ON THE STRONG DENSITY CONJECTURE FOR INTEGRAL APOLLONIAN CIRCLE PACKINGS

JEAN BOURGAIN AND ALEX KONTOROVICH

ABSTRACT. We prove that a set of density one satisfies the Strong Density Conjecture for Apollonian Circle Packings. That is, for a fixed integral, primitive Apollonian gasket, almost every (in the sense of density) sufficiently large, admissible (passing local obstructions) integer is the curvature of some circle in the packing.

CONTENTS

FIGURE 1. The Apollonian gasket with root quadruple $v_0 = (-11, 21, 24, 28)^t$.

1. INTRODUCTION

1.1. **The Strong Density Conjecture.**

Let $\mathscr{P}$ be an Apollonian circle packing, or gasket, see Fig. 1. The number $b(C)$ shown inside a circle $C \in \mathscr{P}$ is its "bend" or curvature, that is, the reciprocal of its radius (the bounding circle has negative orientation). Soddy [Sod37] first observed the existence of *integral* packings, meaning that there are gaskets $\mathscr{P}$ so that $b(C) \in \mathbb{Z}$ for all $C \in \mathscr{P}$. Let

$$\mathscr{B} = \mathscr{B}_{\mathscr{P}} := \{b(C) : C \in \mathscr{P}\}$$

be the set of all bends in $\mathscr{P}$. We call a packing *primitive* if $\gcd(\mathscr{B}) = 1$. From now on, we restrict our attention to a fixed primitive integral Apollonian gasket $\mathscr{P}$.

Lagarias, Graham et al [LMW02, GLM$^+$03] initiated a detailed study of Diophantine properties of $\mathscr{B}$, with two separate families of problems: studying $\mathscr{B}$ with multiplicity (that is, studying circles), or without multiplicity (studying the integers which arise). In the present paper, we are concerned with the latter (see e.g. [KO11, FS10] for the former).

In particular, the following striking local-to-global conjecture for $\mathscr{B}$ is given in [GLM$^+$03]. Let $\mathscr{A} = \mathscr{A}_{\mathscr{P}}$ denote the *admissible* integers, that is, those passing local (congruence) obstructions:

$$\mathscr{A} := \{n \in \mathbb{Z} : n \in \mathscr{B}(\mathrm{mod}\, q), \text{ for all } q \geq 1\}.$$

Conjecture 1.1 (Strong Density Conjecture [GLM$^+$03], p. 37)**.** *If $n \in \mathscr{A}$ and $n \gg 1$, then $n \in \mathscr{B}$. That is, every sufficiently large admissible number is the bend of a curvature in the packing.*

We restate the conjecture in the following way. For $N \geq 1$, let

$$\mathscr{B}(N) := \mathscr{B} \cap [1, N], \quad \text{and} \quad \mathscr{A}(N) := \mathscr{A} \cap [1, N].$$

Then the Strong Density Conjecture is equivalent to the assertion that

$$\#\mathscr{B}(N) = \#\mathscr{A}(N) + O(1), \quad \text{as } N \to \infty. \tag{1.2}$$

In her thesis, Fuchs [Fuc10] proved that $\mathscr{A}$ is a union of arithmetic progressions with modulus dividing 24, that is,

$$n \in \mathscr{A} \quad \Longleftrightarrow \quad n(\mathrm{mod}\,24) \in \mathscr{A}(24), \tag{1.3}$$

cf. Lemma 2.21. Hence obviously

$$\#\mathscr{A}(N) = \frac{\#\mathscr{A}(24)}{24} \cdot N + O(1).$$

In fact the local obstructions are easily determined from the so-called *root quadruple*

$$v_0 = v_0(\mathscr{P}), \tag{1.4}$$

that is, the column vector of the four smallest curvatures in $\mathscr{B}$, as in Fig. 1; see Remark 2.23.

1.2. **Partial Progress and Statement of the Main Theorem.**

The history of this problem is as follows. The first progress towards (1.2) was already made in [GLM$^+$03], who showed that

$$\#\mathscr{B}(N) \gg N^{1/2}. \tag{1.5}$$

Sarnak [Sar07] improved this to

$$\#\mathscr{B}(N) \gg \frac{N}{(\log N)^{1/2}}, \tag{1.6}$$

and then Fuchs [Fuc10] showed

$$\#\mathscr{B}(N) \gg \frac{N}{(\log N)^{0.150\ldots}}.$$

Finally Bourgain and Fuchs [BF10] settled the so-called "Positive Density Conjecture," that

$$\#\mathscr{B}(N) \gg N.$$

The purpose of this paper is to prove that a set of density one (within the admissibles) satisfies the Strong Density Conjecture.

Theorem 1.7. *Let $\mathscr{P}$ be a primitive, integral Apollonian circle packing, and let $\mathscr{A}(N), \mathscr{B}(N)$ be as above. Then there exists some absolute, effectively computable $\eta > 0$, so that as $N \to \infty$,*

$$\#\mathscr{B}(N) = \#\mathscr{A}(N) + O(N^{1-\eta}). \tag{1.8}$$

1.3. **Methods.**

Our main approach is through the Hardy-Littlewood-Ramanujan circle method, combining two new ingredients. The first, applied to the major arcs, is effective bisector counting in infinite volume hyperbolic 3-folds, recently achieved by I. Vinogradov [Vin12], as well as the uniform spectral gap over congruence towers of such, established in [Var10, BV11, BGS09]. The second ingredient is the minor arcs analysis, inspired by that given recently by the first-named author in [Bou11], where it was proved that the prime curvatures in a packing constitute a positive proportion of the primes. (Obviously (1.8) implies that 100% of the admissible prime curvatures appear.)

1.4. **Plan for the Paper.**

A more detailed outline of the proof, as well as the setup of some relevant exponential sums, is given in §4. Before we can do this, we need to recall the Apollonian group and some of its subgroups in §2, and also some technology from spectral and representation theory of infinite volume hyperbolic quotients, described in §3. Some lemmata are reserved for §5, the major arcs are estimated in §6, and the minor arcs are dealt with in §§7-9.

1.5. **Notation.**

We use the following standard notation. Set $e(x) = e^{2\pi i x}$ and $e_q(x) = e(\frac{x}{q})$. We use $f \ll g$ and $f = O(g)$ interchangeably; moreover $f \asymp g$ means $f \ll g \ll f$. Unless otherwise specified, the implied constants may depend at most on the packing $\mathscr{P}$ (or equivalently on the root quadruple v_0), which is treated as fixed. The symbol $\mathbf{1}_{\{.\}}$ is the indicator function of the event $\{\cdot\}$. The greatest common divisor of n and m is written (n, m), their least common multiple is $[n, m]$, and $\omega(n)$ denotes the number of distinct prime factors of n. The cardinality of a finite set S is denoted $|S|$ or $\#S$. The transpose of a matrix g is written g^t. The prime symbol $'$ in $\sum_{r(q)}'$ means the range of $r(\mathrm{mod}\, q)$ is restricted to $(r, q) = 1$. Finally, $p^j \| q$ denotes $p^j \mid q$ and $p^{j+1} \nmid q$.

2. PRELIMINARIES I: THE APOLLONIAN GROUP AND ITS SUBGROUPS

2.1. **Descartes Theorem and Consequences.**

Descartes' Circle Theorem states that a quadruple v of (oriented) curvatures of four mutually tangent circles lies on the cone

$$F(v) = 0, \tag{2.1}$$

where F is the Descartes quadratic form:

$$F(a, b, c, d) = 2(a^2 + b^2 + c^2 + d^2) - (a + b + c + d)^2. \tag{2.2}$$

Note that F has signature $(3, 1)$, and let

$$G := \mathrm{SO}_F(\mathbb{R}) = \{g \in \mathrm{SL}(3, \mathbb{R}) : F(gv) = F(v), \text{ for all } v \in \mathbb{R}^3\}$$

be the real special orthogonal group preserving F.

It follows immediately that for b, c and d fixed, there are two solutions a, a' to (2.1), and

$$a + a' = 2(b + c + d).$$

Hence we observe that a can be changed into a' by a reflection, that is,

$$(a, b, c, d)^t = S_1 \cdot (a', b, c, d)^t,$$

where the reflections

$$S_1 = \begin{pmatrix} -1 & 2 & 2 & 2 \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 1 & & & \\ 2 & -1 & 2 & 2 \\ & & 1 & \\ & & & 1 \end{pmatrix},$$

$$S_3 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ 2 & 2 & -1 & 2 \\ & & & 1 \end{pmatrix}, \qquad S_4 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ 2 & 2 & 2 & -1 \end{pmatrix},$$

generate the so-called *Apollonian group*

$$\mathcal{A} = \langle S_1, S_2, S_3, S_4 \rangle. \tag{2.3}$$

It is a Coxeter group, free save the relations $S_j^2 = I$, $1 \leq j \leq 4$. We immediately pass to the index two subgroup

$$\Gamma := \mathcal{A} \cap \mathrm{SO}_F$$

of orientation preserving transformations, that is, even words in the generators. Then Γ is freely generated by $S_1 S_2$, $S_2 S_3$ and $S_3 S_4$. Note that Γ is Zariski dense in G but *thin*, that is, of infinite index in $G(\mathbb{Z})$; equivalently, the Haar measure of $\Gamma \backslash G$ is infinite.

2.2. **Using Unipotents.**

Now we review the arguments from [GLM$^+$03, Sar07] which lead to (1.5) and (1.6), as our setup depends critically on them.

Recall that for any fixed packing $\mathscr{P}$, there is a root quadruple v_0 of the four smallest bends in $\mathscr{P}$, cf. (1.4). It follows from (2.1) and (2.3) that the set $\mathscr{B}$ of all bends can be realized as the orbit of the root quadruple v_0 under $\mathcal{A}$. Let

$$\mathscr{O} = \mathscr{O}_{\mathscr{P}} := \Gamma \cdot v_0$$

be the orbit of v_0 under Γ. Then the set of all bends certainly contains

$$\mathscr{B} \supset \bigcup_{j=1}^{4} \langle e_j, \mathscr{O} \rangle = \bigcup_{j=1}^{4} \langle e_j, \Gamma \cdot v_0 \rangle, \tag{2.4}$$

where $e_1 = (1,0,0,0)^t, \ldots, e_4 = (0,0,0,1)^t$ constitute the standard basis for $\mathbb{R}^4$, and the inner product above is the standard one. Recall we are treating $\mathscr{B}$ as a set, that is, without multiplicities.

It was observed in [GLM$^+$03] that Γ contains unipotent elements, and hence one can use these to furnish an injection of affine space in the otherwise intractable orbit $\mathscr{O}$, as follows. Note first that

$$C_1 := S_4 S_3 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ 2 & 2 & -1 & 2 \\ 6 & 6 & -2 & 3 \end{pmatrix} \in \Gamma, \tag{2.5}$$

and after conjugation by

$$J := \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ -1 & 1 & -2 & 1 \\ -1 & & & 1 \end{pmatrix},$$

we have

$$\tilde{C}_1 := J^{-1} \cdot C_1 \cdot J = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 2 & 1 \\ & & 4 & 4 & 1 \end{pmatrix}.$$
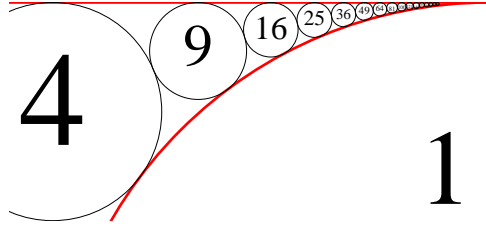
FIGURE 2. Circles tangent to two fixed circles.

Recall the spin homomorphism $\rho : \mathrm{SL}_2 \to \mathrm{SO}(2,1)$, embedded for our purposes in SL_4, given explicitly by

$$\rho : \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \mapsto \frac{1}{\alpha\delta - \beta\gamma} \begin{pmatrix} 1 & & & \\ & \alpha^2 & 2\alpha\gamma & \gamma^2 \\ & \alpha\beta & \alpha\delta + \beta\gamma & \gamma\delta \\ & \beta^2 & 2\beta\delta & \delta^2 \end{pmatrix}. \qquad (2.6)$$

In fact SL_2 is a double cover of $\mathrm{SO}(2,1)$ under ρ, and PSL_2 is a bijection. It is clear from inspection that

$$\rho : \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} =: T_1 \mapsto \tilde{C}_1.$$

The powers of T_1 are elementary to describe, $T_1^n = \begin{pmatrix} 1 & 2n \\ 0 & 1 \end{pmatrix}$, and hence for each $n \in \mathbb{Z}$, Γ contains the element

$$C_1^n = J \cdot \rho(T_1^n) \cdot J^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 4n^2 - 2n & 4n^2 - 2n & 1 - 2n & 2n \\ 4n^2 + 2n & 4n^2 + 2n & -2n & 2n + 1 \end{pmatrix}.$$

(Of course this can be seen directly from (2.5); these transformations will be more enlightening below.)

Thus if $v = (a, b, c, d)^t \in \mathscr{O}$ is a quadruple in the orbit, then $\mathscr{O}$ also contains $C_1^n \cdot v$ for all n. From (2.4), we then have that the set $\mathscr{B}$ of curvatures contains

$$\mathscr{B} \ni \langle e_4, C_1^n \cdot v \rangle = 4(a + b)n^2 + 2(a + b - c + d)n + d. \qquad (2.7)$$

The circles thus generated are all tangent to two fixed circles, which explains the square bends in Fig. 2. Of course (2.7) immediately implies (1.5).

Observe further that

$$C_2 := S_2 S_3 = \begin{pmatrix} 1 & & & \\ 6 & 3 & -2 & 6 \\ 2 & 2 & -1 & 2 \\ & & & 1 \end{pmatrix}$$

is another unipotent element, with

$$\tilde{C}_2 := J^{-1} \cdot C_2 \cdot J = \begin{pmatrix} 1 & & & \\ & 1 & 4 & 4 \\ & & 1 & 2 \\ & & & 1 \end{pmatrix},$$

and

$$\rho : \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} =: T_2 \mapsto \tilde{C}_2.$$

Since T_1 and T_2 generate $\Lambda(2)$, the principal 2-congruence subgroup of $\mathrm{PSL}(2, \mathbb{Z})$, we see that the Apollonian group Γ contains the subgroup

$$\Xi := \langle C_1, C_2 \rangle = J \cdot \rho\Big(\Lambda(2)\Big) \cdot J^{-1} < \Gamma. \tag{2.8}$$

In particular, whenever $(2x, y) = 1$, there is an element

$$\begin{pmatrix} * & 2x \\ * & y \end{pmatrix} \in \Lambda(2),$$

and thus Ξ contains the element

$$\xi_{x,y} := J \cdot \rho \begin{pmatrix} * & 2x \\ * & y \end{pmatrix} \cdot J^{-1} \tag{2.9}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ * & * & * & * \\ * & * & * & * \\ 4x^2 + 2xy + y^2 - 1 & 4x^2 + 2xy & -2xy & 2xy + y^2 \end{pmatrix}.$$

Write

$$\begin{aligned} w_{x,y} &= \xi_{x,y}^t \cdot e_4 \tag{2.10} \\ &= (4x^2 + 2xy + y^2 - 1, 4x^2 + 2xy, -2xy, 2xy + y^2)^t. \end{aligned}$$

Then again by (2.4), we have shown the following

Lemma 2.11 ([Sar07]). *Let $x, y \in \mathbb{Z}$ with $(2x, y) = 1$, and take any element $\gamma \in \Gamma$ with corresponding quadruple*

$$v_\gamma = (a_\gamma, b_\gamma, c_\gamma, d_\gamma)^t = \gamma \cdot v_0 \in \mathscr{O}. \tag{2.12}$$

Then the number

$$\langle e_4, \xi_{x,y} \cdot \gamma \cdot v_0 \rangle = \langle w_{x,y}, \gamma \cdot v_0 \rangle = 4A_\gamma x^2 + 4B_\gamma xy + C_\gamma y^2 - a_\gamma \tag{2.13}$$

is the curvature of some circle in $\mathscr{P}$, where we have defined

$$\begin{aligned}
A_\gamma &:= a_\gamma + b_\gamma, \\
B_\gamma &:= \frac{a_\gamma + b_\gamma - c_\gamma + d_\gamma}{2}, \\
C_\gamma &:= a_\gamma + d_\gamma.
\end{aligned} \tag{2.14}$$

Note from (2.1) *that B_γ is integral.*

Observe that, by construction, the value of a_γ is unchanged under the orbit of the group (2.8), and the circles whose bends are generated by (2.13) are all tangent to the circle corresponding to a_γ. It is classical (see [IK04, (1.87)]) that the number of distinct primitive values up to N assumed by a binary quadratic form is of order at least $N(\log N)^{-1/2}$, proving (1.6).

To fix notation, we define the binary quadratic appearing in (2.13) and its shift by

$$f_\gamma(x, y) := A_\gamma x^2 + 2B_\gamma xy + C_\gamma y^2, \qquad \mathfrak{f}_\gamma(x, y) := f_\gamma(x, y) - a_\gamma, \tag{2.15}$$

so that

$$\langle w_{x,y}, \gamma \cdot v_0 \rangle = \mathfrak{f}_\gamma(2x, y). \tag{2.16}$$

Note from (2.14) and (2.1) that the discriminant of f_γ is

$$\Delta_\gamma = 4(B_\gamma^2 - A_\gamma C_\gamma) = -4a_\gamma^2. \tag{2.17}$$

When convenient, we will drop the subscripts γ in all the above.

2.3. **Congruence Subgroups.**

For each $q \geq 1$, define the principal q-congruence subgroup

$$\Gamma(q) := \{\gamma \in \Gamma : \gamma \equiv I \,(\mathrm{mod}\, q)\}. \tag{2.18}$$

These groups all have infinite index in $G(\mathbb{Z})$, but finite index in Γ. The quotients $\Gamma/\Gamma(q)$ have been determined explicitly by Fuchs [Fuc10] via Strong Approximation [MVW84], as we explain below. Of course G does not have Strong Approximation, but its connected spin double cover $\mathrm{SL}(2, \mathbb{C})$ does. We will need the covering map explicitly later, so record it here.

First change variables from the Descartes form F to

$$\tilde{F}(x, y, z, w) := xw + y^2 + z^2.$$

Then there is a homomorphism $\iota_0 : \mathrm{SL}(2, \mathbb{C}) \to \mathrm{SO}_{\tilde{F}}(\mathbb{R})$, sending

$$g = \begin{pmatrix} a + \alpha i & b + \beta i \\ c + \gamma i & d + \delta i \end{pmatrix} \in \mathrm{SL}(2, \mathbb{C})$$

to

$$\frac{1}{|\det(g)|^2} \begin{pmatrix} a^2 + \alpha^2 & 2(ac + \alpha\gamma) & 2(c\alpha - a\gamma) & -c^2 - \gamma^2 \\ ab + \alpha\beta & bc + ad + \beta\gamma + \alpha\delta & d\alpha + c\beta - b\gamma - a\delta & -cd - \gamma\delta \\ a\beta - b\alpha & -d\alpha + c\beta - b\gamma + a\delta & -bc + ad - \beta\gamma + \alpha\delta & d\gamma - c\delta \\ -b^2 - \beta^2 & -2(bd + \beta\delta) & 2(b\delta - d\beta) & d^2 + \delta^2 \end{pmatrix}.$$

To map from $\mathrm{SO}_{\tilde{F}}$ to SO_F, we apply an obvious conjugation, see [GLM$^+$03, (4.1)]. Let

$$\iota : \mathrm{SL}(2, \mathbb{C}) \to \mathrm{SO}_F(\mathbb{R}) \qquad (2.19)$$

be the composition of this conjugation with ι_0. Let $\tilde{\Gamma}$ be the preimage of Γ under ι.

Lemma 2.20 ([GLM$^+$05, Fuc10]). *The group $\tilde{\Gamma}$ is generated by*

$$\pm \begin{pmatrix} 1 & 4i \\ & 1 \end{pmatrix}, \quad \pm \begin{pmatrix} -2 & i \\ i & \end{pmatrix}, \quad \pm \begin{pmatrix} 2 + 2i & 4 + 3i \\ -i & -2i \end{pmatrix}.$$

One can thus determine explicitly the images of $\tilde{\Gamma}$ in $\mathrm{SL}(2, \mathbb{Z}[i]/(q))$, and hence understand the quotients $\Gamma/\Gamma(q)$.

Lemma 2.21 ([Fuc10]).

(1) The quotient groups $\Gamma/\Gamma(q)$ are multiplicative, that is, if q factors as

$$q = p_1^{\ell_1} \cdots p_r^{\ell_r},$$

then

$$\Gamma/\Gamma(q) \cong \Gamma/\Gamma(p_1^{\ell_1}) \times \cdots \times \Gamma/\Gamma(p_r^{\ell_r}).$$

(2) If $(q, 6) = 1$ then the quotient is onto,

$$\Gamma/\Gamma(q) \cong \mathrm{SO}_F(\mathbb{Z}/q\mathbb{Z}). \qquad (2.22)$$

(3) If $q = 2^\ell$, $\ell \geq 3$, then $\Gamma/\Gamma(q)$ is the full preimage of $\Gamma/\Gamma(8)$ under the projection $\mathrm{SO}_F(\mathbb{Z}/q\mathbb{Z}) \to \mathrm{SO}_F(\mathbb{Z}/8\mathbb{Z})$. That is, the powers of 2 stabilize at 8. Similarly, the powers of 3 stabilize at 3, meaning that for $q = 3^\ell$, $\ell \geq 1$, the quotient $\Gamma/\Gamma(q)$ is the preimage of $\Gamma/\Gamma(3)$ under the corresponding projection map.

Remark 2.23. This of course explains all local obstructions, cf. (1.3). The admissible numbers are precisely those residue classes (mod 24) which appear as some entry in the orbit of v_0 under $\Gamma/\Gamma(24)$.
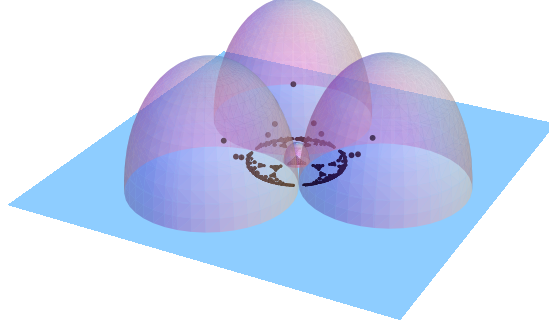
FIGURE 3. The orbit of a point in hyperbolic space under the Apollonian group.

3. PRELIMINARIES II: AUTOMORPHIC FORMS AND REPRESENTATIONS

3.1. Spectral Theory.

Recall the general spectral theory in our present context. We abuse notation (in this section only), passing from $G = \mathrm{SO}_F(\mathbb{R})$ to its spin double cover $G = \mathrm{SL}(2, \mathbb{C})$. Let $\Gamma < G$ be a geometrically finite discrete group. Then Γ acts discontinuously on the upper half space $\mathbb{H}^3$, and any Γ orbit has a limit set Λ_Γ in the boundary $\partial \mathbb{H}^3 \cong S^2$ of some Hausdorff dimension $\delta = \delta(\Gamma) \in [0, 2]$. We assume that Γ is non-elementary (not virtually abelian), so $\delta > 0$, and moreover that Γ is not a lattice, that is, the quotient $\Gamma \backslash \mathbb{H}^3$ has infinite hyperbolic volume; then $\delta < 2$. The hyperbolic Laplacian Δ acts on the space $L^2(\Gamma \backslash \mathbb{H})$ of functions automorphic under Γ and square integrable on the quotient; we choose the Laplacian to be positive definite. The spectrum is controlled via the following, see [Pat76, Sul84, LP82].

Theorem 3.1 (Patterson, Sullivan, Lax-Phillips). *The spectrum above* 1 *is purely continuous, and the spectrum below* 1 *is purely discrete. The latter is empty unless $\delta > 1$, in which case, ordering the eigenvalues by*

$$0 < \lambda_0 < \lambda_1 \leq \cdots \leq \lambda_{max} < 1, \tag{3.2}$$

the base eigenvalue λ_0 is given by

$$\lambda_0 = \delta(2 - \delta).$$

Remark 3.3. Of course in our application to the Apollonian group, the limit set is precisely the underlying gasket, see Fig. 3. It has dimension

$$\delta \approx 1.3... > 1. \tag{3.4}$$

Corresponding to λ_0 is the Patterson-Sullivan base eigenfunction, φ_0, which can be realized explicitly as the integral of a Poisson kernel against the so-called Patterson-Sullivan measure μ. Roughly speaking, μ is the weak* limit as $s \to \delta^+$ of the measures

$$\mu_s(x) := \frac{\sum_{\gamma \in \Gamma} \exp(-s\, d(\mathfrak{o}, \gamma \cdot \mathfrak{o})) \mathbf{1}_{x=\gamma\mathfrak{o}}}{\sum_{\gamma \in \Gamma} \exp(-s\, d(\mathfrak{o}, \gamma \cdot \mathfrak{o}))}, \tag{3.5}$$

where $d(\cdot, \cdot)$ is the hyperbolic distance, and $\mathfrak{o}$ is any fixed point in $\mathbb{H}^3$.

3.2. **Spectral Gap.**

We assume henceforth that Γ is moreover "arithmetic," in the sense that $\Gamma < \mathrm{SL}(2, \mathcal{O})$, where $\mathcal{O} = \mathbb{Z}[i]$. Then we have a tower of congruence subgroups: for any integer $q \geq 1$, define $\Gamma(q)$ to be the kernel of the projection map $\Gamma \to \mathrm{SL}(2, \mathcal{O}/\mathfrak{q})$, with $\mathfrak{q} = (q)$ the principal ideal. As in (3.2), write

$$0 < \lambda_0(q) < \lambda_1(q) \leq \cdots \leq \lambda_{max(q)}(q) < 1, \tag{3.6}$$

for the discrete spectrum of $\Gamma(q)$. The groups $\Gamma(q)$, while of infinite covolume, have finite index in Γ, and hence

$$\lambda_0(q) = \lambda_0 = \delta(2 - \delta). \tag{3.7}$$

But the second eigenvalues $\lambda_1(q)$ could *a priori* encroach on the base. The fact that this does not happen is the spectral gap property for Γ.

Theorem 3.8 ([BGS09, Var10, BV11])**.** *Given Γ as above, there exists some $\varepsilon = \varepsilon(\Gamma) > 0$ such that for all $q \geq 1$,*

$$\lambda_1(q) \geq \lambda_0 + \varepsilon. \tag{3.9}$$

Remark 3.10. The above was proved for the corresponding Cayley graphs over number fields for square-free ideals $\mathfrak{q}$ in [Var10]. The extension to all q is executed as in [BV11], and the derivation for an archimedean gap from that of the graph Laplacian is given in [BGS09].

3.3. **Representation Theory and Mixing Rates.**

By the Duality Theorem of Gelfand, Graev, and Piatetski-Shapiro [GGPS66], the spectral decomposition above is equivalent to the decomposition into irreducibles of the right regular representation acting on $L^2(\Gamma \backslash G)$. That is, we identify $\mathbb{H}^3 \cong G/K$, with $K = \mathrm{SU}(2)$ a maximal compact, and lift functions from $\mathbb{H}^3$ to (right K-invariant) functions on G. Corresponding to (3.2) is the decomposition

$$L^2(\Gamma \backslash G) = V_{\lambda_0} \oplus V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_{max}} \oplus V_{temp}. \tag{3.11}$$

Here V_{temp} contains the tempered spectrum, and each V_{λ_j} is an infinite dimensional vector space, isomorphic as a G-representation to a complementary series representation with parameter $s_j \in (1,2)$ determined by $\lambda_j = s_j(2-s_j)$. Obviously, a similar decomposition holds for $L^2(\Gamma(q)\backslash G)$, corresponding to (3.6).

We also have the following well-known general fact about mixing rates of matrix coefficients, see e.g. [CHH88, BR02]. First we recall the relevant Sobolev norm. Let (π, V) be a unitary G-representation, and let $\{X_j\}$ denote an orthonormal basis of the Lie algebra $\mathfrak{k}$ of K with respect to an Ad-invariant scalar product. For a smooth vector $v \in V^\infty$, define the (second order) Sobolev norm $\mathcal{S}$ of v by

$$\mathcal{S}v := \|v\|_2 + \sum_j \|d\pi(X_j).v\|_2 + \sum_j \sum_{j'} \|d\pi(X_j)d\pi(X_{j'}).v\|_2.$$

Theorem 3.12. *Let $\Theta > 1$ and (π, V) be a unitary representation of G which does not weakly contain any complementary series representation with parameter $s > \Theta$. Then for any smooth vectors $v, w \in V^\infty$,*

$$|\langle \pi(g).v, w \rangle| \ll \|g\|^{-(2-\Theta)} \cdot \mathcal{S}v \cdot \mathcal{S}w. \tag{3.13}$$

Here $\| \cdot \|$ is the standard Frobenius matrix norm.

3.4. **Effective Bisector Counting.**

The next ingredient which we require is the recent work by Vinogradov [Vin12] on effective bisector counting for such infinite volume quotients. Recall the following sub(semi)groups of G:

$$A = \left\{ a_t := \begin{pmatrix} e^{t/2} & \\ & e^{-t/2} \end{pmatrix} : t \in \mathbb{R} \right\}, A^+ = \{a_t : t \geq 0\},$$

$$M = \left\{ \begin{pmatrix} e^{2\pi i\theta} & \\ & e^{-2\pi i\theta} \end{pmatrix} : \theta \in \mathbb{R}/\mathbb{Z} \right\}, K = \mathrm{SU}(2).$$

We have the Cartan decomposition $G = KA^+K$, unique up to the normalizer M of A in K. We require it in the following more precise form. Identify K/M with the sphere $S^2 \cong \partial\mathbb{H}^3$. Then for every $g \in G$ not in K, there is a unique decomposition

$$g = s_1(g) \cdot a(g) \cdot m(g) \cdot s_2(g)^{-1}. \tag{3.14}$$

with $s_1, s_2 \in K/M$, $a \in A^+$ and $m \in M$, corresponding to

$$G = K/M \times A^+ \times M \times M\backslash K.$$

Theorem 3.15 ([Vin12])**.** *Let* $\Phi, \Psi \subset S^2$ *be sectors with* $\mu(\partial\Phi) = \mu(\partial\Psi) = 0$, *and let* $\mathcal{I} \subset \mathbb{R}/\mathbb{Z}$ *be an interval. Then under the above hypotheses on* Γ *(in particular* $\delta > 1$*), and using the decomposition* (3.14), *we have*

$$\sum_{\gamma \in \Gamma} \mathbf{1} \left\{ \begin{array}{c} s_1(\gamma) \in \Phi \\ s_2(\gamma) \in \Psi \\ \|a(\gamma)\|^2 < T \\ m(\gamma) \in \mathcal{I} \end{array} \right\} = c_\delta \cdot \mu(\Phi)\mu(\Psi)\ell(\mathcal{I})T^\delta + O\big(T^\Theta\big), \quad (3.16)$$

as $T \to \infty$. *Here* $\|\cdot\|$ *is the Frobenius norm,* ℓ *is Lebesgue measure,* μ *is Patterson-Sullivan measure (cf.* (3.5)*),* c_δ *is an absolute constant depending only on* δ, *and*

$$\Theta < \delta \qquad\qquad (3.17)$$

depends only on the spectral gap for Γ. *The implied constant does not depend on* Φ, Ψ, *or* $\mathcal{I}$.

This generalizes from $\mathrm{SL}(2, \mathbb{R})$ to $\mathrm{SL}(2, \mathbb{C})$ the main result of [BKS10], which is itself a generalization to non-lattices of [Goo83, Thm 4].

4. Setup and Outline of the Proof

In this section, we introduce the main exponential sum and give an outline of the rest of the argument. Recall the fixed packing $\mathscr{P}$ having curvatures $\mathscr{B}$ and root quadruple v_0. Let Γ be the Apollonian subgroup of critical exponent $\delta \approx 1.3$, with subgroup Ξ, see (2.8). Recall also from (2.13) that for any $\gamma \in \Gamma$ and $\xi \in \Xi$,

$$\langle e_4, \xi\gamma v_0 \rangle \in \mathscr{B}.$$

Our approach, mimicing [BK10, BK11], is to exploit the bilinear (or multilinear) structure above.

We first give an informal description of the main ensemble from which we will form an exponential sum. Let N be our main growing parameter. We construct our ensemble by decomposing a ball in Γ of norm N into two balls, a small one in all of Γ of norm T, and a larger one of norm X^2 in Ξ, corresponding to $x, y \asymp X$. Specifically, we take

$$T = N^{1/100} \quad \text{and} \quad X = N^{99/200}, \qquad \text{so that} \qquad TX^2 = N. \quad (4.1)$$

See (9.10) and (9.14) where these numbers are used.

We further need the technical condition that in the T-ball, the value of $a_\gamma = \langle e_1, \gamma v_0 \rangle$ (see (2.12)) is of order T. This is used crucially in (7.10) and (9.17).

Finally, for technical reasons (see Lemma 5.12 below), we need to further split the T-ball into two: a small ball of norm T_1, and a big ball of norm T_2. Write

$$T = T_1 T_2, \qquad T_2 = T_1^{\mathcal{C}}, \qquad (4.2)$$

where $\mathcal{C}$ is a large constant depending only on the spectral gap for Γ; it is determined in (5.13). We now make formal the above discussion.

4.1. Introducing the Main Exponential Sum.

Let N, X, T, T_1, and T_2 be as in (4.1) and (4.2). Define the family

$$\mathfrak{F} = \mathfrak{F}_T := \left\{ \gamma = \gamma_1 \gamma_2 : \begin{array}{c} \gamma_1, \gamma_2 \in \Gamma, \\ T_1 < \|\gamma_1\| < 2T_1, \\ T_2 < \|\gamma_2\| < 2T_2, \\ \langle e_1, \gamma_1\, \gamma_2\, v_0 \rangle > T/100 \end{array} \right\}. \qquad (4.3)$$

Trivially from (3.16) (or just [LP82]), we have the bound

$$\#\mathfrak{F}_T \ll T^\delta. \qquad (4.4)$$

From (2.16), we can identify $\gamma \in \mathfrak{F}$ with a shifted binary quadratic form $\mathfrak{f}_\gamma$ of discriminant $-4a_\gamma^2$ via

$$\mathfrak{f}_\gamma(2x, y) = \langle w_{x,y}, \gamma\, v_0 \rangle.$$

Recall from (2.13) that whenever $(2x, y) = 1$, the above is a curvature in the packing. We sometimes drop γ, writing simply $\mathfrak{f} \in \mathfrak{F}$; then the latter can also be thought of as a family of shifted quadratic forms. Note also that the decomposition $\gamma = \gamma_1 \gamma_2$ in (4.3) need not be unique, so some forms may appear with multiplicity.

One final technicality is to smooth the sum on $x, y \asymp X$. To this end, we fix a smooth, nonnegative function Υ, supported in $[1, 2]$ and having unit mass, $\int_{\mathbb{R}} \Upsilon(x)dx = 1$.

Our main object of study is then the representation number

$$\mathcal{R}_N(n) := \sum_{\mathfrak{f} \in \mathfrak{F}_T} \sum_{(2x,y)=1} \Upsilon\left(\frac{2x}{X}\right) \Upsilon\left(\frac{y}{X}\right) \mathbf{1}_{\{n=\mathfrak{f}(2x,y)\}}, \qquad (4.5)$$

and the corresponding exponential sum, its Fourier transform

$$\widehat{\mathcal{R}_N}(\theta) := \sum_{\mathfrak{f} \in \mathfrak{F}} \sum_{(2x,y)=1} \Upsilon\left(\frac{2x}{X}\right) \Upsilon\left(\frac{y}{X}\right) e(\theta\,\mathfrak{f}(2x, y)). \qquad (4.6)$$

Clearly $\mathcal{R}_N(n) \neq 0$ implies that $n \in \mathscr{B}$. Note also from (4.4) that the total mass satisfies

$$\widehat{\mathcal{R}_N}(0) \ll T^\delta X^2. \qquad (4.7)$$

The condition $(2x, y) = 1$ will be a technical nuisance, so we introduce another parameter

$$U, \qquad (4.8)$$

a small power of N, depending only on the spectral gap of Γ; it is determined in (6.3). Then by truncating Möbius inversion, define

$$\widehat{\mathcal{R}_N^U}(\theta) := \sum_{\mathfrak{f} \in \mathfrak{F}} \sum_{x,y \in \mathbb{Z}} \Upsilon\left(\frac{2x}{X}\right) \Upsilon\left(\frac{y}{X}\right) e(\theta\,\mathfrak{f}(2x, y)) \sum_{\substack{u|(2x,y) \\ u<U}} \mu(u), \qquad (4.9)$$

with corresponding "representation number" $\mathcal{R}_N^U$ (which could be negative).

4.2. Reduction to the Circle Method.

We are now in position to outline the argument in the rest of the paper. Recall the set $\mathscr{A}$ of admissible numbers. We first reduce our main Theorem 1.7 to the following

Theorem 4.10. *There exists an $\eta > 0$ and a function $\mathfrak{S}(n)$ with the following properties. For $\frac{1}{2}N < n < N$, the singular series $\mathfrak{S}(n)$ is nonnegative, vanishes only when $n \notin \mathscr{A}$, and is otherwise $\gg_{\varepsilon} N^{-\varepsilon}$ for any $\varepsilon > 0$. Moreover, for $\frac{1}{2}N < n < N$ and admissible,*

$$\mathcal{R}_N^U(n) \gg \mathfrak{S}(n) T^{\delta-1}, \tag{4.11}$$

except for a set of cardinality $\ll N^{1-\eta}$.

Proof of Theorem 1.7 assuming Theorem 4.10:

We first show that the difference between $\mathcal{R}_N$ and $\mathcal{R}_N^U$ is small in ℓ^1. Using (4.4) we have

$$
\begin{aligned}
\sum_{n<N} |\mathcal{R}_N(n) - \mathcal{R}_N^U(n)| &= \sum_{n<N} \left| \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{x,y\in\mathbb{Z}} \Upsilon\left(\frac{2x}{X}\right) \Upsilon\left(\frac{y}{X}\right) \mathbf{1}_{\{n=\mathfrak{f}(2x,y)\}} \sum_{\substack{u|(2x,y)\\u\geq U}} \mu(u) \right| \\
&\ll \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{y<X} \sum_{\substack{u|y\\u\geq U}} \sum_{\substack{x<X\\2x\equiv 0(\mathrm{mod}\,u)}} 1 \\
&\ll_{\varepsilon} T^{\delta} X\, X^{\varepsilon} \frac{X}{U},
\end{aligned}
$$

for any $\varepsilon > 0$. Recall from (4.8) that U is a fixed power of N, so the above saves a power from the total mass (4.7).

Now let Z be the "exceptional" set of admissible $n < N$ for which $\mathcal{R}_N(n) = 0$. Futhermore, let W be the set of admissible $n < N$ for which (4.11) is satisfied. Then

$$
\begin{aligned}
T^{\delta} X^2 \frac{X^{\varepsilon}}{U} &\gg_{\varepsilon} \sum_{n<N} |\mathcal{R}_N^U(n) - \mathcal{R}_N(n)| \geq \sum_{n\in Z\cap W} |\mathcal{R}_N^U(n) - \mathcal{R}_N(n)| \\
&\gg_{\varepsilon} |Z\cap W| \cdot T^{\delta-1} N^{-\varepsilon}.
\end{aligned}
$$

Note also from Theorem 4.10 that $|Z \cap W^c| \ll N^{1-\eta}$. Hence by (4.1),

$$|Z| = |Z\cap W^c| + |Z\cap W| \ll_{\varepsilon} N^{1-\eta} + \frac{N^{1+\varepsilon}}{U}, \tag{4.12}$$

which is a power savings since $\varepsilon > 0$ is arbitrary. This completes the proof. □

To establish (4.11), we decompose $\mathcal{R}_N^U$ into "major" and "minor" arcs, reducing Theorem 4.10 to the following

Theorem 4.13. *There exists an $\eta > 0$ and a decomposition*

$$\mathcal{R}_N^U(n) = \mathcal{M}_N^U(n) + \mathcal{E}_N^U(n) \tag{4.14}$$

with the following properties. For $\frac{1}{2}N < n < N$ and admissible, $n \in \mathscr{A}$, we have

$$\mathcal{M}_N^U(n) \gg \mathfrak{S}(n)T^{\delta-1}, \tag{4.15}$$

except for a set of cardinality $\ll N^{1-\eta}$. The singular series $\mathfrak{S}(n)$ is the same as in Theorem 4.10. Moreover,

$$\sum_{n<N} |\mathcal{E}_N^U(n)|^2 \ll N\,T^{2(\delta-1)}N^{-\eta}. \tag{4.16}$$

Proof of Theorem 4.10 assuming Theorem 4.13:

We restrict our attention to the set of admissible $n < N$ so that (4.15) holds (the remainder having sufficiently small cardinality). Let Z denote the subset of these n for which $\mathcal{R}_N^U(n) < \frac{1}{2}\mathcal{M}_N^U(n)$; hence for $n \in Z$,

$$1 \ll \frac{|\mathcal{E}_N^U(n)|}{N^{-\varepsilon}T^{\delta-1}}.$$

Then by (4.16),

$$|Z| \quad \ll_\varepsilon \quad \sum_{n<N} \frac{|\mathcal{E}_N^U(n)|^2}{N^{-\varepsilon}T^{2(\delta-1)}} \ll N^{1-\eta+\varepsilon},$$

whence the claim follows, since $\varepsilon > 0$ is arbitrary. $\qquad\square$

4.3. **Decomposition into Major and Minor Arcs.**

Next we explicate the decomposition (4.14). Let M be a parameter controlling the depth of approximation in Dirichlet's theorem: for any irrational $\theta \in [0,1]$, there exists some $q < M$ and $(r,q) = 1$ so that $|\theta - r/q| < 1/(qM)$. We will eventually set

$$M = XT, \tag{4.17}$$

see (7.9) where this value is used. (Note that M is a bit bigger than $N^{1/2} = XT^{1/2}$.)

Writing $\theta = r/q + \beta$, we introduce parameters

$$Q_0, K_0, \tag{4.18}$$

small powers of N as determined in (6.2), so that the "major arcs" correspond to $q < Q_0$ and $|\beta| < K_0/N$. In fact, we need a smooth version of this decomposition.

To this end, recall the "hat" function and its Fourier transform

$$\mathfrak{t}(x) := \min(1 + x, 1 - x)^+, \qquad \widehat{\mathfrak{t}}(y) = \left(\frac{\sin(\pi y)}{\pi y}\right)^2. \qquad (4.19)$$

Localize $\mathfrak{t}$ to the width K_0/N, periodize it to the circle, and put this spike on each fraction in the major arcs:

$$\mathfrak{T}(\theta) = \mathfrak{T}_{N,Q_0,K_0}(\theta) := \sum_{q < Q_0} \sum_{(r,q)=1} \sum_{m \in \mathbb{Z}} \mathfrak{t}\left(\frac{N}{K_0}\left(\theta + m - \frac{r}{q}\right)\right). \qquad (4.20)$$

By construction, $\mathfrak{T}$ lives on the circle $\mathbb{R}/\mathbb{Z}$ and is supported within K_0/N of fractions r/q with small denominator, $q < Q_0$, as desired.

Then define the "main term"

$$\mathcal{M}_N^U(n) := \int_0^1 \mathfrak{T}(\theta)\widehat{\mathcal{R}_N^U}(\theta)e(-n\theta)d\theta, \qquad (4.21)$$

and "error term"

$$\mathcal{E}_N^U(n) := \int_0^1 (1 - \mathfrak{T}(\theta))\widehat{\mathcal{R}_N^U}(\theta)e(-n\theta)d\theta, \qquad (4.22)$$

so that (4.14) obviously holds.

Since $\mathcal{R}_N^U$ could be negative, the same holds for $\mathcal{M}_N^U$. Hence we will establish (4.15) by first proving a related result for

$$\mathcal{M}_N(n) := \int_0^1 \mathfrak{T}(\theta)\widehat{\mathcal{R}_N}(\theta)e(-n\theta)d\theta, \qquad (4.23)$$

and then showing that $\mathcal{M}_N$ and $\mathcal{M}_N^U$ cannot differ by too much for too many values of n. This is the same (but in reverse) as the transfer from $\mathcal{R}_N$ to $\mathcal{R}_N^U$ in (4.12). See Theorem 6.7 for the lower bound on $\mathcal{M}_N$, and Theorem 6.10 for the transfer.

To prove (4.16), we apply Parseval and decompose dyadically:

$$\begin{aligned}
\sum_n |\mathcal{E}_N^U(n)|^2 &= \int_0^1 |1 - \mathfrak{T}(\theta)|^2 \left|\widehat{\mathcal{R}_N^U}(\theta)\right|^2 d\theta \\
&\ll \mathcal{I}_{Q_0,K_0} + \mathcal{I}_{Q_0} + \sum_{\substack{Q_0 \le Q < M \\ \text{dyadic}}} \mathcal{I}_Q,
\end{aligned}$$

where we have dissected the circle into the following regions (using that $|1 - \mathfrak{t}(x)| = |x|$ on $[-1, 1]$):

$$\mathcal{I}_{Q_0, K_0} \;\; := \int\limits_{\substack{\theta = \frac{r}{q} + \beta \\ q < Q_0, (r,q)=1, |\beta| < K_0/N}} \left| \beta \frac{N}{K_0} \right|^2 \left| \widehat{\mathcal{R}_N^U}(\theta) \right|^2 d\theta, \qquad (4.24)$$

$$\mathcal{I}_{Q_0} \;\; := \int\limits_{\substack{\theta = \frac{r}{q} + \beta \\ q < Q_0, (r,q)=1, K_0/N < |\beta| < 1/(qM)}} \left| \widehat{\mathcal{R}_N^U}(\theta) \right|^2 d\theta, \qquad (4.25)$$

$$\mathcal{I}_Q \;\; := \int\limits_{\substack{\theta = \frac{r}{q} + \beta \\ Q \le q < 2Q, (r,q)=1, |\beta| < 1/(qM)}} \left| \widehat{\mathcal{R}_N^U}(\theta) \right|^2 d\theta. \qquad (4.26)$$

Bounds of the quality (4.16) are given for (4.24) and (4.25) in §7, see Theorem 7.11. Our estimation of (4.26) decomposes further into two cases, whether $Q < X$ or $X \le Q < M$, and are handled separately in §8 and §9; see Theorems 8.14 and 9.20, respectively.

4.4. **The Rest of the Paper.**

The only section not yet described is §5, where we furnish some lemmata which are useful in the sequel. These decompose into two categories: one set of lemmata is related to some infinite-volume counting problems, for which the background in §3 is indispensable. The other lemma is of a classical flavor, corresponding to a local analysis for the shifted binary form $\mathfrak{f}$; this studies a certain exponential sum which is dealt with via Gauss and Kloosterman/Salié sums.

This completes our outline of the rest of the paper.

5. Some Lemmata

5.1. **Infinite Volume Counting Statements.**

Equipped with the tools of §3, we isolate here some consequences which will be needed in the sequel. We return to the notation $G = \mathrm{SO}_F$, with F the Descartes from (2.2), $\Gamma = \mathcal{A} \cap G$, the orientation preserving Apollonian subgroup, and $\Gamma(q)$ its principal congruence subgroups. Moreover, we import all the notation from the previous section.

First we use the spectral gap to see that summing over a coset of a congruence group can be reduced to summing over the original group.

Lemma 5.1. *Fix $\gamma_1 \in \Gamma$, $q \geq 1$, and any "congruence" group $\Gamma_0(q)$ satisfying*

$$\Gamma(q) < \Gamma_0(q) < \Gamma. \tag{5.2}$$

Then as $Y \to \infty$,

$$\#\{\gamma \in \Gamma_0(q) : \|\gamma_1\gamma\| < Y\} \tag{5.3}$$

$$= \frac{1}{[\Gamma : \Gamma_0(q)]} \cdot \#\{\gamma \in \Gamma : \|\gamma\| < Y\} + O(Y^{\Theta_0}), \tag{5.4}$$

where $\Theta_0 < \delta$ depends only on the spectral gap for Γ. The implied constant above does not depend on q or γ_1. The same holds with $\gamma_1\gamma$ in (5.3) replaced by $\gamma\gamma_1$.

This simple lemma follows from a more-or-less standard argument. We give a sketch below, since a slightly more complicated result will be needed later, cf. Lemma 5.24, but with essentially no new ideas. After proving the lemma below, we will use the argument as a template for the more complicated statement.

Sketch of Proof.

Denote the left hand side (5.3) by $\mathcal{N}_q$, and let $\mathcal{N}_1/[\Gamma : \Gamma_0(q)]$ be the first term of (5.4). For $g \in G$, let

$$f(g) = f_Y(g) := \mathbf{1}_{\{\|g\|<Y\}}, \tag{5.5}$$

and define

$$F_q(g, h) := \sum_{\gamma \in \Gamma_0(q)} f(g^{-1}\gamma h), \tag{5.6}$$

so that

$$\mathcal{N}_q = F_q(\gamma_1^{-1}, e). \tag{5.7}$$

By construction, F_q is a function on $\Gamma_0(q)\backslash G \times \Gamma_0(q)\backslash G$, and we smooth F_q in two copies of $\Gamma_0(q)\backslash G$, as follows. Let $\psi \geq 0$ be a smooth

bump function supported in a ball of radius $\eta > 0$ (to be chosen later) about the origin in G with $\int_G \psi = 1$, and automorphize it to

$$\Psi_q(g) := \sum_{\gamma \in \Gamma_0(q)} \psi(\gamma g).$$

Then clearly Ψ_q is a bump function in $\Gamma_0(q) \backslash G$ with $\int_{\Gamma_0(q) \backslash G} \Psi_q = 1$. Let

$$\Psi_{q,\gamma_1}(g) := \Psi_q(g\gamma_1).$$

Smooth the variables g and h in F_q by considering

$$\mathcal{H}_q := \langle F_q, \Psi_{q,\gamma_1} \otimes \Psi_q \rangle = \int_{\Gamma_0(q) \backslash G} \int_{\Gamma_0(q) \backslash G} F_q(g,h) \Psi_{q,\gamma_1}(g) \Psi_q(h) dg\, dh$$

$$= \sum_{\gamma \in \Gamma(q)} \int_{\Gamma(q) \backslash G} \int_{\Gamma(q) \backslash G} f(\gamma_1 g^{-1} \gamma h) \Psi_q(g) \Psi_q(h) dg\, dh.$$

First we estimate the error from smoothing:

$$\mathcal{E} = |\mathcal{N}_q - \mathcal{H}_q|$$

$$\leq \sum_{\gamma \in \Gamma} \int_{\Gamma_0(q) \backslash G} \int_{\Gamma_0(q) \backslash G} |f(\gamma_1 g^{-1} \gamma h) - f(\gamma_1 \gamma)| \Psi_q(g) \Psi_q(h) dg\, dh,$$

where we have increased γ to run over all of Γ. The analysis splits into three ranges.

(1) If γ is such that

$$\|\gamma_1 \gamma\| > Y(1 + 10\eta), \tag{5.8}$$

then both $f(\gamma_1 g^{-1} \gamma h)$ and $f(\gamma_1 \gamma)$ vanish.

(2) In the range

$$\|\gamma_1 \gamma\| < Y(1 - 10\eta), \tag{5.9}$$

both $f(\gamma_1 g^{-1} \gamma h)$ and $f(\gamma_1 \gamma)$ are 1, so their difference vanishes.

(3) In the intermediate range, we apply [LP82], bounding the count by

$$\ll Y^\delta \eta + Y^{\delta - \varepsilon}, \tag{5.10}$$

where $\varepsilon > 0$ depends on the spectral gap for Γ.

Thus it remains to analyze $\mathcal{H}_q$.

Use a simple change of variables (see [BKS10, Lemma 3.7]) to express $\mathcal{H}_q$ via matrix coefficients:

$$\mathcal{H}_q = \int_G f(g) \langle \pi(g) \Psi_q, \Psi_{q,\gamma_1} \rangle_{\Gamma_0(q) \backslash G}\, dg.$$

Decompose the matrix coefficient into its projection onto the base irreducible V_{λ_0} in (3.11) and an orthogonal term, and bound the remainder by the mixing rate (3.13) using the uniform spectral gap $\varepsilon > 0$ in (3.9). The functions ψ are bump functions in six real dimensions, so can be chosen to have second-order Sobolev norms bounded by $\ll \eta^{-5}$. Of course the projection onto the base representation is just $[\Gamma : \Gamma_0(q)]^{-1}$ times the same projection at level one, cf. (3.7). Running the above argument in reverse at level one (see [BKS10, Prop. 4.18]) gives:

$$\mathcal{N}_q = \frac{1}{[\Gamma : \Gamma(q)]} \cdot \mathcal{N}_1 + O(\eta Y^\delta + Y^{\delta-\varepsilon}) + O(Y^{\delta-\varepsilon}\eta^{-10}). \qquad (5.11)$$

Optimizing η and renaming $\Theta_0 < \delta$ in terms of the spectral gap ε gives the claim. $\qquad \square$

Next we exploit the previous lemma and the product structure of the family $\mathfrak{F}$ in (4.3) to save a small power of q in the following modular restriction. Such a bound is needed at several places in §8.

Lemma 5.12. *Fix $1 \le q < N$ and any $r(\mathrm{mod}\, q)$. Let Θ_0 be as in (5.4). Define $\mathcal{C}$ in (4.2) by*

$$\mathcal{C} := \frac{10^{30}}{\delta - \Theta_0}, \qquad (5.13)$$

hence determining T_1 and T_2. Then there exists some $\eta_0 > 0$ depending only on the spectral gap of Γ so that

$$\sum_{\gamma \in \mathfrak{F}} \mathbf{1}_{\{\langle e_1, \gamma v_0 \rangle \equiv r(\mathrm{mod}\, q)\}} \ll \frac{1}{q^{\eta_0}} T^\delta. \qquad (5.14)$$

The implied constant is independent of r.

Proof. Dropping the condition $\langle e_1, \gamma_1 \gamma_2 v_0 \rangle > T/100$ in (4.3), bound the left hand side of (5.14) by

$$\sum_{\substack{\gamma_1 \in \Gamma \\ \|\gamma_1\| \asymp T_1}} \sum_{\substack{\gamma_2 \in \Gamma \\ \|\gamma_2\| \asymp T_2}} \mathbf{1}_{\{\langle e_1, \gamma_1 \gamma_2 v_0 \rangle \equiv r(\mathrm{mod}\, q)\}} \qquad (5.15)$$

We decompose the argument into two ranges of q.

Case 1: q small. In this range, we fix γ_1, and follow a standard argument for γ_2. Let $\Gamma_0(q) < \Gamma$ denote the stabilizer of $v_0(\mathrm{mod}\, q)$, that is

$$\Gamma_0(q) := \{\gamma \in \Gamma : \gamma v_0 \equiv v_0(\mathrm{mod}\, q)\}. \qquad (5.16)$$

Clearly (5.2) is satisfied, and it is elementary that

$$[\Gamma : \Gamma_0(q)] \asymp q^2, \qquad (5.17)$$

cf. (2.22). Decompose $\gamma_2 = \gamma_2'\gamma_2''$ with $\gamma_2'' \in \Gamma_0(q)$ and $\gamma_2' \in \Gamma/\Gamma_0(q)$. Then by (5.4) and [LP82], we have

$$(5.15) \quad = \sum_{\substack{\gamma_1 \in \Gamma \\ \|\gamma_1\| \asymp T_1}} \sum_{\gamma_2' \in \Gamma/\Gamma_0(q)} \mathbf{1}_{\{\langle e_1, \gamma_1\gamma_2'v_0\rangle \equiv r(\mathrm{mod}\, q)\}} \sum_{\substack{\gamma_2'' \in \Gamma_0(q) \\ \|\gamma_2'\gamma_2''\| \asymp T_2}} 1$$

$$\ll \; T_1^\delta \, q \, \left(\frac{1}{q^2}\, T_2^\delta + T_2^{\Theta_0}\right).$$

Hence we have saved a whole power of q, as long as

$$q < T_2^{(\delta-\Theta_0)/2}. \tag{5.18}$$

Case 2: $q \geq T_2^{\frac{\delta-\Theta_0}{2}}$. Then by (5.13) and (4.2), q is actually a huge power of T_1,

$$q \geq T_1^{10^{29}}. \tag{5.19}$$

In this range, we exploit Hilbert's Nullstellensatz and effective versions of Bezout's theorem; see a related argument in [BG09, Proof of Prop. 4.1].

Fixing γ_2 in (5.15) (with $\ll T_2^\delta$ choices), we set

$$v := \gamma_2 v_0,$$

and play now with γ_1. Let S be the set of γ_1's in question (and we now drop the subscript 1):

$$S = S_{v,q}(T_1) := \{\gamma \in \Gamma : \|\gamma\| \asymp T_1, \langle e_1, \gamma v\rangle \equiv r(\mathrm{mod}\, q)\}.$$

This congruence restriction is to a modulus much bigger than the parameter, so we

Claim: There is an integer vector $v_* \neq 0$ and an integer z_* such that

$$\langle e_1, \gamma v_*\rangle = z_* \tag{5.20}$$

holds for all $\gamma \in S$. That is, the modular condition can be lifted to an exact equality.

First we assume the Claim and complete the proof of (5.14). Let q_0 be a prime of size $\asymp T_1^{(\delta-\Theta_0)/2}$, say, such that $v_* \not\equiv 0(\mathrm{mod}\, q_0)$; then

$$|S| \; \ll \; \#\{\|\gamma_1\| < T_1 : \langle e_1, \gamma v_*\rangle \equiv z_*(\mathrm{mod}\, q_0)\}$$

$$\ll \; q_0 \left(\frac{1}{q_0^2}T_1^\delta + T_1^{\Theta_0}\right) \ll \frac{1}{q_0}T_1^\delta,$$

by the argument in Case 1. Recall we assumed that $q < N$. Since q_0 above is a small power of N, the above saves a tiny power of q, as desired.

It remains to establish the Claim. For each $\gamma \in S$, consider the condition

$$\langle e_1, \gamma\, v \rangle = \sum_{1 \leq j \leq 4} \gamma_{1,j}\, v_j \equiv r(\mathrm{mod}\, q).$$

First massage the equation into one with no trivial solutions. Since v is a primitive vector, after a linear change of variables we may assume that $(v_1, q) = 1$. Then multiply through by $\bar{v}_1$, where $v_1 \bar{v}_1 \equiv 1 (\mathrm{mod}\, q)$, getting

$$\gamma_{1,1} + \sum_{2 \leq j \leq 4} \gamma_{1,j}\, v_j \bar{v}_1 \equiv r \bar{v}_1 (\mathrm{mod}\, q). \tag{5.21}$$

Now, for variables $V = (V_2, V_3, V_4)$ and Z, and each $\gamma \in S$, consider the (linear) polynomials $P_\gamma \in \mathbb{Z}[V, Z]$:

$$P_\gamma(V, Z) := \gamma_{1,1} + \sum_{2 \leq j \leq 4} \gamma_{1,j}\, V_j - Z,$$

and the affine variety

$$\mathcal{V} := \bigcap_{\gamma \in S} \{P_\gamma = 0\}.$$

If this variety $\mathcal{V}(\mathbb{C})$ is non-empty, then there is clearly a rational solution, $(V^*, Z^*) \in \mathcal{V}(\mathbb{Q})$. Hence we have found a rational solution to (5.20), namely $v^* = (1, V_2^*, V_3^*, V_4^*) \neq 0$ and $z^* = Z^*$. Since (5.20) is projective, we may clear denominators, getting an integral solution, v_*, z_*.

Thus we henceforth assume by contradiction that the variety $\mathcal{V}(\mathbb{C})$ is empty. Then by Hilbert's Nullstellensatz, there are polynomials $Q_\gamma \in \mathbb{Z}[V, Z]$ and an integer $\mathfrak{d} \geq 1$ so that

$$\sum_{\gamma \in S} P_\gamma(V, Z)\, Q_\gamma(V, Z) = \mathfrak{d}, \tag{5.22}$$

for all $(V, Z) \in \mathbb{C}^4$. Moreover, Hermann's method [Her26] (see [MW83, Theorem IV]) gives effective bounds on the heights of Q_γ and $\mathfrak{d}$ in the above Bezout equation. Recall the height of a polynomial is the logarithm of its largest coefficient (in absolute value); thus the polynomials P_γ are linear in four variables with height $\leq \log T_1$. Then Q_γ and $\mathfrak{d}$ can be found so that

$$\mathfrak{d} \leq e^{8^{4 \cdot 2^{4-1}-1}(\log T_1 + 8 \log 8)} \ll T_1^{10^{28}}. \tag{5.23}$$

(Much better bounds are known, see e.g. [BY91, Theorem 5.1], but these suffice for our purposes.)

On the other hand, reducing (5.22) modulo q and evaluating at

$$V_0 = (v_2 \bar{v}_1, v_3 \bar{v}_1, v_4 \bar{v}_1), \qquad Z_0 = r \bar{v}_1,$$

we have

$$\sum_{\gamma \in S} P_\gamma(V_0, Z_0) Q_\gamma(V_0, Z_0) \equiv 0 \equiv \mathfrak{d} (\mathrm{mod}\, q),$$

by (5.21). But then since $\mathfrak{d} \geq 1$, we in fact have $\mathfrak{d} \geq q$, which is incompatible with (5.23) and (5.19). This furnishes our desired contradiction, completing the proof. $\square$

Next we need a slight generalization of Lemma 5.1, which will be used in the major arcs analysis, see (6.6).

Lemma 5.24. *Let $1 < K \leq T_2^{1/10}$, fix $|\beta| < K/N$, and fix $x, y \asymp X$. Then for any $\gamma_0 \in \Gamma$, any $q \geq 1$, and any group $\Gamma_0(q)$ satisfying (5.2), we have*

$$\sum_{\gamma \in \mathfrak{F} \cap \{\gamma_0 \Gamma_0(q)\}} e\left(\beta \, \mathfrak{f}_\gamma(2x, y) \right) \;=\; \frac{1}{[\Gamma : \Gamma_0(q)]} \sum_{\gamma \in \mathfrak{F}} e\left(\beta \, \mathfrak{f}_\gamma(2x, y) \right)$$
$$+ O(T^\Theta K), \quad (5.25)$$

where $\Theta < \delta$ depends only on the spectral gap for Γ, and the implied constant does not depend on q, γ_0, β, x or y.

Proof. The proof follows with minor changes that of Lemma 5.1, so we give a sketch; see also [BKS10, §4].

According to the construction (4.3) of $\mathfrak{F}$, the γ's in question satisfy $\gamma = \gamma_1 \gamma_2 \in \gamma_0 \Gamma_0(q)$, and hence we can write

$$\gamma_2 = \gamma_1^{-1} \gamma_0 \gamma_2',$$

with $\gamma_2' \in \Gamma_0(q)$. Then $\gamma_2' = \gamma_0^{-1} \gamma_1 \gamma_2$, and using (2.16), we can write the left hand side of (5.25) as

$$\sum_{\substack{\gamma_1 \in \Gamma \\ T_1 < \|\gamma_1\| < 2T_1}} \sum_{\substack{\gamma_2' \in \Gamma_0(q) \\ T_2 < \|\gamma_1^{-1} \gamma_0 \gamma_2'\| < 2T_2}} \mathbf{1}_{\{\langle e_1, \gamma_0 \gamma_2' \, v_0 \rangle > T/100\}} \, e\left(\beta \, \langle w_{x,y}, \gamma_0 \gamma_2' \, v_0 \rangle \right).$$

Now we fix γ_1 and mimic the proof of Lemma 5.1 in γ_2'.

Replace (5.5) by

$$f(g) := \mathbf{1}_{\{T_2 < \|\gamma_1^{-1} g\| < 2T_2\}} \mathbf{1}_{\{\langle e_1, g\, v_0 \rangle > T/100\}} \, e\left(\beta \, \langle w_{x,y}, g\, v_0 \rangle \right).$$

Then (5.6)-(5.8) remains essentially unchanged, save cosmetic changes such as replacing (5.7) by $F_q(\gamma_1 \gamma_0^{-1}, e)$. Then in the estimation of the difference $|\mathcal{N}_q - \mathcal{H}_q|$ by splitting the sum on γ_2' into ranges, the argument now proceeds as follows.

(1) The range (5.8) should be replaced by

$$\|\gamma_1\gamma_0^{-1}\gamma_2'\| < T_2(1 - 10\eta), \text{ or } \|\gamma_1\gamma_0^{-1}\gamma_2'\| > 2T_2(1 + 10\eta),$$

$$\text{or } \langle e_1, \gamma_1\gamma_0^{-1}\gamma_2' v_0 \rangle < \frac{T}{100}(1 - 10\eta).$$

(2) The range (5.9) should be replaced by the buffered range in which f is differentiable,

$$T_2(1+10\eta) < \|\gamma_1\gamma_0^{-1}\gamma_2'\| < 2T_2(1-10\eta), \text{ and } \langle e_1, \gamma_1\gamma_0^{-1}\gamma_2' v_0 \rangle > \frac{T}{100}(1+10\eta).$$

Here instead of the difference $|f(\gamma_1\gamma_0^{-1}g\gamma_2'h) - f(\gamma_1\gamma_0^{-1}\gamma_2')|$ vanishing, it is now bounded by

$$\ll \eta K,$$

for a net contribution to the error of $\ll \eta K T^\delta$.

(3) In the remaining range, (5.10) remains unchanged, using $|f| \le 1$.

The error in (5.11) is then replaced by

$$O(\eta\,K\,T_2^\delta + T_2^{\delta-\varepsilon}\eta^{-10}).$$

Optimizing η and renaming Θ gives the bound $O(T_2^\Theta K^{10/11})$, which is better than claimed in the power of K. Rename Θ once more using (4.2) and (5.13), giving (5.25).   □

The following is our last counting lemma, showing a certain equidistribution among the values of $\mathfrak{f}_\gamma(2x, y)$ at the scale N/K. This bound is used in the major arcs, see the proof of Theorem 6.7.

Lemma 5.26. *Fix $N/2 < n < N$, $1 < K \le T_2^{1/10}$, and $x, y \asymp X$. Then*

$$\sum_{\gamma \in \mathfrak{F}} \mathbf{1}_{\left\{|\mathfrak{f}_\gamma(2x,y)-n| < \frac{N}{K}\right\}} \gg \frac{T^\delta}{K} + T^\Theta, \tag{5.27}$$

where $\Theta < \delta$ only depends on the spectral gap for Γ. The implied constant is independent of x, y, and n.

Sketch. The proof is a tedious explicit calculation nearly identical to the one given in [BKS10, §5], so we give a sketch. Write the left hand side of (5.27) as

$$\sum_{\substack{\gamma_1 \in \Gamma \\ T_1 < \|\gamma_1\| < 2T_1}} \sum_{\substack{\gamma_2 \in \Gamma \\ T_2 < \|\gamma_2\| < 2T_2}} \mathbf{1}_{\{\langle e_1, \gamma_1\gamma_2 v_0 \rangle > T/100\}} \mathbf{1}_{\{|\langle w_{x,y}, \gamma_1\gamma_2 v_0 \rangle - n| < N/K\}}.$$

Fix γ_1 and express the condition on γ_2 as $\gamma_2 \in R \subset G$, where R is the region

$$R = R_{\gamma_1, x, y, n} := \left\{ g \in G : \begin{array}{c} T_2 < \|g\| < 2T_2 \\ \langle \gamma_1^t e_1, g\, v_0 \rangle > T/100 \\ |\langle \gamma_1^t w_{x,y}, g\, v_0 \rangle - n| < \frac{N}{K} \end{array} \right\}.$$

Lift $G = \mathrm{SO}_F(\mathbb{R})$ to its spin cover $\tilde{G} = \mathrm{SL}_2(\mathbb{C})$ via the map ι of (2.19). Let $\tilde{R} \subset \tilde{G}$ be the corresponding pullback region, and decompose $\tilde{G}$ into Cartan KAK coordinates according to (3.14). Note that ι is quadratic in the entries, so, e.g., the condition

$$\|g\|^2 \asymp T \text{ gives } \|\iota(g)\| \asymp T, \tag{5.28}$$

explaining the factor $\|a(g)\|^2$ appearing in (3.16).

Then chop $\tilde{R}$ into sectors and apply Theorem 3.15. The same argument as in [BKS10, §5] then leads to (5.27), after renaming Θ; we suppress the details. $\qquad\square$

5.2. **Local Analysis Statements.**

In this subsection, we study a certain exponential sum which arises in a crucial way in our estimates. Fix $\mathfrak{f} \in \mathfrak{F}$, and write $\mathfrak{f} = f - a$ with

$$f(x, y) = Ax^2 + 2Bxy + Cy^2$$

according to (2.15). Let $q_0 \geq 1$, fix r with $(r, q_0) = 1$, and fix $n, m \in \mathbb{Z}$. Define the exponential sum

$$\mathcal{S}_f(q_0, r; n, m) := \frac{1}{q_0^2} \sum_{k(q_0)} \sum_{\ell(q_0)} e_{q_0}\left(rf(k, \ell) + nk + m\ell \right). \qquad (5.29)$$

This sum appears naturally in many places in the minor arcs analysis, see e.g. (7.5) and (9.2).

Lemma 5.30. *With the above conditions,*

$$|\mathcal{S}_f(q_0, r; n, m)| \leq q_0^{-1/2}. \qquad (5.31)$$

Proof. Write $\mathcal{S}_f$ for $\mathcal{S}_f(q_0, r; n, m)$. Note first that $\mathcal{S}_f$ is multiplicative in q_0, so we study the case $q_0 = p^j$ is a prime power. Assume for simplicity $(q_0, 2) = 1$; similar calculations are needed to handle the 2-adic case.

First we re-express $\mathcal{S}_f$ in a more convenient form. By Descartes theorem (2.1), primitivity of the packing $\mathscr{P}$, and (2.14), we have that $(A, B, C) = 1$. So we can assume henceforth that $(C, q_0) = 1$, say. Write $\bar{x}$ for the multiplicative inverse of x (the modulus will be clear from context). Recall throughout that $(r, q_0) = 1$.

Looking at the terms in the summand of $\mathcal{S}_f$, we have

$$rf(k, \ell) + nk + m\ell \pmod{q_0}$$
$$\equiv r(Ak^2 + 2Bk\ell + C\ell^2) + nk + m\ell$$
$$\equiv rC(\ell + 2B\bar{C}k)^2 + 4r\bar{C}k^2(AC - B^2) + nk + m\ell$$
$$\equiv rC(\ell + 2B\bar{C}k)^2 + 4a^2 r\bar{C}k^2 + nk + m\ell$$
$$\equiv rC(\ell + 2B\bar{C}k + \overline{2rC}m)^2 - \overline{4rC}m^2 + 4a^2 r\bar{C}k^2 + k(n - 2B\bar{C}m),$$

where we used (2.17). Hence we have

$$\mathcal{S}_f = \frac{1}{q^2} e_{q_0}(-\overline{4rC}m^2) \sum_{k(q_0)} e_{q_0}\left(4a^2 r\bar{C}k^2 + k(n - 2B\bar{C}m) \right)$$
$$\times \sum_{\ell(q_0)} e_{q_0}\left(rC(\ell + 2B\bar{C}k + \overline{2rC}m)^2 \right),$$

and the ℓ sum is just a classical Gauss sum. It can be evaluated explicitly, see e.g. [IK04, eq. (3.38)]. Let

$$\varepsilon_{q_0} := \begin{cases} 1 & \text{if } q_0 \equiv 1 (\text{mod } 4) \\ i & \text{if } q_0 \equiv 3 (\text{mod } 4). \end{cases}$$

Then the Gauss sum on ℓ is $\varepsilon_{q_0} \sqrt{q_0} \left(\frac{rC}{q_0} \right)$, where $\left(\frac{\cdot}{q_0} \right)$ is the Legendre symbol. Thus we have

$$\mathcal{S}_f = \frac{\varepsilon_{q_0}}{q_0^{3/2}} \left(\frac{rC}{q_0} \right) e_{q_0}(-\overline{4rC}m^2) \sum_{k(q_0)} e_{q_0}\left(4a^2 r\bar{C}k^2 + k(n - 2B\bar{C}m) \right).$$

Let

$$\tilde{q}_0 := (a^2, q_0), \qquad q_1 := q_0/\tilde{q}_0, \qquad \text{and} \qquad a_1 := a^2/\tilde{q}_0, \qquad (5.32)$$

so that $a^2/q_0 = a_1/q_1$ in lowest terms. Break the sum on $0 \le k < q_0$ according to $k = k_1 + q_1\tilde{k}$, with $0 \le k_1 < q_1$ and $0 \le \tilde{k} < \tilde{q}_0$. Then

$$\begin{aligned} \mathcal{S}_f = \;& \frac{\varepsilon_{q_0}}{q_0^{3/2}} \left(\frac{rC}{q_0} \right) e_{q_0}(-\overline{4rC}m^2) \\ & \times \sum_{k_1(q_1)} e_{q_1}\left(4a_1 r\bar{C}(k_1)^2 \right) e_{q_0}\left(k_1(n - 2B\bar{C}m) \right) \\ & \times \sum_{\tilde{k}(\tilde{q}_0)} e_{\tilde{q}_0}\left(\tilde{k}(n - 2B\bar{C}m) \right). \end{aligned}$$

The last sum vanishes unless $n - 2B\bar{C}m \equiv 0 (\text{mod } \tilde{q}_0)$, in which case it is $\tilde{q}_0$. In the latter case, define L by

$$L := (Cn - 2Bm)/\tilde{q}_0. \qquad (5.33)$$

Then we have

$$\begin{aligned} \mathcal{S}_f = \;& \mathbf{1}_{nC \equiv 2mB(\tilde{q}_0)} \frac{\varepsilon_{q_0}}{q_0^{3/2}} \left(\frac{rC}{q_0} \right) e_{q_0}(-\overline{4rC}m^2) \\ & \times e_{q_1}\left(-\overline{16a_1 rC}L^2 \right) \left[\sum_{k_1(q_1)} e_{q_1}\left(4a_1 r\bar{C}(k_1 + \overline{8a_1 r}L)^2 \right) \right] \tilde{q}_0. \end{aligned}$$

The Gauss sum in brackets is again evaluated as $\varepsilon_{q_1} q_1^{1/2} \left(\frac{4a_1 r \bar{C}}{q_1} \right)$, so we have

$$
\mathcal{S}_f(q_0, r; n, m) = \mathbf{1}_{nC \equiv 2mB(\tilde{q}_0)} \frac{\varepsilon_{q_0} \varepsilon_{q_1} \tilde{q}_0^{1/2}}{q_0} e_{q_0}(-\overline{4rC}m^2) \tag{5.34}
$$

$$
\times e_{q_1} \left(-\overline{16a_1 rC} L^2 \right) \left(\frac{rC}{q_0} \right) \left(\frac{a_1 r\bar{C}}{q_1} \right).
$$

The claim then follows trivially.   □

Next we introduce a certain average of a pair of such sums. Let f, q_0, r, n, and m be as before, and fix $q \equiv 0 \pmod{q_0}$ and $(u_0, q_0) = 1$. Let $\mathfrak{f}' \in \mathfrak{F}$ be another shifted form $\mathfrak{f}' = f' - a'$, with

$$
f'(x, y) = A'x^2 + 2B'xy + C'y^2.
$$

Also let $n', m' \in \mathbb{Z}$. Then define

$$
\mathcal{S} = \mathcal{S}(q, q_0, f, f', n, m, n', m'; u_0) \tag{5.35}
$$

$$
:= \sideset{}{'}\sum_{r(q)} \mathcal{S}_f(q_0, ru_0; n, m) \overline{\mathcal{S}_{f'}(q_0, ru_0; n', m')} e_q(r(a' - a)).
$$

This sum also appears naturally in the minor arcs analysis, see (8.2) and (9.4).

Lemma 5.36. *With the above notation, we have the estimate*

$$
|\mathcal{S}| \ll (q/q_0)^2 \frac{\{(a^2, q_0) \cdot ((a')^2, q_0)\}^{1/2}}{q^{5/4}} (a - a', q)^{1/4}. \tag{5.37}
$$

Proof. Observe that $\mathcal{S}$ is multiplicative in q, so we again consider the prime power case $q = p^j$, $p \neq 2$; then q_0 is also a prime power, since $q_0 \mid q$. As before, we may assume $(C, q_0) = (C', q_0) = 1$.

Recall $a_1, \tilde{q}_0$, and L given in (5.32) and (5.33), and let $a'_1, \tilde{q}'_0$ and L' be defined similarly. Inputting the analysis from (5.34) into both $\mathcal{S}_f$ and $\mathcal{S}_{f'}$, we have

$$
\mathcal{S} = \mathbf{1}_{\substack{nC \equiv 2mB(\tilde{q}_0) \\ n'C' \equiv 2m'B'(\tilde{q}'_0)}} \frac{\varepsilon_{q_1} \bar{\varepsilon}_{q'_1} (\tilde{q}_0 \tilde{q}'_0)^{1/2}}{q_0^2} \left(\frac{CC'}{q_0} \right) \left(\frac{a_1 u_0 \bar{C}}{q_1} \right) \left(\frac{a'_1 u_0 \bar{C}'}{q'_1} \right)
$$

$$
\times \left[\sideset{}{'}\sum_{r(q)} \left(\frac{r}{q_1} \right) \left(\frac{r}{q'_1} \right) e_q(r\{a' - a\}) \right. \tag{5.38}
$$

$$
\left. \times e_{q_0} \left(\bar{r}\bar{u}_0 \left\{ \bar{4}(\overline{C'}(m')^2 - \overline{C}m^2) + \overline{16}(\overline{a'_1 C'}(L')^2 \tilde{q}' - \overline{a_1 C} L^2 \tilde{q}) \right\} \right) \right].
$$

The term in brackets $\left[\,\cdot\,\right]$ is a Kloosterman- or Salié-type sum, for which we have an elementary bound to the power $3/4$:

$$|\mathcal{S}| \ \ll\ \frac{(\tilde{q}_0\tilde{q}_0')^{1/2}}{q_0^2}q^{3/4}(a-a',q)^{1/4},$$

giving the claim. $\hfill\square$

In the case $a = a'$, (5.37) only saves one power of q, and in §9 we will need slightly more; see the proof of (9.13). We get a bit more cancellation in the special case $f(m,-n) \neq f'(m',-n')$ below.

Lemma 5.39. *Assuming $a = a'$ and $f(m,-n) \neq f'(m',-n')$, we have the estimate*

$$|\mathcal{S}| \ll (q/q_0)^5 \frac{(a^2,q_0)}{q^{9/8}} \cdot |f(m,-n) - f'(m',-n')|^{1/2}. \qquad (5.40)$$

Proof. Assume first that q (and hence q_0) is a prime power, continuing to omit the prime 2. Returning to the definition of $\mathcal{S}$ in (5.35), it is clear in the case $a = a'$ that

$$\sideset{}{'}\sum_{r(q)} = (q/q_0)\sideset{}{'}\sum_{r(q_0)}.$$

Hence we again apply Kloosterman's $3/4$th bound to (5.38), getting

$$|\mathcal{S}| \ \ll\ \mathbf{1}_{\substack{nC\equiv 2mB(\tilde{q}_0)\\ n'C'\equiv 2m'B'(\tilde{q}_0')}} (q/q_0)^{9/2}\frac{(a^2,q_0)}{q^{5/4}} \qquad (5.41)$$

$$\times \prod_{p^j\|q_0}\left(p^j,\overline{4}(\overline{C'}(m')^2 - \overline{C}m^2) + \overline{16a_1}(a^2,p^j)(\overline{C'}(L')^2 - \overline{C}L^2)\right)^{1/4},$$

where we have now dropped the assumption that q_0 is a prime power. (Here a_1 satisfies $a^2 = a_1(a^2,p^j)$ as in (5.32), and L is given in (5.33), so both depend on p^j.)

Break the primes diving q_0 into two sets, $\mathcal{P}_1$ and $\mathcal{P}_2$, defining $\mathcal{P}_1$ to be the set of those primes p for which

$$\overline{C}m^2 + \overline{C}L^2\overline{4a_1}(a^2,p^j) \equiv \overline{C'}(m')^2 + \overline{C'}(L')^2\overline{4a_1}(a^2,p^j) \pmod{p^{\lfloor j/2\rfloor}},$$
$$(5.42)$$

and $\mathcal{P}_2$ the rest. For the latter, the gcd in $(p^j,\cdots)$ of (5.41) is at most $p^{j/2}$, so we clearly have

$$\prod_{\substack{p^j\|q_0\\ p\in\mathcal{P}_2}} (p^j,\cdots)^{1/4} \le \prod_{p^j\|q_0} p^{j/8} = q_0^{1/8}. \qquad (5.43)$$

For $p \in \mathcal{P}_1$, we multiply both sides of (5.42) by
$$4a^2 = 4(AC - B^2) = 4(A'C' - (B')^2) = 4a_1(a^2, p^j),$$
giving
$$4(AC - B^2)\overline{C}m^2 + \overline{C}L^2(a^2, p^j)^2 \tag{5.44}$$
$$\equiv \ 4(A'C' - (B')^2)\overline{C'}(m')^2 + \overline{C'}(L')^2(a^2, p^j)^2 \qquad (\mathrm{mod}\, p^{\lfloor j/2 \rfloor}).$$
Using (5.33) that
$$nC - 2mB = (a^2, p^j)L, \qquad n'C' - 2m'B' = (a^2, p^j)L'$$
and subtracting a from both sides of (5.44), we have shown that
$$f'(m', -n') \equiv f(m, -n) \qquad (\mathrm{mod}\, p^{\lfloor j/2 \rfloor}). \tag{5.45}$$
Let
$$Z = |f(m, -n) - f'(m', -n')|.$$
By assumption $Z \neq 0$. Moreover (5.45) implies that
$$\left(\prod_{p \in \mathcal{P}_1} p^{\lfloor j/2 \rfloor} \right) \mid Z,$$
and hence
$$\prod_{\substack{p^j \| q_0 \\ p \in \mathcal{P}_1}} p^{j/4} \leq Z^{1/2}. \tag{5.46}$$
Combining (5.46) and (5.43) in (5.41) gives the claim. $\qquad\square$

6. Major Arcs

We return to the setting and notation of §4 with the goal of establishing (4.15). Thanks to the counting lemmata in §5.1, we can now define the major ars parameters Q_0 and K_0 from (4.18). First recall the two numbers $\Theta < \delta$ appearing in (5.25), (5.27), and define

$$1 < \Theta_1 < \delta \tag{6.1}$$

to be the larger of the two. Then set

$$Q_0 = T^{(\delta - \Theta_1)/20}, \qquad K_0 = Q_0^2. \tag{6.2}$$

We may now also set the parameter U from (4.8) to be

$$U = Q_0^{(\eta_0)^2/100}, \tag{6.3}$$

where $0 < \eta_0 < 1$ is the number which appears in Lemma 5.12.

Let $\mathcal{M}_N^{(U)}(n)$ denote either $\mathcal{M}_N(n)$ or $\mathcal{M}_N^U(n)$ from (4.23), (4.21), respectively. Putting (4.20) and (4.6) (resp. (4.9)) into (4.23) (resp. (4.21)), making a change of variables $\theta = r/q + \beta$, and unfolding the integral from $\sum_m \int_0^1$ to $\int_{\mathbb{R}}$ gives

$$\mathcal{M}_N^{(U)}(n) = \sum_{x,y \in \mathbb{Z}} \Upsilon\left(\frac{2x}{X}\right) \Upsilon\left(\frac{y}{X}\right) \cdot \mathfrak{M}(n) \cdot \sum_u \mu(u), \tag{6.4}$$

where in the last sum, u ranges over $u \mid (2x, y)$ (resp. and $u < U$). Here we have defined

$$\mathfrak{M}(n) = \mathfrak{M}_{x,y}(n) \tag{6.5}$$
$$:= \sum_{q < Q_0} \sideset{}{'}\sum_{r(q)} \sum_{\gamma \in \mathfrak{F}} e_q(r(\langle w_{x,y}, \gamma v_0 \rangle - n)) \int_{\mathbb{R}} \mathfrak{t}\left(\frac{N}{K_0}\beta\right) e(\beta(\mathfrak{f}_\gamma(2x, y) - n)) d\beta,$$

using (2.16).

As in (5.16), let $\Gamma_0(q)$ be the stabilizer of $v_0 \pmod q$. Decompose the sum on $\gamma \in \mathfrak{F}$ in (6.5) as a sum on $\gamma_0 \in \Gamma/\Gamma_0(q)$ and $\gamma \in \mathfrak{F} \cap \gamma_0 \Gamma_0(q)$. Applying Lemma 5.24 to the latter sum, using the definition of Θ_1 in (6.1), and recalling the estimate (5.17) gives

$$\mathfrak{M}(n) = \mathfrak{S}_{Q_0}(n) \cdot \mathfrak{W}(n) + O\left(\frac{T^{\Theta_1}}{N} K_0^2 Q_0^4\right), \tag{6.6}$$

where

$$\mathfrak{S}_{Q_0}(n) \;:=\; \sum_{q<Q_0} \sideset{}{'}\sum_{r(q)} \sum_{\gamma_0\in\Gamma/\Gamma_0(q)} \frac{e_q(r(\langle w_{x,y},\gamma_0 v_0\rangle - n))}{[\Gamma:\Gamma_0(q)]},$$

$$\mathfrak{W}(n) \;:=\; \frac{K_0}{N} \sum_{\mathfrak{f}\in\mathfrak{F}} \widehat{\mathfrak{t}}\left((\mathfrak{f}(2x,y)-n)\frac{K_0}{N}\right).$$

Clearly we have thus split $\mathfrak{M}$ into "modular" and "archimedean" components. It is now a simple matter to prove the following

Theorem 6.7. *For $\frac{1}{2}N < n < N$, there exists a function $\mathfrak{S}(n)$ as in Theorem 4.10 so that*

$$\mathcal{M}_N(n) \gg \mathfrak{S}(n)T^{\delta-1}. \tag{6.8}$$

Proof. First we discuss the modular component. Write $\mathfrak{S}_{Q_0}$ as

$$\mathfrak{S}_{Q_0}(n) = \sum_{q<Q_0} \frac{1}{[\Gamma:\Gamma_0(q)]} \sum_{\gamma_0\in\Gamma/\Gamma_0(q)} c_q(\langle w_{x,y},\gamma_0 v_0\rangle - n),$$

where c_q is the Ramanujan sum, $c_q(m) = \sum'_{r(q)} e_q(rm)$. By (2.22), this reduces to a classical estimate for the singular series. We may use the transitivity of the γ_0 sum to replace $\langle w_{x,y},\gamma_0 v_0\rangle$ by $\langle e_4,\gamma_0 v_0\rangle$, extend the sum on q to all natural numbers, and use multiplicativity to write the sum as an Euler product. Then the resulting singular series

$$\mathfrak{S}(n) := \prod_p \left[1 + \sum_{k\geq 1} \frac{1}{[\Gamma:\Gamma_0(p^k)]} \sum_{\gamma_0\in\Gamma/\Gamma_0(p^k)} c_{p^k}\left(\langle e_4,\gamma_0\,v_0\rangle - n\right)\right]$$

vanishes only on non-admissible numbers, and can easily be seen to satisfy

$$N^{-\varepsilon} \ll_\varepsilon \mathfrak{S}(n) \ll_\varepsilon N^\varepsilon, \tag{6.9}$$

for any $\varepsilon > 0$. See, e.g. [BK10, §4.3].

Next we handle the archimedean component. By our choice of $\mathfrak{t}$ in (4.19), specifically that $\widehat{\mathfrak{t}} > 0$ and $\widehat{\mathfrak{t}}(y) > 2/5$ for $|y| < 1/2$, we have

$$\mathfrak{W}(n) \gg \frac{K_0}{N} \sum_{\mathfrak{f}\in\mathfrak{F}} \mathbf{1}_{\{|\mathfrak{f}(2x,y)-n|<\frac{N}{2K_0}\}} \gg \frac{T^\delta}{N} + \frac{T^{\Theta_1}K_0}{N},$$

using Lemma 5.26.

Putting everything into (6.6) and then into (6.4) gives (6.8), using (6.2) and (4.1). $\qquad\square$

Next we derive from the above that the same bound holds for $\mathcal{M}_N^U$ (most of the time).

Theorem 6.10. *There is an $\eta > 0$ such that the bound* (6.8) *holds with $\mathcal{M}_N$ replaced by $\mathcal{M}_N^U$, except on a set of cardinality $\ll N^{1-\eta}$.*

Proof. Putting (6.6) into (6.4) gives

$$\sum_{n<N} |\mathcal{M}_N(n) - \mathcal{M}_N^U(n)| \ll \sum_{x,y \asymp X} \sum_{n<N} |\mathfrak{M}(n)| \sum_{\substack{u|(2x,y) \\ u \geq U}} 1$$

$$\ll_\varepsilon \sum_{\substack{y<X \\ u|y \\ u \geq U}} \sum_{\substack{x<X \\ 2x \equiv 0 (\mathrm{mod}\, u)}} \left\{ N^\varepsilon \sum_{\mathfrak{f} \in \mathfrak{F}} \frac{K_0}{N} \left[\sum_{n<N} \widehat{\mathfrak{t}} \left((\mathfrak{f}(2x,y) - n) \frac{K_0}{N} \right) \right] + K_0^2 Q_0^4 T^{\Theta_1} \right\}$$

$$\ll \quad N^\varepsilon X \frac{X}{U} T^\delta,$$

using (6.9) and (6.2). The rest of the argument is identical to that leading to (4.12). $\square$

This establishes (4.15), and hence completes our Major Arcs analysis; the rest of the paper is devoted to proving (4.16).

7. Minor Arcs I: Case $q < Q_0$

We keep all the notation of §4, our goal in this section being to bound (4.24) and (4.25). First we return to (4.9) and reverse orders of summation, writing

$$\widehat{\mathcal{R}_N^U}(\theta) = \sum_{u < U} \mu(u) \sum_{\mathfrak{f} \in \mathfrak{F}} e(-a\theta) \widehat{\mathcal{R}}_{f,u}(\theta), \tag{7.1}$$

where $\mathfrak{f} = f - a$ according to (2.15), and we have set

$$\widehat{\mathcal{R}}_{f,u}(\theta) := \sum_{2x \equiv 0(u)} \sum_{y \equiv 0(u)} \Upsilon\left(\frac{2x}{X}\right) \Upsilon\left(\frac{y}{X}\right) e\left(\theta f(2x, y)\right).$$

We restrict our attention henceforth to the case when u is even; then we have

$$\widehat{\mathcal{R}}_{f,u}(\theta) = \sum_{x,y \in \mathbb{Z}} \Upsilon\left(\frac{xu}{X}\right) \Upsilon\left(\frac{yu}{X}\right) e\left(\theta f(xu, yu)\right), \tag{7.2}$$

and a similar expression holds for u odd. We first massage $\widehat{\mathcal{R}}_{f,u}$ further.

Since f is homogeneous quadratic, we have

$$f(xu, yu) = u^2 f(x, y).$$

Hence expressing $\theta = \frac{r}{q} + \beta$, we will need to write u^2/q as a reduced fraction; to this end, introduce the notation

$$\tilde{q} := (u^2, q) \qquad u_0 := u^2/\tilde{q}, \qquad q_0 := q/\tilde{q}, \tag{7.3}$$

so that $u^2/q = u_0/q_0$ in lowest terms, $(u_0, q_0) = 1$.

Lemma 7.4. *Recalling the notation* (5.29), *we have*

$$\widehat{\mathcal{R}}_{f,u}\left(\frac{r}{q} + \beta\right) = \frac{1}{u^2} \sum_{n,m \in \mathbb{Z}} \mathcal{J}_f\left(X, \beta; \frac{n}{uq_0}, \frac{m}{uq_0}\right) \mathcal{S}_f(q_0, ru_0; n, m), \tag{7.5}$$

where we have set

$$\mathcal{J}_f\left(X, \beta; \frac{n}{uq_0}, \frac{m}{uq_0}\right) := \iint_{x,y \in \mathbb{R}} \Upsilon\left(\frac{x}{X}\right) \Upsilon\left(\frac{y}{X}\right) e\left(\beta f(x, y) - \frac{n}{uq_0}x - \frac{m}{uq_0}y\right) dx dy. \tag{7.6}$$

Proof. Returning to (7.2), we have

$$\widehat{\mathcal{R}}_{f,u}\left(\frac{r}{q}+\beta\right) = \sum_{x,y\in\mathbb{Z}} \Upsilon\left(\frac{ux}{X}\right)\Upsilon\left(\frac{uy}{X}\right)e_{q_0}\left(ru_0 f(x,y)\right)e\left(\beta u^2 f(x,y)\right)$$

$$= \sum_{k(q_0)}\sum_{\ell(q_0)} e_{q_0}\left(ru_0 f(k,\ell)\right)$$

$$\times\left[\sum_{\substack{x\in\mathbb{Z}\\x\equiv k(q_0)}}\sum_{\substack{y\in\mathbb{Z}\\y\equiv\ell(q_0)}} \Upsilon\left(\frac{ux}{X}\right)\Upsilon\left(\frac{uy}{X}\right)e\left(\beta u^2 f(x,y)\right)\right].$$

Apply Poisson summation to the bracketed term above:

$$\left[\,\cdot\,\right] = \sum_{x,y\in\mathbb{Z}} \Upsilon\left(\frac{u(q_0 x+k)}{X}\right)\Upsilon\left(\frac{u(q_0 y+\ell)}{X}\right)e\left(\beta u^2 f(q_0 x+k, q_0 y+\ell)\right)$$

$$= \sum_{n,m\in\mathbb{Z}}\iint_{x,y\in\mathbb{R}} \Upsilon\left(\frac{u(q_0 x+k)}{X}\right)\Upsilon\left(\frac{u(q_0 y+\ell)}{X}\right)e\left(\beta u^2 f(q_0 x+k, q_0 y+\ell)\right)$$

$$\times e(-nx-my)\,dxdy$$

$$= \frac{1}{u^2 q_0^2}\sum_{n,m\in\mathbb{Z}} e_{q_0}(nk+m\ell)\mathcal{J}_f\left(X,\beta;\frac{n}{uq_0},\frac{m}{uq_0}\right),$$

Inserting this in the above, the claim follows immediately. □

We are now in position to prove the following

Proposition 7.7. *With the above notation,*

$$\left|\widehat{\mathcal{R}}_{f,u}\left(\frac{r}{q}+\beta\right)\right| \ll u(\sqrt{q}|\beta|T)^{-1}. \tag{7.8}$$

Proof. By (non)stationary phase, the integral in (7.6) has negligible contribution unless

$$\frac{|n|}{uq_0},\frac{|m|}{uq_0} \ll |\beta|\cdot|\nabla f| \ll |\beta|\cdot TX,$$

so the n,m sum can be restricted to

$$|n|,|m| \ll |\beta|\cdot TX\cdot uq_0 \ll u. \tag{7.9}$$

Here we used $|\beta| \ll (qM)^{-1}$ with M given by (4.17). In this range, stationary phase gives

$$\left|\mathcal{J}_f\left(X,\beta;\frac{n}{uq_0},\frac{m}{uq_0}\right)\right| \ll \min\left(X^2,\frac{1}{|\beta|\cdot|\operatorname{discr}(f)|^{1/2}}\right) \ll \min\left(X^2,\frac{1}{|\beta|T}\right),$$
$$\tag{7.10}$$

using (2.17) and (4.4) that $|\operatorname{discr}(f)| = 4|B^2 - AC| = 4a^2 \gg T^2$.

Putting (7.9), (7.10) and (5.31) into (7.5), we have

$$\left|\widehat{\mathcal{R}}_{f,u}\left(\frac{r}{q} + \beta\right)\right| \ll \frac{1}{u^2} \sum_{|n|,|m| \ll u} \frac{1}{|\beta|T} \cdot \frac{1}{\sqrt{q_0}},$$

from which the claim follows, using (7.3). $\qquad\square$

Finally, we prove the desired estimates of the strength (4.16).

Theorem 7.11. *Recall the integrals $\mathcal{I}_{Q_0,K_0}$, $\mathcal{I}_{Q_0}$ from (4.24), (4.25). There is an $\eta > 0$ so that*

$$\mathcal{I}_{Q_0,K_0},\ \mathcal{I}_{Q_0} \ll N\, T^{2(\delta-1)}\, N^{-\eta},$$

as $N \to \infty$.

Proof. We first handle $\mathcal{I}_{Q_0,K_0}$.

Returning to (7.1) and applying (7.8) gives

$$\left|\widehat{\mathcal{R}_N^U}\left(\frac{r}{q} + \beta\right)\right| \ll \sum_{u<U}\sum_{\mathfrak{f}\in\mathfrak{F}} u(\sqrt{q}|\beta|T)^{-1} \ll U^2 T^{\delta-1}(\sqrt{q}|\beta|)^{-1}.$$

Inserting this into (4.24) gives

$$\begin{aligned}
\mathcal{I}_{Q_0,K_0} &\ll \sum_{q<Q_0}\sideset{}{'}\sum_{r(q)} \int_{|\beta|<K_0/N} \left|\beta\frac{N}{K_0}\right|^2 U^4 T^{2(\delta-1)}\frac{1}{q|\beta|^2}d\beta \\
&\ll Q_0\frac{N}{K_0}U^4 T^{2(\delta-1)} \ll NT^{2(\delta-1)}N^{-\eta},
\end{aligned}$$

using (6.2), (6.3).

Next we handle

$$\begin{aligned}
\mathcal{I}_{Q_0} &\ll \sum_{q<Q_0}\sideset{}{'}\sum_{r(q)} \int_{\frac{K_0}{N}<|\beta|<\frac{1}{qM}} U^4 T^{2(\delta-1)}\frac{1}{q|\beta|^2}d\beta \\
&\ll Q_0 U^4 T^{2(\delta-1)}\left(\frac{N}{K_0} + Q_0 M\right) \\
&\ll NT^{2(\delta-1)}\frac{Q_0 U^4}{K_0},
\end{aligned}$$

which is again a power savings. $\qquad\square$

8. MINOR ARCS II: CASE $Q_0 \leq Q < X$

Keeping all the notation from the last section, we now turn our attention to the integrals $\mathcal{I}_Q$ in (4.26). It is no longer sufficient just to get cancellation in $\widehat{\mathcal{R}}_{f,u}$ alone, as in (7.8); we must use the fact that $\mathcal{I}_Q$ is an L^2-norm.

To this end, recall the notation (7.3), and put (7.5) into (7.1), applying Cauchy-Schwarz in the u-variable:

$$\left|\widehat{\mathcal{R}_N^U}\left(\frac{r}{q}+\beta\right)\right|^2 \ll U\sum_{u<U}\left|\sum_{\mathfrak{f}\in\mathfrak{F}}e_q(-ra)e(-a\beta)\right. \tag{8.1}$$

$$\left.\times\frac{1}{u^2}\sum_{n,m\in\mathbb{Z}}\mathcal{J}_f\left(X,\beta;\frac{n}{uq_0},\frac{m}{uq_0}\right)\mathcal{S}_f(q_0,ru_0;n,m)\right|^2.$$

Recall from (2.15) that, as throughout, $\mathfrak{f} = f - a$. Insert (8.1) into (4.26) and open the square, setting $\mathfrak{f}' = f' - a'$. This gives

$$\mathcal{I}_Q \ll U\sum_{u<U}\frac{1}{u^4}\sum_{q\asymp Q}\sideset{}{'}\sum_{r(q)}\int_{|\beta|<\frac{1}{qM}}\left|\sum_{\mathfrak{f}\in\mathfrak{F}}e_q(-ra)e(-a\beta)\right.$$

$$\left.\times\sum_{n,m\in\mathbb{Z}}\mathcal{J}_f\left(X,\beta;\frac{n}{uq_0},\frac{m}{uq_0}\right)\mathcal{S}_f(q_0,ru_0;n,m)\right|^2 d\beta$$

$$= U\sum_{u<U}\frac{1}{u^4}\sum_{n,m,n',m'\in\mathbb{Z}}\sum_{\mathfrak{f},\mathfrak{f}'\in\mathfrak{F}}\sum_{q\asymp Q} \tag{8.2}$$

$$\times\left[\sideset{}{'}\sum_{r(q)}\mathcal{S}_f(q_0,ru_0;n,m)\overline{\mathcal{S}_{f'}(q_0,ru_0;n',m')}e_q(r(a'-a))\right]$$

$$\times\left[\int_{|\beta|<\frac{1}{qM}}\mathcal{J}_f\left(X,\beta;\frac{n}{uq_0},\frac{m}{uq_0}\right)\overline{\mathcal{J}_{f'}\left(X,\beta;\frac{n'}{uq_0},\frac{m'}{uq_0}\right)}e(\beta(a'-a))d\beta\right].$$

Note that again the sum has split into "modular" and "archimedean" pieces (collected in brackets, respectively), with the former being exactly equal to $\mathcal{S}$ in (5.35).

Decompose (8.2) as

$$\mathcal{I}_Q \ll \mathcal{I}_Q^{(=)} + \mathcal{I}_Q^{(\neq)}, \tag{8.3}$$

where, once $\mathfrak{f}$ is fixed, we collect $\mathfrak{f}'$ according to whether $a' = a$ (the "diagonal" case) and the off-diagonal $a' \neq a$.

Lemma 8.4. *Assume $Q < X$. For $\square \in \{=, \neq\}$, we have*

$$\mathcal{I}_Q^{(\square)} \ll U^6 \frac{X^2}{T} \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{\substack{\mathfrak{f}'\in\mathfrak{F} \\ a'\square a}} \sum_{q\asymp Q} \frac{\{(a^2,q)\cdot((a')^2,q)\}^{1/2}(a-a',q)^{1/4}}{q^{5/4}}. \quad (8.5)$$

Proof. Apply (5.37) and (7.9), (7.10) to (8.2), giving

$$\begin{aligned}
\mathcal{I}_Q^{(\square)} &\ll U \sum_{u<U} \frac{1}{u^4} \sum_{|n|,|m|,|n'|,|m'|\ll u} \sum_{\substack{\mathfrak{f},\mathfrak{f}'\in\mathfrak{F} \\ a'\square a}} \sum_{q\asymp Q} \\
&\quad \times \frac{u^4\{(a^2,q)\cdot((a')^2,q)\}^{1/2}(a-a',q)^{1/4}}{q^{5/4}} \\
&\quad \times \int_{|\beta|<1/(qM)} \min\left(X^2, \frac{1}{|\beta|T}\right)^2 d\beta,
\end{aligned}$$

where we used (7.3). The claim then follows immediately from (4.17) and $Q < X$. $\qquad\square$

We treat $\mathcal{I}_Q^{(=)}$, $\mathcal{I}_Q^{(\neq)}$ separately, starting with the former; we give bounds of the quality claimed in (4.16).

Proposition 8.6. *There is an $\eta > 0$ such that*

$$\mathcal{I}_Q^{(=)} \ll N\, T^{2(\delta-1)} N^{-\eta}, \quad (8.7)$$

as $N \to \infty$.

Proof. From (8.5), we have

$$\begin{aligned}
\mathcal{I}_Q^{(=)} &\ll U^6 \frac{X^2}{T} \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{\substack{\mathfrak{f}'\in\mathfrak{F} \\ a'=a}} \sum_{q\asymp Q} \frac{(a^2,q)}{q} \\
&\ll \frac{U^6 X^2}{QT} \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{\substack{\tilde{q}_1|a^2 \\ \tilde{q}_1\ll Q}} \tilde{q}_1 \sum_{\substack{q\asymp Q \\ q\equiv 0(\tilde{q}_1)}} \sum_{\substack{\mathfrak{f}'\in\mathfrak{F} \\ a'=a}} 1 \\
&\ll_\varepsilon \frac{U^6 X^2}{T} \sum_{\mathfrak{f}\in\mathfrak{F}} T^\varepsilon \sum_{\substack{\mathfrak{f}'\in\mathfrak{F} \\ a'=a}} 1.
\end{aligned}$$

Recalling that $a = a_\gamma = \langle e_1, \gamma v_0\rangle$, replace the condition $a' = a$ with $a' \equiv a(\mathrm{mod}\lfloor Q_0\rfloor)$, and apply (5.14):

$$\mathcal{I}_Q^{(=)} \ll_\varepsilon \frac{U^6 X^2}{T} T^\delta T^\varepsilon \frac{T^\delta}{Q_0^{\eta_0}}.$$

Then (6.3) and (4.1) imply the claimed power savings. $\qquad\square$

Next we turn our attention to $\mathcal{I}_Q^{(\neq)}$, the off-diagonal contribution. We decompose this sum further according to whether $\gcd(a, a')$ is large or not. To this end, introduce a parameter H, which we will eventually set to

$$H = U^{10/\eta_0} = Q_0{}^{\eta_0/10}, \tag{8.8}$$

where, as in (6.3), the constant $\eta_0 > 0$ comes from Lemma 5.12. Write

$$\mathcal{I}_Q^{(\neq)} = \mathcal{I}_Q^{(\neq,>)} + \mathcal{I}_Q^{(\neq,\leq)}, \tag{8.9}$$

corresponding to whether $(a, a') > H$ or $(a, a') \leq H$, respectively. We deal first with the large gcd.

Proposition 8.10. *There is an $\eta > 0$ such that*

$$\mathcal{I}_Q^{(\neq,>)} \ll N\,T^{2(\delta-1)}N^{-\eta}, \tag{8.11}$$

as $N \to \infty$.

Proof. Writing $(a, a') = h > H$, $\tilde{q}_1 = (a^2, q)$, $\tilde{q}_1' = ((a')^2, q)$, and using $(a - a', q) \leq q$ in (8.5), we have

$$
\begin{aligned}
\mathcal{I}_Q^{(\neq,>)} \;\ll\;& U^6 \frac{X^2}{T} \sum_{\mathfrak{f} \in \mathfrak{F}} \sum_{\substack{\mathfrak{f}' \in \mathfrak{F} \\ a' \neq a, (a,a') > H}} \sum_{q \asymp Q} \frac{\{(a^2, q) \cdot ((a')^2, q)\}^{1/2}(a - a', q)^{1/4}}{q^{5/4}} \\
\ll\;& U^6 \frac{X^2}{T} \sum_{\mathfrak{f} \in \mathfrak{F}} \sum_{\substack{h \mid a \\ h > H}} \sum_{\substack{\mathfrak{f}' \in \mathfrak{F} \\ a' \equiv 0 (\bmod\, h)}} \sum_{\substack{\tilde{q}_1 \mid a^2 \\ \tilde{q}_1 \ll Q}} \sum_{\substack{\tilde{q}_1' \mid (a')^2 \\ [\tilde{q}_1, \tilde{q}_1'] \ll Q}} (\tilde{q}_1 \tilde{q}_1')^{1/2} \sum_{\substack{q \asymp Q \\ q \equiv 0([\tilde{q}_1, \tilde{q}_1'])}} \frac{1}{Q} \\
\ll_\varepsilon\;& U^6 \frac{X^2}{T} T^\varepsilon \sum_{\mathfrak{f} \in \mathfrak{F}} \sum_{\substack{h \mid a \\ h > H}} \sum_{\substack{\mathfrak{f}' \in \mathfrak{F} \\ a' \equiv 0 (\bmod\, h)}} 1,
\end{aligned}
$$

where we used $[n, m] > (nm)^{1/2}$. Apply (5.14) to the innermost sum, getting

$$\mathcal{I}_Q^{(\neq,>)} \;\ll_\varepsilon\; U^6 \frac{X^2}{T} T^\varepsilon T^\delta \frac{1}{H^{\eta_0}} T^\delta.$$

By (8.8) and (6.3), this is a power savings, as claimed. $\qquad\square$

Finally, we handle small gcd.

Proposition 8.12. *There is an $\eta > 0$ such that*

$$\mathcal{I}_Q^{(\neq,\leq)} \ll N\,T^{2(\delta-1)}N^{-\eta}, \tag{8.13}$$

as $N \to \infty$.

Proof. First note that

$$
\begin{aligned}
\mathcal{I}_Q^{(\neq,\leq)} &= U^6 \frac{X^2}{T} \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{\substack{\mathfrak{f}'\in\mathfrak{F}\\ a'\neq a,(a,a')\leq H}} \sum_{q\asymp Q} \frac{\{(a^2,q)\cdot((a')^2,q)\}^{1/2}(a-a',q)^{1/4}}{q^{5/4}} \\
&\ll U^6 \frac{X^2}{T} \frac{1}{Q^{5/4}} \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{\substack{\mathfrak{f}'\in\mathfrak{F}\\ a'\neq a,(a,a')\leq H}} \sum_{q\asymp Q} (a,q)(a',q)(a-a',q)^{1/4}.
\end{aligned}
$$

Write $g = (a,q)$ and $g' = (a',q)$, and let $h = (g,g')$; observe then that $h \mid (a,a')$ and $h \ll Q$. Hence we can write $g = hg_1$ and $g' = hg_1'$ so that $(g_1,g_1') = 1$. Note also that $h \mid (a-a',q)$, so we can write $(a-a',q) = h\tilde{g}$; thus g_1, g_1', and $\tilde{g}$ are pairwise coprime, implying

$$
[hg_1, hg_1', h\tilde{g}] \geq g_1 g_1' \tilde{g}.
$$

Then we have

$$
\begin{aligned}
\mathcal{I}_Q^{(\neq,\leq)} &\ll U^6 \frac{X^2}{T} \frac{1}{Q^{5/4}} \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{\substack{\mathfrak{f}'\in\mathfrak{F}\\ a'\neq a,(a,a')\leq H}} \sum_{\substack{h\mid(a,a')\\ h\leq H}} \sum_{\substack{g_1\mid a\\ g_1\ll Q}} \sum_{\substack{g_1'\mid a'\\ g_1'\ll Q}} \\
&\quad\times \sum_{\substack{\tilde{g}\mid(a-a')\\ [hg_1,hg_1',h\tilde{g}]\ll Q}} (hg_1)(hg_1')(h\tilde{g})^{1/4} \sum_{\substack{q\asymp Q\\ q\equiv 0([hg_1,hg_1',h\tilde{g}])}} 1 \\
&\ll_\varepsilon U^6 \frac{X^2}{T} \frac{H^{9/4}}{Q^{5/4}} \sum_{\mathfrak{f},\mathfrak{f}'\in\mathfrak{F}} T^\varepsilon \sum_{\substack{g_1\mid a\\ g_1\ll Q}} \sum_{\substack{g_1'\mid a'\\ g_1'\ll Q}} \sum_{\substack{\tilde{g}\mid(a-a')\\ \tilde{g}\ll Q}} g_1\, g_1'\, \tilde{g}^{1/4} \frac{Q}{g_1 g_1' \tilde{g}} \\
&\ll U^6 \frac{X^2}{T} \frac{H^{9/4}}{Q^{1/4}} \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{\tilde{g}\ll Q} \frac{1}{\tilde{g}^{3/4}} T^\varepsilon \sum_{\substack{\mathfrak{f}'\in\mathfrak{F}\\ a'\equiv a(\mathrm{mod}\,\tilde{g})}} 1.
\end{aligned}
$$

To the last sum, we again apply Lemma 5.12, giving

$$
\mathcal{I}_Q^{(\neq,\leq)} \ll_\varepsilon U^6 \frac{X^2}{T} \frac{H^{9/4}}{Q^{1/4}} T^\delta \sum_{\tilde{g}\ll Q} \frac{1}{\tilde{g}^{3/4}} T^\varepsilon \frac{1}{\tilde{g}^{\eta_0}} T^\delta \ll U^6 \frac{X^2}{T} \frac{H^{9/4}}{Q_0^{\eta_0}} T^\delta T^\varepsilon T^\delta,
$$

since $Q \geq Q_0$. By (8.8) and (6.3), this is again a power savings, as claimed. $\square$

Putting together (8.3), (8.7), (8.9), (8.11), and (8.13), we have proved the following

Theorem 8.14. *For $Q_0 \leq Q < X$, there is some $\eta > 0$ such that*

$$
\mathcal{I}_Q \ll N\, T^{2(\delta-1)}\, N^{-\eta},
$$

as $N \to \infty$.

9. Minor Arcs III: Case $X \leq Q < M$

In this section, we continue our analysis of $\mathcal{I}_Q$ from (4.26), but now we need different methods to handle the very large Q situation. In particular, the range of x, y in (7.2) is now such that we have incomplete sums, so our first step is to complete them.

To this end, recall the notation (7.3) and introduce

$$\lambda_f\left(X, \beta; \frac{n}{q_0}, \frac{m}{q_0}, u\right) := \sum_{x,y \in \mathbb{Z}} \Upsilon\left(\frac{ux}{X}\right) \Upsilon\left(\frac{uy}{X}\right) e\left(-\frac{n}{q_0}x - \frac{m}{q_0}y\right) e\left(\beta u^2 f(x,y)\right),$$
(9.1)

so that, using (5.29), an elementary calculation gives

$$\widehat{\mathcal{R}}_{f,u}\left(\frac{r}{q} + \beta\right) = \sum_{n(q_0)} \sum_{m(q_0)} \lambda_f\left(X, \beta; \frac{n}{q_0}, \frac{m}{q_0}, u\right) \mathcal{S}_f(q_0, ru_0; n, m). \quad (9.2)$$

Put (9.2) into (7.1) and apply Cauchy-Schwarz in the u-variable:

$$\left|\widehat{\mathcal{R}}_N^U\left(\frac{r}{q} + \beta\right)\right|^2 \ll U \sum_{u<U} \left| \sum_{\mathfrak{f} \in \mathfrak{F}} e_q(-ra)e(-a\beta) \right. \tag{9.3}$$

$$\times \left. \sum_{0 \leq n,m < q_0} \lambda_f\left(X, \beta; \frac{n}{q_0}, \frac{m}{q_0}, u\right) \mathcal{S}_f(q_0, ru_0; n, m) \right|^2.$$

As before, open the square, setting $\mathfrak{f}' = f' - a'$, and insert the result into (4.26):

$$\mathcal{I}_Q \ll U \sum_{u<U} \sum_{q \asymp Q} \sum_{n,m,n',m'<q_0} \sum_{\mathfrak{f},\mathfrak{f}' \in \mathfrak{F}} \tag{9.4}$$

$$\times \left[\sum_{r(q)}' \mathcal{S}_f(q_0, ru_0; n, m) \overline{\mathcal{S}_{f'}(q_0, ru_0; n', m')} e_q(r(a' - a)) \right]$$

$$\times \left[\int_{|\beta|<1/(qM)} \lambda_f\left(X, \beta; \frac{n}{q_0}, \frac{m}{q_0}, u\right) \overline{\lambda_{f'}\left(X, \beta; \frac{n'}{q_0}, \frac{m'}{q_0}, u\right)} e(\beta(a - a'))d\beta \right].$$

Yet again the sum has split into modular and archimedean components with the former being exactly equal to $\mathcal{S}$ in (5.35). As before, decompose $\mathcal{I}_Q$ according to the diagonal $(a = a')$ and off-diagonal terms:

$$\mathcal{I}_Q \ll \mathcal{I}_Q^{(=)} + \mathcal{I}_Q^{(\neq)}. \tag{9.5}$$

Lemma 9.6. *Assume $Q \geq X$. For $\square \in \{=, \neq\}$, we have*

$$\mathcal{I}_Q^{(\square)} \ll \frac{UX^3}{QT} \sum_{u<U} \frac{1}{u^4} \sum_{q \asymp Q} \sum_{n,m,n',m' \ll \frac{UQ}{X}} \sum_{\mathfrak{f} \in \mathfrak{F}} \sum_{\substack{\mathfrak{f}' \in \mathfrak{F} \\ a' \square a}} |\mathcal{S}|. \qquad (9.7)$$

Proof. Consider the sum λ_f in (9.1). Since $x, y \asymp X/u$, $|\beta| < 1/(qM)$, $X \leq Q$, and using (4.17), we have that

$$|\beta u^2 f(x,y)| \ll \frac{1}{QM} u^2 T \left(\frac{X}{u}\right)^2 = \frac{X}{Q} \leq 1.$$

Hence there is contribution only if $nx/q_0, my/q_0 \ll 1$, that is, we may restrict to the range

$$n, m \ll uq_0/X.$$

In this range, we give λ_f the trivial bound of X^2/u^2. Putting this analysis into (9.4), the claim follows. $\qquad \square$

We handle the off-diagonal term first.

Proposition 9.8. *Assuming $X \leq Q < M$, there is some $\eta > 0$ such that*

$$\mathcal{I}_Q^{(\neq)} \ll N\, T^{2(\delta-1)} N^{-\eta}, \qquad (9.9)$$

as $N \to \infty$.

Proof. Since (5.37) is such a large savings in $q > X$, we can afford to lose in the much smaller variable T. Hence put (5.37) into (9.7), estimating $(a - a', q) \leq |a - a'|$ (since $a \neq a'$):

$$\begin{aligned}
\mathcal{I}_Q^{(\neq)} &\ll \frac{UX^3}{QT} \sum_{u<U} \frac{1}{u^4} \sum_{q \asymp Q} \sum_{n,m,n',m' \ll \frac{UQ}{X}} \sum_{\mathfrak{f},\mathfrak{f}' \in \mathfrak{F}} u^4 \frac{a \cdot a'}{q^{5/4}} |a - a'|^{1/4} \\
&\ll \frac{U^6 X^3}{T} \left(\frac{Q}{X}\right)^4 T^{2\delta} \frac{T^2}{Q^{5/4}} T^{1/4} \\
&\ll U^6 X^{7/4} T^{2\delta} T^4 = X^2 T\, T^{2(\delta-1)} \left(U^6 X^{-1/4} T^5\right),
\end{aligned}$$

where we used (7.3), $Q < M$, and (4.17). Using (4.1) we have that

$$X^{-1/4} T^5 = N^{-59/800}, \qquad (9.10)$$

so together with (6.3), this is clearly a substantial power savings. $\quad \square$

Lastly, we deal with the diagonal term. We no longer save enough from $a = a'$ alone. But recall that here more cancellation can be gotten from (5.40) in the special case that $\mathfrak{f}(m, -n) \neq \mathfrak{f}'(m', -n')$. Hence we return to (9.7) and, once n, m, and $\mathfrak{f}$ are determined, separate n', m',

and $\mathfrak{f}'$ into cases corresponding to whether $\mathfrak{f}(m, -n) = \mathfrak{f}'(m', -n')$ or not. Accordingly, write

$$\mathcal{I}_Q^{(=)} \ll \mathcal{I}_Q^{(=,=)} + \mathcal{I}_Q^{(=,\neq)}. \tag{9.11}$$

We now estimate $\mathcal{I}_Q^{(=,\neq)}$ using the extra cancellation in (5.40).

Proposition 9.12. *Assuming $Q < XT$, there is some $\eta > 0$ such that*

$$\mathcal{I}_Q^{(=,\neq)} \ll N\, T^{2(\delta-1)}\, N^{-\eta}, \tag{9.13}$$

as $N \to \infty$.

Proof. Returning to (9.7), apply (5.40):

$$
\mathcal{I}_Q^{(=,\neq)} \;\ll\; \frac{UX^3}{QT} \sum_{u<U} \frac{1}{u^4} \sum_{\mathfrak{f}\in\mathfrak{F}} \sum_{\substack{\mathfrak{f}'\in\mathfrak{F}\\ a'=a}} \sum_{q\asymp Q} \sum_{n,m\ll\frac{UQ}{X}} \sum_{\substack{n',m'\ll\frac{UQ}{X}\\ \mathfrak{f}(m,-n)\neq\mathfrak{f}'(m',-n')}} |\mathcal{S}|
$$

$$
\ll\; \frac{UX^3}{QT} \sum_{u<U} \frac{1}{u^4} \sum_{\mathfrak{f},\mathfrak{f}'\in\mathfrak{F}} \sum_{\substack{\tilde{q}_1|a^2\\ \tilde{q}_1\ll Q}} \sum_{\substack{q\asymp Q\\ q\equiv 0(\tilde{q}_1)}} \sum_{n,m,n',m'\ll\frac{UQ}{X}} u^{10}\frac{\tilde{q}_1}{Q^{9/8}} \left(T\left(\frac{UQ}{X}\right)^2\right)^{1/2}
$$

$$
\ll_\varepsilon\; \frac{U^8 X^3}{T} T^{2\delta}\, T^\varepsilon \left(\frac{UQ}{X}\right)^4 \frac{1}{Q^{9/8}} T^{1/2}\frac{UQ}{X}
$$

$$
\ll\; X^2 T\, T^{2(\delta-1)} \left(X^{-1/8}T^{35/8}U^{13}T^\varepsilon\right),
$$

where we used that $\mathfrak{f}(m,n) \ll T(UQ/X)^2$ and $Q < XT$. From (4.1), we have

$$X^{-1/8}T^{35/8} = N^{-29/1600}, \tag{9.14}$$

so we have again a power savings, as claimed. $\qquad\square$

Lastly, we turn to the case $\mathcal{I}_Q^{(=,=)}$, with $\mathfrak{f}(m, -n) = \mathfrak{f}'(m', -n')$. We exploit this condition to get savings as follows.

Lemma 9.15. *For $Q < XT$, and n, m, and $\mathfrak{f}$ fixed as above,*

$$\#\left\{ \begin{array}{c} \mathfrak{f}' \in \mathfrak{F} \\ n', m' \in \mathbb{Z} \end{array} \,\middle|\, \begin{array}{c} a' = a \\ \mathfrak{f}(m, -n) = \mathfrak{f}'(m', -n') \end{array} \right\} \ll_\varepsilon T^\varepsilon, \tag{9.16}$$

for any $\varepsilon > 0$.

Proof. From (2.14), (4.3), and (2.17), we have that $\mathfrak{f} = f - a$ where $f(x, y) = Ax^2 + 2Bxy + Cy^2$ has coefficients of size

$$A, B, C \ll T,$$

and $AC - B^2 = a^2$, with $a \asymp T$. It follows that $AC \asymp T^2$, and hence

$$A, C \asymp T. \tag{9.17}$$

The same holds for $\mathfrak{f}' = f' - a$.

Set $z := \mathfrak{f}(m, -n)$ and observe that

$$z \ll T \left(\frac{u q_0}{X} \right)^2 \ll u^2 T^3,$$

using $q_0 \ll Q < XT$. Now $z+a$ is represented by at most $2^{\omega(z+a)} \ll_\varepsilon T^\varepsilon$ classes of binary quadratic forms with discriminant $-4a^2$. We claim that within each class, the number of equivalent forms f' is bounded.

Indeed if $\mathfrak{f}'$ and $\mathfrak{f}''$ have $a = a' = a''$ and the corresponding quadratic forms $A'x^2 + 2B'xy + C'y^2$ and $A''x^2 + 2B''xy + C''y^2$ are equivalent, then there is an element $\begin{pmatrix} g & h \\ i & j \end{pmatrix} \in \mathrm{GL}(2, \mathbb{Z})$ so that

$$\begin{aligned} A'' &= g^2 A' + 2gi B' + i^2 C', \\ B'' &= gh A' + (gj + hi) B' + ij C', \\ C'' &= h^2 A' + 2hj B' + j^2 C'. \end{aligned}$$

The first line can be rewritten as

$$A'' = C'(i + gB'/C')^2 + g^2 \frac{4a^2}{C'},$$

so

$$g^2 \leq A'' \frac{C'}{4a^2} \ll 1.$$

Similarly,

$$(i + gB'/C')^2 \leq \frac{A''}{C'} \ll 1,$$

and so $|i| \ll 1$. In a similar fashion, we see that $|h|$ and $|j|$ are also bounded, thus the number of equivalent forms is bounded, as claimed.

Hence $z + a$ is represented by at most T^ε classes, and the number of equivalent forms in each class is $O(1)$. For each fixed form f', the equation $f'(m', -n') = z + a$ has $\ll_\varepsilon T^\varepsilon$ solutions in n', m', as claimed. $\qquad\square$

We can finally estimate $\mathcal{I}_Q^{(=,=)}$.

Proposition 9.18. *Assuming $Q < XT$, there is some $\eta > 0$ such that*

$$\mathcal{I}_Q^{(=,=)} \ll N\, T^{2(\delta-1)}\, N^{-\eta}, \tag{9.19}$$

as $N \to \infty$.

Proof. Returning to (9.7), apply (5.37), and (9.16):

$$
\mathcal{I}_Q^{(=,=)} \quad \ll \quad \frac{UX^3}{QT} \sum_{u<U} \frac{1}{u^4} \sum_{q \asymp Q} \sum_{n,m \ll \frac{UQ}{X}} \sum_{\mathfrak{f} \in \mathfrak{F}} \sum_{\substack{\mathfrak{f}' \in \mathfrak{F} \\ a'=a}} \sum_{\substack{n',m' \\ \mathfrak{f}(m,-n)=\mathfrak{f}'(m',-n')}} u^4 \frac{(a^2,q)}{q^{5/4}} q^{1/4}
$$

$$
\ll \quad \frac{UX^3}{Q^2T} \sum_{u<U} \sum_{\mathfrak{f} \in \mathfrak{F}} \sum_{\substack{\tilde{q}_1 | a^2 \\ \tilde{q}_1 \ll Q}} \tilde{q}_1 \sum_{\substack{q \asymp Q \\ q \equiv 0(\tilde{q}_1)}} \sum_{n,m \ll \frac{UQ}{X}} \left[\sum_{\substack{\mathfrak{f}' \in \mathfrak{F} \\ a'=a}} \sum_{\substack{n',m' \\ \mathfrak{f}(m,-n)=\mathfrak{f}'(m',-n')}} 1 \right]
$$

$$
\ll_\varepsilon \quad \frac{U^4 X^3}{Q^2 T} T^\delta Q \frac{Q^2}{X^2} T^\varepsilon = X^2 T\, T^{2(\delta-1)} \left(T^{1-\delta} U^4 T^\varepsilon \right).
$$

From (3.4), this is a power savings. $\qquad\square$

Combining (9.5), (9.9), (9.11), (9.13), and (9.19), we have the following

Theorem 9.20. *If $X \le Q < M$, then there is some $\eta > 0$ so that*
$$
\mathcal{I}_Q \ll N\, T^{2(\delta-1)}\, N^{-\eta},
$$
as $N \to \infty$.

Finally, Theorems 7.11, 8.14, and 9.20 together complete the proof of (4.16), and hence Theorem 1.7 is proved.

References

[BF10]   J. Bourgain and E. Fuchs. A proof of the positive density conjecture for integer Apollonian circle packings, 2010. Preprint, `arXiv:1001.3894`. 3

[BG09]   Jean Bourgain and Alex Gamburd. Expansion and random walks in $SL_d(\mathbb{Z}/p^n\mathbb{Z})$. II. *J. Eur. Math. Soc. (JEMS)*, 11(5):1057–1103, 2009. With an appendix by Bourgain. 25

[BGS09]   Jean Bourgain, Alex Gamburd, and Peter Sarnak. Generalization of Selberg's theorem and Selberg's sieve, 2009. Preprint. 4, 13

[BK10]   J. Bourgain and A. Kontorovich. On representations of integers in thin subgroups of $SL(2, \mathbf{Z})$. *GAFA*, 20(5):1144–1174, 2010. 16, 36

[BK11]   J. Bourgain and A. Kontorovich. On Zaremba's conjecture, 2011. Preprint, `arXiv:1107.3776`. 16

[BKS10]   J. Bourgain, A. Kontorovich, and P. Sarnak. Sector estimates for hyperbolic isometries. *GAFA*, 20(5):1175–1200, 2010. 15, 23, 24, 27, 28, 29

[Bou11]   J. Bourgain. Integral Apollonian circle packings and prime curvatures, 2011. Preprint. `arXiv:1105.5127v1`. 4

[BR02]   J. Bernstein and A. Reznikov. Sobolev norms of automorphic functionals. *IMRN*, 40:2155–2174, 2002. 14

[BV11]   J. Bourgain and P. Varju. Expansion in $SL_2(\mathbf{Z}/q\mathbf{Z})$, q arbitrary, 2011. To appear, *Invent. Math.* `arXiv:1006.3365v1`. 4, 13

[BY91]   Carlos A. Berenstein and Alain Yger. Effective Bezout identities in $\mathbf{Q}[z_1, \cdots, z_n]$. *Acta Math.*, 166(1-2):69–120, 1991. 26

[CHH88]   M. Cowling, U. Haagerup, and R. Howe. Almost L^2 matrix coefficients. *J. Reine Angew. Math.*, 387:97–110, 1988. 14

[FS10]   E. Fuchs and K. Sanden. Some experiments with integral apollonian circle packings, 2010. To appear, Exp. Math. 2

[Fuc10]   E. Fuchs. Arithmetic properties of Apollonian circle packings, 2010. Princeton University Thesis. 3, 9, 10

[GGPS66]   I. M. Gelfand, M. I. Graev, and I. I. Pjateckii-Shapiro. *Teoriya predstavlenii i avtomorfnye funktsii*. Generalized functions, No. 6. Izdat. "Nauka", Moscow, 1966. 13

[GLM+03]   Ronald L. Graham, Jeffrey C. Lagarias, Colin L. Mallows, Allan R. Wilks, and Catherine H. Yan. Apollonian circle packings: number theory. *J. Number Theory*, 100(1):1–45, 2003. 2, 3, 6, 10

[GLM+05]   Ronald L. Graham, Jeffrey C. Lagarias, Colin L. Mallows, Allan R. Wilks, and Catherine H. Yan. Apollonian circle packings: geometry and group theory. I. The Apollonian group. *Discrete Comput. Geom.*, 34(4):547–585, 2005. 10

[Goo83]   Anton Good. *Local analysis of Selberg's trace formula*, volume 1040 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1983. 15

[Her26]   Grete Hermann. Die Frage der endlich vielen Schritte in der Theorie der Polynomideale. *Math. Ann.*, 95(1):736–788, 1926. 26

[IK04]   Henryk Iwaniec and Emmanuel Kowalski. *Analytic number theory*, volume 53 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2004. 9, 31

[KO11]   Alex Kontorovich and Hee Oh. Apollonian circle packings and closed
          horospheres on hyperbolic 3-manifolds. *Journal of the American Math-
          ematical Society*, 24(3):603–648, 2011. 2
[LMW02]   Jeffrey C. Lagarias, Colin L. Mallows, and Allan R. Wilks. Beyond the
          Descartes circle theorem. *Amer. Math. Monthly*, 109(4):338–361, 2002.
          2
[LP82]   P.D. Lax and R.S. Phillips. The asymptotic distribution of lattice points
          in Euclidean and non-Euclidean space. *Journal of Functional Analysis*,
          46:280–350, 1982. 12, 16, 23, 25
[MVW84]   C. Matthews, L. Vaserstein, and B. Weisfeiler. Congruence properties
          of Zariski-dense subgroups. *Proc. London Math. Soc*, 48:514–532, 1984.
          9
[MW83]   D. W. Masser and G. Wüstholz. Fields of large transcendence degree
          generated by values of elliptic functions. *Invent. Math.*, 72(3):407–464,
          1983. 26
[Pat76]   S.J. Patterson. The limit set of a Fuchsian group. *Acta Mathematica*,
          136:241–273, 1976. 12
[Sar07]   P. Sarnak. Letter to J. Lagarias, 2007. Online at
          http://www.math.princeton.edu/sarnak. 3, 6, 9
[Sod37]   F. Soddy. The bowl of integers and the hexlet. *Nature*, 139:77–79, 1937.
          2
[Sul84]   D. Sullivan. Entropy, Hausdorff measures old and new, and limit sets
          of geometrically finite Kleinian groups. *Acta Math.*, 153(3-4):259–277,
          1984. 12
[Var10]   P. Varju. Expansion in $SL_d(O_K/I)$, I square-free, 2010.
          arXiv:1001.3664v1. 4, 13
[Vin12]   I. Vinogradov. Effective bisector estimate with application to
          Apollonian circle packings, 2012. Princeton University Thesis,
          arxiv:1204.5498v1. 4, 14, 15

E-mail address: bourgain@ias.edu

IAS, PRINCETON, NJ

E-mail address: alex.kontorovich@yale.edu

YALE UNIVERSITY, NEW HAVEN, CT