GEOMETRY OF FERMAT'S SUM OF SQUARES

GREG MCSHANE AND VLAD SERGESCIU

Abstract. We prove Fermat's sum of two squares theorem using well known calculations from hyperbolic geometry and considerations of automorphisms of the three punctured sphere.

1. Introduction

Consider the following pair of well known theorems from 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.

These results are intimately linked and often one deduces the second as a corollary of the first, for example, by using unique factorisation in the Gaussian integers. 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 in Zagier's remarkable proof [15] (see also [8, 12, 2, 6] for closely related constructions and discussion) both results 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

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

1.1. **Organisation, Remarks.** In Section 2 we recall the statement of Burnsides Lemma and apply it to a Klein four group generated by involutions of $\mathbb{F}_p^*$ yielding a proof of Theorem 1.1. In Section 3 we introduce $\Gamma(2)$ and the associated Riemann surface $\mathbb{H}/\Gamma(2)$. In Section 4, for each prime p we study how the automorphisms of $\mathbb{H}/\Gamma(2)$ act on a family of geodesic on this surface obtained in a natural way from the rationals k/p. In particular we show (Lemma 5.3) that if p is congruent to 1 modulo 4 there is always an orientation preserving involution that leaves one of our 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, presented by Uspensky and Heaslet in the 70s. Zagier's proof in [15] is a clever reformulation of ths argument.

Like our proof it is based on 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

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

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

$$\begin{array}{rcl} z & \mapsto & -\bar{z} \\ z & \mapsto & 1/\bar{z} \end{array} .$$

1.1.2. *Burnside and signatures.* The astute reader will surely realise that Burnside 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 [8].

1.1.3. *Farey tessalation.* The argument in Lemma 5.3 is inspired by the definition of the *Farey tessalation*.

1.1.4. *Lambda lengths.* The idea of associating a length to a geodesic joining cusps (paragraph 4.1) appears in Penner's work on moduli [11]. He defined the λ-length of simple bicuspidal geodesic on a punctured surface to be the length of the portion outside of some fixed system of cusp regions.

By using calculations in Wolpert [14] 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. Then, using the fact that each arc is invariant under the elliptic involution one can show, using Lemma 4.3, that every Markoff number is the sum of two squares. In fact this was the observation that was the starting point for this paper and we give a short exposition of it in the appendix.

1.1.5. *Bezouts Theorem.* We are implicitly using Bezout's Theorem (and in particular in the proof of Lemma 4.4) when we assert that

- $\mathrm{SL}(2,\mathbb{Z})$ is transitive on $\mathbb{Q}\cup\infty$ (which is equivalent to Bezout's Theorem.)
- $\Gamma(2)$ has exactly three orbits on $\mathbb{Q}\cup\infty$.

In fact Lemma 4.4 can be proved without using our notion of length for a bicuspidal geodesics but instead by studying the action of the lifts U', V' of the generators of our group K^0 and applying Bezout's Theorem.

1.1.6. *References.* Almost all of the material in Sections 3 and 4 can be found in Serre's book [13] and the reader should not need any other references to understand this paper if they are already familiar with the Burnside Lemma.

1.2. **Thanks.** The first author thanks Louis Funar and the second author 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.

2. Burnside Lemma

We give a proof of Theorem 1.1 using the Burnside Lemma. Recall that if G is a group acting on a finite set X then the Burnside Lemma says

$$|G||X/G| = \sum_g |X^g| \tag{1}$$

where, as usual, X^g denotes the set of fixed points of the element g and X/G the orbit space.

Let $p \neq 2$, $X = \mathbb{F}_p^*$ and G be the group generated by the two involutions

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

The group G has exactly four elements namely:

- the trivial element which has $p-1$ fixed points
- $x \mapsto -x$ which has no fixed points
- $x \mapsto 1/x$ has exactly two fixed points namely 1 and -1.
- $g : x \mapsto -1/x$ is the remaining element and the theorem is equivalent to the existence of a fixed point for it.

Note that since $\mathbb{F}_p$ is a field $|X^g| = \sharp\{x^2 = -1,\, x \in \mathbb{F}_p^*\}$ is either 0 or 2. Now for our choice of X and G equation (1) yields

$$4|X/G| = (p-1) + 2 + |X^g|. \tag{2}$$

The LHS is always divisible by 4 so the RHS is too and it follows from this that

$$|X^g|. = \begin{cases} 0 & (p-1) = 2 \mod 4 \\ 2 & (p-1) = 0 \mod 4 \end{cases}$$

This proves Theorem 1.1.

Note. As was noted in the introduction one can obtain the same conclusion by calculating the signature of $x \mapsto -1/x$ using the fact that it is the composition of $x \mapsto -x$ and $x \mapsto 1/x$.

3. Automorphisms of 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 2). 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 1. 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 2).

3.1. **Automorphism groups.** 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 $\mathbb{H}/\Gamma(2)$.

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. **Another Klein four group.** 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

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

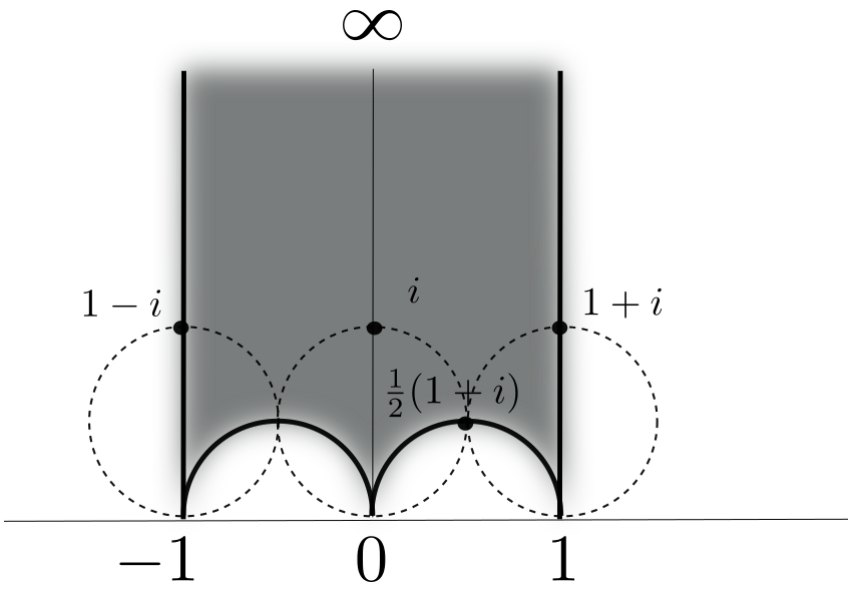

FIGURE 1. Standard fundamental domain for $\Gamma(2)$ and its decomposition into ideal triangles.

and we note that

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

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.*

Strictly speaking, for the proof of Theorem 1.2 this lemma is irrelevant as all we require is that the group contains a suitable 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

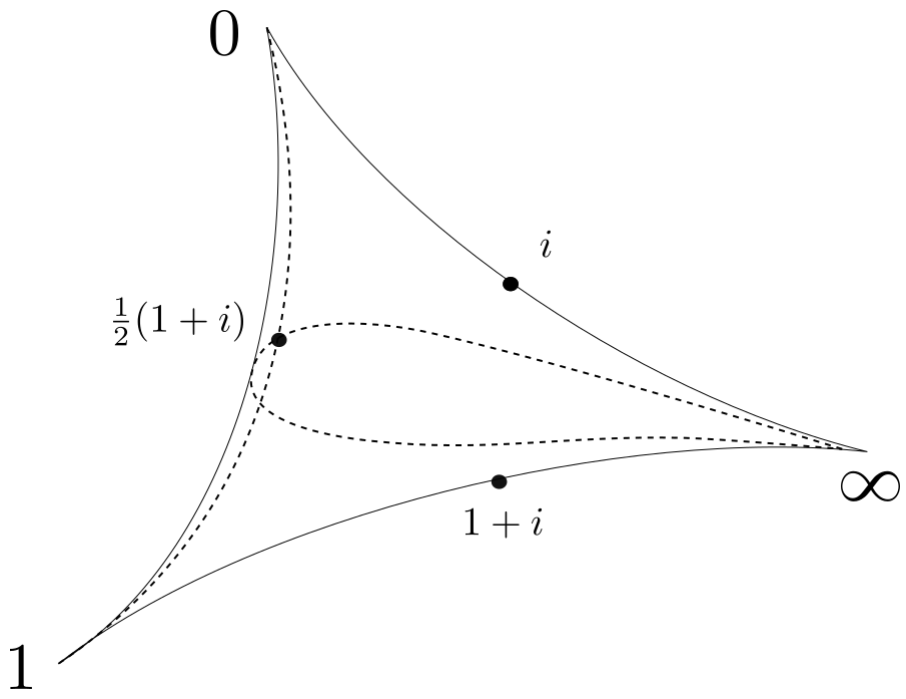

FIGURE 2. Three punctured sphere with cusps and midpoints labelled. The dotted loop is the fixed point set of the automorphism induced by V.

- U fixes the vertical line $\{it, t \in \mathbb{R}\}$
- V fixes the vertical line $\{\frac{1}{2} + it, t \in \mathbb{R}\}$
- $U \circ V$ 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 2) 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 automorphism induced by $U \circ V$ consists of 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 1) 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 geodesics

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

$$\{k/pn + it,\ t \in \mathbb{R}\},\ k \in N'.$$

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.
- those joining the cusps labelled ∞ and 0.

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

$$\{k/n + it,\ t \in \mathbb{R}\},\ k \in N',\ k\ odd,$$

and it is on this set that we study the action of a suitable Klein four group. Let

$$\hat{\mathcal{G}_n} := \{k/n + it,\ t \in \mathbb{R}\},\ k \in N',\ k\ odd,$$

and $\mathcal{G}_n$ denote the image of this family on the surface.

4.1. **Ford circles, lengths, midpoints.** We denote by F the set $\{z, \operatorname{Im} z > 1\}$ 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 circles, the so-called *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.

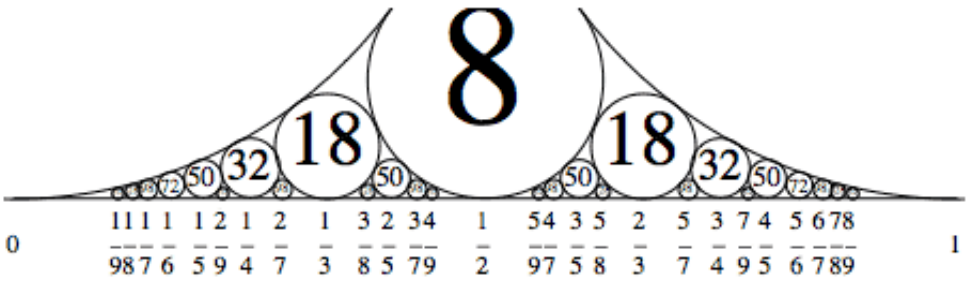

FIGURE 3. Ford circles with tangent points and curvatures. Recall that the curvature of a euclidean circle is the reciprocal of its radius.

The following is well known and is easily checked:

Lemma 4.1. *The Ford circle tangent to the real line at m/n has Euclidean diameter $1/n^2$.*

We define the *length* of the vertical line $\{k/p + it,\, t \in \mathbb{R}\}$ to be the length of the sub arc joining F to the Ford circle tangent at k/p. Further we define its *mid point* to be the midpoint of this sub arc. We remark that if the projection of the line to $\mathbb{H}/\Gamma(2)$ is invariant by an automorphism then the midpoint is necessarily a fixed point of the automorphism.

The following is a restatement of Lemma 4.1 in terms of these notions:

Lemma 4.2. . *Let m/n be a rational. Then the hyperbolic geodesic $\{m/n + it,\, t \in \mathbb{R}\}$*

- *has length $2 \log n$,*
- *has its midpoint at $\frac{1}{n}(m + i)$.*

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

Lemma 4.3. *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 integers is equal to the number the integers $0 \leq k < n - 1$ coprime to n such that the line

$$\{k/n + it,\, t \in \mathbb{R}\}$$

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 4.2 one has:

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

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 $0 \leq \operatorname{Re} w < 1$. 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.

□

4.1.1. *Cusp regions.* The image of a Ford circle on $\mathbb{H}/\Gamma(2)$ is a *cusp region* around one of the three cusps $0, 1, \infty$. 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 $\mathbb{H}$.

4.2. **The Group action.** 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 and further:

Lemma 4.4. *The group* K^0 *permutes the geodesics of* $\mathcal{G}_n$.

Proof. Let $g \in K^0$ and $\gamma \in \mathcal{G}_n$ a geodesic. Choose a lift $\tilde{g} : \mathbb{H} \mapsto \mathbb{H}$ of g. By Lemma 4.2 the length of any lift $\hat{\gamma} \subset \mathbb{H}$ is $2 \log n$ and, since $\tilde{g}$ normalises $\Gamma(2)$, it preserves the Ford circles so that $\tilde{g}(\hat{\gamma})$ has the same length. Since $g(\gamma)$ joins the cusps labelled ∞ and 1 there is a lift of this geodesic which is a vertical line and the other endpoint is a rational m/n. The length of the lift is again $2 \log n$ so the diameter of this Ford circle is $1/n^2$ and by Lemma 4.1 its center is a multiple of $1/n$. □

5. Proof of Fermat's Theorem

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 $\mathcal{G}_p$ 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 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 $sl2$ orbit of $i¿$ Then one has as in Lemma 4.3:

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

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 geodesics.** We will 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 $\mathcal{G}_p$
(2) the automorphism induced by V' preserves at most two geodesics in $\mathcal{G}_p$

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. Let us start by saying which geodesics are preserved by the automorphism: they are unsurprisingly the pair with endpoints $\pm 1/p$. To see

this consider the map

$$z \mapsto \frac{\bar{z}}{p\bar{z}-1}, \tag{3}$$

and observe that on setting $p = 1$ the resulting map coincides with V'. This map fixes 0 and $2/p$ and permutes $1/p$ and ∞ so that it maps the geodesic of $\hat{\mathcal{G}}_n$ with endpoint $1/p$ to itself. Moreover the map is an inversion in the semi circle joining 0 and $2/p$ and is conjugate to V' by an element of $\Gamma(2)$. The following is an elementary exercise in (hyperbolic) geometry:

Lemma 5.2. *Let ϕ_1 (resp. ϕ_2) be an of inversion of $\mathbb{H}$ with fixed point set $L_1 \subset \partial\mathbb{H}$ (resp L_2). Then ϕ_1 and ϕ_2 are conjugate by an isometry f (i.e. $\phi_1 = f \circ \phi_2 \circ f^{-1}$) if and only if $f(L_1) = L_2$.*

So it suffices to find a map that takes one fixed point set to the other. To do this it proves convenient to represents the fixed point set, which is a geodesic in $\mathbb{H}$, by its endpoints, which are a pair of rational numbers, and encode this pair as a matrix whose entries are the numerators and denominators of the fractions.. Concretely, to an ordered pair of rationals $(m/n, m'/n')$, we associate the following matrix:

$$\begin{pmatrix} m & m' \\ n & n' \end{pmatrix}.$$

Determining the conjugation in Lemma 5.2 is reduced to solving a matrix equation. For example, for the pair $0/1, 2/1$ and $0/1, 2/p$ one has the matrix equation:

$$\begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -(p-1)/2 & 1 \end{pmatrix} \begin{pmatrix} 0 & 2 \\ 1 & p \end{pmatrix}. \tag{4}$$

The first factor of the LHS is an element of $\Gamma(2)$ iff $(p-1)/2$ is even and from it, using Lemma 5.2, we can obtain an isometry of $\mathbb{H}$ conjugating the inversion (3) to V'.

Conjugating the map defined in (3) above by $z \mapsto z + 1$ one obtains an inversion which permutes ∞ and $1 + 1/p$. It follows that geodesic in $\mathcal{G}_n$ with endpoint $1 + 1/p$ projects to a second geodesic on $\mathbb{H}/\Gamma(2)$ preserved by the automorphism induced by V'.

5.3. **Exactly two fixed geodesics.** Having established the existence of suitable geodesics in the preceding paragraph it suffices to show that no other geodesic is preserved.

Lemma 5.3. *Let p be a prime. The automorphism induced by V' preserves two and exactly two geodesics in $\mathcal{G}_p$.*

Proof. We give a proof for p be a prime congruent to 1 modulo 4 the proof of the other case is similar.

Let $1 < k < p - 1$ be an integer. It suffices to show that the inversion in the semi circle with endpoints $(k-1)/p$ and $(k+1)/p$

(i.e. the one that permutes F and the Ford circle tangent at k/p) is not conjugate to V' via a hyperbolic isometry induced by an element of $\mathrm{SL}(2,\mathbb{Z})$. Consider the matrix equation we must solve to find the required element of $\mathrm{SL}(2,\mathbb{Z})$:

$$\begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} k+1 & k-1 \\ p & p \end{pmatrix}. \tag{5}$$

Suppose that a solution exists. Taking determinants one obtains a contradiction immediately:

$$2 = 1 \times 2p \Rightarrow p = 1.$$

□

5.3.1. *Composite integers.* It is well known that the set of integers n which can be written as a sum of squares $c^2 + d^2$ with c, d co prime is closed under multiplication. For example one has

$$50 = 5^2 \times 2 = |(1+2i)^2(1+i)|^2 = |-7+i|^2 = 7^2 + 1^2$$

and

$$65 = 5 \times 13 = |(1+2i)(2+3i)|^2 = |-4+7i|^2 = 7^2 + 4^2.$$

It seems difficult to prove this directly, that is without decomposing the integer into prime factors as above but by considering the set X of bi cuspidal geodesics of length $2\log 65$. The Burnside Lemma for the group of automorphisms we consider above yields:

$$4|X/G| = \phi(65) + |\{x, x^2 = 1\}| + |\{x, x^2 = -1\}| = 48 + 4 + |\{x, x^2 = -1\}|.$$

It is easy to check that the kernel of the morphism

$$x \mapsto x^2,\ (\mathbb{Z}/65)^* \to (\mathbb{Z}/65)^*$$

acts freely on the orbit space X/G and from this one can deduce that $|X/G|$ is even. It is an immediate consequence that

$$4 + |\{x, x^2 = -1\}| = 0 \mod 8$$

so $x^2 = -1$ has solutions in $(\mathbb{Z}/65)^*$ (these are precisely $8, 18, 47, 57$.) In fact 65 can be written as a sum of two squares in two essentially different ways towit

$$65 = 7^2 + 4^2 = 8^2 + 1.$$

It may be possible that this kind of argument, that is showing $|X/G|$ is divisible by some power of 2, can be adapted to the geometric setting to show this.

6. Appendix: 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 6.5) following Baragar and Button see also [10, 16, 17] for alternative approaches. The content of this appendix is 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.

Theorem 6.1. *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.*

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

$$X^2 + Y^2 + Z^2 - 3XYZ = 0. \tag{6}$$

A *Markoff number* is an integer in a Markoff triple. H. Cohn showed that Markoff numbers are related to 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. More precisely, if γ is such a geodesic then :

$$X = \frac{2}{3} \cosh\left(\frac{\ell_\gamma}{2}\right), \tag{7}$$

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

6.2. **Character Variety.** It is convenient to change variables and study solutions off

$$X^2 + Y^2 + Z^2 - XYZ = 0. \tag{8}$$

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

$$2 + tr\,(\alpha\beta\alpha^{-1}\beta^{-1}) = X^2 + Y^2 + Z^2 - XYZ. \tag{9}$$

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.

6.3. λ **lengths.** There is an embedded cusp region H of area 2 on the punctured torus $\mathbb{H}/\rho(\mathbb{Z} * \mathbb{Z})$ (see [9] 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 $\hat{H} = \{\mathrm{Im}\, z > 3\}$. 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 6.2. *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 [9] but here we give a more conceptual proof using the computations from [14].

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 [14] Wolpert divides the Markoff cubic by XYZ to obtain

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

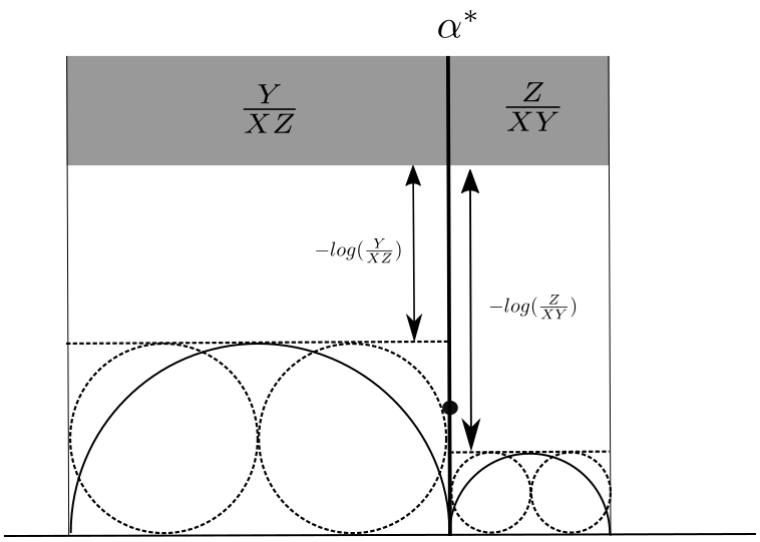

FIGURE 4. Calculating the hyperbolic length of α^* in the upper half plane the λ-length is the exponential of this. The mid point of α^* is marked by a circle and the two corners adjacent to α^* are shaded

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 4 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:

$$X^2.$$

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

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

6.4. **Sum of squares.** In the proof of Lemma 6.2 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 6.2, 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 6.3. *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 an arc on the modular torus.*

Proof. The Markoff triples form a binary tree with a preferred vertex corresponding to the fundamental triple $(1, 1, 1)$. 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.

□

6.5. **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 6.4 (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 6.5 (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:

$$a = \mathrm{Re}\,(c \pm id)^k \tag{10}$$

$$b = \mathrm{Im}\,(c \pm id)^k. \tag{11}$$

Since a, b are unique up to permutation then, by Lemma 4.3, 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 6.4.

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.

□

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] Heath-Brown, Roger. *Fermat's two squares theorem.* Invariant (1984)

[9] G. McShane, *Simple geodesics and a series constant over Teichmuller space* Invent. Math. (1998)

[10] M.L. Lang, S.P Tan, *A simple proof of the Markoff conjecture for prime powers* Geometriae Dedicata volume 129, pages15–22 (2007)

[11] R. C. Penner, *The decorated Teichmueller space of punctured surfaces*, Communications in Mathematical Physics 113 (1987), 299–339.

[12] Northshield, Sam. *A Short Proof of Fermat's Two-square Theorem.* The American Mathematical Monthly. 127. 638-638. (2020).

[13] J-P. Serre, *A Course in Arithmetic*, Graduate Texts in Mathematics, Springer-Verlag New York 1973

[14] Scott Wolpert, *On the Kahler form of the moduli space of once-punctured tori*, Comment. Math. Helv. 58(1983)246-256

[15] D. Zagier, *A one-sentence proof that every prime $p = 1 \pmod 4$ is a sum of two squares*, American Mathematical Monthly, 97 (2): 144

[16] Y. Zhang, *An elementary proof of uniqueness of Markoff numbers* preprint, arXiv:math.NT/0606283

[17] 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