GEODESICS AND VALUES OF QUADRATIC FORMS

GREG MCSHANE

1. Introduction

Recall that $\bar{m} \in \mathbb{F}_p$ is a *quadratic residue* iff there is $\bar{x} \in \mathbb{F}_p$ such that $\bar{x}^2 = \bar{m}$. The first results concerning quadratic residues were stated by Fermat.

Determining the primes p for which a particular m is a quadratic residue is a fundamental problem, studied by Fermat, Euler, Legendre and Gauss. The case of -1 is particularly famous:

Theorem 1.1. *Let p be a prime then the equation*

$$\bar{x}^2 = -1$$

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

Classically, as observed by Euler this result is the first step in proving an a priori stronger result:

Theorem 1.2 (Euler)**.** *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.

So there are two parts to Euler's proof:

(1) the *reciprocity step* which consists of proving Theorem 1.1. This tells us that there are $a, k \in \mathbb{N}$ with $kp = a^2 + 1^2$.
(2) the descent step which consists of showing there is a monotonically decreasing sequence $k_i \in \mathbb{N}$ with $k_i p = a_i^2 + b_i^2$ that terminates with some $k_i = 1$.

There are myriad proofs of these theorems but the approach initiated by Heath-Brown in [13] has inspired many admirers if not imitators [11, 15, 18, 7], probably The most elegant of these is probably Dolan's proof [6]. but see Elsholtz's very nice account [8] of the methd in general. In some sense this manuscript is a companion to Elsholtz's where, instead of looking at the number theory as combinatorics, we work in an explicitly geometric context. As such we refer the reader to Elsholtz for historical perspective and the like.

1.1. **Involutions.** The Heath-Brown proof is intriguing as it appears so different from Euler's method in that it avoids both reciprocity and descent. The essential ingredients in the Heath-Brown paper are a finite set X equipped with a pair of involutions such that:

- Any fixed point of the one of the involutions, should it exist, is a solution of the equation;
- The other involution has a unique fixed point which is easy to compute.

The existence of the unique fixed point of the second involution allows one to conclude that the set X has an odd number of elements and so that any involution has a fixed point.

Generalov [11] has given a proof (see also Elsholtz [8]) in the style of Zagier of one part of another result conjectured by Fermat:

Theorem 1.3. *Let p be a prime then the equation*

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

has a solution in integers iff $p = 2$ or p is congruent to 1 or 3 mod 8.

Obviously the case $p = 2$ is trivial. Of the other two cases both Generalov and Elsholtz give proofs for $p = 3 \mod 8$ though $p = 1 \mod 8$ resisted them. We will give an entirely geometric proof of this result and a "hybrid" proof of the other case.

Our hybrid proof is based on an analogue of Theorem 1.1 (see the discussion in Section 2.1 for details):

Lemma 1.4. *Let p be a prime then -2 is a quadratic residue, iff p is congruent to 1 or 3 mod 8.*

1.2. **Arcs on a surface with cusps.** What inspired us was the insistence of V Sergesciu that one could understand quadratic forms like x^2+y^2 and more generally x^2+my^2 from a geometric point of view (see Cox's book [5] for the classical approach). Certainly Conway [4] has developed a geometric theory of the values of a quadratic form using what he calls the *topograph of a quadratic form* (see Hatcher's book [12] for a complete account with back ground) but he is in the main concerned with indefinite forms.

Our approach, like Conway's, is based on properties of Γ acting on $\mathbb{H}$. Whilst the finite set X employed in the variations of Heath-Brown's method is readily described in terms of solutions in positive integers to some equation (see Dolan [6]) the geometric interpretation of our set X requires some knowledge of the elementary theory of Fuchsian groups acting on $\mathbb{H}$ and its ideal boundary $\partial\mathbb{H}$ which we will now sketch. The reader, unfamiliar with the relationship between hyperbolic geometry and number theory, should consult Springborn's articles [24, 25] where a dictionary between geometric and number theoretic notions is presented.

Let $\Gamma = \mathrm{SL}(2,\mathbb{Z})$ and recall that it admits an action on $\mathbb{H} \cup \partial\mathbb{H}$ via

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma, z \mapsto \frac{az+b}{cz+d}.$$

The restriction of this action to $\mathbb{H}$ is properly discontinuous but not free and the quotient $\mathbb{H}/\Gamma$ is a non compact singular surface called the *modular surface*. The Γ-orbit of ∞ is exactly the extended rationals $\mathbb{Q} \cup \{\infty\}$. Let $G < \Gamma$ be a finite index subgroup, we will be concerned with involutions of the cover $\mathbb{H}/G \to \mathbb{H}/\Gamma$ and their actions on the arcs on the cover. Evidently $\mathbb{H}/G$ is non compact and has a finite number of ends called *cusps* which are in 1-1 correspondence with G-orbits of $\mathbb{Q} \cup \{\infty\}$. In fact Γ possesses many subgroups of interest amongst which:

- Γ' its commutator subgroup
- the family of principle congruence subgroups $\Gamma(N)$.
- the Hecke congruence subgroups of level N denoted $\Gamma_0(N)$ and their normalisers $\Gamma_0^+(N)$.

In [17] we studied $\Gamma(2)$ and the geometry of arcs on its quotient surface $\mathbb{H}/\Gamma(2)$ to give a proof of Theorem 1.2. We define an *arc* to be the image on $\mathbb{H}/G$ of any Poincaré geodesic joining a pair of distinct elemements of $\mathbb{Q}\cup\{\infty\}$. As such an arc is a complete geodesic on $\mathbb{H}/G$ with both of its ends "terminating" at cusps, so it has infinite length, but if we only consider the portion outside some family of sufficiently small uniform neighbors of the cusps then this will be finite. The λ-length of the arc is the exponential of half the length of this finite portion. Whilst this definition works well for simple arcs, since the portion of the arc outside the uniform cusp region is connected, more care is needed if the arc is not simple. Since Penner [19] first defined λ-length there has been much work on their applications

- Weil-Peterson volume of moduli space [19]
- cluster algebras [9].
- Conway-Coxeter frieze patterns [3].

1.3. **Counting arcs.** Our approach to sums of squares in [17] (see Section 4 for a sketch) depended on determining the number of arcs on a surface of λ-length p. Fortunately, the surface we were lead to study $\mathbb{H}/\Gamma(2)$ is probably the easiest surface to count arcs on.

Theorem 1.5. *Let p be a prime number then the number of arcs of λ-length p on the surface $\mathbb{H}/\Gamma(2)$ is $3(p-1)$.*

The surface $\mathbb{H}/\Gamma(2)$ has three cusps since the extended rationals splits as three $\Gamma(2)$-orbits

$$\mathbb{Q} \cup \{\infty\} = \Gamma(2).\{0\} \sqcup \Gamma(2).\{1\} \sqcup \Gamma(2).\{\infty\}.$$

This is a manifestation of the fact that the $\Gamma(2)$ action preseres parity. By *parity* we mean one of the three classes obtained by taking

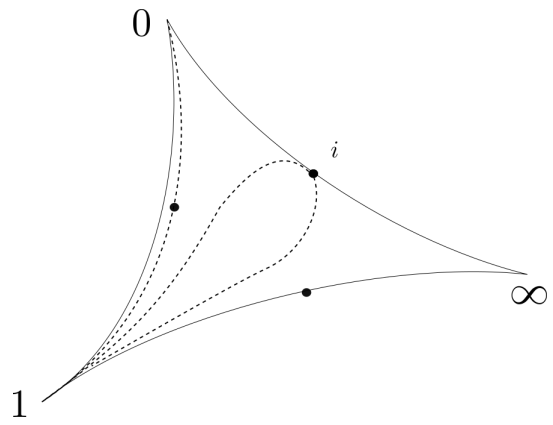

FIGURE 1. The surface $\mathbb{H}/\Gamma(2)$. The dotted lines are arcs one of which is a loop of λ-length 2, and the other, which joins cusps labelled 0 and 1, has λ-length 1.

the numerator and denominator of a fraction $\frac{a}{c}$ modulo 2 where, by convention, $\infty = \frac{1}{0}$. The proof of Theorem 1.5 is basically the observation that the Poincaré geodesics of $\mathbb{H}$ (vertical lines) ending at $2k/p \in \mathbb{Q}$, $1 \leq k \leq p-1$ is a set of lifts for the arcs joining the cusps labeled 0 and ∞ on $\mathbb{H}/\Gamma(2)$ (see Figure 1.)

From a topological point of view the arcs on $\mathbb{H}/\Gamma(2)$ evidently fall into two families:

- arcs that join distinct cusps.
- arcs with both ends terminating at the same cusp

We will refer to the second kind of arc as a *loop* (see Figure 1). Amusingly one can characterise loops (algebraically) using λ-length.

Lemma 1.6. *An arc on $\mathbb{H}/\Gamma(2)$ is a loop if and only if its λ-length is even.*

1.4. **Representing arcs as matrices.** To give a proof of Lemma 1.6 we clearly need an algebraic characterisation of λ-lengths. It is fortunate then that computing the λ-length of an arc α on $\mathbb{H}/\Gamma(2)$ with respect to the canonical choice of cusp regions is relatively simple. One considers a lift $\hat{\alpha} \subset \mathbb{H}$ of the arc. This is a Poincaré geodesic joining a pair of distinct rationals $a/c, b/d \in \mathbb{Q}$ and the λ-length is just the absolute value of the determinant of the matrix

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

Now Lemma 1.6 is immediate, for if $\frac{a}{c}$ and $\frac{b}{d}$ have the same parity then reducing modulo the matrix modulo 2 then both columns are the same and so the parity of the determinant is 0. Conversely, if the parity of the determinant is 0 then columns modulo 2 are linearly dependent

over $\mathbb{Z}/2\mathbb{Z}$ and this means they have the same parity and so represent the same cusp on $\mathbb{H}/\Gamma(2)$.

In what follows we use matrices over $\mathbb{Z}$, up to an equivalence relation denoted $\sim$, to represent unoriented arcs. Explicitly for a pair of invertible matrices

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

iff $\{\frac{a}{c}, \frac{b}{d}\} = \{\frac{a'}{c'}, \frac{b'}{d'}\} \subset \mathbb{Q} \cup \{\infty\}$. We think of the matrices

$$\begin{pmatrix} 1 & k \\ 0 & p \end{pmatrix}, \begin{pmatrix} k & 1 \\ p & 0 \end{pmatrix}.$$

as representing an arc joining ∞ to k/p but with opposite orientations.

This matrix representation will prove very convenient when later (see Section 6.1) have to check that two arcs are in the same G-orbit for some $G < \mathrm{SL}(2, \mathbb{Z})$ and so project to the same arc on the quotient surface $\mathbb{H}/G$.

1.5. **Philosophy.** In some sense we adopt the same approach to (positive definite) integer quadratic forms as Sarnak [21] and Penner in [19] (see also [24].) Such a quadratic form is associated to an involution of the upper half space $\mathbb{H}$:

- if the form is (positive) definite then the involution is orientation preserving, for example $x^2 + y^2$ is associated with $z \mapsto -1/z$
- if the form is indefinite then the involution is orientation reversing, for exampleg $x^2 - y^2$ is associated with $z \mapsto 1/\bar{z}$.

What will be of interest to us here is when our involution normalises some group G and so induces an automorphism on the quotient surface $\mathbb{H}/G$. If $G = \Gamma(2)$ then is the case for both of the involutions above.

More generally, we will show that questions concerning the values of the quadratic form $x^2 + my^2$ can be interpreted of terms of arcs on the surface $\mathbb{H}/\Gamma_0^t(m)$ where $\Gamma_0^t(2)$ is the *anti Hecke congruence group*:

$$\Gamma_0^t(m) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ * & 1 \end{pmatrix} \mod m \right\} < \Gamma(m).$$

This group is normalised by the involution $z \mapsto -m/\bar{z}$ which fixes $i\sqrt{m}$. For $\Gamma_0^t(2)$ we will work through all the calculations and the geometry in detail. This will allow us to give a geometric proof of the fact thatany prime congruent to 3 modulo 8 is represented by $x^2 + 2y^2$. Of course, it is also true (see Lemma 1.4) any prime congruent to 1 modulo 8 is represented by $x^2 + 2y^2$ but unfortunately we are unable to give a purely geometric proof of this.

2. Klein four group and the 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

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

The group G has exactly four elements namely:

- the trivial element which has $p-1$ fixed points
- $\bar{x} \mapsto -\bar{x}$ which has no fixed points
- $\bar{x} \mapsto 1/\bar{x}$ has exactly two fixed points namely 1 and -1.
- $g : \bar{x} \mapsto -1/\bar{x}$ 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

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

This proves Theorem 1.1.

2.1. **Extending.** Thus we have shown that $-1 \in \mathbb{F}_p$ is a quadratic residue if p is of the form $4k+1$. It is natural to consider the other questions considered by Fermat: namely for which values of p are -2 and -3 residues?

In fact -2 is a residue if p is 2 or of the form $8k+1$ or $8k+3$ (Lemma 1.4). Showing this in the spirit of Heath-Brown requires one to consider a group generated by the involutions

$$\begin{aligned} \bar{x} &\mapsto -\bar{x} \\ \bar{x} &\mapsto 2/\bar{x}. \end{aligned}$$

One immediately sees that things are more complicated as the second involution has fixed points if and only if 2 is a quadratic residue whereas $\bar{x} \mapsto 1/\bar{x}$ always had exactly two fixed points. Thus there are two cases:

- $p = 8k+1$ and both 2 and -2 are residues
- $p = 8k+3$ and -2 is a residue but 2 is not.

To prove this second assertion one must show that $x \mapsto 2/x$ has no fixed point so that, by Burnside, $x \mapsto -2/x$ has two fixed points both of which are square roots of -2. Thus one must show that the only solution of the associated diophantine equation

$$np = x^2 - 2y^2,$$

is the trivial solution $n = x = y = 0$. Now using the fact that $x^2 - 2y^2$ is the norm of $x + y\sqrt{2} \in \mathbb{Z}[\sqrt{2}]$ which is a euclidean ring for this norm, one reduces as Euler did to considering just the solutions of

$$p = x^2 - 2y^2.$$

Finally, one concludes by showing that if x, y are integers then $x^2 - 2y^2$ never takes the value 3 mod 8.

2.2. **The case $p = 11$.** The first case of genuine interest in understanding $x \mapsto -2/x$ is that of $\mathbb{F}_{11}^*$. Evidently, $11 = 3^2 + 2 \times 1^2$ so that $\bar{3}$ and $-\bar{3} = \bar{8}$ are the fixed points of the involution.

The reduction homomorphism $x \mapsto \bar{x}$ allows one to identify the elements of $\mathbb{F}_p$ with the equivalence classes that constitute the quotient $\mathbb{Z}/p\mathbb{Z}$. It is usual to choose the integers $0, 1, 2 \ldots p-1$ as representatives for the latter, however, we shall find it convenient to work with another set of representatives, namely the even integers $0, 2, 4 \ldots 2p-2$. Using the euclidean algorithm to compute $\bar{x}^{-1} \in \mathbb{F}_{11}$ we have the following table:

x	12	2	14	4	16	6	18	8	20	10
$\bar{x}$	1	2	3	4	5	6	7	8	9	10
$\bar{x}^{-1}$	12	6	4	14	20	2	8	18	16	10
$-2\bar{x}^{-1}$	20	10	14	16	4	18	6	8	12	2

One notes that there are two fixed points of $-\bar{x} \mapsto -2\bar{x}^{-1}$ namely $\overline{14} = \bar{3}$ and $\bar{8}$.

3. Counting sums of squares

We count solutions for the diophantine problem $n = c^2 + d^2$ in two ways:

- Firstly by showing that solutions are naturally associated to the Γ orbit of i;
- Secondly by counting arcs of λ-length n.

We extend these results to counting solutions of $n = mc^2 + d^2$ using the Γ-orbit of $i\sqrt{m}$.

3.1. **From solutions to Γ-orbits.** The transformation $z \mapsto z+1$ generates an infinite cyclic group acting on $\mathbb{H}$. The standard fundamental domain for this group is an infinite strip, which we will refer to as the *fundamental strip*, consisting of all the $z \in \mathbb{C}$ such that the real part is between 0 and 1.

Lemma 3.1. *Let $n \geq 2$ be an 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 of points of $\Gamma.\{i\}$, the $\mathrm{SL}(2,\mathbb{Z})$ orbit of i, in the fundamental strip such that the imaginary part (euclidean height) is $\frac{1}{n}$.

Note that we are counting $c^2 + d^2$ and $d^2 + c^2$ as *different* representations of n.

Proof. Suppose there is a point w verifying the hypotheses, in particular

$$w = \frac{ai + b}{ci + d}, \text{ for some } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma.$$

Then one has:

$$\frac{1}{n} = \operatorname{Im} w = \operatorname{Im}\left(\frac{ai + b}{ci + d}\right) = \frac{\operatorname{Im} i}{c^2 + d^2}, \tag{3}$$

so $n = c^2 + d^2$ as claimed.

Conversely if c, d are coprime integers such that $n = c^2 + d^2$ then there exists a, b such that

$$ad - bc = 1 \Rightarrow A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma.$$

By applying a suitable iterate of the parabolic transformation $z \mapsto z+1$, if necessary one can choose w such that $0 \leq \operatorname{Re} w < 1$.

□

3.2. **From $\Gamma(2)$-orbits to arcs.** Suppose that n can be written as a sum of squares $c^2 + d^2$ and w is the corresponding point in the fundamental strip then we can associate a Poincaré geodesic to w in a natural way: we simply take the vertical line that passes through w. This geodesic joins two points in the ideal boundary of $\mathbb{H}$ namely ∞ and $\frac{ac+bd}{n} \in \mathbb{Q}$. This geodesic projects to an arc on the surface $\mathbb{H}/\Gamma(2)$ and, using the definition above, its λ-length is n.

3.3. **Real part and reduction modulo n.** By a similar argument as for Lemma 3.1 one has a slightly more general result:

Lemma 3.2. *Let $n \geq 2$ be an integer and $m < n$ a square free integer. The number of ways of writing n as a sum of squares*

$$n = mc^2 + d^2$$

with c, d coprime integers is equal to the number of points of $\Gamma.\{i\sqrt{m}\}$ the $\mathrm{SL}(2,\mathbb{Z})$ orbit of $i\sqrt{m}$, in the fundamental strip at euclidean height $\frac{\sqrt{m}}{n}$.

Thus the imaginary part of $w = \left(\frac{ai\sqrt{m}+b}{ci\sqrt{m}+d}\right)$ arises naturally in relation to counting the number of ways of representing an integer n as mc^2+d^2. Now the real part of w is

$$\frac{mac+bd}{mc^2+d^2},$$

and it also plays and important role: it is a fraction whose denominator is n and whose numeratoris nothing other than a square root of $-\bar{m}$ in $\mathbb{Z}/n\mathbb{Z}$. In Section 6 we will give a converse to this result in certain cases, showing how to "lift" a square root in $\mathbb{Z}/n\mathbb{Z}$ to a solution of $mc^2+d^2=n$.

Lemma 3.3. *Let $a, b, c, d, m \in \mathbb{Z}/n\mathbb{Z}$ and suppose that they satisfy the relations*

$$mc^2 + d^2 = 0 \tag{4}$$

$$ad - bc = 1. \tag{5}$$

Then

$$(mac+bd)^2 = -m.$$

Proof. To prove this begin by noting that:

$$\begin{aligned}(mac+bd)^2 &= (ma^2)(mc^2) + b^2d^2 + 2m(abcd)\\ m(ad-bc)^2 &= (ma^2)d^2 + b^2(mc^2) - 2m(abcd) = m\end{aligned}$$

and adding these:

$$(mac+bd)^2 + m = (ma^2)(mc^2+d^2) + b^2(mc^2+d^2) = 0.$$

□

4. Inversions and the proof of Theorem 1.2

We present a sketch of the geometric proof given in our previous work with Sergiescu[17].

Let $\mathbb{H}$ denote the Poincaré upper half plane and $\partial\mathbb{H}$ its ideal boundary ie $\mathbb{R} \cup \{\infty\}$. Recall that an *inversion* is an orientation reversing isometry of $\mathbb{H} \cup \partial\mathbb{H}$. A Poincaré geodesic is either a vertical line or a semicircle orthogonal to $\mathbb{R}$. In both cases it is uniquely determined by its endpoints in the ideal boundary. To each Poincaré geodesic is associated a unique inversion which fixes it pointwise. The inversion $\phi_h : z \mapsto -\bar{z}$ fixes 0 and ∞ and so the geodesic joining them. The group of isometries acts transitively on pairs of distinct points $a, b \in \partial\mathbb{H}$ and so there is an inversion that fixes the geodesic joining a, b which is in fact conjugate to ϕ_h. The inversion fixing $1, -1$ is easily seen to be $\phi_v : z \mapsto \frac{1}{\bar{z}}$.

Note that if a, b are coprime integers then:

- The image of $\frac{a}{b}$ under ϕ_h is $-\frac{a}{b}$ and the λ-length of the geodesic joining them is $2|ab|$.

- The image of $\frac{a}{b}$ under ϕ_v is $\frac{b}{a}$ and the λ-length of the geodesic joining them is $|a^2 - b^2|$.

It follows from these remarks that:

Lemma 4.1. *Let $p > 2$ be a prime then:*

- *There is no arc of λ-length p invariant under ϕ_h.*
- *There are exactly two arcs of λ-length p invariant under ϕ_v;*

Proof. The first part is easy because the λ-length of such an arc is an even integer. The second part follows from the fact that p factorises as

$$p = |(a - b)||(a + b)|,$$

so up to permutation and change of sign a, b are the integers $\frac{1}{2}(p \pm 1)$. □

Our proof of Theorem 1.2 consists in noting that the above involutions normalise $\Gamma(2)$ and so induce involutions of the quotient surface $\mathbb{H}/\Gamma(2)$. We take X to be the set of arcs of λ-length p joining the cusps marked 0 and ∞ (see Figure 1.5) and using Lemma 4.1 compute the terms in the Burnside formula:

$$4|X/G| = (p - 1) + 2 + |X^g|,$$

where g is the involution induced by $z \mapsto -1/z$. We conclude that there is an arc that meets $\Gamma.\{i\}$ and obtain p as a sum of squares by Lemma 3.1.

Later we will need to consider arcs invariant under the inversion $z \mapsto \frac{m}{\bar{z}}$ for $m = 2, 3$. This is an inversion in a semi circle with endpoints $\pm\sqrt{m} \in \mathbb{R}$. If $\frac{a}{b} \in \mathbb{Q}$ then it is the endpoint of exactly one arc is invariant under this involution and the of this arc is λ-length $|a^2 - mb^2|$. For $m = 2$ one can show that no invariant arc has λ-length congruent to 3 or 5 mod 8.

5. Congruence subgroups

The Hecke congruence subgroup $\Gamma_0(N)$ of level N is the subgroup of $\Gamma = \mathrm{SL}(2, \mathbb{Z})$ consisting of the matrices satisfying:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \mod N.$$

For $N = 2$ this is generated by just two elements namely:

$$P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } Q = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}.$$

The product $P^{-1}Q$ is an element of order 2:

$$P^{-1}Q = \begin{pmatrix} -1 & -1 \\ 2 & 1 \end{pmatrix}.$$

The quotient $\mathbb{H}/\Gamma_0(2)$ is a non-compact orbifold with two cusps, corresponding to the cyclic groups generated by P and Q, and a single

cone point corresponding to $P^{-1}Q$. This orbifold admits a Klein four group as its group of automorphisms and the quotient by this group is a hyperbolic triangle with angles $0, \pi/2, \pi/4$.

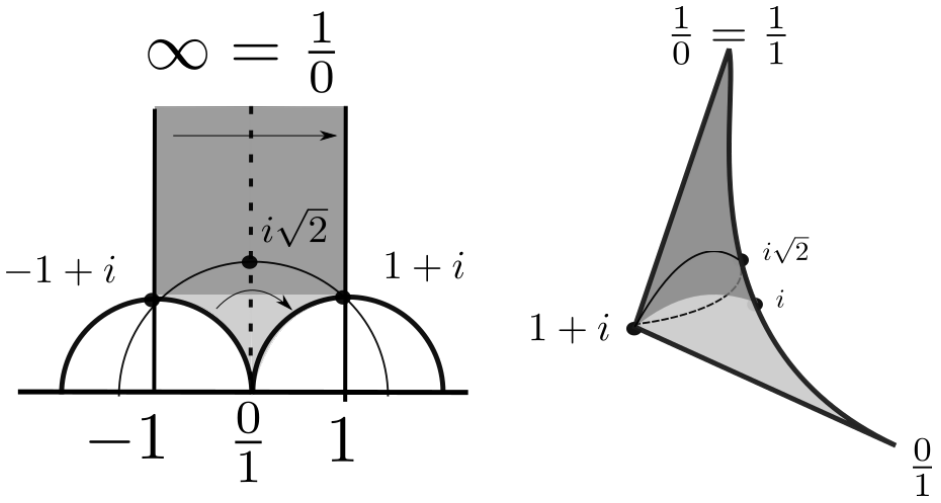

FIGURE 2. On the left a fundamental domain for $\Gamma_0^t(2)$ with side pairings. On the right the quotient surface $\mathbb{H}/\Gamma_0^t(2)$, the dark region is a cusp region.

The action of $\Gamma_0(2)$ on $\mathbb{Q} \cup \{\infty\}$ is not transitive and there are two orbits. Now $\Gamma(2) < \Gamma_0(2)$ so each of these orbits is a union of $\Gamma(2)$-orbits. Since $\Gamma(2)$ preserves the parity of the numerator and denominator of a fraction there are exactly three $\Gamma(2)$ orbits corresponding to $\frac{0}{1} = 0$, $\frac{1}{1} = 1$, $\frac{1}{0} = \infty$. Now since P maps 0 to 1:

$$\Gamma_0(2)\{0\} = \Gamma(2)\{0\} \cup \Gamma(2)\{1\}.$$

In the previous section we considered the action of the involution $x \mapsto -2/x$ on $\mathbb{F}_p^*$. It is natural to study the action of the corresponding involution of $\mathbb{H}$ that is $z \mapsto -2/z$ but unfortunately this does not normalise $\Gamma_0(2)$. However it does normalise the *anti Hecke congruence-group*:

$$\Gamma_0^t(2) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ * & 1 \end{pmatrix} \mod N \right\} < \Gamma(2).$$

The Hecke congruence group and the anti Hecke group are isomorphic and in fact $z \mapsto -1/z$ conjugates them in Γ . We can determine the orbits of this group on $\mathbb{Q}$ using this conjugation and we have:

$$\Gamma_0^t(2)\{\infty\} = \Gamma(2)\{\infty\} \cup \Gamma(2)\{1\}.$$

5.1. **Canonical cusp regions.** 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. If $G < \mathrm{SL}(2,\mathbb{Z})$ is any finite index subgroup then each Ford circle projects to a cusp region on $\mathbb{H}/G$ and we call this system the *canonical system of cusp regions*.

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

Corollary 5.2. *The λ-length of the arc joining $a/c, b/d \in \mathbb{Q}$ is the absolute value of the determinant of the associated matrix*

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

Proof. There exists $b' \in \mathbb{Z}$ and a matrix $A' \in \mathrm{SL}(2,\mathbb{Z})$ such that the product $A'A$ is an upper triangular matrix:

$$A'A = \begin{pmatrix} 1 & b' \\ 0 & \det A \end{pmatrix}.$$

The image of a/c under the Mobius transformation associated to A' is ∞ and the image of b/d is $b'/\det A$. The Ford circle at ∞ is F and the diameter of the circle tangent at $b'/\det A$ is $(\det A)^2$.

□

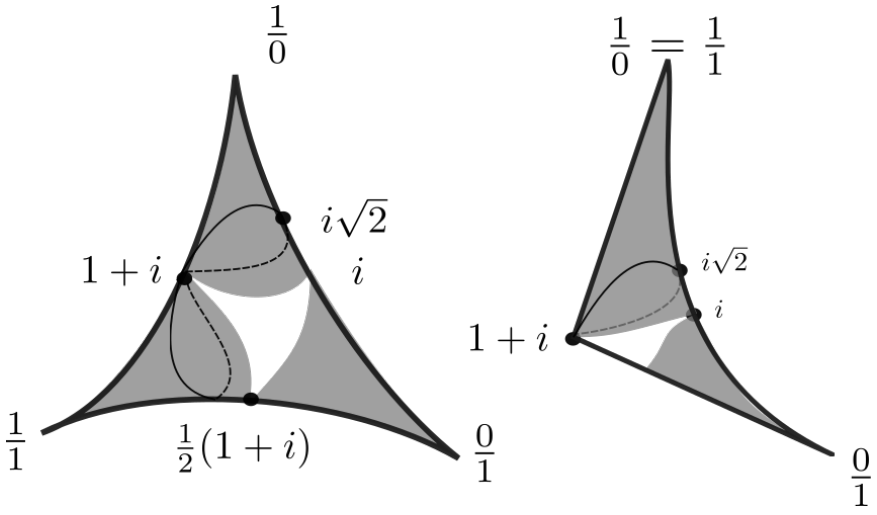

FIGURE 4. On the left $\mathbb{H}/\Gamma(2)$ with the cusp regions inherited from the Ford circles $\mathbb{H}$. On the right $\mathbb{H}/\Gamma_0^t(2)$ with the unmodified cusp regions.

5.2. **Cusp regions on $\mathbb{H}/\Gamma_0^t(2)$.** The canonical system on $\mathbb{H}/\Gamma(2)$ consists of three cusp regions one for each of the three cusps $0, 1, \infty$. The map $z \mapsto z/z+1$ fixes 0 and normalises $\Gamma(2)$, so induces an automorphism, in fact an involution of $\mathbb{H}/\Gamma(2)$ which fixes the cusp labeled 0. The quotient of $\mathbb{H}/\Gamma(2)$ by the involution is naturally identified with the surface $\mathbb{H}/\Gamma_0^t(2)$ inherits a system of cusp regions from $\mathbb{H}/\Gamma(2)$ via the quotient map. The involution $z \mapsto -2/z$ normalises $\Gamma_0^t(2)$ so induces an automorphism of $\mathbb{H}/\Gamma_0^t(2)$ which fixes the points labelled $1+i$ and $i\sqrt{2}$, swaps the cusps labelled $\frac{1}{0}$ and $\frac{0}{1}$ but which does not swap the cusp regions inherited from $\mathbb{H}/\Gamma(2)$. In fact a computation shows that the cusp region for $\frac{1}{0}$ has area 2 whilst the cusp region for $\frac{0}{1}$ has area 1. We remedy this by choosing a pair of cusp regions which are tangent at the fixed point of the automorphism and which both have area $\sqrt{2}$. To do this

- the cusp region for $1/0$ shrinks by a factor of $\sqrt{2}$
- whilst the cusp region for $0/1$ expands by $\sqrt{2}$.

The lifts of this modified pair of cusp regions to $\mathbb{H}$ form a family of circles each of which, like the Ford circles, is tangent to the real line at a rational $\frac{m}{n} \in \mathbb{Q}$. However, the diameter of the circle tangent at m/n is no longer $\frac{1}{n^2}$ as in Lemma 5.1 rather we have:

Lemma 5.3. *The diameter of the circle tangent at m/n:*

- $\sqrt{2} \times 1/n^2$ *if m is even.*
- $1/\sqrt{2} \times 1/n^2$ *if m is odd.*

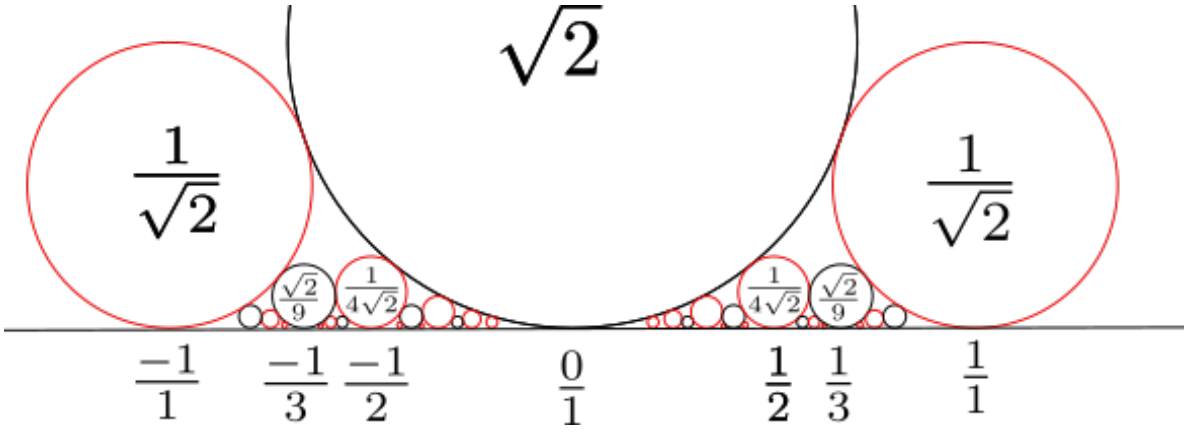

FIGURE 5. Modified Farey circles

5.3. **Arcs on $\mathbb{H}/\Gamma_0^t(2)$.** The surface has two cusps and so there are two kind of arc

- arcs that join distinct cusps $0/1$ and $1/0$
- arcs that have both ends at the same either cusps $0/1$ or $1/0$.

It follows from Lemma 5.3:

Lemma 5.4. *Arcs of the first kind, that is those which join distinct cusps $0/1$ and $1/0$, have the same λ-length for the inherited cusp regions and our modified cusp regions.*

Now any arc of the first kind lifts to a vertical line ending at some rational $\frac{m}{n} \in \mathbb{Q}$. For each $p > 2$ prime the projections to $\mathbb{H}/\Gamma_0^t(2)$ of the Poincaré geodesics $\infty, \frac{2k}{p}$ with $k = 1, 2 \dots p-1$ are distinct arcs of the first kind and each of these has λ-length p for our choice of cusp regions:

Corollary 5.5. *If p is an odd prime then there are exactly $p-1$ arcs of λ-length p.*

Since our involution swaps cusps only arcs of the first kind can be invariant for it.

Theorem 5.6. *If p is a prime of the form $8k+3$ then there are integers x, y such that*

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

Proof. As before we consider an action of a Klein four group, that is the automorphisms of the surface $\mathbb{H}/\Gamma_0(2)$, on a set X of cardinality $p-1$ namely the set of arcs of the first kind having λ-length p.

Under the hypothesis the orientation reversing involutions induced by the inversions in the circles fix no elements of X so the Burnside formula gives:

$$4|X/G| = (p-1) + |X^g|,$$

where g denotes the orientation preserving involution. Now $p-1$ is congruent to 2 modulo 4 so $|X^g|$ is too and g fixes an arc.

□

6. Lifting involutions

In order to simplify the exposition we suppose that p is a prime and $m < p$ a square free integer such that $p-1$ is divisible by m. Note that if $m = 2$ then this condition is essentially trivial.

Let $\mathbb{G}$ denote the set of $p-1$ arcs joining ∞ and $\frac{mk}{p}$ where $1 \leq k < p$. These arcs project to pairwise distinct arcs of λ-length p on the surface $\mathbb{H}/\Gamma_0(m)$. There is a natural bijection taking an arc to the congruence class of the denominator modulo p:

$$\begin{array}{rcl} \pi : \mathbb{G} & \to & \mathbb{F}_p \\ mk : & \mapsto & \overline{mk}. \end{array}$$

Evidently, since π is a bijection any map $f : \mathbb{F}_p \to \mathbb{F}_p$ lifts to a map of $\mathbb{G}$

$$\begin{array}{ccc} \mathbb{G} & \xrightarrow{F} & \mathbb{G} \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ \mathbb{F}_p & \xrightarrow{f} & \mathbb{F}_p \end{array}$$

and an obvious question is: when does an automorphism f of $\mathbb{F}_p$ lift to an automorphism of the quotient surface?

In particular for our three involutions

- $z \mapsto -\bar{z}$
- $z \mapsto \frac{m}{\bar{z}}$
- $z \mapsto \frac{-m}{z}$

wish to show that each of them

(1) induces an involution of $\mathbb{G}$,
(2) the map lifts the corresponding involutions on $\mathbb{F}_p$ (see Lemma 6.1 below).

We think of this second point as extending the calculation in the previous section 3.3. The first of our involutions $z \mapsto \bar{z}$ is easily dealt with since $-(mk) = m(-k)$. The third is the composition of the first and the second, so it suffices to prove (1) and (2) for $z \mapsto \frac{m}{\bar{z}}$.

6.1. **Some matrix algebra.** Since k and p are coprime there is a Bezout identity, that is there are integers k', p' such that,

$$k'k + p'p = 1$$

Note that we may replace the pair k', p' by another pair $k' + rp, p' - rk$ for any integer r and, by chosing r suitably, may assume that k' is divisible by m. This observation allows us to define maps on $\mathbb{G}$ as

follows. The map $z \mapsto \frac{m}{\bar{z}}$ is an inversion in the half circle with end points $\pm\sqrt{m}$ and is is represented by the matrix

$$\begin{pmatrix} 0 & m \\ 1 & 0 \end{pmatrix}$$

The image of the arc $\infty, \frac{mk}{p}$ under this map is $0, \frac{p}{k}$ and we have the equation

$$\begin{pmatrix} 0 & m \\ 1 & 0 \end{pmatrix} \begin{pmatrix} mk & 1 \\ p & 0 \end{pmatrix} = \begin{pmatrix} mp & 0 \\ mk & 1 \end{pmatrix} \sim \begin{pmatrix} p & 0 \\ k & 1 \end{pmatrix} \tag{6}$$

Consider the matrix identity

$$\begin{pmatrix} p' & k' \\ -k & p \end{pmatrix} \begin{pmatrix} p & 0 \\ k & 1 \end{pmatrix} = \begin{pmatrix} 1 & k' \\ 0 & p \end{pmatrix} \tag{7}$$

This last matrix represents the arc joining ∞ to k'/p and, since k' is divisible m, this is an element of $\mathbb{G}$. Thus we may define a map $\mathbb{G} \to \mathbb{G}$ that takes the arc $\infty, \frac{mk}{p}$ to $\infty, \frac{k'}{p}$. Note further that, since $p = 1 \mod m$,

$$\begin{pmatrix} p' & k' \\ -k & p \end{pmatrix} \in \Gamma_0^t(m)$$

so this map induces a map on the arcs on the quotient surface $\mathbb{H}/\Gamma_0^t(m)$.

Theorem 6.1. *Let f denote the involution on $\mathbb{F}_p$ defined by $x \mapsto mx^{-1}$ and F the map described above then*

$$f(\overline{mk}) = \overline{F(mk)}.$$

Proof. This follows from the observation that on reducing the Bezout identity modulo p one has

$$\bar{1} = \overline{k'k} + \overline{p'p} = \overline{k'k}$$

so that $\overline{k}^{-1} = \overline{k'}$. □

Corollary 6.2. *There is an involution of $\mathbb{G}$ that lifts the involution $\bar{x} \mapsto -m\bar{x}^{-1}$ on $\mathbb{F}_p$.*

Proof. The involution in question is the composition of a pair of involutions that lift. One obtains the required lift G as the composition of the lifts of these involutions. □

6.2. **Explicit formulas for square roots in $\mathbb{F}_p$.** In many cases one can find a formula for the square root of a quadratic residue modulo p.

- If $p = 1 \mod 4$ then, by Wilson's theorem $\left(\frac{p-1}{2}\right)!$ is a square root of -1
- If $p = 1 \mod 4$ and x is a solution of the equation $X^4 + 1 = 0$ then $x^2 + x^{-2} = 0$ and $(x \pm x^{-1})^2 = \pm 2$

- If $p = 1 \mod 6$ and x is a solution of the equation $X^6 + 1 = 0$ then $x^3 + x^{-3} = 0$ and, expanding $(x \pm x^{-1})^3$ using the binomial formula, one has $(x \pm x^{-1})^2 = \pm 3$
- If $p = 3 \mod 4$ and x a quadratic residue then $x^{\frac{p+1}{2}}$ is a square root of x.

7. Appendix: elementary hyperbolic geometry

Let $\Gamma = \mathrm{SL}(2, \mathbb{Z})$ then it admits a action on $\mathbb{H} \cup \partial\mathbb{H}$ via

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma, z \mapsto \frac{az + b}{cz + d}.$$

The action decomposes into two quite different pieces:

(1) the restriction of this action to $\mathbb{H}$ is properly discontinuous but not free and the quotient is a singular surface called the *modular surface* $\mathbb{H}/\Gamma$
(2) the stabiliser of $i \in \mathbb{H}$ is non trivial and is generated by $z \mapsto -1/z$
(3) the action on $\partial\mathbb{H}$ admits dense orbits for example the orbit of ∞ is exactly $\mathbb{Q} \cup \{\infty\}$
(4) the stabiliser of $\infty \in \partial\mathbb{H}$ is the infinite cyclic group generated by $z \mapsto z + 1$
(5) $F := \{z \in \mathbb{H}, \mathrm{Im}\, z > 1\}$ is invariant under the group $\langle z \mapsto z+1 \rangle$ and embeds in $\mathbb{H}/\Gamma$ as a *cusp region*.

The *principal congruence subgroup of level* N is:

$$\Gamma(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \pm \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \mod N \right\}$$

Any (finite index) subgroup $G < \Gamma$ gives rise to a cover $\mathbb{H}/G \to \mathbb{H}/\Gamma$. If g is an automorphism of $\mathbb{H}$ then g induces an automorphism of $\mathbb{H}/G$ iff g normalises G

$$gGg^{-1} = G.$$

The *Hecke congruence subgroup of level* N is:

$$\Gamma_0(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \pm \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \mod N \right\}.$$

Evidently $\Gamma(N) < \Gamma_0(N) \trianglelefteq \Gamma$ and in fact Γ is the normaliser of the principal congruence subgroup but the normaliser $\Gamma_0^+(N)$ of the Hecke group is considerably more difficult to compute.

References

[1] Aigner M., Ziegler G.M. *Representing numbers as sums of two squares.* In: Proofs from the book. Springer, Berlin, Heidelberg. (2010)
[2] J. O. Button, *The uniqueness of the prime Markoff numbers*, J. London Math. Soc. (2) 58 (1998), 9–17.

[3] Ilke Canakci, Anna Felikson, Ana Garcia Elsener, Pavel Tumarkin, *Friezes for a pair of pants*, Proceedings of the Formal Power Series and Algebraic Combinatorics 2022 - Séminaire Lotharingien de Combinatoire
[4] John Horton Conway *The Sensual Quadratic Form* Carus Mathematical Monographs Volume: 26 1997
[5] D. A. Cox, Primes of the Forms $x^2 + ny^2$: Fermat, Class Field Theory, and Complex Multiplication, John Wiley, New York, 1989.
[6] Dolan, S., *A very simple proof of the two-squares theorem.* The Mathematical Gazette, 106(564), 511-511. (2021) doi:10.1017/mag.2021.120
[7] Guillaume Dubach, Fabian Muehlboeck Formal verification of Zagier's one-sentence proof `https://arxiv.org/abs/2103.11389`
[8] 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)
[9] S. Fomin and D. Thurston, *Cluster algebras and triangulated surfaces* Part II: Lambda lengths, `arXiv:1210.5569v1.`
[10] Lester R Ford, *Automorphic Functions*
[11] Generalov, A.I. *A combinatorial proof of Euler-Fermat's theorem on the representation of the primes p=8k+3 by the quadratic form $x^2 + 2y^2$.* J Math Sci 140, 690–691 (2007). https://doi.org/10.1007/s10958-007-0008-6
[12] Allen Hatcher, *Topology of Numbers* American Mathematical Society. 2022 `https://pi.math.cornell.edu/~hatcher/TN/TNbook.pdf`
[13] Heath-Brown, Roger. *Fermat's two squares theorem.* Invariant (1984)
[14] Neil Herriot, Communication with Jim Propp `https://faculty.uml.edu/jpropp/reach/Herriot/ptolemywriteup.html`
[15] Jackson, Terence H.. "A Short Proof That Every Prime $p = 3 \mod 8$ Is of the Form $x^2 + 2y^2$." The American Mathematical Monthly 107 (2000): 447 - 447.
[16] G. McShane, *Simple geodesics and a series constant over Teichmuller space* Invent. Math. (1998)
[17] Greg McShane, Vlad Sergiescu, *Geometry of Fermat's sum of squares*
[18] Northshield, Sam. *A Short Proof of Fermat's Two-square Theorem.* The American Mathematical Monthly. 127. 638-638. (2020).
[19] R. C. Penner, *The decorated Teichmueller space of punctured surfaces*, Communications in Mathematical Physics 113 (1987), 299–339.
[20] James Propp, The combinatorics of frieze patterns and Markoff numbers, in Integers, Volume 20 (2020) `http://math.colgate.edu/~integers/u12/u12.pdf`
[21] P. Sarnak *Class numbers of indfinite binary quadra tic forms* J. Number Theory, 15 (1982), pp. 229-247
[22] J-P. Serre, *A Course in Arithmetic*, Graduate Texts in Mathematics, Springer-Verlag New York 1973
[23] H. Shimizu, *On discontinuous groups operating on the product of half spaces*, Ann. of Math. (2)77(1963), 33-71.
[24] B. Springborn. The hyperbolic geometry of Markov's theorem on Diophantine approximation and quadratic forms. Enseign. Math., 63(3-4):333–373, 2017.
[25] Boris Springborn, *The worst approximable rational numbers* `https://arxiv.org/abs/2209.15542`
[26] 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

[27] Ying Zhang *Representing Primes as $x^2 + 5y^2$: An Inductive Proof that Euler Missed* https://arxiv.org/abs/math/0606547

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