Proceedings of the 12th European Conference on Combinatorics, Graph Theory and Applications
EUROCOMB'23
Prague, August 28 - September 1, 2023

Raising the roof on the threshold for Szemerédi's theorem with random differences

(Extended abstract)

Jop Briët* 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.

DOI: `https://doi.org/10.5817/CZ.MUNI.EUROCOMB23-032`

1 Introduction

Szemerédi [14] 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 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. Szemerédi's theorem implies

*CWI & QuSoft, Science Park 123, 1098 XG Amsterdam, The Netherlands. Supported by the Dutch Research Council (NWO) as part of the NETWORKS programme (grant no. 024.002.003).

†CWI & QuSoft, Science Park 123, 1098 XG Amsterdam, The Netherlands. Supported by the Dutch Research Council (NWO) as part of the NETWORKS programme (grant no. 024.002.003).

that for large enough N_0, the set $\{0, 1, \ldots, N_0\}$ is ℓ-intersective for $N \geq N_0$. Non-trivial examples include a special case of a result of Bergelson and Leibman [3] showing that the perfect squares are ℓ-intersective for every ℓ, and a special case of a result of Wooley and Ziegler [17] showing the same for the prime numbers minus one.

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 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 suddenly 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 [11] 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 [10] 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 [6]. 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 peculiar 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 [13]. It was shown in [6] 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 [7] 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

$$m^*_2(N) \ll N^{\frac{1}{3}+o(1)}.$$

[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 $f(n) \ll g(n)$.

For technical reasons, their techniques do not directly generalize to improve the bounds for q-query LDCs with $q \geq 4$, although they could potentially lead to improvements for all odd $q \geq 3$ (but not for even q). 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 $\ell \geq 2$. In particular, we remove the ceiling (raise the roof) in (2).

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

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

2 Outline of the argument

We now give an outline of the proof of Theorem 1.1. Fix an integer $k \geq 3$ and a positive parameter $\varepsilon > 0$, and suppose N is sufficiently large relative to k and ε. Given a sequence of differences $D = (d_1, \ldots, d_m) \in [N]^m$ and some set $A \subseteq [N]$, let $\Lambda_D(A)$ be the normalized count of k-APs with common difference in D which are contained in A:

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

Let $m \geq 1$ be an integer, and suppose

$$\Pr_{D\in[N]^m}\big(\exists A \subseteq [N] :\ |A| \geq \varepsilon N,\, \Lambda_D(A) = 0\big) \geq 1/2. \tag{3}$$

By a standard averaging argument originally due to Varnavides [16], we can conclude from Szemerédi's theorem that

$$\Lambda_{[N]}(A) \gg_{k,\varepsilon} 1 \quad \text{for all } A \subseteq [N] \text{ with } |A| \geq \varepsilon N \tag{4}$$

(where we identify $[N]$ with the sequence $(1, 2, \ldots, N) \in [N]^N$). Noting that $\mathbb{E}_{D'\in[N]^m}\Lambda_{D'}(A) = \Lambda_{[N]}(A)$, by combining inequalities (3) and (4) we conclude that

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

From this last inequality, a simple "symmetrization argument" given in [6] implies

$$\mathbb{E}_{D\in[N]^m}\mathbb{E}_{\sigma\in\{-1,1\}^m} \max_{A\subseteq[N]:\, |A|\geq\varepsilon N} \left|\mathbb{E}_{i\in[m]}\mathbb{E}_{x\in[N]}\, \sigma_i \prod_{\ell=0}^{k-1} A(x+\ell d_i)\right| \gg_{k,\varepsilon} 1;$$

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 $A \subseteq [N]$ (which can be seen as a vector in $\{0,1\}^N$) by a vector $Z \in \{-1,1\}^N$:

$$\mathbb{E}_{D\in[N]^m}\mathbb{E}_{\sigma\in\{-1,1\}^m} \max_{Z\in\{-1,1\}^N} \left|\mathbb{E}_{i\in[m]}\mathbb{E}_{x\in[N]}\, \sigma_i \prod_{\ell=0}^{k-1} Z(x+\ell d_i)\right| \gg_{k,\varepsilon} 1; \tag{5}$$

here and in what follows we use the convention that $Z(y) = 0$ for all $y > N$ when $Z \in \{-1,1\}^N$. The change from $\{0,1\}^N$ to $\{-1,1\}^N$ is a convenient technicality so we can ignore terms which get squared in a product.

This last inequality (5) 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 terms inside the product. The main idea from [1] to deal with this case is to apply a "Cauchy-Schwarz trick" to pass from (5) to the inequality

$$\mathbb{E}_{D\in[N]^m}\mathbb{E}_{\sigma\in\{-1,1\}^m} \max_{Z\in\{-1,1\}^N} \sum_{i\in L, j\in R} \sum_{x\in[N]} \sigma_i\sigma_j \prod_{\ell=1}^{k-1} Z(x+\ell d_i)Z(x+\ell d_j) \gg_{k,\varepsilon} m^2 N, \tag{6}$$

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 now on we assume that k is odd,[2] and write $k = 2r+1$. For $i, j \in [m]$, denote $P_i(x) = \{x+d_i, x+2d_i, \ldots, x+2rd_i\}$ and $P_{ij}(x) = P_i(x) \cup P_j(x)$. From inequality (6) it follows that we can fix a "good" set $D \in [N]^m$ satisfying

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

and for which we have the technical conditions

$$\left|\left\{i\in L, j\in R :\ |P_{ij}(0)| \neq 4r\right\}\right| \ll_k m^2/N \quad \text{and} \tag{8}$$

$$\max_{x\in[N]} \sum_{i=1}^{m} \sum_{\ell=1}^{2r} \mathbf{1}\{\ell d_i = x\} \ll_k \log N, \tag{9}$$

which are needed to bound the probability of certain bad events later on.

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

$$\mathbb{E}_{\sigma\in\{-1,1\}^m} \left\| \sum_{i\in L, j\in R} \sigma_i\sigma_j M_{ij} \right\|_{\infty\to 1} \tag{10}$$

is related to the expression on the left-hand side of inequality (7). 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 (10); 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 [13] (see also [8]).

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 $M_{ij} \in \mathbb{R}^{\binom{[N]}{s}\times\binom{[N]}{s}}$ by

$$M_{ij}(S,T) = \sum_{x\in[N]} \mathbf{1}\left\{|S\cap P_i(x)| = |S\cap P_j(x)| = r,\ S\triangle T = P_{ij}(x)\right\}$$

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

if $|P_{ij}(0)| = 4r$, and $M_{ij}(S,T) = 0$ if $|P_{ij}(0)| \neq 4r$. From the definition of this matrix, it is not hard to deduce from inequality (7) a lower bound on the expectation (10): one can show that

$$\mathbb{E}_{\sigma\in\{-1,1\}^m}\Big\|\sum_{i\in L, j\in R}\sigma_i\sigma_j M_{ij}\Big\|_{\infty\to 1} \gg_{k,\varepsilon} \binom{N-4r}{s-2r} m^2 N. \tag{11}$$

Now we need to compute an upper bound for the expectation above. The key ingredient for this is the following non-commutative version of Khintchine's inequality, which can be extracted from a result of Tomczak-Jaegermann [15]:

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*

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

In order to apply this inequality, 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 $i \in L$, $\sigma_R \in \{-1,1\}^R$, we define the matrix

$$M_i^{\sigma_R} = \sum_{j\in R}\sigma_j M_{ij}.$$

Applying Theorem 2.1 to the expression

$$\mathbb{E}_{\sigma\in\{-1,1\}^L}\Big\|\sum_{i\in L}\sigma_i M_i^{\sigma_R}\Big\|_2$$

(for some fixed $\sigma_R \in \{-1,1\}^R$) and using properties (8) and (9) to bound the sum $\sum_{i\in L}\|M_i^{\sigma_R}\|_2^2$, one can show (with some effort) that

$$\mathbb{E}_{\sigma\in\{-1,1\}^L}\Big\|\sum_{i\in L}\sigma_i M_i^{\sigma_R}\Big\|_2 \ll_{k,\varepsilon} \sqrt{\log\binom{N}{s}}\cdot m^{1/2}(\log N)^k\frac{m}{N^{1-2/k}} \tag{12}$$

holds whenever $m \geq N^{1-2/k}$ (recall that we choose $s = \lfloor N^{1-2/k}\rfloor$).

Finally, we note that

$$\Big\|\sum_{i\in L, j\in R}\sigma_i\sigma_j M_{ij}\Big\|_{\infty\to 1} = \Big\|\sum_{i\in L}\sigma_i M_i^{\sigma_R}\Big\|_{\infty\to 1} \leq \binom{N}{s}\Big\|\sum_{i\in L}\sigma_i M_i^{\sigma_R}\Big\|_2.$$

Averaging over all signs $\sigma \in \{-1,1\}^m$ and combining inequalities (11) and (12), we conclude that $m \ll_{k,\varepsilon} N^{1-2/k}(\log N)^{2k+1}$. As we started with the assumption (3), this shows that $m_{k-1}^*(N) \ll N^{1-2/k}(\log N)^{2k+1}$ as wished.

References

[1] Omar Alrabiah, Venkatesan Guruswami, Pravesh Kothari, and Peter Manohar, *A near-cubic lower bound for 3-query locally decodable codes from semirandom CSP refutation*, 2022, Electronic Colloquium on Computational Complexity (ECCC). Report no. TR22-101.

[2] Avraham Ben-Aroya, Oded Regev, and Ronald de Wolf, *A hypercontractive inequality for matrix-valued functions with applications to quantum computing and ldcs*, 2008 IEEE 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, 2008, pp. 477–486.

[3] V. Bergelson and A. Leibman, *Polynomial extensions of van der Waerden's and Szemerédi's theorems*, J. Amer. Math. Soc. **9** (1996), no. 3, 725–753. MR 1325795

[4] B. Bollobás and A. Thomason, *Threshold functions*, Combinatorica **7** (1987), no. 1, 35–38. MR 905149

[5] J. Bourgain, *Ruzsa's problem on sets of recurrence*, Israel J. Math. **59** (1987), no. 2, 150–166. MR 920079

[6] Jop Briët, Zeev Dvir, and Sivakanth Gopi, *Outlaw distributions and locally decodable codes*, Theory Comput. **15** (2019), Paper No. 12, 24. MR 4028880

[7] Jop Briët and Sivakanth Gopi, *Gaussian width bounds with applications to arithmetic progressions in random settings*, Int. Math. Res. Not. IMRN (2020), no. 22, 8673–8696. MR 4216700

[8] Jop Briët, *On embeddings of ℓ_1^k from locally decodable codes*, 2016.

[9] Michael Christ, *On random multilinear operator inequalities*, 2011.

[10] Nikos Frantzikinakis, Emmanuel Lesigne, and Máté Wierdl, *Random sequences and pointwise convergence of multiple ergodic averages*, Indiana Univ. Math. J. **61** (2012), no. 2, 585–617. MR 3043589

[11] ———, *Random differences in Szemerédi's theorem and related results*, J. Anal. Math. **130** (2016), 91–133. MR 3574649

[12] Harry Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256. MR 498471

[13] Iordanis Kerenidis and Ronald de Wolf, *Exponential lower bound for 2-query locally decodable codes via a quantum argument*, J. Comput. System Sci. **69** (2004), no. 3, 395–420, Preliminary version in STOC'03. MR 2087942

[14] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27** (1975), 199–245. MR 369312

[15] Nicole Tomczak-Jaegermann, *The moduli of smoothness and convexity and the Rademacher averages of trace classes $S_p(1 \leq p < \infty)$*, Studia Math. **50** (1974), 163–182. MR 355667

[16] P. Varnavides, *Note on a theorem of Roth*, J. London Math. Soc. **30** (1955), 325–326. MR 76797

[17] Trevor D. Wooley and Tamar D. Ziegler, *Multiple recurrence and convergence along the primes*, Amer. J. Math. **134** (2012), no. 6, 1705–1732. MR 2999293