Hyperbolic geometry on Diophantine approximation

Xiao Yang

0 Introduction

The aim of this note is to provide a new proof of Hurwitz's theorem using hyperbolic geometry:

Theorem 0.1. (Hurwitz) *1. For all $x \in \mathbb{R} \smallsetminus \mathbb{Q}$ there are infinitely many pairs of integers r and s with $s > 0$ such that*

$$\left| x - \frac{r}{s} \right| < \frac{1}{\sqrt{5} s^2}. \tag{1}$$

2. Furthermore, the constant $\frac{1}{\sqrt{5}}$ is optimal. This means, that for $(\sqrt{5} + \epsilon)^{-1}$, there are irrationals $x \in \mathbb{R} \smallsetminus \mathbb{Q}$ such that the Inequality (1) *only holds for finitely many pairs of integers.*

Later we could prove the Markov's theorem using hyperbolic geometry as well.

1 Hyperbolic Plane and Horocycles

We denote $\mathbb{H} := \{z \in \mathbb{C} \mid \Im(z) > 0\} \subseteq \mathbb{C}$. It can be also viewed a 2 dimensional manifold since locally we could write the coordinates as $(\Re(z), \Im(z)) := (x, y)$.

Definition 1.1. The *hyperbolic plane* is $\mathbb{H}$ equipped with metric tensor

$$g = \frac{\mathrm{d}x \otimes \mathrm{d}x + \mathrm{d}y \otimes \mathrm{d}y}{y^2}.$$

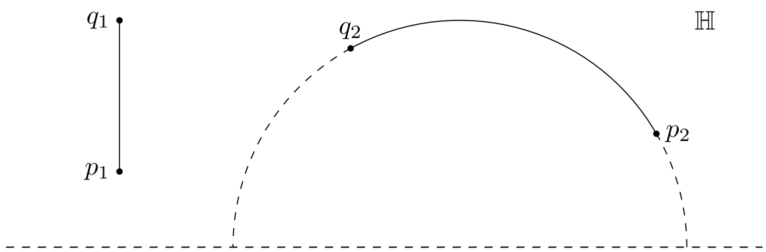
This means the length of a given piecewise smooth curve $c : [0, 1] \longrightarrow \mathbb{H}$ can be computed in the following way

$$L(c) := \int_0^1 \sqrt{g(\dot{c}(t), \dot{c}(t))} \mathrm{d}t = \int_0^1 \frac{|\dot{c}(t)|}{\Im(c(t))} \mathrm{d}t.$$

We should anticipate that the "lines" (geodesics) on $\mathbb{H}$ may not be "straight" anymore. But from the form of the metric tensor, we can see that the model is conformal, i.e., the angle between two intesecting lines is the same as the angles between their tangent lines as in Euclidean geometry.
We provide the two following lemmas to show how the geodesics on $\mathbb{H}$ look like and what the isometry group of $\mathbb{H}$ is. The proofs can be found in [THY12].

Lemma 1.2. *Let p, q be two distinct points in $\mathbb{H}$. When $\Re(p) = \Re(q)$, the geodesic connecting them is the (unique) line segment (in Euclidean sense) passing through them; when $\Re(p) \neq \Re(q)$, the geodesic connecting them is the arc between p and q of the (unique) semi circle (in Euclidean sense) passing through p, q whose center is lying on the real axis (of the complex plane). Precisely, the distance between p and q is*

$$d(p,q) = \log\left(\frac{|p-q| + |p+\bar{q}|}{2\sqrt{\Im(p)\Im(q)}}\right). \tag{2}$$



Pic. 1

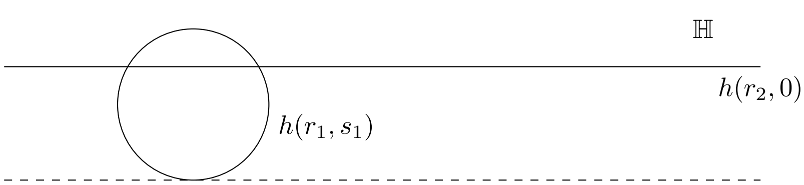
Let $\mathrm{Isom}(\mathbb{H})$ denote the group of all isometries of $\mathbb{H}$ and $\mathrm{Isom}^+(\mathbb{H})$ denote the group of all orientation-preserving isometries of $\mathbb{H}$. We have:

Lemma 1.3. $\mathrm{Isom}(\mathbb{H}) \cong \mathrm{PGL}_2(\mathbb{R})$ *and* $\mathrm{Isom}^+(\mathbb{H}) \cong \mathrm{PSL}_2(\mathbb{R})$. *Precisely, every isometry M of $\mathbb{H}$ is given by the action of a matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PGL}_2(\mathbb{R})$, on $\mathbb{H}$:*

$$M_A : z \longmapsto A.z := \begin{cases} \dfrac{az+b}{cz+d} & \text{if } \det(A) > 0 \\ \dfrac{a\bar{z}+b}{c\bar{z}+d} & \text{if } \det(A) < 0 \end{cases}.$$

$\mathrm{PGL}_2(\mathbb{R}) := \mathrm{GL}_2(\mathbb{R})/\{\lambda \,\mathrm{Id} \mid \lambda \in \mathbb{R}^\times\}$ and $\mathrm{PSL}_2(\mathbb{R}) := \mathrm{SL}_2(\mathbb{R})/\{\pm \mathrm{Id}\}$. $\mathrm{PSL}_2(\mathbb{R})$ is a proper subgroup of $\mathrm{PGL}_2(\mathbb{R})$. E.g, the reflection over the y-axis is an isometry but not orientation-preserving one.

Definition 1.4. Let $r, s \in \mathbb{R}$ but cannot be zero simultaneously. A *horocycle* $h(r,s)$ with *parameters* (r,s) is:
1. $h(r,s) := \{z \in \mathbb{C} \mid \Im(z) = r^2\}$ when $s = 0$.
2. $h(r,s) := \{z \in \mathbb{C} \mid |z - \frac{r}{s} - \frac{1}{2s^2} i| = \frac{1}{2s^2}\}$ when $s \neq 0$.

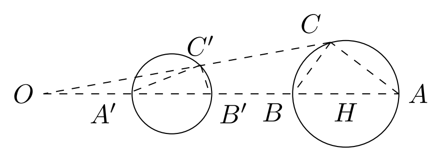


Pic. 2

We can see a horocycle is either a Euclidean circle tangent to the real axis or a Euclidean horizontal line. The case when $s = 0$ can be thought as a limit case of the case when $s \neq 0$, i.e., a circle of radius infinity. The point ∞ and the tangent point of circle on the real axis do not belong to $\mathbb{H}$. However, it is sometimes more convenient to consider the complex plane is *compactified* by ∞. So sometimes we call the Euclidean line with infinity and a Euclidean circle a *generalized circle*. And some naïve computations involved with ∞ are allowed: $\infty + b = \infty$, $\lambda\infty = \infty$, $\overline{\infty} = \infty$, $\frac{1}{\infty} = 0$, and $\frac{1}{0} = \infty$ with $b \in \mathbb{C}$ and $\lambda \in \mathbb{C}^\times$.

Lemma 1.5. *Generalized circles are sent to generalized circles under translation* $z \longmapsto z + b$*, dilation* $z \longmapsto \lambda z$*, conjugation* $z \longmapsto \bar{z}$*, and inversion* $z \longmapsto \frac{1}{\bar{z}}$ *with* $\lambda \in \mathbb{C}^\times, b \in \mathbb{C}$.

Proof. It is trivial to see that the lemma is true for the first three cases thanks to (primary) school geometry.
For the inversion, we could prove it with some elementary computation. But there is a nice geometry trick (inversion with respect to a circle). Given a circle $\odot H$, we want to compute its image under the inversion $z \longmapsto \frac{1}{\bar{z}}$. Link the center H and the origin O. The line $\overline{OH}$ intersect $\odot H$ at A and B. Pick an arbitrary point C on $\odot H$. Denote the inverted point of A, B, C by A', B', C' respectively. See Pic. 3.



Pic. 3

By definition of the inversion map, we have the points, its inverted image and the origin are colinear. Furthermore, $|OA'||OA| = |OB'||OB| = |OC'||OC| = 1$. Consider $\triangle OCA$ and $\triangle OA'C'$. They share one angle, i.e., $\angle COA = \angle A'OC'$. Rearranging the equality of sides, we have $\frac{|OC'|}{|OA|} = \frac{|OA'|}{|OC|}$. Hence, $\triangle OCA \sim \triangle OA'C'$. Then $\angle CAO = \angle OC'A'$. Similarly, $\triangle OCB \sim \triangle OB'C'$. Then $\angle CBO = \angle OC'B'$. Note that $\angle CBO = \angle CAO + \angle BCA = \angle CAO + 90°$. Thus, $\angle CAO + 90° = \angle OC'A' + \angle A'C'B'$. Then $\angle A'C'B'$ is a right angle. Thus, C' lies on the circle with diameter $\overline{A'B'}$. ■

Lemma 1.6. *For* $A \in \mathrm{GL}_2(\mathbb{R})$ *with* $\det(A) = \pm 1$ *and for* $v \in \mathbb{R}^2 \smallsetminus \{(0,0)\}$ *the hyperbolic isometry* M_A *maps the horocycle* $h(v)$ *to* $h(Av)$ (*with at most one point missing*) *where we identify* Av *is the canonical matrix-vector multiplication (regarding* v *as a column vector).*

Proof. An isometry of $\mathbb{H}$ can be considered as compositions of the $z \longmapsto z + b$, $z \longmapsto \lambda z$, $z \longmapsto -\bar{z}$ and $z \longmapsto \frac{1}{\bar{z}}$ which correspond to the following matrices

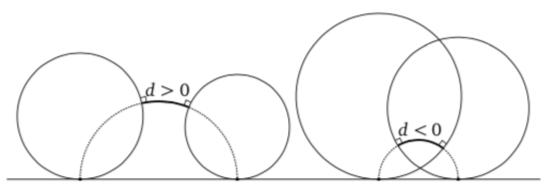
$$\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} \sqrt{\lambda} & 0 \\ 0 & \sqrt{\lambda^{-1}} \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \text{ and } \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

where $b \in \mathbb{R}$ and $\lambda \in \mathbb{R}^+$. With the help of Lemma 1.5, the rest will be elementary computation. For example, $z \longmapsto z+b$ is the translation to the right by b. If $s=0$, the translation does nothing on the horocycle and $A(r,0)=(r,0)$. So $A.h(v)=h(Av)$. If $s \neq 0$, the translation shifts the horocycle $h(v)$ to the left for b and $A(r,s)=(r+bs,s)$. The radius is the same $\frac{1}{2s^2}$ as before and the center is $\frac{r+bs}{s}+\frac{i}{2s^2}=\frac{r}{s}+b+\frac{i}{2s^2}$. So $A.h(v)=h(Av)$ again. ■

Definition 1.7. Let $h_1 := h(r_1,s_1), h_2 := (r_2,s_2)$ be two horocycles. The *signed distance* between h_1 and h_2 is defined to as

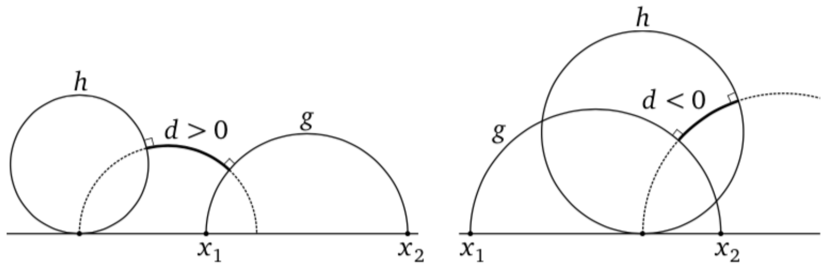
$$d(h_1,h_2) := \begin{cases} 2\log(|r_1s_2 - s_1r_2|) & \text{if } r_1s_2 - s_1r_2 \neq 0 \\ -\infty & \text{otherwise} \end{cases}.$$

Remark 1.8. The signed distance between two horocycles h_1 and h_2 can be elaborated in the following way:
1. If the two horocycles osculate at the same point on the real axis or one of them is a Euclidean horizontal line, the signed distance is $-\infty$.
2. If the two horocycles osculate at different points on the real axis and are disjoint, the signed distance is the length between the two intersecting points on $\mathbb{H}$ between the horocycle the hyperbolic geodesic (semi-circle in the Euclidean sense) with the two osculating points as limit points.
3. If the two horocycles osculate at the different points on the real axis and intersect, the signed distance is the same length as in 2 but taken negative.
4. If the two horocycles osculate at the different points on the real axis and tangent, the signed distance is zero.



Pic. 4

Definition 1.9. For a horocycle h and a geodesic g, the signed distance $d(h,g)$ is defined as
1. If h and g do not intersect, then $d(h,g)$ is the length of the geodesic segment connecting h and g and orthogonal to both.
2. If h and g do intersect, then $d(h,g)$ is the length of that geodesic segment taken negative.
3. If h and g are tangent then $d(h,g)=0$.
4. If g ends in the osculating point of h on the real axis then $d(h,g)=-\infty$.

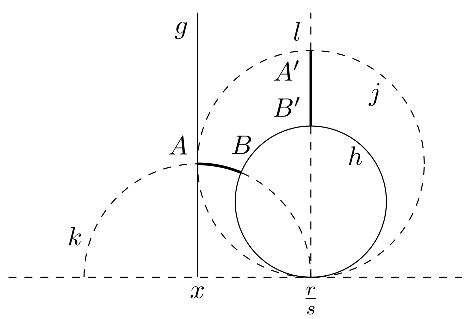


Pic. 5

Proposition 1.10. *Let $h = h(r, s)$ with $s \neq 0$ and a vertical geodesic g on $\mathbb{H}$ emanating from $x \in \mathbb{R}$. Their signed distance is*

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

Proof. First, consider the case when the line and the horocycle are disjoint. As is depicted in Pic. 6, we want to show that the distance measured by the geodesic segment $\overline{AB}$ is the same as the vertical one $\overline{A'B'}$. We need to find an isometry sends $\overline{AB}$ to $\overline{A'B'}$.

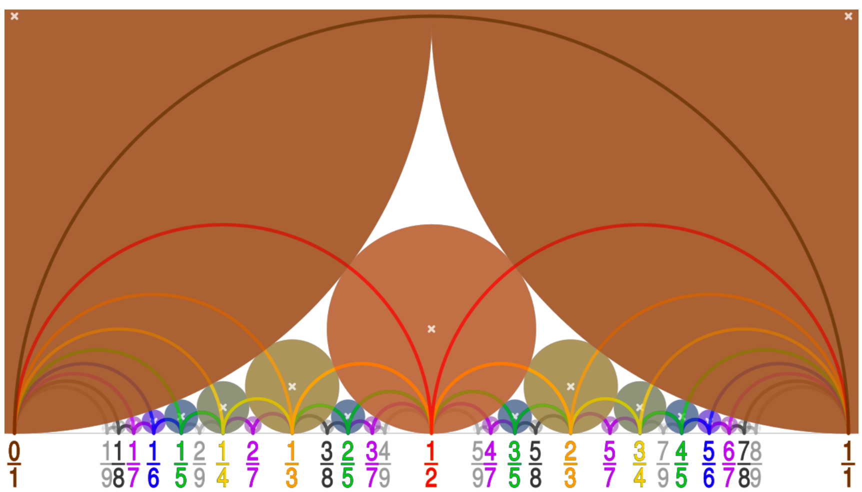


Pic. 6

We can always apply a translation and a dilation. Hence, we reduce to the case $r = 0, s = 1$. We claim that the desired isometry should be $z \longmapsto \frac{z}{\frac{1}{2}z+1}$. Applying Lemma 1.6, we can see that the inner and outer horocycles h and j are preserved. The geodesic semicircle k centered at x is sent to the geodesic vertical line l emanating from $\frac{r}{s} = 0$ which can be computed by Lemma 1.5. After transformation the images of endpoint A must still lie on the horocycle j and the vertical geodesic l. Hence, A' is the only possibility. So is for B and B'. After knowing this, we can set the number r, s, x back (without assuming $r = 0, s = 1$). By some simple Euclidean geometry, we can get the coordinates $A' = (\frac{r}{s}, 2|x - \frac{r}{s}|)$ and $B' = (\frac{r}{s}, \frac{1}{s^2})$. Then the distance easily by (2).
The case when the line and the horocycle are intersecting is analogous. The only difference is the sign and the position of these circles and lines. ■

Now, we take the parameter of horocycles (r, s) from $\mathbb{Z}^2 \setminus \{(0,0)\}$. After doing this, we could link geometry to number theory. Let $h(r_1, s_1), h(r_2, s_2)$ be two horocycles with integral parameters. By Definition 1.7 and Remark 1.8 if $s_1, s_2 \neq 0$ and $\frac{r_1}{s_1} = \frac{r_2}{s_2}$ the two horocycles have a common osculating on the real axis. If $s_1, s_2 \neq 0$ and $\frac{r_1}{s_1} \neq \frac{r_2}{s_2}$, they will never intersect and are tangent to each other if and only if $r_1 s_2 - r_2 s_1 = \pm 1$ which happens only if (r_1, s_1) and (r_2, s_2) coprime. Hence, $\frac{r_1}{s_1}$ and $\frac{r_2}{s_2}$ represent reduced fraction of rational numbers. We, hence, will remove the cases where the horocycles osculate at the same point on the real axis.

Definition 1.11. A *Ford circle* is a horocycle with integral parameters $h(r, s)$ such that r, s are coprime.
A *triangle* on $\mathbb{H}$ is a subset of $\mathbb{H}$ surrounded by three geodesic segments.
An *ideal triangle* on $\mathbb{H}$ is a subset of $\mathbb{H}$ surrounded by three geodesic with three distinct points from $\mathbb{R} \cup \{\infty\}$ as their limit end points.
The Farey tessellation of $\mathbb{H}$ is an ideal triangulation of $\mathbb{H}$ with vertex set $\mathbb{Q} \cup \{\infty\}$.



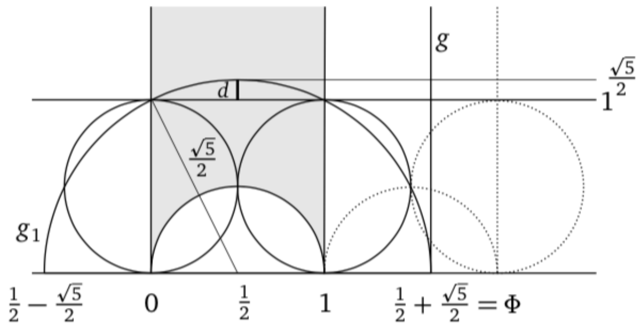
Pic. 7

Lemma 1.12. *Suppose a geodesic g crosses an ideal triangle T of the Farey tessellation. If g is one of the three geodesics bisecting two sides of T, then*

$$d(h, g) = -\log\left(\frac{\sqrt{5}}{2}\right)$$

for all three Ford circles h at the vertices of T.

Proof. After applying a proper isometry, we can reduce to the case of ideal triangle with end points $0, 1, \infty$.



Pic. 8

The rest of the argument is simply Euclidean (school) geometry task and translation into Hyperbolic geometry as is shown in Pic. 8. ■

Now we are prepared to prove the Hurwitz's theorem (Theorem 0.1).

Proof. 1. We first proof part 1 of Theorem 0.1. Let x be an irrational and g be the vertical geodesic from x to ∞. This part of Hurwitz's theorem is equivalent to: there are infinitely many Ford circles h satisfying

$$d(g,h) < -\log\left(\frac{\sqrt{5}}{2}\right). \tag{3}$$

Note that g passes through infinitely many (ideal) triangles of the Farey tessellation. For each of these triangles, at least one of its Ford circles h satisfies Inequality (3) by Lemma 1. For consecutive triangles that g crosses, the same horocycle may satisfy Inequality (3). But this can happen only finitely many times (otherwise x would be rational), and then the geodesic will never again intersect a triangle incident with this horocycle. Hence, infinitely many Ford circles satisfy Inequality (3), and this completes the proof of part 1. □

2. For the second part, we will again show for $\Phi := \frac{\sqrt{5}+1}{2}$, g the geodesic from Φ to ∞. and $\epsilon > 0$ there are only finitely many Ford circles h that satisfy

$$d(g,h) < -\log\left(\frac{\sqrt{5}}{2}\right) - \epsilon. \tag{4}$$

To this end, let g_1 be the geodesic from Φ to $1-\Phi$. For every Ford circle h,

$$d(h,g_1) \geqslant -\log\left(\frac{\sqrt{5}}{2}\right).$$

Indeed, the distance is equal to $-\log\left(\frac{\sqrt{5}}{2}\right)$ for all Ford circles that g_1 intersects, and positive for all others. Because the geodesics g and g_1 converge at the common end Φ, there is a point $p \in g$ such that all Ford circles h intersecting the ray from p to Φ satisfy

$$|d(g, \Phi) - d(g_1, \Phi)| < \epsilon,$$

and hence

$$d(g, \Phi) \geqslant -\log\left(\frac{\sqrt{5}}{2}\right) - \epsilon.$$

On the other hand, the complementary ray of g, from p to ∞, intersects only finitely many Ford circles. Hence, only finitely many Ford circles satisfy Inequality (4), and this completes the proof of part 2. ■

References

[SPR19] B. Springborn, *The hyperbolic geometry of Markov's theorem on Diophantine approximation and quadratic forms*, arxiv.org, 2019

[THY12] D. Thurston & Q. Yuan, *274 Curves on Surfaces, Lecture 4*, math.berkeley.edu/~qchu/Notes/274/Lecture4.pdf, 2012