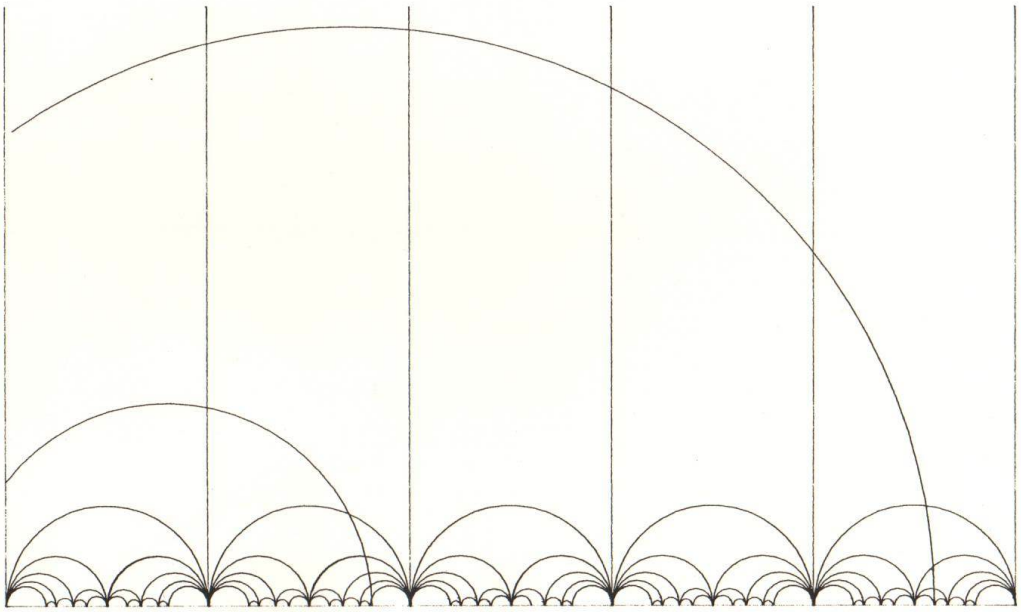
Continued Fractions and Hyperbolic Geometry

Caroline Series

Loughborough LMS Summer School

July 2015



Outline

Why is it that 22/7 and 355/113 are chosen as good approximations to π? In fact 355/113 $= 3 + 1/(7 + 1/16)$ approximates π to six decimal places. They are examples of continued fractions, which are used to get 'best approximations' to an irrational number for a given upper bound on the denominator, so-called Diophantine approximation.

There is a beautiful connection between continued fractions and the famous tiling of the hyperbolic (non-Euclidean) plane shown Figure 1. It is called the Farey tessellation and its hyperbolic symmetries are the 2x2 matrices with integer coefficients and determinant one, important in number theory. We shall use the Farey tessellation to learn about both continued fractions and hyperbolic geometry, leading to geometrical proofs of some classical results about Diophantine approximation.

Lecture 1 We describe the Farey tessellation $\mathcal{F}$ and give a very quick introduction to the basic facts we need from hyperbolic geometry, using the upper half plane model.

Lecture 2 We introduce continued fractions and explain the relationship between continued fractions and $\mathcal{F}$.

Lecture 3 We use $\mathcal{F}$ to visualise some classical results about continued fractions and outline a few of the many applications and further developments.

Everything needed about continued fractions and hyperbolic geometry will be explained in the lectures, but to prepare in advance you could look at any of the many texts on these subjects. Here are a few sources:

G. H. Hardy and E. M. Wright. *The Theory of Numbers.* Oxford University Press, Many editions.

A. Ya. Khinchin *Continued Fractions.* University of Chicago Press, 1935.

C. Series. *Hyperbolic geometry notes MA448.* Unpublished lecture notes, available at `homepages.warwick.ac.uk/~masbb/`

For an introduction to the Farey tessellation and continued fractions from a slightly different viewpoint see

A. Hatcher. *Toplogy of Numbers.* Unpublished draft book, available at `www.math.cornell.edu/~hatcher/TN/TNpage.html`

1 The Farey tessellation and the hyperbolic plane

Fractions $p/q, r/s \in \mathbb{Q}$ are called *neighbours* if $|ps - rq| = 1$. Their *Farey sum*, denoted $p/q \oplus_F r/s$, is defined to be $(p+r)/(q+s)$. Note that if $p/q < r/s$ are neighbours, then so are $p/q < p/q \oplus_F r/s$ and $p/q \oplus_F r/s < r/s$. Figure 1, drawn in the complex plane, is formed by the following procedure:

- Draw vertical lines from n to ∞ at each integer point $n \in \mathbb{R}$. Label these points $n/1$. Note that for each $n \in \mathbb{Z}$, the pair $(n/1, (n+1)/1)$ are neighbours.
- Join each adjacent pair $(n/1, (n+1)/1)$ by a semicircle with its centre on $\mathbb{R}$.
- Mark the point $n/1 \oplus_F (n+1)/1 = (2n+1)/2$. Join the adjacent neighbours $n/1, (2n+1)/2$ and $(2n+1)/2, (n+1)/1$ by semicircles centred on $\mathbb{R}$.
- Inductively, suppose that $p/q < r/s$ are Farey neighbours joined by an arc. Join p/q to $(p+r)/(q+s)$ and $(p+r)/(q+s)$ to r/s by semicircles.
- Continue in this way.

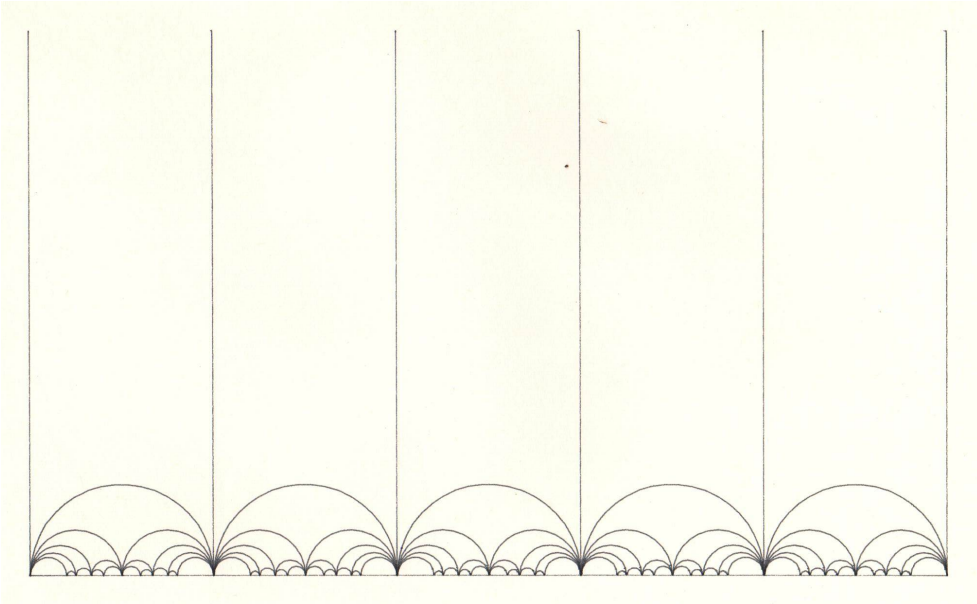


Figure 1: The Farey tessellation

Exercise 1.1. *Check by induction that if $p/q, r/s$ are joined by an arc of $\mathcal{F}$ then* $\begin{pmatrix} p & r \\ q & s \end{pmatrix}$ *has determinant* ± 1.

The Farey tessellation is a tessellation or tiling of the *hyperbolic plane*. This means there is a basic figure, a so-called *ideal triangle*, whose images under some group of symmetries cover the hyperbolic plane without overlaps.

To understand this we need a bit of background on hyperbolic geometry. Everything we shall use is worked through in detail in the first few chapters of [11], but we explain what we need briefly here. Hyperbolic geometry originated as geometry in which Euclid's parallel postulate fails. It is the geometry of space with constant curvature -1. All we need to know is that 2-dimensional hyperbolic geometry can be *modelled* as the upper half plane $\mathbb{H} = \{z \in \mathbb{C} : \Im z > 0\}$ with the *metric* $ds^2 = (dx^2 + dy^2)/y^2$, where $z = x + iy$. What this means is that to find the length of an arc γ joining points A, B we have to integrate: $\ell(\gamma) = \int_\gamma ds = \int_\gamma \sqrt{dx^2 + dy^2}/y$ and $d_{\mathbb{H}}(A, B) = \inf_\gamma \ell(\gamma)$.

Here is an example. Let $A = ai$ and $B = bi$ so that A, B are on the imaginary axis $\mathbb{I}$, and assume $b > a$. Let γ be any arc joining A to B. Then

$$\ell(\gamma) = \int_\gamma ds = \int_\gamma \sqrt{dx^2 + dy^2}/y \geq \int_\gamma dy/y = \int_{y=a}^{y=b} dy/y = \log b/a.$$

Moreover if we take γ_0 to be the vertical path from A to B then $\ell(\gamma_0) = \log b/a$. Hence $d_{\mathbb{H}}(ai, bi) = \log b/a$. Note that this shows that the vertical path γ_0 is a shortest distance path, otherwise called a *geodesic* or a *hyperbolic line*.

The boundary at infinity The above formula shows that $d_{\mathbb{H}}(i, ti) \to \infty$ as $t \to 0$. Thus the real axis is at infinite distance from a point in $\mathbb{H}$. Notice that the real axis $\mathbb{R}$ is *not* included in $\mathbb{H}$. Clearly the point ∞ is also at infinite distance from any point in $\mathbb{H}$. We view $\mathbb{R} \cup \infty$ as a circle, known as the *boundary (or circle) at infinity*.

1.1 Isometries of $\mathbb{H}$

To understand a geometry and its tilings we need to understand its *isometries*, that is, its distance preserving maps. The isometries of $\mathbb{H}$ have a very nice description in terms of the group $SL(2, \mathbb{R})$. This is the group of 2×2 matrices with real entries and determinant 1, i.e. $\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{R}, ad - bc = 1 \right\}$. $SL(2, \mathbb{R})$ acts on $\mathbb{H}$ in the following way. Let $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R})$ and $z \in \mathbb{H}$. Then $T(z) = (az + b)/cz + d)$. By convention, $T(\infty) = a/c$ and $T(-d/c) = \infty$.

Exercise 1.2. *Show that:*

a. if $\Im z > 0$ then $\Im(az + b)/cz + d) > 0$.

b. T maps the circle at infinity to itself.

c. *if* $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *and* $T' = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$ *then* $T'(T(z)) = (T'T)(z)$, *where* $T'T$ *is the matrix product of* T' *with* T *and* $T'(T(z))$ *is the image of* $T(z)$ *under* T'.

d. *if* $(az+b)/(cz+d) \equiv z$ *then* $a = d = \pm 1, b = c = 0$.

Exercise 1.2 (c) shows that to compose maps we simply need to multiply matrices. (d) shows that the group $PSL(2,\mathbb{R}) = SL(2,\mathbb{R}) / \pm \mathrm{Id}$ (where $\mathrm{Id} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$) acts freely on $\mathbb{H}$, that is, if $T(z) = z$ then $T = \mathrm{id}$ as an element of $PSL(2,\mathbb{R})$. Where it won't lead to confusion, we often use $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ to represent a transformation in $PSL(2,\mathbb{R})$.

Proposition 1.1. *$PSL(2,\mathbb{R})$ acts by isometries on $\mathbb{H}$. In other words, if $T \in PSL(2,\mathbb{R})$, then $d_{\mathbb{H}}(T(P), T(Q)) = d_{\mathbb{H}}(P,Q)$ for any $P, Q \in \mathbb{H}$.*

Proof. To abbreviate, write $|dz| = \sqrt{dx^2 + dy^2}$. Let $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{R})$ and let $w = T(z) = (az+b)/cz+d)$. We claim that $|dw|/\Im w = |dz|/\Im z$ and consequently $\int_\gamma |dz| = \int_{T(\gamma)} |dw|$.

Exercise 1.3. *Finish the proof!*

Linear fractional transformations

A mapping of the form $z \mapsto (az+b)/cz+d)$, where $a, b, c, d \in \mathbb{C}$ and $ad - bc \neq 0$ is called a *linear fractional transformation* or *Möbius map*. Möbius maps carry circles to circles and preserve angles. Here 'circle' is interpreted to mean either an ordinary circle or a line through infinity. For more details and a proof see [11] Chapter 1.

Exercise 1.4. a. *Show that under the action of $PSL(2,\mathbb{R})$, a vertical line in $\mathbb{H}$ is carried either to another vertical line or to a semicircle centred on $\mathbb{R}$.*

b. *Show that* $T = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ *maps the imaginary axis $\mathbb{I}$ to the semicircle with centre $1/2$ joining 0 to 1.*

c. *Let $\xi < \eta \in \mathbb{R}$. Find a map $T \in PSL(2,\mathbb{R})$ which maps 0 to ξ and ∞ to η.*

d. *Why is any semicircle with centre on $\mathbb{R}$ a geodesic (straight line) in $\mathbb{H}$?*

e. *Show that there is a* unique *geodesic joining any two points in $\mathbb{H}$, namely the semicircle through the two points with centre on $\mathbb{R}$.*

The group $SL(2,\mathbb{Z})$

The group $SL(2,\mathbb{Z})$ is the subgroup of $SL(2,\mathbb{R})$ all of whose entries are integers. We define $PSL(2,\mathbb{Z}) = SL(2,\mathbb{Z}) \,/\, \pm Id$.

Exercise 1.5. *a. Show that* $J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ *is the unique non-trivial element of* $SL(2,\mathbb{Z})$ *which maps* $\mathbb{I}$ *to itself.*

b. Check that, as an isometry of $\mathbb{H}$*,* J *has order* 2 *and fixes* i*.*

The 'tiles' of $\mathcal{F}$ are all *ideal triangles*. This means that each tile has three geodesic sides, which meet in pairs on the boundary at infinity, i.e. $\mathbb{R} \cup \infty$. We denote the triangle with vertices $0, 1, \infty$ by Δ, called the *basic triangle*. When we need to be strict, we consider that Δ is the closed triangle including its sides (but excluding the 3 vertices which lie outside $\mathbb{H}$) and we let Δ° denote its interior, that is, Δ excluding its sides.

Exercise 1.6. *a. Find the element* $S \in PSL(2,\mathbb{Z})$ *which sends* $0 \to 1, 1 \to \infty, \infty \to 0$*.*

b. Conclude that the stabiliser of Δ *in* $PSL(2,\mathbb{Z})$ *has order* 3*.*

c. Show that S *has a unique fixed point in* $\mathbb{H}$*, and find it.*

The following proposition allows us to prove the key facts about $\mathcal{F}$.

Proposition 1.2. *The ideal triangles in the Farey tessellation* $\mathcal{F}$ *cover the hyperbolic plane without overlaps (except of their boundaries). Moreover if* $g \in SL(2,\mathbb{Z})$*, then* $g(\Delta)$ *is a triangle in* $\mathcal{F}$*.*

Proof. From the construction, it is clear that every point in $\mathbb{H}$ is contained in at least one (closed) ideal triangle of the construction. We have to show that no two triangles overlap.

First note that every triangle in the tessellation is the image of Δ under some element in $SL(2,\mathbb{Z})$. In fact by Exercise 1.1, if $p/q, r/s$ are joined by an arc of $\mathcal{F}$ and if we assume that $p/q > r/s$ then $\det \begin{pmatrix} p & r \\ q & s \end{pmatrix} = 1$ so that $T = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2,\mathbb{Z})$. By Exercise 1.4, T carries the positive imaginary axis $\mathbb{I}$ to the hyperbolic line joining p/q to r/s, in other words, the semicircle with these endpoints. Moreover T carries 1 to the point $p/q \oplus r/s$ so that it takes the other two sides of Δ to semicircular arcs joining these new neighbours.

Let $\mathcal{T}$ be the set of triangles in $\mathcal{F}$. If $E \in \mathcal{T}$, let E° denote its interior. We have to show that $E_1^\circ \cap E_2^\circ = \emptyset$ for any $E_1, E_2 \in \mathcal{T}$. We have just shown that $E_i = g_i(\Delta)$ for some $g_i \in SL(2,\mathbb{Z})$. So it is enough to show that $\Delta^\circ \cap g(\Delta^\circ) = \emptyset$ for any $g \in SL(2,\mathbb{Z})$. (Why?) Let $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ so that $a/c > b/d$. By translating and rotating Δ if needed (using the transformation S of Exercise 1.6), we may assume that the side of $g(\Delta)$ joining

a/c to b/d cuts the imaginary axis $\mathbb{I}$ (why?), so that $a/c > 0 > b/d$. We claim this is impossible for $g \in SL(2,\mathbb{Z})$. Note that without loss of generality we can take $d > 0$, (why?) so automatically $b < 0$. Then a, c have the same sign. If both are positive then $1 = ad - bc \geq 1 + 1 = 2$ which is impossible. The other case is similar.

The same argument shows that $g(\Delta) \in \mathcal{T}$ for any $g \in SL(2,\mathbb{Z})$. This completes the proof. □

Here are some important consequences of Proposition 1.2.

Corollary 1.3. *1. Every pair of neighbouring rationals are the endpoints of some side of $\mathcal{F}$.*

2. Every point $p/q \in \mathbb{Q}$ is a vertex of $\mathcal{F}$.

3. The Farey tessellation $\mathcal{F}$ is invariant under the action of $PSL(2,\mathbb{Z})$.

Exercise 1.7. *Prove Corollary 1.3. Hint for (2): Use the Euclidean algorithm!*

Exercise 1.8. *a. Let $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ as in Exercise 1.5. Suppose that $g \in PSL(2,\mathbb{Z})$ carries $\mathbb{I}$ to another side s of $\mathcal{F}$, so that $g(i) \in s$. Prove that gJg^{-1} is the unique non-trivial element in $PSL(2,\mathbb{Z})$ which fixes s.*

b. Find an element $g \in SL(2,\mathbb{Z})$ which carries $\mathbb{I}$ to the hyperbolic line from 0 to 1 and hence or otherwise, find the unique non-trivial element of $PSL(2,\mathbb{Z})$ which fixes the point $(1+i)/2$.

Exercise 1.9. *a. Explain why J maps any hyperbolic line through i to itself, interchanging endpoints.*

b. With g as in Exercise 1.8, prove that $T = gJg^{-1}J = \pm \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ maps the hyperbolic line L joining i to $(1+i)/2$ to itself. Hint: T is the product of two π rotations about points on L.

c. What are the end points of this line? Check they are fixed by T.

We will come back to this transformation T later.

Finally, here is an exercise on hyperbolic geometry which we will need in the last lecture.

Exercise 1.10. *a. Let H be the region above the horizontal line $\Im z = h$. Explain why the image of H under $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is the region inside a disk tangent to $\mathbb{R}$ at a/c.*

b. Prove that the radius of this disk is $1/2hc^2$. Hint: Suppose the disk has radius r, so its highest point is $a/c + 2ir$. Explain why h is the imaginary part of $T^{-1}(a/c + 2ir)$ and hence find the formula relating r and h.

Fundamental domain for $SL(2,\mathbb{Z})$.

According to Exercise 1.6, the basic tile Δ is fixed by a non-trivial element S in $SL(2,\mathbb{Z})$, where $S^3 = \mathrm{id}$. (The element S was computed in Exercise 1.6.) A *fundamental domain* for the action of a group G on $\mathbb{H}$ is a region R such that the images of the closure of R cover the plane, and such that $\mathrm{Int}\, R \cap g(\mathrm{Int}\, R) = \emptyset$ for every non-trivial element of G. To find a fundamental domain for $SL(2,\mathbb{Z})$ we have to subdivide Δ into three parts, which are mapped one onto the other by S. This is illustrated in Figure 2 in which the basic triangle Δ is subdivided into three four sided regions, each of which is a fundamental domain for the action of $SL(2,\mathbb{Z})$ on $\mathbb{H}$. The uppermost dotted arcs are parts of semicircles of radius 1. The three dotted arcs in Δ meet in the point $(1+i\sqrt{3})/2$ and $S = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}$ rotates the three quadrilaterals in Δ one onto another.

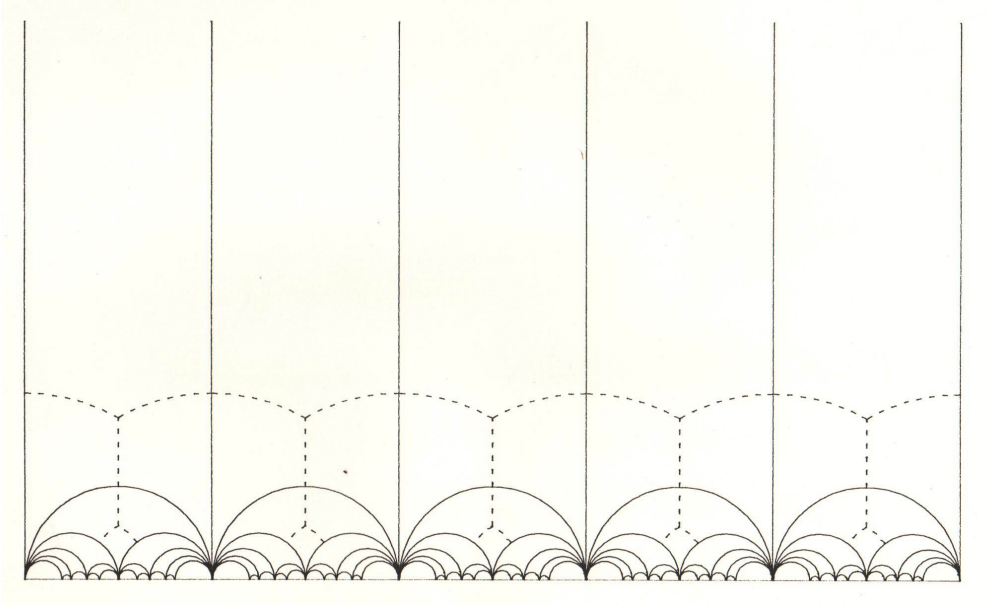


Figure 2: Fundamental domain for the action of $SL(2,\mathbb{Z})$ on $\mathbb{H}$.

2 The Farey tessellation and continued fractions

A continued fraction is an expression

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots}}}$$

where $x \in \mathbb{R}$ and $a_i \in \mathbb{N}$. Usually we shorten this to $x = [a_0; a_1, a_2, a_3, \ldots]$. We also require that $a_i > 0$ for all $i \geq 1$. (For conventions on negative numbers x, see also 2.0.1 below.) It is not hard to see that any rational x can be expressed in this way. For example, if $x = 3/4$ then: $1/x = 4/3 = 1 + 1/3$ so $x = [0; 1, 3]$ or alternatively $x = [0; 1, 2, 1]$. If $x \notin \mathbb{Q}$ then a similar procedure leads to an infinite continued fraction.

Example 2.1. Suppose $x = (1 + \sqrt{5})/2$. Recall that if $y \in \mathbb{R}$ then $[y]$ denotes the *integer part* of y, that is $y = [y] + t$ where $0 \leq t < 1$. Since $1 < x < 2$ we have $[x] = 1$ and $x - 1 = (-1 + \sqrt{5})/2$ so that $1/(x-1) = 2/(-1+\sqrt{5}) = 2(1+\sqrt{5})/4 = x$. Unwinding we get $x = 1 + 1/x$ from which it follows that $x = [1; 1, 1, 1, \ldots]$.

Exercise 2.1. *a. Show that* $\sqrt{2} = [1; 2, 2, 2, 2, \ldots]$. *Hint: Use* $(\sqrt{2}+1)(\sqrt{2}-1) = 1$.

b. Show that $x > 0$ is rational iff its continued fraction terminates. Show that in this case there are always two continued fraction expressions for x, namely $[a_0; a_1, \ldots, a_n]$ with $a_n > 1$ and $[a_0; a_1, \ldots, a_n - 1, 1]$

c. Use a calculator to find the first few terms in the continued fraction expansion for π.

The continued fraction for $x \in \mathbb{R}$ can be read off from the Farey tessellation $\mathcal{F}$ as follows. Join x to any point on the imaginary axis $\mathbb{I}$ by a hyperbolic geodesic (semicircular arc) γ. This arc cuts a succession of tiles of $\mathcal{F}$. Each tile is an ideal triangle, so γ cuts exactly two sides of $\mathcal{F}$. These sides meet in a vertex which is either to the left, or to the right, of the oriented arc γ. Label this segment of γ by L or R accordingly. (In the exceptional case in which γ terminates in a vertex of the triangle, choose either label.) The resulting sequence $L^{n_0}R^{n_1}L^{n_3}\ldots, n_1 \in \mathbb{N}$ is called the *cutting sequence* of x. If $x > 1$ the sequence begins with L, while if $0 < x < 1$ the sequence begins with R. (For $x < 0$ see 2.0.1). Note that the cutting sequence is independent of the initial point of γ on $\mathbb{I}$. The key observation is:

Proposition 2.2. *Let $x > 1$ have cutting sequence $L^{n_0}R^{n_1}L^{n_3}\ldots, n_i \in \mathbb{N}$. Then $x = [n_0; n_1, n_2, \ldots]$. Likewise if $0 < x < 1$ has cutting sequence $R^{n_1}L^{n_3}\ldots, n_i \in \mathbb{N}$ then $x = [0; n_1, n_2, \ldots]$.*

Proof. Let γ be an oriented geodesic starting at a point on $\mathbb{I}$ and ending at x, thus defining the cutting sequence of x. Assume that $x > 1$. Since $[x] = n_0 > 0$ we see that γ begins its descent to $\mathbb{R}$ (i.e. first cuts a non-vertical side of $\mathcal{F}$) in the interval $n_0 \leq x < n_0 + 1$. Thus the cutting sequence of γ begins $L^{n_0}R$.

Let $P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. By Corollary 1.3, both P and J map $\mathcal{F}$ to itself preserving orientation[1]. Thus γ cuts an ideal triangle T with the same symbol as $P(\gamma)$ cuts

[1]This means that the order, clockwise or anticlockwise, of points round the boundary of a circle or triangle is preserved by the action of any element in $SL(2,\mathbb{Z})$. It works because every element in $SL(2,\mathbb{Z})$ has positive determinant.

$P(T)$ and similarly for J, in other words, both P and J preserve cutting sequences as long as we keep track of starting points correctly.

Let $u_{-1}, u_0, u_1, \ldots$ be unit vectors based at the points $z_{-1}, z_0, z_1, \ldots$ on sides of $\mathcal{F}$ which mark the changes from L segments to R segments, and pointing along γ, where $z_{-1} \in \mathbb{I}$. (We start the numbering from -1 for convenience later.)

Consider the effect of JP^{-n_0} on γ. Note that $P^{-n_0}(z_0) \in \mathbb{I}$. The geodesic $P^{-n_0}(\gamma)$ starting from $P^{-n_0}(z_0)$ ends at $P^{-n_0}(x) \in [0,1)$, hence the initial term of the cutting sequence of $P^{-n_0}(\gamma)$ is R. Now J maps $\mathbb{I}$ to itself and $J(x-n_0) = -1/(x-n_0) < -1$, while $JP^{-n_0}(z_0)$ points into the left half plane. From the continued fraction expansion of x, $[1/(x-n_0)] = n_1$ and so $-(n_1+1) < -1/(x-n_0) \leq -n_1$. This means that the cutting sequence of $JP^{-n_0}(\gamma)$, starting from the point $JP^{-n_0}(u_0)$ where it crosses $\mathbb{I}$, is $R^{n_1}L$.

Now apply P^{n_1} followed by J. The geodesic $JP^{n_1}JP^{-n_0}(\gamma)$ meets $\mathbb{I}$ at $JP^{n_1}JP^{-n_0}(z_1)$ and ends in the point $1/(1/(x-n_0)-n_1) > 1$. From the continued fraction expansion of x we have $[1/(1/(x-n_0)-n_1)] = n_2$. Therefore $JP^{n_1}JP^{-n_0}(\gamma)$ has a cutting sequence which starts $L^{n_2}R$. This sequence is the same as the sequence of γ read starting from z_1, so γ itself has sequence $L^{n_0}R^{n_1}L^{n_2}R\ldots$. Now the argument repeats.

The reasoning for $0 < x < 1$ is similar. □

Example 2.3. By Exercise 1.9, $T = \begin{pmatrix} 1 & 1 \\ -2 & 1 \end{pmatrix}$ maps the hyperbolic line α joining i to $(1+i)/2$ to itself. Then α cuts $\mathbb{I}$ and ends at the point $(\sqrt{5}-1)/2 \in \mathbb{R}$. As in the exercise, T is the product of order two rotations each of which are symmetries of $\mathcal{F}$. It is not hard to see that every time α cuts a side s of $\mathcal{F}$, it does so in the unique $SL(2,\mathbb{Z})$ image of i on s (see Exercise 1.8). Reading along α, we see that starting from i, α has cutting sequence $LRLRLR\ldots$. Hence $(\sqrt{5}-1)/2 = [0;1,1,1,\ldots]$, agreeing with Example 2.1.

Example 2.4. We will show that $\sqrt{2} = [1;2,2,2,\ldots]$ (c.f. Exercise 2.1). Let $x = \sqrt{2}+1$ and let γ be the semicircle centre 1 radius $\sqrt{2}$, with endpoints $\pm x$. The segment of γ joining the point where it cuts $\mathbb{I}$ to x is the cutting sequence of x. Clearly the sequence begins LLR. Applying JP^{-2} we find $JP^{-2}(x) = 1/(1-\sqrt{2}) = -(1+\sqrt{2}) = -x$. By symmetry the cutting sequence from $\mathbb{I}$ to $-x$ begins RRL. Thus x has sequence $LLRRL\ldots$. Now applying JP^2 we get $JP^2JP^{-2}(x) = x$. Thus x has the sequence $LLRRLLRR\ldots$ and hence $1+\sqrt{2} = [1;2,2,2,\ldots]$.

2.0.1 Negative numbers

To handle negative numbers there are differing conventions. In [6], negative numbers are written $x = [a_0;a_1,a_2,a_3,\ldots]$ with $a_0 < 0$ and $a_i > 0$ for $i \geq 1$. Dealing with cutting sequences, we observe that x and $J(x)$ have the same cutting sequence and so it is usually enough to replace a negative number y by $-1/y = [b_0;b_1,b_2,b_3,\ldots]$ with $b_0 \geq 0$.

More on continued fractions

Let $x = [a_0; a_1, a_2, \ldots]$ and $p_n/q_n = [a_0; a_1, a_2, \ldots, a_n]$. The fractions p_n/q_n (always assumed to be in lowest terms) are called the *convergents* of x. They can be interpreted in terms of the passage of a geodesic γ ending at x across $\mathcal{F}$. We shall see that $p_n/q_n, p_{n+1}/q_{n+1}$ are the two ends of the side of $\mathcal{F}$ cut by γ at the moment the sequence of γ changes from L to R or vice versa, so that x lies between p_n/q_n and p_{n+1}/q_{n+1}.

We start with the following theorem, which contains some basic facts about continued fractions.

Theorem 2.5. *Let $x = [a_0; a_1, a_2, \ldots]$ and $p_n/q_n = [a_0; a_1, a_2, \ldots, a_n]$, so particular $p_0 = a_0, q_0 = 1$. Then*

1. $\begin{pmatrix} p_n & p_{n+1} \\ q_n & q_{n+1} \end{pmatrix}$ *has determinant ± 1 for all $n \geq 0$.*
2. *For $n \geq 1$, $p_{n+1} = a_{n+1}p_n + p_{n-1}$, $q_{n+1} = a_{n+1}q_n + q_{n-1}$ where $p_{-1} = 1, q_{-1} = 0$.*
3. *$p_{2n}/q_{2n} \leq p_{2n+2}/q_{2n+2} \leq x \leq p_{2n+1}/q_{2n+1} \leq p_{2n-1}/q_{2n-1}$ for all $n \geq 0$, with equality on one or other side iff x is rational and the sequence terminates.*

Proof. For definiteness take $x > 0$ and as above let $P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Then $P^{-a_0}(x) = [0; a_1, a_2, \ldots]$ and $JP^{-a_0}(x) = -[a_1; a_2, \ldots]$. Likewise $JP^{a_1}JP^{-a_0}(x) = [a_2; a_3, \ldots]$ and so on. Let $M = JP^{-a_{2n}} \ldots JP^{a_1}JP^{-a_0}$ (so that M depends on $a_0, \ldots, a_{2n}$). By unwinding the continued fraction, one sees that for any ξ,

$$M([a_0; a_1, a_2, \ldots, a_{2n} + \xi]) = \xi. \tag{1}$$

Setting $\xi = 0, \infty$ respectively, we find $M([a_0; a_1, a_2, \ldots, a_{2n}]) = 0$ and $M([a_0; a_1, a_2, \ldots, a_{2n} + \infty]) = M([a_0; a_1, a_2, \ldots, a_{2n-1}] = \infty$. Thus $M^{-1}(0) = p_{2n}/q_{2n}$ and $M^{-1}(\infty) = p_{2n-1}/q_{2n-1}$. We deduce that $M^{-1} = \begin{pmatrix} \lambda p_{2n-1} & \lambda p_{2n} \\ \lambda q_{2n-1} & \lambda q_{2n} \end{pmatrix}$ for some $\lambda \in \mathbb{R}$. Since $p_{2n}, q_{2n}, \lambda p_{2n}, \lambda q_{2n} \in \mathbb{Z}$ and since p_{2n}, q_{2n} have no common factor we deduce that $\lambda \in \mathbb{Z}$ (why?). Moreover $\det M = 1$ so that $\lambda^2 = 1$ hence $\lambda = \pm 1$. So $M^{-1} = \pm \begin{pmatrix} p_{2n-1} & p_{2n} \\ q_{2n-1} & q_{2n} \end{pmatrix} \in SL(2, \mathbb{Z})$.

Now the claims of the theorem follow easily. (1) holds because $\det M = 1$; this also shows that $p_{2n}/q_{2n} < p_{2n-1}/q_{2n-1}$. To prove (2), note that

$$p_{2n+1}/q_{2n+1} = [a_0; a_1, a_2, \ldots, a_{2n}, a_{2n+1}] = [a_0; a_1, a_2, \ldots, a_{2n} + 1/a_{2n+1}]$$

so by Equation (1) with $\xi = 1/a_{2n+1}$ we have $M^{-1}(1/a_{2n+1}) = p_{2n+1}/q_{2n+1}$. On the other hand $M^{-1}(1/a_{2n+1}) = (p_{2n-1} + a_{2n+1}p_{2n})/(q_{2n-1} + a_{2n+1}q_{2n})$. Comparing coefficients and using the fact that all the integers involved are relatively prime, the result follows.

To prove (3) note that $x = [a_0; a_1, a_2, \ldots, a_{2n} + t]$ where $0 \leq t < 1 < \infty$. Now M^{-1} maps the interval $[0, \infty)$ to the interval $[p_{2n}/q_{2n}, p_{2n-1}/q_{2n-1})$ and so $M^{-1}(0) \leq M^{-1}(t) < M^{-1}(\infty)$. But $M^{-1}(t) = x$ by Equation (1) and the result follows. □

Corollary 2.6. *With the notation of Theorem 2.5, $p_n/q_n \to x$ as $n \to \infty$. Moreover $|x - p_n/q_n| \leq 1/q_n q_{n+1}$ and $q_n \geq n$ for all $n \in \mathbb{N}$.*

The following result interprets Theorem 2.5 in terms of cutting sequences.

Corollary 2.7. *Let $s_0, s_1, s_2, \ldots$ be the sides of $\mathcal{F}$ which mark the changes in the cutting sequence of γ from L to R and vice versa, starting with s_0 being the vertical line from a_0 to ∞. Then for $n \geq 0$, the endpoints of s_{2n} are the points $p_{2n}/q_{2n} < p_{2n-1}/q_{2n-1}$ while the endpoints of s_{2n+1} are the points $p_{2n}/q_{2n} < p_{2n+1}/q_{2n+1}$.*

Proof. By definition s_0 has endpoints a_0 and ∞. As in Theorem 2.5, define $p_{-1} = 1, q_{-1} = 0$ so that $\infty = p_{-1}/q_{-1}$ and $a_0 = p_0/q_0$. Let $z_0, z_1, z_2, \ldots$ be the points where γ cuts sides $s_0, s_1, s_2, \ldots$. After z_0 there are a_1 segments of γ labelled R, up to the point z_1. The left hand end of s_1 is still $a_0 = p_0/q_0$. Using repeated Farey addition, we see the right hand endpoint of s_1 is $(a_1 p_0 + p_{-1})/q_0 = p_1/q_1$ as claimed. Now let's find s_2. The cutting sequence progresses through a_2 segments labelled L. Thus the right hand endpoint of s_2 is still p_1/q_1. The left hand endpoint moves through a_2 steps from p_0/q_0 towards p_1/q_1. Thus its endpoint is $(p_0 + a_2 p_1)/(q_0 + a_2 q_1) = p_2/q_2$ by Theorem 2.5, proving our claim. Continuing in the way completes the proof. □

Corollary 2.6 explains why (p_n/q_n) are called the *convergents* of x while Corollary 2.7 gives a nice geometrical interpretation. The relations $|x - p_n/q_n| \leq 1/q_n q_{n+1}$ and $q_{n+1} = a_{n+1} q_n + q_{n-1}$ show that $|x - p_n/q_n| \leq 1/q_n^2 a_{n+1} \leq 1/q_n^2$. If a_{n+1} is large, p_n/q_n gives an extremely good approximation to x. For example, $\pi = [3; 7, 15, 1, 292, 1, \ldots]$ has convergents $3, 22/7, 333/106, 355/113, \ldots$. In particular $p_3/q_3 = 355/113 = 3.141592653\ldots$ and $|\pi - 355/113| < 1/(292 \times (113)^2) < 10^{-6}$.

3 Applications

3.1 Equivalence under $SL(2, \mathbb{Z})$

Two numbers $x = [a_0; a_1, \ldots], y = [b_0; b_1, \ldots]$ are said to *have the same tails* if there exist $k, l \in \mathbb{N}$ such that $a_{k+r} = b_{l+r}$ for all $r \geq 1$. They have the same tails mod 2 if $k + l$ is even. To see the significance of this idea, we need the following lemma:

Lemma 3.1. *Let γ, γ' be oriented geodesics in $\mathbb{H}$ with the same positive endpoint x. Then the cutting sequences of γ, γ' eventually coincide.*

Proof. We can obviously assume that $x \notin \mathbb{Q}$. Pick a side s of $\mathcal{F}$ which is cut by both γ and γ'. Use the method of Proposition 2.2 to find $A \in SL(2,\mathbb{Z})$ such that $A(s) = \mathbb{I}$. Note that the unit tangent vectors u, u' along γ, γ' pointing towards x map under A to unit vectors pointing from $\mathbb{I}$ into the same half plane (right if $x > 0$, left otherwise). The result now follows from the observation that, starting from $\mathbb{I}$, the cutting sequence of $x > 0$ is independent of the initial point on $\mathbb{I}$ of the choice of geodesic on $\mathbb{I}$ ending at x. □

Now we can prove a classic result about continued fractions, see for example [6].

Proposition 3.2. *$x = [a_0; a_1, \ldots], y = [b_0; b_1, \ldots]$ have the same tails mod 2 if and only if there exists $g \in SL(2,\mathbb{Z})$ such that $g(x) = y$.*

Proof. Suppose x, y have the same tails and $k = 2n, l = 2m$. Then

$$JP^{a_{2n-1}} \ldots JP^{a_1} JP^{-a_0}(x) = [a_{2n}; a_{2n+1}, \ldots] = JP^{b_{2m-1}} \ldots JP^{b_1} JP^{-b_0}(y).$$

If both k, l are odd the argument is similar. So if $k+l$ is even, $g(x) = y$ for some $g \in SL(2,\mathbb{Z})$. We leave the case $k + l$ odd to the reader.

Now suppose $g(x) = y$ for some $g \in SL(2,\mathbb{Z})$. Assume first $x, y > 0$. Pick $w < 0$ and let γ_x, γ_y be the oriented geodesics with endpoints (w, x) and (w, y) respectively, so that γ_x, γ_y cut $\mathbb{I}$ in points z_x, z_y respectively. Thus the continued fraction expansions of x, y can be read off from the cutting sequences of γ_x, γ_y, starting from z_x, z_y. Now $g(\gamma_x)$ ends in the point $g(x) = y$, so by Lemma 3.1 the cutting sequences of $g(\gamma_x), \gamma_y$ eventually coincide. On the other hand, after choosing appropriate starting points, the cutting sequence of $g(\gamma_x)$ is the same of that of γ_x. So the tails of the cutting sequences of γ_x, γ_y must agree. Moreover $k + l$ must be even because of the alternating L, R symbols in the cutting sequences must match.

If $x < 0$ then we can replace it by $J(x) > 0$ and apply the same argument, noting that $J(x)$ has the same cutting sequence as x and also that the tail of the continued fraction expansion of $J(x) = -1/x$ is the same as that of x, and similarly for y. □

3.2 Periodic continued fractions

It is not hard to see algebraically that any number whose continued fraction is eventually periodic is quadratic[2]. Conversely, any quadratic number has an eventually periodic expansion. This result, due to Lagrange, is slightly tricky to prove, see for example [6], [4] or [9][3].

Example 3.3. The continued fraction of $\sqrt{n}$ has a particularly nice form: its continued fraction coefficients are palindromic. The reason for this can be understood by considering the semicircle centre 0 and radius $\sqrt{n}$. It can be proved that γ is mapped to itself

[2] A quadratic number is one which satisfies a quadratic equation with integer coefficients.

[3] There is a gap in the proof in [9] which is corrected in [12] §5.4.3.

by some element of $SL(2,\mathbb{Z})$. Why does it follow that the whole doubly infinite cutting sequence of γ is periodic? Now using the symmetry in the imaginary axis, show that $\sqrt{n} = [a_0, \overline{a_1, a_2, \ldots a_2, a_1, 2a_0}]$ where the overline indicates infinite repetition. As an example, check that $\sqrt{7} = [2; \overline{1, 1, 4}]$.

3.3 Diophantine approximation

Corollary 2.6 gives a hint that continued fractions give good approximants to irrationals. This is the subject of *Diophantine approximation*. In fact, the best rational approximation to an irrational for a given bound on the denominator is given by the convergents of its continued fraction, see for example [7] and [6].

Diophantine approximation has a beautiful geometrical interpretation. Since we don't have time to go into this in detail here, let us focus on the example of the golden mean $\omega = (1+\sqrt{5})/2$. The key is to look at *horocycles*, that is, circles in $\mathbb{H}$ tangent to the real axis, or horizontal lines in $\mathbb{H}$ (which can be interpreted as circles tangent to ∞), and to use the result of Exercise 1.10. A *horodisk* is the region enclosed by the horocycle.

Lemma 3.4. *Let $\omega = (1+\sqrt{5})/2 = [1; 1, 1, 1, \ldots]$ and let p_n/q_n be its convergents. Then*

$$\inf\{c : |\omega - p_n/q_n| \le c/q_n^2 \text{ for infinitely many } n\} = 1/\sqrt{5}.$$

Proof. We want to investigate $|\omega - p_n/q_n|$. Define the *height* $ht(\gamma)$ of a hyperbolic geodesic γ to be its maximum Euclidean height above $\mathbb{R}$, that is, its Euclidean radius. Let α be the geodesic with endpoints $(1 \pm \sqrt{5})/2$ studied in Exercise 1.9. Every time α crosses a tile of $\mathcal{F}$, it enters and leaves through one of the $SL(2,\mathbb{Z})$ images of the special points i. The same must be true of every image $g(\alpha), g \in SL(2,\mathbb{Z})$. Therefore $\sup\{ht(g(\alpha) : g \in SL(2,\mathbb{Z})\} = ht(\alpha) = \sqrt{5}/2$.

Let H be the open horodisk bounded by the line $\Im z = \sqrt{5}/2$ and let $p_{n-1}/q_{n-1}, p_n/q_n$ be successive convergents to ω. Assume for definiteness that n is even; the argument if not is similar. Then the image of H under $A = \begin{pmatrix} p_{n-1} & p_n \\ q_{n-1} & q_n \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix} \in SL(2,\mathbb{Z})$ is a horodisk tangent to $\mathbb{R}$ at p_n/q_n. By Exercise 1.10, it has radius $1/\sqrt{5}q_n^2$. (The first matrix in this product sends ∞ to the odd convergent p_{n-1}/q_{n-1}, and 0 to p_n/q_n. To fix this we first apply S which sends ∞ to 0.) Since $A^{-1}(\alpha) \cap H = \emptyset$ we have $\alpha \cap A(H) = \emptyset$. Therefore the vertical line V_ω in $\mathbb{H}$ ending at ω does not intersect $A(H)$. (Why?) It follows that $|\omega - p_n/q_n| \ge 1/\sqrt{5}q_n^2$. Since this holds for all convergents, we have $c(\omega) \ge 1/\sqrt{5}$.

Let $T = \begin{pmatrix} 1 & 1 \\ -2 & 1 \end{pmatrix}$. By Exercise 1.9, T maps α to itself and no power of T fixes ∞. Let H' be a horizontal horocycle of height $\sqrt{5}/2 - \epsilon$ where $\epsilon > 0$. Then $T^k(\alpha) \cap H' \neq \emptyset$ for $k \in \mathbb{N}$. So $\alpha \cap T^{-k}(H') \neq \emptyset$ for $k \in \mathbb{N}$. Since T^k does not fix ∞, we see that $E = T^{-k}(H')$ is a horocycle of larger radius than $T^{-k}(H)$ based at p_k/q_k for some $p_k/q_k \in \mathbb{Q}$. We would like

to claim that, since α intersects E, so does V_ω, because it would follow that for some $d > 1$ (depending on ϵ), $|\omega - p_k/q_k| \leq d/\sqrt{5}q_k^2$. This is not quite correct. However for any $p \in \alpha$, the points $T^n(p)$ approach ω as $n \to \infty$. Moreover α and V_ω are asymptotic. Choose $p \in \alpha$ so that a small neighbourhood B of p is contained in E. Then $T^n(p) \in T^n(B) \subset T^n(E)$. Note that the sets $T^n(B)$ all have the same diameter since T is an isometry. It follows that for large enough n, $V_\omega \cap T^n(B) \neq \emptyset$ so that also $V_\omega \cap T^n(E) \neq \emptyset$. We conclude that for any $d > 1$, $|\omega - p_n/q_n| \leq d/\sqrt{5}q_n^2$ for infinitely many $n \in \mathbb{N}$. See [5] for details. The result follows. □

The number $1/\sqrt{5}$ is called the *Hurwitz constant*. This is just the beginning of the story. For $x \in \mathbb{R}$, let

$$c(x) = \inf\{c : |x - p/q| \leq c/q^2 \quad \text{for infinitely many } p/q\}.$$

It can be shown that $c(x)$ depends only on the equivalence class of x modulo $SL(2, \mathbb{Z})$, that is, on the tail of the continued fraction of x, see [**?**] Chapter 7 Lemma 1. We have just shown that $c(\omega) = 1/\sqrt{5}$. In fact $\omega = (1 + \sqrt{5})/2 = [1; 1, 1, 1, \ldots]$ and its images under $SL(2, \mathbb{Z})$ are the *worst approximated* numbers, in the sense that for any number whose tail does not end in an infinite string of 1's, $c(x) < c(\omega)$.

The set of all possible values of $c(x), x \in \mathbb{R}$ is called the *Lagrange spectrum*. The following lemma shows there is a definite gap in the spectrum between the Hurwitz constant and the next possible value of $c(x)$:

Lemma 3.5. *Let $x = [a_0; a_1, a_2, a_3, \ldots]$ and suppose that $a_n \neq 1$ for infinitely many n. Then $c(x) \leq 1/2$.*

Proof. Let γ be a geodesic starting on $\mathbb{I}$ and ending at x. The hypothesis shows that the cutting sequence of γ contains L^2 or R^2 infinitely often. Thus infinitely many of the images of γ have height at least 1. An argument similar to that above shows that $c(x) \leq 1/2$. For details see [10] or [5]. □

It turns out there is a very special sequence of quadratic numbers x_n for which $c(x_n)$ decreases monotonically to $1/3$, while there are uncountably many $SL(2, \mathbb{Z})$ classes of points x with $c(x) = 1/3$. This all has a beautiful geometrical explanation, see [10]. One can also show by geometrical arguments that there exists a number $c_0 > 0$ such that the Lagrange spectrum contains the entire interval $(0, c_0]$. These points form what is called the *Hall ray*, see [8] and [1]. The precise best value of c_0 was discovered by G. Freiman in 1973:

$$1/c_0 = 4 + \frac{253589820 + 283748\sqrt{462}}{491993569} = 4 + [0, 3, 2, 1, 1, \overline{3, 1, 3, 1, 2, 1}] + [0, 4, 3, 2, 2, \overline{3, 1, 3, 1, 2, 1}].$$

For more information and references on Diophantine approximation, see [3].

3.4 Dense geodesics and ergodic theory

The connection between continued fractions and geodesics in the hyperbolic plane was used in a famous paper by E. Artin [2] to provide the first example of a geodesic trajectory on a Riemann surface with a dense trajectory. Here is a quick sketch of how it works. Consider the surface Σ obtained by gluing the sides of the fundmental domain for $SL(2,\mathbb{Z})$ shown in Figure 2. The two vertical sides are matched by the transformation $P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ while the two finite sides are matched by the rotation $S = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}$, see Exercise 1.6. (Check that $S^2(i) = i + 1$.) This produces a *hyperbolic surface* Σ which is topologically a sphere with one missing point, namely the point at infinity is a 'missing point' or *cusp*. There are two other special points which are *cone points*, one of order 2 and one of order 3, corresponding to the points i and $(1+i)/2$ respectively. (For more background on hyperbolic surfaces and covering spaces, see [11].)

The sides of $\mathcal{F}$ all project to the geodesic $\mathcal{L}$ on Σ which joins the cusp to the order 2 cone point; $\mathcal{L}$ is the image of the line in $\mathbb{H}$ joining i to ∞. (The segment from i to 0 also projects to $\mathcal{L}$.) Every other infinite oriented geodesic $\bar{g}$ on Σ must cross $\mathcal{L}$ infinitely often. Starting from a point where $\bar{g}$ cuts $\mathcal{L}$, $\bar{g}$ lifts to a unique geodesic γ in $\mathbb{H}$ which cuts $\mathbb{I}$ on the line between i and ∞. After possibly applying J and a suitable power of P, we can arrange that γ has endpoints $\eta \in (-1, 0]$ and $\xi \in [1, \infty)$. Let $\xi = [a_0; a_1, a_2, \ldots]$ and $-1/\eta = [b_0; b_1, \ldots]$. Represent the geodesic with endpoints (η, ξ) by the doubly infinite sequence $\ldots, b_1, b_0, |a_0, a_1, \ldots$ where the bar $|$ indicates the position where γ cuts $\mathbb{I}$. Using the ideas we have already outlined, one can show that γ' is another lift of $\mathcal{L}$ with its endpoints (η', ξ') in the same intervals iff its representative sequence $\ldots, b'_1, b'_0, |a'_0, a'_1, \ldots$ is a shift of $\ldots, b_1, b_0, |a_0, a_1, \ldots$[4].

Observe that if geodesics γ, δ with endpoints $(\eta, \xi), (\alpha, \beta)$ respectively are close then γ and δ are close for a long section of their trajectories. Equally, (η, ξ) is close to (α, β) if the corresponding sequences agree over a long block surrounding the zero bar $|$.

Now choose a doubly infinite sequence in which every possible finite block of positive integers occurs infinitely often. This sequence represents infinitely many distinct geodesics in $\mathbb{H}$, depending on where we place the zero bar. Denote the set of all such geodesics by Ξ. From the above discussion, one sees that all possible initial positions and directions for geodesics cutting $\mathbb{I}$ are approximated arbitrarily well by geodesics in Ξ. The geodesics in Ξ are all images of one another under $SL(2,\mathbb{Z})$, and hence project to a single geodesic on the surface Σ whose trajectory is dense on Σ.

Introducing measure theory into the above discussion leads to some very important ideas in ergodic theory. For example, the first proof that the geodesic flow on a Riemann

[4]To get the geodesic with the opposite orientation, just read the sequence backwards.

surface can be ergodic[5] was done using this method. We can also use this picture to give an easy derivation of the famous Gauss measure for continued fractions, see [7] for the definitions and classical results and [9] for the proof using hyperbolic geometry. The idea of using symbols to study dynamical systems, of which this is one of the very earliest examples, is fundamental in the theory of chaos.

3.5 More general groups and surfaces

There are far reaching generalisations of many of the above ideas, obtained when the group $SL(2,\mathbb{Z})$ is replaced by any discrete subgroup of $SL(2,\mathbb{R})$ containing a translation $z \mapsto z+c, c \in \mathbb{R}$. The study of Diophantine approximation on the quotient surfaces becomes the study of how far trajectories 'go up the cusp', see for example [5], [8] and [1].

References

[1] M. Artigiani, L. Marchese, C. Ulcigrai. The Lagrange spectrum of a Veech surface has a Hall ray. *Preprint*, arXiv:1409.7023 [math.DS].

[2] E. Artin. Ein mechanisches System mit quasiergodischen Bahnen. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg 1*, 3, 1924, 170 – 175.

[3] T. Cusick and M. Flahive. *The Markoff and Lagrange spectra.* American Math. Surveys and Monographs 30, American Math. Soc., Providence, 1989.

[4] A. Hatcher. *Toplogy of Numbers.* Unpublished draft book, available at `www.math.cornell.edu/~hatcher/TN/TNpage.html`

[5] A. Haas and C. Series. The Hurwitz constant and Diophantine approximation on Hecke groups. *J. London Math. Soc. 2*, 34, 1986, 219 – 334.

[6] G. H. Hardy and E. M. Wright. *The Theory of Numbers.* Oxford University Press, 1938.

[7] A. Ya. Khinchin *Continued Fractions.* University of Chicago Press, 1935.

[8] T. Schmidt and M. Scheingorn. Riemann surfaces have Hall rays at each cusp. *Illinois J. Math. Soc. 3*, 41, 1997, 378 – 397.

[9] C. Series. The modular surface and continued fractions. *J. London Math. Soc. 2*, 31, 1985, 69 – 85.

[10] C. Series. The Geometry of Markoff Numbers. *Math. Intelligencer 7*, 1985, 20 – 29.

[5] This means that the space cannot be decomposed into two invariant measurable sets each of positive measure.

[11] C. Series. *Hyperbolic geometry notes MA448.* Unpublished lecture notes, available at `homepages.warwick.ac.uk/~masbb/`

[12] Geometrical methods of symbolic coding, in *Ergodic Theory and Symbolic Dynamics in Hyperbolic Spaces*, T. Bedford, M. Keane and C. Series eds., Oxford Univ. Press 1991, 125 – 151.

4 Solutions to exercises

Exercise 1.1 Easy.

Exercise 1.2

a. Let $z = x+iy$. Then $\Im(ax+aiy+b)/(cx+ciy+d) = \Im(ax+aiy+b)(cx-ciy+d)/|cz+d|^2 = y/|cz+d|^2$.

b. Obvious using the conventions $T(\infty) = a/c, T(-d/c) = \infty$.

c. $T'(T(z)) = T'(\dfrac{az+b}{cz+d}) = \dfrac{a'(az+b)+b'(cz+d)}{c'(az+b)+d'(cz+d)}$. Check the coefficients in this fraction are the same as the matrix coefficients of the matrix product $T'T$.

d. If $az+b \equiv cz^2+dz \ \forall \ z \in \mathbb{H}$ then $b = 0, a = d, c = 0$. Now use $ad-bc = 1$.

Exercise 1.3 Check $dw = dz/(cz+d)^2$ and $\Im w = \Im z/|cz+d|^2$. (See Exercise 1.2 (a).) For more detail see [11] Ch. 2.

Exercise 1.4

a. $T \in SL(2,\mathbb{R})$ maps lines to lines or circles and preserves angles of intersection. A vertical line in $\mathbb{H}$ is a 'circle' orthogonal to $\mathbb{R} \cup \infty$. By Exercise 1.2 (b), T maps $\mathbb{R} \cup \infty$ to itself. So it maps A vertical line in $\mathbb{H}$ to the part of a line or circle orthogonal to $\mathbb{R} \cup \infty$ in $\mathbb{H}$.

b. $T(0) = 0, T(\infty) = 1$, now use (a).

c. $\begin{pmatrix} \eta & \xi \\ 1 & 1 \end{pmatrix}$ would work except its determinant is $\eta - \xi \neq 1$ (in general). To remedy this divide all the matrix coefficients by $\sqrt{\eta - \xi}$.

d. By (a) and (b) we can find $T \in SL(2,\mathbb{R})$ which carries the semicircle to $\mathbb{I}$. We already know $\mathbb{I}$ is a geodesic and T is an isometry.

e. Given $z_1, z_2 \in \mathbb{H}$, the point where the perpendicular bisector of the line from z_1 to z_2 meets $\mathbb{R}$ is the centre of the required semicircle. It is a geodesic by (c).

Exercise 1.5

a. J interchanges 0 and ∞ so maps $\mathbb{I}$ to itself. If $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z})$ fixes $0, \infty$ then $b = 0 = c$. So $ad = 1$ which since $a, d \in \mathbb{Z}$ forces $a = d = \pm 1$.

b. $J^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ is the identity in $PSL(2,\mathbb{R})$. If $J(z) = z$ then $-1/z = z$ which gives $z = \pm i$; only $i \in \mathbb{H}$.

Exercise 1.6

a. Let $S = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2,\mathbb{Z})$. Then $r/s = 1, (p+r)/(q+s) = \infty, p/q = 0$. So $r = s, q = -s, p = 0$. Using $ps - rq = 1$, $r^2 = 1$ so $S = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}$.

b. Check that $S^3 = -\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Any other non-trivial element fixing Δ must fix one vertex and hence one side. By applying S or S^2 we may assume the side fixed is $\mathbb{I}$. The only non-trivial element fixing $\mathbb{I}$ (Exercise 1.5) is J but J interchanges left and right half planes so does not map Δ to itself.

c. $S(z) = z$ implies $1/(-z+1) = z$ i.e. $z^2 - z + 1 = 0$ so $z = (1 \pm i\sqrt{3})/2$. Exactly one of these solutions lies in $\mathbb{H}$.

Exercise 1.7

a. Let $p/q, r/s$ be neighbours and say $ps - rq = 1$. Then $T = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2,\mathbb{Z})$ and by Proposition 1.2 $T(\mathbb{I})$ is a side of $\mathcal{F}$.

b. Use the Euclidean algorithm to find $a, b \in \mathbb{Z}$ so that $ap - bq = 1$. Then $T = \begin{pmatrix} p & b \\ q & a \end{pmatrix} \in SL(2,\mathbb{Z})$ and $T(\infty) = p/q$. Now use Proposition 1.2.

c. Similar reasoning shows that each triangle in $\mathcal{F}$ maps to another such, and the map on triangles in $\mathcal{T}$ is bijective.

Exercise 1.8

a. $g(\mathbb{I}) = s$ so $gJg^{-1}(s) = gJg^{-1}g(\mathbb{I}) = gJ(\mathbb{I}) = g(\mathbb{I}) = s$. If another non-trivial element h fixed s then $g^{-1}hg$ would fix $\mathbb{I}$. Now use Exercise 1.5.

b. Let $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then $g(0) = b/d = 0$ so $b = 0$; $g(\infty) = a/c = 1$ so $a = c$; $ad - bc = 1$ so $ad = 1$. Since $a, d \in \mathbb{Z}$, $a = d = \pm 1$. Thus one such matrix is $g = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$.

Calculate $gJg^{-1} = \begin{pmatrix} 1 & -1 \\ 2 & -1 \end{pmatrix}$. Check: $gJg^{-1}((1+i)/2) = \dfrac{(1+i)/2 - 1}{1 + i - 1} = \dfrac{1+i}{2}$.

Exercise 1.9

a. J is an order 2 isometry which fixes i, so it is an order 2 rotation on the circle of tangent directions at i, so it must rotate by π. Alternatively, check that a hyperbolic line L passes through i iff its endpoints are $-1/\eta$ and η for some $\eta > 0$. Since J interchanges its endpoints it maps L to itself.

b. Let L be the line through $i, (1+i)/2$. By similar reasoning to (a), gJg^{-1} also maps L to itself. So $T = gJg^{-1}J$ maps L to itself fixing endpoints. The matrix for T is found by multiplying out.

c. Endpoints are $(-1 \pm \sqrt{5})/2$. Either do a direct Euclidean computation or find the fixed points of T: $T(z) = z$ iff $(z+1)/(z+2) = z$ iff $z^2 + z - 1 = 0$.

Exercise 1.10 H is a disk tangent to $\mathbb{R} \cup \infty$ at ∞. T maps circles to circles and preserves angles, so $T(H)$ is a disk tangent to $T(\mathbb{R} \cup \infty) = \mathbb{R} \cup \infty$ at $T(\infty) = a/c$. $T^{-1}(a/c + 2ir)$ must lie on the boundary of H so $\Im T^{-}(a/c + 2ir) = h$. $T^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ so

$$T^{-1}(a/c + 2ir) = \frac{d(a/c + 2ir) - b}{-c(a/c + 2ir) + a} = \frac{1 + 2ircd}{-2irc^2} \text{ and } \Im T^{-}(a/c + 2ir) = \frac{1}{2rc^2}.$$