Annals of Mathematics, **167** (2008), 601–623

Growth and generation in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$

By H. A. Helfgott*

Abstract

We show that every subset of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ grows rapidly when it acts on itself by the group operation. It follows readily that, for every set of generators A of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$, every element of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ can be expressed as a product of at most $O((\log p)^c)$ elements of $A \cup A^{-1}$, where c and the implied constant are absolute.

1. Introduction

1.1. *Background.* Let G be a finite group. Let $A \subset G$ be a set of generators of G. By definition, every $g \in G$ can be expressed as a product of elements of $A \cup A^{-1}$. We would like to know the length of the longest product that might be needed; in other words, we wish to bound from above the diameter $\mathrm{diam}(\Gamma(G, A))$ of the Cayley graph of G with respect to A. (The *Cayley graph* $\Gamma(G, A)$ is the graph (V, E) with vertex set $V = G$ and edge set $E = \{(ag, g) : g \in G, a \in A\}$. The *diameter* of a graph $X = (V, E)$ is $\max_{v_1, v_2 \in V} d(v_1, v_2)$, where $d(v_1, v_2)$ is the length of the shortest path between v_1 and v_2 in X.)

If G is abelian, the diameter can be very large: if G is cyclic of order $2n + 1$, and g is any generator of G, then g^n cannot be expressed as a product of length less than n on the elements of $\{g, g^{-1}\}$. However, if G is non-abelian and simple, the diameter is believed to be quite small:

Conjecture (Babai, [BS]). *For every non-abelian finite simple group G,*

$$\mathrm{diam}(\Gamma(G, A)) \ll (\log |G|)^c, \tag{1.1}$$

where c is some absolute constant and $|G|$ is the number of elements of G.

This conjecture is far from being proved. Even for the basic cases, viz., $G = A_n$ and $G = \mathrm{PSL}_2(\mathbb{Z}/p\mathbb{Z})$, the conjecture has remained open until now; these two choices of G seem to present, already, many of the main difficulties of the general case.

*The author was supported by a fellowship from the Centre de Recherches Mathématiques at Montréal. Travel was partially funded by the Clay Mathematics Institute.

Work on both kinds of groups long predates the general conjecture in [BS]. Let us focus[1] on $G = \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. There are some classical results for certain specific generators. Let

$$A = \left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \right\}. \tag{1.2}$$

Selberg's spectral-gap theorem for $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ ([Se]) implies that $\{\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), A)\}_{p\geq 5}$ is a family of expander graphs (see., e.g., [Lu, Thm. 4.4.2, (i)]). It follows easily that

$$\mathrm{diam}(\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), A)) \ll \log p.$$

Unfortunately, this argument works only for a few other choices of A. For example, no good bounds were known up to now for $\mathrm{diam}(\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), A))$ with, say,

$$A = \left\{ \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \right\}, \tag{1.3}$$

let alone for general A, uniformly on A or not.

1.2. *Results.* We prove the conjecture for $G = \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$.

Main Theorem. *Let p be a prime. Let A be a set of generators of $G = \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. Then the Cayley graph $\Gamma(G, A)$ has diameter $O((\log p)^c)$, where c and the implied constant are absolute.*

The theorem is a direct consequence of the following statement.

Key Proposition. *Let p be a prime. Let A be a subset of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ not contained in any proper subgroup.*

(a) *Assume that $|A| < p^{3-\delta}$ for some fixed $\delta > 0$. Then*

$$|A \cdot A \cdot A| > c|A|^{1+\varepsilon}, \tag{1.4}$$

where $c > 0$ and $\varepsilon > 0$ depend only on δ.

(b) *Assume that $|A| > p^{\delta}$ for some fixed $\delta > 0$. Then there is an integer $k > 0$, depending only on δ, such that every element of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ can be expressed as a product of at most k elements of $A \cup A^{-1}$.*

The crucial fact here is that the constants c, ε and k do not depend on p or on A.

It follows immediately from the main theorem (via [DSC, §2, Lemma 2, §3, Cor. 3.1, and §3, Cor. 3.2]) that the *mixing time* of $\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), A)$ is $O(|A|(\log p)^{2c+1})$, where c and the implied constant are absolute, and c is as in the main theorem. (The *mixing time* is the least t for which a lazy random walk of length t starting at the origin of the Cayley graph has a distribution of destinations close to the uniform distribution in the ℓ_1 norm; see §6.)

[1] While $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ is not simple, the statement (1.1) for $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ is trivially equivalent to (1.1) for $\mathrm{PSL}_2(\mathbb{Z}/p\mathbb{Z})$, and treating the former group is both slightly more conventional and notationally simpler.

If A equals the projection of a fixed set of generators of a free group in $\mathrm{SL}_2(\mathbb{Z})$ (take, e.g., A as in (1.2) or (1.3)) it follows by a simple argument that A must grow rapidly at first when multiplied by itself. In such a situation, we obtain a bound of

$$\operatorname{diam}(\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), A)) \ll \log p,$$

where the implied constant depends on the elements of $\mathrm{SL}_2(\mathbb{Z})$ of which A is a projection. For (1.3) and most other examples, this bound is new; for A as in (1.2), it is, of course, known, and the novelty lies in the proof.[2]

If A is a random pair of generators, then, with probability tending to 1 as $p \to \infty$, the graph $\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), A)$ does not have small loops (see §6). It then follows from the key proposition that $\operatorname{diam}(\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), A)) \ll \log p$, as ventured by Lubotzky ([Lu, Prob. 10.3.3]). The implied constant is absolute.

1.3. *Techniques.* The tools used are almost exclusively additive-combinatorial. Fourier analysis over finite fields and Ruzsa distances are used repeatedly. Both Gowers's effective version of the Balog-Szemerédi theorem ([Go1]) and the sum-product estimates in [BKT] and [Ko] play crucial roles. It is only through [Ko] that arithmetic strictly speaking plays a role, viz., in the guise of an estimate proved in [HBK] with techniques derived from Stepanov's elementary proof of the Weil bounds. The Weil bounds themselves are not used, and even the use of [Ko] becomes unnecessary when auxiliary results suffice to ensure the growth of A small (namely, in the cases of fixed or random generators).

Estimates on growth in $\mathbb{Z}/p\mathbb{Z}$ will be proved in Section 3, and part (a) of the key proposition will be reduced thereto in Section 4. Given part (a), it suffices to prove (b) for very large A – and this is a relatively simple task (§5), yielding to the use of growth estimates coming from Fourier analysis.

1.4. *Work to do.* A natural next step would be to generalise the main results to the group $\mathrm{SL}_2(\mathbb{F}_{p^\alpha})$, $\alpha > 1$. At first sight, this does not seem too hard; however, there seem to be actual difficulties in making the result uniform on α.

A generalisation to $\mathrm{SL}_n(\mathbb{Z}/p\mathbb{Z})$ for $n \geq 3$ is likely to require a great deal of original work. The arguments in Sections 4.1–4.3 should carry over, but those in Section 3 and Section 4.4 do not. It is possible that the basic approach in Sections 4.1–4.3 will eventually prove itself valid for all simple[3] groups of Lie type, but it is too soon to tell whether something will be found to replace Section 3 and Section 4.4 in a general context.

[2] What is given here is not, however, the first elementary proof for the choice of A in (1.2); see [SX]. The proof in [SX] works for all projections of sets generating finite-index subgroups of $\mathrm{SL}_2(\mathbb{Z})$. Gamburd [Ga1] succeeded in extending the method to projections of sets generating subgroups of $\mathrm{SL}_2(\mathbb{Z})$ whose limit sets have Hausdorff dimension greater than 5/6.

[3] The diameter of a Cayley graph $\Gamma(G, A)$ of a solvable linear algebraic group G can be large: for example, G could be generated by the set A of all elements of G all of whose eigenvalues lie in B, where $B \subset (\mathbb{F}_{p^\alpha})^*$ is a set that grows very slowly when multiplied by itself. By the Lie-Kolchin theorem, the eigenvalues of $A \cdot A \cdots A$ will lie in $B \cdot B \cdots B$, which, by assumption, is only slightly larger than B. (See also [ET].) It is unclear whether the present paper's approach will be directly applicable to groups that are neither solvable nor simple (nor almost simple).

No attempt has been made to optimize – or compute – the constant c in the main theorem, though, like the implied constant, it is effective and can be made explicit. Actual numerical constants will sometimes be used in the argument for the sake of notational clarity.

1.5. *Further remarks.* There is a rich literature on the growth of sets in linear algebraic groups over fields of characteristic zero: see, most recently, [EMO]. In such a situation, one has access to topological arguments without clear analogues in $\mathbb{Z}/p\mathbb{Z}$. It is possible, nevertheless, to adapt the vocabulary of growth on infinite groups to the finite case. For example, one can say the key proposition implies immediately that A does not have *moderate growth* ([DSC2]).

The problem of bounding the diameter of $\Gamma(\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z}), A)$ for p fixed and k variable is fundamentally different from that of bounding the diameter of $\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), A)$ for p variable. From a p-adic perspective, the problem for $\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z})$ is analogous to that for SU(2), which was treated by Solovay and Kitaev [NC]. Dinai [Di] has succeeded in giving a polylogarithmic bound for $\mathrm{diam}(\Gamma(\mathrm{SL}_2(\mathbb{Z}/p^k\mathbb{Z}), A))$, p fixed, in part by adapting Solovay and Kitaev's procedure.

Consider the family $\mathscr{F} = \{\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), A)\}_{p,A}$, where both p and A vary: p ranges across the primes and A ranges across all sets that generate $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. If we could prove that $\mathscr{F}$ is an expander family, we would obtain the main theorem with the constant c set to 1. We are still far from proving that $\mathscr{F}$ is an expander family, and we will not, of course, assume such a hypothesis; rather, we will obtain a weaker statement as an immediate consequence of the main theorem (Cor. 6.1). It seems unjustified for now to hope for a purely combinatorial proof that a family of Cayley graphs $\{\Gamma(G, A)\}$ where both G and A vary quite freely is an expander family: we would need, not estimates on the growth of a set A when added to or multiplied by itself, but, instead, estimates on the growth of a set A under the action of addition or multiplication by a small, fixed set S, or under the action of a small set of operations. (Here "small" means "of cardinality less than a constant".) Such estimates are outside of the reach of the already remarkably strong sum-product techniques of [BKT] and [Ko].

1.6. *Acknowledgments.* I would like to thank A. Venkatesh for having first called this problem to my attention and for shedding light spontaneously. His Clay Mathematics Institute grant paid for a trip during which the present subject and many other interesting things were discussed. I was otherwise funded by the Centre de Recherches Mathématiques and the Institut de Sciences Mathématiques (Montréal).

Thanks are also due to N. Anantharaman, E. Breuillard, O. Dinai, U. Hadad, C. Hall and G. Harcos, for their careful reading and several helpful comments, to A. Gamburd, A. Lubotzky and I. Pak, for their instructive remarks and references, and to A. Granville, for his encouragement and advice, and for access to an unpublished set of lecture notes.

2. Background and preliminaries

2.1. *General notation.* As is customary, we denote by $\mathbb{F}_{p^\alpha}$ the finite field of order p^α. We write $|f|_r$ for the L_r-norm of a function f. Given a set A, we denote its cardinality by $|A|$, and its characteristic function by A itself. Thus, $|A| = |A|_1$. By $A + B$ (resp. $A \cdot B$), we shall always mean $\{x + y : x \in A, y \in B\}$ (resp. $\{x \cdot y : x \in A, y \in B\}$), or the characteristic function thereof; cf. $(A * B)(x) = |\{(y, z) \in A \times B : y + z = x\}|$. By $A + \xi$ and $\xi \cdot A$ we mean $\{x + \xi : x \in A\}$ and $\{\xi \cdot x : x \in A\}$, respectively.

For us, A^r means $\{x^r : x \in A\}$; in general, if f is a function on A, we take $f(A)$ to mean $\{f(x) : x \in A\}$. Given a positive integer r and a subset A of a group G, we define A_r to be the set of all products of at most r elements of $A \cup A^{-1}$:

$$A_r = \{g_1 \cdot g_2 \cdots g_r : g_i \in A \cup A^{-1} \cup \{1\}\}.$$

Finally, we write $\langle A \rangle$ for the group generated by A.

2.2. *Fourier analysis over* $\mathbb{Z}/p\mathbb{Z}$. We will review some basic facts, in part to fix our normalizations. The Fourier transform $\widehat{f}$ of a function $f : \mathbb{Z}/p\mathbb{Z} \to \mathbb{C}$ is given by

$$\widehat{f}(y) = \sum_{x \in \mathbb{Z}/p\mathbb{Z}} f(x) e^{-2\pi i x y / p}.$$

The Fourier transform is an isometry:

$$\sum_{x \in \mathbb{Z}/p\mathbb{Z}} |\widehat{f}(x)|^2 = p \cdot \sum_{x \in \mathbb{Z}/p\mathbb{Z}} |f(x)|^2.$$

For any $f, g : \mathbb{Z}/p\mathbb{Z} \to \mathbb{C}$, we have $\widehat{f * g} = \widehat{f} \cdot \widehat{g}$. If $A, B \subset \mathbb{Z}/p\mathbb{Z}$, then $|A * B|_1 = |A||B|$.

2.3. *Additive combinatorics, abelian and non-abelian.* Some basic concepts and proofs of additive combinatorics transfer effortlessly to the non-abelian case; some do not. In the following, G need not be an abelian group, except, of course, when it is explicitly said to be one.

Definition 1. Let A and B be finite subsets of a group G. We define the *Ruzsa distance*

$$d(A, B) = \log\left(\frac{|AB^{-1}|}{\sqrt{|A||B|}}\right).$$

If G is an abelian group whose operation is written additively, we denote the Rusza distance by $d_+(A, B)$.

The Ruzsa distance, while not truly a distance function ($d(A, A) \neq 0$ in general), does satisfy the triangle inequality.

Lemma 2.1. *Let* A, B *and* C *be finite subsets of a group* G. *Then*

$$d(A, C) \leq d(A, B) + d(B, C). \tag{2.1}$$

Proof (*Ruzsa*). It is enough to prove that

$$|AC^{-1}||B| \leq |AB^{-1}||BC^{-1}|. \tag{2.2}$$

We will do as much by constructing an injection $\iota : AC^{-1} \times B \hookrightarrow AB^{-1} \times BC^{-1}$. For every $d \in AC^{-1}$, choose once and for all a pair $(a_d, c_d) \in A \times C$ such that $d = a_d c_d^{-1}$. Define $\iota(d, b) = (a_d b^{-1}, b c_d^{-1})$. We can recover $d = a_d c_d^{-1}$ from $\iota(d, b)$; since (a_d, c_d) depends only on d, we recover (a_d, c_d) thereby. From $\iota(d, b)$ and (a_d, c_d) we can tell b. Thus, ι is an injection. □

In particular, we have

$$d(A, A) \leq d(A, A^{-1}) + d(A^{-1}, A) = 2d(A, A^{-1}). \tag{2.3}$$

If G is abelian, then, by [Ru2, Thm. 2],

$$d(A, A^{-1}) \leq 3d(A, A). \tag{2.4}$$

This need not hold if G is not abelian: if A is a coset gH of a large nonnormal subgroup $H \subset G$, we have $|AA^{-1}| = |H| = |A|$, but $|AA| = |HgH|$ may be much larger than $|A|$, and thus $d(A, A^{-1})$ is unbounded while $d(A, A) = 0$.

Another peculiarity of the abelian case is that, if $A \cdot \dots \cdot A$ is large, then $A \cdot A$ must be large. If G is not abelian, and A is of the form $H \cup \{g\}$, where H is a large subgroup of G, then $|A \cdot A| \leq 3|H| + 1 < 3|A|$, while $A \cdot A \cdot A$ contains HgH, and thus may be very large. However, the following auxiliary result does hold even for G non-abelian.

Lemma 2.2. *Let $n > 2$ be an integer. Let A be a finite subset of a group G. Suppose that*

$$|A_n| > c|A|^{1+\varepsilon}$$

for some $c > 0$, $\varepsilon > 0$. Then

$$|A \cdot A \cdot A| > c'|A|^{1+\varepsilon'},$$

where $c' > 0$, $\varepsilon' > 0$ depend only on c, ε and n.

Proof. By (2.2),

$$\frac{|A_{n-2}A_2|}{|A|} \leq \frac{|A_{n-2} \cdot A^{-1}|}{|A|} \frac{|A \cdot A_2|}{|A|} \leq \frac{|A_{n-1}|}{|A|} \frac{|A_3|}{|A|}.$$

Proceeding by induction on n, we obtain that

$$\frac{|A_n|}{|A|} \leq \left(\frac{|A_3|}{|A|} \right)^{n-2}.$$

It remains to bound $|A_3|/|A|$ from above by a power of $|A \cdot A \cdot A|/|A|$. Again by (2.2),

$$\begin{aligned} &|AAA^{-1}||A| = |AAA^{-1}||A^{-1}| \leq |AAA||A^{-1}A^{-1}| \leq |AAA|^2, \\ &|AA^{-1}A||A| \leq |AA^{-1}A^{-1}||AA| = |AAA^{-1}||AA| \leq |AAA^{-1}||AAA|. \end{aligned} \tag{2.5}$$

Bound $|AA^{-1}A^{-1}|, |A^{-1}AA|, \dots, |A^{-1}A^{-1}A^{-1}|$ in terms of $|AAA|$ and $|A|$ by reducing them to either case of (2.5): take inverses and replace A by A^{-1} as needed. □

2.4. *Regularity.* The following is a special case of the Gowers-Balog-Szemerédi theorem.

Theorem 2.3. *Let A be a finite subset of an additive abelian group. Let S be a subset of $A \times A$ with cardinality $|S| \geq |A|^2/K$. Suppose there exists the bound*

$$|\{a + b : (a, b) \in S\}| \leq K|A|.$$

Then there is a subset A' of A such that $|A'| \geq cK^{-C}|A|$ and

$$|A' + A'| \leq CK^C|A|,$$

where $c > 0$ and $C > 0$ are absolute.

Proof. By [Go1, Prop. 12], with $B = A$, there are sets $A', B' \subset A$ such that $|A'|, |B'| \geq cK^{-C}|A|$ and $|A' - B'| \leq CK^C|A|$. By the pigeonhole principle, there is a z such that $a - b = z$ for at least $C^{-1}c^2K^{-3C}|A|$ pairs $(a, b) \in A' \times B'$. Thus, $|V| \geq C^{-1}c^2K^{-3C}|A|$, where we define $V = A' \cap (B' + z)$. At the same time, $V - V \subset (A' - B') - z$, and so $|V - V| \leq CK^C|A|$. By (2.4), $d(V, -V) \leq 3d(V, V)$, and so $|V + V| \leq \frac{C^6}{c^6}K^{12C}|V|$. We redefine A' to be V and we are done. □

2.5. *Sum-product estimates in finite fields.*

2.5.1. *Estimates for small sets.* It is a simple matter to generalize the main result in [Ko] to finite fields other than $\mathbb{F}_p$.

Theorem 2.4. *Let $q = p^\alpha$ be a prime power. Let $\delta > 0$ be given. Then, for any $A \subset \mathbb{F}_q^*$ with $C < |A| < p^{1-\delta}$,*

$$\max(|A \cdot A|, |A + A|) > |A|^{1+\varepsilon},$$

where $C > 0$ and $\varepsilon > 0$ depend only on δ.

Explicit values of C and ε can be computed for any given $\delta > 0$.

Proof. The proofs of [HBK, Lemma 5], [Ko, Lemma 5], and [Ko, Thm. 2], work for any finite field $\mathbb{F}_q^*$ without any changes. (In the statements of [Ko, Lemma 5 and Thm. 3], the conditions $|A| < \sqrt{|F|}$ and $|B| < \sqrt{|F|}$ need to be replaced by $|A| < \sqrt{p}$ and $|B| < \sqrt{p}$.) For the range $|A| \geq p^{1/2}$, use [BKT, Thm. 4.3]. □

Note the condition $|A| < p^{1-\delta}$ in Theorem 2.4, where one might expect $|A| < q^{1-\delta}$. A subset A of $\mathbb{F}_q^*$ may be of size about p and fail to grow larger under multiplication by itself: take, for instance, $A = (\mathbb{F}_p)^*$, viewed as a subset of $\mathbb{F}_q^*$. One can prove a version of Theorem 2.4 in the range $p^{1-\delta} \leq A < q^{1-\delta}$ (see [BKT, Thm. 4.3]), but we will not need to work in such a range, hence also the condition $|A| < p^{1-\delta}$ in Propositions 3.1 and 3.3.

2.5.2. *Estimates for large sets.*

Lemma 2.5. *Let p be a prime, A a subset of $\mathbb{F}_p$, S a subset of $\mathbb{F}_p^*$. Then there is an element $\xi \in S$ such that*

$$|A + \xi A| \geq \left(\frac{1}{p} + \frac{1}{|S||A|^2/p}\right)^{-1} \geq \frac{1}{2}\min\left(p, \frac{|S||A|^2}{p}\right).$$

Furthermore, for every $c \in (0,1]$, *there are at least* $(1-c)|S|$ *elements* $\xi \in S$ *such that*

$$|A + \xi A| \geq c \left(\frac{1}{p} + \frac{1}{|S||A|^2/p}\right)^{-1}.$$

Cf. [Ko, Lemma 2], which is stronger when $|A| < p^{1/2}$.

Proof. Let us take Fourier transforms and proceed as in the beginning of the proof of Theorem 6 in [BGK]:

$$\begin{aligned} p \cdot \sum_{\xi \in S} |A * \xi A|_2^2 &= \sum_{\xi \in S} |\widehat{A * \xi A}|_2^2 = \sum_{\xi \in S} |\hat{A} \cdot \widehat{\xi A}|_2^2 = \sum_{\xi \in S} \sum_{x \in \mathbb{F}_p} |\hat{A}(x) \widehat{A}(\xi x)|^2 \\ &\leq |S||\hat{A}(0)|^4 + \sum_{x \in \mathbb{F}_p^*} \sum_{y \in \mathbb{F}_p^*} |\hat{A}(x)\hat{A}(y)|^2 = |S||A|^4 + \left(\sum_{x \in \mathbb{F}_p^*} |\hat{A}(x)|^2\right)^2 \\ &= |S||A|^4 + p^2(|A|_2^2)^2 = |S||A|^4 + p^2|A|^2. \end{aligned}$$

Hence, there is an element $\xi_0 \in S$ such that

$$|A * \xi_0 A|_2^2 \leq \left(\frac{|A|^4}{p} + \frac{p|A|^2}{|S|}\right),$$

and for every $c \in (0,1]$, there are at least $(1-c)|S|$ elements $\xi \in S$ such that

$$|A * \xi A|_2^2 \leq \frac{1}{c}\left(\frac{|A|^4}{p} + \frac{p|A|^2}{|S|}\right).$$

By Cauchy's inequality,

$$|A * \xi A|_1^2 \leq |A + \xi A| \cdot |A * \xi A|_2^2.$$

As $|A * \chi A|_1 = |A|^2$ for every $\chi \in \mathbb{F}_p^*$, we obtain that

$$|A + \xi_0 A| \geq \frac{|A * \xi_0 A|_1^2}{|A * \xi_0 A|_2^2} \geq \frac{|A|^4}{\frac{|A|^4}{p} + \frac{p|A|^2}{|S|}} = \left(\frac{1}{p} + \frac{1}{|S||A|^2/p}\right)^{-1}$$

for at least one $\xi_0 \in S$, and

$$|A + \xi A| \geq \frac{|A * \xi A|_1^2}{|A * \xi A|_2^2} \geq \frac{c|A|^4}{\frac{|A|^4}{p} + \frac{p|A|^2}{|S|}} = c\left(\frac{1}{p} + \frac{1}{|S||A|^2/p}\right)^{-1}$$

for at least $(1-c)|S|$ elements $\xi \in S$. □

3. Expanding functions on $\mathbb{F}_q$

Let f be a fairly unexceptional polynomial on x and y (or on x, x^{-1}, y and y^{-1}). It is natural to expect a result of the following type to hold: for every $\delta > 0$ and some r, $\varepsilon > 0$ and $C > 0$ depending only on δ, every set $A \subset \mathbb{F}_p$ with $C < |A| < p^{1-\delta}$ must fulfill $|f(A_r, A_r)| > |A|^{1+\varepsilon}$. The work in [BKT] and [Ko] amounts to such a result for $f(x,y) = x + y$. We will now see how to derive therefrom a result of the same type for some other choices of $f(x,y)$.

Proposition 3.1. *Let $q = p^\alpha$ be a prime power. Let $\delta > 0$ be given. Then, for any $A \subset \mathbb{F}_q^*$ with $C < |A| < p^{1-\delta}$,*

$$|\{(x + x^{-1}) \cdot (y + y^{-1}) : x, y \in A_2\}| > |A|^{1+\varepsilon},$$

where $C > 0$ and $\varepsilon > 0$ depend only δ.

Proof. Let $w(x) = x + x^{-1}$. Suppose $|\{w(x)w(y) : x, y \in A_2\}| \leq |A|^{1+\varepsilon}$. It follows directly that $|A_2| \leq \frac{1}{2}|A|^{1+\varepsilon}$. Since $w(x)w(y) = w(xy) + w(xy^{-1})$, and the cardinality of $S = \{(w(xy), w(xy^{-1})) : x, y \in A\}$ is at least $|A|^2/16$, we may apply Theorem 2.3, and obtain that there is an $A' \subset A_2$ (which may be taken to be closed under inversion) such that $|A'| > c'|A|^{1-C'\varepsilon}$ and $|w(A')+w(A')| < C'|A|^{1+C'\varepsilon}$. At the same time, $|w(A')w(A')| \leq |w(A_2)w(A_2)| \leq |A|^{1+\varepsilon}$. By Theorem 2.4, we have a contradiction, provided that ε is small enough and C is large enough. □

Lemma 3.2. *Let A and B be subsets of a group G. Then A can be covered by at most $|A \cdot B|/|B|$ cosets $a_j B_2$ of B_2, where $a_j \in A$.*

This is the noncommutative version of an argument of Ruzsa's ([Ru]).

Proof. Let $\{a_1, a_2, \ldots, a_k\}$ be a maximal subset of A with the property that the cosets $a_j B$, $1 \leq j \leq k$, are all disjoint. It is clear that $k \leq |A \cdot B|/|B|$. Let $x \in A$. Since $\{a_1, a_2, \ldots, a_k\}$ is maximal, there is a j such that $a_j B \cap xB$ is nonempty. Then $x \in a_j B B^{-1} \subset a_j B_2$. Thus, the sets $a_j B_2$ cover A. □

Proposition 3.3. *Let $q = p^\alpha$ be a prime power. Let $\delta > 0$ and $a_1, a_2 \in \mathbb{F}_q^*$ be given. Then, for any $A \subset \mathbb{F}_q^*$ with $C < |A| < p^{1-\delta}$,*

$$|\{a_1(xy + x^{-1}y^{-1}) + a_2(x^{-1}y + xy^{-1}) : x, y \in A_{20}\}| > |A|^{1+\varepsilon},$$

where $C > 0$ and $\varepsilon > 0$ depend only on δ.

Proof. By Lemma 3.2, we may cover A_4 with at most $|A_4 \cdot A^2|/|A^2|$ cosets $a_1 A_2^2, \ldots, a_k A_2^2$ of A_2^2, where $a_j \in A_4$. Given $x, y \in A_2$ such that $xy \in a_j A_2^2$, we know that $xy^{-1} = (xy)y^{-2} \in a_j A_4^2$. By Proposition 3.1 and the pigeonhole principle, there is an index j such that

$$|\{(r + r^{-1}) + (s + s^{-1}) : r, s \in a_j A_4^2\}| > \frac{|A|^{1+\varepsilon}}{|A_4 \cdot A^2|/|A^2|}. \tag{3.1}$$

Since $|A_4 \cdot A^2|/|A^2| \leq 2|A_6|/|A|$, we have either $2|A_6| > |A|^{1+\varepsilon/4}$ or

$$\frac{|A|^{1+\varepsilon}}{|A_4 \cdot A^2|/|A^2|} > |A|^{1+3\varepsilon/4}.$$

In the former case, we are already done. So, let us assume $2|A_6| \leq |A|^{1+\varepsilon/4}$.

Write $B = a_j A_4^2 \subset A_{12}$. Since $|B| \leq |A_4| \leq |A|^{1+\varepsilon/4}$, inequality (3.1) implies that

$$d_+(w(B), -w(B)) \geq \frac{\varepsilon}{2} \log |A|.$$

By (2.4), we obtain

$$d_+(w(B), w(B)) \geq \frac{\varepsilon}{6} \log |A|.$$

Then, by the triangle inequality (2.1),

$$d_+(a_1 w(B), -a_2 w(B)) \geq \frac{1}{2} d_+(w(B), w(B)) \geq \frac{\varepsilon}{12} \log |A|.$$

In other words,

$$|\{a_1(r+r^{-1}) + a_2(s+s^{-1}) : r, s \in B\}| \geq |B||A|^{\varepsilon/12} \geq \frac{1}{2}|A|^{1+\varepsilon/12}. \tag{3.2}$$

For any $r, s \in B$, the ratio r/s is in $A_4^2 A_4^{-2} \subset A_8^2$. Let $y \in A_8$ be such that $y^2 = r/s$; define $x = r/y \in A_{20}$. Then $r = xy$ and $s = x/y$. Therefore

$$\{a_1(r+r^{-1}) + a_2(s+s^{-1}) : r, s \in B\} \subset \{a_1(xy + xy^{-1}) + a_2(xy^{-1} + x^{-1}y) : x, y \in A_{20}\}.$$

By (3.2), we are done. □

4. Traces and growth

In Section 4.1 we will see how, if $A \subset \mathrm{SL}_2(\mathbb{F}_p)$ fails to grow, it must commute with itself to a fair extent, so to speak. The arguments in Section 4.2 are familiar from the study of growth in complex groups. The results in Section 4.3 will follow from those in Section 4.1 by means of simple combinatorial arguments. We will be able to prove the main part of the key proposition in Section 4.4, using the results in Section 3 and Sections 4.1–4.3.

4.1. *Growth and commutativity.* We will first see that, if a subset A of any group G does not grow rapidly under multiplication by itself, there must be an element g of A with which many elements of A commute. We shall then use the fact that, in a linear algebraic group, two elements h_1, h_2 that commute with a given g with distinct eigenvalues $\lambda_{g,1}, \ldots, \lambda_{g,n}$ must also commute with each other. Since nonunipotent elements are easy to produce in $\mathrm{SL}_2(K)$ (Lemma 4.2), we conclude that every given subset A of $\mathrm{SL}_2(K)$ either grows rapidly or contains a large simultaneously diagonalizable subset (Cor. 4.3).

PROPOSITION 4.1. *Let G be a group and A a nonempty finite subset thereof. Let Λ_A be the set of conjugacy classes of G with nonzero intersection with A. For $g \in G$, let $C_G(g)$ be the centralizer of g in G. Then there is a $g \in A$ such that*

$$|C_G(g) \cap (A^{-1}A)| \geq \frac{|\Lambda_A||A|}{|A \cdot A \cdot A^{-1}|}.$$

Proof. Let $g, h_1, h_2 \in A$. If $h_1 g h_1^{-1} = h_2 g h_2^{-1}$, then $h_2^{-1} h_1 \in A^{-1}A$ commutes with g. Hence, for any $g \in G$,

$$|\{hgh^{-1} : h \in A\}| \geq \frac{|A|}{|C_G(g) \cap A^{-1}A|}.$$

Let $\Upsilon \subset A$ be a set of representatives of Λ_A. Then

$$|AAA^{-1}| \geq |\{hgh^{-1} : h \in A, g \in \Upsilon\}| \geq \sum_{g\in\Upsilon} \frac{|A|}{|C_G(g) \cap A^{-1}A|}.$$

If $|C_G(g) \cap (A^{-1}A)| < \frac{|\Lambda_A||A|}{|A\cdot A\cdot A^{-1}|}$ for every $g \in \Upsilon$, then

$$\sum_{g\in\Upsilon} \frac{|A|}{|C_G(g) \cap A^{-1}A|} > |\Upsilon| \frac{|A \cdot A \cdot A^{-1}|}{|\Lambda_A|} = |A \cdot A \cdot A^{-1}|,$$

and we reach a contradiction. □

LEMMA 4.2. *Let K be a field and A be a finite subset of $\mathrm{SL}_2(K)$ not contained in any proper subgroup of $\mathrm{SL}_2(K)$. Then A_2 has at least $\frac{1}{4}|A| - 1$ elements with trace other than ± 2.*

Proof. Let $g \in A$ be an element of trace 2 or -2 other than $\pm I$. Let $B \subset A$ be the set of all elements of A with trace ± 2 and an eigenvector in common with g. Suppose $|B| \leq \frac{1}{4}|A| + 3$. Let $h \in A \setminus B$. If h has trace ± 2, then either gh or $g^{-1}h$ does not. Therefore $A \cup A \cdot A \cup A^{-1}A$ has at least $\frac{1}{3}|A \setminus B| \geq \frac{1}{4}|A| - 1$ elements with trace other than 2. Suppose now $|B| > \frac{1}{4}|A| + 3$. Let h be an element of A that does not have an eigenvector in common with g. Then there are at most two elements g' of B such that $g'h$ has trace 2. Hence $A \cdot A$ has more than $\frac{1}{4}|A| + 1$ elements with trace other than ± 2. □

COROLLARY 4.3. *Let K be a field. Let A be a nonempty finite subset of $\mathrm{SL}_2(K)$ not contained in any proper subgroup of $\mathrm{SL}_2(K)$. Assume $|\mathrm{Tr}(A)| \geq 2$, $|A| \geq 4$. Then there are at least $\frac{(|\mathrm{Tr}(A)|-2)(\frac{1}{4}|A|-1)}{|A_6|}$ simultaneously diagonalizable matrices in A_4.*

Proof. Let B be the set of elements of A_2 with trace other than ± 2. By Lemma 4.2, $|B| \geq \frac{1}{3}|A| - 1$. We may apply Proposition 4.1, and obtain that there is a $g \in B$ such that

$$|C_G(g) \cap (B^{-1}B)| \geq \frac{|\Lambda_B||B|}{|B \cdot B \cdot B^{-1}|} \geq \frac{|\mathrm{Tr}(B)||B|}{|B \cdot B \cdot B^{-1}|} \geq \frac{(|\mathrm{Tr}(A)| - 2)(\frac{1}{4}|A| - 1)}{|A_6|}.$$

All elements of $V = C_G(g) \cap (B^{-1}B)$ commute with g; since $\mathrm{Tr}(g) \neq \pm 2$, it follows that, when g is diagonalized, so is all of V. □

4.2. *Escaping from subvarieties.* The following lemma[4] is based closely on [EMO, Prop. 3.2].

LEMMA 4.4. *Let G be a group. Consider a linear representation of G on a vector space V over a field K. Let W be a union $W_1 \cup W_2 \cup \ldots \cup W_n$ of proper subspaces of V.*

[4] Thanks are due to N. Anantharaman for pointing out an inaccuracy in a previous version of this paper, and to both N. Anantharaman and E. Breuillard for help with the current phrasing.

Let A be a subset of G; let $\mathscr{O}$ be an $\langle A\rangle$-orbit in V not contained in W. Then there are constants $\eta > 0$ and m depending only on n and $\dim V$ such that, for every $x \in \mathscr{O}$, there are at least $\max(1, \eta|A|)$ elements $g \in A_m$ such that $gx \notin W$.

This may be phrased as follows: one can escape from W by the action of the elements of A. One can give stronger and more general statements of this kind; the spaces W_n could very well be taken to be varieties instead. However, what we have just stated will do.

Proof. Let us begin by showing that there are elements $g_1, \ldots, g_l \in A_r$ such that, for every $x \in \mathscr{O}$, at least one of the $g_i \cdot x$'s is not in W. (Here l and r are bounded in terms of n and $d = \dim V$ alone.) We will proceed by induction on (d_W, s_W), where d_W is the maximal dimension of the spaces $W_1, \ldots, W_n$ (i.e., $d_W = \max_{1\leq j\leq n} \dim(W_j)$) and s_W is the number of spaces of dimension d_W among $W_1, \ldots, W_n$. We shall always pass from W to a union of the form $W' = W_1' \cup \cdots \cup W_n'$, where either (a) $d_{W'} < d_W$ or (b) $d_{W'} = d_W$ and $s_{W'} < s_W$. The base case of the inductive process will be $(d_W, s_W) = (0, 0)$.

Let W_+ be the union of subspaces W_j, $1 \leq j \leq n$, of dimension d_W (the maximal dimension). If W_+ and $\mathscr{O}$ are disjoint, we set $W' = W \setminus W_+$. Suppose otherwise. Since $\mathscr{O}$ is not contained in W_+, we can find $x_0 \in W_+ \cap \mathscr{O}$, $g \in A \cup A^{-1}$ such that $gx_0 \notin W_+$. Hence the set of subspaces of maximal dimension in W is not the same as the set of subspaces of maximal dimension in W'. It follows that $W' = gW \cap W$ does not contain W_+, and thus has fewer subspaces W_j' of dimension d_W (the maximal dimension) than W has.

We have thus passed from W to W', where either (a) $d_W' < d_W$ or (b) $d_W' = d_W$ and $s_W' < s_W$. By the inductive hypothesis, we already know that there are $g_1', \ldots, g_{l'}' \in A_{r'}$ such that, for every $x \in \mathscr{O}$, at least one of the $g_i' \cdot x$'s is not in W'. (Here l' and r' are bounded in terms of n' and $d = \dim V$ alone; the number n' of subspaces $W_1', W_2', \ldots, W_{n'}'$ is bounded by n^2.) Since at least one of the $g_i' \cdot x$'s is not in $W' = gW \cap W$, either one of the $g_i' \cdot x$'s is not in W or one of the $g_i' \cdot x$'s is not in gW, i.e., one of the $g^{-1}g_i' \cdot x$'s is not in W. Set

$$g_1 = g_1',\ g_2 = g_2',\ \ldots,\ g_l = g_l'$$
$$g_{l+1} = g^{-1}g_1',\ g_{l+2} = g^{-1}g_2',\ \ldots,\ g_{2l} = g^{-1}g_l', \qquad l' = 2l.$$

(As can be seen, $g_i \in A_r$, where $r = r' + 1$.) We conclude that, for every $x \in \mathscr{O}$, at least one of the $g_i \cdot x$'s is not in W.

The rest is easy: for each $x \in \mathscr{O}$ and each $g \in A$, at least one of the elements $g_i g \cdot x$, $1 \leq i \leq l$ ($g_i \in A_r$) will not be in W. Each possible $g_i g$ can occur for at most l different elements $g \in A$; thus, there are at least $\min(1, |A|/l)$ elements $h = g_i g$ of A_{r+1} such that $hx \notin W$. □

We derive some immediate consequences.

Corollary 4.5. *Let K be a field. Let A be a finite subset of $\mathrm{SL}_2(K)$ not contained in any proper subgroup of $\mathrm{SL}_2(K)$. If $|K| > 3$, the following holds: for any basis $\{v_1, v_2\}$ of $\overline{K}^2$, there is a $g \in A_k$ such that $gv_i \neq \lambda v_j$ for all choices of $\lambda \in \overline{K}$, $i, j \in \{1, 2\}$, where k is an absolute constant.*

Proof. Consider $G = \mathrm{SL}_2(K)$ and its natural action on the vector space $V = M_2(\overline{K})$ of 2-by-2 matrices. Let W be the subset of V consisting of all $h \in V$ such that $hv_i = v_j$ for some $i, j \in \{1, 2\}$. Let x be the identity in $M_2(\overline{K})$. Apply Lemma 4.4.

Before Lemma 4.4 can be applied, we must verify[5] that the orbit $\mathscr{O} = \mathrm{SL}_2(K)$ of x is not contained in W. Let $G_{i,j}$ be the set of matrices g in $\mathrm{SL}_2(K)$ such that gv_i is a multiple of v_j. Since $W(K) \cap \mathscr{O} = G_{1,1} \cup G_{1,2} \cup G_{2,1} \cup G_{2,2}$, we would like to bound $|G_{i,j}|$. Let $g \in G_{i,j}$. Choose a vector $v \in K^2$ (one of $v = (1, 0)$ or $v = (0, 1)$, say) that is not a multiple of v_i. It is clear that gv and gv_i determine g. At the same time, we already know that $gv_i = \lambda v_j$, and, if gv is fixed, two different values of λ determine two matrices g with different determinants; in particular, at most one $\lambda \in \overline{K}$ gives us a $g \in \mathrm{SL}_2(K)$. Thus gv actually determines g. Since gv must be non-zero and lie in K^2, we conclude that $|G_{i,j}| \leq |K|^2 - 1$.

The sets $G_{1,1}$ and $G_{2,2}$ intersect at the identity. Thus, $|W(K) \cap \mathscr{O}| \leq 4(|K|^2 - 1) - 1$. Since $|\mathrm{SL}_2(K)| = |K| \cdot (|K|^2 - 1)$, it is enough to assume $|K| \geq 4$ to conclude that $|W(K) \cap \mathscr{O}| < |\mathrm{SL}_2(K)|$. In particular, for $|K| \geq 4$, the set $\mathscr{O} = \mathrm{SL}_2(K)$ is not contained in W. We are entitled to apply Lemma 4.4, after all. □

Corollary 4.6. *Let K be a field. Let A be a finite subset of* $\mathrm{SL}_2(K)$ *not contained in any proper subgroup of* $\mathrm{SL}_2(K)$. *Then there are absolute constants $k, c > 0$ such that, given any two non-zero vectors $v_1, v_2 \in \overline{K}^2$,*

$$|A_k \setminus (H_{v_1} \cup H_{v_2})| > c|A|,$$

where $H_v = \{g \in \mathrm{SL}_2(K) : v \text{ is an eigenvector of } g\}$.

Proof. Consider $G = \mathrm{SL}_2(K)$ and its natural action on $V = M_2(\overline{K})$. Let $W = H'_{v_1} \cup H'_{v_2}$, where $H'_v = \{g \in M_2(\overline{K}) : v \text{ is an eigenvector of } g\}$. Let $x = I$.

Before we apply Lemma 4.4, we need to check that $\mathrm{SL}_2(K)$ is not contained in $W(K)$. Since the matrices $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ share no eigenvectors, there is no pair of eigenvectors v_1, v_2 such that each of three matrices has at least one of v_1, v_2 as an eigenvector. Thus $\mathrm{SL}_2(K) \not\subset W(K)$. Now apply Lemma 4.4. □

Lemma 4.2 could be derived from Lemma 4.4 as well, but, since the proof of Lemma 4.2 is simple as it is, we will not bother.

4.3. *Size from trace size.* Given a large set V of diagonal matrices and a matrix $g \notin V$ with only nonzero entries, one can multiply V and g to obtain at least $\gg |V|^3$ different matrices.

Lemma 4.7. *Let K be a field. Let $V \subset \mathrm{SL}_2(K)$ be a finite set of simultaneously diagonalizable matrices; call their common eigenvectors v_1 and v_2. Let $g \in \mathrm{SL}_2(K)$ be*

[5] Thanks to O. Dinai for the counting argument about to be used.

such that $gv_i \neq \lambda v_j$ for any $\lambda \in \overline{K}$, $i, j \in \{1, 2\}$. Then

$$|VgVg^{-1}V| \geq \frac{1}{2}\left(\frac{1}{4}|V| - 5\right)|V|^2.$$

Proof. Diagonalize V, conjugating by an element of $\mathrm{SL}_2(\overline{K})$ if necessary. Write $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. By assumption, $abcd \neq 0$. Then

$$g\begin{pmatrix} r & 0 \\ 0 & r^{-1} \end{pmatrix} g^{-1} = \begin{pmatrix} rad - r^{-1}bc & (r^{-1} - r)ab \\ (r - r^{-1})cd & r^{-1}ad - rbc \end{pmatrix}, \tag{4.1}$$

the product of whose upper-right and lower-left entries is $-(r - r^{-1})^2 abcd$. The map $r \mapsto -(r - r^{-1})^2 abcd$ cannot send more than four distinct elements of K^* to the same element of K. Thus, the set $\{h_{12}h_{21} : h \in gVg^{-1}\}$ has cardinality at least $|V|/4$. The upper-left and lower-right entries of the matrix in the right-hand side of (4.1) can be both equal to 0 only if $r^2 - r^{-2} = 0$, and that can happen for at most four values of r. Let $U = \{h \in gVg^{-1} : (h_{11}h_{12}h_{21} \neq 0) \wedge (h_{22}h_{12}h_{21} \neq 0)\}$; we have that $|\{h_{12}h_{21} : h \in U\}| \geq \frac{1}{4}|V| - 5$.

Let $h \in U$ be fixed. Define

$$f_h(s, t) = \begin{pmatrix} s & 0 \\ 0 & s^{-1} \end{pmatrix}\begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix}\begin{pmatrix} t & 0 \\ 0 & t^{-1} \end{pmatrix} = \begin{pmatrix} sth_{11} & st^{-1}h_{12} \\ s^{-1}th_{21} & s^{-1}t^{-1}h_{22} \end{pmatrix}.$$

The product of the upper-right and lower-left entries of $f_h(s, t)$ is $h_{12}h_{21}$, which is independent of s and t. Since $h \in U$, we may recover s^2, t^2 and st from h and $f_h(s, t)$. Thus, for h fixed, there cannot be more than two pairs (s, t) sharing the same value of $f_h(s, t)$. For each element of $\{h_{12}h_{21} : h \in U\}$, choose an h corresponding to it; let s and t vary. We obtain at least $\frac{1}{2}|\{h_{12}h_{21} : h \in U\}||V|^2$ different values of $f_h(s, t) \in VgVg^{-1}V$. We conclude that $\{VgVg^{-1}V\}$ has cardinality at least $\frac{1}{2}|\{h_{12}h_{21} : h \in U\}||V|^2 = \frac{1}{2}(\frac{1}{4}|V| - 5)|V|^2$. □

We will now use Corollaries 4.3 and 4.5 and Lemma 4.7 to show that, unless A grows substantially under multiplication by itself, the cardinality of A_k cannot be much smaller than the cube of the cardinality of the set of traces $\mathrm{Tr}(A)$ of A.

PROPOSITION 4.8. *Let K be a field. Let A be a finite subset of $\mathrm{SL}_2(K)$ not contained in any proper subgroup of $\mathrm{SL}_2(K)$. Assume $|\mathrm{Tr}(A)| \geq 2$, $|A| \geq 4$ and $|K| > 3$. Then*

$$|A_k| \geq \frac{1}{2}\left(\frac{1}{4}\frac{(|\mathrm{Tr}(A)| - 2)(\frac{1}{4}|A| - 1)}{|A_6|} - 5\right)\left(\frac{(|\mathrm{Tr}(A)| - 2)(\frac{1}{4}|A| - 1)}{|A_6|}\right)^2,$$

where k is an absolute constant.

Proof. By Corollary 4.3, there is a simultaneously diagonalizable subset $V \subset A_4$ with $|V| \geq \frac{(|\mathrm{Tr}(A)|-2)(\frac{1}{4}|A|-1)}{|A_6|}$; call its common eigenvectors v_1 and v_2. Since A is not contained in any proper subgroup of $\mathrm{SL}_2(K)$, Corollary 4.5 yields a $g \in A_k$ such that $gv_i \neq \lambda v_j$ for all $\lambda \in K$, $i, j \in \{1, 2\}$. Hence, by Lemma 4.7, $|VgVg^{-1}V| \geq \frac{1}{2}\left(\frac{1}{4}|V| - 5\right)|V|^2$. □

We must now prove that, unless A grows substantially when multiplied by itself, the cardinality of $\mathrm{Tr}(A_k)$ cannot be much smaller than the cube root of the cardinality of A. A preparatory lemma is needed. Like Lemma 4.7, it is of a very simple type – the cardinality of a set is bounded from below by virtue of its being contained in the image of a map that has a large enough domain and is not too far from being injective.

LEMMA 4.9. *Let K be a field. Let A be a finite subset of $\mathrm{SL}_2(K)$. Write the matrices in $\mathrm{SL}_2(K)$ with respect to a basis $\{v_1, v_2\}$ of $\overline{K}^2$. Suppose $g_{12}g_{21} \neq 0$ for every $g \in A$. Then*

$$|\mathrm{Tr}(AA^{-1})| \geq \frac{|A|}{2 \cdot |\{(g_{11}, g_{22}) : g \in A\}|}.$$

Proof. Let $D = \{(g_{11}, g_{22}) : g \in A\}$. Consider any two distinct $g, g' \in B$ with $g_{11} = g'_{11}$, $g_{22} = g'_{22}$. Then gg'^{-1} has trace

$$\mathrm{Tr}(gg'^{-1}) = g_{11}g'_{22} + g_{22}g'_{11} - g_{12}g'_{21} - g_{21}\left(\frac{g'_{11}g'_{22} - 1}{g'_{21}}\right).$$

Thus, given $g \in B$, there can be at most two $g' \in B$ with $g_{11} = g'_{11}$, $g_{22} = g'_{22}$ such that $\mathrm{Tr}(gg'^{-1})$ is equal to a given value. Choose g such that $|\{g' \in B : g'_{11} = g_{11}, g'_{22} = g_{22}\}|$ is maximal. □

PROPOSITION 4.10. *Let K be a field. Let A be a finite subset of $\mathrm{SL}_2(K)$ not contained in any proper subgroup of $\mathrm{SL}_2(K)$. Then*

$$|\mathrm{Tr}(A_k)| \geq c|A|^{1/3},$$

where k and $c > 0$ are absolute constants.

Proof. If A has an element of trace other than ±2, let h be one such element. Otherwise, choose any $g_1 \in A$ other than $\pm I$, and any $g_2 \in A$ not in the unique Borel subgroup in which g_1, being parabolic, lies; then either $g_1g_2 \in A \cdot A$ or $g_1^{-1}g_2 \in A^{-1}A$ has trace $\neq \pm 2$; choose $h \in A_2$, $\mathrm{tr}(h) \neq \pm 2$, to be one of the two. From now on, write all matrices with respect to the two eigenvectors v_1, v_2 of h. We denote by r and r^{-1} the two eigenvalues of h.

By Corollary 4.6, $|X| \geq c|A|$, where $X = A_{k_0} \setminus (H_{v_1} \cup H_{v_2})$ and k, $c > 0$ are absolute constants. Lemma 4.9 now implies that

$$|\mathrm{Tr}(A_{2k_0})| \geq |\mathrm{Tr}(XX^{-1})| \geq \frac{|X|}{2 \cdot |\{(g_{11}, g_{22}) : g \in X\}|}. \tag{4.2}$$

For $t \in K$, let $D_t = |\{(g_{11}, g_{22}) : g_{11} + g_{22} = t, g \in X\}|$. Let $t \in K$ be such that $|D_t|$ is maximal. For any $(a, d) \in D_t$, we have $ra + r^{-1}d = (r - r^{-1})a + r^{-1}t$. Thus, for any two distinct pairs $(a, d), (a', d') \in D_t$, the two values $ra + r^{-1}d$, $ra' + r^{-1}d'$ must be distinct. Thus

$$|\mathrm{Tr}(A_{k_0+2})| \geq |\mathrm{Tr}(hX)| \geq |D_t| \geq \frac{|\{(g_{11}, g_{22}) : g \in X\}|}{|\mathrm{Tr}(X)|}.$$

Multiplying by (4.2), we obtain

$$|\operatorname{Tr}(A_{k_0+2})||\operatorname{Tr}(A_{2k_0})| \geq \frac{|X|}{2|\operatorname{Tr}(X)|},$$

and so $|\operatorname{Tr}(A_{2k_0})|^3 \geq |\operatorname{Tr}(A_{k_0+2})||\operatorname{Tr}(A_{2k_0})||\operatorname{Tr}(X)| \geq \frac{1}{2}|X|$, where we assume, as we may, that $k_0 \geq 2$. Hence

$$|\operatorname{Tr}(A_{2k_0})| \geq \left(\frac{1}{2}|X|\right)^{1/3} \geq \frac{c_0^{1/3}}{2^{1/3}}|A|^{1/3}. \qquad \square$$

4.4. *Growth of small sets.* The statements in the section up to now reduce the main problem to a question in $\mathbb{F}_{p^2}$, and that question can be answered using the results in Section 3.

Proof of part (a) *of the key proposition.* We may assume that p is larger than an absolute constant; otherwise we may make (1.4) true simply by adjusting the constant c therein. By the same token, we may assume that $|A|$ is larger than an absolute constant.

By Proposition 4.10, $|\operatorname{Tr}(A_{k_0})| \geq c_0|A|^{1/3}$, where k_0 and c_0 are absolute constants. As before, we may assume that $|A| \geq \max((4/c_0)^3, 8)$. Thus, by Corollary 4.3, there are at least

$$\frac{(c_0|A|^{1/3}-2)(\frac{1}{4}|A_{k_0}|-1)}{|A_{6k_0}|} \geq \frac{c_0|A|^{1/3}|A_{k_0}|}{16|A_{6k_0}|}$$

simultaneously diagonalizable matrices in A_{4k_0}; denote by V the set of the eigenvalues of $\lceil \frac{c_0|A|^{1/3}|A_{k_0}|}{16|A_{6k_0}|} \rceil$ such matrices. Since we may assume that $c_0 < 1$, we have $|V| < |A|^{1/3} < p^{1-\delta/3}$. We also take for granted that $|A_{6k_0}| < |A|^{7/6}$; otherwise, by Lemma 2.2, we are already done. Thus $|V| > \frac{c_0}{16}|A|^{1/6}$, and so, given a C depending only on δ, we may assume that $|V| > C$ by adjusting the constant c in (1.4) accordingly.

By Corollary 4.5, there is a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in A_{k_1}$ such that $abcd \neq 0$, where k_1 is an absolute constant. Now, for any scalars x, y, the trace of

$$\begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} y & 0 \\ 0 & y^{-1} \end{pmatrix}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

is $ad(xy + x^{-1}y^{-1}) - bc(x^{-1}y + xy^{-1})$. Letting x, y range on all of V, we see that

$$\begin{aligned} \operatorname{tr}(A_{160k_0+2k_1}) &= \operatorname{tr}(A_{20\cdot 4k_0+k_1+20\cdot 4k_0+k_1}) \\ &\supset \{ad(xy + x^{-1}y^{-1}) - bc(x^{-1}y + xy^{-1}) : x, y \in V_{20}\}. \end{aligned}$$

Now we apply Proposition 3.3 with $q = p^2$, and obtain that

$$|\operatorname{tr}(A_{160k_0+2k_1})| > |V|^{1+\varepsilon},$$

where $\varepsilon > 0$ depends only on δ. Here we have assumed, as we may, that $|V| > C$, where C is the constant in the statement of Proposition 3.3, with δ equal to one-third of our δ.

By the same argument as when we took $|V| > \frac{c_0}{16}|A|^{1/6}$, we may assume that

$$\frac{|\operatorname{Tr}(A_{160k_0+2k_1})||A_{160k_0+2k_1}|}{|A_{6(160k_0+2k_1)}|} \geq 40.$$

(Otherwise we are already done.) We proceed by applying Proposition 4.8, and obtain

$$\begin{aligned}|A_{k_2(160k_0+2k_1)}| &\geq \frac{1}{2^{16}}\frac{|\operatorname{Tr}(A_{160k_0+2k_1})|^3|A_{160k_0+2k_1}|^3}{|A_{6(160k_0+2k_1)}|^3} > \frac{1}{2^{16}}\frac{|A_{160k_0+2k_1}|^3}{|A_{6(160k_0+2k_1)}|^3}|V|^{3(1+\varepsilon)}\\ &\geq \frac{1}{2^{16}}\frac{|A_{160k_0+2k_1}|^3}{|A_{6(160k_0+2k_1)}|^3}\frac{c_0^3|A_{k_0}|^3}{2^{12}|A_{6k_0}|^3}|A|^{1+\varepsilon} \geq \frac{c_0^3}{2^{28}}\frac{|A|^6}{|A_{6(160k_0+2k_1)}|^6}|A|^{1+\varepsilon},\end{aligned}$$

where k_2 is an absolute constant. Hence, either $|A_{6(160k_0+2k_1)}|$ or $|A_{k_2(160k_0+2k_1)}|$ must be greater than $\frac{c_0^{3/7}}{16}|A|^{1+\varepsilon/7}$. By Lemma 2.2, we are done. □

5. Generating the whole group

Since we have proved part (a) of the key proposition, we know how to attain a set of cardinality $p^{3-\delta}$, $\delta > 0$, by multiplying a given set of generators A by itself $(\log(p/|A|))^c$ times. It remains to show how to produce the group $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ in a bounded number of steps from a set almost as large as $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ itself. As might be expected, instead of the sum-product estimates for small sets (§2.5.1), we will use the estimates for large sets (§2.5.2). We first focus on what happens in the Borel subgroups.

LEMMA 5.1. *Let p be a prime. Let H be a Borel subgroup of* $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. *Let $A \subset H$ be given with $|A| > 2p^{5/3} + 1$. Then A_8 contains all elements of H with trace* 2.

Proof. We may as well assume that H is the set of upper-triangular matrices. Define $P_r(A) = \left\{x \in \mathbb{Z}/p\mathbb{Z} : \begin{pmatrix} r & x \\ 0 & r^{-1}\end{pmatrix} \in A\right\}$. By the pigeonhole principle, there is an $r \in (\mathbb{Z}/p\mathbb{Z})^*$ such that $|P_r(A)| > 2p^{2/3}$. Let $\begin{pmatrix} t & u \\ 0 & t^{-1}\end{pmatrix}$ be any element of A with $t \neq r$. Then

$$\begin{pmatrix} t & u \\ 0 & t^{-1}\end{pmatrix}\begin{pmatrix} r & x \\ 0 & r^{-1}\end{pmatrix}\begin{pmatrix} t^{-1} & -u \\ 0 & t\end{pmatrix}\begin{pmatrix} r^{-1} & -x' \\ 0 & r\end{pmatrix}$$

equals

$$\begin{pmatrix} r & t^2x + (r^{-1} - r)ut \\ 0 & r^{-1}\end{pmatrix}\begin{pmatrix} r^{-1} & -x' \\ 0 & r\end{pmatrix} = \begin{pmatrix} 1 & r(-x' + t^2x) + (1 - r^2)ut \\ 0 & 1\end{pmatrix}.$$

Therefore, $P_1(AAA^{-1}A^{-1})$ is a superset of $r(-P_r(A) + t^2P_r(A)) + (1 - r^2)ut$. Define $S = \{t \in (\mathbb{Z}/p\mathbb{Z})^*, t \neq r : \exists\, u \in \mathbb{Z}/p\mathbb{Z} \text{ s.t.} \begin{pmatrix} t & u \\ 0 & t^{-1}\end{pmatrix} \in A, u \in \mathbb{Z}/p\mathbb{Z}, t \neq r\}$. Clearly

$|S| > \frac{1}{p}(2p^{5/3} - p) > p^{2/3}$. By Lemma 2.5, there is a $t \in S$ such that

$$\begin{aligned} |r(-P_r(A) + t^2 P_r(A)) + (1-r^2)ut| &= |P_r(A) - t^2 P_r(A)| \\ &\geq \frac{1}{\frac{1}{p} + \frac{p}{\frac{1}{2}|S||P_r(A)|^2}} > \frac{1}{\frac{1}{p} + \frac{1}{2p}} = \frac{2}{3}p. \end{aligned}$$

Thus,

$$(r(P_r(A) + t^2 P_r(A)) + (1-r^2)ut) + (r(P_r(A) + t^2 P_r(A)) + (1-r^2)ut) = \mathbb{Z}/p\mathbb{Z}.$$

It follows that $AAA^{-1}A^{-1}AAA^{-1}A^{-1}$ contains all matrices $\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}$, $x \in \mathbb{Z}/p\mathbb{Z}$. □

Proof of part (b) *of the key proposition.* By part (a) of the main theorem, we may assume that $|A| > 6p^{8/3} > (2p^{5/3} + 1)(p + 1)$. By the pigeonhole principle, there are at least $(2p^{5/3} + 1)$ matrices in A with the same lower row up to multiplication by a scalar in $(\mathbb{Z}/p\mathbb{Z})^*$; the same holds, of course, for the upper row. Thus, there are at least $2p^{5/3} + 1$ upper-diagonal matrices and at least $2p^{5/3} + 1$ lower-diagonal matrices in $C = AA^{-1}$. By Lemma 5.1, C_8 contains all matrices of the form $\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}$, $\begin{pmatrix} 1 & 0 \\ y & 1 \end{pmatrix}$, $x, y \in \mathbb{Z}/p\mathbb{Z}$. Every element of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ can be written in the form

$$\begin{pmatrix} 1 & 0 \\ y & 1 \end{pmatrix}\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ y' & 1 \end{pmatrix}\begin{pmatrix} 1 & x' \\ 0 & 1 \end{pmatrix},$$

where $x, y, x', y' \in \mathbb{Z}/p\mathbb{Z}$. Hence $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}) = C_8 C_8 C_8 C_8 \subset A_{64}$. □

Note added in proof. A far more elegant proof of part (b) given part (a) may be obtained by an approach due to Gowers [Go2]; see [NP]. In brief: in the present context, it is cleaner and simpler to do Fourier analysis on $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ itself, rather than to prove and use results based on Fourier analysis over $\mathbb{Z}/p\mathbb{Z}$ (§2.5.2, §5).

6. The main theorem and further consequences

Proof of Main Theorem. The statement of the theorem follows immediately from the key proposition, parts (a) and (b), when $|A|$ is larger than an absolute constant. Since $|A \cup A \cdot A| \geq |A| + 1$ for any A not a subgroup of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$, we may increase the cardinality of A by an absolute constant C simply by multiplying A by itself C times. □

Let G be a finite group and $A \subset G$ a set of generators of G. Let ψ be a probability distribution on G whose support contains A. We will assume throughout that ψ is symmetric; i.e., $\psi(g) = \psi(g^{-1})$ for every $g \in G$. We define the transition matrix $T_\psi(G, A) = \{\psi(y^{-1}x)\}_{x,y \in G}$. The largest eigenvalue of $T_\psi(G, A)$ is clearly 1.

Consider a family $\{G_j, A_j\}_{j \in J}$ of finite groups G_j and sets of generators A_j of G_j such that $d = |A_j \cup A_j^{-1}|$ is constant. Let $\psi_j(g) = \frac{1}{d}$ if $g \in A_j \cup A_j^{-1}$ and $\psi_j(g) = 0$ otherwise. If the difference between the largest and the second largest eigenvalue of $T_{\psi_j}(G_j, A_j)$ is bounded from below by a constant $\varepsilon > 0$, then $\{\Gamma(G_j, A_j)\}_{j\in J}$ is a *family of expander graphs.* Now let $\{(G_j, A_j)\}_{j\in J}$ be the family of all pairs (G, A) with $G = \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$, p varying over all primes, and A varying over all sets of generators of G with $d = |A \cup A^{-1}|$ fixed. The question of whether this is a family of expander graphs may still be far from being answered. We can prove a weaker property that has certain consequences of its own.

Corollary 6.1 (of the main theorem). *Let p be a prime. Let A be a set of generators of $G = \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. Let ψ be a symmetric probability distribution on G whose support contains A; let $\eta = \min_{g\in A\cup A^{-1}} \psi(g)$. Then the second largest eigenvalue of $T_\psi(G, A)$ is at most $1 - \frac{C}{\eta(\log p)^{2c}}$, where c and $C > 0$ are absolute constants.*

Here c is the same as in the main theorem.

Proof. This is immediate from the main theorem and the standard bound for the spectral gap in terms of η and the diameter (see, e.g., [DSC, Cor. 1]). □

From now on, assume for notational convenience that $A = A^{-1}$, and choose the following probability distribution on G:

$$\psi(g) = \begin{cases} \frac{1}{2|A|}\delta_A(g) & \text{if } g \text{ is not the identity,} \\ \frac{1}{2|A|}\delta_A(g) + \frac{1}{2} & \text{if } g \text{ is the identity,} \end{cases} \tag{6.1}$$

where δ_A is the characteristic function of A. For every positive integer n and every $g_0 \in G$, let ϕ_{n,g_0} be the probability distribution on G defined as a vector $\phi_{n,g_0} = (T_\psi(G, A))^n\delta_{g_0}$, where the transition matrix $T_\psi(G, A)$ is as before and δ_{g_0} is the characteristic function of g_0 seen as a vector of length $|G|$. We may regard ϕ_{n,g_0} as the outcome of a so-called *lazy random walk*: start at a vertex g_0 of $\Gamma(G, A)$ and do the following n times – throw a coin into the air, take a random edge out of your current vertex if it is heads, but stay in place if it is tails.

The *mixing time* $\mathrm{mix}_{G,A}$ of the lazy random walk on $\Gamma(G, A)$ is defined to be the smallest positive integer n such that

$$\sum_{g\in G} \left|\phi_{n,g_0}(g) - \frac{1}{|G|}\right| \le \frac{1}{2}. \tag{6.2}$$

It is clear that $\mathrm{mix}_{G,A}$ is independent of g_0. The constant $\frac{1}{2}$ in (6.2) is conventional; if it were changed to 1/1000000, the mixing time would change by at most a constant factor.

Corollary 6.2 (of Corollary 6.1). *Let p be a prime. Let A be a set of generators of $G = \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. Then the mixing time $\mathrm{mix}_{G,A}$ is $O(|A|(\log p)^{2c+1})$, where c and the implied constant are absolute.*

Again, the constant c is as in the main theorem.

Proof. This is immediate from Corollary 6.1 via [DSC, Lemma 2]. (For ψ as in (6.1), the transition matrix $T_\psi(G, A)$ has no negative eigenvalues; see [DSC, Lemma 1].) □

By a *word* on the symbols $x_1, x_2, \ldots, x_n$ we mean, as is usual, a product of finitely many copies of $x_1, x_1^{-1}, x_2, x_2^{-1}, \ldots, x_n^{-1}$. A *trivial word* is a product of finitely many terms of the form gg^{-1}, where g is any word.

Corollary 6.3 (of the key proposition, part (b)). *Let A be a set of generators of a free subgroup of $\mathrm{SL}_2(\mathbb{Z})$. Let p be any prime for which the reduction $\bar{A} \subset \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ of A modulo p generates a free subgroup of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. Then the diameter of the Cayley graph $\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}), \bar{A})$ is $O_A(\log p)$, where the implied constant depends only on A.*

We may take, for example, A as in (1.2) or (1.3), with $p \geq 5$.

Proof. Let $g_1, g_2, \ldots, g_n \in \mathrm{SL}_2(\mathbb{Z})$ be the elements of A. Let $w(x_1, x_2, \ldots, x_n)$ be a nontrivial word on $x_1, x_2, \ldots, x_n$. Since A generates a free group, $w(g_1, g_2, \ldots, g_n) \neq I$. Suppose that $w(\bar{g}_1, \bar{g}_2, \ldots, \bar{g}_n)$ equals the identity in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$, where $\bar{g}_1, \ldots, \bar{g}_n$ are the reductions mod p of $g_1, \ldots, g_n$. Then at least one of the entries of $w(g_1, g_2, \ldots, g_n)$ must have absolute value at least $p - 1$. Yet it is clear that this is impossible if w is of length $\leq k \log p$, where $k > 0$ is a constant depending only on A. (Cf. [Ma].)

We thus have that any two distinct products of length at most $\frac{k}{2} \log p$ on the symbols $x_1, \ldots, x_n$ must take distinct values in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ for $x_1 = \bar{g}_1, \ldots, x_n = \bar{g}_n$. We obtain that $|\bar{A}^{\lfloor \frac{k}{2} \log p \rfloor}| \geq n^{\lfloor \frac{k}{2} \log p \rfloor}$. For all p larger than an absolute constant, we have $n^{\lfloor \frac{k}{2} \log p \rfloor} \geq p^{\varepsilon}$, where $\varepsilon > 0$ depends only on k, and hence only on A. We apply part (b) of the key proposition to $\bar{A}^{\lfloor \frac{c}{2} \log p \rfloor}$, and conclude that $\mathrm{diam}(\Gamma(\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}))) \leq C \log p$ for some constant C depending only on A. □

The following lemma seems to be folkloric. A more general statement was proved in unpublished work by A. Shalev [Lu2]. Similar results have been discovered independently by others; in particular, a generalization will appear in a paper by Gamburd et al. [Ga2]. We give a proof for the sake of completeness.

Lemma 6.4. *Let p be a prime. Let $G = \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. Let $\mathscr{C}_p$ be the set of all pairs $(g, h) \in G^2$ such that g and h generate G. There is an absolute constant $c > 0$ such that $\Gamma(G, \{g, h\})$ has loops of length $\leq c \log p$ for at most $o(|\mathscr{C}_p|)$ pairs $(g, h) \in \mathscr{C}_p$, where the rate of convergence to 0 of $o(|\mathscr{C}_p|)$ is absolute.*

Proof. Let $w(g, h)$ be a nontrivial word. Let $f_{12}, f_{21} \in \mathbb{Z}[x_1, x_2, \ldots, x_n]$ be the upper-right and lower-left entries of the matrix obtained by formally replacing all occurrences of g, h, g^{-1}, h^{-1} in $w(g, h)$ by the matrices

$$\begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}, \begin{pmatrix} x_5 & x_6 \\ x_7 & x_8 \end{pmatrix}, \begin{pmatrix} x_4 & -x_2 \\ -x_3 & x_1 \end{pmatrix}, \begin{pmatrix} x_8 & -x_6 \\ -x_7 & x_5 \end{pmatrix},$$

respectively. Either f_{12} or f_{21} is not identically equal to zero: let A be as in (1.2), and denote its elements by X and Y; since X and Y generate a free subgroup of $\mathrm{SL}_2(\mathbb{Z})$, at least one of the upper-right and lower-left entries of $w(X, Y)$ or $w(Y, X)$ must be nonzero. (We cannot have $w(X, Y) = -I = w(Y, X)$, and neither $w(X, Y) = I$ nor $w(Y, X) = I$ is possible.)

Assume henceforth that the length ℓ of w is at most $\frac{\log(p-2)}{\log 2}$. The coefficients of f_{12} and f_{21} are bounded above in absolute value by $2^\ell \leq p-2$. Hence at least one of the reductions $\bar{f}_{12}, \bar{f}_{21} \in (\mathbb{Z}/p\mathbb{Z})[x_1, x_2, \dots, x_8]$ is nonzero. Choose one of the nonzero reductions and call it P.

Since P is a nonzero polynomial of degree at most ℓ, there are at most $8\ell p^7$ tuples $(x_1, \dots, x_8) \in (\mathbb{Z}/p\mathbb{Z})^8$ such that $P(x_1, \cdots, x_8) = 0$. (While this follows immediately from the Lang-Weil estimates, it is also quite easy to give an elementary proof. For every tuple $(x_2, \dots, x_8) \in (\mathbb{Z}/p\mathbb{Z})^7$, either there are no more than ℓ values of x_1 with $P(x_1, \dots, x_8) = 0$, or $f_{(1)}(x_2, \dots, x_8) = 0$, where $f_{(1)}$ is the leading coefficient of f considered as a polynomial on x_1. If $f_{(1)}(x_2, \dots, x_8) = 0$, repeat the argument with $f_{(1)}$ instead of f and $(x_2, \dots, x_8)$ instead of $(x_1, \dots, x_8)$.) Take any $g, h \in \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ such that $w(g, h) = I$. Then, for all $c_1, c_2 \in (\mathbb{Z}/p\mathbb{Z})^*$, both the upper-right and lower-right entries of $w(c_1 g, c_2 h)$ are 0. Moreover, each pair $c_1 g, c_2 h \in M_2(\mathbb{Z}/p\mathbb{Z})$ can arise from at most four different pairs $g, h \in \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. Since every pair $c_1 g$, $c_2 h$ gives a distinct solution to $P(x_1, \dots, x_8) = 0$, there are at most $32\ell p^5$ pairs $g, h \in \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ such that $w(g, h) = I$.

There are at most $4^l + 4^{l-1} + \cdots + 1 < 4^{l+1}$ distinct words w on g and h of length at most l. We conclude that, for every $l \leq \frac{\log(p-2)}{\log 2}$, there are fewer than $32l4^{l+1}p^5$ pairs $g, h \in \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ such that $w(g, h) = I$ for some nontrivial word w of length at most l. Set $l = \frac{\log p}{2 \log 4}$. Our aim is to show that $32l4^{l+1}p^5 \ll p^{5.5} \log p$ is small compared to $|\mathscr{C}_p|$; it will suffice to show that few of the $((p^2-1)p)^2$ pairs $(g, h) \in (\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}))^2$ are not in $\mathscr{C}_p$.

Every proper subgroup of $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$ is contained in at least one of (a) $O(p)$ subgroups of $\mathrm{SL}_2(\mathbb{Z}/p)$ of order $O(p^2)$, (b) $O(p^2)$ subgroups of order $O(p)$, or (c) $O(p^3)$ subgroups of order $O(1)$, where the implied constants are absolute. Tautologically, a pair of elements of a group G fail to generate G if and only if they are both contained in some proper subgroup of G. Hence there are at most $O(p^5)$ pairs $(g, h) \in (\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z}))^2$ not in $\mathscr{C}_p$.

We conclude that there are at most $O(|\mathscr{C}_p|(\log p)/p^{1/2})$ pairs $(g, h) \in \mathscr{C}_p$ for which the graph $\Gamma(G, \{g, h\})$ has loops of length $< \frac{\log p}{2 \log 4}$. (A trivial change in the argument would give the bound $O_\varepsilon(|\mathscr{C}_p|(\log p)/p^{1-\varepsilon})$ for $\varepsilon > 0$ arbitrary.) □

We can now answer in the affirmative a question of Lubotzky's ([Lu, Prob. 10.3.3]).

COROLLARY 6.5 (of the key proposition, part (b)). *Let p be a prime. Let $G = \mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. Let $\mathscr{C}_p$ be the set of all pairs $(g, h) \in G^2$ such that g and h generate G. There is an absolute constant $C > 0$ such that* $\mathrm{diam}(\Gamma(G, \{g, h\})) \leq C \log p$ *for all pairs $(g, h) \in \mathscr{C}_p$ outside a subset of $\mathscr{C}_p$ of cardinality $o(|\mathscr{C}_p|)$, where the rate of convergence to 0 of $o(|\mathscr{C}_p|)$, is absolute.*

Proof. By Lemma 6.4, all pairs $(g, h) \in \mathscr{C}_p$ outside a subset of $\mathscr{C}_p$ of cardinality $o(|\mathscr{C}_p|)$ yield graphs $\Gamma(G, \{g, h\})$ without loops of length $\leq c \log p$, where $c > 0$ is absolute. Let (g, h) be any such pair. Then $|\{g, h\}^{\lfloor \frac{c}{2} \log p \rfloor}| = |2^{\lfloor \frac{c}{2} \log p \rfloor}| \ll p^{\frac{c \log 2}{2}}$. (Cf. the proof of Cor. 6.3.) Applying part (b) of the key proposition to $A = \{g, h\}^{\lfloor \frac{c}{2} \log p \rfloor}$, we are done. □

In Corollaries 6.3 and 6.5, only the second part of the key proposition was directly invoked. Of course, the proof of part (b) of the key proposition does use part (a), but only with $|A| > p^{\delta}$, where $\delta > 0$ is fixed. This means in turn that the sum-product estimate (Theorem 2.4) is used only for subsets of $\mathbb{F}_q^*$ whose cardinality is greater than p^{ε}, where $\varepsilon > 0$ is fixed. Thus, the results in [Ko] are not used. Since the sum-product estimates in [BKT] are purely combinatorial, the proofs of Corollaries 6.3 and 6.5 are ultimately free of arithmetic.

Note added in proof. (a) Bourgain and Gamburd have recently derived results much stronger than Corollaries 6.3 and 6.5 from the key proposition of the present paper; see [BG]. (b) There is now a proof ([TV, §2.8]) of the sum-product theorem that does not involve Stepanov's method even for subsets of $\mathbb{F}_q^*$ of cardinality smaller than p^{ε}. Thus, all that is not additive combinatorics has disappeared from what is employed in this paper.

University of Bristol, Bristol, United Kingdom
E-mail address: h.andres.helfgott@bristol.ac.uk

References

[BG] J. Bourgain and A. Gamburd, Uniform expansion bounds for Cayley graphs of $\mathrm{SL}_2(\mathbb{F}_p)$, *Ann. of Math.* **167**, 000–000.

[BGK] J. Bourgain, A. A. Glibichuk, and S. V. Konyagin, Estimate for the number of sums and products and for exponential sums in fields of prime order, *J. London Math. Soc.* **73** (2006), 380–398 (electronic).

[BKT] J. Bourgain, N. Katz, and T. Tao, A sum-product estimate in finite fields, and applications, *Geom. Funct. Anal.* **14** (2004), 27–57.

[BS] L. Babai and Á. Seress, On the diameter of permutation groups, *European J. Combin.* **13** (1992), 231–243.

[D] L. E. Dickson, *Linear Groups, with an Exposition of the Galois Field Theory*, Teubner, Leipzig, 1901.

[Di] O. Dinai, Poly-log diameter bounds for some families of finite groups, *Proc. Amer. Math. Soc.* **134** (2006), 3137–3142 (electronic).

[DSC] P. Diaconis and L. Saloff-Coste, Comparison techniques for random walk on finite groups, *Ann. Probab.* **21** (1993), 2131–2156.

[DSC2] ———, Moderate growth and random walk on finite groups, *Geom. Funct. Anal.* **4** (1994), 1–36.

[ET] J. Ellenberg and J. Tymoczko, A sharp diameter bound for unipotent groups of classical type over $\mathbb{Z}/p\mathbb{Z}$, preprint, arXiv:math.GR/0510506.

[EMO] A. Eskin, S. Mozes, and H. Oh, On uniform exponential growth for linear groups, *Invent. Math.* **160** (2005), 1–30.

[Ga1] A. Gamburd, Spectral gap for infinite index "congruence" subgroups of $\mathrm{SL}_2(\mathbb{Z})$, *Israel J. Math.* **127** (2002), 157–200.

[Ga2] ———, personal communication.

[Go1] W. T. Gowers, A new proof of Szemerédi's theorem for arithmetic progressions of length four, *Geom. Funct. Anal.* **8** (1998), 529–551.

[Go2] ———, Quasirandom groups, preprint, arXiv:0710.3877.

[HBK] D. R. Heath-Brown and S. V. Konyagin, New bounds for Gauss sums derived from kth powers, and for Heilbronn's exponential sums, *Quart. J. Math.* **51** (2000), 221–235.

[Ko] S. V. Konyagin, A sum-product estimate in fields of prime order, preprint, math.NT/03042147.

[Lu] A. Lubotzky, *Discrete Groups*, *Expanding Graphs and Invariant Measures* (With an appendix by Jonathan D. Rogawski), *Progress in Math.* **125**, Birkäuser Verlag, Basel, 1994.

[Lu2] ———, personal communication.

[LPS] A. Lubotzky, R. Phillips, and P. Sarnak, Ramanujan graphs, *Combinatorica* **8** (1988), 261–277.

[Ma] G. A. Margulis, Explicit constructions of graphs without short cycles and low density codes, *Combinatorica* **2** (1982), 71–78.

[NC] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge Univ. Press, Cambridge, 2000.

[NP] N. Nikolov and L. Pyber, Product decompositions of quasirandom groups and a Jordan-type theorem, preprint, arXiv:math/0703.5343.

[Ru] I. Z. Ruzsa, An analog of Freiman's theorem in groups, in *Structure Theory of Set Addition*, *Astérisque* **258** (1999), 323–326.

[Ru2] ———, On the cardinality of $A+A$ and $A-A$, *Combinatorics*, *Proc. Fifth Hungarian Colloq.* (Keszthely, 1976), Vol. II, 933–938, North-Holland, New York, 1978.

[SX] P. Sarnak and X. Xue, Bounds for multiplicities of automorphic representations, *Duke Math. J.* **64** (1991), 207–227.

[Se] A. Selberg, On the estimation of Fourier coefficients of modular forms, *Proc. Sympos. Pure Math.* **III**, 1–15, A.M.S., Providence, RI, 1965.

[St] S. A. Stepanov, The number of points of a hyperelliptic curve over a prime field, *Izv. Akad. Nauk. SSSR Ser. Mater.* **33** (1969), 1171–1181.

[TV] T. Tao and V. Vu, *Additive Combinatorics*, *Cambridge Studies in Adv. Math.* **105**, Cambridge Univ. Press, Cambridge, 2006.

(Received November 7, 2005)