GEOMETRY OF FERMAT'S SUM OF SQUARES

GREG MCSHANE AND VLAD SERGIESCU

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 [11] (see also [5, 7, 1, 2] 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 every odd prime can be written in an essentially unique way as the difference of two squares $a^2 - b^2$.

1.1. **Organisation and Remarks.** In Section 2 we recall the statement of Burnside's 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 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 [5] 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 [11] is a clever reformulation of ths argument.

Like our proof it is based on an 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 generated by

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

1.1.2. *Burnside's lemma and signatures.* The astute reader will surely realise that Burnside's lemma 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 [5].

1.1.3. *Farey tessellation, λ-lengths.* Much of our inspiration for this approach comes from endless contemplation of the *Farey tessellation*.

The idea of associating a length to a geodesic joining cusps (paragraph 3.1) appears in Penner's work on moduli [6], see also [9] for a more recent account. He defined the λ-length of a simple bicuspidal geodesic on a punctured surface to be the exponential of half the length of the portion outside of some fixed system of cusp regions. Though not strictly necessary, we will phrase parts of proof in terms of *ambda*-lengths, as this is the origin of the intuition behind our approach. Lemma 3.2 shows that in the context we consider, a λ-length is always the determinant of an integer matrix.

In [6] Penner presents an approach to the Markoff equation using λ-lengths. Alternatively, using the existing calculations in Wolpert [10] 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 3.3, that every Markoff number is the sum of two squares. In fact this was the observation that was the starting point for this paper.

1.1.4. *Bézouts Lemma.* We are implicitly using Bézout's Lemma which states that for any pair of integers a and b, if d is the greatest common divisor of a and b, then there exist integers x and y such that:

$$ax + by = d$$

and in particular if a, b are coprime then

$$ax + by = 1.$$

For example in the proof of Lemma 5.1 when we assert that

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

we are relying on Bézout's Lemma. In fact Lemma 5.1 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 G_4 and applying Bézout's Theorem.

1.1.5. *References.* Almost all of the material in Sections 3 and 4 can be found in Serre's book [8] and the reader should not need any other references to understand this paper if they are already familiar with Burnside's 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's Lemma

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

Since $\mathbb{F}_p$ is a an integral domain the equation $x^2 = -1$ has at most two solutions. If there is a solution α then $-\alpha$ is also a solution and these are distinct unless $p = 2$ which we have excluded above so $|X^g| = \sharp\{x^2 = -1,\, x \in \mathbb{F}_p^*\}$ is either 0 or 2. Now for our choice of X and G equation (2) yields

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

The LHS is always divisible by 4 so the RHS is too and hence

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

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. 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:

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

The key lemma that relates this $\mathrm{SL}(2,\mathbb{Z})$ action to sums of squares is Lemma 3.3.

3.1. **Ford circles, lengths, midpoints.** Lemma 3.3 illustrates the connexion between sums of squares, the orbit $\mathrm{SL}(2,\mathbb{Z}) \cdot i$ and Poincaré 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 Poincaré 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 $\{z, \operatorname{Im} z > 1\}$ 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 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 Euclidean radius tangent to the extended real line at $\infty = 1/0$.

FIGURE 1. Ford circles with tangent points and curvatures. Recall that the curvature of a euclidean circle is the reciprocal of the square of its radius.

The following is well known and is easily checked:

Lemma 3.1. .

(1) *The Ford circle tangent to the real line at m/n has Euclidean diameter $1/n^2$.*
(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 Poincaré metric on $\mathbb{H}$, of the portion of this arc outside of the Ford circles tangent at $a/c, b/d$.

Following Penner [6], see also [9] for a more recent account, we define the λ-length of the arc to be the exponential of half of this length and define its *midpoint* to be the midpoint of this sub arc.

Lemma 3.2. *Let $a/c, b/d$ be a pair of distinct extended rationals. Then the λ-length of the arc joining them is the absolute value of the determinant of the matrix*

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

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

Proof. By transitivity of the $\mathrm{SL}(2,\mathbb{Z})$ action on the extended rationals we may suppose $a/b = 1/0 = \infty$. The arc joining $a/c, b/d$ is a vertical line and the determinant of the matrix in the statement is just d. By Lemma 3.1 the Ford circle tangent at b/d has diameter $1/d^2$ so the portion of the arc outside F (the Ford circle tangent at ∞) and this circle

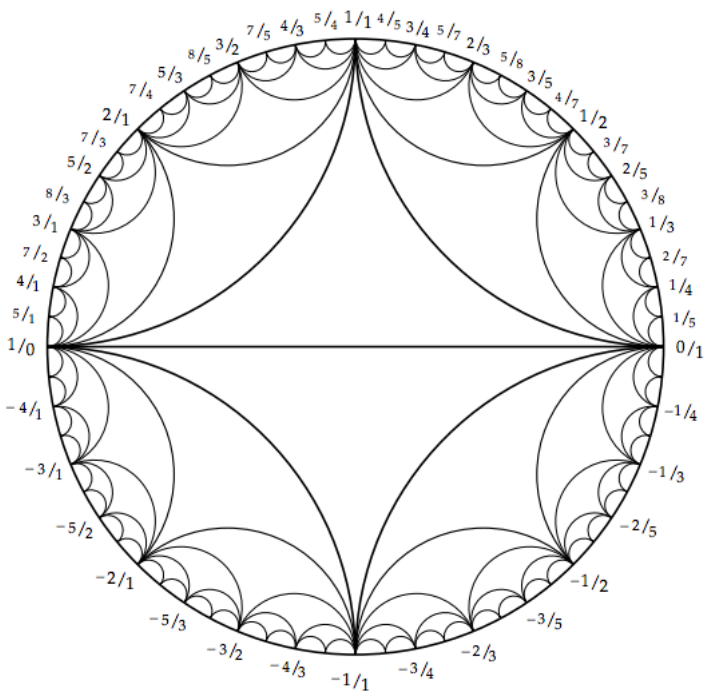

FIGURE 2. Farey diagram.

is a segment between z_1 with imaginary part 1 and z_2 with imaginary part $1/d^2$. A standard calculation using the Poincaré metric $\frac{dz}{\mathrm{Im} z}$ shows that the length of this segment is indeed $\log(d^2)$. The second part of the statement follows from a similar calculation. □.

The arcs of λ-length 1 are the edges in the so-called *Farey diagram* (see Figure 2). Recall that the Farey diagram is the graph whose vertices are the rationals and where two vertices are joined by an edge if the determinant of the matrix

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

is ± 1.

3.2. **Sums of squares.**

Lemma 3.3. *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 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 $F = \{z, \operatorname{Im} z > 1\}$ 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 3.2 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.

□

3.3. **Relation with square roots of -1.** As we saw in the proof of Theorem 1.1 under the hypothesis of that p is a prime and $p-1$ is a multiple of 4 there is $\bar{m} \in \mathbb{F}_p$ such that $\bar{m}^2 = -1$. In fact the real part of $\frac{ai+b}{ci+d}$ is related to the square this $\bar{m}$ in a simple way which, although we will not need it later, we explain now. Begin by writing

$$\frac{ai+b}{ci+d} = \frac{m}{p} + \frac{i}{p}$$

then

$$\frac{m^2+1}{p^2} = \left(\frac{ai+b}{ci+d}\right)\overline{\left(\frac{ai+b}{ci+d}\right)} = \frac{a^2+b^2}{c^2+d^2}$$

now since $p = c^2 + d^2$ one has:

$$m^2 + 1 = p(a^2+b^2). \tag{3}$$

Thus m determines a square root of -1 in $\mathbb{F}_p$.

4. The three punctured sphere

In this section we are concerned with the geometry of the surface $\mathbb{H}/\Gamma(2)$ associated to $\Gamma(2)$, the principal level 2 congruence subgroup of $\mathrm{SL}(2,\mathbb{Z})$ that is the kernel of the canonical (reduction) homomorphism $\mathrm{SL}(2,\mathbb{Z}) \to \mathrm{SL}(2,\mathbb{Z})/2\mathbb{Z}$. 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:

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

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 4). 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 3. 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 4).

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

4.1. **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. $\Gamma(2)$. Since $\Gamma(2)$ is a normal subgroup of $\mathrm{SL}(2, \mathbb{Z})$, the quotient group

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

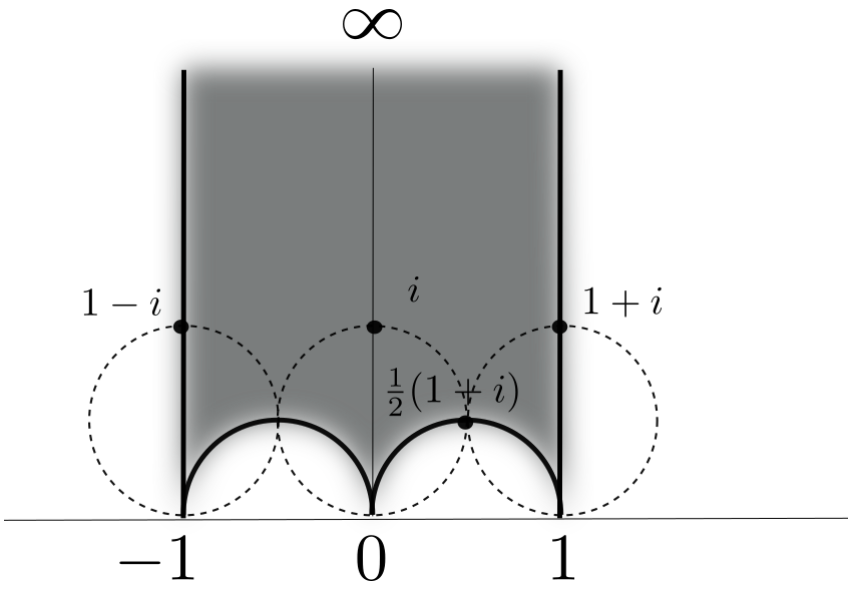

FIGURE 3. Standard fundamental domain for $\Gamma(2)$ and its decomposition into ideal triangles.

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 $\mathbb{H}/\Gamma(2)$.

4.2. **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:

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

generate a group which we denote G_4. Note that the composition of U and V is none other than the involution

$$z \mapsto -1/z,$$

so we have a Klein 4-group as in Section 2. Both U and V normalise $\Gamma(2)$ so induce automorphisms of $\mathbb{H}/\Gamma(2)$.

Lemma 4.1. *The group G_4 descends to a group of automorphisms on the three punctured sphere which preserve the cusp labelled* 1 *(see Figure 4).*

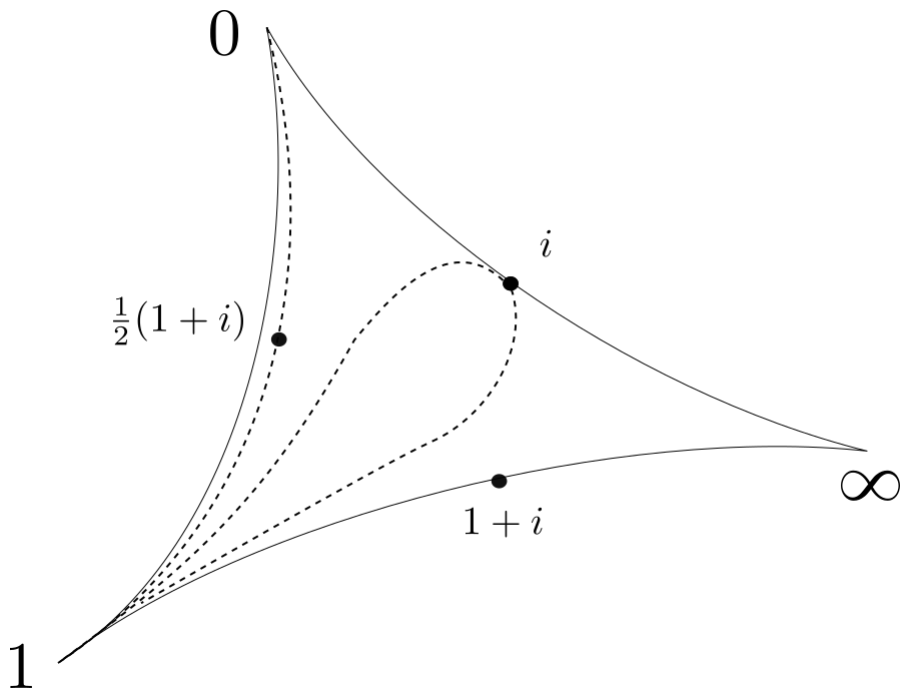

FIGURE 4. A depiction of the three punctured sphere with cusps and midpoints labelled. The dotted loop is the fixed point set of the automorphism induced by V'.

Proof. One checks that each of the generators U and V preserve the standard fundamental domain and further they preserve the set $\{1, -1\}$. Note that this pair of points belong to the same $\Gamma(2)$ orbit and so project to the same cusp on $\mathbb{H}/\Gamma(2)$. It follows that the generators preserve this cusp ie the cusp labelled 1.

□

4.3. **Fixed point sets.** The group generated by U, V which we will denoted G_4 in paragraph 4.2. Consider the fixed point sets of the elements:

- U fixes the vertical line $\{it, t > 0\}$;
- V fixes the semi circle joining -1 and 1;
- the composition $U \circ V(z)$ is just $z \mapsto -1/z$ and so fixes i.

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 4) 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 these arcs are sides of the standard fundamental domain for $\Gamma(2)$:

- $U(1 + it) = -1 + it = f(1 + it)$, where $z \mapsto z - 2$ is the side-pairing for these sides.
- U exchanges the semi circle joining 0 to 1 and the semi circle joining 0 to -1 and these sides are paired by $z \mapsto z/(2z + 1)$

This proves the first part of:

Lemma 4.2. *Any arc of the Farey diagram projects to a component of the fixed point set of the automorphism of* $\mathbb{H}/\Gamma(2)$ *induced by* U.

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* i *on* $\mathbb{H}/\Gamma(2)$

Proof. The first part of the statement follows from the discussion above: the automorphism of U exchanges the pair of ideal triangles in the standard decomposition of $\mathbb{H}/\Gamma(2)$ leaving their edges fixed.

The second part of lemma follows from the observation that $U \circ V$ leaves the standard fundamental domain for $\Gamma(2)$ invariant swapping the two ideal triangles in the complement of the the fixed point set of U. One sees from this that any fixed point of $U \circ V$ is contained in the fixed point set of V. Further, $U \circ V$ fixes the cusp labeled 1 but exchanges the cusps labeled 0 and 1. It follow that any fixed point of $U \circ V$ is contained in the projection of the vertical line $\{it, t > 0\}$. This line contains i and the lemma is proven.

□

5. Action on a family of arcs

Now G_4 permutes the cusps labelled ∞ and 0 on $\mathbb{H}/\Gamma(2)$ and 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 0 and so γ has a well defined *ambda*-length. For any integer n G_4 permutes the set of geodesics joining the cusps labelled ∞ and 0 on $\mathbb{H}/\Gamma(2)$ of λ-length n^2. This will be our set X.

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

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

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

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

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

$$\hat{X} := \{2k/n + it,\, t > 0\},\, k \in N'$$

Lemma 5.1. *Let p be a prime then the set X consists of $p-1$ elements.*

Proof. The $z \mapsto z + 2$ is a generator of the subgroup of $\Gamma(2)$ that stabilises ∞ so to have a complete set of representatives of $\hat{X}$ we need only consider $2k/n$ with $0 < k < n$. If n is prime then for any such k $2k$ and n are coprime so $\hat{X}$ contains exactly $n - 1$ elements.

□

6. Proof of Fermat's Theorem

Throughout this section the integer n is a prime which we denote p. Theorem 1.2 follows from Lemma 3.3 and the following result:

Lemma 6.1. *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 $U \circ V$. For $p = 2$ this can be done explicitly and for the general case by the argument using Burnside's lemma in Section 2 it suffices to show that:

- U fixes no element of X
- V fixes at most two.

6.1. **The singular case of Lemma 6.1.** 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

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

so this point is in the $\mathrm{SL}(2, \mathbb{Z})$ orbit of $i¿$ Then one has as in Lemma 3.3:

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

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

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

6.2. **Inversions and fixed points in X.** We will finish the proof of Lemma 6.1 by showing that there is a geodesic invariant by the orientation preserving automorphism in G_4, obtaining the required midpoint as the fixed point of the automorphism induced by $U \circ V$. Our argument is exactly the same as for Theorem 1.1. More precisely, we show that, for any $p > 2$:

(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.

Let us start by considering the action of V on the set of rationals, recall that

$$V : z \mapsto 1/\bar{z}$$

so that for coprime integers a, b

$$a/b \mapsto b/a.$$

If $a/b \neq \pm 1$ then this map preserves the geodesic joining a/b and b/a and the λ-length of this arc can be computed using the determinant formula:

$$\lambda - \mathrm{length} = \left| \begin{pmatrix} a & b \\ b & a \end{pmatrix} \right| = a^2 - b^2 = \pm p$$

So if this is to be a lift of one of our arcs and since p is prime the only solutions must satisfy (after possibly permuting a, b):

$$a + b = \pm p,\ a - b = \pm 1.$$

Thus up to changing sign

$$a = \pm(p+1)/2,\ b = (p-1)/2.$$

Thus we have:

Lemma 6.2. *The automorphism induced by V preserves two and exactly two geodesics in X.*

6.3. **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 coprime is (almost) closed under multiplication: if p, q are sums of squares and at least one is odd then the product is also a sum of squares. For example one has

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

and

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

but of course

$$20 = 2 \times 10 = |(1+i)(1+3i)|^2 = |-2+4i|^2 = 2^2 + 4^2.$$

Despite this multiplication by 2 seems to have a nice geometric interpretation in the spirit of our analysis of arcs but multiplication by 5 does not seem to have one. Let p be a prime such that $p-1$ is a multiple of 4 and m a square root of -1 then $\frac{m}{p} + \frac{i}{p}$ is a point of the $\Gamma(2)$-orbit of i. The semi circle joining $\frac{m-1}{p}$ to $\frac{m+1}{p}$ has λ-length $2p$ and passes through $\frac{m}{p} + \frac{i}{p}$. So by Lemma 3.3 $2p$ is a sum of squares.

It seems difficult to prove that 65 is a sum of squares directly using our approach with Burnside's lemma and considering the set X of arcs of λ-length 65. Naively applying Burnside's lemma yields:

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

6.3.1. *The Ptolemy Identity and multiplicativity.* Looking at this equation modulo 4 is not helpful. However, by using the fact that λ-lengths satisfy the Ptolemy Identity one can prove a weaker form of the closure property quite easily.

The *Ptolemy Identity* concerns the λ-lengths of an ideal quadrilateral see Figure 5. If $X^{\pm}, Y^{\pm}$ are the λ-lengths of pairs of opposite sides and $Z^{\pm}$ of the diagonals then

$$X^+X^- + Y^+Y^- = Z^+Z^- \tag{4}$$

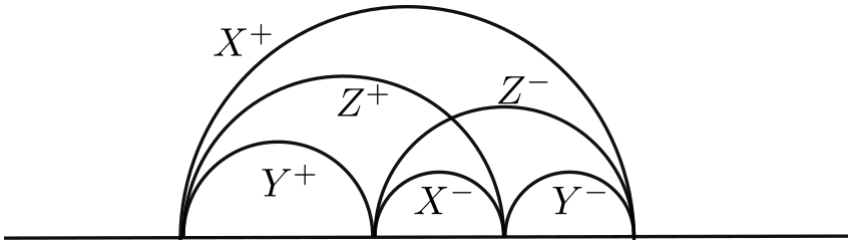

FIGURE 5. Ptolemy Identity diagram.

Now suppose that p, q are distinct integers which can be written as sums of squares, by Lemma 3.3 there is a pair of arcs on $\mathbb{H}/\Gamma(2)$ with λ-lengths p and q respectively which meet in the projection of i to the surface. In the upper half space $\mathbb{H}$ there is a corresponding pair of lifts that intersect in i. The convex hull of these arcs is an ideal quadrilateral in $\mathbb{H}$ to which we can apply the Ptolemy Identity. Observe that in addition this quadrilateral is invariant under $z \mapsto -1/z$ so $X^+ = X^-$ and $Y^+ = Y^-$, such a quadrilateral is a *parallelogram* for λ-length, and we have

$$(X^+)^2 + (Y^+)^2 = pq. \tag{5}$$

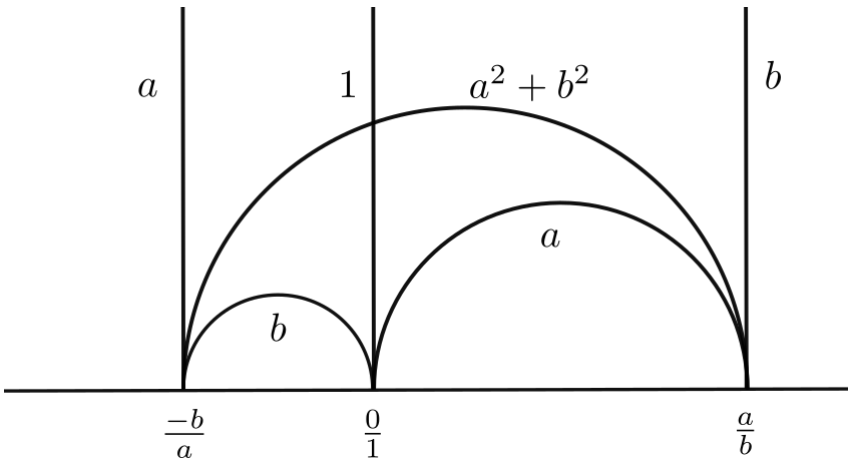

FIGURE 6. A λ-length parallelogram invariant under $z \mapsto -1/z$. Its vertices are $\infty, -b/a, 0, a/b$. This illustrates a special case of the Ptolemy Identity.

Note that if p, q are square free and coprime then X^+ and Y^+ are automatically coprime though this may not be the case in general so more work is required to prove the strong form of closure under multiplication.

If $q = p > 2$ then find an arc in $\mathbb{H}$ passing through i and of λ-length p. The image of the arc under $z \mapsto \bar{z}$ also passes through i and has λ-length p. Now proceed as before to obtain an ideal parallelogram which represents p^2 as a sum of squares. Note that this fails for 2 as the initial arc joins -1 to 1 and so is invariant under $z \mapsto \bar{z}$.

7. Concluding remarks

We have given a geometric treatment of Fermat's theorem using the automorphisms of the surface $\mathbb{H}/\Gamma(2)$ and Penner's λ-lengths. In another work we will study the relation between the quadratic form $x^2 + xy + y^2$ and $\mathbb{H}/\Gamma(2)$.

References

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

[2] 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)

[3] *Ptolemy Relation and Friends,* Anna Felikson, preprint arXiv:2302.06379

[4] Lester R Ford, *Automorphic Functions*

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

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

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

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

[9] B. Springborn. *The hyperbolic geometry of Markov's theorem on Diophantine approximation and quadratic forms.* Enseign. Math., 63(3-4):333–373, 2017. doi:10.4171/LEM/63-3/4-5.

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

[11] 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

Institut Fourier 100 rue des maths, BP 74, 38402 St Martin d'Hères cedex, France

Email address: mcshane at univ-grenoble-alpes.fr