre-periodic. [12]

Proof: Let α and α^* be a pair of algebraically conjugate quadratic irrationals fixed by the loxodromic map $f \in \Gamma$. We know from Lemma 4.18 that Γ_α is a cyclic group, so we may choose f to be the generator of the subgroup of Γ_α that contains the *loxodromic* modular maps that fix α and α^*. Without loss of generality, we may assume that $\alpha^* < \alpha$ and that α is the attracting fixed point of f. Let ℓ be the hyperbolic geodesic with end points α and α^*. Then ℓ is fixed set-wise by the maps f and f^{-1}. Since α and α^* are irrational, they cannot be the endpoints of any Farey geodesic. Thus ℓ must be crossed by some Farey geodesic, say σ.

Since $\mathbb{I}$ cuts the axis $g(\ell)$ joining $g(\alpha^*)$ and $g(\alpha)$, we have that $\sigma = g^{-1}(\mathbb{I})$ cuts the axis ℓ joining α and α^*.

We have shown in Theorem 3.20 that the convergents of the simple continued fraction expansion of a point $g(\alpha) > 0$ correspond to the nodes of a semi-infinite $\mathbb{T}_0$-path that converges to $g(\alpha)$. The Farey tessellation, the $\rho - \tau$ Farey tree, and the axis $g(\ell)$ are invariant under the loxodromic modular map $f_0 = gfg^{-1}$ that fixes $g(\alpha^*)$ and $g(\alpha)$ with $tr^2(f_0) = tr^2(f)$.

If $\mathbb{T}_j$ is any vertex of the $\mathbb{T}_0$-path that converges to $g(\alpha)$, then $f_0^r(\mathbb{T}_j)$ is also on this $\mathbb{T}_0$-path for all $r \in \mathbb{Z}$. In particular, since $g(\alpha)$ and $g(\alpha^*)$ are in standard normal form, we know that $f_0^r(\mathbb{T}_0)$ is also on this $\mathbb{T}_0$-path for all $r \in \mathbb{Z}$. Now let

$$\mathbb{T}_1 = f_0(\mathbb{T}_0) = \tau^{b_0}\rho^{b_1}\tau^{b_2}\rho^{b_3}\cdots\tau^{b_t}(\mathbb{T}_0).$$

Since $f_0(\mathbb{T}_0)$ lies on the $\mathbb{T}_0$-path that converges to $g(\alpha) > 1$, we know that $b_0 \geq 1$.

Now

$$f_0^r(\mathbb{T}_0) = (\tau^{b_0}\rho^{b_1}\tau^{b_2}\rho^{b_3}\cdots\tau^{b_t})^r(\mathbb{T}_0)$$

for all $r \in \mathbb{Z}$.

Thus

$$g(\alpha) = \lim_{r\to\infty}(\tau^{b_0}\rho^{b_1}\tau^{b_2}\rho^{b_3}\cdots\tau^{b_t})^r(\infty)$$

since $g(\alpha)$ is fixed by f_0. Thus $g(\alpha)$ has a periodic simple continued fraction expansion with period given by

$$(\tau^{b_0}\rho^{b_1}\tau^{b_2}\rho^{b_3}\cdots\tau^{b_t}).$$

If $-1 < g(\alpha^*) < 0$ then $1 < \varphi(g(\alpha^*))$, where $\varphi(z) = -\frac{1}{z}$. Then $\varphi(g(\alpha^*))$ is fixed by the loxodromic map $\varphi f_0 \varphi$. By the same argument as above we see that $\varphi(g(\alpha^*))$ has a periodic simple continued fraction expansion. Hence both α and α^* have pre-periodic simple continued fraction expansions. That is

$$\alpha = \lim_{r\to\infty} g^{-1}(\tau^{b_0}\rho^{b_1}\tau^{b_2}\rho^{b_3}\cdots\tau^{b_t})^r(\infty)$$

where α and $g(\alpha)$ have the same tails.

Further, if

$$\varphi(g(\alpha^*)) = \lim_{r\to\infty}(\tau^{n_0}\rho^{n_1}\tau^{n_2}\cdots\rho^{n_k})^r(\infty)$$

then

$$g(\alpha^*) = \lim_{r\to\infty}\varphi(\tau^{n_0}\rho^{n_1}\tau^{n_2}\cdots\rho^{n_k})^r(\infty)$$

and

$$\alpha^* = \lim_{r\to\infty} g^{-1}\varphi(\tau^{n_0}\rho^{n_1}\tau^{n_2}\cdots\rho^{n_k})^r(\infty)$$

where α^* and $g(\alpha^*)$ have the same tails (Chapter 2.2). ■

Theorem 4.23 A positive number α has an infinite periodic simple continued fraction expansion if and only if α is a quadratic irrational and $\alpha \in (1,\infty)$ and $\alpha^* \in (-1,0)$. [12]

Proof: If α is a quadratic irrational in $(1,\infty)$ and the conjugate α^* is in $(-1,0)$, then α and α^* are in standard normal forms and so Theorem 4.22 establishes that the simple continued fraction expansion of α is periodic.

Conversely we suppose that $\alpha > 0$ has a periodic simple continued fraction expansion given by

$$\alpha = \lim_{r\to\infty}(\tau^{b_0}\rho^{b_1}\tau^{b_2}\rho^{b_3}\cdots\tau^{b_t})^r(\infty) = \lim_{r\to\infty}(s_{b_0}s_{b_1}\cdots s_{b_t})^r(\infty).$$

Thus $b_0 \geq 1$ since $b_0 < 1$ cannot occur as a partial quotient in a simple periodic continued fraction expansion. Thus $\alpha > 1$. Since any periodic simple continued fraction is trivially pre-periodic we see by Theorem 4.21 that α is a quadratic irrational.

It remains to show that the conjugate $\alpha^* \in (-1,0)$. We note that for each $j = 0, 1, 2, \cdots, t$, we have that the composition of maps of the form $s_{b_j}(z) = b_j + \frac{1}{z}$ or $s_{b_j}^{-1}(z) = \frac{1}{z-b_j}$, with $b_j \geq 1$ for all j, are both loxodromic in $\tilde{\Gamma}$, by Lemma 4.20, and

$$\alpha^* = \lim_{r\to\infty}(s_{b_0}s_{b_1}\cdots s_{b_t})^{-r}(\infty) = \lim_{r\to\infty}(s_{b_t}^{-1}\cdots s_{b_1}^{-1}s_{b_0}^{-1})^r(\infty).$$

Finally we note that each map $s_{b_j}^{-1}(z) = \frac{1}{z-b_j}$ maps the interval $[-1,0]$ into itself and the map $s_{b_j}(z) = b_j + \frac{1}{z}$ maps the interval $[1,\infty)$ into itself. Thus we have that $\alpha^* \in (-1,0)$. ■

Bibliography

[1] G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers* , 2nd ed. Oxford University Press, 1945, pp. 128 - 152.

[2] A.Ya. Khintchine, *Continued Fractions (translated by Peter Wynn)*, Groningen, The Netherlands: P. Noordhoff, 1963, pp. 6 - 56.

[3] Andrew M. Rockett and Peter Szüsz, *Continued Fractions*, World Scientific Publishing, 1992, pp. 1 - 57.

[4] L.R. Ford, *Fractions*, The American Mathematical Monthly, vol. 45, no. 9, Nov. 1938, pp 586 - 601.

[5] W.B. Jones and W.J. Thron, *Continued Fractions: Analytic Theory and Applications*, Encyclopedia of Mathematics and its Applications, vol. 11, Addison-Wesley, Reading, Massachusetts, 1980, pp. 56 - 59.

[6] J.F. Paydon and H.S. Wall, *The Continued Fraction as a Sequence of Linear Transformations*, Duke Mathematical Journal, vol. 9, 1942, pp. 360 - 372.

[7] Alan F. Beardon, *Continued Fractions, Discrete Groups and Complex Dynamics*, Computational Methods and Function Theory, vol. 1, no. 2, 2001, pp 319, 535 - 594.

[8] Caroline Series, *The Modular Surface and Continued Fractions*, Journal of the London Mathematical Society, vol. 2, no. 31, 1985, pp. 69 - 80.

[9] Svetlana Katok, *Continued Fractions, Hyperbolic Geometry and Quadratic Forms*, Mass Selecta, American Mathematical Society, 2003, pp. 121 - 160.

[10] Gareth A. Jones and David Singerman, *Complex Functions - An Algebraic and Geometric Viewpoint*, Cambridge University Press, 1987, pp. 17 - 40, 224, 231 - 232, 242 - 244.

[11] Alan F. Beardon, *Algebra and Geometry*, Cambridge University Press, 2005, pp. 254-273.

[12] Alan F. Beardon, Meira Hockman and Ian M. Short, *The Geometry of Continued Fractions*, unpublished draft, April 2010.

[13] James W. Anderson, *Hyperbolic Geometry*, 2nd ed. London, Springer-Verlag, 2007, pp. 3, 5, 6, 73, 86, 92, 93.

[14] Alan F. Beardon, *Geometry of Discrete Groups*, Springer, 1983, pp. 22, 23, 33 - 36, 204.

[15] Linda Keen and Nikola Lakic, *Hyperbolic Geometry from a Local Viewpoint*, London Mathematical Society Student Texts, Cambridge University Press, vol. 68, 2007, pp. 98, 103.

[16] Svetlana Katok, *Fuchsian Groups*, Chicago Lectures in Mathematics, University of Chicago Press, 1992, pp. 26, 49, 55, 56.

[17] Arlan Ramsay and Robert D. Richtmyer, *Introduction to Hyperbolic Geometry*, New York: Springer-Verlag, 1995, pp. 228 - 231.

[18] Ronald Gould, *Graph Theory*, Menlo Park, California: The Benjamin/Cummings Publishing Company, 1988, pp. 3, 4, 9, 10, 65.

[19] Clifford W. Marshall, *Applied Graph Theory*, John Wiley & Sons, 1971, pg 4.

[20] Kenneth R. Goodearl, *Partially Ordered Abelian Groups with Interpolation*, American Mathematical Society, 2010, pg. 70.

[21] John Stillwell, *Naive Lie Theory*, Springer, 2008, pg. 179.