Invent math (2010) 179: 559–644
DOI 10.1007/s00222-009-0225-3

Affine linear sieve, expanders, and sum-product

Jean Bourgain · Alex Gamburd · Peter Sarnak

Received: 29 September 2008 / Accepted: 30 October 2009 /
Published online: 26 November 2009

Abstract Let $\mathcal{O}$ be an orbit in $\mathbb{Z}^n$ of a finitely generated subgroup Λ of $\mathrm{GL}_n(\mathbb{Z})$ whose Zariski closure $\mathrm{Zcl}(\Lambda)$ is suitably large (e.g. isomorphic to SL_2). We develop a Brun combinatorial sieve for estimating the number of points on $\mathcal{O}$ at which a fixed integral polynomial is prime or has few prime factors, and discuss applications to classical problems, including Pythagorean triangles and integral Apollonian packings. A fundamental role is played by the expansion property of the "congruence graphs" that we associate with $\mathcal{O}$. This expansion property is established when $\mathrm{Zcl}(\Lambda) = \mathrm{SL}_2$, using crucially sum-product theorem in $\mathbb{Z}/q\mathbb{Z}$ for q square-free.

The first author was supported in part by the NSF. The second author was supported in part by DARPA, the NSF and Sloan Foundation. The third author was supported in part by Veblen Fund (IAS) and the NSF.

J. Bourgain · P. Sarnak
School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540, USA

J. Bourgain
e-mail: bourgain@math.ias.edu

A. Gamburd
Department of Mathematics, University of California at Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA
e-mail: agamburd@ucsc.edu

P. Sarnak (✉)
Department of Mathematics, Princeton University, Princeton, NJ 08544, USA
e-mail: sarnak@math.princeton.edu

1 Introduction

This paper is concerned with the following general problem. For $j = 1, 2, \ldots, k$ let A_j be invertible integer coefficient polynomial maps of $\mathbb{Z}^n$ to $\mathbb{Z}^n$ (here $n \geq 1$ and the inverses of A_j's are assumed to be of the same type). Let Λ be the group generated by $A_1, \ldots, A_k$ and let $\mathcal{O} = \mathcal{O}_b = b \cdot \Lambda$ be the orbit of some $b \in \mathbb{Z}^n$ under Λ. Given a polynomial $f \in \mathbb{Q}[x_1, \ldots, x_n]$ which is integral on $\mathcal{O}$ our aim is to show that there are many points $x \in \mathcal{O}$ at which $f(x)$ has few or even the least possible number of prime factors, in particular that such points are Zariski dense in the Zariski closure,[1] $\mathrm{Zcl}(\mathcal{O})$ of $\mathcal{O}$. Let $\mathcal{O}(f, r)$ denote the set of $x \in \mathcal{O}$ for which $f(x)$ has at most r prime factors. As $r \to \infty$ the sets $\mathcal{O}(f, r)$ increase and potentially at some point become Zariski dense. Define the saturation number $r_0(\mathcal{O}, f)$ to be the least integer r such that $\mathrm{Zcl}(\mathcal{O}(f, r)) = \mathrm{Zcl}(\mathcal{O})$. It is by no means obvious that $r_0(\mathcal{O}, f)$ is finite or even if one should expect it to be so in general. If it is finite, we say that the pair $(\mathcal{O}, f)$ saturates.

Many classical results and conjectures are concerned with this problem in the case that Λ is a subgroup of $\mathbb{Z}^n$ acting by translations, that is $A_j(x) = x + b_j$. For example if $\Lambda = q\mathbb{Z}$, $\mathcal{O} = b + \Lambda$ and $f(x) = x$ one checks that Dirichlet's Theorem [17] is equivalent to $r_0(\mathcal{O}, f) = 1 + \nu((b, q))$, where $\nu(m)$ is the number of prime divisors of m. Another example is $\Lambda = \mathbb{Z}$, $\mathcal{O} = \mathbb{Z}$ and $f(x) = x(x + 2)$. Brun [11] invented the combinatorial sieve to show that this pair $(\mathcal{O}, f)$ saturates; the twin prime conjecture is equivalent to $r_0(\mathcal{O}, f) = 2$. One can use the classical combinatorial sieve in $\mathbb{Z}^n$ along the lines of Sect. 3 below, to show that any pair $(\mathcal{O}, f)$ with $\Lambda \subset \mathbb{Z}^n$ acting by translations saturates. One of the main goals of this paper is to study the case that Λ acts by affine linear transformations $(A_j(x) = a_j x + b_j)$. By increasing the dimension of the underlying space we can assume without loss of generality that $\Lambda \subset \mathrm{GL}_n(\mathbb{Z})$. We develop tools to attack the problem of $(\mathcal{O}, f)$ saturation at least if the radical of G, the Zariski closure of Λ in GL_n, contains no tori.[2] It turns out that in this context multiplication is much more problematic than addition and in extending the elementary combinatorial sieve to this affine-linear setting a number of novel problems present themselves, the most interesting and difficult being the proof that certain graphs (see Sect. 4) associated with reduction of the orbit mod q, form an expander family. A large part of the paper is concerned with proving this expander property in cases of the simplest semisimple groups. As a consequence we prove that $(\mathcal{O}, f)$ saturates when $G = \mathrm{Zcl}(\Lambda)$ is a $\mathbb{Q}$-morphic image of SL_2 or the unit group of a quaternion algebra over $\mathbb{Q}$. This already has a number of classical applications (see Sect. 6).

[1] Unless indicated otherwise, the Zariski closure is in affine space $\mathbb{A}^n$.

[2] The difficulties with tori are discussed in Sect. 2.1.

In investigating the finer aspects, such as the exact value of $r_0(\mathcal{O}, f)$, we need to take possible local congruence obstructions into account. If $q \geq 2$ is an integer and there is no $x \in \mathcal{O}$ such that $(f(x), q) = 1$, then any $f(x)$ is divisible by at least one of the prime factors of q. Since we demand Zariski density in the definition of the saturation number (and we assume that f is not constant when restricted to $\mathrm{Zcl}(\mathcal{O})$) it follows that r_0 will be larger than expected. For example, in this case $r_0(\mathcal{O}, f) \geq 2$. We say that $(\mathcal{O}, f)$ is primitive if there is no such local obstruction, that is if for every $q \geq 2$ there is $x \in \mathcal{O}$ such that $(f(x), q) = 1$. We will show (see Sect. 2) that in our setting this condition is easy to check and only involves finitely many q's. Note that being primitive is stronger than the condition that $\gcd(f(\mathcal{O})) = 1$. We give examples demonstrating this (see Sect. 2.4), and for this reason we will henceforth assume that $(\mathcal{O}, f)$ is always primitive.

In dimension 1, the affine linear motions preserving $\mathbb{Z}$ are $x \to \pm x + m$, $m \in \mathbb{Z}$ and a set is Zariski dense iff it is infinite. Hence Dirichlet's theorem [17] asserts that if Λ is a nontrivial (infinite) group of such motions of $\mathbb{A}^1$ and $\mathcal{O}$ an orbit and $f(x) = \lambda x + \beta$ with $\alpha, \beta \in \mathbb{Q}$, $\alpha \neq 0$ and primitive for $\mathcal{O}$, then $r_0(\mathcal{O}, f) = 1$. For f of degree 2 or higher, r_0 is not known but there is the following strong conjecture of Schinzel:

Conjecture 1.1 (Schinzel [55]) *Let $\mathcal{O}$ be an orbit of a nontrivial subgroup Λ of $\mathbb{Z}$ acting on $\mathbb{Z}$ by translations. Let $f \in \mathbb{Q}[x]$ with f integral and primitive on $\mathcal{O}$. If f has t irreducible factors in $Q[x]$ then $r_0(\mathcal{O}, f) = t$.*

Note that this implies that for $f(x) = f_1(x) \cdots f_t(x)$ with $f_j \in \mathbb{Q}[x]$ and irreducible, if there are no local congruence obstructions then there are infinitely many x at which $f_j(x)$ are simultaneously prime.

One can formulate the Hardy-Littlewood k-tuple conjectures [29] which are concerned with simultaneous linear equations for primes in two or more variables as follows:

Conjecture 1.2 (Hardy-Littlewood [29]) *Let Λ be a subgroup of $\mathbb{Z}^n$ acting by translations on $\mathbb{Z}^n$. Assume that for each j the j-th coordinate function x_j is nonconstant when restricted to Λ. If $\mathcal{O} = b + \Lambda$ is the orbit of b under Λ and $f(x) = x_1 x_2 \cdots x_n$ is $\mathcal{O}$ primitive then $r_0(\mathcal{O}, f) = n$. That is, the set of $x \in \mathcal{O}$ for which x_j are simultaneously prime, is Zariski dense in the affine linear subspace $b + \mathrm{Zcl}(\Lambda)$.*

The general case of Conjecture 1.2 follows from its special case when $\mathrm{rank}(\Lambda) = 1$, which is the exact from in which it was formulated in [29]. Moreover this rank one case is equivalent to the special case of Conjecture 1.1 when f factors into linear factors over $\mathbb{Q}$. Progress in this rank one case has been very slow. However, if $\mathrm{rank}(\Lambda) \geq 2$ there has been significant progress.

Vinogradov's fundamental method of bilinear forms, introduced in [63] allows one to establish Conjecture 1.2 for a nondegenerate (if $x \in \Lambda$ and x has two coordinates equal to 0 then $x = 0$) rank two subgroup of $\mathbb{Z}^3$. Recently Green and Tao [26] have established Conjecture 1.2 for a non-degenerate rank two subgroups Λ of $\mathbb{Z}^4$. In Sect. 6 we give an application of [26] to compute the saturation number of the pair $(\mathcal{O}, f)$, where $\mathcal{O}$ consists of Pythagorean triples in the affine cone in $\mathbb{A}^3$ and f is the area of the corresponding triangle. The case of rank one Λ in $\mathbb{Z}^2$ is a much sought-after case of Conjecture 1.2 since it implies the twin prime conjecture.

Before putting forth our general conjecture concerning primes in orbits of a linear action we explicate the simplest such case, one which could be viewed as an "$\mathrm{SL}_2(\mathbb{Z})$ analogue of Dirichlet's Theorem".

Conjecture 1.3 *Let Λ be a non-elementary subgroup of $\mathrm{SL}_2(\mathbb{Z})$ (equivalently, $\mathrm{Zcl}(\Lambda) = \mathrm{SL}_2$), b a primitive vector in $\mathbb{Z}^2$, $\mathcal{O} = b \cdot \Lambda$ the corresponding orbit and $\pi(\mathcal{O})$ the points $x \in \mathcal{O}$ with x_1 and x_2 both prime. Then*

$$\mathrm{Zcl}(\pi(\mathcal{O})) = \mathrm{Zcl}(\mathcal{O})(= \mathbb{A}^2)$$

iff $f(x) = x_1 x_2$ is $\mathcal{O}$ primitive.

The non-elementary condition in the above formulation is necessary. Clearly we must avoid finite subgroups of $\mathrm{SL}_2(\mathbb{Z})$ (and finite orbits $\mathcal{O}$ more generally) but Conjecture 1.3 can be false for cyclic toral subgroups. We discuss the difficulties connected with tori in Sect. 2.1 and explain the connection in the torus action case to Mersenne and Fibonacci primes. The methods of this paper don't apply to such torus actions and we need to avoid them. In fact, even the question of saturation is questionable for tori; see the discussion in Sect. 2.1. In a forthcoming paper [8] we will give a quantitative version of Conjecture 1.3, as well as some numerical evidence.

We turn to our general setting. The study of the pair $(\mathcal{O}, f)$ with $\mathcal{O} = b\Lambda$ reduces, by passing to the universal covering group, to the fundamental case that $G = \mathrm{Zcl}(\Lambda)$ is simply connected and the orbit is a subgroup of the group variety G. In this case we put forth a prescriptive Conjecture (or Hypothesis). We assume that $G \subset \mathrm{GL}_n$ is connected, simply connected and is absolutely almost simple and defined over $\mathbb{Q}$. The coordinate ring $\mathbb{Q}[G]$ is a unique factorization domain (see [19] and [49, Lemme 6.9]). The following is a generalization of Schinzel's Conjecture 1.1 above.

Conjecture 1.4 *Let $G \subset \mathrm{GL}_n$ be connected, simply connected, absolutely almost simple and defined over $\mathbb{Q}$. Let Λ be a subgroup of $G \cap \mathrm{GL}_n(\mathbb{Z})$ such that $G = \mathrm{Zcl}(\Lambda)$. Fix $f \in \mathbb{Q}[G]$ which is neither a unit nor zero and which factors into t irreducibles in $\mathbb{Q}[G]$. Let $\mathcal{O} = \Lambda$ (in the affine space of $n \times n$ matrices) and assume that $(\mathcal{O}, f)$ is primitive. Then $r_0(\mathcal{O}, f) = t$.*

As with Conjecture 1.1, this conjecture implies that if $f = f_1 f_2 \cdots f_t$ with f_j irreducible in $\mathbb{Q}[G]$ and integral on $\mathcal{O}$, then the set of $x \in \mathcal{O}$, $f_j(x)$ are all prime, is Zariski dense in G.

Conjecture 1.4 implies Conjecture 1.3 by a general pull back argument. The group $G = \mathrm{SL}_2 \subset \mathrm{Mat}_{2\times 2}$ is realized in the standard way with coordinates $x_{ij}, i, j = 1, 2$. If $\varphi : G \to \mathbb{A}^2$ is the morphism $\varphi(g) = (b_1, b_2)g$ then by composition φ^* maps $\mathbb{Q}[\mathbb{A}^2] \to \mathbb{Q}[G]$ and, in particular,

$$\varphi^*(x_1) = b_1 x_{11} + b_2 x_{21}, \qquad \varphi^*(x_2) = b_1 x_{12} + b_2 x_{22},$$

which are prime in $\mathbb{Q}[G]$, and one applies Conjecture 1.4.

In Sect. 2.2 we give examples which show that Conjecture 1.4 need not hold for G's which are not simply connected. The determination of the saturation number in such cases can be gotten by applying Conjecture 1.4 to the pull back of the data to the universal cover $\tilde{G}$. Also in Sect. 2.4 we investigate the local obstructions in Conjecture 1.4 and show that there is a $q = q(\mathcal{O})$ such that if the local condition is valid for $q(\mathcal{O})$ than it is valid for all $q \geq 2$.

We should clarify at this point that in the above conjecture, as well as elsewhere in this paper, by $f(x)$ being a prime number we mean that $f(x)$ generates a prime ideal in $\mathbb{Z}$ (i.e. $f(x) = \pm p$ where p is a positive prime). The reason for this is that in the several variable context we cannot restrict to $f(x) > 0$ since otherwise Conjecture 1.4 and the theorems below can be false. This is related to the negative solution of Hilbert's tenth problem and we explain this in Sect. 2.3.

As with Conjecture 1.2 some special cases of Conjecture 1.4 can be proven. For example, in [46] the following is proven using Vinogradov's methods. Let $n \geq 3$ and Λ finite index subgroup of $\mathrm{SL}_n(\mathbb{Z})$. The group Λ acts on $(\mathrm{Mat}_{n\times n}(\mathbb{Z}) \cong \mathbb{Z}^{n^2})$ by left multiplication. Fix $A \in \mathrm{Mat}_{n\times n}(\mathbb{Z})$ with $\det(A) = m \neq 0$ and let

$$V_m = A \cdot G = A \cdot \mathrm{SL}_n = \{X : X \in \mathrm{Mat}_{n\times n}, \det(X) = m\}.$$

Then Conjecture 1.4 is true for $\mathcal{O} = A \cdot \Lambda$ when f's are coordinate functions, $f_{ij}(X) = x_{ij}$ for $i, j = 1, 2, \ldots, n$. In particular if $\mathcal{O}^m$ consists of all integral matrices of determinant equal to m, then an analysis of the local congruence obstructions shows that the subset of $\mathcal{O}^m$ all of whose coordinates are prime, is Zariski dense in V_m iff $m \equiv 0(2^{n-1})$. Another instance of conjecture 1.4 was proven recently in [20]. Their Theorem 5 implies the conjecture for $\Lambda = \mathrm{SL}_2(\mathbb{Z})$, $\mathcal{O} = \Lambda$ in $\mathbb{A}^4$ via the standard realization of SL_2 and $f(x_1, x_2, x_3, x_4) = x_1^2 + x_2^2 + x_3^2 + x_4^2 - 2$. Some cases where Conjecture 1.4 is proven for Λ which are "thin" (see below for a definition) are given in Sect. 6 in connection with integral Apollonian packings.

We turn now to what we can prove, that being the $(\mathcal{O}, f)$ saturation in many cases. In the setting of Λ acting by translations and in particular Conjecture 1.1, it is well known that one can use the combinatorial sieve of Brun to prove that r_0 is finite. The bound for r_0 depends on the setting and much effort has gone into reducing this number in special cases [21, 27]. To prove $(\mathcal{O}, f)$ saturation we develop a combinatorial sieve in the setting of linear actions. To do so in this generality we need to make a further hypothesis (which as we discuss below can be established in many cases) about "congruence graphs" associated to Λ and $\mathcal{O}$. Let Λ and G be as in Conjecture 1.4. For $q \geq 1$ let $\Lambda(q)$ be the finite index "congruence" subgroup of Λ given as the kernel of the reduction of Λ mod q. The following conjecture is due to Lubotzky; in the special case that $G = \mathrm{SL}_2$ it has been popularized as his "1-2-3" question [42].

Conjecture 1.5 *Let $G \subset \mathrm{GL}_n$ and $\Lambda \subset \mathrm{GL}_n(\mathbb{Z}) \cap G$ with Λ Zariski dense in G, be as in Conjecture 1.4 and S be a finite symmetric set of generators of Λ. Then for q square-free the family of Cayley graphs $\mathcal{G}(\Lambda/\Lambda(q), S)$ is an expander family.*

See Sect. 4 and [31, 50] for definitions and properties of expanders.

We can now state our main saturation result. For simplicity we assume that $f = f_1 f_2 \cdots f_t$ with f_j irreducible in $\bar{\mathbb{Q}}[G]$.

Theorem 1.1 *Let Λ, G, f be as in Conjecture 1.4, $\mathcal{O} = \Lambda$, and, as always, assume that $(\mathcal{O}, f)$ is primitive. Then, assuming Conjecture 1.5 for Λ, it follows that $(\mathcal{O}, f)$ saturates. Moreover, the bound for $r_0(\mathcal{O}, f)$ is given explicitly and effectively in terms of the spectral gap in the expander family.*[3]

From Theorem 1.1 we can deduce $r_0(\mathcal{O}, f)$ finiteness when G is almost simple (but not necessarily simply-connected) as well as for orbits of such G's. We don't state this here as a general theorem because it still depends on Conjecture 1.5. This process is carried out in Sect. 6: see Theorem 6.2 for the cases where we have established Conjecture 1.5.

We turn to Conjecture 1.5. Progress on the general Ramanujan conjectures for $G(\mathbb{Q})\backslash G(\mathbb{A})$ (see [15, 51, 57]) establish Conjecture 1.5 when Λ is a congruence subgroup of $G(\mathbb{Q})$. When Λ is Zariski dense in G but of infinite index in $G(\mathbb{Z})$ it is apparently much more difficult to establish the conjecture as we cannot appeal to the theory of automorphic forms. We call this the "thin case". A large body of this paper is devoted to doing so for $G = \mathrm{SL}_2$ and Λ a thin subgroup of $\mathrm{SL}_2(\mathbb{Z})$ or a subgroup of $G(\mathbb{Z})$ where $G \subset \mathrm{GL}_n$ is a $\mathbb{Q}$-form

[3] The explicit bound for $r_0(\Lambda, f)$ is given in (3.51).

of SL_2. We expect these methods will eventually settle the general case of Conjecture 1.5.

Theorem 1.2 *Let Λ be a subgroup of $\mathrm{SL}_2(\mathbb{Z})$ which is Zariski dense in SL_2 and let S be a finite symmetric set of generators for Λ. Then for q square-free the family of Cayley graphs $\mathcal{G}(\Lambda/\Lambda(q), S)$ is an expander family.*[4]

In the recent paper [7] Theorem 1.2 is proven in the case that q is restricted to be prime, using Helfgott's result [30], which in turn builds crucially on sum-product theorem in $\mathbb{Z}/p\mathbb{Z}$ [9, 10]. For the application at hand it is crucial to allow q to be square-free and we need, among other things, the following sum-product theorem in $\mathbb{Z}/q\mathbb{Z}$, a result which is of independent interest.

Theorem 1.3 *Let $1 > \delta_2 \geq \delta_1 > 0$ be fixed. Let $q = \prod_{j=1}^{J} p_j$ be a product of distinct primes. Let $\pi_{q'}$ denote the projection $\mathbb{Z}/q\mathbb{Z} \to \mathbb{Z}/q'\mathbb{Z}$ for $q'|q$. Let $A \subset \mathbb{Z}/q\mathbb{Z}$ and assume that*

$$|A| < q^{1-\delta_1} \tag{1.1}$$

and

$$|\pi_{q_1}(A)| > q_1^{\delta_2} \quad \text{for all } q_1|q \text{ with } q_1 > q^{\eta}, \text{ where } \eta = \eta(\delta_1) = \frac{\delta_1}{3}. \tag{1.2}$$

Then

$$|A + A| + |A \cdot A| > q^{\delta_3}|A| \tag{1.3}$$

where $\delta_3 = \delta_3(\delta_1, \delta_2) > 0$.

The original sum-product theorem [10] establishes the above when $q = p$ is prime and $|A| > q^{\delta_1}$. The removal of this lower bound assumption for $q = p$ was established in [9], while when q is a product of a fixed number of large primes, Theorem 1.3 is proven in [6].

We end the introduction with a brief outline of the contents of the sections and the proofs. In Sect. 2 we examine various obstructions to the existence of points on an orbit for which $f = f_1 f_2 \cdots f_t$ is a product of t primes. The analysis of the local obstructions in the setting of Conjecture 1.4 makes use of a theorem of Matthews, Vaserstein and Weisfeiler [45] which asserts that for all but finitely many p, the projection of Λ on $G(\mathbb{Z}/p\mathbb{Z})$ is onto. In Sect. 3 we explain the fundamental lemma of the combinatorial sieve and the set up

[4] In fact we prove more (and this is crucial in our in our applications of Theorem 1.2), namely we show that $\mathcal{G}(\Lambda/\Lambda(q), S)$ form a family of absolute expanders (see Definition 4.1).

in our context of an orbit of Λ. Most of the work goes into verifying the axioms of the sieve. An interesting point here is that we do not probe the orbit in the usual way of ordering points by an Archimedean height (cf. [8, 37]). The reason for this being that in this context of thin orbits we don't know how to count the points asymptotically according to such an ordering. Instead, we order the points according to word length in generators of Λ (as is commonly done in combinatorial group theory). The resulting main terms in the sieving process are analyzed using the algebraic theorem in [45] mentioned above, coupled with more standard techniques from arithmetic geometry (specifically [40]). The expander property of the congruence graphs is used to control the remainder terms in the sieve and to establish a sufficiently strong form of level distribution. In the more familiar setting of sieving in $\mathbb{Z}$ (or $\mathbb{Z}^n$) the expander feature does not appear. In that setting the number of integer points in arithmetic progressions which are contained in a large interval, may be estimated accurately in the obvious way. However, when $\mathbb{Z}$ is replaced by, say, a free nonabelian group, the boundary of a big set is at least as large as the set, and a new ingredient is needed in order to give a suitably sharp estimate for the number of points of $\mathcal{O}$ in a large "ball". This ingredient is the expander property. In this connection we note that if Λ in Theorem 1.1 contains unipotent elements then one can approach the sieving problem in a more classical fashion. Using unipotent subgroups one produces nonconstant polynomial maps from $\mathbb{Z}$ into $\mathcal{O}$. In this way one can sieve in the familiar classical setting of $\mathbb{Z}$. If however Λ contains no unipotent elements, then, as far as we can see, our approach, and in particular expander property, is necessary.

The Zariski density of the points in Theorem 1.1 follows from the quantitative lower bound for the number of such points (when ordered combinatorially) that the fundamental lemma of the sieve, provides. This lemma also provides upper bounds in this ordering and these yield sharp (up to a multiplicative constant) upper bounds for the number of points in $\mathcal{O}$ for which $f_1, \ldots, f_t$ are prime.

Sections 4 and 5 are devoted to proving Theorems 1.2 and 1.3 respectively. The proof of the expander property follows the method in [24, 54] which is based on an upper bound on the number of closed cycles combined with exploiting the large dimensionality of a nontrivial irreducible representation of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ (the latter is due to Frobenius [22]). The extension of the required multiplicity bound in $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ is straightforward, proceeding inductively on the number of prime factors of q. The problem then reduces to giving a sharp upper bound for the number of closed walks of length l (for l is a suitable range) in the graphs $\mathcal{G}(\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z}), S)$. As in [7] this is achieved by an l^2-flattening lemma (Proposition 4.1) of Sect. 4. The proof of this proposition makes use of various results from additive combinatorics and in partic-

ular a noncommutative version of Balog-Szemeredi-Gowers Lemma, due to Tao [59].

An important input in [7] is Helfgott's result [30] asserting that subsets of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ grow under multiplication. His proof makes use of the sum-product theorem [10] for $\mathbb{Z}/p\mathbb{Z}$. Both of these need to be extended to $\mathbb{Z}/q\mathbb{Z}$ and this turns out to be quite involved. Proposition 4.3 of Sect. 4 is the appropriate extension of [30] to $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$, while Theorem 1.3 is the $\mathbb{Z}/q\mathbb{Z}$ sum-product theorem. The proof of Theorem 1.3, given in Sect. 5, can be read independently of the rest of the paper. It uses the techniques and results developed in the proofs of the special cases of the theorem [6, 9, 10], as well as the analytic tools for general modulus exponential sums which were developed in [5].

In the final Sect. 6 of the paper we give explicit examples of applications of Theorem 1.1, in particular unconditional ones coming under the purview of Theorem 1.2. Theorem 6.2 establishes saturation for a class of $(\mathcal{O}, f)$'s, while example A shows that $r_0(\mathcal{O}, f) < \infty$ for the pair in Conjecture 1.3; example B concerns orbits of orthogonal groups in 3-variables and example C deals with the cone of Pythagorean triples. In example D we apply our theory to integral Apollonian packings which are governed by a thin subgroup of an orthogonal group in four variables.

Finally we note that if the group Λ is a congruence subgroup (that is the non-thin case) one can develop the affine linear sieve of this paper in a much sharper quantitative fashion by appealing to some advanced results in automorphic forms. This is carried out in [46] and [41] and the bounds for $r_0(\mathcal{O}, f)$ that are established are comparable in quality to those of the classical one-variable sieve [16, 27]. For a leisurely overview of these sieving problems see [53].

2 Algebraic preliminaries

In this section, which prepares the way for the sieve analysis in Sect. 3, we collect some algebraic tools and discuss some diophantine obstructions to producing primes on orbits.

2.1

We begin with the difficulties connected with a torus. To demonstrate this in $\mathbb{A}^1$ consider the ring $R = \mathbb{Z}[\frac{1}{2}, \frac{1}{3}]$. It is a unique factorization domain and has a unit group U consisting of numbers of the form $\pm 2^a 3^b$ with $a, b \in \mathbb{Z}$. The prime ideals are $Rp = (p)$ with p a prime $p \neq 2$ or 3. Let Λ be the subgroup of $\mathrm{GL}_1(R)$ generated by 4, i.e. $\Lambda = \{4^m : m \in \mathbb{Z}\}$. Λ is Zariski dense in GL_1 and the polynomial $f(x) = x - 1 \in \mathbb{R}[x]$ is irreducible. Since

$f(4) = 3$ which is in $(R/qR)^*$ for all ideals qR of R, there are no local congruence obstructions to making $f(x)$ prime in R and (Λ, f) is primitive. However $f(x)$ can be a prime in R for at most a finite number of x in Λ since $(4^n - 1) = (2^n - 1)(2^n + 1)$ and $2^n \pm 1 = \pm 2^a 3^b$ has only finitely many solutions in n, a, b (this is elementary but more generally it follows from the finiteness to the S-unit equation; see [1], Chap. 5). Thus the local to global principle in Conjecture 1.4 fails for this multiplicative group. The reason being that Λ is too thin in that it consists of squares. One can try to remedy this by taking for Λ the bigger group $\langle 2 \rangle$. The question of whether $\langle 2 \rangle$ contains a Zariski dense set of points (i.e. infinitely many in this case) with $f(x)$ a prime in R, is the well known Conjecture of Mersenne: that $2^p - 1$ is prime for infinitely many primes p.[5]

However these and more general tori probably present much more serious problems even as far as saturation goes. Consider the torus A in SL_2 given as follows. Let

$$\Lambda = \left\{ \begin{bmatrix} 3 & 1 \\ -1 & 0 \end{bmatrix}^m : m \in \mathbb{Z} \right\} \subset \mathrm{SL}_2(\mathbb{Z}).$$

The group Λ is infinite cyclic; $\mathrm{Zcl}(\Lambda) = A$ is a torus and if $\mathcal{O} = (1, 0) \cdot \Lambda$ then $\mathrm{Zcl}(\mathcal{O})$ in $\mathbb{A}^2$ is the hyperbola $\{(x_1, x_2) : x_1^2 - 3x_1x_2 + x_2^2 = 1\}$. The orbit consists of pairs (F_{2n}, F_{2n-2}) with $n \in \mathbb{Z}$ where F_n is the n-th Fibonacci number. As with the previous example of a torus, this sequence is too sparse both from an algebraic and an analytic point of view to execute any kind of sieve to establish saturation. In fact, while it is conjectured that F_n is prime for infinitely many n, as pointed out to us by Lagarias, standard heuristics suggest a very different behavior for F_{2n}. We have $F_{2n} = F_n L_n$ where L_n is the n-th Lucas number and assuming a probabilistic model for the number of prime factors of a large integer in terms of its size and that F_n and L_n are independent, leads to the conjecture that F_{2n} has an unbounded number of prime factors as $n \to \infty$. A precise conjecture along these lines is put forward in [12] (see Conjecture 5.1). In our language this asserts that if $\mathcal{O}$ is as above and $f(x_1, x_2) = x_1$ then $r_0(\mathcal{O}, f) = \infty$. It would be very interesting to produce an example of a pair $(\mathcal{O}, f)$ for which one can prove that $r_0(\mathcal{O}, f)$ is infinite. In view of this and also in terms of the setting in which our methods apply, we must keep away from tori which occur in $\mathrm{rad}(\mathrm{Zcl}(\Lambda))$.

2.2

The prescriptive local to global Conjecture 1.4 also fails for semisimple groups which are not simply connected (of course r_0 should be finite in these

[5] Some things can be said about high divisibility of $2^n - 1$ for most n, as well as for similar questions about the denominators of rational points on elliptic curves, see Sect. 10 in [38].

cases). Consider the special orthogonal group $G = \mathrm{SO}_F$ where

$$F(x_1, x_2, x_3) = x_1 x_3 - x_2^2. \tag{2.1}$$

G is contained in GL_3, it is simple, and over $\mathbb{Q}$ it is given by the equations in 9-dimensional space

$$\begin{cases} X^t A X = A, \\ \det X = 1, \end{cases} \tag{2.2}$$

where

$$A = \begin{bmatrix} 0 & 0 & 1/2 \\ 0 & -1 & 0 \\ 1/2 & 0 & 0 \end{bmatrix}.$$

Let $\mathbb{Q}[G]$ be the corresponding coordinate ring. The simply connected double cover $\tilde{G}$ of G is SL_2. This is realized explicitly by the group homomorphism π of GL_2 onto G:

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \to \frac{1}{(\alpha\delta - \beta\gamma)} \begin{bmatrix} \alpha^2 & 2\alpha\gamma & \gamma^2 \\ \alpha\beta & \alpha\delta + \beta\gamma & \gamma\delta \\ \beta^2 & 2\beta\delta & \delta^2 \end{bmatrix}. \tag{2.3}$$

The homomorphism π restricts to a morphism of SL_2 onto G (as group varieties over $\mathbb{Q}$) with kernel $\pm I$. It is classical (see [13, pp. 301–302]) that $\pi(\mathrm{GL}_2(\mathbb{Z})) = G(\mathbb{Z})$ while $\Lambda = \pi(\mathrm{SL}_2(\mathbb{Z}))$ is of index 2 in $G(\mathbb{Z})$. The polynomial $f(x) = x_{11} - 1$ is prime in $\mathbb{Q}[G]$ and there are no local obstructions to $f(x)$ being prime on Λ. However since x_{11} is a square when $x \in \Lambda$ we see that $f(x)$ is prime only if $x_{11} = 3$. The source of the difficulty here is that G is not simply connected, and, in particular, $G(\mathbb{Z})$ fails to satisfy strong approximation, that is $G(\mathbb{Z}) \to G(\mathbb{Z}/p\mathbb{Z})$ is not onto for half of the primes. Thus Conjecture 1.4 is false for $G = \mathrm{SO}_F$. However unlike the torus case this is not a serious issue, at least in terms of understanding r_0. Let $f \in \mathbb{Q}[x_{ij}]$ for which we seek to understand $r_0(\Gamma, f)$, with $\Gamma \subset G(\mathbb{Z})$ and $\mathrm{Zcl}(\Gamma) = G$. The morphism π from $\tilde{G}$ onto G induces by composition an injective ring homomorphism $\pi^* : \mathbb{Q}[G] \to \mathbb{Q}[\tilde{G}]$. Thus it suffices to examine π^* and its values on the group $\Lambda = \pi^{-1}(\Gamma)$ in $\tilde{G}$. The factorization of $\pi^*(f)$ in $\mathbb{Q}[\tilde{G}]$ is in this way the critical issue. This reduces the study of the group variety G or, more generally an orbit $V = b \cdot G$ of G, to understanding the simply connected setting. Thus Conjecture 1.4 is the central one. This strategy is pursued in Sect. 6 where we establish the almost prime theorem for non simply connected cases as well as for orbits thereof by invoking Theorem 1.1.

2.3

We pointed out in the introduction that when looking for primes or almost primes $f(x)$, we cannot insist that $f(x)$ be positive because of difficulties associated with the negative solution of Hilbert's 10-th problem. In several variables the condition $f(x) > 0$, $f \in \mathbb{Z}[x_1, x_2, \ldots, x_n]$ can encode the general diophantine equation (for example if $f(x) = 1 - g^2(x)$ then $f(x) > 0$ is equivalent to $g(x) = 0$). The work of Matiyasevich et al. [44] on Hilbert's 10-th problem shows that given any recursively enumerable subset S of the positive integers $\mathbb{N}$ there is an $f \in \mathbb{Z}[x_0, x_1, \ldots, x_n]$ (one can take $n = 10$), such that S is exactly the set of $t \in \mathbb{N}$ for which $f(t, x_1, \ldots, x_n) = 0$ has a solution $x_1, x_2, \ldots, x_n \in \mathbb{Z}$. From this it is not difficult to construct a $g \in \mathbb{Z}[x_1, x_2, \ldots, x_n]$ such that the set of positive values assumed by g is exactly S. Now suppose that our orbit $\mathcal{O}$ is all of $\mathbb{Z}^n$ (say $\Gamma = \mathbb{Z}^n$ acting by translations). We can choose S so as to make $g(x)$ behave very singularly as far as its positive values. For example, let S consists of the numbers in the sequence

$$a_m = \prod_{m<p<2m} p,$$

for $m > 2$. Then we have

(i) S is recursively enumerable.
(ii) There are no local obstructions to making the corresponding $g(x)$ a prime.
(iii) For any r the set of $x \in \mathbb{Z}^n$ such that $g(x) > 0$ and is a product of at most r primes lies in a finite union of closed sets of the form $A_a = \{x : g(x) = a\}$. Hence this set is not Zariski dense in $\mathrm{Zcl}\{x \in \mathbb{Z}^n : g(x) > 0\}$.

Thus the pair $(\mathbb{Z}^n, g)$ does not saturate when restricted to values of $g(x)$ which are positive. Without the positivity condition the pair $(\mathbb{Z}^n, g)$ saturates by a simple version of Theorem 1.1.

At the other extreme of this phenomenon of positive values is the well known example of S being the subset of $\mathbb{N}$ consisting of the prime numbers. A corresponding explicit g of degree 25 in 26 variables is given in [34]. In this case the set of positive values assumed by g is exactly all primes, and from our point of view this is too many primes. The combinatorial sieve is based on the equidistribution of points in the orbit mod q, for any q and clearly restricted to $g(x) > 0$; this is far from true here.

2.4

We turn to the main setting of the paper. $G \subset \mathrm{GL}_n$ is a connected, simply connected absolutely almost simple group defined over $\mathbb{Q}$. Λ is a subgroup of

$\mathrm{GL}_n(\mathbb{Z})$ for which $\mathrm{Zcl}(\Lambda) = G$. For $d \geq 1$ an integer we denote by Λ_d the image in $\mathrm{GL}_n(\mathbb{Z}/d\mathbb{Z})$ of the reduction of Λ modulo d. Let $\Lambda(d)$ be the kernel of this reduction so that $\Lambda/\Lambda(d) \cong \Lambda_d$. For $(d_1, d_2) = 1$, Λ reduces diagonally into a subgroup of $\Lambda_{d_1} \times \Lambda_{d_2}$ and we need to know the extent to which this is a surjection or, at least, is a product. By Noether's theorem [47], outside a finite set $S = S(G)$ of primes the reduction of G mod p is an (absolutely) irreducible variety over $\mathbb{F}_p \cong \mathbb{Z}/p\mathbb{Z}$ and we denote the corresponding $\mathbb{F}_p$ points by $G(\mathbb{F}_p)$, $p \notin S$. The key stabilization property that is needed for sieving is the following, which is due to Matthews et al. [45].

Theorem 2.1 *Given an integer M there is $q_1 = q_1(\Lambda, M)$ containing the primes in S and also $M|q_1$, such that*

(i) *For p a prime, $p \nmid q_1$*

$$\Lambda_p = G(\mathbb{F}_p).$$

(ii) *For $d = p_1 p_2 \cdots p_l$ square-free and $(d, q_1) = 1$, the diagonal reduction*

$$\Lambda \to \Lambda_d \to \Lambda_{p_1} \times \Lambda_{p_2} \times \cdots \times \Lambda_{p_l}$$

is surjective.

(iii) *For $(d, q_1) = 1$ square-free*

$$\Lambda \to \Lambda_{q_1} \times \Lambda_d$$

is surjective.

Remark While in this paper we use Theorem 2.1 for square free moduli and for convenience state it for such, it is valid for arbitrary moduli.

Proof Parts (i) and (ii) are proved in [45] with a suitable q_0 (in place of q_1) depending on Λ. Their proof makes use of the classification of finite simple groups. The treatment of this theorem in [48] does not make use of this classification. To see that part (iii) is true, choose q_1 with $q_0|q_1$ and $M|q_1$, and so that the following holds: the center Z of G is finite and for p large enough $G(\mathbb{F}_p)/Z(\mathbb{F}_p)$ are distinct finite simple groups. So we can clearly arrange for a large enough q_1 so that Λ_{q_1} has no composition factors in common with $G(\mathbb{Z}/q\mathbb{Z})$ with q square free and $(q, q_1) = 1$. Now if $(d, q_1) = 1$ and d is square free then the image of Λ in $\Lambda_{q_1} \times \Lambda_d$ surjects onto each factor and hence by Goursat's Lemma (see [39, p. 75]) and the above remarks, the image surjects onto the product. □

Theorem 2.1 says that a Λ which is Zariski dense in G can be deficient at only a finite number of primes (if it is thin then it is automatically deficient at

infinity). To sieve out all primes we need a little more in terms of projections onto products. For this we pass to a finite index subgroup and the following Proposition follows from Theorem 2.1.

Proposition 2.1 *Let Λ, G and $M = N^2$ be as in Theorem 2.1 and $q_1 = q_1(\Lambda, M)$ be the resulting integer in the Theorem. Let $\Gamma = \Lambda(q_1)$ be the corresponding principal congruence subgroup of Λ. Then for $d = d_1 d_2$ of the form $N^\beta t$ with $\beta = 0$ or 1, t square free and $(d_1, d_2) = 1$ we have that $\Gamma \to \Gamma_{d_1} \times \Gamma_{d_2}$ is surjective and $\Gamma_p = \Lambda_p$ for $p \nmid q_1$.*

Next we discuss the primitivity condition in this context. Let $f \in \mathbb{Q}[G]$, f integral on $\mathcal{O} = \Lambda$. We can write $f = g/N$ where $g \in \mathbb{Z}[G]$ and $N \geq 1$, $N | \gcd(g(\mathcal{O}))$. Since we are assuming that f is primitive, we have that $\gcd(f(\mathcal{O})) = 1$ (which we call weakly primitive), and hence $N = \gcd(g(\Lambda))$. Note that if our given f is not weakly primitive then $f/\gcd(f(\mathcal{O}))$ is, and it is clear that weak primitivity is easily checked and involves only finitely many congruences. Concerning primitivity we have

Proposition 2.2 *With the above notations $(\mathcal{O}, f)$ is primitive iff there is a $\xi \in \mathcal{O}$ such that $(f(\xi), q_1) = 1$ where $q_1 = q_1(\Lambda, N^2)$ as in Theorem 2.1.*

Proof By definition of primitive the condition is satisfied with $d = q_1$. To prove the converse, let $d \geq 1$; we seek an $x \in \mathcal{O}$ such that $(f(x), d) = 1$. We may assume that d is square-free and that $d = d_1 d_2$ with $d_1 | q_1$ and that $(d_2, q_1) = 1$. Consider the orbit $\mathcal{O}' = \xi\Gamma$ where ξ is given in Proposition 2.2 for q_1 and $\Gamma = \Lambda(q_1)_\xi$ as in Proposition 2.1. By this proposition

$$\Gamma \to \Gamma_{q_1} \times \Gamma_{p_1} \times \cdots \times \Gamma_{p_\nu} \tag{2.4}$$

is onto, where $p_1 p_2 \cdots p_\nu$ is the prime factorization of d_2. For each p_j there is a $y_j \in \mathcal{O}$ such that $p_j \nmid f(y_j)$ since $(\mathcal{O}, f)$ is weakly primitive. Hence, using (2.4) and $\Gamma_{p_j} = \Lambda_{p_j}$ we can find a $\gamma \in \Gamma$ such that

$$\gamma\xi \equiv y_j \bmod p_j.$$

Hence $f(\gamma\xi) \equiv f(y_j) \neq 0 \bmod p_j$. Also

$$Nf(\gamma\xi) = g(\gamma\xi) \equiv g(\xi) \bmod q_1$$

since $\gamma \equiv 1 \bmod q_1$. Hence $Nf(\gamma\xi) \equiv Nf(\xi) \bmod q_1$, and therefore $f(\gamma\xi) \equiv f(\xi) \bmod q_1/N$. But $d_1 | q_1$, d_1 is square free and $N^2 | q_1$, hence

$$f(\gamma\xi) \equiv f(\xi) \bmod d_1.$$

Therefore if $x = \gamma\xi$ then $(f(x), d_1 d_2) = 1$ as needed. □

To end this section we give a simple example in this setting of G simply connected and simple and a pair (Λ, f) which is weakly primitive but not primitive. Let

$$\Lambda = \left\langle \begin{pmatrix} 2 & 1 \\ 15 & 8 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \right\rangle \leq \mathrm{SL}_2(\mathbb{Z}).$$

Then $\mathrm{Zcl}(\Lambda) = \mathrm{SL}_2$ and if

$$f(x_{11}, x_{12}, x_{21}, x_{22}) = (x_{11} - 29)(x_{11} - 11),$$

note that $15 | q_1(\Lambda)$. One checks that (Λ, f) is weakly primitive but that for any $x \in \Lambda$, $f(x)$ is 3 or -5 mod 15. Thus f is not primitive. The problem of course is that $\Lambda \to \Lambda_3 \times \Lambda_5$ is not a product.

3 Sieving: proof of Theorem 1.1

3.1 Combinatorial sieve

We will make use of the simplest combinatorial sieve which is turn is based on the Fundamental Lemma in the theory of elementary sieve, see [33] and [27]. Our formulation is tailored for the applications below.

Let A denote a finite sequence a_n, $n \geq 1$ of nonnegative numbers. Denote by X the sum

$$\sum_n a_n = X. \tag{3.1}$$

X will be large, in fact tending to infinity. For a fixed finite set of primes B let z be a large parameter (in our applications z will be a small power of X and B will usually be empty). Let

$$P = P_z = \prod_{\substack{p \leq z \\ p \notin B}} p. \tag{3.2}$$

Under suitable assumptions about sums of A over n's in progressions with moderate-size moduli d, the sieve gives upper and lower estimates which are of the same order of magnitude for sums of A over the n's which remain after sifting out numbers with prime factors in P.

More precisely, let

$$S(A, P) := \sum_{(n,P)=1} a_n. \tag{3.3}$$

The assumptions on sums in progressions are as follows:

(A_0) For d square-free, and having no prime factors in B ($d < X$), we assume that the sums over multiples of d take the form

$$\sum_{n \equiv 0(d)} a_n = \beta(d)X + r(A, d), \tag{3.4}$$

where $\beta(d)$ is a multiplicative function of d and

$$\text{for } p \notin B, \quad \beta(d) \leq 1 - \frac{1}{c_1} \quad \text{for a fixed } c_1.$$

The understanding being that $\beta(d)X$ is the main term and that the remainder $r(A, d)$ is smaller, at least on average (see the next axiom).

(A_1) A has level distribution $D = D(X)$, $(D < X)$ that is

$$\sum_{d \leq D} |r(d, A)| \ll X^{1-\varepsilon_0} \quad \text{for some } \varepsilon_0 > 0.$$

(A_2) A has sieve dimension $t > 0$, that is for a fixed c_2 we have

$$\left| \sum_{\substack{w \leq p \leq z \\ p \notin B}} \beta(p) \log p - t \log \frac{z}{w} \right| \leq c_2$$

for $2 \leq w \leq z$.

In terms of these conditions (A_0), (A_1), (A_2) the elementary combinatorial sieve yields:

Theorem 3.1 *Assume* (A_0), (A_1) *and* (A_2) *for* $s > 9t$ *and* $z = D^{1/s}$ *and* X *large we have*

$$\frac{X}{(\log X)^t} \ll S(A, P_z) \ll \frac{X}{(\log X)^t}. \tag{3.5}$$

The implied constants depend explicitly on $t, \varepsilon_0, c_1, c_2$.

3.2 Arithmetic and geometry of the orbit

We review the setting in Theorem 1.1. G is a connected, simply connected semisimple matrix group in GL_n which is defined over $\mathbb{Q}$. f is a nonunit in the coordinate ring $\mathbb{Q}[G]$. We are assuming further that in this unique factorization domain f factors as $f_1 f_2 \cdots f_t$ with f_j irreducible in $\bar{\mathbb{Q}}[G]$. Hence the varieties G and

$$W_k = G \cap \{x : f_k(x) = 0\} \quad \text{for } k = 1, \ldots, t$$

are defined over $\mathbb{Q}$ and are absolutely irreducible. For our purposes of sieving we will assume further (without loss of generality) that the f_j's are distinct in $\mathbb{Q}[G]$. In particular, since the W's are connected, $\dim(W_i \cap W_j) < \dim(G) - 2$ for $i \neq j$, while $\dim(W_j) = \dim(G) - 1$.

Consider now the reduction of G and W_j modulo p for p a large prime. According to Noether's Theorem [47] for p outside a set $S_1 = S_1(G, f)$ these reduce to absolutely irreducible varieties G and W_k is defined over $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$.

By the Lang-Weil Theorem ([40], see also [56] for an elementary treatment) we have that for $p \notin S_1$

$$\begin{cases} |G(\mathbb{F}_p)| = p^{\dim G} + O(p^{\dim G - \frac{1}{2}}), \\ |W_k(\mathbb{F}_p)| = p^{\dim G - 1} + O(p^{\dim G - \frac{3}{2}}), \\ |W_k \cap W_l(\mathbb{F}_p)| \ll p^{\dim G - 2} \text{ if } k \neq l, \end{cases} \tag{3.6}$$

where the implied constants depend on G and f.

Recall that $\Lambda \subset \mathrm{GL}_n(\mathbb{Z})$ and $\mathrm{Zcl}(\Lambda) = G$. Our sieve will be carried out on the orbit (in this case a coset) of a subgroup Γ of Λ. For the purposes of counting on the orbit it is convenient to work with a free group. By Tits Theorem [61] there is a subgroup L of Λ which is free on two generators and is Zariski dense in G; clearly Theorem 1.1 follows from Theorem 1.1 for L so it is sufficient to establish Theorem 1.1 for L. Applying Proposition 2.1 with L (in place of Λ) and $M = N^2$ where $f = g/N$ as in Proposition 2.1, we arrive at the subgroup $\Gamma = L(q_1)$ ($q_1 = q_1(L, N^2)$ and $N^2 | q_1$) which satisfies

$$\Gamma \text{ is Zariski dense in } G; \tag{3.7a}$$

$$\Gamma \text{ is free on } k \geq 2 \text{ generators}; \tag{3.7b}$$

$$\begin{aligned} &\text{outside a finite set of primes } S_2 = S_2(\Gamma) \\ &\text{we have } \Gamma_p = \Lambda_p \cong G(\mathbb{F}_p); \end{aligned} \tag{3.7c}$$

$$\begin{aligned} &\Gamma \to \Gamma_{d_1} \times \Gamma_{d_2} \text{ is surjective for } (d_1, d_2) = 1 \\ &\text{and } d_1 d_2 = N^\beta t \text{ with } \beta = 0 \text{ or } 1 \text{ and } t \text{ square free.} \end{aligned} \tag{3.7d}$$

Now since (Λ, f) is primitive, given

$$\nu = q_1(L, N^2) \prod_{p \in S_1} p \prod_{p \in S_2} p, \tag{3.8}$$

we can find $x \in \Lambda$ such that

$$(f(x), \nu) = 1. \tag{3.9}$$

Let $\mathcal{O} = x\Gamma \subset \mathrm{GL}_n(\mathbb{Z})$. We will sieve on the orbit $\mathcal{O}$. For $d \geq 1$ denote by $\mathcal{O}_d$ the reduction of $\mathcal{O}$ in $\mathrm{GL}_n(\mathbb{Z}/d\mathbb{Z})$. Clearly

$$\mathcal{O}_d = x\Gamma_d \quad (\text{in } \mathrm{GL}_n(\mathbb{Z}/d\mathbb{Z})). \tag{3.10}$$

Also

$$|\mathcal{O}_d| = |\Gamma_d|, \tag{3.11}$$

since the stabilizer of x in Γ_d is trivial (since $\det x = 1$).

From (3.10) and (3.7d) it follows that $\mathcal{O}_d$ inherits the product structure. That is, for $(d_1, d_2) = 1$ and $d = d_1 d_2 = N^\beta t$ as in (3.7)

$$\mathcal{O} \to \mathcal{O}_d \to \mathcal{O}_{d_1} \times \mathcal{O}_{d_2} \quad \text{is surjective.} \tag{3.12}$$

For our given $g \in \mathbb{Z}[G]$ where $f = g/N$ let

$$\mathcal{O}_d^{(g)} = \{x \in \mathcal{O}_d : g(x) \equiv 0 \bmod d\}. \tag{3.13}$$

These sets are well-defined and by the ordinary Chinese remainder theorem we have that

$$\mathcal{O}_d^{(g)} \to \mathcal{O}_{d_1}^{(g)} \times \mathcal{O}_{d_2}^{(g)} \tag{3.14}$$

is a bijection for d_1, d_2 as in (3.12). Finally, since $g(x) \equiv 0 \bmod N$ for $x \in \mathcal{O}$ we note that

$$\mathcal{O}_N^{(g)} = \mathcal{O}_N. \tag{3.15}$$

3.3 Sieving on an orbit

Continuing with the notation and setup of the previous two sections we have $\mathcal{O} = x\Gamma$ where Γ is a free group on k generators ($k \geq 2$) which we denote by $A_1, \ldots, A_k$. Since Γ acts simply transitively on $\mathcal{O}$ we can identify Γ and $\mathcal{O}$. In this way we turn $\mathcal{O}$ into a $2k$ regular tree by joining y in $\mathcal{O}$ to $y \cdot A_j$ and $y \cdot A_j^{-1}$ for $j = 1, 2, \ldots, k$. Another way of saying this is that $\mathcal{O}$ is identified with the Cayley graph of Γ with respect to the generating set $S = \{A_1, A_1^{-1}, \ldots, A_k, A_k^{-1}\}$. For $x, y \in \mathcal{O}$ let $w(x, y)$ denote the distance in the tree from x to y. The key nonnegative sequence a_n to which we apply the combinatorial sieve in Sect. 3.1 is defined as follows: for $n \geq 0$ and $L \geq 0$ let

$$a_n(L) = \#\{y \in \mathcal{O} : w(y, x) \leq L, |f(y)| = n\}. \tag{3.16}$$

Let $r = 2k - 1$. It is elementary that the number of points on a $2k$-regular tree whose distance to a given vertex is at most L is equal to

$$1 + (r+1)\sum_{i=1}^{L} r^{i-1} = \frac{(r+1)r^L - 2}{r-1}.$$

Hence

$$X := \sum_n a_n(L) = \sum_{\substack{y\in\mathcal{O}\\ w(y,x)\le L}} 1 = \frac{(r+1)r^L - 2}{r-1}. \tag{3.17}$$

We need to study the sums of $a_n(L)$ for n in progressions. For $d \ge 1$ we have

$$\sum_{n\equiv 0(d)} a_n(L) = \sum_{\substack{y\in\mathcal{O}\\ w(y,x)\le L f(y)\equiv 0(d)}} 1. \tag{3.18}$$

Clearly we have

$$\sum_{n\equiv 0(d)} a_n(L) = \sum_{\substack{y\in\mathcal{O}\\ w(y,x)\le L g(y)\equiv 0(Nd)}} 1 = \sum_{\rho\in\mathcal{O}_{Nd}^{(g)}} \sum_{\substack{\delta\in\Gamma(Nd)\\ w(\rho'\delta,x)\le L}} 1, \tag{3.19}$$

where $\rho' \in \mathcal{O}$ is any point in $\mathcal{O}$ which reduces to ρ in $\mathcal{O}_{Nd}$.

To analyze the inner sum in (3.19) we make use of the $2k$-regular quotient graphs $\mathcal{G}_{Nd} = \mathcal{O}/\Gamma(Nd)$. The size of this graph is $|\mathcal{O}_{Nd}| = |\Gamma_{Nd}|$ which we denote by F. Let $\varphi_0, \dots, \varphi_{F-1}$ be an orthonormal basis of $\Gamma(Nd)$-periodic functions on $\mathcal{O}$ (i.e. functions ϕ satisfying $\varphi(y\gamma) = \varphi(y)$ for $\gamma \in \Gamma(Nd)$) which are eigenfunctions of the discrete "Laplacian" Δ

$$\Delta\varphi(y) = \sum_{\eta\sim y} \varphi(\eta). \tag{3.20}$$

Denote by λ_j the eigenvalue of φ_j;

$$\Delta\varphi_j = \lambda_j\varphi_j. \tag{3.21}$$

The indices are chosen so that

$$\lambda_0 = 2k \quad \text{and} \quad \varphi_0(y) = \frac{1}{\sqrt{F}}. \tag{3.22}$$

The graph $\mathcal{G}_{Nd}$ is isomorphic to the Cayley graph $\mathcal{G}(\Gamma/\Gamma(Nd), S) = \mathcal{G}(\Gamma_{Nd}, S)$. The assumption about these that is made in Theorem 1.1 is that they are a family of absolute expanders. That is, any eigenvalue λ of $\mathcal{G}_{Nd}$ with $|\lambda| \ne 2k$ satisfies

$$|\lambda| \le \kappa, \quad \text{with } \kappa < 2k \text{ independent of } d. \tag{3.23}$$

This is our key analytic input into controlling the level distribution in the sieve. The smaller κ the better this level, and we keep track of the dependence on κ in the ensuing estimates.

Using the basis φ_j we can expand the function in the inner sum in (3.19) in the following form (see [43]): for $y, \eta \in \mathcal{O}$

$$\sum_{\substack{w(y\delta,\eta)\le L\\ \delta\in\Gamma(Nd)}} 1 = \sum_{j=0}^{F-1} P_L\left(\frac{\lambda_j}{2\sqrt{r}}\right)\varphi_j(y)\varphi_j(\eta), \tag{3.24}$$

where P_L is the degree L polynomial

$$P_L(\cos\theta) = r^{L/2}\left(\frac{\sin(L+1)\theta}{\sin\theta} + \frac{\sin L\theta}{\sqrt{r}\sin\theta}\right). \tag{3.25}$$

In particular

$$P_L\left(\frac{\lambda_0}{2\sqrt{r}}\right) = P_L\left(\frac{r+1}{2\sqrt{r}}\right) = \frac{r^{L+1}-1}{r-1} + \frac{r^L-1}{r-1} = X. \tag{3.26}$$

Thus the contribution from $j = 0$ to (3.24) is (using (3.22) and (3.11))

$$\frac{X}{|F|} = \frac{X}{|\Gamma_{Nd}|} = \frac{X}{|\mathcal{O}_{Nd}|}. \tag{3.27}$$

For $j \neq 0$, $|\lambda_j| \le \kappa$ and hence

$$\left|P_L\left(\frac{\lambda_j}{2\sqrt{r}}\right)\right| \le \left(\frac{\kappa+\sqrt{\kappa^2-4r}}{2}\right)^L \ll X^\tau, \tag{3.28}$$

where

$$\tau = \frac{\log(\frac{\kappa+\sqrt{\kappa^2-4r}}{2})}{\log r} < 1. \tag{3.29}$$

Also $\sum_{j=0}^{F-1}|\varphi_j(y)|^2$ is independent of y since Γ acts isomorphically and transitively on $\mathcal{G}_{Nd}$. This coupled with φ_j being an orthonormal basis of $L^2(\mathcal{G}_{Nd})$ gives

$$\sum_{j=0}^{F-1}|\varphi_j(y)|^2 = 1. \tag{3.30}$$

Hence, uniformly for $y, \eta \in \mathcal{O}$, we have

$$\sum_{j=0}^{F-1} P_L\left(\frac{\lambda_j}{2\sqrt{r}}\right)\varphi_j(y)\varphi_j(\eta) = \frac{X}{|\mathcal{O}_{Nd}|} + O(X^\tau). \tag{3.31}$$

Hence

$$\sum_{w(y\delta,\eta)\le L\delta\in\Gamma(Nd)} 1 = \frac{X}{|\mathcal{O}_{Nd}|} + O(X^{\tau}). \tag{3.32}$$

Substituting this in (3.19) yields

$$\sum_{\substack{y\in\mathcal{O}\\ w(y,x)\le L\\ g(y)\equiv 0(d)}} 1 = \sum_{\rho\in\mathcal{O}_{Nd}^{(g)}} \left(\frac{X}{|\mathcal{O}_{Nd}|} + O(X^{\tau})\right) = \frac{|\mathcal{O}_{Nd}^{(g)}|}{|\mathcal{O}_{Nd}|}X + O\big(|\mathcal{O}_{Nd}^{(g)}|X^{\tau}\big). \tag{3.33}$$

We have therefore shown that for $a_n(L)$ as in (3.16) we have

$$\sum_{n\equiv 0(d)} a_n(L) = \beta(d)X + r(d,A), \tag{3.34}$$

where

$$\beta(d) = \frac{|\mathcal{O}_{Nd}^{(g)}|}{|\mathcal{O}_{Nd}|} \tag{3.35}$$

and

$$|r(d,A)| \ll |\mathcal{O}_{Nd}^{(g)}|X^{\tau}. \tag{3.36}$$

The following proposition verifies (A_0) of Sect. 3.1 (with B the empty set).

Proposition 3.1 *For d square free $\beta(d)$ is multiplicative and there is $c_1 > 0$ fixed (depending only on Γ and N) such that $\beta(p) \le 1 - \frac{1}{c_1}$ for all p.*

Proof Let $(d_1,d_2) = 1$ and d_1, d_2 square free. Write $N = N_1N_2$ with $(N_1,d_2) = 1$, $(N_2,d_1) = 1$ and $(N_1,N_2) = 1$. The product structure (3.14) for $d_1d_2 = N^{\beta}t$, $\beta = 0$ or 1, t square free shows that

$$\begin{aligned} |\mathcal{O}_{Nd}^{(g)}| &= |\mathcal{O}_{N_1N_2d_1d_2}^{(g)}| = |\mathcal{O}_{N_1d_1}^{(g)}||\mathcal{O}_{N_2d_2}^{(g)}| \\ &= \frac{|\mathcal{O}_{N_1N_2d_1}^{(g)}|}{|\mathcal{O}_{N_2}^{(g)}|}\frac{|\mathcal{O}_{N_1N_2d_2}^{(g)}|}{|\mathcal{O}_{N_1}^{(g)}|} = \frac{|\mathcal{O}_{Nd_1}^{(g)}||\mathcal{O}_{Nd_2}^{(g)}|}{|\mathcal{O}_{N}^{(g)}|}. \end{aligned} \tag{3.37}$$

Similarly from (3.12) we have

$$|\mathcal{O}_{Nd}| = \frac{|\mathcal{O}_{Nd_1}||\mathcal{O}_{Nd_2}|}{|\mathcal{O}_{N}|}. \tag{3.38}$$

Since, as noted in (3.15), $\mathcal{O}_N^{(g)} = \mathcal{O}_N$, we have

$$\beta(d) = \beta(d_1 d_2) = \frac{|\mathcal{O}_{Nd}^{(g)}|}{|\mathcal{O}_{Nd}|} = \frac{|\mathcal{O}_{Nd_1}^{(g)}|}{|\mathcal{O}_{Nd_1}|} \frac{|\mathcal{O}_{Nd_2}^{(g)}|}{|\mathcal{O}_{Nd_2}|} = \beta(d_1)\beta(d_2), \tag{3.39}$$

establishing the multiplicativity.

For p prime with $p|\nu$ where ν is given in (3.8), by our choice of x in (3.9) we have $(f(x), p) = 1$ and hence $\mathcal{O}_{Np}^{(g)} \neq \mathcal{O}_{Np}$, and so $\beta(p) < 1$. If $p \nmid \nu$ then by (3.7c) we have $\mathcal{O}_{Np}^{(g)} \neq \mathcal{O}_{Np}$ since f is weakly primitive. Thus we have that $\beta(p) < 1$ for all p. To establish the required uniformity note that for $p \nmid \nu$ we have

$$\mathcal{O}_p = \Gamma_p = G(\mathbb{F}_p) \tag{3.40}$$

and

$$\mathcal{O}_p^{(f)} = \bigcup_{k=1}^{t} W_k(\mathbb{F}_p). \tag{3.41}$$

From (3.7) it follows that

$$p\frac{|\mathcal{O}_p^{(f)}|}{|\mathcal{O}_p|} = t + O(p^{-1/2}), \tag{3.42}$$

where the implied constant depends on G, Λ and f. We have therefore verified that $\beta(p) < 1 - \frac{1}{c_1}$ for c_1 fixed and for all primes p and this completes the proof of Proposition 3.1. □

We turn to the level of distribution axiom (A_1) of Sect. 3.1. From the product structure (3.42), (3.6), (3.10), (3.11) we have

$$\left|\mathcal{O}_{Nd}^{(g)}\right| \ll d^{\dim(G)-1} \tag{3.43}$$

with an implied constant depending only on Λ and f. Hence from (3.36) we have that

$$\sum_{d \leq D} |r(d, A)| \ll X^{\tau} D^{\dim G}. \tag{3.44}$$

Thus our level of distribution in (A_1) is

$$D = X^{(1-\tau)/\dim G - \varepsilon} \tag{3.45}$$

for some ε.

The third axiom concerns the sieve dimension. From (3.42) we have that

$$\sum_{w\le p\le z}\beta(p)\log p=\sum_{w\le p\le z}\left(\frac{t\log p}{p}+O\left(\frac{\log p}{p^{3/2}}\right)\right)=t\log\frac{z}{w}+O(1). \tag{3.46}$$

This establishes (A_2) with the sieve dimension being t.

We are ready to use the elementary sieve Theorem 3.1 except that in our analysis of the sums on progressions we included $n=0$. According to Proposition 3.2 below this term can be omitted as can any fixed term a_{n_0}. Applying the sieve we have shown that for $z=X^{(1-\tau)/9t\dim G}$

$$\frac{X}{(\log X)^t}\ll S(A,P)\ll\frac{X}{(\log X)^t}. \tag{3.47}$$

The n's that remain after this sieving satisfy $(n,P_z)=1$ and hence all the prime factors p of n must be bigger than z. Also if $y\in\mathcal{O}$ with $f(y)=n$ then $y=xA_{i_1}A_{i_2}\cdots A_{i_r}$ with $A_{i_j}\in\{A_1^{\pm1},\dots,A_k^{\pm1}\}$ and $r\le L$. Hence the Hilbert-Schmidt norm of y ($\|y\|=(\sum|y_{ij}|^2)^{1/2}$) satisfies

$$\|y\|\le C^{L+1}, \tag{3.48}$$

where

$$C=\max\left\{\|x\|,\left\|A_1^{\pm1}\right\|,\dots,\left\|A_k^{\pm1}\right\|\right\}. \tag{3.49}$$

Let $\deg(f)$ be the total degree of f. Then for a point y as above we have

$$|f(y)|\ll C^{(L+1)\deg(f)}. \tag{3.50}$$

Thus our points $y\in\mathcal{O}$ which contribute to the sum $S(A,P_z)$ satisfy (3.50) and all prime factors of $f(y)$ are at least

$$X^{(1-\tau)/(9t\dim G)}\gg r^{(L+1)(1-\tau)/(9t\dim G)}.$$

That is each such $f(y)$ has at most

$$\frac{9t\dim(G)\deg(f)\log C}{(1-\tau)\log r} \tag{3.51}$$

prime factors. In order to complete the proof of Theorem 1.1 with the saturation number $r_0(\Lambda,f)$ at most the number[6] in (3.51) (plus 1 if it is an integer) we need to show that the y's produced above are Zariski dense in G. For this we use the expander property the second time.

[6] To be precise, the bound for the saturation number given in (3.51) is valid in the case when Λ is free. In the case when Λ is not free, we first pass to a free subgroup $\tilde{\Lambda}$ which is Zariski dense in G, provided by the Tits theorem as discussed on p. 575; clearly $r_0(\Lambda,f)\le r_0(\tilde{\Lambda},f)$.

Proposition 3.2 *Let W be a proper subvariety of G defined over $\mathbb{Q}$. Then*

$$|\{y \in \mathcal{O} : w(y,x) \le L, y \in W\}| \ll X^{1-\delta},$$

where $\delta = (1-\tau)/\dim G$ and the implied constant depends on W.

With this proposition we can complete the proof of Theorem 1.1. If the points y produced by the sieve in the discussion leading to (3.51) are not Zariski dense then they lie in a proper subvariety W of G which is defined over $\mathbb{Q}$. Hence by the proposition their number is at most $O(X^{1-\delta})$. However the sieve produces at least $c_1 X/(\log X)^t$ such points with $c_1 > 0$ and fixed.

Proof of Proposition 3.2 Since G is irreducible the W in question is defined over $\mathbb{Q}$ and has dimension at most $\dim(G) - 1$. Let $W = \bigcup_{j=1}^{k} W_j$ be the decomposition into irreducible components of W. These are defined over a fixed finite extension K of $\mathbb{Q}$ and each W_j has dimension at most $\dim(G) - 1$. For P outside a finite set of prime ideals of the integers $\mathcal{O}_K$ of K, W_j is absolutely irreducible over the finite field $\mathcal{O}_K/P$. Hence by Lang-Weil Theorem [40]

$$|W_j(\mathcal{O}_K/P)| \ll N(P)^{\dim W_j} \le N(P)^{\dim G - 1}. \tag{3.52}$$

Choosing p a large rational prime (of size to be determined momentarily) so that (p) splits completely in K so that $\mathcal{O}_K/P \cong \mathbb{F}_p$ for any $P|(p)$ and hence

$$|W(\mathbb{Z}/p\mathbb{Z})| \le \sum_{j=1}^{k} |W_j(\mathcal{O}_K/P)| \ll p^{\dim G - 1}. \tag{3.53}$$

Note that for any p (large)

$$\sum_{\substack{y \in \mathcal{O} \\ w(y,x) \le L \\ y \in W}} 1 \le \sum_{\substack{y \in \mathcal{O} \\ w(y,x) \le L \\ y \in W(\mathbb{Z}/p\mathbb{Z})}} 1. \tag{3.54}$$

According to (3.53) and the analysis leading to (3.33) which uses the expander property, we have that

$$\sum_{\substack{y \in \mathcal{O} \\ w(y,x) \le L \\ y \in W(\mathbb{Z}/p\mathbb{Z})}} 1 \ll \sum_{y \in W(\mathbb{Z}/p\mathbb{Z})} \left(\frac{X}{|\mathcal{O}_p|} + X^{\tau} \right) \ll \frac{X}{p} + p^{\dim G - 1} X^{\tau}. \tag{3.55}$$

We are ready to choose p. By the Chebotarev density theorem [14] we can choose p which splits completely in K and p satisfies

$$\frac{X^{(1-\tau)/\dim G}}{2} \le p \le 2X^{(1-\tau)/\dim G}.$$

With this the right hand side of (3.55) is $O(X^{1-\delta})$ with $\delta = (1-\tau)/\dim G$ and coupled with (3.54) this proves Proposition 3.2. □

In applying the sieve to the saturation problem we only made use of the lower bound for $S(A, P)$. With our ordering of the orbit $\mathcal{O} = x\Gamma$ in terms of the $2k$-regular tree, the upper bound provided by the sieve for $S(A, P)$ is certainly meaningful and is sharp up to multiplicative constant. However as far as upper bounds go this says nothing about the original orbit Λ since after applying Tits Theorem Γ is possibly of infinite index in Λ. To obtain meaningful upper bound for Λ with the analysis that we have developed it is more natural to do the counting in the language of random walks. This means that we don't pay attention to whether we visit a point $x \in \Lambda$ repeatedly but in the present context this does not cost much. Let S be a finite symmetric set of generators of Λ and perform a random walk on Λ by starting on e and each step moving by multiplying by an element of S chosen at random with probability $1/|S|$. For $\rho \ge 1$ let $P_\rho(L)$ be the probability that after L steps of the walk on Λ one is a point s for which $f(x)$ has at most ρ prime factors. Our analysis shows (by restricting to the finite index subgroup $\Lambda(q_1 N^2)$) that if ρ is at least as large as the quantity in (3.50) then as $L \to \infty$

$$P_\rho(L) \ge C_1 L^{-t} \quad \text{with } C_1 > 0. \tag{3.56}$$

On the other hand, Proposition 3.2 shows that given a proper subvariety W of G that the probability $P_W(L)$ that after L steps of the walk (on Λ) on lies in W satisfies

$$P_W(L) \ll_W e^{-\beta L} \tag{3.57}$$

for a positive $\beta = \beta(\Lambda)$. Thus Theorem 1.1 can be established more directly this way without passing to the subgroup Λ_1 of Λ and hence Theorem 1.1 can be established as stated assuming Conjecture 1.5 for Λ itself rather than for Γ.

With this language we can give a meaningful and sharp upper bound for $P_t(L)$, that is the probability of $f(y)$ having exactly t factors (its minimal number on a Zariski dense set). We apply the upper bound sieve to the walk on Λ where this time we take for B in Theorem 3.1 the set of all primes p which divide the number $q_1(\Lambda, N^2)$ in Theorem 2.1. We also need the upper bound in Proposition 3.2 to hold uniformly in m for the varieties

$$V_m = \{x \in G : f(x) = m\}.$$

This follows easily from the considerations in Lang-Weil [40]. With this and the analysis in Sect. 3 the upper bound sieve yields

$$P_t(L) \ll \frac{1}{L^t}. \tag{3.58}$$

This upper bound is of the correct order of magnitude in that we expect a "prime number theorem" which can be viewed as a quantitative form of Conjecture 1.4:

$$\lim_{L\to\infty} L^t P_t(L) = C(\Lambda, f) \neq 0. \tag{3.59}$$

We have not determined a conjectured value for $C(\Lambda, f)$ but it will clearly involve the local probabilities $|\Lambda_{Nd}^{(g)}|/|\Lambda_{Nd}|$ as well as the Lyapunov exponent for the random walk in $\mathrm{GL}_n(\mathbb{R})$ determined by the measure $\mu = \frac{1}{|S|}\sum_{g\in S}\delta_g$ (see [3]).

4 Expanders: proof of Theorem 1.2

The *adjacency matrix* of a graph $\mathcal{G}$, $A(\mathcal{G})$ is the $|\mathcal{G}|$ by $|\mathcal{G}|$ matrix, with rows and columns indexed by vertices of $\mathcal{G}$, such that the x, y entry is 1 if and only if x and y are adjacent and 0 otherwise. For a d-regular graph on n vertices the adjacency matrix is a symmetric matrix having n real eigenvalues which we can list in the decreasing order:

$$d = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{n-1} \geq -d;$$

$d = \lambda_0$ is strictly greater than λ_1 iff the graph is connected (which we assume from now on). The smallest eigenvalue λ_{n-1} is equal to $-d$ if and only if the graph is bipartite, in the latter case it occurs with multiplicity one. A family of d-regular graphs $\mathcal{G}_{n,d}$ is said to form an expander family (see [31, 50]) if

$$\limsup_{n\to\infty} \lambda_1(A(\mathcal{G}_{n,d})) < d.$$

For our applications we need (and prove) a slightly stronger property:

Definition 4.1 A family of connected d-regular graphs $\mathcal{G}_{n,d}$ forms a family of absolute expanders if, denoting by $\lambda(A(\mathcal{G}))$ an eigenvalue different from $\pm d$ of greatest absolute modulus, we have

$$\limsup_{n\to\infty} |\lambda(A(\mathcal{G}_{n,d}))| < d.$$

Given a finite group G with a symmetric set of generators S, the Cayley graph $\mathcal{G}(G, S)$, is a graph which has elements of G as vertices, and which has an edge from x to y if and only if $x = \sigma y$ for some $\sigma \in S$.

For a Cayley graph $\mathcal{G}(G, S)$ with $S = \{g_1, g_1^{-1}, \ldots, g_k, g_k^{-1}\}$ the adjacency matrix A can be written as

$$A(\mathcal{G}(G, S)) = \Pi_R(g_1) + \Pi_R\big(g_1^{-1}\big) + \cdots + \Pi_R(g_k) + \Pi_R\big(g_k^{-1}\big), \tag{4.1}$$

where Π_R is a regular representation of G given by the permutation action of G on itself by right multiplication. Every irreducible representation $\rho \in \hat{G}$ appears in Π_R with the multiplicity equal to its dimension

$$\Pi_R = \rho_0 \oplus \bigoplus_{\substack{\rho \in \hat{G} \\ \rho \neq \rho_0}} \underbrace{\rho \oplus \cdots \oplus \rho}_{d_\rho}, \tag{4.2}$$

where ρ_0 denotes the trivial representation and $d_\rho = \dim(\rho)$ is the dimension of the irreducible representation ρ.

Let $N = |G|$. The adjacency matrix $A(\mathcal{G}(G, S))$ is a symmetric matrix having N real eigenvalues which we can list in the decreasing order:

$$2k = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{N-1} \geq -2k;$$

the eigenvalue $2k$ corresponds to the trivial representation in the decomposition (4.2). The strict inequality

$$2k = \lambda_0 > \lambda_1$$

follows from connectivity of our graphs $\mathcal{G}(q)$ (for q sufficiently large), which is a consequence of strong approximation (and, in the case of SL_2, can be also established elementarily as in Sect. 4.1 of [7]). Denoting by W_{2m} the number of closed walks from identity to itself of length $2m$, the trace formula takes form

$$\sum_{j=0}^{N-1} \lambda_j^{2m} = N W_{2m}. \tag{4.3}$$

We now fix $S = \{g_1, g_1^{-1}, \ldots, g_k, g_k^{-1}\}$ such that $\langle S \rangle$ is a free subgroup of $\mathrm{SL}_2(\mathbb{Z})$, and consider, for q square-free, $\mathcal{G}(q) = \mathcal{G}(\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z}), S_q)$, where S_q is a projection of S modulo q. Let $N(q) = |\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})|$. Let $\Omega(q)$ denote the nontrivial spectrum of the adjacency matrix $A(q)$ of $\mathcal{G}(q)$ (that is, all the eigenvalues of $A(\mathcal{G}(q))$ except for $\pm 2k$) and let $\lambda(q)$ be the eigenvalue of maximum modulus in $\Omega(q)$.

Denote by ν the probability measure on $\mathrm{SL}_2(\mathbb{Z})$ supported on S,

$$\nu = \frac{1}{|S|} \sum_{g \in S} \delta_g,$$

and denote by ν_q the probability measure on $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ supported on S_q,

$$\nu_q = \frac{1}{|S|} \sum_{g \in S_q} \delta_g.$$

Let $\nu^{(l)}$ denote the l-fold convolution of ν:

$$\nu^{(l)} = \underbrace{\mu * \cdots * \mu}_{l},$$

where

$$\mu * \nu(x) = \sum_{g \in G} \mu(xg^{-1})\nu(g). \tag{4.4}$$

Note that we have

$$\nu_q^{(2l)}(1) = \frac{W_{2l}}{(2k)^{2l}}. \tag{4.5}$$

For a measure μ on G we let

$$\|\mu\|_2 = \left(\sum_{g \in G} \mu^2(g) \right)^{1/2},$$

and

$$\|\mu\|_\infty = \max_{g \in G} \mu(g).$$

The following proposition is proved in Sect. 4.1.

Proposition 4.1 *Notation being as above, for any $\eta > 0$ there is $C(S, \eta)$ such that if q is square-free for $l > C(S, \eta) \log q$*

$$\left\| \nu_q^{(l)} \right\|_2 < q^{-\frac{3}{2}+\eta}. \tag{4.6}$$

Now observe that since S is a symmetric generating set, we have

$$\nu_q^{(2l)}(1) = \sum_{g \in G} \nu^{(l)}(g)\nu^{(l)}(g^{-1}) = \sum_{g \in G} (\nu^{(l)}(g))^2 = \|\nu^{(l)}\|_2^2,$$

therefore, keeping in mind (4.5), we conclude that (4.6) implies that for

$$l > C(\eta)\log_{2k} q$$

we have

$$W_{2l} < \frac{(2k)^{2l}}{q^{3-2\eta}}. \tag{4.7}$$

Let $q = p_1 \cdot \cdots \cdot p_J$ where p_j are primes. Each irreducible representation of $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$, $\rho(q)$ is given by the tensor product of irreducible representations $\rho(p_i)$ of $\mathrm{SL}_2(\mathbb{Z}/p_i\mathbb{Z})$:

$$\rho(q) = \rho(p_1) \otimes \cdots \otimes \rho(p_J). \tag{4.8}$$

Our proof proceeds by induction on the number of prime factors J.

For $J = 1$ a result going back to Frobenius [22], asserts that for $G = \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ with p prime we have

$$d_\rho(p) \geq \frac{p-1}{2} \tag{4.9}$$

for all *nontrivial* irreducible representations.

Denoting by $m_p(\lambda)$ the multiplicity of $\lambda(p)$, we clearly have

$$\sum_{j=0}^{N(p)-1} \lambda_j^{2l} > m_p(\lambda)\lambda(p)^{2l}, \tag{4.10}$$

since the other terms on the left-hand side of (4.10) are positive.

Combining (4.10) with the Frobenius bound (4.9), and the bound on the number of closed paths (4.7), we obtain, using the trace formula (4.3), that for $l > C(\eta)\log p$ we have

$$\frac{p-1}{2}\lambda(p)^{2l} < |\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})|\frac{(2k)^{2l}}{p^{3-2\eta}}. \tag{4.11}$$

Since $|\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})| = p(p^2-1) < p^3$, this implies that

$$\lambda(p)^{2l} \ll \frac{(2k)^{2l}}{p^{1-2\eta}}, \tag{4.12}$$

and therefore

$$|\lambda(p)| < (2k)^{1-\frac{(1-2\eta)}{C(\eta)}} = \beta(S) < 2k; \tag{4.13}$$

here $\beta(S)$ depends only on the Archimedean norm of elements in S.

Now suppose that Theorem 1.2 is established for $q \in \mathcal{Q}(J-1)$, where $\mathcal{Q}(J-1)$ consists of square-free numbers given by a product of $J-1$ prime factors, that is we have

$$|\lambda(q)| < \beta(S) < 2k \quad \forall q \in \mathcal{Q}(J-1); \tag{4.14}$$

we want to extend (4.14) to the square-free $q \in \mathcal{Q}(J)$, that is, we want to extend it to the square-free numbers $q(J)$ given by a product of J prime factors. The irreducible representations $\rho_{q(J)}$ can be split into two classes: the "old" ones, of the form (4.8) with at least one of the ρ_{p_j} being the trivial representation, and the "new" ones, where all of the factors ρ_{p_j} are given by nontrivial irreducible representations of $\mathrm{SL}_2(\mathbb{Z}/p_j\mathbb{Z})$. Corresponding to this split we have the decomposition

$$\Omega(q) = \Omega_{\mathrm{old}}(q) \cup \Omega_{\mathrm{new}}(q),$$

where

$$\Omega_{\mathrm{old}}(q) = \bigcup_{r \in Q_{J-1}(q)} \Omega(r),$$

with $Q_{J-1}(q)$ being the set of all products of $J-1$ distinct primes in the decomposition of $q = \prod_{j=1}^{J} p_j$. Either $\lambda(q) \in \Omega_{\mathrm{old}}(q)$, or $\lambda(q) \in \Omega_{\mathrm{new}}(q)$. In the "old" case, $\lambda(q)$ occurs as an eigenvalue for a square-free modulus given by a product of at most $J-1$ prime factors and the spectral gap bound is established by the induction hypothesis (4.14). In the "new" case, we have that

$$\mathrm{mult}(\lambda(q)) = \dim(\rho_{\mathrm{new}}(q))$$

for some

$$\rho_{\mathrm{new}}(q) = \rho(p_1) \otimes \cdots \otimes \rho(p_J),$$

with $\rho(p_j)$ being nontrivial for all $1 \le j \le J$. Consequently, in the "new" case we have, using Frobenius bound (4.9),

$$\begin{aligned} \mathrm{mult}(\lambda(q)) = \dim(\rho_{\mathrm{new}}(q(J)) &= \dim(\rho(p_1)) \times \cdots \times \dim(\rho(p_J)) \\ &\ge \prod_{j=1}^{J} \frac{p_j - 1}{2}. \end{aligned} \tag{4.15}$$

Therefore, for $l > C(\eta) \log q$, we obtain

$$\lambda(q)^{2l} \prod_{j=1}^{J} \frac{p_j - 1}{2} < |\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})| \frac{(2k)^{2l}}{q^{3-2\eta}}. \tag{4.16}$$

Since $|\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})| = O(q^3)$, this implies that

$$\lambda(q)^{2l} \ll \frac{(2k)^{2l}}{q^{1-2\eta}}, \tag{4.17}$$

and therefore

$$|\lambda(q)| < (2k)^{1-\frac{(1-2\eta)}{C(\eta)}} = \beta(S) < 2k, \tag{4.18}$$

completing the proof of Theorem 1.2.

4.1 The measure convolution on $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$, q square-free.

In this section we prove Proposition 4.1, which follows immediately from the following

Proposition 4.2 *Let $\mu = \pi_q[\nu^{(\ell)}]$ with $\ell = c\log q$ for some $c > 0$ and assume that for some γ, $0 < \gamma < \frac{3}{4}$ we have*

$$\|\mu\|_2 > q^{-\frac{3}{2}+\gamma}. \tag{4.19}$$

Then

$$\|\mu * \mu\|_2 < q^{-\eta}\|\mu\|_2, \tag{4.20}$$

where $\eta = \eta(\gamma) > 0$ depends only on γ.

We now proceed to prove Proposition 4.2 following the approach in [7]. Assume (4.20) fails, that is, suppose that for any $\eta > 0$ we have that

$$\|\mu * \mu\|_2 > q^{-\eta}\|\mu\|_2. \tag{4.21}$$

We will prove that by choosing η sufficiently small we can find a set A violating Proposition 4.3.

Set

$$J = 10\log q \tag{4.22}$$

and let

$$\tilde{\mu} = \sum_{j=1}^{J} 2^{-j}\chi_{A_j}, \tag{4.23}$$

where A_j are the level sets of the measure μ: for $1 \le j \le J$

$$A_j = \{x \,|\, 2^{-j} < \mu(x) \le 2^{-j+1}\}. \tag{4.24}$$

Setting

$$A_{J+1} = \{x \,|\, 0 < \mu(x) \le 2^{-J}\},$$

we have, for any $x \in G$,

$$\tilde{\mu}(x) \le \mu(x) \le 2\tilde{\mu}(x) + \frac{1}{2^J}\chi_{A_{J+1}}(x),$$

hence, keeping in mind (4.22) we obtain

$$\tilde{\mu}(x) \le \mu(x) \le 2\tilde{\mu}(x) + \frac{1}{q^{10}}. \tag{4.25}$$

Note also, that for any j satisfying $1 \le j \le J$, we have

$$|A_j| \le 2^j. \tag{4.26}$$

By our assumption, (4.21) holds for arbitrarily small η, consequently, in light of (4.25), so does

$$\|\tilde{\mu} * \tilde{\mu}\|_2 > q^{-\eta}\|\tilde{\mu}\|_2. \tag{4.27}$$

Using triangle inequality

$$\|f + g\|_2 \le \|f\|_2 + \|g\|_2,$$

we obtain

$$\begin{aligned}\|\tilde{\mu} * \tilde{\mu}\|_2 &= \Big\| \sum_{1 \le j_1, j_2 \le J} 2^{-j_1-j_2} \chi_{A_{j_1}} * \chi_{A_{j_2}} \Big\|_2 \\ &\le \sum_{1 \le j_1, j_2 \le J} 2^{-j_1-j_2} \|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2.\end{aligned}$$

Therefore, by the pigeonhole principle, for some j_1, j_2, satisfying

$$J \ge j_1 \ge j_2 \ge 1,$$

we have

$$J^2 2^{-j_1-j_2} \|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2 \ge \|\tilde{\mu} * \tilde{\mu}\|_2. \tag{4.28}$$

On the other hand,

$$\begin{aligned}\|\tilde{\mu}\|_2 = \left(\sum_{j=1}^{J} \frac{1}{2^{2j}} |\chi_{A_j}|\right)^{1/2} &\ge \left(\frac{1}{2^{2j_1}}|A_{j_1}| + \frac{1}{2^{2j_2}}|A_{j_2}|\right)^{1/2} \\ &\ge \left(2^{-j_1-j_2}|A_{j_1}|^{1/2}|A_{j_2}|^{1/2}\right)^{1/2},\end{aligned}$$

therefore

$$\|\tilde{\mu}\|_2 \geq 2^{-j_1/2}2^{-j_2/2}|A_{j_1}|^{1/4}|A_{j_2}|^{1/4}. \tag{4.29}$$

Note that by (4.27) we also have

$$J^2 2^{-j_1-j_2}\|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2 \geq q^{-\eta}\max\big(2^{-j_1}|A_{j_1}|^{\frac{1}{2}}, 2^{-j_2}|A_{j_2}|^{\frac{1}{2}}\big),$$

and since, using Young's inequality

$$\|f * g\|_2 \leq \|f\|_1\|g\|_2, \tag{4.30}$$

we have

$$|A_{j_1}|^{\frac{1}{2}}|A_{j_2}|^{\frac{1}{2}}\min\big(|A_{j_1}|^{\frac{1}{2}}, |A_{j_2}|^{\frac{1}{2}}\big) \geq \|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2,$$

we obtain

$$\min\big(2^{-j_1}|A_{j_1}|, 2^{-j_2}|A_{j_2}|\big) \geq \frac{q^{-\eta}}{J^2}. \tag{4.31}$$

Now combining (4.27), (4.28) and (4.29) we have

$$J^2 2^{-j_1-j_2}\|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2 \geq \|\tilde{\mu} * \tilde{\mu}\|_2 \geq q^{-\eta}2^{-j_1/2}2^{-j_2/2}|A_{j_1}|^{1/4}|A_{j_2}|^{1/4},$$

yielding

$$\|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2 \geq \frac{q^{-\eta}}{J^2}2^{j_1/2}2^{j_2/2}|A_{j_1}|^{1/4}|A_{j_2}|^{1/4};$$

recalling (4.22) and (4.26), we obtain

$$\|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2 \geq q^{-2\eta}|A_{j_1}|^{3/4}|A_{j_2}|^{3/4}. \tag{4.32}$$

Let

$$A = A_{j_1} \quad \text{and} \quad B = A_{j_2}. \tag{4.33}$$

Given two multiplicative sets A and B in an ambient group G, their *multiplicative energy* is given by

$$E(A, B) = |\{(x_1, x_2, y_1, y_2) \in A^2 \times B^2 | x_1 y_1 = x_2 y_2\}| = \|\chi_A * \chi_B\|_2^2. \tag{4.34}$$

Inequality (4.32) means that for the sets A and B, defined in (4.33), we have

$$E(A, B) \geq q^{-4\eta}|A|^{3/2}|B|^{3/2}. \tag{4.35}$$

We are ready to apply the noncommutative version of Balog-Szemerédi-Gowers theorem, established by Tao [59] (Corollary 2.46 [60]), which implies that there exists $A_1 \subset A$ such that

$$|A_1| > q^{-\eta_1}|A|, \tag{4.36}$$

where

$$\eta_1 = 4C_1\eta \quad \text{with an absolute constant } C_1, \tag{4.37}$$

such that

$$|A_1(A_1)^{-1}| < q^{\eta_1}|A_1|, \tag{4.38}$$

which means that

$$d(A_1, A_1) < \eta_1 \log q, \tag{4.39}$$

where

$$d(A, B) = \log \frac{|A \cdot B^{-1}|}{|A|^{1/2}|B|^{1/2}}$$

is *Ruzsa distance* between two multiplicative sets.

By definition, a *multiplicative K-approximate group* is any multiplicative set H which is symmetric,

$$H = H^{-1}, \tag{4.40}$$

contains the identity, and is such that there exists a set X of cardinality

$$|X| \le K, \tag{4.41}$$

such that we have the inclusions

$$H \cdot H \subseteq X \cdot H \subseteq H \cdot X \cdot X; \tag{4.42}$$

$$H \cdot H \subseteq H \cdot X \subseteq X \cdot X \cdot H. \tag{4.43}$$

Note, that (4.41), (4.42), (4.43) imply

$$|H^3| = |H \cdot H^2| \le |H^2 \cdot X| < |H \cdot X^2| < K^2|H|. \tag{4.44}$$

Now by Theorem 2.43 [60] (established by Tao in [59]), connecting Ruzsa distance with the notion of approximate group in noncommutative setting,

(4.39) implies that there exists a q^{η_2}-approximative group H, where

$$\eta_2 = C_2\eta_1 \quad \text{with an absolute constant } C_2, \tag{4.45}$$

satisfying the following properties:

$$|H| < q^{\eta_2}|A_1| \tag{4.46}$$

and

$$A_1 \subset XH, \qquad A_1 \subset HY \quad \text{with } |X||Y| < q^{\eta_2}. \tag{4.47}$$

Now since $A_1 \subset \bigcup_{x\in X} xH$ and $|X| < q^{\eta_2}$, there is $x_0 \in X$ such that

$$|A_1 \cap x_0H| > q^{-\eta_2}|A_1|. \tag{4.48}$$

Since $A_1 \subset A = A_{j_1}$, by definition (4.24) of A_j, we have

$$\begin{aligned}\mu(x_0H) &> \mu(A_1 \cap x_0H) > \frac{1}{2^{j_1}}|A_1 \cap x_0H| \overset{(4.48)}{>} \frac{1}{2^{j_1}}q^{-\eta_2}|A_1| \\ &\overset{(4.36)}{>} \frac{1}{2^{j_1}}q^{-\eta_2}q^{-\eta_1}|A_{j_1}|,\end{aligned}$$

and consequently, keeping in mind (4.31), we have

$$\mu(x_0H) > q^{-\eta_3} \tag{4.49}$$

with

$$\eta_3 = \eta_1 + \eta_2 + 2\eta. \tag{4.50}$$

Now (4.46) combined with $A_1 \subset A_{j_1}$ and (4.26) implies that

$$|H| \le q^{\eta_2}2^{j_1}. \tag{4.51}$$

Using Young's inequality (4.30), we have

$$\|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2 \le |A_{j_2}||A_{j_1}|^{1/2},$$

therefore

$$2^{j_2}|A_{j_1}|^{1/2} \ge |A_{j_2}||A_{j_1}|^{1/2} \ge \|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2$$

and

$$2^{-j_1}|A_{j_1}|^{1/2} \ge 2^{-j_1-j_2}\|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2. \tag{4.52}$$

Since by (4.26)

$$2^{-j_1/2} \ge 2^{-j_1}|A_{j_1}|^{1/2}$$

and since by (4.22), (4.25), (4.27), (4.28) we have

$$2^{-j_1-j_2}\|\chi_{A_{j_1}} * \chi_{A_{j_2}}\|_2 \geq q^{-2\eta}\|\mu\|_2,$$

(4.52) implies that

$$2^{-j_1/2} \geq q^{-2\eta}\|\mu\|_2,$$

which combined with (4.19) yields

$$2^{j_1} \leq q^{4\eta}\|\mu\|_2^{-2} \leq q^{3-2\gamma+4\eta}. \tag{4.53}$$

Therefore, keeping in mind (4.51), we have

$$|H| \leq q^{\eta_2}2^{j_1} \leq q^{3-2\gamma+4\eta+\eta_2}. \tag{4.54}$$

Now recall the following result of Kesten [35].

Lemma 4.1 *Let F_k denote the free group on k generators $\{\tilde{g}_1, \ldots, \tilde{g}_k\}$. Denote by $\tilde{\mu}$ the probability measure on F_k supported on $\tilde{g}_i$'s and their inverses,*

$$\tilde{\mu} = \frac{1}{2k}\sum_{i=1}^{k}(\delta_{\tilde{g}_i} + \delta_{\tilde{g}_i^{-1}}). \tag{4.55}$$

Denoting by $\tilde{p}^{(l)}(x, y)$ the probability of being at y after starting at x and performing a random walk according to $\tilde{\mu}$ for l steps, we have

$$\limsup_{l\to\infty} \tilde{p}^{(l)}(x, x)^{1/l} = \frac{\sqrt{2k-1}}{k}. \tag{4.56}$$

In particular,

$$\tilde{p}^{(l)}(x, y) \leq \tilde{p}^{(l)}(x, x) \leq \left(\frac{\sqrt{2k-1}}{k}\right)^l. \tag{4.57}$$

Using the fact that the group $\langle S\rangle$ is free and applying lemma 4.1 as in [7] we obtain that

$$\|\mu\|_\infty < q^{-\gamma_1}. \tag{4.58}$$

Combining (4.49) with (4.58) we have

$$|H| > q^{\gamma_1-\eta_3}. \tag{4.59}$$

Since H is a q^{η_2}-approximate group, it follows from (4.44) that

$$|H \cdot H \cdot H| < q^{2\eta_2}|H|, \tag{4.60}$$

and, therefore, using (4.59), we have

$$|H \cdot H \cdot H| < |H|^{1+\frac{2\eta_2}{\gamma_1-\eta_3}}. \tag{4.61}$$

We now apply the following product theorem for $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$, proved in Sect. 4.2; it is a generalization of Helfgott's result [30].

Proposition 4.3 *Let q be square-free. Let A be a subset of $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ satisfying the following properties for some $\kappa_0 > 0$ and $\kappa_1 > 0$*

$$q^{\kappa_0} < |A| < q^{3-\kappa_0}; \tag{4.62}$$

$$|\pi_{q_1}(A)| > q_1^{\kappa_1} \quad \text{for all } q_1|q \text{ with } q_1 > q^{\omega(\kappa_0)}, \text{ where } \omega(\kappa_0) = \frac{\kappa_0}{40}. \tag{4.63}$$

For all $t \in \mathbb{Z}/q\mathbb{Z}$, for all $g \in \mathrm{Mat}_2(q)$ with $\pi_p(g) \neq \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ for all primes $p|q$ we have

$$\#\{x \in A \mid \gcd(q, (\mathrm{Tr}(gx) - t)) > q^{\kappa_2}\} < o(|A|), \tag{4.64}$$

where[7] *$\kappa_2 = \kappa_2(\kappa_0, \kappa_1) > 0$.*

Then

$$|A \cdot A \cdot A| > q^{\kappa_3}|A| \tag{4.65}$$

with $\kappa_3 = \kappa_3(\kappa_0, \kappa_1) > 0$.

We now show that by choosing η sufficiently small we can ensure that the set H satisfies the conditions (4.62), (4.63), (4.64) in Proposition 4.3, while violating (4.65).

The condition (4.62) is satisfied for η sufficiently small in light of (4.54) and (4.59).

We turn to verifying condition (4.63) for $q_1|q$ with $q_1 > q^{\omega(\kappa_0)}$. For a matrix L define its norm by

$$\|L\| = \sup_{x \neq 0} \frac{\|Lx\|}{\|x\|},$$

[7] To be precise, $\kappa_2(\kappa_0, \kappa_1)$ must satisfy

$$0 < \kappa_2(\kappa_0, \kappa_1) < \min\left(\frac{7}{300}\kappa_0, \frac{7}{70 + 5400\kappa_0^{-1}\kappa_1^{-1}}, \frac{\kappa_0\gamma(\kappa_0, \kappa_1)}{28 + 16\gamma(\kappa_0, \kappa_1)}\right),$$

where $\gamma(\kappa_1, \kappa_1) = \delta_3(\frac{\kappa_0}{10}, \frac{\kappa_1}{10})$ with $\delta_3(\delta_1, \delta_2)$ determined by (1.3) in sum-product theorem (Theorem 1.3).

where the norm of $x = (x_1, x_2)$ is the standard Euclidean norm $\|x\| = \sqrt{x_1^2 + x_2^2}$; let

$$D(g_1, \ldots, g_k) = \max_{1 \le i \le k} \|g_i\|. \tag{4.66}$$

Let $D = D(g_1, \ldots, g_k)$ be as in (4.66) and choose l_0 such that

$$\pi_{q_1} | \operatorname{supp} \nu^{(\ell_0)} \to \mathrm{SL}_2(\mathbb{Z}/q_1\mathbb{Z})$$

is one-to-one and

$$D^{l_0} < q_1 < D^{2l_0}. \tag{4.67}$$

We will make use of the following elementary observation.

Lemma 4.2 *Let μ, μ_1, μ_2 be probability measures on a group G, and suppose that $\mu = \mu_1 * \mu_2$ and $\mu(X) > \alpha$ for some $X \subset G$. Then for some $g \in G$ we have $\mu_2(gX) > \alpha$.*

Proof of Lemma 4.2 Write

$$\mu(X) = \sum_{g \in G} \mu_1(g) \mu_2(g^{-1}X).$$

Suppose $\mu_2(g^{-1}X) < \alpha$ for all $g \in G$. Then, since $\sum_{g \in G} \mu_1(g) = 1$ we obtain a contradiction. □

Writing

$$\nu^{(l)} = \nu^{(l-l_0)} * \nu^{(l_0)},$$

keeping in mind (4.49) and applying Lemma 4.2, we obtain that for some $x_1 \in G$ we have

$$\nu_q^{(l_0)}(x_1 H) > q^{-\eta_3}; \tag{4.68}$$

recall that η_3 can be chosen arbitrarily small. Hence

$$\begin{aligned} |\pi_{q_1}(H)| = |\pi_{q_1}(x_1 H)| &\ge |x_1 H \cap (\operatorname{supp} \nu^{(l_0)})| \\ &\ge \frac{\nu_q^{(l_0)}(x_1 H)}{\|\nu^{(l_0)}\|_\infty} > q^{-\eta_3} \left(\frac{k}{\sqrt{2k-1}} \right)^{l_0}, \end{aligned}$$

were we applied Kesten's bound (4.57) for the random walk on free group.

Consequently,

$$|\pi_{q_1}(H)| > q_1^{\frac{\log(k/\sqrt{2k-1})}{2 \log D} - \frac{\eta_3}{\omega(k_0)}},$$

and so for sufficiently small η the condition (4.63) is satisfied.

It remains to verify condition (4.64). It clearly suffices to show that for all $t \in \mathbb{Z}/q\mathbb{Z}$, for all $b \in \mathrm{Mat}_2(q)$ with $\pi_p(b) \neq \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ for all $p|q$, and for all $q_1|q$ satisfying $q_1 > q^{\kappa_2}$ we have

$$\#\{x \in H \mid \mathrm{Tr}(bx) \equiv t \pmod{q_1}\} < q^{-\varepsilon}|H| \tag{4.69}$$

for some $\varepsilon > 0$.

Assume, that (4.69) fails, that is, assume that for some $b \in \mathrm{Mat}_2(\mathbb{Z}/q\mathbb{Z})$ such that $b \neq \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right) \pmod{p}$ for all $p|q$, for some $t \in \mathbb{Z}/q\mathbb{Z}$, and for some q_1 satisfying $q_1|q$, $q_1 > q^{\kappa_2}$ we have

$$\#\{x \in H \mid \mathrm{Tr}(bx) \equiv t \pmod{q_1}\} = \Omega_\varepsilon(q^{-\varepsilon})|H|$$

for all $\varepsilon > 0$. Recalling (4.24), (4.31), (4.36), (4.48), we have

$$\mu\big[x \mid \mathrm{Tr}\big(bx_0^{-1}x\big) \equiv t \pmod{q_1}\big] > \Omega_\varepsilon(q^{-\varepsilon})q^{-\eta_3}. \tag{4.70}$$

Let $\ell_1 \sim \log q$, to be specified below (see (4.77)). Writing again $\nu^{(\ell)} = \nu^{(\ell-\ell_1)} * \nu^{(\ell_1)}$, and applying Lemma 4.2 we get some $y_0 \in G$ such that

$$\nu^{(\ell_1)}\big[x|\mathrm{tr}\big(bx_0^{-1}y_0x\big) \equiv t \pmod{q_1}\big] > \Omega_\varepsilon(q^{-\varepsilon})q^{-\eta_3}. \tag{4.71}$$

Let $b' = bx_0^{-1}y_0$. Since $b \neq \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right) \pmod{p}$ for all $p|q$ and $x_0, y_0 \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ we have

$$b' \neq \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \pmod{p} \text{ for all } p|q. \tag{4.72}$$

Denote $W(m) = \mathrm{supp}\,\nu^{(m)}$. Let T denote the set

$$T = \{x \in W(l_1) \mid \mathrm{Tr}(b'x) \equiv t \pmod{q_1}\}; \tag{4.73}$$

we have

$$\nu_q^{(\ell_1)}(T) > \Omega_\varepsilon(q^{-\varepsilon})q^{-\eta_3}. \tag{4.74}$$

For any quintuple $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x$ in T we have

$$\mathrm{Tr}b'(x^{(j)} - x) \equiv 0 \pmod{q_1} \quad (1 \leq j \leq 4). \tag{4.75}$$

Viewing Mat_2 as a four-dimensional vector space, that is, identifying $b = \left(\begin{smallmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{smallmatrix}\right)$ with $(b_{11}, b_{12}, b_{21}, b_{22})$, $\mathrm{Tr}(ab)$ is identified with the inner product,

$$\mathrm{Tr}(ab) = a_{11}b_{11} + a_{12}b_{12} + a_{21}b_{21} + a_{22}b_{22}.$$

Consequently, in light of (4.72) and (4.75) we have that for all $p|q_1$

$$\det \begin{pmatrix} x_{11}^{(1)} - x_{11} & x_{11}^{(2)} - x_{11} & x_{11}^{(3)} - x_{11} & x_{11}^{(4)} - x_{11} \\ x_{12}^{(1)} - x_{12} & x_{12}^{(2)} - x_{12} & x_{12}^{(3)} - x_{12} & x_{12}^{(4)} - x_{12} \\ x_{21}^{(1)} - x_{21} & x_{21}^{(2)} - x_{21} & x_{21}^{(3)} - x_{21} & x_{21}^{(4)} - x_{21} \\ x_{22}^{(1)} - x_{22} & x_{22}^{(2)} - x_{22} & x_{22}^{(3)} - x_{22} & x_{22}^{(4)} - x_{22} \end{pmatrix} = 0 \pmod{p}. \tag{4.76}$$

By submultiplicativity of the norm of product of matrices, the elements of $W(l_1) \subset \mathrm{SL}_2(\mathbb{Z})$ have entries bounded by $D_1^{l_1}$ for some $D_1(g_1, \ldots, g_k)$. Let $D_2 = D_1^6$ and choose l_1 so that

$$D_2^{\ell_1} < q_1 < D_2^{2\ell_1}. \tag{4.77}$$

Hence the determinant of the matrix on the left-hand side of (4.76) is an integer bounded by $5^5 D_1^{5\ell_1}$, which is less than q_1. Consequently, by (4.76) we have

$$\det \begin{pmatrix} x_{11}^{(1)} - x_{11} & x_{11}^{(2)} - x_{11} & x_{11}^{(3)} - x_{11} & x_{11}^{(4)} - x_{11} \\ x_{12}^{(1)} - x_{12} & x_{12}^{(2)} - x_{12} & x_{12}^{(3)} - x_{12} & x_{12}^{(4)} - x_{12} \\ x_{21}^{(1)} - x_{21} & x_{21}^{(2)} - x_{21} & x_{21}^{(3)} - x_{21} & x_{21}^{(4)} - x_{21} \\ x_{22}^{(1)} - x_{22} & x_{22}^{(2)} - x_{22} & x_{22}^{(3)} - x_{22} & x_{22}^{(4)} - x_{22} \end{pmatrix} = 0 \quad \text{in } \mathbb{Z}. \tag{4.78}$$

We now proceed as follows. Choose a prime P satisfying $\log P \sim l_1$. Applying expander result for prime modulus [7] to Cayley graph of $\mathrm{SL}_2(\mathbb{Z}/P\mathbb{Z})$ with respect to $\pi_P(g_1, \ldots, g_k)$ we have

$$\left\| \nu_P^{(l_1)} \right\|_\infty = O_\varepsilon(P^\varepsilon) P^{-3}. \tag{4.79}$$

It follows form (4.74), (4.79) that

$$|\pi_P(T)| \geq \frac{\nu_P^{(l_1)}(T)}{\|\nu_P^{(l_1)}\|_\infty} > \Omega_\varepsilon(q^{-\varepsilon}) q^{-\eta_3} \Omega_\varepsilon(P^{-\varepsilon}) P^3. \tag{4.80}$$

Now since $q_1 > q^{\kappa_2}$ we have

$$\log P > \frac{\kappa_2 \log q}{\log D_2(g_1, \ldots, g_k)},$$

therefore (4.80) implies that

$$|\pi_P(T)| > \Omega_\varepsilon(P^{-\varepsilon}) P^{3 - \frac{\log D_2 \eta_3}{\kappa_2}}. \tag{4.81}$$

Recalling (4.78), valid for all $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x \in T$, (4.81) implies that

$$\left|\left\{(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x) \in \mathrm{SL}_2(\mathbb{Z}/P\mathbb{Z})^5 : \right.\right.$$
$$\left.\left. \det \begin{pmatrix} x_{11}^{(1)} - x_{11} & x_{11}^{(2)} - x_{11} & x_{11}^{(3)} - x_{11} & x_{11}^{(4)} - x_{11} \\ x_{12}^{(1)} - x_{12} & x_{12}^{(2)} - x_{12} & x_{12}^{(3)} - x_{12} & x_{12}^{(4)} - x_{12} \\ x_{21}^{(1)} - x_{21} & x_{21}^{(2)} - x_{21} & x_{21}^{(3)} - x_{21} & x_{21}^{(4)} - x_{21} \\ x_{22}^{(1)} - x_{22} & x_{22}^{(2)} - x_{22} & x_{22}^{(3)} - x_{22} & x_{22}^{(4)} - x_{22} \end{pmatrix} = 0 \quad \text{in } \mathbb{F}_P \right\}\right|$$
$$> \Omega_\varepsilon(P^{-\varepsilon}) P^{-5\frac{\log D_2 \eta_3}{\kappa_2}} |\mathrm{SL}_2(\mathbb{Z}/P\mathbb{Z})|^5 \tag{4.82}$$

for any $\varepsilon > 0$. If the polynomial

$$f(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x)$$
$$= \det \begin{pmatrix} x_{11}^{(1)} - x_{11} & x_{11}^{(2)} - x_{11} & x_{11}^{(3)} - x_{11} & x_{11}^{(4)} - x_{11} \\ x_{12}^{(1)} - x_{12} & x_{12}^{(2)} - x_{12} & x_{12}^{(3)} - x_{12} & x_{12}^{(4)} - x_{12} \\ x_{21}^{(1)} - x_{21} & x_{21}^{(2)} - x_{21} & x_{21}^{(3)} - x_{21} & x_{21}^{(4)} - x_{21} \\ x_{22}^{(1)} - x_{22} & x_{22}^{(2)} - x_{22} & x_{22}^{(3)} - x_{22} & x_{22}^{(4)} - x_{22} \end{pmatrix}$$

were not identically zero, the number of solutions to $f \equiv 0 \bmod P$ as $(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x)$ varies over $\mathrm{SL}_2(\mathbb{Z}/P\mathbb{Z})^5$ would be bounded by $O(P^{14})$ [56]. By choosing η sufficiently small, (4.82) therefore implies that f vanishes identically on $\mathrm{SL}_2(\mathbb{F}_P)^5$, implying that $\mathrm{SL}_2(\mathbb{F}_P) \subset \mathbb{F}_P^4$ is contained in a hyperplane, obtaining a contradiction and completing the proof of Proposition 4.2.

4.2 Product theorem in $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$, q square-free

In this section we establish Proposition 4.3, which generalizes the result of Helfgott [30] in the case of prime modulus.

4.2.1 Outline of the proof

We begin by giving a very rough outline of the proof. Denote by $A^{(n)}$ the n-fold product set

$$A^{(n)} = \underbrace{(A \cup A^{-1}) \cdots (A \cup A^{-1})}_{n}. \tag{4.83}$$

Assume (4.65) fails, that is assume that

$$|A \cdot A \cdot A| < O_\varepsilon(q^\varepsilon)|A| \tag{4.84}$$

for any $\varepsilon > 0$. By Proposition 2.40 [60] we then have

$$|A^{(n)}| < O_\varepsilon(q^{n\varepsilon})|A| \tag{4.85}$$

for all $n \geq 1$.

In the outline below we denote by k_i (small) absolute integer constants, and by ρ_i positive constants depending on κ_0, κ_1; these are detailed in the course of the proof.

(1) If A fails to grow, that is, if it satisfies (4.85), and if it contains two elements a, b such that for a (large) divisor q_1 of q the projections $\pi_{q_1}(a)$ and $\pi_{q_1}(b)$ are in "general position" (do not have common eigenvectors), and such that most of the elements of $\pi_{q_1}(A)$, $\pi_{q_1}(aA)$, $\pi_{q_1}(bA)$ are non-unipotent, we deduce (Lemma 4.6) that $A^{(k_1)}$ contains a large subset V, whose projection modulo q_1 consists of simultaneously diagonalizable matrices.

(2) Using sum-product theorem, we deduce (Lemma 4.8) that given a simultaneously diagonalizable set V with the set $\mathrm{Tr}(V)$ satisfying the assumptions of sum-product theorem, and a matrix c with non-zero entries (in the chosen basis), the set of traces of $V^{(8)}cV^{(8)}c$ grows substantially. Applying this result to the set $V \subset A^{(k_1)}$ constructed in the preceding step, results in a set $A^{(k_1k_2)}$ with $\mathrm{Tr}(A^{(k_1k_2)}) \gtrsim |A|^{\frac{1}{3}+\rho_2}$.

(3) We now apply Lemma 4.5, which says, roughly speaking, that if a subset A of $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ does not grow much under multiplication (that is, if it satisfies (4.85)) then $A^{(2)}$ contains a subset W of matrices whose projections modulo q' (a large divisor of q) are simultaneously diagonalizable, and whose size is not much less than the size of traces of matrices in A with non-unipotent projections modulo q'. This allows us to deduce that the set $A^{(2k_1k_2)}$ contains a subset W of simultaneously diagonalizable matrices (modulo a large divisor q_3 of q) of size $|W| \gtrsim |A|^{\frac{1}{3}+\rho_3}$.

(4) Finally, we apply Lemma 4.7, asserting that if W is a simultaneously diagonalizable set of matrices and d is a matrix with non-zero entries then $|WdWdW| > \Omega_\varepsilon(q^{-\varepsilon})|W|^3$, to obtain $|A^{(2k_1k_2k_3)}| \gtrsim |A|^{1+\rho_4}$ implying a contradiction with (4.85).

As detailed at the beginning of Sect. 4.2.5, the existence of matrices a, b, c, d, needed in the course of proof, is ensured using condition (4.64) in the product theorem.

4.2.2 Trace size from size

Our first goal it to show (Corollary 4.2) that for any set $A \subset \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$, given two matrices g, h, whose projections modulo a large divisor q' of q ($q' > q^{1-\tau}$) are "in general position" (that is, have no common eigenvector), the size of the set of traces of one of the sets A, gA, hA is not much smaller than $q^{-\tau}|A|^{\frac{1}{3}}$.

Lemma 4.3 *Let p be a prime. Let $\{g, h\}$ be elements in $\mathrm{SL}_2(\mathbb{F}_p)$ with no common eigenvector over $\overline{\mathbb{F}_p}$. Then the map*

$$\mathrm{SL}_2(\mathbb{F}_p) \longrightarrow \mathbb{F}_p^3 : x \mapsto (\mathrm{Tr}(x), \mathrm{Tr}(gx), \mathrm{Tr}(hx)) \tag{4.86}$$

has multiplicity at most 2.

Proof of Lemma 4.3 Assume first $\mathrm{Tr}(g) \neq \pm 2$. Diagonalize g in $\mathbb{F}_p$ or in an extension field $K \simeq \mathbb{F}_{p^2}$. Specify the basis, so as to make g diagonal. Thus we can write

$$g = \begin{bmatrix} r & 0 \\ 0 & r^{-1} \end{bmatrix} \quad \text{and} \quad h = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix},$$

where, from our assumption, $r \in K \backslash \{1, -1\}$ and $\beta\gamma \neq 0 \pmod{p}$.

For $x = \left[\begin{smallmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{smallmatrix}\right] \in \mathrm{SL}_2(K)$ we get

$$\mathrm{Tr}(x) = x_{11} + x_{22}, \tag{4.87}$$

$$\mathrm{Tr}(gx) = rx_{11} + r^{-1}x_{22}, \tag{4.88}$$

$$\mathrm{Tr}(hx) = \alpha x_{11} + \beta x_{21} + \gamma x_{12} + \delta x_{22}. \tag{4.89}$$

Let $\mathrm{Tr}(x), \mathrm{Tr}(gx), \mathrm{Tr}(hz)$ be given. From (4.87), (4.88) we recover x_{11} and x_{22}. Since $x_{11}x_{22} - x_{12}x_{21} = 1$, (4.89) implies

$$x_{12}(\gamma\beta^{-1}x_{12} + \beta^{-1}(\alpha x_{11} + \delta x_{22} - \mathrm{Tr}(hx))) = 1 - x_{11}x_{22} \tag{4.90}$$

and therefore x_{12} is determined up to multiplicity 2. If $x_{12} \neq 0$, also x_{21} and hence x are determined. If $x_{12} = 0$, (4.89) determines x_{21}.

Next, suppose that $\mathrm{Tr}(g) \in \{2, -2\}$. In an appropriate basis we obtain

$$g = \begin{bmatrix} \pm 1 & b \\ 0 & \pm 1 \end{bmatrix} \quad \text{and} \quad h = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$$

with $b \neq 0$ and $\gamma \neq 0$, again from our assumption. Hence

$$\mathrm{Tr}(gx) = \pm(x_{11} + x_{22}) + bx_{21} = \pm\mathrm{Tr}(x) + bx_{21}, \tag{4.91}$$

determining x_{21}. We obtain the equation

$$1 = x_{11}(\mathrm{Tr}(x) - x_{11}) - x_{21}\gamma^{-1}(\mathrm{Tr}(hx) - \beta x_{21} - \alpha x_{11} - \delta(\mathrm{Tr}(x) - x_{11})), \tag{4.92}$$

that determines x_{11} up to multiplicity 2. From (4.87), x_{22} is obtained and (4.89) gives x_{12}. This completes the proof of Lemma 4.3. □

Now let q be square-free, $q = \prod_{i\in I} p_i$; thus $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ is isomorphic to the product $\prod \mathrm{SL}_2(\mathbb{Z}/p_i\mathbb{Z})$. The following result is an immediate consequence of Lemma 4.3.

Lemma 4.4 *Let $g, h \in \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ and assume that for each $p|q$*

$$\{\pi_p(g), \pi_p(h)\} \text{ do not have a common eigenvector.} \tag{4.93}$$

Then the map

$$\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z}) \to (\mathbb{Z}/q\mathbb{Z})^3 \ : x \mapsto (\mathrm{Tr}(x), \mathrm{Tr}(gx), \mathrm{Tr}(hx)) \tag{4.94}$$

has multiplicity at most $2^{|I|}$.

The following corollary is an immediate consequence of Lemma 4.4.

Corollary 4.1 *Let g, h be elements of $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ satisfying (4.93). For any subset A of $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ we have*

$$|\mathrm{Tr}(A)| + |\mathrm{Tr}(gA)| + |\mathrm{Tr}(hA)| > \Omega_\varepsilon(q^{-\varepsilon})|A|^{1/3}. \tag{4.95}$$

Corollary 4.2 *Assume that g, h are elements of $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ such that for some $\tau > 0$ we have*

$$\gcd(q, \mathrm{Tr}(ghg^{-1}h^{-1}) - 2) < q^\tau. \tag{4.96}$$

Then for any subset A of $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ we have

$$|\mathrm{Tr}(A)| + |\mathrm{Tr}(gA)| + |\mathrm{Tr}(hA)| > \Omega_\varepsilon(q^{-\varepsilon})q^{-\tau}|A|^{1/3}. \tag{4.97}$$

Proof of Corollary 4.2 Let $q_1 = \gcd(q, \mathrm{Tr}(ghg^{-1}h^{-1}) - 2) < q^\tau$ and $q' = \frac{q}{q_1}$. Thus if $p|q'$, then $\{\pi_p(g), \pi_p(h)\} \subset \mathrm{SL}_2(p)$ don't have a common eigenvector. Applying Corollary 4.1 to $\pi_{q'}(A)$, it follows that

$$\begin{aligned} &|\mathrm{Tr}(A)| + |\mathrm{Tr}(gA)| + |\mathrm{Tr}(hA)| \\ &\quad \geq |\mathrm{Tr}(\pi_{q'}(A))| + |\mathrm{Tr}(\pi_{q'}(gA))| + |\mathrm{Tr}(\pi_{q'}(hA))| \\ &\quad > \Omega_\varepsilon(q^{-\varepsilon})|\pi_{q'}(A)|^{1/3} > \Omega_\varepsilon(q^{-\varepsilon})q^{-\tau}|A|^{1/3}. \end{aligned} \tag{4.98}$$

□

4.2.3 Growth and simultaneously diagonalizable subsets

Lemma 4.5 *Let $A \subset \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$, q square free. Let $T \subset \mathrm{Tr}(A) \subset \mathbb{Z}/q\mathbb{Z}$ such that for some $\tau > 0$ we have*

$$\gcd(q, t^2 - 4) < q^{\tau} \quad \text{for all } t \in T. \tag{4.99}$$

Then there is a subset $V \subset A^{-1}A$ and $q'|q$, such that

$$q' > q^{1-\tau}, \tag{4.100}$$

$$\pi_{q'}(V) \subset \mathrm{SL}_2(\mathbb{Z}/q'\mathbb{Z}) \quad \text{are simultaneously diagonalizable over } \prod_{p|q'} \overline{\mathbb{F}_p}, \tag{4.101}$$

$$|V| > |T| \frac{|A|}{|A^2A^{-1}|}. \tag{4.102}$$

Proof of Lemma 4.5. For each $t \in T$, take an element $g_t \in A$ with $\mathrm{Tr}(g_t) = t$. Define sets C_t by

$$C_t = \{x g_t x^{-1} | x \in A\} \subset A^2A^{-1}; \tag{4.103}$$

these sets are clearly disjoint. Hence, by the pigeonhole principle, there is $t \in T$ such that

$$|C_t| \leq \frac{|A^2A^{-1}|}{|T|}. \tag{4.104}$$

Split $A = A_1 \cup \cdots \cup A_k$ into disjoint subsets A_j such that $x g_t x^{-1} = y g_t y^{-1}$ for $x, y \in A_j$. Again, by the pigeonhole principle, for some j we have $|A_j| \geq \frac{|A|}{k} \geq \frac{|A|}{|C_t|}$. Setting $A_0 = A_j$, we have

$$|A_0| \geq \frac{|A|}{|C_t|} \geq \frac{|A|}{|A^2A^{-1}|}|T|. \tag{4.105}$$

Choose $x_0 \in A$ such that

$$x g_t x^{-1} = x_0 g_t x_0^{-1} \quad \text{for } x \in A_0 \tag{4.106}$$

and set $V = x_0^{-1} A_0$.

From (4.99), there is $q'|q$ satisfying (4.100) and such that $\mathrm{Tr} g_t \neq \pm 2 \pmod{p}$ for all $p|q'$. For $p|q'$, diagonalize $\pi_p(g_t)$ over $\overline{\mathbb{F}}_p$. Thus, in this basis

$$\pi_p(g_t) = \begin{pmatrix} r_p & 0 \\ 0 & r_p^{-1} \end{pmatrix} \quad \text{with } r_p \neq \pm 1. \tag{4.107}$$

If $g \in V$, (4.106) implies that g and g_t commute. Thus, writing in the chosen basis

$$\pi_p(g) = \begin{pmatrix} \alpha_p & \beta_p \\ \gamma_p & \delta_p \end{pmatrix},$$

it follows that

$$\left(r_p - r_p^{-1}\right)\beta_p \equiv 0 \equiv \left(r_p - r_p^{-1}\right)\gamma_p \pmod{p};$$

hence $\beta_p \equiv 0 \equiv \gamma_p \pmod{p}$. Therefore $\pi_{q'}(g)$ is diagonal in this basis for all $g \in V$. □

Lemma 4.6 *Let $A \subset \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$. Assume there are elements g, h in A such that the following properties are satisfied*:

$$\gcd(q, \mathrm{Tr}(ghg^{-1}h^{-1}) - 2) < q^{\tau} \tag{4.108}$$

and

$$\gcd(q, ((\mathrm{Tr}x)^2 - 4)((\mathrm{Tr}(gx))^2 - 4)((\mathrm{Tr}(hx))^2 - 4)) < q^{\tau} \tag{4.109}$$

for all $x \in A' \subset A$, where $|A'| > |A| - o(|A|)$. Then there is $q_1|q$ and $V \subset A^{-1}A$ such that

$$q_1 > q^{1-\tau}, \tag{4.110}$$

$$\pi_{q'}(V) \text{ are simultaneously diagonalizable}, \tag{4.111}$$

$$|V| > \Omega_{\varepsilon}(q^{-\varepsilon})q^{-\tau}\frac{|A|^{4/3}}{|A^3A^{-1}|}. \tag{4.112}$$

Proof By Corollary 4.2, assumption (4.108) implies that there is $g_0 \in \{1, g, h\}$, such that $|\mathrm{Tr}g_0A'| > \Omega_{\varepsilon}(q^{-\varepsilon})q^{-\tau}|A|^{\frac{1}{3}}$. Next, apply Lemma 4.5 to the set g_0A' with $\mathrm{Tr}(g_0A') = T$. Assumption (4.109) implies that condition (4.99) holds. The conclusion is clear from (4.100)–(4.102). □

Lemma 4.7 *Let $V \subset \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ be a set of diagonal elements (in a specified basis). Let $g = \left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right)$ with*

$$\alpha\beta\gamma \neq 0 \pmod{p} \quad \text{for all } p|q. \tag{4.113}$$

Then

$$|VgVgV| > \Omega_{\varepsilon}(q^{-\varepsilon})|V|^3. \tag{4.114}$$

Proof For $p|q$, denote

$$S_p = \{x \in \mathbb{F}_p^* | \pi_p(\alpha)x + \pi_p(\delta)x^{-1} = 0 \text{ or } \pi_p(\alpha^2)x + \pi_p(\beta\gamma)x^{-1} = 0\},$$

which has at most 4 elements, since $\pi_p(\alpha) \neq 0$. Now partitioning for each $p|q$

$$\mathbb{F}_p^* = (\mathbb{F}_p^* \backslash S_p) \cup S_p,$$

and keeping in mind that $|S_p| \leq 4$ we obtain $q_1|q$ and a subset $V' \subset V$ such that:

$$\pi_p(V') \cap S_p = \emptyset \quad \text{if } p|q_1,$$

$$|V'| > 5^{-|I|}|V| > \Omega_\varepsilon(q^{-\varepsilon})|V|,$$

$$|\pi_{q/q_1}(V')| = 1.$$

Thus $\pi_{q_1}|V'$ is one to one.

Next, we show that the map $\pi_{q_1}(V')^3 \to \mathrm{Mat}_2(q_1)$ given by

$$\begin{pmatrix} x_1 & 0 \\ 0 & x_1^{-1} \end{pmatrix} \times \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix} \times \begin{pmatrix} x_2 & 0 \\ 0 & x_2^{-1} \end{pmatrix} \mapsto \begin{pmatrix} x_1 & 0 \\ 0 & x_1^{-1} \end{pmatrix} \pi_{q_1}(g) \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix} \pi_{q_1}(g) \begin{pmatrix} x_2 & 0 \\ 0 & x_2^{-1} \end{pmatrix} \quad (4.115)$$

has multiplicity at most $10^{|I|}$, which by the preceding will imply (4.114).

It clearly suffices to show that for each prime $p|q_1$ the map

$$\mathbb{F}_p^* \times \mathbb{F}_p^* \times (\mathbb{F}_p^* \backslash S_p) : (x_1, x_2, x) \mapsto \begin{pmatrix} x_1 & 0 \\ 0 & x_1^{-1} \end{pmatrix} \pi_p(g) \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix} \pi_p(g) \begin{pmatrix} x_2 & 0 \\ 0 & x_2^{-1} \end{pmatrix} \quad (4.116)$$

is of bounded multiplicity. Fix $p|q_1$, and denote again

$$\pi_p(g) = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}, \quad \alpha\beta\gamma \neq 0 \pmod{p}.$$

Then expression on the right hand side of (4.116) is equal to $\begin{pmatrix} ax_1x_2 & b\frac{x_1}{x_2} \\ c\frac{x_2}{x_1} & \frac{d}{x_1x_2} \end{pmatrix}$ with

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \alpha^2 x + \beta\gamma x^{-1} & \beta(\alpha x + \delta x^{-1}) \\ \gamma(\alpha x + \delta x^{-1}) & \delta^2 x^{-1} + \beta\gamma x \end{pmatrix}.$$

We have that $bc = \beta\gamma(\alpha x + \delta x^{-1})^2$, hence x is determined up to multiplicity 4. Since $x \notin S_p$, $a = \alpha^2 x + \beta\gamma x^{-1} \neq 0$ and $b = \beta(\alpha x + \delta x^{-1}) \neq 0 \pmod p$; therefore both $x_1 x_2$ and $\frac{x_1}{x_2}$ are determined $(\text{mod } p)$. This completes the proof of Lemma 4.7 □

We remark that Lemma 4.7 remains valid if $\mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ is replaced by $\mathrm{SL}_2\big(\prod_{p|q} \overline{\mathbb{F}}_p\big)$ with $\overline{\mathbb{F}}_p = \mathbb{F}_p$ or $\overline{\mathbb{F}} = \mathbb{F}_{p^2}$.

4.2.4 Trace amplification

Lemma 4.8 *Let $V \subset \mathrm{SL}_2(\mathbb{Z}/q\mathbb{Z})$ be a set of simultaneously diagonalizable elements which for each $p|q$ we diagonalize over $\overline{\mathbb{F}}_p$ in an appropriate basis. Let in this basis*

$$g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \mathrm{SL}_2\Big(\prod_{p|q} \overline{\mathbb{F}}_p\Big) \tag{4.117}$$

with

$$\alpha\beta\gamma\delta \neq 0 \ (\text{mod } p) \quad \textit{for all } p|q. \tag{4.118}$$

Assume

$$|V| > q^{\delta_2}. \tag{4.119}$$

For all $0 < \delta_1, \delta_2 < \frac{1}{10}$, there is $\gamma = \gamma(\delta_1, \delta_2) > 0$, such that one of the following properties holds:

$$|V| > q^{1-\delta_1}, \tag{4.120}$$

$$\textit{There is } q_1|q \textit{ such that } q_1 > q^{\frac{\delta_1}{3}} \textit{ and } |\pi_{q_1}(V)| < q_1^{\delta_2}, \tag{4.121}$$

$$|\mathrm{Tr}(V^{(8)} g V^{(8)} g)| > |V|^{1+\gamma}, \tag{4.122}$$

where we denote by $V^{(n)}$ the n-fold product set defined in (4.83).

Proof Let $V = \{\left(\begin{smallmatrix} x & 0 \\ 0 & x^{-1} \end{smallmatrix}\right) | x \in M\}$ where $M \subset \prod_{p|q} (\overline{\mathbb{F}}_p)^*$. We have

$$\begin{aligned} &\mathrm{Tr}\begin{pmatrix} x_1 & 0 \\ 0 & x_1^{-1} \end{pmatrix}\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}\begin{pmatrix} x_2 & 0 \\ 0 & x_2^{-1} \end{pmatrix}\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \\ &\quad = \alpha^2 x_1 x_2 + \delta^2 \frac{1}{x_1 x_2} + \beta\gamma\left(\frac{x_1}{x_2} + \frac{x_2}{x_1}\right). \end{aligned} \tag{4.123}$$

Suppose (4.122) fails, that is, suppose that for all $\varepsilon > 0$ we have

$$|\mathrm{Tr}(V^{(8)} g V^{(8)} g)| < |V|^{1+\varepsilon}. \tag{4.124}$$

Using (4.123) we then have

$$\left|\left\{\alpha^2 x_1 x_2 + \delta^2 \frac{1}{x_1 x_2} + \beta\gamma\left(\frac{x_1}{x_2} + \frac{x_2}{x_1}\right) \middle| x_1, x_2 \in M^{(8)}\right\}\right| < O_\varepsilon(q^\varepsilon)|M|. \tag{4.125}$$

Letting $x_1 = y_1 y_2, x_2 = \frac{y_1}{y_2}$ with $y_1, y_2 \in M^{(4)}$, it follows that for all $\varepsilon > 0$

$$\left|\left\{\left(\alpha^2 y_1^2 + \delta^2 y_1^{-2}\right) + \beta\gamma\left(y_2^2 + y_2^{-2}\right) \middle| y_1, y_2 \in M^{(4)}\right\}\right| < O_\varepsilon(q^\varepsilon)|M|. \tag{4.126}$$

Let

$$B = \{\alpha^2 y^2 + \delta^2 y^{-2} | y \in M^{(4)}\},$$

$$C = \{y^2 + y^{-2} | y \in M^{(4)}\},$$

$$C' = \beta\gamma C.$$

By Ruzsa's sumset inequality (see [60]) we have

$$|C' + C'| \leq \frac{|B + C'|^2}{|B|}. \tag{4.127}$$

For $ab \neq 0 \pmod p$ the map

$$(\overline{\mathbb{F}}_p)^* \to \overline{\mathbb{F}}_p : y \mapsto ax^2 + bx^{-2}$$

has multiplicity at most 4; therefore

$$|B| > \Omega_\varepsilon(q^{-\varepsilon})|M^{(4)}|, \tag{4.128}$$

$$|C'| = |C| > \Omega_\varepsilon(q^{-\varepsilon})|M^{(4)}|. \tag{4.129}$$

Consequently, we conclude from (4.126), (4.127), (4.128), (4.129), that

$$|C' + C'| < O_\varepsilon(q^\varepsilon)|M|. \tag{4.130}$$

Let

$$T_s = \{x^2 + x^{-2} | x \in M^{(s)}\}. \tag{4.131}$$

By (4.130) we have that

$$|T_4 + T_4| < O_\varepsilon(q^\varepsilon)|M|. \tag{4.132}$$

Since $1 \in M^{(2)}$, we have that $T_2 \subset T_4$. Further, using identity

$$(x^2 + x^{-2})(y^2 + y^{-2}) = (xy)^2 + (xy)^{-2} + (xy^{-1})^2 + (xy^{-1})^{-2},$$

we conclude that $T_2 \cdot T_2 \subset T_4 + T_4$. Consequently (4.132) implies that for all $\varepsilon > 0$ we have

$$|T_2 + T_2| + |T_2 \cdot T_2| < O_\varepsilon(q^\varepsilon)|M|. \tag{4.133}$$

Since clearly $|M^{(2)}| \geq |M|$, and since by the remark following (4.127) we have that

$$|T_s| > \Omega_\varepsilon(q^{-\varepsilon})|M^{(s)}|, \tag{4.134}$$

we obtain that

$$|T_2 + T_2| + |T_2 \cdot T_2| < O_\varepsilon(q^\varepsilon)|T_2|. \tag{4.135}$$

Note that

$$T_2 \subset \mathrm{Tr}(V \cdot V \cdot V \cdot V) \subset \prod_{p|q} \mathbb{F}_p = \mathbb{Z}/q\mathbb{Z},$$

so that we may invoke the sum-product theorem in $\mathbb{Z}/q\mathbb{Z}$ (Theorem 1.3). Since the conclusion of Theorem 1.3 fails by (4.135), either assumption (1.1) or (1.2) from Theorem 1.3 fails. If $|T_2| > q^{1-\delta_1}$, (4.120) holds. Next assume $q_1|q, q_1 > q^{\frac{\delta_1}{3}}$ and $|\pi_{q_1}(T_2)| < q_1^{\delta_2}$. Then also $|\pi_{q_1}(M)| < q_1^{\delta_2}$, and therefore the alternative (4.121) holds. This completes the proof of Lemma 4.8.

4.2.5 Set amplification

We are ready to complete the proof of Proposition 4.3. Assume (4.65) fails; as discussed in Sect. 4.2.1, this implies that (4.85) holds.

To see how condition (4.64) implies the existence of matrices g_i mentioned in the outline, note that we can re-express this condition as follows. Given $t \in \mathbb{Z}/q\mathbb{Z}$ and $g \in \mathrm{Mat}_2(q)$ with $\pi_p(g) \neq \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ for all $p|q$, let $\xi_{g,t}(x)$ denote the affine form given by $\xi_{g,t}(x) = \mathrm{Tr}(gx - t)$. Then

$$\#\left\{x \in A \,\middle|\, \prod_{\substack{p|q \\ \xi_{g,t}(x)=0 \pmod p}} p > q^{\kappa_2}\right\} < o(|A|). \tag{4.136}$$

This assumption also implies that for a fixed number r, given $g_1, \ldots, g_r$ in $\mathrm{Mat}_2(q)$ with $\pi_p(g_j) \neq \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ for all $p|q$ and all $1 \leq j \leq r$, and $t_1, \ldots, t_r$ in $\mathbb{Z}/q\mathbb{Z}$, we have

$$\#\left\{x \in A \,\middle|\, \prod_{\substack{p|q \\ \xi_{g_1,t_1}(x)\ldots\xi_{g_r,t_r}(x)=0 \pmod p}} p > q^{r\kappa_2}\right\} < o(|A|), \tag{4.137}$$

or, equivalently,

$$\#\left\{x \in A \,\middle|\, \gcd\left(q, \prod_{j=1}^{r}[\mathrm{Tr}(g_j x) - t_j]\right) > q^{r\kappa_2}\right\} < o(|A|). \tag{4.138}$$

Next, letting $r = 2$, $g_1 = g_2 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and $t_1 = 2$, $t_2 = -2$ we have

$$\#\left\{x \in A \,\middle|\, \prod_{\substack{p|q \\ \mathrm{Tr}(x)=\pm 2 \pmod p}} p > q^{2\kappa_2}\right\} < o(|A|);$$

consequently there is an element $g \in A$ such that

$$\tilde{q} = \prod_{\substack{p|q \\ \mathrm{Tr}\, g=\pm 2 \pmod p}} p < q^{2\kappa_2}. \tag{4.139}$$

Let $q' = \frac{q}{\tilde{q}}$. We have

$$q' > q^{1-2\kappa_2}, \tag{4.140}$$

and for each $p|q'$ in an appropriate basis the matrix g may be diagonalized over $\overline{\mathbb{F}}_p$:

$$\pi_p(g) = \begin{pmatrix} r_p & 0 \\ 0 & r_p^{-1} \end{pmatrix} \quad \text{with } r_p \neq \pm 1. \tag{4.141}$$

Letting $r = 2$, $t_1 = t_2 = 0$ and choosing g_1, g_2 corresponding, in the chosen basis, to the linear forms

$$\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \mapsto x_{12} \quad \text{and} \quad \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \mapsto x_{21}$$

on $\mathrm{Mat}_2(\prod \overline{\mathbb{F}}_p)$, another application of (4.138) yields $h \in A$ such that

$$\pi_p(h) = \begin{pmatrix} \alpha_p & \beta_p \\ \gamma_p & \delta_p \end{pmatrix} \quad \text{with } \beta_p \gamma_p \neq 0 \tag{4.142}$$

for all $p|q''$ with $q''|q'$ such that

$$q'' > q' q'^{-2\kappa_2} > q^{1-4\kappa_2}. \tag{4.143}$$

Hence for $p|q''$ we have

$$\det(gh - hg) = \beta_p \gamma_p \left(\frac{1}{r_p} - r_p\right) \neq 0 \pmod p$$

and therefore

$$\gcd(q, \mathrm{Tr}(ghg^{-1}h^{-1}) - 2) < \frac{q}{q''} < q^{4\kappa_2}. \tag{4.144}$$

Hence condition (4.108) of Lemma 4.6 holds with $\tau = 4\kappa_2$.

Applying (4.138) with $r = 6$, $g_1 = g_2 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $g_3 = g_4 = g$, $g_5 = g_6 = h$ and $t_j = \pm 2$, condition (4.109) is obtained with $\tau = 4\kappa_2$. Application of Lemma 4.6 therefore yields a subset $V \subset A^{-1}A$ and $q_1|q$ such that

$$q_1 > q^{1-6\kappa_2}. \tag{4.145}$$

$$\text{The elements of } \pi_{q_1}(V) \text{ are simultaneously diagonalizable,} \tag{4.146}$$

$$|V| > \Omega_\varepsilon(q^{-\varepsilon}) q^{-6\kappa_2} \frac{|A|^{4/3}}{|A^3 A^{-1}|}. \tag{4.147}$$

Now since by (4.85) we have

$$|A^3 A^{-1}| = O_\varepsilon(q^\varepsilon)|A|, \tag{4.148}$$

combining (4.147) and (4.148) we obtain

$$|V| > \Omega_\varepsilon(q^{-\varepsilon}) q^{-6\kappa_2} |A|^{1/3}, \tag{4.149}$$

which combined with the left-hand side of the inequality (4.62) ($|A| > q^{\kappa_0}$) yields

$$|V| > \Omega_\varepsilon(q^{-\varepsilon}) q^{\frac{\kappa_0}{3} - 6\kappa_2} \tag{4.150}$$

and

$$|V| > \Omega_\varepsilon(q^{-\varepsilon}) |A|^{\frac{1}{3} - \frac{6\kappa_2}{\kappa_0}}. \tag{4.151}$$

Perform a basis change to make $\pi_{q_1}(V)$ diagonal. Another application of (4.138) yields $g_0 \in A$ and $q_2|q_1$ s.t.

$$q_2 > q_1^{1-4\kappa_2} > q^{1-10\kappa_2}, \tag{4.152}$$

and, in the basis diagonalizing $\pi_{q_1}(V)$, we have

$$g_0 = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \quad \text{with } \pi_p(\alpha\beta\gamma\delta) \neq 0 \text{ for all } p|q_2. \tag{4.153}$$

Apply Lemma 4.8 with q replaced by q_2 to the set $\pi_{q_2}(V)$; condition (4.118) is implied by (4.153) and condition (4.119) is implied by (4.150). Set

$$\delta_1 = \frac{\kappa_0}{10}, \qquad \delta_2 = \frac{\kappa_1}{10}. \tag{4.154}$$

We now consider in turn the three possibilities (4.120), (4.121), (4.122) and show that in each case we obtain a contradiction.

Case 1: We have

$$|\pi_{q_2}(V)| > q_2^{1-\delta_1}. \tag{4.155}$$

Application of Lemma 4.7 gives

$$\begin{aligned} |\pi_{q_2}(Vg_0Vg_0V)| &> \Omega_\varepsilon(q^{-\varepsilon})|\pi_{q_2}(V)|^3 > \Omega_\varepsilon(q^{-\varepsilon})q_2^{3(1-\delta_1)} \\ &> \Omega_\varepsilon(q^{-\varepsilon})q^{3(1-10\kappa_2)(1-\delta_1)} > \Omega_\varepsilon(q^{-\varepsilon})q^{3-\kappa_4} \end{aligned} \tag{4.156}$$

with

$$\kappa_4 = \frac{3}{10}\kappa_0 + 30\kappa_2 - 3\kappa_1\kappa_2.$$

Now since $V \subset A^{-1}A$, we have $Vg_0Vg_0V \subset A^{(8)}$ and therefore (4.156) implies that

$$|A^{(8)}| > \Omega_\varepsilon(q^{-\varepsilon})q^{3-\kappa_4}.$$

On the other hand, by our assumption (4.85), we have

$$|A^{(8)}| < O_\varepsilon(q^\varepsilon)|A|$$

and by (4.62) we have

$$|A| < q^{3-\kappa_0},$$

yielding

$$|A^{(8)}| < O_\varepsilon(q^\varepsilon)q^{3-\kappa_0}.$$

Consequently we obtain a contradiction for $\kappa_4 < \kappa_0$, that is for

$$\kappa_2 < \frac{7}{300}\kappa_0. \tag{4.157}$$

Case 2: Alternative (4.121) holds, that is, there is $q_3|q_2$ with $q_3 > q_2^{\frac{\delta_1}{3}}$, such that

$$|\pi_{q_3}(V)| < q_3^{\delta_2}. \tag{4.158}$$

Hence we may specify a subset V_1 of V, such that

$$|V_1| > q_3^{-\delta_2}|V|, \tag{4.159}$$

and

$$|\pi_{q_3}(V_1)| = 1. \tag{4.160}$$

Applying Lemma 4.7 with q replaced by q_2/q_3 to the set $\pi_{q_2/q_3}(V)$, we obtain

$$|\pi_{q_2/q_3}(V_1 g_0 V_1 g_0 V_1)| > \Omega_\varepsilon(q^{-\varepsilon})|V_1|^3 > \Omega_\varepsilon(q^{-\varepsilon})q_3^{-3\delta_2}|V|^3, \tag{4.161}$$

where the set

$$W = V_1 g_0 V_1 g_0 V_1$$

satisfies by (4.160)

$$|\pi_{q_3}(W)| = 1. \tag{4.162}$$

At this point, invoke assumption (4.63) on A. Keeping in mind (4.152) and (4.154) we have

$$q_3 > q_2^{\frac{\delta_1}{3}} > q^{\frac{\kappa_0(1-10\kappa_2)}{30}}, \tag{4.163}$$

and therefore, provided

$$\kappa_2 < \frac{1}{10} - \frac{3\omega(\kappa_0)}{\kappa_0} = \frac{1}{40}, \tag{4.164}$$

we have

$$|\pi_{q_3}(A)| > q_3^{\kappa_1}. \tag{4.165}$$

It then follows from (4.161)–(4.165) that

$$|A^{(9)}| \geq |\pi_{q_2}(W \cdot A)| \geq |\pi_{q_2/q_3}(W)||\pi_{q_3}(A)| \geq \Omega_\varepsilon(q^{-\varepsilon})q_3^{\kappa_1-3\delta_2}|V|^3.$$

Recalling equation (4.149) we therefore have

$$|A^{(9)}| > \Omega_\varepsilon(q^{-\varepsilon})q_3^{\kappa_1-3\delta_2}q^{-18\kappa_2}|A|,$$

and hence, using (4.163) and (4.154), we obtain

$$|A^{(9)}| > \Omega_\varepsilon(q^{-\varepsilon})|A|q^{\frac{7\kappa_0\kappa_1-\kappa_2(5400+70\kappa_0\kappa_1)}{300}}.$$

Consequently, using the left-hand side of the inequality (4.62) ($|A| > q^{\kappa_0}$) we obtain a contradiction with (4.85) provided

$$\kappa_2(\kappa_0, \kappa_1) < \frac{7}{70 + 5400\kappa_0^{-1}\kappa_1^{-1}}. \tag{4.166}$$

Case 3: Alternative (4.122) holds, that is for some $\gamma > 0$, $\gamma(\delta_1, \delta_2) = \gamma(\kappa_0, \kappa_1)$ we have mod q_2

$$|\mathrm{Tr}(V^{(8)} g V^{(8)} g)| > |\pi_{q_2}(V)|^{1+\gamma}.$$

Since $V^{(8)} g V^{(8)} g \subset A^{34}$, using (4.149) and (4.152) we obtain

$$|\mathrm{Tr}(A^{(34)})| > \Omega_\varepsilon(q^{-\varepsilon}) q^{-16\kappa_2(1+\gamma)} |A|^{\frac{1}{3}(1+\gamma)}. \tag{4.167}$$

Let $T = \mathrm{Tr}(A^{(34)})$. With the aim of applying Lemma 4.5 to $A^{(34)}$, we pass to a divisor of q, so that condition (4.99) is fulfilled. Partitioning for each $p|q$

$$\mathbb{F}_p = \{2\} \cup \{-2\} \cup (\mathbb{F}_p \backslash \{2, -2\}),$$

we obtain $q_4|q$ and $T_0 \subset T$ such that the following holds:

$$|T_0| > 3^{-|I|}|T| > \Omega_\varepsilon(q^{-\varepsilon}) q^{-16\kappa_2(1+\gamma)} |A|^{\frac{1}{3}(1+\gamma)}, \tag{4.168}$$

$$|\pi_{q/q_4}(T_0)| = 1, \tag{4.169}$$

and

$$\pi_p(T_0) \cap \{2, -2\} = \emptyset \quad \text{for all } p|q_4. \tag{4.170}$$

Now apply Lemma 4.5 with q replaced by q_4 to the set $\pi_{q_4}(A^{(34)}) \subset \mathrm{SL}_2(\mathbb{Z}/q_4\mathbb{Z})$. By (4.170) we have

$$\gcd(q_4, t^2 - 4) = 1 \quad \text{for all } t \in \pi_{q_4}(T_0),$$

consequently Lemma 4.5 yields a subset $W \subset A^{(68)}$ such that $\pi_{q_4}(W)$ is simultaneously diagonalizable and

$$|\pi_{q_4}(W)| > |T_0| \frac{|\pi_{q_4}(A)|}{|\pi_{q_4}(A^{(102)})|}. \tag{4.171}$$

Diagonalize $\pi_{q_4}(W) \subset \mathrm{SL}_2(\mathbb{Z}/q_4\mathbb{Z})$ in an appropriate basis. By (4.138), there is $g_1 \in A$ and $q_5|q_4$, satisfying

$$q_5 > q_4^{1-4\kappa_2}, \tag{4.172}$$

such that in the chosen basis we have

$$g_1 = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \quad \text{with } \pi_p(\alpha\beta\gamma\delta) \neq 0 \text{ for } p|q_5. \tag{4.173}$$

Applying Lemma 4.7 to $\pi_{q_5}(W) \subset \mathrm{SL}_2(\mathbb{Z}/q_5\mathbb{Z})$, we obtain

$$\begin{aligned} |\pi_{q_5}(WgWgW)| &> \Omega_\varepsilon(q^{-\varepsilon})|\pi_{q_5}(W)|^3 > \Omega_\varepsilon(q^{-\varepsilon})q_4^{-12\kappa_2}|\pi_{q_4}(W)|^3 \\ &\overset{(4.171)}{>} \Omega_\varepsilon(q^{-\varepsilon})q_4^{-12\kappa_2}|T_0|^3\left(\frac{|\pi_{q_4}(A)|}{|\pi_{q_4}(A^{(102)})|}\right)^3 \\ &\overset{(4.168)}{>} \Omega_\varepsilon(q^{-\varepsilon})q^{-28\kappa_2-16\kappa_2\gamma}|A|^{1+\gamma}\left(\frac{|\pi_{q_4}(A)|}{|\pi_{q_4}(A^{(102)})|}\right)^3. \end{aligned} \tag{4.174}$$

Now since $WgWgW \subset A^{(206)}$, the left-hand side of (4.174) is no greater than $|A^{(206)}|$. By our assumption (4.85), we have

$$|A^{(206)}| < O_\varepsilon(q^\varepsilon)|A|. \tag{4.175}$$

So combining (4.174) and (4.175) we have

$$|A| > \Omega_\varepsilon(q^{-\varepsilon})q^{-28\kappa_2-16\kappa_2\gamma}|A|^{1+\gamma}\left(\frac{|\pi_{q_4}(A)|}{|\pi_{q_4}(A^{(102)})|}\right)^3,$$

and therefore, since $|A| > q^{\kappa_0}$, we obtain

$$|\pi_{q_4}(A^{(102)})| > \Omega_\varepsilon(q^{-\varepsilon})q^{\frac{\gamma\kappa_0-28\kappa_2-16\kappa_2\gamma}{3}}|\pi_{q_4}(A)|. \tag{4.176}$$

Now choose $\xi \in \mathbb{Z}/q_4\mathbb{Z}$ and $A_1 \subset A$ such that

$$\pi_{q_4}(A_1) = \{\xi\} \tag{4.177}$$

and

$$|\pi_{q/q_4}(A_1)| = |A_1| \geq \frac{|A|}{|\pi_{q_4}(A)|}. \tag{4.178}$$

Then from (4.176)–(4.178) we have for all $\varepsilon > 0$

$$\begin{aligned} |A^{(103)}| &> |A_1 \cdot A^{(102)}| > |\pi_{q_4}(A^{(102)})| \cdot |\pi_{q/q_4}(A_1)| \\ &> \Omega_\varepsilon(q^{-\varepsilon})q^{\frac{\gamma\kappa_0-28\kappa_2-16\kappa_2\gamma}{3}}|A|. \end{aligned}$$

Therefore, provided

$$\kappa_2 < \frac{\kappa_0\gamma(\kappa_0,\kappa_1)}{28+16\gamma(\kappa_0,\kappa_1)}, \tag{4.179}$$

we obtain a contradiction to (4.85).

This completes the proof of Proposition 4.3. □

5 Sum-product theorem in $\mathbb{Z}/q\mathbb{Z}$ (q square-free)

This section is devoted to the proof of Theorem 1.3. Recall that $q = \prod_{j=1}^{J} p_j$ is a product of distinct primes; for $q'|q$ we let $\pi_{q'}$ denote the projection $\mathbb{Z}/q\mathbb{Z} \to \mathbb{Z}/q'\mathbb{Z}$. Let $\mathbb{Z}_q^*$ denote the units of $\mathbb{Z}_q$, where $\mathbb{Z}_q = \mathbb{Z}/q\mathbb{Z}$.

5.1 Outline of the proof

We begin by giving a rough outline of the proof.

Reduction to a subset of $\mathbb{Z}_q^*$. Assuming A satisfies assumptions (1.1), (1.2) of Theorem 1.3 but fails (1.3), we first show that there is a large subset A_1 of A, and a large divisor q_1 of q such that $\pi_{q_1}(A_1) \subset \mathbb{Z}_{q_1}^*$ and $\pi_{q_1}(A_1)$ satisfies (1.1), (1.2) but fails (1.3) in $\mathbb{Z}_{q_1}$. For $A_1 \subset \mathbb{Z}_{q_1}^*$ the failure of (1.3) implies (using Lemma 5.1 established in [5]) that for a large subset A_2 of A_1 all the polynomial expressions do not grow, so (after passing to a large subset of A and a large divisor of q) the failure of (1.3) implies that for any $k > 0$ and any $\varepsilon > 0$

$$|kA^k| < O_\varepsilon(q^\varepsilon)|A|. \tag{5.1}$$

From now on our task is to establish a contradiction with (5.1).

Application of sum-product estimate in $\mathbb{Z}_p$ for prime p. Sum-product estimate in $\mathbb{Z}_p$ for prime p [9] implies that for $A \subset \mathbb{Z}_p$ satisfying $|A| > p^\tau$ we have $rA^r = \mathbb{Z}_p$ for $r = r(\tau)$ (see Lemma 1 in [4]). A slight generalization of the exponential sum bound in [9] implies that the same conclusion also holds for different sets $A_{i,j} \subset \mathbb{Z}_p$ satisfying $|A_{i,j}| > p^\tau$, that is, given $\tau > 0$, there is $r = r(\tau)$ such that we have

$$\sum_{i=1}^{r} \prod_{j=1}^{r} A_{i,j} = \mathbb{Z}_p. \tag{5.2}$$

"Regularization". With the aim of applying (5.2) we perform the following "regularization" of A. Naturally associated with a subset A of $\mathbb{Z}_q$ is a directed tree $\mathcal{T}(A)$, consisting of J levels, with vertices on level j consisting of elements in $\pi_{p_1 \cdots p_j}(A)$, and with each vertex corresponding to the element x at level j, connected to those vertices at level $j+1$, for which there is $t \in \mathbb{Z}_{p_{j+1}}$, such that $(x, t) \in \pi_{p_1 \cdots p_{j+1}}(A)$. After mild "pruning" we obtain a "regularized" subset of A which is of comparable size with A, such that the degrees of vertices in $\mathcal{T}(A)$ are constant at each level.

Preserving large subfactors. Letting q_1 be the product of those primes for which the degrees of $\mathcal{T}(A)$ constructed in the previous step are greater than

$p^{\delta_1/3}$, we obtain using (5.1), (5.2) (applied with $\tau = \delta_1/3$), that $q_1 > q^{\delta_2/2}$ and

$$\pi_{q_1}(rA^r) = \mathbb{Z}_{q_1} \tag{5.3}$$

where $r = r(\delta_2)$, and with the same property (5.3) holding for all $q'|q$ with $q' > q^{\delta_1/3}$ (with q replaced by q').

"Gluing" different factors. Finally, we combine different factors obtained in the preceding step to iteratively increase the value of q_1, thereby obtaining a contradiction with (5.1) and (1.1). This is accomplished using Proposition 5.1, asserting that if for some subset $S \subset \mathbb{Z}_q$ and $q_1|q$, $q_2|\frac{q}{q_1}$ we have $\pi_{q_1}(S) = \mathbb{Z}_{q_1}, \pi_{q_2}(S) = \mathbb{Z}_{q_2}$, then there is $Q|q_1q_2$, such that $Q > q_1q_2^{\rho}$ (with $\rho > 0$) and $\pi_Q(400S^2) = \mathbb{Z}_Q$. Proposition 5.1 is proved using techniques developed in [5], combined with "very dense graph" analogue of Balog-Szemerédi-Gowers Lemma (Lemma 5.9) and near-exact sum set theorem (Lemma 5.7), which is a consequence of Kneser's theorem [36].

5.2 Reduction to a subset of $\mathbb{Z}_q^*$

Assume (1.3) fails, that is, suppose that for all $\varepsilon > 0$ we have

$$|A + A| + |A \cdot A| < q^{\varepsilon}|A|. \tag{5.4}$$

The aim of this section is to show that assuming (5.4) and (1.1), (1.2), there is a divisor q' of q, $q' > q^{1-\eta}$ and a large subset B of A ($|B| > \Omega_{\varepsilon}(q^{-\varepsilon})|A|$) for any $\varepsilon > 0$), such that $\pi_{q'}(B) \subset \mathbb{Z}_{q'}^*$, which satisfies condition (1.2) (with q replaced by q', η replaced by 2η and δ_2 replaced by $\frac{\delta_2}{2}$), and such that for any $k \geq 1$ we have

$$|k\pi_{q'}(B)^k| < O_{k,\varepsilon}(q'^{\varepsilon})|\pi_{q'}(B)|.$$

We begin by constructing a large subset A' of A, such that $\pi_{q'}(A') \subset \mathbb{Z}_{q'}^*$ with $q'|q$, $q' > q^{1-\eta}$, and having a small sum-set $A' + A'$ and a small product-set $A' \cdot A'$. Let $A_0 = A$, $q_0' = 1$, $q_0'' = 1$. Let

$$A_1' = \{x \in A_0 \,|\, \pi_{p_1}(x) \neq 0\}.$$

If $|A_1'| \geq \frac{p_1-1}{p_1}|A_0|$, let $A_1 = A_1'$ and let $q_1' = q_0'p_1$, $q_1'' = q_0''$. If $|A_1'| < \frac{p_1-1}{p_1}|A_0|$, let $A_1 = A_0$ and let $q_1' = q_0'$, $q_1'' = q_0''p_1$. Proceeding iteratively, at step $i+1$ let

$$A_{i+1}' = \{x \in A_i \,|\, \pi_{p_{i+1}}(x) \neq 0\}.$$

If $|A'_{i+1}| \geq \frac{p_{i+1}-1}{p_{i+1}}|A_i|$, let $A_{i+1} = A'_{i+1}$ and let $q'_{i+1} = q'_i p_{i+1}$, $q''_{i+1} = q''_i$. If $|A'_{i+1}| < \frac{p_{i+1}-1}{p_{i+1}}|A_i|$, let $A_{i+1} = A_i$ and let $q'_{i+1} = q'_i$, $q''_{i+1} = q''_i p_{i+1}$. After J steps we obtain a subset A' of A, $A' = A'_J$ and $q' = q'_J$, $q'' = q''_J$, satisfying the following properties:

$$\pi_{q'}(A') \subset \mathbb{Z}^*_{q'}, \tag{5.5}$$

$$\pi_{q''}(A') = \{0\}, \tag{5.6}$$

$$|\pi_{q'}(A')| = |A'| > 2^{-J}|A|. \tag{5.7}$$

Hence, keeping in mind (1.2) we have

$$|A'| > \Omega_\varepsilon(q^{-\varepsilon})|A| > \Omega_\varepsilon(q^{-\varepsilon})q^{\delta_2}. \tag{5.8}$$

We claim that $q'' \leq q^\eta$ (and hence $q' > q^{1-\eta}$). Otherwise, (1.2) would imply

$$|\pi_{q''}(A)| > (q'')^{\delta_2} > q^{\eta\delta_2}$$

and, since by (5.5), (5.6), we have

$$|A + A'| \geq |\pi_{q''}(A)||\pi_{q'}(A')|,$$

we would obtain

$$|A + A| \geq |A + A'| \geq |\pi_{q''}(A)|\,|A'| > \Omega_\varepsilon(q^{-\varepsilon})q^{\eta\delta_2}|A|$$

contradicting (5.4).

By (5.4), (5.7) we have for any $\varepsilon > 0$

$$|A' + A'| + |A' \cdot A'| < O_\varepsilon(q^\varepsilon)|A'|. \tag{5.9}$$

We now make use of the following result:

Lemma 5.1 (Lemma 3 in [5]) *Let $A \subset \mathbb{Z}^*_q$ satisfy*

$$|A + A| + |A \cdot A| < K|A|. \tag{5.10}$$

Fix $k \in \mathbb{Z}_+$. Then there is a subset $A_1 \subset A$ such that

$$|A_1| > K^{-2}|A|, \tag{5.11}$$

$$|kA_1^k| < K^C|A_1| \tag{5.12}$$

with $C = C(k)$.

Applying Lemma 5.1, a further reduction to a subset A_1 of A', $|A_1| > \Omega_\varepsilon(q^{-\varepsilon})|A'|$, permits us to ensure that moreover

$$\left|kA_1^k\right| < O_{k,\varepsilon}(q^\varepsilon)|A_1| \tag{5.13}$$

for any given $k \in \mathbb{Z}_+$ and any $\varepsilon > 0$. We denote here by kB (respectively B^k) the k-fold sum (respectively product) set of B.

Next, let $q_1|q'$ and $q_1 > (q')^{2\eta} > q^\eta$. Assume $|\pi_{q_1}(A_1)| < q_1^{\delta_2/2}$. We may then specify $x_0 \in \mathbb{Z}_{q_1}$ and $A_0 \subset A_1$ such that $\pi_{q_1}(A_0) = \{x_0\}$ and $|A_0| > q_1^{-\delta_2/2}|A_1|$. Write again

$$|A + A| \geq |A + A_0| \geq |\pi_{q_1}(A)|\,|A_0| > q_1^{\delta_2} q_1^{-\delta_2/2}\Omega_\varepsilon(q^{-\varepsilon})|A|,$$

which contradicts (5.4). Therefore $|\pi_{q_1}(A_1)| \geq q_1^{\delta_2/2}$.

In summary, the set $B = \pi_{q'}(A_1) \subset \mathbb{Z}_{q'}^*$ satisfies the following properties:

$$|B| > \Omega_\varepsilon(q^{-\varepsilon})|A|; \tag{5.14}$$

$$|kB^k| < O_{k,\varepsilon}(q^\varepsilon)|B|; \tag{5.15}$$

$$\text{if } q_1|q',\ q_1 > (q')^{2\eta}, \text{ then } |\pi_{q_1}(B)| > q_1^{\delta_2/2}. \tag{5.16}$$

Replacing q by q', and A by B, we may thus assume that A satisfies the conditions

$$A \subset \mathbb{Z}_q^*, \tag{5.17}$$

$$|kA^k| < O_{k,\varepsilon}(q^\varepsilon)|A|, \tag{5.18}$$

in addition to (1.1), (1.2).

5.3 Construction of a regular subset

Naturally associated with a subset A of $\mathbb{Z}_q$ is a directed tree $\mathcal{T}(A)$, consisting of J levels, with vertices on level j consisting of elements in $\pi_{p_1\cdots p_j}(A)$, and with each vertex corresponding to the element x at level j connected to those vertices at level $j+1$, for which there is $t \in \mathbb{Z}_{p_{j+1}}$ such that $(x,t) \in \pi_{p_1\cdots p_{j+1}}(A)$. Our aim in this section is to show that by performing "regularization" of A we can pass to a large subset A' of A, such that the degrees of vertices in $\mathcal{T}(A')$ are constant at each level.

Lemma 5.2 *There exists a subset A' of A satisfying the following properties*:

For all $1 \le j \le J$ *and* $x \in \pi_{p_1 \cdots p_j}(A')$ *we have*

$$|\{t \in \mathbb{Z}_{p_{j+1}} | (x,t) \in \pi_{p_1 \cdots p_{j+1}}(A')\}| = K_{j+1}, \tag{5.19}$$

where $\{K_j\}_{1<j\le J}$ *is a sequence of positive integers*;

$$|A'| > \left[\prod_{j=1}^{J} (2\log p_j)^{-1}\right]|A|. \tag{5.20}$$

Proof of Lemma 5.2 Recall that $q = p_1 \cdots p_J$. We perform the regularization in a straightforward way, starting from the bottom so as to preserve the regularization performed at an earlier stage. For $x \in \mathbb{Z}_{q/p_J}$ consider the subset $A(x) \subset \mathbb{Z}_{p_J}$ for which obviously $0 \le |A(x)| \le p_J$. Partitioning $\pi_{q/p_J}(A)$ into $\log p_J$ subsets, we may specify $A_J \subset A$ and a positive integer K_J, such that for $x \in \pi_{q/p_J}(A_J)$, we have

$$K_J \le |A(x)| = |A_J(x)| < 2K_J, \tag{5.21}$$

$$|A_J| > (\log p_J)^{-1}|A|. \tag{5.22}$$

A further restriction of A_J (at the cost of an extra factor $\frac{1}{2}$ in (5.22)) permits us to ensure that

$$|A_J(x)| \in \{0, K_J\} \quad \text{for } x \in \mathbb{Z}_{q/p_J}. \tag{5.23}$$

Next, consider for $x \in \mathbb{Z}_{q/p_{J-1}p_J}$ the sets $\pi_{p_{J-1}}(A_J(x)) \subset \mathbb{Z}_{p_{J-1}}$. We may specify an integer K_{J-1} and a further subset $A_{J-1} \subset A_J$ with $A_{J-1}(x) = A_J(x)$ for $x \in \pi_{q/p_{J-1}p_J}(A_{J-1})$, such that

$$|A_{J-1}| > (2\log p_{J-1})^{-1}|A_J|, \tag{5.24}$$

$$|\pi_{p_{J-1}}(A_{J-1}(x))| \in \{0, K_{J-1}\} \quad \text{for } x \in \mathbb{Z}_{q/p_{J-1}p_J}. \tag{5.25}$$

In light of (5.23), we also have

$$|A_{J-1}(x)| = K_{J-1}K_J \quad \text{for } x \in \pi_{q/p_{J-1}p_J}(A_{J-1}). \tag{5.26}$$

The continuation of the process is clear; as a result we obtain a set A' such that the degrees of vertices in $\mathcal{T}(A')$ are constant at each level. □

5.4 Sum-product sets in $\mathbb{Z}_p$ for prime p

We will need the following property:

Lemma 5.3 *For all $\tau > 0$, there is $r = r(\tau) \in \mathbb{Z}_+$ such that the following holds*:

Let $(A_{s,s'})_{1\le s,s'\le r}$ be subsets of $\mathbb{Z}_p$ with

$$|A_{s,s'}| > p^{\tau} \quad \text{for all } 1 \le s, s' \le r. \tag{5.27}$$

Then the sum-product set of $(A_{s,s'})_{1\le s,s'\le r}$ equals all of $\mathbb{Z}_p$:

$$\sum_{s=1}^{r} \prod_{s'=1}^{r} A_{s,s'} = \mathbb{Z}_p. \tag{5.28}$$

Proof of Lemma 5.3 Our aim is to show that we can find $r = r(\tau)$ such that for any $y \in \mathbb{Z}_p$

$$\#\left\{ (x_{s,s'}) \in \prod A_{s,s'} \,\middle|\, y = \sum_{s=1}^{r} x_{s,1} \cdots x_{s,r} \right\} > 0. \tag{5.29}$$

Note that

$$\begin{aligned} &\#\left\{ (x_{s,s'}) \in \prod A_{s,s'} \,\middle|\, y = \sum_{s=1}^{r} x_{s,1} \cdots x_{s,r} \right\} \\ &\quad = \frac{1}{p} \sum_{a=0}^{p-1} \sum_{x_{s,s'} \in A_{s,s'}} e_p\left(a\left(y - \sum_{s=1}^{r} x_{s,1} \cdots x_{s,r} \right)\right), \end{aligned} \tag{5.30}$$

where $e_p(x) = \exp(\frac{2\pi i x}{p})$, and consequently it is enough to show that for some $r = r(\tau)$ we have

$$\frac{1}{p} \sum_{a=0}^{p-1} \sum_{x_{s,s'} \in A_{s,s'}} e_p\left(a\left(y - \sum_{s=1}^{r} x_{s,1} \cdots x_{s,r} \right)\right) > 0. \tag{5.31}$$

From the exponential sum result in [9] (to be precise, from a slightly more general version of Theorem 5 in [9]), for any $\tau > 0$ there is $r_1 = r_1(\tau)$ and $\tau_1 = \tau_1(\tau) > 0$ such that

$$\max_{(a,p)=1} \left| \sum_{x_i \in A_i} e_p(a x_1 \cdots x_{r_1}) \right| < p^{-\tau_1} |A_1| \cdots |A_{r_1}|, \tag{5.32}$$

whenever $A_1, \ldots, A_{r_1} \subset \mathbb{Z}_p$, $|A_i| > p^{\tau}$.

Consequently we have

$$\begin{aligned}
&\frac{1}{p}\sum_{a=0}^{p-1}\sum_{x_{s,s'}\in A_{s,s'}} e_p\left(a\left(y-\sum_{s=1}^{r}x_{s,1}\cdots x_{s,r}\right)\right)\\
&\quad=\frac{1}{p}\prod_{s,s'}|A_{s,s'}|+O\left(\max_{a\neq 0}\prod_{s=1}^{r}\left|\sum_{x_{s,s'}\in A_{s,s'}} e_p(ax_{s,1}\cdots x_{s,r})\right|\right)\\
&\quad>\left(\frac{1}{p}-p^{-r\tau_1}\right)\prod_{s,s'}|A_{s,s'}|>0, \qquad (5.33)
\end{aligned}$$

provided we take $r>\max(r_1,\frac{1}{\tau_1})$. □

5.5 Preserving large subfactors

Identify $\mathbb{Z}_q$ with $\prod_{j=1}^{J}\mathbb{Z}_{p_j}$. Fix $\tau=\eta$ and decompose $\{1,\dots,J\}=\mathcal{J}_1\cup\mathcal{J}_2$ where

$$\mathcal{J}_1=\left\{1\le j\le J\,|\,K_j>p_j^{\tau}\right\}, \qquad (5.34)$$

with $\{K_j\}_{1<j\le J}$ a sequence of positive integers in Lemma 5.2. Let $q=q_1\cdot q_2$ with $q_1=\prod_{j\in\mathcal{J}_1}p_j$ and $q_2=\prod_{j\in\mathcal{J}_2}p_j$. Take $r=r(\tau)$ according to Lemma 5.3.

We claim that

$$\pi_{q_1}(r(A')^r)=\pi_{q_1}(rA^r)=\mathbb{Z}_{q_1}. \qquad (5.35)$$

Denote A' by A. Let $j_1<j_2<\cdots<j_\beta$ be an enumeration of elements in $\mathcal{J}_1$. Fix $\xi_\alpha\in\mathbb{Z}_{p_\alpha}$, where $1\le\alpha\le\beta$. Since $\pi_{p_{j_1}}(A)\ge K_{j_1}>p_{j_1}^{\tau}$, applying Lemma 5.3 we have

$$\pi_{p_{j_1}}(rA^r)=r\pi_{p_{j_1}}(A)^r=\mathbb{Z}_{p_{j_1}}.$$

Therefore, there are elements $x_{s,s'}^{(1)}\in\pi_{p_1\cdots p_{j_1}}(A)$, such that

$$\pi_{p_{j_1}}\left(\sum_{s=1}^{r}\prod_{s'=1}^{r}x_{s,s'}^{(1)}\right)=\xi_1. \qquad (5.36)$$

Take $x_{s,s'}^{[1]}\in\pi_{p_1\cdots p_{j_2-1}}(A)$ with

$$\pi_{p_1\cdots p_{j_1}}\left(x_{s,s'}^{[1]}\right)=x_{s,s'}^{(1)}. \qquad (5.37)$$

Consider next the sets $\pi_{p_{j_2}}(A(x^{[1]}_{s,s'})) \subset \mathbb{Z}_{p_{j_2}}$ that are each of cardinality $K_{j_2} > p^{\tau}_{j_2}$, by (5.19). Hence again by Lemma 5.3

$$\pi_{p_{j_2}}\left(\sum_{s=1}^{r}\prod_{s'=1}^{r} A\big(x^{[1]}_{ss'}\big)\right) = \mathbb{Z}_{p_{j_2}}.$$

We can therefore obtain elements $x^{(2)}_{s,s'} \in \pi_{p_1\cdots p_{j_2}}(A)$ satisfying

$$\pi_{p_{j_2}}\left(\sum_{s=1}^{r}\prod_{s'=1}^{r} x^{(2)}_{s,s'}\right) = \xi_2.$$

Take again $x^{[2]}_{s,s'} \in \pi_{p_1\cdots p_{j_3-1}}(A)$ such that

$$\pi_{p_1\cdots p_{j_2}}\big(x^{[2]}_{ss'}\big) = x^{(2)}_{ss'}. \tag{5.38}$$

Consider the sets $\pi_{p_{j_3}}(A(x^{[2]}_{ss'})) \subset \mathbb{Z}_{p_{j_3}}$ of cardinality $K_{j_3} > p^{\tau}_{j_3}$ and repeat the construction.

After β steps, we obtain elements $x_{ss'} \in A$ $(1 \le s, s' \le r)$, such that for all $1 \le \alpha \le \beta$

$$\pi_{p_j\cdots p_{j_\alpha}}(x_{ss'}) = x^{(\alpha)}_{ss'} \tag{5.39}$$

with

$$\pi_{p_{j_\alpha}}\left(\sum_{s}\prod_{s'} x^{(\alpha)}_{ss}\right) = \xi_\alpha. \tag{5.40}$$

Hence

$$\pi_{p_{j_\alpha}}\left(\sum_{s}\prod_{s'} x_{ss'}\right) = \xi_\alpha \quad \text{for } 1 \le \alpha \le \beta, \tag{5.41}$$

where $\sum_s \prod_{s'} x_{ss'} \in r\,A^r$. This proves validity of (5.35).

Recalling (5.20), (5.19), we have

$$\prod_{j=1}^{J} K_j > \frac{|A|}{\prod_{j=1}^{J}(2\log p_j)} > \Omega_\varepsilon(q^{-\varepsilon})|A|, \tag{5.42}$$

and by (5.34) the left hand side of (5.42) is at most

$$q_1 \cdot \prod_{j\in\mathcal{J}_2} p^{\tau}_j < q_1 \cdot q^{\tau}.$$

Therefore, recalling that $\tau = \eta = \frac{\delta_1}{3}$ and that $\delta_2 \geq \delta_1$, we have

$$q_1 > \Omega_\varepsilon(q^{-\varepsilon}) q^{-\frac{\delta_1}{3}} |A| \overset{(1.2)}{>} \Omega_\varepsilon(q^{-\varepsilon}) q^{\delta_2 - \frac{\delta_1}{3}} > q^{\delta_2/2}. \tag{5.43}$$

Hence we have proved

Lemma 5.4 *There is $q_1|q$ such that $q_1 > q^{\delta_2/2}$ and*

$$\pi_{q_1}(r A^r) = \mathbb{Z}_{q_1}, \tag{5.44}$$

where $r = r(\delta_2)$.

Recalling assumption (1.2), the same claim holds for sets $\pi_{q'}(A)$ with $q'|q$ and $q' > q^\eta$ (just apply the preceding argument with q replaced by q' and A by $\pi_{q'}(A)$). Hence we have

Lemma 5.5 *Let $q'|q$ and $q' > q^\eta$. There is $q''|q'$ s.t. $q'' > (q')^{\delta_2/2}$ and*

$$\pi_{q''}(r A^r) = \mathbb{Z}_{q''}, \tag{5.45}$$

where $r = r(\delta_2)$.

5.6 Completion of the proof

Applying Lemma 5.4, we find $q_1|q$, $q_1 > q^{\delta_2/2}$ such that

$$\pi_{q_1}(r_1 A^{r_1}) = \mathbb{Z}_{q_1} \quad (r_1 = r_1(\delta_2)). \tag{5.46}$$

Recalling (5.18) and (1.1), we have

$$q_1 = |r_1 A^{r_1}| < O_{\delta_2,\varepsilon}(q^\varepsilon)|A| < O_{\delta_2,\varepsilon}(q^\varepsilon) q^{1-\delta_1}. \tag{5.47}$$

Write $q = q_1 \cdot q_1'$, where $q_1' > q^{\delta_1/2}$. Since $\eta = \frac{\delta_1}{3} < \frac{\delta_1}{2}$, we can apply Lemma 5.5 and obtain $q_1''|q_1'$, $q_1'' > (q_1')^{\delta_2/2}$, such that we also have

$$\pi_{q_1''}(r_1 A^{r_1}) = \mathbb{Z}_{q_1''}, \tag{5.48}$$

where $(q_1, q_1'') = 1$. The next problem we encounter is how, knowing (5.46), (5.48), we may significantly enlarge q_1 to a divisor q_2 of q, $q_1|q_2$, so that again

$$\pi_{q_2}(r_2 A^{r_2}) = \mathbb{Z}_{q_2} \quad (\text{with } r_2 = r_2(\delta_2)).$$

A (bounded) number of iterations will then lead to the required contradiction with (5.18) and (1.1).

This problem is taken care of in Sect. 8 of [5]; following the argument there closely, in Sect. 5.7 we prove the following

Proposition 5.1 *Let* $q_1|q$, $q_2|\frac{q}{q_1}$ *and* $S \subset \mathbb{Z}_q$ *such that* $\pi_{q_1}(S) = \mathbb{Z}_{q_1}$, $\pi_{q_2}(S) = \mathbb{Z}_{q_2}$. *Then there is* $Q|q_1q_2$, *such that*

$$Q > q_1 q_2^{\frac{1}{2}10^{-4}} \tag{5.49}$$

and

$$\pi_Q(400S^2) = \mathbb{Z}_Q. \tag{5.50}$$

Now taking q_1 as above,

$$q^{\frac{\delta_1}{2}} < q_1 < O_{\delta_2,\varepsilon}(q^{\varepsilon})q^{1-\delta_1},$$

and $q_2 = q_1''$ with $q_1''|\frac{q}{q_1}$, $q_1'' > q^{\frac{\delta_1}{2}}$ and $S = r_1 A^{r_1}$, using Proposition 5.1, we obtain Q_1 dividing q such that

$$Q_1 > q_1\left(\frac{q}{q_1}\right)^{\frac{\delta_2}{40000}}$$

and

$$\pi_{Q_1}(400S^2) = \mathbb{Z}_{Q_1}.$$

Now since $400S^2 \subset r_2 A^{r_2}$ with $r_2 = 400r_1^2$, we can repeat the procedure with q_1, r_1 replaced by Q_1, r_2.

Proceeding iteratively, we obtain at step i a divisor Q_i of q such that

$$Q_i > Q_{i-1}\left(\frac{q}{Q_{i-1}}\right)^{\frac{\delta_2}{40000}}$$

and

$$\pi_{Q_i}(r_{i+1}A^{r_{i+1}}) = \mathbb{Z}_{Q_i}.$$

Now choose i so that $Q_i > q^{1-\frac{\delta_1}{2}}$. Then, since (5.18) and (1.1) yield

$$Q_i = |r_{i+1}A^{r_{i+1}}| < O_{\varepsilon}(q^{\varepsilon})|A| < O_{\varepsilon}(q^{\varepsilon})q^{1-\delta_1},$$

we obtain a contradiction. This completes the proof of Theorem 1.3.

5.7 Proof or Proposition 5.1

We will make use of the following Lemmas, proven in Sect. 5.8.

Lemma 5.6 *Let A be a finite subset of an additive group Z and $\mathcal{G} \subset A \times A$, $0 < \alpha < \frac{1}{4}$, such that*

$$|\mathcal{G}| > (1-\alpha)|A|^2.$$

Then there exists a subset A' of A satisfying

$$|A'| > (1-\sqrt{\alpha})|A|$$

and

$$|A' + A'| < \frac{|A \overset{\mathcal{G}}{+} A|^4}{(1-\sqrt{\alpha})(1-2\sqrt{\alpha})^2|A|^3}.$$

The following lemma is Corollary 5.6 on p. 202 in [60] and is a consequence of Kneser's theorem (Theorem 5.5 on p. 200 in [60]).

Lemma 5.7 (Near-exact inverse sum set theorem) *Let A be a finite subset of an additive group Z such that*

$$|A + A| < \frac{3}{2}|A|.$$

Then there are $x \in Z$ and a subgroup G of Z, such that

$$A \subset x + G$$

and

$$|G| < \frac{3}{2}|A|.$$

Lemma 5.8 *Let q be square-free and suppose that $A \subset \mathbb{Z}_q$ satisfies $|A| > \gamma q$ with $\gamma > q^{-2/5}$. Then there is $q'|q$ such that*

$$\frac{q}{q'} < \gamma^{-\frac{20}{9}} \tag{5.51}$$

and

$$\pi_{q'}(100A \cdot A) = \mathbb{Z}_{q'}. \tag{5.52}$$

We now proceed to the proof of Proposition 5.1. The argument given below is slightly simpler than the one appearing in [5] and relies on Lemmas 5.6 and 5.7 (that were not used in [5]).

Let q_1 be a divisor of q with $\pi_{q_1}(S) = \mathbb{Z}_{q_1}$. Given $x \in \mathbb{Z}_{q_1}$ and a prime divisor p of $\frac{q}{q_1}$, let $\psi_p(x)$ denote an element of $\mathbb{Z}_p$ such that $(x, \psi_p(x)) \in \pi_{q_1 p}(S)$.

Claim 5.1 *For each divisor p of $\frac{q}{q_1}$ one of the following alternatives holds: either*

$$|\{(x, y) \in \mathbb{Z}_{q_1} \times \mathbb{Z}_{q_1} | \psi_p(x + y) \neq \psi_p(x) + \psi_p(y)\}| > 10^{-4} q_1^2; \tag{5.53}$$

or there is a subset $B \subset \mathbb{Z}_{q_1}$ such that

$$|B| > \frac{99}{100} q_1 \quad \text{and} \quad |\psi_p(B)| = 1. \tag{5.54}$$

Proof of Claim 5.1 For a prime divisor p of $\frac{q}{q_1}$ denote

$$\mathcal{G}_+ = \{(x, y) \in \mathbb{Z}_{q_1} \times \mathbb{Z}_{q_1} | \psi_p(x + y) = \psi_p(x) + \psi_p(y)\}.$$

Assume

$$|\mathcal{G}_+| > (1 - 10^{-4}) q_1^2. \tag{5.55}$$

Apply Lemma 5.6 taking $Z = \mathbb{Z}_{q_1} \times \mathbb{Z}_p \simeq \mathbb{Z}_{q_1 p}$ and

$$A = \{(x, \psi_p(x)) | x \in \mathbb{Z}_{q_1}\}, \qquad |A| = q_1,$$

$$\mathcal{G} = \{((x, \psi_p(x)), (y, \psi_p(y))) | (x, y) \in \mathcal{G}_+\} \subset A \times A.$$

By (5.55) we have

$$|\mathcal{G}| > (1 - 10^{-4}) |A|^2$$

and from the definition of $\mathcal{G}_+$

$$|A \overset{\mathcal{G}}{+} A| \leq |A|.$$

According to Lemma 5.6 applied with $\alpha = 10^{-4}$, we obtain a subset B of $\mathbb{Z}_{q_1}$ such that

$$|B| > \frac{99}{100} q_1 \tag{5.56}$$

and

$$|\{(x + y, \psi_p(x) + \psi_p(y)) | x, y \in B\}| < \beta |B| \tag{5.57}$$

where

$$\beta = \frac{1}{(1 - \frac{1}{100})(1 - \frac{1}{50})^2} \frac{100}{99} < \frac{3}{2}. \tag{5.58}$$

Next apply Lemma 5.7 to the set

$$A' = \{(x, \psi_p(x)) | x \in B\} \subset \mathbb{Z}_{q_1} \times \mathbb{Z}_p$$

for which by (5.57), (5.58) we have

$$|A' + A'| < \frac{3}{2}|A'|.$$

Hence A' is contained in a translate of a subgroup H of $\mathbb{Z}_{q_1} \times \mathbb{Z}_p$ with $|H| < \frac{3}{2}|A'| \le \frac{3}{2}q_1$. Since p and q_1 are relatively prime, H is of the form $H = H_1 \times H_0$ with $H_1 < \mathbb{Z}_{q_1}$, $H_0 < \mathbb{Z}_p$. Also

$$|H_1| = |\pi_{q_1}(H)| \ge |\pi_{q_1}(A')| = |B| > \frac{99}{100}q_1$$

so that $H_1 = \mathbb{Z}_{q_1}$. Consequently $|H_0| \le \frac{3}{2}q_1|H_1|^{-1} = \frac{3}{2}$ and $H_0 = \{0\}$; therefore $|\psi_p(B)| = 1$. This completes the proof of Claim 5.1. □

Take next $q_2 | \frac{q}{q_1}$, such that also $\pi_{q_2}(S) = \mathbb{Z}_{q_2}$ and let $q_2 = p_1 \cdots p_\ell$. For each p_i dividing q_2, one of the alternatives in Claim 5.1 holds, yielding a factorization $q_2 = q_2^{(1)} q_2^{(2)}$, with $q_2^{(1)}$ being a product of primes satisfying (5.53), and $q_2^{(2)}$ being a product of primes satisfying (5.54). Clearly, either $q_2^{(1)} \ge q_2^{1/2}$ or $q_2^{(2)} > q_2^{1/2}$. We now show that the conclusion of Proposition 5.1 holds in each of these two cases.

Case 1. Assume $q_2^{(1)} \ge q_2^{1/2}$. Let

$$D_i = \{(x, y) \in \mathbb{Z}_{q_1} \times \mathbb{Z}_{q_1} | \psi_{p_i}(x + y) \ne \psi_{p_i}(x) + \psi_{p_i}(y)\}, \tag{5.59}$$

so that $|D_i| > 10^{-4} q_1^2$ for $p_i | q_2^{(1)}$. Thus

$$\sum_{p_1 | q_2^{(1)}} \log p_i |D_i| > 10^{-4} q_1^2 \sum_{p_1 | q_2^{(1)}} \log p_i,$$

which we can rewrite as

$$\frac{1}{|\mathbb{Z}_{q_1} \times \mathbb{Z}_{q_1}|} \sum_{x \in \mathbb{Z}_{q_1} \times \mathbb{Z}_{q_1}} \sum_{p_1 | q_2^{(1)}} \log p_i \chi_{D_i}(x) > 10^{-4} \log q_2^{(1)}.$$

Therefore, for some $x \in \mathbb{Z}_{q_1} \times \mathbb{Z}_{q_1}$ we have

$$\sum_{p_1 | q_2^{(1)}} \log p_i \chi_{D_i}(x) > 10^{-4} \log q_2^{(1)}.$$

Consequently, denoting $x = (a, b)$ with $a \in \mathbb{Z}_{q_1}$, $b \in \mathbb{Z}_{q_1}$, and letting $I = \{i | (a, b) \in D_i\}$ we obtain

$$\sum_{\substack{p_i | q_2^{(1)} \\ i \in I}} \log p_i > 10^{-4} \log q_2^{(1)}.$$

Hence, keeping in mind our assumption $q_2^{(1)} \geq q_2^{1/2}$, we have

$$\bar{q}_2 = \prod_{\substack{p_i | q_2^{(1)} \\ i \in I}} p_i > \left(q_2^{(1)}\right)^{10^{-4}} > (q_2)^{\frac{1}{2}10^{-4}}. \tag{5.60}$$

Denoting

$$\bar{a} = (a, \psi(a)) \in S \subset \mathbb{Z}_{q_1} \times \mathbb{Z}_{q/q_1}, \tag{5.61}$$

$$\bar{b} = (b, \psi(b)) \in S, \tag{5.62}$$

$$\overline{a+b} = (a+b, \psi(a+b)) \in S, \tag{5.63}$$

it follows from the definition (5.59) of D_i that

$$\overline{a+b} - \bar{a} - \bar{b} \neq 0 \bmod (p_i) \quad \text{for } i \in I,$$

while obviously

$$\overline{a+b} - \bar{a} - \bar{b} = 0 \bmod (q_1).$$

Thus, since

$$\pi_{q_1}(S^2) = \pi_{q_1}(S) = \mathbb{Z}_{q_1} \quad \text{and} \quad \pi_{\bar{q}_2}(S) = \mathbb{Z}_{\bar{q}_2},$$

we have

$$\begin{aligned} &\pi_{q_1\bar{q}_2}(S^2 + (\overline{a+b} - \bar{a} - \bar{b})S) \\ &\quad = \{(\pi_{q_1}(xx'), \pi_{\bar{q}_2}(xx') + \pi_{\bar{q}_2}(\overline{a+b} - \bar{a} - \bar{b})\pi_{\bar{q}_2}(y)) | x, x', y \in S\} \\ &\quad = \mathbb{Z}_{q_1} \times \mathbb{Z}_{\bar{q}_2} \end{aligned} \tag{5.64}$$

since $\pi_{\bar{q}_2}(\overline{a+b} - \bar{a} - \bar{b}) \in \mathbb{Z}_{\bar{q}_2}^*$.

Therefore

$$\pi_{q_1\bar{q}_2}(S^2 + (S - S - S)S) = \mathbb{Z}_{q_1\bar{q}_2},$$

and the conclusion of Proposition 5.1 is established in this case.

Case 2. $q_2^{(2)} > q_2^{1/2}$. Let $I = \{i | p_i \text{ divides } q_2^{(2)}\}$. For $i \in I$, there is $B_i \subset \mathbb{Z}_{q_1}$ such that $|B_i| > \frac{99}{100}\, q_1$ and

$$|\psi_{p_i}(B_i)| = 1. \tag{5.65}$$

Therefore we have

$$\frac{1}{|\mathbb{Z}_{q_1}|} \sum_{x \in \mathbb{Z}_{q_1}} \sum_{i \in I} (\log p_i) \chi_{B_i}(x) > \frac{99}{100} \big(\log q_2^{(2)}\big).$$

Applying Jensen's inequality we obtain

$$\begin{aligned} \frac{1}{|\mathbb{Z}_{q_1}|} \sum_{x \in \mathbb{Z}_{q_1}} \left[\prod_{i \in I} p_i^{\chi_{B_i}(x)} \right] &= \frac{1}{|\mathbb{Z}_{q_1}|} \sum_{x \in \mathbb{Z}_{q_1}} \exp\left\{ \sum_{i \in I} (\log p_i) \chi_{B_i}(x) \right\} \\ &> \big[q_2^{(2)}\big]^{99/100}. \end{aligned}$$

Decomposing for each $i \in I$, $\mathbb{Z}_{q_1} = B_i \cup B_i^c$, we can rewrite the expression on the left-hand side as the sum of $2^{|I|}$ terms:

$$\frac{1}{|\mathbb{Z}_{q_1}|} \sum_{x \in \mathbb{Z}_{q_1}} \sum_{\varepsilon_i \in \{0,1\}^{|I|}} \left[\prod_{i \in I} p_i^{\varepsilon_i} (\varepsilon_i \chi_{B_i}(x) + (1 - \varepsilon_i) \chi_{B_i^c}(x)) \right];$$

consequently, by the pigeonhole principle, for some choice of $(\tilde{\varepsilon}_1, \ldots \tilde{\varepsilon}_I) \in \{0, 1\}^I$ we have

$$\begin{aligned} &\frac{1}{|\mathbb{Z}_{q_1}|} \sum_{x \in \mathbb{Z}_{q_1}} \left[\prod_{i \in I} p_i^{\tilde{\varepsilon}_i} (\tilde{\varepsilon}_i \chi_{B_i}(x) + (1 - \tilde{\varepsilon}_i) \chi_{B_i^c}(x)) \right] \\ &\quad > 2^{-|I|} \big[q_2^{(2)}\big]^{99/100} > \big[q_2^{(2)}\big]^{9/10}. \end{aligned} \tag{5.66}$$

Thus, letting

$$B = \bigcap_{i \in I, \tilde{\varepsilon}_i = 1} B_i \quad \text{and} \quad \bar{q}_2 = \prod_{\tilde{\varepsilon}_i = 1} p_i, \quad \bar{q}_2 | q_2^{(2)},$$

it follows from (5.66) that

$$|B| \bar{q}_2 > \big[q_2^{(2)}\big]^{9/10} \cdot q_1. \tag{5.67}$$

If $p_i | \bar{q}_2$, then by (5.65)

$$|\psi_{p_i}(B)| = 1.$$

Therefore we may specify for each $p_i|\bar{q}_2$ an element $u_i \in \mathbb{Z}_{p_i}$, such that

$$\pi_{p_i}\psi(x) = u_i \quad \text{for } p_i|\bar{q}_2 \text{ and } x \in B. \tag{5.68}$$

Consider next $\bar{B} = \{(x, \psi(x))\big|x \in B\}$ and write

$$\pi_{q_1\bar{q}_2}(\bar{B} + S) = \{(x + \pi_{q_1}(y), u + \pi_{\bar{q}_2}(y))|x \in B, y \in S\}, \tag{5.69}$$

where $\pi_{\bar{q}_2}(S) = \mathbb{Z}_{\bar{q}_2}$. Hence, by (5.67), (5.68)

$$|\pi_{q_1\bar{q}_2}(S + S)| \geq |B|\bar{q}_2 > \left[q_2^{(2)}\right]^{9/10}.q_1. \tag{5.70}$$

Applying Lemma 5.8 to the set $\pi_{q_1q_2^{(2)}}(2S) \subset \mathbb{Z}_{q_1q_2^{(2)}}$ with $\gamma = [q_2^{(2)}]^{-\frac{1}{10}}$, we obtain Q dividing $q_1q_2^{(2)}$ such that

$$Q > q_1\left[q_2^{(2)}\right]^{\frac{7}{9}} > q_1q_2^{\frac{7}{18}} \tag{5.71}$$

and

$$\pi_Q(400S^2) = \mathbb{Z}_Q. \tag{5.72}$$

Therefore the conclusions (5.49), (5.50) hold in this case as well, and the proof of Proposition 5.1 is complete.

5.8 Proofs of Lemmas 5.6–5.8

The proof of Lemma 5.6 is based on the following lemma (Lemma 5.9), which is Exercise 2.5.4 on p. 82 of [60]; for completeness we supply the proof.

Lemma 5.9 *Let A, B, C be additive sets in an ambient group Z, let $0 < \alpha < 1/4$, and let $G \subset A \times B$, $H \subset B \times C$ be such that $|G| \geq (1-\alpha)|A||B|$ and $|H| \geq (1-\alpha)|B||C|$. Then there are subsets $A' \subseteq A$ and $C' \subseteq C$ with $|A'| \geq (1-\sqrt{\alpha})|A|$ and $|C'| \geq (1-\sqrt{\alpha})|C|$ such that*

$$|A' - C'| \leq \frac{|A \overset{G}{-} B|\,|B \overset{H}{-} C|}{(1-2\sqrt{\alpha})|B|}, \tag{5.73}$$

where

$$A \overset{\mathcal{G}}{-} B = \{a - b|(a, b) \in \mathcal{G}\}.$$

Proof of Lemma 5.9 We first show that at most $\sqrt{\alpha}|B|$ elements of B have a G-degree of less than $(1-\sqrt{\alpha})|A|$. Let m be the number of elements b in B of G-degree, $\deg(b) > (1-\sqrt{\alpha})|A|$. Then

$$(1-\alpha)|A||B| \le |G| = \sum_{b:\deg(b)\le(1-\sqrt{\alpha})|A|} \deg(b) + \sum_{b:\deg(b)>(1-\sqrt{\alpha})|A|} \deg(b)$$

$$\le (1-\sqrt{\alpha})|A|(|B|-m) + m|A|,$$

therefore

$$(1-\alpha)|B| \le (1-\sqrt{\alpha})(|B|-m) + m$$

and

$$(1-\sqrt{\alpha})|B| \le m.$$

Similarly, we have that at most $\sqrt{\alpha}|B|$ elements of B have an H-degree of less than $(1-\sqrt{\alpha})|C|$. Consequently, at least $(1-2\sqrt{\alpha})|B|$ elements of B have a G-degree of at least $(1-\sqrt{\alpha})|A|$ and an H-degree of at least $(1-\sqrt{\alpha})|C|$; let B' be a subset of B satisfying these properties and let A' (respectively C') be a subset of A (respectively of C) connected to elements of B' in G (respectively in H). Clearly, we have that $|A'| \ge (1-\sqrt{\alpha})|A|$ and $|C'| \ge (1-\sqrt{\alpha})|C|$. From the identity

$$a'-c' = (a'-b') + (b'-c')$$

we see that every element $a'-c'$ in $A'-C'$ has at least $|B'| = (1-2\sqrt{\alpha})|B|$ distinct representations of the form $x+y$ with $(x,y) \in (A \overset{G}{-} B) \times (B \overset{H}{-} C)$, completing the proof of Lemma 5.9. □

Proof of Lemma 5.6 Take $A = C$, $B = -A$ and $G = \{(x,-y)|(x,y) \in \mathcal{G}\} \subset A \times B$, $H = \{(-x,y)|(x,y) \in \mathcal{G}\} \subset B \times C$. Using Lemma 5.9 we obtain subsets $A' \subset A$, $C' \subset A$ such that $|A'| > (1-\sqrt{\alpha})|A|$, $|C'| > (1-\sqrt{\alpha})|A|$ and

$$|A'-C'| \le \frac{|A \overset{G}{-} B|\,|B \overset{H}{-} C|}{(1-2\sqrt{\alpha})|B|} = \frac{|A \overset{\mathcal{G}}{+} A|^2}{(1-2\sqrt{\alpha})|A|}.$$

Applying Ruzsa's triangle inequality

$$|A'+A'| \le \frac{|A'-C'|^2}{|C'|}$$

the statement follows. □

In order to prove Lemma 5.8 we first establish the following result.

Lemma 5.10 *Let A be a subset of $\mathbb{Z}_q$ (q arbitrary) satisfying the following property*:

$$\forall q_1|q, \text{ and } z \in \mathbb{Z}_{q_1}, \ \#\{x \in A|\pi_{q_1}(x) = z\} \leq q_1^{-11/20}|A|. \tag{5.74}$$

Then

$$\mathbb{Z}_q = 100A \cdot A. \tag{5.75}$$

Proof of Lemma 5.10 Note that

$$100A \cdot A = \{x_1x_2 + \cdots + x_{199}x_{200}|x_i \in A\}.$$

Our aim is to show that for all $\xi \in \mathbb{Z}_q$

$$\#\Big\{(x_1, \ldots, x_{200}) \in \underbrace{A \times \cdots \times A}_{200}|x_1x_2 + \cdots + x_{199}x_{200} = \xi\Big\} > 0. \tag{5.76}$$

Proceeding by the circle method we have

$$\begin{aligned}
&\#\Big\{(x_1, \ldots, x_{200}) \in \underbrace{A \times \cdots \times A}_{200}|x_1x_2 + \cdots + x_{199}x_{200} = \xi\Big\} \\
&= \frac{1}{q}\sum_{0\leq z<q}\Big[\sum_{x,y\in A} e_q(zxy)\Big]^{100} e_q(-\xi z) \\
&> \frac{1}{q}|A|^{200} - \frac{1}{q}\sum_{\substack{q_1|q \\ q_1>1}}\sum_{z\in\mathbb{Z}_{q_1}^*}\Big|\sum_{x,y\in A} e_{q_1}(zxy)\Big|^{100}.
\end{aligned} \tag{5.77}$$

Fix $q_1|q$ and denote, for $z \in \mathbb{Z}_{q_1}$,

$$\eta(z) = \#\{x \in A|\pi_{q_1}(x) = z\}.$$

For any $z \in \mathbb{Z}_m^*$, and any two functions f, g on $\mathbb{Z}_m$, a simple application of Cauchy-Schwarz inequality yields

$$\Big|\sum_{x\in\mathbb{Z}_m}\sum_{y\in\mathbb{Z}_m} f(x)g(y)e_m(xyz)\Big| \leq \Big(m\sum_{x\in\mathbb{Z}_m} f^2(x)\sum_{y\in\mathbb{Z}_m} g^2(x)\Big)^{\frac{1}{2}}. \tag{5.78}$$

Applying (5.78) with $f = g = \eta$ and $m = q_1$ we obtain for any $z \in \mathbb{Z}_{q_1}^*$:

$$\begin{aligned}\left|\sum_{x,y\in A} e_{q_1}(zxy)\right| &= \left|\sum_{x_1,y_1\in\pi_{q_1}(A)} \eta(x_1)\eta(y_1)e_{q_1}(zx_1y_1)\right| \\ &\le \left|\sum_{x_1\in\mathbb{Z}_{q_1}}\sum_{y_1\in\mathbb{Z}_{q_1}} \eta(x_1)\eta(y_1)e_{q_1}(zx_1y_1)\right| \\ &\le \sqrt{q_1}\sum_{x_1\in\mathbb{Z}_{q_1}} \eta(x_1)^2. \end{aligned} \tag{5.79}$$

Now assumption (5.74) implies

$$\|\eta\|_\infty < q_1^{-11/20}|A|, \tag{5.80}$$

while we clearly have

$$\|\eta\|_1 = \sum_{z\in\mathbb{Z}_{q_1}} \eta(z) \le |A|. \tag{5.81}$$

Consequently, using

$$\|\eta\|_2 \le \|\eta\|_1\|\eta\|_\infty,$$

we obtain

$$\left|\sum_{x,y\in A} e_{q_1}(zxy)\right| \le q_1^{-1/20}|A|^2. \tag{5.82}$$

Substitution of (5.82) in (5.77) implies that the right-hand side in (5.77) is greater than

$$\frac{1}{q}|A|^{200} - \frac{1}{q}\sum_{\substack{q_1|q\\ q_1>1}} \varphi(q_1)q_1^{-5}|A|^{200} > \frac{1}{q}|A|^{200}\left(1 - \sum_{j=2}^{\infty} j^{-4}\right) > 0,$$

establishing (5.76) and completing the proof of Lemma 5.10 □

Proof of Lemma 5.8 If the condition (5.74) of Lemma 5.10 holds, then the conclusion of Lemma 5.10 clearly follows. Assume that condition (5.74) of Lemma 5.10 fails. Then for some $q_1|q$ there is $z_1 \in \mathbb{Z}_{q_1}$, such that $|A_1| > q_1^{-11/20}|A|$, where $A_1 = \{x \in A | \pi_{q_1}(x) = z_1\}$. Since $|\pi_{q_1}(A)| = 1$ and $|A_1| > q_1^{-11/20}\gamma q$, we have that

$$\frac{q}{q_1} > \gamma q q_1^{-11/20}$$

and so

$$q_1 < \gamma^{-20/9} < q^{8/9}.$$

Replace q by $\frac{q}{q_1}$ and A by $A' = \pi_{\frac{q}{q_1}}(A_1)$. If (5.74) fails again, there is $q_2 | \frac{q}{q_1}$ and $z_2 \in \mathbb{Z}_{q_2}$, such that $|A_2| > q_2^{-11/20}|A_1|$, where $A_2 = \{x \in A_1 | \pi_{q_2}(x) = z_2\}$. If $q_1, \ldots, q_s$ are the consecutive divisors of q obtained after s steps, then by construction

$$\frac{q}{q_1 \cdots q_s} \geq |A_s| \geq (q_1 \cdots q_s)^{-\frac{11}{20}} |A| > \gamma (q_1 \cdots q_s)^{-\frac{11}{20}} q,$$

implying that

$$q_1 \cdots q_s < \gamma^{-\frac{20}{9}} < q^{8/9}. \tag{5.83}$$

Clearly this construction terminates after finitely many steps s, resulting in $q' = \frac{q}{q_1 \ldots q_s}$ satisfying the bound (5.51), and in a set A_s satisfying

$$\#\{x \in A_s | \pi_{q_{s+1}}(x) = z\} \leq q_{s+1}^{-11/20} |A_s| \quad \text{for all } q_{s+1} | q' \text{ and } z \in \mathbb{Z}_{q_{s+1}}.$$

Application of Lemma 5.10 to $\pi_{q'}(A_s) \subset \mathbb{Z}_{q'}$ gives

$$\mathbb{Z}_{q'} = 100 \pi_{q'}(A_s) \cdot \pi_{q'}(A_s),$$

implying (5.52) and completing the proof of Lemma 5.8. □

6 Explicit applications

We give explicit applications of our main theorems. We stick to forms of SL_2 and their orbits since for the time being these are the only cases for which we have established Conjecture 1.5. Once the general form of Conjecture 1.5 is proven, then using Theorem 1.1 and the passage from simply connected groups to other groups and their orbits (as is done below for SL_2) one can establish saturation for quite general pairs $(\mathcal{O}, f)$.

Our basic example is SL_2 itself. That is $G = \mathrm{SL}_2$ sitting in Mat_2, the affine 4 dimensional space of 2×2 matrices. As we have noted with $G = \{X : \det X = 1\}$, $\mathbb{Q}[G]$ is a unique factorization domain.

Theorem 6.1 *Let Λ be a subgroup of $\mathrm{SL}_2(\mathbb{Z})$ which is Zariski dense in G and let $f \in \mathbb{Q}[G]$ be integral and primitive on Λ. Assume that f is nonconstant when restricted to G and that the factors of f are irreducible in $\bar{\mathbb{Q}}[G]$. Then $r_0(\Lambda, f) < \infty$.*

Proof This is an immediate consequence of Theorems 1.1 and 1.2 coupled with the fact that the expander property in Theorem 1.2 is valid for q of the form $N^\beta d$ with $\beta = 0$ or 1 and d is squarefree (these being the q's that are used in the proof of Theorem 1.1). This more general case follows from the proof of the squarefree case.

We also note that the assumption that the factors of f are absolutely irreducible was made for convenience. One can drop this assumption and still deduce Theorem 6.1. This involves using the Chebotarev theorem in a more quantitative way than in the proof of Proposition 3.2, so as to determine the behavior of the sum over p in (3.46) coming from a modified (according to the finite extension of $\mathbb{Q}$ which splits f) form of (3.6) and (3.42). □

A related basic example which we can handle is that of a quaternion division algebra in place of the matrix algebra $\mathrm{Mat}_2/\mathbb{Q}$. Let $D/\mathbb{Q}$ be such an algebra. D is linearly generated over $\mathbb{Q}$ by $1, \omega, \Omega, \omega\Omega$, where $\omega^2 = a$, $\Omega^2 = b$ with a, b nonzero integers. The elements $1, \omega, \Omega, w\Omega$ satisfy the usual rules for multiplication of quaternions. Let N denote the reduced norm on D and let D_1 denote the elements α with $N(\alpha) = 1$. By $D(\mathbb{Z})$ we mean the subring of elements $\alpha \in D$ of the form $\alpha = x_1 + x_2\omega + x_3\Omega + x_4\omega\Omega$ with $x_j \in \mathbb{Z}$. This is not a maximal order, but it is of finite index in such and this suffices for our purposes. Let $D_1(\mathbb{Z})$ be the corresponding unit group, that is elements $\alpha \in D(\mathbb{Z})$ with $N(\alpha) = 1$. This group is infinite iff $D \otimes \mathbb{R}$ is the matrix algebra $M_2(\mathbb{R})$ which we will assume is the case. D_1 is an algebraic group defined over $\mathbb{Q}$ and in terms of the coordinates $(x_1, x_2, x_3, x_4) \in \mathbb{A}^4$ it is given by $N(x) = x_1^2 - ax_2^2 - bx_2^3 + abx_4^2 = 1$.

Theorem 6.1′ *Let Λ be a subgroup of $D_1(\mathbb{Z})$ and assume Λ is Zariski dense in D_1. Let $f \in \mathbb{Q}[D_1]$ be primitive integral and nonconstant on Λ, then $r_0(\Lambda, f) < \infty$.*

Proof The proof is the same as that of Theorem 6.1. Note that D_1 is connected and simply connected so that Theorem 1.1 applies. While Theorem 1.2 does not apply directly to $D_1(\mathbb{Z})$ the proof of that theorem does. That is for p outside a finite set of primes we have $D_1(\mathbb{Z})_p = D_1(\mathbb{F}_p) \approxeq \mathrm{SL}_2(\mathbb{F}_p)$ and an inspection of the proof of Theorem 1.2 shows that this and the product structure for $D_1(\mathbb{Z})_d$ for $d = d_1d_2$, $(d_1, d_2) = 1$ is all that was used.

When the quaternion algebra D splits over $\mathbb{Q}$, that is when it is the full matrix algebra $\mathrm{Mat}_{2\times 2}$ then D_1 is essentially SL_2 and Theorem 6.1′ becomes Theorem 6.1. We allow both of these cases for D and D_1 in what follows. Let $\pi : D_1 \to \mathrm{GL}_n$ be a rational representation of D_1 into GL_n defined over $\mathbb{Q}$. The matrix entries of $\pi(g)$ are polynomials with rational coefficients in the coordinates (x_1, x_2, x_3, x_4) of g. Denote by G the matrix algebraic group $\pi(D_1)$. It is a subgroup of GL_n, defined over $\mathbb{Q}$ and it is connected. Let Γ be

a subgroup of $G(\mathbb{Z}) = G \cap \mathrm{GL}_n(\mathbb{Z})$ which is Zariski dense in G. Fix $b \in \mathbb{Z}^n$ and denote by $\mathcal{O}$ the orbit $b\Gamma$ in $\mathbb{A}^n$. □

Theorem 6.2 *Let G, Γ and $\mathcal{O}$ be as above and let $f \in \mathbb{Q}[x_1, \ldots, x_n]$ with f integral, primitive, and nonconstant on $\mathcal{O}$. Then $r_0(\mathcal{O}, f) < \infty$.*

Proof By composition we have that $F(x_1, x_2, x_3, x_4) = f(b\pi(g))$ is in $\mathbb{Q}[x_1, x_2, x_3, x_4]$. Now $\pi(D_1(\mathbb{Z}))$ is commensurable with $G(\mathbb{Z})$ (see [2]) and hence $\Delta = \pi(D_1(\mathbb{Z})) \cap \Gamma$ is of finite index in Γ and is Zariski dense in G. Thus without loss of generality we can assume that $\Gamma \subset \pi(D_1(\mathbb{Z}))$. Set $\Lambda = \pi^{-1}(\Gamma)$, then Λ is a subgroup of $D_1(\mathbb{Z})$ and F is integral primitive and nonconstant on Λ. Now $\ker \pi$ is finite (since it is a proper normal subgroup of D_1 and we are assuming that G is not trivial). Hence Λ is Zariski dense in D_1. Applying Theorem 6.1 (or 6.1′) yields that either F is constant on Λ, or there is an $r < \infty$ such that the set of $x \in \Lambda$, call it P, for which $F(x)$ is a product of at most r primes, is Zariski dense in D_1. If F is constant on Λ, then it is constant on D_1 and hence f is constant of $\mathrm{Zcl}(\mathcal{O})$ which we assumed was not the case. Hence we are in the first case and $\pi(P)$ is contained in Γ and $b\pi(P)$ is contained in $\mathcal{O}$. To complete the proof we need only to show that $\mathrm{Zcl}(b\pi(P)) = \mathrm{Zcl}(bG)$ in $\mathbb{A}^n$ since f is a product of at most r primes at these points. Now in the topology of GL_n we have that

$$G = \pi(D_1) = \pi(\mathrm{Zcl}(P)) \subset \mathrm{Zcl}(\pi(P)) \subset \mathrm{Zcl}(\pi(D_1)) = G.$$

Hence $\mathrm{Zcl}(\pi(P)) = G$. Also

$$\mathrm{Zcl}(b\pi(P)) \supset b\mathrm{Zcl}(\pi(P)) = bG.$$

Hence

$$\mathrm{Zcl}(b\pi(P)) = \mathrm{Zcl}(bG),$$

which completes the proof of Theorem 6.2. □

We explicate some instances of Theorem 6.2 with concrete examples.

Example A This is connected with Conjecture 1.3. Let π be the standard representation of SL_2 by linear action. If $b \in \mathbb{Z}^2$, $b \neq 0$ and Λ is a non-elementary subgroup of $\mathrm{SL}_2(\mathbb{Z})$ then the orbit $\mathcal{O} = b \cdot \Lambda$ is Zariski dense in $\mathbb{A}^2$. Let $f \in \mathbb{Q}[x_1, x_2]$ be a nonconstant polynomial which is integral and primitive on $\mathcal{O}$. Then according to Theorem 6.2, $(\mathcal{O}, f)$ saturates. With $f(x) = x_1 x_2$ this yields an approximation ("almost prime") to Conjecture 1.3.

Example B The next set of examples are associated with ternary integral quadratic forms. Let $F(x_1, x_2, x_3)$ be a regular such quadratic form which

is indefinite over $\mathbb{R}$. Let $G = \mathrm{SO}_F \subset \mathrm{GL}_3$ be the corresponding special orthogonal group preserving F. If $F(x) = x^t Ax$ with A symmetric, then G is given as a matrix group defined over $\mathbb{Q}$ by the 3×3 matrices X satisfying

$$\left.\begin{array}{c} X^t AX = A \\ \det X = 1 \end{array}\right\}. \tag{6.1}$$

G is not simply connected, but the simply connected covering group $\tilde{G}$ is a double cover. It can be realized as the norm 1 group in a quaternion algebra D defined over $\mathbb{Q}$. This is described explicitly in (2.3) and the general case is described in [13]. D is Mat_2 if F is isotropic over $\mathbb{Q}$ and it is a division algebra in the anisotropic case. In either case, $G = \pi(D_1)$ with π this covering morphism, and Theorem 6.2 can be applied.

Let Γ be a subgroup of $G(\mathbb{Z})$ which is Zariski dense in G. Let $b \in \mathbb{Z}^3$, $b \neq 0$, for which $F(b) = k$ and let $\mathcal{O} = b \cdot \Gamma$. Then if $k \neq 0$, $\mathrm{Zcl}(\mathcal{O}) = b \cdot G$ and is the affine quadric V_k given by $\{x : F(x) = k\}$. As usual we conclude that if $f \in \mathbb{Q}[x_1, x_2, x_3]$ is nonconstant, primitive and integral then $(\mathcal{O}, f)$ saturates.

This result is even interesting when applied to the full group $\Lambda = G(\mathbb{Z})$. In this case if $V_k(\mathbb{Z})$ is a finite union of $G(\mathbb{Z})$ orbits and one deduces that $r_0(V_k(\mathbb{Z}), f) < \infty$, In as much as our proof of Theorem 1.2 gives no explicit bound for the expansion our proof yields no explicit bound for $r_0(V_k(\mathbb{Z}), f)$. In this case where $\Lambda = G(\mathbb{Z})$ one can use the theory of automorphic forms to address the expansion. Instead of using combinatorial ordering of the orbit as in Theorem 1.1 one can apply a much more efficient Archimedean weighted ordering on the quadric and a corresponding sharp quantitative analysis. This is carried out in [41] where it is shown that for $f(x) = x_1 x_2 x_3$, $r_0(V_k(\mathbb{Z}), f) \leq 26$ (for any F) as long as $V_k(\mathbb{Z}) \neq \emptyset$.

Example C In the case that $k = 0$ in (B) and F is isotropic over $\mathbb{Q}$, V_0 is an affine cone. We restrict to the specific case that

$$F(x_1, x_2, x_3) = x_1^2 + x_2^2 - x_3^2 \tag{6.2}$$

and in the tradition of Fermat examine what Theorem 6.2 gives in this case. A point in $V_0(\mathbb{Z})$ with $\gcd(x_1, x_2, x_3) = 1$ is a Pythagorean triple (or a Pythagorean triangle if x_1, x_2, x_3 are positive) (see [28, 58]). The group $\mathrm{SO}_F(\mathbb{Z})$ acts transitively on the set T of all such Pythagorean triples. Consider the ancient problem of the divisibility properties of the area $A(x_1, x_2, x_3) = \frac{x_1 x_2}{2}$ of such a triangle. It is elementary (see below) that $f = A/6$ is integral on the set T. Note that $f(3, 4, 5) = 1$ and hence $(\mathcal{O}, f)$ is integral and primitive for any orbit $\mathcal{O} = (3, 4, 5)\Lambda$ where Λ is a Zariski dense subgroup of $G = \mathrm{SO}_F$. Hence by Theorem 6.2 we have that $r_0(\mathcal{O}, f) < \infty$ for any such orbit $\mathcal{O}$.

The question of the value of $r_0(\mathcal{O}, f)$ for this orbit and f is of interest as it gives the minimal divisibility of the areas of a Zariski dense (in V_0) set of Pythagorean triangles in $\mathcal{O}$. We show that Conjecture 1.4 implies that $r_0(\mathcal{O}, f) = 4$, that is given any $\mathcal{O}$ as above, the set of $x \in \mathcal{O}$ whose areas have 6 prime factors (including 2 and 3) is Zariski dense in V_0, while those with 5 or fewer prime factors is not Zariski dense. To see this recall the standard parametrization of triples x in T (after switching x_1 and x_2 if need be):

$$x_1 = m^2 - n^2, \qquad x_2 = 2mn, \qquad x_3 = m^2 + n^2, \tag{6.3}$$

where $(m, n) = 1$ and m and n are of different parity. From this it is clear that x_2 is divisible by 4 and one of x_1 or x_2 is divisible by 3. Hence $A = \frac{x_1 x_2}{2}$ is divisible by 6 and hence f is integral on T. From (6.3) it is also clear that the set of $x \in T$ for which $f(x) = \frac{1}{6}(m-n)(m+n)mn$ has at most 2 prime factors is finite. The set of $x \in T$ for which $f(x)$ has at most 3 prime factors is probably infinite, in fact this would follow from Conjecture 1.2 of Hardy and Littlewood for the case of Λ of rank 1 in $\mathbb{Z}^4$. However even if this set is infinite it is not Zariski dense in V_0 since such triangles are of special form and lie on a finite number of curves in V_0. Hence $r_0(T, f) \geq 4$ and a fortiori $r_0(\mathcal{O}, f) \geq 4$.

In order to apply Conjecture 1.4 we proceed by pull-back from $G = \mathrm{SO}_F$ to its double cover SL_2. We can describe $\pi : \mathrm{SL}_2 \to G$ in coordinates similar to those in (2.3) and we find that the pullback $f^* \in \mathbb{Q}[\mathrm{SL}_2]$ is given by

$$f^*(x_1, x_2, x_3, x_4) = \frac{(2x_1 + x_2 + 2x_3 - x_4)(2x_1 + x_2 + 2x_3 + x_4)(2x_1 + x_2)(2x_3 + x_4)}{6} \tag{6.4}$$

and $\Gamma = \pi^{-1}(\Lambda) \leq \mathrm{SL}_2(\mathbb{Z})$ is Zariski dense in SL_2. f^* is integral and primitive on Γ (since $f(1, 0, 0, 1) = 1$) and f^* factors into 4 factors in $\mathbb{Q}[\mathrm{SL}_2]$. Hence according to Conjecture 1.4 $r_0(\Gamma, f^*) = 4$ and thus $r_0(\mathcal{O}, f) = 4$. While a proof that $r_0(\mathcal{O}, f) = 4$ for a thin such orbit of triples is well out of reach of present technology it is interesting that the recent advance of Green and Tao [26] mentioned after Conjecture 1.2, allows one to prove that if $\mathcal{O} = T$ is the full orbit then $r_0(T, f) = 4$. Using the morphism of $\mathbb{A}^2$ into V_0 given by the parametrization (6.3) the problem is reduced to finding a Zariski dense (in $\mathbb{A}^2$) set of points $x, y \in \mathbb{Z}^2$ for which the 4 homogeneous linear forms $x, y, 2x + 3y$ and $2x - 3y$ are all prime. In the terminology of [26] this linear system has complexity 2 and this is exactly the new case beyond Vinogradov that their method can handle. Their lower bound for the count of the number of x, y satisfying the above (there are no local obstructions) implies by a simple analogue of Proposition 3.2 that the points produced are

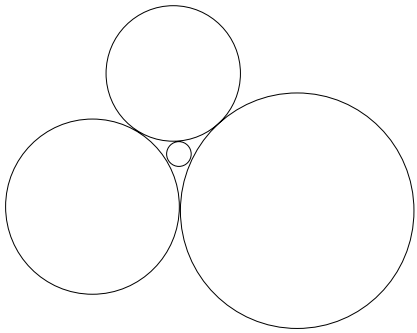

Fig. 1 Descarte's configuration

Zariski dense in $\mathbb{A}^2$. We state the result explicitly as it resolves the minimal divisibility question for the areas of Pythagorean triangles:

> *The set of Pythagorean triangles whose areas have at most r prime factors is Zariski dense in the affine cone V_0 iff $r \geq 6$.*

Example D Our final example is concerned with an orthogonal group in four variables and a naturally thin subgroup which governs integral Apollonian packings. A theorem of Descarte asserts that (a_1, a_2, a_3, a_4) in $\mathbb{R}^4$ are the curvatures of 4 mutually tangent circles in the plane (see Fig. 1) iff $F(a) = 0$ where

$$F(x_1, x_2, x_3, x_4) = 2\big(x_1^2 + x_2^2 + x_3^2 + x_4^2\big) - (x_1 + x_2 + x_3 + x_4)^2. \quad (6.5)$$

For details concerning this and the related basic facts that we record below see [25]. Given an initial configuration of 4 such circles in generation 1 of Fig. 2 (note that by convention the outer circle has curvature -6) we fill in repeatedly the lune regions with the unique circle which is tangent to 3 sides (which is possible by a theorem of Apollonius). In this way we get a packing of the outside circle by circles giving an Apollonian packing. The interesting diophantine feature is that if the initial curvatures are integral then so are the curvatures of the entire packing. The numbers in the circles in Fig. 2 indicate their curvatures. There are many questions (most being difficult) that one can ask about the integers that appear in this way.

The connection to groups is that such a packing is associated to an orbit of the Apollonian group A, which is the group of 4×4 integral matrices generated by the involutions S_j, $j = 1, 2, 3, 4$ where $S_j(e_k) = -3e_k + 2(e_1 + e_2 + e_3 + e_4)$ if $k = j$ and $S_j(e_k) = e_k$ if $k \neq j$ (e_1, e_2, e_3, e_4 are the standard basis vectors). The configurations of 4 mutually tangent circles in the packing with initial configuration $a = (a_1, a_2, a_3, a_4)$ consists of points x in the orbit $\mathcal{O}^a = a \cdot A$ of A. The elements S_j preserve F and hence $A \leq O_F(\mathbb{Z})$. A is Zariski dense in O_F but it is thin in $O_F(\mathbb{Z})$. For example if $|\ |$ is a matrix

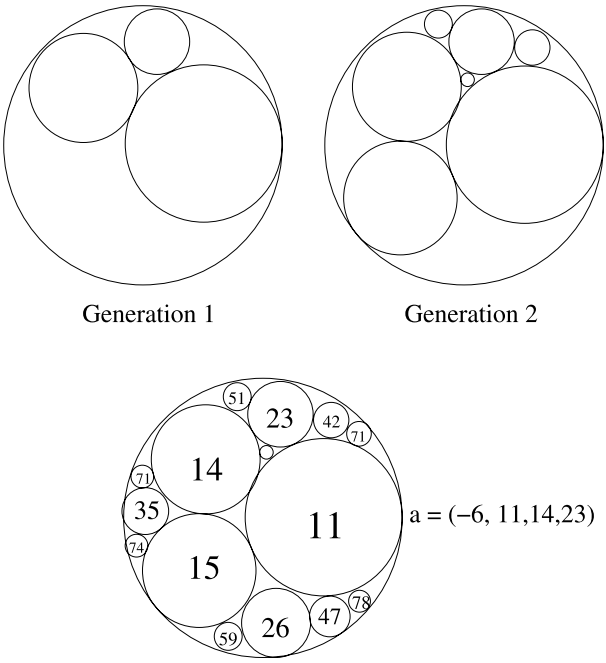

Fig. 2 Integral Apollonian packing

norm on $\mathrm{Mat}_{4\times 4}(\mathbb{R})$, then $|\{\gamma \in A : |\gamma| \le T\}| \sim c_1 T^{\delta}$ as $T \to \infty$ where $\delta = 1.3\ldots$ is the Hausdorff dimension of the limit set of A (see [25] and [52]), while $|\{\gamma \in O_F(\mathbb{Z}) : |\gamma| \le T\}| \sim c_2 T^2$. It is this thinness which makes the diophantine analysis of the orbit $\mathcal{O}^a$ problematic. $\mathcal{O}^a$ is Zariski dense in the cone $V_0 = \{x : F(x) = 0\}$. If a is primitive (which we assume henceforth), that is $\gcd(a_1, a_2, a_3, a_4) = 1$, then the same is true of every member of $\mathcal{O}^a$. The primitive points in $V_0(\mathbb{Z})$ decompose into infinitely many A-orbits, each corresponding to a different Apollonian packing (see [25]).

A modification of Theorem 6.2 implies that if $f \in \mathbb{Q}[x_1, x_2, x_3, x_4]$ and f is nonconstant and primitive on $\mathcal{O}^a$ then the pair $(\mathcal{O}^a, f)$ saturates. To see this we follow the recipe of passing to the spin double cover of SO_F. This can be realized as SL_2/K where $K = \mathbb{Q}[\sqrt{-1}]$ (see [18] and also [23]; note that our form F has signature (3, 1) over $\mathbb{R}$ and it is isotropic). In this way the key expander property follows from the following version of Conjecture 1.5 (see [62]):

Theorem 6.3 *Let Γ be a subgroup of $\mathrm{SL}_2(\mathbb{Z}[\sqrt{-1}])$ which is Zariski dense in SL_2 and such that the traces of elements of Γ generate the field $\mathbb{Q}(\sqrt{-1})$. Then as $\mathcal{A}$ varies over squarefree ideals in $\mathbb{Z}[\sqrt{-1}]$ the Cayley graphs $\mathrm{SL}_2(\mathbb{Z}[\sqrt{-1}])/\mathcal{A}, S)$, where S is a fixed symmetric generating set of generators of Γ, is a family of absolute expanders.*

According to Weisfeiler [64], outside a finite set of primes $\mathcal{P}$ of $\mathbb{Z}[\sqrt{-1}]$, Γ projects onto $\mathrm{SL}_2(\mathbb{Z}[\sqrt{-1}])/P \cong \mathrm{SL}_2(\mathbb{F}_p) \times \mathrm{SL}_2(\mathbb{F}_p)$ if p splits (that is, if $p \equiv 1 \bmod 4$) and is isomorphic to $\mathrm{SL}_2(\mathbb{F}_{p^2})$ otherwise (that is, if $p \equiv 3 \bmod 4$). Our proof of Theorem 1.2 extends without much trouble to this case. This implies that the Cayley graphs (A_d, S) where A_d is the reduction of A in $\mathrm{Mat}_{4\times 4}(\mathbb{Z}/d\mathbb{Z})$, d a square-free integer, $S = \{S_1, S_2, S_3, S_4\}$, are an expander family. This completes our discussion of the saturation of $(\mathcal{O}^a, f)$.

As far as determining the exact value of $r_0(\mathcal{O}^a, f)$ for certain f's, some progress can be made. If $f(x) = x_1$ then f is integral and primitive on $\mathcal{O}^a$ and the pullback f^* to $\mathbb{Q}(\mathrm{Spin}_1 G)$ is prime. Hence Conjecture 1.4 asserts that $r_0(\mathcal{O}^a, f) = 1$. In [52] this is proven using ad-hoc methods which among other things employ Fuchsian subgroups of A as well as Iwaniec's work in half-dimensional sieves [32]. In particular, it follows that in any integral Apollonian packing (for example the one in Fig. 2) there are infinitely many circles whose curvature is a prime number. For $f(x) = x_1 x_2$, Conjecture 1.4 implies that $r_0(\mathcal{O}^a, x_1 x_2) = 2$. This can be proven by the same methods, as is shown in [52]. In particular it follows that the set of pairs of tangent circles in an integral Apollonian packing, for which both curvatures are prime, is infinite (in fact they are pairs in quadruples of mutually tangent circles of the packing which form a Zariski dense set in V_0).

Consider next $f(x) = x_1 x_2 x_3 x_4$. That is, we are looking for quadruples of mutually tangent circles such that the product of their curvatures has few prime factors. f is not primitive on $\mathcal{O}^a$ since each primitive $a \in V_0(\mathbb{Z})$ has two components even and two odd. Still our discussion yields that $r_0(\mathcal{O}, f) < \infty$, though with no explicit bound. For the purpose of an explicit bound a simpler approach to this saturation problem can be taken by using the unipotent elements $S_i S_j$, $i \neq j$ in A as indicated in the discussion on p. 566. This and a number of related things have been carried out in [23] where it is shown that $r_0(\mathcal{O}^a, x_1 x_2 x_3 x_4) \leq 28$.

If we order the circles in a given integral Apollonian packing by the generation in which they are produced, that is by reduced word length with respect to generators S_1, S_2, S_3, S_4, then applying the upper bound sieve as in (3.58) and using Theorem 6.3 we get

$$|\{\text{circles } C \text{ at generation } n \,:\, \text{curvature}(C) \text{ is prime}\}| \ll 3^n/n. \tag{6.6}$$

This bound is of the correct order of magnitude and we expect a "prime number theorem" for integral Apollonian packings; that is the left hand side of (6.6) is asymptotic to $\frac{c_1(a)3^n}{n}$ as $n \to \infty$. The proof that $r_0(\mathcal{O}^a, x_1) = 1$ when quantified produces an exponential number of such circles but far fewer than what is predicted by this prime number theorem.

Acknowledgements We thank B. Conrad, N. Katz, J. Lagarias, E. Lindenstrauss, A. Lubotzky, B. Mazur for discussions on various aspects of this work.

References

1. Bombieri, E., Gubler, W.: Heights in Diophantine Geometry. Cambridge University Press, Cambridge (2006)
2. Borel, A., Harish-Chandra: Arithmetic subgroups of algebraic groups. Ann. Math. **75**, 485–535 (1962)
3. Bougerol, P., Lacroix, J.: Products of Random Matrices with Applications to Schrödinger Operators. Progress in Probability and Statistics, Vol. 8. Birkhäuser, Basel (1985)
4. Bourgain, J.: Mordell's exponential sum estimate revisited. J. Am. Math. Soc. **18**, 477–499 (2005)
5. Bourgain, J.: Exponential sum estimates over subgroups of $\mathbb{Z}_q^*$, q arbitrary. J. Anal. **97**, 317–355 (2005)
6. Bourgain, J., Chang, M.-C.: Exponential sum estimates over subgroups and almost subgroups of $\mathbb{Z}_q^*$, where q is composite with few prime factors. Geom. Funct. Anal. **16**, 327–366 (2006)
7. Bourgain, J., Gamburd, A.: Uniform expansion bounds for Cayley graphs of $\mathrm{SL}_2(\mathbb{F}_p)$. Ann. Math. **167**, 625–642 (2008)
8. Bourgain, J., Gamburd, A., Sarnak, P.: Generalization of Selberg's theorem and affine sieve. Preprint
9. Bourgain, J., Glibichuk, A., Konyagin, S.: Estimate for the number of sums and products and for exponential sums in fields of prime order. J. Lond. Math. Soc. **73**, 380–398 (2006)
10. Bourgain, J., Katz, N., Tao, T.: A sum-product estimate in finite fields and applications. Geom. Funct. Anal. **14**, 27–57 (2004)
11. Brun, V.: Le crible d'Eratosthène et le théoreme de Goldbach. Skr. Nor. Vidensk. Akad. Kristiania I **3**, 1–36 (1920)
12. Bugeaud, Y., Luca, F., Mignotte, M., Siksek, S.: On Fibonacci numbers with few prime divisors. Proc. Jpn. Acad., Ser. A Math. Sci. **81**(2), 17–20 (2005)
13. Cassels, J.W.S.: Rational Quadratic Forms. Academic Press, San Diego (1978)
14. Chebotarev, N.G.: Opredelenie plotnosti sovokupnosti prostykh chisel, prinadlezhashchikh zadannomu klassu podstanovok. Izv. Ross. Akad. Nauk. **17**, 205–250 (1923)
15. Clozel, L.: Demonstration de la conjecture τ. Invent. Math. **151**, 297–328 (2003)
16. Diamond, H., Halberstam, H.: Some applications of sieves of dimension exceeding 1. In: London Math. Soc. Lecture Note Ser., vol. 237, pp. 101–107. Cambridge University Press, Cambridge (1997)
17. Lejeune Dirichlet, G.: Démonstration d'un théorème sur la progression arithmètique. Abh. Preuss. Akad. Wiss. 45–71 (1837)
18. Elstrodt, J., Grunewald, F., Mennicke, J.: Groups Acting on Hyperbolic Space. Harmonic Analysis and Number Theory. Springer Monographs in Mathematics. Springer, Berlin (1998)
19. Fossum, R., Iversen, B.: On Picard groups of algebraic fibre spaces. J. Pure Appl. Algebra **3**, 269–280 (1973)
20. Friedlander, J., Iwaniec, H.: Hyperbolic prime number theorem. Acta Math. **202**, 1–19 (2009)
21. Friedlander, J., Iwaniec, H.: In preparation
22. Frobenius, G.: Über Gruppencharaktere. Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin, pp. 985–1021 (1896)
23. Fuchs, E.: Ph.D. thesis, in preparation

24. Gamburd, A.: Spectral gap for infinite index "congruence" subgroups of $SL_2(\mathbb{Z})$. Israel J. Math. **127**, 157–200 (2002)
25. Graham, R., Lagarias, J., Mallows, C., Wilks, A., Yan, C.: Apollonian circle packings: number theory. J. Number Theory **100**(1), 1–45 (2003)
26. Green, B., Tao, T.: Linear equations in primes. Preprint
27. Halberstam, H., Richert, H.: Sieve Methods. Academic Press, San Diego (1974)
28. Hall, A.: Geneology of Pythagorean Triads. Math. Gazette **54**(390), 377–379 (1970)
29. Hardy, G.H., Littlewood, J.E.: Some problems of 'Partitio Numerorum': III. On the expression of a number as a sum of primes. Acta Math. **44**, 1–70 (1922)
30. Helfgott, H.: Growth and generation in $SL_2(\mathbb{Z}/p\mathbb{Z})$. Ann. Math. **167**, 601–623 (2008)
31. Hoory, S., Linial, N., Wigderson, A.: Expander graphs and their applications. Bull. Am. Math. Soc. **43**, 439–561 (2006)
32. Iwaniec, H.: Primes represented by quadratic polynomials in two variables. Acta Arith. **24**, 435–459 (1973/74)
33. Iwaniec, H., Kowalski, E.: Analytic Number Theory. AMS, Providence (2004)
34. Jones, J., Sato, D., Wada, H., Wiens, D.: Diophantine representation of the set of prime numbers. Am. Math. Monthly **83**, 449–464 (1976)
35. Kesten, H.: Symmetric random walks on groups. Trans. Am. Math. Soc. **92**, 336–354 (1959)
36. Kneser, M.: Abschätzungen asymptotischen Dichte von Summenmengen. Math. Z. **58**, 459–484 (1953)
37. Kontorovich, A., Oh, H.: Apollonian circle packings and closed horospheres on hyperbolic 3-manifolds. Preprint
38. Kowalski, E.: The Large Sieve and Its Applications. Arithmetic Geometry, Random Walks and Discrete Groups. Cambridge Tracts in Mathematics, vol. 175. Cambridge University Press, Cambridge (2008)
39. Lang, S.: Algebra, 3rd edn. Springer, New York (2002)
40. Lang, S., Weil, A.: Number of points of varieties in finite fields. Am. J. Math. **76**, 819–827 (1954)
41. Liu, J., Sarnak, P.: Almost primes on quadrics in 3 variables. Preprint
42. Lubotzky, A.: Cayley graphs: eigenvalues, expanders and random walks. In: Rowbinson, P. (ed.) Surveys in Combinatorics. London Math. Soc. Lecture Note Ser., vol. 218, pp. 155–189. Cambridge University Press, Cambridge (1995)
43. Lubotzky, A., Phillips, R., Sarnak, P.: Ramanujan graphs. Combinatorica **8**, 261–277 (1988)
44. Matiyasevich, Yu.V.: Hilbert's Tenth Problem. MIT Press, Cambridge (2004)
45. Matthews, C., Vaserstein, L., Weisfeiler, B.: Congruence properties of Zariski-dense subgroups. Proc. Lond. Math. Soc. **48**, 514–532 (1984)
46. Nevo, A., Sarnak, P.: Prime and almost prime integral points on principal homogeneous spaces. Preprint
47. Noether, E.: Ein algebraisches Kriterium für absolute Irreduzibilität. Math. Ann. **85**, 26–40 (1922)
48. Nori, M.V.: On subgroups of $GL_n(F_p)$. Invent. Math. **88**, 257–275 (1987)
49. Sansuc, J.J.: Groupe de Brauer et arithmétique des groupes algébriques linéaires sur un corps de nombres. J. Reine Angew. Math. **327**, 12–80 (1981)
50. Sarnak, P.: What is an expander? Not. Am. Math. Soc. **51**, 762–763 (2004)
51. Sarnak, P.: Notes on the generalized Ramanujan conjectures. Clay Math. Proc. **4**, 659–685 (2005)
52. Sarnak, P.: Letter to Lagarias on integral Apollonian packings. Available at http://www.math.princeton.edu/sarnak/
53. Sarnak, P.: Equidistribution and Primes. (2007) PIMS Lecture. Available at http://www.math.princeton.edu/sarnak/

54. Sarnak, P., Xue, X.: Bounds for multiplicities of automorphic representations. Duke Math. J. **64**, 207–227 (1991)
55. Schinzel, A., Sierpinksi, W.: Sur certaines hypotheses concernant les nombres premiers. Acta Arith. **4**, 185–208 (1958)
56. Schmidt, W.: Equations over Finite Fields: An Elementary Approach. Kendrick Press, Heber City (2004)
57. Selberg, A.: On the estimation of Fourier coefficients of modular forms. In: Proc. Symp. Pure Math., Vol. VII, pp. 1–15. AMS, Providence (1965)
58. Sierpinski, W.: Pythagorean Triangles. Dover, New York (2003)
59. Tao, T.: Product sets estimates for non-commutative groups. Combinatorica **28**, 547–594 (2008)
60. Tao, T., Vu, V.: Additive Combinatorics. Cambridge University Press, Cambridge (2006)
61. Tits, J.: Free subgroups in linear groups. J. Algebra **20**, 250–270 (1972)
62. Varju, P.: Expansion in $\mathrm{SL}_d(\mathcal{O}/\mathcal{I})$, $\mathcal{I}$ square-free. Preprint
63. Vinogradov, I.M.: Representations of an odd number as a sum of three primes. Dokl. Akad. Nauk SSSR **15**, 291–294 (1937)
64. Weisfeiler, B.: Strong approximation for Zariski-dense subgroups of semisimple algebraic groups. Ann. Math. **120**(2), 271–231 (1984)