NUMBER THEORY OF BI CUSPIDAL GEODESICS

GREG MCSHANE

ABSTRACT. We discuss the relationship between Penner's λ-lengths, with both Fermat's theorem on representation of a prime as the sum of squares and the Markoff spectrum. The text follows pretty closely the two talks I gave in June 2021 for the online meeting. We have included an informal discussion of some work on sums of squares which was suggested by questions raised at the meeting.

1. INTRODUCTION

The *Farey tessalation* is a fundamental object in the theory of Fuchsian groups. It is a tessalation of hyperbolic space by *ideal triangles*.

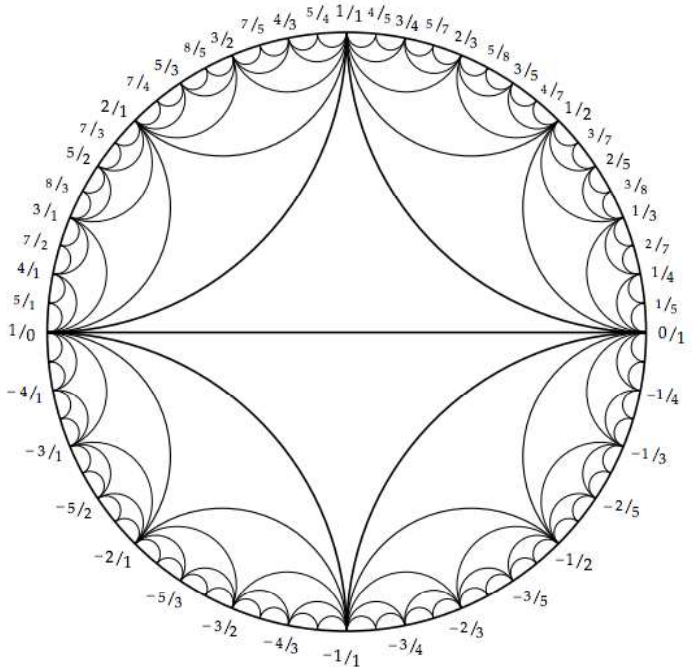

FIGURE 1. Farey diagram.

The tessalation is invariant under the action of the modular group $\Gamma = \mathrm{PSL}(2,\mathbb{Z})$. This group has a pair of (torsion free) normal subgroups of index 6 namely:

- the principal congruence subgroup $\Gamma(2)$
- the derived subgroup $\Gamma' = [\Gamma, \Gamma]$.

This work was partially funded by the Equipe Action ToFu part of Persyval-Lab.

The quotient of the upper half space by the modular group is the *modular orbifold* which has a single cusp and two cone points. The modular orbifold admits two degree 6 covers corresponding to the normal subgroups above

- the three punctured sphere $\mathbb{H}/\Gamma(2)$
- the once punctured modular torus $\mathbb{H}/\Gamma'$.

Both of these surfaces are non compact but have finite hyperbolic area equal to 2π. The non compactness is due to the presence of *cusps* - $\mathbb{H}/\Gamma(2)$ has exactly 3 cusps and $\mathbb{H}/\Gamma'$ has just one. These surfaces have many interesting geometric properties well documented in the literature but in this text we will be concerned principally with their relation to problems in number theory. The first of these is Fermat's theorem on primes which are the sums of two squares and the second is Frobenius' unicity conjecture for Markoff numbers. In each case we will see that the question can be stated in terms of lengths of bi cuspidal geodesics on one of these surfaces. The geodesic edges of the Farey tessalation yield a set of three such geodesic, in fact the shortest bi cuspidal geodesics, on the quotient surfaces. The geodesics obtained from projecting the Farey tessalation are *simple*, that is they have no self intersections, and their complement consists of a pair of ideal triangles. Of course, a bi cuspidal geodesic α has infinite length with respect to the Poincaré metric but one can define a useful geometric quantity by truncating the geodesic ie removing a portion of infinite length which is inside the cusp regions of the surface. This idea of associating a finite length to an *arc*, that is a simple bi cuspidal geodesic, (paragraph 2.1) appears in Penner's work on moduli [15]. He defined the λ-length of simple bi cuspidal geodesic on a punctured surface to be the exponential of the length of the portion outside of some fixed system of cusp regions. Lemma 2.3 shows that in the context we consider a λ length is always the square of the determinant of an integer matrix.

In the following text we show how considerations of λ- length lead to proofs of:

- Fermats's theorem on which primes are the sum of two squares
- A result of Baragar-Button-Zhang on the uniqueness of certain Markoff numbers.

The connections between continued fractions and hyperbolic geodesics on the modular surface have been known since Hedlund, Hopf and others began investigating the geodesic flow of hyperbolic surfaces.

1.1. **Sums of squares.** The following pair of results are the basis of many undergraduate courses on elementary number theory:

Theorem 1.1. Let p be a prime then the equation

$$x^2 = -1$$

admits a solution in $\mathbb{F}_p$ iff $p = 2$ or $p - 1$ is a multiple of 4.

Theorem 1.2 (Fermat)**.** Let p be a prime then the equation

$$x^2 + y^2 = p$$

has a solution in integers iff $p = 2$ or $p - 1$ is a multiple of 4.

They are intimately linked. The first is an immediate corollary of the second but it is also possible to deduce the second as a corollary of the first, for example, by using unique factorisation in the Gaussian integers. In an article with Vlad Sergesciu [13] we present a unified geometric approach to these results using the theory of group actions and in particular an application of Burnsides's Lemma. As with Zagier's remarkable proof

[20] (see also [10, 16, 2, 6] for closely related constructions and discussion) both follow from showing that a certain involution has a fixed point. Amusingly Burnsides's Lemma reduces this to showing that another involution has exactly two fixed points:

- In the proof of Theorem 1.1 this is a consequence of the fact that a quadratic equation over a field has at most two solutions.
- In the proof of Theorem 1.2 this follows from some geometry and the fact that

$$\begin{pmatrix} k+1 & k-1 \\ p & p \end{pmatrix} = 2p \neq 2. \tag{1}$$

We will see later that this determinant can be interpreted as (essenially) the length of a bi cuspidal geodesic on the three punctured sphere $\mathbb{H}/\Gamma(2)$.

For each integer n, the automorphisms of $\mathbb{H}/\Gamma(2)$ act on the set of bi cuspidal geodesic on this surface of λ length n. In [13] we show (Lemma 5.2) that if p is congruent to 1 modulo 4 there is a certain orientation preserving involution which leaves one of these geodesics invariant and from this we deduce Theorem 1.2.

1.1.1. *Heath-Brown's proof.* In 1984 Heath-Brown published a proof of Theorem 1.2 apparently in the journal of the Oxford University undergraduate mathematics society. His proof arose from a study of the account of Liouville's papers on identities for parity functions. Zagier's celebrated one line proof [20] is a clever reformulation of this argument. Heath-Brown studies the action of a Klein four group on a finite set and considerations of parity. To define his set Heath-Brown introduces an auxiliary equation namely

$$p = 4xy + z^2$$

whereas in our proof the sum of squares decomposition arises directly as the result of a geometric construction. As such, the motivation for our work is to show that the finite sets involved in the proof can be chosen to be both natural and have a geometric interpretation. For example, in Section 2 we give a proof of Theorem 1.1 using a group generated by

$$\begin{aligned} x &\mapsto -x \\ x &\mapsto 1/x \end{aligned}$$

and in the proof of Theorem 1.2 our group is conjugate to a group generated by

$$\begin{aligned} z &\mapsto -\overline{z} \\ z &\mapsto 1/\overline{z} \end{aligned}.$$

Although we use the Burnside lemma in [13] it is not essential to our argument and that one can achieve the same reduction by considering the signature of the permutations associated to the involutions we consider. This approach is closer to the parity arguments in Heath-Brown [10].

1.2. **Markoff numbers.** A *Markoff triple* is a solution (X, Y, Z) in positive integers to the *Markoff cubic*

$$X^2 + Y^2 + Z^2 - 3XYZ = 0. \tag{2}$$

A *Markoff number* is an integer in a Markoff triple. It is classical that one can associate a tree, the *Markoff tree* (see Figure 2) to the set of all Markoff triples - the set of vertices is just the set of triples and two triples are joined by an edge iff they share a pair of Markoff number.

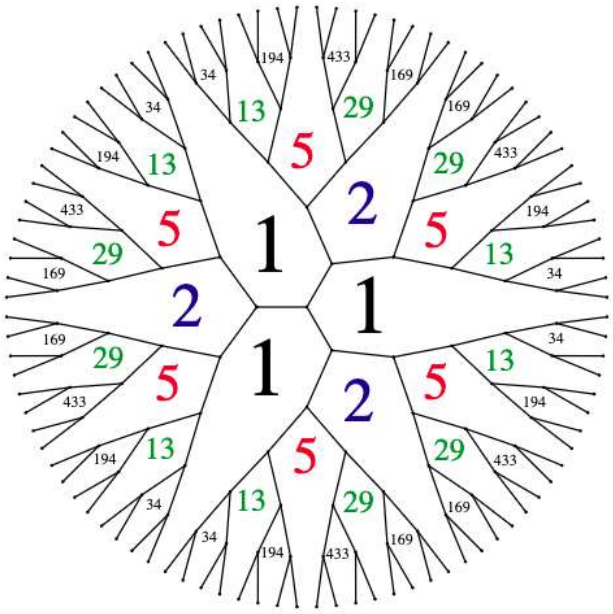

FIGURE 2. Markoff tree.

H. Cohn initiated an approach which culminated in a correspondence (see [8, 9]) between Markoff numbers and the lengths of simple closed geodesics on the *modular torus* $\mathbb{H}/\Gamma'$, where $\Gamma' < P\mathrm{SL}(2,\mathbb{Z})$ is the commutator subgroup. More precisely, if γ is such a geodesic then :

$$X = \frac{2}{3}\cosh\left(\frac{\ell_\gamma}{2}\right), \tag{3}$$

is a Markoff number where ℓ_γ is the length of γ. Conversely, every Markoff number arises as the length of such a geodesic.

Less well known is that there is a connection between λ-lengths and Markoff numbers showing that every such number is the sum of two squares without applying Theorem 1.2. By using calculations in Wolpert [19] (for background on the various coordinate systems he introduced see also [11]) one can show that, for a suitable choice of cusp region on the modular torus the λ-lengths of arcs coincide with the squares of Markoff numbers. We give a proof of the following result which is implicit in [19]:

Theorem 1.3. For each Markoff triple (X, Y, Z) there is a (unique) ideal triangulation of the modular torus such that the λ-lengths of the arcs are X^2, Y^2, Z^2.

Then, using the fact that each arc is invariant under the elliptic involution of the torus one can show, using Lemma 2.1, that every Markoff number is the sum of two squares. In fact this was the observation that was the starting point for this work. We then proceed to show in Theorem 6.4 the uniqueness for Markoff numbers satisfying certain arithmetic conditions following Baragar and Button (see also [14, 21, 22] for alternative approaches.)

1.3. **Thanks.** The first author thanks Louis Funar, Hidetoshi Masai and Vlad Sergesiu for many useful conversations over the years concerning this subject. He would also like to thank Xu Binbin for reading early drafts of the manuscript.

He would also like to thank the organisers of the conference and appreciates their work during the emergency situation.

2. Reciprocals of sums of squares and arcs

This group of integer matrices $\mathrm{SL}(2,\mathbb{Z})$ acts on $\mathbb{H}$ by linear fractional transformations that is:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}),\ z \in \mathbb{H},\ \begin{pmatrix} a & b \\ c & d \end{pmatrix}.z = \frac{az+b}{cz+d}.$$

The key lemma that relates this $\mathrm{SL}(2,\mathbb{Z})$ action to sums of squares is:

Lemma 2.1. Let n be a positive integer. The number of ways of writing n as a sum of squares

$$n = c^2 + d^2$$

with c, d coprime positive integers is equal to the number of integers $0 \leq k < n-1$ coprime to n such that the line

$$\{k/n + it,\ t > 0\}$$

contains a point in the $\mathrm{SL}(2,\mathbb{Z})$ orbit of i.

Proof. Suppose there is such a point which we denote w. The point w is a fixed point of some element of order 2 in $\mathrm{SL}(2,\mathbb{Z})$. Since the Ford circles are $\mathrm{SL}(2,\mathbb{Z})$ invariant this element must permute F with the Ford circle tangent to the real line at the real part of w. So, in particular, w is the midpoint of the line that it lies on and by Lemma 2.3 one has:

$$\frac{1}{n} = \mathrm{Im}\,\frac{1}{n}(k+i) = \mathrm{Im}\,\frac{ai+b}{ci+d} = \frac{\mathrm{Im}\, i}{c^2+d^2}.$$

Conversely if c, d are coprime integers then there exists a, b such that

$$ad - bc = 1 \Rightarrow \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}).$$

By applying a suitable iterate of the parabolic transformation $z \mapsto z+1$, one can choose w such that $0 \leq \mathrm{Re}\, w < 1$. So if $n = c^2 + d^2$ then $\frac{ai+b}{ci+d}$ is on one of the lines of the family in the statement.

□

2.1. **Ford circles, lengths, midpoints.** Lemma 2.1 illustrates the connexion between sums of squares, the orbit $\mathrm{SL}(2,\mathbb{Z}).i$ and Poincare geodesics. We now recall some standard ideas from hyperbolic geometry which in particular will allow us to give an intuitive definition of our set X in the next section. We define an *arc* to be a Poincare geodesic with endpoints in $\partial\mathbb{H}$ a pair of extended rationals, that is elements of $\mathbb{Q} \cup \infty$.

We denote by F the set $\{z, \mathrm{Im}\, z > 1\}$ this is a *horoball in* $\mathbb{H}$ centered at ∞. The image of F under the $\mathrm{SL}(2,\mathbb{Z})$ action consists of F and infinitely many disjoint discs, which we will refer to as *Ford circles*, each tangent to the real line at some rational m/n. We adopt the convention that F is also a Ford circle of infinite radius tangent to the extended real line at $\infty = 1/0$.

The following is well known and is easily checked:

Lemma 2.2. .

(1) The Ford circle tangent to the real line at m/n has Euclidean diameter $1/n^2$.

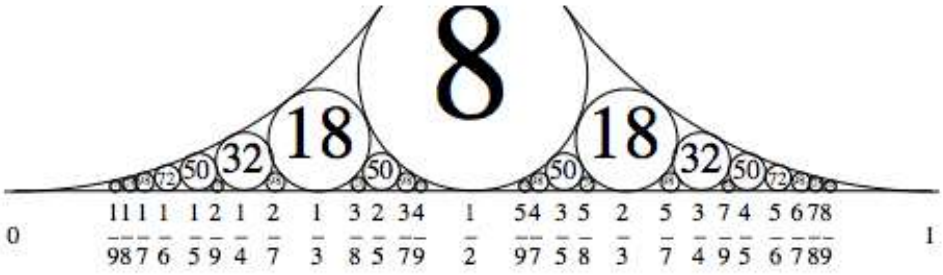

FIGURE 3. Ford circles with tangent points and curvatures. Recall that the curvature of a euclidean circle is the reciprocal of the square of its radius.

(2) The closures of a pair of distinct Ford circles are either disjoint or meet in a point of the $\mathrm{SL}(2,\mathbb{Z})$-orbit of i.

Let $a/c, b/d$ be a pair of distinct rationals. We define the *length* of the arc joining these rationals to be the length, with respect to the Poinacre metric on $\mathbb{H}$, of the portion of this arc outside of the Ford circles tangent at $a/c, b/d$. Further we define its *mid point* to be the midpoint of this sub arc.

Following Penner [15] we define the λ-length of the arc to be the exponential of this length. It is a consequence of Lemma 2.3 below that the arcs of λ-length 1 are the edges in the so-called *Farey diagram* (see Figure 1). The lemma is a simple exercise left to the reader:

Lemma 2.3. Let $a/c, b/d$ be a pair of distinct extended rationals. Then the λ-length of the arc joining them is the square of the determinant of the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Further if $a/b = 1/0$ then the arc is a vertical line whose midpoint has imaginary part equal to $1/d$.

2.2. **Continued fractions.** Though we will not use we feel that we must mention the work of Caroline Series [18] which studies the combinatorics of hyperbolic geodesics γ starting at the point $i \in \mathbb{H}$ and with an end point on the real line γ^+. She relates the *cutting sequence* for γ to the continued fraction expansion of γ. In particular she introduces the idea of *pivot sequence* for a geodesic. As the geodesic γ crosses the Farey triangulation then it separates each triangle in its path into two components one of which has a single spike the other has two. The pivot sequence is a sequence of integers which count the number of consecutive single spike components on the left and on the right of γ. If γ^+ is rational then there is an ambiguity because the final trangle of the Farey tessalation is cut into two components which each have a single spike but this is doesn't pose any serious problems in practice.

3. The three punctured sphere

We consider $\Gamma(2)$, the principal level 2 congruence subgroup of $\mathrm{SL}(2,\mathbb{Z})$. This group acts on $\mathbb{Z}^2$, that is pairs of integers, preserving parity.

It also acts on $\mathbb{H}$ by linear fractional transformations that is:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}),\ z \in \mathbb{H},\ \begin{pmatrix} a & b \\ c & d \end{pmatrix}.z = \frac{az+b}{cz+d}.$$

The quotient $\mathbb{H}/\Gamma(2)$ is conformally equivalent to the Riemann sphere minus three points which we will refer to as *cusps* (see Figure 5). Following convention we label these cusps $0, 1, \infty$ respectively corresponding to the three $\Gamma(2)$ orbits of $\mathbb{Q} \cup \infty$. Finally, the *standard fundamental domain* for $\Gamma(2)$ is the convex hull of the points $\infty, -1, 0, 1$. This region can be decomposed into two ideal triangles $\infty, -1, 0$ and $0, 1, \infty$ as in Figure 4. The edges of the ideal triangles project to three disjoint simple geodesics on $\mathbb{H}/\Gamma(2)$ and each edge has a *midpoint* which is a point of the $\mathrm{SL}(2,\mathbb{Z})$ orbit of i (see Figure 5).

3.0.1. *Cusp regions.* The image of a Ford circle on $\mathbb{H}/\Gamma(2)$ is a *cusp region* around one of the three cusps $0, 1, \infty$. Pairs of these cusp regions are tangent at one of the midpoints labelled $i, 1+i, \frac{1}{2}(1+i)$. It is not difficult to see that these cusp regions are permuted by the automorphisms of $\mathbb{H}/\Gamma(2)$. It follows that if an automorphism preserves a geodesic joining cusps on $\mathbb{H}/\Gamma(2)$ then it must permute the Ford regions at each end of a lift to $\mathbb{H}$.

3.1. **Automorphism groups of $\mathbb{H}/\Gamma(2)$.** From covering theory an isometry of $\mathbb{H}$ induces an automorphism of $\mathbb{H}/\Gamma(2)$ iff it normalises the covering group i.e. $\Gamma(2)$. It follows that, since $\Gamma(2)$ is a normal subgroup of $\mathrm{SL}(2,\mathbb{Z})$, the quotient group

$$H^+ := \mathrm{SL}(2,\mathbb{Z})/\Gamma(2)$$

acts as a group of (orientation preserving) automorphisms of the surface $\mathbb{H}/\Gamma(2)$. More generally, $\Gamma(2)$ is normal in $\mathrm{GL}(2,\mathbb{Z})$ and

$$H := \mathrm{GL}(2,\mathbb{Z})/\Gamma(2)$$

acts as a group of possibly orientation reversing automorphisms of the surface $\mathbb{H}/\Gamma(2)$.

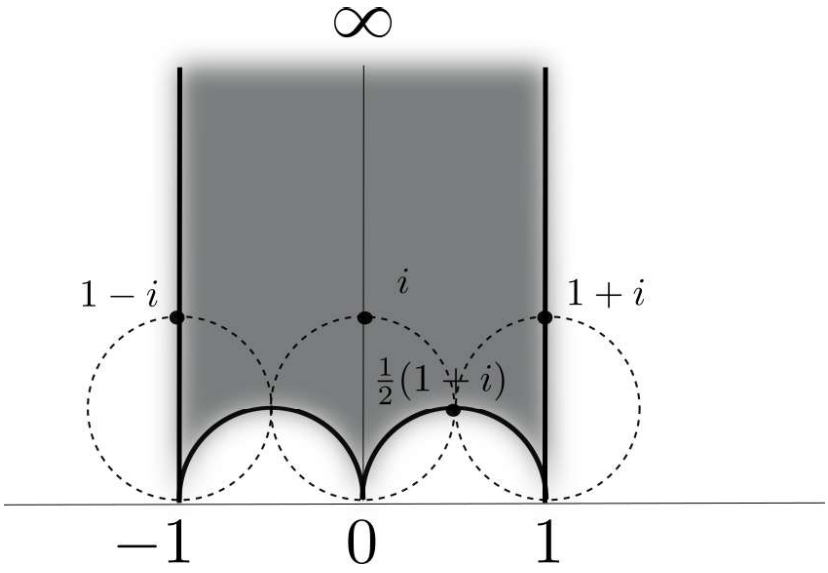

FIGURE 4. Standard fundamental domain for $\Gamma(2)$ and its decomposition into ideal triangles.

3.2. **Orientation reversing automorphisms.** To prove Theorem 1.2 we will have to work with automorphisms that do not preserve the orientation and in particular those induced by the involutions:

$$\begin{aligned} U : z &\mapsto -\bar{z} \\ V : z &\mapsto 1 - \bar{z}. \end{aligned}$$

Both U and V normalise $\Gamma(2)$ so induce automorphisms of $\mathbb{H}/\Gamma(2)$. In fact, since V is the composition of U and $z \mapsto z+1$, it suffices to show that U normalises $\Gamma(2)$. This is easy to check, for if $a, b, c, d \in \mathbb{Z}$ and $f(z) = (az+b)/(cz+d)$ then one has:

$$U \circ f \circ U^{-1}(z) = -\overline{f(\bar{-z})} = -f(-z) = \frac{az-b}{-cz+d},$$

so conjugation does not change the parity of a, b, c, d and it follows that U normalises $\Gamma(2)$.

3.3. **Klein four group of automorphisms.** The pair of involution U, V generate a group of isometries of $\mathbb{H}$, which we denote by $\hat{K}^{\infty}$, isomorphic to the infinite dihedral group D_{∞} infinite dihedral group. One checks that

$$U \circ V(z) = V \circ U(z) = z + 1$$

and we note that

$$z + 1 = \frac{z+1}{0+1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} .z,$$

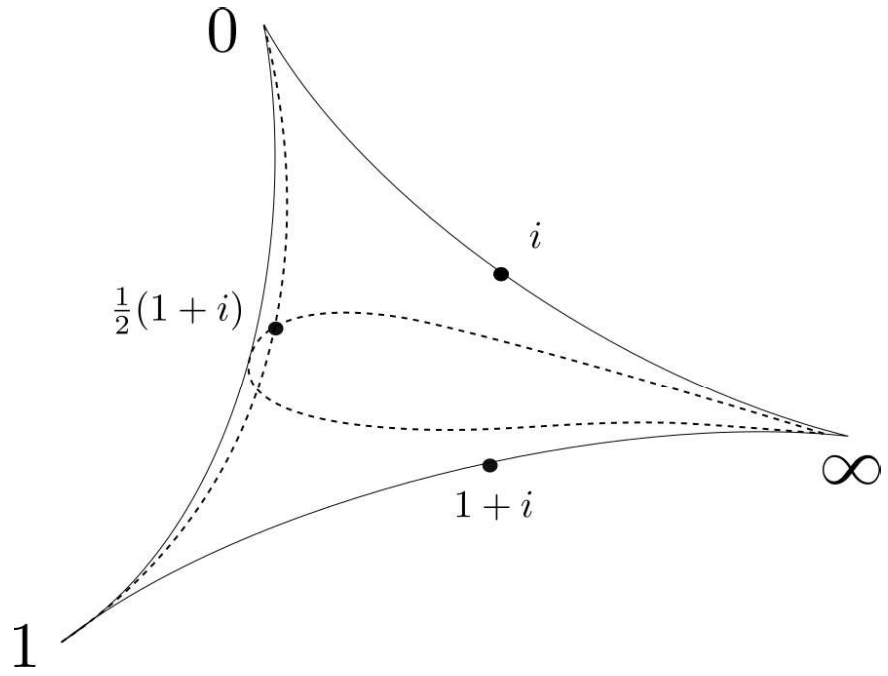

FIGURE 5. Three punctured sphere with cusps and midpoints labelled. The dotted loop is the fixed point set of the automorphism induced by V'.

so the composition is not covered by an element of $\Gamma(2)$ though its square is. One sees from this that U, V induce a group of automorphisms of $\mathbb{H}/\Gamma(2)$ isomorphic to a Klein four group.

Consider the subgroup K^∞ of automorphisms that preserve the puncture ∞. If $g \in K^\infty$

- preserves both 0 and 1 then it is induced by U
- permutes 0 and 1 then it is induced by either V or $U \circ V$

Thus we have proved:

Lemma 3.1. The group of automorphisms that preserves a cusp on the three punctured sphere is a Klein four group.

3.4. **Fixed point sets.** Recall that $\hat{K}^\infty$, the group generated by U, V, it is isomorphic to the dihedral group D_∞. Consider the fixed point sets of the elements

- U fixes the vertical line $\{it, t > 0\}$
- V fixes the vertical line $\{\frac{1}{2} + it, t > 0\}$
- $U \circ V$ is a translation and has no fixed points in $\mathbb{H}$ as such.

From this we may deduce that the automorphisms of $\mathbb{H}/\Gamma(2)$ induced by U and V each fix a pair of lines on the surface. The fixed point set of V projects to a geodesic on $\mathbb{H}/\Gamma(2)$ (depicted as a dotted loop in Figure 5) separating the surface into two pieces which are permuted by the corresponding automorphism, so the fixed point set is exactly this geodesic. For U the fixed point set of the induced automorphism is strictly bigger as it will also fix the images on the surface of $\{1 + it, t \in \mathbb{R}\}$ and the semi circle joining 0 to 1. This is because

$$U(1 + it) = -1 + it = f(1 + it),$$

where $f : z \mapsto z - 2$ is induced by an element of $\Gamma(2)$.

Lemma 3.2. The fixed point set of the automorphism induced by $U \circ V$ is exactly the intersection of the fixed point sets of the automorphisms induced by U and V. This is a single point namely the image of $\frac{1}{2}(1 + i)$ on $\mathbb{H}/\Gamma(2)$

Proof. The standard fundamental domain for the action of $\Gamma(2)$ is the convex hull of $\infty, -1, 0, 1$. This can be decomposed into two ideal triangles (as in Figure 4) with vertices $\infty, -1, 0$ and $0, 1, \infty$ respectively. The map $U \circ V$ takes the first of these onto the second which means that if the induced automorphism has fixed points then they can only arise from points on the semi circle joining 0 to 1. Now

$$U \circ V\left(\frac{1}{2}(-1 + i)\right) = \frac{1}{2}(1 + i) = f\left(\frac{1}{2}(-1 + i)\right),$$

where $f(z) = \frac{z}{2z+1}$ which is clearly induced by an element of $\Gamma(2)$.

□

4. Action on a family of arcs

Let K^0 denote the subgroup of automorphisms that preserves the cusp labelled 0 on $\mathbb{H}/\Gamma(2)$. This group is generated by automorphisms induced by the maps

$$U' : z \mapsto 2 - \bar{z},\ V' : z \mapsto \bar{z}/(\bar{z} - 1)$$

so that their composition is

$$U' \circ V' : z \mapsto z \mapsto (-z + 2)/(z + 1)$$

whose fixed point is $i+1$.

Now K^0 permutes the cusps labelled ∞ and 1 on $\mathbb{H}/\Gamma(2)$. and further will obviously permute the geodesics joining them. If γ is such a geodesic then any lift $\hat{\gamma} \subset \mathbb{H}$ is an arc joining a point in the $\Gamma(2)$ orbit of ∞ to another in the $\Gamma(2)$ orbit of 1 and so γ has a well defined λ length. So for any integer n K^0 permutes the set of geodesics joining the cusps labelled ∞ and 1 on $\mathbb{H}/\Gamma(2)$ of λ-length n^2. This will be our set X.

4.1. **Canonical lifts.** Let n be an integer and N' the set of integers coprime with n. Consider the family of geodesics of $\mathbb{H}$.

$$\{k/pn + it,\, t > 0\},\, k \in N'.$$

The image of this family on the quotient surface $\mathbb{H}/\Gamma(2)$ consists of $2\phi(n)$ geodesics and, since $\Gamma(2)$ preserves parity, these split into two sub families namely:

- those joining the cusps labelled ∞ and 1 that is belonging our set X
- those joining the cusps labelled ∞ and 0.

The first of these sub families consists of projections of the lines

$$\hat{X} := \{k/n + it,\, t > 0\},\, k \in N',\, k \textit{ odd}.$$

Obviously one has:

Lemma 4.1. Let p be a prime then the set X consists of $p-1$ elements.

5. Sketch of Proof of Fermat's Theorem

We discuss the idea of the proof in [13] without going into the details. Throughout this section the integer n is a prime which we denote p. We can deduce Theorem 1.2 from:

Lemma 5.1. Let p be a prime congruent to 1 or 2 modulo 4. Then there is always a geodesic in the family $\hat{X}$ that has as its midpoint a point in the $\mathrm{SL}(2,\mathbb{Z})$ orbit of i.

This is equivalent to saying that, on projecting to the surface $\mathbb{H}/\Gamma(2)$, there is always a geodesic in X which passes through the fixed point of the map induced by $U' \circ V'$.

5.1. **The singular case of Lemma 5.1.** The case $p = 2$ is exceptional and we will deal with it first. From the preceding paragraph there is a single geodesic namely the projection of the line

$$\{1/2 + it,\, t \in \mathbb{R}\}$$

and this contains the point $\frac{1}{2}(1+i)$ Note that one has

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}),\ \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.i = \frac{1}{2}(1+i)$$

so this point is in the $\mathrm{SL}(2,\mathbb{Z})$ orbit of i¿ Then one has as in Lemma 2.1:

$$\mathrm{Im}\,\frac{1}{2}(1+i) = \frac{1}{2} = \mathrm{Im}\,\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.i = \frac{\mathrm{Im}\, i}{1^2+1^2}$$

So, in a rather roundabout way, we obtain 2 as a sum of squares by comparing denominators:

$$2 = 1^2 + 1^2.$$

5.2. **Inversions and fixed points in X.** We finish the proof of Lemma 5.1 by showing that there is a geodesic invariant by the orientation preserving automorphism in K^0, obtaining the required midpoint as the fixed point of the automorphism. Our argument is exactly the same as for Theorem 1.1. More precislely, we show that, for $p > 2$:

(1) the automorphism induced by U' preserves no geodesic in X
(2) the automorphism induced by V' preserves at most two geodesics in X

The first point is rather easy (the automorphism induced by U' fixes three disjoint geodesics joining cusps and permutes the pair of ideal triangles in their complement) but the second requires establishing the analogue of the fact that the equation

$$x^2 = 1$$

has at most two solutions in any field or integral domain for that matter. Explicity the analogous result from [13] is:

Lemma 5.2. Let p be a prime. The automorphism induced by V' preserves two and exactly two geodesics in X.

For the inversions we consider the fixed point sets are arcs joining rationals so a necessary condition for the inversions to be conjugate by an element of $\mathrm{SL}(2,\mathbb{Z})$ is that the fixed point sets have the same λ-length. Observe that, if we consider the fixed point set of V' as an arc, then it has λ-length 4 but for any rational k/p the inversion of $\mathbb{H}$ that swaps the Ford circles based at k/p and ∞ must fix the rationals $(k \pm 1)/p$. The λ-length of the arc joining these points is seen to be strictly greater than 4 from equation (1) unless $k = \pm 1$ so it is not a lift of V'.

5.3. **Eisenstein integers.** The Eisenstein integers $\mathbb{Z}[\omega]$ where ω is an irrational cubic root of unity are the ring of integers of a quadratic extension of $\mathbb{Q}$. If $a+b\omega$ is an Eisenstein integer then its norm is

$$a^2 - ab + b^2.$$

Obviously there is an analogue of Theorem 1.2 in this setting:

Theorem 5.3. Let p be a prime then the equation

$$a^2 - ab + b^2 = p \tag{4}$$

has a solution in integers iff $p = 3$ or $p - 1$ is a multiple of 6.

It is easy to see why this condition is necessary since if one has equation (4) then in $\mathbb{F}_p$

$$\bar{a}^2 - \bar{a}\bar{b} + \bar{b}^2 = 0$$

and so , for $p > 3$, $\bar{a}/\bar{b}$ is a cubic root of -1 ie an element of order 6 in the group of units and it follows that 6 divides the order of this group which is just $p - 1$.

Similarly to Theorem 1.2 this result has an interpretation in terms of bi cuspidal geodesics on $\mathbb{H}/\Gamma(2)$ and fixed points of automorphisms. Begin by observing that the element

$$\begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$$

induces an automorphism ϕ of order 3 on $\mathbb{H}/\Gamma(2)$ which permutes the cusps cyclically. Note that ω projects to one of the fixed points of this automorphism and Theorem 5.3 is equivalent to:

There is some member of the family of geodesics of $\mathbb{H}$

$$\{k/p + it,\, t > 0\}$$

which projects to a bi cuspidal geodesic on $\mathbb{H}/\Gamma(2)$ passing through a fixed points of the automorphism ϕ.

Unfortunately there does not seem to be a straightforward geometric argument giving a proof of this fact.

6. Markoff numbers

We discuss the connection between λ-lengths and Markoff numbers showing that every such number is the sum of two squares without applying Theorem 1.2. We then proceed to show uniqueness for Markoff numbers satisfying certain arithmetic conditions (Theorem 6.4) following Baragar and Button see also [14, 21, 22] for alternative approaches. The content is almost purely expository and, as such, we make no claims of originality. We will assume that the reader has some familiarity with the theory of Fuchsian groups.

Recall, from the introduction, that the lengths of simple closed geodesics on the *modular torus* that is $\mathbb{H}/\Gamma'$ where $\Gamma' < P\mathrm{SL}(2,\mathbb{Z})$ is the commutator subgroup and Markoff numbers are related by a simple formula. If γ is such a geodesic then :

$$X = \frac{2}{3}\cosh\left(\frac{\ell_\gamma}{2}\right), \tag{5}$$

is a Markoff number where ℓ_γ is the length of γ.

We will now develop the correspondence between λ-lenghts of arcs and Markoff numbers.

6.1. **Character Variety.**

It is convenient to change variables and study solutions off

$$X^2 + Y^2 + Z^2 - XYZ = 0. \tag{6}$$

By the work of Fricke the set of solutions in positive real numbers can be identified with a certain slice of the *relative character variety of* $\mathbb{Z} * \mathbb{Z}$. This is the set of representations

$$\rho : \mathbb{Z} * \mathbb{Z} \to SL(2,\mathbb{R})$$

such that the trace of the image of the commutator of the generators is -2 up to conjugation. The key point in Fricke's work is that an (irreducible) representation ρ is determined up to conjugation by the three numbers

$$\begin{aligned} X &= tr\rho(\alpha), \\ Y &= tr\rho(\beta), \\ Z &= tr\rho(\alpha\beta)), \end{aligned}$$

where α, β are generators of $\mathbb{Z} * \mathbb{Z}$. Fricke calculates the trace of the commutator and shows that

$$2 + tr\left(\alpha\beta\alpha^{-1}\beta^{-1}\right) = X^2 + Y^2 + Z^2 - XYZ. \tag{7}$$

The quotient surface $\mathbb{H}/\rho(\mathbb{Z} * \mathbb{Z})$ is invariably a once punctured torus and we identify $\mathbb{Z} * \mathbb{Z}$ with its fundamental group. The $\alpha\beta\alpha^{-1}\beta^{-1}$ is a loop around the puncture and the condition of the trace means that the monodromy around this loop is parabolic.

6.2. λ **lengths.** There is an embedded cusp region H of area 2 on the punctured torus $\mathbb{H}/\rho(\mathbb{Z}*\mathbb{Z})$ (see [12] for a discussion). By replacing ρ by a conjugate representation we may assume $\rho(\mathbb{Z}*\mathbb{Z})$ that

$$\rho(\alpha\beta\alpha^{-1}\beta^{-1}) : z \mapsto z+6,$$

it follows that H lifts to the set $\hat{H} = \{\operatorname{Im} z > 3\}$. Let $\alpha*$ be an*arc* that is a bicuspidal geodesic without self intersetions. There is a lift of $\alpha*$ to $\mathbb{H}$ which is a vertical line which evidently meets $\hat{H}$, we claim that any lift of of $\alpha*$ which meets $\hat{H}$ is a vertical line and not a semi circle. For, if C is a semi circle that meets $\hat{H}$ its diameter is strictly greater than 6 and it follows that C and $C+6$ meet transversely in some point$x$. Such a point gives rise to a self intersection on the quotient surface It follows that, the portion of α^* outside of H is connected, and define and we define λ length to be the exponential of the length of this sub arc.

Lemma 6.1. Let $\alpha*$ be an arc on a once punctured torus and α the unique simple closed geodesic disjoint it. Then the square root of the λ-length of the arc α is equal to $\frac{2}{3}\cosh \ell_\alpha/2$.

It is possible to prove this directly using hyperbolic trigonometry following the same schema as in [12] but here we give a more conceptual proof using the computations from [19].

Given an arc α^* one may extend it to an ideal triangulation off the punctured torus: that is there is a pair of arcs β^*, γ^*, each disjoint from α^* and their complement is a pair of ideal tirangles. Let X denote $2\cosh \ell_\alpha/2$ where α is the unique closed simple geodesic disjoint from α.

$$\begin{aligned} Y &= 2\cosh \ell_\beta/2 \\ Z &= 2\cosh \ell_\gamma/2 \end{aligned}$$

where β resp γ is the unique closed simple geodesic disjoint from β^* resp. γ^*.

In [19] Wolpert divides the Markoff cubic by XYZ to obtain

$$\frac{X}{YZ} + \frac{Y}{XZ} + \frac{Y}{XZ} = 1.$$

The three terms in this relation have a geometric interpretation which we will exploit to compute the λ-length of α^*. Let H denote the cusp region of area 2. A*corner* of an ideal triangle is one of the three components of its intersection with H. Every torus admits an *elliptic involution* which leaves each of the arcs of the ideal triangulation invariant and swaps the triangles. So, in fact, to each triangulation we can associate three numbers namely the areas of the corners of one of the ideal triangles and these coincide with Wolpert's three numbers.

Lifting the ideal triangulation to $\mathbb{H}$ as in Figure 6 one sees that α^* decomposes into two arcs of length $-\log(Y/XZ)$ and $-\log(Z/XY)$ respectively so that its is of length $2\log X$.

So, on any hyperbolic punctured torus, the λ-length of α^* wrt the cusp region of area 2 is the exponential of this, that is:

$$X^2.$$

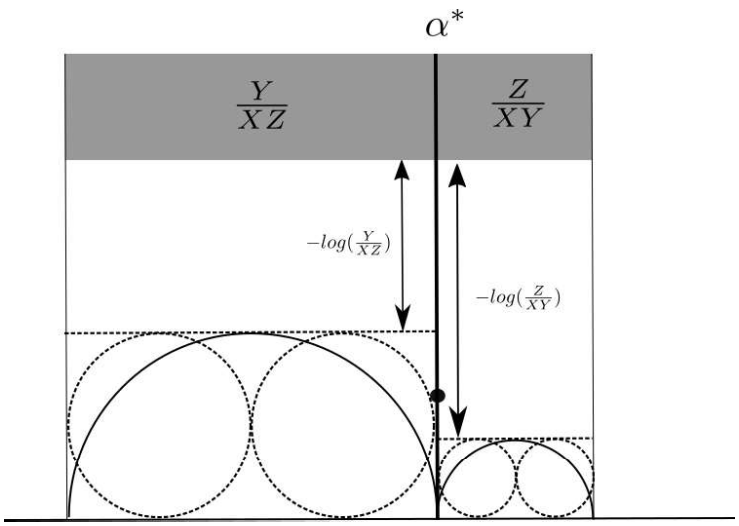

FIGURE 6. Calculating the hyperbolic length of α^* in the upper half plane the λ-length is the exponential of this. The mid point of α^* is marked by a circle and the two corners adjacent to α^* are shaded

Now on the modulaire torus $\mathbb{H}/\Gamma'$ there is an embedded cusp region of area 6 and the λ-length of α^* wrt this cusp region is

$$\frac{X^2}{9}.$$

6.3. **Sum of squares.** In the proof of Lemma 6.1 we used the fact that every torus admits an *elliptic involution* which leaves each of the arcs of the ideal triangulation invariant and swaps the triangles. For the modular torus the involution is covered by $z \mapsto -1/z$ and this means that for any arc α^* every lift contains a point of the $\mathrm{SL}(2,\mathbb{Z})$-orbit of i. In particular, by Lemma 6.1, a lift which is a vertical line ends at a rational which has as denominator a Markoff number and so this Markoff number is a sum of two squares. Conversely, every Markoff number arises as the square of a λ-length of some arc α^* and so must be the sum of two squares. By extending this reasoning slightly one may show:

Theorem 6.2. Frobenius' conjecture is equivalent to: Let m be a Markoff number then exactly one of the vertical lines with endpoint k/m, where $1 \leq k \leq m-1$ is coprime to m, projects to a simple arc on the modular torus.

Proof. The Markoff triples form a binary tree with a preferred vertex corresponding to the fundamental triple $(1,1,1)$ (see Figure 2). Define the multiplicity of a Markoff number to be the number of triples for which it appears as the largest integer. One can easily check that for the so-called singular Markoff numbers 1 and 2 their multiplicity is 3 and, since, group of automorphisms of the tree that fix the fundamental triple of order 6 that the multiplicity of any other Markoff number is at least 6. Thus Frobenius' conjecture can be restated as: multiplicity of any other Markoff number is at most 6.

Using Cohn's correspondence it follows that Frobenius' conjecture is equivalent to: the number of oriented closed simple geodesics on the modular torus of any given length is at most 6. Each (unoriented) closed simple geodesic is disjoint from exactly one arc so that there can be at most three arcs of any given λ-length.

The group of orientation preserving automorphisms of the modular torus is canonically isomorphic to

$$\mathrm{SL}(2,\mathbb{Z})/\Gamma' \simeq \mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/2 \; ZZ \simeq \mathbb{Z}/6\mathbb{Z}.$$

The commutator of the generators of Γ' is $z \mapsto z+6$ and since each automorphism ϕ must leave the cusp invariant it lifts to a map of the form $\hat{\phi} : z \mapsto z+k$, $k = 0, \ldots 5$. Now consider the lift of some arc on the modular torus which, WLOG, is a vertical line. After applying (the lift of) an automorphism $\hat{\phi}$ we may assume it has its end point in $\mathbb{R}$ between 0 and 1. The statement now follows by counting multiplicities as before.

□

6.4. **Uniqueness of Markoff Numbers.** Frobenius' conjecture says that the largest number in a Markoff triple determines the remaining two numbers [1]. Button and Baragar (see chapter 10 of Aigner [1]) used basic algebraic algebraic number to show that certain Markoff numbers satisfied the uniqueness conjecture. Subsequently Aigner extended used this approach showing:

Theorem 6.3 (Aigner)**.** Let m be a Markoff number of the form

$$m = Np^k$$

where p is an odd prime and $N \leq 10^{35}$ is another Markoff number. Then m is unique.

This is a strengthening a result from Button's thesis:

Theorem 6.4 (Baragar, Button, Schmutz)**.** Let m be a Markoff number of the form $m = p^k$ or $m = 2p^k$ then it is unique if p is an odd prime.

We give a short proof of this using the fact that the Gaussian integers is a unique factorisation domain.

Proof. : Suppose that $m = p^k$ is a Markoff number. By the previous paragraph there are coprime integers a, b so that

$$p^k = a^2 + b^2 \Rightarrow a^2 b^{-2} = -1 \in \mathbb{F}_p.$$

It follows that p is either 2 or 1 mod 4 and so by Theorem 1.2 there are coprime positive integers c, d, unique up to permutation, so that

$$p = c^2 + d^2 = (c+id)(c-id).$$

It is well known that the RHS is the unique factorisation of p in the Gaussian integers and it follows that the unique factorisation of m is

$$p^k = (c+id)^k (c-id)^k.$$

A consequence of this is that the pair coprime positive integers a, b such that $p^k = a^2 + b^2$ is unique up to permutation. Explicitly we have:

$$a = \mathrm{Re}\,(c \pm id)^k \tag{8}$$

$$b = \mathrm{Im}\,(c \pm id)^k. \tag{9}$$

Since a, b are unique up to permutation then, by Lemma 2.1, there can only be a single geodesic of the family of vertical lines ending at k/p^k which meets the $\mathrm{SL}(2,\mathbb{Z})$-orbit of i. The result follows immediately from the Paragraph 6.3.

Now suppose that $m = 2p^k$ is a Markoff number. By the above p^k can be written as a sum of squares $a^2 + b^2 = |a + ib|^2$ essentially uniquely. Observe that 2 factors as

$$2 = i(1 + i)^2.$$

Observe that $2p^k$ can also be written as a sum of squares essentially uniquely namely

$$2p^k = |(1 + i)(a + ib)|^2 = (a - b)^2 + (a + b)^2,$$

so that the result follows in this case too.

□

6.5. **Fibonacci numbers, primes, sums of squares.** It is quite difficult to say anything useful about the set of Markoff numbers. Though Theorem 6.4 is quite elegant it is not known if there are infinitely many Markoff numbers which satisfy the hypothesis of the theorem. We discuss this problem in a more restricted context.

Recall that the Fibonacci numbers F_n are defined recursively by

$$F_0 = F_1 = 1$$

and

$$F_{n+2} \quad = \quad F_{n+1} + F_n$$

The odd-indexed Fibonacci numbers F_n, and the of odd-indexed Pell numbers P_n give rise to two families of Markoff triples. It is easy to check that, for any integer n,

$$(F_{2n+1}, F_{2n-1}, 1),$$

is a Markoff triples. So F_{2n+1} is a Markoff number: it is conjectured that there are infinitely many prime Fibonacci numbers but at the time of writing only 51 of them are known to be prime.

The Fibonacci numbers satisfy many identities. Among the most useful of these is the *Cassini identity*

$$F_{n+1}F_{n-1} - F_n^2 = (-1)^n. \tag{10}$$

and *Ocagne's identity*

$$F_{2n} = F_{n+1}^2 - F_{n-1}^2 = F_n(F_{n+1} + F_{n-1}). \tag{11}$$

Cassini's identity is easy to prove by interpreting the LHS as the determinant of a matrix namely:

$$\begin{pmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^n.$$

In fact, setting $A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ and using the factorisation $A^{2n} = A^n A^n$, one has

$$\begin{pmatrix} F_{2n+1} & F_{2n} \\ F_{2n} & F_{2n-1} \end{pmatrix} = \begin{pmatrix} F_{n+1}^2 + F_n^2 & F_n(F_{n+1} + F_{n-1}) \\ * & * \end{pmatrix}$$

so that comparing coefficients in the first rows one obtains the following identities:

$$F_{2n+1} \quad = \quad F_n^2 + F_{n+1}^2 \tag{12}$$

$$F_{2n} \quad = \quad F_n(F_{n+1} + F_{n-1}). \tag{13}$$

The first of these gives an explicit representation, it is tempting to say this is the canonical expression, for an odd indexed Fibonacci number as a sum of squares. In the previous

section the proof of Button's Theorem followed from the fact that $2p^k$ had an essentially unique (ie canonical) decomposition as a sum of squares.

Obviously (12) shows that there is a recursion for integers such that F_{2n+1}. In fact this is true more generally for all Markoff numbers and there appears to be a "canonical" way to write each of them as a sum of squares whether they are prime which we will outline now. Let's look at this recursion from a slightly different perspective. For the odd indexed numbers one has:

$$\det A^{2n} = F_{2n+1}F_{2n-1} - F_{2n}^2 = 1, \tag{14}$$

or

$$F_{2n+1}F_{2n-1} = F_{2n}^2 + 1 = |iF_{2n} + \pm 1|^2. \tag{15}$$

The key observation is that one can take the "square root" of this relation as follows. Our starting point is the relation (12) from which it follows that:

$$\begin{aligned} F_{2n-1} &= |iF_n + F_{n-1}|^2 \\ F_{2n+1} &= |iF_n + F_{n+1}|^2 \end{aligned}$$

so that considering the product of the Gaussian integers on the RHS one obtains

$$\begin{aligned} (iF_n + F_{n-1})(iF_n + F_{n+1}) &= (F_{n+1}F_{n-1} - F_n^2) + iF_n(F_{n+1} + F_{n-1}) \\ &= (F_{n+1}F_{n-1} - F_n^2) + iF_{2n} \\ &= (-1)^n + iF_{2n} \end{aligned}$$

where the second last line follows by Ocagne's identity and the last from Cassini's.

For a general Markoff number we begin by considering the Markoff cubic as a quadratic equation in Z:

$$Z^2 - (3XY)Z + (X^2 + Y^2) = 0.$$

Let $Z^{\pm}$ denote the roots of this equation, one has the Vieta formulas

$$Z^+ + Z^- = 3XY \tag{16}$$

$$Z^+ \times Z^- = X^2 + Y^2 = |X + iY|^2. \tag{17}$$

The Markoff triples $(X, Y, Z^{\pm})$ are adjacent vertices of the Markoff tree as defined above. In fact the first Vieta formula (17) can be used to enumerate all the Markoff triples but it is the second that will concern us here and we will show how to extract it's "square root" as we did for (15) in the preceding paragraph.

We have implemented this algorithm as a computer program and are working on its connections to moduli of ideal triangulations of the once punctured torus.

7. Concluding remarks

We have presented an approch to some classical problems of number theory from a geometric point of view and indicated that there are still open questions.

References

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

[2] Aigner M., Ziegler G.M. *Representing numbers as sums of two squares.* In: Proofs from THE BOOK. Springer, Berlin, Heidelberg. (2010)

[3] A. Baragar, *On the Unicity Conjecture for Markoff Numbers* Canadian Mathematical Bulletin , Volume 39 , Issue 1 , 01 March 1996 , pp. 3 - 9

[4] J. O. Button, *The uniqueness of the prime Markoff numbers*, J. London Math. Soc. (2) 58 (1998), 9–17.
[5] Ilke Canakci, Ralf Schiffler *Snake graphs and continued fractions* European Journal of Combinatorics Volume 86, May 2020, 103081
[6] Elsholtz C.A *Combinatorial Approach to Sums of Two Squares and Related Problems.* In: Chudnovsky D., Chudnovsky G. (eds) Additive Number Theory. Springer, New York, NY. (2010)
[7] Lester R Ford, *Automorphic Functions*
[8] Andrew Haas. *Diophantine approximation on hyperbolic Riemann surfaces.* Acta Math. 156 33 - 82, 1986.
[9] *The geometry of Markoff forms.* Number Theory, New York pp 232-244 Lecture notes in math 1240, 1988
[10] Heath-Brown, Roger. *Fermat's two squares theorem.* Invariant (1984)
[11] Yi Huang *Moduli Spaces of Surfaces* Ph.D. Thesis, The University of Melbourne (2014)
[12] G. McShane, *Simple geodesics and a series constant over Teichmuller space* Invent. Math. (1998)
[13] G. McShane, V. Sergesciu *Geometry of Fermat's sum of squares* in preparation.
[14] M.L. Lang, S.P Tan, *A simple proof of the Markoff conjecture for prime powers* Geometriae Dedicata volume 129, pages15–22 (2007)
[15] R. C. Penner, *The decorated Teichmueller space of punctured surfaces*, Communications in Mathematical Physics 113 (1987), 299–339.
[16] Northshield, Sam. *A Short Proof of Fermat's Two-square Theorem.* The American Mathematical Monthly. 127. 638-638. (2020).
[17] J-P. Serre, *A Course in Arithmetic*, Graduate Texts in Mathematics, Springer-Verlag New York 1973
[18] Series, C. (1985), *The Modular Surface and Continued Fractions. Journal of the London* Mathematical Society, s2-31: 69-80.
[19] Scott Wolpert, *On the Kahler form of the moduli space of once-punctured tori*, Comment. Math. Helv. 58(1983)246-256
[20] D. Zagier, *A one-sentence proof that every prime p = 1 (mod 4) is a sum of two squares*, American Mathematical Monthly, 97 (2): 144
[21] Y. Zhang, *An elementary proof of uniqueness of Markoff numbers* preprint, arXiv:math.NT/0606283
[22] Y. Zhang, *Congruence and uniqueness of certain Markoff numbers* Acta Arithmetica, Volume: 128, Issue: 3, page 295-301

Institut Fourier 100 rue des maths, BP 74, 38402 St Martin d'Hères cedex, France
Email address: mcshane at univ-grenoble-alpes.fr