ISRAEL JOURNAL OF MATHEMATICS **209** (2015), 397–420
DOI: 10.1007/s11856-015-1223-3

ALMOST PRIME COORDINATES FOR ANISOTROPIC AND THIN PYTHAGOREAN ORBITS

BY

Jiuzu Hong

Department of Mathematics, Yale University, New Haven, CT 06520 USA
e-mail: jiuzu.hong@yale.edu

AND

Alex Kontorovich*

Department of Mathematics, Yale University, New Haven, CT 06520 USA
and
School of Mathematics, IAS, Princeton, NJ 08540 USA
e-mail: alex.kontorovich@yale.edu

ABSTRACT

We make an observation which doubles the exponent of distribution in certain Affine Sieve problems, such as those considered by Liu–Sarnak, Kontorovich, and Kontorovich–Oh. As a consequence, we decrease the known bounds on the saturation numbers in these problems.

Contents

* A.K. is partially supported by an NSF CAREER grant DMS-1254788, an Alfred P. Sloan Research Fellowship, a Yale Junior Faculty Fellowship, and support at IAS from The Fund for Math and The Simonyi Fund.
Received January 26, 2014

1. Introduction

The purpose of this paper is to make a simple observation about the execution of the Affine Sieve, which has the effect of doubling the exponent of distribution in many natural sieve problems. For concreteness, we will illustrate the method on problems studied by Liu–Sarnak [LS10], Kontorovich [Kon07, Kon09], and Kontorovich–Oh [KO12]. We first briefly recall the general setup, then specialize to these particular problems, explain what is known, and finally describe our results.

1.1. The General Affine Sieve. Roughly speaking, the Affine Sieve inputs a pair $(\mathcal{O}, f)$ consisting of (i) an integer orbit $\mathcal{O} \subset \mathbb{Z}^n$ by a linear group and (ii) a polynomial function f which is integral on $\mathcal{O}$, and outputs a number $R \leq \infty$ so that there are "many" points $\mathbf{x} \in \mathcal{O}$ with $f(\mathbf{x})$ having at most R prime divisors. Let us make this precise.

Let $\Gamma < \mathrm{GL}_n(\mathbb{Z})$ be a finitely-generated group of invertible $n \times n$ integer matrices, let $\mathbb{G} := \mathrm{Zcl}(\Gamma)$ be its Zariski closure, and denote the real points of $\mathbb{G}$ by $G := \mathbb{G}(\mathbb{R})$. When the Haar measure of $\Gamma\backslash G$ is infinite, we refer to Γ as **thin**. For a fixed primitive vector $\mathbf{y} \in \mathbb{Z}^n$, we consider the orbit

$$\mathcal{O} := \mathbf{y} \cdot \Gamma \subset \mathbb{Z}^n; \tag{1.1}$$

we refer to $\mathcal{O}$ as thin when Γ is. Given a polynomial f in n variables which is integral on $\mathcal{O}$, we say that the pair $(\mathcal{O}, f)$ is **strongly primitive**[1] if, for all integers $q \geq 1$, there is an $\mathbf{x} \in \mathcal{O}$ so that $f(\mathbf{x}) \in (\mathbb{Z}/q\mathbb{Z})^\times$. We assume henceforth that this is the case.

For an integer $R \geq 1$, let

$$\mathcal{P}_R \subset \mathbb{Z}$$

denote the set of R-almost primes, that is, numbers with at most R prime divisors. We allow $R = \infty$, in which case $\mathcal{P}_R = \mathbb{Z}$. Let

$$\mathcal{O}(f, R) := \{\mathbf{x} \in \mathcal{O} : f(\mathbf{x}) \in \mathcal{P}_R\}.$$

The goal of the Affine Sieve is to study the so-called **saturation number**, given by

$$R_0(\mathcal{O}, f) := \min\{R \leq \infty : \mathrm{Zcl}(\mathcal{O}(f, R)) = \mathrm{Zcl}(\mathcal{O})\}.$$

[1] This corrects the typo in [KO12, Definition 1.3].

That is, R_0 is the minimal R for which $\mathcal{O}(f, R)$ is Zariski dense in the Zariski closure of $\mathcal{O}$. (Here $\mathrm{Zcl}(\cdot)$ refers to the Zariski closure in affine space $\mathbb{A}^n_{\mathbb{Q}}$.)

The program initiated by Bourgain–Gamburd–Sarnak [BGS06, BGS10] and completed by Salehi Golsefidy–Sarnak [SGS11] shows in essentially the greatest generality possible that pairs $(\mathcal{O}, f)$ are **factor finite**, meaning that we have the strict inequality $R_0(\mathcal{O}, f) < \infty$. Beyond factor-finiteness, one would like to actually determine the saturation number of any given pair $(\mathcal{O}, f)$. As stated, this problem is completely hopeless, as it includes all classical sieve problems (see the discussion in, e.g., [BGS10]). Nevertheless, there is an ongoing program of determining, or at least giving strong estimates for, the saturation number in certain specific cases, where more structure can be exploited. We give a few natural examples below.

1.2. Thin Pythagorean Orbits. Let $G = \mathrm{SO}_{\mathcal{Q}}(\mathbb{R}) < \mathrm{SL}_3(\mathbb{R})$ be the real special orthogonal group preserving the "Pythagorean" quadratic form

$$\mathcal{Q}(\mathbf{x}) := x^2 + y^2 - z^2, \tag{1.2}$$

where $\mathbf{x} = (x, y, z)$. We define a Pythagorean triple to be a primitive integer vector on the cone $\mathcal{Q} = 0$.

Let $\Gamma < G(\mathbb{Z})$ be a finitely generated subgroup of the integer matrices in G, and assume Γ is non-elementary, or equivalently, that its Zariski closure is $\mathrm{SO}_{\mathcal{Q}}$. For a fixed Pythagorean triple, e.g., $\mathbf{y} = (3, 4, 5)$, let $\mathcal{O}$ be its corresponding Γ-orbit, as in (1.1). We allow Γ, and hence $\mathcal{O}$, to be thin, in which case we refer to $\mathcal{O}$ as a thin Pythagorean orbit.

A measure of this thinness is the critical exponent

$$\delta = \delta_\Gamma \in [0, 1]$$

of Γ; this is the abscissa of convergence of the Poincaré series of Γ, or equivalently, the Hausdorff dimension of the limit set of Γ. Since Γ is non-elementary, δ is strictly positive; moreover, Γ is thin if and only if $\delta < 1$. The role played by this geometric invariant is illustrated by the easy fact that

$$\#\{\mathbf{x} \in \mathcal{O} : \|\mathbf{x}\| < T\} = T^{\delta + o(1)}, \tag{1.3}$$

as $T \to \infty$, where $\|\cdot\|$ is the standard Euclidean norm.

For various choices of the polynomial f, one can consider the problem of estimating the saturation number $R_0(\mathcal{O}, f)$. Three natural choices for f considered

in [Kon07, Kon09, KO12] are

$$\begin{cases} f_{\mathscr{H}}(\mathbf{x}) = z, & \text{the "hypotenuse"}, \\ f_{\mathscr{A}}(\mathbf{x}) = \frac{1}{12}xy, & \text{the "area"}, \\ f_{\mathscr{C}}(\mathbf{x}) = \frac{1}{60}xyz, & \text{the product of coordinates.} \end{cases} \tag{1.4}$$

Recall that we assume, as throughout, that the pair $(\mathcal{O}, f)$ is strongly primitive; the fractions in (1.4) are to remove extraneous prime factors (e.g., it is elementary that the product of coordinates xyz in a Pythagorean triple is always divisible by 60).

We will refer to the pairs $(\mathcal{O}, f)$ above with $f \in \{f_{\mathscr{H}}, f_{\mathscr{A}}, f_{\mathscr{C}}\}$ as Examples A, B, and C, respectively.

THEOREM 1.5 ([Kon07, Kon09, KO12]): *Assume the critical exponent δ of Γ is sufficiently close to* 1. *Then we have*

$$R_0(\mathcal{O}, f) \ \leq \ \begin{cases} 13, & \textit{for Example } A, \\ 40, & \textit{for Example } B, \\ 58, & \textit{for Example } C. \end{cases} \tag{1.6}$$

Remark 1.7: We have taken this opportunity to correct the values of R in the statement of [KO12, Theorem 1.5], which were improperly computed; see Remark 2.38.

The upper bounds on R_0 given in (1.6) are based on Gamburd's spectral gap [Gam02] (see §2.1); the lower bounds, and expected true values of R_0, are the so-called "sieve dimensions" (see Remark 2.19), given by

$$\kappa = \kappa(\mathcal{O}, f) \ := \ \begin{cases} 1, & \text{for Example } A, \\ 4, & \text{for Example } B, \\ 5, & \text{for Example } C. \end{cases} \tag{1.8}$$

In Example A, the upper bound on the saturation number has been reduced significantly in [BK13] to

$$R_0(\mathcal{O}, f_{\mathscr{H}}) \leq 4 \tag{1.9}$$

by quite different methods from those discussed here, so we will not focus on this case. For the other two choices of f, an easy consequence of our method is the following improvement.

THEOREM 1.10: *Theorem 1.5 holds with* (1.6) *replaced by*

$$R_0(\mathcal{O}, f) \le \begin{cases} 25, & \text{for Example } B, \\ 37, & \text{for Example } C. \end{cases} \tag{1.11}$$

Remark 1.12: For Example A, our method gives $R_0(\mathcal{O}, f_{\mathscr{H}}) \le 7$; see §3.3.1. This is an improvement over (1.6), but does not compete with (1.9).

Remark 1.13: In all the statements above (and below), the Zariski density is an easy consequence of a lower bound on the cardinality of $\mathcal{O}(f, R)$ restricted roughly to an archimedean ball; see Remark 2.48.

1.3. ANISOTROPIC ORBITS. In [LS10], Liu–Sarnak consider a related problem. Instead of the isotropic Pythagorean form $\mathcal{Q}$ in (1.2), they let $\mathcal{Q}$ be an anistropic (over $\mathbb{Q}$) indefinite integral ternary quadratic form, e.g., $\mathcal{Q}(\mathbf{x}) = x^2 + y^2 - 3z^2$. This means that there are no rational points on the cone $\mathcal{Q} = 0$, and so one instead considers the affine quadric

$$V = V_{\mathcal{Q},t} := \{\mathbf{x} : \mathcal{Q}(\mathbf{x}) = t\}, \tag{1.14}$$

for a fixed non-zero integer t, chosen so that $V(\mathbb{Z})$ is non-empty. For simplicity, assume that $t \cdot \Delta(\mathcal{Q})$ is square-free, where $\Delta(\mathcal{Q})$ is the discriminant of $\mathcal{Q}$. The study of the vectors in $V(\mathbb{Z})$ reduces (see [LS10, §2]) to that of orbits $\mathcal{O} := \mathbf{y} \cdot \Gamma$, where $\mathbf{y} \in V(\mathbb{Z})$, and $\Gamma = \mathrm{SO}_{\mathcal{Q}}(\mathbb{Z})$ is the integer matrix group preserving $\mathcal{Q}$. (Such an orbit is not thin, as Γ is a lattice in $G = \mathrm{SO}_{\mathcal{Q}}(\mathbb{R})$, with critical exponent $\delta = 1$.) Let $f(\mathbf{x}) = xyz$ be the product of coordinates, and recall our assumption that the pair $(\mathcal{O}, f)$ is strongly primitive (for example, this is guaranteed if $\mathbf{y} = (1, 1, 1)$).

We refer to this pair $(\mathcal{O}, f)$ as Example D.

THEOREM 1.15 ([LS10]): *We have the following bound on the saturation number in Example D:*

$$R_0(\mathcal{O}, f) \le 26. \tag{1.16}$$

Assuming the Selberg Eigenvalue Conjecture (see Theorem 2.2), we have

$$R_0(\mathcal{O}, f) \le 22. \tag{1.17}$$

Remark 1.18: Note that for the product of coordinates here the sieve dimension is $\kappa = 3$ (see §2.4), rather than $\kappa = 5$ in (1.8) for Example C, that is, for

$f = f_{\mathscr{C}}$. This is because in the isotropic case, there are non-constant polynomial parametrizations of the integer points of the corresponding orbits which can be (and, in the Pythagorean case, are) reducible; see Remark 2.32.

As a consequence of our method, we have

THEOREM 1.19: *Theorem 1.15 holds unconditionally with* (1.16) *replaced by*

$$R_0(\mathcal{O}, f) \leq 16.$$

Assuming the Selberg Eigenvalue Conjecture, (1.17) *may be replaced by*

$$R_0(\mathcal{O}, f) \leq 14.$$

1.4. NEW OBSERVATION. Our key new observation is that, for all the problems above (indeed for nearly all natural Affine Sieve problems in the literature), the polynomial f is homogeneous. Roughly speaking, this allows us, in the modular/archimedean decomposition of the Affine Sieve, to projectivize, taking a larger stabilizer group (see §3.2). As a result, we have no modular loss in the error terms, whereas in the previous approaches, a power of the level was lost; see Remark 3.6. The upshot is an improvement by a factor of two in the level of distribution (see §2.3) in the above problems, which translates to the above-claimed improved bounds on saturation numbers. In fact our main observation is a general principle, applying to many other settings, e.g., the pairs $(\mathcal{O}, f)$ considered in [NS10] with f homogeneous; we will not bother with other applications here.

1.5. OUTLINE. In §2 we collect some relevant background. In particular, we recall facts on spectral gaps, counting, levels of distribution, and the Diamond–Halberstam–Richert sieve. We also sketch proofs of Theorems 1.5 and 1.15, since our proofs of Theorems 1.10 and 1.19 are nearly identical. In §3 we explain our new observation, and use it to prove Theorems 1.10 and 1.19.

1.6. NOTATION. We use the standard notation $f = O(g)$ and $f \ll g$ synonymously to mean $f(x) \leq Cg(x)$ for an implied constant $C > 0$ and all x sufficiently large. Unless otherwise specified, C may depend only on the pair $(\mathcal{O}, f)$, which is treated as fixed. The little-oh notation $f = o(g)$ means $f/g \to 0$.

ACKNOWLEDGEMENTS. The authors thank Shamgar Gurevich, Nick Katz, and Peter Sarnak for enlightening discussions.

2. Background

2.1. Spectral Gap. We take the following as our definition of a spectral gap for the cases of interest to us here. For $\mathcal{Q}$ a ternary indefinite integral quadratic form (either isotropic or anisotropic over $\mathbb{Q}$), let

$$G = \mathrm{SO}_{\mathcal{Q}}(\mathbb{R}) \cong \mathrm{SO}_{2,1}(\mathbb{R})$$

be its stabilizer group, and let $\Gamma < G(\mathbb{Z})$ be a finitely generated (and hence geometrically finite) integer subgroup with critical exponent

$$\delta > 1/2.$$

The decomposition of the right regular representation of G on $L^2(\Gamma\backslash G)$ is of the form [GGPS66, LP82]

$$L^2(\Gamma\backslash G) = V_0 \oplus V_1 \oplus \cdots \oplus V_J \oplus V_{temp}.$$

Here V_{temp} is a (reducible) subspace consisting of the tempered spectrum; the V_j, $j = 1, \ldots, J$ are isomorphic as G-representations to complementary series representations with corresponding parameters

$$1/2 < s_J \leq \cdots \leq s_1 < \delta \leq 1$$

(in our normalization, the principal series representations lie on the critical line $\mathfrak{Re}(s) = 1/2$); and V_0 is either the trivial representation if Γ is a lattice, or a complementary series representation of parameter $s_0 = \delta$ if $\delta < 1$ [Pat76, Sul84]. We say a number $s \in (1/2, 1)$ **appears** in $L^2(\Gamma\backslash G)$ if it arises as one of the s_j above.

For a square-free integer $q \geq 1$, define the level q principal congruence subgroup of Γ as

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

We have a similar decomposition for $L^2(\Gamma(q)\backslash G)$, and the inclusion $\Gamma(q) < \Gamma$ induces the reverse inclusion on spectrum; that is, any parameter s which appears in $L^2(\Gamma\backslash G)$ also appears in $L^2(\Gamma(q)\backslash G)$. We say $s \in (1/2, 1)$ is the **new** spectrum at level q, if the parameter s appears in $L^2(\Gamma(q)\backslash G)$ but does not arise in this way as a lift from $L^2(\Gamma\backslash G)$; let $\mathrm{Spec}^{new}(q)$ denote the new spectra at level q.

We say that Γ has a **uniform spectral gap**

$$\frac{1}{2} \leq \theta = \theta(\Gamma) < \delta$$

if there exists an integer

$$\mathfrak{B} \geq 1 \tag{2.1}$$

so that, for all q coprime to $\mathfrak{B}$,

$$\mathrm{Spec}^{new}(q) \subset (1/2, \theta].$$

In particular, the "base" parameter δ remains isolated as q ranges through square-free numbers coprime to the "bad" modulus $\mathfrak{B}$.

In our archimedean (as opposed to combinatorial, for which see [SGV12]) setting, the following is the current state of affairs on spectral gaps.

THEOREM 2.2: *Assume $\delta > 1/2$. Then:*

- *Γ has some spectral gap $\theta \in [1/2, \delta)$* [BG08, BGS10, BGS11].
- *If moreover $\delta > 5/6$, then we can take $\theta = 5/6$* [Gam02].
- *If moreover Γ is a congruence group (and hence $\delta = 1$), then we can take $\theta = 1/2 + 7/64 = 39/64$ and $\mathfrak{B} = 1$* [JL70, KS03].
- *If moreover we assume the Selberg Eigenvalue Conjecture, then we can take $\theta = 1/2$ and $\mathfrak{B} = 1$* [Sel65].

Remark 2.3: In the case $\mathcal{Q}$ is anisotropic over $\mathbb{Q}$, that is, for Example D, the quotient $\Gamma\backslash G$ is compact, and the Jacquet–Langlands correspondence is used to apply the best-known bounds towards the Selberg Eigenvalue (or Generalized Ramanujan) Conjecture in the statement of Theorem 2.2.

2.2. EFFECTIVE COUNTING ON CONGRUENCE TOWERS. With the spectral gap in place, we state the following now-standard smooth counting theorem (see, e.g., [BKS10] or [BK13, Theorem 2.9]). We define a norm on G via $\|g\|^2 = \mathrm{tr}\, g^t g$.

THEOREM 2.4: *Assume $\Gamma < G$ is a finitely-generated discrete group as above with critical exponent $\delta > 1/2$ and spectral gap $1/2 \leq \theta < \delta$. Then for $T \to \infty$, there is a function $\Upsilon_T : G \to \mathbb{R}_{\geq 0}$ with the following properties.*

(i) *Υ_T is a smoothed indicator of $\|g\| < T$, in the sense that*

$$\Upsilon_T(g) = \begin{cases} 1, & \text{if } \|g\| < \frac{1}{2}T, \\ 0, & \text{if } \|g\| > 2T, \\ \in [0,1], & \text{otherwise,} \end{cases} \tag{2.5}$$

and[2]

$$\sum_{\gamma\in\Gamma}\Upsilon_T(\gamma) = T^{\delta+o(1)}. \tag{2.6}$$

Moreover,

(ii) *for any* $\gamma_0 \in \Gamma$, *any square-free* $q \geq 1$ *coprime to* $\mathfrak{B}$ *in* (2.1), *and any* $\Xi(q)$ *satisfying* $\Gamma(q) \leqq \Xi(q) \leqq \Gamma$, *we have*

$$\sum_{\gamma_1\in\Xi(q)}\Upsilon_T(\gamma_1\gamma_0) = \frac{1}{[\Gamma:\Xi(q)]}\sum_{\gamma\in\Gamma}\Upsilon_T(\gamma) + O(T^{\theta+o(1)}). \tag{2.7}$$

The implied constant above is independent of q *and* γ_0.

Remark 2.8: The interpretation of (2.7) is that one has effective (with power savings down to the spectral gap) equidistribution of Γ along congruence towers mod q. It is important here (in fact absolutely crucial to our observation!) to have the flexibility to choose any $\Xi(q)$ lying between Γ and the full level q principal congruence subgroup $\Gamma(q)$.

2.3. Level of Distribution. We now define a certain finite sequence

$$\mathcal{A} = \{a_n(T)\}$$

of non-negative real numbers depending on a parameter

$$T \to \infty,$$

which will play a key role in the analysis. This sequence is supported on values of $f(\mathbf{x})$, with $\mathbf{x} \in \mathcal{O}$, where the pair $(\mathcal{O}, f)$ is one of the pairs discussed in §1.2 or §1.3, that is, Examples *A*–*D*. For ease of exposition, we assume henceforth that $\mathfrak{B} = 1$; minor adjustments are needed in the general case. Using the smooth counting function from the previous subsection, we define

$$a_n(T) := \sum_{\gamma\in\Gamma}\Upsilon_T(\gamma)\cdot\mathbf{1}_{\{f(\mathbf{y}\cdot\gamma)=n\}}. \tag{2.9}$$

Thus $a_n(T)$ counts roughly the number of representations of n of the form $f(\mathbf{y}\cdot\gamma)$, for γ restricted to an archimedean ball. (In the case that $\mathbf{y}$ has a non-trivial stabilizer in Γ, this will be an over-count; statements about the Zariski closure of $\mathcal{O}(f, R)$ are not sensitive to this over-counting.)

[2] Throughout, only the exponents will be relevant to our analysis, so we will be quite crude with such statements, even when much more information is available.

We first determine the total amount of "mass" contained in $\mathcal{A}$, that is, we have from (2.6) the approximation

$$|\mathcal{A}| := \sum_n a_n(T) = T^{\delta+o(1)}. \tag{2.10}$$

Next we introduce a parameter N which controls the number of terms in $\mathcal{A}$ that are non-zero, setting

$$N := \max\{n \geq 1 : a_n \neq 0\}. \tag{2.11}$$

Since $\mathbf{y}$ is treated as fixed and $\gamma \in \Gamma$ is of size T, we have roughly that $|f(\mathbf{y} \cdot \gamma)| \leq N$, where

$$N = T^{\deg(f)+o(1)}. \tag{2.12}$$

For a square-free integer $q \geq 1$ called the **level**, we will need to understand the distribution of the sequence $\mathcal{A}$ along multiples of q. To this end, we define

$$|\mathcal{A}_q| := \sum_{n \equiv 0(q)} a_n(T). \tag{2.13}$$

The following key theorem is used to determine for how large we can take the level and still prove equi-distribution.

THEOREM 2.14: *Let $(\mathcal{O}, f)$ be as in Examples A–D, with Γ having critical exponent $\delta > 1/2$ and spectral gap $\theta < \delta$. For any square-free integer $q \geq 1$, we have the estimate*

$$|\mathcal{A}_q| = \omega(q) \cdot |\mathcal{A}| + O(q \cdot T^{\theta} (qT)^{o(1)}), \tag{2.15}$$

where $\omega(q)$ is a "local density" function with the following properties. It is a multiplicative function on square-free q's with

(1) $\omega(1) = 1$,

(2) *for all primes $p \geq 2$,*

$$0 \leq \omega(p) < 1, \tag{2.16}$$

and

(3) *there are constants $K \geq 2$ and $\kappa \geq 1$ so that we have the local density bound*

$$\prod_{z_1 \leq p \leq z} \frac{1}{1-\omega(p)} \leq \Big(\frac{\log z}{\log z_1}\Big)^{\kappa} \Big(1 + \frac{K}{\log z_1}\Big) \tag{2.17}$$

for any $2 \leq z_1 < z$.

Remark 2.18: Versions of Theorem 2.14 are proved in [Kon09, Proposition 4.3], [KO12, §5.2], and [LS10, Theorem 2.1] for Examples *A*–*D*, but we repeat a sketch of the proof below, as it will be relevant to us later.

Remark 2.19: One can interpret (2.17) as insisting that the local density at primes be roughly

$$\omega(p) \approx \frac{\kappa}{p}, \tag{2.20}$$

at least on average; see Lemma 2.30. The number κ appearing in (2.17) is called the "sieve dimension" for $\mathcal{A}$; see (1.8). Note that κ is not uniquely defined by (2.17), as any larger value also satisfies (2.17); in practice one typically takes the least allowable value.

Sketch of Proof. To prove Theorem 2.14, we first insert the definition (2.9) into (2.13):

$$|\mathcal{A}_q| = \sum_{n \equiv 0(q)} a_n(T) = \sum_{\gamma \in \Gamma} \Upsilon_T(\gamma) \cdot \mathbf{1}_{\{f(\mathbf{y}\cdot\gamma) \equiv 0 (\mathrm{mod}\, q)\}}.$$

The first most basic Affine Sieve observation is that the condition

$$f(\mathbf{y} \cdot \gamma) \equiv 0 (\mathrm{mod}\, q) \tag{2.21}$$

can be captured by breaking the sum according to the residue of γ mod q. In other words, we can decompose

$$\Gamma \;\cong\; \Gamma(q) \times (\Gamma(q) \backslash \Gamma). \tag{2.22}$$

Using this decomposition and following the procedure below, one would obtain (2.15) with the worse error term $O(q^2 T^\theta)$, ignoring $o(1)$'s. This would lead to (2.36) being replaced by the exponent of distribution $\alpha = (\delta - \theta)/(3 \deg(f))$.

Instead, what is done in [Kon09, LS10, KO12] is to capture the condition (2.21) by decomposing $\mathbf{y} \cdot \gamma$ (rather than just γ) into residue classes mod q. To this end, let $\Gamma_{\mathbf{y}}(q)$ be the stabilizer group of $\mathbf{y} (\mathrm{mod}\, q)$, that is, define

$$\Gamma_{\mathbf{y}}(q) := \{\gamma \in \Gamma : \mathbf{y} \cdot \gamma \equiv \mathbf{y} (\mathrm{mod}\, q)\},$$

and write $\gamma \in \Gamma$ uniquely as

$$\gamma = \gamma_1 \gamma_0,$$

with $\gamma_1 \in \Gamma_{\mathbf{y}}(q)$ and $\gamma_0 \in \Gamma_{\mathbf{y}}(q) \backslash \Gamma$. Then since $\mathbf{y}\gamma_1 \equiv \mathbf{y} (\mathrm{mod}\, q)$, we have that

$$f(\mathbf{y} \cdot \gamma) = f(\mathbf{y} \cdot \gamma_1 \gamma_0) \equiv f(\mathbf{y} \cdot \gamma_0) \pmod{q}. \tag{2.23}$$

Hence applying (2.7) with $\Xi(q) = \Gamma_{\mathbf{y}}(q)$, we have

$$|\mathcal{A}_q| = \sum_{\gamma_0 \in \Gamma_{\mathbf{y}}(q)\backslash\Gamma} \sum_{\gamma_1 \in \Gamma_{\mathbf{y}}(q)} \Upsilon_T(\gamma_1\gamma_0) \cdot \mathbf{1}_{\{f(\mathbf{y}\cdot\gamma_1\gamma_0)\equiv 0 (\mathrm{mod}\, q)\}}$$

$$= \sum_{\gamma_0 \in \Gamma_{\mathbf{y}}(q)\backslash\Gamma} \mathbf{1}_{\{f(\mathbf{y}\cdot\gamma_0)\equiv 0 (\mathrm{mod}\, q)\}} \Big[\sum_{\gamma_1 \in \Gamma_{\mathbf{y}}(q)} \Upsilon_T(\gamma_1\gamma_0) \Big]$$

$$\overset{(2.7)}{=} \sum_{\substack{\gamma_0 \in \Gamma_{\mathbf{y}}(q)\backslash\Gamma \\ f(\mathbf{y}\cdot\gamma_0)\equiv 0 (\mathrm{mod}\, q)}} \Big[\frac{1}{[\Gamma : \Gamma_{\mathbf{y}}(q)]} |\mathcal{A}| + O(T^{\theta+o(1)}) \Big] \tag{2.24}$$

$$= \frac{\mathcal{C}(\Gamma_{\mathbf{y}}(q); f)}{[\Gamma : \Gamma_{\mathbf{y}}(q)]} |\mathcal{A}| + O(\mathcal{C}(\Gamma_{\mathbf{y}}(q); f) \cdot T^{\theta+o(1)}). \tag{2.25}$$

Here we have defined

$$\mathcal{C}(\Xi(q); f) := \#\{\gamma_0 \in \Xi(q)\backslash\Gamma : f(\mathbf{y}\cdot\gamma_0) \equiv 0 (\mathrm{mod}\, q)\}, \tag{2.26}$$

where $\Xi(q)$ is any group with $\Gamma(q) \leqq \Xi(q) \leqq \Gamma$, for which the above makes sense, that is, whenever the condition $f(\mathbf{y}\cdot\gamma_0) \equiv 0 (\mathrm{mod}\, q)$ is left-$\Xi(q)$ invariant.

Now we can set the local density function to be

$$\omega(q) := \frac{\mathcal{C}(\Gamma_{\mathbf{y}}(q); f)}{[\Gamma : \Gamma_{\mathbf{y}}(q)]}, \tag{2.27}$$

whence we have a decomposition of the form (2.15).

It is straightforward to compute that the index

$$[\Gamma : \Gamma_{\mathbf{y}}(q)] = q^{2+o(1)}, \tag{2.28}$$

and moreover that, very roughly,

$$\mathcal{C}(\Gamma_{\mathbf{y}}(q); f) < q^{1+o(1)}. \tag{2.29}$$

Inserting (2.29) into the error term of (2.25) confirms the error term in (2.15).

It remains to verify the properties of ω. The condition (1), that is, that $\omega(1) = 1$, is clear, and multiplicativity follows from Strong Approximation and Goursat's Lemma. It follows from the strong primitivity assumption that $\omega(p) < 1$ for all primes. Verification of the key property (2.17) is postponed to the next Lemma, whence the proof of Theorem 2.14 is complete. ∎

The following Lemma verifies (2.20), from which the local density bound (2.17) follows by classical methods.

LEMMA 2.30: *For primes p sufficiently large, we have the following estimates on $\omega(p)$. In the "thin Pythagorean" cases, we have that (see* [KO12, Lemma 5.4]*)*

$$\omega(p) = \begin{cases} \frac{2}{p+1}, & \text{if } p \equiv 1 (\mathrm{mod}\, 4), \\ 0, & \text{if } p \equiv 3 (\mathrm{mod}\, 4), \end{cases} \quad \text{for Example } A,$$

$$\omega(p) = \frac{4}{p+1}, \quad \text{for Example } B,$$

and[3]

$$\omega(p) = \begin{cases} \frac{6}{p+1}, & \text{if } p \equiv 1 (\mathrm{mod}\, 4), \\ \frac{4}{p+1}, & \text{if } p \equiv 3 (\mathrm{mod}\, 4), \end{cases} \quad \text{for Example } C.$$

In the "anisotropic" case, we have [LS10, (6.4)] *that*

$$\omega(p) = \frac{3}{p} + O\Big(\frac{1}{p^2}\Big), \quad \text{for Example } D.$$

In particular, (2.17) *holds with*

$$\kappa = \begin{cases} 1, & \text{in Example } A, \\ 4, & \text{in Example } B, \\ 5, & \text{in Example } C, \\ 3, & \text{in Example } D. \end{cases} \tag{2.31}$$

Remark 2.32: It is only here in the local density estimate that the sieve can distinguish the sieve dimensions κ for the product of coordinates in the isotropic Example C and the anisotropic Example D (see Remark 1.18). The form $\mathcal{Q}$ being isotropic is equivalent to the cone $\mathcal{Q} = 0$ (and other level sets) being parametrizable by non-constant polynomial maps. In particular, if $\mathcal{Q}$ is isotropic, then there exist rational binary quadratic forms G_1, G_2, G_3 so that

$$\mathcal{Q}(G_1(c,d), G_2(c,d), G_3(c,d)) = 0.$$

If the G_j are reducible, then the product of coordinates $f(\mathbf{x}) = xyz$ can be the product of more than 3 irreducible factors, and this is exactly what happens in the Pythagorean case. On the other hand, no such parametrization exists if $\mathcal{Q}$ is anisotropic, whence the product of three coordinates always has sieve dimension $\kappa = 3$.

[3] This corrects a typo in [KO12, (5.6)].

To make this completely concrete for the form $\mathcal{Q} = x^2 + y^2 - z^2$, recall the ancient parametrization of Pythagorean triples $\mathbf{x} = (x, y, z)$ with y even as

$$\begin{cases} x = G_1(c,d) = c^2 - d^2, \\ y = G_2(c,d) = 2cd, \\ z = G_3(c,d) = c^2 + d^2. \end{cases}$$

Both G_1 and G_2 factor into products of two linear forms, and so in Example C,

$$f_{\mathscr{C}}(\mathbf{x}) = \frac{1}{60}xyz = \frac{1}{30}(c+d)(c-d)cd(c^2+d^2)$$

is a product of $\kappa = 5$ irreducible factors.

On the other hand, the form

$$\mathcal{Q}(\mathbf{x}) = x^2 + y^2 - 2z^2$$

is also isotropic over $\mathbb{Q}$, but the cone $\mathcal{Q} = 0$ has a parametrization

$$\begin{cases} x = G_1(c,d) = c^2 + 2cd - d^2, \\ y = G_2(c,d) = c^2 - 2cd - d^2, \\ z = G_3(c,d) = c^2 + d^2, \end{cases}$$

in which all three forms G_j are irreducible. In this example, the product of coordinates would have sieve dimension $\kappa = 3$.

In light of (2.20) and (2.10), the "main" term in the approximation (2.15) is roughly of size T^δ/q, while the "error" is about qT^θ. Balancing these terms, we can take q almost as large as $T^{(\delta-\theta)/2}$. Converting to the parameter N in (2.12), we see that the approximation (2.15) is a true asymptotic whenever

$$q < N^{(\delta-\theta)/(2\deg(f))-\varepsilon}, \tag{2.33}$$

for any fixed $\varepsilon > 0$. For later reference, we record the following estimate, which follows immediately from (2.15).

COROLLARY 2.34: *For any fixed $\varepsilon > 0$, there is an $\eta = \eta(\varepsilon) > 0$, so that*

$$\sum_{\substack{q < N^{\alpha-\varepsilon} \\ q \text{ square-free}}} ||\mathcal{A}_q| - \omega(q) \cdot |\mathcal{A}|| \ll_\varepsilon |\mathcal{A}|^{1-\eta}, \tag{2.35}$$

as $T \to \infty$, where

$$\alpha := \frac{\delta - \theta}{2\deg(f)} \tag{2.36}$$

is the exponent in (2.33).

Remark 2.37: The quantity N^α is called a **level of distribution** for $\mathcal{A}$, and the exponent α in (2.36) is called an **exponent of distribution**. This is not a quantity intrinsic to $\mathcal{A}$ but is rather a function of what one can prove about $\mathcal{A}$. In particular, any smaller value of α is also an exponent of distribution, but in applications, one wishes to take α as large as possible.

Remark 2.38: We are correcting here a typo in [KO12, (2.23)], where $\deg(f)$ was omitted from α (our α is $1/\mu$ in the notation of [KO12]); hence the values of R computed in [KO12] are only accurate in the case $f = f_{\mathscr{H}}$ of Example *A*; see Remark 1.7.

Remark 2.39: In sieve applications, one only needs the average estimate (2.35) and not the estimate for individual q discussed before (2.33). In Example *A*, it is exactly this averaging which is exploited in [BK13] to prove (1.9). In Examples *B*–*D*, we do not currently know how to exploit this average, and so the level of distribution just follows from the individual estimate (2.15). See [Mar10] for some sharp levels of distribution for non-thin isotropic (and hence parametrizable; cf. Remark 2.32) orbits, also obtained by exploiting the average on q.

We now have all the properties we need from the sequence $\mathcal{A}$. In the next subsection, we recall the high-dimensional weighted sieve used in applications.

2.4. Diamond–Halberstam–Richert Sieve. Recall that $\mathcal{P}_R$ is the set of R-almost primes. Sieve theory produces an estimate for

$$\sum_{n \in \mathcal{P}_R} a_n(T),$$

given knowledge of the distribution of $\mathcal{A}$ along arithmetic progressions. Adapted to our setting, we have the following

Theorem 2.40 ([DHR88, DH97]): *Let $\mathcal{A}$, N, ω, κ, and α be as described in* (2.9), (2.11), (2.27), (2.31), *and* (2.36)*; in particular, they satisfy the key conditions* (2.17) *and* (2.35)*. It is convenient to define another parameter*

$$\tau := \frac{\alpha \log N}{\log |\mathcal{A}|} = \frac{\alpha \cdot \deg(f)}{\delta} + o(1). \tag{2.41}$$

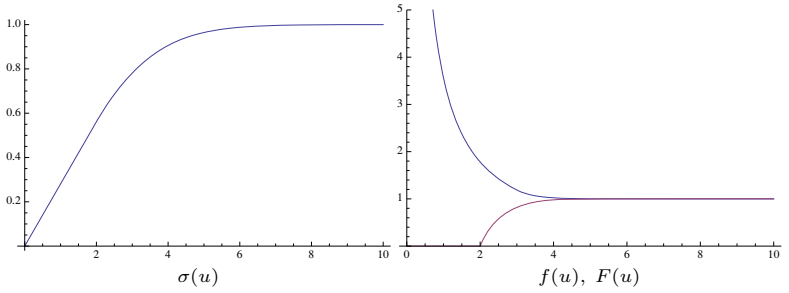

Figure 1. Plots of $\sigma(u)$, $f(u)$ and $F(u)$ for $\kappa = 1$.

(i) *Let $\sigma(u) = \sigma_\kappa(u)$ be the continuous solution of the differential-difference problem*

$$\begin{cases} u^{-\kappa}\sigma(u) = A_\kappa^{-1}, & \text{for } 0 < u \le 2,\ A_\kappa = (2e^\gamma)^\kappa \Gamma(\kappa+1), \\ (u^{-\kappa}\sigma(u))' = -\kappa u^{-\kappa-1}\sigma(u-2), & \text{for } u > 2, \end{cases} \tag{2.42}$$

where γ is the Euler constant and Γ is the Gamma function[4]*. Then there exist numbers*

$$\alpha_\kappa \ge \beta_\kappa \ge 2 \tag{2.43}$$

so that the following simultaneous differential-difference system has continuous solutions $F(u) = F_\kappa(u)$ and $f(u) = f_\kappa(u)$ which satisfy

$$F(u) = 1 + O(e^{-u}), \quad f(u) = 1 + O(e^{-u}),$$

and F (resp. f) decreases (resp. increases) monotonically towards 1 as $u \to \infty$:

$$\begin{cases} F(u) = 1/\sigma(u), & \text{for } 0 < u \le \alpha_\kappa, \\ f(u) = 0, & \text{for } 0 < u \le \beta_\kappa, \\ (uF(u))' = f(u-1), & \text{for } u > \alpha_\kappa, \\ (uf(u))' = F(u-1), & \text{for } u > \beta_\kappa. \end{cases} \tag{2.44}$$

See Figure 1 for plots of σ, f and F in the case $\kappa = 1$.

(ii) *For any two real numbers u and v with*

$$\tau^{-1} < u \le v, \quad \beta_\kappa < \tau v, \tag{2.45}$$

[4] There should be no confusion here with the discrete group Γ.

and assuming that

$$R > \frac{\tau u}{\alpha} - 1 + \frac{\kappa}{f(\tau v)} \int_1^{v/u} F(\tau v - s)\Big(1 - \frac{u}{v}s\Big)\frac{ds}{s}, \tag{2.46}$$

we have

$$\sum_{n \in \mathcal{P}_R} a_n(T) \gg |\mathcal{A}| \prod_{p < N^\alpha} (1 - \omega(p)) \gg \frac{|\mathcal{A}|}{(\log T)^\kappa}. \tag{2.47}$$

Remark 2.48: The statements in Examples *A–D* on the Zariski density of $\mathcal{O}(f, R)$ are then proved easily from the archimedean bounds in (2.47); see, e.g., the proof of [LS10, Corollary 2.3].

Remark 2.49: The sieve dimensions relevant to us are $\kappa = 1, 3, 4$, and 5, and we will need the corresponding values of the constant β_κ for (2.45). These are computed in [DHR88, p. 345], and we reproduce them in Table 1.

κ	1	3	4	5
β_κ	2	6.6408...	9.0722...	11.5347...

Table 1. Values of β_κ for $\kappa = 1, 3, 4, 5$.

While the expression on the right-hand side of (2.46) is unwieldy, it can certainly be estimated by one's favorite software package. That said, the following simplification is quite effective in practice (see [LS10, (6.15)]): for any $0 < \zeta < \beta_\kappa$, the expression is maximized by any value of

$$m_{\alpha,\kappa}(\zeta) := \frac{1}{\alpha}\Big(1 + \zeta - \frac{\zeta}{\beta_\kappa}\Big) - 1 + (\kappa + \zeta)\log\frac{\beta_\kappa}{\zeta} - \kappa + \zeta\frac{\kappa}{\beta_\kappa}. \tag{2.50}$$

2.5. Proofs of Theorems 1.5 and 1.15. It remains to insert the specific values of α, κ, and τ, and compute the resulting values of R for each of our examples.

2.5.1. *Example A.* To obtain as small a value of R as possible, we take δ as large as possible, that is, near 1, to take advantage of Gamburd's $\theta = 5/6$ gap in Theorem 2.2. At first, we just set $\delta = 1$. The degree of $f = f_{\mathscr{H}}$ is $\deg(f) = 1$, so the exponent of distribution (2.36) is

$$\alpha = \frac{1 - 5/6}{2} = \frac{1}{12},$$

and the sieve dimension is $\kappa = 1$. With these values of α and κ, the minimal value of $m(\zeta)$ in (2.50) is $m(0.12) = 13.93$, leading to the bound $R_0(\mathcal{O}, f_{\mathscr{H}}) \le 14$ for the saturation number. Letting δ be slightly less than 1, we can still ensure that α is large enough that the minimal value of $m(\zeta)$ is < 14.

But in fact, better methods are known to estimate R-values for linear (that is, dimension $\kappa = 1$) sieve problems using essentially identical assumptions; see, e.g., Richert's weights in [FI10, §25.3]. From the exponent of distribution $\alpha_A = 1/12$, these produce $R = 13$-almost primes, with room to perturb δ to a little below 1. This is the R value we stated in Theorem 1.5 for Example A. Regardless, none of these values are relevant anymore, in light of (1.9).

2.5.2. *Example B.* Because the Pythagorean form $\mathcal{Q}(\mathbf{x}) = x^2 + y^2 - z^2$ is isotropic with reducible parametrizing forms (see Remark 2.32), the sieve dimension for the "area" function $f(\mathbf{x}) = f_{\mathscr{A}}(\mathbf{x}) = \frac{1}{12}xy$ is $\kappa = 4$ (rather than $\kappa = 2$). The degree is $\deg(f) = 2$. As above, we begin by taking $\delta = 1$ and using Gamburd's gap $\theta = 5/6$. This gives the exponent of distribution

$$\alpha = \frac{1 - 5/6}{2 \cdot 2} = \frac{1}{24}.$$

Optimizing $m(\zeta)$ with these values gives $m(0.16) = 39.28$. Again, letting δ be slightly below 1 still recovers the value $R = 40$, as claimed in Theorem 1.5.

2.5.3. *Example C.* For $f = f_{\mathscr{C}}$, the degree is $\deg(f) = 3$ and sieve dimension is $\kappa = 5$. Again we take $\delta = 1$ and $\theta = 5/6$, giving the exponent of distribution

$$\alpha = \frac{1 - 5/6}{2 \cdot 3} = \frac{1}{36}.$$

Now optimizing $m(\zeta)$ gives $m(0.136) = 57.3$. For δ slightly below 1, we still recover $R = 58$.

2.5.4. *Example D.* In this non-thin anisotropic example, we have $\delta = 1$, $\deg(f) = 3$, and sieve dimension $\kappa = 3$. Using the Kim–Sarnak spectral gap $\theta = 39/64$ in Theorem 2.2, we obtain the exponent of distribution

$$\alpha = \frac{1 - 39/64}{2 \cdot 3} \approx \frac{1}{15.36}.$$

Optimizing $m(\zeta)$ gives $m(0.186) = 25.26$, giving the claimed value $R = 26$. Assuming the Selberg Eigenvalue Conjecture, we can take $\theta = 1/2$ and

$$\alpha = \frac{1 - 1/2}{2 \cdot 3} = \frac{1}{12}.$$

Then $m(\zeta)$ is optimized at $m(0.23) = 21.3$, giving $R = 22$, as claimed in Theorem 1.15.

These are the values of R produced in [Kon09], [LS10], and [KO12]. In the next section, we make one further simple observation, which has the effect of doubling the exponent of distribution over that in (2.36). Using the same methods as here, we then conclude Theorems 1.10 and 1.19.

3. Proofs of Theorems 1.10 and 1.19

We keep all the same notation from the previous section, first describing our initial aim in rough terms, before explaining our new observation.

3.1. Initial Idea. The goal of this project was to try to improve the level of distribution by exploiting the γ_0 sum in (2.24), which was estimated trivially to arrive at (2.25). Of course this requires us to keep track of all the lower order terms in (2.7), rather than estimating them in absolute value. We proceed as follows.

We will want $\Xi(q)\backslash\Gamma$ to be a group (i.e., $\Xi(q)$ to be normal in Γ), so return to the decomposition (2.22); that is, we set $\Xi(q) = \Gamma(q)$, rather than $\Xi(q) = \Gamma_{\mathbf{y}}(q)$. (So the length of the γ_0 sum in (2.24) is now about q^2 instead of q, but we hope to recover this loss and more.) Assume for simplicity that $\Gamma(q)\backslash\Gamma \cong \mathrm{PSL}_2(q)$ and that q is prime. The space $L^2(\Gamma(q)\backslash G)$ carries not only a right (regular) G-action, but also a left (Hecke-like) $\Gamma(q)\backslash\Gamma$-action. Decomposing with respect to the latter action, the estimate (2.7) can be obtained from an expansion of the form (see the discussion after [BK13, (2.12)])

$$\sum_{\gamma_1 \in \Gamma(q)} \Upsilon_T(\gamma_1\gamma_0) = \sum_{\rho \in \widehat{\mathrm{PSL}_2(q)}} \mathcal{M}_\rho(T, q; \gamma_0), \tag{3.1}$$

where ρ ranges over irreducible unitary representations of the finite group $\mathrm{PSL}_2(q)$, and $\mathcal{M}_\rho$ is the contribution coming from ρ. The first term in (2.7) comes from $\rho = \mathbf{1}$, that is, the trivial representation; the other ρ's come from the new spectrum, and (2.7) is obtained by controlling these terms in totality by the spectral gap.

Instead of estimating the error terms in absolute value, we will want to capitalize on the full decomposition (3.1). So we insert it into the analogue of (2.24), capture the condition $f(\mathbf{y} \cdot \gamma_0) \equiv 0(q)$ by abelian harmonic analysis, and carry out the $\gamma_0 \in \mathrm{PSL}_2(q)$ sum on each irreducible. Expanding out the terms, one

faces the following problem, which seems to be new (see the somewhat related questions arising in [SA87, Kat93]): Given an irreducible unitary representation (ρ, V) of a finite non-abelian group, e.g., $\mathrm{PSL}_2(\mathbb{F}_q)$, an additive character ψ on $\mathbb{F}_q$, and a polynomial f on $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{PSL}_2$, capture cancellation in the matrix coefficients of ρ against the character of the polynomial; that is, give a non-trivial estimate for a sum of the form

$$\sum_{\gamma \in \mathrm{PSL}_2(\mathbb{F}_q)} \langle \rho(\gamma).v, w \rangle \cdot \psi(f(\gamma)),$$

for $v, w \in V$. While the initial aim of the project was somewhat sophisticated, after computing several explicit examples of the above type, we stumbled upon a completely elementary observation that had been previously overlooked. Its effect, in our applications, is to make the γ_0 sum in the analogue of (2.24) have length q^{ε}, rather than q, leading to a level of distribution twice as large as before. So while the above general problem is still interesting and may have other applications, in the end it is of no consequence to our current results.

3.2. The Observation. The key new observation is that one can use a larger group than $\Gamma_{\mathbf{y}}(q)$ in capturing the condition (2.21). To this end, we introduce the group $\Gamma_{\langle \mathbf{y} \rangle}(q)$ which stabilizes the linear span of $\mathbf{y} \bmod q$. That is, we define

$$\begin{aligned} \Gamma_{\langle \mathbf{y} \rangle}(q) :=& \{\gamma \in \Gamma : \mathbf{y} \cdot \gamma \in \langle \mathbf{y} \rangle \, (\mathrm{mod}\, q)\} \\ =& \{\gamma \in \Gamma : \exists a \in (\mathbb{Z}/q\mathbb{Z})^{\times} \text{ with } \mathbf{y} \cdot \gamma \equiv a\mathbf{y} (\mathrm{mod}\, q)\}. \end{aligned}$$

Clearly

$$\Gamma(q) \leqq \Gamma_{\langle \mathbf{y} \rangle}(q) \leqq \Gamma.$$

Note that, because the functions f in all the Examples A–D are homogeneous, we have

$$f(\mathbf{y} \cdot \gamma_1 \gamma_0) \equiv a^{\deg(f)} f(\mathbf{y} \cdot \gamma_0) \ (\mathrm{mod}\, q), \quad \text{for some } a \in (\mathbb{Z}/q\mathbb{Z})^{\times},$$

whenever $\gamma_1 \in \Gamma_{\langle \mathbf{y} \rangle}(q)$. Hence we can replace (2.23) by the fact that

$$f(\mathbf{y} \cdot \gamma_1 \gamma_0) \equiv 0 (\mathrm{mod}\, q) \quad \text{if and only if} \quad f(\mathbf{y} \cdot \gamma_0) \equiv 0 (\mathrm{mod}\, q). \tag{3.2}$$

Remark 3.3: We emphasize here that it is not the cone $\mathcal{Q} = 0$ which takes advantage of this homogeneity (since we also consider other level sets, see (1.14)), but rather the sieve, which only asks for the distribution of $a_n(T)$ on multiples of q, see (2.13). If for other applications one wants to capture residue classes other than $0 (\mathrm{mod}\, q)$, then the homogeneity of f will not help.

Beyond this simple observation, we proceed exactly as described in §2.

3.3. The Proofs.

Theorem 3.4: *Theorem 2.14 holds exactly as stated, but with* (2.15) *replaced by*

$$|\mathcal{A}_q| = \omega(q) \cdot |\mathcal{A}| + O(T^\theta\,(qT)^{o(1)}). \tag{3.5}$$

Remark 3.6: The only difference to notice is that the error term in (3.8) is essentially T^θ, rather than qT^θ in (2.15); that is, we have recovered a power of q which was lost in the previous approach. This explains our comment in §1.4.

Sketch of Proof. We start with the same definition of $a_n(T)$ as in (2.9). Replacing the decomposition (2.22) with

$$\Gamma \cong \Gamma_{\langle \mathbf{y}\rangle}(q) \times (\Gamma_{\langle \mathbf{y}\rangle}(q)\backslash\Gamma), \tag{3.7}$$

we now write

$$\begin{aligned}
|\mathcal{A}_q| &= \sum_{\gamma_0 \in \Gamma_{\langle \mathbf{y}\rangle}(q)\backslash\Gamma} \sum_{\gamma_1 \in \Gamma_{\langle \mathbf{y}\rangle}(q)} \Upsilon_T(\gamma_1\gamma_0) \cdot \mathbf{1}_{\{f(\mathbf{y}\cdot\gamma_1\gamma_0)\equiv 0 (\mathrm{mod}\, q)\}} \\
&\overset{(3.2)}{=} \sum_{\gamma_0 \in \Gamma_{\langle \mathbf{y}\rangle}(q)\backslash\Gamma} \mathbf{1}_{\{f(\mathbf{y}\cdot\gamma_0)\equiv 0 (\mathrm{mod}\, q)\}} \Big[\sum_{\gamma_1 \in \Gamma_{\langle \mathbf{y}\rangle}(q)} \Upsilon_T(\gamma_1\gamma_0) \Big] \\
&\overset{(2.7)}{=} \sum_{\substack{\gamma_0 \in \Gamma_{\langle \mathbf{y}\rangle}(q)\backslash\Gamma \\ f(\mathbf{y}\cdot\gamma_0)\equiv 0 (\mathrm{mod}\, q)}} \Big[\frac{1}{[\Gamma : \Gamma_{\langle \mathbf{y}\rangle}(q)]} |\mathcal{A}| + O(T^\theta) \Big] \\
&= \frac{\mathcal{C}(\Gamma_{\langle \mathbf{y}\rangle}(q); f)}{[\Gamma : \Gamma_{\langle \mathbf{y}\rangle}(q)]} |\mathcal{A}| + O(\mathcal{C}(\Gamma_{\langle \mathbf{y}\rangle}(q); f) \cdot T^\theta),
\end{aligned} \tag{3.8}$$

where we applied (2.7) with $\Xi(q) = \Gamma_{\langle \mathbf{y}\rangle}(q)$, and used the definition (2.26). Note we are allowed to use $\Xi(q) = \Gamma_{\langle \mathbf{y}\rangle}(q)$ in (2.26); indeed, the observation (3.2) says precisely that $f(\mathbf{y} \cdot \gamma_0) \equiv 0 (\mathrm{mod}\, q)$ (as a condition on γ_0) is left-$\Gamma_{\langle \mathbf{y}\rangle}(q)$ invariant.

The new "local density" function

$$\omega(q) := \frac{\mathcal{C}(\Gamma_{\langle \mathbf{y}\rangle}(q); f)}{[\Gamma : \Gamma_{\langle \mathbf{y}\rangle}(q)]} \tag{3.9}$$

is then actually the same as that in (2.27). Indeed, just fix q and take $T \to \infty$, comparing (3.8) with (2.25). Thus the sieve dimensions are the same as before.

On the other hand, the index $[\Gamma : \Gamma_{\langle \mathbf{y} \rangle}(q)]$ is now of size $q^{1+o(1)}$ instead of (2.28). Thus comparing (3.9) to Lemma 2.30 gives

$$\mathcal{C}(\Gamma_{\langle \mathbf{y} \rangle}(q); f) < q^{o(1)} \tag{3.10}$$

instead of (2.29). Inserting (3.10) into (3.8) gives (3.5), as claimed. ∎

Then we obtain the same Corollary 2.34 but with the exponent of distribution

$$\alpha = \frac{\delta - \theta}{\deg(f)}, \tag{3.11}$$

instead of (2.36). That is, the effect of the simple observation (3.2) is to double the exponent of distribution.

With all the other ingredients in place, it remains to estimate the new values of R.

3.3.1. *Example A.* As before, we start by taking $\delta = 1$ with Gamburd's $\theta = 5/6$ spectral gap. The sieve dimension and degree are both $\kappa = \deg(f) = 1$. Inserting these values into (3.11) gives the exponent of distribution

$$\alpha = \frac{1 - 5/6}{1} = \frac{1}{6}.$$

Linear sieve methods then produce the value $R = 7$, with room to allow δ a little below 1; see Remark 1.12.

3.3.2. *Example B.* Again we take $\delta = 1$, $\theta = 5/6$, and $\deg(f) = 2$. The exponent of distribution is

$$\alpha = \frac{1 - 5/6}{2} = \frac{1}{12}$$

for this dimension $\kappa = 4$ problem. Optimizing the function $m(\zeta)$ in (2.50) gives $m(0.295) = 24.99$, or $R = 25$.

3.3.3. *Example C.* We set $\delta = 1$, with $\theta = 5/6$ and $\deg(f) = 3$. The exponent of distribution is then

$$\alpha = \frac{1 - 5/6}{3} = \frac{1}{18}$$

for a sieve of dimension $\kappa = 5$. Optimizing $m(\zeta)$ gives $m(0.25) = 36.3$, or $R = 37$. This completes the proof of Theorem 1.10.

3.3.4. *Example D.* Unconditionally, we have $\delta = 1$, and the Kim–Sarnak gap $\theta = 39/64$ with $\deg(f) = 3$. The exponent of distribution is then

$$\alpha = \frac{1 - 39/64}{3} \approx \frac{1}{7.7}.$$

The sieve dimension is $\kappa = 3$, and optimizing $m(\zeta)$ gives $m(0.33) = 15.9$, or $R = 16$.

Assuming the Selberg Eigenvalue Conjecture, we can take $\theta = 1/2$ with exponent

$$\alpha = \frac{1 - 1/2}{3} = \frac{1}{6}.$$

Now optimizing $m(\zeta)$ gives $m(0.4) = 13.7$, or $R = 14$.

This completes the proof of Theorem 1.19.

References

[BG08] J. Bourgain and A. Gamburd, *Uniform expansion bounds for Cayley graphs of* $\mathrm{SL}_2(\mathbb{F}_p)$, Annals of Mathematics **167** (2008), 625–642.

[BGS06] J. Bourgain, A. Gamburd and P. Sarnak, *Sieving and expanders* Comptes Rendus Mathématique. Académie des Sciences. Paris **343** (2006), 155–159.

[BGS10] J. Bourgain, A. Gamburd and P. Sarnak, *Affine linear sieve, expanders, and sum-product*, Inventiones Mathematicae **179** (2010), 559–644.

[BGS11] J. Bourgain, A. Gamburd and P. Sarnak, *Generalization of Selberg's 3/16th theorem and affine sieve*, Acta Mathematica **207** (2011), 255–290.

[BK13] J. Bourgain and A. Kontorovich, *The affine sieve beyond expansion I: thin hypotenuses*, 2013, International Mathematics Research Notices (2014). doi:10.1093/imrn/rnu222. `http://imrn.oxfordjournals.org/ content/early/ 2014/11/30/imrn.rnu222.full.pdf`

[BKS10] J. Bourgain, A. Kontorovich and P. Sarnak, *Sector estimates for hyperbolic isometries*, Geometric and Functional Analysis **20** (2010), 1175–1200.

[DH97] H. Diamond and H. Halberstam, *Some applications of sieves of dimension exceeding 1*, in *Sieve Methods, Exponential Sums, and their Applications in Number Theory (Cardiff, 1995)*, London Mathematical Society Lecture Note Series, Vol. 237, Cambridge University Press, Cambridge, 1997, pp. 101–107.

[DHR88] H. Diamond, H. Halberstam and H.-E. Richert, *Combinatorial sieves of dimension exceeding one*, Journal of Number Theory **28** (1988), 306–346.

[FI10] J. Friedlander and H. Iwaniec, *Opera de cribro*, American Mathematical Society Colloquium Publications, Vol. 57, American Mathematical Society, Providence, RI, 2010.

[Gam02] A. Gamburd, *On the spectral gap for infinite index "congruence" subgroups of* $\mathrm{SL}_2(\mathbf{Z})$, Israel Journal of Mathematics **127** (2002), 157–200.

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

[JL70] H. Jacquet and R. P. Langlands, *Automorphic Forms on* GL(2), Lecture Notes in Mathematics, Vol. 114. Springer-Verlag, Berlin, 1970.

[Kat93] N. M. Katz, *Estimates for Soto-Andrade sums*, Journal für die Reine und Angewandte Mathematik **438** (1993), 143–161.

[KO12] A. Kontorovich and H. Oh, *Almost prime Pythagorean triples in thin orbits*, Journal für die Reine und Angewandte Mathematik **667** (2012), 89–131; `arXiv:1001.0370`.

[Kon07] A. V. Kontorovich, *The Hyperbolic Lattice Point Count in Infinite Volume with Applications to Sieves*, Columbia University Thesis, 2007.

[Kon09] A. Kontorovich, *The hyperbolic lattice point count in infinite volume with applications to sieves*, Duke Mathematical Journal **149** (2009), 1–36; `arXiv:0712.1391`.

[KS03] H. Kim and P. Sarnak, *Refined estimates towards the Ramanujan and Selberg conjectures*, Journal of the American Mathematical Society **16** (2003), 175–181.

[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** (1982), 280–350.

[LS10] J. Liu and P. Sarnak, *Integral points on quadrics in three variables whose coordinates have few prime factors*, Israel Journal of Mathematics **178** (2010), 393–426.

[Mar10] G. Marasingha, *Almost primes represented by binary forms*, Journal of the London Mathematical Society **82** (2010), 295–316.

[NS10] A. Nevo and P. Sarnak, *Prime and almost prime integral points on principal homogeneous spaces*, Acta Mathematica **205** (2010), 361–402.

[Pat76] S. J. Patterson, *The limit set of a Fuchsian group*, Acta Mathematica **136** (1976), 241–273.

[SA87] J. Soto-Andrade, *Geometrical Gelfand models, tensor quotients, and Weil representations*, in *The Arcata Conference on Representations of Finite Groups (Arcata, Calif., 1986)* Proceedings of Symposia in Pure Mathematics, Vol. 47, American Mathematical Society, Providence, RI, 1987, pp. 305–316.

[Sel65] A. Selberg, *On the estimation of Fourier coefficients of modular forms*, in *Proceedings of Symposia in Pure Mathematics, Vol. VII*, American Mathematical Society, Providence, RI, 1965, pp. 1–15.

[SGS11] A. Salehi Golsefidy and P. Sarnak, *Affine sieve*, Journal of the American Mathematical Society, 2011, to appear.

[SGV12] A. Salehi Golsefidy and P. P. Varjú, *Expansion in perfect groups*, Geometric and Functional Analysis **22** (2012), 1832–1891.

[Sul84] D. Sullivan, *Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups*, Acta Mathematica **153** (1984), 259–277.