The number of subsets of integers with no k-term arithmetic progression

József Balogh * Hong Liu † Maryam Sharifzadeh ‡

May 11, 2016

Abstract

Addressing a question of Cameron and Erdős, we show that, for infinitely many values of n, the number of subsets of $\{1, 2, \ldots, n\}$ that do not contain a k-term arithmetic progression is at most $2^{O(r_k(n))}$, where $r_k(n)$ is the maximum cardinality of a subset of $\{1, 2, \ldots, n\}$ without a k-term arithmetic progression. This bound is optimal up to a constant factor in the exponent. For all values of n, we prove a weaker bound, which is nevertheless sufficient to transfer the current best upper bound on $r_k(n)$ to the sparse random setting. To achieve these bounds, we establish a new supersaturation result, which roughly states that sets of size $\Theta(r_k(n))$ contain superlinearly many k-term arithmetic progressions.

For integers r and k, Erdős asked whether there is a set of integers S with no $(k+1)$-term arithmetic progression, but such that any r-coloring of S yields a monochromatic k-term arithmetic progression. Nešetřil and Rödl, and independently Spencer, answered this question affirmatively. We show the following density version: for every $k \geq 3$ and $\delta > 0$, there exists a reasonably dense subset of primes S with no $(k+1)$-term arithmetic progression, yet every $U \subseteq S$ of size $|U| \geq \delta|S|$ contains a k-term arithmetic progression.

Our proof uses the hypergraph container method, which has proven to be a very powerful tool in extremal combinatorics. The idea behind the container method is to have a small certificate set to describe a large independent set. We give two further applications in the appendix using this idea.

*Department of Mathematical Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. Email: jobal@math.uiuc.edu. Research is partially supported by NSA Grant H98230-15-1-0002, NSF DMS-1500121 and Arnold O. Beckman Research Award (UIUC Campus Research Board 15006).

†Mathematics Institute and DIMAP, University of Warwick, Coventry, CV4 7AL, UK. Email: h.liu.9@warwick.ac.uk. This research was done while HL was at University of Illinois at Urbana-Champaign.

‡Department of Mathematical Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. Email: sharifz2@illinois.edu.

1 Introduction

Enumerating discrete objects in a given family with certain properties is one of the most fundamental problems in extremal combinatorics. In the context of graphs, this was initiated by Erdős, Kleitman and Rothschild [18] who studied the family of triangle-free graphs. For recent developments, see e.g. [2, 34] and the references therein. In this paper, we investigate counting problems in the arithmetic setting. In this direction, one of the major open problems, raised by Cameron and Erdős [11], was to prove that the number of sum-free sets[1] in $\{1, 2, \ldots, n\}$ is $O(2^{n/2})$. This conjecture was proven independently by Green [24] and Sapozhenko [41]. See also [3, 4] for the proof of another conjecture of Cameron and Erdős [12] concerning the family of maximal sum-free sets.

Our main results are given in Section 1.1: the first is on counting subsets of integers without an arithmetic progression of fixed length (see Theorem 1.2); the second is a new supersaturation result for arithmetic progressions (see Theorem 1.6). In Section 1.2, we show the existence of a set of primes without a $(k+1)$-term arithmetic progression which however is very rich in k-term arithmetic progressions (see Theorem 1.7[2]).

1.1 Enumerating sets with no k-term arithmetic progression

A subset of $[n] := \{1, 2, \ldots, n\}$ is *k-AP-free* if it does not contain a k-term arithmetic progression. Denote by $r_k(n)$ the maximum size of a k-AP-free subset of $[n]$. Cameron and Erdős [11] raised the following question: How many subsets of $[n]$ do not contain a k-term arithmetic progression? In particular, they asked the following question.

Question 1.1 (Cameron-Erdős)**.** *Is it true that the number of k-AP-free subsets of $[n]$ is*

$$2^{(1+o(1))r_k(n)}?$$

Since every subset of a k-AP-free set is also k-AP-free, one can easily obtain $2^{r_k(n)}$ many k-AP-free subsets of $[n]$. In fact, Cameron and Erdős [11] slightly improved this obvious lower bound: writing $R_k(n)$ for the number of k-AP-free subsets of $[n]$, they proved that

$$\limsup_{n\to\infty} \frac{R_k(n)}{2^{r_k(n)}} = \infty. \tag{1}$$

Until recently, the only progress on the upper bound in the last 30 years was improving the bounds on $r_k(n)$. Then Balogh, Morris and Samotij [5], and independently Saxton and Thomason [42], proved the following: for any $\beta > 0$ and integer $k \geq 3$, there exists $C > 0$ such that for $m \geq Cn^{1-1/(k-1)}$, the number of k-AP-free m-sets in $[n]$ is at most $\binom{\beta n}{m}$. This deep counting result implies the sparse random analogue of Szemerédi's theorem [48] which was proved earlier by Conlon and Gowers [13] and independently by Schacht [43]. However, this bound is far from settling Question 1.1.

[1] A set S is *sum-free*, if for any $x, y \in S$, $x + y \notin S$.

[2] A similar result is discussed in the Appendix as well as an additional problem about sumsets.

One of the reasons for the difficulty in finding good upper bounds on $R_k(n)$ is our limited understanding of $r_k(n)$. Indeed, despite much effort, the gap between the current known lower and upper bounds on $r_3(n)$ is still rather large; closing this gap remains one of the most difficult problems in additive number theory. For the lower bound on $r_3(n)$, the celebrated construction of Behrend [6] shows that

$$r_3(n) = \Omega\left(\frac{n}{2^{2\sqrt{2}\sqrt{\log_2 n}} \cdot \log^{1/4} n}\right).$$

This was recently improved by Elkin [15] by a factor of $\sqrt{\log n}$, see also Green and Wolf [29]. Roth [39] gave the first non-trivial upper bound on $r_3(n)$, followed by the improvements of Heath-Brown [31], Szemerédi [47], Bourgain [10] and a recent breakthrough of Sanders [40]. The current best bound is due to Bloom [8]:

$$r_3(n) = O\left(\frac{n(\log\log n)^4}{\log n}\right). \tag{2}$$

For $k \geq 4$ there exist $c_k, c'_k > 0$ such that

$$\frac{n}{2^{c_k(\log n)^{1/(k-1)}}} \leq r_k(n) \leq \frac{n}{(\log\log n)^{c'_k}}, \tag{3}$$

where the lower bound is due to Rankin [46] and the upper bound is by Gowers [22, 23].

Notice that, using the lower bound in (3), we obtain the following trivial upper bound for $R_k(n)$:

$$R_k(n) \leq \sum_{i=0}^{r_k(n)} \binom{n}{i} < 2\binom{n}{r_k(n)} < 2\left(\frac{en}{r_k(n)}\right)^{r_k(n)} = 2^{O\left(r_k(n)\cdot(\log n)^{\frac{1}{k-1}}\right)}.$$

We show that, for infinitely many n, the $(\log n)^{\frac{1}{k-1}}$ term in the exponent is not needed, i.e. our result is optimal up to a constant factor in the exponent.

Theorem 1.2. *The number of k-AP-free subsets of $[n]$ is $2^{O(r_k(n))}$ for infinitely many values of n.*

An immediate corollary of Theorem 1.2 is the following.

Corollary 1.3. *For every $\varepsilon > 0$, there exists a constant $b > 0$ such that the following holds. Let $A(b) \subseteq \mathbb{Z}$ consist of all integers n such that the number of k-AP-free subsets of $[n]$ is at most $2^{b \cdot r_k(n)}$. Then*

$$\limsup_{n\to\infty} \frac{|A(b) \cap [n]|}{n} \geq 1 - \varepsilon.$$

Enumerating discrete structures with certain local constraints is a central topic in combinatorics. Theorem 1.2 is the first such result in which the order of magnitude of the corresponding extremal function is not known.

It is also worth mentioning that two other natural conjectures of Erdős are false: it was conjectured that the number of Sidon sets[3] in $[n]$, denoted by $S(n)$, is $2^{(1+o(1))s(n)}$, where $s(n)$ denotes the size of a maximum Sidon set. However, it is known that $2^{1.16s(n)} \leq S(n) = 2^{O(s(n))}$, where the lower bound is by Saxton and Thomason [42] and the upper bound is by Kohayakawa, Lee, Rödl and Samotij [33] (see also [42]). Another conjecture of Erdős states that the number of C_6-free[4] graphs on vertex set $[n]$, denoted by $H(n)$, is $2^{(1+o(1))\mathrm{ex}(n,C_6)}$. However, $2^{1.0007\mathrm{ex}(n,C_6)} \leq H(n) = 2^{O(\mathrm{ex}(n,C_6))}$, where the lower bound is by Morris and Saxton [34] and the upper bound is by Kleitman and Wilson [32]. In view of these examples and (1), it is not inconceivable that the answer to Question 1.1 is no.

For all values of n, we obtain the following weaker counting estimate, which is nevertheless sufficient to improve previous transference theorems for Szemerédi's theorem, in particular implies Corollary 1.5.

Theorem 1.4. *If $r_k(n) \leq \frac{n}{h(n)}$, where $h(n) \leq (\log n)^c$ for some $c > 0$, then the number of k-AP-free subsets of $[n]$ is at most $2^{O(n/h(n))}$. Furthermore, for any $\gamma > 0$, there exists $C = C(k, c, \gamma) > 0$ such that for any $m \geq n^{1-\frac{1}{k-1}+\gamma}$, the number of k-AP-free m-subsets of $[n]$ is at most*

$$\binom{Cn/h(n)}{m}.$$

Theorem 1.4 improves the counting result of Balogh-Morris-Samotij [5] and Saxton-Thomason [42] with a slightly weaker bound on m. We say that a set $A \subseteq \mathbb{N}$ is (δ, k)-*Szemerédi* if every subset of A of size at least $\delta|A|$ contains a k-AP. Denote by $[n]_p$ the p-random subset of $[n]$, where each element of $[n]$ is chosen with probability p independently of others. As mentioned earlier, the counting result of [5] and [42] implies the following sparse analogue of Szemerédi's theorem, which was only recently proved by a breakthrough transference theorem of Conlon and Gowers [13] and Schacht [43]: For any constant $\delta > 0$ and integer $k \geq 3$, there exists $C > 0$, such that almost surely $[n]_p$ is (δ, k)-Szemerédi for $p \geq Cn^{-\frac{1}{k-1}}$. As an easy corollary of Theorem 1.4, we obtain the following sharper version, in which δ could be taken as a function of n. In fact, it transfers the current best bounds on $r_k(n)$ of Bloom [8] and Gowers [22, 23] to the random setting. Proving Corollary 1.5 from Theorem 1.4 is similar as in [5] and [42], thus we omit the proof here. We remark that the bound on p is optimal up to the additive error term γ in the exponent.

Corollary 1.5. *If $r_k(n) \leq \frac{n}{h(n)}$, where $h(n) \leq (\log n)^c$ for some constant $c > 0$, then for any $\gamma > 0$, there exists $C = C(k, c, \gamma) > 0$ such that the following holds. If $p_n \geq n^{-\frac{1}{k-1}+\gamma}$ for all*

[3] A set $A \subseteq [n]$ is a *Sidon* set if there do not exist distinct $a, b, c, d \in A$ such that $a + b = c + d$.

[4] Denote by C_k the cycle of length k. Given a graph H, a graph G is H-*free* if G does not contain H as a subgraph. Denote by $\mathrm{ex}(n, G)$ the maximum number of edges a G-free graph can have.

sufficiently large n, then

$$\lim_{n\to\infty} \mathbb{P}\left([n]_{p_n} \text{ is } \left(\frac{C}{h(n)}, k\right)\text{-Szemerédi}\right) = 1.$$

Combining the upper bounds in (2) and (3) with Corollary 1.5, for some $C > 0$, we have that almost surely $[n]_p$ is $\left(\frac{C(\log\log n)^4}{\log n}, 3\right)$-Szemerédi for $p \geq n^{-\frac{1}{2}+o(1)}$; and for $k \geq 4$ that almost surely $[n]_p$ is $\left(\frac{C}{(\log\log n)^{c'_k}}, k\right)$-Szemerédi for $p \geq n^{-\frac{1}{k-1}+o(1)}$.

The proof of Theorem 1.2 uses the hypergraph container method, developed by Balogh, Morris and Samotij [5], and independently by Saxton and Thomason [42]. In order to apply the hypergraph container method, we need a supersaturation result. Supersaturation problems are reasonably well-understood if the extremal family is of positive density. For example, the largest sum-free subset of $[n]$ has size $\lceil n/2 \rceil$, while any set of size $(\frac{1}{2}+\varepsilon)n$ has $\Omega(n^2)$ triples satisfying $x+y=z$ (see [26]). In the context of graphs, the Erdős-Stone theorem gives[5] $\mathrm{ex}(n, G) = (1 - \frac{1}{\chi(G)-1} + o(1))\frac{n^2}{2}$, while any n-vertex graph with $(1 - \frac{1}{\chi(G)-1} + \varepsilon)\frac{n^2}{2}$ edges contains $\Omega(n^{|V(G)|})$ copies of G. However, the degenerate case is significantly harder. Indeed, a famous unsolved conjecture of Erdős and Simonovits [19] in extremal graph theory asks whether an n-vertex graph with $\mathrm{ex}(n, C_4) + 1$ edges has at least two copies of C_4.

For arithmetic progressions, the supersaturation result concerned only sets of size linear in n, more precisely Varnavides [49] proved that any subset of $[n]$ of size $\Omega(n)$ has $\Omega(n^2)$ k-APs. (see also [14]). More recently, Croot and Sisask [14] proved a nice formula, which is unfortunately not helping when $|A| \leq O(r_k(n))$ and $r_k(n) \ll n/f(n)$ where $f(n)$ is a polylogarithmic function. Their formula is that for every $A \subset [n]$, and every $1 \leq M \leq n$, the number of 3-APs in A is at least

$$\left(\frac{|A|}{n} - \frac{r_3(M)+1}{M}\right) \cdot \frac{n^2}{M^4}.$$

We need a supersaturation for sets of size $\Theta(r_k(n))$. Our second main result shows that the number of k-APs in any set A of size constant factor times greater than $r_k(n)$ is superlinear in n.

Theorem 1.6. *Given $k \geq 3$, there exists a constant $C' = C'(k) > 0$ and an infinite sequence $\{n_i\}_{i=1}^{\infty}$, such that the following holds. For any $n \in \{n_i\}_{i=1}^{\infty}$ and any $A \subseteq [n]$ of size $C' r_k(n)$, the number of k-APs in A is at least*

$$\log^{3k-2} n \cdot \left(\frac{n}{r_k(n)}\right)^{k-1} \cdot n.$$

[5]The *chromatic number* of G, denoted by $\chi(G)$, is the minimum number of colors needed to color the vertices of G such that no two adjacent vertices receive the same color.

1.2 Arithmetic progressions in the primes

The study of arithmetic progressions in the set of primes has witnessed great advances in the last decade. Extending the seminal result of Szemerédi [48], Green and Tao in their landmark paper [28] proved that any subset of the primes with positive relative density contains arbitrarily long arithmetic progression. In fact, they showed the following supersaturation version. Denote by $\mathbf{P}_{\leq n}$ the set of primes which are at most n. Then the number of k-APs in any subset $U \subseteq \mathbf{P}_{\leq n}$ with $|U| = \Omega(|\mathbf{P}_{\leq n}|)$ is $\Theta(n^2/\log^k n)$. We are interested in the following question: does there exist a subset of primes that is $(k+1)$-AP-free, yet any subset of it with positive density contains a k-AP? A priori, it is not even clear whether such a set exists in $\mathbb{Z}$, since intuitively a $(k+1)$-AP-free set is unlikely to be rich in k-APs. It is worth mentioning that Erdős [16] asked whether, for every integer r, there is a set of integers with no $(k+1)$-AP, but any r-coloring of it yields a monochromatic k-AP. Spencer [45] proved the existence of such a set and Nešetřil and Rödl [37] constructed such a set. The question raised above is a strengthening of Erdős' in two aspects: it is a density version and asks for a set of primes. Our next result gives an affirmative answer to this question.

Theorem 1.7. *For any $\delta > 0$ and $k \geq 3$, there exists a set of primes $S \subseteq \mathbf{P}_{\leq n}$ of size $n^{1-1/k-o(1)}$ such that S is $(k+1)$-AP-free and (δ, k)-Szemerédi.*

One might attempt to find a set of integers with the desired properties and then apply it to a very long arithmetic progression in the primes guaranteed by the Green-Tao theorem. However the subset of primes obtained in this way would be extremely sparse in $\mathbf{P}_{\leq n}$. To obtain a fairly dense subset in $\mathbf{P}_{\leq n}$ with the desired properties given Theorem 1.7, we will instead do things in the "reverse" order. We first use the supersaturation version of the Green-Tao theorem and the container method to get the following counting result, which together with a standard application of the probabilistic method (see for similar and earlier applications of Nenadov and Steger [35] and of Rödl, Rucinski and Schacht [38]) will establish Theorem 1.7.

Theorem 1.8. *For any $\beta > 0$, $\gamma > 0$ and $k \geq 3$, the number of k-AP-free m-subsets of $\mathbf{P}_{\leq n}$ with $m \geq n^{1-\frac{1}{k-1}+\gamma}$ is at most*

$$\binom{\beta|\mathbf{P}_{\leq n}|}{m}.$$

Consequently, the number of k-AP-free subsets of $\mathbf{P}_{\leq n}$ is at most $2^{o(|\mathbf{P}_{\leq n}|)}$.

We omit the proof of Theorem 1.8 since it follows along the same line as that of Theorem 1.4. The only difference is that, for the supersaturation, we use the Green-Tao theorem instead of Lemma 2.4. Similarly to Theorem 1.4, Theorem 1.8 implies the following sparse random analogue of the Green-Tao theorem.

Corollary 1.9. *For any $\delta > 0$ and $\gamma > 0$, if $p_n \geq n^{-\frac{1}{k-1}+\gamma}$ for all sufficiently large n and S_n is a p-random subset of $\mathbf{P}_{\leq n}$, then*

$$\lim_{n\to\infty} \mathbb{P}\left(S_n \text{ is } (\delta, k)\text{-Szemerédi}\right) = 1.$$

Organization. The rest of the paper will be organized as follows. In Section 2, we introduce the hypergraph container method and some lemmas needed for proving supersaturation. In Section 3, we prove our main result, Theorem 1.2, and also Corollary 1.3, Theorem 1.4, Theorem 1.6 and Theorem 1.7.

Notation. We write $[a, b]$ for the interval $\{a, a+1, \ldots, b\}$ and $[n] := [1, n]$. Given a set $A \subseteq [n]$, denote by $\Gamma_k(A)$ the number of k-APs in A. Denote by $\min(A)$ the smallest element in A. We write log for logarithm with base 2. Throughout the paper we omit floors and ceilings where they are not crucial.

2 Preliminaries

In the next subsection, we present the hypergraph container theorem and derive a version tailored for arithmetic progressions. We then prove some supersaturation results needed for the proof of Theorem 1.6 in Section 2.2.

To see how they work, we give a quick overview of the proof of Theorem 1.2. We first apply the hypergraph container theorem (Corollary 2.2) to obtain a small collection of containers covering all k-AP-free sets in $[n]$, each of these containers having only few copies of k-APs. Then we apply the supersaturation result (Theorem 1.6) to show that every container necessarily has to be small in size ($O(r_k(n))$), from which our main result follows.

2.1 The hypergraph container theorem

An r-uniform hypergraph $\mathcal{H} = (V, E)$ consists of a vertex set V and an edge set E, in which every edge is a set of r vertices in V. An *independent* set in $\mathcal{H}$ is a set of vertices inducing no edge in E. The *independence number* $\alpha(\mathcal{H})$ is the maximum cardinality of an independent set in $\mathcal{H}$. Denote by $\chi(\mathcal{H})$ the *chromatic number* of $\mathcal{H}$, i.e., the minimum integer ℓ, such that $V(\mathcal{H})$ can be colored by ℓ colors with no monochromatic edge.

Many classical theorems in combinatorics can be phrased as statements about independent sets in a certain auxiliary hypergraph. For example, the celebrated theorem of Szemerédi [48] states that for $V(\mathcal{H}) = [n]$ and $E(\mathcal{H})$ consisting of all k-term arithmetic progressions in $[n]$, $\alpha(\mathcal{H}) = o(n)$. The cornerstone result of Erdős and Stone [20] in extremal graph theory characterizes the structure of all maximum independent sets in $\mathcal{H}$, where $V(\mathcal{H})$ is the edge set of K_n and $E(\mathcal{H})$ is the edge set of copies of some fixed graph G.

We will use the method of hypergraph containers for the proof of Theorem 1.2. This powerful method was recently introduced independently by Balogh, Morris and Samotij [5], and by Saxton and Thomason [42]. Roughly speaking, it says that if a hypergraph $\mathcal{H}$ has a somewhat uniform edge-distribution, then one can find a relatively small collection of sets covering all independent sets in $\mathcal{H}$. Among others, this method provides an alternative proof of a recent breakthrough transference theorem of Conlon and Gowers [13] and Schacht [43] for extremal results in sparse random setting. We refer the readers to [5, 42] for more details and applications, see also [3, 4] for more recent applications of container-type results in the arithmetic setting.

Let $\mathcal{H}$ be an r-uniform hypergraph with average degree d. For every $S \subseteq V(\mathcal{H})$, its co-degree, denoted by $d(S)$, is the number of edges in $\mathcal{H}$ containing S, i.e.,

$$d(S) = |\{e \in E(\mathcal{H}) : S \subseteq e\}|.$$

For every $j \in [r]$, denote by Δ_j the j-th maximum co-degree of $\mathcal{H}$, i.e.,

$$\Delta_j = \max\{d(S) : S \subseteq V(\mathcal{H}), |S| = j\}.$$

For any $\tau \in (0,1)$, define the following function which controls simultaneously the maximum co-degrees Δ_j's for all $j \in \{2, \dots, r\}$:

$$\Delta(\mathcal{H}, \tau) = 2^{\binom{r}{2}-1} \sum_{j=2}^{r} 2^{-\binom{j-1}{2}} \frac{\Delta_j}{d\tau^{j-1}}.$$

Note that we are interested to have small τ, as smaller τ means smaller family of containers. Here, as the codegrees are relatively small, only the $\Delta_r/(d \cdot \tau^{r-1})$ part matters.

We need the following version of the hypergraph container theorem (Corollary 3.6 in [42]).

Theorem 2.1. *Let $\mathcal{H}$ be an r-uniform hypergraph on vertex set $[n]$. Let $0 < \varepsilon, \tau < 1/2$. Suppose that $\tau < 1/(200r!^2 r)$ and $\Delta(\mathcal{H}, \tau) \le \varepsilon/(12r!)$. Then there exists $c = c(r) \le 1000r!^3 r$ and a collection of vertex subsets $\mathcal{C}$ such that*

(i) every independent set in $\mathcal{H}$ is a subset of some $A \in \mathcal{C}$;

(ii) for every $A \in \mathcal{C}$, $e(\mathcal{H}[A]) \le \varepsilon e(\mathcal{H})$;

(iii) $\log |\mathcal{C}| \le cn\tau \log(1/\varepsilon) \log(1/\tau)$.

Given an integer $k \ge 3$, consider the k-uniform hypergraph $\mathcal{H}_k$ encoding the set of all k-APs in $[n]$: $V(\mathcal{H}_k) = [n]$ and the edge set of $\mathcal{H}_k$ consists of all k-tuples that form a k-AP. It is easy to check that the number of k-APs in $[n]$ is $n^2/(2k) < e(\mathcal{H}_k) < n^2/k$. Note that $\Delta_1 \le k \cdot \frac{n}{k-1} < 2n$ and

$$d = d(\mathcal{H}_k) \ge \frac{n}{2}, \quad \Delta_k = 1, \quad \Delta_i \le \Delta_2 \le \binom{k}{2} < k^2 \ \text{ for } 2 \le i \le k-1. \tag{4}$$

Using the k-AP-hypergraph $\mathcal{H}_k$, we obtain the following adaption of Theorem 2.1 to the arithmetic setting.

Corollary 2.2. *Fix an arbitrary integer $k \ge 3$ and let $0 < \varepsilon, \tau < 1/2$ be such that*

$$\tau < 1/(200k^{2k}) \quad \text{and} \quad \varepsilon n \tau^{k-1} > k^{3k}. \tag{5}$$

Then for sufficiently large n, there exists a collection $\mathcal{C}$ of subsets of $[n]$ such that

(i) every k-AP-free subset of $[n]$ is contained in some $F \in \mathcal{C}$;

(ii) for every $F \in \mathcal{C}$, the number of k-APs in F is at most εn^2;

(iii) $\log |\mathcal{C}| \le 1000k^{3k} n\tau \log(1/\varepsilon) \log(1/\tau)$.

Proof. Consider the k-AP hypergraph $\mathcal{H}_k$. Fix any $0 < \varepsilon, \tau < \frac{1}{2}$ such that $\tau < \frac{1}{200k^{2k}} < 2^{-3k}$ and $\varepsilon n \tau^{k-1} > k^{3k}$. Define $\alpha_j := 2^{-\binom{j-1}{2}} \cdot \tau^{-(j-1)}$ for $2 \le j \le k$. Since $\tau < 2^{-3k}$, we have that for $2 \le j \le k-2$,

$$\frac{\alpha_j}{\alpha_{j+1}} = \frac{2^{\binom{j}{2}} \cdot \tau^j}{2^{\binom{j-1}{2}} \cdot \tau^{j-1}} = 2^{j-1}\tau < 2^k \tau < 1 \quad \text{and} \quad \frac{k^3 \alpha_{k-1}}{\alpha_k} = k^3 2^{k-2}\tau < 1. \tag{6}$$

Note that for any $k \ge 3$, we have that $\tau < 1/(200k^{2k}) < 1/(200k!^2 k)$. We now bound the function $\Delta(\mathcal{H}_k, \tau)$ from above as follows:

$$\begin{aligned}
\Delta(\mathcal{H}_k, \tau) &= 2^{\binom{k}{2}-1} \sum_{j=2}^{k} \alpha_j \frac{\Delta_j}{d} \overset{(4)}{\le} 2^{\binom{k}{2}-1} \left(\sum_{j=2}^{k-1} \alpha_j \frac{k^2}{d} + \frac{\alpha_k}{d} \right) \overset{(6)}{\le} 2^{\binom{k}{2}-1} \left((k-2)\alpha_{k-1} \frac{k^2}{d} + \frac{\alpha_k}{d} \right) \\
&\overset{(6)}{\le} 2^{\binom{k}{2}-1} \cdot \frac{2\alpha_k}{d} = \frac{2^{k-1}}{d\tau^{k-1}} \overset{(4)}{\le} \frac{2^k}{n\tau^{k-1}} \overset{(5)}{\le} \frac{\varepsilon}{12k!}.
\end{aligned}$$

We now apply Theorem 2.1 on $\mathcal{H}_k$ to obtain $\mathcal{C}$. Then the conclusions follow from the observation that every independent set in $\mathcal{H}_k$ is a k-AP-free subset of $[n]$. □

2.2 Supersaturation

In this subsection, we present the second main ingredient for the proof of Theorem 1.2: a supersaturation result, Lemma 2.4, which states that many k-APs start to appear in a set once its size is larger than $r_k(n)$.

First notice that for any $A \subseteq [n]$ of size $K \cdot r_k(n)$, the following greedy algorithm gives

$$\Gamma_k(A) \ge (K-1) \cdot r_k(n). \tag{7}$$

Set $B := A$. Repeat the following process $(K-1) \cdot r_k(n)$ times: since $|B| > r_k(n)$, there is a k-AP in B; update B by removing an arbitrary element in this k-AP. We use a random sparsening trick to improve this simple argument.

Lemma 2.3. *For every $A \subseteq [n]$ of size $K \cdot r_k(n)$ with $K \ge 2$, we have*

$$\Gamma_k(A) \ge \left(\frac{K}{2} \right)^k \cdot r_k(n).$$

Proof. Let T be a set chosen uniformly at random among all subsets of A of size $2r_k(n)$. Then the expected number of k-APs in T is

$$\mathbb{E}[\Gamma_k(T)] = \frac{\binom{|A|-k}{|T|-k}}{\binom{|A|}{|T|}} \cdot \Gamma_k(A) \le \left(\frac{|T|}{|A|} \right)^k \cdot \Gamma_k(A) = \frac{\Gamma_k(A)}{(K/2)^k}.$$

Thus, there exists a choice of T such that $\Gamma_k(T) \le \frac{\Gamma_k(A)}{(K/2)^k}$. On the other hand, from (7), $\Gamma_k(T) \ge r_k(n)$, hence $\Gamma_k(A) \ge \left(\frac{K}{2}\right)^k \cdot r_k(n)$ as desired. □

However, the bound given above is still linear in $|A|$, which is not sufficient for our purposes. A superlinear bound is provided in the following lemma, which implies that $\Gamma_k(A) \geq |A| \cdot \mathrm{polylog}(n)$ for infinitely many values of n (as in Theorem 1.6). A key new idea in our proof is that an averaging argument is carried out only over a set of carefully chosen arithmetic progressions with prime common differences. To obtain a superlinear bound, we will apply the following lemma with roughly $M \sim |A| \cdot \left(\frac{|A|}{n}\right)^{k+1}$, and $|A| \geq r_k(n)$.

Lemma 2.4. *For any $1 \leq M \leq n$ and $A \subseteq [n]$, if $|A|/M$ is sufficiently large and $|A|/n \geq 8K \cdot r_k(M)/M$ with $K \geq 2$, then*

$$\Gamma_k(A) \geq \frac{|A|^2}{M^2} \cdot \frac{K^k \cdot r_k(M)}{2^{k+4} \log^2 n}.$$

Proof. Define $x = |A|/(4M)$, and assume that it is sufficiently large. Then the Prime Number Theorem (see e.g. [44]) implies that the number of prime numbers less than x is at least $x/\log x$ and at most $2x/\log x$. Denote $\mathcal{B}_d$ the set of M-term arithmetic progressions with common difference d in $[n]$ and set

$$\mathcal{B} := \bigcup_{\substack{d \text{ is prime} \\ d \leq x}} \mathcal{B}_d,$$

that is, $\mathcal{B}$ consists of all M-APs whose common difference is a prime number not larger than x. We notice first that any k-AP can occur in at most $M \log n$ many members of $\mathcal{B}$. Indeed, fix an arbitrary k-AP, say Q', with common difference d'. Note that every M-AP Q containing Q' can be constructed in two steps:

(i) choose $1 \leq i \leq M$ and set the i-th term of Q to $\min(Q')$;

(ii) choose the common difference d for Q.

There are clearly at most M choices for (i). As for (ii), in order to have $Q' \subseteq Q$, we need $d|d'$. Since $Q \in \mathcal{B}$, d must be a prime divisor of d'. Using that the number of prime divisors of d' is at most $\log d' \leq \log n$, the number of such choices is at most $\log n$. As a consequence, we have that

$$\Gamma_k(A) \geq \frac{1}{M \log n} \sum_{B \in \mathcal{B}} \Gamma_k(A \cap B). \tag{8}$$

Let $\mathcal{G} \subseteq \mathcal{B}$ consists of all $B \in \mathcal{B}$ such that $|A \cap B| \geq K \cdot r_k(M)$. Then by Lemma 2.3, we have $\Gamma_k(A \cap B) \geq (K/2)^k \cdot r_k(M)$ for every $B \in \mathcal{G}$. Together with (8), this gives that

$$\Gamma_k(A) \geq \frac{1}{M \log n} \sum_{B \in \mathcal{G}} \Gamma_k(A \cap B) \geq |\mathcal{G}| \cdot \frac{K^k \cdot r_k(M)}{2^k M \log n}. \tag{9}$$

Our next goal is to give a lower bound on $|\mathcal{G}|$, to achieve this, we will do a double-counting on $\sum_{B \in \mathcal{B}} |A \cap B|$.

For each $d \le x$, define $I_d := [(M-1)d+1, n-(M-1)d]$. Then every $z \in I_d$ appears in exactly M members of $\mathcal{B}_d$. Since $x = |A|/(4M)$,

$$|A \cap I_d| = |A| - 2(M-1)d \ge |A| - 2Mx \ge \frac{|A|}{2}.$$

As an immediate consequence of the Prime Number Theorem, the number of primes less than x, which is the number of choices for d, is at least $x/\log x$ and at most $2x/\log x$ for sufficiently large x. Therefore,

$$\sum_{B\in\mathcal{B}} |A\cap B| = \sum_{\substack{d \text{ is prime} \\ d\le x}} \sum_{B\in\mathcal{B}_d} |A\cap B| \ge M \sum_{\substack{d \text{ is prime} \\ d\le x}} |A\cap I_d| \ge M\cdot\frac{x}{\log x}\cdot\frac{|A|}{2}. \tag{10}$$

On the other hand, since $|\mathcal{B}_d| < n$, for each d we have $|\mathcal{B}| \le \frac{2x}{\log x}\cdot n$, hence

$$\sum_{B\in\mathcal{B}} |A\cap B| \le M|\mathcal{G}| + K\cdot r_k(M)\cdot|\mathcal{B}\setminus\mathcal{G}| \le M|\mathcal{G}| + K\cdot r_k(M)\cdot\frac{2xn}{\log x}. \tag{11}$$

Combining (10) and (11), we get

$$\begin{aligned} |\mathcal{G}| &\ge \frac{x}{\log x}\cdot\frac{|A|}{2} - K\cdot\frac{r_k(M)}{M}\cdot\frac{2xn}{\log x} = \frac{x}{\log x}\left(\frac{|A|}{2} - 2K\cdot\frac{r_k(M)}{M}\cdot n\right) \\ &\ge \frac{x}{\log n}\cdot\frac{|A|}{4} = \frac{|A|^2}{16M\log n}, \end{aligned}$$

where the last inequality follows from $\frac{|A|}{n} \ge 8K\cdot\frac{r_k(M)}{M}$. Thus, by (9), we have

$$\Gamma_k(A) \ge \frac{|A|^2}{16M\log n}\cdot\frac{K^k\cdot r_k(M)}{2^k M\log n} = \frac{|A|^2}{M^2}\cdot\frac{K^k\cdot r_k(M)}{2^{k+4}\log^2 n}.$$

□

3 Proof of Theorem 1.2

Throughout this section, we fix k a positive integer and write $r(n)$ instead of $r_k(n)$ and define $f(n) = r(n)/n$. We will use the following functions:

$$M(n) = \frac{n}{\log^{3k} n}\left(\frac{r(n)}{n}\right)^{k+2},\ \varepsilon(n) = \frac{\log^{3k-2} n}{n}\left(\frac{n}{r(n)}\right)^{k-1},\ \tau(n) = \frac{r(n)}{n}\frac{1}{\log^3 n}. \tag{12}$$

We first observe a simple fact about the function $r(n)$. Since the property of having no k-AP is invariant under translation, for any given $m < n$, if we divide $[n]$ into consecutive intervals of length m, then any given k-AP-free subset of $[n]$ contains at most $r(m)$ elements from each interval. Thus,

$$r(n) \le \left\lceil \frac{n}{m} \right\rceil \cdot r(m). \tag{13}$$

Since $\frac{1}{n}\cdot\left\lceil\frac{n}{m}\right\rceil < \frac{2}{m}$ for any $m < n$, dividing by n on both sides of (13) yields:

Fact 3.1. *For every $m < n$, $f(n) < 2f(m)$.*

For the proof of Theorem 1.6, we will apply Lemma 2.4 with M defined as in (12). To do so, it requires that the function $r(n)$ is "smooth". This is proved in the following lemma.

Lemma 3.2. *Given $k \geq 3$, there exists $C := C(k) > 4$ and an infinite sequence $\{n_i\}_{i=1}^{\infty}$, such that*

$$C\frac{r(n_i)}{n_i} \geq \frac{r(M(n_i))}{M(n_i)}$$

for all $i \geq 1$, where $M(n)$ is defined as in (12).

Proof. Fix $C = C(k) > 4$ a sufficiently large constant. From Behrend's construction, we know that $f(n) > 2^{-5\sqrt{\log n}}$. We need to show that, for infinitely many n, $Cf(n) \geq f(M(n)) = f\left(\frac{n}{\log^{3k} n} f(n)^{k+2}\right)$. Suppose to the contrary, that for all but finitely many n, $f(n) \leq C^{-1} f(M(n))$. Let n_0 be the largest integer such that $f(n) > C^{-1} f(M(n))$.

Define a decreasing function $g(x) = 2^{-(5k+11)\sqrt{\log x}}$ for $x \geq 1$. Note that for sufficiently large n, since $f(n) > 2^{-5\sqrt{\log n}}$,

$$M(n) = \frac{n}{\log^{3k} n} f(n)^{k+2} > \frac{n}{\log^{3k} n} \cdot 2^{-5(k+2)\sqrt{\log n}} > n \cdot 2^{-(5k+11)\sqrt{\log n}} = n \cdot g(n).$$

Then by Fact 3.1, we have $f(M(n)) < 2f(n \cdot g(n))$. Therefore, by the definition of n_0, we have that for any $n > n_0$,

$$f(n) \leq C^{-1} f(M(n)) < \left(\frac{C}{2}\right)^{-1} f(n \cdot g(n)). \tag{14}$$

Fix an integer $n > n_0^2$ and set $t = \lfloor \frac{1}{2} \frac{\sqrt{\log n}}{5k+11} \rfloor$. We will show by induction that for every $1 \leq j \leq t$,

$$f(n) < \left(\frac{C}{4}\right)^{-j} f(n \cdot g(n)^j). \tag{15}$$

The base case, $j = 1$, is given by (14). Suppose (15) holds for some $1 \leq j < t$. Define $n' := n \cdot g(n)^j$. Then

$$n' > n \cdot g(n)^t = n \cdot 2^{-(5k+11)\sqrt{\log n} \cdot \lfloor \frac{1}{2} \frac{\sqrt{\log n}}{5k+11} \rfloor} \geq n \cdot 2^{-\frac{1}{2}\log n} = \sqrt{n} > n_0.$$

So by (14), $f(n') < \left(\frac{C}{2}\right)^{-1} f(n' \cdot g(n'))$. Since $n' < n$ and $g(x)$ is decreasing, $n' \cdot g(n') > n' \cdot g(n)$. Then by Fact 3.1, $f(n' \cdot g(n')) < 2f(n' \cdot g(n))$. Hence, $f(n') < \left(\frac{C}{4}\right)^{-1} f(n' \cdot g(n))$. Thus by the induction hypothesis

$$\begin{aligned} f(n) &< \left(\frac{C}{4}\right)^{-j} f(n \cdot g(n)^j) = \left(\frac{C}{4}\right)^{-j} f(n') \\ &< \left(\frac{C}{4}\right)^{-j} \left(\frac{C}{4}\right)^{-1} f(n' \cdot g(n)) = \left(\frac{C}{4}\right)^{-(j+1)} f(n \cdot g(n)^{j+1}). \end{aligned}$$

This proves (15) for $j = t$ and note that $f(n) \le 1$ and that $f(n \cdot g(n)^t) < 2f(\sqrt{n})$ by Fact 3.1, hence

$$f(n) < \left(\frac{C}{4}\right)^{-t} f(n \cdot g(n)^t) < \left(\frac{C}{4}\right)^{-t} \cdot 2f(\sqrt{n}) \le 2\left(\frac{C}{4}\right)^{-t} = 2\left(\frac{C}{4}\right)^{-\lfloor \frac{1}{2}\frac{\sqrt{\log n}}{5k+11} \rfloor} < 2^{-5\sqrt{\log n}}$$

for C sufficiently large, a contradiction. □

Theorem 1.6 follows immediately from Lemmas 2.4 and 3.2.

Proof of Theorem 1.6. Let K be the constant from Lemma 2.4. Let C be the constant and $\{n_i\}_{i=1}^{\infty}$ be the sequence from Lemma 3.2. Define $C' = 8CK$. Fix an arbitrary $n \in \{n_i\}_{i=1}^{\infty}$ and write $M = M(n)$ as defined in (12). Let $A \subseteq [n]$ be an arbitrary set of size $C'r(n)$. Then by Lemma 3.2,

$$\frac{|A|}{n} = \frac{8CK \cdot r(n)}{n} \ge 8K\frac{r(M(n))}{M(n)}.$$

By Fact 3.1, $\frac{2r(M)}{M} > \frac{r(n)}{n}$. Thus by Lemma 2.4 and that $K \ge 2$, $C > 4$, we have

$$\begin{aligned}\Gamma_k(A) &> \frac{|A|^2}{M^2} \cdot \frac{K^k \cdot r(M)}{2^{k+4}\log^2 n} = \frac{(8CK)^2 r(n)^2}{M^2} \cdot \frac{K^k \cdot r(M)}{2^{k+4}\log^2 n} = \frac{r(n)^2}{M\log^2 n} \cdot \frac{2r(M)}{M} \cdot \frac{(8CK)^2 K^k}{2^{k+5}} \\ &> \frac{r(n)^2}{M\log^2 n} \cdot \frac{2r(M)}{M} > \frac{r(n)^2}{M\log^2 n} \cdot \frac{r(n)}{n} = \log^{3k-2} n \left(\frac{n}{r(n)}\right)^{k-1} n.\end{aligned}$$

□

Proof of Theorem 1.2. Let $\{n_i\}_{i=1}^{\infty}$ be the infinite sequence guaranteed by Lemma 3.2. We will show that the conclusion holds for this sequence of values of n. Let $M = M(n), \varepsilon = \varepsilon(n)$ and $\tau = \tau(n)$ be as defined in (12). For sufficiently large n, we have that $\tau < \frac{1}{200k^{2k}}$ and

$$\varepsilon n \tau^{k-1} = \frac{\log^{3k-2} n}{n}\left(\frac{n}{r(n)}\right)^{k-1} \cdot n \cdot \left(\frac{r(n)}{n}\frac{1}{\log^3 n}\right)^{k-1} = \log n > k^{3k}.$$

Thus by Corollary 2.2, there is a family $\mathcal{C}$ of containers such that every k-AP-free subset of $[n]$ is a subset of some container in $\mathcal{C}$. By (12), $\log\frac{1}{\varepsilon}\log\frac{1}{\tau} < \log^2 n$, thus

$$\log|\mathcal{C}| \le 1000k^{3k} n\tau \log\frac{1}{\varepsilon}\log\frac{1}{\tau} < 1000k^{3k} n \cdot \frac{r(n)}{n}\frac{1}{\log^3 n} \cdot \log^2 n = o(r(n)).$$

Since for every container $A \in \mathcal{C}$, the number of k-APs in A is at most εn^2, then by Theorem 1.6, $|A| < C'r(n)$ for every $A \in \mathcal{C}$. Recall that every k-AP-free subset is contained in some member of $\mathcal{C}$. Hence, the number of k-AP-free subsets of $[n]$ is at most

$$\sum_{A\in\mathcal{C}} 2^{|A|} \le |\mathcal{C}| \cdot \max_{A\in\mathcal{C}} 2^{|A|} < 2^{o(r(n))} \cdot 2^{C'r(n)} = 2^{O(r(n))}.$$

□

Proof of Corollary 1.3. Let $\{n_i\}_{i=1}^{\infty}$ be a sequence of integers for which the conclusion of Theorem 1.2 holds. Fix an arbitrary $\varepsilon > 0$ and n_i. From Theorem 1.2, we know that the number of k-AP-free subsets of $[n_i]$ is at most $2^{c \cdot r(n_i)}$ for some absolute constant $c > 0$. For any $\varepsilon n_i \le m < n_i$, by (13), we have that $r(n_i) \le \left\lceil \frac{1}{\varepsilon} \right\rceil \cdot r(m) < \frac{2}{\varepsilon} \cdot r(m)$. Therefore, by setting $b = 2c/\varepsilon$, we have that the number of k-AP-free subsets of $[m]$ is at most $2^{c \cdot r(n_i)} \le 2^{b \cdot r(m)}$. It then follows that $m \in A(b)$ for any $\varepsilon n_i \le m < n_i$ and that $|A(b) \cap [n_i]|/n_i \ge 1 - \varepsilon$ as desired. □

The proof of Theorem 1.4 is along the same line as of the proof of Theorem 1.2, hence we provide here only a sketch of it. The difference is that to prove this weaker bound, we only need supersaturation results for sets of size $n/\mathrm{polylog}\, n$. For sets of this size, we do not need the technical condition in Lemma 3.2 and we can invoke Lemma 2.4 with $M = n^{o(1)}$ for all values of n instead of $M = n^{1-o(1)}$ as in proof of Theorem 1.2.

Proof of Theorem 1.4. Fix an arbitrary $0 < \gamma < 1$. We apply Corollary 2.2 with $\varepsilon = n^{-\gamma/2}$, $\tau = n^{-\frac{1}{k-1}+\gamma/2}$ and let $\mathcal{C}$ be the family of containers of size $\log|\mathcal{C}| = o(n^{1-\frac{1}{k-1}+\gamma})$. Each container contains at most $\varepsilon n^2 = n^{2-\gamma/2}$ many k-APs. It follows that for every $A \in \mathcal{C}$, $|A| \le \frac{C'n}{h(n)}$ for some $C' = C'(k, c, \gamma)$, since otherwise applying Lemma 2.4 on A with $M = n^{\gamma/4}$ would imply $\Gamma_k(A) > n^{2-\gamma/3} > \varepsilon n^2$, a contradiction. Thus, the number of k-AP-free subsets of $[n]$ is at most $|\mathcal{C}| \cdot 2^{C'n/h(n)} = 2^{2C'n/h(n)}$, as desired. Similarly, the number of k-AP-free m-subsets of $[n]$ is at most $|\mathcal{C}| \cdot \binom{C'n/h(n)}{m} \le 2^m \cdot \binom{C'n/h(n)}{m} \le \binom{2C'n/h(n)}{m}$, where the first inequality follows from $m \ge n^{1-\frac{1}{k-1}+\gamma} \ge \log|\mathcal{C}|$. □

Proof of Theorem 1.7. Fix an arbitrary $\delta > 0$ and an integer $k \ge 3$. Let $S \subseteq \mathbf{P}_{\le n}$ be a random subset of $\mathbf{P}_{\le n}$, in which each element is chosen with probability $p = n^{-1/k}$ independently of others. A standard application of Chernoff bound implies that $|S| \ge p|\mathbf{P}_{\le n}|/2$ with probability $1 - o(1)$.

Set $\beta = \delta/50, \gamma = 1/(10k^2)$ and $m = \delta p|\mathbf{P}_{\le n}|/5$. Then by the Prime Number Theorem, we have

$$m = \Omega\left(\frac{n^{1-1/k}}{\log n}\right) > n^{1-\frac{1}{k-1}+\gamma}.$$

Thus by Theorem 1.8 the number of k-AP-free m-subsets in $\mathbf{P}_{\le n}$ is at most $\binom{\beta|\mathbf{P}_{\le n}|}{m}$.

Let X be the number of k-AP-free m-subsets in S, and Y be the number of $(k+1)$-APs in S. Then,

$$\mathbb{E}[X] \le \binom{\beta|\mathbf{P}_{\le n}|}{m} p^m \le \left(\frac{e \cdot \beta|\mathbf{P}_{\le n}|}{m} \cdot p\right)^m = \left(\frac{e \cdot (\delta/50)|\mathbf{P}_{\le n}| \cdot p}{\delta p|\mathbf{P}_{\le n}|/5}\right)^m = \left(\frac{e}{10}\right)^m = o(1).$$

Thus by Markov's inequality, $X = 0$ with probability at least $2/3$.

By the Green-Tao theorem, the number of $(k+1)$-APs in $\mathbf{P}_{\le n}$ is $\Theta\left(\frac{n^2}{(\log n)^{k+1}}\right)$. Thus,

$$\mathbb{E}[Y] \le \frac{n^2}{\log^k n} \cdot p^{k+1} = \frac{n^{1-1/k}}{\log^k n}.$$

We have, by Markov's inequality, that $Y \leq \frac{3n^{1-1/k}}{\log^k n}$ with probability at least $2/3$. Therefore, with positive probability, there is a choice of S such that $|S| \geq p|\mathbf{P}_{\leq n}|/2$, $X = 0$ and $Y \leq \frac{3n^{1-1/k}}{\log^k n}$. Let S' be the set obtained from S by deleting one element from every $(k+1)$-AP in S. Then

$$|S'| = |S| - Y \geq \frac{p|\mathbf{P}_{\leq n}|}{2} - \frac{3n^{1-1/k}}{\log^k n} \geq \frac{n^{1-1/k}}{3\log n} - \frac{3n^{1-1/k}}{\log^k n} \geq \frac{n^{1-1/k}}{4\log n}.$$

We claim that S' has the desired property. Indeed, clearly S' is $(k+1)$-AP-free. Suppose S' is not (δ, k)-Szemerédi, then there exists a k-AP-free subset $U \subseteq S'$ of size

$$|U| \geq \delta|S'| \geq \delta \cdot \frac{n^{1-1/k}}{4\log n} > \frac{\delta p|\mathbf{P}_{\leq n}|}{5} = m.$$

However, this contradicts that $X = 0$. □

Acknowledgement

We would like to thank József Solymosi who suggested considering the set of primes instead of integers in Theorem 1.7. We also thank the anonymous referee for the helpful comments which greatly improved the presentation of this paper.

References

[1] N. Alon, Large sets in finite fields are sumsets, *J. Number Theory*, (126) 2007, 110–118.

[2] J. Balogh, H. Liu, S. Petříčková and M. Sharifzadeh, The typical structure of maximal triangle-free graphs, *Forum of Mathematics, Sigma*, (3) 2015.

[3] J. Balogh, H. Liu, M. Sharifzadeh and A. Treglown, The number of maximal sum-free subsets of integers, *Proc. Amer. Math. Soc.*, (143) 2015, 4713–4721.

[4] J. Balogh, H. Liu, M. Sharifzadeh and A. Treglown, Sharp bound on the number of maximal sum-free subsets of integers, submitted.

[5] J. Balogh, R. Morris and W. Samotij, Independent sets in hypergraphs, *J. Amer. Math. Soc.*, (28) 2015, 669–709.

[6] F.A. Behrend, On sets of integers which contain no three terms in arithmetical progression, *Proc. Nat. Acad. Sci. U.S.A.*, (32) 1946, 331–332.

[7] E.R. Berlekamp, A construction for partitions which avoid long arithmetic progressions, *Canad. Math. Bull.*, (11) 1968, 409–414.

[8] T. Bloom, A quantitative improvement for Roth's theorem on arithmetic progressions, *Arxiv:1405.5800*, submitted.

[9] B. Bollobás, Random Graphs, *London: Academic Press*, 1985.

[10] J. Bourgain, On triples in arithmetic progression, *Geom. Funct. Anal.*, (9) 1999, 968–984.

[11] P. Cameron and P. Erdős, On the number of sets of integers with various properties, in Number Theory (R.A. Mollin, ed.), 61–79, Walter de Gruyter, Berlin, 1990.

[12] P. Cameron and P. Erdős, Notes on sum-free and related sets, *Combin. Probab. Comput.*, 8, (1999), 95–107.

[13] D. Conlon and W.T. Gowers, Combinatorial theorems in sparse random sets, *Ann. of Math.*, to appear.

[14] E. Croot and O. Sisask, A new proof of Roth's Theorem on Arithmetic progressions, *Proc. Amer. Math. Soc.*, (137) 2009, 805–809.

[15] M. Elkin, An improved construction of progression-free sets, *Israel J. Math.*, (184) 2011, 93–128.

[16] P. Erdős, Problems and results in combinatorial number theory, *Journees Arithmétiques de Bordeaux (Conf., Univ. Bordeaux, Bordeaux, 1974)*, 295–310. Astérisque, Nos. 24-25.

[17] P. Erdős and A. Hajnal, Research problem 2-5, *J. Combinatorial Theory*, (2) 1967, 104–105.

[18] P. Erdős, D. J. Kleitman and B. L. Rothschild, Asymptotic enumeration of K_n-free graphs, *Colloquio Internazionale sulle Teorie Combinatorie (Rome, 1973), Tomo II* 19–27. Atti dei Convegni Lincei, 17, Accad. Naz. Lincei, Rome.

[19] P. Erdős, M. Simonovits, Cube-supersaturated graphs and related problems, Progress in graph theory (Waterloo, Ont., 1982), pp. 203–218, Academic Press, Toronto, ON, 1984.

[20] P. Erdős and A.H. Stone, On the structure of linear graphs, *Bull. Amer. Math. Soc.*, (52) 1946, 1087–1091.

[21] J. Folkman, Graphs with monochromatic complete subgraphs in every edge coloring, *SIAM J. Appl. Math.*, (18) 1970, 19–24.

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

[23] W.T. Gowers, A new proof of Szemerédi's theorem, *Geom. Func. Anal.*, (11) 2001, 465–588.

[24] B. Green, The Cameron-Erdős conjecture, *Bull. London Math. Soc.*, 36, (2004), 769–778.

[25] B. Green, Essay submitted for the Smith's Prize. Cambridge University, 2001.

[26] B. Green, A Szemerédi-type regularity lemma in abelian groups, with applications, *Geom. Funct. Anal.*, (15) 2005, 340–376.

[27] B. Green, Counting sets with small sumset, and the clique number of random Cayley graphs. *Combinatorica*, (25) 2005, 307–326.

[28] B. Green and T. Tao, The primes contain arbitrarily long arithmetic progressions, *Ann. of Math. (2)*, (167) 2008, 481–547.

[29] B. Green and J. Wolf, A note on Elkin's improvement of Behrend's construction, *Additive number theory*, Springer, New York, 2010, 141–144.

[30] H. Hàn, T. Retter, V. Rödl and M. Schacht, Ramsey-type numbers involving graphs and hypergraphs with large girth, *arXiv*: 1604.05066.

[31] D.R. Heath-Brown, Integer sets containing no arithmetic progressions, *J. London Math. Soc. (2)*, (35) 1987, 385–394.

[32] D. Kleitman and D. Wilson, On the number of graphs which lack small cycles, manuscript, 1996.

[33] Y. Kohayakawa, S. Lee, V. Rödl and W. Samotij, The number of Sidon sets and the maximum size of Sidon sets contained in a sparse random set of integers, *Random Structures and Algorithms*, (46) 2015, 1–25.

[34] R. Morris and D. Saxton, The number of C_{2k}-free graphs, *Advances in Math.*, to appear.

[35] R. Nenadov and A. Steger, A short proof of the Random Ramsey theorem, *Combinatorics, Probability, and Computing*, Volume 25, January 2016, 130–144.

[36] J. Nešetřil and V. Rödl, The Ramsey property for graphs with forbidden complete subgraphs, *J. Combinatorial Theory Ser. B*, (20) 1976, 243–249.

[37] J. Nešetřil and V. Rödl, Van der Waerden theorem for sequences of integers not containing an arithmetic progression of k terms, *Comment. Math. Univ. Carolinae*, (17) 1976, 675–681.

[38] V. Rödl, A. Rucinski, M. Schacht, An exponential-type upper bound for Folkman numbers, *Combinatorica*, to appear.

[39] K.F. Roth, On certain sets of integers, *J. London Math. Soc.*, (28) 1953, 104–109.

[40] T. Sanders, On Roth's theorem on progressions, *Ann. of Math. (2)*, (174) 2011, 619–636.

[41] A. A. Sapozhenko, The Cameron-Erdős conjecture, (Russian) *Dokl. Akad. Nauk.*, (393) 2003, 749–752.

[42] D. Saxton and A. Thomason, Hypergraph containers, *Invent. Math.*, (201) 2015, 925–992.

[43] M. Schacht, Extremal results for random discrete structures, *Ann. of Math. (2)*, (184) 2016, 331–363.

[44] A. Selberg, An elementary proof of the prime-number theorem, *Ann. of Math. (2)*, (50) 1949, 305–313.

[45] J. Spencer, Restricted Ramsey configurations, *J. Combinatorial Theory Ser. A*, (19) 1975, 278–286.

[46] R.A. Rankin, Sets of integers containing not more than a given number of terms in arithmetical progression, *Proc. Roy. Soc. Edinburgh Sect. A*, (65) 1960/1961, 332–344.

[47] E. Szemerédi, Integer sets containing no arithmetic progressions, *Acta Math. Hungar.*, (56) 1990, 155–158.

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

[49] P. Varnavides, On certain sets of positive density, *J. London Math. Soc.*, (34) 1959, 358–360.

[50] B.L. van der Waerden, Beweis einer Baudetschen Vermutung, *Nieuw Archief voor Wiskunde*, (15) 1927, 212–216.

4 Appendix

The idea behind the container method is to have a small certificate set to describe a large independent set. We give two further applications using this idea.

4.1 A variation of van der Waerden's theorem

Van der Waerden's theorem [50], a classical result in Ramsey theory, says that the set of integers is rich in arithmetic progressions: for any positive integers k and r, there exists $n_0 > 0$ such that every r-coloring of $[n]$ with $n > n_0$ yields a monochromatic k-term arithmetic progression. Denote by $W(k;r)$ the r-colored van der Waerden number, i.e., the minimum integer n such that every r-coloring of $[n]$ contains a monochromatic k-AP. The best known upper bound on $W(k;r)$ is due to Gowers [23]: $W(k;r) \leq 2^{2^{r^{2^{2^{k+9}}}}}$. For two colors, the best lower bound is due to Berlekamp [7]: $W(p+1;2) \geq p2^p$, where p is a prime number.

By setting $\delta = 1/r$ in Theorem 1.7, we immediately obtain the following extension of results of Spencer [45], Nešetřil and Rödl [37] in primes on restricted van der Waerden's theorem.

Corollary 4.1. *For any $r \geq 2$ and $k \geq 3$, there exists a set of primes $S \subseteq \mathbf{P}_{\leq n}$ such that S is $(k+1)$-AP-free and any r-coloring of S yields a monochromatic k-AP.*

This type of question was first raised by Erdős and Hajnal [17]: They asked whether there exists a K_{k+1}-free graph such that every r-edge-coloring of it induces a monochromatic K_k. This was answered in the affirmative by Folkman [21] for $r = 2$, and by Nešetřil and Rödl [36] for arbitrary r, see also Rödl-Ruciński-Schacht [38] for recent developments.

We will use the hypergraph container method to give an alternative proof of Corollary 4.1 without invoking Theorem 1.7. This proof draws on ideas from Nenadov-Steger [35] and Rödl-Ruciński-Schacht [38].

If we work in the set of all integers instead of just the primes, we are able to get the following quantitative result, Proposition 4.2.[6] A set $S \subseteq [n]$ is $(k;r)$-*Folkman* if S is $(k+1)$-AP-free and every r-coloring of S contains a monochromatic k-AP. Define

$$g(k;r) := \min\{n : \exists S \subseteq [n],\ S \text{ is } (k;r)\text{-Folkman}\}.$$

Clearly $g(k;r) \geq W(k;r)$. To see this, simply notice that if $n = W(k;r) - 1$, then there exists an r-coloring of $[n]$ with no monochromatic k-AP and any $(k+1)$-AP-free subset of it inherits this property, implying that $g(k;r) > n$.

Proposition 4.2. *For any $r \geq 2$ and $k \geq 40$,*

$$g(k;r) \leq k^{4k^3} W(k;r)^{5k^2}.$$

The proof of Proposition 4.2 follows along the same line as Corollary 4.1. We omit its proof.

For the proof of Corollary 4.1, we need the following supersaturation lemma. Given a coloring ϕ, denote by $\phi^{(i)} := \phi^{-1}(i)$ the i-th color class. We write $W := W(k;r)$.

Lemma 4.3. *Given any coloring $\phi : \mathbf{P}_{\leq n} \to [r+1]$ with n sufficiently large, if $\sum_{i \leq r} \Gamma_k(\phi^{(i)}) \leq n^2/(\log n)^{W+1}$, then $|\phi^{(r+1)}| \geq n/(\log n)^{W+1}$.*

Proof. Fix an arbitrary $(r+1)$-coloring ϕ of $\mathbf{P}_{\leq n}$ such that $\sum_{i \leq r} \Gamma_k(\phi^{(i)}) \leq n^2/(\log n)^{W+1}$. Recall that the number of W-term arithmetic progressions in $[n]$ is at least $c_W n^2/(\log n)^W$ for some constant $c_W > 0$. Let $x \cdot c_W n^2/(\log n)^W$ be the number of W-APs colored completely by one of the first r colors. Then by the definition of $W(k;r)$, each of these W-APs induces a monochromatic k-AP in the first r colors. We claim that every k-AP is contained in at most W^2 many W-APs. Indeed, given any k-AP $\{a_1, \ldots, a_k\}$, a W-AP $\{b_1, \ldots, b_W\}$ containing it will be uniquely determined once we fix $a_1 = b_i$, $a_2 = b_j$ for $1 \leq i \neq j \leq W$. The number of

[6] We remark that a very recent paper [30] achieves essentially the same quantitative bound, but with a stronger "large girth" property.

choices of the two indices i and j is at most W^2. Therefore, the number of monochromatic k-APs in the first r colors is at least $(xc_W n^2/(\log n)^W)/W^2$. On the other hand, this number is at most $n^2/(\log n)^{W+1}$, thus for sufficiently large n we have

$$\frac{xc_W n^2}{(\log n)^W W^2} \leq \frac{n^2}{(\log n)^{W+1}} \quad \Rightarrow \quad x \leq \frac{W^2}{c_W \log n} \leq \frac{1}{2}.$$

So the number of W-APs containing at least one element from $\phi^{(r+1)}$ is at least $c_W n^2/2(\log n)^W$. Note that each element in $[n]$ is in at most $W \cdot \frac{n}{W-1} < 2n$ many W-APs. Indeed, there are at most W choices to decide which term in a W-AP an element in $[n]$ will be and at most $\frac{n}{W-1}$ many choices to choose the common difference. Therefore, $|\phi^{(r+1)}| \geq (c_W n^2/2(\log n)^W)/2n \geq n/(\log n)^{W+1}$, as desired. □

Proof of Corollary 4.1. We set the parameters as follows:

$$p = \frac{1}{n^{1/k}(\log n)^W}, \qquad \varepsilon = \frac{1}{(\log n)^{2W}}, \qquad \tau = \frac{1}{n^{1/k}(\log n)^{3W}}. \tag{16}$$

For any $k \geq 3$, and p, ε, τ defined in (16), and sufficiently large n, we have that

$$1000rk^{3k} \log \frac{1}{\tau} \log \frac{1}{\varepsilon} = 1000rk^{3k} \cdot \left(\frac{\log n}{k} + 3W \log\log n \right) \cdot 2W \log\log n \leq \log^2 n.$$

Thus, we have

$$\frac{np}{10(\log n)^{W+1}} = \frac{n^{1-1/k}}{10(\log n)^{2W+1}} \geq \frac{n^{1-1/k}}{(\log n)^{3W-2}} = n\tau \cdot \log^2 n \geq 1000rk^{3k} n\tau \log \frac{1}{\tau} \log \frac{1}{\varepsilon}. \tag{17}$$

We also need the following inequality.

$$\frac{np}{10(\log n)^{W+1}} = \frac{n^{1-1/k}}{10(\log n)^{2W+1}} \geq \frac{n^{1-1/k}}{(\log n)^{k+(k+1)W}} = p^{k+1} \frac{n^2}{(\log n)^k}. \tag{18}$$

Let $S \subseteq \mathbf{P}_{\leq n}$ be a random subset of $\mathbf{P}_{\leq n}$, in which every element is chosen with probability p as defined in (16), independently of each other. Denote by B_1 the event that every r-coloring of S contains a monochromatic k-AP and by B_2 the event that S is $(k+1)$-AP-free. We will show that $\mathbb{P}[B_1] + \mathbb{P}[B_2] > 1$, which then implies that with positive probability there is a choice of $S \subseteq \mathbf{P}_{\leq n}$ with the desired properties.

To estimate $\mathbb{P}[B_2]$, we apply the FKG inequality (see e.g. [9]). Note that the number of $(k+1)$-APs in $\mathbf{P}_{\leq n}$ is at most $Cn^2/(\log n)^{k+1}$ for some constant $C > 0$ depending only on k.

$$\mathbb{P}[B_2] \geq \left(1 - p^{k+1}\right)^{Cn^2/(\log n)^{k+1}} > \exp\left\{ -p^{k+1} \cdot \frac{n^2}{(\log n)^k} \right\}. \tag{19}$$

Let ε, τ be as defined in (16). We claim that ε, τ satisfy (5). It follows immediately from the definition of τ that $\tau < \frac{1}{200k^{2k}}$. For the other inequality,

$$\varepsilon n \tau^{k-1} \;=\; \frac{n}{(\log n)^{2W}} \cdot \frac{1}{n^{\frac{k-1}{k}} (\log n)^{3(k-1)W}} = \frac{n^{1/k}}{(\log n)^{(3k-1)W}} \geq k^{3k}.$$

Thus we can apply Corollary 2.2 with ε, τ to obtain $\mathcal{C}$, the set of containers.

We now bound $\mathbb{P}[\overline{B_1}]$ from above. Note that $\overline{B_1}$ implies that there is an r-coloring of S with no monochromatic k-AP. The idea is that if such a coloring exists, then necessarily there is a fairly dense set disjoint from the random set S, which is highly unlikely. Fix one such coloring $\sigma : S \to [r]$ with color classes $X_1, \ldots, X_r$. Since every X_i is k-AP-free, $X_i \subseteq F_i$ for some container $F_i \in \mathcal{C}$. Define $T = \mathbf{P}_{\leq n} \setminus \bigcup_i F_i$, so $S \cap T = \emptyset$. Notice that T is independent of the initial coloring σ, and the number of choices for T is at most $|\mathcal{C}|^r$.

We claim that the set T obtained above is large: $|T| \geq n/(\log n)^{W-1}$. To see this, define an auxiliary $(r+1)$-coloring $\phi : \mathbf{P}_{\leq n} \to [r+1]$ as follows: $\forall x \in T$, $\phi(x) = r+1$, and $\forall x \notin T$, $\phi(x) = \min\{i : x \in F_i\}$. By Corollary 2.2 (ii), the number of monochromatic k-APs in the first r colors of ϕ is at most $r \cdot \varepsilon n^2 = rn^2/(\log n)^{2W} \leq n^2/(\log n)^{W+1}$. Then by Lemma 4.3, $|T| \geq n/(\log n)^{W-1}$ as desired.

Applying the union bound over all possible choices of T, we obtain that

$$\begin{aligned}
\mathbb{P}[\overline{B_1}] &\leq \mathbb{P}[\cup_T S \cap T = \emptyset] \leq |\mathcal{C}|^r \cdot (1-p)^{n/(\log n)^{W+1}} \\
&\leq \exp\left\{ r \cdot 1000 k^{3k} n\tau \log\frac{1}{\tau} \log\frac{1}{\varepsilon} - \frac{np}{(\log n)^{W+1}} \right\} \\
&\overset{(17)}{\leq} \exp\left\{ -\frac{np}{2(\log n)^{W+1}} \right\} \overset{(18)}{\leq} \exp\left\{ -p^{k+1}\frac{n^2}{(\log n)^k} \right\} < \mathbb{P}[B_2].
\end{aligned}$$

Thus we have $\mathbb{P}[B_1] + \mathbb{P}[B_2] = 1 - \mathbb{P}[\overline{B_1}] + \mathbb{P}[B_2] > 1$ as desired. □

4.2 d-fold sumset

Denote by dA the d-fold sumset of A: $dA = A + \ldots + A$, where $A \subseteq \mathbb{Z}_p$. How large does a set in $\mathbb{Z}_p$ have to be so that it is a d-fold sumset? Define $f_d(p)$ to be the maximum integer ℓ such that for any set F of size ℓ, $\mathbb{Z}_p - F = dA$ for some $A \subseteq \mathbb{Z}_p$. Green and Gowers [25], using the discrete Fourier method, showed that $\Omega(\log p) = f_2(p) = O(p^{2/3}\log^{1/3} p)$. It was later improved by Alon [1] using eigenvalues of the Cayley sum-graphs:

$$\Omega\left(\frac{\sqrt{p}}{\sqrt{\log p}}\right) = f_2(p) = O\left(\frac{p^{2/3}}{\log^{1/3} p}\right).$$

It remains a difficult open question to close the gap above. In this section, we investigate this function for the d-fold sumset for every $d \geq 2$. Green and Gowers' proof in fact works for all $d \geq 2$, giving an upper bound[7]

$$f_d(p) = O(p^{2/3}\log^{1/3} p).$$

Our next result gives an improvement when $d \geq 3$.

[7]Perhaps a more involved argument using the Fourier technique can give improvement on this bound obtained directly from their argument.

Theorem 4.4. *For any $d \geq 2$,*

$$f_d(p) = O(p^{\frac{d}{2d-1}+o(1)}).$$

Here, to prove Theorem 4.4, we use Proposition 4.6 instead to find such a small certificate. Though our bound for $d = 2$ is weaker than the previous bounds by a polylog factor, it easily works for any $d \geq 2$. We think that the bound above is far from best possible. We conjecture that $f_d(p) = p^{c_d+o(1)}$, where $c_d \to 0$ as $d \to \infty$.

First, we need to define an auxiliary hypergraph, from which the upper bound on $f_d(p)$ can be derived. Given a set $T \subseteq \mathbb{Z}_p$, we define the d-uniform Cayley sum-hypergraph $G(\mathbb{Z}_p, T)$ generated by T as follows: $V(G(\mathbb{Z}_p, T)) = \mathbb{Z}_p$ and its edge set consists of all d-tuples $\{x_1, \ldots, x_d\}$ such that $x_1 + \ldots + x_d = t$ for some $t \in T$. The following claim, due to Green [25] and Alon [1], gives a way to obtain upper bound, we repeat here their short proof.

Claim 4.5. *If there exists a set T of size t such that $t > 2\alpha(G(\mathbb{Z}_p, T))$, then $f_d(p) \leq 2t$.*

Proof. If we have a set $F \subseteq \mathbb{Z}_p$ such that $\mathbb{Z}_p - F$ is not a sumset, then $f_d(p) \leq |F|$. We will find such a set $F = T \cup T'$ in two steps with $|T| = |T'| = t$, where T is the set guaranteed by the hypothesis, i.e., $t > 2\alpha(G(\mathbb{Z}_p, T))$. Notice that if any set $S \subseteq \mathbb{Z}_p - T$ is a d-fold sumset for some set $A \subseteq \mathbb{Z}_p$, then A has to be an independent set in $G(\mathbb{Z}_p, T)$. Note also that the number of d-fold sumsets S cannot be larger than the number of sets A that we generate $S = dA$ from. We then choose another set T' of size t, there are $\binom{p-t}{t}$ many choices. Suppose that for each of these choices, $S = \mathbb{Z}_p - T - T'$ is a d-fold sumset dA, then from the observation above we have that the number of choices for A is at least

$$\binom{p-t}{t} > \binom{p}{\alpha(G(\mathbb{Z}_p, T))}.$$

This is impossible since A is an independent set in $G(\mathbb{Z}_p, T)$. Thus there exists a T' such that $F = T \cup T'$ is the desired set. □

We use the following slight variation of Proposition 19 in [27].

Proposition 4.6. *For any $d \geq 2$, there exists $c := c(d)$ such that the following holds. For any set $S \subseteq \mathbb{Z}_p$ of size m with m sufficiently large, there exists a set $R \subseteq S$ with $|R| \leq m^{1/d}/c$ and $|\widehat{dR}| \geq cm$, where $\widehat{dR} := \{x : x = a_1 + \ldots + a_d \text{ with distinct } a_i \in R\}$.*

Proof of Theorem 4.4. Let T be a random t-set of $\mathbb{Z}_p$ for $t = \frac{2}{c^2}(p \log p)^{d/(2d-1)}$. Let $G := G(\mathbb{Z}_p, T)$. We will show that with high probability $\alpha(G) < t/2$, then the bound on $f_d(p)$ follows from Claim 4.5.

For every set $S \subseteq G$ of size $m = t/2$, by Proposition 4.6, S contains a set R of size at most $m^{1/d}/c$ and $|\widehat{dR}| \geq cm$. Clearly, $\widehat{dR} \subseteq dS$. It then follows that if every R of size at most $m^{1/d}/c$ with $|\widehat{dR}| \geq cm$ is not an independent set, then $\alpha(G) < m$. Fix such a choice of R, then R is an independent set only when $\widehat{dR} \cap T = \emptyset$. Thus the probability that R is

independent is at most $(1-\frac{cm}{p})^t \le e^{-cmt/p}$. Applying the union bound, we obtain that the probability that there exists an R that is independent is at most

$$\sum_{i=1}^{m^{1/d}/c} \binom{p}{i} \cdot e^{-cmt/p} \le \exp\left\{\frac{1}{c}t^{1/d}\log p - \frac{ct^2}{2p}\right\} = o(1),$$

where the last equality follows from $t = \frac{2}{c^2}(p\log p)^{d/(2d-1)}$. Thus $\alpha(G) < t/2$ as desired. □

Here is a related conjecture of Alon [1].

Conjecture 4.7. *There exist constants c_1, c_2 so that the following holds. Let Γ be an abelian group of odd order n. Then for every $1 \le t \le n$, there is a subset $T \subseteq \Gamma$ of size t so that for $G := G(\Gamma, T)$,*

$$\alpha(G) \le c_1 \frac{n}{t}(\log n)^{c_2}.$$

The conjecture was known to be true for $t = \Omega(n)$. Alon [1] gave a bound $\alpha(G) \le \frac{n}{t^{1/2}}\log n$. Using the idea in the proof of Theorem 4.4, we can establish the following bound:

$$\alpha(G) \le \left(\frac{n}{t}\right)^2 (\log n)^2,$$

which implies Conjecture 4.7 for $t \ge n/(\log n)^{O(1)}$.