MARKOV NUMBERS AND THEIR RELATION TO HYPERBOLIC GEOMETRY

Bachelor thesis

Ana Marija Vego
mail: anamarija.vego@math.ethz.ch
ETH Zürich

Abstract. *In this bachelor thesis we will introduce Hurwitz's and Markov's theorems using an approach in hyperbolic geometry based on the paper of B. Springborn. At the end we briefly establish a connection between Markov's theorem and Fricke's trace identity and continued fractions.*

Supervised by:
Prof. Dr. Özlem Imamoglu

Zürich, 2021

Contents

1 Introduction

One of the oldest questions posed in number theory is how well can real numbers be approximated by rational numbers. Classical theorems of Dirichlet, Liouville and Roth from Diophantine approximation show that algebraic and transcendental numbers can be distinguished by how well they can be approximated.

In this bachelor thesis we are going to study two other important theorems of Diophantine approximation, namely the theorems of Hurwitz and Markov. Markov's theorem is a central result that establishes a beautiful connection between indefinite binary quadratic forms, Diophantine approximation and one (seemingly unrelated) Diophantine equation, which is now called Markov's equation.

Our main goal is to approach Hurwitz's and Markov's theorems from the point of hyperbolic geometry.

In the next section we start by giving the classical theorems of Dirichlet, Louville, Roth and Hurwitz. We give 2 proofs of Hurwitz's theorem. The first one uses continued fractions. For the second proof we translate the problem into one in hyperbolic geometry, and prove it from that point of view.

In Section 3 we turn to Markov's theorem. We start by giving the necessary background in hyperbolic geometry to understand and prove the statement. In this section we follow the paper "*The hyperbolic geometry of Markov's theorem on Diophantine approximation and quadratic forms*" by B. Springborn [S].

In the last section we will first prove Fricke's trace identity and relate it to Markov's equation. We finish by giving the relation of Markov's theorem to continued fractions. In this section we rely mainly on the paper "*The Geometry of Markoff Numbers*" by Caroline Series [Caroline S.].

2 Hurwitz's theorem

There are various ways to approximate a real number α, by rational numbers. For example, one can write α in its decimal expansion and only look at finitely many decimal places. In that way one obtains a sequence of rationals, from these decimal numbers, which can be written in the form $\{\frac{p_n}{q_n}\}_{n\in\mathbb{N}}$ and which approximate α arbitrarily well. But for such a sequence, the denominators q_n usually grow rapidly. Now, what if one reformulates the problem as follows:

How well can a real number α be approximated by rational numbers $\frac{p}{q}$, such that the denominator q only varies slightly?

The term *varies slightly* means in this context that $|\alpha - \frac{p}{q}|$ can be well bounded by a bound depending on q. The question is now how strong the bound can be chosen. There is a classical theorem by Dirichlet that gives an answer to that question.

Theorem 2.1 (Dirichlet)**.** *Let $\alpha \in \mathbb{R}$ and $N \in \mathbb{N}$. There exists $\frac{p}{q} \in \mathbb{Q}$ with $q \leq N$ such that*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{qN}$$

Proof: The proof consists of an application of the pigeonhole principle. For a real number β write $\beta = \lfloor\beta\rfloor + \{\beta\}$, where $\lfloor\beta\rfloor$ is the integer part, and $0 \leq \{\beta\} < 1$. Consider the partition

$$[0,1) = \bigcup_{i=0}^{N-1} \left[\frac{i}{N}, \frac{i+1}{N}\right)$$

of the interval $[0,1)$ into N parts, and the $N+1$ numbers $\{\alpha\}, \{2\alpha\}, \ldots, \{(N+1)\alpha\}$. By the pigeonhole principle, there are integers k, ℓ, with $1 \leq k < \ell \leq N+1$, such that $\{k\alpha\}$ and $\{\ell\alpha\}$ lie in the same part, hence $|\{\ell\alpha\} - \{k\alpha\}| < \frac{1}{N}$. Setting $q = \ell - k \leq N, p = \lfloor\ell\alpha\rfloor - \lfloor k\alpha\rfloor$, one has

$$q\alpha = \ell\alpha - k\alpha = \lfloor\ell\alpha\rfloor - \lfloor k\alpha\rfloor + \{\ell\alpha\} - \{k\alpha\} = p + \{\ell\alpha\} - \{k\alpha\}$$

and thus

$$|q\alpha - p| = |\{\ell\alpha\} - \{k\alpha\}| < \frac{1}{N}$$

Division by q yields the result. □

Dirichlet's theorem can be strengthened as follows:

Corollary 2.1.1. *If $\alpha \notin \mathbb{Q}$, then there are infinitely many $\frac{p}{q} \in \mathbb{Q}$ with*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}.$$

Theorem 2.2 (Liouville)**.** *Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be algebraic with degree d. Then there is a constant $C > 0$ such that*

$$\frac{C}{q^d} < \left|\alpha - \frac{p}{q}\right|$$

for all $\frac{p}{q} \in \mathbb{Q}$.

Definition 2.1. A real number α can be ***approximated to order d*** if there exists a constant $C(\alpha)$, depending on α, and infinitely many $\frac{p}{q} \in \mathbb{Q}$ with $|\alpha - \frac{p}{q}| < \frac{C(\alpha)}{q^d}$.

Corollary 2.2.1. *An algebraic number of degree d can be approximated to order at most d.*

Proof of corollary 2.2.1: Let α be an algebraic number of degree d. Let $t > d$ with $\left|\alpha - \frac{p}{q}\right| < \frac{C(\alpha)}{q^t}$ for infinitely many $\frac{p}{q} \in \mathbb{Q}$. By Liouville's theorem there exists a constant $C > 0$ s.t. $\frac{C}{q^d} < \left|\alpha - \frac{p}{q}\right|$ for every $\frac{p}{q}$. Then one has that

$$\left|\alpha - \frac{p}{q}\right| < \frac{C(\alpha)}{q^t}\frac{Cq^d}{Cq^d} < \frac{C(\alpha)}{C}q^{d-t}\left|\alpha - \frac{p}{q}\right|$$

and hence $1 < \frac{C(\alpha)}{C}q^{d-t}$. Since there are infinitely many $\frac{p}{q} \in \mathbb{Q}$ satisfying this inequality, this leads to a contradiction. □

From Liouville's theorem one obtains the following inequalities:

$$\frac{C}{q^d} < \left|\alpha - \frac{p}{q}\right| < \frac{C(\alpha)}{q^{d+\epsilon}}.$$

They imply that $q^\epsilon < \frac{C(\alpha)}{C}$, and so there are only finitely many $\frac{p}{q}$ for any $\epsilon > 0$.

Proof of Liouville's theorem: Let $f(x) = a_d x^d + a_{d-1}x^{d-1} + \cdots + a_0 \in \mathbb{Z}[x]$ with $f(\alpha) = 0$, and let $\beta \neq \alpha$ be a real root of $f(x)$ closest to α, $|\alpha - \beta| = s$. Consider the open interval $(\alpha - s, \alpha + s)$. In case there are no further real roots, take $s = 1$. Now let $\frac{p}{q} \in \mathbb{Q}$ be arbitrary. We distinguish 2 cases:
Case 1. $\left|\alpha - \frac{p}{q}\right| \geq s$. Then one has $\left|\alpha - \frac{p}{q}\right| > \frac{s}{2q^d}$.
Case 2. $\left|\alpha - \frac{p}{q}\right| < s$. Then one has that $\left|f\left(\frac{p}{q}\right)\right| = \left|\frac{a_d p^d + a_{d-1}p^{d-1}q + \cdots + a_0 q^d}{q^d}\right| \geq \frac{1}{q^d}$ By the mean value theorem, there exists $\gamma \in (\alpha - s, \alpha + s)$ with

$$\frac{f\left(\frac{p}{q}\right) - f(\alpha)}{\frac{p}{q} - \alpha} = \frac{f\left(\frac{p}{q}\right)}{\frac{p}{q} - \alpha} = f'(\gamma) \neq 0$$

where γ lies between $\frac{p}{q}$ and α, and we conclude that

$$\left|\alpha - \frac{p}{q}\right| |f'(\gamma)| = \left|f\left(\frac{p}{q}\right)\right| \geq \frac{1}{q^d}.$$

Take any M with $M > |f'(x)|$ for all $x \in (\alpha - s, \alpha + s)$. Then

$$\left|\alpha - \frac{p}{q}\right| > \frac{1}{Mq^d}$$

and $C = \min\left(\frac{s}{2}, \frac{1}{M}\right)$ will fulfill the requirement of the theorem. □

Dirichlet's and Liouville's theorems give a concrete method for distinguishing between algebraic and transcendental numbers: by looking at how well they can be approximated. If α can be approximated to order > 1, it must be irrational, and if α can be approximated to every order n, it is transcendental. Now, given an algebraic number of degree d, can one sharpen Liouville's theorem to an order of approximation *smaller* than d? This question was answered by Roth in "Rational approximations to algebraic numbers", *Mathematika 2*, for which he also received the Fields medal.

Theorem 2.3 (Roth)**.** *Let α be a real number and $\epsilon > 0$. If there are infinitely many $\frac{p}{q} \in \mathbb{Q}$ with*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^{2+\epsilon}}$$

then α is transcendental.

Roth's theorem says that in general one can not improve on the exponent 2. But, keeping the exponent 2 for q, can one *improve the constant 1?*
It turns out we can improve the constant 1 and Hurwitz's theorem describes how.

Theorem 2.4 (Hurwitz)**.** *Let $\alpha \notin \mathbb{Q}$.*
(i) Let $k = \sqrt{5}$. Then there are infinitely many rational numbers $\frac{p}{q}$ such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{kq^2}. \tag{1}$$

(ii) If $k > \sqrt{5}$, then there exist infinitely many irrational numbers (which are dense everywhere on the real axis), and each one satisfies the inequality above only for a finite number of rational fractions.

The classical proof of Hurwitz's theorem relies on continued fractions, which we will review in the next subsection.

2.1 Continued fractions

Let $\frac{p}{q} \in \mathbb{Q}$, $q > 0$. Then one can write $\frac{p}{q}$ in the form of a continued fraction using the Euclidean algorithm:

$$\frac{p}{q} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_k}}} =: [a_0, a_1, \ldots, a_k]$$

with $a_i \in \mathbb{Z}$.
Similarly, one can write an irrational number α in the form of infinite continued fractions:

$$\alpha = [a_0, a_1, \ldots].$$

If one has that

$$\lim_{n \longrightarrow \infty} [a_0, a_1, \ldots, a_n] = \alpha$$

then we call $\frac{p_n}{q_n} := [a_0, \ldots, a_n]$ a *convergent*.
Here I list some useful properties about continued fractions, whose proof can be found in [A].

Lemma 2.1. *Let $a_0, a_1, a_2, \ldots$ be a sequence of integers with $a_i > 0$ for $i \geq 0$, and set $r_n = \frac{p_n}{q_n} = [a_0, a_1, \ldots, a_n]$. Then the following hold:*

1. $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n+1} \quad (n \geq 0)$,
2. $r_n - r_{n-1} = \frac{(-1)^{n+1}}{q_{n-1} q_n} \quad (n \geq 1)$,
3. $p_n q_{n-2} - p_{n-2} q_n = (-1)^n a_n \quad (n \geq 1)$,
4. $r_n - r_{n-2} = \frac{(-1)^n a_n}{q_{n-2} q_n} \quad (n \geq 2)$,
5. $\gcd(p_n, q_n) = 1 \quad (n \geq 0)$,
6. $q_1 = a_1 \geq 1, q_n \geq q_{n-1} + q_{n-2}, 1 = q_0 \leq q_1 < \cdots$,
7. $r_0 < r_2 < r_4 < \cdots, \cdots < r_5 < r_3 < r_1$,
8. $\lim_{n \longrightarrow \infty} r_n = \alpha$ *exists, and we have* $r_{2i} < \alpha < r_{2j+1}$, $\forall i, j \geq 0$.

Lemma 2.1 will be useful to prove the next proposition.

Proposition 2.1. *Let $\alpha \notin \mathbb{Q}$ and $\frac{p_n}{q_n} = [a_0, a_1, \ldots, a_n]$, such that*

$$\alpha = [a_0, a_1, \ldots] = \lim_{n \longrightarrow \infty} [a_0, a_1, \ldots, a_n].$$

Whenever $\frac{p}{q} \in \mathbb{Q}$ satisfies $|\alpha - \frac{p}{q}| < \frac{1}{2q^2}$, then $\frac{p}{q} = \frac{p_n}{q_n}$ for some n.

Proof: Assume that $\frac{p}{q}$ is not a convergent, with $q_n \leq q < q_{n+1}$. From Lemma 1.1.8 it follows that

$$|q_n \alpha - p_n| \leq |q\alpha - p| < \frac{1}{2q},$$

hence

$$\left| \alpha - \frac{p_n}{q_n} \right| < \frac{1}{2qq_n}.$$

Since $\frac{p}{q} \neq \frac{p_n}{q_n}$, one has $|pq_n - qp_n| \geq 1$, and therefore

$$\frac{1}{qq_n} \leq \frac{|pq_n - qp_n|}{qq_n} = \left| \frac{p}{q} - \frac{p_n}{q_n} \right| \leq \left| \alpha - \frac{p}{q} \right| + \left| \alpha - \frac{p_n}{q_n} \right| < \frac{1}{2q^2} + \frac{1}{2qq_n}.$$

But this means that $\frac{1}{qq_n} < \frac{1}{q^2}$ or $q < q_n$, contradiction. $\square$

2.2 The proof of Hurwitz's theorem

In this section we will show 2 different proofs of Hurwitz's theorem.

Proof using continued fractions: (i) Let $\frac{p_n}{q_n}$ be as in Proposition 2.1 It suffices to show that for every $j \geq 1$, at least one

$$\frac{p}{q} \in \left\{ \frac{p_{j-1}}{q_{j-1}}, \frac{p_j}{q_j}, \frac{p_{j+1}}{q_{j+1}} \right\} \text{ satisfies } \left| \alpha - \frac{p}{q} \right| < \frac{1}{\sqrt{5}q^2}.$$

Assume

$$\left| \alpha - \frac{p_{j-1}}{q_{j-1}} \right| \geq \frac{1}{\sqrt{5}q_{j-1}^2}, \quad \left| \alpha - \frac{p_j}{q_j} \right| \geq \frac{1}{\sqrt{5}q_j^2}.$$

Then

$$\left| \alpha - \frac{p_{j-1}}{q_{j-1}} \right| + \left| \alpha - \frac{p_j}{q_j} \right| = \left| \frac{p_{j-1}}{q_{j-1}} - \frac{p_j}{q_j} \right| = \frac{1}{q_{j-1}q_j} \geq \frac{1}{\sqrt{5}q_{j-1}^2} + \frac{1}{\sqrt{5}q_j^2}.$$

It follows that

$$1 \geq \frac{q_j}{\sqrt{5}q_{j-1}} + \frac{q_{j-1}}{\sqrt{5}q_j}, \text{ or } \frac{q_j}{q_{j-1}} + \frac{q_{j-1}}{q_j} \leq \sqrt{5}$$

and therefore, $\frac{q_j}{q_{j-1}} < \frac{\sqrt{5}+1}{2}$. (Note that one has a strict inequality, since $\frac{\sqrt{5}+1}{2}$ is irrational.) Now if also $\left| \alpha - \frac{p_{j+1}}{q_{j+1}} \right| \geq \frac{1}{\sqrt{5}a_{j+1}^2}$, then by the same argument, $\frac{q_{j+1}}{q_j} < \frac{\sqrt{5}+1}{2}$. Using $q_{j+1} = a_{j+1}q_j + q_{j-1}$, we thus conclude that

$$\frac{\sqrt{5}+1}{2} > \frac{q_{j+1}}{q_j} = a_{j+1} + \frac{q_{j-1}}{q_j} \geq 1 + \frac{q_{j-1}}{q_j} > 1 + \frac{\sqrt{5}-1}{2} = \frac{\sqrt{5}+1}{2},$$

contradiction. □

The second part (ii) of Hurwitz's theorem can also be proved using continued fractions, but the proof won't be given here. It can be found in [A]. Next, we will take a look at Hurwitz's theorem from a different point of view. The problem above can also be stated *geometrically*. Consider the upper half plane $\mathcal{H}$ and a point $z \in \mathbb{R} \setminus \mathbb{Q}$ on the x-axis. Through the point z draw a perpendicular L to the x-axis. At each rational point $\frac{p}{q}$ of the x-axis, draw a *horocycle* $S(\frac{p}{q}, h)$, $h > 0$, which is a circle tangent to the x-axis at the point $\frac{p}{q}$ with radius $\frac{1}{2hq^2}$.
If $S(\frac{p}{q}, h)$ is intersected by L, the distance between $\frac{p}{q}$ and z is less than the radius, i.e.

$$\left| \frac{p}{q} - z \right| < \frac{1}{2hq^2}.$$

Therefore Hurwitz's theorem boils down to figuring out how large h can be chosen s.t. L would intersect infinitely many S-circles.

Geometric proof: (i) For $\frac{p}{q} \in \mathbb{Q}$ and $h > 0$ let $S(\frac{p}{q}, h)$ be a horocycle. Consider the line $y = h$, parallel to the x-axis. Let $b, d \in \mathbb{Z}$ be such that $ad - bc = 1$, i.e. the matrix $\begin{pmatrix} p & b \\ q & d \end{pmatrix}$ is an element of the modular group Γ. Then, the modular transformation

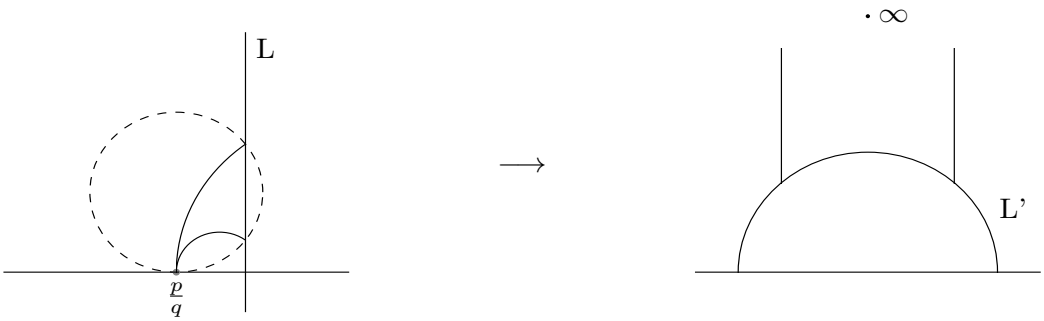
$$z \mapsto \begin{pmatrix} p & b \\ q & d \end{pmatrix} z$$

maps the line $y = h$ to the horocycle $S(\frac{p}{q}, h)$.
Consider the fundamental domain

$$\mathcal{D} = \{ z \in \mathcal{H} \mid |Re z| \leq \frac{1}{2}, \ |z| \geq 1 \}.$$

This triangle region tessellates the upper half plane. Now, consider the line L in the geometric statement of the theorem. L passes through infinitely many of those triangles.
For p, q coprime, let $\frac{p}{q}$ be a cusp of a triangle that L passes through. Using the inverse modular

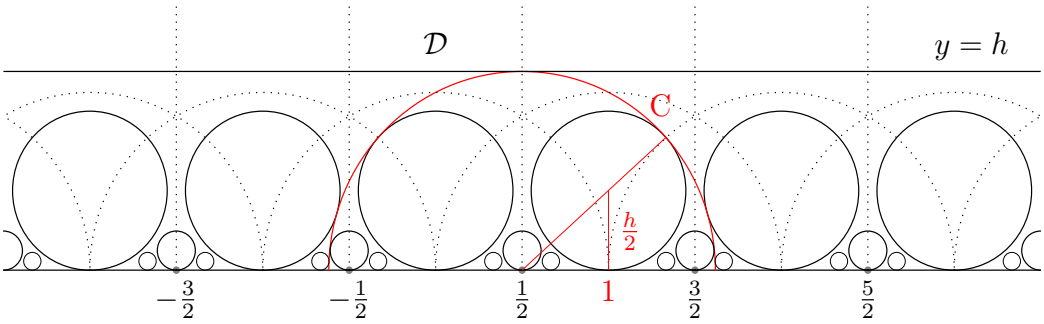
transformation to the one defined above, one can map $S(\frac{p}{q}, h)$ back to the line $y = h$, and map the line L to the geodesic L' orthogonal to the real line.



By applying a translation $\begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}$, $n \in \mathbb{Z}$, assume that L' intersects the base of D.

Next, consider the following question: *How large can h be chosen such that every geodesic intersecting the base of $\mathcal{D}$ will intersect either $y = h$ or an S-circle at one of the integral points?*

Let G be an arbitrary geodesic like that. First, note that G has at least 2 integral points on its interior; otherwise it would not intersect the base of $\mathcal{D}$. One can see that the most favourable position of the geodesic, such that it avoids intersections with the S-circles, is when its center is at $\frac{1}{2}$ (or $\frac{-1}{2}$), and it contains only 2 integral points in its interior.



Let C be a geodesic with "center" $\frac{1}{2}$, touching $y = h$, $S(\frac{0}{1}, h)$ and $S(\frac{1}{1}, h)$. Using some elementary geometry, one obtains that the radius of C is given by

$$\sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2h}\right)^2} + \frac{1}{2h} \tag{2}$$

Setting (2) equal to h, one obtains

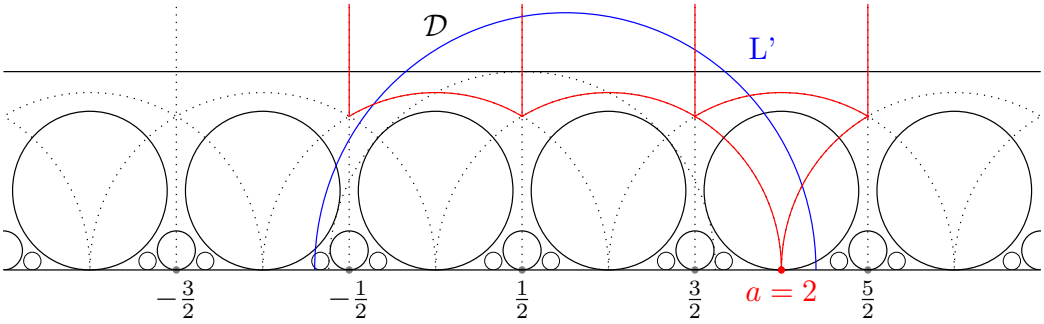
$$h = \frac{1}{2}\sqrt{5}.$$

Now, that we have obtained h, lets return to the geodesic L', which has the only restriction to intersect the base of $\mathcal{D}$. Note that L' can not coincide with C, since the endpoints of C are $\frac{1}{2} - \frac{\sqrt{5}}{2}$, $\frac{1}{2} + \frac{1}{\sqrt{5}}$, which are both irrational. Since L is carried to L', one of the endpoints should be rational; contradiction!

Recall that C was the most favorable choice for a geodesic that does not intersect (but only barely touches) $S(\frac{0}{1}, \frac{\sqrt{5}}{2}), y = \frac{\sqrt{5}}{2}$ and $S(\frac{1}{1}, \frac{\sqrt{5}}{2})$.

Since L' intersects the base of $\mathcal{D}$, L' also intersects at least one hyperbolic triangle with cusp at infinity. If we look at the triangles L' intersects, going from left to right and starting at $\mathcal{D}$, then we see that L' intersects one to a couple of triangles with cusps at infinity. At some point L' 'leaves' the triangles with cusps at infinity and enters one with a cusp on the x-axis. Denote this cusp (which is an integer) as a.

One concludes that when L is mapped to L', L' intersects at least one of the circles $S(\frac{0}{1}, \frac{\sqrt{5}}{2})$, $y = \frac{\sqrt{5}}{2}$, $S(\frac{a}{1}, \frac{\sqrt{5}}{2})$.

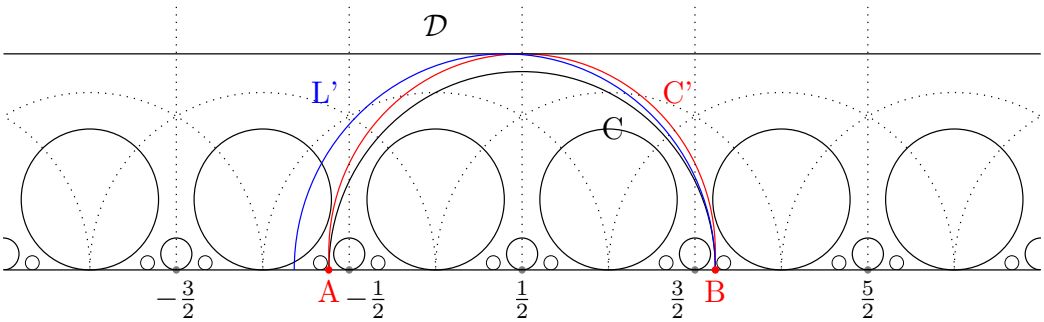


That means that at least one of the S-circles, of these three successive cusps of the triangles through which L' passes, must be *intersected*.
If one now maps L' back to L (as before), we get that for any three successive cusps of the triangles through which L passes, *at least one* is intersected when $h = \frac{\sqrt{5}}{2}$. Finally, recall that L is a line orthogonal to the x-axis at z, and L intersects an S-circle at $\frac{p}{q}$ if and only if $\frac{p}{q}$ and z satisfy the inequality $|\frac{p}{q} - z| < \frac{1}{2hq^2}$. Also note that we chose $\frac{p}{q}$ *arbitrarily* at the start, and then constructed all the S-circles and found at least one S-circle (with a corresponding cusp) around the one with center at $\frac{p}{q}$ which intersects L'. Since there are infinitely many cusps like that, L intersects infinitely many S-circles, and part (i) follows. □

(ii) Consider again the geodesic C. Its endpoints are satisfying the equation $z^2 - z - 1$. Rewriting this as

$$z = \frac{2z+1}{z+1}$$

shows that the endpoints of C are fixed points of the hyperbolic isometry $g := \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ in Γ. Hence, any geodesic passing through those 2 endpoints is mapped to itself. Also, note that this transformation maps $S(\frac{0}{1}, \frac{\sqrt{5}}{2})$ to $S(\frac{1}{2}, \frac{\sqrt{5}}{2})$.
Let P be the point of tangency between $S(\frac{0}{1}, \frac{\sqrt{5}}{2})$ with C and Q the point of tangency between $S(\frac{1}{1}, \frac{\sqrt{5}}{2})$ and C. Then g transforms the arc PQ on C into another arc on C, starting in Q and going in the direction of $\frac{1}{2} + \frac{\sqrt{5}}{2} =: B$. By continued application of g the whole arc QB is covered by an infinite number of transforms of PQ.
Suppose that $h > \frac{\sqrt{5}}{2}$. Let C' be a circle passing through $A := \frac{1}{2} - \frac{\sqrt{5}}{2}, B$, which is tangent to $y = h$.



This circle is fixed by g. By a similar procedure as before, one obtains that the arc between the points of tangency with $y = h$ and $S(\frac{2}{1}, h)$ covers the whole C' between A and B. Furthermore, there are no S-circles between C' and C.
Then, the idea is the following. As we have seen before, the point z on the x-axis corresponds to a line L (or L'), and $\frac{p}{q}$ and h correspond to S-circles in the upper half plane $\mathcal{H}$. We want to construct a geodesic L', that has, on one end, the endpoint z and, on the other end, a *rational* number, such that L' intersects only a finitely many S-circles.

Now, take B for the point z, from the geometric statement of the theorem. The line L', perpendicular to the x-axis at B (since L' is a geodesic), lies in the neighbourhood of B between C' and C, and intersects there no S-circles. The line L' lies in the area between C and C' until it 'leaves' that area at some point (going from right to left) and intersects C'. Since the point $B = z$ is irrational, the other endpoint of L' is rational, by part (i). The S-circles L' intersects lie between

the intersection of L' and C' and that rational endpoint, and there is *only a finite number of them.* Hence, when $h > \frac{\sqrt{5}}{2}$, there are, for the irrational point B, only finitely many fractions satisfying equation (1). The same is true for the point A. By letting the modular group Γ act on $\mathcal{H}$ we get infinitely many numbers for which only a finite number of fractions satisfy the equation (1), when $h > \frac{\sqrt{5}}{2}$. Note that a modular transformation $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$ maps the endpoint B (and A) again into an irrational number, to which the reasoning above applies again. These numbers, i.e. $\frac{aB+b}{cB+d}$, are irrational and can be found in any interval of the x-axis. This finishes the proof of part (ii) $\square$

3 Markov's theorem

The following section is concerned with a more general result than Hurwitz's theorem. The main topic, which is Markov's theorem, establishes a connection between Diophantine approximation and Diophantine equations. We will later state two versions of Markov's theorem, a Diophantine approximation and a quadratic forms version of the theorem.

Definition 3.1 (Markov triple)**.** A ***Markov triple*** is a triple (a, b, c) of positive integers satisfying Markov's equation

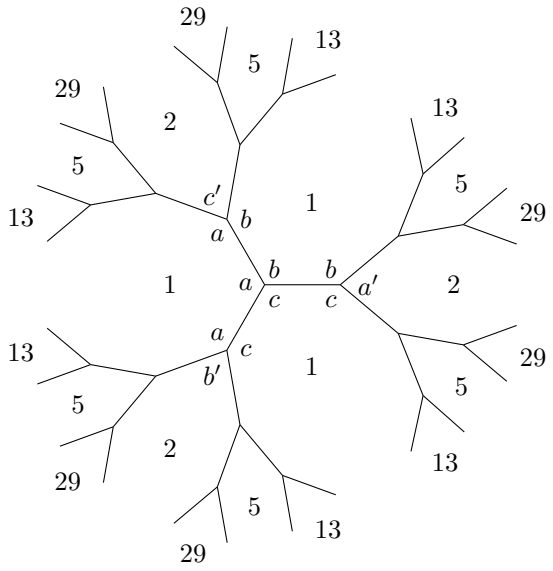
$$a^2 + b^2 + c^2 = 3abc.$$

A Markov number is a number that appears in some Markov triple.

One can obtain new solutions of Markov's equation using known ones.
Assume that (a, b, c) and (a', b, c) are solutions of Markov's equation, i.e. $a^2 + b^2 + c^2 = 3abc$, ${a'}^2 + b^2 + c^2 = 3a'bc$. Now, if we subtract the second equation from the first one and then divide by $(a - a')$ we obtain a formula for calculating new Markov numbers, namely

$$a' = 3bc - a = \frac{b^2 + c^2}{a}. \tag{3}$$

Note that since Markov's equation is quadratic in every variable, the obtained equation is linear in a'. Analogously, we can obtain formulas for b' and c'. Then there are three involutions, say σ_k for $k \in \{1, 2, 3\}$, acting on the set of Markov triples that map (a, b, c) to either $\sigma_1(a, b, c) = (a', b, c)$, $\sigma_2(a, b, c) = (a, b', c)$ or $\sigma_3(a, b, c) = (a, b, c')$. Every Markov triple can be obtained from a single Markov triple by applying a composition of these involutions. Note that the sequence of Markov triples is uniquely determined if we demand that no triple is visited twice. These last two results can be obtained by looking at the zeros of the polynomial $f(x) = x^2 - (3bc)x + b^2 + c^2$ (which has a as one of its roots) and ordering different Markov triples obtained from the involutions σ_k acting on, say $(1, 1, 1)$, by 'size'. Namely, we order the different Markov triplets according to the size of *the largest term* and *the smallest term* in the triple into a tree. Then every Markov number appears as maximum of some Markov triple. To obtain uniqueness we use certain properties of the unique structure of that tree. A detailed description of this procedure and a proof can be found in [A] (Chap. 3, Theorems 3.3-3.5).
Therefore the solutions of Markov's equation form a trivalent tree (which is a tree for which each node has vertex degree smaller than 4), called the Markov tree. Here is a visual representation of the tree: (Start with $(1, 1, 1)$ in the middle, and then apply recursively the involutions σ_k to every triple (a, b, c) to obtain $(a', b, c), (a, b', c), (a, b, c')$. The tree below connects (a, b, c) to $\sigma_k(a, b, c)$.)



We will denote the sequence of Markov numbers with $\mathcal{M}$, and they can be ordered into an infinite increasing sequence of integers which starts as follows:

$$\mathcal{M} = \{1, 2, 5, 13, 29, \dots\}$$

One can establish a connection between Markov numbers and Hurwitz's theorem (Theorem 2.4). By Hurwitz's theorem we have that for the golden ratio $\gamma_1 := \frac{1}{2}(1 + \sqrt{5})$, we have that only finitely

many fractions $\frac{p}{q} \in \mathbb{Q}$ satisfy

$$\left|\gamma_1 - \frac{p}{q}\right| < \frac{1}{kq^2}$$

for $k = \sqrt{5}$. Moreover, for every irrational number α that is equivalent to γ_1 via a modular transformation $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$, there are only finitely many fractions $\frac{p}{q} \in \mathbb{Q}$ satisfying

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{kq^2}. \tag{4}$$

Hurwitz also stated the following result: If $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ is not equivalent to γ_1, then infinitely many fractions satisfy (4) with $k := \gamma_2 = 2\sqrt{2}$. If α is not equivalent to γ_1 and $k > \gamma_2 := 2\sqrt{2}$, then we will see that there are only finitely many fractions satisfying (4).
Another result Hurwitz obtained is that for any $k < 3$, there are only finitely many equivalence classes of numbers for which there are only finitely many fractions satisfying (4). But, for $k = 3$ there are infinitely many classes of numbers that cannot be well approximated, i.e. for which there are only finitely many fractions satisfying (4). We will call those equivalence classes of irrational numbers that can't be well approximated, for a fixed $k < 3$, the *worst case of irrational numbers*.

Hurwitz stopped here, but Markov took over. Namely, Markov's theorem established a bijection between the equivalence classes of worst irrational numbers and sorted Markov triplets. One can already start to wonder whether there is a pattern behind the choice of γ_1 and then γ_2, and this is exactly the content of Markov's theorem. Later we will see that γ_1, γ_2 are only the first two numbers in a sequence which is uniquely determined by Markov's numbers.

Definition 3.2. The ***Lagrange number*** $L(\alpha)$ of an irrational number α is defined by:

$$L(\alpha) = \sup\{k \in \mathbb{R} \mid \left|\alpha - \frac{p}{q}\right| < \frac{1}{kq^2}, \text{ for infinitely many } \frac{p}{q} \in \mathbb{Q}\}$$

and the set of Lagrange numbers $\{L(\alpha) \mid \alpha \in \mathbb{R} \setminus \mathbb{Q}\}$ is called the ***Lagrange spectrum***.

Markov's theorem is a central result about the Lagrange spectrum below 3. One can state Markov's theorem in two equivalent ways as follows:

Theorem 3.1 (Markov, Diophantine approximation version)**.** *Let (a, b, c) be any Markov triple, let p_1, p_2 be integers satisfying*

$$p_2 b - p_1 a = c \tag{5}$$

and let

$$x = \frac{p_2}{a} + \frac{b}{ac} - \frac{3}{2} + \sqrt{\frac{9}{4} - \frac{1}{c^2}}. \tag{6}$$

Then there are infinitely many fractions $\frac{p}{q}$ satisfying

$$\left|x - \frac{p}{q}\right| < \frac{1}{kq^2} \tag{7}$$

with

$$k = \sqrt{\frac{9c^2 - 4}{c^2}},$$

but only finitely many for any larger value of k, i.e. $L(x) = k = \sqrt{\frac{9c^2-4}{c^2}}$.

Conversely, suppose x' is an irrational number such that only finitely many fractions $\frac{p}{q}$ satisfy (7) for some $k < 3$. then there exists a unique sorted Markov triple (a, b, c) such that x' is equivalent to x defined by equation (6).

Remark 3.1. It may seem unsymmetric that only p_2 (and not p_1) appears in equation (6). If we use Markov's equation and substitute $p_2 b = c + p_1 a$ into equation (6), we obtain that

$$x = \frac{p_1}{b} - \frac{a}{bc} + \frac{3}{2} + \sqrt{\frac{9}{4} - \frac{1}{c^2}}.$$

Next, one can ask: what if we consider two different pairs of integers $(p_1, p_2), (p_1', p_2')$ that satisfy (5), for the same triple (a, b, c), how will this affect x? Substituting $p_2' = p_2 - \frac{a}{b}(p_1 - p_1')$ into (6)

leads to another equivalent value of x, that differs from x by $\frac{p_1'-p_1}{b}$, which is an integer.

Before continuing to the second version of Markov's theorem, lets introduce some useful terminology.

3.1 Quadratic forms

Consider binary quadratic forms of the following form:

$$f(p,q) = Ap^2 + 2Bpq + Cq^2,$$

such that $A, B, C \in \mathbb{R}$. Markov's theorem is only concerned with indefinite forms, namely the ones with positive determinant, i.e. with $\det f = B^2 - AC > 0$. In that case, the quadratic polynomial

$$f(p,1) = Ap^2 + 2Bp + C$$

has two distinct roots:

$$\frac{-B \pm \sqrt{\det f}}{2A}$$

if $A \neq 0$. For $A = 0$ take $\frac{-C}{2B}$ and ∞ as two roots in the real projective line $\mathbb{R} \cup \infty$. Then, the following two statements are equivalent:

(i) The polynomial $f(p,1)$ has at least one root in $\mathbb{Q} \cup \infty$.

(ii) There exist integers p, q, not both zero, such that $f(p,q) = 0$.

On the other hand, one can look for indefinite forms f for which the set

$$\{f(p,q) \mid (p,q) \in \mathbb{Z}^2, (p,q) \neq (0,0)\}$$

stays farthest away from 0. Since the values $f(p,q)$ can be very large, it makes sense to normalize the forms f such that $\det f = 1$. In other words, we can ask for which form f is

$$M(f) = \inf_{\substack{(p,q)\in\mathbb{Z}^2 \\ (p,q)\neq 0}} \frac{|f(p,q)|}{\sqrt{|\det f|}} \tag{8}$$

maximal.

Korkin and Zolotarev obtained a result in regards to the question; for which f is $\mathcal{M}(f)$ maximal.

Theorem 3.2 (Korkin and Zolotarev)**.** *Let f be an indefinite binary quadratic form with real coefficients. If f is equivalent to the form*

$$p^2 - pq - q^2,$$

then

$$M(f) = \frac{2}{\sqrt{5}}.$$

Otherwise

$$M(f) < \frac{1}{\sqrt{2}}.$$

Remark 3.2. Korkin and Zolotarev's theorem is only the first result in line that can be obtained from Markov's theorem, corresponding to the first Markov triple $(1,1,1)$ (in the sequence of Markov numbers). Markov's theorem tells us that Hurwitz's theorem is, roughly speaking, the Diophantine approximation version of Korkin and Zolotarev's theorem and this will be visible from the second version of Markov's theorem (stated later).

Two binary quadratic forms are called ***equivalent*** if there exists $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$ such that

$$\tilde{f}(p,q) = f(ap + bq, cp + dq).$$

Now that we have introduced the necessary tools, we will state another version of Markov's theorem.

Theorem 3.3 (Markov, quadratic forms version)**.** *Let (a, b, c) be any Markov triple, let p_1, p_2 be integers satisfying the equation $p_2 b - p_1 a = c$ and let*

$$x_0 = \frac{p_2}{a} + \frac{b}{ac} - \frac{3}{2}.$$

Let

$$r = \sqrt{\frac{9c^2 - 4}{4c^2}}$$

and let f be the indefinite quadratic form

$$f(p, q) = p^2 - 2x_0 pq + (x_0^2 - r^2)q^2. \tag{9}$$

Then $\mathcal{M}(f) = \frac{1}{r}$ and the infimum in equation (8) is attained.
Conversely, suppose that $\tilde{f}$ is an indefinite binary quadratic form with $\mathcal{M}(\tilde{f}) > \frac{2}{3}$. Then there is an unique sorted Markov triple (a, b, c) such that $\tilde{f}$ is equivalent to a multiple of the form f in equation (9).

Remark 3.3. The number x defined by (6) is a root of the quadratic form (9) given in the theorem above. Also note that $\mathcal{M}(f) = 2L(x)^{-1}$, where $L(x)$ is the Langrange number of x.

3.2 Horocycles and Farey tessellation

We have seen in Section 2 how to assign horocycles to fractions. To every $(p, q) \in \mathbb{R}^2 \setminus (0, 0)$ one can assign a *horocycle* as follows:

(i) If $q \neq 0$, let $h(p, q)$ be the horocycle at $\frac{p}{q}$ with euclidean diameter $\frac{1}{q^2}$

(ii) If $q = 0$, let $h(p, q)$ be the horocycle at ∞ with height p^2

Then one has that:

Proposition 3.1. *For $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL_2(\mathbb{R})$ with $|detA| = 1$ and for $v \in \mathbb{R}^2 \setminus \{0\}$, the hyperbolic isometry $M_A = \frac{az+b}{cz+d}$ maps the horocycle $h(v)$ to $h(Av)$.*

Note that this means that the map is equivariant w.r.t. the $PGL_2(\mathbb{R})$-action.

Proof: Every isometry of the upper-half plane $\mathcal{H}$ can be represented as a composition of isometries of the following types: a translation $n_b := \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}$, a scalar multiplication $a_\lambda := \begin{pmatrix} \lambda^{\frac{1}{2}} & 0 \\ 0 & \lambda^{\frac{1}{2}} \end{pmatrix}$, an involution $f_1 := \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ or $f_2 := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Then the proof boils down to a direct calculation and it is enough to consider only those 4 simpler transformations. For a visual example of how those 4 transformations act on a horocycle, consider Figure 1.

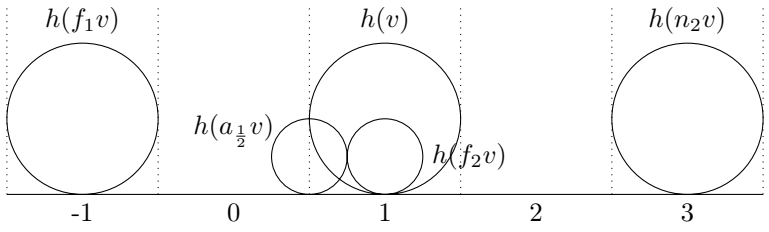


Figure 1: $n_2, a_{\frac{1}{2}}, f_1, f_2$ acting on $h(v) = h(1, 1)$

□

Next, we will introduce ***Ford circles***. Ford circles are horocycles $h(p, q)$ with integer and coprime parameters (p, q). There is exactly one Ford circle centered at each rational number $\frac{p}{q}$ and at ∞.

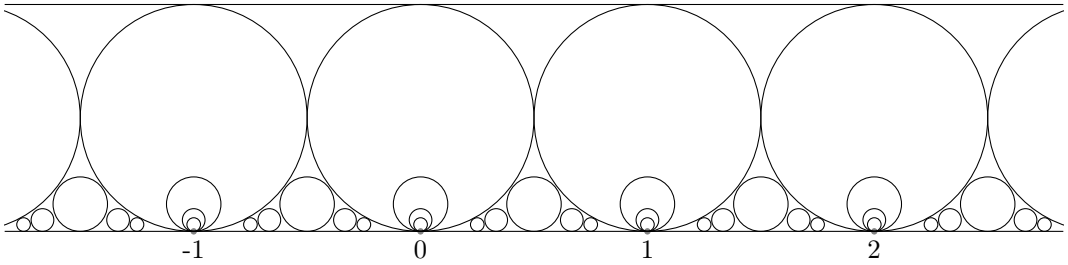


Figure 2: Ford circles

If one connects the points where the tangent Ford circles touch the real axis with geodesics, one obtains the so-called ***Farey tessellation***. The Farey tessellation is an ideal triangulation of the hyperbolic plane with vertex set $\mathbb{Q} \cup \{\infty\}$ (see Figure 3).

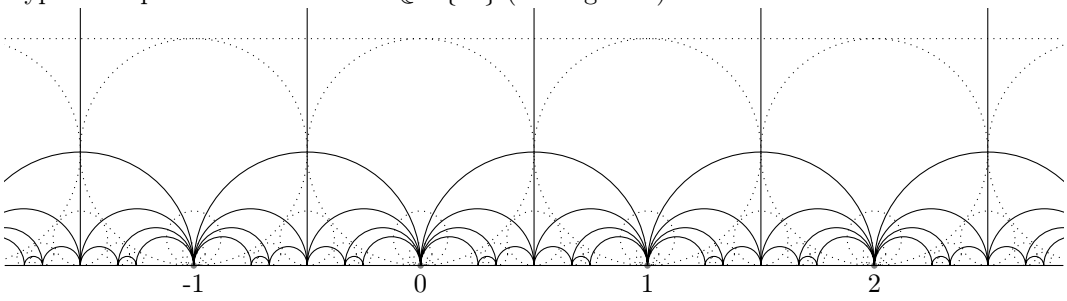


Figure 3: Farey tessellation

Definition 3.3. The ***signed distance*** $d(h_1, h_2)$ of the horocycles h_1, h_2 is defined as the hyperbolic length of the geodesic segment connecting the horocycles which is orthogonal to both. if h_1 and h_2 intersect, it is defined to be *negative*, while if they don't intersect, it is *positive*. If h_1 and h_2 have the same center, then $d(h_1, h_2) = -\infty$ (see Figure 4).

Figure 4: Signed distance between horocycles

Before continuing to the next proposition, we will recall some hyperbolic geometry. Consider the upper-half plane $\mathcal{H}$ equipped with the hyperbolic length, where the hyperbolic length of a parametrized curve $\lambda : [a, b] \longrightarrow \mathcal{H}$, $\lambda(t) = x(t) + iy(t)$ is given by

$$L(\lambda) = \int_a^b \frac{\sqrt{x'(t)^2 + y'(t)^2}}{y(t)} dt. \tag{10}$$

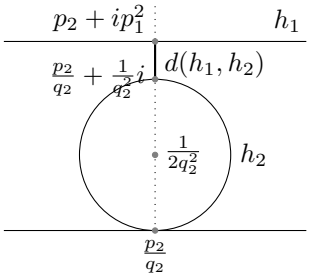
Then, if $a < b$, $x \in \mathbb{R}$, the hyperbolic length of the segment $[x + ai, x + bi]$ is given by

$$\int_a^b \frac{dt}{t} = \log \frac{b}{a}.$$

Proposition 3.2. *The signed distance of two horocycles $h_1 = h(p_1, q_1)$ and $h_2 = h(p_1, q_2)$ is given by*

$$d(h_1, h_2) = 2 \log|p_1 q_2 - p_2 q_1|.$$

Proof: First, assume that the center of the horocycle $h_1 = h_1(p_1, 0)$ is centered at ∞ and $h_2 = h_2(p_2, q_2)$. Then $d(h_1, h_2) = \log\left(\frac{p_1^2}{\frac{1}{q_2^2}}\right) = 2\log(p_1 q_2)$, and the formula above holds. Next, let $h_1(p_1, q_1)$ be an arbitrary horocycle. Then the hyperbolic isometry $\begin{pmatrix} 0 & 1 \\ 1 & \frac{-p_1}{q_1} \end{pmatrix}$ maps one horocycle to a horocycle centered at ∞, and the conclusion follows from 3.1 □



Definition 3.4. The ***signed distance*** $d(h, g)$ between a horocycle h and a geodesic g is defined as the length of the geodesic segment connecting h and g that is orthogonal to both. It is defined negative if h and g intersect; otherwise positive, and if g ends in the center of h, then let $d(h, g) = -\infty$ (see Figure 5).

Figure 5: Signed distance between a horocycle and a geodesic

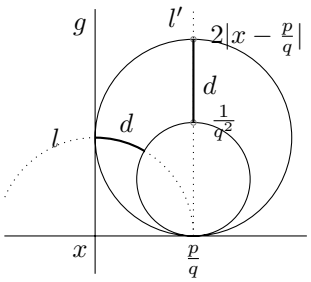
The next proposition introduces a geometric interpretation of Hurwitz's theorem and the Diophantine approximation version of Markov's theorem.

Proposition 3.3. *Let $h(p, q)$ be a horocycle with $q \neq 0$ and let g be a vertical geodesic from $x \in \mathbb{R}$ to ∞. Then their signed distance is*

$$d(h, g) = \log\left(2q^2\left|x - \frac{p}{q}\right|\right).$$

Remark 3.4. A fraction $\frac{p}{q} \in \mathbb{Q}$ satisfies the inequality $\left|x - \frac{p}{q}\right| < \frac{1}{kq^2}$ if and only if $d(h(p, q), g) < -\log\frac{k}{2}$.

Proof: Let $h(p, q)$ and g be as in the statement of the proposition. Since the signed distance between g and h is the length of a geodesic *orthogonal* to both, it has to intersect g at its highest point and have the point of tangency of the horocycle h and the real axis, as one endpoint. Then the signed distance $d(h, g)$ is given by the hyperbolic distance between the highest point of a geodesic, say l, and the intersection point of that geodesic with $h(p, q)$, where l is a geodesic with center x and endpoints $\frac{p}{q}, 2x - \frac{p}{q}$. One can apply a parabolic isometry to the geodesic l, which would keep the endpoint $\frac{p}{q}$ fixed, and send $2x - \frac{p}{q}$ to ∞ (see picture right). Now, draw another horocycle at $\frac{p}{q}$, which is tangent to g, and look at the length between the intersection point of this new horocycle and the transformed line l, denoted l', and $h(p, q)$ and l'.



Then the signed distance $d(h, g)$ is equal to the hyperbolic length of that segment on l' which was previously described. And that is equal to

$$\log\left(\frac{2\left|x - \frac{p}{q}\right|}{\frac{1}{q^2}}\right)$$

which is exactly the statement of our proposition. □

Remark 3.5. Let $x \in \mathbb{R} \setminus \mathbb{Q}$ and let g be the vertical geodesic at base x. Then, by Proposition 3.3, the first part of Hurwitz's theorem is equivalent to saying that there exist infinitely many horocycles h such that $d(h, g) < -\log\frac{\sqrt{5}}{2}$.

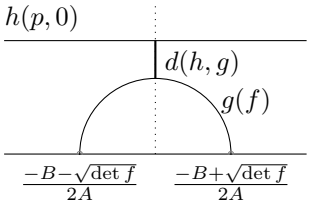
Now, lets turn our attention to indefinite quadratic forms. One can assign a geodesic $g(f)$ to every indefinite binary quadratic form f with real coefficients A, B, C. Namely, one assigns the geodesic that connects the zeros of the polynomial $f(x,1) = Ax^2 + 2Bx + C$ and if $A = 0$, then let $g(f)$ be a vertical geodesic connecting $\frac{-C}{2B}$ to ∞.

Proposition 3.4. *Let f be an indefinite quadratic form. The signed distance of the horocycle $h(p,q)$ and the geodesic $g(f)$ is given by*

$$d(h(p,q), g(f)) = \log \frac{|f(p,q)|}{\sqrt{\det f}}. \tag{11}$$

Proof: First, consider the case of horizontal horocycles ($q = 0$). If $g(f)$ is a vertical geodesic, i.e. $f(p,0) = 0$, then equation (11) is immediate.

Otherwise, consider the picture on the right.
Note that $\frac{\sqrt{\det f}}{A}$ is half the distance between the zeros of $g(f)$ and from $\frac{\sqrt{\det f}}{A} = \frac{\sqrt{\det f}p^2}{|f(p,0)|}$ one obtains that



$$d(h(p,0), g(f)) = \log \left(\frac{p^2}{\frac{\sqrt{\det f}p^2}{|f(p,0)|}} \right) = \log \frac{|f(p,0)|}{\sqrt{\det f}}.$$

For the general case, consider an isometry $A \in SL_2(\mathbb{R})$ with

$$A \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} \tilde{p} \\ 0 \end{pmatrix}.$$

Then

$$\begin{aligned} d(h(p,q), g(f)) &= d(Ah(p,q), Ag(f)) \\ &= d(h(\tilde{p},0), g(f \circ A^{-1})) \\ &= \log \frac{|(f \circ A^{-1})(\tilde{p},0)|}{\sqrt{\det(f \circ A^{-1})}} \\ &= \log \frac{|f(p,q)|}{\sqrt{\det f}}. \end{aligned}$$

□

3.3 Decorated ideal triangles and the modular torus

Next, we want to establish a direct bijection between Markov triples and certain lengths of, so called, *decorated ideal triangles*. That will be a crucial step in translating Markov's theorem from an algebraic into a purely geometric problem.

Definition 3.5. An ***ideal point*** is a point on the boundary. An ***ideal triangle*** is a closed region of the hyperbolic plane that is bounded by three geodesics connecting three ideal points. A ***decorated ideal triangle*** is an ideal triangle together with a horocycle at each vertex (see Figure 6).

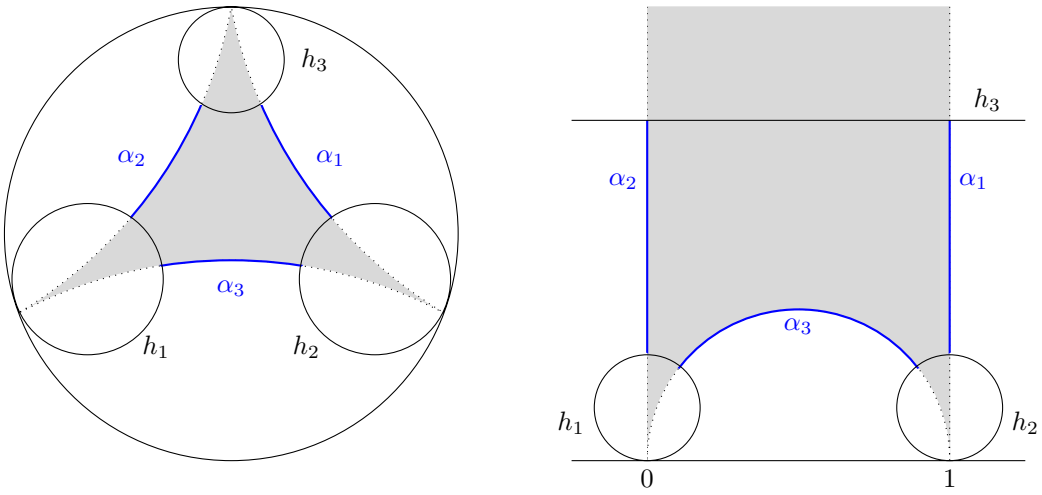


Figure 6: Decorated ideal triangle in the Poincare disk model (left) and in the half-plane model (right)

Next, we will study the distance between pairs of horocycles on a geodesic, i.e. sides of decorated ideal triangles.

Definition 3.6. Let g be a geodesic decorated with 2 horocycles h_1,h_2 at its ends. Define the ***truncated length*** of the decorated geodesic g as the signed distance of the horocycles h_1,h_2 (recall Definition 3.3):

$$\alpha := d(h_1, h_2)$$

and let $a := e^{\frac{\alpha}{2}}$ be defined as its ***weight***.

Remark 3.6. Figure 6 shows the truncated lengths $\alpha_1, \alpha_2, \alpha_3$. Note that any triple $(\alpha_1, \alpha_2, \alpha_3) \in \mathbb{R}^3$ of truncated lengths (or equivalently of weights) determines a unique decorated ideal triangulation up to isometry.

Proposition 3.5. *Let T be a decorated ideal triangle with truncated lengths α_k and weights a_k, for $k \in \{1,2,3\}$. Its horocycles intersect the triangle in three finite arcs, denote their hyperbolic lengths by c_k, $k \in \{1,2,3\}$ (see Figure 7). Then the truncated side lengths α_k determine the horocyclic arc lengths c_k, and vice versa, via the relation*

$$c_k = \frac{a_k}{a_i a_j} = e^{\frac{1}{2}(-\alpha_i - \alpha_j + \alpha_k)} \tag{12}$$

where (i,j,k) is a permutation of $(1,2,3)$.

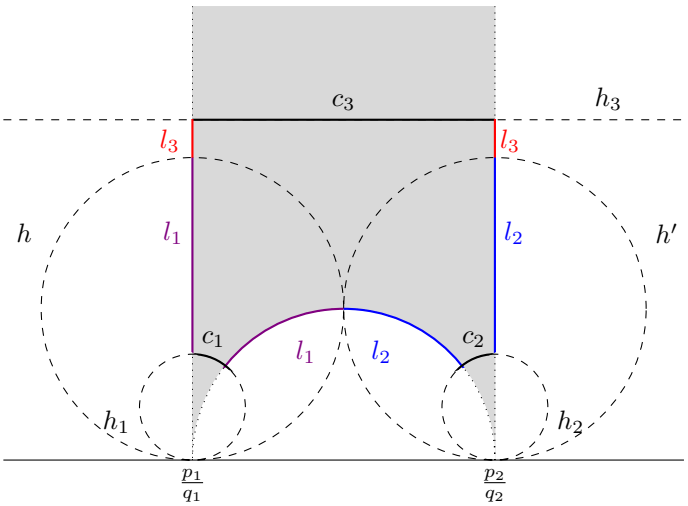


Figure 7: Relation between c_k and α_k

Proof: First, consider a decorated ideal triangle in the half-plane model from From figure 6. Let h_k,c_k,α_k and a_k, for $k \in \{1,2,3\}$, be as stated in the proposition. Next, we are going to 'expand' the horocycles h_1 and h_2 such that they are still tangent to the real axis, but are also tangent to each other. Namely, we will increase the radiuses of h_1 and h_2 until they both have the same radius and are tangent to each other (see Figure 7). Denote those two new circles by h and h'. Let $\frac{p_1}{q_1}$ (respectively $\frac{p_2}{q_2}$) be the tangent point of h_1 (respectively h_2) to the x-axis. Then h and h' intersect the geodesic connecting the points $\frac{p_1}{q_1}$ and $\frac{p_2}{q_2}$ precisely in the middle. Moreover, since that geodesic is orthogonal to h and h' it must intersect the horocycles h and h' through the point where h,h' are tangent to each other.
Now, write the signed distances $d(h_i, h_j) = \alpha_k = l_i + l_j$, $i,j,k \in \{1,2,3\}$ where l_1 is the length of the segment contained in the circle h and l_3 is the length of the segment outside of it. Analogously define l_2 with respect to h' (see Figure 7). Then we have that

$$\alpha_3 = l_1 + l_2 = \alpha_1 + \alpha_2 - 2l_3. \tag{13}$$

Next, we will express the length l_3 in terms of c_3. Let y denote the (Euclidean) height of the side of the decorated ideal triangle T connecting $\frac{p_1}{q_1}$ to c_3. Then the hyperbolic length l_3 is given by

$$\log\left(\frac{y}{\frac{p_2}{q_2} - \frac{p_1}{q_1}}\right), \tag{14}$$

since the radius of h is given by the Euclidean length $(\frac{p_2}{q_2} - \frac{p_1}{q_1})/2$. Note that the hyperbolic length of c_3, by equation (10) is given by

$$\frac{\frac{p_2}{q_2} - \frac{p_1}{q_1}}{y}.$$

Then by equation 13 and 14 we have that

$$\log\left(\frac{1}{c_3}\right) = l_3 = \frac{1}{2}(\alpha_1 + \alpha_2 - \alpha_3),$$

and hence $c_3 = e^{\frac{1}{2}(-\alpha_1 - \alpha_2 + \alpha_3)}$. The proof for c_1 and c_2 is completely analogous. □

We will use the ideal triangles introduced above to work with ***ideal quadrilaterals***. Ideal quadrilaterals and *decorated ideal quadrilaterals* can be constructed from two decorated ideal triangles (see Figure 8).

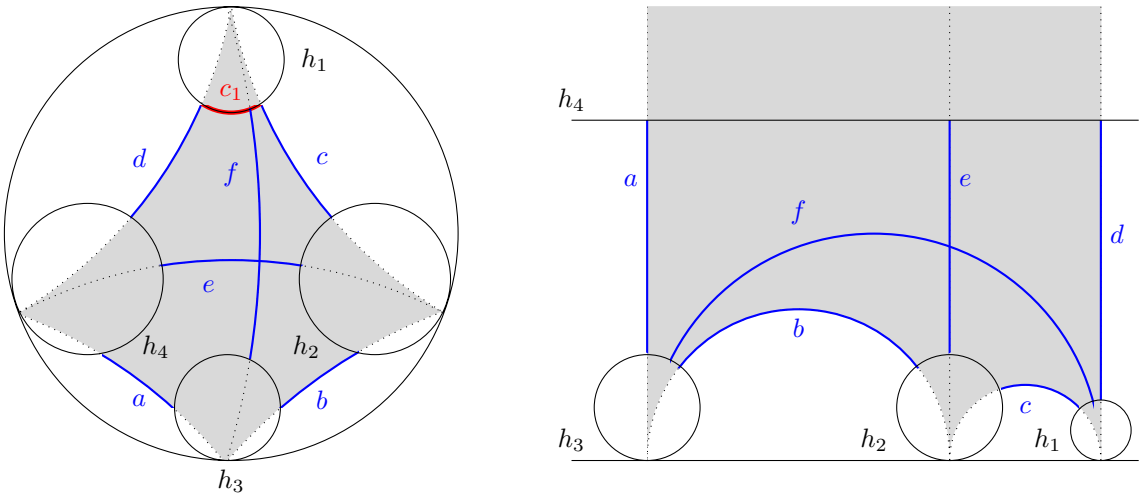


Figure 8: Decorated ideal quadrilateral in the Poincare disk model (left) and in the half-plane model (right)

Let Q be a decorated ideal quadrilateral. Then it can be decomposed into two decorated ideal triangles in *two* different ways (see Figure 8). Assume that Q is constructed by gluing 2 decorated ideal triangles with weights a, d, f and b, c, f along the edge with weight f, where the weight is given by $e^{\frac{\alpha}{2}}$ with α being the signed distance between corresponding horocycles.
Similarly decompose Q into two decorated ideal triangles with weights d, c, e and a, b, e. Using equation 12, one can derive the following relation:

Proposition 3.6 (***Ptolemy's relation***). *For the weights a, b, c, d, e, f we have that:*

$$ef = ac + bd. \tag{15}$$

Proof of Ptolemy's relation: Let Q be a decorated ideal quadrilateral with corresponding horocycles h_1, h_2, h_3, h_4, and a horocyclic arc length c_1 as in Figure 8 (left). Then, from equation 12 one obtains that

$$c_1 = \frac{e}{dc} \quad \text{and} \quad c_1 = \frac{a}{df} + \frac{b}{cf} \tag{16}$$

(since the arc c_1 on the circle h_1 can be written as a sum of two circle arcs separated into the two decorated ideal triangles a, f, d and b, c, f as in Figure 8). Putting those two equations together and multiplying by cdf yields Ptolemy's relation. □

Definition 3.7. Let G be the group of orientation-preserving hyperbolic isometries generated by

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}. \tag{17}$$

The ***modular torus*** is given by the orbit space $M = \mathcal{H}/G$, where $\mathcal{H}$ is the upper-half plane.

A fundamental domain $\mathcal{F}$ of the group G together with A and B acting on it is illustrated in Figure 9.

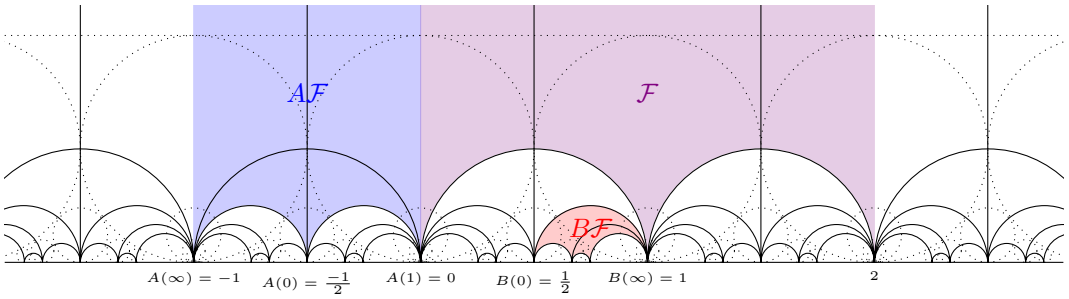


Figure 9: A and B acting on a fundamental domain $\mathcal{F}$ of G

The group G is the commutator subgroup of the modular group $PSL_2(\mathbb{Z})$ with $[G : PSL_2(\mathbb{Z})] = 6$. Moreover, the quotient group $PSL_2(\mathbb{Z})/G$ is the group of orientation preserving isometries of the modular torus $\mathcal{H}/G$.

3.4 Correspondence between Markov triples and ideal triangles

Definition 3.8. Let T be a connected orientable surface with genus $g \geq 0$, without boundary and with $n \geq 1$ punctures that has negative Euler characteristic (in our case the modular torus). An ***ideal triangulation*** of T is a maximal collection of disjoint *essential arcs* that are pairwise non-homotopic, where essential arcs are defined as geodesics with endpoint in the punctures (in our case $\mathbb{Q} \cup \{\infty\}$) which are not homotopic (relating to its endpoints) to a point in $\overline{T}$ (see [T] for more details).

Let T be an ideal quadrilateral as depicted in Figure 8, but with edge weights a, b, c and a' (as depicted in the Figure 10). Next, let T' be the ideal triangulation obtained from T by flipping the edge with weight a, i.e. by replacing this edge with the other diagonal, namely a', in the ideal quadrilateral formed by the other edges (see Figure 10). By equation 3 and Ptolemy's relation, the edge weights of T' are $(a', b, c) = \sigma_1(a, b, c)$. One can obtain analogous equations if a different edge is flipped (consider Example 3.1 for a visualisation of this).

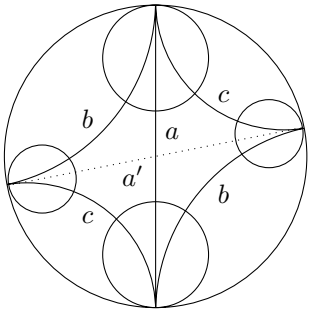


Figure 10: Triangulations T and T' of a once punctured torus

Theorem 3.4. *(i) A triple (a, b, c) of positive integers is a Markov triple if and only if there is an ideal triangulation of the decorated modular torus whose three edges have the weights a, b, c. This triangulation is unique up to the 12-fold symmetry of the torus.*

(ii) If T is an ideal triangulation of the decorated modular tours with edge weights (a, b, c), and if T' is an ideal triangulation obtained from T by performing a single edge flip, then the edge weights of T' are given by one of $(a', b, c), (a, b', c)$ or (a, b, c'); depending on which edge was flipped.

To understand the theorem above, we need to introduce some more terminology from topology.

Definition 3.9. A ***once punctured hyperbolic torus*** is a torus with one point removed, equipped with a complete metric of constant curvature -1 and finite volume.

One can obtain a once punctured hyperbolic torus by accordingly gluing 2 decorated ideal triangles along their edges such that the congruent horocycles fit together (along the vertical line l in Figure 11).
Conversely, every ideal triangulation of a hyperbolic torus with one puncture decomposes into two ideal triangles. A good way to see this is to consider Figure 11.

Consider the decorated ideal triangles glued together as in Figure 11. Denote their edges by a_1, a_2, a_3, a_4, as illustrated in Figure 11. Then we can use the following gluing isometries:

$$\begin{cases} \varphi_1 : a_1 \longrightarrow a_4 \\ \quad z \longmapsto B(z) \end{cases} \qquad \begin{cases} \varphi_3 : a_3 \longrightarrow a_2 \\ \quad z \longmapsto A(z) \end{cases}$$

to obtain a once punctured hyperbolic torus. We have that φ_1 sends -1 to 0 and ∞ to 1, while φ_3 sends 1 to 0 and ∞ to -1.

Remark 3.7. Note that Figure 11 also shows a fundamental domain of the group G.

Definition 3.10. A ***decorated* once punctured hyperbolic torus** is a once punctured hyperbolic torus together with a choice of horocycle at the cusp.

Remark 3.8. Let $(a, b, c) \in \mathbb{R}^3_{>0}$ be a triple of weights, then it determines a decorated once punctured hyperbolic torus up to isometry, together with an ideal triangulation. Conversely, a decorated once punctured hyperbolic torus, together with an ideal triangulation, determines such a triple of edge weights.

Next, we will finally make the connection between Markov's equation

$$a^2 + b^2 + c^2 = 3abc$$

and decorated once punctured hyperbolic toruses.
Let $\mathcal{T}$ be a decorated once punctured hyperbolic torus with an ideal triangulation T with edge weights $(a, b, c) \in \mathbb{R}^3_{>0}$. Now, consider again the gluing from Figure 11, but pay attention to the

horocycles that decorate the ideal triangles. Those arcs of decorated horocycles 'build' a circle after they are mapped to the once punctured hyperbolic torus (see Figure 11).

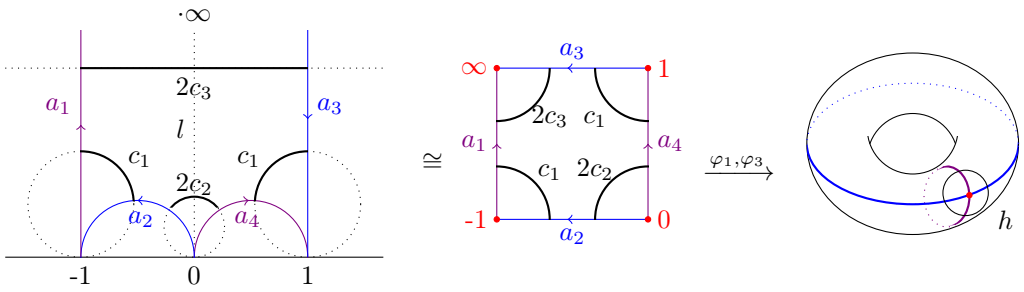


Figure 11: Gluing the punctured hyperbolic torus

Let c_1, c_2, c_3 denote the lengths of those arcs, as in Figure 11. From equation 12 we have that the total length l of that circle h is given by

$$l = 2(c_1 + c_2 + c_3) = 2\left(\frac{a}{bc} + \frac{b}{ac} + \frac{c}{ab}\right).$$

This equation is equivalent to

$$a^2 + b^2 + c^2 = \frac{l}{2}abc.$$

Thus the weights satisfy Markov's equation if and only if the horocycle h has length $l = 6$. In Markov's theorem we are only interested in the case where $l = 6$, therefore we will decorate all once punctured hyperbolic tori with the horocycle of length 6.

The modular torus M, decorated with a horocycle of length 6, is obtained by gluing two decorated ideal triangles with weights $(1, 1, 1)$. Lifting this triangulation and decoration to the hyperbolic upper-half plane, one obtains the Farey tessellation with Ford circles (Figure 3). This implies that for every Markov triple (a, b, c) there is an ideal triangulation of the decorated modular torus with edge weights a, b, c. To see this, follow the path in the Markov tree leading from $(1, 1, 1)$ to (a, b, c) and perform the corresponding edge flips on the projected Farey tessellation.

On the other hand, the *flip graph* of a complete hyperbolic surface with punctures is also connected (see [Hatcher],[Mosher, pg. 35-37] for the proof). The flip graph has the ideal triangulations as vertices, and edges connect triangulations related by a single edge flip. This implies the converse statement: If a, b, c are the weights of an ideal triangulation of the modular torus, then (a, b, c) is a Markov triple.

Example 3.1. *Flipping edges to obtain an ideal triangulation of the decorated modular torus with weights* $(2, 5, 1)$.

To obtain an ideal triangulation of the decorated modular torus with weights (a, b, c), we have to follow the Markov tree leading from $(1, 1, 1)$ to (a, b, c) and perform corresponding edge flips on the projected Farey tessellation (in the upper-half plane).

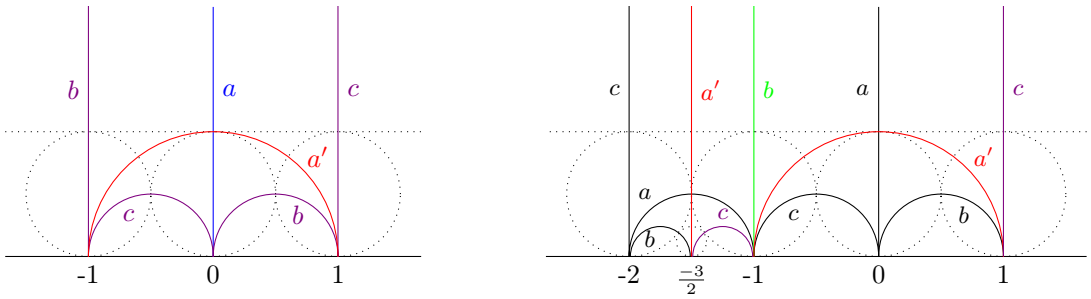


Figure 12: Ideal triangulation of the decorated modular torus with weights $(1, 1, 1)$ (left), Ideal quadrilateral that is a fundamental domain for the torus with b as diagonal (right)

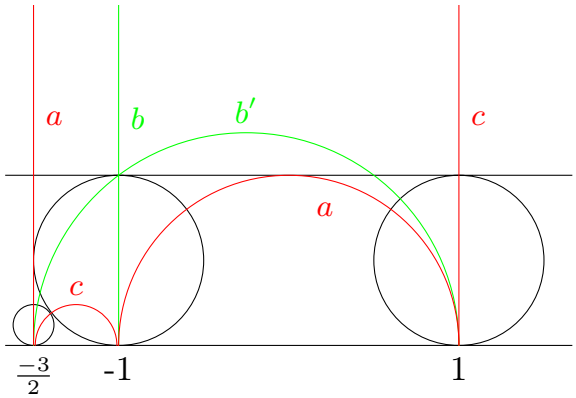
Start with $(a, b, c) := (1, 1, 1)$. Using Ptolemy's relation we get:

$$(a', b, c) = \sigma_1(a, b, c) = (\frac{b^2 + c^2}{a}, b, c) = (2, 1, 1).$$

Next, we want to flip the edge b.
We are looking for an ideal quadrilateral that is a fundamental domain for the modular torus **and** has the edge b as a diagonal. The isometry

$$\beta := \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

sends the ideal triangle with vertices $-1, 0, 1$ to the ideal triangle with vertices $\frac{-3}{2}, -1, \infty$. Note that the two ideal triangles with vertices $-1, 0, 1$ and $-1, 1, \infty$ form a fundamental domain $-1, 0, 1, \infty$ of the modular torus. Thus, the ideal triangle with vertices $-1, 1, \infty$ together with the ideal triangle $\frac{-3}{2}, -1, \infty$ form an ideal quadrilateral with vertices $\frac{-3}{2}, -1, 1, \infty$ which is a fundamental domain of the modular torus (see Figure 12).



Using Ptolemy's relation again for $(a, b, c) := (2, 1, 1)$ one obtains:

$$(a, b', c) = \sigma_2(a, b, c) = (a, \frac{a^2 + c^2}{b}, c) = (2, 5, 1),$$

and we get an ideal triangulation of the decorated modular torus with weights $(2, 5, 1)$.

Remark 3.9. There exists only one ideal triangulation of the modular torus with weights $(1, 1, 1)$, namely the ideal triangulations that lifts to the Farey tessellation. Then, the symmetries of the modular torus permute its edges. Since the Markov tree and the flip graph are isomorphic, we have that the two ideal triangulations with the same weights are related by an isometry of the modular torus $M = \mathcal{H}/G$. Therefore, we obtain Theorem 3.4.

Note: The wording in this chapter is mainly borrowed from [S], so we encourage the reader to take a look at [S] for more details.

3.5 Geodesics that stay away from horocycles

The goal of the next section is to study how far a geodesic, which crosses a decorated ideal triangle, can stay away from the horocycles at the vertices. We will say that a geodesic ***bisects*** a side of a decorated ideal triangle if it intersects the side in the point at equal distance to the two horocycles at the ends of the side. Similarly like in the proof of Hurwitz's theorem, we are again looking for geodesics crossing a decorated ideal triangle that stay away from horocycles. More precisely, we are looking at the following problem:

Let T be a decorated ideal triangle with three sides, say a_1, a_2, a_3. If we look at all geodesics crossing a_1 and a_2, which geodesic maximises the minimum of signed distances to the three horocycles at the vertices?

The following proposition gives the solution to this optimizing problem.

Proposition 3.7. *(i) Let T be a decorated ideal triangle with horocycles h_1, h_2, h_3, and let a_1, a_2, a_3 denote both its sides and their weights. If*

$$a_1^2 \leq a_2^2 + a_3^2 \quad \text{and} \quad a_2^2 \leq a_1^2 + a_3^2 \tag{18}$$

then the geodesic g bisecting the sides a_1 and a_2 is the unique geodesic that maximizes the minimum of signed distances to the three horocycles at the vertices.

(ii) Let $(j,k) \in \{(1,2),(2,1)\}$ with

$$a_j^2 \geq a_k^2 + a_3^2. \tag{19}$$

Then the perpendicular bisector g' of the side $a_k$is the unique solution to the optimizing problem stated above. In this case, the minimal distance is attained for h_j and h_3, namely

$$d(h_j, g') = d(h_3, g') = \frac{\alpha_k}{2} \leq d(h_k, g').$$

Proof: Here we will only give a sketch of the proof, since a very detailed proof is given by [S]. The proof of Proposition 3.7 relies on a couple of geometric facts about bisecting geodesics. First, we can conclude from the 180° rotational symmetry around the bisecting *intersection* point, that any geodesic bisecting a side has equal distance from the two horocycles at the end. Now, let v_k, for $k \in \{1,2,3\}$ be the three vertices of our decorated ideal triangle. Let g be the geodesic bisecting a_1 and a_2. Let P_k, for $k \in \{1,2,3\}$ be the foot of the perpendicular from vertex v_k to the geodesic g. Next, we can distinguish between a few cases of different positions of P_1, P_2, P_3 on g (for example P_3 could lie between P_1 and P_2) and deduce (using some hyperbolic geometry) that the geodesic g is indeed the unique solution to the optimizing problem. Lastly, one can show that the order of the points P_k on g depends on the fact whether the weights satisfy the inequality (18) or one of the inequalities (19). □

Next, we will use Proposition 3.7 to explicitly compute the signed distance between g and its horocycles. It turns out that the geodesic g has a quite similar form to the indefinite quadratic form defined in (9).

Proposition 3.8. *Let g be the geodesic bisecting a_1 and a_2 of a decorated ideal triangle. Then the common signed distance of g and the horocycles is*

$$d(h_1, g) = d(h_2, g) = d(h_3, g) = -\log r$$

with

$$r = \sqrt{\frac{\delta^2}{4} - \frac{1}{a_3^2}},$$

and δ is thee sum of the lengths of the horocyclic arcs,

$$\delta = c_1 + c_2 + c_3 = \frac{a_1}{a_2 a_3} + \frac{a_2}{a_1 a_3} + \frac{a_3}{a_1 a_2}.$$

Moreover, suppose the vertices of the decorated ideal triangle are

$$v_1 < v_2, \quad v_3 = \infty$$

and the horocycle h_3 has height 1. Then the endpoints $x_{1,2}$ of g are

$$x_{1,2} = x_0 \pm r,$$

where $x_0 = v_2 + \frac{a_2}{a_3 a_1} - \frac{\delta}{2}$.

Proof: Assume for the vertices of the decorated ideal triangle that $v_1 < v_2$, $v_3 = \infty$ and $h_3 = h(1, 0)$. Then, by Proposition 3.7, g has equal signed distance to all horocycles, i.e.

$$d(h_1, g) = d(h_2, g) = d(h_3, g).$$

To compute this distance explicitly, consider only $d(h_3, g)$. Note that the (Euclidean) height from the center x_0 of the geodesic g to h_3 (consider figure Figure 13) is 1 and if r denotes the radius of the geodesic g and we have (by (10)) that

$$d(h_3, g) = \log\left(\frac{1}{r}\right) = -\log r.$$

Next, we will compute the radius r of g explicitly.
Let c_k, for $k \in \{1, 2, 3\}$ denote the arc-length of the horocycle h_k which is contained inside the decorated ideal triangle. Let $c_3 = s_1 + s_2$, where c_3 is split by the vertical geodesic connecting x_1 to v_3 (Figure 13) into the two arcs s_1 and s_2.

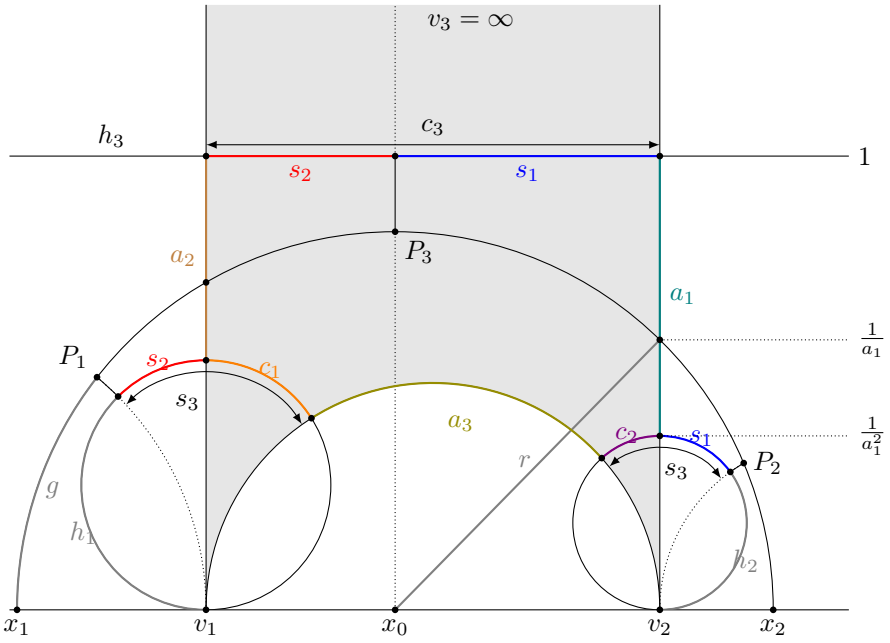


Figure 13: A decorated ideal triangle with a geodesic g going through the midpoints of sides a_1 and a_2

Similarly, draw a geodesic starting at v_k, for $k \in \{1, 2\}$, and orthogonal to the geodesic g, as shown in Figure 13, and let P_k, for $k \in \{1, 2\}$, be the foot of that geodesic on g. Let s_3 denote the arclength on the horocycle h_2, which is an extension of the arclength c_3 to the orthogonal on g ending in P_1. Then $s_3 = s_2 + c_1$ (see Figure 13). Similarly one can see that $s_3 = s_1 + c_2$. Alltogether, we get

$$c_1 = -s_2 + s_3, \quad c_2 = -s_1 + s_3, \quad c_3 = s_1 + s_2,$$

which implies that

$$s_3 = \frac{1}{2}(c_1 + c_2 + c_3) = \frac{1}{2}\left(\frac{a_1}{a_2 a_3} + \frac{a_2}{a_3 a_1} + \frac{a_3}{a_1 a_2}\right).$$

Using the relations above together with Pythagoras's theorem:

$$r^2 = s_i^2 + \frac{1}{a_i^2}, \quad i = 1, 2,$$

one obtains:

$$\begin{aligned} s_3^2 &= \left(\frac{1}{2}(\frac{a_1}{a_2a_3}+\frac{a_2}{a_3a_1}+\frac{a_3}{a_1a_2})\right)^2 \\ &= \frac{1}{4}\left(c_1^2+c_2^2+c_3^2+2\left(\frac{1}{a_3^2}+\frac{1}{a_2^2}+\frac{1}{a_1^2}\right)\right) \\ &= \frac{1}{4}\left(c_1^2+c_2^2+c_3^2+\frac{2}{a_3^2}+2\left(r^2-s_1^2+r^2-s_2^2\right)\right) \\ &= \frac{1}{2a_3^2}+r^2+\frac{1}{2}s_3^2+\frac{1}{2}\left(s_1s_2-s_2s_3-s_1s_3\right). \end{aligned}$$

Thus

$$\begin{aligned} s_3^2 &= \frac{1}{a_3^2}+2r^2+(s_1s_2-s_2s_3-s_1s_3) \\ &= \frac{1}{a_3^2}+2r^2+\frac{-1}{2}((c_1+s_2+c_2+s_1)(s_1+s_2)-s_1s_2) \\ &= \frac{1}{a_3^2}+2r^2+(-1)\left(r^2+\frac{1}{2}\left(c_1c_3+c_2c_3-\frac{1}{a_1^2}-\frac{1}{a_2^2}\right)\right) \\ &= \frac{1}{a_3^2}+2r^2+(-1)\left(r^2+\frac{1}{2}\left(\frac{a_1}{a_2a_3}\frac{a_3}{a_1a_2}+\frac{a_2}{a_1a_3}\frac{a_3}{a_1a_2}-\frac{1}{a_1^2}-\frac{1}{a_2^2}\right)\right) \\ &= \frac{1}{a_3^2}+r^2. \end{aligned}$$

Set $\delta = 2s_3$, then

$$r = \sqrt{\frac{\delta^2}{4}-\frac{1}{a_3^2}}.$$

The last thing to obtain is the formula for x_0 from Proposition 3.8. From Figure 13 one sees that

$$\begin{aligned} x_0 &= v_2 - s_1 \\ &= v_2 - (s_3 - c_2) \\ &= v_2 - \left(\frac{\delta}{2} - c_2\right) \\ &= v_2 + \frac{a_2}{a_3a_1} - \frac{\delta}{2}. \end{aligned}$$

□

3.6 The topology of a once punctured hyperbolic torus

The goal of the next section is to introduce some topological facts about a once punctured hyperbolic torus, since we have seen in Section 3.4 that the fundamental domain of our group G can be glued to become a once punctured hyperbolic torus. Therefore we can focus on projections of geodesics on the once punctured hyperbolic torus, instead of looking at them in the whole upper-half plane $\mathcal{H}$.

Definition 3.11. A ***simple geodesic*** is a geodesic whose image on the once punctured torus has no self intersections. An ***ideal arc*** in a complete hyperbolic surface with cusps is a simple geodesic connecting two punctures or a puncture with itself.

Remark 3.10. Ideal triangulations are maximal sets of non-intersecting ideal arcs.

We will see that ideal arcs are in a one-to-one correspondence with simple closed geodesics (Proposition 3.9).

Proposition 3.9. *Let T be a fixed once punctured hyperbolic torus.*

(i) Let c be an ideal arc. Then there exists a unique simple closed geodesic g that does not intersect c.

(ii) Every other geodesic not intersecting c has either two ends in the puncture, or one end in the puncture and the other end approaching (by distance) the closed geodesic g.

(iii) If a, b, c are the edges of an ideal triangulation T, then the simple closed geodesic g that does not intersect c intersects each of the two triangles of T in a geodesic segment bisecting the edges a and b.

(iv) For every simple closed geodesic g, there is a unique ideal arc c that does not intersect g.

Proof:

(i) First cut the torus T along the ideal arc c (see Figure 14 on the right). We are going to obtain a hyperbolic cylinder which has two boundary curves that are complete geodesics with both endpoints in the cusp (which is split into two points after the cut). There is up to orientation a unique non-trivial free homotopy class that contains simple curves (i.e. curves that don't cross themselves), and this class contains a unique simple closed geodesic. That geodesic can be imagined pulling all those simple closed curves (from the homotopy class) tight on the torus, relative to the hyperbolic metric.

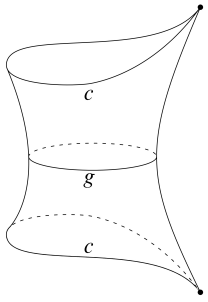


Figure 14: Once punctured hyperbolic torus cut along c

(ii) Let g' be a geodesic different from g which does not intersect c. Now, consider the universal cover of the cylinder in the hyperbolic plane. Then we can imagine the cut along c as a rule to choose g' in $\mathcal{H}$ such that it does not cross c, nor any geodesic Fc, where $F \in G$. First, note that g' has to have one endpoint in the puncture (i.e. a rational number) to avoid all of the geodesics Fg, $F \in G$. W.l.o.g. assume that this endpoint is either $-1, 0, 1$ or ∞.
If we choose the other end of g' to also be in the puncture, but avoiding c, then we are done. Assume that is not the case. Then we want to show that g' approaches g.
First, note that g' can not cross g, because if we would project g' on the fundamental domain $\mathcal{F}$ of the group G, g' would intersect Fc, for some $F \in G$ after it crosses the intersection point between g and g'. If we look at the height of g' projected on $\mathcal{F}$, we see that this height must either increase (if g' is located under g) or decrease (if g' is located above g), to avoid intersecting c or Fc. Therefore g' has one endpoint in the puncture, and the other endpoint approaches one endpoint of g (in $\mathcal{H}$) i.e. the projection of g' on the punctured hyperbolic torus approaches g.

(iii) Consider the three edge midpoints of a triangulated once punctured torus. Those three midpoints are the fixed points of an orientation preserving isometric involution. Moreover, every ideal arc passes through one of these points.
More precisely, an ideal triangulation of a once punctured torus is symmetric with respect to a 180° rotation around the edge midpoint. That rotation swaps the geodesic segments bisecting edges a and b in the two ideal triangles, so they connect smoothly.

Hence they form a simple closed geodesic, which does not intersect c.

(iv) First cut the torus along the simple closed geodesic g. We obtain a cylinder with a cusp and two geodesic boundary circles (Figure 15 on the right). Now, take the puncture point as the base point for the homotopy group. Then there is up to orientation a unique non-trivial homotopy class containing simple closed curves and this class contains an unique ideal arc. □

Note: The illustrations in Figure 14 and 15 were taken from [S].

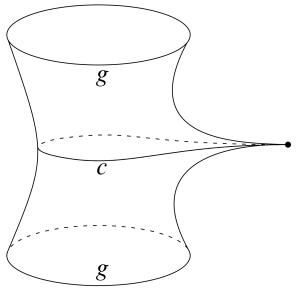


Figure 15: Once punctured hyperbolic torus cut along g

3.7 The proof of Markov's theorem

After building up all the required theory in the past sections, in this section we are going to prove both versions of Markov's theorem (recall Theorem 3.1 and 3.3).

Definition 3.12. Two geodesics g, g' in $\mathcal{H}$ are $GL_2(\mathbb{Z})$***-related*** if, for some $A \in GL_2(\mathbb{Z})$ we have $M_A g := \begin{pmatrix} a & b \\ c & d \end{pmatrix} g = g'$, where $\det(M_A) = 1$ (recall Proposition 3.1).

Proposition 3.10. *Let g be a complete geodesic in the hyperbolic plane, and let $\pi(g)$ be its projection on the modular torus. Then the following three statements are equivalent:*

(i) $\pi(g)$ is a simple closed geodesic.

(ii) There is a Markov triple (a, b, c) so that for one (hence any) choice of integers p_1, p_2 satisfying $p_2 b - p_1 a = c$, the geodesic g is $GL_2(\mathbb{Z})$-related to the geodesic ending in $x_0 \pm r$ with x_0 and r defined by

$$x_0 = \frac{p_2}{a} + \frac{b}{ac} - \frac{3}{2} \tag{20}$$

and

$$r = \sqrt{\frac{9}{4} - \frac{1}{c^2}}. \tag{21}$$

(iii) The greatest lower bound for the signed distances of g and a Ford circle is greater than $-\log \frac{3}{2}$.

If g satisfies one (hence all) of the statements (i),(ii),(iii), then

(iv) the minimal signed distances of g and a Ford circle is $-\log r$,

(v) among all Markov triples (a, b, c) that verify (ii), there is a unique sorted Markov triple.

Proof: $'(i) \Rightarrow (ii)'$: If $\pi(g)$ is a simple closed geodesic, then there is a unique ideal arc c not intersecting $\pi(g)$ (Proposition 3.9 (iv)). Pick an ideal triangulation T of the modular torus that contains c, and let a and b be the other edges. By Theorem 3.4 (a, b, c) is a Markov triple (here a, b, c denote both the ideal arcs and their weights). The geodesic $\pi(g)$ intersects each of the two triangles of T in a geodesic segment bisecting the edges a and b (by Proposition 3.9 (iii)).
Now let p_1, p_2 be integers satisfying $p_2 b - p_1 a = c$ and consider the decorated ideal triangle in $\mathcal{H}$ with vertices

$$v_1 = \frac{p_1}{b}, \quad v_2 = \frac{p_2}{a}, \quad v_3 = \infty, \tag{22}$$

and their resp. Ford circles

$$h_1 = h(p_1, b), \quad h_2 = h(p_2, a), \quad h_3 = h(1, 0). \tag{23}$$

Such integers p_1, p_2 exist because the numbers a, b, c of a Markov triple are pairwise coprime. Moreover, this implies that the fractions in (22) are reduced, and v_1 and v_2 are determined up to addition of a common integer. By Proposition 3.2, this decorated ideal triangle has edge weights

$$a_1 = a, \quad a_2 = b, \quad a_3 = c, \tag{24}$$

since the signed distance between two horocycles $h_1 = h(p_1, q_1)$ and $h_2 = h(p_2, q_2)$ is given by $d(h_1, h_2) = 2\log|p_1 q_2 - p_2 q_1|$.
Conversely, every ideal triangle with vertices $\tilde{v}_1, \tilde{v}_2, \tilde{v}_3$ with $\tilde{v}_3 = \infty$ and rational vertices $\tilde{v}_1, \tilde{v}_2$, that is decorated with the respective Ford circles, has weights (24), and satisfies $\tilde{v}_1 < \tilde{v}_2$ is obtained this way. To get the triangles with $\tilde{v}_1 > \tilde{v}_2$, change c to $-c$ in $p_2 b - p_1 a = c$.
This implies that any lift of a triangle of T to the hyperbolic plane is $GL_2(\mathbb{Z})$-related to v_1, v_2, v_3. Finally, use Proposition 3.8 with $\delta = 3$ to deduce that g is $GL_2(\mathbb{Z})$-related to the geodesic ending in $x_0 \pm r$.

$'(ii) \Rightarrow (iv)'$: Let T be a triangulation of the modular torus that contains c, and let a and b be the other edges. Let $\hat{T}$ be the lift of the triangulation T to $\mathcal{H}$. The geodesic g crosses an infinite strip of triangles of $\hat{T}$. By Proposition 3.8, the signed distance of g and any Ford circle centered at a vertex incident with this strip is $-\log r$.
We claim that the signed distance to any other Ford circle is larger. To see this, consider a vertex $v \in \mathbb{Q} \cup \{\infty\}$ that is not incident with the triangle strip, and let ρ be a geodesic ray from v to a point $p \in g$. Note that the projected ray $\pi(\rho)$ on the modular torus intersects $\pi(g)$ at least once

before it ends in $\pi(p)$, and that the signed distance to the first intersection is at least $-\log r$.

$'(ii) \wedge (iv) \Rightarrow (iii)'$: This follows directly from $r = \sqrt{\frac{9}{4} - \frac{1}{c^2}} < \frac{3}{2}$, since then by (iv), the minimum signed distance of g and a Ford circle is $-\log r > -\log \frac{3}{2}$.

$'(iii) \Rightarrow (i)'$: We will show the contrapositive: If the geodesic g does not project to a simple closed geodesic, then there is a Ford circle with signed distance between the circle and g smaller than $-\log \frac{3}{2} + \epsilon$, for every $\epsilon > 0$.

There is nothing to show if at least one end of g is in $\mathbb{Q} \cup \{\infty\}$, because then the Ford circle at this end has signed distance $-\infty$. So assume g does not project to a simple closed geodesic and both ends of g are irrational.

We will recursively define a sequence $(T_n)_{n \geq 0}$ of ideal triangulations of the modular torus, with edges labeled a_n, b_n, c_n, such that the following holds:

(1) The geodesic $\pi(g)$ has at least one pair of consecutive intersections with the edges a_n, b_n.

(2) The edge weights, which we also denote by a_n, b_n, c_n, satisfy

$$a_n \leq b_n \leq c_n,$$

so that (a_n, b_n, c_n) is a sorted Markov triple.

(3) $c_{n+1} > c_n$

This proves the claim, since Propositions 3.7 and 3.8 imply that for each n, there is a horocycle with signed distance to g less than

$$-\frac{1}{2} \log \left(\frac{9}{4} - \frac{1}{c_n^2} \right),$$

which tends to $-\log \frac{3}{2}$ from above as $n \to \infty$.

We are going to define the sequence (T_n) inductively. Let T_0 be the triangulation with edge weights $(1, 1, 1)$, with edges labeled so that (1) holds.

Suppose the triangulation T_n with labeled edges is already defined for some $n \geq 0$. Define the labeled triangulation T_{n+1} as follows: since $\pi(g)$ is *not* a simple closed geodesic, it intersects all three edges. Because g has an irrational end (in fact, both ends are assumed to be irrational), there are infinitely many edge intersections. Hence, there is a pair of intersections with a_n and b_n next to an intersection with c_n. If the sequence of intersections is $a_n b_n c_n$, let T_{n+1} be the triangulation with edges

$$(a_{n+1}, b_{n+1}, c_{n+1}) = (a_n, c_n, b_n'),$$

and if the sequence is $b_n a_n c_n$, let T_{n+1} be the triangulation with

$$(a_{n+1}, b_{n+1}, c_{n+1}) = (b_n, c_n, a_n'),$$

where a_n' and b_n' are the ideal arcs obtained by flipping (recall Section 3.4) the edges a_n or b_n in T_n, respectively. By induction on n, one sees that (1),(2),(3) are satisfied for all $n \geq 0$.

$'(i) \wedge (ii) \Rightarrow (v)'$: The Markov triples (a, b, c) verifying (ii) are precisely the triples of edge weights of ideal triangulations containing the ideal arc c not intersecting $\pi(g)$. The triangulations containing the ideal arc c form a doubly infinite sequence (i.e. a sequence infinite in both directions) in which neighbors are related by a single edge flip fixing c. In this sequence, there is a unique triangulation for which the weight c is largest. □

Proposition 3.11. *Let g be a complete geodesic in the hyperbolic plane, and let $X \subset \mathbb{R} \backslash \mathbb{Q}$ be the set of ends of lifts of simple closed geodesics in the modular torus. Then the following two statements are equivalent:*

(a) The ends of g are contained in $\mathbb{Q} \cup \{\infty\} \cup X$.

(b) For some $M > -\log \frac{3}{2}$ there are only finitely many (possibly zero) Ford circles h with signed distance $d(g, h) < M$.

Proof: $'(a) \Rightarrow (b)'$: Consider the ends x_k of g, for $k \in \{1, 2\}$. If $x_k \in \mathbb{Q} \cup \{\infty\}$, then g contains a ray ρ_k that is contained inside the Ford circle at x_k. In this case, let $M_k = 0$. If $x_k \in X$, then

x_k is also the end of a geodesic $\tilde{g}$ that projects to a simple closed geodesic in the modular torus. By Proposition 3.10

$$\inf d(h, \tilde{g}) > -\log \frac{3}{2},$$

where the infimum is taken over all Ford circles h. Since g and $\tilde{g}$ converge at x_k, there is a constant $M_k > -\log \frac{3}{2}$ and a ray ρ_k contained in g and ending in x_k such that $d(h, \rho_k) > M_k$ for all Ford circles h. The part of g not contained in ρ_1 or ρ_2 is empty or of finite length, so it can intersect the interiors of at most finitely many Ford circles. This implies (ii) with $M = \min(M_1, M_2)$.

$'(b) \Rightarrow (a)'$: To show the contrapositive, assume (a) is false: At least one end of g is irrational but not the end of a lift of a simple closed geodesic in the modular torus.
This implies that the projection $\pi(g)$ intersects every ideal arc in the modular torus infinitely many times. Adapt the argument for the implication $'(iii) \Rightarrow (i)'$ in the proof of Proposition 3.10 to show that there is a sequence of horocycles (h_n) and an increasing sequence of Markov numbers (c_n) such that $d(g, h_n) < -\frac{1}{2} \log\left(\frac{9}{4} - \frac{1}{c_n^2}\right)$. This implies that (b) is false. □

The quadratic forms version of Markov's theorem follows from Proposition 3.10.
The indefinite quadratic form

$$f(p, q) = p^2 - 2x_0pq + (x_0^2 - r^2)q^2$$

given in equation (9) has precisely $x_0 \pm r$ as its zeros. Then (recall Section 3.1)

$$\det f = x_0^2 - (x_0^2 - r^2) = r^2, \tag{25}$$

and hence

$$\begin{aligned} M(f) &= \inf_{\substack{(p,q)\in\mathbb{Z}^2 \\ (p,q)\neq 0}} \frac{|f(p, q)|}{\sqrt{|\det f|}} \\ &= \frac{1}{r}. \end{aligned}$$

The converse statement in Theorem 3.3 follows from Proposition 3.10 $'(iii) \Rightarrow (v)'$.

The Diophantine approximation version of Markov's theorem follows from Proposition 3.10 together with Proposition 3.11.
We know from Proposition 3.10 that the geodesic centered at x with endpoints $x_0 \pm r$ is a simple closed geodesic in the modular torus. Then by Proposition 3.11, for some

$$M > -\log \frac{3}{2},$$

there are only finitely many ford circles h with signed distance $d(g, h) < M$. Recall (Proposition 3.3) that a fraction $\frac{p}{q} \in \mathbb{Q}$ satisfies the inequality $\left|x - \frac{p}{q}\right| < \frac{1}{\tilde{k}q^2}$ if and only if $d(h(p, q), g) < -\log \frac{\tilde{k}}{2}$. Now, if we set $M := -\log \frac{k}{2}$, with $k = \sqrt{9 - \frac{4}{c^2}} < 3$ (as in Theorem 3.1), we get that

$$M > -\log \frac{3}{2},$$

and hence there are only finitely many fractions $\frac{p}{q} \in \mathbb{Q}$ satisfying

$$\left|x - \frac{p}{q}\right| < \frac{1}{kq^2}.$$

4 A different approach to Markov's theorem

4.1 Fricke's trace identity

In this chapter we are going to look at an alternative approach to Markov's theorem. For this, let us go back to the end of Section 3.3, namely Definition (3.7), where we defined the modular torus as

$$M = \mathcal{H}/G$$

where G is generated by the two matrices

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}. \tag{26}$$

Next, we will look at certain properties of traces of those matrices.

Proposition 4.1 (***Fricke's trace identity***)**.** *Let $A, B \in SL_2(\mathbb{R})$ be arbitrary and let $C := AB$. Then*

$$(trA)^2 + (trB)^2 + (trC)^2 = trA\ trB\ trC + tr(ABA^{-1}B^{-1}) + 2. \tag{27}$$

Proof: By Cayley-Hamilton's theorem, any matrix $Y \in SL_2(\mathbb{R})$ satisfies its characteristic equation $\lambda + \lambda^{-1} = trY$, which means that

$$Y + Y^{-1} = (trY)\mathcal{I}. \tag{28}$$

Multiplying (28) by $X \in SL_2(\mathbb{R})$ yields

$$XY + XY^{-1} = tr(Y)X,$$

and after taking the trace we obtain the so-called *skein relation*:

$$tr(XY) + tr(XY^{-1}) = trX\ trY. \tag{29}$$

By several applications of the skein relation, one can obtain Fricke's identity.
Set $X := ABA^{-1}$ and $Y = B$, then

$$trABA^{-1}B + trABA^{-1}B^{-1} = trABA^{-1}\ trB = trB^2.$$

Next, apply the skein relation to $X = AB$ and $Y = A^{-1}B$:

$$trABA^{-1}B + trABB^{-1}A = trAB\ trA^{-1}B.$$

Since $AB = C$ and $trA^{-1}B = -trAB + trA\ trB$ (again by the skein relation), one has

$$trABA^{-1}B = -tr(A^2) - tr(C^2) + trA\ trB\ trC.$$

Finally, using that $tr(A^2) = (trA)^2 - 2$ holds for every $A \in SL_2(\mathbb{R})$, one obtains Fricke's identity. □

Denote $x := trA, z := trB, z := trC$ and $j_c = trABA^{-1}B^{-1}$, then Fricke's identity becomes the following cubic equation

$$x^2 + y^2 + z^2 - xyz - (j_c + 2) = 0,$$

with positive x, y, z.
To make the connection to Markov's theorem, consider now hyperbolic matrices A, B with parabolic commutator $ABA^{-1}B^{-1}$. Then we have that $j_c = trABA^{-1}B^{-1} = -2$, and hence Fricke's identity takes the form

$$x^2 + y^2 + z^2 - xyz = 0.$$

If we rescale the variables by $\frac{1}{3}$, we obtain Markov's equation.
Therefore the simplest solution $(1, 1, 1)$ of Markov's equation corresponds to the A, B generators of the group G (26).

4.2 On the relation between Markov's theorem and continued fractions

On another note, similarly like in the first proof of Hurwitz's theorem we gave, one can approach Markov's theorem with the help of continued fractions. While this approach is well explained in [Aigner], we will only explain a short connection here (see [Caroline S.] for more details). For that, we will introduce *cutting sequences*.

For simplicity, first assume we have a square grid in the (Euclidean) plane, there the grid lines are all spaced with equal distance apart (such that a translation preserves the vertical, resp. horizontal, grid lines). Assume that a is the distance between the horizontal grid lines, while b is the distance between the vertical grid lines.
Now, draw a straight line l passing through the grid with some slope λ.

Definition 4.1. The ***cutting sequence*** of the line l is defined as the sequence of sides of a's and b's that l meets going from left to right.

Remark 4.1. We are looking at doubly infinite sequences of the line l, but the pattern of the cutting sequence often turns out to be periodic, so that we can write it more compactly. We will use the cutting sequence to study the *slope* λ of l.

If β denotes the distance along l between two vertical segments and α denotes the distance between horizontal segments, then $\lambda = \beta/\alpha$.
The cutting sequence of l we started with is not necessary to define the slope, but by analysing the cutting sequence we can determine the continued fraction expansion of l.

First, assume that the slope $\lambda > 1$. Then by looking at the grid one observes:

(1) between any two a's, there is at least one b, more precisely
(2) between any two a's, there are either $\lfloor\lambda\rfloor$ or $\lfloor\lambda\rfloor + 1$ b's.

If $\lambda < 1$, we replace a by b and $\lfloor\lambda\rfloor$ by $\lfloor\frac{1}{\lambda}\rfloor$ (in (2)).

Definition 4.2. Let λ be the slope of a line l with a cutting sequence satisfying (1) and (2). Then let

$$n := \begin{cases} \lfloor\lambda\rfloor, & \lambda > 1 \\ \lfloor\frac{1}{\lambda}\rfloor, & \lambda < 1 \end{cases}$$

be defined as the ***value*** of l.

Next, assume that we are looking at a cutting sequence satisfying (1) and (2) (with the same notation with a's and b's as before). Then we can rewrite that sequence if we set

$$\begin{pmatrix} a' \\ b' \end{pmatrix} := \begin{pmatrix} ab^n \\ b \end{pmatrix}. \tag{30}$$

Note that a cutting sequence of a line l can be rewritten in this way arbitrarily many times.
If we look at the successively obtained *values* $n_0, n_1, n_2, \dots$ of l, we get that the slope has the continued fraction form

$$\lambda = [n_0, n_1, n_2, \dots].$$

Note that we take the substitution

$$\begin{pmatrix} b' \\ a' \end{pmatrix} := \begin{pmatrix} a^n b \\ a \end{pmatrix} \tag{31}$$

if $\lambda < 1$.
The idea of how to prove this is by looking at the linear map

$$\begin{pmatrix} 1 & \\ n_k & 1 \end{pmatrix}$$

and how it acts on the square grid. The proof can be found in [Caroline S.].

A similar idea to the one above can be realised in *hyperbolic geometry*, where our grid will be the *Farey tessellation* of the upper-half plane $\mathcal{H}$.

Recall that the Fundamental domain $\mathcal{F}$ of the group G is given by an ideal quadrilateral with edges $-1, 0, 1$ and ∞. Now, consider the tessellation of $\mathcal{H}$ obtained by letting the isometries A, B (recall (26)) act on $\mathcal{F}$. Just as the maps a, b of the Euclidean plane generate the abelian group $\mathbb{Z}^2$, which

is the fundamental group of the torus, so the maps A, B of the hyperbolic plane generate the free group G of the once punctured torus. With that in mind, one can pose the same question as before:

Which A, B sequences occur as the cutting sequences of lines across l?

First note that the cutting sequences may now contain A^{-1} and B^{-1} additionally to A, B, depending on the direction in which we cut sides of l. Then every sequence occurs as a cutting sequence of some geodesic in $\mathcal{H}$, terminating sequences correspond to lines beginning or ending in the puncture (with one exception).
The exception is the periodic sequence $\ldots ABA^{-1}B^{-1}ABA^{-1}B^{-1}\ldots$. This corresponds to a loop encircling the puncture, which is a homotopy class with no geodesic representative.

One can use this to show that Markov irrationalities are exactly the endpoints of lifts of simple closed geodesics which are limits of simple closed geodesics on the punctured torus, but this requires more work. Here we would only like to explain the first step: namely how to use the Farey tessellation to obtain the continued fraction expansion of a number.

Consider the Farey tessellation $\mathcal{F}$ in $\mathcal{H}$, which is obtained by letting the group $\Gamma(2)$ act on a fundamental region, say $\mathcal{R}$, of $\mathcal{F}$. The sides of the fundamental region $\mathcal{R}$ are mapped to each other by the maps (see Figure 16)

$$Q := \begin{pmatrix} & -1 \\ 1 & \end{pmatrix} \quad \text{and} \quad W := \begin{pmatrix} 2 & 1 \\ 1 & \end{pmatrix}. \tag{32}$$

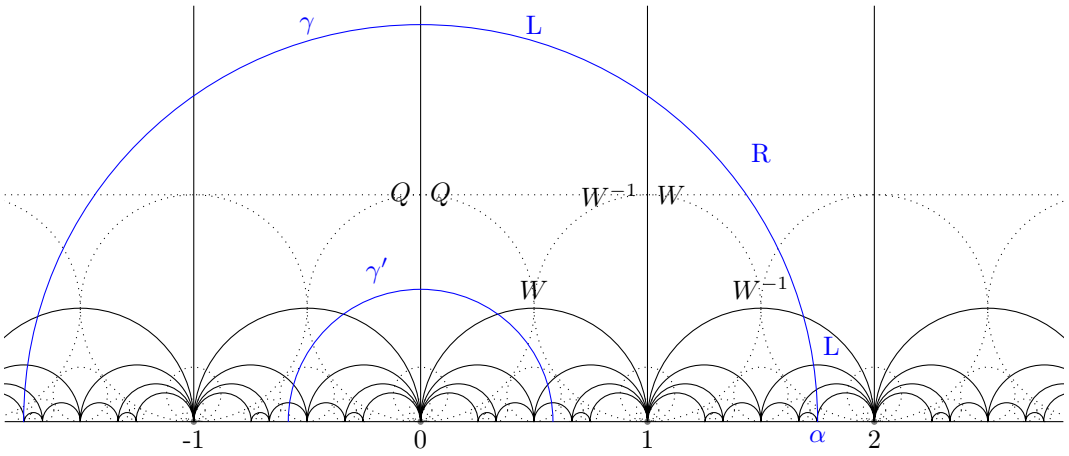


Figure 16: Cutting sequence

The cutting sequences of geodesics relative to $\mathcal{F}$ are of the form $\ldots QW^{n_1}QW^{n_2}Q\ldots$, where $n_i \in \mathbb{Z}$. Notice that Q^2 never appears since $Q^{-1} = Q$.
Since $SL_2(\mathbb{Z})$ is generated by Q, W and $\Omega := \begin{pmatrix} & -1 \\ 1 & -1 \end{pmatrix}$, its action preserves the tessellation $\mathcal{F}$ although not the Q, W labelling. One can label segments of geodesics cutting across the triangles in $\mathcal{F}$ so as to be invariant under $SL_2(\mathbb{Z})$, by labelling a segment L or R according to whether *the vertex of the triangle cut off by the segment is to the left or right*, as we have done in Figure 14. It is easy to write down a 'recipe' for conversion from Q, W to L, R sequences:

$$\begin{cases} QW \\ W^{-1}W^{-1} \\ WQ \end{cases} \longrightarrow L, \qquad \begin{cases} QW^{-1} \\ WW \\ W^{-1}Q \end{cases} \longrightarrow R.$$

It now follows that *two geodesics in $\mathcal{H}$ are equivalent under $SL_2(\mathbb{Z})$ is and only if their L, R sequences agree.*
But this is not the end; the L, R sequences bring us back to continued fractions. Let α be any positive real number, and let $\gamma(\alpha)$ be a geodesic ray joining *any* point on the imaginary axis to α

(see Figure 16). Reading off the L, R sequence of $\gamma(\alpha)$ we obtain a sequence

$$L^{n_0} R^{n_1} L^{n_2} \dots$$

(if $\alpha < 1$ the sequence begins with R, not L). Then

$$[n_0, n_1, n_2, \dots]$$

is the continued fraction expansion of α!

Proof: First, note that $n_0 = \lfloor \alpha \rfloor$. Let D be the point where $\gamma(\alpha)$ cuts $\alpha = n_0$.
Applying the map

$$\tau_1 : z \longrightarrow \frac{-1}{z} - n_0,$$

D is mapped to a point D' on the imaginary axis and $\gamma(\alpha)$ becomes a ray γ' through D' pointing in the negative direction with endpoint at $\frac{-1}{\alpha} - n_0$. The n_1 segments of type R in $\gamma(\alpha)$ which follow the initial n_0 segments of type L are now apparent as the n_1 vertical strips crossed by $\tau_1(\gamma)$ before it descends to $\tau_1(\alpha)$.
Thus $n_1 = \frac{1}{\alpha} - n_0$, so that

$$\alpha = n_0 + \frac{1}{n_1} + r, \;\; 0 < r < 1.$$

Now apply

$$\tau_1 : z \longrightarrow \frac{-1}{z + n_1}$$

to γ' and proceed as before. □

References

[S] Boris Springborn, *The hyperbolic geometry of Markov's theorem on Diophantine approximation and quadratic forms*. Technische Universitat Berlin, 2017

[F] Dr Lester R. Ford, *A Geometrical Proof of a Theorem of Hurwitz*. 1917

[A] Martin Aigner, *Markov's Theorem and 100 Years of the Uniqueness Conjecture*. Springer, Cham (2013)

[MP] Mareike Pfeil, *The Farey Tessellation*. Seminar: Geometric Structures on manifolds, 2015

[Markoff] A. Markoff, *Sur les formes quadratiques binaires indéfinies*. Mathematische Annalen, St. Petersbourg, 1879.

[Fricke] V.M. Buchstaber and A.P. Veselov, *Fricke identities, frobenius k-characters and Markov equation*. 2019

[HC] Harvey Cohn, *Approach to Markoff's minimal forms through modular functions*. Annals of Mathematics, U.S.A., 1955

[Caroline S.] Caroline Series, *The Geometry of Markoff Numbers*. The mathematical intelligencer vol. 7, no. 3, Springer-Verlag New York, 1985

[T] Mustafa Korkomaz and Athanase Papadopoulos, *On the ideal triangulation graph of a punctures surface* Annales de l'Institut Fourier, Association des Annales de l'Institut Fourier, 2012, 62 (4), p. 1367-1382.

[Hatcher] Hatcher, Allen E., *On triangulations of surfaces*. Topology Appl. 40, No. 2, 189-194 (1991)

[Mosher] L. Mosher, *Tiling the projective foliation space of a punctured surface*. Trans. Amer. Math. Soc. 306 (1988)

[Ex] M. Wohlfender, *Markov's Problem from a Hyperbolic Geometry point of view*. Quadratic forms, Markov numbers and Diophantine approximation, ETH Zurich (2020)