ON THE THRESHOLD FOR SZEMERÉDI'S THEOREM WITH RANDOM DIFFERENCES

JOP BRIËT AND DAVI CASTRO-SILVA

Abstract. Using recent developments on the theory of locally decodable codes, we prove that the critical size for Szemerédi's theorem with random differences is bounded from above by $N^{1-\frac{2}{k}+o(1)}$ for length-k progressions. This improves the previous best bounds of $N^{1-\frac{1}{\lceil k/2 \rceil}+o(1)}$ for all odd k.

1. Introduction

Szemerédi [18] proved that dense sets of integers contain arbitrarily long arithmetic progressions, a result which has become a hallmark of additive combinatorics. Multiple proofs of this result were found over the years, using ideas from combinatorics, ergodic theory and Fourier analysis over finite abelian groups.

Furstenberg's ergodic theoretic proof [12] opened the floodgates to a series of powerful generalizations. In particular, it led to versions of Szemerédi's theorem where the allowed common differences for the arithmetic progressions are restricted to very sparse sets. We say that a set $D \subseteq [N]$ is *ℓ-intersective* if any positive-density set $A \subseteq [N]$ contains an $(\ell+1)$-term arithmetic progression with common difference in D. For example, a special case of a result of Bergelson and Leibman [3] shows that the perfect squares are ℓ-intersective for every ℓ, and a special case of a result of Wooley and Ziegler [21] shows that the prime numbers minus one are as well.

The existence of such sparse intersective sets motivated the problem of showing whether, in fact, random sparse sets are typically intersective. The task of making this quantitative falls within the scope of research on threshold phenomena. We say that a property of subsets of $[N]$, given by a family $\mathcal{F} \subseteq 2^{[N]}$, is *monotone* if $A \in \mathcal{F}$ and $A \subseteq B \subseteq [N]$ imply $B \in \mathcal{F}$. The *critical size* $m^* = m^*(N)$ of a property is the least m such that a uniformly random m-element subset of $[N]$ has the property with probability at least $1/2$. (This value exists if $\mathcal{F}$ is non-empty and monotone, as this probability then increases monotonically with m). A famous result of Bollobás and Thomason [4] asserts that

This work was supported by the Dutch Research Council (NWO) as part of the NETWORKS programme (grant no. 024.002.003).

every monotone property has a threshold function; this is to say that the probability

$$p(m) = \Pr_{A \in \binom{[N]}{m}}[A \in \mathcal{F}]$$

spikes from $o(1)$ to $1 - o(1)$ when m increases from $o(m^*)$ to $\omega(m^*)$.[1] In general, it is notoriously hard to determine the critical size of a monotone property.

This problem is also wide open for the property of being ℓ-intersective, which is clearly monotone, and for which we denote the critical size by $m_\ell^*(N)$. Bourgain [5] showed that the critical size for 1-intersective sets is given by $m_1^*(N) \asymp \log N$; at present, this is the only case where precise bounds are known. It has been conjectured [10] that $\log N$ is the correct bound for all fixed ℓ, and indeed no better lower bounds are known for $\ell \geq 2$. It was shown by Frantzikinakis, Lesigne and Wierdl [11] and independently by Christ [9] that

$$m_2^*(N) \ll N^{\frac{1}{2}+o(1)}. \tag{1}$$

The same upper bound was later shown to hold for $m_3^*(N)$ by the first author, Dvir and Gopi [7]. More generally, they showed that

$$m_\ell^*(N) \ll N^{1-\frac{1}{\lceil(\ell+1)/2\rceil}+o(1)}, \tag{2}$$

which improved on prior known bounds for all $\ell \geq 3$. The appearance of the ceiling function in these bounds is due to a reduction for even ℓ to the case $\ell+1$. The reason for this reduction originates from work on locally decodable error correcting codes [14]. It was shown in [7] that lower bounds on the block length of $(\ell+1)$-query locally decodable codes (LDCs) imply upper bounds on m_ℓ^*. The bounds (2) then followed directly from the best known LDC bounds; see [8] for a direct proof of (2), however.

For the same reason, a recent breakthrough of Alrabiah et al. [1] on 3-query LDCs immediately implies an improvement of (1) to

$$m_2^*(N) \ll N^{\frac{1}{3}+o(1)}.$$

For technical reasons, their techniques do not directly generalize to improve the bounds for q-query LDCs with $q \geq 4$. Here, we use the ideas of [1] to directly prove upper bounds on m_ℓ^*. Due to the additional arithmetic structure in our problem, it is possible to simplify the exposition and, more importantly, apply the techniques to improve the previous best known bounds for all even $\ell \geq 2$.

Theorem 1.1. *For every integer $\ell \geq 2$, we have that*

$$m_\ell^*(N) \ll N^{1-\frac{2}{\ell+1}+o(1)}.$$

[1] Our (standard) asymptotic notation is defined as follows. Given a parameter n which grows without bounds and a function $f : \mathbb{R}_+ \to \mathbb{R}_+$, we write: $g(n) = o(f(n))$ to mean $g(n)/f(n) \to 0$; $g(n) = \omega(f(n))$ to mean $g(n)/f(n) \to \infty$; $g(n) \ll f(n)$ to mean that $g(n) \leq Cf(n)$ holds for some constant $C > 0$ and all n; and $g(n) \asymp f(n)$ to mean both $g(n) \ll f(n)$ and $f(n) \ll g(n)$.

The arguments presented here in fact work in greater generality, and hold for any finite additive group G whose size is coprime to $\ell!$ (so as not to incur in divisibility issues when considering $(\ell+1)$-tem arithmetic progressions).

Let G be a finite additive group, $\ell \geq 1$ be an integer and $\varepsilon \in (0,1)$. We say that a set $S \subseteq G$ is (ℓ,ε)-*intersective* if every subset $A \subseteq G$ of size $|A| \geq \varepsilon|G|$ contains an $(\ell+1)$-term arithmetic progression with common difference in D. We denote the critical size for the property of being (ℓ,ε)-intersective in G by $m^*_{\ell,\varepsilon}(G)$. Our main result is the following:

Theorem 1.2. *For every $\ell \geq 2$ and $\varepsilon \in (0,1)$, there exists $C(\ell,\varepsilon) > 0$ such that*

$$m^*_{\ell,\varepsilon}(G) \leq C(\ell,\varepsilon)(\log|G|)^{2\ell+3}|G|^{1-\frac{2}{\ell+1}}$$

for every additive group G whose size is coprime to $\ell!$.

Note that Theorem 1.1 follows easily from this last result by embedding $[N]$ into a group of the form $\mathbb{Z}/p\mathbb{Z}$, where p is a prime between $(\ell+1)N$ and $2(\ell+1)N$. We omit the standard details.

2. Preliminaries

Our arguments will rely heavily on the analysis of high-dimensional matrices. Here we recall the matrix inequalities which will be needed.

If $M \in \mathbb{R}^{d\times d}$ is a matrix, we define its operator norms

$$\|M\|_2 = \max\{u^T M v : \|u\|_2 = \|v\|_2 = 1\}$$
$$\|M\|_{\infty\to 1} = \max\{u^T M v : \|u\|_\infty = \|v\|_\infty = 1\}$$
$$\|M\|_{1\to 1} = \max\{u^T M v : \|u\|_\infty = \|v\|_1 = 1\}.$$

We will make use of the following simple inequalities:

$$\|M\|_{\infty\to 1} \leq d\|M\|_2, \quad \|M\|_{\infty\to 1} \leq \sum_{i=1}^{d} \|M(i,\cdot)\|_1$$

and, when M is symmetric,

$$\|M\|_2 \leq \|M\|_{1\to 1}.$$

We will also use the following noncommutative version of Khintchine's inequality, which can be extracted from a result of Tomczak-Jaegermann [19]:

Theorem 2.1. *Let $n,d \geq 1$ be integers, and let $A_1,\ldots,A_n$ be any sequence of $d\times d$ real matrices. Then*

$$\mathbb{E}_{\sigma\in\{-1,1\}^n}\left\|\sum_{i=1}^{n}\sigma_i A_i\right\|_2 \leq 10\sqrt{\log d}\left(\sum_{i=1}^{n}\|A_i\|_2^2\right)^{1/2}.$$

Furthermore, we will need a well-known concentration inequality for polynomials due to Kim and Vu [15], which requires the introduction of some extra notation. Let $H = (V, E)$ be a hypergraph, where we allow for repeated edges (so E may be a multiset), and let $f : \{0,1\}^V \to \mathbb{R}$ be the polynomial given by

$$f(x) = \sum_{e \in E} \prod_{v \in e} x_v. \tag{3}$$

For a set $A \subseteq V$, define

$$f_A(x) = \sum_{e \in E \colon A \subseteq e} \prod_{v \in e \setminus A} x_v,$$

where the monomial corresponding to the empty set is defined to be 1. For $p \in (0,1)$, we say that X is a *p-Bernoulli random variable on* $\{0,1\}^V$, denoted $X \sim \mathrm{Bern}(p)^V$, if its coordinates are all independent and each equals 1 with probability p (and equals 0 with probability $1-p$). For each $i \in \{0, 1, \dots, |V|\}$, define

$$\mu_i = \max_{A \in \binom{V}{i}} \mathbb{E}_{X \sim \mathrm{Bern}(p)^V} f_A(X).$$

Note that μ_0 is just the expectation of $f(X)$. Define also the quantities

$$\mu = \max_{i \in \{0,1,\dots,|V|\}} \mu_i \qquad \text{and} \qquad \mu' = \max_{i \in \{1,2,\dots,|V|\}} \mu_i.$$

The polynomial concentration inequality of Kim and Vu is given as follows:

Theorem 2.2. *For every $k \in \mathbb{N}$, there exist constants $C, C' > 0$ such that the following holds. Let $H = (V, E)$ be an n-vertex hypergraph whose edges have size at most k, and let f be given by (3). Then, for any $\lambda > 1$, we have*

$$\Pr\big[|f(X) - \mu_0| > C\lambda^{k-\frac{1}{2}}\sqrt{\mu\mu'}\big] \le C' \exp\big(-\lambda + (k-1)\log n\big).$$

To suit our needs, we will use a slight variant of this result, which follows easily from it and the following basic proposition.

Proposition 2.3. *Let $f : \{0,1\}^n \to \mathbb{R}_+$ be a monotone increasing function and $p \in (\frac{16}{n}, 1)$. Then, for any integer $0 \le t \le pn/2$,*

$$\mathbb{E}_{S \in \binom{[n]}{t}} f(1_S) \le \frac{1}{2} \mathbb{E}_{X \sim \mathrm{Bern}(p)^n} f(X).$$

Proof: By direct calculation,

$$\begin{aligned}\mathbb{E}_{X\sim\mathrm{Bern}(p)^n} f(X) &= \sum_{i=0}^{n} p^i(1-p)^{n-i} \sum_{S\in\binom{[n]}{i}} f(1_S)\\ &= \sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i}\,\mathbb{E}_{S\in\binom{[n]}{i}} f(1_S)\\ &\geq \sum_{i\geq t}\binom{n}{i}p^i(1-p)^{n-i}\,\mathbb{E}_{S\in\binom{[n]}{t}} f(1_S)\\ &\geq \frac{1}{2}\,\mathbb{E}_{S\in\binom{[n]}{t}} f(1_S),\end{aligned}$$

where in the third line we used monotonicity of f and the fourth line follows from the Chernoff bound. □

Corollary 2.4. *For every $k \in \mathbb{N}$, there exist constants $C, C' > 0$ such that the following holds. Let $H = (V, E)$ be an k-uniform hypergraph on n vertices, let f be given as in (3) and let $p \in (\frac{16}{n}, 1)$. Then, for any integer $0 \leq t \leq pn/2$, we have*

$$\Pr_{S\in\binom{V}{t}}\left[f(1_S) \geq C(\log n)^{k-\frac{1}{2}}\mu\right] \leq \frac{C'}{n^4}.$$

Proof: For a sufficiently large constant $C = C(k) > 0$, let $g : \{0,1\}^n \to \{0,1\}$ be the indicator function

$$g(1_S) = \mathbf{1}\left[f(1_S) \geq C(\log n)^{k-\frac{1}{2}}\mu\right].$$

Since f is monotone, so is g. Setting $\lambda = (3+k)\log n$, it follows from Theorem 2.2 that

$$\mathbb{E}_{X\sim\mathrm{Bern}(p)^n}\, g(X) \leq \frac{C'}{n^4}.$$

The result now follows from Proposition 2.3. □

3. The main argument

Fix an integer $k \geq 3$ and a positive parameter $\varepsilon > 0$. Let G be an additive group with N elements, where N is coprime to $(k-1)!$ and is assumed to be sufficiently large relative to k and ε for our arguments to hold.

For convenience, instead of considering random intersective sets, we will consider random *intersective sequences*, where a sequence in G^m is ℓ-intersective if the set of its distinct elements is. Clearly, the probability that a uniformly random m-element sequence is ℓ-intersective is a most the probability that a uniform m-element set is. Since we are interested in proving upper bounds on the critical size, it suffices to bound the minimal m such that a random sequence in G^m is ℓ-intersective with probability at least $1/2$.

Given a sequence of differences $D = (d_1, \dots, d_m) \in G^m$ and some set $A \subseteq G$, let $\Lambda_D(A)$ be the normalized count of k-APs with common difference in D which are contained in A:

$$\Lambda_D(A) = \mathbb{E}_{i\in[m]}\mathbb{E}_{x\in G} \prod_{\ell=0}^{k-1} A(x + \ell d_i).$$

Similarly, we denote by $\Lambda_G(A)$ the proportion of all k-APs which are contained in A:

$$\Lambda_G(A) = \mathbb{E}_{d\in G}\mathbb{E}_{x\in G} \prod_{\ell=0}^{k-1} A(x + \ell d).$$

By a suitable generalization of Szemerédi's theorem, we know that

$$\Lambda_G(A) \gg_{k,\varepsilon} 1 \quad \text{for all } A \subseteq G \text{ with } |A| \geq \varepsilon|G|. \tag{4}$$

This can be proven, for instance, by using the *hypergraph removal lemma* of Gowers [13] and Nagle, Rödl, Schacht and Skokan [17, 16]. It can also be obtained via a standard averaging argument (originally due to Varnavides [20]) applied to a version of Szemerédi's theorem valid for the specific group G in consideration (though the bound obtained might then depend on the structure of G).

Now suppose $m \in [N]$ is an integer for which

$$\Pr_{D\in G^m}\big(\exists A \subseteq G :\ |A| \geq \varepsilon|G|,\ \Lambda_D(A) = 0\big) \geq 1/2. \tag{5}$$

Noting that $\mathbb{E}_{D'\in G^m}\Lambda_{D'}(A) = \Lambda_G(A)$, by combining inequalities (5) and (4) we conclude that

$$\mathbb{E}_{D\in G^m} \max_{A\subseteq G\colon |A|\geq\varepsilon N} \big|\Lambda_D(A) - \mathbb{E}_{D'\in G^m}\Lambda_{D'}(A)\big| \gg_{k,\varepsilon} 1.$$

We next apply a simple symmetrization argument given in [8, page 8690] to write this in a more convenient form:

Lemma 1 (Symmetrization)**.** *Let $c > 0$, and suppose that*

$$\mathbb{E}_{D\in G^m} \max_{A\subseteq G\colon |A|\geq\varepsilon|G|} \big|\Lambda_D(A) - \mathbb{E}_{D'\in G^m}\Lambda_{D'}(A)\big| \geq c.$$

Then

$$\mathbb{E}_{D\in G^m}\mathbb{E}_{\sigma\in\{-1,1\}^m} \max_{A\subseteq G\colon |A|\geq\varepsilon|G|} \left|\mathbb{E}_{i\in[m]}\mathbb{E}_{x\in G}\, \sigma_i \prod_{\ell=0}^{k-1} A(x + \ell d_i)\right| \geq \frac{c}{2}.$$

The appearance of the expectation over signs $\sigma \in \{-1, 1\}^m$ is crucial to our arguments. By an easy multilinearity argument, we can replace the set $A \subseteq G$ (which can be seen as a vector in $\{0, 1\}^G$) by a vector $Z \in \{-1, 1\}^G$. In combination with (5) and Lemma 1, this gives

$$\mathbb{E}_{D\in G^m}\mathbb{E}_{\sigma\in\{-1,1\}^m} \max_{Z\in\{-1,1\}^G} \left|\mathbb{E}_{i\in[m]}\mathbb{E}_{x\in G}\, \sigma_i \prod_{\ell=0}^{k-1} Z(x + \ell d_i)\right| \gg_{k,\varepsilon} 1. \tag{6}$$

The change from $\{0,1\}^G$ to $\{-1,1\}^G$ is a convenient technicality so we can ignore terms which get squared in a product.

This last inequality (6) is what we need to prove the result for even values of k using the arguments we will outline below. For odd values of k, however, this inequality is unsuited due to the odd number of factor inside the product. The main idea from [1] to deal with this case is to apply a "Cauchy-Schwarz trick" to obtain a better suited inequality:

Lemma 2 (Cauchy-Schwarz trick)**.** *Let $c > 0$, and suppose $m \geq 2/c^2$ is an integer for which*

$$\mathbb{E}_{D\in G^m}\mathbb{E}_{\sigma\in\{-1,1\}^m} \max_{Z\in\{-1,1\}^G} \left|\mathbb{E}_{i\in[m]}\mathbb{E}_{x\in G}\, \sigma_i \prod_{\ell=0}^{k-1} Z(x+\ell d_i)\right| \geq c.$$

Then there exists a partition $[m] = L \mathbin{\dot\cup} R$ such that

$$\mathbb{E}_{D\in G^m}\mathbb{E}_{\substack{\sigma\in\{-1,1\}^L\\ \tau\in\{-1,1\}^R}} \max_{Z\in\{-1,1\}^G} \sum_{\substack{i\in L\\ j\in R}}\sum_{x\in G} \sigma_i\tau_j \prod_{\ell=1}^{k-1} Z(x+\ell d_i)Z(x+\ell d_j) \geq \frac{c^2m^2N}{8}.$$

Proof: By Cauchy-Schwarz, for any $Z\in\{-1,1\}^G$ we have

$$\begin{aligned}\left|\mathbb{E}_{i\in[m]}\mathbb{E}_{x\in G}\, \sigma_i \prod_{\ell=0}^{k-1} Z(x+\ell d_i)\right|^2 &= \left|\mathbb{E}_{x\in G}\, Z(x)\cdot\left(\mathbb{E}_{i\in[m]}\sigma_i \prod_{\ell=1}^{k-1} Z(x+\ell d_i)\right)\right|^2\\ &\leq \left(\mathbb{E}_{x\in G}\, Z(x)^2\right)\mathbb{E}_{x\in G}\left(\mathbb{E}_{i\in[m]}\sigma_i \prod_{\ell=1}^{k-1} Z(x+\ell d_i)\right)^2\\ &= \mathbb{E}_{x\in G}\mathbb{E}_{i,j\in[m]}\, \sigma_i\sigma_j \prod_{\ell=1}^{k-1} Z(x+\ell d_i)Z(x+\ell d_j).\end{aligned}$$

Applying Cauchy-Schwarz again, we conclude from our assumption that

$$\begin{aligned}c^2 &\leq \mathbb{E}_{D\in G^m}\mathbb{E}_{\sigma\in\{-1,1\}^m} \max_{Z\in\{-1,1\}^G} \left|\mathbb{E}_{i\in[m]}\mathbb{E}_{x\in G}\, \sigma_i \prod_{\ell=0}^{k-1} Z(x+\ell d_i)\right|^2\\ &\leq \mathbb{E}_{D\in G^m}\mathbb{E}_{\sigma\in\{-1,1\}^m} \max_{Z\in\{-1,1\}^G} \mathbb{E}_{x\in G}\mathbb{E}_{i,j\in[m]}\, \sigma_i\sigma_j \prod_{\ell=1}^{k-1} Z(x+\ell d_i)Z(x+\ell d_j).\end{aligned}$$

Now consider a uniformly random partition $[m] = L \,\dot\cup\, R$, so that for any $i, j \in [m]$ with $i \neq j$ we have $\Pr_{L,R}(i \in L,\ j \in R) = 1/4$; then

$$\begin{aligned}\mathbb{E}_{i,j\in[m]}\, &\sigma_i\sigma_j \prod_{\ell=1}^{k-1} Z(x+\ell d_i)Z(x+\ell d_j)\\ &= \frac{1}{m^2}\sum_{\substack{i,j=1\\ i\neq j}}^{m} \sigma_i\sigma_j \prod_{\ell=1}^{k-1} Z(x+\ell d_i)Z(x+\ell d_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sigma_i^2\prod_{\ell=1}^{k-1} Z(x+\ell d_i)^2\\ &= \frac{4}{m^2}\mathbb{E}_{L,R}\sum_{i\in L, j\in R}\sigma_i\sigma_j\prod_{\ell=1}^{k-1} Z(x+\ell d_i)Z(x+\ell d_j) + \frac{1}{m}.\end{aligned}$$

It follows that

$$c^2 \leq \frac{1}{m} + \frac{4}{m^2}\mathbb{E}_{L,R}\mathbb{E}_{D\in G^m}\mathbb{E}_{\sigma\in\{-1,1\}^m}\max_{Z\in\{-1,1\}^G}\mathbb{E}_{x\in G}\sum_{i\in L, j\in R}\sigma_i\sigma_j\prod_{\ell=1}^{k-1}Z(x+\ell d_i)Z(x+\ell d_j).$$

Using that $m \geq 2/c^2$, we conclude there exists a choice of partition $[m] = L \,\dot\cup\, R$ satisfying the conclusion of the lemma. □

From now on we assume that k is odd, and write $k = 2r + 1$.[2] For $i, j \in [m]$, denote $P_i(x) = \{x + d_i, x + 2d_i, \ldots, x + 2rd_i\}$ and $P_{ij}(x) = P_i(x) \cup P_j(x)$, where we hide the dependence on the difference set D for ease of notation. From inequality (6) and Lemma 2 we conclude that

$$\mathbb{E}_{D\in G^m}\mathbb{E}_{\substack{\sigma\in\{-1,1\}^L\\ \tau\in\{-1,1\}^R}}\max_{Z\in\{-1,1\}^G}\sum_{\substack{i\in L\\ j\in R}}\sum_{x\in G}\sigma_i\tau_j\prod_{y\in P_{ij}(x)}Z(y) \gg_{k,\varepsilon} m^2N, \tag{7}$$

where (L, R) is a suitable partition of the index set $[m]$ and we assume (without loss of generality) that m is sufficiently large depending on ε and k.

From inequality (7) it follows that we can fix a "good" set $D \in G^m$ satisfying

$$\mathbb{E}_{\substack{\sigma\in\{-1,1\}^L\\ \tau\in\{-1,1\}^R}}\max_{Z\in\{-1,1\}^G}\sum_{\substack{i\in L\\ j\in R}}\sigma_i\tau_j\sum_{x\in G}\prod_{y\in P_{ij}(x)}Z(y) \gg_{k,\varepsilon} m^2N \tag{8}$$

and for which we have the technical conditions

$$\left|\left\{i\in L, j\in R:\ |P_{ij}(0)|\neq 4r\right\}\right| \ll_k m^2/N \quad \text{and} \tag{9}$$

$$\max_{x\neq 0}\sum_{i=1}^{m}\sum_{\ell=-2r}^{2r}\mathbf{1}\{\ell d_i = x\} \ll_k \log N, \tag{10}$$

[2] The even case is similar but simpler. We focus on the odd case here because this is where we get new bounds.

which are needed to bound the probability of certain bad events later on. Indeed, for $\ell, \ell' \in [k-1]$ and independent uniform $d_i, d_j \in G$, we have that $\Pr[\ell d_i = \ell' d_j] = 1/N$. Hence, the expectation of the left-hand side of (9) (taken with respect to independent d_i, d_j for $i \in L$ and $j \in R$) is at most $O_k(m^2/N)$. It then follows from Markov's inequality that (9) holds with probability at least 3/4. It follows from the Chernoff bound and a union bound that (10) also holds with probability at least 3/4. Finally, since the maxima in the expectation of (8) are bounded by $m^2 N$, it follows that also this condition holds with probability at least 3/4. Hence, with positive probability, all the conditions hold.

The next key idea is to construct matrices M_{ij} for which the quantity

$$\mathbb{E}_{\substack{\sigma \in \{-1,1\}^L \\ \tau \in \{-1,1\}^R}} \Big\| \sum_{i \in L, j \in R} \sigma_i \tau_j M_{ij} \Big\|_{\infty \to 1} \tag{11}$$

is related to the expression on the left-hand side of inequality (8). The reason for doing so is that this allows us to use strong *matrix concentration inequalities*, which can be used to obtain a good upper bound on the expectation (11); this in turn translates to an upper bound on m as a function of N, which is our goal. Such uses of matrix inequalities go back to work of Ben-Aroya, Regev and de Wolf [2], in turn inspired by work of Kerenidis and de Wolf [14] (see also [6]).

The matrices we will construct are indexed by sets of a given size s, where (with hindsight) we choose $s = \lfloor N^{1-2/k} \rfloor$. For $i \in L$, $j \in R$, define the matrix $M_{ij} \in \mathbb{R}^{\binom{G}{s} \times \binom{G}{s}}$ by

$$M_{ij}(S,T) = \sum_{x \in G} \mathbf{1}\big\{ |S \cap P_i(x)| = |S \cap P_j(x)| = r,\, S \triangle T = P_{ij}(x) \big\}$$

if $|P_{ij}(0)| = 4r$, and $M_{ij}(S,T) = 0$ if $|P_{ij}(0)| \neq 4r$; note that, despite the asymmetry in their definition, these matrices are in fact symmetric. We will next deduce from inequality (8) a lower bound on the expectation (11).

For a vector $Z \in \{-1,1\}^G$, denote by $Z^{\odot s} \in \{-1,1\}^{\binom{G}{s}}$ the "lifted" vector given by

$$Z^{\odot s}(S) = \prod_{y \in S} Z(y) \quad \text{for all } S \in \binom{G}{s}.$$

If $|P_{ij}(0)| = 4r$, then for all $Z \in \{-1,1\}^G$ we have

$$\begin{aligned}\sum_{S,T\in\binom{G}{s}} M_{ij}(S,T)Z^{\odot s}(S)Z^{\odot s}(T) &= \sum_{S,T\in\binom{G}{s}} M_{ij}(S,T)\prod_{y\in S\triangle T} Z(y)\\ &= \sum_{x\in G}\sum_{S\in\binom{G}{s}} \mathbf{1}\big\{|S\cap P_i(x)| = |S\cap P_j(x)| = r\big\}\prod_{y\in P_{ij}(x)} Z(y)\\ &= \binom{2r}{r}^2\binom{N-4r}{s-2r}\sum_{x\in G}\prod_{y\in P_{ij}(x)} Z(y), \qquad (12)\end{aligned}$$

since there are $\binom{2r}{r}^2\binom{N-4r}{s-2r}$ ways of choosing a set $S \in \binom{G}{s}$ satisfying $|S \cap P_i(x)| = |S\cap P_j(x)| = r$ and, once such a set S is chosen, there is only one set $T \in \binom{G}{s}$ for which $S\triangle T = P_{ij}(x)$. It follows that

$$\begin{aligned}&\mathbb{E}_{\substack{\sigma\in\{-1,1\}^L\\ \tau\in\{-1,1\}^R}}\Big\|\sum_{i\in L,j\in R}\sigma_i\tau_j M_{ij}\Big\|_{\infty\to 1}\\ &\qquad\geq \mathbb{E}_{\substack{\sigma\in\{-1,1\}^L\\ \tau\in\{-1,1\}^R}}\max_{Z\in\{-1,1\}^G}\sum_{S,T\in\binom{G}{s}}\sum_{i\in L,j\in R}\sigma_i\tau_j M_{ij}(S,T)Z^{\odot s}(S)Z^{\odot s}(T)\\ &\qquad= \mathbb{E}_{\substack{\sigma\in\{-1,1\}^L\\ \tau\in\{-1,1\}^R}}\max_{Z\in\{-1,1\}^G}\binom{2r}{r}^2\binom{N-4r}{s-2r}\sum_{\substack{i\in L,j\in R\\ |P_{ij}(0)|=4r}}\sigma_i\tau_j\sum_{x\in G}\prod_{y\in P_{ij}(x)} Z(y);\end{aligned}$$

combining this with inequalities (8) and (9), we conclude the lower bound

$$\mathbb{E}_{\substack{\sigma\in\{-1,1\}^L\\ \tau\in\{-1,1\}^R}}\Big\|\sum_{i\in L,j\in R}\sigma_i\tau_j M_{ij}\Big\|_{\infty\to 1} \gg_{k,\varepsilon} \binom{N-4r}{s-2r}m^2N. \qquad (13)$$

Now we need to compute an upper bound for the expectation above. The main idea here is to use the non-commutative version of Khintchine's inequality given in Theorem 2.1. Intuitively, this inequality shows that the sum in the last expression incurs many cancellations due to the presence of the random signs σ_i, and thus the expectation on the left-hand side of (13) is much smaller than one might expect.

To apply Theorem 2.1, it is better to collect the matrices M_{ij} into groups and use only one half of the random signs σ_i (another idea from [1]). For $i \in L$, $\tau \in \{-1,1\}^R$, we define the matrix

$$M_i^\tau = \sum_{j\in R}\tau_j M_{ij}.$$

We will then provide an upper bound for the expression

$$\max_{\tau\in\{-1,1\}^R}\mathbb{E}_{\sigma\in\{-1,1\}^L}\Big\|\sum_{i\in L}\sigma_i M_i^\tau\Big\|_{\infty\to 1}$$

which is itself an upper bound for the expectation in (13).

Towards this goal, we will prune the matrices M_i^τ by removing remove all rows and columns whose ℓ_1-weight significantly exceeds the average. By symmetry and non-negativity of these matrices, the ℓ_1-weight of a row or column indexed by a set $S \in \binom{G}{s}$ is bounded by

$$\begin{aligned}\sum_{T\in\binom{G}{s}} \Big|\sum_{j\in R} \tau_j M_{ij}(S,T)\Big| &\leq \sum_{T\in\binom{G}{s}} \sum_{j\in R} M_{ij}(S,T) \\ &= \sum_{\substack{j\in R \\ |P_{ij}(0)|=4r}} \sum_{x\in G} \mathbf{1}\big\{|S\cap P_i(x)| = |S\cap P_j(x)| = r\big\}.\end{aligned}$$

To show that pruning makes little difference to the final bounds, we show that only a small proportion of the rows and columns have large ℓ_1-weight. To this end, let U be a uniformly distributed $\binom{G}{s}$-valued random variable and, for each $i \in L$, define the random variable corresponding to the last expression above,

$$X_i := \sum_{\substack{j\in R \\ |P_{ij}(0)|=4r}} \sum_{x\in G} \mathbf{1}\big\{|U\cap P_i(x)| = |U\cap P_j(x)| = r\big\}.$$

The calculation done in (12), with Z the all-ones vector, shows that

$$\mathbb{E}[X_i] = \frac{1}{\binom{N}{s}} \sum_{\substack{j\in R \\ |P_{ij}(0)|=4r}} \binom{2r}{r}^2 \binom{N-4r}{s-2r} N \ll_k \frac{m}{N^{1-2/k}} \tag{14}$$

where we used our chosen value for s in the inequality. The following lemma gives an upper-tail estimate on X_i, provided m is sufficiently large.

Lemma 3. *Suppose that $m \geq N^{1-2/k}$. Then, for every $i \in L$, we have that*

$$\Pr\Big[X_i \geq (\log N)^k \frac{m}{N^{1-2/k}}\Big] \leq \frac{1}{N^4}.$$

Proof: Fix an $i \in L$. Consider the hypergraph H_i on vertex set G and with edge set

$$E(H_i) = \biguplus_{\substack{j\in R \\ |P_{ij}(0)|=4r}} \biguplus_{x\in G} \binom{P_i(x)}{r} \times \binom{P_j(x)}{r},$$

and let $f : \mathbb{R}^G \to \mathbb{R}$ be the polynomial associated with H_i as in (3),

$$f(t) = \sum_{e\in E(H_i)} \prod_{v\in e} t_v.$$

Note that $X_i = f(1_U)$, where U is uniformly distributed over $\binom{G}{s}$ and $1_U \in \mathbb{R}^G$ denotes its (random) indicator vector.

For each $0 \leq \ell \leq 2r$, we wish to bound the quantity

$$\mu_\ell := \max_{A \in \binom{G}{\ell}} \mathbb{E}_{t \sim \mathrm{Bern}(s/N)^G} f_A(t).$$

(Recall the notation introduced in Section 2.) By (14), we have that $\mu_0 \ll_k mN^{-(1-2/k)}$. For a set $A \in \binom{G}{\ell}$, define its degree in H_i by

$$\deg(A) = |\{e \in E(H_i) : e \supseteq A\}|,$$

where we count multiplicities of repeated edges. Note that for any $B \subseteq A$, we have that $\deg(A) \leq \deg(B)$. Then,

$$\mu_\ell = \max_{A \in \binom{G}{\ell}} \Big(\frac{s}{N}\Big)^{2r-\ell} \deg(A).$$

For any $v \in G$, we have that $\deg(v) \ll_k m$, since v is contained in $O_k(1)$ arithmetic progressions of length k with a fixed common difference. It follows that for $\ell \in [r]$, we have that

$$\mu_\ell \;\leq\; \Big(\frac{s}{N}\Big)^{2r-\ell} \max_{v \in G} \deg(v) \;\ll_k\; mN^{-2r/(2r+1)} \;=\; \frac{m}{N^{1-1/k}}.$$

Let $A \subseteq G$ be a set of size $\ell \in \{r+1, \ldots, 2r\}$ and

$$e \in \binom{P_i(x)}{r} \times \binom{P_j(x)}{r}$$

be an edge of $E(H_i)$ that contains A. By the Pigeonhole principle, A contains an element $a \in P_i(x)$ and an element $b \in P_j(x)$. Knowing a limits x to a set of size at most $2r$. Moreover, it follows from (10) that for each x, there are at most $O_k(\log N)$ possible values of $j \in R$ such that $b \in P_j(x)$. Therefore,

$$\mu_\ell \ll_k \Big(\frac{s}{N}\Big)^{2r-\ell} \log N \leq \log N.$$

Using our assumption on m, it follows that for each $\ell \in \{0, \ldots, 2r\}$, we have that $\mu_\ell \ll_k mN^{-(1-2/k)} \log N$. The result now follows directly from Corollary 2.4. □

Lemma 3 shows that for each matrix M_i^τ, at most an N^{-4} fraction of all rows and columns have ℓ_1-weight exceeding $(\log N)^k mN^{-(1-2/k)}$. Now define $\widetilde{M_i^\tau}$ as the 'pruned' matrix obtained from M_i^τ by zeroing out all such heavy rows and columns. Note that $\widetilde{M_i^\tau}$ is symmetric, and so

$$\|\widetilde{M_i^\tau}\|_2 \leq \|\widetilde{M_i^\tau}\|_{1\to 1} = \max_{S \in \binom{G}{s}} \|\widetilde{M_i^\tau}(S, \cdot)\|_1 \leq (\log N)^k \frac{m}{N^{1-2/k}};$$

this bound on the operator norm is what makes the pruned matrices more convenient for us to work with.

We first show that replacing the original matrices by their pruned versions has negligible effect on our bounds. Indeed, from the definition of X_i we see that its maximum value is bounded by mN, and so

$$\begin{aligned}\big\|M_i^\tau - \widetilde{M}_i^\tau\big\|_{\infty\to1} &\le \sum_{S\in\binom{G}{s}} \big\|M_i^\tau(S,\cdot) - \widetilde{M}_i^\tau(S,\cdot)\big\|_1\\ &\le 2\binom{N}{s}\cdot \mathbb{E}\big[X_i\,\mathbf{1}\big\{X_i \ge (\log N)^k mN^{-(1-2/k)}\big\}\big]\\ &\le 2\binom{N}{s}\cdot mN\Pr\big[X_i \ge (\log N)^k mN^{-(1-2/k)}\big].\end{aligned}$$

(The multiplication by 2 in the second inequality happens because we must take into account both heavy rows and heavy columns.) By Lemma 3 we conclude that

$$\big\|M_i^\tau - \widetilde{M}_i^\tau\big\|_{\infty\to1} \le \frac{2m}{N^3}\binom{N}{s} \quad \text{for all } i\in L,\ \tau\in\{0,1\}^R. \tag{15}$$

Next we apply the concentration inequality from Theorem 2.1 to the pruned matrices $\widetilde{M}_i^\tau$; we obtain

$$\begin{aligned}\mathbb{E}_{\sigma\in\{-1,1\}^L}\Big\|\sum_{i\in L}\sigma_i\widetilde{M}_i^\tau\Big\|_{\infty\to1} &\le \binom{N}{s}\mathbb{E}_{\sigma\in\{-1,1\}^L}\Big\|\sum_{i\in L}\sigma_i\widetilde{M}_i^\tau\Big\|_2\\ &\le \binom{N}{s}\sqrt{\log\binom{N}{s}}\Big(\sum_{i\in L}\|\widetilde{M}_i^\tau\|_2^2\Big)^{1/2}\\ &\le \binom{N}{s}\sqrt{\log\binom{N}{s}}\Big(\sum_{i\in L}\|\widetilde{M}_i^\tau\|_{1\to1}^2\Big)^{1/2}\\ &\le \binom{N}{s}\sqrt{s\log N}\cdot m^{1/2}(\log N)^k\frac{m}{N^{1-2/k}}.\end{aligned}$$

By the triangle inequality and our previous bounds, we conclude that

$$\begin{aligned}\mathbb{E}_{\sigma\in\{-1,1\}^L}\Big\|\sum_{i\in L}\sigma_iM_i^\tau\Big\|_{\infty\to1} &\le \mathbb{E}_{\sigma\in\{-1,1\}^L}\Big\|\sum_{i\in L}\sigma_i\widetilde{M}_i^\tau\Big\|_{\infty\to1} + \sum_{i\in L}\big\|M_i^\tau - \widetilde{M}_i^\tau\big\|_{\infty\to1}\\ &\le \binom{N}{s}\sqrt{s\log N}\cdot m^{1/2}(\log N)^k\frac{m}{N^{1-2/k}} + \frac{2m^2}{N^3}\binom{N}{s}.\end{aligned}$$

Combining this with inequality (13) gives

$$\binom{N-4r}{s-2r}m^2N \ll_{k,\varepsilon} \binom{N}{s}\sqrt{ms\log N}(\log N)^k\frac{m}{N^{1-2/k}}.$$

Rearranging and using that $\binom{N}{s}/\binom{N-4r}{s-2r} \ll_k (N/s)^{2r} = N^{2-2/k}$, we conclude that

$$m \ll_{k,\varepsilon} s(\log N)^{2k+1} = N^{1-2/k}(\log N)^{2k+1}.$$

As we started with the assumption (5), this shows that $m^*_{k-1,\varepsilon}(G) \ll_{k,\varepsilon} N^{1-2/k}(\log N)^{2k+1}$ as wished.

References

[1] Omar Alrabiah et al. *A Near-Cubic Lower Bound for 3-Query Locally Decodable Codes from Semirandom CSP Refutation.* Electronic Colloquium on Computational Complexity (ECCC). Report no. TR22-101. 2022.

[2] Avraham Ben-Aroya, Oded Regev, and Ronald de Wolf. "A Hypercontractive Inequality for Matrix-Valued Functions with Applications to Quantum Computing and LDCs". In: *2008 IEEE 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS).* IEEE Computer Society, 2008, pp. 477–486. DOI: 10.1109/FOCS.2008.45.

[3] V. Bergelson and A. Leibman. "Polynomial extensions of van der Waerden's and Szemerédi's theorems". In: *J. Amer. Math. Soc.* 9.3 (1996), pp. 725–753. ISSN: 0894-0347. DOI: 10.1090/S0894-0347-96-00194-4.

[4] B. Bollobás and A. Thomason. "Threshold functions". In: *Combinatorica* 7.1 (1987), pp. 35–38. ISSN: 0209-9683. DOI: 10.1007/BF02579198.

[5] J. Bourgain. "Ruzsa's problem on sets of recurrence". In: *Israel J. Math.* 59.2 (1987), pp. 150–166. ISSN: 0021-2172. DOI: 10.1007/BF02787258.

[6] Jop Briët. *On Embeddings of ℓ_1^k from Locally Decodable Codes.* 2016. DOI: doi.org/10.48550/arXiv.1611.06385. arXiv: 1611.06385 [cs.CC].

[7] Jop Briët, Zeev Dvir, and Sivakanth Gopi. "Outlaw distributions and locally decodable codes". In: *Theory Comput.* 15 (2019), Paper No. 12, 24. DOI: 10.4086/toc.2019.v015a012.

[8] Jop Briët and Sivakanth Gopi. "Gaussian width bounds with applications to arithmetic progressions in random settings". In: *Int. Math. Res. Not. IMRN* 22 (2020), pp. 8673–8696. ISSN: 1073-7928. DOI: 10.1093/imrn/rny238.

[9] Michael Christ. *On random multilinear operator inequalities.* 2011. DOI: 10.48550/ARXIV.1108.5655.

[10] Nikos Frantzikinakis, Emmanuel Lesigne, and Máté Wierdl. "Random differences in Szemerédi's theorem and related results". In: *J. Anal. Math.* 130 (2016), pp. 91–133. ISSN: 0021-7670. DOI: 10.1007/s11854-016-0030-z.

[11] Nikos Frantzikinakis, Emmanuel Lesigne, and Máté Wierdl. "Random sequences and pointwise convergence of multiple ergodic averages". In: *Indiana Univ. Math. J.* 61.2 (2012), pp. 585–617. ISSN: 0022-2518. DOI: 10.1512/iumj.2012.61.4571.

[12] Harry Furstenberg. "Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions". In: *J. Analyse Math.* 31 (1977), pp. 204–256. ISSN: 0021-7670. DOI: 10.1007/BF02813304.

[13] W Timothy Gowers. "Hypergraph regularity and the multidimensional Szemerédi theorem". In: *Annals of Mathematics* 166 (2007), pp. 897–946. DOI: 10.4007/annals.2007.166.897.

[14] Iordanis Kerenidis and Ronald de Wolf. "Exponential lower bound for 2-query locally decodable codes via a quantum argument". In: *J. Comput. System Sci.* 69.3 (2004). Preliminary version in STOC'03, pp. 395–420. ISSN: 0022-0000. DOI: 10.1016/j.jcss.2004.04.007.

[15] Jeong Han Kim and Van H. Vu. "Concentration of multivariate polynomials and its applications". In: *Combinatorica* 20.3 (2000), pp. 417–434. ISSN: 0209-9683. DOI: 10.1007/s004930070014.

[16] Brendan Nagle, Vojtěch Rödl, and Mathias Schacht. "The counting lemma for regular k-uniform hypergraphs". In: *Random Structures Algorithms* 28.2 (2006), pp. 113–179. ISSN: 1042-9832. DOI: 10.1002/rsa.20117. URL: https://doi.org/10.1002/rsa.20117.

[17] Vojtěch Rödl and Jozef Skokan. "Regularity lemma for k-uniform hypergraphs". In: *Random Structures Algorithms* 25.1 (2004), pp. 1–42. ISSN: 1042-9832. DOI: 10.1002/rsa.20017. URL: https://doi.org/10.1002/rsa.20017.

[18] E. Szemerédi. "On sets of integers containing no k elements in arithmetic progression". In: *Acta Arith.* 27 (1975), pp. 199–245. ISSN: 0065-1036. DOI: 10.4064/aa-27-1-199-245.

[19] Nicole Tomczak-Jaegermann. "The moduli of smoothness and convexity and the Rademacher averages of trace classes $S_p(1 \leq p < \infty)$". In: *Studia Math.* 50 (1974), pp. 163–182. ISSN: 0039-3223.

[20] P. Varnavides. "Note on a theorem of Roth". In: *J. London Math. Soc.* 30 (1955), pp. 325–326. ISSN: 0024-6107. DOI: 10.1112/jlms/s1-30.3.325. URL: https://doi.org/10.1112/jlms/s1-30.3.325.

[21] Trevor D. Wooley and Tamar D. Ziegler. "Multiple recurrence and convergence along the primes". In: *Amer. J. Math.* 134.6 (2012), pp. 1705–1732. ISSN: 0002-9327. DOI: 10.1353/ajm.2012.0048.

CWI & QuSoft, Science Park 123, 1098 XG Amsterdam, The Netherlands

Email address: j.briet@cwi.nl

CWI & QuSoft, Science Park 123, 1098 XG Amsterdam, The Netherlands

Email address: davi.silva@cwi.nl