Counting X-free sets

by

Ashna Keaton Wright
B.Sc., University of Victoria, 2022

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

We acknowledge with respect the Lək̓ʷəŋən Peoples (Songhees and Esquimalt) on whose traditional territory the university stands, and the Lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Counting X-free sets

by

Ashna Keaton Wright
B.Sc., University of Victoria, 2022

Supervisory Committee

Dr. Natasha Morrison, Co-Supervisor
(Department of Mathematics and Statistics)

Dr. Jonathan A. Noel, Co-Supervisor
(Department of Mathematics and Statistics)

Dr. Anthony Quas, Departmental Member
(Department of Mathematics and Statistics)

ABSTRACT

Let X be a finite subset of $\mathbb{Z}^d$. A set $A \subseteq [n]^d$ is *X-free* if it does not contain a *copy of* X, that is subset of the form $\boldsymbol{b} + r \cdot X$ for any $r > 0$ and $\boldsymbol{b} \in \mathbb{R}^d$. Let $r_X(n)$ denote the cardinality of the largest X-free subset of $[n]^d$. In this thesis we explore X-free sets in three ways. Firstly, we give an exposition of a standard multidimensional extension of Behrend's [7] construction that gives a lower bound on $r_X(n)$ for all $|X| \geq 3$. Next, we lower bound the number of copies of X guaranteed in subsets with cardinality larger than $r_X(n)$. This is a *supersatuation* result that uses the previously demonstrated lower bound on $r_X(n)$. Finally, using our supersaturation result, we show that for infinitely many values of n the number of X-free subsets is $2^{O(r_X(n))}$. This result is obtained using the powerful hypergraph container method. Further, it generalizes previous work of Balogh, Liu, and Sharifzadeh [3] and Kim [41].

This thesis includes joint work with Natalie Behague, Joseph Hyde, Natasha Morrison, and Jonathan A. Noel.

Table of Contents

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my wonderful supervisors, Jon and Natasha, without whom the last two years would not be possible. Jon, thank you for encouraging me to pursue graduate studies in the first place and for having an open door policy (even for silly questions). Natasha, thank you for your always offering words of wisdom, encouragement, and Taylor Swift. Having the privilege to learn from and work with both of you is such an honour.

Beyond my supervisors, I want to thank all the other amazing professors at UVic who shaped and supported me during both my undergraduate degree and Master's. In particular, a big thank you to Kieka Mynhardt, Gary MacGillivray, Gourab Ray, Peter Dukes, Jane Butterfield, and, of course, Chris Eagle. I also owe so much to the community of graduate students and postgraduates in the Math and Stats Department. Thank you all helping me have fun and sticking around even for the not-so-fun parts. Without you all, this certainly would not have been possible.

Thank you to my collaborators, Joseph Hyde and Natalie Behague. I am so grateful for all I learned from both of you.

I wish to thank my entire family for their endless love and support even from far away. Lastly, to Jacob – for everything, always.

Chapter 1
Introduction

Let $X \subseteq \mathbb{Z}^d$, such that $|X|$ is finite and $|X| \geq 3$. We will call X' a *copy of* X if there exists $\boldsymbol{b} \in \mathbb{R}^d$ and $q \geq 0$ such that

$$X' = \boldsymbol{b} + q \cdot X := \{\boldsymbol{b} + q\boldsymbol{x} : \boldsymbol{x} \in X\}.$$

Moreover, a copy if *non-trivial* if $q > 0$. We say that $A \subseteq [n]^d$ is X*-free* is it does not contain any non-trivial copies of X. This thesis considers the following three questions:

1. How large can an X-free subset of $[n]^d$ be?
2. How many copies of X must be present in a subset of $[n]^d$ larger than the extremal threshold?
3. How many X-free subsets of $[n]^d$ are there?

The general forms of these three questions are central to extremal combinatorics and have been studied in a myriad of settings with varying forbidden substructures. That is, one may consider: maximizing or minimizing the size of a combinatorial object without some forbidden substructure, the number of copies of a forbidden substructure guaranteed above the "extremal threshold", and the number of combinatorial objects without some forbidden substructure.

This introduction will serve as a motivation for our questions, with a brief history for each in both the general setting and for X-free sets. A more in-depth history in our specific setting is placed at the beginning of the appropriate chapter to place us in the proper context.

1.1 Large X-free subsets

It is typical in extremal combinatorics to ask for the maximum or minimum size of a combinatorial object without some forbidden substructure. For example, consider the classical result of Mantel [47] which states that the maximum number of edges in a triangle-free graph on n vertices is $\left\lfloor \frac{n^2}{4} \right\rfloor$ edges. This result, often referred to as Mantel's Theorem [47], also states there is exactly one tight construction: the complete bipartite graph $K_{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$. As it is simple to observe that this graph is triangle-free and is dense, it is a fairly natural triangle-free graph to consider.

In Chapter 2 we will explore constructing X-free subsets of $[n]^d$. Let $r_X(n)$ denote the size of the largest X-free subset. For general sets X, constructing a large X-free set is an intimidating problem as, unlike with Mantel's Theorem [47], there does not appear to be an immediately obvious candidate to consider. We can make some simple observations about $r_X(n)$. If $|X| \in \{1, 2\}$, then $r_X(n)$ is rather trivial. Next, observe that if $X' \subseteq X$, then any X'-free subset is also X-free. The reverse implication however does not hold, as a set may not contain a copy of X but have a copy of X'. Therefore, we have that

$$r_{X'}(n) \leq r_X(n), \text{ for all } X' \subseteq X.$$

Let us assume for a moment that X is a set of k numbers in arithmetic progression, or a *k-AP*. Observe that forbidding k-term arithmetic progressions is equivalent to forbidding copies of $[k]$, as any k-term arithmetic progression is a scaled and translated copy of $[k]$. We will denote $r_{[k]}(n)$ simply by $r_k(n)$. Most natural constructions of sets without k-APs are rather sparse, for instance the set

$$\{2, 4, 8, 16, \dots\} \subseteq [n]$$

has density of only $\frac{\lfloor \log(n) \rfloor}{n}$. Determining the largest $[k]$-free subset has shown to be an extremely difficult problem that remains open after over 100 years of study. It was shown by Szemerédi [79] in 1975 that $r_k(n) = o(n)$, for all finite $k \geq 1$. This result is often referred to as Szemerédi's Theorem. We note that the case $k = 3$ was handled by Roth [62] and, consequently, is referred to as Roth's Theorem.

The ideas in the proof of Szemerédi's Theorem led to Szemerédi's celebrated Regularity Lemma. Szemerédi's Regularity lemma essentially says that sufficiently large graphs may be partitioned into clusters such that the edges between most clusters are "random-like". The power of this lemma comes from the fact that random graphs are quite "well-behaved" and as a result are often much easier to work with. We will explore the history of $r_k(n)$ in much greater detail in Chapter 2.

Returning to general X, can we say something similar to Szemerédi's Theorem, i.e for a sufficiently large n, does any dense subset of $[n]^d$ contain a copy of X? The following theorem of Furstenberg and Katznelson [29] resolves this question.

Theorem 1.1.1 (Multidimensional Szemerédi Theorem [29]; see also [35, 51, 61, 80])**.** *If $d \geq 1$ and X is a finite subset of $\mathbb{N}^d$, then $r_X(n) = o(n^d)$.*

Furstenberg and Katznelson's [29] proof uses ergodic theory and so does not yield quantitative bounds. This sparked interest in the formulation of a combinatorial proof of Theorem 1.1.1. This goal was accomplished independently by Gowers [35], Tao [80] and Rödl, Nagle, Schacht, and Skokan [51, 61] whose results provided bounds, albeit they are rather weak.

The path to obtaining combinatorial proofs of Theorem 1.1.1 was long. It was observed by Ruzsa and Szemerédi [63] that the following consequence of Szemerédi's Regularity Lemma proves Szemerédi's Theorem in the case that $k = 3$.

Lemma 1.1.2 (Triangle Removal Lemma, [64])**.** *For every constant $\varepsilon > 0$, there exists a constant $\delta > 0$ such that if G is a graph with n vertices containing at most δn^3 triangles, then it is possible to make G triangle-free by removing at most εn^2 edges.*

Solymosi [78] also showed that Lemma 1.1.2 implies Theorem 1.1.1 when $X = \{0, e_1, e_2\} \subseteq \mathbb{Z}^2$, where e_1, e_2 are the standard basis vectors of $\mathbb{R}^2$. It was observed by Frankl and Rödl [26] that a generalization of the triangle-removal lemma to the k-uniform hypergraph setting would imply Szemerédi's Theorem in full. It was then observed by Solymosi [78] that modification of this argument of Frankl and Rödl would actually prove Theorem 1.1.1.

The proof method laid out by Frankl and Rödl [26] is essentially what is employed in [35, 51, 61, 80]. To achieve a hypergraph removal lemma, the main ingredient of the proofs in [35, 51, 61, 80] is a hypergraph analogue of Szemerédi's Regularity Lemma for k-uniform hypergraphs. We note that the cases $k = 3$ and $k = 4$ was first handled Ruzsa and Szemerédi [63] and by Frankl and Rödl [26], respectively.

1.2 Supersaturation

As discussed in Section 1.1, in extremal combinatorics we often look to maximize or minimize the size of a structure while avoiding some forbidden substructure. But, what happens above this extremal threshold? Consider again Mantel's Theorem [47]. A natural follow-up question asks: how many triangles must be present in a graph with more than $\left\lfloor \frac{n^2}{4} \right\rfloor$ edges? It is easy to see a graph with $\left\lfloor \frac{n^2}{4} \right\rfloor + x$ edges must contain at least x triangles (consider selecting an edge to delete from a triangle one at a time until the graph is of size $\left\lfloor \frac{n^2}{4} \right\rfloor$). Rademacher was the first to improve on this trivial bound; he showed in an unpublished proof that a graph with just $\left\lfloor \frac{n^2}{4} \right\rfloor + 1$ edges must contain at least $\left\lfloor \frac{n}{2} \right\rfloor$ triangles. Erdős [18, 19] looked at lower bounding the number of triangles in graphs with much larger size than this, attracting much attention to the problem, see [25, 33, 46, 54]. Exact asymptotic values for the minimum number of triangles in supersaturated graphs with size $\Omega(n^2)$ above the extremal threshold were found by Razborov [59] in one of the first applications of his method of flag algebras [58]. Nikiforov [53] and Reiher [60] extended Razborov's result to bounds on the minimum number of copies of K_4 and K_k for $k \geq 5$ respectively using Lagrange multipliers; while obtaining their results, they each showed their method could be used to provide an alternative proof to Razborov's result. An exact solution for the minimum number of triangles in a supersaturated graph was found by Liu, Pikhurko, and Staden [45] in 2020 for large graphs that are "not too dense".

The question of what happens above the extremal threshold has been studied as a variation on many classical problems; these problems are called *supersaturation problems*, or Erdős-Rademacher type problems. Let's briefly look at a few examples in different classical combinatorial results. First, we already explored an example in extremal graph theory for the minimum number of triangles in graphs with more than $\left\lfloor \frac{n^2}{4} \right\rfloor$ edges. Next, in extremal set theory, Erdős, Ko, and Rado proved that if $k \leq \frac{n}{2}$ and a family of subsets of $[n]$ all of size k pairwise intersect, then the family is of cardinality at most $\binom{n-1}{k-1}$. Therefore, if such a family is of cardinality strictly greater than $\binom{n-1}{k-1}$, i.e is supersaturated, there must be at least one pair of sets that are disjoint. Das, Gan, and Sudakov [15] provided lower bounds the number of disjoint pairs in supersaturated families. Finally, we look at an example in group theory. A subset A of a finite abelian group $\mathbb{G}$ is *sum-free* if it does not contain a triple x, y, z such that $x + y = z$. The order of the largest sum-free set was established, after many years of progress, by Green and Ruzsa [37]. Samotij and Sudakov [66] explored how many triples of this form must exist in a subset above the threshold established by Green and Ruzsa for various groups.

In Chapter 3 we will explore supersaturation by lower bounding the number of copies of X in sets of cardinality larger than $r_X(n)$. These results are not only interesting in their own regard, but they can be used as a key ingredient in obtaining enumeration results. We will show the connection between supersaturation results and enumeration results in Chapter 4.

1.3 Enumeration

Instead of considering what happens above an extremal threshold, we may instead consider how many combinatorial objects exist that do not contain some forbidden substructure. Chapter 4 will focus on bounding the number of X-free subsets of $[n]^d$. It is trivial to see that the number of X-free subsets of $[n]^d$ is at least $2^{r_X(n)}$ as any subset of an X-free set must itself be X-free. In 1990, Erdős and Cameron [12] asked whether this bound is approximately correct in the case that X is a k-term arithmetic progression.

Question 1.3.1 (Cameron and Erdős [12, Section 4.2]). *Is the number of subsets of $[n]$ not containing a k-term arithmetic progressions equal to $2^{(1+o(1))r_k(n)}$?*

Question 1.3.1 is still open; there are historical examples that may lend one to believe the answer is "yes" or "no". In the literature, there are many examples which assert that, up to a $(1+o(1))$ factor in the exponent, the number of combinatorial structures satisfying a certain set of constraints is equal to the number of subsets of the largest such structure. For example, the results of Erdős, Kleitman, and Rothschild [21] on triangle-free graphs, Kleitman [42] on antichains, Dong, Mani, and Zhao [16] on t-error correcting codes, and Balogh, Das, Delcourt, Liu, and Sharifzadeh [1] on intersecting families (see also the work of Balogh, Garcia, Li, and Wagner [2]).

Conversely, similar natural questions have been asked for other combinatorial objects that have in turn been answered with "no". That is, there are historical examples in which the number of combinatorial objects with no forbidden substructure is asymptotically greater than the number of subsets of the largest substructure. For example, the results of Morris and Saxton [49] on the number of C_6-free graphs, and Saxton and Thomason [73] on the number of Sidon sets.

In Chapter 4, we will prove the following result which offers some progress towards answering the more general form of Question 1.3.1, where we instead consider any finite set $X \subseteq \mathbb{Z}^d$ such that $|X| \geq 3$.

Theorem 1.3.2. *Let d be a positive integer and let $X \subseteq \mathbb{N}^d$ be a finite set such that $|X| \geq 3$. For infinitely many $n \in \mathbb{N}$, the number of X-free subsets of $[n]^d$ is $2^{O(r_X(n))}$.*

Theorem 1.3.2 was previously handled in the case that X is a k-AP by Balogh, Liu, and Sharifzadeh [3] and in the case that X is a corner by Kim [41]. *A corner* is a set of the form

$$\{\mathbf{0}, \boldsymbol{e_1}, \boldsymbol{e_2} \ldots, \boldsymbol{e_d}\},$$

where e_i represents the i-th standard basis vector for $\mathbb{R}^d$. It may be surmised that Kim [41] focused on the case where X is a corner due to the importance of corners in proving Theorem 1.1.1. It may be shown that Theorem 1.1.1 in the special case that X is a corner implies the theorem in full. However, there does not appear to be an obvious reduction of Theorem 1.3.2 to the result of Kim [41] due to the fact that the function $r_X(n)$ for a generic set X may have a very different growth rate when compared with the maximum size of a corner-free set.

1.3.1 The hypergraph container method

Our main tool in proving Theorem 1.3.2 is the hypergraph container method. This method was developed in 2015 by Balogh, Morris, and Samotij [4] and, independently, Saxton and Thomason [73]. This work was highly influenced by work of Kleitman and Winston [43]. After the work of Kleitman and Winston [43] and before the development of the hypergraph container method, other results using similar methods appeared. Notably, Sapozhenko [68–72] first began developing a generalized container method for graphs and was the first to use the term "containers". Hypergraph container-like proof styles were also used by Conlon and Gowers [13], Friedgut, Rödl, and Schacht [27], and Schacht [74] in their work on Ramsey-type problems in sparse random settings before the development of the hypergraph container method. We give an overview of how the method works in Chapter 4.

The hypergraph container method is very powerful and has been used in a wide variety of settings in combinatorics to achieve enumeration results. For instance, the results of Morris and Saxton [49] on the number of C_6-free graphs and, Saxton and Thomason [73] on the number of Sidon sets were proved used the hypergraph container method directly. Further, the results of Dong, Mani and Zhao [16] on the number of t-error correcting codes [1, 16] and Balogh, Das, Delcourt, Liu, and Sharifzadeh [1] on intersecting families used variations on the hypergraph container method. The hypergraph container method has also been used to reprove several classical results, often offering a more simple proof. Relevantly, the aforementioned enumeration results of Erdős, Kleitman, and Rothschild [21] on triangle-free graphs and Kleitman [42] on antichains have be re-proved using the hypergraph container method [4, 73].

1.4 Overview of thesis

This thesis explores three separate, though deeply related, questions. These three ideas build on one another, with the results of each chapter using the results of the previous chapters.

In Chapter 2, we will look at constructing large X-free sets. More specifically, we use standard methods to generalize a lower bound of Behrend [7] for sets without 3-APs to obtain a lower bound on the size of X-free subsets.

In Chapter 3, we will prove a supersaturation result for sets of size $C \cdot r_X(n)$, where C is a large, positive constant independent of n and n is an element of an infinite subset of $\mathbb{N}$. As we do not know exact asymptotics of $r_X(n)$, we will employ our lower bound from Chapter 2 to give us some sense of $r_X(n)$.

Finally, in Chapter 4 we will show that for infinitely many values of n, the number of X-free subsets of $[n]^d$ is $2^{O(r_X(n))}$. We will achieve this bound through an application of the seminal hypergraph container method [4, 73]. As is typical with applications of the hypergraph container method, we will use the supersaturation result achieved in Chapter 3.

The results of Chapter 2 and Section 4.1 are expository. The results of Chapters 3 and Section 4.2 are joint work with Natalie Behague, Joseph Hyde, Jonathan A. Noel, and Natasha Morrison; see [6].

Chapter 2
Behrend Type Constructions in Every Dimension

2.1 Introduction

The study of arithmetic progressions in subsets of $[n]$ is central to additive combinatorics. The type of questions one can ask about arithmetic progressions is vast. For example: what is the largest subset of $[n]$ not containing an arithmetic progression of length k? Recall we denote this quantity with $r_k(n)$. A deeply related question asks: does there always exist a number $w(r, k)$ such that for all r colourings of $[w(r, k)]$ there exists a monochromatic k-term arithmetic progression? The question was answered in the affirmative by van der Waerden [81] in 1927. Finding exact values of $w(r, k)$, also called van der Waerden numbers, has turned out to be hard; there are only 7 known values of $w(r, k)$ to date. The best upper bound on $w(r, k)$ comes from Gowers [34] and states

$$w(r, k) \leq 2^{2^{k^{2^{2^{r+9}}}}}.$$

In 1936, Erdős and Turán [24], working towards a conjecture of Szekeres related to van der Waerden numbers, proved the following upper bound on $r_3(n)$: for sufficiently large n and any $\varepsilon > 0$, $r_3(n) < (\frac{4}{9} + \varepsilon)n$. They conjectured any sufficiently dense subset of $[n]$ contains arbitrarily long arithmetic progressions; that is they conjectured for all k, $r_k(n) = o(n)$. Further, they conjectured that for all k there exists a constant $c_k > 0$ such that $r_k(n) < n^{1-c_k}$.

In 1942, Salem and Spencer [65] showed that $r_3(n) > n^{1-\frac{c}{\log \log n}}$, disproving the second conjecture of Erdős and Turán. Modifying their method, Behrend [7] found an even larger sub-linear lower bound on $r_3(n)$. In 1955, Roth [62] proved that $r_3(n) = o(n)$ using analytic methods, a result now called Roth's Theorem. In 1967, Szemerédi extended Roth's Theorem to show $r_4(n) = o(n)$ and then in 1975 [79] proved $r_k(n) = o(n)$ for all k. This result is typically referred to as Szemerédi's Theorem. This theorem is much stronger than van der Waerden's Theorem, as it allows us to ignore the underlying colouring that is partitioning $[n]$ completely. Szemerédi's Theorem was proved using purely combinatorial arguments and led to the discovery of Szemerédi's celebrated Regularity Lemma. A few years later, Furstenberg [28] found an alternative proof of Szemerédi's Theorem using ergodic theory. While this approach relies on the Axiom of Choice and, therefore, does not yield effective bounds, it has the benefit of being more amenable to generalization. For example, the first proof of the Density Hales–Jewett Theorem, a multidimensional generalization of Szemerédi's Theorem, was obtained via an extension of Furstenberg's ideas [30]. Later, another proof of Szemerédi's Theorem was given by Gowers [34] using Fourier analysis.

With Szemerédi's Theorem in hand, it is natural to wonder about the exact asymptotics of $r_k(n)$. This problem has been shown to be extraordinarily difficult, even in the $k = 3$ case. As previously mentioned, in 1946 Behrend [7] presented a construction of large sets that do not contain 3-term arithmetic progressions, giving the following lower bound on $r_3(n)$:

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

We say that a set $A \subseteq \mathbb{R}^d$ is *convexly independent* if no element in A may be written as a convex combination of other elements in A. Behrend's proof utilizes that a sphere in any dimension is convexly independent and so cannot contain three points all on the same line; this proof will be the focus of Section 2.2 of this chapter. In 2011, this bound was finally improved upon by a factor of $\Theta(\sqrt{\log n})$ by Elkin [17]. In a similar nature to Behrend's construction, Elkin's construction utilizes a subset of a thin annulus that is convexly independent (Elkin considers a subset of a thin annulus as an annulus itself may not be convexly independent). Shortly after Elkin's paper, Green and Wolf [38] found a simpler, less constructive proof of the same bound using similar ideas to Elkin. In 2024, Hunter [39] found the first quasipolynomial improvement of Behrend's construction, proving that

$$r_3(n) \gg n \cdot 2^{-(c+o(1))2\sqrt{2\log_2(n)}},$$

where $c = \sqrt{\log_2\left(\sqrt{\frac{32}{9}}\right)}$. We note that Hunter's proof does not optimize over c, so a more careful proof may obtain a smaller c. There have been many improvements on the first upper bound of Roth [62] throughout the years, see [8–10, 40, 67, 75]. Currently, the best known upper bound is

$$r_3(n) \leq n \cdot e^{-O((\log n)^{\frac{1}{11}})},$$

for some absolute constant $c > 0$. This upper bound was proved by Bloom and Sisask [11] by making minor simplifications to a result of Kelley and Meka [40].

The case for $k > 3$ has also been studied extensively. Notably, Rankin [57] in 1960 obtained a lower bound on $r_k(n)$ by generalizing Behrend's construction in a natural way. Currently, the best known bounds on $r_k(n)$ are:

$$\frac{Cn \sqrt[2a]{\log n}}{2^{a2^{(a-1)/2} \sqrt[a]{\log n}}} \leq r_k(n) \leq \frac{n}{(\log \log n)^{C'}},$$

where $a = \lceil \log k \rceil$ and C, C' are positive constants depending only on k. The lower bound was obtained by O'Bryant [55] in 2011 by incorporating into Rankin's [57] method the generalizations made to Behrend's construction made by Elkin [17], and Green and Wolf [38]. The upper bound is due to Gowers [34] and uses Fourier analysis.

Although much of the study of k-AP-free-sets focuses on the asymptotics of $r_k(n)$, there is some work done on exact numerical values of $r_k(n)$ for small values of n. For the $k = 3$ case, the exact values for $1 \leq n \leq 21$ were found by Erdős and Turán [24], with corrections given by Mąkowski [48] and Moser [50]. In 1967, Wagstaff [83] computed exact values for $1 \leq n \leq 52$. In 2010, Garsarch, Glenn, and Kruskal [32] found exact values of $r_3(n)$ for $1 \leq n \leq 187$. The values up to $n = 150$ were previously found by Wróblewski [85] and posted to his website. For $4 \leq k \leq 8$, exact values up to around 50 were calculated by Wagstaff [83, 84]. In 2011, further values of $r_k(n)$ for $4 \leq k \leq 8$ and up to around $n = 110$ were shown by Shao, Deng, Liang, and Xu [77], with some additional values of n found for specific k. Excluding the small values Erdős and Turán calculated by hand, exact values are found with computer assistance. Exact values of $r_k(n)$ are important as they have applications outside of pure mathematics such as in the multiplication of large matrices; a short survey of applications in the $k = 3$ case can be found in [32].

2.2 Classical Behrend Construction

In this section, we give an exposition of Behrend's [7] construction. We will present a proof of the following lemma, which relaxes the conditions of Behrend's original construction slightly. Consequently, we will give a construction of an X-free set, where X is any finite subset of $\mathbb{N}$ such that $|X| \geq 3$. This will give us a lower bound for $r_X(n)$ analogous to Behrend's original lower bound for $r_3(n)$.

Lemma 2.2.1. *If $X \subseteq \mathbb{N}$ such that $|X| \geq 3$, then there exists $c_X > 0$ depending only on X such that, for all $n \in \mathbb{N}$,*

$$r_X(n) \geq n \cdot \exp(-c_X \sqrt{\log n}).$$

We note that the statement and proof are fundamentally unchanged from Behrend's construction after this small and standard generalization. The main idea of this construction is to "project" a sphere down to a one-dimensional set in a clever way such that we are able to guarantee that any 3-AP necessarily corresponds to three points on a line. So, as spheres cannot contain three points on the same line, under our projection the sphere will correspond to a 3-AP-free set. The following small example showcases this projection:

Example 2.2.2. We will take the intersection of $[7]^3$ with a sphere in $\mathbb{R}^3$ and project it down to $\mathbb{N}$ to obtain a 3-AP free set. Take the sphere

$$S = \{(x, y, z) : x^2 + y^2 + z^2 = 54\}.$$

Define $A = S \cap [7]^3$, see Figure 2.1.

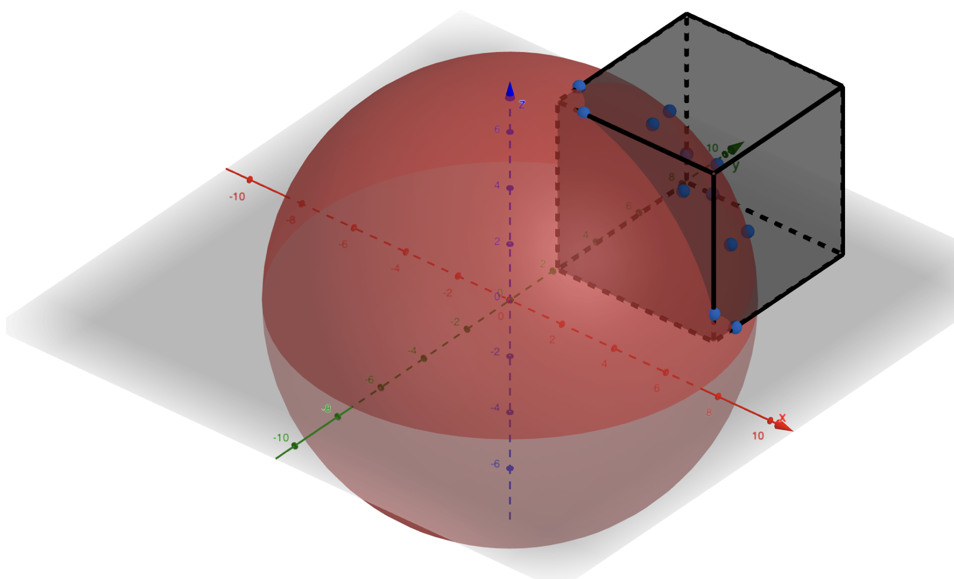

Figure 2.1: A sphere centred at $\mathbf{0}$ with radius $\sqrt{54}$ intersecting with $[7]^3$.

Define $f : \mathbb{R}^3 \to \mathbb{R}$ by $f(\boldsymbol{x}) = x_1 + 14x_2 + 196x_3$. The key property of f is that, because of the large difference between the coefficients of x_1, x_2, x_3, it is injective on the set $\{\boldsymbol{x} - \boldsymbol{y} : x, y \in [7]^3\}$. See Figure 2.2. More details about the function f are given in the proof of Theorem 2.2.1.

We have $f(A) = \{231, 296, 413, 467, 491, 636, 675, 1013, 1052, 1221, 1388, 1401\}$ and it can easily be shown this set is in fact 3-AP free.

Before proceeding with the proof of Lemma 2.2.1, we make the following observation:

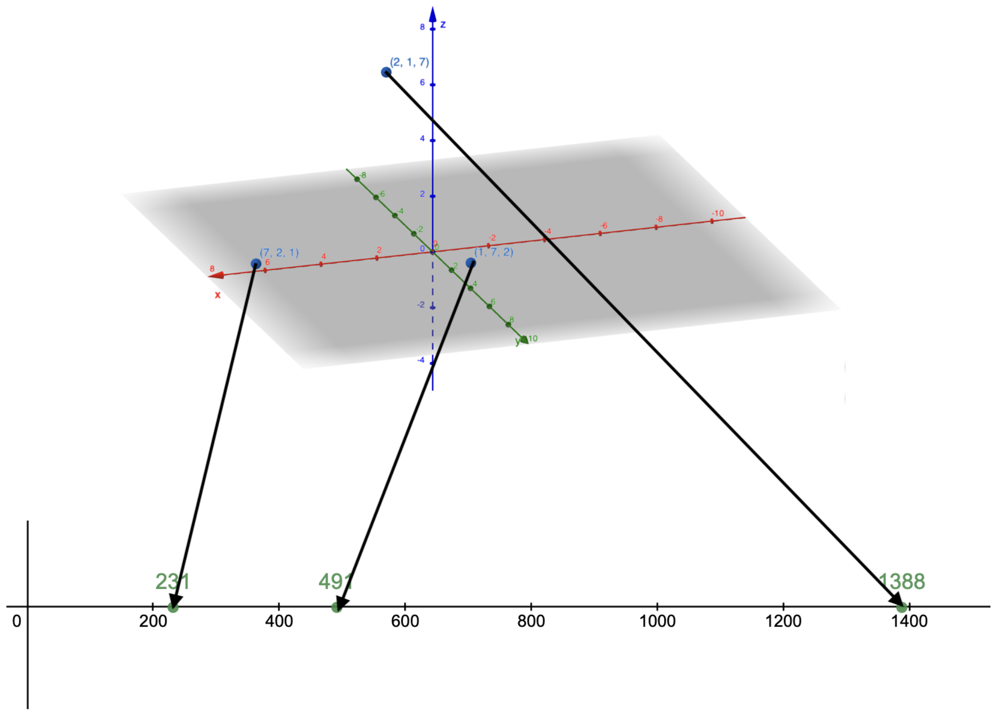

Figure 2.2: Projection of three points of A onto $\mathbb{Z}$ using f.

Observation 2.2.3. *If X' is a copy of X, then S is X-free if and only if it is X'-free.*

Therefore, we may choose to work with a copy of X so that the points in X have convenient coordinates.

We note that Lemma 2.2.1 excludes the cases where $|X| \in \{1, 2\}$ as in these cases, an X-free subset must be rather trivial. If $|X| = 1$, then clearly $r_X(n) = 0$. If $|X| = 2$, then we cannot have any two points on a line parallel to the line between the points in X. Therefore, we again are very constricted and are not able to obtain a lower bound as in Lemma 2.2.1.

Proof of Lemma 2.2.1. As $r_X(n) \geq r_{X'}(n)$ for all $X' \subseteq X$, then we may assume that $|X| = 3$. Let X' be a copy of X of the form $\{0, \alpha, \beta\}$ where $\alpha, \beta \in \mathbb{N}$, $0 < \alpha < \beta$, and α, β are as small as possible. Note that for any $x < y < z$, the set $\{x, y, z\}$ is a copy of X' if and only if $(\beta - \alpha)(y - x) = \alpha(z - y)$. We will build an X'-free set $S \subseteq [n]$, where n is sufficiently large, such that $|S| \geq n \cdot e^{-c_X \sqrt{\log n}}$, where c_X is a constant depending on X. Our result will follow immediately by Observation 2.2.3 and definition of $r_X(n)$.

For $n \geq 1$, define $M := \left\lfloor \frac{n^{1/N}}{(\alpha+\beta)} \right\rfloor$ and $N := \left\lfloor \sqrt{\log n} \right\rfloor$. Consider the points

$$\boldsymbol{x} = (x_1, x_2, \dots, x_N) \in [M]^N,$$

of which there are M^N. For each point $\boldsymbol{x}$, $r^2 := x_1^2 + x_2^2 + \cdots + x_N^2$ is an integer in the range $[N, N\cdot M^2]$. By the pigeonhole principle, there exists a sphere that contains at least $\left\lceil \frac{M^N}{N\cdot M^2-N+1} \right\rceil$ points in $[M]^N$; let S' be the intersection of this sphere with $[M]^N$. Observe, by definition of M and N,

$$|S'| \geq \left\lceil \frac{M^N}{N\cdot M^2 - N + 1} \right\rceil \geq \frac{M^{N-2}}{N} = \frac{\left(\left\lfloor \frac{n^{1/N}}{(\alpha+\beta)} \right\rfloor\right)^N}{N\cdot M^2} \geq \frac{n}{N\cdot M^2\cdot (2(\alpha+\beta))^N}.$$

By definition of M and N, we have for some $c > 0$

$$M^2 \leq \frac{n^{2/N}}{(\alpha+\beta)^2} \leq e^{\frac{2\log n}{\sqrt{\log n}}} \leq e^{c\cdot\sqrt{\log n}}.$$

Trivially, by definition of N,

$$N \leq e^{\sqrt{\log(n)}}.$$

Further, by definition of N, for some $c' > 0$,

$$(2(\alpha+\beta))^N \leq e^{(\sqrt{\log n}+1)(\log(2(\alpha+\beta)))} \leq e^{c'(\sqrt{\log n})}.$$

So, we can find a large constant $c_X > 0$ depending only on X such that

$$|S'| \geq n\cdot e^{-c_X\sqrt{\log n}}.$$

Next, we will define a function f that maps S' injectively to an X-free subset of $[n]$. As we have shown $|S'|$ is as desired, we will then be done. Define $f : \mathbb{Z}^N \to \mathbb{Z}$ such that

$$f(\boldsymbol{x}) := \sum_{i=1}^{N} x_i((\alpha+\beta)M)^{i-1}.$$

The function f will act as a "projection" of S' onto $\mathbb{Z}$. To show $f(S')$ is X-free, we will show a correspondence between copies of X' in $f(S')$ and three points on a line in S'. As all points in S' lie on a sphere, no three points in S' may lie on a line. By our correlation, this will imply $f(S')$ is X'-free.

Observe, as α, β, M are integers, $f(x)$ is an integer for all $x \in \mathbb{Z}^N$. Therefore, $\mathbb{Z}$ is the image of $f(x)$. Further, it is easy to see that f is a linear function.

The following claim will be used to show f is injective and X'-free:

Claim 2.2.4. *The function $f(\boldsymbol{x})$ is 0 on the domain $(-(\alpha+\beta)M, (\alpha+\beta)M)^N \cap \mathbb{Z}^N$ if and only if $\boldsymbol{x} = 0$.*

Proof. The backwards direction is immediate. Suppose that $f(\boldsymbol{x}) = 0$. We will show that $x_i = 0$ for all $i \in [N]$. First consider x_1. Define $s_1 = -\frac{x_1}{(\alpha+\beta)M}$. As $f(\boldsymbol{x}) = 0$, we have

$$-x_1 = \sum_{i=2}^{N} x_i((\alpha+\beta)M)^{i-1} = ((\alpha+\beta)M)\cdot s_1.$$

As $\boldsymbol{x} \in (-(\alpha+\beta)M, (\alpha+\beta)M)^N \cap \mathbb{Z}^N$, it follows that $0 \le |s_1| < 1$. But, it is trivial to see that as $\alpha, \beta, M \in \mathbb{N}$ and $x_i \in \mathbb{Z}$ for all $i \in [N]$,

$$s_1 = \sum_{i=2}^{N} x_i((\alpha+\beta)M)^{i-2} \in \mathbb{Z}.$$

This implies that $s_1 = 0$. Hence $x_1 = 0$. We may repeat this argument for x_2, x_3, and so on, observing that at step j we have that

$$-x_j = \sum_{i=j+1}^{N} x_i((\alpha+\beta)M)^{i-1-(j-1)} = ((\alpha+\beta)M)\cdot s_j.$$

This will imply that $x_j = 0$ using analogous observations as in the x_1 case. Therefore, $\boldsymbol{x} = 0$ as desired.

□

We will now show that $f(S')$ is X'-free. If $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in [M]^N$, then

$$(\alpha-\beta)\boldsymbol{x} + \beta\boldsymbol{y} - \alpha\boldsymbol{z} \in (-(\alpha+\beta)M, (\alpha+\beta)M)^N \cap \mathbb{Z}^N. \tag{2.1}$$

Suppose $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ form a copy of X', where $f(\boldsymbol{x}) < f(\boldsymbol{y}) < f(\boldsymbol{z})$. Then, $(\alpha-\beta)f(\boldsymbol{x}) + \beta f(\boldsymbol{y}) - \alpha f(\boldsymbol{z}) = 0$. By linearity of f, if $(\alpha-\beta)f(\boldsymbol{x}) + \beta f(\boldsymbol{y}) - \alpha f(\boldsymbol{z}) = 0$ then $f\left((\alpha-\beta)(\boldsymbol{x}) + \beta(\boldsymbol{y}) - \alpha(\boldsymbol{z})\right) = 0$. By Claim 2.2.4 and (2.1), it follows that

$$(\alpha-\beta)(\boldsymbol{x}) + \beta(\boldsymbol{y}) - \alpha(\boldsymbol{z}) = 0.$$

This implies that $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ are all on the same line. As S' is a sphere, at least one of these points is not in S'. It follows that $f(S')$ is X'-free.

We now show that $|f(S')| = |S'|$. Suppose $\boldsymbol{x}, \boldsymbol{y} \in [M]^N$ and $f(\boldsymbol{x}) = f(\boldsymbol{y})$. As f is linear, this implies $f(\boldsymbol{x}-\boldsymbol{y}) = 0$. Observe that $(\boldsymbol{x}-\boldsymbol{y}) \in (-(\alpha+\beta)M, (\alpha+\beta)M)^N \cap \mathbb{Z}^N$. It follows from Claim 2.2.4 that $\boldsymbol{x} = \boldsymbol{y}$. Therefore, f is injective on $[M]^N$ and so $|f(S')| = |S'|$.

Finally, we show that $f(S') \subseteq [n]$ to ensure our $f(S')$ is as desired. For any $\boldsymbol{x} \in [M]^N$, we have $f(\boldsymbol{x}) \ge 1$ as the coefficients of f are positive. Next, observe that

$$f(\boldsymbol{x}) \le \sum_{i=1}^{N} M((\alpha+\beta)M)^{i-1} = M \cdot \frac{((\alpha+\beta)\cdot M)^N - 1}{(\alpha+\beta)\cdot M - 1} \le ((\alpha+\beta)\cdot M)^N.$$

Therefore, $f(S') \subseteq f([M]^N) \subseteq [((\alpha+\beta)\cdot M)^N]$. Define $M = \left\lfloor \frac{n^{1/N}}{(\alpha+\beta)} \right\rfloor$, so $f(S') \subseteq [n]$. Therefore, $f(S')$ is as desired as it is an X-free subset of $[n]$ with cardinality at least $n \cdot e^{-c_X\sqrt{\log n}}$, where c_X is a positive constant dependent only on X.

□

2.3 Lower-bound on X-free sets

In this section, we extend Lemma 2.2.1 for any set finite $X \subseteq \mathbb{Z}^d$ such that $|X| \geq 3$ and $d \geq 1$. To do so, we will use Lemma 2.2.1 to project up a large 3-AP-free set to a higher dimension. We give an example of this process below.

Example 2.3.1. Let $X = \{(0,0),(0,1),(1,0)\}$ and $S' = \{1,2,4,9,11,12\} \subset [\frac{24}{2}]$. It can easily be shown that S' does not contain a 3-AP. We will use S' to construct an X-free subset of $[24]^2$. Consider

$$S = \{(x,y) \in [24]^2 : y - x \in S'\},$$

Figure 2.3 is a visual representation of how the set S is formed.

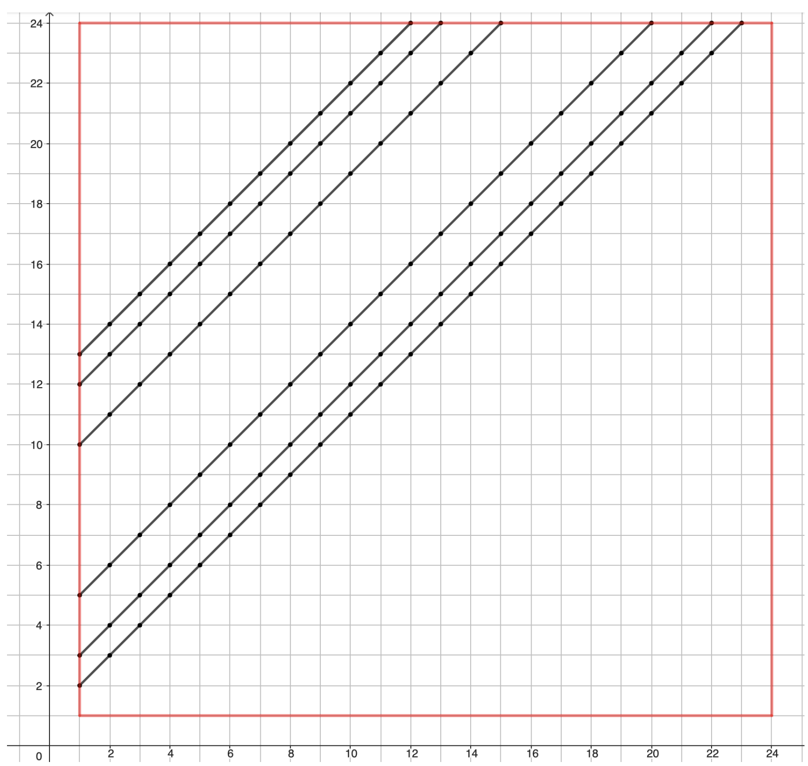

Figure 2.3: The black points form the set S.

Observe that S = {(1, 3), (2, 4), (3, 5), (4, 6), (5, 7), (6, 8), (7, 9), (8, 10), (9, 11), (10, 12), (11, 13), (12, 14), (13, 15), (14, 16), (15, 17), (16, 18), (17, 19), (18, 20), (19, 21), (20, 22), (21, 23), (22, 24), (1, 3), (2, 4), (3, 5), (4, 6), (5, 7), (6, 8), (7, 9), (8, 10), (9, 11), (10, 12), (11, 13), (12, 14), (13, 15), (14, 16), (15, 17), (16, 18), (17, 19), (18, 20), (19, 21), (20, 22), (21, 23), (22, 24), (1, 5), (2, 6), (3, 7), (4, 8), (5, 9), (6, 10), (7, 11), (8, 12), (9, 13), (10, 14), (11, 15), (12, 16), (13, 17), (14, 18), (15, 19), (16, 20), (17, 21), (18, 22), (19, 23), (20, 24), (1, 10), (2, 11), (3, 12), (4, 13), (5, 14), (6, 15), (7, 16), (8, 17), (9, 18), (10, 19), (11, 20), (12, 21), (13, 22), (14, 23), (15, 24)(1, 12), (2, 13), (3, 14), (4, 15), (5, 16), (6, 17), (7, 18), (8, 19), (9, 20), (10, 21), (11, 22), (12, 23), (13, 24), (1, 13), (2, 14), (3, 15), (4, 16), (5, 17), (6, 18), (7, 19), (8, 20), (9, 21), (10, 22), (11, 23), (12, 24)}.

Suppose S is not X-free. Then, there exists $x, y, d \in [24]$ such that $\{(x, y), (x + d, y), (x, y + d)\} \subseteq S$. But, then $\{y - x - d, y - x, y - x + d\} \subseteq S'$. This is not possible, as S' does not contain a 3-AP. So, S is X-free.

Let's briefly discuss the set X in Example 2.3.1, which is called a *corner in* $\mathbb{Z}^2$. An X-free set for this particular choice of X is said to be *corner-free*. Using a generalized version of the technique above, we are able to see that a dense 3-AP-free set in $[n]$ gives rise to a dense corner-free set in $[n]^2$. Therefore, a lower bound on the size of the largest corner-free subset of $\mathbb{Z}^2$ follows from Lemma 2.2.1. Lemma 2.3.2 demonstrates that for any finite $X \subseteq \mathbb{Z}^d$ where $|X| \geq 3$ we can achieve an analogous lower bound on $r_X(n)$ to Lemma 2.2.1 using Lemma 2.2.1 and a generalization of the technique in Example 2.3.1.

Lemma 2.3.2. *Let $d \geq 1$ and $X \subseteq \mathbb{N}^d$ be fixed. If $|X| \geq 3$, then there exists $c_X > 0$ depending only on X such that*

$$r_X(n) \geq n^d \cdot \exp(-c_X \sqrt{\log n}).$$

Proof of Lemma 2.3.2. If $d = 1$, then we are done by Lemma 2.2.1, and so we assume that $d \geq 2$. Since $r_X(n) \geq r_{X'}(n)$ for any subset X' of X, we may assume that $X = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$. Our goal is to construct, for each $n \geq 1$, a "large" set $S \subseteq [n]^d$ which is X-free. We may assume throughout the proof that n is divisible by $4(d-1)$; this is because for any $n' \leq n$, an X-free subset of $[n']$ is also an X-free subset of $[n]$. Therefore, we may extend to all n by picking a sufficiently large c_X.

We start by making numerous (valid) assumptions about the vectors $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$. Let T be the convex hull of $\{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$ and note that T is a (possibly degenerate) triangle. By relabelling $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ if necessary, we can assume that neither of the interior angles of T at $\boldsymbol{x}$ nor $\boldsymbol{y}$ are equal to $\pi/2$; this is because any triangle can contain at most one interior angle of $\pi/2$. Note that, in the special case that $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ are collinear, T is a degenerate triangle and the interior angles at $\boldsymbol{x}$ and $\boldsymbol{y}$ are either 0 or π and so this property holds automatically. Now, by translating X if necessary, we may assume that $\boldsymbol{z}$ is the zero vector. Let the entries of $\boldsymbol{x}$ and $\boldsymbol{y}$ be given by $\boldsymbol{x} = (x_1, \dots, x_d)$ and

$\boldsymbol{y} = (y_1, \ldots, y_d)$. Given that $\boldsymbol{z} = \mathbf{0}$, the condition that the interior angles of T at $\boldsymbol{x}$ and $\boldsymbol{y}$ are not equal to $\pi/2$ can be expressed algebraically as follows:

$$\langle \boldsymbol{x}, \boldsymbol{x} - \boldsymbol{y} \rangle = \sum_{i=1}^{d} x_i(x_i - y_i) \neq 0, \tag{2.2}$$

$$\langle \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle = \sum_{i=1}^{d} y_i(x_i - y_i) \neq 0, \tag{2.3}$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^d$. Since $\boldsymbol{x} \neq \boldsymbol{y}$, by reordering the indices, rescaling X, and swapping $\boldsymbol{x}$ and $\boldsymbol{y}$ if necessary, we can assume that $x_d - y_d = 1$ and that $|x_i - y_i| \leq 1$ for all $i \in [d]$. Note that the rescaling may put $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{Q}^d$. For each $i \in [d-1]$, let $x_i - y_i = r_i/q_i$ such that r_i and $q_i > 0$ are integers with q_i as small as possible. Note that $q := \max_{i \in [d-1]} q_i$ is a constant depending on X. These assumptions about $\boldsymbol{x}$ and $\boldsymbol{y}$ will be convenient later on, as they make it easier to work with the coordinates of $\boldsymbol{x} - \boldsymbol{y}$.

Next, we define

$$t := \frac{\sum_{i=1}^{d} y_i(y_i - x_i)}{\sum_{i=1}^{d} (x_i - y_i)^2}.$$

Note that t is well-defined because $\boldsymbol{x} \neq \boldsymbol{y}$, and that $t \in \mathbb{Q}$ because $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Q}^d$. Moreover, t can be regarded as a quantity which depends only on X. Also, (2.3) implies that $t \neq 0$ and (2.2) implies that $t \neq 1$. So, $X' := \{0, t, 1\}$ is a subset of $\mathbb{Q}$ of cardinality 3. Using Lemma 2.2.1, we let S' be a subset of $[n/4]$ which is X'-free and satisfies

$$|S'| \geq (n/4) \cdot e^{-c_{X'}\sqrt{\log(n/4)}}.$$

Note that X' depends on t which, in turn, depends on X, and so $c_{X'}$ is a constant that depends on X. Define $S \subseteq [n]^d$ by

$$S := \left\{ \boldsymbol{s} = (s_1, \ldots, s_d) \in [n]^d : \sum_{i=1}^{d} \left(s_i - \frac{n}{2}\right)(x_i - y_i) \in S' \right\}.$$

To complete the proof, we need to analyze the cardinality of S and show that it is X-free. We start with its cardinality. For each $i \in [d-1]$, since $|x_i - y_i| \leq 1$, the number of choices for $s_i \in [n]$ such that

$$-\frac{n}{4(d-1)} \leq \left(s_i - \frac{n}{2}\right)(x_i - y_i) \leq 0$$

is at least $\frac{n}{4(d-1)}$. Since $x_i - y_i$ is a rational with denominator bounded by the constant q for all $i \in [d-1]$, the number of such s_i for which $\left(s_i - \frac{n}{2}\right)(x_i - y_i)$ is also an integer is at least $\left\lfloor \frac{n}{4q(d-1)} \right\rfloor$. Therefore, the number of choices for $(s_1, \ldots, s_{d-1}) \in [n]^{d-1}$ such that

$$\sum_{i=1}^{d-1} \left(s_i - \frac{n}{2}\right)(x_i - y_i) \in \mathbb{Z} \cap \left[-\frac{n}{4}, 0\right]$$

is $\Omega(n^{d-1})$, where the constant factor depends only on d and q (which, in turn, depends on X). Recalling that $x_d - y_d = 1$ and that $S' \subseteq [n/4]$, we see that, for any such $(s_1, \ldots, s_{d-1})$ and any $a \in S'$, there is a unique choice of $s_d \in [n]$ so that

$$\sum_{i=1}^{d} \left(s_i - \frac{n}{2}\right)(x_i - y_i) = a.$$

Thus, we get that

$$|S| = \Omega(n^{d-1}|S'|) = n^d \cdot e^{-c_X\sqrt{\log n}}$$

for a sufficiently small constant $c_X > 0$ depending on X, as desired.

Finally, we show that S is X-free. Suppose, to the contrary, that there exists a non-trivial copy of X in S. In other words, there exists $\boldsymbol{b} \in \mathbb{Q}^d$ and $r \in \mathbb{Q} \cap (0, \infty)$ such that

$$\{\boldsymbol{b}, \boldsymbol{b} + r\boldsymbol{x}, \boldsymbol{b} + r\boldsymbol{y}\} \subseteq S.$$

Let $\boldsymbol{b} = (p_1, \ldots, p_d)$ and define

$$a := \sum_{i=1}^{d} \left(p_i - \frac{n}{2}\right)(x_i - y_i),$$

$$b := \sum_{i=1}^{d} \left(p_i + rx_i - \frac{n}{2}\right)(x_i - y_i),$$

$$c := \sum_{i=1}^{d} \left(p_i + ry_i - \frac{n}{2}\right)(x_i - y_i).$$

The condition $\{\boldsymbol{b}, \boldsymbol{b} + r\boldsymbol{x}, \boldsymbol{b} + r\boldsymbol{y}\} \subseteq S$ is equivalent to $a, b, c \in S'$. We claim that $\{a, b, c\}$ is a non-trivial copy of $\{0, t, 1\}$, which will contradict the definition of S' and complete the proof. Since $r \neq 0$, we have

$$\frac{a-c}{b-c} = \frac{r\sum_{i=1}^{d} y_i(y_i - x_i)}{r\sum_{i=1}^{d}(x_i - y_i)^2} = t.$$

Therefore

$$c + (b-c)\{0, t, 1\} = c + \{0, a-c, b-c\} = \{a, b, c\}$$

and so $\{a, b, c\}$ is a copy of $\{0, t, 1\}$. It is non-trivial because $r \neq 0$ and $\boldsymbol{x} \neq \boldsymbol{y}$ implies

$b - c = r \cdot \sum_{i=1}^{d}(x_i - y_i)^2$ is non-zero. This contradiction completes the proof. $\square$

Chapter 3
Copies of X in supersaturated sets

3.1 Introduction

Recall from Section 1.2 that a supersaturation result in the context of X-free sets lower bounds the number of copies of X in a set of size larger than $r_X(n)$. In this chapter, we will prove three progressively stronger supersaturation results for subsets of $[n]^d$ of cardinality greater than $r_X(n)$. As discussed at length in Chapter 2, we do not know exact asymptotics for even $r_3(n)$. So, supersaturation results for X-free sets are difficult as we are working without full understanding of the structure or even cardinality of extremal X-free sets. Despite these difficulties, some work has been done on the supersaturation problem for sets without k-APs. Varnavides [82] proved in 1959 that any subset of size $\Omega(n)$ contains at least $\Omega(n^2)$ k-APs. Croot and Sisask [14], in giving a new proof of Roth's Theorem, used an application of a quantitative version of Varnavides' [82] result. With this, they were able to show that any $A \subset [n]$ and every $1 \leq M \leq n$, the number of 3-APs contained in A is at least

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

The following theorem of Balogh, Liu, and Sharifzadeh [3] gives a supersaturation result for an infinite subset of $\mathbb{N}$.

Theorem 3.1.1 (Theorem 1.6 in [3]). *Given $k \geq 3$, there exists a constant $C' = C'(k) > 1$ and an infinite set $N \subseteq \mathbb{N}$ such that the following holds. For any $n \in N$ and any $A \subset [n]$ of size $C' \cdot r_k(n)$, the number of non-trivial k-APs in A is at least*

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

The main result of this chapter is a generalization of Theorem 3.1.1 in Section 3.3

to all finite sets $X \subseteq \mathbb{N}$ of cardinality at least 3.

3.2 Supersaturation Lemmas

In this subsection we prove some preliminary supersaturation results that will be used to prove our main supersaturation result in Section 3.3. For $A \subseteq [n]^d$, let $\Gamma_X(A)$ denote the number of non-trivial copies of X in A. Trivially, if $|A| \subseteq [n]^d$ such that $|A| > r_X(n)$ then A must contain at least one copy of X. We warm-up with the following only slightly less trivial lower bound on $\Gamma_X(A)$.

Lemma 3.2.1. *For any $A \subseteq [n]^d$,*

$$\Gamma_X(A) \geq |A| - r_X(n).$$

Proof. We proceed by induction on $|A|$. If $|A| \leq r_X(n)$, then the right side of the inequality is at most zero so there is nothing to prove. Now, assume $|A| \geq r_X(n) + 1$. Then A must contain at least one copy of X, say X'. For any $x \in X'$, the set $A \setminus \{x\}$ contains at least $|A| - 1 - r_X(n)$ copies of X by the induction hypothesis. These copies, along with X' which is not in $A \setminus \{x\}$, yield $|A| - r_X(n)$ copies of X in A. □

The argument to obtain Lemma 3.2.1 is, obviously, very simple. However, as we see in Lemma 3.2.3 below, we can obtain a seemingly stronger bound by applying Lemma 3.2.1 within copies of $[M]^d$ inside of $[n]^d$, where $1 \leq M \leq n$, and using an "averaging argument". The copies of $[M]^d$ that we consider are translates of $p \cdot [M]^d$ where p is a small prime; the following consequence of the prime number theorem (see, e.g., [76]) is useful for counting the choices of p. For $\ell \geq 2$, let $\pi(\ell)$ be the number of primes p such that $2 \leq p \leq \ell$.

Proposition 3.2.2. *There exists $\ell_0 \in \mathbb{N}$ such that $\pi(\ell) \geq \frac{\ell}{2\log(\ell)}$ for all $\ell \geq \ell_0$.*

Lemma 3.2.3. *Let M and n be integers such that $2 \leq M \leq n$. If $A \subseteq [n]^d$ such that*

$$|A| > \max\{4\ell_0 d n^{d-1} M, r_X(n)\} \tag{3.1}$$

where ℓ_0 is as in Proposition 3.2.2, then

$$\Gamma_X(A) \geq \frac{|A|}{23d\log^2(n)} \cdot \frac{n}{M}\left(\frac{|A|}{2n^d} - \frac{r_X(M)}{M^d}\right).$$

Proof. Recall that Observation 2.2.3 states that if X' is a (non-trivial) copy of X, then a set is X-free if and only if it is X'-free. So, by translating and scaling X if necessary, we may assume that there does not exist any "smaller" non-trivial copy of X in $\mathbb{N}^d$.

Specifically, we can assume that there does not exist an integer $t > 1$ which divides the coordinates of the vector $\boldsymbol{x} - \boldsymbol{y}$ for all $\boldsymbol{x}, \boldsymbol{y} \in X$.

Let p be a prime chosen uniformly at random from all primes $p \leq \frac{|A|}{4dn^{d-1}M}$. Next, choose a "base point" $\boldsymbol{b}$ uniformly at random from the set $\{0, \dots, n - Mp - 1\}^d$ (which is non-empty because $Mp \leq \frac{|A|}{4dn^{d-1}} < n$) and define

$$G_{\boldsymbol{b}} = \boldsymbol{b} + p \cdot [M]^d.$$

Let $A_{\boldsymbol{b}} := A \cap G_{\boldsymbol{b}}$. Our goal is to bound $\mathbb{E}[\Gamma_X(A_{\boldsymbol{b}})]$ from below in terms of $|A|$ and from above in terms of $\Gamma_X(A)$; combining these inequalities will give us the desired lower bound on $\Gamma_X(A)$. Note, we choose to scale $[M]^d$ by a prime as it allows us to have a sub-linear upper bound on the number of possible scalars to consider via Proposition 3.2.2.

Note that $G_{\boldsymbol{b}}$ is a non-trivial copy of $[M]^d$, so we can apply Lemma 3.2.1 within $G_{\boldsymbol{b}}$ to yield

$$\mathbb{E}[\Gamma_X(A_{\boldsymbol{b}})] \geq \mathbb{E}[|A_{\boldsymbol{b}}| - r_X(M)] = \mathbb{E}[|A_{\boldsymbol{b}}|] - r_X(M) \tag{3.2}$$

by linearity of expectation. So, to get a lower bound on $\mathbb{E}[\Gamma_X(A_{\boldsymbol{b}})]$, it suffices to bound $\mathbb{E}[|A_{\boldsymbol{b}}|]$. Define $I_p := \{Mp+1, \dots, n - Mp - 1\}^d \cap [n]^d$; that is I_p is obtained from $[n]^d$ by removing all points that are "too close" to the boundary in some coordinate. As there are fewer than $2dn^{d-1}Mp$ such points and $p \leq \frac{|A|}{4dn^{d-1}M}$, we have that

$$|A \cap I_p| > |A| - 2dn^{d-1}Mp \geq |A| - 2dn^{d-1}M\left(\frac{|A|}{4dn^{d-1}M}\right) = \frac{|A|}{2}$$

for every possible choice of p. Given the choice of p, for any $v \in I_p$ there are exactly M^d choices of base point $\boldsymbol{b} \in \{0, \dots, n - Mp - 1\}^d$ with the property that $v \in G_{\boldsymbol{b}}$. Thus, conditioned on the choice of p, the probability that any given $v \in I_p$ is in $G_{\boldsymbol{b}}$ is at least $\frac{M^d}{(n-Mp)^d} \geq \frac{M^d}{n^d}$. Therefore, $\mathbb{E}[|A_{\boldsymbol{b}}|] \geq \mathbb{E}[|A \cap I_p \cap G_{\boldsymbol{b}}|] \geq \frac{|A|}{2} \cdot \frac{M^d}{n^d}$. Substituting this lower bound into (3.2) gives

$$\mathbb{E}[\Gamma_X(A_{\boldsymbol{b}})] \geq \frac{|A|}{2} \cdot \frac{M^d}{n^d} - r_X(M). \tag{3.3}$$

Next, we will obtain an upper bound on $\mathbb{E}[\Gamma_X(A_{\boldsymbol{b}})]$ in terms of $\Gamma_X(A)$. To do this, we let Y be an arbitrary fixed non-trivial copy of X in A and bound the probability that $Y \subseteq G_{\boldsymbol{b}}$ from above. Let $\boldsymbol{a} \in \mathbb{R}^d$ and $r > 0$ so that $Y = \boldsymbol{a} + r \cdot X$ and note that, since $|X| \geq 2$, the choice of $\boldsymbol{a}$ and r is unique. Then, clearly, Y is a subset of $G_{\boldsymbol{b}}$ if and only if $(\boldsymbol{a} - \boldsymbol{b}) + r \cdot X$ is a subset of $p \cdot [M]^d$. Thus, if Y is a subset of $G_{\boldsymbol{b}}$, then, for any $\boldsymbol{x}, \boldsymbol{y} \in X$, every coordinate of

$$(\boldsymbol{a} - \boldsymbol{b} + r\boldsymbol{x}) - (\boldsymbol{a} - \boldsymbol{b} + r\boldsymbol{y}) = r(\boldsymbol{x} - \boldsymbol{y})$$

is a multiple of p. By our assumption that X is the smallest copy of itself in $\mathbb{N}^d$, this implies that r is an integer and p divides r. Thus, for any non-trivial copy $Y = \boldsymbol{a} + rX$ of X in A, the probability that Y is a subset of $G_{\boldsymbol{b}}$ is at most the probability that p is a divisor of r, multiplied by the probability that, for an arbitrary $\boldsymbol{y}_0 \in Y$, the base point $\boldsymbol{b}$ is chosen to be one of the at most M^d base points for which $\boldsymbol{y}_0 \in G_{\boldsymbol{b}}$. Let us now bound these two probabilities individually. Since, for any such Y, the integer r has at most $\log_2(r) \leq \log_2(n)$ divisors, the probability that p divides r is at most

$$\frac{\log_2(n)}{\pi(|A|/4dn^{d-1}M)} \leq \frac{\log_2(n) 8dn^{d-1}M \log(|A|/4dn^{d-1}M)}{|A|} \leq \frac{8\log_2^2(n) \cdot Mdn^{d-1}}{|A|}$$

where the first inequality uses Proposition 3.2.2 and the fact that $|A| > 4\ell_0 dn^{d-1}M$. After fixing the choice of p, for any $\boldsymbol{y}_0 \in Y$, the probability that $\boldsymbol{y}_0 \in G_{\boldsymbol{b}}$ is at most

$$\frac{M^d}{(n - Mp)^d} \leq \frac{M^d}{\left(n - \frac{|A|}{4dn^{d-1}}\right)^d}.$$

Observe that

$$\left(\frac{M}{n - \frac{|A|}{4dn^{d-1}}}\right)^d = \left(\frac{M}{n}\right)^d \left(1 - \frac{|A|}{4dn^d}\right)^{-d} \leq \left(\frac{M}{n}\right)^d \left(1 - \frac{1}{4d}\right)^{-d} \leq \left(\frac{M}{n}\right)^d \cdot \frac{4}{3}.$$

Therefore, for any non-trivial copy Y of X in A, we have

$$\mathbb{P}[Y \subseteq A_{\boldsymbol{b}}] \leq \frac{8\log_2^2(n) \cdot Mdn^{d-1}}{|A|} \cdot \left(\frac{M}{n}\right)^d \cdot \frac{4}{3} \leq \frac{11 \cdot \log_2^2(n) M^{d+1} d}{|A| n}.$$

Using this, we get the following upper bound on $\mathbb{E}[\Gamma_X(A_{\boldsymbol{b}})]$ by linearity of expectation:

$$\mathbb{E}[\Gamma_X(A_{\boldsymbol{b}})] \leq \Gamma_X(A) \cdot \frac{11 \cdot \log_2^2(n) M^{d+1} d}{|A| n}. \tag{3.4}$$

Putting together (3.3) and (3.4) and solving for $\Gamma_X(A)$ gives,

$$\Gamma_X(A) \geq \frac{n|A|}{11\log_2^2(n) M^{d+1} d} \left(\frac{|A| M^d}{2n^d} - r_X(M)\right) \geq \frac{|A|}{23d\log^2(n)} \cdot \frac{n}{M}\left(\frac{|A|}{2n^d} - \frac{r_X(M)}{M^d}\right),$$

as desired. □

Let's pause for a second to make sense of the statement given in Lemma 3.2.3. Observe that to obtain a non-trivial lower bound on $\Gamma_X(A)$, we require that

$$\frac{|A|}{n^d} \geq 2 \cdot \frac{r_X(M)}{M^d}.$$

So, for Lemma 3.2.3 to give a non-trivial lower bound on $\Gamma_X(A)$ we require that A has at least twice the density in $[n]^d$ that $r_X(M)$ has in $[M]^d$. If we wish to be able to apply Lemma 3.2.3 in the case that $|A| = C \cdot r_X(n)$, for some constant $C \geq 1$ then we require that the density of $r_X(M)$ is at most a constant times $r_X(n)$. While it is conceivable that this is true for many choices of n and M, without exact asymptotics of $r_X(n)$ we cannot make such a statement in generality. We know from Lemma 2.3.2 that the density of $r_X(n)$ does not decrease too rapidly asymptotically, but we cannot say the decrease in density is "smooth". In the next section, we will look to find an infinite set of values of n and M such that we have the desired relation between the densities of $r_X(n)$ and $r_X(M)$.

A careful reader will observe that Lemma 3.2.3 does not require that $|A| > r_X(n)$. As it is not completely obvious why such a relaxation should be true, let's briefly consider why Lemma 3.2.3 holds if we remove the condition $|A| > r_X(n)$. First, observe that if $M \geq \frac{r_X(n)}{4\ell_0 dn^{d-1}}$ it is trivial to remove the condition $|A| > r_X(n)$ as this is implied by the condition $|A| > 4\ell_0 dn^{d-1}M$. Therefore, we only need to consider why Lemma 3.2.3 should hold when $M < \frac{r_X(n)}{4\ell_0 dn^{d-1}}$.

Consider a set A such that $|A| \leq r_X(n)$; from the cardinality of A, it does not need to contain copies of X. As we may have $\Gamma_X(A) = 0$, for Lemma 3.2.3 to hold it is sufficient to have

$$\frac{|A|}{2n^d} \leq \frac{r_X(n)}{2n^d} \leq \frac{r_X(M)}{M^d}. \tag{3.5}$$

The left inequality in (3.5) is true by definition of A. For the right inequality, we use Proposition 3.2.4.

Proposition 3.2.4. *Let M and n be integers such that $2 \leq M \leq n$ then,*

$$\frac{(r_X(n) - 2dn^{d-1}M)}{n^d} \leq \frac{r_X(M)}{M^d}.$$

Before proceeding with the proof, let us demonstrate why Proposition 3.2.4 implies the right-hand side of (3.5) when $M < \frac{r_X(n)}{4\ell_0 dn^{d-1}}$. Substituting this upper bound for M in the inequality given by Proposition 3.2.4 implies

$$\frac{\left(r_X(n) - 2d\ell_0 n^{d-1} \cdot \frac{r_X(n)}{4\ell_0 dn^{d-1}}\right)}{n^d} = \frac{r_X(n)}{2n^d} \leq \frac{(r_X(n) - 2dn^{d-1}M)}{n^d} \leq \frac{r_X(M)}{M^d},$$

as desired. We now prove Proposition 3.2.4.

Proof of Proposition 3.2.4. Let $A \subseteq [n]^d$ such that $|A| = r_X(n)$ and $\Gamma_X(A) = 0$, where such a set A must exist by definition of $r_X(n)$. Choose a "base point" $\boldsymbol{b}$ uniformly at random from the set $\{0, \ldots, n - M - 1\}^d$, noting that this is non-empty as $M \leq n$.

Define

$$G_{\boldsymbol{b}} = \boldsymbol{b} + [M]^d.$$

Define $I = \{M+1, \dots n-M-1\}$; that is, let I is obtained by removing all points in $[n]^d$ "too close" to the boundary. As there are fewer than $2dn^{d-1}M$ such points, we have that

$$|A \cap I| > |A| - 2dn^{d-1}M = r_X(n) - 2dn^{d-1}M.$$

For each point $\boldsymbol{v} \in I$, there are exactly M^d choices of $\boldsymbol{b}$ such that $\boldsymbol{v} \in G_{\boldsymbol{b}}$, as such a point $\boldsymbol{v}$ may play any "role" in $G_{\boldsymbol{b}}$. So, each point $v \in I \cap A$ is contained in $G_{\boldsymbol{b}}$ with probability $\frac{M^d}{(n-M)^d} \geq \frac{M^d}{n^d}$. So,

$$\mathbb{E}[A \cap G_{\boldsymbol{b}}] \geq \mathbb{E}[A \cap I \cap G_{\boldsymbol{b}}] \geq (r_X(n) - 2dn^{d-1}M) \cdot \frac{M^d}{n^d}. \tag{3.6}$$

As A is X-free the set $A \cap G_{\boldsymbol{b}}$ must also be X-free. Therefore, as $A \cap G_{\boldsymbol{b}}$ is an X-free subset of a translated copy of $[M]^d$, it contains at most $r_X(M)$ points. So,

$$\mathbb{E}[A \cap G_{\boldsymbol{b}}] \leq r_X(M). \tag{3.7}$$

Putting together (3.6) and (3.7) and rearranging gives the result. □

3.3 Stronger Supersaturation on an Infinite Set

The goal of this section is to find an infinite set $N \subseteq \mathbb{N}$ such that for each $n \in N$ we may carefully choose $m(n)$ such that the density of $r_X(n)$ is at least the density of $r_X(m(n))$ multiplied by a small constant $\alpha > 0$. This will allow us to apply Lemma 3.2.3 to the elements of N and achieve a supersaturation result for subsets A with cardinality at least $C \cdot r_X(n)$.

We first prove the following more general relationship between sequences of a certain form. More specifically, we show that for sequences with a specific lower bound there exists an infinite sequence $N \subseteq \mathbb{N}$ where the function is relatively "smooth". We note that the lower bound on $r(n)$ in Lemma 3.3.1 is motivated by the lower bound on $r_X(n)$ obtained by Lemma 2.3.2.

Lemma 3.3.1. *For $\beta_r, \beta_m > 0$, let $0 < \alpha < \exp(-\beta_r\beta_m/2)$. For $d \geq 1$, suppose that $r : \mathbb{N} \to \mathbb{N}$ and $m : \mathbb{N} \to \mathbb{N}$ are such that*

$$r(n) \geq \frac{n^d}{\exp(\beta_r\sqrt{\log(n)})}$$

and

$$n \geq m(n) \geq \frac{n}{\exp(\beta_m \sqrt{\log(n)})}$$

for all $n \geq 1$. *Then there exists an infinite subset* N *of* $\mathbb{N}$ *such that*

$$\frac{r(n)}{n^d} \geq \frac{\alpha \cdot r(m(n))}{m(n)^d}$$

for all $n \in N$.

Proof. Take $\varepsilon > 0$ so that $\alpha = \frac{\exp(\frac{-\beta_r \beta_m}{2})}{(1+\varepsilon)^3}$. By definition of $m(n)$,

$$\begin{aligned} \sqrt{\log(m(n))} - \sqrt{\log(n)} &\geq \sqrt{\log\left(\frac{n}{\exp(\beta_m \sqrt{\log(n)})}\right)} - \sqrt{\log(n)} \\ &= \sqrt{\log(n) - \beta_m \sqrt{\log(n)}} - \sqrt{\log(n)}. \end{aligned} \tag{3.8}$$

Observe that the expression in (3.8) converges to $\frac{-\beta_m}{2}$ as $n \to \infty$. Therefore, there exists some $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$\sqrt{\log(m(n))} - \sqrt{\log(n)} \geq \frac{-\beta_m}{2} - \frac{\log(1+\varepsilon)}{\beta_r}.$$

This implies that for all $n \geq n_0$,

$$\exp\left(\beta_r \left(\sqrt{\log(m(n))} - \sqrt{\log(n)}\right)\right) \geq \frac{\exp\left(\frac{-\beta_r \beta_m}{2}\right)}{(1+\varepsilon)}. \tag{3.9}$$

Define

$$Y := \left\{ \frac{r(n) \exp(\beta_r \sqrt{\log(n)})}{n^d} : n \in \mathbb{N} \right\}.$$

Let Z be the set of all cluster points of Y in the extended reals, $\mathbb{R} \cup \{-\infty, \infty\}$. Note that, by the lower bound on $r(n)$, we have $Z \subseteq [1, \infty]$. Let $z = \sup(Z)$. We now proceed by cases depending on the value of z.

Suppose $z = \infty$. Define

$$N := \left\{ n \in \mathbb{N} : n \geq n_0 \text{ and } \frac{r(n) \exp(\beta_r \sqrt{\log(n)})}{n^d} \geq \frac{r(m) \exp(\beta_r \sqrt{\log(m)})}{m^d}, \forall m \leq n \right\}.$$

As $z = \infty$, it follows that N is an infinite set. As $m(n) \leq n$, for all $n \in N$ we have,

$$\frac{r(n)}{n^d} \geq \frac{r(m(n)) \cdot \exp(\beta_r \sqrt{\log(m(n))})}{m(n)^d \cdot \exp(\beta_r \sqrt{\log(n)})} \geq \frac{r(m(n))}{m(n)^d} \cdot \frac{\exp(\frac{-\beta_m \beta_r}{2})}{(1+\varepsilon)} \geq \alpha \cdot \frac{r(m(n))}{m(n)^d},$$

where the second inequality follows from (3.9). Therefore, N is as desired.

We may now assume that $z < \infty$. Choose n_0' large enough such that for all $n \geq n_0'$,

$$r(m(n)) \cdot \frac{\exp(\beta_r \sqrt{\log(m(n))})}{m(n)^d} \leq (1+\varepsilon)z. \tag{3.10}$$

Note that such an n_0' exists as $m(n) \to \infty$ as $n \to \infty$. Define

$$N := \left\{ n \in \mathbb{N} : n \geq \max\{n_0, n_0'\} \text{ and } \frac{r(n) \cdot \exp(\beta_r \sqrt{\log(n)})}{n^d} \geq \frac{z}{(1+\varepsilon)} \right\}.$$

By definition of z, the set N is infinite. By definition of N, for all $n \in N$,

$$\begin{aligned}\frac{r(n)}{n^d} &\geq \frac{z}{(1+\varepsilon)\cdot\exp(\beta_r\sqrt{\log(n)})} \overset{(3.10)}{\geq} \frac{r(m(n))}{m(n)^d} \cdot \frac{\exp(\beta_r\sqrt{\log(m(n))})}{\exp(\beta_r\sqrt{\log(n)})\cdot(1+\varepsilon)^2} \\ &\overset{(3.9)}{\geq} \frac{\exp(-\frac{\beta_r\beta_m}{2})}{(1+\varepsilon)^3} \cdot \frac{r(m(n))}{m(n)^d} = \alpha \cdot \frac{r(m(n))}{m(n)^d}.\end{aligned}$$

Therefore, N is as desired. □

In proving the main result of this chapter, we will apply Lemma 3.3.1 with $r(n) = r_X(n)$. Recall the following result that we proved in Chapter 2:

Lemma 2.3.2. *Let $d \geq 1$ and $X \subseteq \mathbb{N}^d$ be fixed. If $|X| \geq 3$, then there exists $c_X > 0$ depending only on X such that*

$$r_X(n) \geq n^d \cdot \exp(-c_X\sqrt{\log n}).$$

In order to apply Lemma 3.3.1, we require another function $m(n)$. Define, for the rest of this chapter,

$$m(n) := \left\lfloor \frac{n}{\log^{3|X|}(n)} \cdot \left(\frac{r_X(n)}{n^d}\right)^{|X|+2} \right\rfloor.$$

Observe, as $m(n) \to \infty$ as $n \to \infty$, for sufficiently large n we have

$$m(n) \geq \frac{1}{2} \cdot \frac{n}{\log^{3|X|}(n)} \cdot \left(\frac{r_X(n)}{n^d}\right)^{|X|+2} \tag{3.11}$$

Using Lemma 2.3.2 and (3.11), observe that, for sufficiently large n,

$$m(n) = \left\lfloor \frac{n}{\log^{3|X|}(n)} \cdot \left(\frac{r_X(n)}{n^d}\right)^{|X|+2} \right\rfloor \geq \frac{n}{\exp\left(-2c_X|X|\sqrt{\log(n)}\right)} \cdot \left(\exp(-c_X\sqrt{\log n}\right)^{|X|+2}$$
$$> n \cdot \left(\exp\left(-(5c_X|X|)\sqrt{\log n})\right)\right). \tag{3.12}$$

Let α be chosen such that $0 < \alpha < \exp\left(-\frac{5c_X^2|X|}{2}\right)$. This choice of α is fixed for the remainder of this chapter.

It is interesting to note how perfectly the bound in Lemma 2.3.2 works in our context. Analyzing the proof of Lemma 3.3.1, one can observe that much of the proof relies on the well-known fact that the sequence $(n + n^\beta)^\beta - n^\beta$ converges if and only if $\beta \leq 1/2$. So, if the lower bound in Lemma 2.3.2 was even slightly weaker (i.e. if we only had $\frac{r_X(n)}{n^d} \geq \exp(-\beta_r(\log(n))^{2/3})$ instead), we could not apply this sort of argument. In fact, it is not clear if such a statement should hold in that case. As this bound was derived from the analogous result for 3-APs from Behrend [7] proved in 1946, it is somewhat surprising something stronger is not needed.

We now prove our final lower bound for $\Gamma_X(A)$ for large sets $A \subseteq [n]^d$, where $n \in N$ for some infinite subset $N \subseteq \mathbb{N}$.

Lemma 3.3.2. *There exists an infinite subset N of $\mathbb{N}$ such that, if $n \in N$ and $A \subseteq [n]^d$ where $|A| \geq \frac{4}{\alpha} \cdot r_X(n)$, then*

$$\Gamma_X(A) \geq \frac{\log^{3|X|-2}(n)}{6d} \cdot \left(\frac{n^d}{r_X(n)}\right)^{|X|} \cdot n^d.$$

Proof. Let ℓ_0 be as defined in Proposition 3.2.2. By Lemma 2.3.2 and (3.12), we may apply Lemma 3.3.1 with $m(n)$, $r_X(n)$, and α to obtain an infinite subset N' of $\mathbb{N}$, such that for all $n \in N'$,

$$\frac{r_X(n)}{\alpha \cdot n^d} \geq \frac{r_X(m(n))}{m(n)^d}. \tag{3.13}$$

Note that $\alpha < 1$ and so the condition $|A| \geq \frac{4}{\alpha} \cdot r_X(n)$ implies that $|A| \geq r_X(n)$. For large enough $n \in N'$, we have

$$4\ell_0 dn^{d-1}m(n) = 4\ell_0 dn^{d-1} \cdot \frac{n}{\log^{3|X|}(n)} \cdot \left(\frac{r_X(n)}{n^d}\right)^{|X|+2} \leq 4\ell_0 dn^d \left(\frac{r_X(n)}{n^d}\right)^2$$
$$= 4\ell_0 d \left(\frac{r_X(n)}{n^d}\right) \cdot r_X(n) < \frac{4}{\alpha} \cdot r_X(n) \leq |A|,$$

where the penultimate inequality holds for sufficiently large n as $r_X(n) = o(n^d)$. Therefore A satisfies the conditions of Lemma 3.2.3 with $M = m(n)$. Applying this we obtain

$$\Gamma_X(A) \geq \frac{1}{23d\log^2(n)} \cdot |A| \cdot \frac{n}{m(n)} \cdot \left(\frac{|A|}{2n^d} - \frac{r_X(m(n))}{m(n)^d} \right).$$

Using $|A| \geq \frac{4}{\alpha} \cdot r_X(n)$ and using (3.13) to bound the final term, we get

$$\begin{aligned}\Gamma_X(A) &\geq \frac{1}{23d\log^2(n)} \cdot |A| \cdot \frac{n}{m(n)} \cdot \left(\frac{2r_X(n)}{\alpha \cdot n^d} - \frac{r_X(n)}{\alpha \cdot n^d} \right) \\ &= \frac{1}{23d\log^2(n)} \cdot |A| \cdot \frac{n}{m(n)} \cdot \left(\frac{r_X(n)}{\alpha \cdot n^d} \right).\end{aligned}$$

Further, we may apply the definition of $m(n)$ and the lower bound on $|A|$ to get

$$\begin{aligned}\Gamma_X(A) &\geq \frac{1}{23d\log^2(n)} \cdot |A| \cdot n \cdot \frac{\log^{3|X|}(n)}{n} \cdot \left(\frac{n^d}{r_X(n)} \right)^{|X|+2} \cdot \left(\frac{r_X(n)}{\alpha \cdot n^d} \right) \\ &\geq \frac{1}{23d\log^2(n)} \cdot \frac{4}{\alpha} \cdot r_X(n) \cdot \log^{3|X|}(n) \cdot \left(\frac{n^d}{r_X(n)} \right)^{|X|+2} \cdot \left(\frac{r_X(n)}{\alpha \cdot n^d} \right) \\ &\geq \frac{\log^{3|X|-2}(n)}{6d\alpha^2} \cdot \left(\frac{n^d}{r_X(n)} \right)^{|X|} \cdot n^d.\end{aligned}$$

Hence, as $\alpha < 1$, we have for large enough $n \in N'$

$$\Gamma_X(A) \geq \frac{\log^{3|X|-2}(n)}{6d} \cdot \left(\frac{n^d}{r_X(n)} \right)^{|X|} \cdot n^d. \tag{3.14}$$

Define N to be the set of all $n \in N'$ which are large enough that (3.14) holds. □

Chapter 4
Approximate Counting Version of the Multidimensional Szemerédi Theorem

In this chapter, we prove Theorem 1.3.2, an approximate counting version of the multidimensional Szemerédi theorem. We begin with a discussion on the hypergraph container method in Section 4.1 which we then apply, in conjunction with results from the previous chapters, to prove Theorem 1.3.2 in Section 4.2.

4.1 A Short Discussion on Hypergraph Containers

Many problems in combinatorics may be viewed as the study of *independent sets*. An independent set in a graph or hypergraph is a collection of vertices that does not contain any edges. Consider the 3-uniform hypergraph $\mathcal{H}$ with vertex-set $V(\mathcal{H}) = E(K_n)$ and edge-set $E(\mathcal{H}) = \{\{e_i, e_j, e_k\} : \ e_i, e_j, e_k \text{ form a triangle in } K_n\}$. Observe that an independent set in $\mathcal{H}$ corresponds to the edge set of a triangle-free subgraph of K_n. Therefore, the classical Mantel's Theorem [47] may be equivalently stated by saying the largest independent set in a 3-uniform hypergraph that encodes copies of K_3 in K_n has order $\left\lfloor \frac{n^2}{4} \right\rfloor$.

The hypergraph container method upper bounds the number of independent sets in a hypergraph. This method was developed by Balogh, Morris, and Samotij [4] and, independently, Saxton and Thomason [73]. A more in-depth history of this method was discussed in Section 1.3.1.

Broadly speaking, the hypergraph container method [4, 73] works by finding a "relatively small" collection of "relatively small" subsets of the vertex set that are chosen to be "close" to independent, called *containers*. The collection of containers is formed such that every independent set is a subset of at least one container. We may then upper bound the number of independent sets by counting the total number of subsets of the containers and applying a union bound. We can, of course, consider our set of containers to be all maximal independent sets and simply count all subsets to achieve an upper bound on the number of independent sets. However, hypergraphs can contain many maximal independent sets meaning the upper bound obtained through this method may be essentially trivial. So, we instead allow containers to have edges such that a container may cover multiple maximal independent sets. We must be careful though as if we allow the containers to have too many edges, then many of the subsets of our containers will not be independent sets. This would result in a poor upper bound as we will count many non-independent sets in our estimation. The hypergraph container method finds a small set of nearly independent containers that covers all independent sets. This is possible as in hypergraphs with relatively even edge distribution, independent sets tend to "cluster" together allowing the containers to align with these clusters.

We now give an exposition of the original Kleitman and Winston [43] algorithm that bounds the number of independent sets in a graph. This algorithm was implicit in the work of Kleitman and Winston. The formulation offered in this thesis was presented by Kohayakawa, Lee, Rödl, and Samotij [44].

This algorithm of Kleitman and Winston is a predecessor to the hypergraph container method of Balogh, Morris, and Samotij [4] and Saxton and Thomason [73]. It is included here to build intuition as to how the hypergraph container method works. We have chosen to include the graph container algorithm rather than the hypergraph container algorithm as the added complexity of working in the hypergraph setting does not yield much of a deeper understanding. The algorithm will use a *max-degree ordering* on a graph G. This is an ordering $\{v_1, v_2, \ldots, v_{|V(G)|}\}$ of $V(G)$ such that v_i is a maximum degree vertex in the graph $G \setminus \{v_1, \ldots, v_{i-1}\}$. Further, if $A \subseteq V(G)$ then $G[A]$ denotes the induced subgraph of G with vertex set A. Finally, for a set $A \subseteq V(G)$ we will use $N_G(A)$ to represent the set of vertices in G adjacent to at least one vertex in A.

Theorem 4.1.1 ([43], see also [44])**.** *Let G be a graph on N vertices. Take $R, q \in \mathbb{N}$ and $0 < \beta < 1$. Suppose that*

$$e^{-\beta q} N \leq R$$

and for every subset $S \subseteq V(G)$ where $|S| \geq R$, we have

$$|E(G[S])| \geq \beta \binom{|S|}{2}.$$

Then, the number of independent sets of size $m \geq q$ is at most

$$\binom{N}{q} \binom{R}{m-q}.$$

Proof. If no independent set of order at least q exists, then the bound is trivially true as the number of independent sets of size $m \geq q$ is 0. Therefore, we may assume such an independent set exists. We will show there exists a collection of sets $C_1, \ldots, C_{\binom{N}{q}}$

such that $|C_i| \leq R + q$ for all $1 \leq i \leq \binom{N}{q}$, where for each independent set $I \subseteq V(G)$ such that $|I| \geq q$ there exists some i such that $I \subseteq C_i$.

Define I to be an arbitrary choice of independent set such that $|I| \geq q$. We run the following algorithm q times. We initialize by setting $A_0 = V(G)$ and $S_0 = \emptyset$. At the i-th step where $1 \leq i \leq q$ we do the following:

Step 1. Label the vertices in $A_{i-1} = \{v_1^{i-1}, v_2^{i-1}, \dots v_{|A_{i-1}|}^{i-1}\}$ such that the labelling corresponds to a maximum-degree ordering on $G[A_{i-1}]$. Define $t_i = \min\{j : 1 \leq j \leq |A_{i-1}|, v_j^{i-1} \in I\}$. Therefore, v_{t_i} is the lowest index vertex in A_{i-1} that is also in I. Define $S_i = S_{i-1} \cup \{v_{t_i}^{i-1}\}$. By definition, $S_i \subseteq I$.

Step 2. Define $A_i = A_{i-1} \setminus (\{v_1^{i-1}, \dots, v_{t_i}^{i-1}\} \cup N_{G[A_{i-1}]}(v_{t_i}^{i-1}))$. That is, A_i is obtained from A_{i-1} by removing all vertices earlier in the max-degree ordering than v_{t_i} as well as all neighbours of v_{t_i}. So, as $v_{t_i} \in I$ and I is an independent set, all vertices of A_{i-1} removed to form A_{i-1} are not contained in I. So, $I \subseteq A_i \cup S_i$. So, we either have that $S_i = I$ and so we have iterated q times and are done, or there are vertices in I still remaining in A_i.

For simplicity at the end of this sequence of q steps, we define $A = A_q$ and $S = S_q$.

We will now show that $|A| \leq R$. Suppose for a contradiction that $|A| > R$. Then, as at each step i we remove vertices from A_{i-1} to form A_i, it must follow that $|A_i| > R$ for all $0 \leq i \leq q$. Consider an arbitrary choice of i such that $0 < i \leq q$. Take the maximum-degree ordering on A_{i-1} where $A_{i-1} = \{v_1^{i-1}, v_2^{i-1}, \dots v_{|A_{i-1}|}^{i-1}\}$, and recall that $v_{t_i}^{i-1}$ is the minimum-indexed vertex in this ordering such that $v_{t_i}^{i-1} \in I$. Now, recall that to form A_i we will remove from A_{i-1} all vertices of lower index than $v_{t_i}^{i-1}$ and all neighbours of $v_{t_i}^{i-1}$. We will lower-bound the number of vertices removed from A_{i-1} to obtain our contradiction.

Define $A' = A_{i-1} \setminus \{v_1^{i-1}, \dots, v_{t_i-1}^{i-1}\}$. Then, $V(A')$ contains $v_{t_i}^{i-1}$ and the set of all vertices with a higher index than t_i under the maximum-degree ordering. Notice that $v_{t_i}^{i-1}$ is, by definition, the vertex of maximum degree in $G[A']$. As $v_{t_i}^{i-1}$ is of maximum degree, its degree must be at least the average degree in $G[A']$. As $|A'| > R$, it then follows by supposition that

$$deg_{G[A']}(v_{t_i}^{i-1}) \geq \frac{2|E(G[A'])|}{|A'|} \geq \frac{2\beta\binom{|A'|}{2}}{|A'|} = \beta(|A'| - 1).$$

Therefore, to form A_i we remove from A_{i-1} at least

$$t_i + \beta(|A'| - 1) = t_i + \beta(|A_{i-1}| - 1 - t_i + 1) = t_i + \beta(|A_{i-1}| - t_i) > \beta(|A_{i-1}|),$$

vertices, where the last inequality holds as $0 < \beta < 1$. Therefore,

$$|A_i| \leq |A_{i-1}| - \beta(|A_{i-1}|) = (1-\beta)|A_{i-1}|.$$

This implies that that at each step i, $|A_i|$ is smaller than $|A_{i-1}|$ by a factor of at least $(1-\beta)$. Hence, recalling $(1-x) \leq e^{-x}$ for all $x \geq 0$, we have

$$|A| \leq (1-\beta)^q N < e^{-\beta q} N \leq R,$$

a contradiction. Therefore, we know that $|A| \leq R$.

We now make two important observations to conclude our proof. Firstly, recall $S \subseteq I$ and $I \setminus S \subseteq A$. Therefore, setting $C = S \cup A$ we have that $I \subseteq C$. Next, we notice that the final set A depends only on S. Therefore, if we know S we can recover A by running the algorithm. That is, we know the ordering to remove vertices from A given that we know S. Further, each independent set I is associated to some set S by our algorithm. By considering all possible sets S_i and their associated A_i, where $S_i \subseteq V(G)$, $|S_i| = q$, and $1 \leq i \leq \binom{N}{q}$, we can form containers $C_i = S_i \cup A_i$. As we showed $|A_i| \leq R$ and $|S_i| = q$ by definition, we have that $|C_i| \leq R + q$. Further, we are guaranteed that each independent set is contained in at least one C_i by our observations. Therefore, we need at most $\binom{N}{q}$ containers to cover all independent sets.

We are now able to obtain an upper bound on the number of independent sets of cardinality m for any $m \geq q$. Associate each independent set I of cardinality m with the container C_i that was obtained via the pair of sets S_i, A_i outputted by the algorithm when I is the input, recalling that $S_i \subseteq I \subseteq S_i \cup A_i = C_i$. Therefore, the number of independent sets associated with C_i of cardinality at m is at most the number of possible subsets of A_i of cardinality $m - q$. As $|A_i| \leq R$, we have that the number of independent sets within C_i is at most $\binom{R}{m-q}$. As the number of containers is $\binom{N}{q}$, we have that the number of independent sets of cardinality m is at most

$$\binom{N}{q}\binom{R}{m-q}.$$

□

We now give some definitions in order to state a hypergraph container lemma of Saxton and Thomason [73]. Consider an r-uniform hypergraph $\mathcal{H} = (V(\mathcal{H}), E(\mathcal{H}))$, where $V(\mathcal{H})$ is the set of vertices and $E(\mathcal{H})$ is the set of hyperedges. We use $\bar{d}(\mathcal{H})$ to denote the average degree of a vertex in $\mathcal{H}$; by a handshaking argument, $\bar{d}(\mathcal{H}) = r|E(\mathcal{H})|/|V(\mathcal{H})|$. For a set $A \subseteq V(\mathcal{H})$, we use $e_{\mathcal{H}}(A)$ to denote the number of hyperedges contained in A. The *co-degree* of a set $A \subseteq V(\mathcal{H})$ is the number of hyperedges of $\mathcal{H}$ that contain A. We use $d_{\mathcal{H}}(A)$, or simply $d(A)$ when the subscript is obvious from context, to denote the co-degree of a set A. We denote the maximum co-degree of all sets of size k in $\mathcal{H}$

as $\Delta_k(\mathcal{H})$; explicitly

$$\Delta_k(\mathcal{H}) := \max\{d_{\mathcal{H}}(A) : |A| = k\}.$$

Further, we define $\Delta(\mathcal{H}, \tau)$ for any $\tau > 0$ as

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

The function $\Delta(\mathcal{H}, \tau)$ is often referred to as the *co-degree function.* Notice that the co-degree function depends on all maximum co-degrees in $\mathcal{H}$, except for $\Delta_1(\mathcal{H})$. In the context of the hypergraph container method, the co-degree function will act as a parameter which depends on the way the edges of $\mathcal{H}$ overlap.

Theorem 4.1.2 (Hypergraph Container Lemma [73])**.** *Let $\mathcal{H}$ be an r-uniform hypergraph. Suppose that $0 < \varepsilon, \tau < 1/2$ satisfy:*

- $\tau < 1/(200 \cdot r \cdot r!^2)$*, and*
- $\Delta(\mathcal{H}, \tau) \leq \varepsilon/(12r!)$.

Then there exists $c = c(r) \leq 1000 \cdot r \cdot r!^3$ and a collection $\mathcal{C}$ of subsets of $V(\mathcal{H})$ such that the following hold:

- *for every independent set I in $\mathcal{H}$, there exists $A \in \mathcal{C}$ such that $I \subseteq A$,*
- $\log|\mathcal{C}| \leq c \cdot |V(\mathcal{H})| \cdot \tau \cdot \log(1/\varepsilon) \cdot \log(1/\tau)$*, and*
- $e_{\mathcal{H}}(A) \leq \varepsilon \cdot |E(\mathcal{H})|$ *for every $A \in \mathcal{C}$.*

Recall in the proof of Theorem 4.1.1, the set S is called the *fingerprint* of the corresponding independent set I. In applications of (hyper)graph containers, a fingerprint is a carefully chosen, small subset of the independent set that corresponds to a container, C. The container C is guaranteed to contain all independent set with fingerprint S. As $|S|$ is small, there is a small number of possible fingerprints S to consider and so, a small number of containers to consider. A similar method is used to prove Theorem 4.1.2. The algorithm described by Saxton and Thomason is derived from similar ideas to those of Sapozhenko [68–72] for regular graphs (i.e. regular 2-uniform hypergraphs) but generalized to work in the r-uniform hypergraph setting for any r. Broadly speaking, the algorithm has two "modes": prune mode, which takes an independent set as an input and outputs its fingerprint; and build mode, which takes a fingerprint as an input and outputs a container. This prune and build mode are in essence the same as the algorithm presented in the proof of Theorem 4.1.1. As in the proof of Theorem 4.1.1, Saxton and Thomason [73] show a vertex is always a subset of the container associated with its fingerprint. Furthermore, they show that each fingerprint (and so the number of containers) is small, the order of each container is bounded away from the order of the hypergraph, and the number of edges in each container is bounded.

Since the development of the hypergraph container method, it has been used to prove a myriad of results in various areas of combinatorics and other fields of mathematics, see the survey [5]. As expected, an application of the hypergraph container method usually requires viewing the object to be enumerated as independent sets in a hypergraph. For example, consider again the hypergraph $\mathcal{H}$ that encodes copies of K_3 as edges. This embedding of triangle-free subgraphs of K_n was used by Balogh, Morris, and Samotij [4] and, independently, Saxton and Thomason [73] to apply the hypergraph container method and re-prove that the number triangle-free graphs on n vertices is $2^{(1+o(1))\frac{n^2}{4}}$. The number of triangle-free graphs was first obtained by Erdős, Kleitman, and Rothschild [22].

In applications of Theorem 4.1.2, the choice of ε and τ largely impact the strength of the result. First, consider the parameter ε which gives an upper bound on the number of edges in each container. In order to relate the number of edges to the order of each container, one typically wishes to apply a supersaturation result, see Chapter 3. In order to make our estimation accurate, we want to choose ε as small as possible to ensure most of the subsets of a container are independent. Therefore, a stronger supersaturation result leads to a better estimation. In this way, our ability to choose ε is often determined completely by the related supersaturation problem. Returning to the problem of enumerating the number of triangle-free graphs, the choice of ε was dictated by the following supersaturation lemma of Erdős and Simonovits [23]:

Lemma 4.1.3 ([23])**.** *For every $\delta > 0$, there exists $\varepsilon(\delta) > 0$ such that if G is a graph on n vertices with at least $(\frac{1}{4} + \delta) \cdot n^2$ edges, then it has at least εn^3 triangles.*

As Theorem 4.1.2 bounds the number of edges in each container, in this setting it tells us the maximum number of triangles in each subgraph corresponding to a container. If the number of triangles in each container is at most εn^3, we may apply the contrapositive of Lemma 4.1.3 to obtain the maximum number of edges in the subgraph of K_n corresponding to each container. Knowing the maximum number of edges in a subgraph H_C corresponding to a container C tells us the maximum number of subgraphs of H_C. Therefore, by a simple union bound we may obtain an upper bound on the number of triangle-free graphs.

We next consider the parameter τ in Theorem 4.1.2. The necessary conditions give an upper and lower bound on τ. The number of containers largely depends on τ. So, the lower bound implies that the more the edges overlap, the more containers we require. That is, τ is chosen large enough to ensure that $\Delta(\mathcal{H}, \tau)$ is small. Intuitively, it makes sense that the more edges overlap the more containers we require as then adding a vertex to a container may force many edges to the container. So, our containers must be "smaller" to be close to independent, which implies we require more of them.

4.2 Proof of an Approximate Multidimensional Szemerédi's Theorem

We will adapt Theorem 4.1.2 to prove Theorem 1.3.2. In order to do so, we restate our problem in terms of independent sets in a hypergraph. Let us consider the $|X|$-uniform hypergraph $\mathcal{G}$ with $V(\mathcal{G}) = [n]^d$ and hyperedge set consisting of all copies of X in $[n]^d$. Clearly independent sets in $\mathcal{G}$ are precisely X-free subsets of $[n]^d$. Therefore, an approximation on the number of independent sets in $\mathcal{G}$ would give us an approximation on the number of X-free subsets of $[n]^d$. Applying Theorem 4.1.2 to this hypergraph $\mathcal{G}$, we immediately get the following corollary.

Corollary 4.2.1. *For a finite set $X \subseteq \mathbb{N}^d$, let $\mathcal{G}$ be the $|X|$-uniform hypergraph on n^d vertices encoding the set of all copies of X in $[n]^d$. Suppose that there exist $0 < \varepsilon$, $\tau < 1/2$ such that*

- $\tau < 1/(200 \cdot |X| \cdot |X|!^2)$
- $\Delta(\mathcal{G}, \tau) \leq \varepsilon/(12|X|!)$.

Then there exists $c = c(|X|) \leq 1000 \cdot |X| \cdot |X|!^3$ and a collection $\mathcal{C}$ of subsets of $V(\mathcal{G})$ such that the following holds:

- *for every X-free subset $I \subseteq [n]^d$, there exists $A \in \mathcal{C}$ such that $I \subseteq A$,*
- $\log |\mathcal{C}| \leq c \cdot n^d \cdot \tau \cdot \log(1/\varepsilon) \cdot \log(1/\tau)$,
- $e_{\mathcal{G}}(A) \leq \varepsilon \cdot |E(\mathcal{G})|$ *for every $A \in \mathcal{C}$.*

We now use Corollary 4.2.1 to prove Theorem 1.3.2, restated below for convenience. We previously noted that in an application of the hypergraph container method, our choice of ε is often dictated by a supersaturation result. We will choose our ε based on Lemma 3.3.2, which will subsequently allow us to obtain an upper bound on the order of each container. We will choose τ such that it satisfies the conditions. As we are not concerned with exact constants, we do not necessarily pick τ to be optimal.

Theorem 1.3.2. *Let d be a positive integer and let $X \subseteq \mathbb{N}^d$ be a finite set such that $|X| \geq 3$. For infinitely many $n \in \mathbb{N}$, the number of X-free subsets of $[n]^d$ is $2^{O(r_X(n))}$.*

Proof. Define N to be the infinite subset of $\mathbb{N}$ guaranteed by Lemma 3.3.2. Consider the $|X|$-uniform hypergraph $\mathcal{G}$ encoding copies of X in $[n]^d$ and choose $\gamma > 0$ so that $\bar{d}(\mathcal{G}) = \frac{|X||E(\mathcal{G})|}{n^d} \geq \gamma n$ for all n sufficiently large. For $n \in N$, define

$$\varepsilon(n) := \frac{1}{6d} \cdot \frac{\log^{3|X|-2}(n)}{n} \cdot \left(\frac{n^d}{r_X(n)}\right)^{|X|},$$

and

$$\tau(n) := \left(\frac{12|X|! \cdot 2^{|X|^2} \cdot |X|^3}{\gamma n \cdot \varepsilon(n)} \right)^{\frac{1}{|X|-1}}.$$

We wish to apply Corollary 4.2.1, so we will verify the conditions for $\mathcal{G}$ with $\varepsilon(n)$, and $\tau(n)$. Clearly, $\varepsilon(n), \tau(n) > 0$ for all n. Using Lemma 2.3.2, observe for sufficiently large n,

$$\begin{aligned}
\varepsilon(n) &= \frac{1}{6d} \cdot \frac{\log^{3|X|-2}(n)}{n} \cdot \left(\frac{n^d}{r_X(n)} \right)^{|X|} \\
&\leq \frac{1}{6d} \cdot \frac{\log^{3|X|-2}(n)}{n} \cdot \left(\frac{n^d}{n^d \cdot \exp(-c_X \sqrt{\log(n)})} \right)^{|X|} \\
&= \frac{1}{n^{1-o(1)}} < \frac{1}{2}. \qquad (4.1)
\end{aligned}$$

Also, we clearly have $\frac{1}{\varepsilon(n)} = o(n)$ and so $\tau(n) = o(1)$. Thus, for n sufficiently large, we have

$$\tau(n) < 1/(200 \cdot |X| \cdot |X|!^2) < \frac{1}{2}. \qquad (4.2)$$

Next, we wish to bound $\Delta(\mathcal{G}, \tau(n))$. By definition,

$$\Delta(\mathcal{G}, \tau(n)) = 2^{\binom{|X|}{2}-1} \sum_{j=2}^{|X|} 2^{-\binom{j-1}{2}} \frac{\Delta_j(\mathcal{G})}{\tau(n)^{j-1} \bar{d}(\mathcal{G})}.$$

For any $2 \leq j \leq |X|$, we have $\Delta_j(\mathcal{G}) \leq \Delta_2(\mathcal{G}) \leq \binom{|X|}{2}$. So, every term in the above summation is at most

$$\frac{\Delta_2(\mathcal{G})}{\tau(n)^{|X|-1} \bar{d}(\mathcal{G})} \leq \frac{|X|^2}{\tau(n)^{|X|-1} \gamma n}$$

and

$$\Delta(\mathcal{G}, \tau(n)) \leq \frac{2^{|X|^2} \cdot |X|^3}{\tau(n)^{|X|-1} \gamma n} = \frac{\varepsilon(n)}{12|X|!} \qquad (4.3)$$

where the equality is by definition of $\tau(n)$.

From (4.1), (4.2) and (4.3), we may apply Corollary 4.2.1 with $\varepsilon = \varepsilon(n)$ and $\tau = \tau(n)$ to get a constant $c = c(X) \leq 1000 \cdot |X| \cdot |X|!^3$ and a collection $\mathcal{C}$ of subsets of $V(\mathcal{G})$ with the following properties:

- for every X-free subset $I \subseteq [n]^d$, there exists $S \in \mathcal{C}$ such that $I \subseteq S$,
- $\log |\mathcal{C}| \leq c \cdot n^d \cdot \tau(n) \cdot \log(1/\varepsilon(n)) \cdot \log(1/\tau(n))$,
- for every $A \in \mathcal{C}$, $e_{\mathcal{G}}(A) \leq \varepsilon(n) \cdot |E(\mathcal{G})|$.

By definition, $\log(1/\varepsilon(n)) = O(\log(n))$ and $\log(1/\tau(n)) = O(\log(n))$. We also have

$\frac{n^d \log^3(n)}{r_X(n)} = o\left((n \cdot \varepsilon(n))^{\frac{1}{|X|-1}}\right)$ which implies that $\tau(n) = o\left(\frac{r_X(n)}{n^d \log^3(n)}\right)$. Therefore,

$$\log |\mathcal{C}| = O\left(n^d \tau(n) \log^2(n)\right) = o(r_X(n)). \tag{4.4}$$

Note that for every container $A \in \mathcal{C}$, the number of copies of X in A is at most $\varepsilon(n) \cdot |E(\mathcal{G})| < \varepsilon(n) \cdot n^{d+1}$. That is,

$$\Gamma_X(A) \leq \varepsilon(n) \cdot n^{d+1} = \frac{\log^{3|X|-2}(n)}{6d} \cdot \left(\frac{n^d}{r_X(n)}\right)^{|X|} \cdot n^d.$$

So, by Lemma 3.3.2, every container $A \in \mathcal{C}$ must satisfy

$$|A| < \frac{4}{\alpha} \cdot r_X(n) \tag{4.5}$$

where α is a constant depending only on X. Since every X-free subset of $[n]^d$ is contained in some container in $\mathcal{C}$, we can use (4.4) and (4.5) to conclude that the number of X-free subsets of $[n]^d$ is at most

$$\sum_{A \in \mathcal{C}} 2^{|A|} \leq |\mathcal{C}| \cdot \max_{A \in \mathcal{C}} 2^{|A|} \overset{(4.4),(4.5)}{<} 2^{o(r_X(n))} \cdot 2^{\frac{4}{\alpha} \cdot r_X(n)} = 2^{O(r_X(n))},$$

as desired.

□

Chapter 5
Conclusion

In this chapter we will discuss some natural questions regarding the results of this thesis and offer some open problems to the interested reader. Recall the main result of Chapter 4, restated below for convenience.

Theorem 1.3.2. *Let d be a positive integer and let $X \subseteq \mathbb{N}^d$ be a finite set such that $|X| \geq 3$. For infinitely many $n \in \mathbb{N}$, the number of X-free subsets of $[n]^d$ is $2^{O(r_X(n))}$.*

The most obvious question following Theorem 1.3.2 is: can this result be extended for all n? The most obvious obstacle in extending Theorem 1.3.2 is extending Lemma 3.3.2, restated below.

Lemma 3.3.2. *There exists an infinite subset N of $\mathbb{N}$ such that, if $n \in N$ and $A \subseteq [n]^d$ where $|A| \geq \frac{4}{\alpha} \cdot r_X(n)$, then*

$$\Gamma_X(A) \geq \frac{\log^{3|X|-2}(n)}{6d} \cdot \left(\frac{n^d}{r_X(n)}\right)^{|X|} \cdot n^d.$$

As discussed in Chapter 3, the reason we must consider only an infinite subset of $\mathbb{N}$ rather than all of $\mathbb{N}$ is because of the unknown behaviour of $r_X(n)$. It does not seem as though there would be an obvious way to extend Lemma 3.3.2 or Theorem 1.3.2 without either a major breakthrough in the understanding of $r_X(n)$, or a different approach entirely.

It's also worth noting that the asymptotics of $r_X(n)$ should depend on the structure of X itself, so finding asymptotics for $r_X(n)$ for general X appears to be a complex and extremely difficult problem. It does not seem inconceivable, though perhaps unlikely, that a stronger understanding of $r_k(n)$, or even $r_3(n)$, could be enough to extend these results using something similar to Lemma 2.3.2 to obtain a lower bound that implies the rate in which the density of $r_X(n)$ decreases is limited. So, research improving upon $r_X(n)$, for any X, is of interest.

We also note that there is no guarantee that an answer to Question 1.3.1 should guarantee that the number of X-free sets for general X behaves the same. Consider the similar question in the graph setting: how many graphs of order n exist that do not contain some forbidden subgraph? Let $\mathrm{ex}(n, H)$ be the maximum size of an H-free graph with order n. It was shown by Erdős, Frankl, and Rödl [20] that if the forbidden subgraph H is non-bipartite, then the number of H-free graphs is $2^{(1+o(1))\mathrm{ex}(n,H)}$. But the number of C_6-free graphs is was shown by Morris and Saxton [49] to be at least $2^{(1+c)\mathrm{ex}(n,C_6)}$ for some $c > 0$. Therefore, it is possible that there are some choices of X such that the number of X-free subsets is $2^{(1+o(1))r_X(n)}$ and other choices of X where the number of X-free subsets is asymptotically greater than this.

5.1 Density Hales-Jewett Theorem

Theorem 1.3.2 is a generalization of a result of Balogh, Liu, and Sharifzadeh [3] in the same way that the multidimensional Szemerédi's Theorem is a generalization of Szemerédi's Theorem. In this way, a logical follow-up to Theorem 1.3.2 would be counting versions of other generalizations of Szemerédi's Theorem.

We will consider the density Hales-Jewett theorem, a generalization of Szemerédi's Theorem. The density Hales-Jewett theorem implies both Szemerédi's Therorem and the multidimensional Szemerédi's Theorem. The density Hales-Jewett Theorem was first proved by Furstenberg and Katznelson [31] in 1991 using ergodic theory. A combinatorial version was first proved by Polymath [56] in 2012. Interestingly, the Polymath proof was birthed in an unusual way; it was a collaborative effort of several researchers who communicated via comments on a blog post of Timothy Gowers. The name "Polymath" is a pseudonym as the work was "open source". For more details about the Polymath project, see the discussions in [36, 52, 56].

Before stating the density Hales-Jewett theorem, we require a definition. A set $L \subseteq [n]^k$ is a *combinatorial line* if there exists a non-empty set $W \subseteq [n]$ and integers $x_i \in [n]$ for all $i \notin W$ such that

$$L = \{(y_1, y_2, \dots, y_k) : y_i = y_j \text{ for all } i, j \in W, y_i = x_i \text{ for all } i \notin W\}.$$

In other words, the points of L have the same entry in all coordinates not in W and a point in L has the same value in every coordinate in W. For example, the following is a combinatorial line where $n = 4$, $k = 5$, and $W = \{4, 5\}$

$$\{(1, 2, 3, 1, 1), (1, 2, 3, 2, 2), (1, 2, 3, 3, 3), (1, 2, 3, 4, 4)\}.$$

Similar to how Szemerédi's Theorem may be thought of as a density version of van der Waerden's theorem, the density Hales-Jewett theorem is, unsurprisingly, a density version of Hales-Jewett theorem. For completeness, the Hales-Jewett theorem is included below.

Theorem 5.1.1. *For all positive integers r and n, there exists an integer $m(r, n)$ such that for all $k \geq m(r, n)$ any r-colouring of $[n]^k$ contains a monochromatic line.*

We now state the density Hales-Jewett theorem.

Theorem 5.1.2. *For every positive integer n and every real number $\delta > 0$, there exists a positive integer $M(n, \delta)$ such that if $k \geq M(k, \delta)$ and A is a subset of $[n]^k$ such that $\frac{|A|}{n^k} \geq \delta$, then A contains a combinatorial line.*

So, similar to the focus of this thesis, we may ask:

1. How large can a subset of $[n]^k$ not containing a combinatorial line be?
2. How many combinatorial lines must exist in a dense subset of $[n]^k$?
3. How many subsets of $[n]^k$ exist without a combinatorial line?

In particular, one may look to answer Question 3 above by generalizing Theorem 1.3.2 to this broader setting using answers to Question 1 and Question 2, as we did in this thesis.

Bibliography

[1] J. Balogh, S. Das, M. Delcourt, H. Liu, and M. Sharifzadeh. Intersecting families of discrete structures are typically trivial. *J. Combin. Theory Ser. A*, 132:224–245, 2015.

[2] J. Balogh, R. I. Garcia, L. Li, and A. Z. Wagner. Intersecting families of sets are typically trivial. *J. Combin. Theory Ser. B*, 164:44–67, 2024.

[3] J. Balogh, H. Liu, and M. Sharifzadeh. The number of subsets of integers with no k-term arithmetic progression. *Int. Math. Res. Not. IMRN*, (20):6168–6186, 2017.

[4] J. Balogh, R. Morris, and W. Samotij. Independent sets in hypergraphs. *J. Amer. Math. Soc.*, 28(3):669–709, 2015.

[5] J. Balogh, R. Morris, and W. Samotij. The method of hypergraph containers. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. IV. Invited lectures*, pages 3059–3092. World Sci. Publ., Hackensack, NJ, 2018.

[6] N. Behague, J. Hyde, N. Morrison, J. A. Noel, and A. Wright. An approximate counting version of the multidimensional Szemerédi theorem. 2023. E-print arXiv:2311.13709.

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

[8] T. F. Bloom. A quantitative improvement for Roth's theorem on arithmetic progressions. *J. Lond. Math. Soc. (2)*, 93(3):643–663, 2016.

[9] T. F. Bloom and O. Sisask. Logarithmic bounds for Roth's theorem via almost-periodicity. *Discrete Anal.*, pages Paper No. 4, 20, 2019.

[10] T. F. Bloom and O. Sisask. Breaking the logarithmic barrier in Roth's theorem on arithmetic progressions. 2021. E-print arXiv:2007.03528v2.

[11] T. F. Bloom and O. Sisask. The Kelley–Meka bounds for sets free of three-term arithmetic progressions. *Essential Number Theory*, 2(1):15–44, Dec. 2023.

[12] P. J. Cameron and P. Erdős. On the number of sets of integers with various properties. In *Number theory (Banff, AB, 1988)*, pages 61–79. de Gruyter, Berlin, 1990.

[13] D. Conlon and W. T. Gowers. Combinatorial theorems in sparse random sets. *Ann. of Math. (2)*, 184(2):367–454, 2016.

[14] E. Croot and O. Sisask. A new proof of Roth's theorem on arithmetic progressions. *Proc. Amer. Math. Soc.*, 137(3):805–809, 2009.

[15] S. Das, W. Gan, and B. Sudakov. The minimum number of disjoint pairs in set systems and related problems. *Combinatorica*, 36(6):623–660, 2016.

[16] D. Dong, N. Mani, and Y. Zhao. On the number of error correcting codes. *Combin. Probab. Comput.*, 32(5):819–832, 2023.

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

[18] P. Erdős. On a theorem of Rademacher-Turán. *Illinois J. Math.*, 6:122–127, 1962.

[19] P. Erdős. On the number of complete subgraphs contained in certain graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 7:459–464, 1962.

[20] P. Erdős, P. Frankl, and V. Rödl. The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent. *Graphs Combin.*, 2(2):113–121, 1986.

[21] P. Erdős, D. J. Kleitman, and B. L. Rothschild. Asymptotic enumeration of K_n-free graphs. In *Colloquio Internazionale sulle Teorie Combinatorie (Roma, 1973), Tomo II*, pages 19–27. Accad. Naz. Lincei, Rome, 1976.

[22] P. Erdős, D. J. Kleitman, and B. L. Rothschild. Asymptotic enumeration of K_n-free graphs. In *Colloquio Internazionale sulle Teorie Combinatorie (Roma, 1973), Tomo II*, pages 19–27. Accad. Naz. Lincei, Rome, 1976.

[23] P. Erdős and M. Simonovits. Supersaturated graphs and hypergraphs. *Combinatorica*, 3(2):181–192, 1983.

[24] P. Erdős and P. Turán. On Some Sequences of Integers. *J. London Math. Soc.*, 11(4):261–264, 1936.

[25] D. C. Fisher. Lower bounds on the number of triangles in a graph. *J. Graph Theory*, 13(4):505–512, 1989.

[26] P. Frankl and V. Rödl. Extremal problems on set systems. *Random Structures Algorithms*, 20(2):131–164, 2002.

[27] E. Friedgut, V. Rödl, and M. Schacht. Ramsey properties of random discrete structures. *Random Structures Algorithms*, 37(4):407–436, 2010.

[28] H. Furstenberg. Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J. Analyse Math.*, 31:204–256, 1977.

[29] H. Furstenberg and Y. Katznelson. An ergodic Szemerédi theorem for commuting transformations. *J. Analyse Math.*, 34:275–291 (1979), 1978.

[30] H. Furstenberg and Y. Katznelson. A density version of the Hales-Jewett theorem. *J. Anal. Math.*, 57:64–119, 1991.

[31] H. Furstenberg and Y. Katznelson. A density version of the Hales-Jewett theorem. *J. Anal. Math.*, 57:64–119, 1991.

[32] W. Gasarch, J. Glenn, and C. P. Kruskal. Finding large 3-free sets. I. The small n case. *J. Comput. System Sci.*, 74(4):628–655, 2008.

[33] A. W. Goodman. On sets of acquaintances and strangers at any party. *Amer. Math. Monthly*, 66:778–783, 1959.

[34] W. T. Gowers. A new proof of Szemerédi's theorem. *Geom. Funct. Anal.*, 11(3):465–588, 2001.

[35] W. T. Gowers. Hypergraph regularity and the multidimensional Szemerédi theorem. *Ann. of Math. (2)*, 166(3):897–946, 2007.

[36] W. T. Gowers. Polymath and the density Hales-Jewett theorem. In *An irregular mind*, volume 21 of *Bolyai Soc. Math. Stud.*, pages 659–687. János Bolyai Math. Soc., Budapest, 2010.

[37] B. Green and I. Z. Ruzsa. Sum-free sets in abelian groups. *Israel J. Math.*, 147:157–188, 2005.

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

[39] Z. Hunter. New lower bounds for $r_3(n)$. 2024. E-print arXiv:2401.16106v2.

[40] Z. Kelley and R. Meka. Strong bounds for 3-progressions. E-print arXiv:2302.05537v4, 2023.

[41] Y. Kim. The number of k-dimensional corner-free subsets of grids. *Electron. J. Combin.*, 29(2):Paper No. 2.53, 19, 2022.

[42] D. Kleitman. On Dedekind's problem: The number of monotone Boolean functions. *Proc. Amer. Math. Soc.*, 21:677–682, 1969.

[43] D. J. Kleitman and K. J. Winston. On the number of graphs without 4-cycles. *Discrete Math.*, 41(2):167–172, 1982.

[44] Y. Kohayakawa, S. J. 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 Algorithms*, 46(1):1–25, 2015.

[45] H. Liu, O. Pikhurko, and K. Staden. The exact minimum number of triangles in graphs with given order and size. *Forum Math. Pi*, 8:e8, 144, 2020.

[46] L. Lovász and M. Simonovits. On the number of complete subgraphs of a graph. II. In *Studies in pure mathematics*, pages 459–495. Birkhäuser, Basel, 1983.

[47] W. Mantel. Problem 28. *Wiskundige Opgaven*, 10, 1907.

[48] A. Mąkowski. Remark on a paper of Erdős and Turán. *J. London Math. Soc.*, 34:480, 1959.

[49] R. Morris and D. Saxton. The number of $C_{2\ell}$-free graphs. *Adv. Math.*, 298:534–580, 2016.

[50] L. Moser. On non-averaging sets of integers. *Canad. J. Math.*, 5:245–252, 1953.

[51] B. Nagle, V. Rödl, and M. Schacht. The counting lemma for regular k-uniform hypergraphs. *Random Structures Algorithms*, 28(2):113–179, 2006.

[52] M. A. Nielsen. Introduction to the Polymath project and "Density Hales-Jewett and Moser numbers". In *An irregular mind*, volume 21 of *Bolyai Soc. Math. Stud.*, pages 651–657. János Bolyai Math. Soc., Budapest, 2010.

[53] V. Nikiforov. The number of cliques in graphs of given order and size. *Trans. Amer. Math. Soc.*, 363(3):1599–1618, 2011.

[54] V. S. Nikiforov. On a problem of P. Erdős. *Annuaire Univ. Sofia Fac. Math. Méc.*, 71(2):157–160, 1976/77.

[55] K. O'Bryant. Sets of integers that do not contain long arithmetic progressions. *Electron. J. Combin.*, 18(1):Paper 59, 15, 2011.

[56] D. H. J. Polymath. A new proof of the density Hales-Jewett theorem. *Ann. of Math. (2)*, 175(3):1283–1327, 2012.

[57] R. A. Rankin. Sets of integers containing not more than a given number of terms in arithmetical progression. *Proc. Roy. Soc. Edinburgh Sect. A*, 65:332–344 (1960/61), 1960/61.

[58] A. A. Razborov. Flag algebras. *J. Symbolic Logic*, 72(4):1239–1282, 2007.

[59] A. A. Razborov. On the minimal density of triangles in graphs. *Combin. Probab. Comput.*, 17(4):603–618, 2008.

[60] C. Reiher. The clique density theorem. *Ann. of Math. (2)*, 184(3):683–707, 2016.

[61] V. Rödl and J. Skokan. Regularity lemma for k-uniform hypergraphs. *Random Structures Algorithms*, 25(1):1–42, 2004.

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

[63] I. Z. Ruzsa and E. Szemerédi. Triple systems with no six points carrying three triangles. In *Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II*, volume 18 of *Colloq. Math. Soc. János Bolyai*, pages 939–945. North-Holland, Amsterdam-New York, 1978.

[64] I. Z. Ruzsa and E. Szemerédi. Triple systems with no six points carrying three triangles. In *Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II*, volume 18 of *Colloq. Math. Soc. János Bolyai*, pages 939–945. North-Holland, Amsterdam-New York, 1978.

[65] R. Salem and D. C. Spencer. On sets of integers which contain no three terms in arithmetical progression. *Proc. Nat. Acad. Sci. U.S.A.*, 28:561–563, 1942.

[66] W. Samotij and B. Sudakov. The number of additive triples in subsets of abelian groups. *Mathematical Proceedings of the Cambridge Philosophical Society*, 160(3):495–512, 2016.

[67] T. Sanders. On Roth's theorem on progressions. *Ann. of Math. (2)*, 174(1):619–636, 2011.

[68] A. A. Sapozhenko. On the number of independent sets in extenders. *Diskret. Mat.*, 13(1):56–62, 2001.

[69] A. A. Sapozhenko. The Cameron-Erdős conjecture. *Dokl. Akad. Nauk*, 393(6):749–752, 2003.

[70] A. A. Sapozhenko. Systems of containers and enumeration problems. In O. B. Lupanov, O. M. Kasim-Zade, A. V. Chaskin, and K. Steinhöfel, editors, *Stochastic Algorithms: Foundations and Applications*, pages 1–13, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[71] A. A. Sapozhenko. Independent sets in quasi-regular graphs. *European J. Combin.*, 27(7):1206–1210, 2006.

[72] A. A. Sapozhenko. An upper bound for the number of independent sets in graphs. *Dokl. Akad. Nauk*, 414(6):743–744, 2007.

[73] D. Saxton and A. Thomason. Hypergraph containers. *Invent. Math.*, 201(3):925–992, 2015.

[74] M. Schacht. Extremal results for random discrete structures. *Ann. of Math. (2)*, 184(2):333–365, 2016.

[75] T. Schoen. Improved bound in Roth's theorem on arithmetic progressions. *Adv. Math.*, 386:Paper No. 107801, 20, 2021.

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

[77] Z. Shao, F. Deng, M. Liang, and X. Xu. On sets without k-term arithmetic progression. *Journal of Computer and System Sciences*, 78(2):610–618, 2012. Games in Verification.

[78] J. Solymosi. Note on a generalization of Roth's theorem. In *Discrete and computational geometry*, volume 25 of *Algorithms Combin.*, pages 825–827. Springer, Berlin, 2003.

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

[80] T. Tao. A variant of the hypergraph removal lemma. *J. Combin. Theory Ser. A*, 113(7):1257–1280, 2006.

[81] B. L. van der Waerden. Beweis einer baudetschen vermutung. *Nieu Arch. Wisk.*, 15:212–216, 1927.

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

[83] S. S. Wagstaff, Jr. On sequences of integers with no 4, or no 5 numbers in arithmetical progression. *Math. Comp.*, 21:695–699, 1967.

[84] S. S. Wagstaff, Jr. On k-free sequences of integers. *Math. Comp.*, 26:767–771, 1972.

[85] J. Wróblewski. Nonaveraging sets search. *https://www.math.uni.wroc.pl/ jwr/non-ave/.*