hexagon:2610.00281v2

ALMOST ALL MARKOFF NUMBERS ARE UNIQUE

WILLIAM Y. CHEN

Abstract. A Markoff number is the largest coordinate of a positive integer solution of $a^2+b^2+c^2 = 3abc$. Frobenius conjectured in 1913 that every Markoff number is the largest coordinate of exactly one such solution with $a \leq b \leq c$. We prove that the conjecture holds for almost all Markoff numbers: the proportion of exceptions among the Markoff numbers up to X tends to 0 as $X \to \infty$.

Two such solutions with the same largest coordinate give simple closed geodesics of equal length on the modular torus, that is, primitive homology classes v, w with equal values $N(v) = N(w)$ of the McShane–Rivin stable norm N. Estimates for the boundary of the unit ball of N reduce the problem, up to an arbitrarily small proportion, to the equations $N(Gv) = N(v)$ for finitely many rational linear maps G. For each G that is not a symmetry of the torus, we use a zero-one law for the 3-adic analytic dynamics of the mapping class group on the Markoff surface to show that the solutions v have density zero.

Contents

1. Introduction 1
2. The modular torus, the stable norm, and Markoff triples 7
3. The boundary of the McShane–Rivin unit ball along Farey sectors 11
4. Equal-length geodesics under a fixed rational map 16
5. Collision pairs lie on rational graphs 24
6. Summation over the graphs and proof of the main theorem 28
7. Concluding remarks 32
References 33

1. Introduction

1.1. **Markoff numbers and the uniqueness conjecture.** A *Markoff triple* is a solution (a, b, c) in positive integers of the Markoff equation

$$a^2 + b^2 + c^2 = 3abc, \tag{1.1}$$

and a *Markoff number* is a coordinate of a Markoff triple, or equivalently, by the tree structure recalled below, the largest coordinate of some Markoff triple. The first few Markoff numbers are $1, 2, 5, 13, 29, 34, 89, 169, 194, 233, 433, 610, 985$.

Markoff introduced the equation in his work on indefinite binary quadratic forms [Mar79, Mar80]. Below 3, which is its first accumulation point, the Markoff spectrum of these forms consists exactly of the numbers $\sqrt{9 - 4/m^2}$ with m a Markoff number, and so does the Lagrange spectrum

Date: October 14, 2026.

of Diophantine approximation of irrational numbers. The forms attaining these values are parametrized by the Markoff triples.

The Markoff equation is preserved by the permutations of the coordinates and by the *Vieta involution* $(a, b, c) \mapsto (a, b, 3ab - c)$, which replaces c by the other root of (1.1) viewed as a quadratic equation in c. Every Markoff triple arises from $(1, 1, 1)$ by these moves. The sorted triples $a \le b \le c$ other than $(1, 1, 1)$ and $(1, 1, 2)$ form a binary tree rooted at $(1, 2, 5)$, in which each triple (a, b, c) has the two children $(a, c, 3ac-b)$ and $(b, c, 3bc-a)$; see Cassels [Cas57, Chapter II] or Aigner [Aig13].

Frobenius [Fro13] asked in 1913 whether a Markoff triple is determined by its largest coordinate.

> **Uniqueness conjecture.** If (a, b, c) and (a', b', c) are Markoff triples with $a \le b \le c$ and $a' \le b' \le c$, then $(a, b) = (a', b')$.

A Markoff number c for which the conclusion holds is called *unique*. The conjecture has been verified numerically for all Markoff numbers below 10^{140} (Baragar [Bar96, §4], extending the search of Borosh [Bor75] up to 10^{105}; see [Aig13] for the history),[1] and it is known for Markoff numbers of special arithmetic shapes:

Button proved uniqueness when c is a prime [But98], and Schmutz [Sch96] and Button [But01] when c is a prime power. Lang and Tan [LT07] gave a short geometric proof of the prime power case, and Srinivasan [Sri09] an elementary proof when c is an odd prime power or twice an odd prime power. Baragar [Bar96] proved uniqueness when c, $3c-2$ or $3c+2$ is a prime, twice a prime or four times a prime, and Zhang [Zha07] when $3c+2$ or $3c-2$ is a prime power, four times a prime power or eight times a prime power. Srinivasan also gave, for odd c, a sufficient condition in terms of ambiguous classes of forms of discriminant $9c^2 - 4$. All these results establish uniqueness under arithmetic conditions on the factorization of c or of $9c^2 - 4 = (3c-2)(3c+2)$. Tao and Yang [TY25] proved unconditionally a consequence of the conjecture: the Markoff equation is not partition regular.

Bourgain, Gamburd and Sarnak [BGS16, BGS26] showed that almost all Markoff numbers, counted with multiplicity, are composite. More precisely, for every ν, the proportion of sorted Markoff triples (a, b, c) with $c \le X$ for which c has at most ν distinct prime factors tends to 0 as $X \to \infty$ [BGS26, Theorem 19].

The Markoff numbers are sparse. Zagier [Zag82] proved that the number of sorted Markoff triples with largest coordinate at most X is

$$C_0 \log^2(3X) + O\big(\log X\,(\log\log X)^2\big), \qquad C_0 = 0.1807\ldots.$$

Gamburd, Magee and Ronan [GMR19, Theorem 3] extended such asymptotic counts to the Markoff–Hurwitz varieties $x_1^2 + \cdots + x_n^2 = ax_1 \cdots x_n + k$ with $n \ge 3$. The uniqueness conjecture would imply that the number of Markoff numbers up to X equals this count of triples for every X. In this paper we prove the asymptotic consequence: the two counting functions have the same leading asymptotic.

1.2. **The main result.** For $X \ge 1$ let

$$\begin{aligned} T(X) &= \#\{(a, b, c) \in \mathbb{Z}^3_{>0} :\ a \le b \le c \le X,\ a^2 + b^2 + c^2 = 3abc\}, \\ D(X) &= \#\{c \le X :\ c \text{ is a Markoff number}\}, \\ D_2(X) &= \#\{c \le X :\ c \text{ is the largest coordinate of at least two sorted Markoff triples}\}. \end{aligned}$$

[1] Perling [Per22, Appendix F] reports a computer check of all Markoff numbers below 10^{15000}.

Thus $T(X) \geq D(X)$, with equality for all X if and only if the uniqueness conjecture holds, and $D_2(X)$ counts the non-unique Markoff numbers up to X.

Theorem 1.1. *As $X \to \infty$,*

$$T(X) - D(X) = o(\log^2 X) \qquad \textit{and consequently} \qquad \frac{D_2(X)}{D(X)} \longrightarrow 0.$$

In words: the proportion of non-unique Markoff numbers among all Markoff numbers up to X tends to zero, and the number of Markoff numbers up to X is asymptotic to $C_0 \log^2(3X)$ with Zagier's constant C_0. The proof gives no rate of decay for $D_2(X)/D(X)$; the uniqueness conjecture is the statement that $D_2(X) = 0$ for all X.

By Theorem 1.1, $T(X) \sim D(X)$, so the theorem of Bourgain, Gamburd and Sarnak quoted above holds also for Markoff numbers counted without multiplicity: for every ν, the Markoff numbers with at most ν distinct prime factors form a set of density zero among all Markoff numbers. This applies in particular to the prime powers and twice prime powers covered by the criteria of Button, Schmutz and Srinivasan.

1.3. **The geometric picture.** The proof uses the interpretation of Markoff numbers through hyperbolic geometry, which goes back to Cohn [Coh55, Coh71] and was developed by Series [Ser85]. Let Γ' be the commutator subgroup of $\mathrm{PSL}_2(\mathbb{Z})$. It is free of rank two and of index six, and the quotient $\mathbb{T} = \mathbb{H}/\Gamma'$ of the upper half-plane is a once-punctured torus, the *modular torus* with a hyperbolic metric. Its fundamental group is the free group F_2, and a primitive element of F_2 (a member of some free basis) determines, up to conjugacy and inversion, a simple closed geodesic on $\mathbb{T}$ [MR95a, Facts 2.2, 2.3 and Lemmas 2.4, 2.5] and a primitive[2] vector $v = (p, q) \in \mathbb{Z}^2$ of the homology group $H_1(\mathbb{T}, \mathbb{Z}) \cong \mathbb{Z}^2$, defined up to sign.

Cohn observed that the traces of the hyperbolic elements of Γ' representing simple closed geodesics are exactly the numbers $3m$ with m a Markoff number, and that the simple closed geodesics represented by the two elements of a free basis and by their product have traces $(3a, 3b, 3c)$ with (a, b, c) a Markoff triple.

In §2 we make this dictionary precise in the form we need. We write $m(v)$ for the Markoff number attached to a primitive vector v, the *label* of v, so that the corresponding geodesic has length

$$N(v) = 2 \operatorname{arcosh}\bigl(3m(v)/2\bigr). \tag{1.2}$$

Let $\mathcal{S} = \{(p, q) \in \mathbb{R}^2 : \ 0 \leq 2p \leq q\}$, the closed sector of vectors of slope p/q between 0 and $1/2$. The isometry group of $\mathbb{T}$ acts on $H_1(\mathbb{T}, \mathbb{Z}) \cong \mathbb{Z}^2$, and hence linearly on $\mathbb{R}^2$, through a dihedral group $\mathcal{K}$ of order twelve (§4.5). The sector $\mathcal{S}$ is a fundamental domain for the action on $\mathbb{R}^2$, and every $\mathcal{K}$-orbit of primitive vectors meets $\mathcal{S}$ in exactly one vector. The sorted Markoff triples correspond bijectively to the primitive vectors of $\mathcal{S}$, and under this bijection the largest coordinate of a triple is the label of the vector (Lemma 2.1).

Consequently, the uniqueness conjecture is equivalent to the statement that two simple closed geodesics of the modular torus of the same length differ by an isometry. Theorem 1.1 says that the number of simple closed geodesics of length $\leq L$, counted modulo isometry, exceeds the number of distinct lengths $\leq L$ by only $o(L^2)$. In particular, the exceptional lengths, those attained by two geodesics that are related by no isometry, form a set of density zero among all lengths. McShane and Parlier [MP08, §7] study this geometric form of the conjecture through the loci in Teichmüller

[2] A vector of $\mathbb{Z}^2$ is *primitive* if it is nonzero and its coordinates are coprime, that is, if it belongs to a basis of $\mathbb{Z}^2$. The vector determined by a primitive element of F_2 is its homology class, which is primitive because the abelianization $F_2 \to H_1(\mathbb{T}, \mathbb{Z})$ maps free bases to bases.

space where two simple closed geodesics have equal length, and Fisac [Fis25] restates the conjecture combinatorially by describing the simple length spectrum of $\mathbb{T}$ explicitly.

McShane and Rivin [MR95a, MR95b] proved that the length function $v \mapsto N(v)$ on primitive vectors extends to a norm N on $\mathbb{R}^2$, the *stable norm* of $\mathbb{T}$, whose unit ball $\mathbb{B}$ is strictly convex. The number of primitive vectors v with $N(v) \le L$, that is, of primitive lattice points in $L\mathbb{B}$, is then asymptotic to $\operatorname{area}(\mathbb{B})L^2/\zeta(2)$ [MR95a, Theorem 0.1 and §5]; the constant comes from Möbius inversion, as in the proof of Lemma 2.1(d). Hence the number of simple closed geodesics of length at most L, each corresponding to a pair $\pm v$, is asymptotic to $\operatorname{area}(\mathbb{B})L^2/(2\zeta(2))$.

In this language Theorem 1.1 says that, among the primitive vectors $v \in \mathcal{S}$ with $N(v) \le L$, whose number is asymptotic to a constant times L^2, all but $o(L^2)$ have a norm $N(v)$ shared by no other primitive vector of $\mathcal{S}$.

The boundary of $\mathbb{B}$ has a corner at every rational direction, with exterior angle decaying exponentially in the height $\max(|p|, |q|)$ of the primitive vector (p, q) in that direction, and it is differentiable at every irrational direction. McShane and Rivin stated both facts, with a sketch of proof [MR95b, Theorem 2.1]. Sorrentino and Veselov [SV19, Theorem 2] proved that $\partial\mathbb{B}$ is differentiable exactly at the irrational directions, in terms of a convex function introduced by Fock [Foc97, §7.3] that parametrizes an arc of $\partial\mathbb{B}$. Doan, Li and Nguyen [DLN26, Theorem 1.2] proved that the exterior angle at a primitive vector $v = (p, q)$ lies between two constant multiples of $\max(|p|, |q|)\, e^{-N(v)}$. For closed surfaces, the unit ball of the stable norm near rational directions was described by Massart [Mas97].

The two one-sided slopes at each corner, equivalently the one-sided derivatives of Fock's function at rational points, were computed by Hines [Hin20] and Gaster [Gas22, Propositions 1.8 and 1.10], answering a question of Sorrentino and Veselov [SV19, §5]. Spalding and Veselov [SV17, §7] relate Fock's function to the growth of Markoff numbers along the paths of the tree, and O'Sullivan [O'S26] extends the function and its one-sided derivatives to the equations $x^2 + y^2 + z^2 - xyz = D$ for all real D. Proposition 3.1(a) below gives the size of each corner in closed form; Doan, Li and Nguyen [DLN26] obtain the same formula (Remark 3.2).

Doan, Li and Nguyen [DLN26] also analyzed the local geometry at irrational directions, and used estimates on the turning of the normal of $\partial\mathbb{B}$ to show that at most $C(\log\log 3m)^2$ simple closed geodesics have length $2\operatorname{arcosh}(3m/2)$; hence each Markoff number m is the largest coordinate of at most $C(\log\log 3m)^2$ sorted triples. We use an estimate of the same nature, Proposition 3.1, to confine all but $o(L^2)$ of the pairs of lattice points of norm at most L at which N takes equal values, outside a sparse set of directions, to the graphs of $(\log L)^{O(1)}$ rational linear maps (Stage 1 in §1.4).

Since $m(v)$ increases with $N(v)$, the stable norm also orders the Markoff numbers. Aigner's three conjectures on this ordering [Aig13] were proved by Rabideau and Schiffler [RS20] (fixed numerator) and, independently, by Lagisquet, Pelantová, Tavenas and Vuillon [LPTV21] and by Lee, Li, Rabideau and Schiffler [LLRS23] (fixed denominator and fixed sum). McShane [McS21] derived all three from the convexity of N, and Gaster [Gas22] determined the slopes of the lines along which the ordering is monotone.

1.4. **Strategy of the proof.** Throughout, L is a large length; in Stages 1 and 3 below, all lattice points v considered lie in the sector $\mathcal{S}$ and satisfy $N(v) \le L$. A *collision pair* is an ordered pair (v, w) of distinct primitive vectors in $\mathcal{S}$ with $N(v) = N(w) \le L$. Theorem 1.1 follows (§6.6) once we show that for every $\varepsilon > 0$ and all large L, the primitive vectors of $\mathcal{S}$ with $N \le L$ can be partitioned into at most εL^2 *deleted* vectors and a set of *retained* vectors among which there are at most εL^2 collision pairs. The deleted set is determined by the geometry of the unit ball alone:

it consists of the lattice points on finitely many rays and in a sparse set of angular sectors. The argument has three stages.

Stage 1: collision pairs lie on rational graphs. Let u, v be *Farey neighbours*, that is, primitive vectors with $|\det(u, v)| = 1$. The vectors u, v and $u + v$ carry the three labels of a Markoff triple, and the Vieta involution of the triple becomes the recurrence

$$m(u + v) + m(u - v) = 3\, m(u)\, m(v). \tag{1.3}$$

In §3 we exploit this recurrence to describe the boundary of $\mathbb{B}$ between two consecutive rays of the *Farey fan of level* H, the set of rays from the origin through the primitive vectors of height $\max(|p|, |q|) \le H$. On each open sector of this fan the supporting lines of $\mathbb{B}$ vary by at most $Ce^{-\kappa H}$, while at the rays themselves $\partial\mathbb{B}$ has corners of exterior angle at least ce^{-CH} (Proposition 3.1).

The decay comes from the label of the *mediant* $u + v$ of the two boundary vectors u, v of such a sector. Substituting $m(u)$ and $m(v)$ for the first two variables of the Markoff equation (1.1) leaves a quadratic equation in the third. Its roots are $m(u + v)$ and, by the recurrence (1.3), $m(u - v)$, and the mediant carries the *larger* one. By the length formula (1.2), $3m(w) = 2\cosh(N(w)/2)$ for every primitive vector w, so the formula for the larger root gives

$$3m(u + v) = 2\cosh\big((N(u) + N(v))/2\big) + O\big(e^{-(N(u)+N(v))/2}\big).$$

Hence $N(u + v) = N(u) + N(v) + O(e^{-N(u)-N(v)})$: up to an exponentially small error, the norm is additive across the sector. This near additivity is what makes $\partial\mathbb{B}$ nearly flat there.

Fix a pair of sectors Σ, Σ' of the fan of level $H \asymp \log L$. Let λ, λ' be supporting functionals of N at points of Σ, Σ', that is, linear forms $\le N$ that agree with N at these points (§3.1), and let $\ell = (\lambda, -\lambda')$ be the corresponding linear form on $\mathbb{R}^4$. For a collision pair $(v, w) \in \Sigma \times \Sigma'$, its *collision vector* $(v, w) \in \mathbb{Z}^4$ satisfies $|\ell(v, w)| \le CLe^{-\kappa H}$, an exponentially small quantity. Four linearly independent collision vectors in $\Sigma \times \Sigma'$ would have an integer determinant of absolute value at least 1, whereas expanding the determinant along ℓ bounds it by $CL^4e^{-\kappa H} < 1$ once the constant in $H \asymp \log L$ is large.

If instead the collision vectors span a hyperplane $W \subset \mathbb{R}^4$, then its primitive integral normal $(\alpha, \beta) \in \mathbb{Z}^2 \times \mathbb{Z}^2$ has size $O(L^3)$, and since ℓ is nearly zero on W, this normal is nearly parallel to ℓ. Hence the supporting functionals of N at points of Σ' are nearly parallel to the integral form β. To rule this out, we excise beforehand the sectors containing the contact points of the integral forms of size $O(L^3)$, the points of $\partial\mathbb{B}$ at which such a form attains its maximum on $\mathbb{B}$. These contact points lie on $O(\log^2 L)$ rays (Lemma 3.3). In each remaining sector, the corners of $\partial\mathbb{B}$ at the two boundary rays keep the supporting functionals away from every such form, by an angle that is exponentially small but larger than the one forced by W once Σ, Σ' are taken from a finer fan than the excised sectors (Lemma 3.5).

The outcome (Lemma 5.1) is that, after deleting $O(L^2/A_0)$ lattice points for a large parameter A_0, the collision vectors of the remaining collision pairs in every sector pair span a subspace of $\mathbb{R}^4$ of dimension at most two, so that, apart from negligible degenerate cases, they satisfy $w = Gv$ for a single matrix $G = G_{\Sigma,\Sigma'} \in \mathrm{GL}_2(\mathbb{Q})$.

We call $\{(v, Gv)\}$ the *graph* of G, and we measure the complexity of G by the index $I(G)$, the maximum of the indices in $\mathbb{Z}^2$ of the lattice $\mathbb{Z}^2 \cap G^{-1}\mathbb{Z}^2$ and of its image under G, and by its entry height $h(G)$, the largest absolute value of its entries. Two elementary lemmas (Lemmas 5.2 and 5.3) show that graphs of index larger than $(\log L)^5$, or of entry height larger than $10(1 + |\det G|)$, carry only $o_{A_0}(L^2)$ collision pairs in total.

Stage 2: a fixed rational map produces few equalities. Stage 1 leaves a family of rational matrices $G \notin \mathcal{K}$ with $I(G) \le (\log L)^5$ and $h(G) \le 10(1 + |\det G|)$, and it remains to count the primitive v with Gv primitive and $N(Gv) = N(v) \le L$ on each graph. For a single fixed G we prove in §4 that there are only $o_G(L^2)$ such v (Proposition 4.1). This is the arithmetic heart of the paper, and it is where the specific structure of the Markoff equation enters.

The equation $N(Gv) = N(v)$ is equivalent to $m(Gv) = m(v)$ by the length formula (1.2). We work on the set

$$\mathcal{X} = \{(a, b, c) \in (\mathbb{Z}_3^\times)^3 : \ a^2 + b^2 + c^2 = 3abc\}$$

of 3-adic unit points of the Markoff surface, on which $\mathrm{GL}_2(\mathbb{Z}) \cong \mathrm{Out}(F_2)$, the extended mapping class group of the torus, acts by the Vieta involutions and the coordinate permutations (the *Nielsen moves*, defined in §4). The label $m(v)$ of a primitive vector v is the value at the base point $o = (1, 1, 1)$ of a trace polynomial Q_v with integer coefficients, and the action satisfies $Q_v(P * \gamma) = Q_{\gamma v}(P)$ for $\gamma \in \mathrm{SL}_2(\mathbb{Z})$.

For h in the finite-index subgroup $H_G = \mathrm{SL}_2(\mathbb{Z}) \cap G^{-1}\mathrm{SL}_2(\mathbb{Z})G$, both h and GhG^{-1} are integral, and for $v, Gv \in \mathbb{Z}^2_{\mathrm{prim}}$ the equality $m(hv) = m(Ghv)$ is the vanishing of the single polynomial

$$F(P, P') = Q_v(P) - Q_{Gv}(P') \qquad \text{at the point} \qquad (o * h,\, o * GhG^{-1}) \in \mathcal{X}^2.$$

The set of such points, as h ranges over a coset of the finite-index subgroup J of H_G that acts trivially on $\mathcal{X}^2$ modulo 9 (§4.7), has a compact closure $Y \subset \mathcal{X}^2$. It carries the J-invariant probability measure μ_Y that is the limit of the uniform measures on the reductions of Y modulo 3^n, whose elements we call *states*.

The key point is a zero-one law: $\mu_Y(F = 0)$ is either 0 or 1 (Lemma 4.3). It rests on p-adic analytic interpolation in the sense of Poonen [Poo14]. Since J acts trivially on $\mathcal{X}^2$ modulo 9, for each $g \in J$ the iterates g^n extend to a $\mathbb{Z}_3$-action $t \mapsto g^t$ given by a single restricted power series in t, that is, one with coefficients in $\mathbb{Z}_3$ tending to 0 (§4.2 and Lemma 4.4). By Strassmann's theorem, along each orbit of this action F either vanishes identically or has finitely many zeros, and the ergodicity of μ_Y turns this into the dichotomy.

The alternative $\mu_Y(F = 0) = 1$ means $N(Gv') = N(v')$ for all v' in the corresponding orbit of v, whose directions are dense, so G would be a linear isometry of the norm N. A rigidity lemma (Lemma 4.5), based on the explicit size of the corners of $\mathbb{B}$, shows that the only rational isometries of N are the twelve elements of $\mathcal{K}$. Hence $\mu_Y(F = 0) = 0$.

Finally, to pass from measure zero on Y to the count of lattice points of norm at most L, we use the asymptotic count, due to Burrin, Nevo, Rühr and Weiss [BNRW20], of the points of an orbit Uv of a finite-index subgroup $U \le \mathrm{SL}_2(\mathbb{Z})$ in the dilates $L\mathbb{B}$. Its leading coefficient, which we obtain from their argument (Theorem 4.7), is proportional to the *cusp width* $[\mathrm{Stab}_{\mathrm{SL}_2(\mathbb{Z})}(v) : \mathrm{Stab}_U(v)]$ divided by the index $[\mathrm{SL}_2(\mathbb{Z}) : U]$.

The J-orbit of v decomposes into sub-orbits indexed by the cycles of the parabolic stabilizer of v on the finite set of states of Y modulo 3^n, and the leading coefficient of a sub-orbit is proportional to the length of its cycle. The equalities occur only in the sub-orbits whose cycles consist of states with $F \equiv 0 \pmod{3^n}$, and summing over these cycles gives the proportion $\mu_Y(F \equiv 0 \bmod 3^n)$, which tends to zero. The count is applied to conjugates of the stabilizers of states modulo 3^n, which need not be congruence subgroups, so it is needed for arbitrary subgroups of finite index.

Stage 3: summation over the graphs. Proposition 4.1 has no uniformity in G, whereas Stage 1 produces up to $(\log L)^{O(1)}$ graphs. Section 6 resolves this by sorting the graphs into dyadic bands $2^j < I(G) \le 2^{j+1}$. For a large integer j_0, we apply Proposition 4.1 to the finitely many graphs with $I(G) \le 2^{j_0}$.

For band j we run the geometric argument of Stage 1 again, with the fan of level j^3 in place of the level $\asymp \log L$, and we excise at once the sectors containing the contact points of all integral forms of size $2^{O(j)}$; the deleted angular widths have sum $O(1/j_0)$ over $j \geq j_0$. In each pair of sectors of a fan of level $\asymp j^3$ finer than the excised sectors, the collision pairs lying on one graph of band j form a *band piece*, which is *large* if it has at least $L^2/2^{20j}$ elements. At most one graph of band j has a large band piece in a given sector pair (Lemma 6.3), because two distinct rational graphs of bounded complexity cannot both be exponentially close to the hyperplane $\ell = 0$.

There are $O(j^{12})$ sector pairs, so the large band pieces of band j contribute $O(L^2 j^{12} 2^{-j})$ by Lemma 5.2, while the $O(2^{9j} j^{12})$ possible small band pieces contribute $O(L^2 j^{12} 2^{-11j})$. The total contribution of band j is therefore summable over j. Letting $L \to \infty$, then $j_0 \to \infty$, then $A_0 \to \infty$ proves Theorem 1.1.

1.5. **Acknowledgements and statement on AI use.** This project was primarily completed using ChatGPT Astra via Codex, and Claude Fable 5.1, using skills developed by the author and Taylor Dupuy. The author initiated the project and gave the model some of his ideas for approaching the problem, but the argument as it appears here is largely due to Astra.

1.6. **Organization and conventions.** Section 2 sets up the modular torus, the stable norm and the dictionary between Markoff triples and lattice points, and fixes the Farey terminology. Section 3 proves the geometric estimates on $\partial\mathbb{B}$. Section 4 carries out Stage 2 (Proposition 4.1); it uses only §2 and 3. Section 5 carries out Stage 1 and §6 Stage 3, ending with the proof of Theorem 1.1. Section 7 discusses quantitative refinements, the role of the prime 3, and the relation to strong approximation.

The letters c, C, κ denote positive constants that depend only on the stable norm N (hence are absolute); their values may change from line to line, and C is large while c, κ are small. Constants depending on further parameters carry those parameters as subscripts, as in C_{A_0} or $o_G(L^2)$. We write $f \ll g$ or $f = O(g)$ for $|f| \leq Cg$, and $f \asymp g$ for $f \ll g \ll f$. The Euclidean norm on $\mathbb{R}^2$ and $\mathbb{R}^4$ is $|\cdot|$, and $|\alpha|$ is the dual Euclidean norm of a linear form. A vector of $\mathbb{Z}^2$ is *primitive* if it is nonzero and its coordinates are coprime; $\mathbb{Z}^2_{\text{prim}}$ is the set of primitive vectors. The *height* of $v = (p, q) \in \mathbb{Z}^2$ is $\operatorname{ht}(v) = \max(|p|, |q|)$. A *direction* is a point of the projective line $\mathbb{RP}^1$, that is, a line through the origin of $\mathbb{R}^2$; a *ray* is a half-line from the origin. We write $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ and $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

2. The modular torus, the stable norm, and Markoff triples

2.1. **The modular torus.** Let $F_2 = \langle x, y \rangle$ be the free group on two generators and let $\rho : F_2 \to \mathrm{SL}_2(\mathbb{Z})$ be the representation defined by

$$A := \rho(x) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad B := \rho(y) = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}.$$

Then $\operatorname{tr} A = \operatorname{tr} B = \operatorname{tr} AB = 3$ and $[A, B] = ABA^{-1}B^{-1} = -T^6$. Let T' be the transpose of T. Then $A = TT'$, $B = T^{-1}T'^{-1}$, $TAT^{-1} = AB$ and $TBT^{-1} = A^{-1}$. Since $\mathrm{PSL}_2(\mathbb{Z})$ is generated by the images of T and T', and $T' = T^{-1}A$, it is generated by the images of T, A and B. As the subgroup generated by A and B is normalized by T, its image Γ' in $\mathrm{PSL}_2(\mathbb{Z})$ is therefore normal.

The group Γ' contains the commutator subgroup of $\mathrm{PSL}_2(\mathbb{Z})$, because $\mathrm{PSL}_2(\mathbb{Z})/\Gamma'$ is generated by the image of T and hence is cyclic. It is contained in the commutator subgroup, because A and B map to the identity in the abelianization $\mathbb{Z}/6$ of $\mathrm{PSL}_2(\mathbb{Z})$: T' is conjugate to T^{-1} in $\mathrm{PSL}_2(\mathbb{Z})$, so the images of T and T' in $\mathbb{Z}/6$ are inverse to each other, and $A = TT'$, $B = T^{-1}T'^{-1}$.

Thus Γ' is the commutator subgroup of $\mathrm{PSL}_2(\mathbb{Z})$, of index six. It is a free group of rank two, torsion free, with one cusp, and the quotient $\mathbb{T} = \mathbb{H}/\Gamma'$ is a complete hyperbolic once-punctured torus, the *modular torus*. The representation ρ identifies F_2 with $\pi_1(\mathbb{T})$ and induces the isomorphism $H_1(\mathbb{T}, \mathbb{Z}) = F_2^{\mathrm{ab}} \cong \mathbb{Z}^2$ sending x, y to the standard basis vectors e_1, e_2.

An element of F_2 is *primitive* if it belongs to some free basis of F_2; its class in $H_1(\mathbb{T}, \mathbb{Z}) = \mathbb{Z}^2$ is then a primitive vector. We use the following classical facts from McShane and Rivin [MR95a].

(F1) ([MR95a, Fact 2.2]) The map sending a primitive element to its homology class induces a bijection between the set of conjugacy classes of primitive elements of F_2, modulo inversion, and the set $\mathbb{Z}^2_{\mathrm{prim}}/\{\pm 1\}$ of primitive vectors modulo sign. For a primitive vector v we write w_v for a primitive element of F_2 in the corresponding class; it is well defined up to conjugacy and inversion.

(F2) ([MR95a, Facts 2.2, 2.3 and Lemmas 2.4, 2.5]) The conjugacy classes of primitive elements, modulo inversion, correspond bijectively to the free homotopy classes of essential, non-peripheral, unoriented simple closed curves on $\mathbb{T}$, and each such class contains a unique closed geodesic. The geodesic in the class of w_v has length $N(v) := 2\,\mathrm{arcosh}\big(|\operatorname{tr}\rho(w_v)|/2\big)$.

(F3) ([MR95a, §5]) The function $v \mapsto N(v)$ on $\mathbb{Z}^2_{\mathrm{prim}}$ extends uniquely to a norm N on $\mathbb{R}^2$, by homogeneity on rational vectors and continuity. For independent integral vectors h, g one has the strict inequality $N(h+g) < N(h) + N(g)$ [MR95a, (5.2)].

Following [DLN26], we call N the *stable norm* of the modular torus, and $\mathbb{B} = \{x \in \mathbb{R}^2 : N(x) \le 1\}$ its unit ball. The strict inequality in (F3) implies that $\mathbb{B}$ is *strictly convex*, that is, its boundary $\partial\mathbb{B}$ contains no line segment: a segment in $\partial\mathbb{B}$ would span a two-dimensional cone on which N is linear, and this cone contains two independent integral vectors h, g, for which then $N(h+g) = N(h)+N(g)$. Since N is a norm, there are constants with $c|x| \le N(x) \le C|x|$.

Two primitive vectors u, v with $\det(u, v) = \pm 1$ are called *Farey neighbours*. If u, v are Farey neighbours, then w_u, w_v can be chosen to form a free basis of F_2 [MR95a, Theorem 2.6]; then $w_u w_v$ is a primitive element with homology class $u + v$, and by (F1) we may take $w_{u+v} = w_u w_v$. For the holonomy matrices $U = \rho(w_u)$, $V = \rho(w_v)$ the trace identity $\operatorname{tr}(UV) + \operatorname{tr}(U^{-1}V) = \operatorname{tr} U \operatorname{tr} V$ and the Fricke identity

$$(\operatorname{tr} U)^2 + (\operatorname{tr} V)^2 + (\operatorname{tr} UV)^2 = \operatorname{tr} U \ \operatorname{tr} V \ \operatorname{tr} UV + \operatorname{tr}[U, V] + 2 \tag{2.1}$$

hold. Since $[U, V]$ is conjugate to $[A, B]^{\pm 1} = -T^{\pm 6}$ (by Nielsen's theorem [MKS76], the commutator of a free basis of F_2 is conjugate to $[x, y]$ or its inverse), $\operatorname{tr}[U, V] = -2$.

2.2. **Labels.** For a primitive vector v define its *label*

$$m(v) := \tfrac{1}{3} \operatorname{tr} \rho(w_v) \in \tfrac{1}{3}\mathbb{Z}.$$

We shall see in Lemma 2.1 that $m(v)$ is a positive integer, in fact a Markoff number, so that the length formula (1.2) is a restatement of (F2).

For Farey neighbours u, v, applying the trace identity to U, V gives

$$m(u+v) + m(u-v) = 3\, m(u)\, m(v), \tag{2.2}$$

and (2.1) with $\operatorname{tr}[U, V] = -2$ gives the *Markoff relation*

$$m(u)^2 + m(v)^2 + m(u+v)^2 = 3\, m(u)\, m(v)\, m(u+v). \tag{2.3}$$

Let

$$r = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}, \qquad s = \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathcal{K} = \langle r, s \rangle \subset \mathrm{GL}_2(\mathbb{Z}),$$

and let $\mathcal{S}$ be the closed sector

$$\mathcal{S} = \{(p,q) \in \mathbb{R}^2 : \ 0 \le 2p \le q\},$$

consisting of the origin and the vectors of slope $p/q \in [0, 1/2]$ in the upper half-plane. By Lemma 2.1 below, $\mathcal{K}$ preserves labels and the norm N, and every $\mathcal{K}$-orbit of primitive vectors meets $\mathcal{S}$ in exactly one vector. A *sorted* Markoff triple is a Markoff triple (a, b, c) with $a \le b \le c$.

2.3. **Farey neighbours, the Farey tree, and Farey fans.** We collect the elementary facts about Farey neighbours that are used throughout; all of them are classical [HW08, Chapter III]. Recall that primitive vectors u, v are Farey neighbours if $\det(u, v) = \pm 1$; then $u + v$ is primitive, and it is a Farey neighbour of both u and v. We call $u + v$ the *mediant* of u and v.

The Farey tree of the sector. The boundary vectors $e_2 = (0, 1)$ and $f := (1, 2)$ of $\mathcal{S}$ are Farey neighbours. Starting from the pair (e_2, f) and repeatedly inserting mediants produces a binary tree of primitive vectors of $\mathcal{S}$: the root is $e_2 + f = (1, 3)$, and a node v, obtained as the mediant of two Farey neighbours u and u' (its *Farey parents*, with u of smaller slope than u'), has the two children $u + v$ (with parents u, v) and $v + u'$ (with parents v, u').

Every primitive vector v of $\mathcal{S}$ with $v \ne e_2, f$ occurs exactly once in this tree. Indeed $v = (p, q)$ with $0 < 2p < q$ can be written $v = \alpha e_2 + \beta f$ with $\alpha = q - 2p > 0$ and $\beta = p > 0$ coprime, and the subtractive Euclidean algorithm applied to (α, β) traces the unique path from the root to v.

Moreover, if v is not the root, then one of its parents is the mediant of the other parent and of a third vector u'', and then $u - u' = \pm u''$; u'' is an earlier node of the tree or one of e_2, f. For the root, $u - u' = e_2 - f = -(1, 1)$.

Farey fans. For an integer $H \ge 1$, the *Farey fan of level H* is the finite set $\mathcal{F}_H$ of rays $\mathbb{R}_{>0}v$ through the primitive vectors v with $\mathrm{ht}(v) \le H$. The connected components of $\mathbb{R}^2 \setminus (\{0\} \cup \bigcup \mathcal{F}_H)$ are open convex cones, which we call the *Farey sectors of level H*. The two primitive vectors u, v of height at most H spanning the boundary rays of a sector are Farey neighbours, and their mediant has height $> H$ (otherwise its ray would lie inside the sector).

In the octant $0 \le p \le q$ the rays of $\mathcal{F}_H$ are the rays of slope p/q with p/q in the Farey sequence of order H, and a sector of level H with boundary slopes $a/b < c/d$ ($bc - ad = 1$, $b, d \le H$, $b + d > H$) has slope width $1/(bd)$. The other octants are obtained by the symmetries $(p, q) \mapsto (\pm p, \pm q), (q, p)$.

The number of sectors of level H is $\asymp H^2$, since the number of primitive vectors of height at most H is $\asymp H^2$. The fan $\mathcal{F}_1$ consists of the eight rays through $\pm e_1, \pm e_2, \pm(e_1 + e_2), \pm(e_1 - e_2)$, the primitive vectors of height one; the first six of these vectors have label 1 and the last two have label 2. The fans are nested: $\mathcal{F}_H \subset \mathcal{F}_{H'}$ for $H \le H'$, so every sector of level H' is contained in a sector of level H.

2.4. **The dictionary.**

Lemma 2.1. *The following statements hold.*

(a) The group $\mathcal{K}$ is dihedral of order twelve, and every $k \in \mathcal{K}$ preserves the label and the norm: $m(kv) = m(v)$ for $v \in \mathbb{Z}^2_{prim}$ and $N(kx) = N(x)$ for $x \in \mathbb{R}^2$.

(b) Every $\mathcal{K}$-orbit in $\mathbb{Z}^2_{prim}$ meets $\mathcal{S}$ in exactly one point.

(c) The map sending $(0, 1)$ to $(1, 1, 1)$, $(1, 2)$ to $(1, 1, 2)$, and every other primitive vector $v \in \mathcal{S}$ to the sorted triple formed by the labels of v and of its two Farey parents, is a bijection from $\mathbb{Z}^2_{prim} \cap \mathcal{S}$ onto the set of sorted Markoff triples. Under this bijection the largest coordinate of the triple is $m(v)$. In particular $m(v)$ is a positive integer for every primitive v, and the length formula (1.2) holds.

(d) As $L \to \infty$,

$$\#\{v \in \mathbb{Z}^2_{prim} \cap \mathcal{S} : \ N(v) \le L\} \sim c_N L^2 \tag{2.4}$$

for a constant $c_N > 0$.

Proof. (a) One checks $r^3 = -I$, $s^2 = I$ and $srs = r^{-1}$, so $\mathcal{K}$ is dihedral of order twelve. The substitutions $(x, y) \mapsto (y^{-1}, xy)$ and $(x, y) \mapsto (x^{-1}, xy)$ define automorphisms ϕ_1, ϕ_2 of F_2, and $\rho \circ \phi_i$ has the same traces $(3, 3, 3)$ as ρ on x, y, xy.

The representation ρ is irreducible (A and B have no common eigenvector), and an irreducible representation of F_2 in SL_2 is determined up to conjugacy by the traces of x, y, xy (the classical theorem of Fricke and Vogt; see [Gol09, Theorem A]).

Hence $\rho \circ \phi_i$ is conjugate to ρ, and $\operatorname{tr} \rho(\phi_i(w)) = \operatorname{tr} \rho(w)$ for every $w \in F_2$. The automorphism ϕ_i sends w_v to a primitive element with homology class $k_i v$, where $k_1 = r$ and $k_2 = s$ are the matrices induced on $H_1 = \mathbb{Z}^2$ (ϕ_1 induces $e_1 \mapsto -e_2$, $e_2 \mapsto e_1 + e_2$, and ϕ_2 induces $e_1 \mapsto -e_1$, $e_2 \mapsto e_1 + e_2$). Therefore $m(rv) = m(v)$ and $m(sv) = m(v)$ for all primitive v, hence $m(kv) = m(v)$ for $k \in \mathcal{K}$, and $N(kv) = N(v)$; by homogeneity and continuity $N \circ k = N$ on $\mathbb{R}^2$.

(b) The element $-I = r^3$ lies in $\mathcal{K}$, and $\mathcal{K}/\{\pm I\}$ acts on slopes by the two maps $\sigma \mapsto 1/(1 - \sigma)$ (induced by r) and $\sigma \mapsto 1 - \sigma$ (induced by s). The six images of the interval $[0, 1/2]$ under this group of order six are the intervals

$$[0, \tfrac{1}{2}],\ [\tfrac{1}{2}, 1],\ [1, 2],\ [2, \infty],\ [\infty, -1],\ [-1, 0]$$

of $\mathbb{RP}^1$, which cover $\mathbb{RP}^1$ and have disjoint interiors. The endpoints form the two orbits $\{0, 1, \infty\}$ and $\{1/2, 2, -1\}$, each of which meets $[0, 1/2]$ once. Hence every direction has exactly one $\mathcal{K}$-image of slope in $[0, 1/2]$, and of the two primitive vectors on that direction exactly one has $q > 0$, since the slope is finite.

(c) We show that the map of (c) sends the Farey tree of $\mathcal{S}$ to the tree of sorted Markoff triples, root to root and children to children; such a map of rooted binary trees is a bijection.

Let $v \in \mathcal{S}$ be primitive, $v \ne e_2, f$, with Farey parents u, u'. We claim: $m(u), m(u'), m(v)$ are positive integers, the sorted triple formed by them is a Markoff triple with largest coordinate $m(v)$, and the triples of the two children of v are the two children of this triple. By (2.2) applied to the Farey neighbours u, u' and the relation $u - u' = \pm u''$ of §2.3,

$$m(v) = 3\, m(u)\, m(u') - m(u'')$$

with u'' an earlier node or one of $e_2, f, (1, 1)$; here $m(e_2) = \operatorname{tr} B/3 = 1$, $m(f) = \operatorname{tr}(AB^2)/3 = 2$ (AB^2 is the holonomy of $f = e_1 + 2e_2$) and $m(1, 1) = \operatorname{tr} AB/3 = 1$. For the root $v = (1, 3)$ this gives $m(v) = 6 - 1 = 5$, and $(1, 2, 5)$ is a Markoff triple with the two children $(1, 5, 13)$ and $(2, 5, 29)$ attached to $(1, 4) = e_2 + v$ and $(2, 5) = v + f$.

For a general node, induction along the tree together with the Markoff relation (2.3) shows that $(m(u), m(u'), m(v))$ is a Markoff triple. The child $u + v$, with parents u, v, has label $3m(u)m(v) - m(u')$, and the child $v + u'$, with parents v, u', has label $3m(v)m(u') - m(u)$. If (a, b, c) is the sorted triple formed by $m(u), m(u'), m(v)$, the triples of the two children of v are the two children $(a, c, 3ac - b)$ and $(b, c, 3bc - a)$ of (a, b, c), and $3ac - b \ge 3c - b \ge 2c > c$ shows that the new label is the largest coordinate.

On the other side, the sorted Markoff triples form a tree under the Vieta involutions. If (a, b, c) is a sorted Markoff triple, then $c' = 3ab - c = (a^2 + b^2)/c$ is a positive integer. If $(a, b) \ne (1, 1)$, then

the quadratic $z^2 - 3abz + a^2 + b^2$, whose roots are c and c', takes at $z = b$ the value

$$a^2 - (3a - 2)b^2 < 0,$$

so $0 < c' < b < c$. Hence the sorted rearrangement of (c', a, b) is a sorted Markoff triple with maximum $b < c$, the *parent* of (a, b, c), and (a, b, c) is one of the two children of its parent.

If $a = b$, the equation forces $a = b = 1$ and $c \in \{1, 2\}$; if $b = c$ it forces $a = b = c = 1$. Hence every sorted triple other than $(1, 1, 1)$ and $(1, 1, 2)$ has $a < b < c$, two distinct sorted children and a unique parent. Descending along parents leads to $(1, 2, 5)$, so the sorted triples other than $(1, 1, 1)$ and $(1, 1, 2)$ form a binary tree rooted at $(1, 2, 5)$ in which each triple occurs exactly once.

The map sending v to the sorted triple formed by $m(u), m(u'), m(v)$ is a map of rooted binary trees that respects the children, so it is a bijection. Together with $e_2 \mapsto (1, 1, 1)$ and $f \mapsto (1, 1, 2)$ this proves (c). In particular every label is a positive integer, and the formula for $N(v)$ in (F2) becomes the length formula (1.2).

(d) The region $\mathcal{S} \cap L\mathbb{B}$ is convex with perimeter $O(L)$, so the number of lattice points in it is $\mathrm{area}(\mathcal{S} \cap \mathbb{B})L^2 + O(L)$. Möbius inversion over the common divisor of the coordinates gives $\#(\mathbb{Z}^2_{\mathrm{prim}} \cap \mathcal{S} \cap L\mathbb{B}) = \mathrm{area}(\mathcal{S} \cap \mathbb{B})L^2/\zeta(2) + O(L \log L)$, which is (2.4) with $c_N = \mathrm{area}(\mathbb{B})/(12\zeta(2))$; here we used that $\mathbb{B}$ is $\mathcal{K}$-invariant by (a) and that the twelve images of $\mathcal{S}$ under $\mathcal{K}$ tile the plane by (b). □

Remark 2.2. Lemma 2.1 identifies $T(X)$ with the number of primitive vectors of $\mathcal{S}$ with $N \le L$, and $D(X)$ with the number of distinct values of N on them, where $L = 2\,\mathrm{arcosh}(3X/2)$; this is the form in which Theorem 1.1 is proved in §6.6.

3. The boundary of the McShane–Rivin unit ball along Farey sectors

3.1. **Supporting functionals and corners.** A *supporting functional* of N at a nonzero vector x is a linear form α on $\mathbb{R}^2$ with

$$\alpha \le N \quad \text{on } \mathbb{R}^2 \qquad \text{and} \qquad \alpha(x) = N(x).$$

Geometrically, $\{\alpha = N(x)\}$ is a supporting line of the convex body $N(x)\mathbb{B}$ at the boundary point x.

The set

$$\partial N(x) := \{\alpha \in (\mathbb{R}^2)^* : \ \alpha \le N, \ \alpha(x) = N(x)\}$$

of supporting functionals at x (the subdifferential of N at x) is a nonempty compact convex set that depends only on the ray through x; it consists of a single functional exactly when $\partial\mathbb{B}$ is differentiable at $x/N(x)$.

We call $\alpha/|\alpha| \in S^1$ the *normal direction* of a supporting functional α at x; it is an outward unit normal of $\partial\mathbb{B}$ at $x/N(x)$. For two linear forms or vectors we write $\measuredangle(\alpha, \beta) \in [0, \pi]$ for the angle between them, and for a set $\{\beta, -\beta\}$ we write $\measuredangle(\alpha, \{\beta, -\beta\}) = \min(\measuredangle(\alpha, \beta), \measuredangle(\alpha, -\beta))$, the angle between the lines they span.

Since $c|y| \le N(y) \le C|y|$, every supporting functional satisfies $c \le |\alpha| \le C$. A supporting functional is determined by its normal direction $n = \alpha/|\alpha|$, as $\alpha = n/\max_{\mathbb{B}} n$, and the support function $n \mapsto \max_{\mathbb{B}} n$ of $\mathbb{B}$ is positive and Lipschitz on the unit circle; hence $|\alpha - \alpha'| \le C\,|\alpha/|\alpha| - \alpha'/|\alpha'||$ for supporting functionals α, α'.

Strict convexity of $\mathbb{B}$ has the following consequence, used repeatedly. For every nonzero linear form β there is a unique point $x_\beta \in \partial\mathbb{B}$ with $\beta(x_\beta) = \max_{\mathbb{B}} \beta$; we call x_β the *contact point* of β.

Equivalently, β is a positive multiple of a supporting functional at x_β and at no other boundary point.

As x runs once around $\partial\mathbb{B}$ in the positive sense, the normal directions of the supporting functionals at x run once around S^1 in the positive sense, monotonically; the normal directions at a fixed boundary point x form a closed arc of S^1, which is a single point unless $\partial\mathbb{B}$ has a corner at x.

The one-sided directional derivative of the convex function N at x in the direction h is

$$D_h N(x) = \lim_{t\downarrow 0} \frac{N(x+th)-N(x)}{t} = \max_{\alpha\in\partial N(x)} \alpha(h).$$

The *corner jump*

$$J_N(x;h) := D_h N(x) + D_{-h} N(x) \ \geq 0$$

measures the opening of the corner of the graph of N at x in the direction h; it vanishes for all h if and only if N is differentiable at x. If $\partial N(x)$ is a segment with endpoints α^+, α^-, the *extreme supporting functionals* at x, then $J_N(x;h) = |(\alpha^+ - \alpha^-)(h)|$ for every h.

3.2. **The corner formula and the exponential estimate.** The next proposition contains the geometric input of the paper. Its first part says that $\partial\mathbb{B}$ has a corner at every rational direction and computes its size; its second part says that between two consecutive rays of the fan of level H the boundary is flat to within $e^{-\kappa H}$.

Proposition 3.1. *There are constants $c, C, \kappa > 0$ with the following properties.*

(a) (Corner formula.) For every primitive u and every $h \in \mathbb{R}^2$,

$$J_N(u;h) = 4\,|\det(u,h)|\,\log\coth\big(N(u)/2\big). \tag{3.1}$$

In particular N is not differentiable at any rational direction, and the two extreme supporting functionals $\alpha^\pm$ at u satisfy $|\det(\alpha^+,\alpha^-)| = 4N(u)\log\coth(N(u)/2)$. If $\mathrm{ht}(u) \leq H$, then the normal directions of α^+ and α^- differ by an angle of at least ce^{-CH}.

(b) (Flatness of sectors.) Let Σ be a Farey sector of level $H \geq 1$ with boundary vectors u, v. Let α be the supporting functional at u with $\alpha(v) = D_v N(u)$, and β the supporting functional at v with $\beta(u) = D_u N(v)$; by (a) each is unique. We call α and β the inward *supporting functionals at u and v (they are the extreme supporting functionals whose normal directions are closest to the normal directions at points of Σ). Then*

$$|\alpha - \beta| \leq Ce^{-\kappa H},$$

and every supporting functional at every point of Σ lies within $Ce^{-\kappa H}$ of α (and of β).

(c) (Total turning.) The sum, over all Farey sectors Σ of level H, of the angular width of the arc of normal directions of supporting functionals at points of Σ is at most $Ce^{-\kappa H}$.

We first fix notation for the holonomy of a pair of Farey neighbours. Let u, v be Farey neighbours with $\det(u,v) = 1$, and let $U = \rho(w_u)$, $V = \rho(w_v)$ be holonomy matrices of a free basis (w_u, w_v) lifting (u,v). Their traces

$$x := \operatorname{tr} U = 3m(u), \qquad y := \operatorname{tr} V = 3m(v), \qquad z := \operatorname{tr} UV = 3m(u+v)$$

are positive by Lemma 2.1. Put $\lambda = e^{N(u)/2}$ and $\mu = e^{N(v)/2}$, so that $x = \lambda + \lambda^{-1}$ and $y = \mu + \mu^{-1}$, and conjugate the pair (U,V) by an element of $\mathrm{SL}_2(\mathbb{R})$ so that $U = \mathrm{diag}(\lambda, \lambda^{-1})$ and $V = \left(\begin{smallmatrix} a & b_1 \\ b_2 & d \end{smallmatrix}\right)$ with $ad - b_1 b_2 = 1$.

Since $\mathrm{tr}[U,V]=-2$, and $\mathrm{tr}[U,V]=2ad-b_1b_2(\lambda^2+\lambda^{-2})$ by direct computation, we get $b_1b_2(\lambda-\lambda^{-1})^2=4$, hence

$$b_1b_2=\frac{4}{(\lambda-\lambda^{-1})^2},\qquad ad=1+b_1b_2=\coth^2\!\big(N(u)/2\big). \tag{3.2}$$

As $a+d=y>0$ and $ad>0$, both a and d are positive. The words $w_u^n w_v$ and $w_u^n w_v^{-1}$ are primitive with homology classes $nu+v$ and $nu-v$, and

$$\mathrm{tr}(U^nV)=\lambda^n a+\lambda^{-n}d,\qquad \mathrm{tr}(U^nV^{-1})=\lambda^n d+\lambda^{-n}a. \tag{3.3}$$

Proof of Proposition 3.1. The corner formula. Let u be primitive and choose v with $\det(u,v)=1$, with the notation above. Since N is convex and homogeneous, the one-sided derivative in the direction v is a limit along the primitive vectors $nu+v$:

$$D_vN(u)=\lim_{n\to\infty}\big(N(nu+v)-nN(u)\big),$$

because $N(nu+v)-nN(u)=n\big(N(u+v/n)-N(u)\big)$.

By (3.3) and $N=2\,\mathrm{arcosh}(\mathrm{tr}\,/2)=2\log\mathrm{tr}+o(1)$,

$$N(nu+v)=2\log(\lambda^n a+\lambda^{-n}d)+o(1)=nN(u)+2\log a+o(1),$$

so $D_vN(u)=2\log a$, and likewise $D_{-v}N(u)=2\log d$. Hence, by (3.2),

$$J_N(u;v)=2\log(ad)=4\log\coth\big(N(u)/2\big).$$

For general h, write $h=\det(u,h)\,v+tu$; adding a multiple of u does not change $J_N(u;\cdot)$, since $\alpha(u)=N(u)$ for all $\alpha\in\partial N(u)$, and $J_N(u;sh)=|s|J_N(u;h)$. This proves (3.1).

Since $J_N(u;v)>0$, N is not differentiable at u, so $\partial N(u)$ is a nondegenerate segment with endpoints $\alpha^\pm$, and $(\alpha^+-\alpha^-)(u)=0$, $|(\alpha^+-\alpha^-)(v)|=J_N(u;v)$ give $\alpha^+-\alpha^-=\pm J_N(u;v)\det(u,\cdot)$. Identifying linear forms with vectors, we get $\det(\alpha^+,\alpha^-)=\det(\alpha^+,\alpha^--\alpha^+)=\mp J_N(u;v)\,\alpha^+(u)$, whose absolute value is $4N(u)\log\coth(N(u)/2)$. As $c\le|\alpha^\pm|\le C$, the angle θ between α^+ and α^- satisfies

$$\sin\theta\ \ge\ cN(u)\log\coth(N(u)/2)\ \ge\ cN(u)e^{-N(u)},$$

using $\log\coth(t/2)\ge 2e^{-t}$. If $\mathrm{ht}(u)\le H$ then $N(u)\le C|u|\le CH$, and $N(u)\ge 2\,\mathrm{arcosh}(3/2)>1$, so $\theta\ge ce^{-CH}$.

The larger root. Let Σ be a Farey sector of level $H\ge 1$ with boundary vectors u,v, $\det(u,v)=1$. We claim that

$$m(u-v)\le\max\big(m(u),m(v)\big),\qquad\text{hence}\qquad m(u+v)\ge 2\max\big(m(u),m(v)\big)>m(u-v). \tag{3.4}$$

The numbers $3m(u+v)$ and $3m(u-v)$ are the two roots of $Z^2-xyZ+x^2+y^2=0$, since the triples $(m(u),m(v),m(u\pm v))$ satisfy the Markoff relation (2.3), applied to u,v and to $u,-v$; so the claim says that $z=3m(u+v)$ is the *larger* root.

The second statement follows from the first by (2.2), since

$$m(u+v)=3m(u)m(v)-m(u-v)\ge 3\max\big(m(u),m(v)\big)-\max\big(m(u),m(v)\big).$$

The first statement holds for the eight sectors of level 1. Their boundary vectors are consecutive among the eight primitive vectors of height one, $\pm e_1,\pm(e_1+e_2),\pm e_2,\pm(e_1-e_2)$, so the difference of the two boundary vectors of each sector is one of $\pm e_1,\pm e_2$, of label 1, while every label is at least 1.

If it holds for a sector with boundary u, v, it holds for the two sectors with boundaries $u, u+v$ and $u+v, v$ obtained by inserting the mediant: for the first, $m(u-(u+v)) = m(v) \le \max(m(u), m(u+v))$ by the second statement for (u, v). Since every sector of level $H \ge 1$ arises from a sector of level 1 by successive mediant insertions, the claim follows by induction.

(The restriction to sectors of fans of level $H \ge 1$ matters: e_1 and e_2 are Farey neighbours, but the ray through $e_1 + e_2$ separates them in every fan, and for this pair the mediant $e_1 + e_2$ carries the smaller root, since $m(e_1 + e_2) = 1 < 2 = m(e_1 - e_2)$.)

The expanding entry. Keep u, v, Σ as above and put

$$P := (\lambda - \lambda^{-1})(\mu - \mu^{-1}) = \sqrt{(x^2 - 4)(y^2 - 4)}.$$

Since $x, y \ge 3$ we have $\lambda, \mu \ge (3 + \sqrt{5})/2$ and $P \ge 5$. The two roots of $Z^2 - xyZ + x^2 + y^2$ are $(xy \pm \sqrt{x^2y^2 - 4x^2 - 4y^2})/2$, and $x^2y^2 - 4x^2 - 4y^2 = P^2 - 16$; by the larger root claim,

$$z = \frac{xy + \sqrt{P^2 - 16}}{2}.$$

Compare z with $z_0 := \lambda\mu + (\lambda\mu)^{-1} = (xy + P)/2$. The difference

$$\xi := z_0 - z = \frac{P - \sqrt{P^2 - 16}}{2} = \frac{8}{P + \sqrt{P^2 - 16}}$$

satisfies $0 < \xi \le \min(1, 5/P)$ for $P \ge 5$.

Solving $a + d = y$ and $\lambda a + \lambda^{-1} d = z$ for the entry a of V gives

$$a = \frac{z - \lambda^{-1} y}{\lambda - \lambda^{-1}} = \mu - \frac{\xi}{\lambda - \lambda^{-1}} = \mu(1 - \delta), \qquad \delta = \frac{\xi}{\mu(\lambda - \lambda^{-1})} \le \frac{5}{P\mu(\lambda - \lambda^{-1})}.$$

A direct evaluation gives $0 < \delta < 1/5$, and $\delta \le C\lambda^{-2}\mu^{-2} = Ce^{-N(u)-N(v)}$. Therefore

$$|2 \log a - N(v)| = 2|\log(1 - \delta)| \le Ce^{-N(u)-N(v)}. \tag{3.5}$$

Flatness of the sector. By (a), $\partial N(u)$ is a segment whose endpoints take values at v differing by $J_N(u; v) > 0$; so α is the endpoint at which the value at v is larger, and $\alpha(v) = D_v N(u) = 2 \log a$, $\alpha(u) = N(u)$. Interchanging the roles of u and v (the computation of the expanding entry used only the traces x, y, z and the trace identities, which are symmetric in u and v) gives $\beta(v) = N(v)$ and $|\beta(u) - N(u)| \le Ce^{-N(u)-N(v)}$. Hence

$$|(\alpha - \beta)(u)| + |(\alpha - \beta)(v)| \le Ce^{-N(u)-N(v)}.$$

The dual basis of (u, v) consists of the forms $\det(\cdot, v)$ and $\det(u, \cdot)$, of Euclidean norm $|v|, |u| \le \sqrt{2}H$, so $|\alpha - \beta| \le CHe^{-N(u)-N(v)}$. Finally $N(u) + N(v) \ge N(u + v) \ge c|u + v| \ge c \operatorname{ht}(u + v) \ge cH$, because the mediant has height $> H$. Thus $|\alpha - \beta| \le CHe^{-cH} \le Ce^{-\kappa H}$.

Let $\mathcal{A}_\Sigma \subset S^1$ be the set of normal directions of supporting functionals at points of Σ, together with the normal directions of α and β. Supporting functionals γ_t at the points $u + tv$ of Σ converge to α as $t \downarrow 0$. Since $\gamma_t \le N$ and $\gamma_t(u + tv) = N(u + tv)$, every limit point of γ_t lies in $\partial N(u)$ and takes at v a value at least $\lim_{t\downarrow 0}(N(u + tv) - N(u))/t = D_v N(u)$, so it is α. Likewise, supporting functionals at the points $v + tu$ converge to β. By the monotonicity of normal directions, $\mathcal{A}_\Sigma$ is therefore the closed arc from $\alpha/|\alpha|$ to $\beta/|\beta|$ traversed in the positive sense.

The arc $\mathcal{A}_{-\Sigma}$ of the opposite sector is $-\mathcal{A}_\Sigma$, and the two arcs have disjoint interiors. Two antipodal arcs with disjoint interiors have length at most π. An arc of length at most π whose endpoints are at chord distance ε has length at most $\pi\varepsilon/2$, and here $\varepsilon \le 2|\alpha - \beta|/|\alpha| \le Ce^{-\kappa H}$.

Hence every normal direction at a point of Σ is within angle $Ce^{-\kappa H}$ of that of α, and by the Lipschitz dependence of a supporting functional on its normal direction (§3.1) every supporting functional at a point of Σ is within $Ce^{-\kappa H}$ of α. This proves (b).

Total turning. The fan of level H has $O(H^2)$ sectors, so the sum of the lengths of the arcs $\mathcal{A}_\Sigma$ is $O(H^2 e^{-\kappa H})$, which is $O(e^{-\kappa' H})$ for a smaller κ'. This proves (c). □

Remark 3.2. Doan, Li and Nguyen [DLN26, Proposition 4.1 and (9)] prove part (a), by essentially the same computation, for every complete hyperbolic once-punctured torus of finite area. Part (a) also follows from the work of Gaster [Gas22, Propositions 1.8 and 2.2], who computes the one-sided derivatives of Fock's function at rational points from the lengths of iterated Dehn twists. In the notation of the proof of (a), his Proposition 2.2 gives $D_v N(u)$ and $D_{-v} N(u)$ in terms of $m(u)$, $m(v)$ and $m(u \mp v)$, and the Markoff relation (2.3) reduces their sum to (3.1). Hines [Hin20, Theorem on p. 6] computed these derivatives earlier, for the function l/q with l the length, but his formulas lose a factor 2 in an exponent: in the limit on his p. 7 the exponent $2q$ becomes q. Corrected, the exponent q of the fractions in both formulas of his Theorem is $2q$, and his formula for the jump of the derivative on p. 8 reads $\coth^{4q}(l/2)$ instead of $\coth^{2q}(l/2)$, in agreement with (3.1).

3.3. **Contact points of integral forms, and the excision of sectors.** A linear form $\beta = (\beta_1, \beta_2)$ on $\mathbb{R}^2$, $\beta(x) = \beta_1 x_1 + \beta_2 x_2$, is *integral* if $\beta_1, \beta_2 \in \mathbb{Z}$, and its size is $|\beta| = (\beta_1^2 + \beta_2^2)^{1/2}$.

Lemma 3.3. *The following statements hold.*

(a) For $M \geq 1$, the contact points x_β of the nonzero integral forms β with $|\beta| \leq M$ lie on at most $C \log^2(2 + M)$ rays.

(b) Let $H \geq 1$ and $k \geq 1$ with $\sqrt{k} \leq cH$ for a sufficiently small constant c. The union of any k Farey sectors of level H has angular width at most $C\sqrt{k}/H$.

Proof. (a) Let β, β' be integral forms of size at most M that are not positive multiples of each other. The angle between them is at least M^{-2}: it is π if β' is a negative multiple of β, and otherwise $|\det(\beta, \beta')| \geq 1$ gives $\measuredangle(\beta, \beta') \geq \sin \measuredangle(\beta, \beta') \geq M^{-2}$. Choose $H_0 = \lceil C \log(2 + M) \rceil$ so large that the total turning bound of Proposition 3.1(c) at level H_0 is smaller than M^{-2}. The normal directions at the points of one open sector of level H_0 then form an arc of length $< M^{-2}$, so the integral forms of size at most M whose contact points lie in that sector are positive multiples of one another, and therefore have one and the same contact point. Hence the contact points lying in open sectors lie on at most $O(H_0^2)$ rays, and the contact points lying on rays of the fan lie on at most $O(H_0^2)$ rays, namely the rays of $\mathcal{F}_{H_0}$.

(b) By the symmetries of the fan it suffices to treat sectors in the octant $0 \leq p \leq q$, where a sector of level H consists of the vectors (p, q) with p/q strictly between consecutive terms $a/b < c/d$ of the Farey sequence of order H; here $bc - ad = 1$, $b, d \leq H$, $b + d > H$, and the sector has slope width $1/(bd)$, which is comparable to its angular width. Fix an integer n with $1 \leq n \leq H/2$. A sector having an endpoint of denominator $b \leq n$ has its other denominator $d > H - b \geq H/2$, hence slope width $O(1/(bH))$; since there are at most b fractions of denominator b in $[0, 1]$, each bounding two sectors, the sectors with an endpoint of denominator at most n have total width $O(n/H)$. Every other sector has both denominators $> n$ and their sum $> H$, so $bd \geq n(H - n) \geq nH/2$ and its width is $O(1/(nH))$. Thus k sectors have total width $O(n/H + k/(nH))$; taking $n \asymp \sqrt{k}$ gives the claim, the hypothesis on k ensuring $n \leq H/2$. □

Definition 3.4 (Excision)**.** Let $M \geq 1$ and let H be an integer with $H \geq C_{\text{exc}} \log(2 + M)$, where C_{exc} is a sufficiently large constant fixed once and for all. The (M, H)-*excision* removes from the plane:

- every ray of the fan $\mathcal{F}_H$;
- every Farey sector of level H whose closure contains the contact point x_β of an integral form β with $0 < |\beta| \le M$.

(When x_β lies on a ray of $\mathcal{F}_H$, both sectors adjacent to that ray are excised.) The sectors removed are called *excised*; the remaining sectors of level H are *retained*. A vector is *retained* if it lies in a retained sector, and *excised* otherwise.

Lemma 3.5. *Let M, H be as in Definition 3.4 and let $L \ge 1$.*

(a) At most $C\log^2(2+M)$ sectors are excised, and their total angular width is at most $C\log(2+M)/H$.

(b) The number of excised primitive vectors v with $N(v) \le L$ is at most

$$C\Big(\frac{\log(2+M)}{H}\,L^2 + L\log^2(2+M) + H^2\Big).$$

(c) (Separation.) There is a constant C_{sep}, depending only on N, such that for every retained primitive vector v, every supporting functional γ at v, and every integral form β with $0 < |\beta| \le M$,

$$\measuredangle(\gamma, \{\beta, -\beta\}) \ge ce^{-C_{\mathrm{sep}}H}.$$

Proof. (a) By Lemma 3.3(a) the contact points in question lie on $k_0 \le C\log^2(2+M)$ rays, and each such ray meets at most two closed sectors of level H; so at most $2k_0$ sectors are excised. By Lemma 3.3(b), which applies because $\sqrt{2k_0} \le cH$ when C_{exc} is large, their total width is at most $C\sqrt{2k_0}/H \le C\log(2+M)/H$.

(b) The intersection of a sector of angular width θ with the ball $L\mathbb{B}$ is convex, of area $O(\theta L^2)$ and perimeter $O(L)$, so it contains $O(\theta L^2 + L + 1)$ lattice points. Summing over the $O(\log^2(2+M))$ excised sectors, and adding one primitive vector for each of the $O(H^2)$ rays of $\mathcal{F}_H$, gives the bound.

(c) Let v lie in a retained sector Σ, with boundary rays R_1, R_2, and let β be integral with $0 < |\beta| \le M$. Both x_β and $x_{-\beta} = -x_\beta$ are contact points of integral forms of size at most M, so neither lies in the closure of Σ.

Hence, moving from v to x_β in the positive direction around the origin one crosses one of the rays R_1, R_2 before reaching x_β, and moving in the negative direction one crosses the other ray first; the same holds for $-x_\beta$.

Since γ is a supporting functional at a point of the open sector Σ and β is a positive multiple of a supporting functional at x_β, the monotonicity of normal directions along $\partial\mathbb{B}$ shows that the normal directions of γ and of β are separated, in each of the two directions around S^1, by the full closed arc of normal directions at the primitive vector spanning the ray crossed. This primitive vector has height at most H, so by Proposition 3.1(a) the arc has length at least ce^{-CH}; by strict convexity, the normal direction of γ, a supporting functional at a point of the open sector, lies outside the arc. Hence $\measuredangle(\gamma, \beta) \ge ce^{-CH}$, and likewise $\measuredangle(\gamma, -\beta) \ge ce^{-CH}$; this is the claim, with C_{sep} the constant C of Proposition 3.1(a). □

4. Equal-length geodesics under a fixed rational map

This section proves the following proposition, the arithmetic input of the proof of Theorem 1.1.

Proposition 4.1. *Let $G \in \mathrm{GL}_2(\mathbb{Q}) \setminus \mathcal{K}$. Then, as $L \to \infty$,*

$$\#\{v \in \mathbb{Z}^2_{prim} : \ Gv \in \mathbb{Z}^2_{prim}, \ N(v) = N(Gv) \le L\} = o_G(L^2).$$

By the length formula (1.2), the condition $N(v) = N(Gv)$ is the equality $m(v) = m(Gv)$ of two labels, and labels are values of integral polynomials on the Markoff surface $a^2 + b^2 + c^2 = 3abc$ (§4.1). The proof combines a zero-one law for the 3-adic dynamics of $\mathrm{SL}_2(\mathbb{Z})$ on this surface, a rigidity statement for rational isometries of N, and a count of lattice points in orbits of finite-index subgroups of $\mathrm{SL}_2(\mathbb{Z})$.

4.1. **The Markoff surface and the mapping class group action.** For a word $w \in F_2$ there is a unique polynomial $\tau_w \in \mathbb{Z}[X, Y, Z]$ such that $\operatorname{tr}\sigma(w) = \tau_w(\operatorname{tr}\sigma(x), \operatorname{tr}\sigma(y), \operatorname{tr}\sigma(xy))$ for every representation $\sigma : F_2 \to \mathrm{SL}_2(R)$ over every commutative ring R; this is the classical trace calculus of Fricke, see [Gol09, §2.2]. The trace relation $X^2 + Y^2 + Z^2 = XYZ$ expresses $\operatorname{tr}\sigma([x, y]) = -2$ by the Fricke identity (2.1). In the scaled coordinates $a = X/3$, $b = Y/3$, $c = Z/3$, a change of coordinates over $\mathbb{Q}$, it becomes the *Markoff surface*

$$\mathcal{M} : \quad a^2 + b^2 + c^2 = 3abc,$$

and we use this equation over $\mathbb{Z}$ as the integral model.

The group $\mathrm{Aut}(F_2)$ acts on trace triples: an automorphism ϕ sends the character of σ to the character of $\sigma \circ \phi$, and on the coordinates (X, Y, Z) this is the polynomial map $(\tau_{\phi(x)}, \tau_{\phi(y)}, \tau_{\phi(xy)})$.

By Nielsen's theorem [MKS76] $\mathrm{Aut}(F_2)$ is generated by the substitutions $(x, y) \mapsto (y, x)$, $(x, y) \mapsto (x^{-1}, y)$ and $(x, y) \mapsto (xy, y)$, which act on (X, Y, Z) by (Y, X, Z), $(X, Y, XY-Z)$ and $(Z, Y, YZ-X)$ respectively (we used $\operatorname{tr}(x^{-1}y) = XY - Z$ and $\operatorname{tr}(xy^2) = YZ - X$). In the scaled coordinates these are the transposition $(a, b, c) \mapsto (b, a, c)$, the map $(a, b, c) \mapsto (a, b, 3ab - c)$, which is the *Vieta involution* in the variable c, and the map $(a, b, c) \mapsto (c, b, 3bc - a)$. We call the group of polynomial automorphisms of $\mathcal{M}$ generated by the coordinate permutations and the Vieta involutions in the three variables the *Nielsen group*, and its elements *Nielsen moves*. Each Nielsen move is a polynomial map with integer coefficients whose inverse is again a Nielsen move.

Inner automorphisms of F_2 act trivially on characters, and $\mathrm{Aut}(F_2)/\operatorname{Inn}(F_2) \cong \mathrm{GL}_2(\mathbb{Z})$ via the action on $H_1 = \mathbb{Z}^2$. We obtain a *right action* of $\mathrm{GL}_2(\mathbb{Z})$ on $\mathcal{M}$ by Nielsen moves, written $P \mapsto P * \gamma$, characterized by

$$Q_w(P * \gamma) = Q_{\phi(w)}(P) \qquad (w \in F_2, \ \phi \in \mathrm{Aut}(F_2) \text{ inducing } \gamma \text{ on } H_1), \tag{4.1}$$

where $Q_w(a, b, c) := \frac{1}{3}\tau_w(3a, 3b, 3c)$ is the scaled trace function of w; it lies in $\frac{1}{3}\mathbb{Z}[a, b, c]$, and in $\mathbb{Z}[a, b, c]$ when w is primitive, see (4.2). (The action is a right action because the character of $\sigma \circ \phi \circ \psi$ is obtained by applying ψ to the character of $\sigma \circ \phi$.) The element $-I$ acts trivially, since it is induced by $(x, y) \mapsto (x^{-1}, y^{-1})$, which preserves the traces of x, y and xy.

For a primitive vector v we put $Q_v := Q_{w_v}$; this is well defined because conjugate words and inverse words have the same trace. Then

$$Q_v \in \mathbb{Z}[a, b, c], \qquad Q_v(o) = m(v) \ \text{ for } o := (1, 1, 1), \qquad Q_v(P * \gamma) = Q_{\gamma v}(P) \quad (\gamma \in \mathrm{GL}_2(\mathbb{Z})). \tag{4.2}$$

The integrality of Q_v holds because w_v is the image of x under an automorphism of F_2, so that Q_v is the first coordinate of the Nielsen move attached to that automorphism, a polynomial with integer coefficients. The value at o is $\operatorname{tr}\rho(w_v)/3 = m(v)$ because o is the scaled character of ρ. The last formula is (4.1) for $w = w_v$, since $\phi(w_v)$ has homology class γv.

Finally, for $G \in \mathrm{GL}_2(\mathbb{Q})$ let

$$H_G := \Gamma \cap G^{-1}\Gamma G = \{h \in \Gamma : \ GhG^{-1} \in \Gamma\},$$

a subgroup of $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ containing $-I$.

Lemma 4.2. *Let $G \in \mathrm{GL}_2(\mathbb{Q})$. Then H_G has finite index in Γ, and the set $\{v \in \mathbb{Z}^2_{prim} : Gv \in \mathbb{Z}^2_{prim}\}$ is H_G-stable and is a finite union of H_G-orbits.*

Proof. Choose an integer $D > 0$ with $\mathbf{A} := DG$ integral, and put $\Delta = |\det \mathbf{A}|$. For $h \in \Gamma$ one has $GhG^{-1} = I + \mathbf{A}(h - I)\operatorname{adj}(\mathbf{A})/\det \mathbf{A}$, which is integral if $h \equiv I \pmod{\Delta}$. Thus H_G contains the principal congruence subgroup $\Gamma(\Delta)$.

If v is primitive with Gv primitive and $h \in H_G$, then $Ghv = (GhG^{-1})(Gv)$ is primitive. For primitive v, the adjugate identity $\operatorname{adj}(\mathbf{A})\mathbf{A}v = (\det \mathbf{A})v$ shows that $\gcd(\mathbf{A}v)$ divides Δ, and $Gv = \mathbf{A}v/D$ is primitive integral if and only if $\gcd(\mathbf{A}v) = D$; this condition depends only on v modulo Δ.

Primitive vectors v, v' with $v \equiv v' \pmod{\Delta}$ lie in the same $\Gamma(\Delta)$-orbit: complete v to a matrix $(v \mid u) \in \Gamma$, then choose $u' \equiv u \pmod{\Delta}$ with $\det(v' \mid u') = 1$ (possible since $\det(v' \mid u) \equiv 1 \pmod{\Delta}$ and v' is primitive), and observe that $(v' \mid u')(v \mid u)^{-1} \in \Gamma(\Delta)$ sends v to v'. Hence the set in question is a union of finitely many $\Gamma(\Delta)$-orbits. □

4.2. **3-adic points and normalized coordinates.** We now pass to 3-adic points. Let

$$\mathcal{X} := \{(a, b, c) \in (\mathbb{Z}_3^\times)^3 : \ a^2 + b^2 + c^2 = 3abc\} \subset \mathbb{Z}_3^3.$$

The base point $o = (1, 1, 1)$ lies in $\mathcal{X}$, and $\mathcal{X}$ is stable under the Nielsen moves, since $3ab - c \equiv -c \pmod 3$ is a unit when c is. On $\mathbb{Z}_3^3$ we use the metric $|P - P'|_3 = \max_i |P_i - P'_i|_3$. Because the Nielsen moves and their inverses are polynomial maps with integer coefficients, they are isometries of $\mathcal{X}$ for this metric; in particular they act on the finite sets $\mathcal{X}(3^n) := \mathcal{X} \bmod 3^n \subset (\mathbb{Z}/3^n)^3$, compatibly with reduction. A *state modulo* 3^n of a point of $\mathcal{X}$, or of $\mathcal{X}^2$, is its image in $\mathcal{X}(3^n)$, or in $\mathcal{X}(3^n)^2$.

The set $\mathcal{X}$ is a disjoint union of the eight *residue balls* $\mathcal{X}_\varepsilon = \{P \in \mathcal{X} : P \equiv \varepsilon \pmod 3\}$, $\varepsilon \in \{\pm 1\}^3$, each of which is nonempty by Hensel's lemma, since $a^2 + b^2 + c^2 \equiv 0 \pmod 3$ for all units and the derivative $2c - 3ab \equiv 2c$ with respect to c is a unit.

Fix a point $P_\varepsilon = (a_0, b_0, c_0)$ in each ball. By the formal implicit function theorem there is a unique power series $c_\varepsilon \in \mathbb{Z}_3[[s_1, s_2]]$ with constant term c_0 such that $(a_0 + s_1, b_0 + s_2, c_\varepsilon(s_1, s_2))$ lies on $\mathcal{M}$ identically; its coefficients lie in $\mathbb{Z}_3$ because the derivative $2c_0 - 3a_0b_0$ is a unit.

For $s \in (3\mathbb{Z}_3)^2$ the series converges, and by the uniqueness in Hensel's lemma $c_\varepsilon(s)$ is the unique $c \equiv c_0 \pmod 3$ with $(a_0 + s_1, b_0 + s_2, c) \in \mathcal{X}$. Hence the *normalized coordinates*

$$\mathcal{X}_\varepsilon \longrightarrow \mathbb{Z}_3^2, \qquad (a, b, c) \longmapsto x = \Big(\frac{a - a_0}{3}, \frac{b - b_0}{3}\Big)$$

form a bijection, and $c = c_\varepsilon(3x_1, 3x_2)$ is given by a power series in x whose coefficient of a monomial of degree d lies in $3^d\mathbb{Z}_3$. We write $\mathbb{Z}_3\langle x_1, \ldots, x_k\rangle$ for the ring of *restricted power series*, that is, power series with coefficients in $\mathbb{Z}_3$ tending to 0; these are exactly the power series that converge on all of $\mathbb{Z}_3^k$, and they are continuous there. Thus $c \in \mathbb{Z}_3\langle x_1, x_2\rangle$. Similarly, a product ball $\mathcal{X}_\varepsilon \times \mathcal{X}_{\varepsilon'}$ of $\mathcal{X}^2$ has normalized coordinates $x \in \mathbb{Z}_3^4$.

4.3. **A zero-one law for analytic dynamics on orbit closures.** We use the following theorem of Strassmann [Str28]; see [Cas86, Chapter 4].

> **Strassmann's theorem.** Let $f(t) = \sum_{n \geq 0} a_n t^n$ with $a_n \in \mathbb{Q}_3$, $a_n \to 0$, and $f \neq 0$. Then f has only finitely many zeros in $\mathbb{Z}_3$ (at most the largest n with $|a_n|_3 = \max_k |a_k|_3$).

Lemma 4.3 (Zero-one law)**.** *Let Z be a compact subset of $\mathbb{Z}_3^d$ and let a finitely generated group J act on Z on the right by isometries of the 3-adic metric. Let $z_0 \in Z$, let Y be the closure of the orbit $z_0 * J$, and let μ_Y be the probability measure on Y whose image on each finite quotient $Y(3^n) := Y \bmod 3^n$ is the uniform measure. Fix a finite generating set of J, and assume that for each g in it there is a continuous map $\Phi_g : \mathbb{Z}_3 \times Y \to Y$ with*

$$\Phi_g(n, z) = z * g^n \quad (n \in \mathbb{Z}_{\geq 0}), \qquad \Phi_g(t+1, z) = \Phi_g(t, z) * g \quad (t \in \mathbb{Z}_3).$$

Let $F : Y \to \mathbb{Z}_3$ be continuous, and assume that for every $z \in Y$ and every g in the generating set the function $t \mapsto F(\Phi_g(t, z))$ on $\mathbb{Z}_3$ is given by a power series with coefficients tending to 0. Then $\mu_Y(F = 0) \in \{0, 1\}$, and if $\mu_Y(F = 0) = 1$ then F vanishes identically on Y.

The measure μ_Y is well defined. The set $Y(3^n)$ is the finite transitive J-set $z_0 * J \bmod 3^n$, and the reduction $Y(3^{n+1}) \to Y(3^n)$ is a J-equivariant surjection between transitive J-sets; so all its fibres have the same cardinality, and the uniform measures are compatible. Since Y is the inverse limit of the $Y(3^n)$, the uniform measures define a Borel probability measure μ_Y on Y. It is J-invariant, and it has full support (every residue ball meeting Y has positive measure).

Proof. Ergodicity. Let $f \in L^2(\mu_Y)$ satisfy $f \circ g = f$ almost everywhere for all $g \in J$. Its conditional expectation f_n with respect to the partition of Y into residue classes modulo 3^n is a function on $Y(3^n)$, and it is J-invariant, because g permutes the residue classes and preserves μ_Y. A J-invariant function on the transitive finite J-set $Y(3^n)$ is constant. Since the residue partitions generate the Borel σ-algebra, $f_n \to f$ in L^2 by the martingale convergence theorem, so f is constant almost everywhere.

Invariance under the flows. Fix g in the generating set and $t \in \mathbb{Z}_3$, and choose nonnegative integers $n_k \to t$ in $\mathbb{Z}_3$. By uniform continuity of Φ_g on the compact set $\mathbb{Z}_3 \times Y$, $\Phi_g(n_k, \cdot) \to \Phi_g(t, \cdot)$ uniformly on Y, so for every continuous ψ on Y

$$\int_Y \psi(\Phi_g(t, z))\, d\mu_Y(z) = \lim_k \int_Y \psi(z * g^{n_k})\, d\mu_Y(z) = \int_Y \psi\, d\mu_Y.$$

Thus $\Phi_g(t, \cdot)$ preserves μ_Y.

The dichotomy. Put $E = \{z \in Y : F(z) = 0\}$, a closed set. Fix g in the generating set and $z \in Y$. By hypothesis $t \mapsto F(\Phi_g(t, z))$ is given by a power series with coefficients tending to zero, so by Strassmann's theorem it either vanishes identically or has finitely many zeros in $\mathbb{Z}_3$. In both cases the function $t \mapsto 1_E(\Phi_g(t+1, z)) - 1_E(\Phi_g(t, z))$ vanishes outside a finite set, hence

$$\int_{\mathbb{Z}_3} \left|1_E(\Phi_g(t+1, z)) - 1_E(\Phi_g(t, z))\right| dt = 0$$

for the Haar probability measure dt on $\mathbb{Z}_3$.

Integrating over z, using Fubini's theorem, the relation $\Phi_g(t+1, z) = \Phi_g(t, z) * g$ and the invariance of μ_Y under $\Phi_g(t, \cdot)$, we obtain

$$0 = \int_{\mathbb{Z}_3} \int_Y \left|1_E(\Phi_g(t, z) * g) - 1_E(\Phi_g(t, z))\right| d\mu_Y(z)\, dt = \int_Y \left|1_E(z * g) - 1_E(z)\right| d\mu_Y(z),$$

that is, $\mu_Y(E \triangle E * g^{-1}) = 0$. The set of $g \in J$ with this property is a subgroup containing the generating set, so it is all of J. Hence 1_E is J-invariant in $L^2(\mu_Y)$, and by ergodicity $\mu_Y(E) \in \{0, 1\}$. If $\mu_Y(E) = 1$, then the closed set E contains the support of μ_Y, which is Y. $\square$

4.4. **Interpolation of iterates.** Let $\varepsilon, \varepsilon' \in \{\pm 1\}^3$ and let $x \in \mathbb{Z}_3^4$ be the normalized coordinates of the product ball $\mathcal{X}_\varepsilon \times \mathcal{X}_{\varepsilon'}$ of $\mathcal{X}^2$. The p-adic interpolation of iterates goes back to Bell [Bel06, Bel08] and to Bell, Ghioca and Tucker [BGT10, Theorem 3.3], for $p \geq 5$; Cantat and Xie [CX18, Theorem 2.4] state it in terms of p-adic analytic flows. The next lemma is Poonen's interpolation theorem [Poo14, Theorem 1], which also covers $p = 3$, written for the prime 3 in the normalization we need; we include the short proof.

Lemma 4.4. *Let $\gamma, \gamma' \in \mathrm{GL}_2(\mathbb{Z})$ and let $g : \mathcal{X}^2 \to \mathcal{X}^2$ be the map $g(P, P') = (P * \gamma, P' * \gamma')$. Assume that g fixes every point of $\mathcal{X}^2$ modulo 9. Then g maps each product ball to itself, and in normalized coordinates it is given by*

$$g(x) = x + 3u(x), \qquad u \in \mathbb{Z}_3\langle x_1, \dots, x_4\rangle^4.$$

Moreover there is $\Phi_g \in \mathbb{Z}_3\langle t, x_1, \dots, x_4\rangle^4$ with $\Phi_g(n, x) = g^n(x)$ for all integers $n \geq 0$, and $\Phi_g(t + t', x) = \Phi_g(t, \Phi_g(t', x))$ for all $t, t' \in \mathbb{Z}_3$. In particular $\Phi_g(t, z)$ lies in the closure of $\{g^n(z) : n \geq 0\}$ for every z.

Proof. We first write g in normalized coordinates, which gives the first claim, and then build Φ_g as a binomial series in the difference operator $\Delta f = f \circ g - f$.

Since g fixes all points modulo 9, it fixes them modulo 3 and maps each product ball to itself. On the ball, in the unnormalized coordinates $s = (a - a_0, b - b_0, a' - a'_0, b' - b'_0) \in (3\mathbb{Z}_3)^4$, the map g is the composite of the implicit functions $c = c_\varepsilon(s_1, s_2)$, $c' = c_{\varepsilon'}(s_3, s_4)$, the polynomial Nielsen moves with integer coefficients, and the projection to the (a, b, a', b') coordinates. Hence, if the image point is also written in unnormalized coordinates, g is given by a power series $\mathbf{G}(s) = \sum_{d \geq 0} \mathbf{G}_d(s)$ with coefficients in $\mathbb{Z}_3$, $\mathbf{G}_d$ homogeneous of degree d; the constant term $\mathbf{G}_0$, the coordinate vector of the image of the point $(P_\varepsilon, P_{\varepsilon'})$ where $s = 0$, lies in $9\mathbb{Z}_3^4$ by the hypothesis. In the normalized coordinates $s = 3x$,

$$g(x) = \frac{\mathbf{G}(3x)}{3} = \frac{\mathbf{G}_0}{3} + \mathbf{G}_1(x) + \sum_{d \geq 2} 3^{d-1} \mathbf{G}_d(x).$$

The terms of degree $d \geq 2$ have coefficients in $3^{d-1}\mathbb{Z}_3 \subset 3\mathbb{Z}_3$, and the constant term lies in $3\mathbb{Z}_3^4$.

The hypothesis also gives $g(x) \equiv x \pmod 3$ for every $x \in \mathbb{Z}_3^4$; reducing the display modulo 3 shows that the linear map $\mathbf{G}_1$ satisfies $\mathbf{G}_1(x) \equiv x \pmod 3$ for all $x \in \mathbb{Z}_3^4$, so its matrix is $\equiv I \pmod 3$. Hence $g(x) - x$ has all coefficients in $3\mathbb{Z}_3$, and they tend to 0; this is the first claim.

Let $\mathcal{R} = \mathbb{Z}_3\langle x_1, \dots, x_4\rangle$ and define $\Delta f := f \circ g - f$ for $f \in \mathcal{R}$; composition with g is well defined on $\mathcal{R}$ since g has coefficients in $\mathbb{Z}_3$. By Taylor's formula, $f(x + 3u) - f(x) = \sum_{|\alpha| \geq 1} 3^{|\alpha|} (D^\alpha f)(x) u^\alpha$, where the divided derivatives $D^\alpha f = \partial^\alpha f / \alpha!$ again lie in $\mathcal{R}$. Hence $\Delta(3^r \mathcal{R}) \subset 3^{r+1}\mathcal{R}$ for all $r \geq 0$. Define

$$\Phi_g(t, x) := \sum_{k \geq 0} \binom{t}{k} \Delta^k x, \qquad \binom{t}{k} = \frac{t(t-1)\cdots(t-k+1)}{k!} \in \mathbb{Q}_3[t],$$

applied to the vector $x = (x_1, \dots, x_4)$ of coordinate functions. The polynomial $\binom{t}{k}$ has coefficients of 3-adic valuation at least $-v_3(k!) \geq -k/2$, while $\Delta^k x \in 3^k \mathcal{R}^4$; so the k-th term has coefficients of valuation at least $k/2$, and the series converges in $\mathbb{Z}_3\langle t, x\rangle^4$. For an integer $n \geq 0$, $\sum_{k \leq n} \binom{n}{k} \Delta^k = (1 + \Delta)^n$ is the operator $f \mapsto f \circ g^n$, so $\Phi_g(n, x) = g^n(x)$.

The identity $\Phi_g(t + t', x) = \Phi_g(t, \Phi_g(t', x))$ holds for $t, t' \in \mathbb{Z}_{\geq 0}$, and both sides are continuous in (t, t'), so it holds on $\mathbb{Z}_3^2$ by density. The last assertion follows by continuity in t. $\square$

4.5. **Rational isometries of the stable norm.**

Lemma 4.5. *If $G \in \mathrm{GL}_2(\mathbb{Q})$ satisfies $N(Gx) = N(x)$ for all $x \in \mathbb{R}^2$, then $G \in \mathcal{K}$.*

Proof. We show, by comparing the corner formula (3.1) at a primitive vector v and at the primitive vector on the ray of Gv, that G maps primitive vectors to primitive vectors and has determinant ± 1; then G permutes the six primitive vectors of minimal norm, which forces $G \in \mathcal{K}$.

Put $\delta = |\det G|$ and $f(t) = \log\coth(t/2)$ for $t > 0$. Since G is a linear isometry of N, one-sided derivatives transform by $D_{Gh}N(Gx) = D_hN(x)$, hence $J_N(Gx; Gh) = J_N(x; h)$; and $J_N(x; h)$ depends only on the ray through x, since N is homogeneous.

Let v be primitive and write $Gv = c_v w$ with $c_v \in \mathbb{Q}_{>0}$ and w primitive. Choose $D \in \mathbb{Z}_{>0}$ with $\mathbf{A} = DG$ integral. Then $Gv = \mathbf{A}v/D$ and $c_v = \gcd(\mathbf{A}v)/D$, and by the adjugate identity $\gcd(\mathbf{A}v)$ divides $\det \mathbf{A}$; so c_v takes only finitely many values as v ranges over $\mathbb{Z}^2_{\mathrm{prim}}$. Put $l = N(v)$, so that $N(w) = l/c_v$. For h with $\det(v, h) \neq 0$, the corner formula (3.1) applied at v and at w gives

$$\begin{aligned} 4|\det(v,h)|f(l) &= J_N(v;h) = J_N(Gv;Gh) = J_N(w;Gh) \\ &= 4|\det(w,Gh)|f(l/c_v) = 4\frac{\delta}{c_v}|\det(v,h)|f(l/c_v), \end{aligned}$$

using $\det(w, Gh) = \det(Gv, Gh)/c_v = (\det G/c_v)\det(v, h)$. Hence

$$f(l) = \frac{\delta}{c_v} f\Big(\frac{l}{c_v}\Big) \qquad (v \in \mathbb{Z}^2_{\mathrm{prim}},\ l = N(v)). \tag{4.3}$$

Some value c_* of c_v is taken for infinitely many primitive v, hence along a sequence with $l = N(v) \to \infty$. Since $f(t) \sim 2e^{-t}$ as $t \to \infty$, (4.3) gives $e^{-l(1-1/c_*)} \to \delta/c_*$ along this sequence, which forces $c_* = 1$ and then $\delta = 1$.

Now (4.3) reads $c_v f(l) = f(l/c_v)$ for every primitive v, or $\varphi(l) = \varphi(l/c_v)$ with $\varphi(t) = tf(t)$. Both $l = N(v)$ and $l/c_v = N(w)$ are norms of primitive vectors, hence at least $2\,\mathrm{arcosh}(3/2) > 1$. On $[1, \infty)$ the function φ is strictly decreasing: $f'(t) = -1/\sinh t$, and

$$f(t) = \int_t^\infty \frac{du}{\sinh u} < \int_t^\infty \frac{\cosh u}{\sinh^2 u}\,du = \frac{1}{\sinh t}, \qquad \text{so} \qquad \varphi'(t) = f(t) - \frac{t}{\sinh t} < \frac{1-t}{\sinh t} \le 0.$$

Therefore $c_v = 1$ for every primitive v: G maps primitive vectors to primitive vectors, so $G \in \mathrm{Mat}_2(\mathbb{Z})$, and $|\det G| = 1$. Thus $G \in \mathrm{GL}_2(\mathbb{Z})$, and it permutes the primitive vectors of minimal norm. By Lemma 2.1, these are the vectors of label 1, which form the single $\mathcal{K}$-orbit $\{\pm e_1, \pm e_2, \pm(e_1 + e_2)\}$ of e_2 (the only sorted Markoff triple with maximum 1 is $(1, 1, 1)$). A linear map permuting this hexagon is determined by the images of e_1 and e_2; there are six choices for Ge_1 and then exactly two choices for Ge_2 for which $Ge_1 + Ge_2$ again lies in the hexagon, so the group of such maps has order twelve and equals $\mathcal{K}$. □

Lemma 4.5 identifies $\mathcal{K}$ with the group of automorphisms of $H_1(\mathbb{T}, \mathbb{Z}) = \mathbb{Z}^2$ induced by the isometries of $\mathbb{T}$. An isometry of $\mathbb{T}$ maps simple closed geodesics to simple closed geodesics of the same length, so the automorphism it induces preserves N and lies in $\mathcal{K}$. Conversely, T normalizes Γ', and the reflection $z \mapsto -\bar{z}$ normalizes $\mathrm{PSL}_2(\mathbb{Z})$ and hence its commutator subgroup Γ'; so both induce isometries of $\mathbb{T}$. Since $TAT^{-1} = AB$, $TBT^{-1} = A^{-1}$ and the reflection exchanges A and B, the two isometries act on H_1 by r^{-1} and sr^{-1}, which generate $\mathcal{K}$.

Lemma 4.6. *Let $J \le \Gamma$ be a subgroup of finite index and $v' \in \mathbb{Z}^2_{prim}$. The directions of the vectors in the orbit Jv' are dense in $\mathbb{RP}^1$.*

Proof. Let u be primitive, $\gamma_u \in \Gamma$ with $\gamma_u e_1 = u$, and $p_u = \gamma_u T \gamma_u^{-1}$; then $p_u x = x + \det(u, x)\, u$ for all x. Since J has finite index, $p_u^k \in J$ for some $k \geq 1$, and $p_u^{kn} v' = v' + kn \det(u, v')\, u$ converges in direction to u as $n \to \infty$ if $\det(u, v') \neq 0$; if $\det(u, v') = 0$ the direction of u is that of v' itself. Hence every rational direction is a limit of directions of Jv', and rational directions are dense. □

4.6. **Counting lattice points in cusp orbits.** For a subgroup $U \leq \Gamma$ of finite index and $v \in \mathbb{Z}^2_{\text{prim}}$, choose $\gamma_v \in \Gamma$ with $\gamma_v e_1 = v$. The stabilizer of v in Γ is $\{\gamma_v T^k \gamma_v^{-1} : k \in \mathbb{Z}\}$, and the *cusp width* of v relative to U is the least integer $\omega(v; U) \geq 1$ with $\gamma_v T^{\omega(v;U)} \gamma_v^{-1} \in U$; it does not depend on the choice of γ_v. We use the following asymptotic count.

Theorem 4.7 (after Burrin–Nevo–Rühr–Weiss)**.** *Let $U \leq \Gamma$ be a subgroup of finite index containing $-I$, let $v \in \mathbb{Z}^2_{prim}$, and let $\Omega \subset \mathbb{R}^2$ be a bounded star-shaped set with respect to the origin whose radial function is piecewise Lipschitz. Then, as $L \to \infty$,*

$$\#(Uv \cap L\Omega) = \frac{\omega(v; U)}{[\Gamma : U]} \frac{\operatorname{area}(\Omega)}{\zeta(2)} L^2 + o_{U,v,\Omega}(L^2). \tag{4.4}$$

The counting theorem [BNRW20, Theorem 2.7] states, for any non-uniform lattice $\Gamma_0 \subset \mathrm{SL}_2(\mathbb{R})$, any discrete orbit $\Gamma_0 v \subset \mathbb{R}^2 \setminus \{0\}$ and any Ω as above, that $\#(\Gamma_0 v \cap L\Omega) = C_{\Gamma_0,v,\Omega} L^2 + O_{\Gamma_0,v,\Omega,\varepsilon}(L^{q+\varepsilon})$ for every $\varepsilon > 0$, with an exponent $q < 2$ that depends only on the spectral gap of Γ_0. A subgroup U of finite index in Γ is a non-uniform lattice in $\mathrm{SL}_2(\mathbb{R})$, and its orbit Uv is discrete, so the theorem applies to U.

The form of the constant in (4.4) comes from the proof of that theorem in [BNRW20, §5.2]. There the orbit map $g \mapsto gv$ carries the elements of U lying in a region of $\mathrm{SL}_2(\mathbb{R})$, built from $L\Omega$ and a fundamental domain of $U \cap \mathrm{Stab}(v)$ in $\mathrm{Stab}_{\mathrm{SL}_2(\mathbb{R})}(v)$, bijectively onto $Uv \cap L\Omega$ [BNRW20, Lemma 5.7]. The counting theorem of Gorodnik and Nevo, quoted as [BNRW20, Theorem 5.6], shows that the number of these elements is asymptotic to the Haar measure of that region divided by the Haar measure of $U \backslash \mathrm{SL}_2(\mathbb{R})$.

Hence the leading constant is the Lebesgue measure of Ω times

$$\frac{\text{Haar measure of a fundamental domain of } U \cap \mathrm{Stab}(v) \text{ in } \mathrm{Stab}_{\mathrm{SL}_2(\mathbb{R})}(v) \cong \mathbb{R}}{\text{Haar measure of } U \backslash \mathrm{SL}_2(\mathbb{R})},$$

and replacing U by Γ divides the numerator by $\omega(v; U)$ and the denominator by $[\Gamma : U]$. The absolute normalization is fixed by the case $U = \Gamma$, $v = e_1$, where $\Gamma e_1 = \mathbb{Z}^2_{\text{prim}}$ and the number of primitive vectors in $L\Omega$ is $\operatorname{area}(\Omega) L^2/\zeta(2) + O_\Omega(L \log L)$.

We apply the theorem with $\Omega = \mathbb{B}$, whose radial function $\theta \mapsto 1/N(\cos\theta, \sin\theta)$ is Lipschitz, for fixed U and v, and use the error term only in the form $o(L^2)$.

4.7. **Proof of Proposition 4.1.** Let $G \in \mathrm{GL}_2(\mathbb{Q}) \setminus \mathcal{K}$. By Lemma 4.2 it suffices to prove, for one H_G-orbit $H_G v_0$ with $G v_0 =: w_0$ primitive, that

$$\#\{v \in H_G v_0 : \ N(v) = N(Gv) \leq L\} = o_{G,v_0}(L^2). \tag{4.5}$$

The paired action. For $z = (P, P') \in \mathcal{X}^2$ and $h \in H_G$ define

$$z \star h := \big(P * h,\ P' * GhG^{-1}\big).$$

This is a right action of H_G on $\mathcal{X}^2$ by isometries, since $GhG^{-1} \in \Gamma$ for $h \in H_G$. Let

$$J := \{h \in H_G : \ z \star h \equiv z \pmod 9 \text{ for all } z \in \mathcal{X}^2\},$$

the kernel of the induced action of H_G on the finite set $\mathcal{X}(9)^2$. It is a normal subgroup of finite index in H_G, it contains $-I$, and it is finitely generated.

Since J is normal in H_G, the orbit $H_G v_0$ is the finite union of the J-orbits $Jh_i v_0 = h_i J v_0$, h_i running over representatives of $J\backslash H_G$; so it suffices to prove (4.5) with $H_G v_0$ replaced by $h_0 J v_0$ for a fixed $h_0 \in H_G$.

Define

$$z_0 := \big(o * h_0,\ o * G h_0 G^{-1}\big) \in \mathcal{X}^2, \qquad F(P, P') := Q_{v_0}(P) - Q_{w_0}(P') \in \mathbb{Z}[a, b, c, a', b', c'].$$

For $j \in J$, the properties (4.2) of the polynomials Q_v give

$$F(z_0 \star j) = Q_{h_0 j v_0}(o) - Q_{(G h_0 j G^{-1}) w_0}(o) = m(h_0 j v_0) - m(G h_0 j v_0), \tag{4.6}$$

where $G h_0 j G^{-1} \in \Gamma$ because $h_0 j \in H_G$. By the length formula (1.2), $F(z_0 \star j) = 0$ if and only if $N(v) = N(Gv)$ for $v = h_0 j v_0$.

The zero-one law applies. Let $Y \subset \mathcal{X}^2$ be the closure of $z_0 \star J$ and μ_Y the measure of Lemma 4.3. Fix a finite generating set of J. For each j in it, the map $g_j : z \mapsto z \star j$ fixes every point of $\mathcal{X}^2$ modulo 9, so Lemma 4.4 provides, on each product ball, a restricted power series $\Phi_{g_j}(t, x)$ interpolating the iterates of g_j. This series is continuous, and $\Phi_{g_j}(t, z) \in Y$ for $z \in Y$ because Y is closed and J-stable.

For $z \in Y$ the function $t \mapsto F(\Phi_{g_j}(t, z))$ is obtained by substituting the restricted power series $\Phi_{g_j}(t, z) \in \mathbb{Z}_3\langle t\rangle^4$ into the polynomial F, whose variables c, c' are themselves restricted power series in the normalized coordinates; so it lies in $\mathbb{Z}_3\langle t\rangle$. Thus Lemma 4.3 applies, and $\mu_Y(F = 0) \in \{0, 1\}$.

If $\mu_Y(F = 0) = 1$, then F vanishes on $Y \supset z_0 \star J$, so by (4.6) $N(Gv) = N(v)$ for every v in the orbit $h_0 J v_0$. By Lemma 4.6 the directions of these v are dense in $\mathbb{RP}^1$; since $N \circ G$ and N are continuous and homogeneous, $N(Gx) = N(x)$ for all $x \in \mathbb{R}^2$, and Lemma 4.5 gives $G \in \mathcal{K}$, a contradiction. Therefore $\mu_Y(F = 0) = 0$, and since the closed sets $\{F \equiv 0 \bmod 3^n\}$ decrease to $\{F = 0\}$,

$$\mu_Y\big(F \equiv 0 \pmod{3^n}\big) \longrightarrow 0 \qquad (n \to \infty). \tag{4.7}$$

From states to cusp orbits. Fix $n \geq 1$ and let

$$S := \{j \in J :\ z_0 \star j \equiv z_0 \pmod{3^n}\},$$

the stabilizer in J of the state of z_0 modulo 3^n; it has finite index in J and contains $-I$. The map $Sj \mapsto z_0 \star j \bmod 3^n$ identifies the finite set $Y(3^n)$ of states of Y with the set $S\backslash J$ of cosets Sj, as right J-sets, and μ_Y induces the uniform measure on $S\backslash J$. Let $\mathcal{E} \subset S\backslash J$ be the set of cosets Sj with $F(z_0 \star j) \equiv 0 \pmod{3^n}$; this condition depends only on the state, so $\mathcal{E}$ is well defined, and

$$\frac{|\mathcal{E}|}{[J : S]} = \mu_Y\big(F \equiv 0 \pmod{3^n}\big). \tag{4.8}$$

Choose $\gamma_0 \in \Gamma$ with $\gamma_0 e_1 = v_0$, let $\omega = \omega(v_0; J)$ be the cusp width of v_0 relative to J, and put $p = \gamma_0 T^\omega \gamma_0^{-1} \in J$. Then $\mathrm{Stab}_J(v_0) = \langle p\rangle$, since $\mathrm{Stab}_\Gamma(v_0) = \{\gamma_0 T^k \gamma_0^{-1}\}$. The element p fixes v_0 and GpG^{-1} fixes w_0, so (4.6) gives $F(z_0 \star jp) = F(z_0 \star j)$ for all j: the function $j \mapsto F(z_0 \star j)$ is invariant under right multiplication by p. Consequently $\mathcal{E}$ is a union of orbits of the right action of $\langle p\rangle$ on $S\backslash J$; we call these orbits *cycles* and write $\ell(Sj)$ for the length of the cycle through Sj.

The map $j \mapsto h_0 j v_0$ from J onto the orbit $h_0 J v_0$ has fibres the cosets $j\langle p\rangle$, and the double cosets $Sj\langle p\rangle$ partition J. The double coset of j corresponds to the cycle through Sj, and its image under this map is the set

$$h_0 Sj\langle p\rangle v_0 = h_0 Sj v_0 = U_S\, v_j, \qquad U_S := h_0 S h_0^{-1}, \quad v_j := h_0 j v_0,$$

the orbit of v_j under the finite-index subgroup $U_S \leq \Gamma$, which contains $-I$ and has index $[\Gamma : U_S] = [\Gamma : J][J : S]$.

We compute the cusp width of v_j relative to U_S. Since $v_j = h_0 j \gamma_0 e_1$, it is the least $k \geq 1$ with $h_0 j \gamma_0 T^k \gamma_0^{-1} j^{-1} h_0^{-1} \in h_0 S h_0^{-1}$, that is, with $j(\gamma_0 T^k \gamma_0^{-1}) j^{-1} \in S$. As $S \subset J$ and $j \in J$, this forces $\gamma_0 T^k \gamma_0^{-1} \in J$, that is, $\omega \mid k$; for $k = \omega k'$ the condition becomes $jp^{k'}j^{-1} \in S$, that is $Sjp^{k'} = Sj$, which holds exactly when $\ell(Sj)$ divides k'. Hence

$$\omega(v_j; U_S) = \omega \cdot \ell(Sj). \tag{4.9}$$

Conclusion. Let $v = h_0 j v_0$ satisfy $N(v) = N(Gv) \leq L$. Then $F(z_0 \star j) = 0$, so $Sj \in \mathcal{E}$, and $v \in U_S v_j$. Choosing one representative j for each cycle contained in $\mathcal{E}$, we obtain

$$\#\{v \in h_0 J v_0 : \ N(v) = N(Gv) \leq L\} \leq \sum_{\text{cycles in } \mathcal{E}} \#(U_S v_j \cap L\mathbb{B}).$$

By Theorem 4.7 and (4.9), each term is

$$\frac{\omega\, \ell(Sj)}{[\Gamma : J][J : S]} \frac{\operatorname{area}(\mathbb{B})}{\zeta(2)} L^2 + o_{U_S, v_j}(L^2),$$

and the cycle lengths add up to $|\mathcal{E}|$. With (4.8),

$$\limsup_{L \to \infty} \frac{1}{L^2} \#\{v \in h_0 J v_0 : \ N(v) = N(Gv) \leq L\} \leq \frac{\omega}{[\Gamma : J]} \frac{\operatorname{area}(\mathbb{B})}{\zeta(2)} \mu_Y(F \equiv 0 \pmod{3^n})$$

for every n. Letting $n \to \infty$ and using the decay (4.7) proves the bound (4.5) for the orbit $h_0 J v_0$, and Proposition 4.1 follows by summing over the finitely many orbits. □

5. Collision pairs lie on rational graphs

From now on we work inside the sector $\mathcal{S}$ of §2 and with a large length L. Recall that a *collision pair* is an ordered pair (v, w) of distinct primitive vectors of $\mathcal{S}$ with $N(v) = N(w) \leq L$, and that its *collision vector* is $z = (v, w) \in \mathbb{Z}^4$. If $w = kv$ with $k \in \mathcal{K}$, then $w = v$ by Lemma 2.1(b); so the two members of a collision pair are never related by an element of $\mathcal{K}$.

Throughout this section and the next, κ and C_{sep} denote the constants of Proposition 3.1(b) and Lemma 3.5(c), and C_{exc} the constant of Definition 3.4.

5.1. The rank reduction.

Fix a parameter $A_0 \geq 1$ and put

$$B_0 := \max\Big\{A_0, \ \frac{C_{\text{sep}} A_0 + 4}{\kappa}\Big\}, \qquad \text{so that} \qquad \kappa B_0 \geq C_{\text{sep}} A_0 + 4; \tag{5.1}$$

define the *coarse level* H_{c}, the *fine level* H_{f} and the *normal bound* M_1 by

$$H_{\text{c}} := \lceil A_0 \log(2 + L) \rceil, \qquad H_{\text{f}} := \lceil B_0 \log(2 + L) \rceil, \qquad M_1 := C_1 L^3,$$

where C_1 is the constant appearing in (5.4) below. If A_0 is sufficiently large, then $H_{\text{c}} \geq C_{\text{exc}} \log(2 + M_1)$ for all $L \geq 1$, so the (M_1, H_{c})-excision of Definition 3.4 is defined.

We call a primitive vector of $\mathcal{S}$ with $N \leq L$ *deleted* if it is excised by the (M_1, H_{c})-excision or lies on a ray of the fan $\mathcal{F}_{H_{\text{f}}}$, and *retained* otherwise. A collision pair is retained if both its members are.

Lemma 5.1. *Let A_0 be sufficiently large.*

(a) The number of deleted vectors is at most $CL^2/A_0 + C_{A_0} L \log^2 L$, with C independent of A_0.

(b) If L is sufficiently large in terms of A_0, then for every ordered pair (Σ, Σ') of Farey sectors of level H_{f}, the collision vectors of the retained collision pairs $(v, w) \in \Sigma \times \Sigma'$ span a subspace of $\mathbb{R}^4$ of dimension at most two.

Proof. (a) By Lemma 3.5(b), the (M_1, H_{c})-excision deletes at most

$$C\,\frac{\log(2+M_1)}{H_{\mathrm{c}}}\,L^2 + CL\log^2(2+M_1) + CH_{\mathrm{c}}^2$$

vectors, and $\log(2+M_1) \le C\log(2+L)$, so this is at most $CL^2/A_0 + C_{A_0}L\log^2 L$. The rays of $\mathcal{F}_{H_{\mathrm{f}}}$ carry at most $CH_{\mathrm{f}}^2 \le C_{A_0}\log^2 L$ primitive vectors.

(b) Fix (Σ, Σ'). Let λ_Σ be the inward supporting functional at one boundary vector of Σ, as in Proposition 3.1(b), and similarly $\lambda_{\Sigma'}$. For a primitive $v \in \Sigma$ with $N(v) \le L$ choose a supporting functional γ at v; then $N(v) = \gamma(v)$ and $|\gamma - \lambda_\Sigma| \le Ce^{-\kappa H_{\mathrm{f}}}$ by Proposition 3.1(b), so

$$|N(v) - \lambda_\Sigma(v)| \le |\gamma - \lambda_\Sigma|\,|v| \le \eta := CLe^{-\kappa H_{\mathrm{f}}}, \tag{5.2}$$

using $|v| \le N(v)/c \le CL$. The same holds for $w \in \Sigma'$ with $\lambda_{\Sigma'}$.

Consider the linear form

$$\ell := (\lambda_\Sigma, -\lambda_{\Sigma'}) \ \text{ on } \mathbb{R}^4, \qquad \ell(v,w) = \lambda_\Sigma(v) - \lambda_{\Sigma'}(w).$$

Its norm satisfies $c \le |\ell| \le C$. If $z = (v, w)$ is the collision vector of a retained collision pair in $\Sigma \times \Sigma'$, then $N(v) = N(w)$ and (5.2) give

$$|z| \le CL, \qquad |\ell(z)| \le 2\eta. \tag{5.3}$$

Dimension four is impossible. Suppose $z_1, \dots, z_4$ are linearly independent collision vectors as above. Their determinant is a nonzero integer. Write $z_i = t_i\hat{\ell} + z_i^{\perp}$ with $\hat{\ell} = \ell/|\ell|$ (viewing ℓ as a vector of $\mathbb{R}^4$), $t_i = \ell(z_i)/|\ell|$ and $z_i^{\perp} \perp \hat{\ell}$. In the expansion of the determinant by multilinearity, the term with no factor $\hat{\ell}$ vanishes because the $z_i^{\perp}$ lie in a three-dimensional space, and so do the terms with two or more factors $\hat{\ell}$. Each of the four remaining terms is bounded by $|t_i|$ times a three-dimensional volume of vectors of norm at most CL. Hence

$$1 \le |\det(z_1, \dots, z_4)| \le 4 \cdot \frac{2\eta}{|\ell|}(CL)^3 \le CL^4 e^{-\kappa H_{\mathrm{f}}}.$$

Since $e^{-\kappa H_{\mathrm{f}}} \le (2+L)^{-\kappa B_0}$ and $\kappa B_0 > 4$ by (5.1), this is impossible for large L.

Dimension three is impossible. Suppose the span W of the collision vectors is three-dimensional. Since W is spanned by integral vectors, it has a primitive integral normal vector $\nu = (\alpha, \beta)$, $\alpha, \beta \in \mathbb{Z}^2$, unique up to sign. Choose linearly independent collision vectors $z_1, z_2, z_3 \in W$. Under the identification $\Lambda^3\mathbb{R}^4 \cong \mathbb{R}^4$, the exterior product $z_1 \wedge z_2 \wedge z_3$ is an integral vector orthogonal to W, hence a nonzero integer multiple of ν. Therefore the three-dimensional volume $\mathrm{vol}_3(z_1, z_2, z_3)$ is at least $|\nu|$, while it is at most $(CL)^3$; so

$$|\nu| \le C_1 L^3 = M_1. \tag{5.4}$$

Every $z \in W$ can be written $z = \sum c_i z_i$, and by Cramer's rule

$$|c_i| \le |z|\,\frac{\mathrm{vol}_2(z_k, z_m)}{\mathrm{vol}_3(z_1, z_2, z_3)} \le \frac{|z|(CL)^2}{|\nu|}, \qquad \{k, m\} = \{1,2,3\} \setminus \{i\}.$$

With (5.3) this gives

$$|\ell(z)| \le \varepsilon|\ell|\,|z| \quad (z \in W), \qquad \varepsilon := \frac{CL^3 e^{-\kappa H_{\mathrm{f}}}}{|\nu|} \le CL^3 e^{-\kappa H_{\mathrm{f}}}. \tag{5.5}$$

Decompose $\ell = \ell_W + \ell_\perp$ with $\ell_W \in W$ and $\ell_\perp \in \mathbb{R}\nu$; then $|\ell_W| \le \varepsilon|\ell|$, so the angle between ℓ and the line $\mathbb{R}\nu$ is at most $\pi\varepsilon/2$, and $\nu/|\nu| = \pm\ell/|\ell| + O(\varepsilon)$ in $\mathbb{R}^4$. Comparing second blocks gives

$$\frac{\beta}{|\nu|} = \mp\frac{\lambda_{\Sigma'}}{|\ell|} + O(\varepsilon).$$

Since $|\lambda_{\Sigma'}|/|\ell| \ge c$ and $\varepsilon \to 0$ as $L \to \infty$ (because $\kappa B_0 > 3$), for large L we obtain $\beta \ne 0$, $|\beta| \ge c|\nu|$, and

$$\measuredangle(\lambda_{\Sigma'}, \{\beta, -\beta\}) \le C\varepsilon \le CL^3 e^{-\kappa H_{\mathrm{f}}}. \tag{5.6}$$

Now let $w \in \Sigma'$ be the second member of one of the retained collision pairs spanning W, and let γ be a supporting functional at w. By Proposition 3.1(b), $|\gamma - \lambda_{\Sigma'}| \le Ce^{-\kappa H_{\mathrm{f}}}$, so by (5.6) $\measuredangle(\gamma, \{\beta, -\beta\}) \le CL^3 e^{-\kappa H_{\mathrm{f}}}$. On the other hand β is a nonzero integral form with $|\beta| \le |\nu| \le M_1$, and w is retained by the (M_1, H_{c})-excision, so Lemma 3.5(c) gives $\measuredangle(\gamma, \{\beta, -\beta\}) \ge ce^{-C_{\mathrm{sep}} H_{\mathrm{c}}}$. Combining the two bounds gives

$$ce^{-C_{\mathrm{sep}} H_{\mathrm{c}}} \le CL^3 e^{-\kappa H_{\mathrm{f}}}, \qquad \text{that is,} \qquad (2+L)^{\kappa B_0 - C_{\mathrm{sep}} A_0 - 3} \le C,$$

which contradicts the choice (5.1) of B_0 for large L. Hence $\dim W \le 2$. □

5.2. **Graph pieces, degenerate pieces, and two complexity bounds.** Keep the notation of Lemma 5.1, and let L be large enough for Lemma 5.1(b). Fix an ordered pair (Σ, Σ') of sectors of level H_{f}, let $W \subset \mathbb{R}^4$ be the span of the collision vectors of the retained collision pairs in $\Sigma \times \Sigma'$, a rational subspace of dimension at most two, and let $\pi_1, \pi_2 : \mathbb{R}^4 \to \mathbb{R}^2$ be the two coordinate projections. There are three cases.

(a) Suppose $\dim W \le 1$. Two proportional collision vectors (v, w) and $t(v, w)$ with v, tv primitive in $\mathcal{S}$ have $t = 1$; so there is at most one retained collision pair in $\Sigma \times \Sigma'$.

(b) Suppose $\dim W = 2$ and one of $\pi_1(W)$, $\pi_2(W)$ is a line. Say $\pi_1(W)$ is a line; it contains at most one primitive vector v of $\mathcal{S}$, and the partners w of v lie on the curve $\{N = N(v)\}$. The unit squares centred at the lattice points of this curve are disjoint and contained in the region $\{N(v) - C \le N \le N(v) + C\}$, which has area $O(L+1)$. Hence there are at most $C(L+1)$ retained collision pairs in $\Sigma \times \Sigma'$. The same bound holds if $\pi_2(W)$ is a line.

(c) Suppose $\dim W = 2$ and both projections are two-dimensional. Then W is the graph of the linear map $G = \pi_2 \circ (\pi_1|_W)^{-1}$, which lies in $\mathrm{GL}_2(\mathbb{Q})$ because W is rational, and every retained collision pair in $\Sigma \times \Sigma'$ has the form (v, Gv). We then call the set of these pairs the *graph piece* of (Σ, Σ') and $G = G_{\Sigma,\Sigma'}$ its matrix; $G \notin \mathcal{K}$ by the remark at the beginning of the section.

In cases (a) and (b) we speak of a *degenerate piece*. The fan of level H_{f} has $O(H_{\mathrm{f}}^2)$ sectors, hence $O(H_{\mathrm{f}}^4) = O_{A_0}(\log^4 L)$ ordered sector pairs; so

the degenerate pieces contain $O_{A_0}(L \log^4 L) = o_{A_0}(L^2)$ retained collision pairs in total. (5.7)

For $G \in \mathrm{GL}_2(\mathbb{Q})$ we define the *input lattice* and the *output lattice*

$$\Lambda_{\mathrm{in}}(G) := \mathbb{Z}^2 \cap G^{-1}\mathbb{Z}^2, \qquad \Lambda_{\mathrm{out}}(G) := G\Lambda_{\mathrm{in}}(G) = \mathbb{Z}^2 \cap G\mathbb{Z}^2;$$

thus $\Lambda_{\mathrm{in}}(G)$ is the set of integral v with Gv integral, and both lattices have finite index in $\mathbb{Z}^2$. We write

$$I_{\mathrm{in}}(G) := [\mathbb{Z}^2 : \Lambda_{\mathrm{in}}(G)], \qquad I_{\mathrm{out}}(G) := [\mathbb{Z}^2 : \Lambda_{\mathrm{out}}(G)], \qquad I(G) := \max\{I_{\mathrm{in}}(G), I_{\mathrm{out}}(G)\}$$

for the two indices and the *index* of G, $h(G) := \max_{i,k} |G_{ik}|$ for its *entry height*, and $q(G)$ for the least positive integer with $q(G)G$ integral.

Lemma 5.2. *Let $G \in \mathrm{GL}_2(\mathbb{Q})$ and $R \geq 1$.*

(a) Let $\mathcal{P}$ be a set of pairs (v, Gv) with $v \in \Lambda_{\mathrm{in}}(G)$, $|v| \leq R$ and $|Gv| \leq R$, and assume that $\mathcal{P}$ contains two pairs whose first members are linearly independent. Then $|\mathcal{P}| \leq CR^2/I(G)$.

(b) The integer $q(G)$ divides $I_{\mathrm{in}}(G)$, and $|\det G| = I_{\mathrm{out}}(G)/I_{\mathrm{in}}(G)$; in particular $|\det G| \leq I(G)$.

Proof. (a) Let $(v_1, Gv_1), (v_2, Gv_2) \in \mathcal{P}$ with v_1, v_2 independent. The lattice $\Lambda_{\mathrm{in}}(G)$ contains the lattice spanned by v_1, v_2, whose covering radius is at most $|v_1| + |v_2| \leq 2R$; so the covering radius of $\Lambda_{\mathrm{in}}(G)$ is at most $2R$, and its Voronoi cells have diameter at most $4R$. The Voronoi cells centred at the first members v of the pairs in $\mathcal{P}$ are pairwise disjoint, have area $I_{\mathrm{in}}(G)$, and lie in the disc of radius $5R$; hence $|\mathcal{P}| \leq 25\pi R^2/I_{\mathrm{in}}(G)$.

The same argument with the second members $Gv \in \Lambda_{\mathrm{out}}(G)$, which include the independent vectors Gv_1, Gv_2, gives $|\mathcal{P}| \leq 25\pi R^2/I_{\mathrm{out}}(G)$.

(b) The index $n = I_{\mathrm{in}}(G)$ annihilates $\mathbb{Z}^2/\Lambda_{\mathrm{in}}(G)$, so $n\mathbb{Z}^2 \subset \Lambda_{\mathrm{in}}(G)$ and $nG\mathbb{Z}^2 \subset \mathbb{Z}^2$; thus nG is integral and $q(G) \mid n$. The determinant identity expresses that G multiplies covolumes by $|\det G|$. □

Lemma 5.3. *Let $G \in \mathrm{GL}_2(\mathbb{Q})$ with $h(G) > 10(1 + |\det G|)$. Then the set of lines*

$$\{\mathbb{R}x : \ x \in \mathbb{R}^2 \setminus \{0\}, \ N(Gx) = N(x)\}$$

has exactly two elements. In particular, at most two primitive vectors $v \in \mathcal{S}$ satisfy $N(Gv) = N(v)$.

Proof. Step 1: the case $|\operatorname{tr} M| > 1 + \det M$. Let M be a real 2×2 matrix with $\det M > 0$ and $|\operatorname{tr} M| > 1 + \det M$. Then $N(Mx) = N(x)$ holds for exactly two directions x. Indeed, the discriminant $(\operatorname{tr} M)^2 - 4\det M > (1 - \det M)^2 \geq 0$ shows that M has two distinct real eigenvalues of the same sign. Replacing M by $-M$, which changes neither $\det M$ nor the equation since N is even, we may take them positive, $\lambda_1 > \lambda_2 > 0$. Then $\lambda_1 + \lambda_2 > 1 + \lambda_1\lambda_2$ means $(\lambda_1 - 1)(1 - \lambda_2) > 0$, so $\lambda_1 > 1 > \lambda_2$.

Let x_1, x_2 be eigenvectors for λ_1, λ_2. On the eigenlines, $N(Mx_1) = \lambda_1 N(x_1) \neq N(x_1)$ and $N(Mx_2) = \lambda_2 N(x_2) \neq N(x_2)$. Every other direction is represented by $\varepsilon t x_1 + x_2$ with $t > 0$ and $\varepsilon = \pm 1$.

Fix ε and put $f(t) := N(\varepsilon t x_1 + x_2)$, a positive, strictly convex, locally Lipschitz function of $t \geq 0$ with $f(t)/t \to N(x_1)$ as $t \to \infty$; let $t_* \geq 0$ be its minimizer. With $k := \lambda_1/\lambda_2 > 1$ the equation $N(M(\varepsilon t x_1 + x_2)) = N(\varepsilon t x_1 + x_2)$ reads $F(t) = 0$ with

$$F(t) := \lambda_2 f(kt) - f(t).$$

For $t \leq t_*/k$ we have $t \leq kt \leq t_*$, so $f(kt) \leq f(t)$ and $F(t) \leq (\lambda_2 - 1)f(t) < 0$. For almost every $t > t_*/k$ we have $F'(t) = \lambda_1 f'(kt) - f'(t) > 0$: if $t \leq t_*$ then $f'(t) \leq 0 < f'(kt)$, and if $t > t_*$ then $f'(kt) \geq f'(t) > 0$ because f' is nondecreasing, so $F'(t) \geq (\lambda_1 - 1)f'(t) > 0$. Hence F is strictly increasing on $[t_*/k, \infty)$, with $F(t_*/k) < 0$ and $F(t)/t \to (\lambda_1 - 1)N(x_1) > 0$. So F has exactly one zero for each ε, and $N(Mx) = N(x)$ has exactly two solution directions.

Step 2: reduction to Step 1. Let G be as in the lemma. If $\det G < 0$, replace G by $M := sG$, where $s = \left(\begin{smallmatrix} -1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ is the generator of $\mathcal{K}$ of determinant -1; since $N \circ s = N$ this does not change the solution set, and $h(G) \leq 2h(M)$ because $G = sM$ and s has entries in $\{0, \pm 1\}$ with at most two nonzero entries per row. Otherwise put $M := G$.

Write $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, $\delta := \det M > 0$, $\sigma := \operatorname{tr} M = a + d$ and $\tau := \operatorname{tr}(rM) = c - b + d$, where $r = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 1 \end{smallmatrix}\right)$ is the generator of $\mathcal{K}$ of order six; then $\operatorname{tr}(r^2M) = \tau - \sigma$, and M, rM, r^2M all have determinant δ and the same solution set, since $N \circ r = N$.

If one of $|\sigma|, |\tau|, |\tau - \sigma|$ exceeds $1 + \delta$, Step 1 applies to the corresponding matrix and we are done. Otherwise $|\sigma|, |\tau| \le 1 + \delta$. Substituting $a = \sigma - d$ and $c = \tau + b - d$ into $\delta = ad - bc$ gives

$$b^2 - bd + d^2 = \sigma d - \tau b - \delta \le (1 + \delta)(|b| + |d|) \le \sqrt{2}\,(1 + \delta)\,|(b, d)|,$$

and $b^2 - bd + d^2 \ge |(b, d)|^2/2$, so $|(b, d)| \le 2\sqrt{2}\,(1 + \delta)$. Then

$$|a| \le |\sigma| + |d| \le (1 + 2\sqrt{2})(1 + \delta), \qquad |c| \le |\tau| + |b| + |d| \le (1 + \delta) + \sqrt{2}\,|(b, d)| \le 5(1 + \delta),$$

so $h(M) \le 5(1 + \delta)$ and $h(G) \le 10(1 + \delta) = 10(1 + |\det G|)$, contrary to the hypothesis. Hence the first case occurs. □

5.3. **Summary of the reduction.** We apply Lemmas 5.2 and 5.3 to the graph pieces of the $O_{A_0}(\log^4 L)$ sector pairs of level $H_{\mathfrak{f}}$. A graph piece contains two pairs with independent first members (its collision vectors span a plane), and every pair (v, Gv) in it satisfies $|v|, |Gv| \le CL$; so by Lemma 5.2(a) it contains at most $CL^2/I(G)$ pairs. Hence the graph pieces with $I(G) > (\log L)^5$ contain $O_{A_0}(L^2/\log L)$ retained collision pairs in total. By Lemma 5.3, a graph piece with $h(G) > 10(1 + |\det G|)$ contains at most two pairs, so these pieces contain $O_{A_0}(\log^4 L)$ pairs in total. Together with the bound (5.7) for the degenerate pieces we conclude:

(5.8) all but $o_{A_0}(L^2)$ of the retained collision pairs lie on graph pieces whose matrix $G \in \mathrm{GL}_2(\mathbb{Q}) \setminus \mathcal{K}$ satisfies

$$I(G) \le (\log L)^5 \qquad \text{and} \qquad h(G) \le 10(1 + |\det G|) \le 10(1 + I(G)).$$

6. Summation over the graphs and proof of the main theorem

Proposition 4.1 bounds the pairs on a single graph with no uniformity in G, while the reduction (5.8) leaves up to $(\log L)^{O(1)}$ graphs. We now sort the graphs by their index into dyadic bands and show that each band contributes a summable amount.

6.1. **Candidates in a band.** For $j \ge 1$ put $Q_j := 2^j$ and let $\mathcal{G}_j$ be the set of $G \in \mathrm{GL}_2(\mathbb{Q}) \setminus \mathcal{K}$ with

$$Q_j < I(G) \le 2Q_j \qquad \text{and} \qquad h(G) \le 10(1 + I(G)).$$

We call $\mathcal{G}_j$ the *band* j; it does not depend on L. For $G \in \mathcal{G}_j$ we have $q(G) \le I_{\mathrm{in}}(G) \le 2Q_j$ by Lemma 5.2(b) and $h(G) \le 30Q_j$, so the integral matrix $\mathbf{A} = q(G)G$ has entries of absolute value at most $60Q_j^2$. Hence

$$|\mathcal{G}_j| \le CQ_j^9. \tag{6.1}$$

For $G = \mathbf{A}/q \in \mathcal{G}_j$ the *graph plane* $W_G := \{(x, Gx) : x \in \mathbb{R}^2\} \subset \mathbb{R}^4$ has a basis consisting of the integral vectors

$$b_i(G) := (qe_i, \mathbf{A}e_i)\ (i = 1, 2), \qquad |b_i(G)| \le R_j := CQ_j^2. \tag{6.2}$$

If $G \ne G'$ are in $\mathcal{G}_j$ and $W_G + W_{G'}$ is three-dimensional, then its primitive integral normal vector ν is a divisor, in $\mathbb{Z}^4$, of the exterior product of three independent vectors among $b_1(G), b_2(G), b_1(G'), b_2(G')$, so

$$|\nu| \le M(j) := CQ_j^6. \tag{6.3}$$

6.2. **One excision for all bands.** Fix a constant $B_1 \geq 1$ with

$$\kappa B_1 \geq C_{\mathrm{sep}} + 1, \tag{6.4}$$

and for $j \geq 1$ put

$$K_{\mathrm{c}}(j) := j^3, \qquad K_{\mathrm{f}}(j) := \lceil B_1 j^3 \rceil.$$

Let j_0 be a large integer, to be chosen at the end, and let $j_{\max} := \lceil 5 \log_2 \log L \rceil$, so that every matrix G allowed by the reduction (5.8) with $I(G) > 2^{j_0}$ lies in a band $\mathcal{G}_j$ with $j_0 \leq j \leq j_{\max}$.

For every j with $j_0 \leq j \leq j_{\max}$ we perform the $(M(j), K_{\mathrm{c}}(j))$-excision of Definition 3.4 and we excise the rays of the fan $\mathcal{F}_{K_{\mathrm{f}}(j)}$; the condition $K_{\mathrm{c}}(j) \geq C_{\mathrm{exc}} \log(2 + M(j))$ follows from $j^3 \geq C_{\mathrm{exc}}(6j \log 2 + C)$ and holds when j_0 is large.

From now on a vector is called *retained* if it is retained in the sense of §5 and is not excised by any of these band excisions, and *deleted* otherwise. For each j, the levels $K_{\mathrm{c}}(j), K_{\mathrm{f}}(j)$ and the j-th excision do not depend on L; the range $j_0 \leq j \leq j_{\max}$ does.

Lemma 6.1. *The following statements hold.*

(a) The band excisions delete at most $CL^2/j_0 + C\,L(\log\log L)^3 + C(\log\log L)^7$ primitive vectors of $\mathcal{S}$ with $N \leq L$, with C independent of j_0.

(b) If v is retained, γ is a supporting functional at v, $j_0 \leq j \leq j_{\max}$, and β is an integral form with $0 < |\beta| \leq M(j)$, then $\measuredangle(\gamma, \{\beta, -\beta\}) \geq ce^{-C_{\mathrm{sep}} j^3}$.

(c) Lemma 5.1(b) and the dichotomy of §5.2 remain valid for the retained collision pairs in the new sense: in every sector pair of the original fine fan $\mathcal{F}_{H_{\mathrm{f}}}$ they form a subset of a degenerate piece or of a graph piece, with the same matrix G.

Proof. (a) By Lemma 3.5(b) with $\log(2 + M(j)) \leq Cj$ and $K_{\mathrm{c}}(j) = j^3$, the j-th excision deletes at most $C(j^{-2}L^2 + Lj^2 + j^6)$ vectors, and the rays of $\mathcal{F}_{K_{\mathrm{f}}(j)}$ carry at most $CK_{\mathrm{f}}(j)^2 \leq Cj^6$ primitive vectors. Summing over $j_0 \leq j \leq j_{\max}$ gives the bound, since $\sum_{j \geq j_0} j^{-2} \leq 2/j_0$. (b) is Lemma 3.5(c) for the $(M(j), K_{\mathrm{c}}(j))$-excision. (c) Removing vectors can only shrink the set of collision pairs in a sector pair, hence the span of their collision vectors. □

6.3. **A population estimate.**

Lemma 6.2. *Let V be a finite set of integral vectors in the disc $|x| \leq R$, and let $D_V := \max_{v,v' \in V} |\det(v, v')|$. If $D_V > 0$ then*

$$|V| \leq 4D_V + CR + C. \tag{6.5}$$

If $D_V = 0$ and all vectors of V are primitive, then $|V| \leq 2$.

Proof. Let $v_1, v_2 \in V$ attain the maximum $D_V > 0$. For $v \in V$ write $v = xv_1 + yv_2$; then $|x|D_V = |\det(v, v_2)| \leq D_V$ and $|y|D_V = |\det(v_1, v)| \leq D_V$, so $|x|, |y| \leq 1$. Thus V lies in the parallelogram $\{xv_1 + yv_2 : |x|, |y| \leq 1\}$, of area $4D_V$ and perimeter $O(R)$, and the unit squares centred at the points of V are disjoint and lie within distance 1 of it; this gives (6.5). If $D_V = 0$, all vectors of V lie on one line through the origin, which contains at most two primitive vectors. □

6.4. **At most one large band piece in a sector pair.** Fix a band j with $j_0 \le j \le j_{\max}$ and an ordered pair (Σ, Σ') of Farey sectors of level $K_{\mathrm{f}}(j)$. For $G \in \mathcal{G}_j$ let

$$\mathcal{P}_G(\Sigma, \Sigma') := \{v \in \Sigma : \ v \text{ and } Gv \text{ are retained}, \ Gv \in \Sigma', \ N(v) = N(Gv) \le L\}$$

and $P_G(\Sigma, \Sigma') := |\mathcal{P}_G(\Sigma, \Sigma')|$. We call the set of pairs (v, Gv) with $v \in \mathcal{P}_G(\Sigma, \Sigma')$ the *band piece* (G, Σ, Σ'), and $P_G(\Sigma, \Sigma')$ its *population*; we write P_G when the sector pair is fixed. (Here retained vectors are primitive vectors of $\mathcal{S}$ by definition.)

The band piece is *large* if $P_G \ge L^2/Q_j^{20}$ and *small* otherwise. Since $Q_j \le 2^{j_{\max}} \le 2(\log L)^5$, the threshold satisfies

$$\frac{L^2}{Q_j^{20}} \ge \frac{L^2}{2^{20}(\log L)^{100}}, \tag{6.6}$$

so that $L = o(L^2/Q_j^{20})$ as $L \to \infty$, uniformly in j.

Lemma 6.3. *If j_0 is sufficiently large (depending only on N) and L is sufficiently large, then for every j with $j_0 \le j \le j_{\max}$ and every ordered sector pair (Σ, Σ') of level $K_{\mathrm{f}}(j)$, at most one $G \in \mathcal{G}_j$ has a large band piece (G, Σ, Σ').*

Proof. As in the proof of Lemma 5.1, let λ_Σ and $\lambda_{\Sigma'}$ be inward supporting functionals at boundary vectors of Σ and Σ', $\ell := (\lambda_\Sigma, -\lambda_{\Sigma'})$, and $\eta_j := CLe^{-\kappa K_{\mathrm{f}}(j)}$; then (5.2) holds at level $K_{\mathrm{f}}(j)$, that is, $|N(v) - \lambda_\Sigma(v)| \le \eta_j$ for $v \in \Sigma$ with $N(v) \le L$, and similarly in Σ'.

A large band piece pins down ℓ on its graph plane. Let (G, Σ, Σ') be large and $V := \mathcal{P}_G(\Sigma, \Sigma')$, a set of $P_G \ge L^2/Q_j^{20}$ primitive vectors in the disc $|x| \le CL$. By Lemma 6.2 and (6.6), $D_V \ge P_G/8$ for large L, so there are $v_1, v_2 \in V$ with $|\det(v_1, v_2)| \ge L^2/(8Q_j^{20})$.

Consider the linear form $\theta_G := \lambda_\Sigma - \lambda_{\Sigma'} \circ G$ on $\mathbb{R}^2$, so that $\ell(x, Gx) = \theta_G(x)$. For $v \in V$, $N(v) = N(Gv)$ gives $|\theta_G(v)| \le 2\eta_j$. Expressing θ_G in the basis dual to (v_1, v_2), whose elements have norm at most $CL/|\det(v_1, v_2)| \le CQ_j^{20}/L$, we get $|\theta_G| \le CQ_j^{20}e^{-\kappa K_{\mathrm{f}}(j)}$, and therefore, since $|x| \le |(x, Gx)|$,

$$|\ell(z)| \le \varepsilon_j |z| \quad (z \in W_G), \qquad \varepsilon_j := CQ_j^{20}e^{-\kappa K_{\mathrm{f}}(j)}. \tag{6.7}$$

Two large band pieces are impossible. Suppose $G \ne G'$ in $\mathcal{G}_j$ both have large band pieces in (Σ, Σ'), and let $W := W_G + W_{G'}$, of dimension $d \in \{3, 4\}$. Choose d linearly independent vectors $y_1, \dots, y_d$ among the basis vectors (6.2) of W_G and $W_{G'}$; each has norm at most $R_j = CQ_j^2$ and lies in W_G or in $W_{G'}$. Their d-dimensional volume is at least 1, being the square root of a positive integer (the Gram determinant). For $z \in W$ write $z = \sum c_i y_i$; Cramer's rule gives $|c_i| \le |z| R_j^{d-1}$, and (6.7) gives $|\ell(y_i)| \le \varepsilon_j R_j$. Hence

$$|\ell(z)| \le d\,\varepsilon_j R_j^d\,|z| \le \varepsilon'_j |z| \quad (z \in W), \qquad \varepsilon'_j := CQ_j^{28}e^{-\kappa K_{\mathrm{f}}(j)} = C\,2^{28j}e^{-\kappa\lceil B_1 j^3\rceil}. \tag{6.8}$$

The quantity ε'_j does not depend on L, and it tends to 0 as $j \to \infty$.

If $d = 4$, then $W = \mathbb{R}^4$ and (6.8) gives $|\ell| \le \varepsilon'_j$, which contradicts $|\ell| \ge c$ once j_0 is large.

If $d = 3$, let $\nu = (\alpha, \beta)$ be the primitive integral normal of W; by (6.3), $|\nu| \le M(j)$. Exactly as in the proof of Lemma 5.1, (6.8) shows that the angle between ℓ and the line $\mathbb{R}\nu$ is $O(\varepsilon'_j)$, hence that $\beta \ne 0$ and $\measuredangle(\lambda_{\Sigma'}, \{\beta, -\beta\}) \le C\varepsilon'_j$ for j_0 large.

Let $w := Gv_1 \in \Sigma'$, which is retained since $v_1 \in \mathcal{P}_G(\Sigma, \Sigma')$, and let γ be a supporting functional at w. By Proposition 3.1(b), $|\gamma - \lambda_{\Sigma'}| \le Ce^{-\kappa K_{\mathrm{f}}(j)} \le C\varepsilon'_j$, so $\measuredangle(\gamma, \{\beta, -\beta\}) \le C\varepsilon'_j$.

But β is a nonzero integral form of size at most $M(j)$ and w is retained, so Lemma 6.1(b) gives $\measuredangle(\gamma, \{\beta, -\beta\}) \geq ce^{-C_{\mathrm{sep}} j^3}$. Therefore

$$ce^{-C_{\mathrm{sep}} j^3} \leq C\, 2^{28j} e^{-\kappa B_1 j^3}, \qquad \text{that is,} \qquad e^{(\kappa B_1 - C_{\mathrm{sep}}) j^3} \leq C\, 2^{28j},$$

which is false for $j \geq j_0$ when j_0 is large, by the choice (6.4) of B_1. □

6.5. **The tails.** Fix j with $j_0 \leq j \leq j_{\max}$. The fan of level $K_{\mathrm{f}}(j)$ has at most $CK_{\mathrm{f}}(j)^2 \leq Cj^6$ sectors, hence at most Cj^{12} ordered sector pairs. By Lemma 6.1(c) and the reduction (5.8), every retained collision pair lying on a graph piece with matrix $G \in \mathcal{G}_j$ lies in the band piece (G, Σ, Σ') for the sector pair (Σ, Σ') of level $K_{\mathrm{f}}(j)$ containing it (its members are retained, hence not on the rays of $\mathcal{F}_{K_{\mathrm{f}}(j)}$). We bound the sum of all populations of band j.

The small band pieces contribute, by the bound (6.1) on $|\mathcal{G}_j|$, at most

$$|\mathcal{G}_j| \cdot Cj^{12} \cdot \frac{L^2}{Q_j^{20}} \leq CL^2 j^{12} 2^{-11j}.$$

By Lemma 6.3 there is at most one large band piece in each sector pair. A large band piece (G, Σ, Σ') contains two pairs with independent first members (Lemma 6.2, as in the proof of Lemma 6.3), and all its pairs (v, Gv) satisfy $|v|, |Gv| \leq CL$; so by Lemma 5.2(a) its population is at most $CL^2/I(G) \leq CL^2/Q_j$. The large band pieces therefore contribute at most $Cj^{12}L^2/Q_j = CL^2 j^{12} 2^{-j}$. Summing over the bands gives

$$\sum_{j_0 \leq j \leq j_{\max}} \sum_{G \in \mathcal{G}_j} \sum_{(\Sigma, \Sigma')} P_G(\Sigma, \Sigma') \leq CL^2\, \tau(j_0), \qquad \tau(j_0) := \sum_{j \geq j_0} j^{12} (2^{-11j} + 2^{-j}), \tag{6.9}$$

and $\tau(j_0) \to 0$ as $j_0 \to \infty$.

Finally, the matrices allowed by the reduction (5.8) with $I(G) \leq 2^{j_0}$ have $q(G) \leq 2^{j_0}$ and $h(G) \leq 10(1 + 2^{j_0})$, so they form a finite set $\mathcal{G}_{\leq j_0}$ of matrices in $\mathrm{GL}_2(\mathbb{Q}) \setminus \mathcal{K}$, independent of L. For each $G \in \mathcal{G}_{\leq j_0}$ the retained collision pairs on graph pieces with matrix G are among the pairs (v, Gv) with v, Gv primitive and $N(v) = N(Gv) \leq L$, whose number is $o_G(L^2)$ by Proposition 4.1. Hence

$$\#\{\text{retained collision pairs on graph pieces with } G \in \mathcal{G}_{\leq j_0}\} = o_{j_0}(L^2). \tag{6.10}$$

6.6. **Proof of Theorem 1.1.** The proof bounds $T(X) - D(X)$ by the number of deleted vectors plus the number of retained collision pairs, and bounds both by the results of §5 and 6.

Let $L = 2 \operatorname{arcosh}(3X/2)$ and

$$\mathcal{V}_L := \mathbb{Z}^2_{\mathrm{prim}} \cap \mathcal{S} \cap L\mathbb{B}, \qquad r_m := \#\{v \in \mathcal{V}_L : \ m(v) = m\} \quad (m \geq 1).$$

By Lemma 2.1(c), the sorted Markoff triples with maximum at most X correspond bijectively to the vectors of $\mathcal{V}_L$, the maximum of the triple being the label of the vector; so $T(X) = |\mathcal{V}_L|$, $D(X) = |m(\mathcal{V}_L)|$, and

$$E_L := T(X) - D(X) = \sum_{m \in m(\mathcal{V}_L)} (r_m - 1).$$

For $\mathcal{V} \subset \mathcal{V}_L$ put $E(\mathcal{V}) := |\mathcal{V}| - |m(\mathcal{V})|$, so that $E_L = E(\mathcal{V}_L)$. Removing one vector from $\mathcal{V}$ lowers $|\mathcal{V}|$ by one and $|m(\mathcal{V})|$ by at most one, hence lowers $E(\mathcal{V})$ by at most one. For the set $\mathcal{V}_{\mathrm{ret}}$ of retained vectors, a label of multiplicity $r \geq 2$ in $\mathcal{V}_{\mathrm{ret}}$ contributes $r - 1 \leq r(r-1)$ to $E(\mathcal{V}_{\mathrm{ret}})$, and $r(r-1)$ is the number of retained collision pairs with that label. Therefore

$$E_L \leq E(\mathcal{V}_{\mathrm{ret}}) + \#\{\text{deleted vectors}\} \leq \#\{\text{deleted vectors}\} + \#\{\text{retained collision pairs}\}.$$

Fix A_0 large and j_0 large; B_0 is then determined by (5.1). By Lemma 5.1(a) and Lemma 6.1(a), the number of deleted vectors is at most

$$\frac{C}{A_0}L^2 + \frac{C}{j_0}L^2 + o_{A_0}(L^2),$$

with C independent of A_0 and j_0. The retained collision pairs are counted by the reduction (5.8) (the exceptional $o_{A_0}(L^2)$ pairs), (6.9) and (6.10). Altogether,

$$\limsup_{L\to\infty} \frac{E_L}{L^2} \le \frac{C}{A_0} + \frac{C}{j_0} + C\tau(j_0).$$

The left-hand side does not depend on A_0 or j_0; letting $j_0 \to \infty$ and then $A_0 \to \infty$ gives $E_L = o(L^2)$.

Since $L = 2\,\mathrm{arcosh}(3X/2) \sim 2\log X$, this says $T(X) - D(X) = o(\log^2 X)$. By the count (2.4) of Lemma 2.1(d), $T(X) = |\mathcal{V}_L| \sim c_N L^2$, so $D(X) \sim T(X)$. A non-unique Markoff number $m \le X$ has $r_m \ge 2$ and contributes at least 1 to E_L, so $D_2(X) \le E_L = o(L^2)$, and $D_2(X)/D(X) \to 0$. □

7. Concluding remarks

Quantitative refinements. Theorem 1.1 leaves open the rate at which $D_2(X)/D(X)$ tends to zero. A rate would follow from a quantitative form of the proof: an explicit decay of

$$\mu_Y(F \equiv 0 \pmod{3^n}) \longrightarrow 0$$

in the proof of Proposition 4.1, together with an explicit error term in Theorem 4.7, both uniformly over the finite family $\mathcal{G}_{\le j_0}$, and a choice of the parameters A_0 and j_0 as functions of L.

The uniqueness conjecture asks for the pointwise statement that no collision pair exists. The zero-one law (Lemma 4.3) permits finitely many zeros of the trace difference F along each orbit of the 3-adic flows, and the excision of §3.3 removes directions rather than individual pairs; the pointwise question therefore requires an additional mechanism.

The role of the prime 3. The 3-adic analysis uses the base point $o = (1, 1, 1)$ of the Markoff surface $a^2 + b^2 + c^2 = 3abc$, on which the 3-adic unit points form eight residue balls that are analytically isomorphic to $\mathbb{Z}_3^2$. Poonen's theorem requires a congruence to the identity modulo a power of 3 with exponent greater than 1/2; in the normalized coordinates of Lemma 4.4 the congruence is $g(x) \equiv x \pmod 3$, with exponent 1, which is why fixation modulo 9 suffices for interpolation. The prime 3 is convenient because the scaled traces of all primitive classes are integral and the base point is the unit point $(1, 1, 1)$, so that the whole orbit of the base point lies in the eight residue balls on which the analytic coordinates are defined.

Relation to strong approximation. The finite spaces used in §4 are the quotients $Y(3^n)$ of the orbit closures of the paired action, which are transitive by construction; this transitivity supplies the compatible uniform measures of Lemma 4.3. Strong approximation, as studied by Bourgain, Gamburd and Sarnak [BGS16] and by the author [Che24], concerns the orbits of the Nielsen group on the full set of points of the Markoff surface over finite fields. In the present argument, the reduction to rational graphs of §5–6 and the cusp-orbit asymptotic of Theorem 4.7 convert the measures μ_Y into the count of lattice points v with $N(v) \le L$.

The Markoff graphs modulo p are connected for every prime $p > 3.448 \cdot 10^{392}$, by Eddy, Fuchs, Litman, Martin and Tripeny [EFL+25] using [BGS16, Che24], and for every prime $p < 10^6$, by a computation of Brown [Bro24]. For $p > 3$, Bourgain, Gamburd and Sarnak [BGS26, Theorem 4] state that connectivity modulo p implies minimality of the action on the p-adic integral points that are nonzero modulo p, and defer the proof to a sequel. Jang [Jan25] gives another proof, by p-adic analytic flows as in §4.4, and extends it to further surfaces $x^2 + y^2 + z^2 = xyz + D$. Cantat

and Jang [CJ24] study orbit closures and stationary measures of automorphism groups of affine surfaces over p-adic fields, including the Markoff-type cubic surfaces [CJ24, §8].

References

[Aig13] Martin Aigner, *Markov's theorem and 100 years of the uniqueness conjecture: A mathematical journey from irrational numbers to perfect matchings*, Springer, Cham, 2013.

[Bar96] Arthur Baragar, *On the unicity conjecture for Markoff numbers*, Canad. Math. Bull. **39** (1996), no. 1, 3–9.

[Bel06] Jason P. Bell, *A generalised Skolem–Mahler–Lech theorem for affine varieties*, J. Lond. Math. Soc. (2) **73** (2006), no. 2, 367–379.

[Bel08] ———, *Corrigendum: A generalised Skolem–Mahler–Lech theorem for affine varieties*, J. Lond. Math. Soc. (2) **78** (2008), no. 1, 267–272.

[BGT10] Jason P. Bell, Dragos Ghioca, and Thomas J. Tucker, *The dynamical Mordell–Lang problem for étale maps*, Amer. J. Math. **132** (2010), no. 6, 1655–1675.

[Bor75] Itshak Borosh, *More numerical evidence on the uniqueness of Markov numbers*, BIT **15** (1975), 351–357.

[BGS16] Jean Bourgain, Alexander Gamburd, and Peter Sarnak, *Markoff triples and strong approximation*, C. R. Math. Acad. Sci. Paris **354** (2016), no. 2, 131–135.

[BGS26] ———, *Strong approximation and Diophantine properties of Markoff triples*, J. Amer. Math. Soc. **39** (2026), no. 1, 177–204.

[Bro24] Colby Austin Brown, *An almost linear time algorithm testing whether the Markoff graph modulo p is connected* (2024). arXiv:2401.00630v3.

[BNRW20] Claire Burrin, Amos Nevo, Rene Rühr, and Barak Weiss, *Effective counting for discrete lattice orbits in the plane via Eisenstein series*, Enseign. Math. **66** (2020), no. 3-4, 259–304. arXiv:1905.01493. The theorem and section numbers cited in the text refer to the authors' version dated 23 May 2020.

[But98] Jack O. Button, *The uniqueness of the prime Markoff numbers*, J. London Math. Soc. (2) **58** (1998), no. 1, 9–17.

[But01] ———, *Markoff numbers, principal ideals and continued fraction expansions*, J. Number Theory **87** (2001), no. 1, 77–95.

[CJ24] Serge Cantat and Seung Uk Jang, *Orbits of automorphism groups of affine surfaces over p-adic fields* (2024). arXiv:2410.08579v1.

[CX18] Serge Cantat and Junyi Xie, *Algebraic actions of discrete groups: the p-adic method*, Acta Math. **220** (2018), no. 2, 239–295.

[Cas57] J. W. S. Cassels, *An introduction to Diophantine approximation*, Cambridge Tracts in Mathematics and Mathematical Physics, vol. 45, Cambridge University Press, 1957.

[Cas86] ———, *Local fields*, London Mathematical Society Student Texts, vol. 3, Cambridge University Press, 1986.

[Che24] William Y. Chen, *Nonabelian level structures, Nielsen equivalence, and Markoff triples*, Ann. of Math. (2) **199** (2024), no. 1, 301–443.

[Coh55] Harvey Cohn, *Approach to Markoff's minimal forms through modular functions*, Ann. of Math. (2) **61** (1955), 1–12.

[Coh71] ———, *Representation of Markoff's binary quadratic forms by geodesics on a perforated torus*, Acta Arith. **18** (1971), 125–136.

[DLN26] Nhat Minh Doan, Xiaobin Li, and Van Nguyen, *Stable-norm boundary geometry on once-punctured tori* (2026). arXiv:2605.14574v3.

[EFL+25] Jillian Eddy, Elena Fuchs, Matthew Litman, Daniel E. Martin, and Nico Tripeny, *Connectivity of Markoff mod-p graphs and maximal divisors*, Proc. Lond. Math. Soc. (3) **130** (2025), no. 2. Paper No. e70027.

[Fis25] David Fisac, *Markov's conjecture on integral necklaces*, Bull. Lond. Math. Soc. **57** (2025), no. 12, 4122–4131.

[Foc97] Vladimir V. Fock, *Dual Teichmüller spaces* (1997). arXiv:dg-ga/9702018v3.

[Fro13] Georg Frobenius, *Über die Markoffschen Zahlen*, Sitzungsber. Königl. Preuss. Akad. Wiss. Berlin (1913), 458–487.

[GMR19] Alex Gamburd, Michael Magee, and Ryan Ronan, *An asymptotic formula for integer points on Markoff–Hurwitz varieties*, Ann. of Math. (2) **190** (2019), no. 3, 751–809.

[Gas22] Jonah Gaster, *Boundary slopes for the Markov ordering on relatively prime pairs*, Adv. Math. **403** (2022). Paper No. 108377; arXiv:2107.13499.

[Gol09] William M. Goldman, *Trace coordinates on Fricke spaces of some simple hyperbolic surfaces*, Handbook of Teichmüller theory. Vol. II, IRMA Lect. Math. Theor. Phys., vol. 13, Eur. Math. Soc., Zürich, 2009, pp. 611–684.

[HW08] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*, 6th ed., Oxford University Press, Oxford, 2008.

[Hin20] Robert Hines, *An infinite product on the Teichmüller space of the once-punctured torus* (2020). arXiv:2001.05557v3.

[Jan25] Seung Uk Jang, *Residual transitivity implies minimality for Markoff surfaces over p-adic integers, by means of p-adic flows* (2025). arXiv:2502.18976v1.

[LPTV21] Clément Lagisquet, Edita Pelantová, Sébastien Tavenas, and Laurent Vuillon, *On the Markov numbers: fixed numerator, denominator, and sum conjectures*, Adv. in Appl. Math. **130** (2021). Paper No. 102227.

[LT07] Mong Lung Lang and Ser Peow Tan, *A simple proof of the Markoff conjecture for prime powers*, Geom. Dedicata **129** (2007), 15–22.

[LLRS23] Kyungyong Lee, Li Li, Michelle Rabideau, and Ralf Schiffler, *On the ordering of the Markov numbers*, Adv. in Appl. Math. **143** (2023). Paper No. 102453.

[MKS76] Wilhelm Magnus, Abraham Karrass, and Donald Solitar, *Combinatorial group theory*, 2nd ed., Dover, New York, 1976.

[Mar79] A. Markoff, *Sur les formes quadratiques binaires indéfinies*, Math. Ann. **15** (1879), 381–406.

[Mar80] ———, *Sur les formes quadratiques binaires indéfinies. II*, Math. Ann. **17** (1880), 379–399.

[Mas97] Daniel Massart, *Stable norms of surfaces: local structure of the unit ball at rational directions*, Geom. Funct. Anal. **7** (1997), no. 6, 996–1010.

[McS21] Greg McShane, *Convexity and Aigner's conjectures* (2021). arXiv:2101.03316v1.

[MP08] Greg McShane and Hugo Parlier, *Multiplicities of simple closed geodesics and hypersurfaces in Teichmüller space*, Geom. Topol. **12** (2008), no. 4, 1883–1919.

[MR95a] Greg McShane and Igor Rivin, *A norm on homology of surfaces and counting simple geodesics*, Internat. Math. Res. Notices **2** (1995), 61–69.

[MR95b] ———, *Simple curves on hyperbolic tori*, C. R. Acad. Sci. Paris Sér. I Math. **320** (1995), no. 12, 1523–1528.

[O'S26] Cormac O'Sullivan, *Patterns in the Markov numbers and their generalizations* (2026). arXiv:2609.14149v1.

[Per22] Markus Perling, *Unfocused notes on the Markoff equation and T-singularities* (2022). arXiv:2210.12982v1.

[Poo14] Bjorn Poonen, *p-adic interpolation of iterates*, Bull. Lond. Math. Soc. **46** (2014), no. 3, 525–527.

[RS20] Michelle Rabideau and Ralf Schiffler, *Continued fractions and orderings on the Markov numbers*, Adv. Math. **370** (2020). Paper No. 107231.

[Sch96] Paul Schmutz, *Systoles of arithmetic surfaces and the Markoff spectrum*, Math. Ann. **305** (1996), 191–203.

[Ser85] Caroline Series, *The geometry of Markoff numbers*, Math. Intelligencer **7** (1985), no. 3, 20–29.

[SV19] Alfonso Sorrentino and Alexander P. Veselov, *Markov numbers, Mather's β function and stable norm*, Nonlinearity **32** (2019), no. 6, 2147–2156.

[SV17] K. Spalding and Alexander P. Veselov, *Lyapunov spectrum of Markov and Euclid trees*, Nonlinearity **30** (2017), no. 12, 4428–4453.

[Sri09] Anitha Srinivasan, *Markoff numbers and ambiguous classes*, J. Théor. Nombres Bordeaux **21** (2009), no. 3, 757–770.

[Str28] Reinhold Strassmann, *Über den Wertevorrat von Potenzreihen im Gebiet der p-adischen Zahlen*, J. Reine Angew. Math. **159** (1928), 13–28.

[TY25] Tianyi Tao and Bohan Yang, *Markov's equation is not partition regular* (2025). arXiv:2510.10151v1.

[Zag82] Don Zagier, *On the number of Markoff numbers below a given bound*, Math. Comp. **39** (1982), no. 160, 709–723.

[Zha07] Ying Zhang, *Congruence and uniqueness of certain Markov numbers*, Acta Arith. **128** (2007), no. 3, 295–301.