Arithmetic Progressions in Sumsets of Random Sets

Rafael Kazuhiro Miyazaki

Dissertation presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
in partial fulfillment
of the requirements
for the degree of
Master of Science

Program: Mathematics
Advisor: Prof. Dr. Yoshiharu Kohayakawa

São Paulo
June, 2023

Arithmetic Progressions in Sumsets of Random Sets

Rafael Kazuhiro Miyazaki

This version of the dissertation includes the corrections and modifications suggested by the Examining Committee during the defense of the original version of the work, which took place on June 23, 2023.

A copy of the original version is available at the Institute of Mathematics and Statistics of the University of São Paulo.

Examining Committee:

Prof. Dr. Yoshiharu Kohayakawa – IME-USP
Prof. Dr. Robert Morris – IMPA
Prof. Dr. Carlos Gustavo Tamm de Araújo Moreira – IMPA

Abstract

Rafael Kazuhiro Miyazaki. **Arithmetic Progressions in Sumsets of Random Sets**. Dissertation (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

Given a set A, its sumset $A + A$ is defined as the set of all sums of two elements, not necessarily distinct, in A. Given a function $p : \mathbb{N} \to [0, 1]$, we consider the sequence of independent random sets $\{A_n\}_{n \in \mathbb{N}}$, where A_n is obtained by choosing independently each integer $1 \leq i \leq n$ with probability $p(n)$. We employ the classical probabilistic tools of the first and second moment methods as well as a recently proven theorem of Park and Pham, formerly known as the Kahn–Kalai Conjecture, regarding the relationship between the threshold function and the expectation threshold of increasing properties in order to find lower and upper bounds for the threshold for the existence of arithmetic progressions of $m(n)$ elements in the sumset of the random set A_n.

Keywords: additive combinatorics, number theory, arithmetic progressions, probabilistic method, combinatorics, threshold, expectation threshold.

Resumo

Rafael Kazuhiro Miyazaki. **Progressões Aritméticas em Conjuntos Soma de Conjuntos Aleatórios**. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Dado um conjunto A, seu conjunto soma $A + A$ é definido como o conjunto das somas de dois elementos, não necessariamente distintos, em A. Dada uma função $p : \mathbb{N} \to [0, 1]$, consideramos a sequência de conjuntos aleatórios independentes $\{A_n\}_{n \in \mathbb{N}}$, onde A_n é obtido pela escolha independente de cada inteiro $1 \leq i \leq n$ com probabilidade $p(n)$. Empregamos as ferramentas probabilisticas clássicas dos métodos do primeiro e do segundo momento tal qual um teorema recentemente provado por Park e Pham, anteriormente conhecido como a Conjectura de Kahn–Kalai, a respeito da relação entre o limiar e o limiar para a esperança de propriedades crescentes, a fim de estabelecer cotas inferiores e superiores para o limiar da existência de progressões aritméticas de $m(n)$ elementos no conjunto soma do conjunto aleatório A_n.

Palavras-chave: combinatória aditiva, teoria dos números, progressões aritméticas, método probabilístico, combinatória, limiar, limiar para esperança.

Contents

Chapter 0

Basic Notation and Definitions

Throughout this dissertation, all floor and ceiling symbols are eclipsed whenever the inclusion is not crucial. All logarithms are taken on the natural base e, unless otherwise indicated. We also define some other basic objects that we shall use frequently.

Definition 0.1. *Let n be a positive integer, we define*

$$[n] = \{i \in \mathbb{N} \,:\, 1 \leq i \leq n\}.$$

Definition 0.2 (Sumset)**.** *Let A and B be sets of integer numbers. Then the sumset $A + B$ is the following set*

$$A + B = \{a + b \,:\, a \in A,\ b \in B\}.$$

Definition 0.3 (Arithmetic progressions)**.** *The* non-trivial arithmetic progression *of m elements,* first element x *and* common difference $d > 0$ *is the set*

$$\{x + (i - 1)d \,:\, i \in [m]\}.$$

*We refer to each of these as an m-*AP of difference d*, or simply an m-*AP *from this point on.*

Definition 0.4 (Longest AP)**.** *Given a finite set $X \subseteq \mathbb{N}$, we let $L(X)$ be the largest number of elements of a non-trivial arithmetic progression in X.*

Definition 0.5 (a.a.s)**.** *Given a sequence of random variables $X = \{X_1, X_2, \dots\}$ and a sequence of properties $P = \{P_1, P_2 \dots\}$, we say that X_n satisfies P_n* asymptoticaly almost surely (a.a.s) *if*

$$\lim_{n\to\infty} \mathbb{P}[X_n \textit{ satisfies } P_n] = 1.$$

Definition 0.6 (Little o and little omega notation)**.** *Let $f, g : \mathbb{N} \to \mathbb{R}$. We say that $f = o(g)$ and $g = \omega(f)$ if*

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$

We also sometimes write $f \ll g$ or $g \gg f$ to denote $f = o(g)$.

Definition 0.7 (Big O and big Omega notation)**.** *Let $f, g : \mathbb{N} \to \mathbb{R}$, with both functions eventually positive. We say that $f = O(g)$ and $g = \Omega(f)$ if*

$$\limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty.$$

Definition 0.8 (Big Theta notation)**.** *Let $f, g : \mathbb{N} \to \mathbb{R}$. We say that $f = \Theta(g)$ if $f = O(g)$ and $f = \Omega(g)$.*

Chapter 1

Introduction

In the field of additive combinatorics, an important object of study is the sumset. In such sets one can look for the existence of certain structures, such as arithmetic progressions. One early result regarding such search was obtained by Bourgain [2], which we present here.

Theorem 1.1. *Let A and B be non-empty subsets of $[n]$. Then,*

$$L(A+B) > \exp\left[c\left(\frac{|A||B|\log n}{n^2}\right)^{1/3} - \log\log n\right],$$

for some positive constant c.

This result was improved when considering APs and sets on the ciclic group by Green [4] in the following theorem.

Theorem 1.2. *Let A and B be non-empty subsets of $\mathbb{Z}/n\mathbb{Z}$. Then,*

$$L(A+B) > \exp\left[c\left(\frac{|A||B|\log n}{n^2}\right)^{1/2} - \log\log n\right],$$

for some positive constant c.

Finally we present the following two Theorems and one Corollary by Croot, Ruzsa and Schoen [3]. These consider sparser sets, although only finite sized arithmetic progressions can then be assured in the sumset.

Theorem 1.3. *Let A be a finite set of integers such that $|A-A| = C|A|$ and $|A-2A| = K|A|$.*

Then,

$$\begin{cases} L(A-A) \geq \operatorname{odd}\left(2\dfrac{\log|A|}{\log K}+1\right) \\ L(A+A) \geq \operatorname{odd}\left(2\dfrac{\log(C^{-1}|A|)}{\log CK}+1\right) \\ L(A+A) \geq \operatorname{odd}\left(\dfrac{\log(C^{-1}|A|)}{2\log C}+1\right). \end{cases}$$

Here $\operatorname{odd}(x)$ *denotes the smallest odd number greater than or equal to* x*.*

Corollary 1.4. *For every odd number* $k > 1$ *and* n *sufficiently large, if*

$$A \subseteq [n], \text{ and } |A| \geq (3n)^{1-1/(k-1)},$$

then $L(A+A) \geq k$*.*

Also, if

$$A,\, B \subseteq [n], \text{ and } |A||B| \geq 6n^{2-2/(k-1)},$$

then $L(A+B) \geq k$*.*

Theorem 1.5. *For every* $\varepsilon > 0$*, there exists* $0 < \theta_0 \leq 1$ *so that if* $0 < \theta < \theta_0 \leq 1$*, then there exist infinitely many integers* n *and sets* $A \subseteq [n]$ *with* $|A| \geq n^{1-\theta}$*, such that*

$$L(A+A) < \exp(c\theta^{-2/3-\varepsilon}),$$

where $c > 0$ *is some absolute constant.*

All these results pertain to the size of the longest arithmetic progression in the sumset of deterministic sets, given conditions on the density of the sets themselves or some function of them, such as $A - A$ and $A - 2A$.

We investigate whether we can say similar things about the sumset of random sets. In particular, we shall study the following problem: for a sequence of probabilities given by a function $p : \mathbb{N} \to [0,1]$, we consider the sequence of independent random sets $\{A_n \subseteq [n]\}_{n\in\mathbb{N}}$, where

$$\mathbb{P}[i \in A_n] = p(n) \text{ for all } i \in [n], \tag{1.1}$$

and these events are mutually independent. Formally, for each natural number n, let $X_n = (x_1, \ldots, x_n)$ be an independent random variable uniformly distributed on the hypercube $[0,1]^n$ and for every $i \in [n]$, let $i \in A_n$ if, and only if, $x_i \leq p(n)$. We shall study $L(A_n + A_n)$ as n goes to infinity. We present two related questions.

Question 1.6. *Given a function* $m : \mathbb{N} \to \mathbb{N}$*, such that* $m(n) \leq 2n$ *for every natural number* n*, what is the (threshold) probability* $t_m : \mathbb{N} \to (0,1)$ *for which there are arithmetic progressions of* $m(n)$ *elements in the sumset* $A_n + A_n$ *with probability* $1/2$*?*

Question 1.7. *For a probability sequence* $p : \mathbb{N} \to (0,1)$ *what is the typical size of* $L(A_n + A_n)$*?*

For the case m constant, Theorem 1.8 below answers Question 1.6 up to a constant factor.

Theorem 1.8. *Let m be a positive integer. If $m \geq 4$, then*

$$t_m(n) = \Theta(n^{-1/2-1/m}). \tag{1.2}$$

If $m \leq 3$, then

$$t_m(n) = \Theta(n^{-1}). \tag{1.3}$$

Since for each given n and m the event $L(A_n + A_n) \geq m$ is increasing, Theorem 1.8 combined with a theorem of Bollobás and Thomason [1] tells us that $n^{1/2-1/m}$ $(m \geq 4)$ is the usual Erdős–Rényi threshold function for this event.

We now focus on the case $p \geq n^{-1/2+o(1)}$. In this regime Question 1.7 leads to more concise answers than Question 1.6. We present these answers next. Our simplest, general upper bound result for $L(A_n + A_n)$ is Theorem 1.9 below.

Theorem 1.9. *If $p = n^{-1/2-o(1)}$, $\limsup_{n\to\infty} p^2 n < 1$ and $\tau < -\log(\limsup_{n\to\infty} p^2 n)$ is a positive constant, then*

$$L(A_n + A_n) \leq \frac{2\log n}{-\log(p^2 n) - \tau} \tag{1.4}$$

asymptotically almost surely.

Theorem 1.9 has the following corollaries, in which we consider the cases $p = o(1/\sqrt{n})$ and $p = \Theta(1/\sqrt{n})$ separately.

Corollary 1.10. *If $p = n^{-1/2-o(1)}$, $p = o(\sqrt{1/n})$, then*

$$L(A_n + A_n) \leq (-2 + o(1))\frac{\log n}{\log(p^2 n)} \tag{1.5}$$

asymptotically almost surely.

Corollary 1.11. *If $p \sim \sqrt{\varepsilon/n}$ for some positive constant $\varepsilon < 1$, then*

$$L(A_n + A_n) \leq \left(\frac{-2}{\log \varepsilon} + o(1)\right)\log n \tag{1.6}$$

asymptotically almost surely.

We now turn to lower bounds for $L(A_n + A_n)$. We start with the following result.

Theorem 1.12. *If $p(n) \leq \sqrt{(\log n)/n}$, then*

$$L(A_n + A_n) \geq \frac{2\log n}{\log\log n + 2\log\log\log n - \log(p^2 n)} \tag{1.7}$$

asymptotically almost surely.

We can simplify (1.7) according to which of $\log\log n$ and $-\log(p^2 n)$ is the main term in the denominator of the right-hand side of (1.7). Doing so we may derive the following two corollaries.

Corollary 1.13. *If* $p = \sqrt{1/n(\log n)^{\omega(1)}}$*, then*

$$L(A_n + A_n) \geq (-2 + o(1)) \frac{\log n}{\log(p^2 n)} \tag{1.8}$$

asymptotically almost surely.

Corollary 1.14. *If* $p = \sqrt{1/n(\log n)^{c+o(1)}}$ *for some nonnegative constant* c*, then*

$$L(A_n + A_n) \geq \left(\frac{2}{1+c} + o(1)\right) \frac{\log n}{\log\log n} \tag{1.9}$$

asymptotically almost surely.

Recall that Theorem 1.12 applies to $p \leq \sqrt{(\log n)/n}$. An alternative approach lets us obtain other lower bounds for $L(A_n + A_n)$ for $p \gg \sqrt{1/n}$.

Theorem 1.15. *If* $p(n) < \sqrt{2(\log n)/n}$ *and* $p = \omega(\sqrt{1/n})$*, then*

$$L(A_n + A_n) \geq e^{(1/2+o(1))p^2 n} \tag{1.10}$$

asymptotically almost surely.

Theorem 1.16. *If* $p(n) = \sqrt{(C + o(1))(\log n)/n}$ *for some constant* $C > 2$*, then*

$$L(A_n + A_n) \geq (2 - 4/C)n \tag{1.11}$$

asymptotically almost surely.

There is an overlap between the ranges of p considered in Theorems 1.12 and 1.15. A straightforward calculation shows that the lower bound in Theorem 1.15 is asymptotically larger then the one in Theorem 1.12 if $p > (\sqrt{2} + o(1))\sqrt{(\log\log n - \log\log\log n)/n}$.

Our results above are summarized in Table 1.1. In that table, k denotes an integer with $k \geq 4$, c denotes a positive constant, ε denotes a constant with $0 < \epsilon < 1$ and C denotes a constant with $C > 2$. The functions m and M are lower and upper bounds for $L(A_n + A_n)$, respectively.

Notice that for $p = \sqrt{1/n(\log n)^{\omega(1)}}$ with $p = n^{-1/2-o(1)}$ and $p = \omega(\sqrt{(\log n)/n})$, our results imply that the random variable $L(A_n + A_n)$ is concentrated in an interval whose endpoints are asymptotically equal. In other words, we know the value of $L(A_n + A_n)$ asymptotically for such p. Notice also that for $p = \sqrt{1/n(\log n)^{c+o(1)}}$ and $p = \sqrt{(C + o(1))(\log n)/n}$, the random variable $L(A_n + A_n)$ is concentrated in an interval whose endpoints have a bounded ratio, that is, we know $L(A_n + A_n)$ up to a multiplicative constant for those p.

Before we proceed, we remark that Theorems 1.8, 1.9, 1.12 above are derived from Theorems 1.17 and 1.18, which are somewhat more technical and presented below. The proof of Theorems 1.8, 1.9, 1.12 are shown in the end of this introduction.

p	m	M	M/m
	$0 \le p = n^{-1/2-\Omega(1)}$		
0	0	0	–
$o(n^{-1/2-1/k})$	–	$k-1$	–
$\omega(n^{-1/2-1/k})$	k	–	–
$\Theta(n^{-1/2-1/k})$	$k-1$	k	–
	$n^{-1/2-o(1)} = p \le \sqrt{1/n}$		
$o(\sqrt{1/n})$	–	$(-2+o(1))\frac{\log n}{\log(p^2 n)}$	–
$\sqrt{1/n(\log n)^{\omega(1)}}$	$(-2+o(1))\frac{\log n}{\log(p^2 n)}$	$(-2+o(1))\frac{\log n}{\log(p^2 n)}$	$1+o(1)$
$\sqrt{1/n(\log n)^{c+o(1)}}$	$\left(\frac{2}{1+c}+o(1)\right)\frac{\log n}{\log\log n}$	$\left(\frac{2}{c}+o(1)\right)\frac{\log n}{\log\log n}$	$\frac{c+1}{c}+o(1)$
$\sqrt{1/n(\log n)^{o(1)}}$	$(2+o(1))\frac{\log n}{\log\log n}$	–	–
$\sqrt{\varepsilon/n}$	$(2+o(1))\frac{\log n}{\log\log n}$	$\left(\frac{-2}{\log\varepsilon}+o(1)\right)\log n$	$\left(\frac{-1}{\log\varepsilon}+o(1)\right)\log\log n$
$\sqrt{1/n}$	$(2+o(1))\frac{\log n}{\log\log n}$	$2n$	$(1+o(1))\frac{n\log\log n}{\log n}$
	$\sqrt{1/n} \ll p \le 1$		
$< \sqrt{2(\log n)/n}$	$\max\left((2+o(1))\frac{\log n}{\log\log n}, e^{(1/2+o(1))p^2 n}\right)$	$2n$	–
$\sqrt{(C+o(1))(\log n)/n}$	$\left(2-\frac{4}{C}\right)n$	$2n$	$\frac{C}{C-2}$
$\omega(\sqrt{(\log n)/n})$	$(2-o(1))n$	$2n$	$1+o(1)$
1	$2n$	$2n$	1

Table 1.1: *Lower & upper bounds for various p*

Theorems 1.17 gives upper bounds for $L(A_n+A_n)$ while Theorems 1.18 gives lower bounds for $L(A_n+A_n)$. Theorems 1.15 and 1.16 are proved in Chapter 5.

Theorem 1.17 (APs are short). *Let $g, m : \mathbb{N} \to \mathbb{R}_{>0}$ and let $p(n) = n^{-0.5-1/m} g(n)$. Let the random set A_n be defined by (1.1). If either*

(a) $m(n) = c$ constant, $c \ge 4$ and $g(n) = o(1)$, or

(b) $1 \ll m(n) = n^{o(1)}$ and $g(n) = 1/2$,

then $L(A_n+A_n) < m(n)$ asymptotically almost surely.

Theorem 1.17 is proved in Chapter 2 using the first moment method as its probabilistic tool. We note that for any function m satisfying the conditions of Theorem 1.17, the probability p considered are smaller than $n^{-0.5}$.

Theorem 1.18 (There are long APs)**.** *Let* $g,\ m:\ \mathbb{N} \to \mathbb{R}_{>0}$ *and let* $p(n) = n^{-1/2-1/m}g(n)$*. If either*

(a) $3 \leq m(n) < 0.48\sqrt{(\log n)/\log\log n}$ *and* $g \gg \log m$ *or*

(b) $0.48\sqrt{(\log n)/\log\log n} \leq m(n) < 0.1 \log n$ *and* $g \gg m^{1/2}\log m$,

then $L(A_n + A_n) \geq m(n)$ *asymptotically almost surely.*

Theorems 1.18(a) and 1.18(b) are proved in Chapter 4 using the Park-Pham Theorem [6].

In Chapter 6 we present a tentative approach to finding lower bounds for $L(A_n + A_n)$ using the second moment method. We later found that these bounds could be improved by Theorem 1.15.

As promised we now prove Theorems 1.8, 1.9, 1.12 as a consequence of Theorems 1.17 and 1.18.

Proof of Theorem 1.8. If $m \geq 4$, Theorems 1.17(a) and 1.18(a) can be applied and yield the desired result. If $m \leq 3$, then

$$\mathbb{P}[L(A_n + A_n) \geq m] = \begin{cases} \mathbb{P}[|A_n| \geq 2] \text{ if } m \in \{2, 3\}, & (1.12) \\ \mathbb{P}[|A_n| \geq 1] \text{ if } m = 1, & (1.13) \end{cases}$$

and Chernoff bounds suffice for the claimed result. □

Proof of Theorem 1.9. Let $m = 2\log n/(-\log(p^2 n) - \tau)$. Notice that $m \to \infty$ as $p = n^{-1/2+o(1)}$ and that

$$p = e^{-\tau/2} n^{-1/2-1/m}. \tag{1.14}$$

Then, by Theorem 1.17, we have $L(A_n + A_n) \leq m(n)$ a.a.s. □

Proof of Theorem 1.12. Let $m = 2\log n/(\log\log n + 2\log\log\log n - \log(p^2 n))$. Notice that

$$\frac{\log n}{m} = \frac{\log\log n}{2} + \log\log\log n - \log p - \frac{\log n}{2}, \tag{1.15}$$

also

$$\frac{\log m}{2} = \frac{1}{2}(\log 2 + \log\log n - \log(\log\log n + 2\log\log\log n - \log(p^2 n))) = \frac{\log\log n}{2} - \omega(1), \tag{1.16}$$

since $\log(p^2 n) < \log\log n$. Finally

$$\log\log m = \log(\log 2 + \log\log n - \log(\log\log n + 2\log\log\log n - \log(p^2 n))) < \log\log\log n \tag{1.17}$$

for sufficiently large n. This in turn implies that

$$\sqrt{m}(\log m)n^{-1/2-1/m} = \exp\left(\frac{\log m}{2} + \log\log m - \frac{\log n}{2} - \frac{\log n}{m}\right) = o(p). \tag{1.18}$$

Then, by Theorem 1.18(b), we have $L(A_n + A_n) \geq m(n)$ a.a.s. $\square$

Chapter 2

When are All Arithmetic Progressions in the Sumset of a Random Set Short?

In this chapter, we prove Theorem 1.17, which pertains to bounds for the probability $p(n)$ that almost guarantee the non-existence of long arithmetic progressions in $A_n + A_n$, with A_n as defined in (1.1).

We prove a more general result stated as Theorem 2.1. Afterwards, we provide a counting lemma that gives an upper bound on the number of arithmetic progressions on the support set of $A_n + A_n$ and that allow us to use Theorem 2.1 to prove Theorem 1.17.

Theorem 2.1. *Let $g, m : \mathbb{N} \to \mathbb{R}_{>0}$ and let $p(n) = \min\left(1, n^{-0.5-1/m} g(n)\right)$. Let $\{\mathcal{R}_n\}_{n\in\mathbb{N}}$ be a sequence of families of $m(n)$-subsets of $[2n]$ such that $|\mathcal{R}_n| = O(n^2/m)$. Let the random set A_n be defined by (1.1). If either*

(a) $m(n) = c$ constant, $c \geq 4$ and $g(n) = o(1)$, or

(b) $1 \ll m(n) = n^{o(1)}$ and $g(n) = 1/2$,

then $A_n + A_n$ does not contain any member of $\mathcal{R}_n$ asymptotically almost surely.

An outline of the proof of Theorem 2.1 is as follows. We will show the existence of a family $\mathcal{B}$ of subsets B of $[n]$ such that in order for no members of $\mathcal{R}$ to be contained in $A_n + A_n$ it suffices that no members of $\mathcal{B}$ are contained in A_n. We then use the first moment method to show that the expected number of members of $\mathcal{B}$ that are subsets of A_n is close to 0, thus Theorem 2.1 holds.

The following definitions will be useful in order to avoid the explicit repetition of conditions used in our proof.

Definition 2.2 (Second order cover)**.** *Let $R \subseteq [2n]$. An R-*second order cover *(or R-*soc*) is any subset B of $[n]$ such that $R \subseteq B + B$.*

Definition 2.3 (Modeling graph)**.** *Let $R \subseteq [2n]$ and $B \subseteq [n]$. Let $G = (V, E)$ be a multigraph without parallel edges, but possibly some loops. If there is an injection $r : E \to R$ and a* *bijection* $b : V \to B$ *such that*

$$b(u) + b(v) = r(uv), \text{ for all } uv \in E,$$

we say that B *has an* R-modeling graph G *with* vertex labeling function b *and* edge labeling function r.

It is important to note that r does not need to be a bijection, that is not all elements of R need to be represented by edges of G.

Lemma 2.4. *Let* $m \geq 3$ *and let* C *be a connected multigraph without parallel edges. Further suppose that* C *is on* k *vertices and has* a *edges, where* $1 \leq a \leq m$. *Define the* m-weight *of* C *to be*

$$w_m(C) = \begin{cases} 1/k + 2a/mk, & \text{if } C \text{ is bipartite,} \\ 2a/mk, & \text{otherwise.} \end{cases}$$

Also define the following multigraphs:

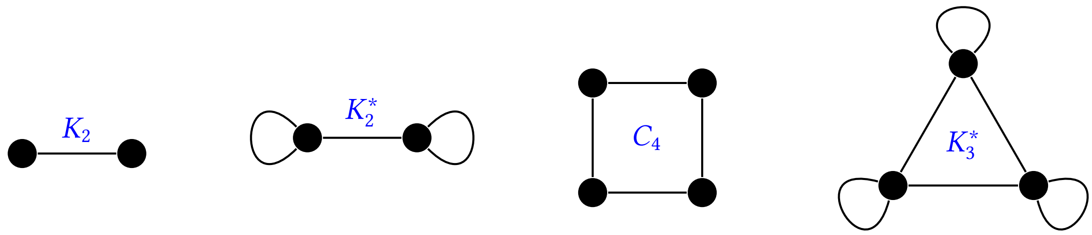

Figure 2.1: *Heavy connected components*

Then

$$w_m(C) \begin{cases} > w_m(K_2), & \text{if } C \in \{K_2^*\} \text{ and } m = 3, \\ = w_m(K_2), & \text{if } C = K_2, \\ = w_m(K_2), & \text{if } C \in \{C_4, K_2^*\} \text{ and } m = 4, \\ = w_m(K_2), & \text{if } C = K_3^* \text{ and } m = 6, \\ < w_m(K_2), & \text{otherwise.} \end{cases}$$

Proof. It is of our interest to find the connected multigraphs C for which the inequality

$$w_m(C) \leq w_m(K_2) = \frac{1}{2} + \frac{1}{m}$$

is false or yields an equality case. Supposing that C is bipartite, $w_m(C) \leq w_m(K_2)$ if, and only if,

$$mk + 2k \geq 2m + 4a,$$

with the same equality conditions. If $k \geq 6$, we have

$$mk + 2k > 6m \geq 2m + 4a,$$

since $m \geq a$. No equality cases exist.

If $k = 5$, then

$$mk + 2k = 5m + 10 \geq 2m + 3a + 10 > 2m + 4a,$$

since $m \geq a$ and $a \leq k^2/4 = 6.25$. Again, no equality cases exist.

If $k = 4$, then

$$mk + 2k = 4m + 8 \geq 2m + 2a + 8 \geq 2m + 4a,$$

since $m \geq a$ and $a \leq k^2/4 = 4$. Equality holds when $m = a = 4$, i.e., when $(C, m) = (C_4, 4)$.

If $k = 3$, then $C = P_3$, the path on 3 vertices, and therefore

$$mk + 2k = 3m + 6 > 2m + a + 6 = 2m + 4a,$$

since $a = 2 < 3 \leq m$. No equality cases exist.

If $k = 2$, then $C = K_2$ and $w_m(C) = w_m(K_2)$.

Finally, if $k = 1$, then $a = 0$, but $a \geq 1$.

Supposing now that C is not bipartite, $w_m(G) \leq w_m(K_2)$ if, and only if,

$$mk + 2k \geq 4a,$$

with the same equality conditions.

If $k \geq 4$, then

$$mk + 2k > 4m \geq 4a,$$

since $m \geq a$. No equality cases exist.

If $k = 3$, then

$$mk + 2k = 3m + 6 \geq 3a + 6 \geq 4a,$$

since $m \geq a$ and $a \leq \binom{k}{2} + k = 6$. Equality holds when $m = a = 6$, i.e., when $(C, m) = (K_3^*, 6)$.

If $k = 2$ and $a \leq 2$, then

$$mk + 2k = 2m + 4 \geq 10 > 8 \geq 4a,$$

since $m \geq 3$. No equality cases exist.

If $k = 2$ and $a \geq 3$, then $C = K_2^*$. Therefore

$$w_m(C) \begin{cases} > w_m(K_2), \text{ if } m = 3, \\ = w_m(K_2), \text{ if } m = 4, \\ < w_m(K_2), \text{ if } m \geq 5. \end{cases}$$

Finally, if $k = 1$, then $a = 1$ and

$$mk + 2k = m + 2 > 4 = 4a,$$

since $m \geq 3$. No equality cases exist. □

Proof of Theorem 2.1(a). For a fixed $R \in \mathcal{R}_n$ and $B \subseteq [n]$ a minimal R-soc, define the hypergraph $G_B = G_B^R = (B, E)$ as follows:

$$xy \in E \iff (x + y \in R) \wedge (|x - y| = \min\{|z - w| \,:\, z + w = x + y,\ z \in B,\ w \in B\}). \quad (2.1)$$

Further, consider the labeling of edges $r : E \to R$, where

$$r(xy) = x + y, \text{ for all } ab \in E. \quad (2.2)$$

Notice that r is a bijective function and G_B is an R-modeling graph with the identity function as the vertex labeling function.

Observe also that $\delta(G_B) \geq 1$, as otherwise the set of vertices of positive degree corresponds to a proper subset of B that is an R-soc. Finally, notice that G_B is an hypergraph on at most $2c$ vertices and c edges. Let $\mathcal{G}$ be the finite family of such multigraphs.

For a fixed $G \in \mathcal{G}$ that has q bipartite connected components and K vertices, we claim that

$$\mathbb{P}[\exists B \subseteq A \,:\, G_B = G] \leq c!n^q p^K. \quad (2.3)$$

Indeed there are $c!$ edge labeling functions r for the modeling graph G. Take a connected component H of G. If H is bipartite, let T_H be an arbitrary spanning tree of H and let e be the edge incident to a leaf that has the least label $r(e)$. Let v be the leaf incident to e (if $H = K_2$, choose v arbitrarily). Now, if $ve_1e_2 \dots e_s w$ is a path from v to w, then

$$b(w) = (-1)^s b(v) + \sum_{t=1}^{s} r_1(e_t)(-1)^{s-t}. \quad (2.4)$$

If H is not bipartite it must contain a loop or an odd cycle with at least 3 vertices. If there is a loop e on a vertex v, then $2b(v) = r(e)$ and $b(v)$ is uniquely defined. If $v_1v_2 \dots v_{2s+1}v_1$ is an odd cycle in G, then, $b(v_1), \dots, b(v_{2s+1})$ must satisfy the linear system

$$b(v_i) + b(v_{i+1}) = r(v_i v_{i+1}),$$

where the index i runs from 1 to $2s + 1$ and we let $v_{2s+2} = v_1$. This system of equations has a unique rational solution given the edge labeling function r, since

$$\det \begin{pmatrix} 1 & 1 & & & \\ & 1 & 1 & & \\ & & \ddots & \ddots & \\ & & & 1 & 1 \\ 1 & & & & 1 \end{pmatrix} = 2 \neq 0$$

whenever the order of the matrix is an odd number larger than 2. Then the vertices in the cycle are uniquely labeled and (2.4) can then be used to find the labels of the remaining vertices of H.

Therefore, once the q leaves connected to the edges with least label in the chosen spanning trees of the bipartite connected components are selected, all vertices in G are uniquely labeled, if at all possible. There are at most n^q such choices. Finally notice that the probability that each B is contained in A is p^K, as $|B|$ is the number of vertices of G.

Now, let $C(G)$ be the set of connected components of G and I_C be the indicator function of the property C is bipartite. Finally notice that, because of Lemma 2.4, we have

$$\begin{aligned} q+2 &= \sum_{C\in C(G)} \left(I_C + \frac{2|E(C)|}{m} \right) \\ &= \sum_{C\in C(G)} (w_m(C)|V(C)|) \\ &\leq \left(\frac{1}{2}+\frac{1}{m}\right) \sum_{C\in C(G)} |V(C)| \\ &= \left(\frac{1}{2}+\frac{1}{m}\right) K \end{aligned}$$

and therefore, by (2.3), for fixed $R \in \mathcal{R}_n$ and $G \in \mathcal{G}$, we have

$$\mathbb{P}[\exists B \subseteq A : G_B^R = G] \leq c! n^q p^K = o(n^{-2} n^{2+q} n^{-K(1/2+1/m)}) = o(1/n^2),$$

since $c!$ is a constant dependent of c and $p = o(n^{-1/2-1/m})$ by the Theorem's conditions.

At last, by the union bound and because every R-soc contains a minimal R-soc, we have

$$\mathbb{P}[\mathcal{P}(A_n + A_n) \cap \mathcal{R}_n \neq \emptyset] \leq \sum_{R\in\mathcal{R}_n} \left(\sum_{G\in\mathcal{G}} \mathbb{P}[\exists B \subseteq A : G_B^R = G] \right) = o(1).$$

since $|\mathcal{G}|$ is a constant dependent of c and $|\mathcal{R}_n| = O(n^2)$ □

Lemma 2.5. *Let $R \subseteq [2n]$ be given, with $|R| \geq 64$. Let $B \subseteq [n]$ be a minimal R-soc. Then, at least one of the following statements is true.*

1. *The set B contains a 5-subset B_0 with R-modeling graph G_0 that has 5 loops as its only edges.*
2. *The set B contains a 7-subset B_1 with an R-modeling graph G_1 that is a 7-vertex tree.*
3. *The set B contains a 15-subset B_2 with an R-modeling graph G_2 that has 5 copies of the 3-vertex path as its connected components.*
4. *The set $B_3 = B$ has an R-modeling graph G_3 that has $|R|$ edges, has at least $|R| - 96$ connected components isomorphic to K_2, has no connected component on more than 6 vertices, has at most 4 connected components on between 3 and 6 vertices, and has at most 4 connected components that are on 2 vertices and have more than one edge or are on 1 vertex.*

Proof. Consider the R-modeling graph $G = G_B = (B, E)$ with edge labeling r given by (2.1) and (2.2). Recall that $\delta(G) \geq 1$, as otherwise the set of vertices of positive degree corresponds to a proper subset of B that is an R-soc.

Let t be the number of connected components of G. Let $s_1 \geq s_2 \geq \cdots \geq s_t > 0$ be the number of vertices in the connected components of G. Let us consider some cases that G might fit.

1. If G contains 5 loops, then the labels of the 5 vertices in those loops form a set B_0 as desired.
2. If $s_1 \geq 7$, then there is a 7-vertex tree that is a subgraph of G, whose set of labels B_1 is as desired.
3. If $t \geq 5$ and $s_5 \geq 3$, then there are at least 5 connected components of at least 3 vertices, and each of those must have a 3-vertex path as a subgraph. Then the set of labels B_2 of the vertices in these 5 paths is as desired.
4. If G does not satisfy any of the previous cases, then G has no connected components with at least 7 vertices, has at most 4 connected components with 3 to 6 vertices, each having at most $\binom{6}{2} = 15$ non-loop edges. Also there are at most 4 connected components on 1 or 2 vertices that contain loops, and each of those contains at most 1 non-loop edge. That means that at least $|R| - 68$ edges of G are in connected components isomorphic to K_2.

□

Proof of Theorem 2.1(b). Recall that $m = n^{o(1)}$ and notice that for sufficiently large n, we have $g(n) = 1/2 < m^{-34/m}$ Let $R \in \mathcal{R}_n$ and for each $i \in \{0, 1, 2, 3\}$, let $\mathcal{B}_i(R)$ be the class of sets B_i as in the statement $i + 1$ of Lemma 2.5. We claim that

$$\mathbb{P}[\exists B_0 \subseteq A : B_0 \in \mathcal{B}_0(R)] \leq m^5 p^5 = O\left(\frac{m^5}{(m^{34}n)^{5/m} n^{2.5}}\right) = o\left(\frac{m}{n^2}\right).$$

Indeed, notice that G_0 is uniquely defined, up to isomorphism, and there are $\binom{m}{5} < m^5$ possible edge labeling functions r_0. Once G_0 and r_0 are defined, B_0 is uniquely defined, since each vertex must be labeled with half of the label of its loop. Finally, the probability that each B_0 is contained in A is p^5. We also claim that

$$\mathbb{P}[\exists B_1 \subseteq A : B_1 \in \mathcal{B}_1(R)] \leq 11 m^6 n p^7 = O\left(\frac{m^6}{(m^{34}n)^{7/m} n^{2.5}}\right) = o\left(\frac{m}{n^2}\right).$$

Indeed, notice that G_1 can be, up to isomorphism, one of 11 possible trees[1], and there are at most m^6 possible edge labeling functions r_1. Let e be the edge incident to a leaf that has the least label $r_1(e)$. Let v be the leaf incident to e. If $v e_1 e_2 \ldots e_s w$ is a path from v to w, then

$$b(w) = (-1)^s b(v) + \sum_{t=1}^{s} r_1(e_t)(-1)^{s-t}.$$

As G_1 is connected, B_1 is uniquely defined by G_1, r_1 and $b(v) \in [n]$. Finally, the probability

[1] There are in fact 11 unlabeled trees on 7 vertices.

that each B_1 is contained in A is p^7. We also claim that

$$\mathbb{P}[\exists B_2 \subseteq A : B_2 \in \mathcal{B}_2(R)] \leq m^{10}n^5p^{15} = O\left(\frac{m^{10}}{(m^{34}n)^{15/m}n^{2.5}}\right) = o\left(\frac{m}{n^2}\right).$$

Indeed, notice that G_2 is uniquely defined, up to isomorphism, and there are $\binom{m}{2,2,2,2,2} < m^{10}$ possible edge labeling functions r_2. As in the previous case, B_2 is uniquely defined by G_2, r_2, and the labeling of the leaves incident to the edges with the least label in each of the five connected components of G_2. There are at most n^5 such labelings. Finally, the probability that each B_2 is contained in A is p^{15}.

Notice that if $B \in \mathcal{B}_3(R)$, then H_B^R, the graph obtained by removing the connected components of G_B^R isomorphic to K_2, must be a hypergraph on at most 68 edges. Let $\mathcal{H}$ be the finite family of such graphs. Then for fixed $H \in \mathcal{H}$, if G, obtained by adding connected components isomorphic to K_2 to H so that the final graph has m edges, has q bipartite connected components and K vertices, we have

$$\mathbb{P}[\exists B \subseteq A : B \in \mathcal{B}_3(R) \wedge H_B^R = H] \leq m^{68}n^qp^K, \tag{2.5}$$

as G_3 is uniquely defined, up to isomorphism, and there are at most m^{68} possible edge labeling functions r_3, as once the edges in H are labeled, the labeling of the other edges is defined up to isomorphism. Similarly to what was done in the proof of Theorem 2.1(a), once the q leaves connected to the edges with least label in the chosen spanning trees of the bipartite connected components are selected, all vertices in G are uniquely labeled, if at all possible. There are at most n^q such choices. Finally notice that the probability that each B is contained in A is p^K, as $|B|$ is the number of vertices of G.

Again, for sufficiently large n, by Lemma 2.4, 2.5, for fixed $R \in \mathcal{R}_n$ and $H \in \mathcal{H}$, we have

$$\begin{aligned}\mathbb{P}[\exists B \subseteq A : B \in \mathcal{B}_3(R) \wedge H_B^R = H] &\leq m^{68}n^{-2}n^{2+q}n^{-K(1/2+1/m)}m^{-34K/m}\\ &\leq mn^{-2}m^{67-34K/m}\\ &\leq mn^{-2}m^{67-34(2m-136)/m}\\ &= mn^{-2}m^{-1+4624/m} = o(m/n^2),\end{aligned}$$

since $K \geq 2(m-68)$ as there are at least $m-68$ non-loop edges in the connected components isomorphic to K_2 and $m \gg 1$.

With this last inequality and the union bound, we have for a fixed $R \in \mathcal{R}_n$ that

$$\mathbb{P}[\exists B_3 \subseteq A_n : B_3 \in \mathcal{B}_3(R)] \leq \sum_{H \in \mathcal{H}} \mathbb{P}[\exists B \subseteq A : B \in \mathcal{B}_3(R) \wedge H_B^R = H] = o(m/n^2),$$

since $\mathcal{H}$ is finite.

At last, by Lemma 2.5, the union bound and because every R-soc contains a minimal R-soc, we have

$$\mathbb{P}[\mathcal{P}(A_n + A_n) \cap \mathcal{R}_n \neq \emptyset] \leq \sum_{R \in \mathcal{R}_n}\left(\sum_{i=0}^{3} \mathbb{P}[\exists B_i \subseteq A_n : B_i \in \mathcal{B}_i(R)]\right) = o(1).$$

□

We now produce upper bounds to the number of m-APs in the set $[2n]$.

Lemma 2.6. *Let $m > 1$ and n be integers. Then there are at most*

$$\frac{2n^2 - n}{m - 1}$$

m-APs that are subsets of $[2n]$.

Proof. Let a be the smallest element of an m-AP that is a subset of $[2n]$. Then the common difference d must satisfy

$$a + (m-1)d \leq 2n,$$

which implies that

$$d \leq \frac{2n - a}{m - 1},$$

from where we conclude that there are at most

$$\sum_{a=1}^{2n} \frac{2n - a}{m - 1} = \frac{(2n-1)n}{m - 1}$$

m-APs in $[2n]$. □

Finally, we can prove Theorem 1.17.

Proof of Theorem 1.17. Let $\mathcal{R}_n$ be the family of $m(n)$-APs in $[2n]$. Because of Lemma 2.6, we have that $|\mathcal{R}_n| = O(n^2/m)$ and the result follows from Theorem 2.1. □

Chapter 3

Theorem of Park and Pham

In this chapter, we introduce the theorem of Park and Pham [6] and the motivations for its application in our problem. In the next chapter we exhibit a proof of Theorems 1.18(a) and 1.18(b) using the techniques presented here.

We begin with some definitions used in the main theorem of this chapter.

Definition 3.1 (Increasing family). *A family $\mathcal{F}$ of subsets of $[n]$ is said to be an* increasing family *if all $B \supseteq A \in \mathcal{F}$ satisfy $B \in \mathcal{F}$. If additionaly $\mathcal{F} \notin \{\emptyset, \mathcal{P}([n])\}$, then $\mathcal{F}$ is a* non-trivial increasing family.

Our interest in this definition is rooted in the fact that for a fixed number m, the family of sets $A \subseteq [n]$ such that $L(A + A) \geq m$ is an increasing family. This is the case as any added elements in A cannot eliminate an m-AP in the set $A + A$.

Definition 3.2 (Product measure). *Let $p \in [0, 1]$ and let n be a positive integer. Then μ_p denotes the* product measure *on $\mathcal{P}([n])$, given by*

$$\mu_p(A) = p^{|A|}(1 - p)^{n-|A|}.$$

Furthermore, if $\mathcal{F} \subseteq \mathcal{P}([n])$ is an increasing family, we let

$$\mu_p(\mathcal{F}) := \sum_{A \in \mathcal{F}} \mu_p(A).$$

It is of note that for a fixed non-trivial and increasing family $\mathcal{F} \subseteq \mathcal{P}([n])$, the product measure $\mu_p(\mathcal{F})$ is strictly increasing in p. Indeed, let us first note that if $x = (x_1, x_2, \dots, x_n)$ is a random vector uniformly distributed in $[0, 1]^n$, then $\mu_p(\mathcal{F})$ is the probability that the random set $A_p = \{i \in [n] : x_i \leq p\}$ is in $\mathcal{F}$. Now notice that if $p < q$, then $A_p \subseteq A_q$. Because $\mathcal{F}$ is an increasing family, we then have

$$\mathbb{P}[A_q \in \mathcal{F}] = \mathbb{P}[A_p \in \mathcal{F}] + \mathbb{P}[A_q \in \mathcal{F}, A_p \notin \mathcal{F}] \geq \mathbb{P}[A_p \in \mathcal{F}] + (q - p)^n,$$

and hence $\mu_p(\mathcal{F})$ is strictly increasing in p, as claimed.

Note furthermore that $\mu_p(\mathcal{F})$ is continuous in p. This motivates the following definition.

Definition 3.3 (Threshold). *For a non-trivial and increasing family $\mathcal{F} \subseteq \mathcal{P}([n])$, let the* threshold $p_c(\mathcal{F})$ *be the unique p for which $\mu_p(\mathcal{F}) = 1/2$.*

Definition 3.4 (Cover). *Given an increasing family $\mathcal{F} \subseteq \mathcal{P}([n])$, we say that $\mathcal{G} \subseteq \mathcal{P}([n])$ is a* cover *of $\mathcal{F}$ when every member of $\mathcal{F}$ contains some member of $\mathcal{G}$.*

Note that if $\mathcal{G}$ is a cover of the increasing family $\mathcal{F}$, then $\mathcal{G}$ acts as a kind of proxy to $\mathcal{F}$. Precisely, instead of looking up if a set A is a member of $\mathcal{F}$, one can look up if there are any members of $\mathcal{G}$ contained in A.

Definition 3.5 (p-small). *Given an increasing family $\mathcal{F} \subseteq \mathcal{P}([n])$ and $p \in [0,1]$, we say that $\mathcal{F}$ is* p-small *if there is a cover $\mathcal{G}$ of $\mathcal{F}$ such that*

$$f_{\mathcal{G}}(p) = \sum_{S \in \mathcal{G}} p^{|S|} \leq \frac{1}{2}.$$

Note that if A is a random set in the probability space given by the product measure μ_p, then $f_{\mathcal{G}}(p)$ denotes the expected number of subsets of A in $\mathcal{G}$. In particular, by Markov's Inequality, if $\mathcal{F}$ is p-small, then $p \leq p_c(\mathcal{F})$.

Notice also that for a fixed cover $\mathcal{G}$ of a non-trivial increasing family $\mathcal{F}$, the function $f_{\mathcal{G}}(p)$ is strictly increasing and continuous in p. This motivates the following definition.

Definition 3.6 (Expectation threshold). *Given a non-trivial increasing family $\mathcal{F} \subseteq \mathcal{P}([n])$, the* expectation threshold *$q(\mathcal{F})$ is the largest p such that $\mathcal{F}$ is p-small.*

We state a celebrated conjecture posed by Kahn and Kalai [5], which was recently proved by Park and Pham [6].

Theorem 3.7. *There is a universal constant K such that*

$$p_c(\mathcal{F}) \leq K q(\mathcal{F}) \log \ell(\mathcal{F})$$

for every n and every increasing family $\mathcal{F} \subseteq \mathcal{P}([n])$, where $\ell(\mathcal{F})$ is the largest size of a minimal element of $\mathcal{F}$.

We extract the following Corollary, which we will use in the next chapter.

Corollary 3.8. *For every positive real δ, there is a constant $L = L(\delta)$ such that if $\mathcal{F} \subseteq \mathcal{P}([n])$ is an increasing family, where $\ell(\mathcal{F})$ is the largest size of a minimal element of $\mathcal{F}$ and*

$$p \geq L q(\mathcal{F}) \log \ell(\mathcal{F}),$$

then $\mu_p(\mathcal{F}) > 1 - \delta$.

Proof. Let $s = \lceil -\log_2(\delta) \rceil$ and $L = Ks$. Then, by Bernoulli's Inequality and Theorem 3.7, we have

$$p \geq L q(\mathcal{F}) \log \ell(\mathcal{F}) \geq s p_c(\mathcal{F}) \geq 1 - (1 - p_c(\mathcal{F}))^s =: p^*.$$

Now let $B = \bigcup_{i=1}^{s} A_i$, where $\{A_i\}_{i=1}^{s}$ is a sequence of independent random subsets of $[n]$ whose distribution are given by the product measure μ_{p_c}. Further notice that, by definition of $p_c(\mathcal{F})$, we have

$$\mathbb{P}[B \in \mathcal{F}] \geq \mathbb{P}[\exists i \in [s] \,:\, A_i \in \mathcal{F}] = 1 - (1/2)^s \geq 1 - \delta$$

and that the distribution of the random set B is given by the product measure μ_{p*}, from which we conclude $\mu_p(\mathcal{F}) \geq \mu_{p^*}(\mathcal{F}) \geq 1 - \delta$, as needed. $\square$

Chapter 4

Finding Long Arithmetic Progressions in the Sumset of a Random Set Using the Expectation Threshold

In this chapter we prove Theorem 1.18(a) and 1.18(b), which are restated as Theorem 4.1(a) and 4.1(b) for convenience.

Theorem 4.1. *For every positive real δ, there is a constant $L = L(\delta)$ for which the following holds. For all functions $m : \mathbb{N} \to \mathbb{N}_{\geq 3}$ and $p : \mathbb{N} \to [0, 1]$ such that either:*

(a) $m(n) < 0.48\sqrt{\log n / \log\log n}$ *and* $p(n) > Ln^{-0.5-1/m(n)} \log m(n)$ *or*

(b) $m(n) < 0.1 \log n$ *and* $p(n) > Ln^{-0.5-1/m(n)} \sqrt{m(n)} \log m(n)$,

the random set A_n defined in (1.1) satisfies

$$L(A_n + A_n) \geq m(n),$$

with probability at least $1 - \delta$.

Similarly to Chapter 2 we prove a counting lemma, this time giving a lower bound on the number of certain arithmetic progressions.

Lemma 4.2. *Let $m : \mathbb{N} \to \mathbb{N}_{\geq 3}$, where $m(n) = O(\log n)$. There exists an integer n_0 such that if $n > n_0$ and $m = m(n)$, then there are at least*

$$\frac{0.9n^2}{m}$$

m-APs that are subsets of $[n/2, 3n/2]$.

Proof. Let $a \in [n/2, 3n/2]$. Then if the common difference d satisfies

$$d \leq \frac{3n - 2a}{2m - 2},$$

it must also satisfy

$$a + (m-1)d \leq \frac{3n}{2},$$

and therefore $\{a, a+d, \dots, a+d(m-1)\}$ is an m-AP in $[n/2, 3n/2]$. Thus, if n is sufficiently large, then there are at least

$$\sum_{a=n/2}^{3n/2} \left[\frac{3n-2a}{2m-2} - 1 \right] = \frac{n^2+n}{m-1} - n > 0.9\frac{n^2}{m}$$

m-APs in $[n/2, 3n/2]$. □

Proof of Theorem 4.1. For fixed n, let $m = m(n)$ and let $\mathcal{G}$ be the family of $2m$-subsets C of $[n]$ that admits a labeling $C = \{x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_m\}$ for which there are $a,\ d \in [2n]$ such that $x_i + y_i = a + id$ for every $i \in [m]$. Note that there could be some sets C that admit multiple labelings that give rise to the same or distinct arithmetic progressions. As we are considering the family of sets and not labelings, those will be accounted for only once in $\mathcal{G}$.

Furthermore, let $\mathcal{F} \subseteq \mathcal{P}([n])$ be the family of subsets B that contain a subset $C \in \mathcal{G}$. Observe that $\mathcal{G}$ is a cover of $\mathcal{F}$.

Finally, let

$$q(n) = r(n)n^{-0.5-1/m},$$

where the function $r : \mathbb{Z}_{>0} \to \mathbb{R}$ is chosen according to the regime covered by m:

1. If $m < 0.48\sqrt{\log n / \log\log n}$, let $r = \sqrt{228}$;
2. otherwise, let $r = e^{9.9}\sqrt{m}$.

Now, because $m \geq 3$, $r \geq \sqrt{65}$ and $m < 0.1 \log n$ we have the following inequalities:

$$\begin{cases} \dfrac{80}{9}\dfrac{m^5 5^m}{n^{2+m}q^{2m}} = \dfrac{80}{9}\dfrac{m^5 5^m}{r^{2m}} < 1, & (4.1) \\[2ex] \log r \leq \dfrac{\log m}{2} + 9.9 < \dfrac{\log m}{2} + \dfrac{\log n}{m}. & (4.2) \end{cases}$$

These conditions are chosen so as to be possible to prove that $\mathcal{F}$ is not q-small and therefore we can use Theorem 3.7.

Claim 4.3. *We have*

$$\sum_{C \in \mathcal{G}} q^{|C|} > \frac{9}{160}\frac{n^{2+m}q^{2m}}{m^5 5^m}$$

for sufficiently large n.

Proof. Notice that for each $C \in \mathcal{G}$ there are at most $(2m)^4$ ways of labeling four elements of C as x_1, x_2, y_1, y_2. For each choice of x_1, x_2, y_1, y_2 there is at most one m-AP that it generates (the one whose first two elements are $x_1 + y_1$ and $x_2 + y_2$).

Additionally, for each m-AP $R = \{a, a + d, \dots, a + (m-1)d\}$ in $[n/2, 3n/2]$ where $m^2 < n/40$, there are at least $(n/5)^m$ subsets $C = \{x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_m\}$ of $[n]$ such that $x_i + y_i = a + id$ for each $i \in [m]$ and each pair $\{x_i, y_i\}$ is uniquely defined, if at all, by C and R. One can achieve this count by constructing C using the following greedy algorithm:

Let $C_0 = \emptyset$ and $A_1 = [n]$. The sets C_i will grow to become C and the sets A_i will be the sets elements that can be added to C_{i-1}.

On step i, where $1 \le i \le m$, choose arbitrarily $x_i < y_i$ with $x_i, y_i \in A_i$ such that $x_i + y_i = a + d(i-1)$, and let

$$C_i = C_{i-1} \cup \{x_i, y_i\}.$$

Also let

$$A_{i+1} = A_i \setminus \{z \in [n] : z + x_i \in R \vee z + y_i \in R\}.$$

Finally let $C = C_m$.

Notice that at each step of this algorithm, the set A_i decreases in at most $2m$ elements and for each $r \in R$ there are at least $n/4$ pairs $\{x, y\} \subseteq [n]$ of different numbers such that $r = x + y$. So at each step there are at least $n/4 - 2m^2 > n/5$ possible choices for $\{x_i, y_i\}$.

Finally, if n is large enough, then, by Lemma 4.2 and because $m = o(\log n)$, there are at least $0.9n^2/m$ m-APs in $[n/2, 3n/2]$ and $m^2 < n/40$. Thus,

$$\sum_{C \in \mathcal{G}} q^{|C|} \ge \sum_{\substack{R \text{ is an } m\text{-AP} \\ R \subseteq [n/2, 3n/2]}} \frac{1}{(2m)^4} \left(\frac{n}{5}\right)^m q^{2m} > 0.9 \frac{n^2}{m} \frac{1}{16m^4} \left(\frac{nq^2}{5}\right)^m = \frac{9}{160} \frac{n^{2+m} q^{2m}}{m^5 5^m}.$$

□

Claim 4.4. *The family $\mathcal{F}$ is not q-small for sufficiently large n.*

Proof. Firstly, notice that any cover $\mathcal{H}$ of $\mathcal{F}$ that contains the empty set satisfies

$$\sum_{C \in \mathcal{H}} q^{|C|} \ge q^0 = 1 > \frac{1}{2}.$$

Suppose that there exists a cover $\mathcal{H}$ of $\mathcal{F}$ that does not contain the empty set and satisfies

$$\sum_{D \in \mathcal{H} \setminus \mathcal{G}} q^{|D|} + \frac{80}{9} \frac{m^5 5^m}{n^{2+m} q^{2m}} \sum_{C \in \mathcal{H} \cap \mathcal{G}} q^{|C|} < \frac{80}{9} \frac{m^5 5^m}{n^{2+m} q^{2m}} \sum_{C \in \mathcal{G}} q^{|C|}. \tag{4.3}$$

We may suppose that $|\mathcal{H} \setminus \mathcal{G}|$ is as small as possible. Then, for an arbitrary set $E = \{e_1, e_2, \dots, e_t\} \in \mathcal{H} \setminus \mathcal{G}$ (such a set must exist otherwise inequality (4.3) does not hold), let

$$\mathcal{G}_E := \{C \in \mathcal{G} : E \subsetneq C\}$$

and

$$\mathcal{H}' := (\mathcal{H} \cup \mathcal{G}_E) \setminus \{E\}.$$

Notice that $\mathcal{H}'$ is a cover of $\mathcal{G}$, and therefore a cover of $\mathcal{F}$, and $|\mathcal{H}' \setminus \mathcal{G}| = |\mathcal{H} \setminus \mathcal{G}| - 1$. Then, by the definition of $\mathcal{H}$,

$$\sum_{D \in \mathcal{H} \setminus \mathcal{G}} q^{|D|} + \frac{80}{9} \frac{m^5 5^m}{n^{2+m} q^{2m}} \sum_{C \in \mathcal{H} \cap \mathcal{G}} q^{|C|} < \frac{80}{9} \frac{m^5 5^m}{n^{2+m} q^{2m}} \sum_{C \in \mathcal{G}} q^{|C|} \leq \sum_{D \in \mathcal{H}' \setminus \mathcal{G}} q^{|D|} + \frac{80}{9} \frac{m^5 5^m}{n^{2+m} q^{2m}} \sum_{C \in \mathcal{H}' \cap \mathcal{G}} q^{|C|}.$$

Hence

$$0 < \left[\sum_{D \in \mathcal{H}' \setminus \mathcal{G}} q^{|D|} + \frac{80}{9} \frac{m^5 5^m}{n^{2+m} q^{2m}} \sum_{C \in \mathcal{H}' \cap \mathcal{G}} q^{|C|} \right] - \left[\sum_{D \in \mathcal{H} \setminus \mathcal{G}} q^{|D|} + \frac{80}{9} \frac{m^5 5^m}{n^{2+m} q^{2m}} \sum_{C \in \mathcal{H} \cap \mathcal{G}} q^{|C|} \right] \leq \frac{80}{9} \frac{m^5 5^m}{n^{2+m}} |\mathcal{G}_E| - q^t.$$

Which in turn implies that

$$|\mathcal{G}_E| > \frac{9}{80} \frac{n^{2+m} q^t}{m^5 5^m}. \tag{4.4}$$

Notice that if $t \geq 2m$, then $\mathcal{G}_E = \emptyset$, as each set in $\mathcal{G}_E$ has $2m$ elements and is a proper superset of E that also contains t elements. This contradicts inequality (4.4). Therefore, $t \leq 2m - 1$. Also notice that $t > 0$, by the definition of $\mathcal{H}$.

Define functions $\mathcal{A}, \mathcal{B}, \mathcal{C} : \mathcal{G}_E \to \mathcal{P}([m])$ as follows: for each $C \in \mathcal{G}_E$, let $C = \{x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_m\}$ be an arbitrary possible element labeling of C such that there exists a and d in $[n]$ for which

$$x_i + y_i = a + id \in [2n] \text{ for all } i \in [m]. \tag{4.5}$$

Then let

$$\begin{cases} \mathcal{A}(C) := \{i \in [m] : \{x_i, y_i\} \subseteq E\}; \\ \mathcal{B}(C) := \{i \in [m] : |\{x_i, y_i\} \cap E| = 1\}; \\ \mathcal{C}(C) := \{i \in [m] : \{x_i, y_i\} \cap E = \emptyset\}. \end{cases}$$

Note that $\mathcal{A}(C)$, $\mathcal{B}(C)$, $\mathcal{C}(C)$ define a partition of $[m]$.

Notice also that it is possible to redefine the labeling of the elements of C without changing the condition given by Equation (4.5) by swapping x_i and y_i for any $i \in \mathcal{B}$ in which $\{x_i, y_i\} \cap E = \{y_i\}$. We shall only consider element labelings of C such that if $i \in \mathcal{B}(C)$, then $\{x_i, y_i\} \cap E = \{x_i\}$.

Finally define

$$\alpha(C) := |\mathcal{A}(C)|, \ \beta(C) := |\mathcal{B}(C)|, \ \gamma(C) := |\mathcal{C}(C)|.$$

Notice that

$$\begin{cases} t = |E| = 2\alpha + \beta \\ m = \alpha + \beta + \gamma, \end{cases}$$

which in turn implies that

$$\begin{cases} \beta = t - 2\alpha \\ \gamma = m - t + \alpha. \end{cases}$$

Now, for each $i \in \{0, 1\}$, define

$$\mathcal{G}_E^i := \{C \in \mathcal{G}_E : \alpha(C) = i\}.$$

Also define

$$\mathcal{G}_E^{\geq 2} := \{C \in \mathcal{G}_E : \alpha(C) \geq 2\}.$$

We claim the following estimates of the sizes of sets $\mathcal{G}_E^0$, $\mathcal{G}_E^1$ and $\mathcal{G}_E^{\leq 2}$, whose proofs involve some double counting and can be found in the end of this chapter.

$$\begin{cases} |\mathcal{G}_E^0| \leq \dfrac{2n^2 - n}{m-1}\dbinom{m}{t} t! \left(\dfrac{n}{2}\right)^{m-t} & (4.6) \\ |\mathcal{G}_E^1| \leq m\dbinom{t}{2}\dfrac{2n}{m-1}\dbinom{m-1}{t-2}(t-2)! \left(\dfrac{n}{2}\right)^{m-t+1} & (4.7) \\ |\mathcal{G}_E^{\geq 2}| \leq \displaystyle\sum_{\alpha=2}^{t/2} \binom{m}{\alpha}\binom{t}{2, 2, t-4}\binom{m-\alpha}{t-2\alpha}(t-2\alpha)! \left(\frac{n}{2}\right)^{m-t+\alpha} & (4.8) \end{cases}$$

Observe that

$$\frac{n^{2+m}q^t}{m^5 5^m} = \frac{r^t}{m^5 5^m} n^{2+m-0.5t-t/m}.$$

Moreover, notice that

$$\left[m^t n^{2+m-t}\right]\left[\frac{m^5 5^m}{n^{2+m}q^t}\right] = n^{t/m-t/2}\frac{m^{5+t}5^m}{r^t} \leq n^{-t/6}\frac{m^{5+t}5^m}{r^t} < n^{-.166+o(1)+0.161} = o(1), \tag{4.9}$$

since $m \geq 3$, $m/r = n^{o(1)}$, $t \geq 1$ and $m < 0.1 \log n$.

Then, Inequalities (4.6) and (4.9) imply that

$$|\mathcal{G}_E^0| < 2m^t n^{2+m-t} = o\left(\frac{n^{2+m}q^t}{m^5 5^m}\right). \tag{4.10}$$

Also, Inequalities (4.7) and (4.9) imply that

$$|\mathcal{G}_E^1| < 2m^{t-2}t^2 n^{m-t+2} < 8m^t n^{m-t+2} = o\left(\frac{n^{2+m}q^t}{m^5 5^m}\right). \tag{4.11}$$

Finally, Inequality (4.8) implies that

$$|\mathcal{G}_E^{\geq 2}| \leq t^4 m^t n^{m-t}\sum_{\alpha=2}^{t/2}\left(\frac{n}{m}\right)^{\alpha} < t^5 n^{m-t/2} m^{t/2} < 32m^{0.5t+5}n^{m-t/2}. \tag{4.12}$$

Let $\lambda = 288 = 32 \cdot 9$. We analyze two cases

1. If $m < 0.48\sqrt{\log n / \log\log n}$ and $r = \sqrt{228}$, then

$$\frac{\log n}{m} + 2m\log r > 4.2m\log m + 2m\log r > m\log 5m + 9.5\log m + \log\lambda + \log r.$$

2. If $m < 0.1\log n$ and $r = e^{9.9}\sqrt{m}$, then

$$(2m-1)\log r = m\log m + 19.8m - \frac{\log m}{2} - 9.9 \geq m\log 5m + 9.5\log m + \log\lambda.$$

Either way, we have

$$\frac{\log n}{m} + 2m\log r \geq m\log 5m + 9.5\log m + \log\lambda + \log r$$

Which is equivalent to

$$2\log n - m\log 5 - 10\log m - \log\lambda \geq (2m-1)\left(\frac{\log m}{2} + \frac{\log n}{m} - \log r\right).$$

Then, Condition (4.2) and the fact that $t \leq 2m-1$ yields

$$t\log r + \left(2 - \frac{t}{m}\right)\log n \geq \log\lambda + m\log 5 + \left(\frac{t}{2} + 10\right)\log m.$$

Finally, recalling Equation(4), this inequality combined with Inequality (4.12) yields

$$|\mathcal{G}_E^{\geq 2}| < 32m^{0.5t+5}n^{m-t/2} \leq \frac{9}{81}\frac{n^{2+m}q^t}{m^5 5^m}. \tag{4.13}$$

Finally, Inequalities (4.10), (4.11) and (4.13) contradict Inequality (4.4), for sufficiently large n, which in turn contradicts the existence of $\mathcal{H}$.

Therefore, because no such $\mathcal{H}$ exists, by Inequality (4.1) and Claim 4.3 we have

$$\sum_{C\in\mathcal{G}'} q^{|C|} \geq \sum_{D\in\mathcal{G}'\setminus\mathcal{G}} q^{|D|} + \frac{80}{9}\frac{m^5 5^m}{n^{2+m}q^{2m}}\sum_{C\in\mathcal{G}'\cap\mathcal{G}} q^{|C|} \geq \frac{80}{9}\frac{m^5 5^m}{n^{2+m}q^{2m}}\sum_{C\in\mathcal{G}} q^{|C|} > \frac{1}{2},$$

for every cover $\mathcal{G}'$ of $\mathcal{F}$ that does not contain the empty set. This completes the proof of Claim 4.4. □

The proof of Theorem 4.1 follows from Corollary 3.8 and Claims 4.3 and 4.4, since all minimal elements of $\mathcal{F}$ have size $2m$. □

We now prove Inequalities (4.6)–(4.8), as promised.

Proof of Inequality (4.6). First, note that, by Lemma 2.6, the *m-AP*

$$\{r_i = x_i + y_i \,:\, i \in [m]\}$$

is one of up to $(2n^2 - n)/(m-1)$ possibilities.

Notice also that, since $\alpha = 0$, we have $\beta = t - 2\alpha = t$ and therefore $\mathcal{B}$ is one of the

$$\binom{m}{t}$$

t-subsets of $[m]$. Additionaly, since $\mathcal{A} = \emptyset$ we also have

$$E = \{x_i : i \in \mathcal{B}\},$$

which limits the labeling of the elements of $E \subsetneq C$ to the $t!$ bijections of E to $\mathcal{B}$. Given such a labeling of the elements of E and the m-AP $\{r_i : i \in [m]\}$, there is, for each $i \in \mathcal{B}$, at most one choice for y_i, that is $r_i - x_i$.

Finally observe that for each $i \in \mathcal{C}$, we have $r_i \in [2n]$ and therefore there are at most $n/2$ choices of $\{x, y\} \subseteq [n]$ such that $x + y = r_i$. Since there are $m - t + \alpha = m - t$ indices in $\mathcal{C}$, there are at most $(n/2)^{m-t}$ possible choices for the set $\{x_i : i \in \mathcal{C}\} \cup \{y_i : i \in \mathcal{C}\}$.

The set C can then only be

$$E \cup \{y_i : i \in \mathcal{B}\} \cup \{x_i : i \in \mathcal{C}\} \cup \{y_i : i \in \mathcal{C}\}.$$

Therefore

$$|\mathcal{G}_E^0| \leq \frac{2n^2 - n}{m-1}\binom{m}{t} t! \left(\frac{n}{2}\right)^{m-t}.$$

□

Proof of Inequality (4.7). First, note that $\mathcal{A}$ has a single element j, and is therefore one of m possible 1-subsets of $[m]$.

Furthermore $\{x_j, y_j\}$ can be any of the $\binom{t}{2}$ 2-subsets of E and there are at most $2n/(m-1)$ possible values for the common difference d. Now, given j, $x_j + y_j$ and d, the m-AP

$$\{r_i = x_i + y_i : i \in [m]\}$$

is uniquely defined, if at all.

Notice also that, since $\alpha = 1$, we have $\beta = t - 2\alpha = t - 2$ and therefore $\mathcal{B}$ is one of the

$$\binom{m-1}{t-2}$$

$(t-2)$-subsets of $[m] \setminus \mathcal{A}$. Additionally,

$$E^* = E \setminus \{x_j, y_j\} = \{x_i : i \in \mathcal{B}\},$$

which limits the labeling of the elements of E^* to the $(t-2)!$ bijections of E^* to $\mathcal{B}$. Given such a labeling of the elements of E^* and the m-AP $\{r_i : i \in [m]\}$, there is, for each $i \in \mathcal{B}$, at most one choice for y_i, that is $r_i - x_i$.

Finally observe that for each $i \in \mathcal{C}$, we have $r_i \in [2n]$ and therefore there are at most $n/2$ choices of $\{x, y\} \subseteq [n]$ such that $x + y = r_i$. Since there are $m - t + \alpha = m - t + 1$ indices in $\mathcal{C}$, there are at most $(n/2)^{m-t+1}$ possible choices for the set $\{x_i : i \in \mathcal{C}\} \cup \{y_i : i \in \mathcal{C}\}$.

The set C can then only be

$$E \cup \{y_i : i \in \mathcal{B}\} \cup \{x_i : i \in \mathcal{C}\} \cup \{y_i : i \in \mathcal{C}\}.$$

Therefore

$$|\mathcal{G}_E^1| \leq m\binom{t}{2}\frac{2n}{m-1}\binom{m-1}{t-2}(t-2)!\left(\frac{n}{2}\right)^{m-t+1}.$$

□

Proof of Inequality (4.8). First, note that, for a fixed $2 \leq \alpha \leq t/2$, the set $\mathcal{A}$ is one of

$$\binom{m}{\alpha}$$

α-subsets of $[m]$.

Furthermore, let $j < k$ be the least elements of $\mathcal{A}$. Then the ordered pair $(\{x_j, y_j\}, \{x_k, y_k\})$ is one of

$$\binom{t}{2, 2, t-4}$$

possible elements of $\binom{E}{2}^2$.

Now, given j, k, $x_j + y_j$ and d, the m-AP

$$\{r_i = x_i + y_i : i \in [m]\}$$

is uniquely defined, if at all. Notice also that, we have $\beta = t - 2\alpha$ and therefore $\mathcal{B}$ is one of the

$$\binom{m-\alpha}{t-2\alpha}$$

$(t-2\alpha)$-subsets of $[m] \setminus \mathcal{A}$. Additionally

$$E^{**} = \{x_i : i \in \mathcal{B}\}$$

is a $(t-2\alpha)$-subset of $E \setminus \{x_j, y_j, x_k, y_k\}$ and, for a fixed set E^{**}, the labeling of the elements of E^{**} is limited to the $(t-2\alpha)!$ bijections from E^{**} to $\mathcal{B}$.

Given such a labeling of the elements of E^{**} and the m-AP $\{r_i : i \in [m]\}$, there is, for each $i \in \mathcal{B}$, at most one choice for y_i, that is $r_i - x_i$.

Finally observe that for each $i \in \mathcal{C}$, we have $r_i \in [2m]$ and therefore there are at most $n/2$ choices of $\{x, y\} \subseteq [n]$ such that $x + y = r_i$. Since there are $m - t + \alpha$ indices in $\mathcal{C}$, there are at most $(n/2)^{m-t+\alpha}$ possible choices for the set $\{x_i : i \in \mathcal{C}\} \cup \{y_i : i \in \mathcal{C}\}$.

The set C can then only be

$$E \cup \{y_i : i \in \mathcal{B}\} \cup \{x_i : i \in \mathcal{C}\} \cup \{y_i : i \in \mathcal{C}\}.$$

Therefore

$$|\mathcal{G}_E^{\geq 2}| \leq \sum_{\alpha=2}^{t/2} \binom{m}{\alpha} \binom{t}{2, 2, t-4} \binom{m-4}{t-2\alpha} (t-2\alpha)! \left(\frac{n}{2}\right)^{m-t+\alpha}.$$

$\square$

Chapter 5

Finding Long Arithmetic Progressions in the Sumset of a Random Set Using the First Moment Method

In this chapter we prove Theorems 1.15 and 1.16. An outline of the proof is as follows. First, sets of consecutive numbers are arithmetic progressions and one can divide an interval of km integers into k disjoint sets of m consecutive integers. Hence it suffices that the number of elements ℓ in this run of km integers that misses the set $A_n + A_n$ is less than k in order for an arithmetic progression of m elements to be contained in $A_n + A_n$. We show that for a carefully chosen interval of km numbers, the expected value of ℓ is $o(k)$ and therefore by Markov's inequality $A_n + A_n$ contains an m-AP.

Proof of Theorems 1.15 and 1.16. For each $i \in [2n]$, consider the random variable $X_i = \mathbb{1}[i \notin A_n + A_n]$. Let $t = \delta n$, where $\delta \in (0, 1)$ will be chosen further down the proof. Define

$$B := [t + 2, 2n - t] \setminus (A_n + A_n), \tag{5.1}$$

the set of integers in the interval $[t + 2, 2n - t]$ that cannot be represented as a sum of two elements of A_n.

For $t + 2 \leq i \leq n$, one can find upper bounds for $\mathbb{E}[X_i]$, namely

$$\mathbb{E}[X_i] = \mathbb{E}[X_{2n+2-i}] \leq (1 - p^2)^{i/2-1} \leq (1 - p^2)^{t/2} \leq e^{-p^2 t/2}. \tag{5.2}$$

Now looking at the random variable

$$|B| = \sum_{i=t+2}^{2n-t} X_i, \tag{5.3}$$

we see that

$$\mathbb{E}[|B|] < 2n(1-\delta)e^{-p^2\delta n/2}. \tag{5.4}$$

If $\delta \in (0,1)$ and $m = m(n)$ are such that

$$(2n(1-\delta)/m \geq 1) \wedge (m = o(e^{p^2\delta n/2})), \tag{5.5}$$

then we can choose $2n(1-\delta)/m$ disjoint intervals I_s of size m in the interval $[t+2, 2n-t]$. If we let b be the number of such intervals that are not disjoint to B, then b is at most $|B|$. By Markov's inequality we also have

$$\mathbb{P}[b = 2n(1-\delta)/m] \leq \mathbb{P}[|B| \geq 2n(1-\delta)/m] \leq \frac{\mathbb{E}[|B|]}{2n(1-\delta)/m} < me^{-p^2\delta n/2} = o(1). \tag{5.6}$$

This means that at least one such interval I_s is a subset of $A_n + A_n$ with probability $1 - o(1)$. Clearly I_s is an m-AP.

Now we choose δ and m appropriate for each case.

If $p(n) < \sqrt{2(\log n)/n}$, then $\delta < 1$ and $m = e^{(\delta-1/2)p^2n}$ satisfy (5.5). Therefore $L(A_n + A_n) \geq e^{(1/2-o(1))p^2n}$.

If $p(n) = \sqrt{(C+o(1))(\log n)/n}$ for some constant $C > 2$, then δ constant in the interval $(2/C, 1)$ and $m = 2n(1-\delta)$ satisfy (5.5). $\square$

Chapter 6

Finding Long Arithmetic Progressions in the Sumset of a Random Set Using the Second Moment Method

In this chapter we introduce Theorem 6.1 that is proved using the second moment method. The proof of Theorem 6.1 is inspired by the proof of Theorem 1.3 given by Croot, Ruzsa and Schoen [3]. We recall that we later found Theorem 6.1 to provide weaker bounds than Theorem 1.15, which was proved in Chapter 5.

Theorem 6.1. *Let $p : \mathbb{N} \to [0,1]$ and $m : \mathbb{N} \to \mathbb{R}$ be given, such that $m = o(\sqrt[4]{n})$, $m = o(\sqrt{pn})$, $m = o(p^2 n)$ and $m(n) \to \infty$. Then the random set A_n defined in (1.1) satisfies*

$$L(A_n + A_n) \geq m(n)$$

asymptotically almost surely.

It is simple to deduce the following corollary of Theorem 6.1

Corollary 6.2. *Let $f, m : \mathbb{N} \to \mathbb{R}_{>0}$ be given, such that $f(n) = o(\log n)$, $f(n) \to \infty$ and $m(n) = o(f(n))$. Set $p(n) = \min(1, \sqrt{f(n)/n})$. Then the random set A_n defined in (1.1) satisfies*

$$L(A_n + A_n) > m(n)$$

asymptotically almost surely.

The outline of the proof of Theorem 6.1 is the following. We show that, asymptotically almost surely, the fixed arithmetic progression $X = \{n-d, n-2d, \ldots, n-md\}$, where $d = \sqrt{n}$, is a subset of $A_n + A_n$. This will done by considering a class $\mathcal{C}$ of sets C of $2d$ elements that generate X (i.e. such that $C + C \supset X$) that are

(P1) plentiful so as to force the expected number of members of $\mathcal{C}$ that are subsets of A_n to be large and

(P2) somewhat independent from the other members of C so as to force the standard deviation of the number of members of C that are subsets of A_n to be small in comparison with its expected value.

This allows us to employ the second moment method to achieve what we need. The following definition will establish C, while Lemmas 6.4 and 6.6 will show that C obeys (P1) and (P2), respectively.

Definition 6.3 ((d,m,n)-dsoc)**.** *For positive integers $m < d < n$, say that a vector $x = (x_1, x_2, \ldots, x_m) \in [n]^m$ is a* diverse modulo d second order cover of an m-arithmetic progression in $[n]$ *(or* (d,m,n)-dsoc *or, when there are no ambiguities, simply dsoc), if*

$$\begin{cases} x_i \not\equiv x_j & \pmod{d}, \\ x_i \not\equiv n - x_j & \pmod{d}, \\ x_i < n - md \end{cases}$$

for all $1 \leq i \neq j \leq m$.

Notice that

$$C_x = \{x_i \,:\, i \in [m]\} \cup \{n - id - x_i \,:\, i \in [m]\}$$

is a $2m$-subset of $[n]$ if x is a dsoc and $\{n-d, n-2d, \ldots, n-md\}$ is an m-AP in $C_x + C_x$.

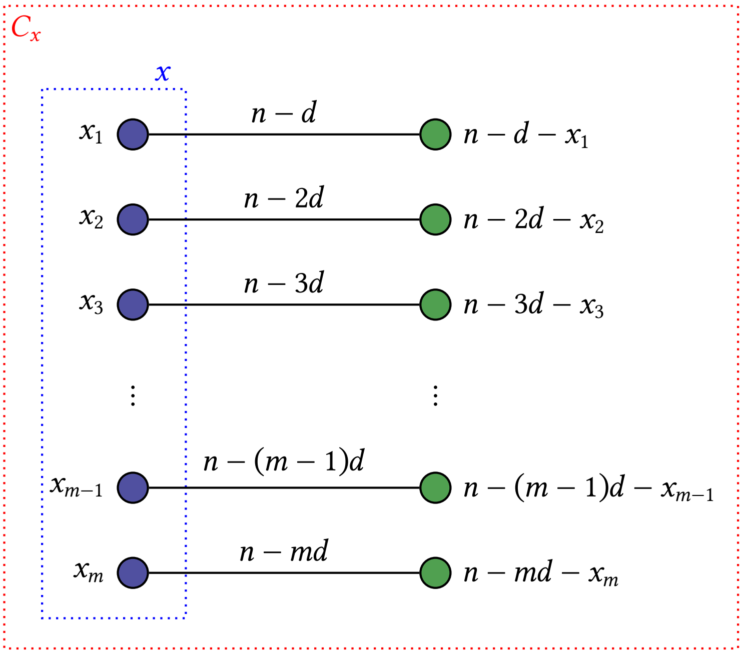

Figure 6.1: *C_x covers exactly $2m$ classes modulo d.*

Lemma 6.4. *There are at least $(n/d - m - 1)^m \prod_{i=1}^{m}(d - 2i + 2)$ dsocs.*

Proof. For fixed $x_1, x_2, \ldots, x_{i-1}$ that can possibly be the first $i-1$ first entries of a dsoc there are at least $d - 2i + 2$ and $n/d - m - 1$ possible remainders and quotients of x_i on its division by n, respectively. □

Definition 6.5. *Let $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_m)$ be dsocs. Define*

$$\alpha(x, y) := \left|\{i \in [m] : |\{x_i, n - id - x_i\} \cap C_y| = 2\}\right|$$

and

$$\beta(x, y) := \left|\{i \in [m] : |\{x_i, n - id - x_i\} \cap C_y| = 1\}\right|.$$

Notice that, if x and y are dsocs, then $\alpha(x, y)$ counts the indices i for which x_i and $n - id - x_i$ are both in C_y. Since y is a dsoc and $x_i + n - id - x_i \equiv n \pmod{d}$, those two numbers have to be, in some order, y_j and $n - jd - y_j$ for an integer j. Therefore

$$x_i + n - id - x_i = y_j + n - jd - y_j,$$

which implies that $i = j$. Therefore

$$\alpha(x, y) = \left|\{i \in [m] : x_i = y_i \vee x_i + y_i = n - id\}\right| = \alpha(y, x)$$

and

$$\beta(x, y) = \left|C_x \cap C_y\right| - 2\alpha(x, y) = \beta(y, x).$$

Lemma 6.6. *Let $\alpha \le m$ and $\beta \le m - \alpha$ be non-negative integers and let x be a dsoc. Then the number of dsocs y such that $\alpha(x, y) = \alpha$ and $\beta(x, y) = \beta$ is at most*

$$\binom{m}{\alpha}\binom{m-\alpha}{\beta}^2 \beta! 2^\alpha 4^\beta n^{m-\alpha-\beta}.$$

Proof. Given y as in the statement of the lemma,

$$S = \{i \in [m] : x_i = y_i \vee x_i + y_i = n - id\}$$

is an α-subsets of $[m]$. For a fixed set S, the sets

$$T_x = \{i \in [m] : |\{x_i, n - id - x_i\} \cap C_y| = 1\},\ T_y = \{i \in [m] : |\{y_i, n - id - y_i\} \cap C_x| = 1\}$$

are β-subsets of $[m] \setminus S$.

Observe that there is a natural bijective function $f : T_x \to T_y$ where

$$\left|\{x_i, n - id - x_i\} \cap \{y_{f(i)}, n - f(i)d - y_{f(i)}\}\right| = 1 \text{ for all } i \in T_x,$$

and for fixed T_x and T_y there are $\beta!$ bijective functions $f : T_x \to T_y$.

Finally notice that

$$y_i \in \{x_i, n - id - x_i\} \text{ for all } i \in S$$

and

$$y_{f(i)} \in \{x_i, n - id - x_i, n - f(i)d - x_i, x_i + (i - f(i))d\} \text{ for all } i \in T_x.$$

Now, for fixed x, S, T_x, T_y and f, each of the α entries of y indexed by elements of S is one of 2 values, each of the β entries of y indexed by elements of T_y is one of at most 4 values, and each of the other $m - \alpha - \beta$ entries of y is one of the n elements of $[n]$. □

We are now ready to prove Theorem 6.1.

Proof of Theorem 6.1. We will employ the second moment method. Let $d = \sqrt{n}$. If x is a dsoc, then

$$\mathbb{P}(C_x \subseteq A_n) = p^{2m}.$$

Set

$$X := \sum_{x \text{ dsoc}} \mathbb{1}[C_x \subseteq A_n].$$

Notice that $X \geq 1$ implies that $A_n + A_n$ contains the m-AP $\{n-d, n-2d, \dots, n-md\}$.

Set $T = \left|\{x \in [n]^m : x \text{ dsoc}\}\right|$. Then, using Lemma 6.4, we have

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \text{ dsoc}} p^{2m} = Tp^{2m} \\
&\geq p^{2m}\left(\frac{n}{d} - m - 1\right)^m \prod_{i=1}^{m}(d - 2i + 2) \\
&\geq \left(p^2 n\left(1 - \frac{(m+1)d}{n}\right)\left(1 - \frac{2m-2}{d}\right)\right)^m \\
&\to \infty,
\end{aligned}$$

as $p^2 n \to \infty$, $m \geq 1$, $m = o(d)$ and $m = o(n/d)$.

Then, using Lemma 6.6, one can estimate the variance as follows:

$$\begin{aligned}
\mathrm{Var}[X] &= \sum_{x \text{ dsoc}} \sum_{y \text{ dsoc}} \mathrm{Cov}\left[\mathbb{1}(C_x \subseteq A), \mathbb{1}(C_y \subseteq A)\right] \\
&\leq \sum_{x \text{ dsoc}} \sum_{\alpha=0}^{m} \sum_{\beta=0}^{m-\alpha} \binom{m}{\alpha}\binom{m-\alpha}{\beta}^2 \beta! 2^\alpha 4^\beta n^{m-\alpha-\beta}(p^{4m-2\alpha-\beta} - p^{4m}) \\
&= \sum_{x \text{ dsoc}} \sum_{\substack{0 \leq \beta \leq m-\alpha \leq m \\ (\alpha,\beta) \neq (0,0)}} \binom{m}{\alpha}\binom{m-\alpha}{\beta}^2 \beta! 2^\alpha 4^\beta n^{m-\alpha-\beta}(p^{4m-2\alpha-\beta} - p^{4m}) \\
&< Tp^{4m} \sum_{\substack{0 \leq \beta \leq m-\alpha \leq m \\ (\alpha,\beta) \neq (0,0)}} \binom{m}{\alpha}\binom{m-\alpha}{\beta}^2 \beta! 2^\alpha 4^\beta n^{m-\alpha-\beta} p^{-2\alpha-\beta}.
\end{aligned}$$

Setting

$$\tau_{\alpha,\beta} = \binom{m}{\alpha}\binom{m-\alpha}{\beta}^2 \beta! 2^\alpha 4^\beta n^{m-\alpha-\beta} p^{-2\alpha-\beta},$$

we have for all $0 \leq \alpha \leq m$, $0 \leq \beta \leq m - \alpha - 1$ and n sufficiently large that

$$\frac{\tau_{\alpha,\beta+1}}{\tau_{\alpha,\beta}} = \frac{4(m-\alpha-\beta)^2}{(\beta+1)np} \leq \frac{4m^2}{np} < \frac{1}{2},$$

as $m^2 = o(np)$.

Using this geometric behaviour and Lemma 6.4, we have, for some constant $C > 0$, that

$$\begin{aligned}
\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2} &\leq \frac{1}{T}\left(\sum_{\beta=1}^{m}\tau_{0,\beta} + \sum_{\alpha=1}^{m}\sum_{\beta=0}^{m-\alpha}\tau_{\alpha,\beta}\right) \\
&< \frac{2}{n^m\left(1-\frac{(m+1)d}{n}\right)^m\left(1-\frac{2m-2}{d}\right)^m}\left(\tau_{0,1} + \sum_{\alpha=1}^{m}\tau_{\alpha,0}\right) \\
&< \frac{2.01}{n^m(1/e)^{(m^2+m)d/n+2(m^2-m)/d}}\left(4m^2n^{m-1}p^{-1} + \sum_{\alpha=1}^{m}\binom{m}{\alpha}n^{m-\alpha}2^\alpha p^{-2\alpha}\right) \\
&< \frac{2.01}{n^m(1/e)^{2m^2(d/n+1/d)}}\left(4m^2n^{m-1}p^{-1} + n^m\left(\left(1+\frac{2}{p^2n}\right)^m - 1\right)\right) \\
&< \frac{Cm^2}{np} + C\left(\left(1+\frac{2}{p^2n}\right)^m - 1\right) = o(1)
\end{aligned}$$

for sufficiently large n, since $m^2 = o(n/d)$, $m^2 = o(d)$, $m^2 = o(np)$ and $m = o(p^2n)$. Now, by Chebyshev's inequality

$$\mathbb{P}[X \geq 1] \to 1$$

as n tends to infinity. $\square$

Chapter 7

Concluding Remarks

We begin by comparing Theorem 1.18 in this dissertation with the deterministic theorems presented in the first chapter. Theorems 1.18 deals with probabilities that are $n^{-1/2+o(1)}$, which allows us to use concentration inequalities such as Chernoff's bounds to prove that $|A_n| = (1 + o(1))pn$ asymptotically almost surely. In this regime Theorem 1.2 and Corollary 1.4 can at best guarantee arithmetic progressions of size 3 in the random set $A_n + A_n$.

We also offer the following possible generalizations of the problem studied in this dissertation. We hope to tackle these problems in a near future.

Problem 7.1. *Let* $p, q : \mathbb{N} \to [0, 1]$. *We consider the independent sequences of independent random sets* $\{A_n \subseteq [n]\}_{n\in\mathbb{N}}$ *and* $\{B_n \subseteq [n]\}_{n\in\mathbb{N}}$, *where for all* $i \in [n]$ *we have*

$$\mathbb{P}[i \in A_n] = p(n) \text{ and } \mathbb{P}[i \in B_n] = q(n)$$

and these $2n$ *events are mutually independent.*

(a) What can we say about the asymptotic behavior of $L(A_n + B_n)$*?*

(b) If $p = q$ *is the asymptotic behavior of* $L(A_n + B_n)$ *similar to the one of* $L(A_n + A_n)$*?*

(c) What are non-trivial upper bounds for $L(A_n + A_n)$ *if* $p = 1/\sqrt{n}$*?*

Problem 7.2. *Let* $p : \mathbb{N} \to [0, 1]$. *Consider the sequence of independent random sets* $\{A_n \subseteq [n]\}_{n\in\mathbb{N}}$, *where for all* $i \in [n]$ *we have*

$$\mathbb{P}[i \in A_n] = p(n)$$

and these events are mutually independent. What can we say about the asymptotic behavior of $L(kA_n)$*? Here* kA_n *denotes the set of sums of* k *not necessarily distinct elements of* A_n.

We feel that in Problems 7.1(a) and 7.1(b) the techniques used in the proofs of the theorems of this dissertation can be slightly modified to prove similar theorems.

Meanwhile in Problem 7.2 the alterations in the proofs needed to find similar theorems to the ones in this dissertations should be more sophisticated, specially the ones regarding upper bounds for $L(kA_n)$. The case $k = 3$ already seems interesting.

Problem 7.1(c) is motivated by the fact that we could not find upper bounds for $L(A_n + A_n)$, other than the one given by the fact that $A_n + A_n \subseteq [2n]$.

References

[1] B. Bollobás and A. Thomason. "Threshold functions". In: *Combinatorica* 7.1 (1987), pp. 35–38. ISSN: 0209-9683. DOI: 10.1007/BF02579198. URL: http://dx.doi.org/10.1007/BF02579198 (cit. on p. 5).

[2] Jean Bourgain. "On arithmetic progressions in sums of sets of integers". In: *A tribute to Paul Erdős*. Cambridge Univ. Press, Cambridge, 1990, pp. 105–109 (cit. on p. 3).

[3] Ernie Croot, Imre Ruzsa, and Tomasz Schoen. "Arithmetic progressions in sparse sumsets". In: *Combinatorial number theory*. de Gruyter, Berlin, 2007, pp. 157–164 (cit. on pp. 3, 35).

[4] Ben Green. "Arithmetic progressions in sumsets". In: *Geom. Funct. Anal.* 12.3 (2002), pp. 584–597. ISSN: 1016-443X. DOI: 10.1007/s00039-002-8258-4. URL: https://doi.org/10.1007/s00039-002-8258-4 (cit. on p. 3).

[5] Jeff Kahn and Gil Kalai. "Thresholds and expectation thresholds". In: *Combin. Probab. Comput.* 16.3 (2007), pp. 495–502. ISSN: 0963-5483. DOI: 10.1017/S0963548307008474. URL: https://doi.org/10.1017/S0963548307008474 (cit. on p. 20).

[6] Jinyoung Park and Huy Tuan Pham. "A proof of the Kahn-Kalai conjecture". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science—FOCS 2022*. IEEE Computer Soc., Los Alamitos, CA, [2022] ©2022, pp. 636–639 (cit. on pp. 8, 19, 20).