Astérisque

GREGORY A. FREIMAN

Structure theory of set addition

Astérisque, tome 258 (1999), p. 1-33

<http://www.numdam.org/item?id=AST_1999__258__1_0>

Astérisque **258**, 1999, p. 1–33

STRUCTURE THEORY OF SET ADDITION

by

Gregory A. Freiman

Abstract. — We review fundamental results in the so-called structure theory of set addition as well as their applications to other fields.

1. 'Structure theory of set addition'[1] is a shorthand for a direction in the study of sets which extracts structures from sets for which some properties of their sums (or products in a non-abelian case) are known.

Here is an indication of what is meant by "structure". The first stage is to build an equivalence relation on sets. Then, by taking well chosen representatives of an equivalence class we are able to reveal its properties and thereby describe its structure (see, for example, the Definition and Theorem in §6).

2. This review is written in the following way. In §§3–8 we explain the main ideas. In §§9–12 we make some historical remarks. Then in §§13–19 we present several concrete problems in additive and combinatorial number theory, showing how new results may be obtained with the help of the described new approach. Further then in §§20–27 we try to show a diversity of fields where the ideas of "Structure Theory" may be applied. Finally in §§28–35 we discuss methods and problems. In the bibliography we include references to a wider spectrum of subjects which may be treated from the point of view of Structure Theory.

3. This approach to additive problems was originally given the name "Inverse problems of additive number theory". A series of nine papers under this heading was published in 1955–1964 (see [85], [86], [87], [88], [89], [90], [91], [92] and [98]).

4. I quote from my lecture in the Fourth All-Union Mathematical Congress, Leningrad, 3-12 July 1961 (see [84]):

1991 Mathematics Subject Classification. — 11 02, 11Z05.

Key words and phrases. — Structure theory of set addition, inverse problems of additive number theory, small doubling property, isomorphism of subsets.

[1] This paper is based on my review lecture given at the conference on *Structure theory of set addition* held at CIRM (Centre International des Rencontres Scientifiques), Luminy, Marseille, on 10 June 1993.

"The term *inverse problems of additive number theory* appeared in 1955 in two of my papers [85][2] and [86]. In [85] the following problem was studied. Let

$$a_1, a_2, \dots, a_r, \dots \tag{1}$$

be an unbounded, monotonically increasing sequence of positive numbers. To have an asymptotic formula

$$\log q(u) \sim Au^\alpha, \qquad \text{where } A > 0, 0 < \alpha < 1$$

it is necessary and sufficient that

$$n(u) \sim B(A, \alpha) u^{\alpha/1-\alpha}$$

where $n(u)$ is the number of terms of a sequence (1) not exceeding u, and $q(u)$ is the number of solutions of the inequality

$$a_1 n_1 + a_2 n_2 + \cdots \leq u.$$

In [86] the case

$$\log q(u) = Au^\alpha + O(u_1^\alpha), \qquad \text{where } 0 < \alpha_1 < \alpha,$$

was studied and an estimate of the error term in the asymptotic formula for $n(u)$ was obtained.

One can easily see that if $q(u)$ is known then (1) is determined in a unique way (see [85]). In 'direct' problems we study $q(u)$ when the sequence (1) is given; a particular case is the classical problem on the representation of positive integers as sums of an unlimited number of positive integers.

Thus a direct problem in additive number theory is a problem in which, given summands and some conditions, we discover something about the set of sums. An inverse problem in additive number theory is a problem in which, using some knowledge of the set of sums, we learn something about the set of summands.

Several cases of inverse problems were studied earlier; see [14] and [67].

Paul Erdős, in 1942, found an asymptotic formula for $n(u)$ when

$$\log p(u) \sim a\sqrt{u}$$

where $p(u)$ is the number of solutions of an equation

$$a_1 n_1 + a_2 n_2 + \cdots = u$$

where $\{a_i\}$ is some sequence of positive integers (see [67]).

In the same paper another inverse problem was studied; if $q(u) \sim Cu^{2\alpha}$, where $q(u)$ is the number of solutions of an inequality

$$a_i + a_j \leq u,$$

[2] The reference numbers given accord with the bibliography of this paper and not the original text.

> then
>
> $$n(u) \sim C_1 u^{\alpha}.$$
>
> In 1960 V. Tashbaev [252] studied the problem of estimating the error term for this inverse problem.
>
> We will now explain how problems on the distribution of prime numbers are connected with inverse problems. If we define
>
> $$q(u) = [e^u]$$
>
> then $a_i = \log p_i$, where p_i denotes the i^{th} prime number. Thus the problem of the distribution of prime numbers may be treated as an inverse problem of additive number theory of the type described above. The study of inverse problems for different $q(u)$ close to $[e^u]$, and also of direct problems when $n(u)$ is close to e^u/u, may give some insight into the problem of the distribution of primes, in a way similar to that in which the behaviour of a function in the vicinity of a point may help to find its value at that point (see A.Beurling [14] and B.M.Bredichin [30], [31], [32] and [33]."

The results of Diamond (see [57], [58], [59], [60] and [61]) should of course be mentioned.

The treatment of prime distribution problems as inverse additive problems have not developed up to now. I still consider this approach very hopeful.

5. We pass on now to the study of additive problems with a fixed number of summands. The majority of papers mentioned in §3 treat the addition of two *equal* sets. The study of this particular case is usually sufficient to develop ideas, methods and results as well as their use in applications.

Let us start with $K \subseteq \mathbb{Z}$ with $|K| = k$. Define

$$2K = K + K = \{x \mid x = a_i + a_j, \quad a_i, a_j \in K\}.$$

We may ask the question what is the minimal cardinality of $2K$? Evidently,

$$|2K| \geq 2k - 1\,. \tag{2}$$

Suppose now that K is such that $|2K|$ is minimal i.e. $|2K| = 2k - 1$. What can be said about such a K? It is clear that,

$$|2K| = 2k - 1\,, \tag{3}$$

only if K is an arithmetic progression.

Suppose now that $|K + K|$ is not much greater than this minimal value. In that case we have the following result [87], describing the structure of K.

Theorem 1. — *Let K be a finite set, $K \subseteq \mathbb{Z}$. If*

$$|K + K| \leq 2k - 1 + b, \quad 0 \leq b \leq k - 3$$

then K is contained in an arithmetic progression of length $k + b$.

Further, suppose that we know that

$$|2K| < Ck, \tag{4}$$

where C is any given positive number, we may ask what then is the structure of K?

6. The theorem answering this question (we will quote it as a main theorem) was proved in a previously mentioned series of papers, expositions of it were given in [81] and [82], and an improved version of a proof was presented in [105]. We are citing here the result of Y. Bilu [16], where he studies a case when C in (4) is a slowly growing function of k.

Definition. — Let A and B be groups, and let $K \subset A$ and $L \subset B$. The map ϕ: $K \to L$ is called an $\mathbb{F}_s$–homomorphism, if for any $x_1, \cdots, x_s$ and $y_1, \cdots, y_s$ in K we have

$$x_1 + \cdots + x_s = y_1 + \cdots + y_s \Rightarrow \phi(x_1) + \cdots + \phi(x_s) = \phi(y_1) + \cdots + \phi(y_s).$$

The $\mathbb{F}_s$–homomorphism ϕ is an $\mathbb{F}_s$–isomorphism if it is invertible and the inverse ϕ^{-1} is also an $\mathbb{F}_s$–homomorphism.

Let $P \subset \mathbb{Z}^n$ be given by

$$P = \{0, \ldots, b_1 - 1\} \times \cdots \times \{0, \ldots, b_n - 1\}.$$

We have $|P|$=$b_1 \ldots b_n$. In this paper we will call P an *n–dimensional parallelepiped.*

Theorem 2. — *Let $K \subset \mathbb{Z}$ and suppose that*

$$|K + K| < \sigma k \tag{5}$$

where

$$k = |K| \geq k_0(\sigma) = \frac{[\sigma][\sigma + 1]}{2([\sigma + 1] - \sigma)} + 1\,,$$

then there exists an n–dimensional parallelepiped, P, such that $n \leq [\sigma - 1]$ and $|P| < ck$, where c depends only on σ and s and there also exists a map ϕ: $P \to \mathbb{Z}$ which is such that $P \to \phi(P)$ is an $\mathbb{F}_s$–isomorphism while $K \subset \phi(P)$.

Let us now return to §1. The equivalence relation that we talked about there, is now seen to be $\mathbb{F}_s$–isomorphism. A representative of an equivalence class is an n–dimensional parallelepiped, P. We now understand that K, a subset of the one-dimensional space $\mathbb{R}$, has, in fact, a multidimensional structure, being a dense subset of an n-dimensional set P (i.e. $\phi^{-1}(K) \subset P$). Consider the numbers

$$a = \phi((0, \ldots, 0)),\ a_1 = \phi((1, 0, \ldots, 0)) - a,\ \ldots,\ a_n = \phi((0, 0, \ldots, 1)) - a.$$

Then,

$$\phi(P) = \{a + a_1x_1 + a_2x_2 + \cdots + a_nx_n, \text{ with } 0 \leq x_i \leq b_i - 1\}\,.$$

Imre Rusza has called such a set $\phi(P)$ a generalized arithmetic progression of rank n. He gave a new and shorter proof, based on new ideas, of the main theorem together with an important generalization; in this the summands A and B may be different, although however the condition $|A| = |B|$ is required (see [233]). His generalization to the case of subsets of abelian groups is to be found in [238].

7. We can now describe an "algorithm" for solving an inverse additive problem, by the following steps.

(i) Choose some (usually numerical) characteristic of the set under study.
(ii) Find an extremal value of this characteristic within the framework of the problem that we are studying.
(iii) Study the structure of the set when its characteristic is equal to its extremal value.
(iv) Study the structure of a set when its characteristic is near to its extremal value.
(v) (vi),... continue, taking larger and larger neighbourhoods for the characteristic.

From estimates obtained by Yuri Bilu it follows that in (5) we can take, for σ, the following very slowly growing function of k,

$$\sigma = c \log\log\log\log k.$$

It will be very important to study the cases

$$\sigma = (\log k)^c \tag{6}$$

and

$$\sigma = k^{\varepsilon},\ \varepsilon > 0, \tag{7}$$

even if ε is a very small number.

Here to simplify this extremely difficult problem a little, it is better to take $|rK|$ as a characteristic value, where r is a fixed, positive, but rather large, integer. So our condition is now

$$|rK| < k^{1+\varepsilon}$$

which is much stronger than (5); rK contains k^r sums, but no more than $k^{1+\varepsilon}$ of them are different.

8. I have here added a playful description of the comparative difficulty of the problems discussed, which should not be taken too literally. To prove (2) took one minute. Condition (3) was studied in three minutes. The proof of the theorem of §5 together with the description of K under the condition $|2K| = 3k - 3$ took one month. Proof of the main theorem took five years. I will be very happy if we will see results for (6) in the next thirty years but I am not certain that for (7) we will have satisfactory results even in the next hundred years.

9. L. Schnirelman [242] was one of the first who passed from studying fixed sets to studying general additive properties. Schnirelman introduced the notion of the density of a sequence.

Definition. — Let $A = (a_1, a_2, \dots, a_n, \dots)$ be an increasing sequence of positive integers and further let,

$$A(x) = \bigl|\{y \in A \mid 0 < y \leq x\}\bigr|,$$

and

$$d(A) = \inf_{x \in \mathbb{N}} A(x)/x\,.$$

The number $d(A)$ is called the *Schnirelman density* of the sequence A (see step (i) of §7).

10. Define

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

and denote

$$\alpha=d(A),\ \beta=d(B),\ \gamma=d(A+B)\,.$$

Schnirelman proved that

$$\gamma\geq\alpha+\beta-\alpha\beta\,.$$

L. Schnirelman and E. Landau conjectured in 1932 and Mann [178] has proved in 1942 that

$$\gamma\geq\alpha+\beta\,. \tag{8}$$

11. The famous $\alpha+\beta$ theorem of Mann cannot be improved. Take a sequence

$$A=\{0,1,\dots,r,l+1,l+2,\dots,l+r,2l+1,2l+2,\dots,2l+r,\dots\}$$

It is clear that if $r\leq l$ then,

$$\alpha=d(A)=r/l\,.$$

However if $2r<l$ then

$$\gamma=d(2A)=2r/l=2\alpha\,.$$

But for $A=B$ we always have from (8) that $\gamma\geq2\alpha$. So step 2 of §7 is now completed.

Thus Mann has entirely solved the problem of increase of the density under summation of sequences. Its solution took ten years. Khinchine [151] writes in his book:

> "The problem has become 'fashionable'. Scientific societies proposed a prize for its solution. My friends from England wrote me in 1935 that half of English mathematicians tried to solve it, putting aside all other obligations"

When Mann had solved the problem, the interest in these subjects disappeared. But what about proving the inequality $\gamma\geq3\alpha$? Or, equivalently, what are the sequences A for which $\gamma<3\alpha$? These questions were not asked.

12. However, Schnirelman density is not a good characteristic. Take $A=\{2,3,4,\dots\}$. For this sequence we have $A(1)=0$ and $d(A)=0$. We feel, however, that the value 1 would be more appropriate for a density. So we arrive at a notion of an asymptotic density:

$$\underline{d}(A)=\liminf_{x\to\infty}A(x)/x\,.$$

In 1953 Martin Kneser [153] proved an analog of the $\alpha+\beta$ theorem for asymptotic densities. He described the structure of A and B in the case when

$$\underline{d}(A)+\underline{d}(B)<\underline{d}(A+B)\,.$$

Recently Yuri Bilu analysed the case when

$$\underline{d}(A+A)\leq\sigma\underline{d}(A)\,,$$

where $\sigma\in[2,5/2]$.

To prove his theorem Kneser had to consider, for some positive integer g, sets of residues A and B modulo g for which

$$|A+B|=|A|+|B|-1\,.$$

Cauchy [38] and Davenport [50] have proved that if $A \subseteq \mathbb{Z}_p$ and $B \subseteq \mathbb{Z}_p$, where p is a prime, then

$$|A+B| \geq \min\bigl(p,\ |A|+|B|-1\bigr)\,.$$

This inequality is analogous to (8).

Vosper [257] proved that if $A, B \subseteq \mathbb{Z}_p$, $|A|+|B|-1 \leq p-2$ and $\min(|A|,|B|) \geq 2$ then from $|A+B| = |A|+|B|-1$ it follows that A and B are arithmetic progressions in $\mathbb{Z}_p$ with the same difference.

Theorems of Kneser, Cauchy-Davenport and Vosper were amongst the first results giving solutions of inverse additive problems.

13. We may ask, are there any applications of the ideas and results described in §§4–8? For an answer to this question we turn now to the extremal combinatorial problems of Paul Erdős.

We begin with the problem raised by Erdős and Freud [68]. Fix some positive integer, ℓ. Denote by A a set of x natural numbers, $\{a_1, a_2, \ldots, a_x\}$, with $1 \leq a_1 < a_2 < \cdots < a_x \leq \ell$. Take the set, $A_0 = \{3, 6, 9, \ldots, 3\left[\frac{\ell}{3}\right]\}$. For each subset $B \subset A_0$ the sum of elements in B, the *subset sum*, is divisible by 3 and thus not equal to any power of 2. In this case $|A_0| = \left[\frac{\ell}{3}\right]$.

However if we take $|A| > \left[\frac{\ell}{3}\right]$ then for sufficiently large ℓ there exist $B \subset A$ and $s \in \mathbb{N}$ such that $\sum\limits_{a_i \in B} a_i = 2^s$. This was proved in [70]. E. Lipkin [167] proved that, for sufficiently large ℓ, a set of maximal cardinality, none of whose subset sums is equal to a power of two, must be exactly the set A_0.

The desired result was achieved with the help of analytical methods. However, there was a difficulty — how to apply them to prove a result which is valid for some integer, say, $\left[\frac{\ell}{3}\right]+1$, but is not valid for an integer which is one less. To cope with this, some conditions were formulated, so that when satisfied an analytical treatment could be used. The case where these conditions were not fulfilled was treated as an inverse additive problem. The structure of such sets was thus determined and it then became possible to finish the proof. (For more details, see §28.)

One might think that the problem of representing powers of two by subset sums is rather special, even artificial and therefore not that interesting. But, Paul Erdős knows how to ask questions. Ideas developed in order to solve the problem explained here, have turned out to be sufficient to solve a wide range of problems in Integer Programming, see §23 and [41]–[44].

14. In the framework of the problem of the previous section we may ask the following questions.

1) Let $|A| > \left[\frac{\ell}{3}\right]$. What is the minimal cardinality $|B|$ of $B \subset A$, whose subset sum is equal to some power of 2?
2) What is the minimal number of summands required in the representation of a power of 2, if equal summands are allowed?

These questions were asked and answered in a paper of M. Nathanson and A. Sárkőzy [201]. The sufficient number of summands required was estimated to be at most 30360 and 3503, respectively. Using the Theorem of §5 it appeared to be possible to improve these estimates to 8 and 6, respectively (see [104]). We will here briefly explain the main ideas. If we apply the Theorem of §5 to some set $A \subset [1,\ell]$, then under doubling the number of elements is multiplied, roughly, by 3 and the length of the segment where the sum $2A$ is situated is multiplied by 2. So, the density is multiplied, roughly, by $\frac{3}{2}$. After the doubling is repeated twice, the density of $4A$ will be $\geq \frac{1}{3}\cdot\frac{3}{2}\cdot\frac{3}{2}=\frac{3}{4}$. One more doubling (or more accurately summing $4A+2A$) will give a long interval, in $8A$ (or even in $6A$), containing then some power of 2.

Noga Alon gave a simple example showing that 4 summands in the case of different and 3 summands in a case of possibly repeating summands are not, in general, sufficient. Recently, Vsevolod Lev [160] found the exact number of summands, in a case of possibly repeating ones. He showed that four summands are sufficient.

The following questions are of interest.

1) For given $|A|$ and s, find, $f(|A|,\ell,s)$, the minimum over all sets $A \subset [1,\ell]$ of order $|A|$, of the maximal length arithmetic progression contained in sA.
2) For given $|A|$ and L, find, $f(|A|,\ell,L)$, the maximum over all sets $A \subset [1,\ell]$ of order $|A|$, of the minimum number of summands, s, such that sA contains an arithmetic progression of length L.

15. Denote by $s^\wedge A$ the set of integers which can be written as a sum of s pairwise distinct elements from A. The set A is called *admissible* if, and only if, $s \neq t$ implies that $s^\wedge A$ and $t^\wedge A$ have no element in common.

E.G. Straus [247] showed that the set $\{N-k+1, N-k+2,\dots,N\}$ is admissible if, and only if, $k \leq 2\sqrt{N+\frac{1}{4}}-1$. He proved that for any admissible set $A \subset [1,N]$ we have $|A| \leq \left(4/\sqrt{3}+o(1)\right)\sqrt{N}$. The constant involved was slightly reduced by P. Erdős, J-L. Nicolas and A. Sárkőzy (cf. [75]). In the paper of J-M. Deshouillers and G. Freiman [52] (see also [51]) Erdős' conjecture was proved, at least when N is sufficiently large.

Theorem 3. — *There exists an integer N_0 such that for any integer $N \geq N_0$ and any admissible subset $A \subset [1,N]$ we have,*

$$|A| \leq 2\sqrt{N+\frac{1}{4}}-1\,.$$

The proof was obtained with the help of methods of the type quoted in §5.

16. Let $A \subset [1,n]$. If $A \cap (A+A) = \varnothing$, the set A is called sum-free. P. Erdős and P.J. Cameron conjectured that for the number I_n of sum-free sets we have,

$$I_n = O(2^{n/2})\,. \tag{9}$$

The typical example of sum-free set $A \subset [1,n]$ is the set $\{1,3,5,\dots\}$ of odd numbers. We can show that $[\frac{n+1}{2}]$ is the maximal cardinality of a sum-free set.

In G. Freiman [101] and the paper of J-M. Deshouillers, G. Freiman, V. Sos and M. Temkin [54], the problem of structure of sum-free sets was raised and studied. It was solved in the case of large cardinality of A, namely, when $|A| > 0.4\ell - c$, where c is some positive constant. An example of such a structure is one in which all the elements of A are congruent to 2 or 3 modulo 5.

The structure of A having been found, the estimate (9) for this class of A, now follows immediately. An open question is to describe the structure of A for smaller cardinalities.

17. In the paper of G. Freiman, L. Low and J. Pitman [106], the following conjecture of Erdős and Heilbronn [73] is proved for sufficiently large primes. *For $A \subset \mathbb{Z}_p$, where p is a prime, $|A| = k < p/50$ and $k > 60$, we have*

$$|A + A| \geq 2k - 3\,.$$

Also, the structure of A was described in the case when $|A + A| < 2.06k - 3$. The conjecture of Erdős and Heilbronn was proved independently by J.A. Dias da Silva and Y.O. Hamidoune, see [246].

18. In the paper of A. Yudin [261], an example of large sets of integers, A, was constructed for which

$$|A + A| < |A - A|^c$$

where $c = 0.756$. The previous example [113] gave only $c = 0.89$. In [113] the estimate $c \geq 0.75$ was proved. The result of A. Yudin puts the important additive characteristic,

$$\liminf \frac{\log|A + A|}{\log|A - A|} = \alpha\,,$$

in a very narrow interval, $0.75 \leq \alpha \leq 0.756$, and allows one to begin to study the structure of sets with values of c which are close to α. Possibly the example of Yudin is not far from an extremal structure (look at §7).

19. In the paper of E. Lipkin [169], the Diderich conjecture [62] was studied. We now describe the conjecture. Let G be a finite Abelian group, $A \subset G$ with $0 \notin A$. Let A^* denote the set of subset sums of the set A. G.T. Diderich called the minimal number n such that, if $|A| \geq n$ then $A^* = G$, the *critical number*, $c(G)$ of the group G.

Let G be an Abelian group of odd order $|G| = ph$ where p is the least prime divisor of $|G|$ and h is a composite integer. Diderich conjectured, and E. Lipkin proved for $G = \mathbb{Z}_q$ when q is sufficiently large, that

$$c(G) = p + h - 2\,.$$

20. In §§21–27 we will give a few examples of problems in different fields which may be looked at and treated as Structure Theory problems. These examples will be chosen from Additive Number Theory (§21), Combinatorial Number Theory (§22), Integer Programming (§23), Probability Theory (§24), Coding Theory (§25), Group Theory (§26) and Mathematical Statistics (§27). Our aim is not so much to enumerate these problems as to show how ideas and methods of Structure Theory may influence their solution and to show their interdependence. Not many examples are chosen and they do not cover the whole stock of related problems.

21. Additive Number Theory. We now present a paper (see [109]) of G. Freiman, H. Halberstam and I.Z. Ruzsa. This paper confronts the problem of how to show that, starting from some set of integers A, the set rA contains an arithmetic progression of integers of length, L, and difference, d.

One obvious set of sufficient conditions is as follows. Firstly, that the set $(r-1)A$ contains an arithmetic progression of length ℓ and difference d. Further that in some arithmetic progression of integers of length $L+2\ell$ and difference d, we have that every part of it which forms an arithmetic progression of length ℓ contains a number from A.

These conditions are very simple and satisfactory but, how may one find such an arithmetic progression of length ℓ, even if ℓ is much smaller than L? It is supplied by results of the paper mentioned! The final result is given below.

Theorem 4. — *Let B be an infinite set of integers such that $\Delta_B(N) \equiv \frac{B(N)}{N} > (\log N)^{-\alpha}$ for every integer $N > N_0$, where α is some fixed number in the interval $\left(0, \frac{1}{3}\right)$, and $N_0 = N_0(\alpha)$. Suppose further that B has the following "local" property.*

Corresponding to each $N > 12N_0$ there exists an integer M with $N_0 \leq M < \frac{1}{12}N$, such that every arithmetic progression modulo q in $[1, N]$ of length $\left[\frac{1}{2}A(M)\right]$ contains an element of $B_N := B \cap [1, N]$, where $2 \leq q \leq M$ and

$$A(M) = e^{\frac{1}{2}C_0(\log M)^{1-3\alpha}}.$$

Then B is an asymptotic basis of order 4.

The first version of this paper was built on methods of [82] and [105], but later changed to methods of [233], proposed by I. Rusza in his proof of the main theorem. The results of [109] were improved by Bourgain [21].

22. Combinatorial Number Theory. See examples given in §§13–19.

23. Integer Programming. Let us discuss problems connected with one linear equation,

$$a_1x_1 + a_2x_2 + \cdots + a_mx_m = b. \tag{10}$$

Suppose that the coefficients in (10) are positive integers, with $a_1 < a_2 < \cdots < a_m < \ell$, and we wish to find a solution in the Boolean case with $x_i \in \{0, 1\}$. Remember that we are dealing here with problems which we would not be encountering in Number Theory. We have to find an algorithm with the help of which a computer has to be able, in a reasonable time, to answer the question, whether or not there exists a solution and then, to find it. And a most important point must be borne in mind, namely that the algorithm has to achieve this task for *any* choice of coefficients in a given range. The number of unknowns in (10) is equal to m, and each unknown may take two values, so the number of possibilities to check, if we decided to do it, is 2^m. Existing methods (branch and bound, partial enumeration, etc.) try to diminish this number but progress has been slow. If the coefficients $a_j \in [1, \ell]$ and $\ell = 10^{12}$, say, then m has to be not bigger than about 100 or 200 for the equation to be solved by today's computers. The dynamic programming approach gives times of $O(\ell m^2)$. If, for example, $m = 10^6$ the time is of order 10^{24} verifications, too long to see results in our lifetime.

A different approach to the problem was outlined in [96]. We began to study the structure of the set of values of a linear form, using Analytic Number Theory. This structure appeared to be rather simple, it is in essence, the union of several arithmetic progressions with the same difference. To characterize an arithmetic progression we have to know its difference d, its first member and its length.

The time required to answer a question of solubility of an equation is $O(m)$ and in our example it is of order 10^6 verifications, a matter of seconds. The main idea is explained in §28. For detailed exposition and literature see a review of Mark Chaimovich [42] and a paper [43].

24. Probability Theory. Estimates for concentration functions and local limit theorems — these are two domains where today there exist applications of the Structure Theory approach to Probability Theory.

Let $\xi_1, \ldots, \xi_n$ be a sequence of independent identically distributed random variables taking values in $\mathbb{Z}$. Further, let $s_n = \sum_{j=1}^{n} \xi_j$. Define

$$Q_\xi(\ell) = Q(\ell) = \sup_x P(x \leq \xi < x + \ell),$$

the concentration function of the random variable ξ, and let $Q_{s_n}(\ell) = Q_n(\ell)$ be the concentration function of s_n.

The paper of J-M. Deshouillers, G. Freiman, A. Yudin [55], gives a new estimate for $Q_n(1)$. Previous results, see for example G. Kesten [150], give an estimate of the type

$$Q_n(1) < \frac{c}{n^{\frac{1}{2}}}, \tag{11}$$

where c is independent of n. In this estimate the exponent $\frac{1}{2}$ cannot in general be replaced by a larger number. Indeed, let us fix some integer valued random variable with variance σ^2. Then by the local limit theorem we have

$$P\{s_n = N\} = \frac{1}{\sigma\sqrt{2\pi n}} \left(\exp\left(-\frac{(\mu n - N)^2}{2n\sigma^2} \right) + o(1) \right).$$

From here we see that the estimate (11) cannot, in general, be improved. If we want to improve (11) we have to impose additional conditions and this is what is done in [55].

Theorem 5. — *Let* $\sigma \in \left(\frac{\log 4}{\log 3}, 2\right)$, $\varepsilon > 0$, $A \geq 1$ *and* $a > 0$ *be given real numbers. Let* n *be a positive integer and let* $\{X_1, \ldots, X_n\}$ *be a set of independent identically distributed integral random variables such that*

$$\max_{q \geq 2} \max_{s (\mathrm{mod}\, q)} \sum_{\ell \equiv s (mod\, q)} P\{X_1 = \ell\} \leq 1 - \varepsilon,$$

$$\forall L \geq A : Q(X_1; L) \leq 1 - aL^{-\sigma}.$$

Then we have

$$Q(S_n; 1) \leq cn^{-1/\sigma},$$

where c *depends at most on* σ,ε,A,a *and* $Q(X_1; 1)$.

We have here two conditions, one excludes the case when the support is a part of some class $\mathrm{mod}\, q$, $q \geq 2$ and the second asks for the tail to be 'heavy'. Conditions of both types are necessary to get results of the form of the Theorem above. In the first version of a paper [55] the condition of type 1 was formulated for a series of random variables as follows. For any $q \in \mathbb{Z}, q \geq 2$

$$\max_r \sum_{k \equiv r (\mathrm{mod}\, q)} p_k < 1 - 10\sqrt{\frac{\ln n}{n}} \,.$$

Let us also stress that the result of Esséen, cited in [55], gives a condition from which the concentration may be estimated from below. All these results give us the possibility to begin to study the distribution of a given random variable ξ, if we know something about the value of $Q_n(1)$, for example if we know that

$$Q_n(1) \asymp \frac{1}{n^{\vartheta}},$$

where $\frac{\ln 4}{\ln 3} < \vartheta \leq 2$. We can ask the same question for series. In this case we have to describe distributions where numbers a_i and numbers p_i may depend on n.

25. Coding Theory. This section and §35 were written jointly with A. Yudin. The connection between coding theory and structure theory was shown by Zemor (see [262] and [263]) and Cohen & Zemor (see [265], [266], [46] and [47]). We will now try to explain that the main problems of coding theory are, in fact, inverse additive problems.

Let $A = \{a_1, \cdots, a_k\}$ be a word in an alphabet of 2 symbols, say, $a_i \in \{0,1\}$. Let A_n be the set of all words in this alphabet of length n, so that we have $|A_n| = 2^n$. The distance, $g(x,y)$, between two words $x = \{x_1, x_2, \ldots, x_n\}$ and $y = \{y_1, y_2, \ldots, y_n\}$ is defined to be

$$g(x,y) = \left|\{i \mid x_i \neq y_i, \quad i = 1, \ldots, n\}\right|,$$

that is, the number of positions in which the symbols in the words x and y differ. It is not difficult to check that $g(x,y)$ satisfies all the axioms for a distance function. The question is how to ensure the correction of possible errors during transmission of information?

Consider some subset, U, of the set of all words A_n. Such a subset is called a *code*. A portion of information has assigned to it some word from U which is then transmitted through the channel. If during the transmission only a small number of mistakes occurred then we are still not far from the code word which was transmitted and thus we can then restore it. Let us put this question in a more precise formulation. We let the word transmitted be $x = \{x_1, \ldots, x_n\}$ and the word received be $\tilde{x} = \{\tilde{x}_1, \ldots, \tilde{x}_n\}$. If during the transmission of a word through a channel no more than t mistakes take place, it means that

$$g(x, \tilde{x}) \leq t \tag{12}$$

and so it is necessary that $\tilde{x}$ be closer to x than to any other word in the code. That is, for any $y \in U$ with $y \neq x$, we have to ensure that

$$g(y, \tilde{x}) > t\,. \tag{13}$$

By the triangle inequality

$$g(x,y) \leq g(x,\tilde{x}) + g(\tilde{x}, y)\,, \tag{14}$$

and when

$$g(x,y) > 2t\,, \tag{15}$$

we can obtain (13) from the inequality (12).

If there exists y such that $g(x,y) = 2t$, then we can find $\tilde{x}$ for which (12) and (13) become equalities and then $g(\tilde{x},y) = t$. Thus, the condition (15) is necessary and sufficient for code correcting t mistakes. We have a set, A_n, and a subset U, but to speak about inverse additive problem is still premature, since an algebraic operation is missing. So we will consider A_n as a vector space over the field $\mathbb{Z}_2$. In this field $-1 = 1$ and for each n-dimensional vector $x \in A_n$ the equality $-x = x$ holds. The distance $g(x,y)$ is equal to the number of 1s in the vector $x - y = x + y$, i.e. to the distance of the element $x + y$ from 0. The condition (15) may now be written as

$$g(x+y,0) > 2t.$$

Thus, a code, correcting t mistakes, is a $U \subset A_n$ such that $\forall z \in 2U$ we have $g(z,0) > 2t$. We have now come to a well known situation, namely, we have a group A_n, a subset U and a condition on $2U$.

In §12 the first results about sums of sets in a group were mentioned. The doubling of sets in groups was studied in the works of Kemperman [146], [147], [148], Freiman [83], Olson [207], [208], [209], Brailovsky & Freiman [27], [29], Brailovsky [22–25] and Hamidoune [124–137]. If n is a minimal number such that for $A \subset G$ we have $nA = G$, A is called a *basis* of G of order n. This theme is reviewed in [9] and [140].

What are the main aims which we are trying to achieve in coding? Atoms of information are transmitted by words of code. Thus, if the quality of a code is fixed, i.e. the number of mistakes to be corrected is fixed, then the code will be the better, the greater the cardinality of the code U. And conversely, if the number $|U|$ is given, how do we choose the best code?

We shall reiterate the formulation of the two problems mentioned above. Let $U \subset A_n = \mathbb{Z}_2^n$ for some fixed $n \in \mathbb{N}$. Assume that for all $z \in 2U$

$$g(z,0) \geq d\,, \tag{16}$$

where $d \in \mathbb{N}$.

Problem I. Let d be fixed. What is the maximum value of $|U|$ for which (16) is valid?

Problem II. Let $|U|$ be fixed. What is the maximum value of d for which (16) is valid for some U of order $|U|$.

We have formulated two inverse additive problems which are the major problems of coding theory but are, in essence, not yet solved satisfactorily. In a paper of Gerard Cohen and Gilles Zemor [47] other inverse additive problems are presented and their connection with coding theory is explained.

26. Group Theory. Results in group theory are reflected in the reviews of M. Herzog [140] and Y. Berkovich [9] and the bibliography to this review. We try now to find an example where our approach gives some progress on a theme which was investigated earlier in group theory.

For a set

$$\{a_1, a_2, a_3\} \tag{17}$$

of elements of a group G, we build all the products,

$$a_1a_2a_3,\ a_1a_3a_2,\ a_2a_1a_3,\ a_2a_3a_1,\ a_3a_1a_2,\ a_3a_2a_1\,. \tag{18}$$

If at least one product in (18) is equal to another one, the set (17) is called *rewritable.* If every 3-element set in G is rewritable, then G is called a *rewritable group*, that is $G \in Q_3$, where by Q_3 we denote the class of rewritable groups. If every product in (18) is equal to some other product, then the set (17) is called *totally rewritable.* If every set (17) in G is totally rewritable, then G is said to be a *totally rewritable group*, written $(G \in P_3)$. The definitions of classes of groups P_n and Q_n are obvious. The problem is to describe all groups in the classes P_n and Q_n. See Kaplansky [145], Blyth & Robinson [19], Freiman & Schein [117] and [118], Longobardy & Maj [170], [171] and [172].

The main tool to use in this study is a notion of 'permutational isomorphism', a realization of the equivalence relation we talked about in §1. This notion is somewhat different from the one introduced in §6, but it is suited very well to the study of this particular problem.

A *permutational isomorphism* of A onto B (where $A \subset S$ and $B \subset R$, while S and T are two sets with binary operations) is a pair of bijections $\varphi\colon A \to B$ and $\psi\colon A^{[3]} \to B^{[3]}$ such that for all pairwise distinct elements $a_1, a_2, a_3 \in A$ we have

$$\psi(a_1a_2a_3) = \varphi(a_1)\varphi(a_2)\varphi(a_3)\,.$$

Here $A^{[3]}$ is the set of all products of triples of distinct elements.

To begin our approach we have only to pay attention to the fact that amongst the six products in (18) there are no more than five distinct ones, if the set (17) is rewritable. Thus, we take as a numerical characteristic, r, the maximal number of different products for all sets (17) in a group G. We thus obtain the classes of groups $P(3,r)$ for $1 \le r \le 6$ (see Freiman & Schein [117]). In [117] all classes of isomorphic triples, 19 classes in all, were obtained and then used to study the classes $P(3,r)$. Similarly one can define the classes of groups $P(4,r)$ of which there are 24. It turns out that $P_3 = P(3,2)$ (see [117]). In [118] the class $P(3,3)$ was described. G. Freiman, D. Robinson and B. Schein [115] partially described the class $P(3,4)$. The next step is the study of $P(3,5) = Q_3$.

27. Mathematical statistics. Let $F = \{f_i\}_{i=0}^n$ be a set of continuous functions on $[a,b]$, and let $F^* = \{f_if_j\}_{i,j=0}^n$. In the paper of B. Granovsky and Eli Passow [120] conditions were determined for the set F^* to consist of exactly $2n+1$ distinct functions. The additional requirement is that F^* has to be a Chebyshev system on $[a,b]$.

A set $\{u_i\}_{i=0}^n$ of continuous functions on $[a,b]$ is said to be a *Chebyshev system* on $[a,b]$ if every nontrivial "polynomial" $\sum_{i=0}^n g_iu_i(x)$ has at most n zeros on $[a,b]$. The number $n+1$ is called the *degree* of the Chebyshev system. In [120] necessary and sufficient conditions were given on the set $\{f_i\}_{i=0}^n$ so that the set $\{f_if_j\}_{i,j=0}^n$ is a Chebyshev system of minimal degree $(2n+1)$. These results have applications to the field of experimental design. See also I. Efrat [66], Kiefer & Wolfowitz [152] and E. Passow [213].

It is clear how this problem can be formulated as a problem of small doubling of a set of real numbers. Given $n+1$ functions pick some fixed argument x_0. Consider the $n+1$ numbers $\{f_i(x_0)\}_{i=0}^n$. Leaving for further investigation the case when they are not all distinct, or some of them are not positive, we have the set D, of logarithms of these numbers, $D = \{\log f_i(x_0)\}_{i=0}^n$ subject to the condition $|2D| = 2n+1$. So D is a set with small doubling and it is very simple to show, not only for integers but also for real numbers, that D is an arithmetic progression. I. Efrat [66] has used the results of Theorem 1 and described all Chebyshev systems with $|F^*| < 3n$.

28. In this section we want to point out the unity of approach and similarity of methods when different problems are treated from the point of view of Structure Theory.

In Combinatorial additive problems we mainly study finite sets of integers. In many of such problems the theorems of §§5 and 6 about a structure of sets of integers with small doubling may be applied directly. In §§13–19 such results were given. These theorems may also be applied to sets in other algebraic systems, such as $\mathbb{Z}_p$, see [88], $\mathbb{T}^1$, see [197], $\mathbb{R}^k$, see [82], page 94, and to functional spaces, see [66]. The sets in $\mathbb{Z}$ may be infinite, see [91] and [82]. The structure of sets with a small product in a nonabelian torsion-free group, see [26], is described with the help of methods developed to prove Mann's theorem.

To solve inverse problems of additive number theory, analytical methods are used. They reveal some unity and similarity when applied to the study of different problems, see §30. Problems in number theory of the evaluation of measure and of the determination of the structure of sets with large trigonometric sum, see [260], [100], [13], and in probability theory, of sets with large characteristic function, see [197] and [55], are often studied by similar methods.

A tool of investigation which can be used in many situations, may be called "multiple use of structural argument". To ensure the existence in Integer Programming, of a solution of an equation (10), see [96], we assume a condition on $A(q) \equiv \{x \in A \mid q|x\}$, namely

$$|A(q)| < |A| - |A|^\delta\,, \tag{19}$$

where $\delta < 1$ is independent q. In analytical number theory it is usual to place such a uniformity condition on the distribution of residues. When it does not hold, the case is not studied. However let us now consider the case when (19) is not valid. Then there exists $q > 1$ such that

$$|A(q)| \geq |A| - |A|^\delta\,.$$

This is a very strong condition to impose on the structure of A and so we can continue our analysis and describe the structure in full. In papers [56] and [197], where problems in probability were studied, a condition of the type (19) is present. This observation opens up the possibility to obtain new results, stronger than those in [56] and [197].

The very notion of a set with small doubling, when brought to group theory, resulted in the appearence of new problems.

The notion of isomorphism which was introduced in the course of proving the main theorem (§6) became a useful tool. In group theory, it provided the possibility of building an equivalence relation on finite sets, describing its equivalence classes and then studying the property of a group in connection with the existence or nonexistence of some classes in this group. In rewritable groups, see §26, a version of isomorphism was given suited to the purpose. In [53] a notion of isomorphism for random variables was introduced, which gave the possibility of describing the behavior of a one-dimensional random variable with the help of a multidimensional one.

29. First results about the structure of sets with small doubling were obtained with the help of elementary methods. Afterwards, the analytic methods were introduced. In fact, there exists an exact dividing point. If $|K+K| \leq 3k-3$, then the elementary approach very quickly gave a full description of K. For larger values the elementary methods did not give results in spite of big efforts.

Very little has been done to get elementary results in the multidimensional case. In [82] the case on the plane of $|K+K| < \frac{10}{3}k-5$ is studied and I. Stanchescu studied the case $|K+K| < (4-\varepsilon)k$. I don't know the range of the doubling coefficient C_n in an inequality $|K+K| < C_n k$, where $K \subset \mathbb{Z}^n$ for which elementary results may be obtained.

To obtain here a clear picture is very desirable and not very difficult. Then it can be used to make the results of the main theorem more precise. Results for doubling coefficients $\frac{10}{3}$ and $4-\varepsilon$ show that the structure of A after it becomes multidimensional may be described more accurately with the help of elementary methods.

Many interesting problems arise from a study of K when two, or more, numerical characteristics are given. A long list of invariants is given in [82], page 41.

30. In direct problems of additive number theory one is usually studying an integral which yields the number of representations of a number expressed as a sum of terms of a certain type. Further, a transform of this integral yields an asymptotic formula for the number of representations. Characteristic of the analytic method in Structure Theory is the fact that an integral with a known value serves as a starting point.

Examples

(i) (See Roth [224].) Sets A without arithmetic progressions of length three. We have

$$\sum_{x\in A}\sum_{y\in A}\sum_{z\in A}\int_0^1 e^{2\pi i\alpha(x+y-2z)}\,d\alpha = |A| = \int_0^1 S^2 S_1\,d\alpha\,,$$

where

$$S=\sum_{x\in A} e^{2\pi i\alpha x},\quad S_1=\sum_{z\in A} e^{-4\pi i\alpha z}\,.$$

(ii) A set K with small doubling (see Freiman [82]). Here

$$\sum_{x\in K}\sum_{y\in K}\sum_{z\in 2K}\int_0^1 e^{2\pi i\alpha(x+y-z)}\,d\alpha = \int_0^1 S^2 S_1\,d\alpha = |K|^2,$$

where

$$S=\sum_{x\in K} e^{2\pi i\alpha x}, \quad S_1=\sum_{x\in 2K} e^{-2\pi i\alpha x}.$$

(iii) Sum-free sets. We have

$$\sum_{x\in A}\sum_{y\in A}\sum_{z\in A}\int_0^1 e^{2\pi i\alpha(x+y-z)}\,d\alpha=\int_0^1 S^2\overline{S}\,d\alpha=0\,,$$

where

$$S=\sum_{x\in A} e^{2\pi i\alpha x}, \quad A\subset[1,l], \quad l\in\mathbb{N}.$$

The next step is to obtain a large trigonometric sum for a certain value (sometimes, for several values) of the argument. Consider an example from Freiman [82], page 48. Let K be a set of residues modulo a prime p. Then

$$I=\sum_{x_1,x_2\in K}\ \sum_{x_3\in 2K}\ \sum_{a=0}^{p-1} e^{2\pi i\frac{a}{p}(x_1+x_2-x_3)}=\sum_{a=0}^{p-1} S^2S_1=k^2p\,,$$

where

$$S=\sum_{x\in K} e^{2\pi i\frac{a}{p}x}, \qquad S_1=\sum_{x\in 2K} e^{-2\pi i\frac{a}{p}x}\,.$$

Let $T=|K+K|$ and assume that $|S|<\frac{3}{5}k$ for every $a\not\equiv 0(p)$ then

$$\begin{aligned}|I|\le k^2T+\sum_{a=1}^{p-1}|S|^2|S_1| &\le k^2T+\frac{3}{5}k\Big(\sum_{a=0}^{p-1}|S|^2\sum_{a=0}^{p-1}|S_1|^2\Big)^{1/2}\\ &= k^2T+\frac{3k}{5}\sqrt{kp\cdot Tp}.\end{aligned}$$

In the example just considered the conditions $T<\frac{12}{5}k$ and $k<\frac{p}{35}$ were assumed, from which it follows that $|I|<k^2p$, a contradiction. We have therefore proved that there exists $a'\not\equiv 0(\mathrm{mod} p)$ such that

$$|S(a')|=\left|\sum_{j=0}^{k-1} e^{2\pi i\frac{a'}{p}a_j}\right|>\frac{3}{5}k\,.$$

The presence of a large trigonometric sum makes it possible to obtain data about the set A which can then be processed using elementary techniques.

31. In the first papers on sets with small doubling information about only one large trigonometric sum was used. In the proof of the main theorem we have used several, but finite number of large sums. The next step was to begin to study a set of all 'large' trigonometric sums. It was first done in 1973 in probability theory field, in the proof of local limit theorems (see D. Moskvin, G. Freiman & A. Yudin [197]). In this case we were dealing with the characteristic function of a lattice random variable,

$$f(\alpha)=\sum_{k\in\mathbb{Z}} p_k e^{2\pi i\alpha k}$$

studying the measure and structure of the sets E, where the characteristic function is large.

The reasoning is, in short, as follows. We use the fact that, if for some α_1 and α_2 we have $|f(\alpha_1)| \geq 1-u$ and $|f(\alpha_2)| \geq 1-u$ then $|f(\alpha_1+\alpha_2)| \geq 1-4u$. We take the set

$$E = \left\{ \alpha \;\middle|\; |f(\alpha)| > 1 - \frac{\sqrt{\log n}}{n}, \quad n \in \mathbb{N} \right\}$$

and begin to double, obtaining sets $2E$, 2^2E, 2^3E, ... If the measure is growing steadily we will cover the set $[0,1)$ very quickly, thus obtaining a contradiction. If at some stage we meet a set with small doubling, we will get a structure. For some $q \in \mathbb{N}$, the arguments $\frac{p}{q}$, with $0 \leq p < q$, will be included in this structure which will lead to the conclusion that almost all the probability measure is concentrated in an arithmetical progression modulo q, which gives a contradiction.

32. We are naturally led to a study of sets with a large measure of large trigonometric sums.

Let k be a positive integer and $u < k$ a positive real. For a set

$$K = \{a_1 < a_2 < \cdots < a_k\}, \quad a_j \in \mathbb{Z}, \quad 1 \leq j \leq k$$

let

$$S_K(\alpha) = \sum_{j=1}^{k} e^{2\pi i \alpha a_j}, \quad s_K(\alpha) = |S_K(\alpha)|,$$

$$E_{K,u} = \{\alpha \in [0,1), \quad \text{for which } s_K(\alpha) \geq k-u\}$$

and

$$\mu_K(u) = \mu(E_{K,u})$$

when μ is the Lebesgue measure on $[0,1]$.

Problem. — Find the set K which maximizes $\mu_K(u)$ and find its maximal value.

We denote by $\mu_{\max}(k,u)$ the supremum of $\mu_K(u)$ over all sets K of size k. The first results on this problem were obtained by Freiman (see [95], page 144) and Yudin (see [260], page 163). I sketched an approach for solving the problem in [100]. In [13] A. Besser carried out and extended this plan very widely. He showed that up to the second order

$$\mu_{\max}(k,u) = 2\beta \simeq \frac{2\sqrt{6}}{\pi}\frac{1}{k}\left(\frac{u}{k}\right)^{\frac{1}{2}}\left(1+\frac{5}{8}\frac{u}{k}\right)$$

and K_{ex} may be described, in the main case, as the union of an arithmetic progression of length $k_0 = k - \frac{5}{12}u$, symmetric around zero, and, for any non-zero integer n, an arithmetic progression of length

$$\frac{1}{2}k_n = \frac{u}{(\pi n)^2}\left(1 - \frac{(-1)^n}{2}\right)$$

centered around $\frac{n}{\beta}$.

We will try to explain from where the structure of K_{ex} comes. If α is small the term $e^{2\pi i\alpha a_j}$ has a value close to 1 if a_j is small. That is why we take an arithmetic progression with difference 1 centered around 0. We have, for $\alpha > 0$,

$$s_k(\alpha) = \frac{\sin(\pi\alpha k)}{\sin(\pi\alpha)} \simeq \frac{\pi\alpha k - \pi^3\alpha^3k^3/6}{\pi\alpha} = k - \frac{\pi^2\alpha^2}{6}k^3 .$$

As α increases, $s_k(\alpha)$ decreases and reaches $k-u$ for α determined by

$$k - \frac{\pi^2\alpha^2}{6}k^3 = k - u$$

that is,

$$\alpha^2 \simeq \frac{6}{\pi^2}\frac{u}{k^3}$$

and thus

$$\alpha_0 \simeq \frac{\sqrt{6}}{\pi}\frac{1}{k}\Big(\frac{u}{k}\Big)^{\frac{1}{2}} .$$

Consider the trigonometric sum at this point α_0. Our set is positioned on the segment $\left[-\frac{k}{2}, \frac{k}{2}\right]$. If we add another number, $\frac{k}{2}+1$, to the arithmetic progression, the term $e^{2\pi i\alpha_0(\frac{k}{2}+1)}$ will be added to the trigonometric sum. If we add $[\frac{1}{\alpha_0}]$, then $e^{2\pi i\alpha_0[1/\alpha_0]}$ will be closer to unity, it will lie in a smaller neighborhood of the x axis and will influence the increasing of $S(\alpha)$ more critically. This consideration explains the appearance of segments near to the points $\frac{n}{\alpha_0}$.

33. An analysis of the remarkable results of A. Besser does not reveal an easy future. The set K_{ex} is of a rather complex two-dimensional structure which becomes more complex as n increases and will, in all likelihood, become multi-dimensional. The structure of K_{ex} has only been found for very small values of u, $u < \frac{k}{32000}$ and an increase is only gained with some effort. Thus, further progress in the problem under consideration would be of great interest, but reaching it is very difficult.

The sets K_{ex} found by Besser have small density for small u's. But in many open problems the situation is different. For example, in the problem of sum-free sets, the density of the set to be considered is close to 0.4. When attempting to strengthen the theorem on the structure of K, with small doubling, outside the bounds $|2K| = 3k-3$, we should begin by considering sets whose densities are close to 0.5. So, we state the problem on measure of large trigonometric sums as follows. Let K be a set of integers in $[0,l]$ with $|K| = k$. Define

$$E_{K,m} = \{\alpha \in [0,1), \text{ for which } s_K(\alpha) \geq m\}$$

and let $\mu_K(m) = \mu(K,m)$ denote the measure of $E_{K,m}$. Also we set

$$\mu_k(m,l) = \max_{K\subset[0,l]} \mu(K,m) .$$

Then if $l = k-1$, the problems is a trivial one. As l increases, it becomes more complex. After the quantities $\mu_k(m,l)$ have been found, one should proceed to describe the structure of those K's for which $\mu(K,m)$ does not differ greatly from $\mu_k(m,l)$.

34. In the problem on sum-free sets, the following integral was being considered,

$$\int_0^1 |S|^2 S\,d\alpha = 0\,,$$

and it follows from this that

$$\int_0^1 |S|^2 \Re(S)\,d\alpha = 0\,.$$

In a neighbourhood of zero the integrand is of order k^3 and its contribution to the integral is of order k^2. Since the integral over the whole interval equals zero, the measure of the set of α's where $|\Re(S)|$ has order k and is negative, should be large.

We come to the following general problem. Let $K \subset \mathbb{Z}$ with $|K| = k$ and set

$$E_{K,-m} = \{\alpha \in [0,1), \text{ for which } \Re(S) \leq -m, \quad 0 < m < k\}\,.$$

Let $\mu(K,-m)$ be the measure of $E_{K,-m}$ and

$$\mu_k(-m,l) = \max_{K\subset[0,l]} \mu(K,-m)\,.$$

The usual questions may be asked once again about the quantities $\mu_k(-m,l)$ and about the structure of the set K for which the measure $\mu(K,-m)$ is close to the maximal value. At the next, deeper stage of study, one may investigate combining two or more numerical characteristics. The first step here should be the study of trigonometric sums when some conditions are imposed not only on $|S|$ but also on $\arg S$.

35. Let G be an abelian group whose operation will be denoted by $+$, and $\widehat{G}$ be the group dual to G, that is the group of characters of G. Let A be a subset of G and define a map

$$f_\chi \colon A \longmapsto \sum_{a\in A} \chi(a), \quad \text{for } \chi \in \widehat{G}\,,$$

that is, to the set A we correspond a function of a character $\chi \in \widehat{G}$.

As is shown in [82], from the fact that $|2A| \leq C|A|$ in the case $G = \mathbb{Z}$ it follows that the set on which $|f(\chi)|$ is rather large has a large measure. With the help of methods from harmonic analysis we can describe the structure of the set A.

It is important to stress that to the set A with small doubling from G corresponds a set

$$\widehat{A}_\alpha = \left\{\chi \in \widehat{G} \text{ such that } \Big|\sum_{a\in A} \chi(a)\Big| > \alpha|A|\right\}, \quad \text{for } \alpha \in \mathbb{R}^+\,,$$

which also has small doubling. From the fact that $\widehat{\widehat{G}} = G$ we may, it seems, suppose that from $B \subset \widehat{G}$ and $|2B| \leq C|B|$ it will follow that $\widehat{B} \subset G$ and $|2\widehat{B}| \leq C|\widehat{B}|$. Note that the constants in different places of this section may differ. For a given additive problem it is possible to find the equivalent problem on the dual group and *vice versa*, and then to study the version which is preferable.

The following observations are also important. Suppose that $A_1 \subset G$ and $A_2 \subset G$ are sets which are structurally 'near' to each other. A natural question to ask is whether $\widehat{A}_1$ and $\widehat{A}_2$ are also 'near' to each other and what kind of topology is induced by the correspondence $A \longmapsto \widehat{A}$. Again, from $\widehat{\widehat{G}} = G$ it follows that these topologies induce one other. It would be very interesting to determine what kind of neighourhoods they define and to what extent these topologies are 'metrisable', because metric characteristics of these topologies will be of great interest during the study of problems of addition of sets.

The analytic tool in the case $G = \mathbb{Z}$ was the equality

$$\int_0^1 S^2 S_1 \, d\alpha = |A|^2 ,$$

where

$$S = \sum_{x \in A} e^{2\pi i \alpha x} \quad \text{and } S_1 = \sum_{x \in 2A} e^{-2\pi i \alpha x} .$$

In the case of a finite abelian group, A, we can write the parallel expression

$$\sum_{\chi} \left(\sum_{a \in A} \chi(a) \right)^2 \sum_{a \in 2A} \overline{\chi}(a) = |A|^2 .$$

Generalization to the nonabelian case should also be studied.

36. I am greatly indebted to Dr. Ruth Lawrence and Mr. Harry Lawrence for their invaluable help in producing this manuscript.

References

[1] Alon N., *Independent sets in regular graphs and sum-free subsets of finite groups*, Israel J. Math. **73** (1991), 247–256.

[2] Alon N., *Subset sums*, J. Number Theory **27** (1987), 196–205.

[3] Alon N., Freiman G.A., *On sums of subsets of a set of integers*, Combinatorica **8(4)** (1988), 297–306.

[4] Alon N., Kleitman D.J., *Sum free subsets* in "A tribute to P. Erdos", edited by A. Baker, B. Bollobas, A. Hajnal, Cambridge University Press, Cambridge, England, (1990), 13–26.

[5] Babai L., Sos V., *Sidon sets in groups and induced subgraphs of Cayley graphs*, Europ. J. Comb. **6** (1985), 101–114.

[6] Balas E., Zemel E., *An algorithm for large zero-one knapsak problems*, Operations Research **28** (1980), 1130–1154.

[7] Bell H.E., Klein A.A., *On rings with redundancy in multiplication*, Arch. Math. **51** (1988), 500–504.

[8] Berkovich Y., *Non-solvable groups with large fraction of involutions*, this volume.

[9] Berkovich Y., *Questions on set squaring in groups*, this volume.

[10] Berkovich Ya.G., Freiman G.A., *On the connection between some numeric characteristics of a finite group and the structure of the group*, (1981), manuscript.

[11] Berkovich Ya.G., Freiman G.A., Praeger C., *Small squaring and cubing properties for finite groups*, Bull. Australian Math. Soc. **44(3)** (1991) 429–450.

[12] Berstein A.A., Freiman G.A., *Analytical methods of discrete optimization*, CEMI (1979), 89–105.

[13] Besser A., *Sets of integers with large trigonometric sums*, this volume,

[14] Beurling A., *Analyse de la loi asymptotique de la distribution des nombres premiers generalises I*, Acta Math. **68** (1937), 255–291.

[15] Bianchi. M., Brandl. R., Mauri A.G., *On the 4-permutational property for groups*, Arch. Math. **48** (1987), 281–285.

[16] Bilu Y., *Structure of sets with small sumset*, this volume,

[17] Blyth R.D., *Rewriting products of group elements I*, J. Alg. **116** (1988), 506–521.

[18] Blyth R.D., *Rewriting products of group elements II*, J. Alg. **119** (1988), 246–259.

[19] Blyth R.D., Robinson D.J.S., *Recent progress on rewritability in groups*, in "Group Theory" (Proc. of the 1987 Singapore Conf.), de Gruyter, Berlin–New York (1989), 77–85.

[20] Bogdanovic S., Ciric M., *Tight semigroups*, Public. de l'Institute Math., **50(64)** (1991), 71–84.

[21] Bourgain J., *On arithmetic progression in sums of sets of integers* in "A tribute to P.Erdos", eds. A. Baker, B. Bollobas, A. Hajnal, Cambridge Univ. Press, Cambridge, England (1990), 105–109.

[22] Brailovsky L.V., *Set multiplication in groups*, Thesis for the degree of Ph.D., Tel Aviv University, 1992.

[23] Brailovsky L.V., *On $(3 - m)$ special elements in groups*, Comm. Algebra **20** (1992), 3301–3320.

[24] Brailovsky L.V., *Structure of quasi-invariant sets*, Arch. Math. (Basel) **59** (1992), 322–326.

[25] Brailovsky L.V., *A characterization of abelian groups*, Proc. Amer. Math. Soc. **117** (1993), 627–629.

[26] Brailovsky L.V., Freiman G.A., *Groups with small cardinality of the cubes of their two-element subsets*, Ann. New York Acad. Sci. **410** (1983), 75–82.

[27] Brailovsky L.V., Freiman G.A., *On the product of finite subsets in a torsion-free group*, J. of Algebra **130** (1990), 462–476.

[28] Brailovsky L.V., Freiman G.A., *On two-element subsets in group*, Ann. New York Acad. Sci. **373** (1981), 183–190.

[29] Brailovsky L.V., Freiman G.A., Herzog M., *Special elements in groups*, in "Group Theory" (Proc. 2nd Internat. Conf., Bressanone, Italy 1989), Suppl. Rend. Circ. Mat. Palermo, II series **23** (1990), 33–42.

[30] Bredihin B.M., *Free numerical semigroups with power densities*, Dokl. Akad. Nauk SSSR (N.S.) **118** (1958), 855–857 [Russian].

[31] Bredihin B.M., *Free numerical semigroups with power densities*, Mat. Sb. (N.S.) **46(88)** (1958), 143–158 [Russian].

[32] Bredihin B.M., *Elementary solutions of inverse problems on bases of free semigroups* Mat. Sb. (N.S.)**50(92)** (1960), 221–232 [Russian].

[33] Bredihin B.M., *The remainder term in the asymptotic formula for $V_G(x)$*, Izv. Vysš. Učebn. Zaved. Matematika **6(19)** (1960), 40–49 [Russian].

[34] Brodsky S., *On groups generated by a pair of elements with small third or fourth power*, this volume,

[35] Buzytsky P.L., Freiman G.A., *Analytical methods in integer programming*, Moscow, CEMI **48** (1980) [Russian]

[36] Cameron P.J., *Portrait of a typical sum free set*, London Math. Soc. Lecture Notes Series **123**(1987), 13–42

[37] Cameron P.J., Erdos P., *On the number of sets of integers with various properties*, in "Number Theory", Banff, Alberta 1988 conference proceedings, de Gruyter Berlin (1990), 61–79.

[38] Cauchy A.L., *Recherches sur les nombres*, J. Ecole Polytechn. **9** (1813), 99–116.

[39] Chaimovich M., *Fast exact and approximate algorithm for k-partition and scheduling independent tasks*, Discrete Mathematics **114** 1993, 87–103.

[40] Chaimovich M., *Solving value–independent knapsack problem with the use of methods of additive number theory*, Congressus Numerantium **72** (1990), 115–123.

[41] Chaimovich M., *Subset sum problem with different summands: Computations*, Discrete Applied Mathematics **27** (1990), 277–282.

[42] Chaimovich, M., *New structural approach to integer programming: a survey*, this volume,

[43] Chaimovich M., *New algorithm for Dense Subset-Sum Problem*, this volume,

[44] Chaimovich M., Freiman G.A, Galil Z., *Solving dense subset-sum problems by using analytic number theory*, J. of Complexity, **5** (1989), 271–282.

[45] Chvatal V., *Hard knapsak problems*, Operations Research **28** (1980), 1402–1411.

[46] Cohen G.D., Zemor G., *Intersetting codes and independent families*, Telecom Paris 92C003, Oct. 1992.

[47] Cohen G.D., Zemor G., *Subset sums and coding theory*, this volume

[48] Corput J.G., Kemperman J.H.B., *The second pearl of the theory of numbers I*, Nederl. Akad. Wetensch., Proc. **52** (1949), 696–704; or Indagationes Math. **11** (1949), 226–234.

[49] Curzio M., Longobardi P., Maj M., *Su di un problema combinatorio in teoria dei gruppi*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. **74(8)** (1983), 136–142.

[50] Davenport H., *On addition of residue sets*, J. London Math. Soc. **10** (1935), 30–32.

[51] Deshouillers J-M., Freiman G.A., *On an additive problem of Erdős and Straus I*, Israel J. Math., **92**, (1995), no. 1–3, 33–43.

[52] Deshouillers J-M., Freiman G.A., *On an additive problem of Erdős and Straus II*, this volume,

[53] Deshouillers J-M., Freiman G.A., Moran W., *On series of discrete random variables 1: Real trinomial distribution with fixed probabilities*, this volume,

[54] Deshouillers J-M., Freiman G.A., Sos V., Temkin M., *On the structure of sum-free sets 2*, this volume,

[55] Deshouillers J-M., Freiman G.A., Yudin A., *On Bounds for the Concentration Function, 1* this volume,

[56] Deshouillers J-M., Freiman G.A., Yudin A., *On a local limit theorem*, manuscript 1992.

[57] Diamond H.G., *The prime number theorem for Beurling's Generalized Numbers*, J. of Number Theory **1(2)** (1969), 200–207.

[58] Diamond H.G., *Asymptotic distribution of Beurling's Generalized Numbers*, Illinois Journal of Mathematics **14(1)** (1970), 12–28.

[59] Diamond H.G., *A set of generalized numbers showing Beurling's theorem to be sharp*, Illinois Journal of Mathematics **14(1)** (1970), 29–34.

[60] Diamond H.G., *Chebyshev estimates for Beurling generalized prime numbers*, Proc. of the American Math. Soc. **39(3)** (1973), 503–508.

[61] Diamond H.G., *When do Beurling's generalized numbers have a density?* J. fur die reine und angewandte Math. **259** (1977), 22–39.

[62] Diderrich G.T., *An addition theorem for abelian groups of order pq*, Journal of Number Theory **7** (1975) 33–48.

[63] Diderrich G.T., Mann H.B., *Combinatorial problems in finite Abelian groups*, in "A Survey of combinatorial Theory", eds. J. N. Srivastava et al., North Holland Publishing Company (1973), 95–100.

[64] Doeblin W., *Sur les sommes d'un grand nombre des variables aleaqtoires independentes* Bull. Sc. Math. **63** (1939), 23–32 and 35–64.

[65] Dyson F., *A theorem on the densities of sets of integers*, J. London Math. Soc. **20** (1945), 8–14.

[66] Efrat I., *Small Chebyshev systems made by products* J. of Approximation theory, **57(3)** (1989), 259–267.

[67] Erdős P., *On an elementary proof of some asymptotic formulas in the theory of partitions*, Ann. of Math. **48(3)** (1942), 437–450.

[68] Erdős P., *Some problems and results on combinatorical number theory*, in "Graph theory and its Applications: East and West (Jinan, 1986)", Ann. New York Acad. Sci. **576** (1989), 132–145

[69] Erdős P., *Some remarks on number theory III*, Math. Lapok **13** (1962), 28–38.

[70] Erdős P., Freiman G.A., *On two additive problems*, J. Number Theory, **34** (1990), 1–12.

[71] Erdős P., Ginzburg A., and Ziv A., *Theorem in the additive number theory*, Bull. Research Council Israel **10F** (1961), 41–43.

[72] Erdős P., Graham R.L., *On a linear diophantine problem of Frobenius*, Acta Arithmetica **XXI** (1972), 399–408.

[73] Erdős P., Heilbronn H., *On the addition of residue classes mod p*, Acta Arithmetica, **9** (1964), 149–159.

[74] Erdős P., Nathanson M.B., Sárkőzy. A., *Sumsets containing infinite arithmetic progressions*, J. Number Theory, **28** (1988), 159–166.

[75] Erdős P., Nicolas J-L., Sárkőzy A., *Sommes de sousensembles*, Sem. Th. Nb. Bord. **3** (1991), 55–72.

[76] Erdős P., Sárkőzy A., *Arithmetic progressions in subset sums*, Discrete Math. **102(3)** (1992), 249–264.

[77] Esseen, C.G., *On the Kolmogorov-Rogosin inequality for the concentration functions*, Z. Wahrscheinlichkeitstheorie und verw. Gebiete **5** (1966), 210–216.

[78] Folkman J., *On the representation of integers as sums of distinct terms from a fixed sequence*, Canad. J. Math. **18** (1966), 643–655.

[79] Freiman G.A., *An analytical method of analysis of linear Boolean equations*, Ann. N.Y. Acad. Sci. **337** (1980) 97–102.

[80] Freiman G.A., *Dense sequences in the theory of partitions*, Elabuz. Gos. Ped. Inst. Učen. Zap. **3** (1958), 120–137 [Russian].

[81] Freiman G.A., "Foundations of a structural theory of set addition", Elabuz. Gos. Ped. Inst., Kazan, 1966 [Russian].

[82] Freiman G.A., "Foundations of a structural theory of set addition", Translations of Mathematical Monographs **37**, Amer. Math. Soc., Providence, R.I., 1973.

[83] Freiman G.A., *Groups and the inverse problems of the additive set theory*, in "Number-theoretic investigations on the Markov spectrum and the structure theory of set addition", Kalinin Gos. Univ. Moscow 1973, 175–183 [Russian].

[84] Freiman G.A., *Inverse problems in additive number theory*, Proc. of the IV All-Union Math. Congr. **2** (1964), 142–146.

[85] Freiman G.A., *Inverse problems in additive number theory*, Uč. Zap. Kazan Univ. **115(14)** (1955), 109–115 [Russian].

[86] Freiman G.A., *Inverse problems in additive theory of numbers*, Izv. Acad. Nauk SSSR Ser. Mat. **19** (1955) 275–284 [Russian].

[87] Freiman G.A., *The addition of finite sets I*, Izv. Vysš. Učebn. Zaved. Matematika**6(13)** (1959), 202–213 [Russian]

[88] Freiman G.A., *Inverse problems of the additive theory of numbers. On the addition of sets of residues with respect to a prime modulus*, Dokl. Akad. Nauk SSSR **141(3)** (1961), 571–573 [Russian]; Soviet Math. Dokl. **2** (1961), 1520–1522 [English translation].

[89] Freiman G.A., *Inverse problems in additive number theory VI. On the addition of finite sets III. Addition of different sets*, Izv. Vysš. Učebn. Zaved. Mathematika **3(28)** (1962), 151–157 [Russian].

[90] Freiman G.A., *Inverse problems in additive number theory VII. On the addition of finite sets IV. The method of trigonometric sums*, Izv. Vysš. Učebn. Zaved. Matematika **6(31)** (1962), 131–134 [Russian].

[91] Freiman G.A., *Inverse problems in additive number theory VIII. On a conjecture of P. Erdős*, Izv. Vysš. Učebn. Zaved. Matematika **3(40)** (1964), 156–169 [Russian].

[92] Freiman G.A., *Inverse problems in additive number theory IX. The addition of finite sets V*, Izv. Vysš. Učebn. Zaved. Matematika **6(43)** (1964), 168–178 [Russian].

[93] Freiman G.A., *New analytical results in subset sum problem*, Proc. of the French-Israeli Conference on Combinatorics and Algorithms, Jerusalem 1988, Discrete Math. **114** (1993), 205–217.

[94] Freiman G.A., *Nonclosed semigroups with cancellations*, Ann. N.Y. Acad. Sci. **410** (1983), 91–98.

[95] Freiman G.A. (Editor), "Number-Theoretic Studies in Markov Spectrum and in the structural theory of set addition", Kalinin Gos. Univ. Moscow 1973 [Russian].

[96] Freiman G.A., *On extremal additive problems of Paul Erdos*, in "The Proceedings of the Second International Conference on Combinatorial Mathematics and Computing, Canberra, 1987", ARS Combinatoria **26B** (1988), 93–114.

[97] Freiman G.A., *On solvability of a system of two boolean linear equations*, Number theory (New York, 1991-1995), 135–150, Springer, New York, 1996.

[98] Freiman G.A., *On the addition of finite sets*, Dokl. Akad. Nauk SSSR **158** (1964), 1038–1041 [Russian].

[99] Freiman G.A., Pitman J., *Partitions into distinct large parts*, J. Austral. Math. Soc. Ser. A **57(3)** (1994), 386–416.

[100] Freiman G.A., *On the measure of large trigonometric sums*, Ann. N.Y. Acad. Sci. **452** (1985), 363–371.

[101] Freiman G.A., *On the structure and the number of sum-free sets*, Asterisque **209** (1992), 195–203.

[102] Freiman G.A., *On two- and three-element subsets of groups*, Aequationes Math. **22** (1981), 140–152.

[103] Freiman G.A., *Subset-sum problem with different summands*, Congressus Numerantium **70** (1990), 207–215.

[104] Freiman G.A., *Sumsets and powers of 2*, Coll. Math. Soc. J. Bolyai **60** [Budapest] (1991), 279–286.

[105] Freiman G.A., *What is the structure of K if $K + K$ is small?*, in "Lecture Notes in Mathematics **1240**", Springer-Verlag, New York 1987, 109–134.

[106] Freiman G.A., Low L., Pitman J., *Sumsets with distinct summands and the conjecture of Erdős'-Heilbronn on sums of residues*, this volume,

[107] Freiman G.A., Heppes A., Uhrin B., *A lower estimation for the cardinality of finite difference sets*, Problems of Computer Science **202** (1987), 63–73.

[108] Freiman G.A., Heppes A., Uhrin B., *A lower estimation for the cardinality of finite difference sets in R^n*, in "Proc. Conf. Number Theory, Budapest 1987", Coll. Math. Soc. J. Bolyai **51**, North-Holland and Bolyai Taursulat, Budapest 1989, 125–139

[109] Freiman G.A., Halberstam H., Ruzsa I.Z., *Integer sum sets containing long arithmetic progressions*, J. London Math. Soc. **46(2)** (1992), 193–201.

[110] Freiman G.A, Yudin A.A., *The general principles of additive number theory*, in "Number theory", Kalinin Gos. Univ. Moscow 1973, 135–147 [Russian].

[111] Freiman G.A, Yudin A.A., Moskvin D.A., *Inverse problems of additive number theory and local limit theorems for lattice random variables*, in"Number Theory", Kalinin Gos. Univ. Moscow 1973, 148–162 [Russian].

[112] Freiman G.A, Yudin A.A., Moskvin D.A., *Structural theory of set addition and local limit theorems for independent lattice random variables*, Teor. Verojatnost. i Primen. **19** (1974), 52–62 [Russian].

[113] Freiman G.A., Pigarev P.A., *The relation between the invariants R and T*, in "Number Theory", Kalinin Gos. Univ. Moscow 1973, 172–174 [Russian].

[114] Freiman G., *Structure theory of set addition*, this volume,

[115] Freiman, G.A., Robinson D., Shein B., *Structure of $R(3,4)$-groups*, manuscript 1995.

[116] Freiman G.A., Shein B.M., *Group and semigroup theoretic considerations inspired by inverse problems of additive number theory*, in "Lecture Notes in Mathematics **1320**", Springer-Verlag, New York 1988, 121–140.

[117] Freiman G.A., Shein B.M., *Interconnections between the structure theory of set addition and rewritability in groups*, Proc. of Amer. Math. Soc. **113(4)** (1991), 899–910.

[118] Freiman G.A., Shein B.M., *Structure of $R(3,3)$-groups*, Israel Journal of Mathematics, **77** (1992), 17–31.

[119] Galil Z., Margalit O., *An almost linear-time algorithm for the dense subset-sum problem*, SIAM J. Comput. **20**, (1991), no. 6, 1157–1189.

[120] Granovsky B.L., Passov E., *Chebyshev systems of minimal degree*, SIAM J. Math. Anal. **15** (1984), 166–169.

[121] Granovsky B.L., *Moment spaces of minimal dimension*, Journal of Approximation Theory, **49(4)**, (1987), 390–397.

[122] Gustafson P.W.H., *What is the probability that two group elements commute?*, Amer. Math. Monthly **80** (1973), 1031–1034.

[123] Hadwiger H., *Minkowskische Addition und Subtraktion beliebiger Punktmengen und die Theoreme von Erhard Schmidt*, Math. Z. **53** (1950), 210–218.

[124] Hamidoune Y.O., *Sur les atomes d'un graphe orienté*, C.R. Acad. Sci. Paris A **284** (1977), 1253–1256.

[125] Hamidoune Y.O., *Quelques problèmes de connexité dans les graphes orienté*, J. Comb. Theory B **30** (1981), 1–10.

[126] Hamidoune Y.O., *An application of connectivity theory in graphes to factorizations of elements in groups*, Europ. J. Comb. **2** (1981), 349–355.

[127] Hamidoune Y.O., *On the connectivity of Cayley digraphs*, Europ. J. Comb. **5** (1984), 309–312.

[128] Hamidoune Y.O., *On a subgroup contained in words with a bounded length*, Discrete Math. **103** (1992), 171–176.

[129] Hamidoune Y.O., *Subsets with small sums in abelian groups, I.*, European J. Combin., **18**, (1997), no. 5, 541–556.

[130] Hamidoune Y.O., Rödseth Ö.J., *On bases in σ-finite groups*, Math. Scand. **78** (1996), no. 2, 246–254.

[131] Hamidoune Y.O., Llàdo A., Serra O., *Vosperian and superconnected abelian Cayley digraphs*, Graphs and Combinatorics **7** (1991), 143–152.

[132] Hamidoune Y.O., *On the representation of some integers as a subset sum*, Bull. London Math. Soc., **26**, (1994), 557–563.

[133] Hamidoune Y.O., *On weighted sums in abelian groups*, Discrete Math., **162**, (1996), 127–132.

[134] Hamidoune Y.O., *On inverse additive problems*, Report Institut Blaise Pascal, EC9501 (1995).

[135] Hamidoune Y.O., *The representation of some integers as a subset sum*, EC 94/03, preprint March 1994.

[136] Hamidoune Y.O., *Subsets with a small product in groups*, this volume.

[137] Hamidoune Y.O., *An Isoperimetric method in Additive Theory*, J. Algebra, **179**, (1996), 622–630.

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

[139] Henstock R., Macbeath A.M., *On the measure of sum-sets I*, Proc. London Math. Soc. **3(3)** (1953), 182–194.

[140] Herzog M., *New results on subset multiplication in groups*, this volume.

[141] Herzog M., Arad Z., *Products of conjugacy classes in groups*, Lecture notes in Mathematics **1112**, Springer-Verlag, 1985.

[142] Herzog M., Longobardi P., Maj M., *On a combinatorial problem in group theory*, Israel J. Math., **82**, (1993), no. 1-3, 329–340.

[143] Jeroslow R.G., *Trivial integer programs unsolvable by branch and bound*, Mathematical Programming bf 6 (1974), 105–109.

[144] Joseph K.S., *Commutativity in non-Abelian groups*, Ph.D. Thesis, University of California, Los-Angeles 1969.

[145] Kaplansky I., *Groups with representations of bounded degree*, Canad. J. Math. **1** (1949), 105–112.

[146] Kemperman J.H.B., *On complexes in a semigroup*, Indagat. Math. **18** (1956), 247–254.

[147] Kemperman J.H.B., *On product sets in locally compact groups*, Fund. Math. **56** (1964), 51–68.

[148] Kemperman J.H.B., *On small sumsets in an abelian group*, Acta Math. **103** (1960), 63–88.

[149] Kemperman J.H.B., Scherk P., *On sums of sets of integers*, Can. J. Math. **6** (1954), 238–252.

[150] Kesten M., *A sharper form of the Doeblin-Levy-Kolmogorov-Rogosin inequality for concentration functions*, Math. Scand. **25** (1969), 133–144.

[151] Khintchine A., in "Three pearls of number theory", Graylock, Rochester, New York, 1952.

[152] Kiefer J., Wolfowitz J., *Optimum designs in regression problems*, Ann. Math. Stat. **30** (1959), 271–294.

[153] Kneser M., *Abschatzung der asymptotischen Dichte von Summenmengen*, Math. Zeit. **58** (1953), 459–484.

[154] Kneser M., *Ein Satz uber Abelsche Gruppen mit Anwendungen auf die Geometrie der Zahlen*, Math. Z. **61** (1955), 429–434.

[155] Kneser M., *Summenmengen in lokalkompakten abelschen Gruppen*, Math. Z. **66** (1956), 88–110.

[156] Kolmogorov A.N., *Sur les propriétés des fonctions de concentrations de M.P. Lévy*, Ann. Inst. H. Poincaré Sect. B. **16(1)** (1958), 27–34.

[157] Lev V.F., P. Smeliansky, *On addition of two distinct sets of integers*, Acta Arithmetica, **LXX.1**, (1995), 85–91.

[158] Lev V.F., *On the structure of sets of integers with small doubling property* ($|A+A| < \frac{10}{3}|A| - 5$), unpublished manuscript.

[159] Lev V.F., *On the extremal aspect of Frobenius problem*, J. Comb. Th. (Series A), **73** (1), (1996), 111–119.

[160] Lev V.F., *Representing powers of 2 by a sum of four integers*, Combinatorica, **16** (3) (1996), 413–416.

[161] Lev V.F., *Structure theorem for multiple addition and the Frobenius problem*, Journal of Number Theory, **58** (1), (1996), 79–88.

[162] Lev V.F., *On small subsets in abelian groups*, this volume.

[163] Lev V., *The structure of multisets with small number of subset sums*, this volume.

[164] Levitin L.B., Hartmann C.R.P., *A new approach to the general minimum distance decoding problem: the zero neighbors algorithm*, IEEE Trans. on Inform. Theory **31(3)** (1985), 378–384.

[165] Levy M.P., Theorie d'addition des variables aleatoires.

[166] Liebeck H., MacHale D., *Groups with automorphisms inverting most elements*, Math. Z. **124** (1972), 51–63.

[167] Lipkin E., *On representation of r–th powers by subset-sums*, Acta Arithmetica **LII** (1989), 353–366.

[168] Lipkin E., *On subset sums of r–sets*, Discrete Mathematics **114** (1993), 1–3 and 367–377.

[169] Lipkin E., *Subset sums of sets of residues*, this volume,

[170] Longobardi P., Maj M., *The classification of groups with the small squaring property on 3–sets*, Bull. Austral. Math. Soc. **46** (1992), 263–269.

[171] Longobardi P.,Maj M., *On groups in which every product of four elements can be reordered*, Arch. Math. **49** (1987), 273–276.

[172] Longobardi P., Maj M., *On the derived length of groups with some permutational properties*, manuscript.

[173] Longobardi P., Maj M., Stonehewer S.E., *Classification of groups in which every product of four elements can be reordered*, Rend. Sem. Mat. Univ. Padova, **93**, (1995), 7–26.

[174] Macbeath A.M., *On the measure of product sets in a topological group*, J. London Math. Soc. **35** (1960), 403–407.

[175] Macbeath A.M., *On the measure of sum sets, II, The sum theorem for the torus*, Proc. Cambridge Philos. Soc. **49** (1953), 40–43.

[176] Macbeath A.M., *On the measure of sum sets, III, The continuous $a - b$ theorem*, Proc. Edinburg Math. Soc. **12(2)** (1960/61), 209–211; correction ibid. 14(1964/65), 165–166.

[177] Maming W.A., *Groups in which a large number of operators may correspond to their inverses*, Trans. Amer. Math. Soc. **7** (1906), 233–240.

[178] Mann H.B., *A proof of the fundamental theorem on the density of sums of sets of positive integers*, Ann. Math. **43(2)** (1942), 523–527.

[179] Mann H.B., *Additive group theory — a progress report*, Bull. Amer. Math. Soc. **79(6)** (1973), 1069–1075.

[180] Mann H.B., *Two addition theorems*, J. Comb. Th. **3** (1967), 233–235.

[181] Mann H.B., Olson J., *Sums of sets in the elementary abelian group of type (p,p)*, J. Comb. Th. **2** (1967), 275–284.

[182] Margalit O., *Efficent elementary methods for the dense subset-sum problem*, M. Sc. Thesis, Computer Science Department, Tel-Aviv University, 1988.

[183] Martello S., Toth P., *A mixture of dynamic programming and branch-and-bound for the subset-sum problem*, Management Science **30** (1984), 765–771.

[184] Martello S., Toth P., *The 0-1 knapsack problem*, in "Combinatorial Optimization", ed: N. Christofides, A. Mingozzi, P. Toth, C. Sandi, Wiley, 1979, 237–279.

[185] McCrudden M., *On product sets in a unimodular group*, Proc. Cambridge Philos. Soc. **64** (1968), 1001–1007.

[186] Mieses R., Giornale dell'Instituto degli Attuari **5** (1934), 483–495.

[187] Miller G.A., *Groups which admit five-eight automorphisms*, Proc. Nat. Acad. Sci. **17** (1931), 39–43.

[188] Miller G.A., *Groups contaning the largest possible number of operators of order two*, Amer. Math. Monthly **12** (1905), 149–151.

[189] Miller G.A., *Non abelian groups admitting more than half inverse corespondences*, Proc. Nat. Acad. Sci. **16** (1930), 168–172.

[190] Milnor J., *A note on curvature and the fundamental group*, J. Diff. Geom. **2** (1968), 1–7.

[191] Milnor J., *Growth of finitely generated solvable groups*, J. Diff. Geom. **2** (1968), 447–449.

[192] Miroshnikov A.L., Rogosin B.A., *Inequalities for the concentration function*, Theory of probability and its applications, **30(1)** (1983), 38–49.

[193] Mitalauscas A., Statulevicius V., *On local limit Theorems I*, Litovski Math. Sbor. Vol. 14 num. 4, 129–144, 1974.

[194] Mitalauscas A., Statulevicius V., *On local limit Theorems II*, Litovski Math. Sbor. **17(4)** (1977), 169–179.

[195] Moran G., *On product equality preserving mappings in groups*, J. Algebra, **182**, (1996), no. 3, 653–663.

[196] Moskvin D.A., *A local limit theorem for large deviations in the case of differently distributed lattice summands*, Theory of Probability and its Applications **17(4)** (1972), 678–684.

[197] Moskvin D.A., Freiman G.A., Yudin A.A., *Inverse problems of additive number theory and local limit theorems for lattice random variables*, in "Number Theory", Kalinin Gos. Univ. Moscow 1973, 148–162 [Russian].

[198] Moskvin D.A., Postnikova L.O., Yudin A.A., *On an arithmetic method of obtaining local limit theorems for lattice random variables*, Prob. Theor. and its applications **15(1)** (1970), 86–96.

[199] Nathanson M.B., *Sumsets of measurable sets*, Proc. Amer. Math. Soc. **78(1)** (1980), 59–63.

[200] Nathanson M.B., "Additive Number Theory. Inverse Problems and the Geometry of Sumsets.", Graduate Texts in Mathematics, **165**, Springer Verlag, New-York, (1996), xiv+293 pp.

[201] Nathanson M.B. and Sárkőzy A., *Sumsets containing long arithmetic progressions and powers of 2*, Acta Arithmetica **46** (1989), 147–154.

[202] Nathanson M., Tenenbaum G., *Inverse theorems and the number of sums and products*, this volume,

[203] Nemhauzer G., Willey L., "Integer and combinatorial optimization", John Wiley & Sons, 1988.

[204] Neuman B.H., *On a problem of Paul Erdős in groups*, J. Austr. Math. Soc. (Ser. A) **21** (1976), 467–472.

[205] Nicolas J-L., *Stratified Sets*, this volume.

[206] Olson J., *An addition theorem modulo p*, J. Comb. Th. **5** (1968), 45–52.

[207] Olson J., *An Addition Theorem for the Elementary Abelian Group*, J. Comb. Th. **5** (1968), 53–58.

[208] OlsonD J., *Sums of sets of group elements*, Acta Arithmetica, **28** (1975), 147–156.

[209] Olson J., *An addition theorem for finite abelian groups*, J. Number Theory **9** (1977), 63–70.

[210] Olson J., *On a combinatorial problem of Erdős, Ginzberg and Ziv*, J. Number theory **8** (1976), 52–57.

[211] Olson J., *A combinatorial problem on finite abelian groups I and II*, J. Number Theory, **1** (1969), 8–11 and 195–199.

[212] Olson J., *On the sum of two sets in a group*, J. Number Theory, **18** (1984), 110–120.

[213] Passow E., *Alternating parity of Chebyshev Systems*, Journal of Approximation Theory **9** (1973), 295–298.

[214] Postnikov A.G., *Introduction to analytic number theory*, Izdat. "Nauka", Moscow, 1971. 416 pp. [Russian].

[215] Postnikov A.G., *Additive problems with growing number of summands*, IAN, Math. Ser., **20** (1956), 751–764.

[216] Postnikova L.P., Yudin A.A., *On the concentration function*, Theory of Probability and its Applications **22(2)** (1977), 371–375.

[217] Postnikova L.P., Yudin A.A., *An analytic method for estimates of the concentration function*, Proceedings of the Steklov Institute of Mathematics **1** (1980).

[218] Postnikova L.P., Yudin A.A., *A sharper form of an inequality for the concentration function*, Theory Prob. Appl. **23** (1978), 359–362.

[219] Pyber L., *The number of pairwise non-commuting elements and the index of the center in a finite group*, J. London. Math. Soc. **35(2)** (1987), 287–295.

[220] Redei L., *Das 'Schiefe Produkt' in der Gruppentheorie*, Comment. Math. Helvet. **20** (1947), 225–264.

[221] Rhemtulla A.H., Street A.P., *Maximal sum free sets in finite abelian groups*, Bull. Austral. Math. Soc. **2** (1970), 289–297.

[222] Rogosin B.A., *An estimate for concentration functions*, Theory of Probability and its Applications **6** (1961), 94–97.

[223] Rohrbach H., *Anwendung eines Satzes der additiven Zahlentheorie auf eine Grupenteoretische Frage*, Math. Z. **42** (1937), 538–542.

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

[225] Roth K.F., *On certain sets of integers II*, J. London Math. Soc. **29** (1954), 20–26.

[226] Rusin D., *What is the probability that two elements of a finite group commute?*, Pac. J. Math. **2(1)** (1979), 237–247.

[227] Ruzsa I.Z., *The density of the set of sums*, Acta Arith.,**58**, (1991), 169–172.

[228] Ruzsa I.Z., *Sums of finite sets*, Number theory (New York seminar, 1991–1995), eds. D. V. Chudnovsky, G. V. Chudnovsky, M. B. Nathanson, Springer, New York, (1996), 281–293.

[229] Ruzsa I.Z., *On the cardinality of $A+A$ and $A-A$*, in "Combinatorics", Eds. A. Hajnal, V.T. Sos, Coll. Math. Soc. J. Bolyai **18**, North Holland 1978, 933–938.

[230] Ruzsa I.Z., *On the number of sums and differences*, Acta Math. Hung. **59** (1992), 439–447.

[231] Ruzsa I.Z., *Sets of sums and differences*, in "Proc. de Seminare de Theorie des nombres de Paris (1982–1983)", Birkhauser, Boston 1984, 267–273.

[232] Ruzsa I.Z., *Sums of sets in several dimensions*, Combinatorica, **14**, (1994), 485–490.

[233] Ruzsa I.Z., *Generalized arithmetical progressions and sumsets*, Acta Math. Hungar., **65**, (1994), 379–388.

[234] Ruzsa I.Z., *Arithmetic progressions in sumsets*, Acta Arith. **60(2)** (1991), 191–202.

[235] Ruzsa I.Z., *An application of graph theory to additive number theory*, Scientia (Series A) Math. Sciences **3** (1989), 97–109.

[236] Ruzsa I.Z., *Sets of sums and commutative graphs*, Proc. of the workshop in combinatorics, Bielefeld 1991, Studia Sci. Math. Hungar., **30**, (1995), 127–148.

[237] Ruzsa I.Z., *Arithmetic progressions and the number of sums*, Period. Math. Hung. **25(1)**[3] (1992), 105–111.

[238] Ruzsa I.Z., An analog of Freiman's theorem in groups, this volume.

[239] Sárkőzy A., *Finite addition theorems I*, J. Number Theory **32(1)** (1989), 114–130

[240] Sárkőzy A., *Finite addition theorems II*, J. Number Theory, **48**, (1994), no. 2, 197–218.

[241] Sárkőzy A., *Finite addition theorems III*, in "Groupe de Travail en Theorie Analytique et Elementaire des Nombres 1989–1990", Publ. Math. Orsay 1992, 105–122.

[242] Schnirelman L.G., *Uber additive Eigenschaften von Zahlen*, Math. Ann. **107** (1933), 649–690.

[243] Semple J.F., Shalev A., *Combinatorial conditions in residually finite groups I*, J. Algebra **157(1)** (1993), 43–50.

[244] Shalev A., *Combinatorial conditions in residually finite groups II*, J. Algebra **157(1)** (1993), 51–62.

[245] Siegel C.L., *Einheiten quadratischer Formen.*

[246] Dias da Silva J.A. and Hamidoune Y.O., *Cyclic spaces for Grassman derivatives and additive theory*, Bull. London Math. Soc. **26** (1994), 140–146.

[247] Straus E.G., *On a problem in combinatorical number theory*, J. Math. Sci. **1** (1966), 77–80.

[3] Vol.25 No.1?

[248] Szemeredi E., *On sets of integers containing no k elemenents in arithmetic progression*, Acta Arithmetica **27** (1975), 199–245.

[249] Szemeredi E., *On a conjecture of Erdős and Heilbronn*, Acta Arithmetica **17** (1970), 227–229.

[250] Szemeredi E., *Integer sets containing no arithmetic progression*, Math. Acad. Sci. Hungar. **56** (1990), 155–158.

[251] Szoni T., Wettl F., *On complexes in a finite abelian group*. Proc. of the Japan Academy **64(7)** (Series A) **7** (1988), 245–246.

[252] Tashbaev V.H., *An inverse additive problem*, Math. Sb. **52(94)** (1960), 947–952 [Russian].

[253] Uhrin B., *On a generalization of the Minkowsky convex body theorem*, J. of Number Theory **13** (1981), 192–209.

[254] Uhrin B., *Some estimations useful in the geometry of numbers*, Period. Math. Hungar. **11** (1980), 95–103.

[255] Uhrin B., *Some remarks about the lattice points in difference sets*, in "Proc of A. Haar Memorial Conf. (Budapest, 1985)", Ed. J. Szabados, Coll. Math. Soc. J. Bolyai **49**, North-Holland, Amsterdam-New York, 1986, 929–937.

[256] Usharov N.G., *Upper estimates of maximum probability for sums of independent random vectors*, Theory of probability and its applications **30(1)** (1983), 38–49 [Russian].

[257] Vosper A.G., *The critical pairs of subsets of a group of prime order*, J. London Math. Soc. **31** (1956), 200–205; see addendum in J. London Math. Soc. **31** (1956), 280–286.

[258] Wall C.T.C., *On groups consisting mostly of involutions*, Proc. Cambridge Philos. Soc. **67(2)** (1970), 251-262.

[259] Wolf J., *Growth of finitely generated solvable groups and curvature of Riemannian manifolds*, J. Diff. Geom. **2** (1968), 421–446.

[260] Yudin A.A., *The measure of the large values of the modulus of a trigonometric sum*, in "Number theoretic studies in the Markov spectrum and in the structural theory of set addition", Kalinin Gos. Univ., Moscow 1973, 163–171 [Russian] .

[261] Hennecart F., Robert G., Yudin A., *On the number of sums and differences*, this volume,

[262] Zemor G., *Subset sums in binary spaces*, Europ. J. Combin., (1992) 13, 221–230.

[263] Zemor G., *A generalisation to non-commutative groups of a theorem of Mann*, Discrete Math.,**126**, (1994), no. 1-3, 365–372.

[264] Zemor G ., *An extremal problem related to the covering radius of binary codes*, in "First French-Soviet Workshop on algebraic coding", Lecture Notes in Computer Science **573**, Springer-Verlag 1992, 42–51.

[265] Zemor G., Cohen G.D., *Error-correcting WOM-codes*, IEEE Trans. on Information Theory **37(3)** (1991), 730–734.

[266] Zemor G., Cohen G., *Applications of coding theory to interconnection networks*, Discrete Applied Math. **37/38** (1992), 553-562.

[267] Zigel G., *Upper estimations for the concentration function in Hilbert space*, Theory of Probability and its applications **26(2)** (1982), 328–343.

[268] Straus E.G., *Non-averaging sets*, in "Combinatorics: conference at Univ. California, Los Angeles, 1968", Proc. Sympos. Pure Math. **XIX**, Amer. Math. Soc., Providence, R.I. 1971, 215–222.

G.A. FREIMAN, School of Mathematical Sciences, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, 69978 Tel Aviv, Israel • *E-mail :* grisha@math.tau.ac.il

AMNON BESSER

Sets of integers with large trigonometric sums

Astérisque, tome 258 (1999), p. 35-76

<http://www.numdam.org/item?id=AST_1999__258__35_0>

Astérisque **258**, 1999, p. 35–76

SETS OF INTEGERS WITH LARGE TRIGONOMETRIC SUMS

by

Amnon Besser

Abstract. — We investigate the problem of optimizing, for a fixed integer k and real u and on all sets $K = \{a_1 < a_2 < \cdots < a_k\} \subset \mathbb{Z}$, the measure of the set of $\alpha \in [0,1]$ where the absolute value of the trigonometric sum $S_K(\alpha) = \sum_{j=1}^{k} e^{2\pi i \alpha a_j}$ is greater than $k - u$. When u is sufficiently small with respect to k we are able to construct a set K_{ex} which is very close to optimal. This set is a union of a finite number of arithmetic progressions. We are able to show that any more optimal set, if one exists, has a similar structure to that of K_{ex}. We also get tight upper and lower bounds on the maximal measure.

1. Introduction

Let k be a positive integer and $u < k$ a positive real. For a set

$$K = \{a_1 < a_2 < \cdots < a_k\}, \quad a_j \in \mathbb{Z}, \quad 1 \leq j \leq k,$$

let

$$S_K(\alpha) = \sum_{j=1}^{k} e^{2\pi i \alpha a_j}, \quad s_K(\alpha) = |S_K(\alpha)|,$$

$$E_{K,u} = \{\alpha \in [0,1) : \quad s_K(\alpha) \geq k - u\}$$

and

$$\mu_K(u) = \mu(E_{K,u}),$$

where μ is the Lebesgue measure on $[0,1]$ normalized so $\mu([0,1]) = 1$.

This work deals with the following problem, first raised at the talk of Freiman and Yudin at the Number Theory Conference (Vladimir, 1968):

Problem 1. — *Find the set K which maximizes $\mu_K(u)$ and find the maximal value.*

1991 ***Mathematics Subject Classification.*** — Primary 11L03; Secondary 42A05.
Key words and phrases. — Trigonometric sums.

We denote by $\mu_{\max}(k,u)$ the supremum of $\mu_K(u)$ on all sets K of size k.

The first results on this problem were obtained by Freiman and Yudin:

***Theorem 1* (Freiman, [2, page 144]).** — *For $u = 1$, $a_1 = 0$ and $a_k < 0.05k^{3/2}$, the maximal measure is*

$$\mu_{\max}(k,u) = \frac{2\sqrt{6}}{\pi}k^{-3/2} + O(k^{-2})$$

and it is attained by K if and only if K is an arithmetic progression.

***Theorem 2* (Yudin, [4]).** — *For $u = o(k)$*

$$\mu_{\max}(k,u) = \frac{2\sqrt{6}}{\pi}\frac{1}{k}\left(\frac{u}{k}\right)^{1/2}(1 + o(1))$$

as $k \longrightarrow \infty$.

In [1] Freiman treated the problem assuming the ratio u/k is small enough. He sketched an approach for attacking the problem and conjectured it would prove that the best set is an arithmetic progression.

The purpose of this work is to carry out and extend Freiman's approach (and also that of [4]). It will turn out that once u is sufficiently large it is no longer true that an arithmetic progression attains $\mu_{\max}(k,u)$. We are unable to find a set which does. Nevertheless, we do describe to some extent the "structure" of the maximal set. To make precise what this means, we will introduce and use the following terminology

Definition 1.1. — Let k and u be as above.

1. For any $\psi \in [0,1]$ we let $\mathbb{K}_\psi$ be the collection of all sets $K \subset \mathbb{Z}$ of size k that satisfy $\mu_K(u) \geq \psi$.
2. A collection $\mathbb{K}$ of sets is said to be "good for ψ", or to be a G_ψ collection, if it satisfies the following two properties:
   (a) We have $\mathbb{K} \subset \mathbb{K}_\psi$,
   (b) For any set $K \subset \mathbb{Z}$ of size k there exists a set $K' \in \mathbb{K}$ such that $\mu_K(u) \leq \mu_{K'}(u)$.

Our main results are of three types. We are able to describe the "structure" of sets in $\mathbb{K}_\psi$ for a ψ which is very close to maximal. In addition we construct a certain subcollection of this which has property G_ψ. The subclass we describe is not a singleton but it does have a rather simple structure: it is essentially the union of arithmetic progressions, and we have a fairly accurate information about the location and length of all these sequences. Lastly, we get a good bound on $\mu_{\max}(k,u)$.

The type of results we get is dictated by our method of proof, which could be describe as an iteration of four steps.

1. We first guess a set K expected to have a large $\mu_K(u)$. We will take $\psi = \mu_K(u)$.
2. We get information about sets K_1 that have an even higher $\mu_{K_1}(u)$. Typically this information consists of knowledge that most elements are contained inside arithmetic progressions of relatively short length. This is what we mean by describing the "structure" of $\mathbb{K}_\psi$.

3. Given a set $K_1 \in \mathbb{K}_\psi$, we give a procedure for obtaining out of K_1 a set K_2 such that $\mu_{K_2}(u) \geq \mu_{K_1}(u)$. The procedure usually involves compressing the elements contained in the short progressions described above to form short progressions, possibly with a single gap. Sets obtained in this way will form a subclass $\mathbb{K}$ that has property G_ψ by construction.
4. We use the knowledge of $\mathbb{K}_\psi$ to get an improved bound on $\mu_{\max}(k, u)$.

Our results apply under the assumptions $k/30000 \geq u > 1$ and $k \geq$ Const, with Const an unspecified constant. We note that this second assumption is only forced on us because we are using lemma 5.2 which is ineffective. If an effective bound is supplied for that lemma, it will be very easy to deduce an effective lower bound on k as well.

We state a simplified version of the results here under the additional assumption $u > k^{2/3}$. With these assumptions, it follows (proposition 3.1) that for an arithmetic progression K of difference 1 and length k there exists some $\beta_{k,u}$(definition 3.2) such that

$$E_{K,u} = [-\beta_{k,u}, \beta_{k,u}] \pmod 1.$$

We will see in proposition 3.4 that

$$\beta_{k,u} \approx \frac{\sqrt{6}}{\pi}\frac{1}{k}\left(\frac{u}{k}\right)^{1/2}.$$

We describe a certain basic set K_{ex} (a more precise description will be given in construction 6.1). Set $m_0 = k - 5u/12$ and $\beta = \beta_{m_0,u}$. To first order, $\beta_{m_0,u} \approx \beta_{k,u}(1 + 5u/8k)$. The set K_{ex} is the union of an arithmetic progression of length m_0, symmetric around 0, and for any non zero integer n an arithmetic progression of length

$$\frac{1}{2}m_n = \frac{u}{(\pi n)^2}\left(1 - \frac{(-1)^n}{2}\right),$$

centered around $\frac{n}{\beta}$. All the arithmetic progressions here have difference 1. The structure of K_{ex} and the particular values of the m_n are chosen in such a way that the contributions of the shorter progressions to $S_{K_{ex}}(\alpha)$ exactly compensates for the decline of the contribution of the large progression when $|\alpha| > \beta/2$. We show in proposition 6.12 that $\mu_{K_{ex}}(u) \approx 2\beta$. The results are now as follows.

1. $$\mu_{\max}(k, u) = 2\beta_{m_0,u}\left(1 + O\left(\left(\frac{u}{k}\right)^2\right)\right).$$

2. A set $K \in \mathbb{K}_{\mu_{K_{ex}}(u)}$ has the following structure (similar to that of K_{ex}).
   (a) All but

$$\frac{5}{12}u + O\left(\frac{u^2}{k}\right)$$

elements of K are contained in a short arithmetic progression of length $1/4\beta$. To state the other results we will assume that this progression is symmetric around 0 and its difference is 1. The general case is essentially the same by translation and dilation which change nothing.

(b) Most other elements are contained in a union of short arithmetic progressions with centres near $\frac{n}{\beta}$ and $-\frac{n}{\beta}$ for $n \in \mathbb{N}$. each such short progression has length $2k$ at the most. The number of elements contained in progressions near $\pm\frac{n}{\beta}$ is

$$m_n + O\left(\frac{u^2}{k}\right).$$

(c) The number of elements not contained in any of the progressions above is $O\left((u/k)^{1/2}u\right)$.

3. The following subclass of $\mathbb{K}_{\mu_{K_{ex}}(u)}$ is of type $G_{\mu_{K_{ex}}(u)}$. It consists of the sets where all the elements *contained* in the progression described in (a) above in fact *form* an arithmetic progression except that one gap may persist.

All O terms can and will be made explicit although no claim for best bounds is made.

Here is a brief summary of the contents of this paper. In section 3 we estimate $\mu(E_{K,u})$ in the case where K is an arithmetic progression, and we prove the lower bound:

$$\mu(E_{K,u}) \leq \frac{2\sqrt{6}}{\pi}\frac{1}{k}\left(\frac{u}{k}\right)^{1/2}$$

for such a progression. In section 4 we prove the upper bound

$$\mu(E_{K,u}) \leq \frac{d}{k}\left(\frac{u}{k}\right)^{1/2}$$

holds for an explicitly given $d \approx 4$ and all sets K under some mild restrictions on k and u. In section 5 we consider a set

$$K \in \mathbb{K}_{\psi}, \quad \psi = \frac{2\sqrt{6}}{\pi}\frac{1}{k}\left(\frac{u}{k}\right)^{1/2}.$$

We show that this implies that $E_{K,u}$ is contained in a union of small intervals and that K has most of its elements contained in an arithmetic progression of short length. We then perform the first of our "compression arguments" mentioned above and construct a G_{ψ} subclass consisting of the sets where these elements form an arithmetic progression with at most one gap. The construction of K_{ex} is described in section 6. In sections 7 to 9 we describe the structure of $\mathbb{K}_{\mu_{K_{ex}}(u)}$ and also describe a $G_{\mu_{K_{ex}}(u)}$ subclass.

Some of the results of this paper appeared in [**1**]. We follow [**1**] very closely in sections 3 to 5. We note that the argument in [**1**, p. 368] may be completed to give the result that, the part of K not in an arithmetic progression is bounded in size by cu with $c \longrightarrow 1/(2-4/\pi)$ as $u/k \longrightarrow 0$. This result is improved here to $c \approx 5/12$.

It is a great pleasure to thank Prof. Freiman for the help and fruitful discussions during the preparation of this work. I would also like to thank the referee for making many valuable remarks.

2. Notation and terminology

An arithmetic progression $\{a, a+q, a+2q, \dots, a+(n-1)q\}$ is said to have length n and difference q. We will often make the distinction between a set being *contained* in an arithmetic progression, which means it is a subset of the above, and set *forming* and arithmetic progression. We will also sometime talk about an interval of integers. This will mean a set of the form $\{x \in \mathbb{Z} : \ s \leq x \leq t\}$. This interval has length $t - s$. Notice that a subset of $\{a, a+q, a+2q, \dots, a+(n-1)q\}$ is contained in an arithmetic progression of length n but in an interval of length $(n-1)q$.

In this paper, excluding the introduction, we will make a non-standard use of the notations $O(1)$ and $o(\epsilon)$. The notation $O(1)$ will mean having absolute value ≤ 1. In particular, we will write $a = b + O(1)c$ to mean $|a - b| \leq c$. We will use $o(\epsilon)$ to refer to a quantity which is very small and will be discarded in the computation by swallowing it in a larger quantity. A typical use of this will be for example $8(1 + o(\epsilon)) \leq 9$. The reader will have to check for himself or herself that such an argument is justified, which should not be too hard. This notation is used because we have been asked to give explicit, while not best possible, upper bounds for everything.

3. The case of arithmetic progressions

As remarked in the introduction, we always assume that $u > 1$. In this section we want to determine $s_K(\alpha), E_{K,u}$ and $\mu_K(u)$ when K is an arithmetic progression. It will be occasionally convenient to write, when K is of length k and difference 1, $s_k(\alpha)$ for $s_K(\alpha)$, $E_{k,u}$ for $E_{K,u}$ and $\mu_k(u)$ for $\mu_K(u)$.

We first note that, for any set K and integers d and m,

$$s_{dK+m}(\alpha) = s_K(d\alpha). \tag{1}$$

Therefore

$$E_{dK+m,u} = \langle d \rangle^{-1}(E_{K,u}). \tag{2}$$

Here, the map $\langle d \rangle : [0,1) \to [0,1)$ is defined by $\langle d \rangle(\alpha) =$ fractional part of $d\alpha$. For $F \subset [0,1)$, $\langle d \rangle^{-1}(F)$ denotes the inverse image of F under $\langle d \rangle$. It is easy to deduce from this that

$$\mu_{dK+m}(u) = \mu_K(u).$$

These observations allow us to reduce to the case of difference 1.

Proposition 3.1. — *Let K be an arithmetic progression of length k and difference* 1. *Assume $k \geq 2u$.*

1. *We have*

$$s_K(\alpha) = \left| \frac{\sin(\pi\alpha k)}{\sin(\pi\alpha)} \right| \quad \textit{when } \alpha \neq 0.$$

2. *The set $E_{k,u}$ is a single interval modulo* 1, *i.e.,*

$$E_{K,u} = [-\beta, \beta] \pmod 1,$$

For some $\beta \in \mathbb{R}$.

It will be worth while to give the number β appearing in the proposition a special notation.

Definition 3.2. — Under the assumptions of proposition 3.1, define $\beta_{k,u} > 0$ by the equality

$$E_{K,u} = [-\beta_{k,u}, \beta_{k,u}] \pmod 1,$$

for K an arithmetic progression of length k and difference 1. Note that

$$s_K(\beta_{k,u}) = k - u.$$

Corollary 3.3. — *If K is an arithmetic progression of length k and difference d, then*

$$E_{K,u} = \bigcup_{q=0}^{d-1} \left[\frac{q}{d} - \frac{\beta_{k,u}}{d}, \frac{q}{d} + \frac{\beta_{k,u}}{d}\right] \pmod 1.$$

Proof. — From proposition 3.1 and (2) it follows that

$$E_{K,u} = \langle d\rangle^{-1}([-\beta_{k,u}, \beta_{k,u}]) = \bigcup_{q=0}^{d-1} \left[\frac{q}{d} - \frac{\beta_{k,u}}{d}, \frac{q}{d} + \frac{\beta_{k,u}}{d}\right] \pmod 1.$$

□

Proof of proposition 3.1. — By (1) it is enough to consider any arithmetic progression of difference 1. In particular, one can take

$$K = \{-\frac{k-1}{2}, \dots, \frac{k-1}{2}\}.$$

Note that this set might be composed of half integers but that makes no difference here. We get for $\alpha \neq 0$,

$$(3) \qquad S_K(\alpha) = \sum_{n=-\frac{k-1}{2}}^{\frac{k-1}{2}} e^{2\pi i \alpha n} = e^{(1-k)\pi i \alpha} \sum_{n=0}^{k-1} e^{2\pi i \alpha n}$$

$$= e^{(1-k)\pi i \alpha} \frac{e^{2\pi i \alpha k} - 1}{e^{2\pi i \alpha} - 1} = \frac{(e^{\pi i \alpha k} - e^{-\pi i \alpha k})/2i}{(e^{\pi i \alpha} - e^{-\pi i \alpha})/2i} = \frac{\sin(\pi \alpha k)}{\sin(\pi \alpha)}.$$

Taking absolute values gives the first assertion. We now have

$$E_{K,u} = \left\{\alpha \in (0,1) : \left|\frac{\sin(\pi \alpha k)}{\sin(\pi \alpha)}\right| \geq k - u\right\} \cup \{0\}.$$

This set is symmetric around $1/2$. It is thus sufficient to consider its intersection with the interval $(0, \frac{1}{2})$. On this set S_K is in fact positive. Indeed, since by assumption $2u \leq k$, we find

$$\frac{1}{2}k \leq k - u \leq \frac{|\sin(\pi \alpha k)|}{\sin(\pi \alpha)} \leq \frac{1}{\sin(\pi \alpha)} \leq \frac{\pi/2}{\pi \alpha},$$

and therefore

$$(4) \qquad \pi \alpha k \leq \pi,$$

which shows that $\sin(\pi\alpha k) \geq 0$. We can also write

$$S_K(\alpha) = \sum_{n=-\frac{k-1}{2}}^{\frac{k-1}{2}} e^{2\pi i\alpha n} = \sum_{n=-\frac{k-1}{2}}^{\frac{k-1}{2}} \cos(2\pi\alpha n).$$

By (4), each term in this sum, and therefore $S_K(\alpha) = |S_K(\alpha)|$, is decreasing in α. Thus, $E_{k,u} \cap [0, \frac{1}{2}]$ is an interval. □

Proposition 3.4. — *When $k > 3u$ we have*

$$\beta_{k,u} = \frac{\sqrt{6}}{\pi}\frac{1}{k}\left(\frac{u}{k}\right)^{1/2}\left(1 + \frac{3u}{20k} + O(1)\left(\frac{u}{k}\right)^2\right).$$

Corollary 3.5. — *When $k > 3u$ we have*

$$\mu_{\max}(k,u) \geq \mu_K(u) \geq \frac{2\sqrt{6}}{\pi}\left(\frac{u}{k}\right)^{\frac{1}{2}}\frac{1}{k}.$$

Proof of proposition 3.4. — Set $\beta = \pi\beta_{k,u}$. Then

$$u = k - \frac{\sin(k\beta)}{\sin(\beta)}.$$

We notice first that

$$\begin{aligned} 0 &\leq \frac{\sin(k\beta)}{\sin(\beta)} - \frac{\sin(k\beta)}{\beta} = \frac{\sin(k\beta)}{\sin(\beta)}\left(\frac{\beta - \sin(\beta)}{\beta}\right) \\ &\leq k\left(\frac{\beta^3/6}{\beta}\right) = \frac{k^3\beta^2}{6}\frac{1}{k^2}. \end{aligned}$$

Now expand

$$\begin{aligned} k - \frac{\sin(k\beta)}{\beta} &= \beta^{-1}(k\beta - \sin(k\beta)) \\ &= \beta^{-1}\left(\frac{(k\beta)^3}{6} - \frac{(k\beta)^5}{120} + \frac{(k\beta)^7}{7!} - \dots\right) \qquad (5) \\ &= \frac{k^3\beta^2}{6}\left(1 - \frac{6(k\beta)^2}{120} + \frac{6(k\beta)^4}{7!} - \dots\right). \end{aligned}$$

It follows that

$$\frac{k^3\beta^2}{6}\left(1 - \frac{(k\beta)^2}{20} + \frac{(k\beta)^4}{840}\right) \geq u \geq \frac{k^3\beta^2}{6}\left(1 - \frac{(k\beta)^2}{20} - \frac{1}{k^2}\right).$$

Thus,

$$\frac{6u}{k^3}\left(1 - \frac{(k\beta)^2}{20} + \frac{(k\beta)^4}{840}\right)^{-1} \leq \beta^2 \leq \frac{6u}{k^3}\left(1 - \frac{(k\beta)^2}{20} - \frac{1}{k^2}\right)^{-1}.$$

We now plug here our first estimate $k\beta \leq \pi$ from (4) to iterate estimates on $(k\beta)^2$. First, since we assume $k > 3u > 3$,

$$1 - \frac{(k\beta)^2}{20} - \frac{1}{k^2} > \frac{1}{3},$$

so

$$(k\beta)^2 < 18\frac{u}{k}.$$

Applying this and using $1/k^2 < u/(3k)$ we now get

$$6\frac{u}{k} \leq (k\beta)^2 < 6\frac{u}{k}\left(1 + \frac{2u}{k}\right).$$

Thus

$$\frac{6u}{k^3}\left(1 - \frac{6u}{20k} + \frac{6^2}{840}\left(\frac{u}{k}\right)^2\right)^{-1} \leq \beta^2$$

and

$$\frac{6u}{k^3}\left(1 - \frac{6u}{20k} - \frac{12}{20}\left(\frac{u}{k}\right)^2 - \left(\frac{u}{k}\right)^2\right)^{-1} \geq \beta^2.$$

The proposition follows easily from this. □

4. An upper bound for $\mu_{\max}(k, u)$

We will prove the upper bound

$$\mu_{\max}(k, u) \leq \frac{d}{k}\left(\frac{u}{k}\right)^{\frac{1}{2}}$$

for a constant $d \approx 4.04$ that will be defined later. Note that this is of the same type as the lower bound we got in corollary 3.5. We will need a few lemmas.

Lemma 4.1. — *For any u and k we have*

$$\mu_{\max}(k, u) \leq \frac{k}{(k-u)^2}.$$

Proof. — Since $s_K(\alpha) \geq k - u$ on a set of measure $\mu_K(u)$, we have

$$(k-u)^2\mu_K(u) \leq \int_0^1 s_K(\alpha)^2 d\alpha.$$

The right hand side can be explicitly computed.

$$\begin{aligned}\int_0^1 s_K(\alpha)^2 d\alpha &= \int_0^1 S_K(\alpha)\overline{S_K(\alpha)}d\alpha \\ &= \int_0^1 \left(\sum_{n=1}^{k} e^{2\pi i\alpha a_n}\right)\left(\sum_{m=1}^{k} e^{-2\pi i\alpha a_m}\right)d\alpha \\ &= \sum_{n,m=1}^{k}\int_0^1 e^{2\pi i\alpha(a_n - a_m)}d\alpha \\ &= \sum_{n=m=1}^{k}\int_0^1 1 d\alpha = k.\end{aligned}$$

This immediately implies the result. □

Lemma 4.2. — *Let $p_1, p_2, \dots, p_k$ be real positive numbers such that $\sum_{i=1}^k p_i = 1$. Let a_i, $i = 1, \dots, k$, be integers. Set $\phi(\alpha) = \sum_{i=1}^k p_i e^{2\pi i \alpha a_i}$. Then,*

$$|\phi(\alpha_1 + \alpha_2)| \geq |\phi(\alpha_1)||\phi(\alpha_2)| - \sqrt{1 - |\phi(\alpha_1)|^2}\sqrt{1 - |\phi(\alpha_2)|^2}.$$

Proof. — We reproduce the proof given in [**5**, Lemma 1]. Let

$$\begin{aligned} v_0 &= (\sqrt{p_1}, \dots, \sqrt{p_k}), \\ v_1 &= (\sqrt{p_1}e^{2\pi i \alpha_1 a_1}, \dots, \sqrt{p_k}e^{2\pi i \alpha_1 a_k}), \\ v_2 &= (\sqrt{p_1}e^{2\pi i \alpha_2 a_1}, \dots, \sqrt{p_k}e^{2\pi i \alpha_2 a_k}), \end{aligned}$$

be three unit vectors in $\mathbb{C}^k$. Then $|\phi(\alpha_1+\alpha_2)| = \cos\theta(v_1, v_2)$ and $|\phi(\alpha_i)| = \cos\theta(v_i, v_0)$ for $i = 1, 2$, where $\theta(v, w)$ is the angle between the vectors v and w. Since $\theta(v_1, v_2) \leq \theta(v_1, v_0) + \theta(v_2, v_0)$ we have

$$\begin{aligned} |\phi(\alpha_1 + \alpha_2)| &\geq \cos\theta(v_1, v_0)\cos\theta(v_2, v_0) - \sin\theta(v_1, v_0)\sin\theta(v_2, v_0) \\ &= \cos\theta(v_1, v_0)\cos\theta(v_2, v_0) \\ &\quad - \sqrt{1 - \cos^2\theta(v_1, v_0)}\sqrt{1 - \cos^2\theta(v_2, v_0)} \\ &= |\phi(\alpha_1)||\phi(\alpha_2)| - \sqrt{1 - |\phi(\alpha_1)|^2}\sqrt{1 - |\phi(\alpha_2)|^2}. \end{aligned}$$

□

Corollary 4.3. — *For any set K and real numbers u_1, u_2 we have $E_{K,u_1} + E_{K,u_2} \subseteq E_{K,2(u_1+u_2)}$.*

Proof. — When $u_1 = u_2$ this was obtained in [**5**]. In the general case, putting in lemma 4.2 $p_i = \frac{1}{k}$ and multiplying by k^2 we get

$$ks_K(\alpha_1 + \alpha_2) \geq s_K(\alpha_1)s_K(\alpha_2) - \sqrt{k^2 - s_K(\alpha_1)^2}\sqrt{k^2 - s_K(\alpha_2)^2}.$$

If we assume $s_K(\alpha_i) \geq k - u_i$ for $i = 1, 2$, then we get

$$\begin{aligned} ks_K(\alpha_1 + \alpha_2) &\geq (k - u_1)(k - u_2) - \sqrt{k^2 - (k - u_1)^2}\sqrt{k^2 - (k - u_2)^2} \\ &= k^2 - k(u_1 + u_2) + u_1u_2 - \sqrt{u_1(2k - u_1)}\sqrt{u_2(2k - u_2)}. \end{aligned}$$

By dropping the term u_1u_2 and replacing $2k - u_i$ by k we see that

$$ks_K(\alpha_1 + \alpha_2) \geq k^2 - k(u_1 + u_2) - k(2\sqrt{u_1}\sqrt{u_2}).$$

Since $2\sqrt{u_1}\sqrt{u_2} \leq u_1 + u_2$ we get

$$ks_K(\alpha_1 + \alpha_2) \geq k^2 - 2k(u_1 + u_2),$$

which implies the result. □

Lemma 4.4. — *If $E \subset [0, 1)$ is closed and $\mu(E) \leq \frac{1}{35}$, then*

$$\mu(E + E \pmod 1) \geq 2\mu(E).$$

Proof. — This is a result of Macbeath and Kneser (see [**6**] for reference). Also, this follows easily from the Theorem in [**3**, p. 46]. The referee informs me that this is also due to Raikov [**7**] with the relaxed condition $\mu(E) \leq \frac{1}{2}$. □

Proposition 4.5. — *Put*

$$c = \frac{\sqrt{2}-1}{4\sqrt{2}-1} \approx 0.09 \ , \ d = \frac{1}{\sqrt{c}(1-c)^2} \approx 4.04 \ .$$

For $k \geq 50$ and $15u < k$ we have

$$\mu_{\max}(k,u) \leq \frac{d}{k}\left(\frac{u}{k}\right)^{\frac{1}{2}} .$$

Proof. — More precise restrictions on k and u are in fact $k \geq 35/(1-c)^2$ and $u < ck$. From corollary 4.3 and lemma 4.4 it follows that

$$\mu_K(4^s u) \geq 2^s \mu_K(u)$$

for every positive integer s for which $\mu_K(4^{s-1}u) \leq \frac{1}{35}$. On the other hand, by lemma 4.1 we have, for any s with $4^s u \leq k$,

$$\mu_K(4^s u) \leq \frac{k}{(k-4^s u)^2} .$$

It follows that

$$\mu_{\max}(k,u) \leq \min\left(\frac{k}{2^s(k-4^s u)^2}\right) ,$$

where the minimum is taken over all the integers s such that

$$(6) \qquad 4^s u \leq k \quad \text{and} \quad \frac{k}{(k-4^{s-1}u)^2} \leq \frac{1}{35} .$$

To get a good upper bound we choose

$$s = [\log_4(ck/u)] + 1,$$

where [ ] is the integral part and $\log_4$ is log in base 4. The conditions of the proposition guarantee that s is in the range (6). Also set $t = \log_4(ck/u) - (s-1)$. Note that $t \in [0,1)$. We have

$$4^s = 4 \cdot 4^{s-1} = 4^{1-t}\left(\frac{ck}{u}\right) .$$

Therefore we obtain the bound

$$\mu_{\max}(k,u) \leq \frac{1}{k}\left(\frac{u}{k}\right)^{\frac{1}{2}} \frac{1}{\sqrt{4^{1-t}c}\,(1-4^{1-t}c)^2} .$$

Consider now $\sqrt{4^{1-t}c}\,(1-4^{1-t}c)^2$ as a function of t. Its maximal value over $[0,1]$ is easily found. It is obtained at $t=1$ and equals $1/d$. This finishes the proof. □

Remark 4.6. — Here is the reasoning behind the choice of s. We are trying to minimize a function of the integer s. The replacement for a differential when computing a "critical value" in this situation is the difference of two successive values, but we may also consider the quotient of two such values, which is more natural here. Therefore, we look for an integer s for which the ratio of the expressions at s and $s-1$ is closest to 1, i.e.,

$$2\left(\frac{k-4\left(4^{s-1}u\right)}{k-\left(4^{s-1}u\right)}\right)^2 \approx 1.$$

Solving this gives

$$4^{s-1} \approx \frac{ck}{u}.$$

Remembering that $s \in \mathbb{Z}$ makes our choice clear. As mentioned above, we have arranged things so that this s will be in the range we are considering.

5. Structure of K with large $E_{K,u}$

In this section we will describe the structure of sets in the class $\mathbb{K}_\psi$, where ψ is roughly the measure attained by an arithmetic progression. More precisely, set $d_1 = \frac{2\sqrt{6}}{\pi}$. We let

$$\psi = \frac{d_1}{k}\left(\frac{u}{k}\right)^{\frac{1}{2}}$$

and consider from now on a set $K \in \mathbb{K}_\psi$. Towards the end of this section we will also describe a subclass satisfying property G_ψ.

Our initial restrictions on k and u in this section are that the restrictions of proposition 4.5 are satisfied for k and 4^3u. It is enough to require $k \geq 50$ and $1000u < k$. These assumptions will be strengthen later. Let d be defined as in proposition 4.5.

Lemma 5.1. — *For a constant $c_1 \approx 0.75$ there exists $i \in \{0,1,2\}$ such that*

$$\mu(E_{K,4^iu} + E_{K,4^iu}) \leq (2+c_1)\mu(E_{K,4^iu}).$$

Proof. — Assume by contradiction that for all $0 \leq i \leq 2$ we have

$$\mu(E_{K,4^iu} + E_{K,4^iu}) > (2+c_1)\mu(E_{K,4^iu}).$$

Applying corollary 4.3 repeatedly we get

$$\mu(E_{K,4^3u}) > (2+c_1)^3\mu(E_{K,u}).$$

substituting the lower bound we imposed on $\mu(E_{K,u})$ and the upper bound of proposition 4.5 on $\mu(E_{K,4^3u})$ we get the inequality

$$2^3d > d_1(2+c_1)^3.$$

We fix c_1 so that this last inequality fails, i.e., $c_1 = 2((d/d_1)^{1/3} - 1)$. The approximation to c_1 is recovered from the estimate $d/d_1 \approx 2.59$. □

For a positive integer q we set

$$E_{q,\delta} = \bigcup_{r=0}^{q-1}[\frac{r}{q} - \frac{\delta}{2}, \frac{r}{q} + \frac{\delta}{2}] \pmod 1.$$

The following lemma is proved in **[5]** on p.154-159.

Lemma 5.2. — *Let $F \subset [0,1)$ be a closed set such that $\mu(F) \leq Const$ for some unspecified constant $Const$. Suppose that there exists $0 < c < 1$ such that*

$$\mu(F+F) \leq (2+c)\mu(F).$$

Then there exist $\beta \in [0,1)$ and a positive integer q such that

$$F \subseteq \beta + E_{q,\delta},$$

where $\delta = \frac{(1+c)}{q}\mu(F)$.

Lemma 5.3. — *Suppose $F \subseteq \beta + E_{q,\delta}$, where $\delta = \frac{(1+c)}{q}\mu(F), 0 < c < 1$, Suppose in addition that $\mu(F) > 0$, that $0 \in F$ and that $-F = F \pmod 1$. Then $F \subseteq E_{q,\delta}$ and*

$$E_{q,2\delta} \subseteq F + F + F \pmod 1.$$

Proof. — To see that $F \subseteq E_{q,\delta}$ note first that $E_{q,\delta}$ is stable under translation by $1/q$. Therefore, we may assume that $|\beta| \leq 1/2q$. We know that $0 \in F \subseteq \beta + E_{q,\delta}$. This implies that $|\beta| \leq \delta/2q$. Finally, as F is stable under negation,

$$F \subseteq (\beta + E_{q,\delta}) \cap (-\beta + E_{q,\delta}) = E_{q,\delta-|\beta|} \subseteq E_{q,\delta}.$$

When $\beta = 0$ the second part of the lemma is proved at the same place the previous lemma was. □

From now, until the end of the paper, excluding section 6, we will be working under the following additional assumption

Assumption 5.4. — Our u and k are such that

$$\frac{4.04}{k}\left(\frac{16u}{k}\right)^{\frac{1}{2}} < Const,$$

where $Const$ is the unknown constant of lemma 5.2.

Making this assumption allows one to use lemma 5.2 for our purposes. It is enough to require that k is big enough, of course. This assumption makes the results of this paper ineffective. It is our hope, however, that one can give an effective bound in lemma 5.2, and thereby for the entire paper.

Proposition 5.5. — *If $K \in \mathbb{K}_\psi$, then there exist integers q and $i \in \{0,1,2\}$ and a positive real number δ such that*

1. *We have the inclusions*

$$E_{K,4^i u} \subseteq E_{q,\delta} \text{ and } E_{K,10\cdot 4^i u} \supseteq E_{q,2\delta}.$$

2. *We have the inequality*

$$\delta \geq q^{-1} 2^i \frac{d_1}{k}\left(\frac{u}{k}\right)^{\frac{1}{2}}.$$

Proof. — Let $i \in \{0,1,2\}$ be the smallest integer for which the assertion of Lemma 5.1 holds. This lemma precisely says that Lemmas 5.2 and 5.3 can be applied in succession to $F = E_{K,4^i u}$. The implication is that there exist some positive integer q and some positive real number δ such that

$$E_{K,4^i u} \subseteq E_{q,\delta} \tag{7}$$

and

$$E_{q,2\delta} \subseteq E_{K,4^i u} + E_{K,4^i u} + E_{K,4^i u}.$$

Corollary 4.3 implies that

$$E_{K,10\cdot 4^i u} \supseteq E_{K,4^i u} + E_{K,4^i u} + E_{K,4^i u}.$$

This gives the first assertion. The inclusion (7) implies

$$q\delta = \mu(E_{q,\delta}) \geq \mu(E_{K,4^i u}) \geq 2^i \mu(E_{K,u}) \geq 2^i \frac{d_1}{k}\left(\frac{u}{k}\right)^{\frac{1}{2}}.$$

This gives the second assertion. □

Proposition 5.6. — *Let q be a positive integer. If*

$$E_{K,u} \supset \{0, q^{-1}, \ldots, \frac{q-1}{q}\},$$

then there exists an integer r such that the set

$$K_r = \{a \in K : \quad a \equiv r \pmod{q}\}$$

satisfies

$$|K_r| \geq k - 2u.$$

Proof. — We have

$$\begin{aligned} q(k-u)^2 &\leq \sum_{r=0}^{q-1} |S_K(r/q)|^2 = \sum_{r=0}^{q-1} \sum_{m,n=1}^{k} e^{2\pi i r(a_n - a_m)/q} \\ &= \sum_{m,n=1}^{k} \sum_{r=0}^{q-1} \left(e^{2\pi i(a_n-a_m)/q}\right)^r = \sum_{a_m \equiv a_n \ (q)} q \\ &= q\sum_{r=0}^{q-1} |K_r|^2 \leq (q \max|K_r|) \sum_{r=0}^{q-1} |K_r| = kq \max|K_r|, \end{aligned}$$

and therefore

$$\max|K_r| \geq k\left(1 - \frac{u}{k}\right)^2 \geq k - 2u.$$

□

Let, for $\theta > 0$,

$$b_\theta = \theta - \int_0^\theta |\cos(\pi\alpha)| d\alpha.$$

Clearly, $b_\theta > 0$. It is also easy to see that b_θ is increasing in θ because its derivative with respect to θ is $1 - |\cos(\pi\theta)| \geq 0$. Finally, one checks that $2b_{1/2} = b_1$.

Proposition 5.7. — *Assume $E_{K,u} \supset [0,\delta]$ and set*

$$\ell_i = a_{k+1-i} - a_i, \quad i = 1, \dots, k.$$

Then, for every $0 < \theta < 1/2$,

$$|\{i : \quad |\ell_i|\delta \geq \theta\}| \leq \frac{u}{b_\theta}.$$

Proof. — We have

$$\begin{aligned}\delta(k-u) &\leq \int_0^\delta s_K(\alpha) d\alpha \leq \frac{1}{2}\sum_{n=1}^k \int_0^\delta \left|e^{2\pi i\alpha a_{k+1-n}} + e^{2\pi i\alpha a_n}\right| d\alpha \\ &= \frac{1}{2}\sum_{n=1}^k \int_0^\delta \left|e^{2\pi i\alpha\ell_n} + 1\right| d\alpha = \sum_{n=1}^k \int_0^\delta |\cos(\pi\alpha\ell_n)| d\alpha.\end{aligned}$$

We bound each term from above. If $|\ell_n|\delta \leq \theta$, we use the trivial estimate

$$\int_0^\delta |\cos(\pi\alpha\ell_n)| d\alpha \leq \delta.$$

Otherwise, we make the change of variables

$$\int_0^\delta |\cos(\pi\alpha\ell_n)| d\alpha = \frac{1}{|\ell_n|}\int_0^{|\ell_n|\delta} |\cos(\pi\alpha)| d\alpha.$$

When $\theta \leq |\ell_n|\delta \leq 1$ we use the estimate

$$\begin{aligned}\frac{1}{|\ell_n|}\int_0^{|\ell_n|\delta} |\cos(\pi\alpha)| d\alpha &= \frac{1}{|\ell_n|}\left(\int_0^\theta |\cos(\pi\alpha)| d\alpha + \int_\theta^{|\ell_n|\delta} |\cos(\pi\alpha)| d\alpha\right) \\ &\leq \frac{1}{|\ell_n|}\left(\theta - b_\theta + |\ell_n|\delta - \theta\right) = \delta - b_\theta\frac{1}{|\ell_n|} \\ &\leq \delta(1 - b_\theta).\end{aligned}$$

When $|\ell_n|\delta > 1$ one finds similarly

$$\begin{aligned}\frac{1}{|\ell_n|}\int_0^{|\ell_n|\delta} |\cos(\pi\alpha)| d\alpha &\leq \delta - b_1\frac{[|\ell_n|\delta]}{|\ell_n|} \\ &\leq \delta - \frac{1}{2}b_1\delta = \delta(1 - b_{\frac{1}{2}}) \leq \delta(1 - b_\theta).\end{aligned}$$

Therefore,

$$\delta(k-u) \leq \delta\left(k - b_\theta \cdot |\{i : \quad |\ell_i|\delta \geq \theta\}|\right),$$

which proves what we wanted. □

Lemma 5.8. — *Suppose $k > 30000u$ and $K \in \mathbb{K}_\psi$. Then there exists a unit vector v such that for any $\alpha \in E_{K,u}$ we have* $\mathrm{Angle}(S_K(\alpha), v) < \pi/2$. *In addition there is a subset K_0 of K with at least $k - 2000u$ elements such that the following is satisfied: For any $a \in K_0$ and any $\alpha \in E_{K,u}$ one has* $\mathrm{Angle}(v, e^{2\pi i \alpha a}) < \pi/4$.

Proof. — By proposition 5.5 there exist $i \in \{0, 1, 2\}$, a positive integer q and a real number δ such that

$$E_{K,u} \subseteq E_{K,4^i u} \subseteq E_{q,\delta}, \quad E_{K,10\cdot 4^i u} \supseteq E_{q,2\delta}.$$

Consider a parameter $\theta < 1/4$ to be set later. According to proposition 5.7 all but $10 \cdot 4^i u/b_{2\theta}$ elements of K are in an interval of length ℓ such that $\ell\delta < 2\theta$, or equivalently $2\pi\ell(\delta/2) < 2\pi\theta$. Further, all but $2 \cdot 4^i u$ elements of these are in the same congruence class modulo q. We may translate K by an integer to make this interval symmetric around 0 and the residue class be that of 0. Now denote by K_0 the intersection of K with the interval and the residue class of 0, and let $\bar{K} = K - K_0$. We will show the lemma with $v = 1$. One easily sees that the condition of the lemma is now equivalent to $\mathrm{Re}\, S_K(\alpha)/|S_K(\alpha)| > \sqrt{2}/2$. We will in fact show this for all α in the bigger set $E_{q,\delta}$. Consider such an α and $a \in K_0$. Suppose first that α is in the interval of $E_{q,\delta}$ around 0. Since $|a| \leq \ell/2$ we have

$$|\arg(e^{2\pi i a\alpha})| = |2\pi a\alpha| \leq 2\pi(\ell/2)(\delta/2) < \pi\theta.$$

Therefore, $\mathrm{Re}\, e^{2\pi i a\alpha} > \cos(\theta\pi)$. Now, since elements of K_0 are divisible by q, it is easily seen that they behave the same on all intervals of $E_{q,\delta}$. Thus, the same estimate is true for any $\alpha \in E_{q,\delta}$. Since K_0 contains at least $k - (10b_{2\theta}^{-1} + 2)4^i u$ elements, this implies that for $\alpha \in E_{q,\delta}$ we have

$$\mathrm{Re}\, S_{K_0} > \cos(\theta\pi)\left(k - (10b_{2\theta}^{-1} + 2)4^i u\right).$$

Therefore

$$\begin{aligned}\frac{\mathrm{Re}\, S_K(\alpha)}{|S_K(\alpha)|} &\geq \frac{\mathrm{Re}\, S_{K_0}(\alpha) - |S_{\bar{K}}(\alpha)|}{k} \\ &> \frac{\cos(\theta\pi)(k - (10b_{2\theta}^{-1} + 2)4^i u) - (10b_{2\theta}^{-1} + 2)4^i u}{k}.\end{aligned}$$

Thus, the lemma will be true if we can find a $\theta < 1/4$ for which the right hand side is larger than $\sqrt{2}/2$. Clearly the worst possible case is when $i = 2$, in which we need to solve

$$\frac{\cos(\theta\pi)(k - (160b_{2\theta}^{-1} + 32)u) - (160b_{2\theta}^{-1} + 32)u}{k} > \frac{\sqrt{2}}{2}.$$

This inequality is equivalent to

$$\frac{k}{u} > (160b_{2\theta}^{-1} + 32)(1 + \cos(\pi\theta))\left(\cos(\pi\theta) - \frac{\sqrt{2}}{2}\right)^{-1}.$$

It remains to numerically find the minimum of the expression on the right over $\theta \in [0, 1/4]$. This is found to be about 29439, located around $\theta = 0.19272$. The result follows with the bound on the number of elements outside K_0 being $160b_{2\theta}^{-1}+32 \approx 1860$ times u. □

Proposition 5.9. — *Suppose $k > 30000u$. Then, the following subclass of $\mathbb{K}_\psi$ has property G_ψ. A set in the class can be written as a disjoint union*

$$M = M_0 \cup \bar{K} \textit{ with } |\bar{K}| < 2000u,$$

where M_0 is an arithmetic progression with at most one gap.

Proof. — Suppose $K \in \mathbb{K}_\psi$. We will show how to find a set M in the subclass described above such that $\mu_M(u) \geq \mu_K(u)$. By the previous lemma we know that we have a decomposition $K = K_0 \cup \bar{K}$ and that there is a unit vector v such that for all $\alpha \in E_{K,u}$ both the sum $S_K(\alpha)$ and any individual term $e^{2\pi i\alpha a}$, with $a \in K_0$, form an angle of $< \pi/4$ with v. In the situation just described it is easily seen that by replacing $a > b \in K_0$ by $c = a - qt, d = b + qt$, such that $c > d$ and $c, d \notin K_0$, we enlarge the value of $S_K(\alpha)$ for $\alpha \in E_{q,\delta}$. This is because the contribution of the pair (c,d) is larger than that of (a, b) and has the same direction which forms an acute angle with the rest of the sum. Therefore, such a change can only increase the value of $\mu_{K,u}$. All that remain to do then is to show that by repeated application of this we transform K_0 into a set M_0 which is an arithmetic progression of difference q with possibly one gap. To see this we may again assume that elements of K_0 are divisible by q. Suppose that $e = \max(K_0)$, $f = \min(K_0)$ and consider the set

$$K_{comp} = \{x \in q\mathbb{Z} : \quad f \leq x \leq e \quad \text{and} \quad x \notin K_0\}.$$

Let $c = \max(K_{comp})$ and $d = \min(K_{comp})$. Suppose that $K_{comp} \neq \phi$ and that $c \neq d$. Since $a = c + q \in K_0$ and $b = d - q \in K_0$, we may perform a transformation as above. It is clear that each step decreases the sum of the absolute values of all the differences between the elements of K_0. Therefore the process has to stop. The computation above shows that it stops only when K_{comp} has at most one element. This means that the resulting set, has at most one gap. □

6. A close to maximal set

In this section we describe a set K_{ex} which we suspect to be very close to maximal. Just how close will become evident later on. We will begin with parameters m_0 and w and construct a set $M(m_0, w)$. This will roughly be our set K_{ex} except that we can not guarantee in general that it will have exactly k elements. We will choose the parameters so that it has about k elements and then take out as many elements as we need to get it to be of the right size.

We assume we are given an odd positive integer m_0 and a real number w which satisfy the assumption $m_0 > 30000w$. Let M_0 be the set $\{-(m_0-1)/2, \ldots, (m_0-1)/2\}$ of size m_0. Clearly $|S_{M_0}| = S_{M_0}$. We write β for $\beta_{m_0,w}$. According to definition 3.2, β satisfies $S_{M_0}(\beta) = m_0 - w$ and furthermore $E_{M_0}(w) = [-\beta, \beta]$.

Construction 6.1. — Given m_0 and w we construct the set $M(m_0, w)$ as follows:

$$M = M(m_0, w) = M_0 \cup \overline{M} = \bigcup_{n\in\mathbb{Z}} M_n, \tag{8}$$

where each M_n is an arithmetic progression of difference 1 centered (as best possible) around $\frac{n}{\beta}$. For $n > 0$, the length of the two progressions $M_{\pm n}$ is the same and is denoted by $\frac{m_n}{2}$. To fully determine $M(m_0, w)$ one only needs to give the number m_n. It will be defined to be the largest even integer smaller than a constant c_n, whose description is given in definition 6.3 below, and which has the approximation, given in proposition 6.8,

$$c_n \approx \frac{2w}{(\pi n)^2}\left(1 - \frac{(-1)^n}{2}\right).$$

We note that $c_n < 2$ for $n >> 0$, hence $m_n = 0$ for all but finitely many values of n. Also m_n is always non-negative (see remark 6.9).

To define the constants c_n we need an auxiliary function f.

Definition 6.2. — We define a function $f = f_{m_0,w}$ on $[0, 1]$ as

$$f_{m_0,w}(r) = \begin{cases} s_{m_0}(\beta r) - m_0 + w & \text{if } \ r \in [\frac{1}{2}, 1] \\ f_{m_0,w}(1-r) & \text{if } r \in [0, \frac{1}{2}]. \end{cases}$$

Note that $f(r)$ is a continuous function such that

$$f(0) = f(1) = s_{m_0}(\beta) - (m_0 - w) = 0$$

by the definition of $\beta = \beta_{m_0,w}$. Also, by construction, f is symmetric around 1/2, which implies that in its real Fourier expansion all the sin functions do not appear. Finally, for all $r \in [0, 1]$, $f(r) \geq 0$. It is enough to check this by symmetry for $r \geq 1/2$, in which case $\beta r \in E_{M_0}(w)$ hence $s_{m_0}(\beta r) \geq m_0 - w$.

Definition 6.3. — Define real numbers $c_n = c_n(m_0, w)$ for $n \geq 0$ in such a way that the real Fourier expansion of f is

$$f(r) = c_0 - \sum_{n=1}^{\infty} c_n \cos(2\pi n r). \tag{9}$$

Define m_n as the largest even number smaller than c_n.

We have the usual integral expansions of c_n,

$$c_0 = \int_0^1 f(r)\ dr = 2\int_{\frac{1}{2}}^1 f(r)\ dr \tag{10}$$

and

$$c_n = -4\int_{\frac{1}{2}}^1 f(r)\cos(2\pi n r)\ dr. \tag{11}$$

The Fourier expansion (9) clearly converges pointwise on $[0,1]$ because f is continuous and piecewise differentiable. Substituting $r=0$ we have

$$c_0 = \sum_{n=1}^{\infty} c_n. \tag{12}$$

Remark 6.4. — The heuristic reasoning behind construction 6.1 is as follows: We are looking for a set M of the form (8). The number β is defined so that $E_{M,w} \subseteq [-\beta,\beta]$ whatever $\overline{M}$ is, so best we can hope for is near equality. We would also like to make $\bar{m} = \sum_{n>0} m_n$ (and therefore $m = |M|$) as large as possible. Since S_{M_0} is real and large, it is easily seen that the best way to enlarge S_M is to contribute to its real part. Thus we assume from the start that $\overline{M}$ is symmetric around 0. Then S_M is real valued. Therefore, the condition for α to be in $E_{M,w}$ becomes $S_M(\alpha) \geq k - w$, which is equivalent to

$$\bar{m} - S_{\overline{M}}(\alpha) \leq S_{M_0}(\alpha) - m_0 + w \tag{13}$$

Suppose it was possible to have $\beta \in E_{M,w}$. Then we get $\bar{m} - S_{\overline{M}}(\beta) = 0$ and thus $e^{2\pi a\beta i} = 1$ for $a \in \overline{M}$. This implies that each $a \in \overline{M}$ is of the form $a = \frac{n}{\beta}$ for some $n \neq 0$. Set

$$e_n = |\{a = \pm\frac{n}{\beta} \in \overline{M}\}|.$$

We can therefore write

$$f_{\overline{M}}(r) := \bar{m} - S_{\overline{M}}(\beta r) = \bar{m} - \sum_{n=1}^{\infty} e_n \cos(2\pi n r).$$

The function $f_{\overline{M}}$ satisfies $f_{\overline{M}} \leq f$ since this is true on $[1/2,1]$ by (13) and since both sides are symmetric for replacing r by $1-r$. Conversely, for any symmetric $f_{\overline{M}} \leq f$ we can, replacing f by $f_{\overline{M}}$ in definition 6.3 and what follows, find constants e_n and create an $\overline{M}$ that will satisfy (13). But since we want the largest $\bar{m} = e_0$ it is clear we should take $f_{\overline{M}} = f$. Then we make the necessary adjustments to get from the c_n to a true candidate for $\overline{M}$ by taking m_n to be the largest even number smaller than c_n and $\overline{M}$ as a union of arithmetic progressions centered on $\pm n/\beta$ and of length $m_n/2$ each, which is just the construction 6.1.

We now derive estimates on the parameters of M and the size of $E_{M,w}$.

Lemma 6.5. — *Let n be an integer. Then,*

$$\sup_{\substack{z\in\mathbb{C}\\ |z|\leq\frac{1}{n}}} \frac{\sin(nz)}{\sin(z)} \leq e^{\frac{1}{n}} \cdot 1.2n.$$

Proof. — We have

$$\frac{\sin(nz)}{\sin(z)} = \frac{e^{inz}-e^{-inz}}{e^{iz}-e^{-iz}} = e^{i(1-n)z}\frac{e^{2inz}-1}{e^{2iz}-1} = e^{i(1-n)z}\sum_{k=0}^{n-1} e^{2ikz}.$$

If $z = x - iy$, then, since $|e^{iz}| = e^y$, we get the upper bound

$$\left|\frac{\sin(nz)}{\sin(z)}\right| \leq e^{-(n-1)y}\sum_{k=0}^{n-1} e^{2ky} = \sum_{\substack{1-n\leq l\leq n-1\\ 2|n-1-l}} e^{ly}.$$

It is enough to find the maximal value of the last expression for $-1/n \leq y \leq 1/n$. It is clear that the expression is symmetric in y. The derivative is given by

$$\sum_{\substack{1\leq l\leq n-1\\ 2|n-1-l}} l(e^{ly} - e^{-ly}),$$

which is clearly positive for positive y. Therefore, the maximal value is obtained at $y = 1/n$ and equals

$$e^{-\frac{n-1}{n}}\sum_{k=0}^{n-1} e^{\frac{2k}{n}} = e^{-\frac{n-1}{n}}\frac{e^2-1}{e^{\frac{2}{n}}-1}.$$

Using the inequality $e^x - 1 \geq x$ we obtain

$$\left|\frac{\sin(nz)}{\sin(z)}\right| \leq e^{-\frac{n-1}{n}}(e^2-1)\frac{n}{2} = e^{\frac{1}{n}}\frac{e^2-1}{2e}n \leq e^{\frac{1}{n}}\cdot 1.2n.$$

□

Let

$$f(r) = w + \sum_{j=1}^{\infty} a_j r^{2j}$$

be the expansion of $f(r)$ on the interval $[1/2, 1]$. In other words, it is the Taylor expansion of $s_{m_0}(\beta r) - m_0 + w$ around 0. Note that the odd coefficients vanish because s_{m_0} is an even function, and that since its value at 0 is m_0 the constant coefficient is indeed w. Let

$$R_2(r) := f(r) - w - a_1 r^2$$

be the error term in the quadratic approximating of $f(r)$.

Lemma 6.6. — *Let $c = 6(w/m_0)(1 + 2w/m_0)$ and define*

$$h(r) := 2m_0\frac{(cr^2)^2}{1-cr^2}.$$

Then, for $r \in [1/2, 1]$ and any $n \geq 1$,

$$\left|\frac{d^n}{dr^n}R_2(r)\right| \leq \frac{d^n}{dr^n}h(r).$$

Proof. — We give an upper bound on the coefficients a_j. From the explicit description of s_{m_0} given in part 1 of proposition 3.1, we see the the complex function $\sin(m_0 z)/\sin(z)$ extends the real function $s_{m_0}(\alpha/\pi)$. Consider the Taylor expansion

$$\sin(m_0 z)/\sin(z) = \sum_{j=0}^{\infty} b_j z^j.$$

By lemma 6.5 we see that since $m_0 \geq 3$, $s_{m_0}(z/\pi) = \sin(m_0 z)/\sin(z)$ is bounded by $2m_0$ when $|z| \leq 1/m_0$. We use the Cauchy integral formula on a circle C_{m_0} of radius $1/m_0$ around 0 to obtain the estimate

$$|b_j| = \left| \frac{1}{2\pi i} \oint_{C_{m_0}} \frac{\sin(m_0 z)/\sin(z)}{z^{j+1}} dz \right| < 2{m_0}^{j+1}.$$

From the definition of f we see that for $j > 0$ we have $a_j = b_{2j}(\pi\beta)^{2j}$. From the bound on β in proposition 3.4 we get for $j > 0$

$$|a_j| < 2{m_0}^{2j+1} \left(\left(\frac{6w(1+2u/m_0)}{m_0} \right)^{\frac{1}{2}} \frac{1}{m_0} \right)^{2j} = 2m_0 c^j.$$

One easily checks that

$$h(r) = 2m_0 \sum_{j=2}^{\infty} c^j r^{2j}.$$

The bound is now clear. □

Corollary 6.7. — *We have the following estimates for $r \in [1/2, 1]$ and $\delta \in [0, 1/2]$.*

(6.7.1) $$R_2(r) \leq 75 \frac{w^2}{m_0} r^4.$$

(6.7.2) $$R_2'(r) \leq 300 \frac{w^2}{m_0} r.$$

(6.7.3) $$|w + a_1| \leq 75 \frac{w^2}{m_0}.$$

(6.7.4) $$|f'(r) + 2wr| \leq 450 \frac{w^2}{m_0} r.$$

(6.7.5) $$f(1-\delta) \leq 2\delta w \left(1 + 225 \frac{w}{m_0}\right).$$

Proof. — To prove (6.7.1) we note that

$$R_2(r) \leq h(r) \leq 2m_0 \frac{c^2}{1-c} r^4 = 2 \cdot 6^2 \frac{w^2}{m_0} r^4 (1 + o(\epsilon)) \leq 75 \frac{w^2}{m_0} r^4.$$

Similarly, we get (6.7.2) because

$$\begin{aligned} h'(r) &= 4m_0c^2r^3\frac{2-r^2c}{(1-cr^2)^2} \leq 8m_0\frac{c^2}{(1-c)^2}r \\ &\leq 8\cdot 36\frac{w^2}{m_0}r(1+o(\epsilon)) \leq 300\frac{w^2}{m_0}r. \end{aligned} \tag{14}$$

Since $f(1)=0$ we find

$$|w+a_1| = |f(1)-w-a_1 1^2| = |R_2(1)| \leq h(1),$$

which is $\leq 75w^2/m_0$ by (6.7.1). This gives (6.7.3). We get (6.7.4) by

$$|f'(r)+2wr| = |R_2'(r)+2(w+a_1)r| \leq h'(r)+2\cdot 75\frac{w^2}{m_0}r \leq 450\frac{w^2}{m_0}r,$$

where the last two inequalities follow from (6.7.3) and (6.7.2) respectively. Finally, (6.7.5) is derived from (6.7.4) by the mean value theorem: Since $f(1)=0$ we can find $1-\delta<\rho<1$ such that $|f(1-\delta)| \leq \delta|f'(\rho)|$. By (6.7.4) we find

$$|f'(\rho)| \leq 2u\rho + 450\frac{w^2}{m_0}\rho \leq 2u + 450\frac{w^2}{m_0} = 2u\left(1+225\frac{w}{m_0}\right),$$

which finishes the proof. □

Proposition 6.8. — *We have the following estimates on the coefficients c_n.*

$$c_0 = \frac{5}{12}w + 75\frac{w^2}{m_0}O(1). \tag{6.8.1}$$

$$c_n = \frac{2w}{(\pi n)^2}\left(1-\frac{(-1)^n}{2}\right) + \frac{1500}{(\pi n)^2}\frac{w^2}{m_0}O(1). \tag{6.8.2}$$

Remark 6.9. — In particular, c_n is positive for every n and hence we can indeed define m_n as we did in construction 6.1.

Lemma 6.10. — *For any C^∞ real valued function g and any nonzero integer n we have*

$$\int_{\frac{1}{2}}^{1}\cos(2\pi nr)g(r)dr = \frac{1}{(2\pi n)^2}\left(g'(1)-(-1)^n g'\left(\frac{1}{2}\right) - \int_{\frac{1}{2}}^{1}\cos(2\pi nr)g''(r)dr\right).$$

Proof. — We use integration by parts twice to get

$$\begin{aligned} &\mathrm{Re}\int_{\frac{1}{2}}^{1} e^{2\pi inr}g(r)dr \\ &= \mathrm{Re}\left(\frac{1}{2\pi ni}\left(g(1)-(-1)^n g\left(\frac{1}{2}\right) - \int_{\frac{1}{2}}^{1} e^{2\pi inr}g'(r)dr\right)\right) \\ &= \mathrm{Re}\left(\frac{-1}{(2\pi ni)^2}\left(g'(1)-(-1)^n g'\left(\frac{1}{2}\right) - \int_{\frac{1}{2}}^{1} e^{2\pi inr}g''(r)dr\right)\right). \end{aligned}$$

Writing out the real part in terms of cos functions gives the result. □

Proof of proposition 6.8. — Taking $g(r)=1$ and $g(r)=r^2$ respectively in lemma 6.10 we get

$$\int_{\frac{1}{2}}^{1}\cos(2\pi nr)dr=0, \tag{15}$$

$$\int_{\frac{1}{2}}^{1}\cos(2\pi nr)r^2dr=\frac{2}{(2\pi n)^2}\left(1-\frac{(-1)^n}{2}\right)\le\frac{3}{(2\pi n)^2}. \tag{16}$$

Now we use $g(r)=R_2(r)$ as defined in lemma 6.6. In the inequality of lemma 6.10 we can replace R_2 and its derivatives by the function h and its derivative, as follows from lemma 6.6. This gives

$$\left|\int_{\frac{1}{2}}^{1}\cos(2\pi nr)R_2(r)dr\right|\le\frac{1}{(2\pi n)^2}\left(h'(1)+h'(\frac{1}{2})+\int_{\frac{1}{2}}^{1}h''(r)dr\right).$$

Evaluating the right hand side we find

$$\left|\int_{\frac{1}{2}}^{1}\cos(2\pi nr)R_2(r)dr\right|\le\frac{2h'(1)}{(2\pi n)^2}\le 150\frac{w^2}{m_0}\frac{1}{(\pi n)^2}, \tag{17}$$

where the inequality follows from the estimate (14). By definition,

$$f(r)=R_2(r)+w+a_1r^2=R_2(r)+w(1-r^2)+(w+a_1)r^2.$$

Here, the main term is $w(1-r^2)$. Therefore, The main terms in the estimates on c_n and c_0 are

$$\begin{aligned}c_n &\approx -4w\int_{\frac{1}{2}}^{1}(1-r^2)\cos(2\pi nr)\,dr=\frac{2w}{(\pi n)^2}\left(1-\frac{(-1)^n}{2}\right),\\ c_0 &\approx\ 2w\int_{\frac{1}{2}}^{1}(1-r^2)dr =\frac{5}{12}w.\end{aligned}$$

The error term for c_n is now obtained by integrating $R_2(r)+(w+a_1)r^2$ multiplied by $\cos(2\pi nr)$ and the appropriate constant, and the integral is estimated using (15),(16) and (17). For $n>0$ we get

$$\begin{aligned}\left|c_n-\frac{2w}{(\pi n)^2}\left(1-\frac{(-1)^n}{2}\right)\right| &=4\left|\int_{\frac{1}{2}}^{1}\cos(2\pi nr)\left(R_2(r)+(w+a_1)r^2\right)dr\right|\\ &\le 4\left(150\frac{w^2}{m_0}\frac{1}{(\pi n)^2}+75\frac{w^2}{m_0}\frac{3}{(\pi n)^2}\right)\\ &=1500\frac{w^2}{m_0}\frac{1}{(\pi n)^2}.\end{aligned}$$

Similarly for c_0 we find

$$\begin{aligned}\left|c_0 - \frac{5}{12}w\right| &= 2\left|\int_{\frac{1}{2}}^{1}(f(r) - w(1-r^2))dr\right| \\ &\leq 2\left|\int_{\frac{1}{2}}^{1} R_2(r)dr\right| + 2\left|\int_{\frac{1}{2}}^{1}(a_1+w)r^2dr\right| \\ &\leq 2\left|\int_{\frac{1}{2}}^{1} 75\frac{w^2}{m_0}r^4dr\right| + 2\cdot\frac{1}{3}\cdot\frac{7}{8}(a_1+w) \\ &\leq \frac{2}{5}\cdot 75\frac{w^2}{m_0} + \frac{2}{3}\cdot\frac{7}{8}\cdot 75\frac{w^2}{m_0} \leq 75\frac{w^2}{m_0}.\end{aligned}$$

□

We now wish to modify the set M to make it our candidate set K_{ex}. We need a few more computations.

Lemma 6.11. — *Let ω be a real constant. Then we have*

$$\sum_{n=1}^{\infty}\min\left(\frac{\omega}{n^2}, 2\right) \leq 4.2\sqrt{\omega}.$$

Proof. — Recall that $\sum 1/n^2 = \pi^2/6$. We begin by considering small values of ω. If $\omega \leq 2$, then the sum is clearly $\omega\pi^2/6 \leq \sqrt{\omega}\sqrt{2}\pi^2/6$. Similar computations show that when $2 < \omega \leq 4$ we get a bound of $\sqrt{\omega}\pi^2/3$ and when $4 < \omega \leq 9$ we get a bound of

$$3\left(\frac{\pi^2}{6} - \frac{1}{4}\right)\sqrt{\omega} \leq 4.2\sqrt{\omega},$$

which is the largest so far. Now suppose that $9 \leq \omega$. Let $x = \sqrt{\omega/2}$. One checks that $2x^2/(x-1) \leq 3x+2$, It is easily seen that the largest n for which $\omega/n^2 \geq 2$ is $[x]$. Thus we have

$$\begin{aligned}\sum_{n=1}^{\infty}\min\left(\frac{\omega}{n^2}, 2\right) &\leq 2[x] + \omega\sum_{n=[x]+1}^{\infty}\frac{1}{n^2} \leq 2[x] + \omega\int_{[x]}^{\infty}\frac{dt}{t^2} = 2[x] + \frac{\omega}{[x]} \\ &\leq 2(x-1) + \frac{\omega}{x-1} = 2x + \frac{2x^2}{x-1} - 2 \\ &\leq 5x = \frac{5}{\sqrt{2}}\sqrt{\omega} \leq 4.2\sqrt{\omega}.\end{aligned}$$

□

Proposition 6.12. — *Suppose k and u satisfy the condition $1 < u \leq k/30000$. Then there exist $m_0, \bar{m} \in \mathbb{Z}$, $w \in \mathbb{R}$ and a set $K_{ex} = K_{ex}(k, u)$, such that the following are* *satisfied.*

$$k = m_0 + \tilde{m}. \tag{6.12.1}$$

$$M_0 \subseteq K_{ex} \subseteq M(m_0, w). \tag{6.12.2}$$

$$E_{K_{ex},u} \supseteq [-\beta, \beta], \quad \textit{with } \beta = \beta_{m_0,w}. \tag{6.12.3}$$

$$c_0(m_0, w) \geq \tilde{m} \geq c_0(m_0, w) - 5.2\sqrt{w}. \tag{6.12.4}$$

$$u = w + \left(\frac{w}{k}\right)^3 w + 2\left(\frac{w}{k}\right)^{3/2}. \tag{6.12.5}$$

Remark 6.13. — Note that $c_0(m_0, w)$ depends on w and m_0 but to first order is just $\frac{5}{12}w \approx \frac{5}{12}u$.

Proof of proposition 6.12. — We start by choosing w such that (6.12.5) is satisfied. For a given m_0 set $\tilde{m} = \sum_{n=1}^{\infty} m_n$ and $m = m_0 + \tilde{m} = |M(m_0, w)|$. We take the smallest m_0 for which $m \geq k$. We now pull elements out of the sets M_n until we obtain our set K_{ex} (the choice of which elements to take is arbitrary, hence K_{ex} is not uniquely defined). By (12) and the construction of the m_n in 6.1 we find the left inequality in (6.12.4). To get the inequality on the right we need to compute the sum of the differences between c_n and m_n and count how many elements we take out of the M_n in the final step. Recall that m_n was taken to be the largest even integer smaller than c_n. Therefore,

$$c_0 - \tilde{m} = \sum_{n=1}^{\infty} c_n - m_n \leq \sum_{n=1}^{\infty} \min(c_n, 2).$$

Let $\omega = (3w/\pi^2)(1 + o(\epsilon))$. It follows from (6.8.2) that

$$c_0 - \tilde{m} \leq \sum_{n=1}^{\infty} \min\left(\frac{\omega}{n^2}, 2\right),$$

which by lemma 6.11 is

$$\leq 4.2\sqrt{\frac{3}{\pi^2}}(1 + o(\epsilon))\sqrt{u} \leq 2.4\sqrt{u}.$$

The number of elements we may have to take out of $M(m_0, w)$ to get K_{ex} is bounded from above by $|M(m_0, w)| - |M(m_0 - 2, w)|$. The difference in size between the two sets is caused by the fact that we have $m_n(m_0, w) - m_n(m_0 - 2, w) = 2$ every time there is a multiple of 2 between $c_n(m_0, w)$ and $c_n(m_0 - 2, w)$. With the constant ω as before, this does not occur once $n > \sqrt{\omega/2}$. Thus,

$$m - k \leq 2 + 2\sqrt{\omega/2} \leq 2 + 0.78\sqrt{u} \leq 2.78\sqrt{u}.$$

This gives (6.12.4) as $2.4 + 2.78 \leq 5.2$. It remains to show (6.12.3). From the definition of f we see that for $r \geq 1/2$,

$$S_{M_0}(\beta r) - m_0 + w = f(r) = c_0 - \sum_{n=1}^{\infty} c_n \cos(2\pi n r) = \sum_{n=1}^{\infty} c_n(1 - \cos(2\pi n r)),$$

where the last equality follows from (12). Since $m_n \leq c_n$ we find

$$S_{M_0}(\beta r) - m_0 + w \geq \sum_{n=1}^{\infty} m_n(1 - \cos(2\pi nr)) = \tilde{m} - \sum_{n=1}^{\infty} m_n \cos(2\pi nr)$$

and, rearranging terms

$$S_{M_0}(\beta r) + \sum_{n=1}^{\infty} m_n \cos(2\pi nr) \geq m_0 + \tilde{m} - w = m - w. \tag{18}$$

Comparing the situation for $r < \frac{1}{2}$, we see that S_{M_0} is larger while the rest is symmetric around $\frac{1}{2}$ and therefore the last inequality holds for $r \in [0,1]$. Each term $m_n \cos(2\pi nr)$ is very close to $S_{M_n}(r\beta) + S_{M_{-n}}(r\beta)$. We measure the difference in the following lemma.

Lemma 6.14. — *Let $A = \{s, s+1, \ldots, t-1, t\}$ be an arithmetic progression of length $l = t - s + 1$. Suppose $\alpha \in [0, 1/l]$ and suppose $x \in \mathbb{R}$ satisfies $|x - x'| \leq 1/2$, with $x' = (s+t)/2$. Then,*

$$|\operatorname{Re}(le^{2\pi ix\alpha} - S_A(\alpha))| \leq \pi l\alpha + \frac{\pi^2}{6} l^3 \alpha^2.$$

Proof. — Suppose first that $x = x'$. Then,

$$|\operatorname{Re}(le^{2\pi ix\alpha} - S_A(\alpha))| = |l - e^{-2\pi ix\alpha} S_A(\alpha)|.$$

It is easy to see that the expression inside the absolute value is real, positive and equal to

$$l - s_l(\alpha) = l - \frac{\sin(\pi l\alpha)}{\sin(\pi\alpha)} \leq \frac{\sin(\pi l\alpha)}{\pi\alpha} \leq \frac{\pi^2}{6} l^3 \alpha^2,$$

where the last inequality follows from (5). To complete the proof all we have to do is to compute

$$|\operatorname{Re}(le^{2\pi ix\alpha} - le^{2\pi ix'\alpha})| = |l(\cos(2\pi x\alpha) - \cos(2\pi x'\alpha))| \leq l|2\pi(x - x')\alpha| \leq \pi l\alpha.$$

□

Applying the last lemma in our situation we see, using the fact that $\sum {m_n}^3 \leq \tilde{m}^3$, that

$$\sum_{n=1}^{\infty} m_n \cos(2\pi nr) - 2\operatorname{Re} S_{M_n}(\beta r) \leq \pi \tilde{m}\beta + \frac{\pi^2}{6} \tilde{m}^3 \beta^2$$

$$\leq \left(\pi \frac{w}{2} \frac{\sqrt{6}}{\pi} \left(\frac{w}{m_0}\right)^{\frac{1}{2}} \frac{1}{m_0} + \frac{\pi^2}{6} \frac{w^3}{8} \frac{6}{\pi^2} \left(\frac{w}{m_0}\right) \frac{1}{{m_0}^2}\right) (1 + o(\epsilon))$$

by (6.8.1) and proposition 3.4

$$\leq \left(\frac{w}{k}\right)^3 w + 2\left(\frac{w}{k}\right)^{3/2}.$$

Thus, our choice of w in (6.12.5) guarantees by (18) that

$$S_M(\beta r) \geq m - w - \left(\frac{w}{k}\right)^3 w - 2\left(\frac{w}{k}\right)^{3/2} = m - u$$

for all $r \in [0,1]$. Therefore, $E_{M,u} \supseteq [-\beta,\beta]$. To finish we just need to note that taking elements out of M does not change the situation as is easily seen. □

7. Structure of the maximal set

In the final three sections we try to determine the structure of a set K in the class $\mathbb{K}_{\mu_{K_{ex}}(u)}$, i.e., a set which is "better" than the example we produced in the last section.

Our assumptions are as usual: $1 < u \leq k/30000$ and u and k satisfy assumption 5.4. Since $\mu_{K_{ex}}(u)$ is greater than the one for arithmetic progressions, we certainly know that our results from section 5 Apply here. Therefore, we can write K in the form $K = K_0 \cup \overline{K}$, where K_0 is the set whose existence is guaranteed by lemma 5.8, $|K_0| = k_0$, $|\overline{K}| = \bar{k}$, and $k_0 + \bar{k} = k$. What we will try to do is determine the structure of $\overline{K}$. Since, as we saw in proposition 5.9, There is, for any set in $\mathbb{K}_{\mu_{K_{ex}}(u)}$, a better one with the corresponding K_0 forming an arithmetic progression with at most a single gap, it is no harm to assume that our sets are already of this type. We will assume in fact that K_0 is an arithmetic progression (without a gap), that it has an odd number of elements and that its difference is 1. The modifications required to cover the general case will be explained in the end. Since by (2) we are always allowed to translate our set, we can assume

$$K_0 = \left\{-\frac{k_0-1}{2}, \ldots, \frac{k_0-1}{2}\right\}.$$

We also make the following *shortcut*:

$$\bar{k} < \frac{1}{2}u. \tag{19}$$

The justification for this is as follows: we have already seen this estimate with $\frac{1}{2}$ replaced by 2000 in proposition 5.9. Shortly (in (7.4.1)) we will see that $\bar{k}$ is to first order $\frac{5}{12}u$, where the second order terms depend on the above mentioned constant. Iterating this we can assume in advance that the constant is say $\frac{1}{2}$.

Let $\beta' = \beta_{k_0,u}$. Set

$$K_n = K \cap [(n - \frac{1}{8})/\beta', (n + \frac{1}{8})/\beta'], \quad k_n = |K_n| + |K_{-n}|. \tag{20}$$

We also write $g = f_{k_0,u}$. By definition 6.2,

$$k_0 - u = S_{K_0}(\beta' r) - g(r). \tag{21}$$

This function has a Fourier expansion similar to (9) and the coefficients $c_n(k_0,u)$ will be denoted d_n. Obviously

$$E_{K,u} \subseteq [-\beta', \beta']. \tag{22}$$

Definition 7.1. — A constant $\delta \geq 0$ (depending on K and u) is defined by the equation

$$\mu_K(u) = 2(1-\delta)\beta'.$$

In this section, δ appears in the error terms for the estimates of the k_n. To get absolute bounds we will bound δ in section 9.

The basis for the estimates is a bound for $\operatorname{Re} S_{\overline{K}}$.

Proposition 7.2. — *If $\alpha = \beta' r \in E_{K,u}$, then*

$$\bar{k} - \operatorname{Re} S_{\overline{K}}(\alpha) \leq g(r)\left(1 + \frac{2\bar{k}}{k}\right) \leq g(r)\left(1 + \frac{u}{k}\right).$$

Corollary 7.3. — *For all $r \in [0,1]$,*

$$\bar{k} - \operatorname{Re} S_{\overline{K}}(\beta' r) \leq \begin{cases} g(r) + \frac{u^2}{k} & \textit{if } \beta' r \in E_{K,u} \\ u & \textit{otherwise.} \end{cases} \tag{23}$$

Proof. — When $\beta' r \in E_{K,u}$ this is clear from the proposition since $g(r) \leq u$. Otherwise we just use the trivial upper bound $2\bar{k}$, which is $< u$ by (19). □

Proof of proposition 7.2. — Assume $\alpha \in E_{K,u}$ but drop α from the notation. Since S_{K_0} is real valued we find

$$(S_{K_0} + \operatorname{Re} S_{\overline{K}})^2 + \operatorname{Im} S_{\overline{K}}^2 = |S_K|^2 \geq (k-u)^2 = (S_{K_0} + \bar{k} - g(r))^2,$$

where the last equality follows from (21). Expanding this, cancelling $\operatorname{Re} S_{\overline{K}}^2 + \operatorname{Im} S_{\overline{K}}^2$ on the left with $\bar{k}^2$ on the right and cancelling out $S_{K_0}^2$ one gets

$$2S_{K_0} \operatorname{Re} S_{\overline{K}} \geq g(r)^2 + 2S_{K_0}\bar{k} - 2\bar{k}g(r) - 2S_{K_0}g(r).$$

Rearranging terms we find

$$2S_{K_0}(\bar{k} - \operatorname{Re} S_{\overline{K}}) \leq g(r)(2\bar{k} + 2S_{k_0} - g(r))$$

and hence

$$\bar{k} - \operatorname{Re} S_{\overline{K}} \leq g(r)\left(1 + \frac{\bar{k} - g(r)/2}{S_{K_0}}\right) \leq g(r)\left(1 + \frac{\bar{k}}{S_{K_0}}\right),$$

because $g \geq 0$. Now we use (19) to get

$$S_{K_0} \geq k_0 - u = k - \bar{k} - u \geq k - \frac{3u}{2} \geq \frac{1}{2}k.$$

Substituting this in the previous inequality we get

$$\bar{k} - \operatorname{Re} S_{\overline{K}} \leq g(r)\left(1 + \frac{2\bar{k}}{k}\right).$$

Using (19) again we get the second inequality. □

Recall that the numbers d_n are the Fourier coefficients of g and that their integral representations are given in (10) and (11), with f replaced by g.

Proposition 7.4. — *We have the following inequalities.*

$$\bar{k} \leq d_0 + \frac{u^2}{k} + 6u\sqrt{\delta}. \tag{7.4.1}$$

$$2\bar{k} \mp k_n \leq 2d_0 \mp d_n + 2\frac{u^2}{k} + 24u\sqrt{\delta}. \tag{7.4.2}$$

Proof. — We multiply (23) by the positive functions $1 \pm \cos(2\pi nr)$ and integrate from $\frac{1}{2}$ to 1. By definition 7.1 the second case in (23) occurs on a set of measure at most δ of r's. On this small set we bound $1 \pm \cos(2\pi nr)$ by 2. We easily obtain

$$\int_{\frac{1}{2}}^{1} (\bar{k} - \operatorname{Re} S_{\overline{K}}(\beta' r))(1 \pm \cos(2\pi nr))dr \leq \frac{1}{2}d_0 \mp \frac{1}{4}d_n + \frac{u^2}{2k} + 2\delta u. \tag{24}$$

Similarly, multiplying by 1 and integrating, we get

$$\int_{\frac{1}{2}}^{1} (\bar{k} - \operatorname{Re} S_{\overline{K}}(\beta' r))dr \leq \frac{1}{2}d_0 + \frac{u^2}{2k} + \delta u.$$

We will derive (7.4.1), the derivation of (7.4.2) being similar and simpler. We expand the left hand side of (24).

$$\begin{aligned}
&\int_{\frac{1}{2}}^{1} (\bar{k} - \operatorname{Re} S_{\overline{K}}(\beta' r))(1 \pm \cos(2\pi nr))dr \\
&= \frac{1}{2}\bar{k} - \int_{\frac{1}{2}}^{1} \operatorname{Re} S_{\overline{K}}(\beta' r)dr \mp \int_{\frac{1}{2}}^{1} \operatorname{Re} S_{\overline{K}}(\beta' r)\cos(2\pi nr)dr = \\
&= \frac{1}{2}\bar{k} - \sum_{a \in \overline{K}} \int_{\frac{1}{2}}^{1} \cos(2\pi |a| \beta' r)dr \\
&\mp \sum_{a \in \overline{K}} \int_{\frac{1}{2}}^{1} \cos(2\pi |a| \beta' r)\cos(2\pi nr)dr.
\end{aligned} \tag{25}$$

We will split the summands in the last sum as

$$\begin{aligned}
&\int_{\frac{1}{2}}^{1} \cos(2\pi |a| \beta' r)\cos(2\pi nr)dr \\
&= \frac{1}{2}\int_{\frac{1}{2}}^{1} \cos(2\pi(|a|\beta' + n)r)dr + \frac{1}{2}\int_{\frac{1}{2}}^{1} \cos(2\pi(|a|\beta' - n)r)dr.
\end{aligned}$$

In case $a \in K_n \cup K_{-n}$ we will further split

$$\frac{1}{2}\int_{\frac{1}{2}}^{1} \cos(2\pi(|a|\beta' - n)r)dr = \frac{1}{4} + \frac{1}{2}\left(\int_{\frac{1}{2}}^{1} \cos(2\pi(|a|\beta' - n)r)dr - \frac{1}{2}\right).$$

We wish to obtain a lower bound for the left hand side of (24). We take all non-constant terms in the last 2 identities, replace them by the negatives of their absolute values and plug into (25). This gives

$$\begin{aligned}
&\int_{\frac{1}{2}}^{1}(\bar{k}-\operatorname{Re}S_{\overline{K}}(\beta' r))(1\pm\cos(2\pi nr))dr\\
&\geq \frac{1}{2}\bar{k}\mp\frac{1}{4}k_n\\
&-\frac{1}{2}\sum_{a\in K_n\cup K_{-n}}\left|\int_{\frac{1}{2}}^{1}\cos(2\pi(|a|\beta'-n)r)dr-\frac{1}{2}\right|\\
&-\sum_{a\in\overline{K}}\left|\int_{\frac{1}{2}}^{1}\cos(2\pi|a|\beta' r)dr\right|\\
&-\frac{1}{2}\sum_{a\in\overline{K}-(K_n\cup K_{-n})}\left|\int_{\frac{1}{2}}^{1}\cos(2\pi(|a|\beta'+n)r)dr\right|\\
&-\frac{1}{2}\sum_{a\in\overline{K}-(K_n\cup K_{-n})}\left|\int_{\frac{1}{2}}^{1}\cos(2\pi(|a|\beta'-n)r)dr\right|.
\end{aligned}\tag{26}$$

The last four terms are error terms. The last three have a common form, namely, they are sums of terms of the form $\int_{1/2}^{1}\cos(2\pi br)\,dr$, with b sufficiently large. In fact, b will be $|a|\beta'$ or $|a|\beta'+n$ for $a\in\overline{K}$ in the second and third terms respectively, or $|a|\beta'-n$ for $a\in\overline{K}-K_{\pm n}$ in the fourth term. By (20) we have in all cases that $|b|>\frac{1}{8}$.

Lemma 7.5. — *Suppose $|b|>\frac{1}{8}$ and that $\delta'<1$. Then*

$$\left|\int_{\frac{1}{2}}^{1}\cos(2\pi br)dr\right|\leq\frac{12}{\pi}|\sin(\pi b(1-\delta'))|+\frac{3}{2}\delta'.$$

Proof. — By using the trivial bound $|\cos(x)|\leq 1$ on $[1-\delta',1]\cup\frac{1}{2}[1-\delta',1]$ we get

$$\left|\int_{\frac{1}{2}}^{1}\cos(2\pi br)dr\right|\leq\left|\int_{\frac{1-\delta'}{2}}^{1-\delta'}\cos(2\pi br)dr\right|+\frac{3}{2}\delta'.$$

Now set $b'=b(1-\delta')$. Then

$$\left|\int_{\frac{1-\delta'}{2}}^{1-\delta'}\cos(2\pi br)dr\right|=\left|\frac{1}{2\pi b}(\sin(2\pi b')-\sin(\pi b'))\right|=$$

$$=\left|\frac{1}{2\pi b}(2\sin(\pi b')\cos(\pi b')-\sin(\pi b'))\right|\leq\frac{2+1}{2\pi(1/8)}|\sin(\pi b')|=\frac{12}{\pi}|\sin(\pi b')|.$$

□

Applying the lemma just proved to the second error term we get

$$\sum_{a\in\overline{K}}\left|\int_{\frac{1}{2}}^{1}\cos(2\pi|a|\beta' r)\,dr\right| \leq \frac{12}{\pi}\sum_{a\in\overline{K}}|\sin(\pi a\beta'(1-\delta'))| + \frac{3}{2}\bar{k}\delta'. \tag{27}$$

By the definition 7.1 of δ and by proposition 7.2 we know that there exists some $\delta' < \delta$ such that

$$\bar{k} - \operatorname{Re} S_{\overline{K}}(\beta'(1-\delta')) \leq g(1-\delta')\left(1+\frac{u}{k}\right). \tag{28}$$

The left hand side can be expanded as

$$\sum_{a\in\overline{K}}(1-\cos(2\pi a\beta'(1-\delta'))) = 2\sum_{a\in\overline{K}}\sin^2(\pi a\beta'(1-\delta')).$$

Using (6.7.5) to bound $g(1-\delta)$ we get the estimate

$$\sum_{a\in\overline{K}}\sin^2(\pi a\beta'(1-\delta')) \leq \delta' u\left(1+225\frac{u}{k_0}\right)\left(1+\frac{u}{k}\right) \leq \delta u\left(1+300\frac{u}{k}\right).$$

Lemma 7.6. — *For any positive real numbers $x_1,\ldots,x_r$ we have*

$$\sum_{i=1}^{r} x_i \leq \sqrt{r\sum_{i=1}^{r} x_i^2}.$$

Proof. — This is just the Cauchy-Schwartz formula for the vectors $(1,\ldots,1)$ and $(x_1,x_2,\ldots,x_r)$. □

Applying this lemma in our situation to the numbers $|\sin(\pi a\beta'(1-\delta'))|$ for $a\in\overline{K}$ we find

$$\sum_{a\in\overline{K}}|\sin(\pi a\beta'(1-\delta'))| \leq \sqrt{\bar{k}\delta u(1+300u/k)}.$$

By (27),

$$\sum_{a\in\overline{K}}\left|\int_{\frac{1}{2}}^{1}\cos(2\pi a\beta' r)\,dr\right| \leq \frac{12}{\pi}\sqrt{\bar{k}u\,(1+300u/k)\,\delta} + \frac{3}{2}\bar{k}\delta.$$

Using (19) we see that the right hand side is smaller than

$$\frac{12}{\pi}\sqrt{\frac{1}{2}u\delta(1+o(\epsilon))} + \frac{3}{4}u\delta \leq 2.9u\delta^{\frac{1}{2}} + \frac{3}{4}u\delta.$$

Clearly the same bound holds for the

$$\sum\left|\int_{\frac{1}{2}}^{1}\cos(2\pi(a\beta' \pm n)r)\,dr\right|$$

and therefore the last three terms in (26) can be bounded by $5.8u\sqrt{\delta} + (3/2)u\delta$.

We now handle the first error term.

Lemma 7.7. — *If $|b| < \frac{1}{8}$ and $\delta' < 1/2$, then*

$$\left| \int_{\frac{1}{2}}^{1} \cos(2\pi br)\ dr - \frac{1}{2} \right| \leq \frac{1}{2}(1 - \cos(2\pi b(1-\delta'))) + \delta'.$$

Proof. — We have

$$\left| \int_{\frac{1}{2}}^{1} \cos(2\pi br)\ dr - \frac{1}{2} \right| = \left| \int_{\frac{1}{2}}^{1} (1 - \cos(2\pi br))\ dr \right|.$$

The estimate of the lemma is obtained by bounding the integrand by $1 - \cos(2\pi b(1 - \delta'))$ on $[\frac{1}{2}, 1-\delta']$ and by 1 on $[1-\delta', 1]$. □

Having the lemma, an argument similar to the one used to bound the last three error terms gives, using (28) and (6.7.5),

$$\sum_{a \in K_n \cup K_{-n}} \left| \int_{\frac{1}{2}}^{1} \cos(2\pi(|a|\beta' - n)r)\ dr - \frac{1}{2} \right|$$
$$\leq \frac{1}{2} g(1-\delta)(1 + o(\epsilon)) + \bar{k}\delta \leq \delta(u(1+o(\epsilon)) + \bar{k})$$

Altogether, the four error terms are bounded by $5.8u\sqrt{\delta} + \delta(u(2.5+o(\epsilon))+\bar{k})$. Because $\delta << 1$ this bound is $\leq 5.9u\sqrt{\delta}$. Now multiply (26) and (24) by 4 and compare them. One gets (7.4.2) up to noticing that we can neglect the δu term by increasing a bit the constant on the $u\sqrt{\delta}$ term. □

Corollary 7.8. — *We have*

$$|k_n - d_n| \leq 2(d_0 - \bar{k}) + \frac{2u^2}{k} + 24u\sqrt{\delta}.$$

Proof. — This is because (7.4.2) implies immediately

$$\mp k_n \pm d_n \leq 2(d_0 - \bar{k}) + \frac{2u^2}{k} + 24u\sqrt{\delta}.$$

□

8. Small perturbations in k_0 and u

In this section we prove bounds on $\beta - \beta'$ and $c_n - d_n$. Recall that

$$\beta = \beta(m_0, w), \quad \beta' = \beta(k_0, u), \quad c_n = c_n(m_0, w), \text{ and } d_n = c_n(k_0, u).$$

Therefore, the bounds will depend on m_0, k_0, w and u and all we need to do is to bound the perturbation of the functions β and c_n in terms of the parameters. Since we are assuming that $K \in \mathbb{K}_{\mu_{K_{ex}}(u)}$ and since $E_{K,u} \subseteq [-\beta', \beta']$ by (22) where as $E_{K_{ex},u} \supseteq [-\beta, \beta]$ by (6.12.3), we find $\beta' \geq \beta$. By the definition 7.1 of the parameter δ we have

$$\delta \leq \frac{\beta' - \beta}{\beta'} \tag{29}$$

Proposition 8.1. — *Suppose we are given m_0, k_0, w and u that satisfy*

$$|k_0 - m_0| \le u, \tag{8.1.1}$$

$$w \le u \le w + \left(\frac{w}{k}\right)^3 w + 2\left(\frac{w}{k}\right)^{3/2}. \tag{8.1.2}$$

Let $\beta = \beta(m_0, w)$, $\beta' = \beta(k_0, u)$ and assume that $\beta' \ge \beta$. Then,

$$\beta' - \beta \le \left[\frac{1}{2}\left(\frac{u-w}{w}\right) + \frac{3(m_0 - k_0)}{2m_0}\left(1 \pm \frac{2w}{m_0}\right)\right]\left(1 \pm 250\frac{u}{k_0}\right)\beta'$$

and

$$\beta' - \beta \ge \left[\frac{1}{2}\left(\frac{u-w}{w}\right) + \frac{3(m_0 - k_0)}{2m_0}\left(1 \pm \frac{2w}{m_0}\right)\right]\left(1 \pm 250\frac{u}{m_0}\right)\beta.$$

Here, the sign depends on the sings of $m_0 - k_0$. The correct signs are those for which the bound is weakest possible.

Proof. — We will assume that $m_0 > k_0$, the other case being similar. By the mean value theorem there exists some $\beta'' \in [\beta, \beta']$ such that

$$s'_{k_0}(\beta'') = \frac{s_{k_0}(\beta) - s_{k_0}(\beta')}{\beta - \beta'}.$$

Therefore,

$$\begin{aligned}\beta' - \beta &= \frac{s_{k_0}(\beta) - s_{k_0}(\beta')}{-s'_{k_0}(\beta'')} \\ &= \frac{m_0 - w - s_{m_0}(\beta) + s_{k_0}(\beta) - (k_0 - u)}{-s'_{k_0}(\beta'')} \\ &= \frac{u - w + (m_0 - s_{m_0}(\beta)) - (k_0 - s_{k_0}(\beta))}{-s'_{k_0}(\beta'')}.\end{aligned}$$

By using the first expression for s_k in (3) we find

$$\begin{aligned}(m_0 - s_{m_0}(\beta)) - (k_0 - s_{k_0}(\beta)) &= \left(\sum_{n=(1-m_0)/2}^{(m_0-1)/2} - \sum_{n=(1-k_0)/2}^{(k_0-1)/2}\right)(1 - \cos(2\pi n\beta)) \\ &\le (m_0 - k_0)(1 - \cos(\pi m_0 \beta)) \le \frac{1}{2}(m_0 - k_0)(\pi m_0 \beta)^2.\end{aligned}$$

By proposition 3.4 we find

$$(\pi m_0 \beta)^2 = 6\left(\frac{w}{m_0}\right)\left(1 + \frac{3w}{20m_0} + O(1)\left(\frac{w}{m_0}\right)^2\right)^2 \le 6\left(\frac{w}{m_0}\right)\left(1 + \frac{2w}{m_0}\right),$$

which implies

$$(m_0 - s_{m_0}(\beta)) - (k_0 - s_{k_0}(\beta)) \le (m_0 - k_0)\frac{3w}{m_0}\left(1 + \frac{2w}{m_0}\right).$$

Using (6.7.4) one gets

$$-s'_{k_0}(\beta'') = -\frac{f'_{k_0,u}\left(\frac{\beta''}{\beta'}\right)}{\beta'} \ge \frac{2u}{\beta'}\left(1-225\frac{u}{k_0}\right)\frac{\beta''}{\beta'} \ge \frac{2w}{\beta'}\left(1-225\frac{u}{k_0}\right)\frac{\beta}{\beta'}$$

and therefore

$$\begin{aligned}\frac{\beta'-\beta}{\beta'} &\le \left[\frac{1}{2}\left(\frac{u-w}{w}\right)+\frac{3(m_0-k_0)}{2m_0}\left(1+\frac{2w}{m_0}\right)\right]\\ &\quad\times\left[\left(1-225\frac{w}{k_0}\right)\left(1-\frac{\beta'-\beta}{\beta'}\right)\right]^{-1}.\end{aligned}$$

To get the first inequality we iterate some trivial estimate on $\beta-\beta'$. Start with $\beta-\beta'\le\beta'/2$. Then we can derive an inequality of the form $\frac{\beta'-\beta}{\beta}\le\frac{4u}{m_0}$ (see the derivation of (8.2.2) below). This in turn implies

$$\left[\left(1-225\frac{w}{k_0}\right)\left(1-\frac{\beta'-\beta}{\beta'}\right)\right]^{-1}\le\left(1-225\frac{w}{k_0}-\frac{4u}{m_0}\right)^{-1}\le 1+250\frac{u}{k_0}.$$

This gives the first inequality. The second is similar. □

Corollary 8.2. — *We have the following inequalities.*

$$\bar{k}\ge\bar{m}-\frac{w^3}{k^2}-\left(\frac{w}{k}\right)^{1/2}. \tag{8.2.1}$$

$$\frac{\beta'-\beta}{\beta}\le\frac{2u}{m_0}. \tag{8.2.2}$$

$$\delta\le\left(\frac{w}{k}\right)^3+\frac{2}{k}\left(\frac{w}{k}\right)^{1/2}+\frac{3|\bar{k}-\bar{m}|}{2k}\left(1+300\frac{u}{k}\right). \tag{8.2.3}$$

Proof. — For the first inequality we only need to check the case $\bar{k}<\bar{m}$. In this case $m_0<k_0$. From proposition 8.1 and the assumption $\beta'\ge\beta$ we find

$$0\le\frac{1}{2}\left(\frac{u-w}{w}\right)+\frac{3(\bar{k}-\bar{m})}{2m_0}\left(1-\frac{2w}{m_0}\right).$$

Therefore,

$$\begin{aligned}\bar{m}-\bar{k}&\le\frac{1}{2}\left(\frac{w-u}{w}\right)\left(\frac{3}{2m_0}(1-o(\epsilon))\right)^{-1}\\ &=\frac{1}{3}m_0\left(\frac{w-u}{w}\right)(1+o(\epsilon))\\ &\le\frac{1}{3}m_0\left(\frac{w}{k}\right)^3(1+o(\epsilon))+\frac{2}{3}\frac{m_0}{w}\left(\frac{w}{k}\right)^{3/2}(1+o(\epsilon))\quad\text{by (8.1.2)}\\ &\le\frac{w^3}{k^2}+\left(\frac{w}{k}\right)^{1/2}.\end{aligned}$$

Next, by proposition 8.1 and by (8.1.2) and (8.1.1) we get

$$\frac{\beta'-\beta}{\beta}\le\left[\frac{1}{2}\left(\frac{w}{k}\right)^3+2w^{-1}\left(\frac{w}{k}\right)^{3/2}+\frac{3u}{2m_0}(1+o(\epsilon))\right](1+o(\epsilon))\le\frac{2u}{m_0},$$

proving (8.2.2). The third inequality follows easily from the proposition and (29). □

Lemma 8.3. — *We have the following inequalities.*

$$|d_n - c_n| \leq \frac{u^2}{m_0(\pi n)^2}. \tag{8.3.1}$$

$$|d_0 - c_0| \leq \frac{u^2}{3m_0}. \tag{8.3.2}$$

Proof. — We follow the proofs of lemma 6.6 and proposition 6.8. Let

$$\sin(k_0 z)/\sin(z) = \sum_{j=0}^{\infty} b'_j z^j$$

be a Taylor expansion around 0, and let

$$g(r) = u + \sum_{j=1}^{\infty} a'_j r^{2j}$$

be the expansion for g converging on $[1/2, 1]$. We have $a'_j = b'_{2j}(\pi\beta')^{2j}$. The identity

$$\frac{\sin(m_0 z)}{\sin(z)} - \frac{\sin(k_0 z)}{\sin(z)} = 2\sin\left(\frac{m_0 - k_0}{2}z\right)\cos\left(\frac{m_0 + k_0}{2}z\right)\Big/\sin(z)$$

easily implies, together with lemma 6.5, that on a circle C_{m_0} of radius $1/m_0$ we have

$$\left|\frac{\sin(m_0 z)}{\sin(z)} - \frac{\sin(k_0 z)}{\sin(z)}\right| \leq 2|m_0 - k_0|.$$

Therefore, using the Cauchy integral formula, we have

$$|b_j - b'_j| \leq 2{m_0}^j |m_0 - k_0|.$$

Recall how the constant c was defined in lemma 6.6 and used in its proof. We find

$$\begin{aligned}
|a_j - a'_j| &= |b_{2j}(\pi\beta)^{2j} - b'_{2j}(\pi\beta')^{2j}| \\
&\leq |b_{2j} - b'_{2j}|(\pi\beta)^{2j} + |b'_{2j}|\pi^{2j}|\beta'^{2j} - \beta^{2j}| \\
&\leq 2c^j|m_0 - k_0| + 2m_0 c^j\left(\frac{\beta'^{2j} - \beta^{2j}}{\beta^{2j}}\right) \\
&= 2c^j|m_0 - k_0| + 2m_0 c^j\left(\left(1 + \frac{\beta' - \beta}{\beta}\right)^{2j} - 1\right).
\end{aligned}$$

By (8.1.1) and (8.2.2) we get

$$\begin{aligned}
|a_j - a'_j| &\leq 2c^j u + 2m_0 c^j\left(u + m_0((1 + 2u/m_0)^{2j} - 1)\right) \\
&\leq 2c^j\left(u + m_0((1 + 2u/m_0)^{2j} - 1)\right).
\end{aligned}$$

When $j = 2$ we use the bound

$$\left(1 + 2\frac{u}{m_0}\right)^4 - 1 \leq 9\frac{u}{m_0}$$

to get

$$|a_2 - a_2'| \le 20c^2 u \le 20 \cdot 6^2(1 + o(\epsilon)) \left(\frac{u}{m_0}\right)^2 u \le 800 \left(\frac{u}{m_0}\right)^2 u.$$

When $j > 2$ we use the bound

$$\left(1 + 2\frac{u}{m_0}\right)^{2j} - 1 < \left(1 + 2\frac{u}{m_0}\right)^{2j} = (1 + o(\epsilon))^{2j}.$$

This gives

$$|a_j - a_j'| \le 2\left(c(1 + o(\epsilon))^2\right)^j (u + m_0) \le 2 \cdot 7^j \left(\frac{u}{m_0}\right)^{j-1} u.$$

Since $f(1) = g(1) = 0$ we see that

$$\begin{aligned}|a_1 - a_1'| &\le |w - u| + \sum_{j=2}^{\infty} |a_j - a_j'| \\ &\le \left(\frac{w}{k}\right)^3 w + 800 \left(\frac{u}{m_0}\right)^2 u + 2 \cdot 7^3 \left(\frac{u}{m_0}\right)^2 u \\ &+ \quad \text{lower order terms} \quad \le 1500 \left(\frac{u}{m_0}\right)^2 u.\end{aligned}$$

Then, imitating the proof of proposition 6.8, we find

$$\begin{aligned}|c_n - d_n| &\le \frac{8}{(2\pi n)^2} \frac{d}{dr} \left(\sum_{j=0}^{\infty} |a_j - a_j'| r^{2j}\right) |_{r=1} \\ &\le \frac{2}{(\pi n)^2}(2 \cdot |a_1 - a_1'| + 4 \cdot |a_2 - a_2'| + 6 \cdot |a_3 - a_3'| + \text{l.o.t.}) \\ &\le \frac{2}{(\pi n)^2} \left(\frac{u}{m_0}\right)^2 u \ (2 \cdot 1500 + 4 \cdot 800 + 6 \cdot 2 \cdot 7^3 + \text{l.o.t.}) \\ &\le 21000 \frac{1}{(\pi n)^2} \left(\frac{u}{m_0}\right)^2 u \le \frac{1}{(\pi n)^2} \left(\frac{u}{m_0}\right) u.\end{aligned}$$

The second estimate is similar. □

9. The main theorems

In the last two sections we gathered many conditional estimates. It is now easy to make these absolute.

Lemma 9.1. — *We have*

$$|\bar{k} - \bar{m}| \le 2\frac{u^2}{k} + 6u\sqrt{\delta} + 5.2\sqrt{u}.$$

Proof. — If $\bar{k} < \bar{m}$, then (8.2.1) implies that $|\bar{k} - \bar{m}| \leq w^3/k^2 + (w/k)^{1/2}$, which is certainly within the bound. Otherwise we have

$$|\bar{k} - \bar{m}| = \bar{k} - \bar{m} = (\bar{k} - d_0) + (d_0 - c_0) + (c_0 - \bar{m}),$$

which, by (6.12.4),(7.4.1) and (8.3.2), is

$$\begin{aligned}&\leq \left(\frac{u^2}{k} + 6u\sqrt{\delta}\right) + \left(\frac{u^2}{3m_0}\right) + (5.2\sqrt{u})\\ &\leq 2\frac{u^2}{k} + 6u\sqrt{\delta} + 5.2\sqrt{u}.\end{aligned}$$

□

Proposition 9.2. — *We have the inequality*

$$\sqrt{\delta} \leq 10\left(\frac{u}{k}\right) + \sqrt{8}\left(\frac{u}{k}\right)^{1/4} k^{-1/4}.$$

Proof. — The last lemma, together with (8.2.3), gives the inequality

$$\begin{aligned}\delta &\leq \left(\frac{w}{k}\right)^3 + \frac{2}{k}\left(\frac{w}{k}\right)^{1/2} + \frac{3}{2k}\left(2\frac{u^2}{k} + 6u\sqrt{\delta} + 5.2\sqrt{u}\right)(1 + o(\epsilon))\\ &\leq 4\left(\frac{u}{k}\right)^2 + 9\left(\frac{u}{k}\right)(1 + o(\epsilon))\sqrt{\delta} + \frac{8\sqrt{u}}{k}.\end{aligned}$$

Here, notice that we were able to swallow the first two terms on the first line, the $(w/k)^3$ term in the $(u/k)^2$ term and the $(2/k)(w/k)^{1/2}$ term by the $\sqrt{u}/k$ term. We treat the last inequality as a quadratic inequality in $x = \sqrt{\delta}$, which reads

$$x^2 \leq 4\left(\frac{u}{k}\right)^2 + 9\left(\frac{u}{k}\right)(1 + o(\epsilon))x + \frac{8\sqrt{u}}{k}.$$

The variable x should lie between the roots of the corresponding equation,

$$x^2 - 9\left(\frac{u}{k}\right)(1 + o(\epsilon))x - \left(4\left(\frac{u}{k}\right)^2 + \frac{8\sqrt{u}}{k}\right) = 0.$$

In particular, it is smaller than the bigger root. This gives

$$\sqrt{\delta} \leq \frac{1}{2}\left(\frac{9u}{k}(1 + o(\epsilon)) + \sqrt{(81 + 12)\left(\frac{u}{k}\right)^2(1 + o(\epsilon)) + \frac{32\sqrt{u}}{k}}\right).$$

Using the fact that for any two positive reals a and b we have $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$, we easily get the required bound. □

Corollary 9.3. — *We have* $\delta < 200\left(\frac{u}{k}\right)^2 + \frac{16\sqrt{u}}{k}$.

Proof. — This just uses the inequality $(a + b)^2 \leq 2(a^2 + b^2)$. □

We now begin to state the main theorems. Recall our assumptions from the beginning of section 7: We have $u > 1$, $k > 30000u$ and k and u satisfy assumption 5.4. We have a set $K \in \mathbb{K}_{\mu_{K_{ex}}(u)}$, written as $K = K_0 \cup \overline{K}$. Recall also that K_0 is contained inside an arithmetic progression. Let q be the difference of this sequence. By translation we can normalize things so that elements of K_0 are divisible by q. Recall that we have a compression procedure for K_0 to an arithmetic progression with at most one gap. Assume also that this progression is symmetric around 0 (as best possible). This fixes K up to shift. With these assumption we associated certain parameters k_n to K, counting roughly the number of elements near $\pm n/\beta(k_0, u)$. When $q \neq 1$ we count only those elements which are divisible by q. We have analogous parameters m_n for our "test set" K_{ex}. The theorem will show that these parameters are pretty close to each other.

Theorem 9.4. — *We have the following inequalities.*

$$m_0 + \frac{u^3}{k^2} + \frac{1}{k}\left(\frac{u}{k}\right)^{1/2} \geq k_0 \geq m_0 - 62\frac{u^2}{k} - 6\sqrt{u} - 20\left(\frac{u}{k}\right)^{1/4} k^{-1/4}u. \tag{9.4.1}$$

$$|k_n - m_n| \leq 250\frac{u^2}{k} + 11\sqrt{u} + 70\left(\frac{u}{k}\right)^{1/4} k^{-1/4}u + 4. \tag{9.4.2}$$

$$|\{a \in K : \quad |a|\beta \geq t \quad \textit{or} \quad q \nmid a\}| \leq \frac{4u}{t} + tu\left(400\left(\frac{u}{k}\right)^2 + \frac{32\sqrt{u}}{k}\right), \quad \textit{for} \quad t > 0. \tag{9.4.3}$$

Proof. — As discussed at the beginning of section 7, we may assume that K_0 is an arithmetic progression with at most one gap. Continue first to assume that it is a progression, has difference 1 and is symmetric around 0. We now write out the results we have so far.

The left inequality in (9.4.1) is just (8.2.1). The right inequality follows since

$$\begin{aligned}|\bar{k} - \bar{m}| &\leq 2\frac{u^2}{k} + 6u\sqrt{\delta} + 5.2\sqrt{u} \quad \text{by lemma 9.1}\\ &\leq 2\frac{u^2}{k} + 5.2\sqrt{u} + 6u\left(10\left(\frac{u}{k}\right) + \sqrt{8}\left(\frac{u}{k}\right)^{1/4} k^{-1/4}\right) \quad \text{by proposition 9.2}\\ &\leq 62\frac{u^2}{k} + 6\sqrt{u} + 20\left(\frac{u}{k}\right)^{1/4} k^{-1/4}u.\end{aligned}$$

We next establish

$$d_0 - \bar{k} \leq 5.5\sqrt{u} + \frac{u^2}{2k}. \tag{30}$$

Indeed

$$\begin{aligned}\bar{k} &\geq \bar{m} - \frac{u^3}{k^2} - \left(\frac{u}{k}\right)^{1/2} \geq c_0 - 5.2\sqrt{u} - \frac{u^3}{k^2} - \left(\frac{u}{k}\right)^{1/2}\\ &\geq d_0 - \frac{u^2}{3m_0} - 5.2\sqrt{u} - \frac{u^3}{k^2} - \left(\frac{u}{k}\right)^{1/2} \geq d_0 - \frac{u^2}{2k} - 5.5\sqrt{u}\end{aligned}$$

by (6.12.4), (8.2.1) and (8.3.2). Here again, like in the proof of proposition 9.2, we are able to swallow the terms $(u/k)^{1/2}$ and u^3/k^2. It now follows that

$$\begin{aligned}|k_n - d_n| &\le 2(d_0 - \bar{k}) + \frac{2u^2}{k} + 24u\sqrt{\delta} \quad \text{by corollary 7.8}\\ &\le 11\sqrt{u} + \frac{u^2}{k} + \frac{2u^2}{k} + 24u\sqrt{\delta} \quad \text{by (30)}\\ &\le 11\sqrt{u} + \frac{3u^2}{k} + 24u\left(10\left(\frac{u}{k}\right) + \sqrt{8}\left(\frac{u}{k}\right)^{1/4} k^{-1/4}\right) \quad \text{by proposition 9.2}\\ &\le 243\frac{u^2}{k} + 11\sqrt{u} + 70\left(\frac{u}{k}\right)^{1/4} k^{-1/4}u.\end{aligned}$$

From (8.3.2) we get

$$|d_n - c_n| \le \frac{u^2}{m_0\pi^2} \le \frac{u^2}{9k}.$$

By the definition of the m_n in (6.1) (taking into account the possible modifications at the proof of (6.12)), we see that $|c_n - m_n| \le 4$. This gives (9.4.2).

To get (9.4.3) we integrate the inequality of proposition 7.2 on the interval $[1-\frac{1}{t}, 1]$. Recalling that there is a subset of size δ on which the inequality fails to hold, and on which bound $\bar{k} - \operatorname{Re} S_{\overline{K}}$ trivially by $2\bar{k} \le u$, we get

$$\int_{1-\frac{1}{t}}^{1} (\bar{k} - \operatorname{Re} S_{\overline{K}}(\beta r)) \le \int_{1-\frac{1}{t}}^{1} g(r)\left(1 + \frac{u}{k}\right) dr + u\delta. \tag{31}$$

The left hand side is

$$\int_{1-\frac{1}{t}}^{1} (\bar{k} - \operatorname{Re} S_{\overline{K}}(\beta r))\, dr = \sum_{a\in\overline{K}} \int_{1-\frac{1}{t}}^{1} (1 - \cos(2\pi a\beta r))\, dr.$$

One finds that if $b \ge t$, then

$$\begin{aligned}\int_{1-\frac{1}{t}}^{1} (1 - \cos(2\pi b r))\, dr &= \frac{1}{t} - \frac{1}{2\pi b}\left(\sin\left(2\pi b\left(1 - \frac{1}{t}\right)\right) - \sin(2\pi b)\right)\\ &\ge \frac{1}{t} - \frac{2}{2\pi b} \ge \frac{1}{2t}.\end{aligned}$$

It follows that the left hand side of (31) is greater than

$$\frac{1}{2t}|\{a \in K : \quad |a|\beta \ge t\}|.$$

Using (6.7.5) the right hand side of (31) is smaller than

$$\left(1 + 225\frac{u}{k_0}\right) u \int_0^{\frac{1}{t}} 2r\, dr + u\delta \le \frac{2u}{t^2} + u\delta.$$

Therefore,

$$\frac{1}{2t}|\{a \in K : \quad |a|\beta \ge t\}| \le \frac{2u}{t^2} + \frac{1}{2}u\delta.$$

Thus,

$$|\{a \in K : \quad |a|\beta \geq t\}| \leq \frac{4u}{t} + 2u\delta,$$

which gives the result after substituting the estimate for δ given in corollary 9.3.

To see that the everything continues to hold for the case $q > 1$, we consider the structure of the proof and see what modifications one must make. Most of the time we have been integrating certain identities on certain intervals. Whenever there is an integral on some interval I involved for the case $q = 1$, take the corresponding integral on the set $\langle q \rangle^{-1}(I)$ and make the obvious adaptations. It is easy to see that all terms in the integration which involve elements that are not congruent to $0 \mod q$ vanish. There was one place, in deriving (28), where we used the value of the sum at a certain point. But there note that elements congruent to $0 \mod q$ behave the same on all intervals and the claim made there about the existence of δ' is certainly true for at least one interval. With these remarks the proof goes through unchanged. □

Now comes a second "compression argument". It is based on the following lemma.

Lemma 9.5. — *Let x_1, x_2, y_1, y_2 and z be vectors in an Euclidean space with scalar product "$\cdot$" and norm $|\ |$, such that $|x_1| = |x_2| = |y_1| = |y_2| = 1$. Let $x = x_1 + x_2$, $y = y_1 + y_2$. Put $x_1 \cdot x_2 = \cos\theta_1$, $y_1 \cdot y_2 = \cos\theta_2$, $x \cdot z = (\cos\theta)|z|$. If*

$$\cos\theta_1 - \cos\theta_2 > 4(1 - \cos\theta),$$

then

$$|x + z| > |y + z|.$$

Proof

$$x \cdot x - y \cdot y = 2 + 2x_1 \cdot x_2 - 2 - 2y_1 \cdot y_2 = 2(\cos\theta_1 - \cos\theta_2) > 8(1 - \cos\theta).$$

Since $|x| + |y| < 4$,

$$|x| - |y| > 2(1 - \cos\theta) \geq (1 - \cos\theta)|x|,$$

and therefore $|x|\cos\theta > |y|$. Thus

$$|z + x|^2 - |z + y|^2 \geq |x|^2 - |y|^2 + 2|z|(|x|\cos\theta - |y|) > 0.$$

□

Corollary 9.6. — *Suppose K is as before, with K_0 and arithmetic progression symmetric around 0. Suppose a, $b \in K$, $q|a$ and $q|b$, and a and b satisfy the inequalities*

$$\beta^{-1} - 2k \geq |a - b|/q \geq 2k.$$

Then, if K' is the set obtained from K by replacing a and b with two elements c and d on the sides of K_0, i.e., $c \approx k_0/2$ and $d = -c$, we get $\mu_{K'}(u) \geq \mu_K(u)$.

Proof. — Assume again that $q = 1$ for simplicity. Observe first that if $\alpha = \beta_{k_0,u}r$, with $r < \frac{1}{\sqrt{6}}$, then

$$|S_{K_0}(\alpha)| - \bar{k} \approx k_0 - ur^2 - \bar{k} \approx k - \frac{5}{6}u - ur^2 > k - \frac{5}{6}u - u\frac{1}{\sqrt{6}}^2 = k - u,$$

because $\bar{k} \approx \frac{5}{12}u$. This implies that such an α belongs to both $E_{K,u}$ and $E_{K',u}$.

Now assume $\alpha > \frac{1}{10}\beta_{k_0,u}$. Set $x_1 = e^{2\pi ic\alpha}$, $x_2 = e^{2\pi id\alpha}$, $y_1 = e^{2\pi ia\alpha}$, $y_2 = e^{2\pi ib\alpha}$ and $z = S_K(\alpha) - (y_1 + y_2)$, so that $z + y$, $z + x$ are the values of $S_K(\alpha)$ and $S_{K'}(\alpha)$ respectively. In our case the conditions of the lemma hold. To see this, notice that we have

$$4\pi k\alpha \leq 2\pi(a-b)\alpha \leq 2\pi(a-b)\beta \leq 2\pi - 4\pi k\beta \leq 2\pi - 4\pi k\alpha.$$

It follows that $\cos(\theta_2) = \cos(2\pi(a-b)\alpha) \leq \cos(4\pi k\alpha)$. Therefore,

$$\begin{aligned}\cos\theta_1 - \cos\theta_2 &\geq \cos(\theta_1) - \cos(4\pi k\alpha) \approx \cos(2\pi k\alpha) - \cos(4\pi k\alpha)\\ &\approx (4\pi\alpha k)^2 - (2\pi\alpha k)^2 = 3(2\pi\alpha k)^2 \geq 3\left(2\pi\frac{1}{10}\beta_{k_0,u}k\right)^2 \geq \frac{u}{2k}.\end{aligned}$$

On the other hand, since x is on the real line, θ is almost the angle between S_K and the real line so

$$|\sin\theta| = \left|\frac{\operatorname{Im} S_K}{\operatorname{Re} S_K}\right| < \frac{\frac{1}{2}u}{k - \frac{3}{2}u} < \frac{u}{k}$$

and therefore $1 - \cos\theta \leq \sin^2\theta \leq (u/k)^2$. Since the lemma applies we see that $\alpha \in E_{K,u}$ implies $\alpha \in E_{K',u}$, hence the result. □

The following two theorems are immediate consequences.

Theorem 9.7. — *Suppose $K \in \mathbb{K}_{\mu_{K_{ex}}(u)}$. Then around each of $\pm n/\beta$, $n \neq 0$, there exists an interval of length $2kq$ such that the total number of elements of $\cup_{n=1}^{\infty}(K_n \cup K_{-n})$ not in any of the intervals is $\leq 63\frac{u^2}{k} + 7\sqrt{u} + 20\left(\frac{u}{k}\right)^{1/4} k^{-1/4}u$.*

Proof. — Indeed, corollary 9.6 implies that any two elements in any K_n that are of distance $\geq 2kq$ apart can be "pushed" into K_0, increasing $\mu_K(u)$ and in particular keeping the set in $\mathbb{K}_{\mu_{K_{ex}}(u)}$. But by (9.4.1) the number of elements one can move into K_0 keeping K in $\mathbb{K}_{\mu_{K_{ex}}(u)}$ is $\leq 63\frac{u^2}{k} + 7\sqrt{u} + 20\left(\frac{u}{k}\right)^{1/4} k^{-1/4}u$ (again we have swallowed some terms). □

Theorem 9.8. — *The following is a $G_{\mu_{K_{ex}}(u)}$ sub-collection of $\mathbb{K}_{\mu_{K_{ex}}(u)}$. It consists of all sets in which K_0 is an arithmetic progression with at most one gap, and where each of the sets K_n, for $n \neq 0$, is contained in an interval of length $2kq$.*

Proof. — Immediate from corollary 9.6. □

We end with an improved bound on $\mu_{\max}$.

Theorem 9.9. — *We have the following inequalities*

$$\mu_{\max}(k,u) \geq \frac{2\sqrt{6}}{\pi}\frac{1}{k}\left(\frac{u}{k}\right)^{1/2}\left(1+\left(\frac{3}{20}+\frac{5}{8}\right)\frac{u}{k}-300\left(\frac{u}{k}\right)^2-\frac{8\sqrt{u}}{k}\right)$$

$$\mu_{\max}(k,u) \leq \frac{2\sqrt{6}}{\pi}\frac{1}{k}\left(\frac{u}{k}\right)^{1/2}\left(1+\left(\frac{3}{20}+\frac{5}{8}\right)\frac{u}{k}+400\left(\frac{u}{k}\right)^2+30\left(\frac{u}{k}\right)^{5/4}k^{-1/4}\right).$$

Proof. — We clearly have $2\beta_{m_0,w} \leq \mu_{\max}(k,u) \leq 2\beta_{k_0,u}$, so the theorem is about estimating the terms on both sides. We show the second inequality only. By (22) and proposition 8.1,

$$\mu_{\max}(k,u) \leq 2\beta_{k_0,u} \leq 2\beta_{k,u}\left[1+\frac{3(k-k_0)}{2k}\left(1+\frac{2u}{k}\right)\left(1+250\frac{u}{k_0}\right)\right].$$

By proposition 3.4,

$$\beta_{k,u} \leq \frac{\sqrt{6}}{\pi}\frac{1}{k}\left(\frac{u}{k}\right)^{1/2}\left(1+\frac{3u}{20k}+\left(\frac{u}{k}\right)^2\right).$$

By (7.4.1),

$$k-k_0=\bar{k}\leq d_0+\frac{u^2}{k}+6u\sqrt{\delta}.$$

It follows from (6.8.1) that

$$d_0 \leq \frac{5}{12}u+75\frac{u^2}{k_0}\leq\frac{5}{12}u+79\frac{u^2}{k}.$$

Hence, using proposition 9.2 we find

$$\begin{aligned} k-k_0 &\leq \frac{5}{12}u+80\frac{u^2}{k}+6u\left(10\left(\frac{u}{k}\right)+\sqrt{8}\left(\frac{u}{k}\right)^{1/4}k^{-1/4}\right)\\ &\leq \frac{5}{12}u+140\frac{u^2}{k}+17u\left(\frac{u}{k}\right)^{1/4}k^{-1/4}. \end{aligned}$$

It follows that

$$\begin{aligned} &\left(1+\frac{3u}{20k}+\left(\frac{u}{k}\right)^2\right)\left[1+\frac{3}{2k}(k-k_0)\left(1+\frac{2u}{k}\right)\left(1+250\frac{u}{k_0}\right)\right]\\ &\leq 1+\left(\frac{3}{20}+\frac{5}{8}\right)\frac{u}{k}+400\left(\frac{u}{k}\right)^2+30\left(\frac{u}{k}\right)^{5/4}k^{-1/4}. \end{aligned}$$

The upper bound is now clear. □

References

[1] Freiman G. A., *On the measure of large trigonometric sums*, In Sixth international conference on collective phenomena: Reports from the Moscow Refusnik seminar. Annals of the New York academy of sciences. New York, New York, 1985.

[2] Freiman G. A., Yudin A. A., *The general principles of additive number theory*, in [6], 135–147.

[3] Freiman G. A., *Foundations of a structural theory of set addition*, In Translation of Mathematical Monographs, vol. **37**, American Mathematical Society, Providence, R.I., 1973.

[4] Yudin A. A., *The measure of the large values of the modulus of a trigonometric sum*, in [6], 163–171.

[5] Moskvin D. A., Freiman G. A. and Yudin A. A., *Inverse problems of additive number theory and local limit theorems for lattice random variables*, in [6], 163–171.

[6] Freiman G. A. edt., *Number-theoretic studies in the Markov spectrum and in the structural theory of set addition* (Russian), Kalinin Gos. University., Moscow, 1973, 191 pp.

[7] Raikov D., *On the addition of point-sets in the sense of Schnirelmann*, Rec. Math. [Mat. Sbornik] N.S., **5(47)**, 1939, 425–440.

A. Besser, Department of Mathematical Sciences, University of Durham, Science Laboratories, South Road, Durham DH1 3LE, England • *Current address :* SFB 478, Geometrische Strukturen in der Mathematik, Einsteinstr. 62, 48189 Münster, Germany
E-mail : `besser@math.uni-muenster.de` • *Url :* `http://fourier.dur.ac.uk:8000/~dma1ab`

YURI BILU

Structure of sets with small sumset

Astérisque, tome 258 (1999), p. 77-108

<http://www.numdam.org/item?id=AST_1999__258__77_0>

Astérisque **258**, 1999, p. 77–108

STRUCTURE OF SETS WITH SMALL SUMSET

by

Yuri Bilu

Abstract. — Freiman proved that a finite set of integers K satisfying $|K+K| \leq \sigma|K|$ is a subset of a "small" m-dimensional arithmetical progression, where $m \leq \lfloor \sigma - 1 \rfloor$. We give a complete self-contained exposition of this result, together with some refinements, and explicitly compute the constants involved.

1. Introduction

This is an exposition of the fundamental theorem due to G. A. Freiman on the addition of finite sets. (It will be referred to as *Main theorem*). Let K be a finite set of integers (more generally, a finite subset of a torsion-free abelian group) of cardinality k. The Main Theorem states that if the sumset $K+K$ is "small", then K possesses a rigid structure. An example of a statement of this type is the following

Proposition 1.1

(i) *Any K satisfies $|K+K| \geq 2k-1$ and the equality $|K+K| = 2k-1$ implies that K is an arithmetical progression .*

(ii) *Assume that $|K+K| = 2k-1+t$, where $0 \leq t \leq k-3$. Then K is a subset of an arithmetical progression of length $k+t$.*

(iii) *Assume that $|K+K| = 3k-3$ and $k \geq 7$. Then either K is a subset of an arithmetical progression of length $2k-1$, or K is a union of two arithmetical progressions with the same difference.*

Here (i) is trivial, for (ii) and (iii) see [**12**, Theorems 1.9 and 1.11], where the result is obtained for subsets of integers. The case of subsets of an arbitrary torsion-free abelian group follows from [**12**, Lemma 1.14], which is Lemma 4.3 of the present paper.

Let us deviate for a while from our main subject, and make a short (and very incomplete) historical account. Item (i) easily generalizes to distinct summands: *if K*

1991 ***Mathematics Subject Classification.*** — 11B25, 11B05.

Key words and phrases. — Addition of finite sets; generalized arithmetical progressions; inverse additive theorems.

and L are finite subsets of a torsion-free abelian group, then $|K+L| \geq |K|+|L|-1$, and the equality $|K+L| = |K|+|L|-1$ implies that K and L are arithmetical progressions with the same difference. Freiman **[10]** extended item (ii) to two distinct summands; see also **[15, 23, 32, 35]**. An important generalization to several (equal or distinct) summands was obtained by Lev **[22]**. Concerning item (iii) see also Hamidoune **[17]**.

Item (i) extends to torsion-free non-abelian groups (Brailovski and Freiman **[4]**). It also has an analogue for cyclic groups of prime order (Cauchy **[6]**, Davenport **[7, 8]**, Vosper **[36]**). Hamidoune **[16]** gave short and conceptual proofs of the theorems of Brailovski-Freiman and Vosper. For general finite (abelian and/or non-abelian) groups see **[20, 18, 37, 38]**. However, we do not know non-commutative analogues of items (ii) and (iii), and we know only partial analogues of these items for cyclic groups of prime order **[11, 12, 2]**.

The first part of item (i) has various continuous analogues, for instance for connected unimodular locally compact groups **[19, 29]**. Item (ii) has a partial analogue for real tori **[1]**.

Many of the results mentioned above are proved in the books of Mann **[24]** and Nathanson **[26]**, where the reader can also find further references.

The Main Theorem, however, develops Proposition 1.1 in a completely different direction. Reformulate item (ii) as follows:

Let $\sigma < 3$ be a positive number. Assume that $|K+K| \leq \sigma k$ and $k > 3/(3-\sigma)$. Then K is a subset of an arithmetical progression of length $(\sigma-1)k+1$.

The Main Theorem extends this to arbitrary σ, without the restriction $\sigma < 3$. To formulate it, we need some definitions. Let A, B be abelian groups, $K \subset A$ and $L \subset B$. The map $\varphi : K \to L$ is *Freiman's homomorphism of order s* or, in the terminology of [28], *F_s-homomorphism*, if for any $x_1, \dots, x_s, y_1, \dots, y_s \in K$ we have

$$x_1 + \cdots + x_s = y_1 + \cdots + y_s \Rightarrow \varphi(x_1) + \cdots + \varphi(x_s) = \varphi(y_1) + \cdots + \varphi(y_s)$$

In the other words, the map

$$\begin{array}{rrcl} \psi: & \overbrace{K + \cdots + K}^{s} & \to & \overbrace{L + \cdots + L}^{s}, \\ & x_1 + \cdots + x_s & \mapsto & \varphi(x_1) + \cdots + \varphi(x_s) \end{array}$$

is well-defined. The F_s-homomorphism φ is an *F_s-isomorphism* if it is invertible and the inverse φ^{-1} is also an F_s-homomorphism; in other words, when both the maps φ and ψ are invertible. (In particular, *F_1-isomorphism* is a synonym to *bijection.*)

It is easy to find an F_s-isomorphism not induced by a group-theoretic homomorphism $A \to B$. A typical example is the map

$$\begin{array}{rcl} \{0, a, \dots, (k-1)a\} & \to & \{0, \dots, k-1\}, \\ xa & \mapsto & x, \end{array}$$

where a generates an additive cyclic group of order $p > (k-1)s$.

A *generalized arithmetical progression* (further *progression*) of rank m in an abelian group A is a set of the form

$$P = P(x_0; x_1, \dots, x_m; b_1, \dots, b_m) = \{x_0 + \beta_1 x_1 + \cdots + \beta_m x_m \ : \ \beta_i = 0, \dots, b_i - 1\},$$

where $x_0, \dots, x_m$ are elements of the group and $b_1, \dots, b_m$ positive integers. We say that P is an *F_s-progression* if the map

$$\begin{array}{rcl} \{0,\dots,b_1-1\} \times \cdots \times \{0,\dots,b_m-1\} & \to & P, \\ (\beta_1,\dots,\beta_m) & \mapsto & x_0+\beta_1x_1+\cdots+\beta_mx_m, \end{array} \tag{1.1}$$

is an F_s-isomorphism. In particular, each F_s-progression is also an $F_{s'}$-progression for any $s' \le s$, and P is an F_1-progression if and only if $|P| = b_1 \cdots b_m$.

Now we are ready to formulate the Main Theorem[1].

***Theorem 1.2* (the Main Theorem).** — *Let σ be a positive real number, s a positive integer, and K a subset of a torsion-free abelian group such that*

$$k := |K| > k_0(\sigma) := \frac{\lfloor \sigma \rfloor \lfloor \sigma+1 \rfloor}{2(\lfloor \sigma+1 \rfloor - \sigma)}$$

and

$$|K+K| \le \sigma k.$$

Then K is a subset of an F_s-progression P of rank $m \le \lfloor \sigma - 1 \rfloor$ and cardinality

$$|P| \le c_{11}(\sigma,s)k. \tag{1.2}$$

It must be pointed out that, unlike Proposition 1.1, this theorem has only very few known analogues for other types of groups, all of them being more or less direct consequences of the Main Theorem; see Chapter 3 of Freiman's book **[12]**.

We also suggest the following more precise version of the Main Theorem, asserting that at most $\lfloor \log_2 \sigma \rfloor$ dimensions of the progression P can be "large"; the others are bounded by a constant, depending on σ.

Theorem 1.3. — *Assuming the hypothesis of Theorem 1.2, write the F_s-progression P as $P(x_0; x_1, \dots, x_m; b_1, \dots, b_m)$, where $b_1 \ge \cdots \ge b_m$. Then*

$$b_i \le c_{12}(\sigma,s) \qquad (i > \lfloor \log_2 \sigma \rfloor). \tag{1.3}$$

(See Subsection 5.5, where Theorem 1.3 is derived from Theorem 1.2.)

The quantitative estimates for the constants involve the function $fr(n,\varepsilon)$, defined in Subsection 5.3. We obtain the estimates

$$c_{11}(\sigma,s) \le (2c_{13}(\sigma)s)^{\sigma^{30\sigma}c_{13}(\sigma)}, \qquad c_{12}(\sigma,s) \le 2c_{11}(\sigma,s')fr\left(\lfloor \log_2 \sigma \rfloor + 1, \varepsilon_0\right),$$

where

$$c_{13}(\sigma) = fr\left(\lceil 8\sigma \log(2\sigma) \rceil, 1\right), \qquad \varepsilon_0 = \lfloor \log_2 \sigma \rfloor + 1 - \log_2 \sigma, \qquad s' = \min(s,2).$$

At present, only a very poor estimate is known (see Subsection 5.3):

$$fr(n,\varepsilon) \le \left(2+\varepsilon^{-1}\right)^{\exp\exp n}.$$

Therefore we have only

$$c_{11} \le (2s)^{\exp\exp\exp(9\sigma\log(2\sigma))}. \tag{1.4}$$

[1] With a few exceptions, we write explicit constants as c_{ij}, where i is the number of the section where the constant is defined, and j is the number of the constant in Section i.

Freiman published two expositions [12, 13] of his proof. Recently a new proof of Freiman's theorem, simpler and more transparent than the original, was found by Ruzsa [30]. Ruzsa's argument implies the estimate $c_{11} \leq (2s)^{\exp(\sigma^c)}$, which is better than (1.4) (here c is an absolute constant). In the final section we briefly review the main points of Ruzsa's proof. A detailed self-contained exposition of Ruzsa's proof is given in [26, Chapter 8]

Our exposition is based on the same principles as Freiman's original proof [12, 13], though the technical details are different. The most substantial innovations are in Subsection 5.1, where we suggest a simpler proof of the Cube Lemma, and in Subsection 8.3, where we apply the Bombieri–Vaaler theorem instead of Freiman's sophisticated elementary argument. We believe that the original argument of Freiman is still of great interest, even after Ruzsa's work.

We tried to make the exposition self-contained. Only three standard results from the Geometry of Numbers, namely, the theorems of Minkowski, Mahler and Bombieri-Vaaler, are quoted without proofs (but with exact references). The other auxiliary facts are provided with complete proofs even if they are available in the literature.

In Section 2 we introduce the notation used throughout the paper. In Sections 3 and 4 we reduce the Main Theorem to certain more technical statements. At the end of Section 4 we give a plan of the remaining part of the article.

Acknowledgments. Gregory Freiman drew my attention to his theorem. Daniel Berend and Henrietta Dickinson read the drafts of the paper at different stages of its preparation and made a number of valuable remarks. Peter Pleasants sent me his unpublished notes on Freiman's theorem and Mel Nathanson put at my disposal a preliminary version of his book [26]. It is a pleasure to thank all of them.

My special gratitude is to Imre Ruzsa, who carefully studied the (pre)final version of this paper. I found his numerous comments and suggestions very useful. Many thanks for the hard job he has done.

The main part of this job was done in Bordeaux and was supported by the *Bourse Chateaubriand du gouvernement français*. I am grateful to Prof. J.-M. Deshouillers, Mrs D. Cooke and Mrs F. Duquesnoy for having done their best to make my work in Bordeaux pleasant and successful.

I must also acknowledge support of IMPA (Rio de Janeiro), Forschungsinstitut für Mathematik (ETH Zürich) and Lise Meitner Fellowship (Austria), during the final stage of my work on this paper.

2. Notation and conventions

For $B, C \subseteq \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ put

$$B \pm C = \{b \pm c \ : \ b \in B, c \in C\}, \quad \alpha B = \{\alpha b \ : \ b \in B\},$$

etc.

A *plane* $\mathcal{L} \subseteq \mathbb{R}^n$ is a set of the form $v + \mathcal{L}'$, where $v \in \mathbb{R}^n$ and $\mathcal{L}'$ is a linear subspace of $\mathbb{R}^n$. By (x, y) we denote the standard inner product in $\mathbb{R}^n$. The Lebesgue measure in $\mathbb{R}^n$ is referred to as *volume* and is denoted by Vol or Vol_n. The standard inner product on $\mathbb{R}^n$ induces an inner product on each subspace, and hence it induces a d-dimensional Lebesgue measure on each d-dimensional plane $\mathcal{L}$. This measure is referred to as *$\mathcal{L}$-volume*, and is denoted by $\mathrm{Vol}_{\mathcal{L}}$, or Vol_d, or simply Vol.

Given a set $S \subset \mathbb{R}^n$, we denote by $\mathcal{L}(S)$ the plain spanned by S. We put $\dim S = \dim \mathcal{L}(S)$, and call it *linear dimension* (or simply *dimension*) of S. The orthogonal complement to the set S is denoted by $S^{\perp}$:

$$S^{\perp} = \{x \in \mathbb{R}^n \ : \ (x,y) = 0 \text{ for all } y \in S\}.$$

Let $\mathcal{L}$ be a subspace of $\mathbb{R}^n$. A *lattice* in $\mathcal{L}$ is a maximal discrete subgroup of $\mathcal{L}$. The $\mathcal{L}$-volume of a fundamental domain of a lattice Γ is denoted by $\Delta(\Gamma)$.

A *convex body* in $\mathcal{L}$ is a bounded convex subset of $\mathcal{L}$ having inner points. A convex body is *symmetric* if it is symmetric with respect to the origin. Given a lattice Γ and a symmetric convex body B in $\mathcal{L}$, we say that B is *Γ-thick* if $\mathcal{L}(B \cap \Gamma) = \mathcal{L}$; in words, if the set $B \cap \Gamma$ generate $\mathcal{L}$ as a vector space.

When $\mathcal{L} = \mathbb{R}^n$ and $\Gamma = \mathbb{Z}^n$, we shall simply say *thick* instead of *$\mathbb{Z}^n$-thick.* Thus, a symmetric convex body $B \subseteq \mathbb{R}^n$ is *thick* if $\dim B \cap \mathbb{Z}^n = n$, where dim is the *linear dimension* defined above.

Let B be a symmetric convex body. The norm associated with B is $\|x\|_B := \inf\{\lambda^{-1} \ : \ \lambda x \in B\}$. Recall the following result of Mahler (see [**5**, Chapter VIII, Corollary of Theorem VII]).

***Lemma 2.1* (Mahler).** — *Let B be a symmetric convex body in $\mathbb{R}^n$. Then there exists a basis $e_1, \dots, e_n$ of $\mathbb{Z}^n$ such that*

$$\begin{array}{rcll} \|e_1\|_B & \leq & \lambda_1, & \\ \|e_i\|_B & \leq & i\lambda_i/2 & (2 \leq i \leq n), \end{array}$$

where $\lambda_1, \dots, \lambda_n$ are the successive minima of B with respect to the lattice $\mathbb{Z}^n$.

(Such a basis will be called a *Mahler basis* of the body B.)

We denote by $\|x\|$ the Euclidean norm of the vector $x = (x_1, \dots, x_r) \in \mathbb{R}^n$, and by $\|x\|_\infty$ its l_∞-norm, i.e.

$$\|x\| = \sqrt{(x,x)} = \sqrt{x_1^2 + \cdots + x_n^2}, \qquad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Finally, given $x \in \mathbb{R}$, we denote by $\lfloor x \rfloor$ (respectively, $\lceil x \rceil$) the maximal integer not exceeding x (respectively, the minimal integer not exceeded by x).

3. A geometric formulation of the Main Theorem

In this section we reformulate the Main Theorem and prove that the new formulation implies the one from the Introduction.

First of all, since K is a finite subset of of a torsion-free abelian group, we may assume that $K \subset \mathbb{Z}^n$ for some natural n.

Further, an F_s-progression may be defined as a set which is F_s-isomorphic to $B \cap \mathbb{Z}^m$, where $B = [0, b_1) \times \cdots \times [0, b_m)$. However, it is more convenient to work with less particular convex bodies than rectangular parallelepipeds. Moreover, since we apply the Geometry of Numbers, it will be preferable to deal with symmetric convex bodies. Therefore we shall assume that $0 \in K$, which does not effect the generality.

Finally, let $\varphi\colon \mathbb{Z}^m \to \mathbb{Z}^n$ be a (group-theoretic) homomorphism. Instead of the condition

$(*)$ φ induces an F_s-isomorphism on the set $B \cap \mathbb{Z}^m$

we prefer a slightly stronger condition

$(**)$ the restriction $\varphi|_{sB\cap\mathbb{Z}^m}$ is one-to-one.

(Actually, $(*)$ and $(**)$ are equivalent if B is the convex hull of its integer points.)

According to the previous paragraphs, we formulate the following theorem.

Theorem 3.1. — *Let K be a finite subset of $\mathbb{Z}^n$ of cardinality $k > k_0(\sigma)$, containing the origin, and satisfying $|K+K| \le \sigma k$. Then for any $T \ge 2$ there exist a positive integer m, a thick symmetric convex body $B \subset \mathbb{R}^m$ and a homomorphism $\varphi\colon \mathbb{Z}^m \to \mathbb{Z}^n$ with the following properties:*

(i) $m \le \lfloor \sigma - 1 \rfloor$;

(ii) $\varphi(B \cap \mathbb{Z}^m) \supseteq K$;

(iii) *the restriction $\varphi|_{TB\cap\mathbb{Z}^m}$ is one-to-one;*

(iv) $\operatorname{Vol} B \le c_{31}(\sigma,T)k$, *where* $c_{31}(\sigma,T) = (c_{13}T)^{\sigma^{25\sigma}c_{13}}$.

Proof of Theorem 1.2 (assuming Theorem 3.1). — If $m = 1$, then $\varphi(B \cap \mathbb{Z})$ is an arithmetical progression of length not exceeding $2c_{31}(\sigma,T)k + 1$, which is less than $c_{11}k$ if $T = 2$.

Now assume that $m \ge 2$. Let $e_1, \dots, e_m$ be a Mahler basis of the body B. Put $\rho_i = \|e_i\|_B$ and define a new norm on $\mathbb{R}^m$:

$$\|x\|_\rho = \max_{1\le i\le m} \rho_i |x_i|,$$

where $x = x_1e_1 + \cdots + x_me_m$. It is a general property of norms in finite dimensional spaces that

$$\|x\|_B \ll \|x\|_\rho \ll \|x\|_B, \tag{3.1}$$

where the implicit constants may *a priori* depend on B. We shall now prove the inequalities (3.1) with constants depending only on the dimension m.

The inequality on the left is easy:

$$\|x\|_B \le |x_1|\|e_1\|_B + \cdots + |x_m|\|e_m\|_B \le m\|x\|_\rho \tag{3.2}$$

The inequality on the right is less trivial. Denote by Δ_i the convex hull of the points $\pm x/\|x\|_B$ and $\pm\rho_j^{-1}e_j$, where $(j \ne i)$. Recall the second inequality of Minkowski [**5**, Chapter VIII, Theorem V]:

$$2^m/m! \le \lambda_1 \cdots \lambda_m \operatorname{Vol} B \le 2^m.$$

Then

$$\operatorname{Vol} B \ge \operatorname{Vol} \Delta_i = \frac{2^m|x_i|\rho_i}{m!\|x\|_B\rho_1\cdots\rho_m} \ge \frac{2^{2m-1}|x_i|\rho_i}{(m!)^2\|x\|_B\lambda_1\cdots\lambda_m} \ge \frac{2^{m-1}}{(m!)^2}\frac{|x_i|\rho_i}{\|x\|_B}\operatorname{Vol} B,$$

whence $|x_i|\rho_i \leq c_{32}(m)\|x\|_B$, where $c_{32}(m) = 2^{1-m}(m!)^2$. This proves that

$$\|x\|_\rho \leq c_{32}(m)\|x\|_B. \tag{3.3}$$

Now put $R = \{x \in \mathbb{R}^m : \|x\|_\rho \leq c_{32}\}$. Then inequalities (3.2) and (3.3) may be rewritten as $B \subseteq R \subseteq mc_{32}B$. Therefore $\varphi(R) \supseteq \varphi(B) \supseteq K$. Further, put $T = smc_{32}$. Then the restriction $\varphi|_{sR\cap\mathbb{Z}^m}$ is one-to-one, whence $P = \varphi(R \cap \mathbb{Z}^m)$ is an F_s-progression.

It remains to estimate the cardinality $|P|$. Since B is thick, we have $\lambda_1 \leq \cdots \leq \lambda_m \leq 1$. Therefore for $1 \leq i \leq m$ we have $\rho_i \leq m/2$ (recall that $m \geq 2$), whence $c_{32}\rho_i^{-1} \geq 1$. Hence

$$|P| = |R \cap \mathbb{Z}^m| = \prod_{i=1}^{m}(2\lfloor c_{32}\rho_i^{-1}\rfloor + 1) \leq \frac{(3c_{32})^m}{\rho_1\cdots\rho_m}. \tag{3.4}$$

Now let $\rho_{i_1} \leq \cdots \leq \rho_{i_m}$ be the rearrangement of $\rho_1, \ldots, \rho_m$ in increasing order. Then, by the definition of successive minima,

$$\rho_{i_1} \geq \lambda_1, \ldots, \rho_{i_m} \geq \lambda_m,$$

whence

$$\rho_1\cdots\rho_m \geq \lambda_1\cdots\lambda_m \geq \frac{2^m}{m!}\frac{1}{\operatorname{Vol} B}.$$

Combining this with (3.4), we obtain finally

$$|P| \leq m!\left(\tfrac{3}{2}c_{32}\right)^m \operatorname{Vol} B \leq c_{33}k$$

with $c_{33} = m!\left(\frac{3}{2}c_{32}\right)^m c_{31}(\sigma, smc_{32}) \leq c_{11}(\sigma, s)$. Thus, the Main Theorem follows from Theorem 3.1. □

4. Iteration step and partial covering

Let K, σ and T be as in Theorem 3.1. We shall deal with triples (m, B, φ), where m is a positive integer, $B \subset \mathbb{R}^m$ is a thick symmetric convex body, and $\varphi : \mathbb{Z}^m \to \mathbb{Z}^n$ is a group homomorphism. *Everywhere in this paper the word "triple" will refer to a triple defined as above, unless the contrary is stated explicitly.* We have to prove that there exists a triple (m, B, φ) satisfying the conditions (i)–(iv) of Theorem 3.1. We construct such a triple iteratively. Namely, we prove

***Proposition 4.1* (the base of iteration).** — *There exist triples (m, B, φ) satisfying the conditions (i)–(iii) of Theorem 3.1.*

(such triples will be called *T-admissible*) and

***Proposition 4.2* (the iteration step).** — *For any T-admissible triple (m, B, φ) there exists another T-admissible triple (m', B', φ') with*

$$\operatorname{Vol} B' \leq c_{41}(\sigma, T)\operatorname{Vol} B\,(k/\operatorname{Vol} B)^{1/c_{42}(\sigma)}. \tag{4.1}$$

Here $c_{41} = (c_{13}T)^{\sigma^{20\sigma}c_{13}}$ and $c_{42} = 20\sigma\log(2\sigma)$.

Proof of Theorem 3.1 (assuming Propositions 4.1 and 4.2). — Put

$$V_0 = \inf\{\operatorname{Vol} B \ : \ (m, B, \varphi) \text{ is } T\text{-admissible}\}.$$

Then there exists a T-admissible triple with $\operatorname{Vol} B \le 2V_0$. By Proposition 4.2

$$c_{41}(\sigma, T)\,(k/\operatorname{Vol} B)^{1/c_{42}(\sigma)} \ge 1/2,$$

whence $\operatorname{Vol} B \le (2c_{41})^{c_{42}}\, k \le c_{31}k$. □

Proposition 4.1 is a consequence of the following important lemma of Freiman [**12**, Lemma 1.14].

***Lemma 4.3* (Freiman).** — *Let $K \subset \mathbb{R}^n$ and* $\dim K = m$. *Then*

$$|K + K| \ge (m+1)k - m(m+1)/2. \tag{4.2}$$

Proof. — Clearly, $k \ge m+1$ and (4.2) is true when $m = 1$ or $k = m+1$. Now fix a pair (m, k) and suppose that (4.2) holds for all pairs (m', k') with $m' < m$ or $m' = m$, $k' < k$. We have to prove (4.2) for the pair (m, k).

Let x be a vertex of the convex polytope spanned by the set K and set $K' = K \setminus x$. There are two possibilities: $\dim K' = m - 1$ and $\dim K' = m$.

In the first case

$$\begin{aligned} |K+K| &= |K'+K'| + |K'+x| + 1 \\ &\ge m(k-1) - m(m-1)/2 + k \\ &= (m+1)k - m(m+1)/2. \end{aligned}$$

In the second case let Π be the convex polytope spanned by K'. There is an $(m-1)$-dimensional face of Π with the following property: if $\mathcal{L}$ is the plain containing this face then x and Π lie in distinct half-spaces with the common boundary $\mathcal{L}$. Since $\dim K' \cap \mathcal{L} = m - 1$, we have $|K' \cap \mathcal{L}| \ge m$. Then

$$\begin{aligned} |K+K| &\ge |K'+K'| + |K' \cap \mathcal{L} + x| + 1 \\ &\ge (m+1)(k-1) - m(m+1)/2 + m + 1 \\ &= (m+1)k - m(m+1)/2. \end{aligned}$$

□

Remark 4.4. — Ruzsa [**31**] obtained an analogue of this result for the sum of two distinct sets: *if* $\dim(K+L) = m$ *and* $|K| \ge |L|$ *then* $|K+L| \ge |K| + m|L| - m(m+1)/2$. The case $L = -K$ was treated earlier in [**14**]. See also [**34**].

Proof of Proposition 4.1. — Without loss of generality $\dim K = n$. Then, since $k > k_0(\sigma)$, Lemma 4.3 implies that $n \le \lfloor \sigma - 1 \rfloor$. We conclude the proof, putting $m = n$, letting B be any thick symmetric convex body, containing K, and letting φ be the identical map. □

The proof of Proposition 4.2 is much more complicated. The main difficulties are concentrated in the following *Lemma on Partial Covering.*

***Lemma 4.5* (Partial Covering).** — *Let (m, B, φ) be a 2-admissible triple. Then there exist a subset $K_0 \subset K$ and a triple (m_0, B_0, φ_0) satisfying*

$$|K_0| \geq k/c_{44}(\sigma), \tag{4.3}$$

$$\operatorname{Vol} B_0 \leq c_{45}(\sigma) \operatorname{Vol} B \,(k/\operatorname{Vol} B)^{1/c_{42}(\sigma)}, \tag{4.4}$$

and having the following properties.

(i)′ $m_0 \leq c_{46}(\sigma)$;

(ii)′ $\varphi_0(B_0 \cap \mathbb{Z}^m) \supseteq K_0 - K_0$.

Here $c_{44} = 2^{9\sigma \log(2\sigma)} c_{13}$, $c_{45} = \exp(33\sigma \log^2(2\sigma))$, $c_{46} = 9\sigma \log(2\sigma)$.

(Intuitively, the triple (m_0, B_0, φ_0) is "not very far" from being admissible).

Note that the statement of the Lemma on Partial Covering does not depend on the parameter T. The dependence on T appears only in Section 9, where we deduce Proposition 4.2 from Lemma 4.5. The deduction involves some computations, but is in fact more or less straightforward. However, the Lemma on Partial Covering itself is a non-trivial combination of several very non-trivial facts. The most important and difficult of the latter is Freiman's 2^n-*theorem*, proved in Section 5. In Section 6 we establish several auxiliary facts, to be used in the proof of the Lemma on Partial Covering. The complete proof of Lemma 4.5 is given in Sections 7-8.

5. Freiman's 2^n-theorem

Lemma 4.3 yields that $|S+S| \geq (n+1-\varepsilon)|S|$ for a sufficiently large n-dimensional set S. However, for such "typical" n-dimensional sets as the set of integer points inside a large cube or ball, one has a stronger inequality $|S+S| \geq 2^n|S|$. In the general case Freiman [**12**, Lemma 2.12] obtained the following result.

***Theorem 5.1* (Freiman).** — *Let S be a finite subset of $\mathbb{R}^n$. Assume that $|S+S| \leq (2^n - \varepsilon)|S|$ for some $\varepsilon > 0$. Then there exists an $(n-1)$-dimensional plane $\mathcal{L}$ such that $|S \cap \mathcal{L}| \geq \delta|S|$, where the positive constant δ depends only on n and ε.*

We apply this remarkable theorem twice. First, in Subsection 5.5 we deduce Theorem 1.3 from Theorem 1.2. Second, the 2^n-theorem plays the key role in the proof of the Lemma on Partial Covering, see Subsection 7.2. (In both cases, instead of Theorem 5.1, we apply a slightly more general Theorem 5.6.)

The presented proof is divided into two steps. First we prove an auxiliary assertion, having some independent interest. We call it *Cube Lemma.* In the second step, which is much simpler, we deduce Theorem 5.1 from the Cube Lemma.

Both steps go back to Freiman's original proof, though they are not specified there explicitly. Our proofs are simpler than Freiman's original, but based on the same ideas.

For another (very long) proof of the 2^n-theorem see [**25**] and [**26**, Chapter 8]. Fishburn [**9**] and Stanchescu [**33**] found new proofs for the case $n = 2$, which give (in this case) better quantitative estimates for δ. Unfortunately, neither Fishburn's nor Stanchescu's argument extends to $n \geq 3$.

5.1. The Cube Lemma. — First we introduce some concepts. An *r-cube* in $\mathbb{R}^n$ is the set

$$\mathcal{C} = \mathcal{C}(b; a_1, \ldots, a_r) = \{b(t) := b + t_1a_1 + \cdots t_ra_r \; : \; t = (t_1, \ldots, t_r) \in [-1; 1]^r\}.$$

Here $b, a_1, \ldots, a_r \in \mathbb{R}^n$ (we do not assume $a_1, \ldots, a_r$ linearly independent). The point b is the *center* of the r-cube $\mathcal{C}$, and the set $V(\mathcal{C}) := \{b(\alpha) \; : \; \alpha \in \{-1; 1\}^r\}$ is *the set of vertices* of $\mathcal{C}$.

***Lemma 5.2* (the Cube Lemma).** — *Let S be a finite subset of $\mathbb{R}^n$ and assume that*

$$|S + S| \leq \tau|S|. \tag{5.1}$$

Put $\delta_1 = \delta_1(n, \tau) = (3\tau)^{-2^n}$. Then there exists an n-cube $\mathcal{C}$ with $V(\mathcal{C}) \subset S$ such that $|\mathcal{C} \cap S| \geq \delta_1|S|$.

It turns out to be more convenient to deal with sets symmetric with respect to a point $b \in \mathbb{R}^n$ (that is, for any $u \in S$ there exist $v \in S$ such that $u + v = 2b$).

Proposition 5.3. — *Let S be a finite subset of $\mathbb{R}^n$ satisfying (5.1). Then there is a subset $S_1 \subset S$ of cardinality $|S_1| \geq |S|/\tau$, symmetric with respect to some $b_1 \in \mathbb{R}^n$.*

Proof. — For any $b \in \mathbb{R}^n$ put $S_b = \{u \in S \; : \; 2b - u \in S\}$. By (5.1), there exist at most $\tau|S|$ non-empty sets S_b. Since any $u \in S$ belongs to exactly $|S|$ sets S_b, we have $\sum |S_b| = |S|^2$. Therefore there exists a set S_b of cardinality at least $|S|^2/\tau|S| = |S|/\tau$. □

The Cube Lemma is an easy consequence of Proposition 5.3 and the following assertion.

Proposition 5.4. — *Let S be a finite subset of $\mathbb{R}^n$, symmetric with respect to $b \in \mathbb{R}^n$. Let also $\mathcal{L}$ be a subspace of $\mathbb{R}^n$ of dimension $n - r$, where $1 \leq r \leq n$. Then there exists an r-cube $\mathcal{C}$ with $V(\mathcal{C}) \subset S$, with center in b and such that $|(\mathcal{C} + \mathcal{L}) \cap S| \geq \delta_2|S|$, where $\delta_2 = \delta_2(r, \tau) = (9\tau)^{-2^{r-1}+1}$.*

Proof. — We use induction in r. Assume first that $r = 1$. For $x \in \mathbb{R}^n$ denote by $\rho(x)$ the (Euclidean) distance from the point $x \in \mathbb{R}^n$ to the plane $b + \mathcal{L}$. Let $b_1 \in S$ satisfy

$$\rho(b_1) = \max_{x \in S} \rho(x).$$

Put $a_1 = b_1 - b$. Then for the 1-cube $\mathcal{C} = \mathcal{C}(b; a_1)$ we have $|(\mathcal{C} + \mathcal{L}) \cap S| = |S| = \delta_2(1, \tau)|S|$.

Now assume that $2 \leq r \leq n$. The argument splits into two cases, depending on how many points from S belong to the plane $b + \mathcal{L}$.

Case 1: $|(b + \mathcal{L}) \cap S| \geq \frac{1}{3}|S|$. — Let a be any element of the set $(b + \mathcal{L}) \cap S$. Then the r-cube $\mathcal{C}(b, a, \ldots, a)$ is as desired, because $1/3 \geq \delta_2(r, \tau)$.

Case 2: $|(b+\mathcal{L})\cap S| \le \frac{1}{3}|S|$. — There exists a subspace $\mathcal{L}'$ of dimension $n-1$ such that $\mathcal{L}\subset\mathcal{L}'$ and

$$(b+\mathcal{L}')\cap S = (b+\mathcal{L})\cap S.$$

At least one of the two open half-spaces with boundary $b+\mathcal{L}'$ contains a subset $S'\subset S$ of cardinality $|S'|\ge \frac{1}{3}|S|$. The set S' need not be symmetric. But since

$$|S'+S'| \le |S+S| \le \tau|S| \le 3\tau|S'|, \tag{5.2}$$

the set S' contains a symmetric subset S_1 of cardinality $|S_1|\ge|S'|/3\tau\ge|S|/9\tau$. As in (5.2), we obtain

$$|S_1+S_1| \le 9\tau^2|S_1|. \tag{5.3}$$

Let b_1 be the center of symmetry of the set S_1. By our construction, $a_1 := b_1-b\notin\mathcal{L}'$, in particular $a_1\notin\mathcal{L}$. Therefore the subspace $\mathcal{L}_1$, generated by $\mathcal{L}$ and a_1, is of dimension $n-r+1$. By induction, there is an $(r-1)$-cube $\mathcal{C}_1$ with center b_1 such that $V(\mathcal{C}_1)\subset S_1$ and

$$|(\mathcal{C}_1+\mathcal{L}_1)\cap S_1| \ge \delta_2(r-1,9\tau^2)|S_1| \ge \delta_2(r,\tau)|S|. \tag{5.4}$$

Write $\mathcal{C}_1=\mathcal{C}(b_1,a_2,\dots,a_r)$ and put $\mathcal{C}=\mathcal{C}(b,a_1,\dots,a_r)$. Each vertex of the cube $\mathcal{C}$ is either a vertex of $\mathcal{C}_1$ or is symmetric to a vertex of $\mathcal{C}_1$ with respect to b. Therefore $V(\mathcal{C})\subset S$. We shall prove that

$$|(\mathcal{C}+\mathcal{L})\cap S| \ge |(\mathcal{C}_1+\mathcal{L}_1)\cap S_1|. \tag{5.5}$$

Together with (5.4) this will complete the proof.

Let u belong to the set $\widehat{S}_1 := (\mathcal{C}_1+\mathcal{L}_1)\cap S_1$. Then the point $v=2b_1-u$ also belongs to $\widehat{S}_1$. We shall see that

(∗) *at least one of the points u, v belongs to the set $\widehat{S} := (\mathcal{C}+\mathcal{L})\cap S$.*

Assume (∗) to be true and consider the map

$$\begin{aligned} \widehat{S}_1 &\to \widehat{S},\\ u &\mapsto \begin{cases} u, & \text{if } u\in\widehat{S},\\ 2b-v, & \text{if } v=2b_1-u\in\widehat{S} \text{ and } u\notin\widehat{S}. \end{cases}\end{aligned}$$

This map is one-to-one[2], whence $|\widehat{S}|\ge|\widehat{S}_1|$, as desired. Thus, it remains to prove the assertion (∗).

So, let u belong to $\widehat{S}_1$ and put $v=2b_1-u$. Then $u=u_1+ta_1+y$ and $v=v_1-ta_1-y$, where $u_1,v_1\in\mathcal{C}_1$ are such that $u_1+v_1=2b_1$, and $y\in\mathcal{L}$. Recall (this is crucial) that, by our construction, the $(r-1)$-cube $\mathcal{C}_1$ and the set $\widehat{S}_1$ belong to the same open half-space with the boundary $b+\mathcal{L}'$. In particular, the points u, v, u_1, v_1 belong to this half-space.

[2] Indeed, let u and u' be two distinct elements of $\widehat{S}_1$. If both are in $\widehat{S}$ or both are not in $\widehat{S}$ then their images are obviously distinct. If one of them belongs to $\widehat{S}$ but the other not, then the images lie in the distinct half-spaces with boundary $b+\mathcal{L}'$.

Since $a_1 \notin \mathcal{L}'$, there exists a linear functional $f\colon \mathbb{R}^n \to \mathbb{R}$ vanishing on $\mathcal{L}'$ and positive at a_1. Then the open half-space mentioned above is defined by the inequality $f(x) > f(b)$. Hence

$$
\begin{aligned}
&(5.6) & f(u) = f(u_1) + tf(a_1) &> f(b),\\
&(5.7) & f(v) = f(v_1) - tf(a_1) &> f(b),\\
&(5.8) & f(u_1) &> f(b),\\
&(5.9) & f(v_1) &> f(b).
\end{aligned}
$$

Since $u_1 + v_1 = 2b + 2a_1$, the latter two inequalities imply that

$$f(u_1), f(v_1) < f(b) + 2f(a_1).$$

Then (5.6) and (5.7) yield that $-2 < t < 2$.

Obviously, $\mathcal{C} = \{x - \theta a_1 \;:\; x \in \mathcal{C}_1, 0 \le \theta \le 2\}$. Therefore $u \in \mathcal{C} + \mathcal{L}$ if $-2 < t \le 0$ and $v \in \mathcal{C} + \mathcal{L}$ if $0 \le t < 2$. The assertion $(*)$ is proved, which completes the proof of Proposition 5.4. □

Proof of Lemma 5.2. — The case $r = n$ of Proposition 5.4 is exactly the assertion of the Cube Lemma for symmetric sets, $\delta_1(n,\tau)$ being replaced by $\delta_2(n,\tau)$. To establish the Cube Lemma for arbitrary sets, apply Proposition 5.4 to the symmetric set S_1 from Proposition 5.3. As in (5.2), we obtain $|S_1 + S_1| \le \tau^2|S_1|$. Since $\delta_2(n,\tau^2)/\tau \ge \delta_1(n,\tau)$, Lemma 5.2 follows. □

5.2. Proof of the 2^n-theorem. — Now we are ready to prove Theorem 5.1. For a positive real number δ put $\varepsilon(\delta,n) = 2^n(4n\delta/\delta_1)^\nu$, where $\nu = \delta_1/10n$ and $\delta_1 = \delta_1(n,2^n)$ is defined in Lemma 5.2.

Let S be a finite subset of $\mathbb{R}^n$ with at most $\delta|S|$ points on every hyperplane. We shall prove that $|S + S| \ge (2^n - \varepsilon)|S|$, where $\varepsilon = \varepsilon(\delta,n)$. Since this is trivial when $\delta \ge \delta_1/4n$, we shall assume that $\delta \le \delta_1/4n$.

Let $\mathcal{C}$ be the n-cube constructed in Lemma 5.2 (where we put $\tau = 2^n$). Since $|\mathcal{C} \cap S| \ge \delta_1|S| > \delta|S|$, we have $\dim \mathcal{C} = n$; in particular, the interior $\mathcal{C}^\circ$ is non-empty. Moreover, since the boundary of $\mathcal{C}$ is contained in a union of $2n$ hyperplanes, we have for the set $S_0 := S \cap \mathcal{C}^\circ$ the estimate

$$|S_0| \ge |S \cap \mathcal{C}| - 2n\delta|S| \ge (\delta_1 - 2n\delta)|S| \ge (\delta_1/2)|S|. \tag{5.10}$$

The $2n$ hyperplanes defined by the faces of the cube $\mathcal{C}$ divide $\mathbb{R}^n \setminus \mathcal{C}^\circ$ into $p := 3^n - 1$ disjoint convex sets. This divides the set $S \setminus S_0$ into p subsets $S_1, \dots, S_p$. We have

$$(S_i + S_i) \cap (S_j + S_j) = \varnothing \qquad (0 \le i < j \le p), \tag{5.11}$$

because S_i and S_j are subsets of disjoint convex sets, and by the same reason

$$(S_0 + V) \cap (S_i + S_i) = \varnothing \qquad (1 \le i \le p), \tag{5.12}$$

where $V = V(\mathcal{C})$ is the set of the vertices of the cube $\mathcal{C}$. (Recall that $V \subseteq S$ by the definition of the cube $\mathcal{C}$.) Further,

$$|S_0 + V| = |V||S_0| = 2^n|S_0|, \tag{5.13}$$

because all the sums $x+v$, where $x \in \mathcal{C}^\circ$ and $v \in V$, are distinct. Also, since $|S_i| < |S|$ for $i = 1, \dots, p$, we have by induction

$$|S_i + S_i| \ge (2^n - \varepsilon_i)|S_j| \quad (1 \le i \le p), \tag{5.14}$$

where $\varepsilon_i = \varepsilon(\delta/\eta_i, n)$ and $\eta_i = |S_i|/|S|$. Now

$$|S+S| \ge |S_0 + V| + \sum_{i=1}^{p} |S_i + S_i| \ge 2^n|S_0| + \sum_{i=1}^{p}(2^n - \varepsilon_i)|S_i| = \left(2^n - \sum_{i=1}^{p} \varepsilon_i\eta_i\right)|S|,$$

and it remains to observe that

$$\sum_{i=1}^{p} \varepsilon_i\eta_i = \varepsilon\sum_{i=1}^{p} \eta_i^{1-\nu} \le \varepsilon p\left(\frac{1}{p}\sum_{i=1}^{p}\eta_i\right)^{1-\nu} = \varepsilon p^{\nu}\left(1 - \frac{|S_0|}{|S|}\right)^{1-\nu} \le \varepsilon e^{\frac{\delta_1}{4}}\left(1 - \frac{\delta_1}{2}\right)^{\frac{1}{2}} \le \varepsilon.$$

(It is a trivial exercise in calculus to show that the function $e^{x/4}(1-x/2)^{1/2}$ decreases on the interval $[0,1]$.) Theorem 5.1 is proved. □

Remark 5.5. — The argument of this section is a version of Freiman's original, with some modifications due to Ruzsa. Ruzsa also noticed that $\mathbb{R}^n \setminus \mathcal{C}^\circ$ can be divided into $2n$ rather than $3^n - 1$ parts, but this does not affect much the final result.

5.3. The function $fr(n,\varepsilon)$. — Put

$$fr(n,\varepsilon) = \sup_S \min_{\mathcal{L}} \frac{|S|}{|S \cap \mathcal{L}|},$$

where S runs over the finite subsets of $\mathbb{R}^n$ satisfying (5.1) and $\mathcal{L}$ runs over the hyperplanes of $\mathbb{R}^n$. Then $fr(n,\varepsilon) \le \delta^{-1}$, where δ is from Theorem 5.1. A calculation shows that

$$fr(n,\varepsilon) \le \left(2 + \varepsilon^{-1}\right)^{\exp\exp n}. \tag{5.15}$$

It would be nice to improve against this extremely weak estimate. Such an improvement would have been possible if Proposition 5.3 were replaced by the following assertion:

Given a finite set $S \subset \mathbb{R}^n$, there exists a symmetric subset $S_1 \subseteq S$ satisfying $|S_1| \ge \tau^{-\alpha}|S|$ and $|S_1 + S_1| \ll \tau|S_1|$. Here α is an absolute constant, and the implied constant is also absolute.

However, Don Coppersmith (private communication) and Imre Ruzsa (private communication) had independently disproved this assertion by similar probabilistic arguments. Moreover, the τ^2-term in (5.3) cannot be replaced even by $\tau^{2-\varepsilon}$, let alone $O(\tau)$. Therefore the estimate (5.15) is probably best possible for the method.

Note in conclusion that Freiman's original argument yields only exponential dependence of $fr(n,\varepsilon)$ in ε^{-1} (when n is fixed). Polynomial dependence in ε^{-1} was achieved due to a suggestion of Ruzsa concerning the argument of Subsection 5.2.

5.4. A generalized 2^n-theorem. — Actually, we need the following simple generalization of Theorem 5.1.

Theorem 5.6. — *Let $1 \le r \le n$. Assume that S satisfies (5.1) with $\tau \le 2^r - \varepsilon$. Put $\delta = (fr(r,\varepsilon))^{-1}$. Then there exists a plane $\mathcal{L} \subset \mathbb{R}^n$ of dimension $\dim \mathcal{L} \le r-1$ such that $|S \cap \mathcal{L}| \ge \delta |S|$.*

Proof. — For any set $T \subseteq \mathbb{R}^n$ denote by $\mathcal{L}_0(T)$ the subspace of the same dimension, parallel to the plane $\mathcal{L}(T)$. We say that the subspace $\Lambda \subseteq \mathbb{R}^n$ of dimension $n-r$ is *generic* if

$$\dim(\Lambda \cap \mathcal{L}_0(S_1)) = \max(0, \dim \mathcal{L}_0(S_1) - r)$$

for any $S_1 \subseteq (S-S)$. Clearly, generic subspaces exist.

Let Λ be a generic subspace and let $\mathbb{R}^n = \Lambda \oplus \mathrm{M}$. Denote by $\pi\colon \mathbb{R}^n \to \mathrm{M}$ the projection along Λ. For any distinct $u, v \in S$ we have $\pi(u) \ne \pi(v)$, because Λ is generic. Hence the finite set $\pi(S) \subset \mathrm{M}$ satisfies

$$|\pi(S) + \pi(S)| = |\pi(S+S)| \le |S+S| \le \tau|S| = \tau|\pi(S)|.$$

Since $\dim \mathrm{M} = r$, we may use Theorem 5.1. Hence for some plane $\mathcal{L}' \subset \mathrm{M}$ of dimension $r-1$ we have $|\mathcal{L}' \cap \pi(S)| \ge \delta|\pi(S)| = \delta|S|$. Put $S_1 = (\mathcal{L}' + \Lambda) \cap S$ and $\mathcal{L} = \mathcal{L}(S_1)$. Then $|S \cap \mathcal{L}| \ge |S_1| \ge \delta|S|$. Since both the subspaces Λ and $\mathcal{L}_0(S_1)$ are parallel to the plane $\mathcal{L}' + \Lambda$ of dimension $n-1$, we have

$$\dim(\Lambda \cap \mathcal{L}_0(S_1)) \ge \dim \mathcal{L}_0(S_1) - r + 1.$$

This is possible only when $\dim \mathcal{L} = \dim \mathcal{L}_0(S_1) \le r-1$. □

5.5. Proof of Theorem 1.3 (assuming Theorem 1.2). — As the first application of the 2^n-theorem, we show that Theorem 1.3 follows from Theorem 1.2. This is an immediate consequence of the following assertion.

Proposition 5.7. — *Let $P = P(x_0; x_1, \dots, x_m; b_1, \dots, b_m)$ be an F_2-progression with $b_1 \ge \dots \ge b_m$ and let K be a subset of P. Assume that $|P| \le \alpha k$ and*

$$|K+K| \le (2^r - \varepsilon)k, \tag{5.16}$$

where $k = |K|$. Then

$$b_i \le 2\alpha fr(r,\varepsilon) \qquad (i \ge r). \tag{5.17}$$

Proof. — Denote by φ the map (1.1). Since φ is an F_2-isomorphism, the set $K' = \varphi^{-1}(K)$ also satisfies (5.16). Put $\delta = (fr(r,\varepsilon))^{-1}$. By Theorem 5.6, there exists a plane $\mathcal{L} \subset \mathbb{R}^m$ of dimension at most $r-1$ such that $|K' \cap \mathcal{L}| \ge \delta k$.

Let now $e_1 = (1, 0, \dots, 0), \dots, e_m = (0, \dots, 0, 1)$ be the standard basis of $\mathbb{Z}^m$. Since $\dim \mathcal{L} \le r-1$, there is an index $j \le r$ such that the vector e_j is not parallel to the plane $\mathcal{L}$. Then the sets

$$\mu e_j + (K' \cap \mathcal{L}) \qquad (0 \le \mu \le b_i - 1) \tag{5.18}$$

are pairwise disjoint. On the other hand, all the sets (5.18) are contained in the progression $P' = P(0; e_1, \ldots, e_m; b_1, \ldots, b_{j-1}, 2b_j, b_{j+1}, b_m)$. Therefore

$$2\alpha k \geq 2|P| = |P'| \geq \sum_{\mu=0}^{b_j-1} |\mu e_j + (K' \cap \mathcal{L})| \geq b_j \delta k,$$

whence $b_j \leq 2\alpha\delta^{-1}$. Since $j \leq r$ and $b_1 \geq \cdots \geq b_m$, we obtain (5.17). □

6. Some lemmas

In this section we prove some auxiliary facts, which, together with Theorem 5.6, will be used the proof of the Lemma on Partial Covering.

Lemma 6.1. — *Let $\gamma_1, \ldots, \gamma_k$ be real numbers, and for any $\beta \in \mathbb{R}$ let $k(\beta) = k(\beta; \gamma_1, \ldots, \gamma_k)$ be the number of indices j satisfying $0 \leq \gamma_j - \beta < 1/2 \pmod 1$. Assume that*

$$\left| \sum_{j=1}^{k} e^{2\pi i \gamma_j} \right| \geq \delta k$$

Then $k(\beta) \geq (1+\delta)k/2$ for some $\beta \in [0,1)$.

Remark 6.2. — This result is due to Freiman [**11**]. A simpler proof was suggested by Postnikova [**27**] and reproduced in [**12**, Lemma 2.2]. We follow this argument with slight modifications.

Proof. — Since $k(\beta)$ is periodic with period 1, it is sufficient to find $\beta \in \mathbb{R}$ with the required property. Also, $k(\beta; \gamma_1, \ldots, \gamma_k) = k(\beta + \gamma; \gamma_1 + \gamma, \ldots, \gamma_k + \gamma)$ for any real γ. Therefore, replacing each γ_i by $\gamma_i + \gamma$, with a suitable $\gamma \in \mathbb{R}$, we may assume that

$$\left| \sum_{j=1}^{k} e^{2\pi i \gamma_j} \right| = \sum_{j=1}^{k} e^{2\pi i \gamma_j} = \sum_{j=1}^{k} \cos 2\pi\gamma_j. \tag{6.1}$$

For $0 \leq x \leq 1$ let $F(x)$ be the number of indices j such that $0 \leq \gamma_j < x \pmod 1$. Then for $0 \leq \beta \leq 1/2$ we have $k(\beta) = F(\beta + 1/2) - F(\beta)$.

Assume that $k(\beta) < (1+\delta)k/2$ for all $\beta \in [0,1)$. Then $k(\beta) > \frac{1}{2}(1-\delta)k$ for all $\beta \in [0,1)$. Estimate now the last sum in (6.1):

$$\begin{aligned} \sum_{j=1}^{k} \cos 2\pi\gamma_j = \int_0^1 \cos 2\pi x dF(x) &= F(x)\cos 2\pi x|_0^1 + 2\pi \int_0^1 F(x) \sin 2\pi x dx \\ &= k + 2\pi \int_0^1 F(x) \sin 2\pi x dx. \end{aligned} \tag{6.2}$$

For the last integral we have

$$\begin{aligned}\int_0^1 F(x)\sin 2\pi x dx &= -\int_0^{1/2}(F(x+1/2)-F(x))\sin 2\pi x dx\\ &= -\int_0^{1/2} k(x)\sin 2\pi x dx\\ &< -\frac{1}{2}(1-\delta)k\int_0^{1/2}\sin 2\pi x dx\\ &= -(2\pi)^{-1}(1-\delta)k.\end{aligned}$$

Substituting this into (6.2), we obtain $\sum\limits_{j=1}^{k}\cos 2\pi\gamma_j < \delta k$, a contradiction. □

Lemma 6.3. — *Let K be a finite set of k elements with $K_1,\ldots,K_r \subset K$ satisfying*

$$|K_i| \geq (1+\delta)k/2 \qquad (1\leq i\leq r),$$

where $0<\delta<1/2$. For $\alpha=(\alpha_1,\ldots,\alpha_r)\in\{0,1\}^r$ put $S_\alpha = \bigcap_{i=1}^r K_i^{\alpha_i}$, where $K_i^1 = K_i$ and $K_i^0 = K\backslash K_i$. Then there exists $\alpha\in\{0,1\}^r$ such that

$$|S_\alpha| \geq (\gamma/2)^r k, \tag{6.3}$$

where $\gamma = (1+\delta)^{(1+\delta)/2}(1-\delta)^{(1-\delta)/2}$.

Remark 6.4. — Note that $\gamma > 1$, and, moreover,

$$\gamma = \exp\left(\sum_{i=1}^{\infty}\frac{\delta^{2i}}{2i(2i-1)}\right) \geq e^{\delta^2/2}.$$

This lemma is also due to Freiman [F1, Lemma 2.11]. He used a probabilistic method, and his result was slightly weaker, with an additional factor $c(\delta)r^{-1/2}$ in the right-hand side of (6.3). The following elegant argument was suggested by Ruzsa (private communication).

Proof. — For $\alpha\in\{0,1\}^r$ write $|\alpha| = \alpha_1+\cdots+\alpha_r$. Notice that

$$\sum|S_\alpha| = k, \quad \sum|\alpha||S_\alpha| = |K_i|+\cdots+|K_r| \geq (1+\delta)kr/2, \tag{6.4}$$

where here and below the summation extends to $\alpha\in\{0,1\}^r$.

Let z be a positive real number, to be specified later. Using (6.4) and the weighted arithmetic and geometric mean inequality(3), we obtain:

$$\sum z^{|\alpha|}|S_\alpha| \geq kz^{(1/k)\sum|\alpha||S_\alpha|} \geq kz^{(1+\delta)r/2}.$$

On the other hand, $\sum z^{|\alpha|} = (1+z)^r$, whence

$$\max|S_\alpha| \geq k\left(z^{(1+\delta)/2}/(1+z)\right)^r.$$

[3] That is, the inequality $a_1b_1+\cdots+a_nb_n \geq a_1^{b_1}\cdots a_n^{b_n}$, where $a_n,\ldots,a_n$ are positive real numbers and $b_1,\ldots,b_n$ non-negative real numbers satisfying $b_1+\cdots+b_n = 1$. It is an immediate consequence of the Jensen inequality for the logarithm.

The optimal choice $z = (1+\delta)/(1-\delta)$ leads to $\max|S_\alpha| \geq k(\gamma/2)^r$. □

In the next two lemmas we state elementary geometric properties of convex bodies.

Lemma 6.5. — *Let $B \subset R^n$ be a convex body. Suppose that its closure $\bar{B}$ contains an n-dimensional ball of radius ρ. Then for any measurable $B_1 \subseteq B$*

$$\mathrm{Vol}_d(B_1) \leq \frac{n!}{d!\rho^{n-d}} \mathrm{Vol}_n(B), \tag{6.5}$$

where $d = \dim(B_1)$.

Proof. — We use induction in $n-d$. When $n-d=0$, the assertion is trivial. Now suppose that $d \leq n-1$. Let Ω be an n-dimensional ball of radius ρ, contained in $\bar{B}$. Then there exists a point $x \in \Omega$ such that the distance between x and $\mathcal{L}(B_1)$ is $\delta \geq \rho$. Put

$$B_2 = \{xt + b(1-t) \ : \ b \in B_1, t \in [0;1]\}.$$

Then $\dim B_2 = d+1$ and by induction

$$\mathrm{Vol}_{d+1}(B_2) \leq \frac{n!}{(d+1)!\rho^{n-d-1}} \mathrm{Vol}_n(B).$$

On the other hand,

$$\mathrm{Vol}_{d+1}(B_2) = \frac{\delta}{d+1} \mathrm{Vol}_d(B_1),$$

which proves (6.5). □

Lemma 6.6. — *Let $w \in \mathbb{R}^n$ be a non-zero vector, $\mathcal{W} = w^\perp$ and $\pi\colon \mathbb{R}^n \to \mathcal{W}$ the orthogonal projection. Then for any symmetric convex body B we have*

$$\mathrm{Vol}_{n-1}(\pi(B)) \leq \frac{n}{2} \frac{\|w\|_B}{\|w\|} \mathrm{Vol}_n(B). \tag{6.6}$$

Proof. — We shall prove the following more general statement.

Let $\mathcal{L}$ be a subspace of $\mathbb{R}^n$ and $\mathcal{W} = \mathcal{L}^\perp$. Denote by $\pi\colon \mathbb{R}^n \to \mathcal{W}$ the orthogonal projection. Then for any symmetric convex body B we have

$$\mathrm{Vol}_m(\pi(B)) \cdot \mathrm{Vol}_l(B \cap \mathcal{L}) \leq \binom{n}{l} \mathrm{Vol}_n(B), \tag{6.7}$$

where $l = \dim \mathcal{L}$, $m = n - l = \dim \mathcal{W}$.

Let $\mathcal{L}$ be the one-dimensional subspace generated by the vector w. Then $\mathrm{Vol}_1(\mathcal{L} \cap B) = 2\|w\|/\|w\|_B$. Therefore inequality (6.6) is the case $l=1$ of inequality (6.7).

Proof of (6.7). — Let S_{m-1} be the unit sphere in $\mathcal{W}$. For any $x \in S_{m-1}$ let $L(x)$ be the $(l+1)$-dimensional half-plane containing x and having $\mathcal{L}$ as the boundary. Put

$$\begin{aligned} r(x) &= \sup\{r > 0 \,:\, rx \in \pi(B)\}, \\ B(x) &= L(x) \cap B, \\ h(x,r) &= \mathrm{Vol}_l(\pi^{-1}(rx) \cap B). \end{aligned}$$

Then

$$\mathrm{Vol}_m(\pi(B)) = \frac{1}{m}\int_{S_{m-1}} r^m(x)dx, \tag{6.8}$$

$$\mathrm{Vol}_n(B) = \int_{S_{m-1}} dx \int_0^{r(x)} r^{m-1}h(x,r)dr. \tag{6.9}$$

Note that $B(x) \supset (B \cap \mathcal{L})$ and $B(x) \cap \pi^{-1}(xr(x)) \neq \varnothing$. Hence

$$h(x,r) \geq \left(\frac{r(x)-r}{r(x)}\right)^l \mathrm{Vol}_l(\mathcal{L} \cap B). \tag{6.10}$$

Combining (6.8)–(6.10) with the well-known equality

$$\int_0^1 t^{m-1}(1-t)^l dt = \frac{(m-1)!l!}{(m+l)!},$$

we obtain (6.7). The lemma is proved. □

Now let B be a symmetric convex body, $X \geq 1$ and $C > 0$.

Definition 6.7. — *The system of vectors $a_1, \ldots, a_r \in \mathbb{R}^n$ is (B,X,C)*-badly approximable *if for any $x \in \mathbb{Z}^n$ and $y = (y_1, \ldots, y_r) \in \mathbb{Z}^r$ satisfying*

$$\|x\|_\infty \leq X, \qquad 0 < \|y\|_\infty \leq X, \tag{6.11}$$

we have

$$\|y_1a_1 + \cdots + y_ra_r - x\|_B \geq C.$$

Lemma 6.8. — *Let $M_1, \ldots, M_r \subseteq \mathbb{R}^n$ be measurable sets, and assume that*

$$\mathrm{Vol}\, M_i > 6^n 3^i X^{n+i} C^n \,\mathrm{Vol}\, B.$$

Then there exists a (B,X,C)-badly approximable system $a_1, \ldots, a_r$ such that $a_1 \in M_1, \ldots, a_r \in M_r$.

Proof. — Use induction in r. Let $r \geq 1$, and suppose that $a_1, \ldots, a_{r-1}$ form a badly approximable system. Estimate the volume of the set

$$M = \{a_r \in \mathbb{R}^n \,:\, a_1, \ldots, a_{r-1}, a_r \text{ is not a badly approximable system}\}.$$

By definition, $a_r \in M$ if and only if there exists x, y satisfying (6.11) and

$$\|y_1a_1 + \cdots + y_{r-1}a_{r-1} + y_ra_r - x\|_B < C. \tag{6.12}$$

Since $a_1, \ldots, a_{r-1}$ is a badly approximable system, we have $y_r \neq 0$. Therefore

$$M = \bigcup_{\substack{\|x\|_\infty, \|y\|_\infty \leq X \\ y_r \neq 0}} M(x,y)$$

where $M(x,y) = \{a_r \in \mathbb{R}^n : \ (6.12) \text{ is true}\}$. We have trivially

$$\operatorname{Vol} M(x,y) = \frac{(2C)^n \operatorname{Vol} B}{|y_r|} \leq (2C)^n \operatorname{Vol} B.$$

whence

$$\begin{aligned} \operatorname{Vol} M &\leq \sum_{\substack{\|x\|_\infty, \|y\|_\infty \leq X \\ y_r \neq 0}} \operatorname{Vol} M(x,y) \\ &\leq (2X+1)^{n+r-1} \cdot 2X \cdot (2C)^n \operatorname{Vol} B \\ &\leq 6^n 3^r X^{n+r} C^n \operatorname{Vol} B \\ &< \operatorname{Vol} M_r. \end{aligned}$$

Therefore we can choose $a_r \in M_r \setminus M$, which proves the lemma. □

For the next lemma we have to define the determinant of the linear map $\varphi \colon \mathcal{L} \to \mathbb{R}^n$, where $\mathcal{L}$ is a subspace of $\mathbb{R}^n$. We put $\det \varphi = 0$ if $\dim \varphi(\mathcal{L}) < \dim \mathcal{L}$. If $\dim \varphi(\mathcal{L}) = \dim \mathcal{L}$, choose orthogonal bases in both $\mathcal{L}$ and $\varphi(\mathcal{L})$ (with respect to the standard inner product in $\mathbb{R}^n$), and let $\det \varphi$ be the determinant of the matrix of φ with respect to these bases (clearly, $\det \varphi$ is independent of the choice of bases).

Lemma 6.9. — *Let $\mathcal{W}$ and $\mathcal{L}$ be proper subspaces of $\mathbb{R}^n$, the subspace $\mathcal{W}$ being of dimension $n-1$. Let w be a non-zero vector orthogonal to $\mathcal{W}$ and l a non-zero vector, orthogonal to $\mathcal{L}$. Denote by $\pi \colon \mathbb{R}^n \to \mathcal{W}$ the orthogonal projection. Then*

$$(6.13) \qquad |\det \pi|_{\mathcal{L}}| \geq \frac{|(w,l)|}{\|w\| \cdot \|l\|}.$$

(Here $\pi|_{\mathcal{L}}$ is the restriction of π on $\mathcal{L}$.)

Proof. — Without loss of generality, $\|l\| = \|w\| = 1$. We may also assume that $\mathcal{L} \not\subset \mathcal{W}$, since otherwise $\pi|_{\mathcal{L}}$ is the identity map, and (6.13) follows from the Cauchy-Schwarz inequality.

Let $e_1, \ldots, e_{d-1}$ be an orthonormal basis of the subspace $\mathcal{L} \cap \mathcal{W}$. Complete it to orthonormal bases $e_1, \ldots, e_{d-1}, e_d$ and $e_1, \ldots, e_{d-1}, e'_d$ of the subspaces $\mathcal{L}$ and $\pi(\mathcal{L})$, respectively. The matrix of the linear map $\pi|_{\mathcal{L}}$ in these bases is

$$\operatorname{diag}\left(1, \ldots, 1, \pm\sqrt{1-(e_d,w)^2}\right)$$

(here the sign of the square root depends on the directions of the vectors e_d and e'_d). Therefore $|\det \pi|_{\mathcal{L}}| = \sqrt{1-(e_d,w)^2}$. But $(e_d,w)^2 + (l,w)^2 \leq \|w\|^2 = 1$ by Bessel's inequality, whence $|\det \pi|_{\mathcal{L}}| \geq |(w,l)|$, as wanted. □

Our final lemma is a well-known result of Bombieri and Vaaler [**3**, Theorem 1].

Lemma 6.10. — *Let $\mathcal{L}$ be a proper subspace of $\mathbb{R}^n$ such that $\Gamma = \mathcal{L} \cap \mathbb{Z}^n$ is a lattice in $\mathcal{L}$. Then there exists a non-zero vector $l \in \mathcal{L}^\perp \cap \mathbb{Z}^n$ such that $\|l\|_\infty \le \Delta(\Gamma)$.*

□

7. Proof of the Lemma on Partial Covering: constructing the triple (m_0, B_0, φ_0)

In this and the next section we prove the Lemma on Partial Covering. Thus, until the end of Section 9 we fix a 2-admissible triple (m, B, φ). If $\operatorname{Vol} B \le c_{45}(\sigma)^{c_{42}(\sigma)}k$, then (4.4) holds with $B_0 = B$, and the assertion of Lemma 4.5 becomes trivial. Therefore we may assume that

$$\operatorname{Vol} B \ge c_{45}(\sigma)^{c_{42}(\sigma)}k \ge \exp(600\sigma^2)k. \tag{7.1}$$

Fix a Mahler basis $e_1, \dots, e_m$ of the body B (see Lemma 2.1). Since B is thick, we have

$$\|e_i\|_B \le \max(1, i/2) \qquad (1 \le i \le m). \tag{7.2}$$

We shall assume that this basis is orthonormal, redefining the inner product if necessary.

Put $K' = \varphi^{-1}(K)$. Since our triple is 2-admissible, the restriction $\varphi|_{K'} : K' \to K$ is an F_2-isomorphism. Therefore

$$|K'| = k, \quad |K' + K'| = |K + K| \le \sigma k.$$

7.1. Freiman's map. — Let r be a positive integer, $a_1, \dots, a_r \in [0,1)^m$ and $b_1, \dots, b_r \in [0,1)$. Define *Freiman's map*

$$\begin{array}{rcl} \Phi\colon \mathbb{Z}^m & \to & \mathbb{Z}^{m+r} \\ x = (x_1, \dots, x_m) & \mapsto & (x_1, \dots, x_m, \lfloor (a_1, x) - b_1 \rfloor, \dots, \lfloor (a_r, x) - b_r \rfloor). \end{array}$$

The map Φ is one-to-one, but it does not induce an F_2-isomorphism $\mathbb{Z}^m \to \Phi(\mathbb{Z}^m)$. However, if for any $\alpha = (\alpha_1, \dots, \alpha_r) \in \{0;1\}^r$ we put

$$\begin{array}{lcl} Z_\alpha & = & \{x \in \mathbb{Z}^m : \alpha_i/2 \le (x, a_i) - b_i < (\alpha_i + 1)/2 \pmod 1 \quad \text{for } 1 \le i \le m\}, \\ (7.3)\ S_\alpha & = & K' \cap Z_\alpha, \end{array}$$

then we obtain the following statement.

Proposition 7.1. — *For any $\alpha \in \{0,1\}^r$ the map $\Phi\colon Z_\alpha \to \Phi(Z_\alpha)$ is an F_2-isomorphism. In particular,*

$$|\Phi(S_\alpha)| = |S_\alpha|, \quad |\Phi(S_\alpha) + \Phi(S_\alpha)| = |S_\alpha + S_\alpha|. \tag{7.4}$$

Proof. — Trivial. □

We put $K'' = \Phi(K')$.

7.2. Distorting vectors. — Fix $\delta > 0$, to be specified later. We say that vector $a \in [0,1)^m$ is *δ-distorting* (or shortly, *distorting*) if

$$\left|\sum_{x\in K'} e^{2\pi i(a,x)}\right| > \delta k.$$

This definition is motivated by Lemma 6.1. Applying this lemma in our situation, we obtain the following assertion.

Proposition 7.2. — *For any δ-distorting vector $a \in [0,1)^m$ there exist $b \in [0,1)$ such that*

$$|\{x \in K' \ : \ 0 \leq (a,x) - b < 1/2 \ (\mathrm{mod} 1)\}| \geq (1+\delta)k/2. \tag{7.5}$$

Return to the construction of Subsection 7.1. We did not yet impose any restrictions on a_i and b_i. Let now all the vectors $a_1,\dots,a_r$ be δ-distorting, and for each a_i let b_i be the b from Proposition 7.2.

Now Lemma 6.3 shows how "small" distortions in (7.5) (where we have $(1+\delta)k/2$ instead of the expected $k/2$) can be combined to obtain "substantial" distortion for one of the sets S_α in (7.3). Applying it, we obtain the following:

Proposition 7.3. — *For any positive integer r there exists $\alpha \in \{0,1\}^r$ such that $|S_\alpha| \geq (\gamma/2)^r k$, where $\gamma = (1+\delta)^{(1+\delta)/2}(1-\delta)^{(1-\delta)/2} \geq e^{\delta^2/2}$.*

Now specify

$$\delta = 1/2\sqrt{\sigma}, \qquad r = \lceil 2\delta^{-2}\log(2\sigma)\rceil = \lceil 8\sigma\log(2\sigma)\rceil.$$

(Our choice of δ will be motivated in Subsection 8.1.) Then $|S_\alpha| \geq 2^{1-r}\sigma k$, whence

$$|S_\alpha + S_\alpha| \leq |K' + K'| \leq \sigma k \leq 2^{r-1}|S_\alpha|,$$

and by (7.4)

$$|\Phi(S_\alpha) + \Phi(S_\alpha)| \leq 2^{r-1}|\Phi(S_\alpha)| \leq (2^r - 1)|\Phi(S_\alpha)|.$$

Now it is the time to apply the 2^n-theorem. By Theorem 5.6, there exists a plane $\mathcal{L} \subset \mathbb{R}^{m+r}$ of dimension $\dim \mathcal{L} \leq r-1$ such that

$$|\mathcal{L} \cap \Phi(S_\alpha)| \geq |\Phi(S_\alpha)|/c_{71} \geq k/c_{72}$$

with $c_{71} = fr(r,1) = c_{13}$ and $c_{72} = 2^r fr(r,1) = 2^r c_{13}$. In particular, putting[4] $K_0'' = K'' \cap \mathcal{L}$ we obtain

$$|K_0''| \geq k/c_{72} \geq k/c_{44}.$$

Without loss of generality

$$\mathcal{L} = \mathcal{L}(K_0''), \tag{7.6}$$

otherwise we can replace $\mathcal{L}$ by the plane $\mathcal{L}(K_0'')$.

[4] Recall that $K'' = \Phi(K')$.

7.3. Constructing the triple (m_0, B_0, φ_0). — Now we are ready to construct the triple (m_0, B_0, φ_0). It will sometimes be notationally convenient to write $\mathbb{R}^m \oplus \mathbb{R}^r$ instead of $\mathbb{R}^{m+r}$ (and $\mathbb{Z}^m \oplus \mathbb{Z}^r$ instead of $\mathbb{Z}^{m+r}$). In these cases, we shall write the elements of $\mathbb{R}^m \oplus \mathbb{R}^r$ as $x \oplus y$, where $x = (x_1, \dots, x_m) \in \mathbb{R}^m$ and $y = (y_1, \dots, y_r) \in \mathbb{R}^r$.

By the definition of the map Φ, the set K'' is contained in the convex body

$$\{x \oplus y \in \mathbb{R}^m \oplus \mathbb{R}^r \ : \ x \in B, \ \ 0 \leq (x, a_i) - b_i - y_i < 1 \ \ (1 \leq i \leq r)\}.$$

Therefore the set $K'' - K''$ is contained in the symmetric convex body

$$\Omega := \{x \oplus y \in \mathbb{R}^m \oplus \mathbb{R}^r \ : \ x \in 2B, \ \ -1 < (x, a_i) - y_i < 1 \ \ (1 \leq i \leq r)\}. \tag{7.7}$$

Proposition 7.4. — *There exist a proper subspace $\mathcal{L}_0$ of $\mathbb{R}^{m+r}$ with the following properties.*

1. *Let K_0'' be the subset of K'' defined at the end of the previous subsection. Then the set $B_0 := \mathcal{L}_0 \cap \Omega$ contains $K_0'' - K_0''$.*
2. *Put $\Gamma_0 = \mathcal{L}_0 \cap \mathbb{Z}^{m+r}$. Then $B_0 \cap \Gamma_0$ generates $\mathcal{L}_0$ as a vector space. In particular, Γ_0 is a lattice in $\mathcal{L}_0$, and B_0 is Γ_0-thick.*
3. *Let $x \oplus y \in \mathbb{R}^m \oplus \mathbb{R}^r$ be a non-zero vector orthogonal to $\mathcal{L}_0$. Then $y \neq 0$.*

Proof. — For every $x \in B$ there exists $y \in \mathbb{Z}^r$ such that $x \oplus y \in \Omega$. (Indeed, we can always find $y_i \in \mathbb{Z}$ satisfying $-1 < (x, a_i) - y_i < 1$.) Since B is thick (by assumption), there exists an m-element subset $M \subset B \cap \mathbb{Z}^m$ of linear dimension m. For any $x \in M$ fix $y \in \mathbb{Z}^r$ such that $x \oplus y \in \Omega$. We obtain an m-element subset $M' \subset \Omega \cap \mathbb{Z}^{m+r}$ of linear dimension m.

Let $\mathcal{L}_1$ be the subspace of $\mathbb{R}^{m+r}$ parallel to the plane $\mathcal{L}$ and of the same dimension[5]. Then $\mathcal{L}_0 := \mathcal{L}_1 + \mathcal{L}(M')$ is a proper subspace of $\mathbb{R}^{m+r}$, because $\dim \mathcal{L}_0 \leq \dim \mathcal{L}_1 + m \leq m + r - 1$.

Proof of item 1. — Since the plane $\mathcal{L}$ contains K_0'', the subspace $\mathcal{L}_1$ contains $K_0'' - K_0''$. Since Ω contains $K'' - K''$, even the set $\mathcal{L}_1 \cap \Omega$ contains $K_0'' - K_0''$. □

Proof of item 2. — Since $B_0 \cap \Gamma_0$ contains both the sets $K_0'' - K_0''$ and M', and since $K_0'' - K_0''$ generates $\mathcal{L}_1$ by (7.6), the set $B_0 \cap \Gamma_0$ generates $\mathcal{L}_0$. □

Proof of item 3. — Let $x \oplus 0 \in \mathbb{R}^m \oplus \mathbb{R}^r$ be orthogonal to $\mathcal{L}_0$. Then it is orthogonal to the set M', whence $x \in \mathbb{R}^m$ is orthogonal to M. Since M generates the whole space $\mathbb{R}^m$, we have $x = 0$. □

Now the Lemma on Partial Covering becomes an easy consequence of the following assertion.

Proposition 7.5. — *We can choose the δ-distorting vectors $a_1, \dots, a_r$ in our construction to have*

$$(\operatorname{Vol} B_0)/\Delta(\Gamma_0) \leq c_{45}(\sigma) \operatorname{Vol} B_0 \left(k / \operatorname{Vol} B_0\right)^{1/2(2m+r)}. \tag{7.8}$$

(Recall that $\delta = 1/2\sqrt{\sigma}$ and $r = \lceil 8\sigma \log(2\sigma) \rceil$.)

[5] Recall that plane $\mathcal{L}$ was defined at the end of the previous subsection.

Proof of Lemma 4.5 (assuming Proposition 7.5). — Let $\pi\colon \mathbb{Z}^m \oplus \mathbb{Z}^r \to \mathbb{Z}^m$ be the projection on the first summand. By the very definition of Freiman's map Φ, we have $\pi \circ \Phi = \mathrm{id}_{\mathbb{Z}^m}$. Therefore π induces a one-to-one map $K'' \to K'$. Hence $\varphi \circ \pi$ induces a one-to-one map $K'' \to K$. It follows that the set $K_0 := \varphi \circ \pi(K_0'')$ satisfies $|K_0| = |K_0''| \geq k/c_{44}$, which is (4.3).

Let $\mathcal{L}_0$, B_0 and Γ_0 be defined from Proposition 7.4, and let φ_0 be the restriction of $\varphi \circ \pi$ to Γ_0. If we change coordinates, identifying Γ_0 with $\mathbb{Z}^{m_0}$ and $\mathcal{L}_0$ with $\mathbb{R}^{m_0}$, then we obtain a triple (m_0, B_0, φ_0), satisfying the requirements of Lemma 4.5. Indeed,

$$m_0 = \dim \mathcal{L}_0 \leq m + r - 1 \leq c_{46}(\sigma),$$

which is the condition (i)$'$ of Lemma 4.5. Further, $K_0'' - K_0'' \subset B_0 \cap \Gamma_0$, whence $\varphi_0(B_0 \cap \Gamma_0) \supset K_0 - K_0$, which is the condition (ii)$'$ of Lemma 4.5. Finally, the left-hand side of (7.8) is independent on the choice of coordinates. Since in the new coordinates we have $\Delta(\Gamma_0) = 1$, we obtain

$$\operatorname{Vol} B_0 \leq c_{45}(\sigma) \operatorname{Vol} B_0 \left(k / \operatorname{Vol} B_0\right)^{1/2(2m+r)}.$$

Since $2(2m + r) \leq c_{42}$, this proves (4.4). □

It remains to prove Proposition 7.5, which will be done in the next section.

8. Proof of the Lemma on Partial Covering: estimating $(\operatorname{Vol} B_0)/\Delta(\Gamma_0)$

8.1. A badly approximable system of distorting vectors. — Let B^* be the convex body *dual to* B, that is

$$B^* = \{x^* \in \mathbb{R}^m \ : \ (x, x^*) \leq 1 \text{ for any } x \in B\}.$$

As proved in [**5**, Chapter IV, Theorem VI],

$$\operatorname{Vol} B^* \leq 4^m V^{-1} = 4^m (\Sigma k)^{-1}, \tag{8.1}$$

where we put $V = \operatorname{Vol} B$ and $\Sigma = V/k$.

We want $a_1, \ldots, a_r$ to be a (B^*, X, C)-badly approximable system of δ-distorting vectors (see Definition 6.7), X and C to be specialized later. First we need to estimate the measure of δ-distorting vectors in the unit cube. We follow the argument of [**12**, Section 2.16].

Proposition 8.1. — *Let δ be a positive real number*[6] *satisfying $\delta < 1/\sqrt{\sigma}$. Then the set $M(\delta)$ of $\delta-$ distorting vectors $a \in [0,1)^m$ satisfies*

$$\operatorname{Vol} M(\delta) \geq \frac{1 - \delta\sqrt{\sigma}}{\sigma} \frac{1}{k}. \tag{8.2}$$

[6] We forget for a while that we have already specified δ.

Proof. — We use the circle method. For $a \in [0,1)^m$ put

$$S(a) = \sum_{x \in K'} e^{2\pi i(a,x)}, \quad S_1(a) = \sum_{x \in K'+K'} e^{-2\pi i(a,x)}.$$

Then

$$k^2 = \int_{[0,1)^m} \sum_{\substack{x,y \in K' \\ z \in K'+K'}} e^{2\pi i(a,x+y-z)} da = \int_{[0,1)^m} S^2(a)S_1(a)da.$$

We have trivially

$$\int_{M(\delta)} S^2(a)S_1(a)da \le k^2 \left|K' + K'\right| \mathrm{Vol} M(\delta) \le \sigma k^3 \, \mathrm{Vol}\, M(\delta),$$

and by the Cauchy-Schwarz inequality

$$\begin{aligned} \int_{[0,1)^m \setminus M(\delta)} S^2(a)S_1(a)da &\le \delta k \int_{[0,1)^m} |S(a)||S_1(a)|da \\ &\le \delta k \sqrt{\int_{[0,1)^m} |S(a)|^2 da} \sqrt{\int_{[0,1)^m} |S_1(a)|^2 da} \\ &\le \delta k \sqrt{k} \sqrt{\sigma k} = \delta \sqrt{\sigma} k^2. \end{aligned}$$

Hence $k^2 \le \sigma k^3 \mathrm{Vol} M(\delta) + \delta\sqrt{\sigma} k^2$, which implies (8.2). □

The next proposition is a direct consequence of Proposition 8.1, Lemma 6.8 and inequality (8.1).

Proposition 8.2. — *Assume that* $0 < \delta < 1/\sqrt{\sigma}$ *and let* $\kappa, \nu > 0$ *satisfy the condition*

$$(m + r)\kappa + m\nu < 1.$$

Also, assume that

$$\Sigma > \left(\frac{\sigma \cdot 24^m \cdot 3^r}{1 - \delta\sqrt{\sigma}} \right)^{\frac{1}{1-(m+r)\kappa+m\nu}}.$$

Then for $X = \Sigma^\kappa$ *and* $C = \Sigma^\nu$ *there exists a* (B^*, X, C)*-badly approximable system of* δ*-distorting vectors* $a_1, \ldots, a_r \in [0,1)^m$.

In particular, specifying $\kappa = \mu = 1/2(2m + r)$, we obtain the following.

Proposition 8.3. — *Assume that* $\Sigma \ge e^{26\sigma \log(2\sigma)}$. *Then for* $\delta = 1/2\sqrt{\sigma}$ *and* $X = C = \Sigma^{1/2(2m+r)}$ *there exists a* (B^*, X, C)*-badly approximable system of* δ*-distorting vectors* $a_1, \ldots, a_r \in [0,1)^m$.

8.2. Estimating $\operatorname{Vol} B_0$. — Since we are going to apply Lemma 6.5, we start with the following assertion.

Proposition 8.4. — *The convex body* Ω *defined in (7.7) contains an* $(m+r)$*-dimensional ball of radius* $(m+1)^{-1}$.

Proof. — Since the Mahler's basis of the body B is orthonormal (see the beginning of Section 7), we obtain

$$\|x\|_B \leq \max(1, m/2)\|x\| \quad (x \in \mathbb{R}^m). \tag{8.3}$$

Define a linear map $A\colon \mathbb{R}^m \to \mathbb{R}^r$ by $Ax = ((x, a_1), \dots, (x, a_r))$. Since $a_1, \dots, a_r \in [0,1)^m$, we have $\|Ax\|_\infty \leq m\|x\|_\infty$. Now fix $x \oplus y \in \mathbb{R}^m \oplus \mathbb{R}^r (\cong \mathbb{R}^{m+r})$. Then

$$\begin{aligned} \|x \oplus y\|_\Omega &= \max\left(\|x\|_{2B}, \|Ax - y\|_\infty\right) \\ &\leq \max\left(\tfrac{1}{2}\|x\|_B, m\|x\|_\infty + \|y\|_\infty\right) \\ &\leq (m+1)\|x \oplus y\|_\infty \\ &\leq (m+1)\|x \oplus y\|. \end{aligned}$$

Therefore Ω contains the $(m+r)$-dimensional ball of radius $(m+1)^{-1}$ with center in the origin. □

Now we are able to estimate $\operatorname{Vol} B_0$.

Proposition 8.5. — *We have*

$$\operatorname{Vol} B_0 \leq c_{81} V, \tag{8.4}$$

where $c_{81} = 2^{m+r}(m+r)!(m+1)^{m+r}$.

Proof. — We have

$$\operatorname{Vol} \Omega = 2^{m+r} V. \tag{8.5}$$

Combining this with Proposition 8.4 and Lemma 6.5, we obtain (8.4). □

8.3. Proof of Proposition 7.5. — Now we are in a position to complete the proof of Proposition 7.5. Let X and C be defined as in Proposition 8.3. By (7.1), the assumption of Proposition 8.3 is satisfied. Therefore we may assume that vectors $a_1, \dots, a_r$ form a (B^*, X, C)-badly approximable system. We shall see that this yields (7.8).

Put $\Delta_0 = \Delta(\Gamma_0)$. The argument splits into two cases (recall that $\Sigma = V/k$).

Case 1: $\Delta_0 \geq \Sigma^{1/2(2m+r)}$. — Since $c_{81} \leq c_{45}$, in this case the inequality (7.8) follows immediately from (8.4). (Note that in this case we did not need the fact that $a_1, \dots, a_r$ form a badly approximable system.)

Case 2: $\Delta_0 \leq \Sigma^{1/2(2m+r)}$. — By Lemma 6.10 there exists a vector $l = \lambda \oplus \mu \in \mathbb{Z}^m \oplus \mathbb{Z}^r$ orthogonal to the subspace $\mathcal{L}_0$ and satisfying $0 < \|l\|_\infty \leq \Delta_0$. By Proposition 7.4 (3), $\mu = (\mu_1, \ldots, \mu_r) \neq 0$. Now, since $\|\mu\| \leq \|l\| \leq \Delta_0 \leq X$ and since $a_1, \ldots, a_r$ form a (B^*, X, C)-badly approximable system, we have $\|\mu_1 a_1 + \cdots + \mu_r a_r + \lambda\|_{B^*} \geq C$. This means that for some $x \in \mathbb{R}^m$ we have

$$(8.6) \qquad |(\mu_1 a_1 + \cdots + \mu_r a_r + \lambda, x)| \geq C\|x\|_B.$$

Put $w = x \oplus Ax \in \mathbb{R}^m \oplus \mathbb{R}^r$ (where A is the linear map defined in the proof of Proposition 8.4). Clearly, $\|w\|_\Omega = \frac{1}{2}\|x\|_B$. Since the left-hand side of (8.6) is equal to $|(w, l)|$, we have

$$(8.7) \qquad |(w,l)| \geq C\|x\|_B = 2C\|w\|_\Omega$$

Let $W = w^\perp$ be the orthogonal complement to w and $\pi\colon \mathbb{R}^{m+r} \to W$ the orthogonal projection. We have the following three inequalities:

$$(8.8) \qquad \mathrm{Vol}_{m+r-1}(\pi(\Omega)) \leq c_{82}\frac{\|w\|_\Omega}{\|w\|}V;$$

$$(8.9) \qquad \mathrm{Vol}_{m_0}(\pi(B_0)) \leq c_{83}\,\mathrm{Vol}_{m+r-1}(\pi(\Omega));$$

$$(8.10) \qquad \mathrm{Vol}_{m_0}(B_0) \leq \frac{\|w\|\|l\|}{|(w,l)|}\mathrm{Vol}_{m_0}(\pi(B_0)).$$

Here $c_{82} = (m+r)2^{m+r}$ and $c_{83} = (m+r-1)!(m+1)^{m+r-1}$.

Indeed, (8.8) follows at once from (8.5) and Lemma 6.6. To prove (8.10) note that by Lemma 6.9

$$\mathrm{Vol}_{m_0}(B_0) = \frac{\mathrm{Vol}_{m_0}(\pi(B_0))}{\det \pi\,|_{\mathcal{L}_0}} \leq \frac{\|w\|\|l\|}{|(w,l)|}\mathrm{Vol}_{m_0}(\pi(B_0)).$$

Finally, as we have seen in Proposition 8.4, the body Ω contains an $(m+r)$-dimensional ball of radius $(m+1)^{-1}$. The projection π maps it onto an $(m+r-1)$-dimensional ball of the same radius. Now we obtain (8.9) applying Lemma 6.5.

Combining the inequalities (8.7)–(8.10), we obtain $\mathrm{Vol}_{m_0}(B_0) \leq c_{84}\|l\|VC^{-1}$ with $c_{84} = \frac{1}{2}c_{82}c_{83}$. Since $\|l\| \leq \sqrt{m+r}\|l\|_\infty \leq \sqrt{m+r}\Delta_0$, we obtain finally

$$\mathrm{Vol}_{m_0}(B_0) \leq c_{85}\Delta_0 V \Sigma^{-1/2(2m+r)}$$

with $c_{85} = \sqrt{m+r}c_{84} \leq c_{45}$, which proves (7.8). This completes the proof of Proposition 7.5 and the Lemma on Partial Covering. □

9. Proof of Proposition 4.2 (the iteration step)

In this section we prove Proposition 4.2. We fix a real number $T \geq 2$, and write "admissible" instead of "T-admissible" in the sequel. Given an admissible triple (m, B, φ), we have to construct another admissible triple (m', B', φ') satisfying (4.1). Note that (4.1) holds (with another constant) for the triple (m_0, B_0, φ_0) constructed in Lemma 4.5. However, instead of conditions (i)–(iii) of Theorem 3.1, we have only (i)$'$ and (ii)$'$, which can be regarded as weaker analogues of (i) and (ii). Our strategy will be to "correct" the triple (m_0, B_0, φ_0) step-by-step, obtaining at the final step the desired (m', B', φ').

Thus, fix an admissible triple (m, B, φ), and again put $V = \operatorname{Vol} B$ and $\Sigma = V/k$.

9.1. The condition (ii)

Proposition 9.1. — *There exists a triple (m_1, B_1, φ_1) satisfying*

$$\operatorname{Vol} B_1 \le c_{91}(\sigma) V \Sigma^{-1/c_{42}(\sigma)} \tag{9.1}$$

and the conditions

(i)″ $m_1 \le c_{92}(\sigma)$;

(ii) $\varphi_1(B_1 \cap \mathbb{Z}^{m_1}) \supset K$.

Here $c_{91} = 2^{\sigma c_{44}} c_{45}$ and $c_{92} = c_{46} + \sigma c_{44}$.

Proof. — What follows is a combination of arguments due to Freiman [**12**, Section 2.24] and Ruzsa [**30**, Section 5]. Let (m_0, B_0, φ_0) and K_0 be constructed in Lemma 4.5. Let $A = \{a_1, \dots, a_s\}$ be a maximal subset of K with the following property:

$$(a_i + K_0) \cap (a_j + K_0) = \varnothing \quad (i \ne j). \tag{9.2}$$

Then

$$s = |A| \le \sigma c_{44}. \tag{9.3}$$

(Indeed, (9.2) yields that $|A + K_0| \ge s|K_0|$, whence

$$\sigma k \ge |K + K| \ge |A + K_0| \ge s|K_0| \ge sk/c_{44},$$

which proves (9.3).) By the maximal choice of the set A, for any $b \in K$ there exist $a_i \in A$ such that $(b + K_0) \cap (a_i + K_0) \ne \varnothing$. In other words,

$$K \subset A + (K_0 - K_0). \tag{9.4}$$

Now put $\Xi = \{x \in \mathbb{R}^s : \|x\|_\infty \le 1\}$, and define a homomorphism $\psi\colon \mathbb{Z}^s \to \mathbb{Z}^n$ by $e_i \mapsto a_i$, where $e_1 = (1, 0, \dots, 0), \dots, e_s = (0, \dots, 0, 1)$ is the standard basis of $\mathbb{R}^s$. Further, put

$$m_1 = m_0 + s, \quad B_1 = B_0 \oplus \Xi,$$

(where we identify $\mathbb{R}^{m_1} \cong \mathbb{R}^{m_0} \oplus \mathbb{R}^s$ and $\mathbb{Z}^{m_1} \cong \mathbb{Z}^{m_0} \oplus \mathbb{Z}^s$) and define $\varphi_1\colon \mathbb{Z}^{m_1} \to \mathbb{Z}^n$ by

$$\varphi_1 |_{\mathbb{Z}^{m_0}} = \varphi_0, \quad \varphi_1 |_{\mathbb{Z}^s} = \psi.$$

Since $\varphi_0(B_0) \supset K_0 - K_0$ and $\psi(\Xi) \supset A$, we have (ii). Estimates (i)″ and (9.1) are obvious. □

Remark 9.2. — We could replace the cube Ξ by the "octahedron"

$$\Xi' = \{x = (x_1, \dots, x_s) \in \mathbb{R}^s : |x_1| + \dots + |x_s| \le 1\}.$$

This would imply a better value for the constant c_{91}. However, this would not have much influence on the final value of the constant c_{11} in the Main Theorem.

9.2. Condition (iii)

Proposition 9.3. — *There exists a triple* (m_2, B_2, φ_2) *satisfying*

$$\operatorname{Vol} B_2 \le c_{93}(\sigma, T) V \Sigma^{-1/c_{42}(\sigma)} \tag{9.5}$$

and the conditions

(i)″ $m_2 \le c_{92}$*;*
(ii) $\varphi_2(B_2 \cap \mathbb{Z}^{m_2}) \supset K$.
(iii) *the restriction* $\varphi_2 \mid_{TB_2 \cap \mathbb{Z}^{m_2}}$ *is one-to-one;*

Here $c_{93} = (2c_{92}T)^{c_{92}} c_{91}$.

Proof. — We follow the argument of [**12**, Lemma 2.26] with some changes. Let (m_1, B_1, φ_1) be the triple constructed in Proposition 9.1. We say that the triple (m_2, B_2, φ_2) is *appropriate* if it satisfies the conditions

$$m_2 \le m_1, \quad \varphi_2(B_2 \cap \mathbb{Z}^{m_2}) \supset K, \quad \operatorname{Vol} B_2 \le (2m_1T)^{m_1 - m_2} \operatorname{Vol} B_1.$$

Appropriate triples exist — for example, the triple (m_1, B_1, φ_1) is such. Fix an appropriate triple (m_2, B_2, φ_2) with the *minimal* value of m_2. To prove the proposition we have to show that this triple satisfies (iii).

Assuming the contrary, we find a non-zero $e \in 2TB_2 \cap \mathbb{Z}^{m_2}$ such that $\varphi_2(e) = 0$. We may assume that the greatest common divisor of the coordinates of vector e is 1. Then there exists a basis $e_1, \ldots, e_{m_2}$ of $\mathbb{Z}^{m_2}$ such that $e_{m_2} = e$. We assume this basis to be orthonormal, redefining the inner product.

Let $\pi \colon \mathbb{R}^{m_2} \to \mathbb{R}^{m_2 - 1}$ be the projection on the first $m_2 - 1$ coordinates. Put $B_2' = \pi(B_2)$. Since $e = e_{m_2} \in \operatorname{Ker}\varphi_2$, there is a uniquely defined map $\varphi_2' \colon \mathbb{Z}^{m_2-1} \to \mathbb{Z}^n$ such that $\varphi_2 = \varphi_2' \circ \pi$. We have

$$\varphi_2'(B_2' \cap \mathbb{Z}^{m_2-1}) = \varphi_2'(\pi(B_2) \cap \pi(\mathbb{Z}^{m_2})) \supset \varphi_2' \circ \pi(B_2 \cap \mathbb{Z}^{m_2}) = \varphi_2(B_2 \cap \mathbb{Z}^{m_2}) \supset K.$$

Also, since $e \in 2TB_2$, we have $\|e\|_{B_2} \le 2T$, and by Lemma 6.6

$$\operatorname{Vol}_{m_2-1} B_2' \le 2Tm_2 \operatorname{Vol}_{m_2} B_2 \le (2m_1T)^{m_1-(m_2-1)} \operatorname{Vol} B_1.$$

Thus, the triple $(m_2 - 1, B_2', \varphi_2')$ is appropriate, which contradicts the minimal choice of m_2. □

9.3. The condition (i). — Now it is easy to complete the proof of Proposition 4.2. Let (m_2, B_2, φ_2) be the triple constructed in Proposition 9.3. Put $K' = \varphi_2^{-1}(K)$. Since $T \ge 2$, it follows from (ii) and (iii) that the map $\varphi_2 \colon K' \to K$ is F_2-isomorphic. Therefore $|K' + K'| \le \sigma |K'|$, whence by Lemma 4.3 we have $m' := \dim K' \le \lfloor \sigma - 1 \rfloor$. Put $\mathcal{L}' = \mathcal{L}(K')$.

We may assume that the Mahler's basis of the body B_2 is orthonormal. Then B_2 contains an m_2-dimensional ball of radius $2/m_2$. Putting $B' = \mathcal{L}' \cap B_2$, we have by Lemma 6.5

$$\operatorname{Vol}_{m'}(B') \le m_2!(m_2/2)^{m_2} \operatorname{Vol}_{m_2}(B_2) \le c_{94}(\sigma, T) V \Sigma^{-1/c_{42}(\sigma)} \tag{9.6}$$

with $c_{94}(\sigma, T) = c_{92}(\sigma)^{2c_{92}(\sigma)} c_{93}(\sigma, T) \le c_{41}$.

Finally, put $\Gamma' = \mathcal{L}' \cap \mathbb{Z}^{m_2}$ and $\varphi' = \varphi_2 |_{\Gamma'}$. When we identify $\mathcal{L}'$ with $\mathbb{R}^{m'}$ and Γ' with $\mathbb{Z}^{m'}$, the volume of B' should be multiplied by $\Delta(\Gamma)^{-1}$. Since $\Delta(\Gamma) \geq 1$, we will still have (9.6).

Thus, the triple (m', B', φ') is admissible and satisfies (4.1). Proposition 4.2 is proved. □

10. Final remarks

10.1. Various formulations of Freiman's theorem. — Both Theorems 1.2 and 1.3 are new, though not very much is added to Freiman's proof. Freiman's original formulation of his theorem is similar to our Theorem 3.1, but with $|B \cap \mathbb{Z}^m|$ instead of $\operatorname{Vol} B$. Ruzsa's result is as follows.

***Theorem 10.1* (Ruzsa [30]).** — *Let K and L be finite subsets of a torsion-free abelian group. Suppose that $|K| = |L| = k$ and $|K+L| \leq \sigma k$. Then K is a subset of a generalized arithmetical progression P of rank $m \leq c_{101}(\sigma)$ and cardinality $|P| \leq c_{102}(\sigma)k$.*

The main advantage of Ruzsa's theorem is that it deals with distinct sets. Ruzsa's proof implies an estimate $c_{102}(\sigma) \leq \exp\exp(\sigma^c)$ with an absolute constant c, which is better than (1.4). However, Ruzsa does not prove that P is an F_s-progression (even for $s = 1$), nor does he obtain the inequality $m \leq \lfloor \sigma - 1 \rfloor$, having only the weaker bound $m \leq \exp(\sigma^c)$.

Both these difficulties can be overcome in the case $K = \pm L$: one should combine Ruzsa's result with the arguments from Sections 9 and 3 of the present paper. (In the case $L = -K$ Lemma 4.3 should be replaced by its analogue for $K - K$ proved in [**14**].) This would give us a new proof of Theorem 1.2, the estimate (1.4) being replaced by $c_{11}(\sigma) \leq (2s)^{\exp(\sigma^c)}$, and an analogue of the Main Theorem with $K - K$ instead of $K + K$.

It is very likely that a similar approach (with some additional ideas) would lead to a complete analogue of Theorem 1.2 for the addition of two distinct sets of the same cardinality.

***Remark 10.2* (added in revision).** — Nathanson [**26**, Section 9.6] posed the following *proper conjecture*:

Let $\alpha \leq 1$ and σ be positive real numbers, let k be a positive integer, and let K and L be finite subsets of a torsion-free abelian group such that

$$\alpha k \leq |K|, |L| \leq k \quad \textit{and} \quad |K+L| \leq \sigma k.$$

Then K is a subset of an F_1-progression P of rank $c(\alpha,\sigma)$ and cardinality $|P| \leq c'(\alpha,\sigma)k$.

It is easy to see that this conjecture is a consequence of Theorem 1.2. Indeed, it follows from [**28**, Lemma 3.3] (reproduced in [**26**] as Theorem 7.8) that

$$|K+K| \le |K+K-K| \le \left(\frac{\sigma k}{|L|}\right)^3 |L| \le (\sigma/\alpha)^3|K|.$$

Applying Theorem 1.2, we prove the conjecture with $c = \lfloor(\sigma/\alpha)^3 - 1\rfloor$ and $c' = c_{11}((\sigma/\alpha)^3, 1)$. (A slightly more accurate argument gives $(\sigma/\alpha)^2$ instead of $(\sigma/\alpha)^3$.)

10.2. Freiman's proof and Ruzsa's proof. — Put $Ks = \overbrace{K+\cdots+K}^{s}$. Ruzsa starts with proving that

$$|K+L| \le \sigma k \Rightarrow |Ks_1 - Ks_2| \le \sigma^{s_1+s_2}k, \tag{10.1}$$

(where s_1 and s_2 are arbitrary positive integers) and then works with the set K only. He shows also that it is sufficient to consider the case $K \subset \mathbb{Z}$.

The first crucial step of his proof is the following nice theorem.

Theorem 10.3 ([**28**]). — *Let K be a finite set of integers, and s a positive integer. Then for any $N \ge 2s|Ks - Ks|$ there is a a subset $K' \subset K$ of cardinality $|K'| \ge k/s$, which is F_s-isomorphic to a subset of the cyclic group of order N.*

Due to this result one may work in a "close environment", which essentially simplifies the reasoning and allows one to avoid iterations.

The second crucial step is the following result:

Theorem 10.4 ([**30**]). — *If A is a subset of a cyclic group of order $N \le \alpha|A|$, then the second difference set $A2 - A2$ contains a progression of rank at most $c_{103}(\alpha)$ and cardinality at least $N/c_{104}(\alpha)$.*

A simple combination of these two theorems shows that there is a progression $P \subset K2-K2$ of bounded rank, satisfying $|P| \ge k/c_{105}(\sigma)$. Now it is easy to complete the proof proceeding in the same manner as in Subsection 9.1 of this paper.

Thus, in both Freiman's and Ruzsa's proofs one first takes care of an "substantial part" of the set in question (or a relative set), and then covers by a progression the whole set. In Freiman's argument the main tools for finding a "partial covering" are the 2^n-theorem and the circle method. In Ruzsa's argument the same role belongs to Theorem 10.4, in the proof of the latter the circle method being crucial too.

Evidently, there are deep interconnections between the two proofs. Revealing them will lead to a much better understanding of the problems connected with Freiman's theorem.

References

[1] Bilu Y., *The $(\alpha + 2\beta)$-inequality on the torus*, J. London Math. Soc., to appear.

[2] Bilu Y., Lev V. F., Ruzsa I. Z., *Rectification principles in additive number theory*, Discr. Comput. Geom., **19**, 1998, 343–353.

[3] Bombieri E., Vaaler J., *On Siegel's Lemma*, Inv. Math., **73**, 1983, 11–31.

[4] Brailovski L.V, Freiman G.A., *On a Product of Finite Subsets in a Torsion-Free Group*, J. Algebra, **130**, 1990, 462–476.

[5] Cassels J.W.S., *An Introduction to the Geometry of Numbers*, Springer, 1959.

[6] Cauchy A.L., *Recherches sur les nombres*, J. Ècole Polytech., **9**, 1813, 99–116.

[7] Davenport H., *On the addition of residue classes*, J. London Math. Soc., **10**, 1935, 30–32.

[8] Davenport H., *A historical note*, J. London Math. Soc., **22**, 1947, 100–101.

[9] Fishburn P.C., *On a contribution of Freiman to additive number theory*, J. Number Theory, **35**, 1990, 325–334.

[10] Freiman G.A., *Inverse problems in additive number theory VI., On the addition of finite sets III* (Russian), Izv. Vysš. Učebn. Zaved. Matem., no. **3 (28)**, 1962, 151–157.

[11] Freiman G.A., *Inverse problems of additive number theory, VII. On addition of finite sets, IV. The method of trigonometric sums* (Russian), Izv. Vysš. Učebn. Zaved. Matem., No **3(28)**, 1962, 131–144.

[12] Freiman G.A., *Foundations of a Structural Theory of Set Addition* (Russian), Kazan', 1959; English Translation: Translation of Mathematical Monographs **37**, Amer. Math. Soc., Providence, 1973.

[13] Freiman G.A., *What is the structure of K if $K+K$ is small?*, Number Theory, New York 1984–1985, Lecture Notes in Math., **1240**, Springer, 1987, 109–134,

[14] Freiman G.A., Heppes A., Uhrin B., *A lower estimation for the cardinality of finite difference sets in $\mathbb{R}^n$*, Proc. Conf. Number Theory, Budapest 1987, Coll. Math. Soc. J. Bolyai, Budapest, **51**, 1989, 125–139.

[15] Hamidoune Y.O., *On inverse additive problems*, Preprint EC95/01, Inst. Blaise Pascal, Paris, January 1995.

[16] Hamidoune Y.O., *An Isoperimetric Method in Additive Theory*, J. Algebra, **179**, 1996, 622–630.

[17] Hamidoune Y.O., *Some Results in Additive Number Theory*, Rapport ECH197, Univ. Paris VI, 1997.

[18] Kemperman I.H.B., *On small sumsets in an abelian group*, Acta Math., **103**, 1960, 63–88.

[19] Kemperman I.H.B., *On products of sets in a locally compact group*, Fund. Math., **56**, 1964, 51–68.

[20] Kneser M., *Ein Satz über abelschen Gruppen mit Anwendungen auf die Geometrie der Zahlen*, Math. Z., **61**, 1955, 429–434.

[21] Kneser M., *Summenmengen in lokalkompakten abelschen Gruppen*, Math. Z., **66**, 1956, 88–110.

[22] Lev V.F., *Structure Theorem for Multiple Addition and Frobenius Problem*, J. Number Th., **58**, 1996, 79–88; addendum: **65**, 1997, 96–100.

[23] Lev V.F., Smeliansky P.Y., *On addition of two distinct sets of integers*, Acta Arith., **70**, 1995, 85-91.

[24] Mann H., *Addition Theorems: the Addition Theorems of Group Theory and Number Theory*, Wiley, New York, 1965.

[25] Nathanson M.B., *An inverse theorem for sums of sets of lattice points*, J. Number Theory, **46**, 1994, 29–59.

[26] Nathanson M.B., *Additive Number Theory: 2. Inverse Theorems and the Geometry of Sumsets*, Graduate Text in Math. **165**, Springer, New York, 1996.

[27] Postnikova L.P., *Fluctuations in the distribution of fractional parts* (Russian), Dokl. Akad. Nauk SSSR, **161**, 1965, 1282–1284; English Translation: Soviet Math. Dokl., **6**, 1965, 597–600.

[28] Ruzsa I.Z., *Arithmetical progressions and the number of sums*, Per. Math. Hung., **25**, 1992, 105–111.

[29] Ruzsa I.Z., *A concavity property of for the measure of product sets in groups*, Fund. Math., **140**, 1992, 247–254.

[30] Ruzsa I.Z., *Generalized arithmetical progressions and sumsets*, Acta Math. Hungar., **65** (1994), 379–388.

[31] Ruzsa I.Z., *Sums of sets in several dimensions*, Combinatorica, **14**, 1994, 485–490.

[32] Stanchescu Y., *On addition of two distinct sets of integers*, Acta Arith., **75**, 1996, 191–194.

[33] Stanchescu Y., *On the structure of sets with small doubling property on the plane (i)*, Acta Arith., **83**, 1998, 127–141.

[34] Stanchescu Y., *On finite difference sets*, Acta Math. Hungar., **79**, 1998, 123–138.

[35] Steinig J., *On Freiman's theorems concerning the sum of two finite sets of integers*, this volume.

[36] Vosper A.G., *The critical pairs of subsets of a group of prime order*, J. London Math. Soc., **31**, 1956, 200–205, 280–282.

[37] Zemor G., *Subset sums for binary spaces*, Europ. J. Combin., **13**, 1992, 221–230.

[38] Zemor G., *A generalization to noncommutative groups of a theorem of Mann*, Discr. Math., **126**, 1994, 365–372.

Y. Bilu, Mathematisches Institut, Universitaet Basel, Rheinsprung 21, CH-4051 Basel, Switzerland
E-mail : yuri@math.unibas.ch

Astérisque

ANDRÁS SÁRKŐZY

On finite addition theorems

Astérisque, tome 258 (1999), p. 109-127

<http://www.numdam.org/item?id=AST_1999__258__109_0>

Astérisque **258**, 1999, p. 109–127

ON FINITE ADDITION THEOREMS

by

András Sárközy

Abstract. — If a finite set A of integers included in $\{1, \ldots, N\}$ has more than N/k elements, one may expect that the set ℓA of sums of ℓ elements of A, contains, when ℓ is comparable to k, a rather long arithmetic progression (which can be required to be homogeneous or not). After presenting the state of the art, we show that some of the results cannot be improved as far as it would be thought possible in view of the known results in the infinite case. The paper ends with lower and upper bounds for the order, as asymptotic bases, of the subsequences of the primes which have a positive relative density.

1. Throughout this paper we use the following notations: $c_1, c_2 \ldots$ denote positive absolute constants. If $f(n) = O(g(n))$, then we write $f(n) \ll g(n)$. The cardinality of the finite set S is denoted by $|S|$. The set of the integers, non-negative integers, resp. positive integers is denoted by $\mathbb{Z}$, $\mathbb{N}_0$ and $\mathbb{N}$. $\mathcal{A}$, $\mathcal{B}$ ... denote (finite or infinite) subsets of $\mathbb{N}_0$, and the counting functions of their positive parts are denoted by $A(n)$, $B(n), \ldots$, so that, *e.g.*, $A(n) = |\mathcal{A} \cap \{1, 2, \ldots n\}|$. The Schnirelmann density of the set $\mathcal{A} \subset \mathbb{N}_0$ is denoted by $\sigma(\mathcal{A})$, while the asymptotic density, asymptotic lower density, resp. asymptotic upper density of it is denoted by $d(\mathcal{A})$, $\underline{d}(\mathcal{A})$ and $\overline{d}(\mathcal{A})$ (see [16] for the definition of these density concepts). $\mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_k$ denotes the set of the integers that can be represented in the form $a_1 + a_2 + \cdots + a_k$ with $a_1 \in \mathcal{A}_1$, $a_2 \in \mathcal{A}_2$, $\ldots$, $a_k \in \mathcal{A}_k$; in particular, we write

$$\mathcal{A} + \mathcal{A} = 2\mathcal{A} = S(\mathcal{A}),$$

$$k\mathcal{A} = \mathcal{A} + (k-1)\mathcal{A} \quad \text{for} \quad k = 3, 4, \ldots,$$

and

$$0\mathcal{A} = \{0\}, \quad 1\mathcal{A} = \mathcal{A}.$$

1991 ***Mathematics Subject Classification.*** — 11B13, 11B25, 11B05.

Key words and phrases. — additive number theory, density, additive bases, structure theory of set addition.

Research partially supported by Hungarian National Foundation for Scientific Research, Grant No. 1811.

If $\mathcal{A} \subset \mathbb{N}$ then $\mathcal{P}(\mathcal{A})$ denotes the set of the distinct positive integers n that can be represented in the form $n = \sum_{a\in\mathcal{A}} \varepsilon_a a$ where $\varepsilon_a = 0$ or 1 for all a and, if $\mathcal{A}$ is infinite, then all but finitely many of the ε's are equal to 0. (This notation will be used only in Section 3, while later the letter $\mathcal{P}$ will be reserved for denoting sets of primes.) An arithmetic progression is said to be *homogeneous* if it consists of the consecutive multiples of a non-zero number, i.e., it is of the form $kd, (k+1)d, \dots, \ell d$ (where $d \neq 0$).

2. The classical Schnirelmann-Mann-Kneser-Folkman theory of the set addition studies sums of *infinite* sets (the density and, in case of Kneser's theorem, the structure of the sum set). However, in many applications we are dealing with *finite* sets; in such a case, we cannot use this classical set addition theory or, in the best case, we have difficulties in applying it. Thus recently I have worked out a theory of addition of *finite* sets (partly jointly with Erdős, resp. Nathanson) which is more or less analogous to the case of infinite sets, and several conclusions and applications of this theory are close to the ones obtained by Freiman using a completely different approach. A considerable part of this work was inspired by a paper of Erdős and Freiman [5]. In this paper, first I will give a brief survey of my papers written on this subject. In the second half of the paper two further related problems will be studied.

3. Nathanson and I [20] proved that if we take "many" integers up to N, and we add the set obtained in this way sufficiently many times, then the sum set contains a long arithmetic progression:

Theorem 1. — *If $N \in \mathbb{N}$, $k \in \mathbb{N}$, $\mathcal{A} \subset \{1, 2, \dots, N\}$ and*

$$|\mathcal{A}| \geq \frac{N}{k} + 1, \tag{3.1}$$

then there exists an integer d with

$$1 \leq d \leq k - 1 \tag{3.2}$$

such that if h and z are any positive integers satisfying the inequality

$$\frac{N}{h} + zd \leq |\mathcal{A}|,$$

then the sum set $(2h)\mathcal{A}$ contains an arithmetic progression with z terms and difference d.

Choosing here $h = 2k$ and $z = [N/2kd]$, we obtain

Corollary 1. — *If $N \in \mathbb{N}$, $k \in \mathbb{N}$, $\mathcal{A} \subset \{1, 2, \dots, N\}$ and $\mathcal{A}$ satisfies (3.1), then there exists an integer d satisfying (3.2) such that $4k\mathcal{A}$ contains an arithmetic progression with difference d and length $[N/2kd] \geq [N/2(k-1)k]$.*

The proof of Theorem 1 was based on Dyson's theorem [3] (which slightly generalizes Mann's theorem [19]). We used Theorem 1 to study a problem of Erdős and Freud on the solvability of the equation

$$a_1 + a_2 + \cdots + a_x = 2^y, \quad a_1, a_2, \dots, a_x \in \mathcal{A} \tag{3.3}$$

in "large" subsets $\mathcal{A}$ of $\{1, 2, \dots, N\}$ (in sets $\mathcal{A}$ with $|\mathcal{A}| > [N/3]$). Indeed, we improved on a result of Erdős and Freiman [5]. Later Freiman [14] found another ingenious approach and he improved further on the result.

Corollary 1 was sufficient to study equation (3.3), however, it is not sharp in the sense that it guarantees an arithmetic progression of length only $\gg N/k^2$ in the sum set while one would expect a longer arithmetic progression and, indeed, later I needed a sharper result of this type. In fact, I proved [21] that having the same assumptions as in Corollary 1, one can guarantee a much longer *homogeneous* arithmetic progression in a sum set $\ell\mathcal{A}$ with $\ell \ll k$ (in many applications, we need the existence of a *homogeneous* arithmetic progression in the sum set, and this fact causes certain difficulties):

Theorem 2. — *If $N \in \mathbb{N}$, $k \in \mathbb{N}$, $\mathcal{A} \subset \{1, 2, \dots, N\}$ and (3.1) holds, then there are integers d, ℓ, m such that (3.2) holds, moreover we have*

$$1 \le \ell < 118k \tag{3.4}$$

and

$$\{(m+1)d, (m+2)d, \dots, (m+N)d\} \subset \ell\mathcal{A}. \tag{3.5}$$

It is easy to see that this theorem is the best possible apart from the constant factor 118 in (3.4). This result can be considered as the finite analog of Kneser's theorem [18] (see Lemma 2 below). The proof of Theorem 2 is complicated, it uses both Dyson's theorem and Kneser's theorem.

One might like to sharpen this result by showing that all the elements of the arithmetic progression in (3.5) can be represented as the sum of possibly few *distinct* elements of $\mathcal{A}$; see [20] and Alon [1] for results of this type. The case when the number of distinct summands is unlimited will be studied later (Theorem 4 below).

Before the famous $\alpha + \beta$ conjecture was proved by Mann [19], Khintchin [17] had settled that most important special case of the conjecture when sum sets of the form $k\mathcal{A}$ are considered; indeed, he proved that

$$\sigma(k\mathcal{A}) \ge \min(1, k\sigma(\mathcal{A})). \tag{3.6}$$

In [23] I proved the following finite analog of this result:

Theorem 3. — *If $N \in \mathbb{N}$, $k \in \mathbb{N}$, $\mathcal{A} \subset \{1, 2, \dots, N\}$ and $|\mathcal{A}| \ge 2$, then there are m, d such that $m \in \mathbb{Z}, d \in \mathbb{N}$,*

$$d < 2\frac{N}{|\mathcal{A}|} \tag{3.7}$$

and

$$|\{m+d, m+2d, \dots, m+Nd\} \cap k\mathcal{A}| \ge \left(\min(1, \frac{1}{800}k\frac{|\mathcal{A}|}{N})\right) N. \tag{3.8}$$

The proof is similar to the proof of Theorem 2, although also further ideas are needed. Again, it is easy to see that this theorem is the best possible apart from the constants 2 in (3.7) and, mostly, $\frac{1}{800}$ in (3.8) (we will return to this question in section 4). Note that an easy consideration shows that here we have to give up the requirement that the arithmetic progression in (3.8) should be homogeneous.

An infinite set $\mathcal{A} \subset \mathbb{N}$ is said to be ***subcomplete*** if it contains an infinite arithmetic progression. Improving on a result of Erdős [4], Folkman [11] proved the following remarkable theorem: if $\mathcal{A} \subset \mathbb{N}$ is an infinite set such that there are $\varepsilon > 0$ and N_0 with

$$A(N) > N^{1/2+\varepsilon} \text{ for } N > N_0,$$

then $\mathcal{P}(\mathcal{A})$ is subcomplete. Improving on a result of Alon and Freiman [2], I proved [22] the following finite analogue of Folkman's theorem:

Theorem 4. — *If $N \in \mathbb{N}$, $N > 2500$, $\mathcal{A} \subset \{1, 2, \dots, N\}$ and*

$$|\mathcal{A}| > 200(N \log N)^{1/2}, \tag{3.9}$$

then there are integers d, y, z such that

$$1 \le d < 10^4 \frac{N}{|\mathcal{A}|},$$

$$z > 7^{-1} 10^{-4} |\mathcal{A}|^2,$$

$$y < 7 \cdot 10^4 N z |\mathcal{A}|^{-2} \tag{3.10}$$

and

$$\{yd, (y+1)d, \dots, zd\} \subset \mathcal{P}(\mathcal{A}).$$

Previously Alon and Freiman had proved a similar result with $N^{2/3+\varepsilon}$ on the right hand side of (3.9) and a slightly weaker inequality in place of (3.10). Moreover, independently and nearly simultaneously Freiman [13] proved a result essentially equivalent to Theorem 4 above. I derived Theorem 4 from Theorem 2; this part of the proof is easier, than the proof of Theorem 2. Freiman's proof is also complicated; he combines methods from the geometry of numbers and exponential sums in the manner of his book [12].

Again, Theorem 4 is the best possible apart from the constant factors and, perhaps, the factor $(\log N)^{1/2}$ on the right hand side of (3.9). Probably this logarithmic factor (or, at least, some of it) is unnecessary, although it is quite interesting and unexpected that exactly the same factor appears also in Freiman's result (obtained by a completely different method).

Theorem 4 has many applications. Alon and Freiman [2] found the first applications of a result of this type. Several further applications are discussed in my paper [22]. Papers [6], [7], [8] and [10] contain further applications.

Erdős and I [9] studied the following problem: what happens, if we replace assumption (3.9) by a slightly weaker one so that $|\mathcal{A}|$ drops below $N^{1/2}$? It turns out that there is a sharp drop in the length of the maximal arithmetic progression that we can guarantee in $\mathcal{P}(\mathcal{A})$, however, still it must contain quite a long one. Indeed, let $u = F(N, t)$ denote the greatest integer u such that for every $\mathcal{A} \subset \{1, 2, \dots, N\}$ with $|\mathcal{A}| = t$, the set $\mathcal{P}(\mathcal{A})$ contains u consecutive multiples of a positive integer d:

$$\{(x+1)d, (x+2)d, \dots, (x+u)d\} \subset \mathcal{P}(\mathcal{A})$$

for some x and d, and let $v = G(n,t)$ denote the greatest integer v such that for every $\mathcal{A} \subset \{1,2,\dots,N\}$ with $|\mathcal{A}| = t$, the set $\mathcal{P}(\mathcal{A})$ contains an arithmetic progression of length v:

$$\{y+(z+1)d, y+(z+2)d, \dots, y+(z+v)d\} \subset \mathcal{P}(\mathcal{A})$$

for some y, z and $d(>0)$. Clearly, $F(N,t) \le G(N,t)$ for all N and $t \le N$, and since

$$\mathcal{P}(\{1,2,\dots,t\}) = \{1,2,\dots,t(t+1)/2\} \subset \{1,2,\dots,t^2\},$$

thus we have

$$(F(N,t) \le)G(N,t) \le t^2$$

for all N and $t \le N$. On the other hand, by Theorem 4 for $t \gg (N\log N)^{1/2}$ we have

$$\begin{aligned} F(N,t) &> z - y > z - 7.10^4 N z|\mathcal{A}|^{-2} \\ &= z(1 - 7.10^4 N|\mathcal{A}|^{-2}) \gg z \gg |\mathcal{A}|^2 = t^2 \end{aligned}$$

if $t \gg (N\log N)^{1/2}$.

Theorem 5. — *If $N \ge N_0$ and $18(\log N)^2 < t \le N$, then we have*

$$(G(N,t) \ge)F(N,t) > \frac{1}{18}\frac{t}{(\log N)^2}.$$

Theorem 6

(i) If $N > N_0$ and $c\log N < t < \frac{1}{3}N^{1/3}$, then we have

$$F(N,t) < 16\frac{t}{\log N}\log\left(\frac{t}{\log N}\right).$$

(ii) If $\varepsilon > O$ and $t_0(\varepsilon) < t < (1-\varepsilon)N^{1/2}$, then we have $F(N,t) < (1+\varepsilon)t$.

Theorem 7

(i) If $N > N_0$ and $\exp(2(\log N)^{1/2}) < t < N^{1/4}$, then we have

$$G(N,t) < t\exp\left(4\max\left(\frac{\log N}{\log t}, \frac{(\log t)^2}{\log N}\right)\right).$$

(ii) $t_0 < t < \frac{1}{2}N^{1/2}$ we have $G(N,t) < 2t^{3/2}$.

Paper [9] contains several further related results.

4. As we mentioned above, Theorem 4 is nearly sharp in the sense that apart from the constant factors and the, perhaps, unnecessary factor $(\log N)^{1/2}$ on the right hand side of (3.9), the theorem is the best possible.

On the other hand, it is easy to see that the other two main theorems Theorem 2 and 3 are the best possible apart from the constants on the right hand side of (3.4) and (3.8) (and, less importantly, (3.7)). One might like to determine or, at least, to estimate these constants. This problem can be considered as the finite analog of the famous $\alpha+\beta$ problem (apart from the fact that here we restrict ourselves to sum sets $\mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_k$ with $\mathcal{A}_1 = \mathcal{A}_2 = \cdots = \mathcal{A}_k$). Since Theorems 2 and 3 are closely related, thus here I will study only the constant in Theorem 3.

If $N \in \mathbb{N}$, $k \in \mathbb{N}$, $\mathcal{A} \subset \{1,2,\dots,N\}$ and $|\mathcal{A}| \geq 2$, then let $E(N,k,\mathcal{A})$ denote the maximal number of elements of $k\mathcal{A}$ contained in an arithmetic progression of length N:

$$E(N,k,\mathcal{A}) = \max_{m\in\mathbb{Z},d\in\mathbb{N}} |\{m+d, m+2d, \dots, m+Nd\} \cap k\mathcal{A}|.$$

For $k \in \mathbb{N}$, $k \geq 2$ let $C(k)$ denote the greatest number such that for all $N \in \mathbb{N}$, $\mathcal{A} \subset \{1,2,\dots,N\}$ and $|\mathcal{A}| \geq 2$ we have

$$E(N,k,\mathcal{A}) \geq \left(\min(1, C(k)k\frac{|\mathcal{A}|}{N})\right) N,$$

and define C by $C = \inf_{k=2,3,\dots} C(k)$ so that C is the greatest number such that for all $N \in \mathbb{N}$, $k \in \mathbb{N}$, $k \geq 2$, $\mathcal{A} \subset \{1,2,\dots,N\}$ and $|\mathcal{A}| \geq 2$ we have

$$E(N,k,\mathcal{A}) \geq \left(\min(1, Ck\frac{|\mathcal{A}|}{N})\right) N.$$

Moreover, for $k \in \mathbb{N}$, $k \geq 2$ let $C_\infty(k)$ denote the greatest number such that for all $\varepsilon > 0$ there is an $L = L(\varepsilon)$ with the property that for all $N \in \mathbb{N}$, $\mathcal{A} \subset \{1,2,\dots,N\}$ and $|\mathcal{A}| > L$ we have

$$E(N,k,\mathcal{A}) \geq \left(\min(1, (C_\infty(k)-\varepsilon)k\frac{|\mathcal{A}|}{N})\right) N,$$

and define C_∞ by $C_\infty = \inf_{k=2,3,\dots} C_\infty(k)$.

By Theorem 3 we have

$$C_\infty \geq C \geq \frac{1}{800}. \tag{4.1}$$

In the proof of Theorem 3, I did not force to give a possibly sharp lower bound for C and C_∞. Correspondingly, by a careful analysis of the proof, the lower bound in (4.1) (mostly the one for C_∞) could be improved considerably; however, to get above, say, $\frac{1}{10}$ with the lower bound, essential new ideas would be needed.

Khintchin's theorem (3.6) may suggest that, perhaps, we have $C = C_\infty = 1$. This is not so; indeed, for $|\mathcal{A}| = 2$, $k \in \mathbb{N}$ clearly we have $k|\mathcal{A}| = k+1$ so that

$$E(N,k,\mathcal{A}) \leq k+1.$$

Thus for $|\mathcal{A}| = 2$, $k \in \mathbb{N}$, $N \geq k+1$ we have

$$E(N,k,\mathcal{A}) \leq k+1 = \left(\min(1, \frac{k+1}{2k} \cdot k\frac{|\mathcal{A}|}{N})\right) N$$

which shows that $C(k) \leq \frac{k+1}{2k}$, $C \leq 1/2$. One might think that this example is the "worst" one so that $C = 1/2$ and, perhaps, $C_\infty = 1$. I will show that this guess is also wrong; the next two sections will be devoted to giving possibly sharp upper bounds for C and C_∞.

5. First it will be proved:

Theorem 8. — *If $N \in \mathbb{N}$,*

$$N \geq k+2 \tag{5.1}$$

and $k \in \mathbb{N}$, then for $\mathcal{A} = \{1,2,N\}$ we have

$$E(N,k,\mathcal{A}) = k+2. \tag{5.2}$$

For $N \geq k+2$ this implies

$$E(N,k,\mathcal{A}) = \left(\min(1, \frac{k+2}{3k} \cdot k\frac{|\mathcal{A}|}{N})\right) N.$$

It follows that

Corollary 2. — *For all $k \in \mathbb{N}$, $k \geq 2$ we have*

$$C(k) \leq \frac{k+2}{3k}$$

so that

$$C \leq \frac{1}{3}.$$

Proof of Theorem 8. — Clearly we have

$$k\mathcal{A} = k\{1,2,N\} = \bigcup_{i=0}^{k} (i\{N\} + (k-i)\{1,2\})$$
$$= \bigcup_{i=0}^{k} \{iN+k-i, iN+k-i+1, \ldots, iN+2(k-i)\} = \bigcup_{i=0}^{k} \mathcal{B}_i,$$

where

$$\mathcal{B}_i = \{iN+k-i, iN+k-i+1, \ldots, iN+2(k-i)\}.$$

Consider now an arithmetic progression $\mathcal{Q}(m,d,N) = \{m+d, m+2d, \ldots, m+Nd\}$ with $m \in \mathbb{Z}$, $d \in \mathbb{N}$. Assume first that $d \geq k+1$. Then for $0 \leq i \leq k$, the difference between the greatest and smallest of $\mathcal{B}_i$ is

$$(iN+2(k-i)) - (iN+k-i) = k-i \leq k < d,$$

thus clearly, $\mathcal{Q}(m,d,N)$ may contain at most one element of each $\mathcal{B}_i$. It follows that

$$|\mathcal{Q}(m,d,N) \cap k\mathcal{A}| = |\mathcal{Q}(m,d,N) \cap \bigcup_{i=0}^{k} \mathcal{B}_i|$$
$$\leq \sum_{i=0}^{k} \left|\mathcal{Q}(m,d,N) \cap \mathcal{B}_i\right| \leq \sum_{i=0}^{k} 1 = k+1 \text{ (for } d \geq k+1).$$

Assume now that

$$d \leq k. \tag{5.3}$$

Clearly, we have

$$\begin{aligned}
&|\mathcal{Q}(m,d,N)\cap\mathcal{B}_i|\\
(5.4)\qquad &\quad\le |\{n : n\equiv m \pmod d,\quad iN+k-i\le n\le iN+2(k-i)\}|\\
&\quad\le \left[\frac{k-i}{d}\right]+1 \qquad \text{for } i=0,1,\dots,k.
\end{aligned}$$

Assume that $0\le i<j\le k$ and both $\mathcal{B}_i$ and $\mathcal{B}_j$ meet $\mathcal{Q}(m,d,N)$. Then the difference between the smallest element of $\mathcal{B}_j$ and the greatest element of $\mathcal{B}_i$ cannot exceed the difference between the greatest and smallest elements of $\mathcal{Q}(m,d,N)$:

$$(jN+k-j)-(iN+2(k-i))\le (N-1)d$$

whence, by (5.1),

$$j-i\le d+\frac{k-i}{N-1}\le d+\frac{k}{N-1}<d+1.$$

Moreover, if $j-i=d$, then denote the greatest element of $\mathcal{Q}(m,d,N)\cap\mathcal{B}_{i+d}$ (where $i+d=j$) by u. Then $v<u-d(N-1)$ implies that $v\notin\mathcal{Q}(m,d,N)$ since $u\in\mathcal{Q}(m,d,N)$, $u-v>d(N-1)$, and the greatest difference between two elements of $\mathcal{Q}(m,d,N)$ is $d(N-1)$. Thus we have

$$\begin{aligned}
&|\mathcal{Q}(m,d,N)\cap\mathcal{B}_i|+|\mathcal{Q}(m,d,N)\cap\mathcal{B}_{i+d}|\\
(5.5)\qquad &\le |\{n : n\equiv m \pmod d,\quad u-d(N-1)\le n\le iN+2(k-i)\}|\\
&\qquad + |\{n' : n'\equiv m \pmod d,\quad (i+d)N+k-(i+d)\le n'\le u\}|.
\end{aligned}$$

To each n' counted in the second term we may assign the integer $n=n'-d(N-1)$ which satisfies $n\equiv m \pmod d$ and $iN+k-i\le n\le u-d(N-1)$. Thus the sum estimated in (5.5) is

$$\begin{aligned}
&\le |\{n : n\equiv m \pmod d,\quad u-d(N-1)\le n\le iN+2(k-i)\}|\\
&\qquad + |\{n : n\equiv m \pmod d,\quad iN+k-i\le n\le u-d(N-1)\}|\\
&= |\{n : n\equiv m \pmod d,\quad iN+k-i\le n\le iN+2(k-i)\}|+1
\end{aligned}$$

(the last term 1 stands for $u-d(N-1)$ counted in both terms of the previous sum) and thus we have

$$\begin{aligned}
&|\mathcal{Q}(m,d,N)\cap\mathcal{B}_i|+|\mathcal{Q}(m,d,N)\cap\mathcal{B}_{i+d}|\\
(5.6)\qquad &\qquad\le \left(\left[\frac{k-i}{d}\right]+1\right)+1=\left[\frac{k-i}{d}\right]+2.
\end{aligned}$$

It follows from (5.4) and the discussion above that if $i_1<i_2<\dots<i_t$ denote the integers i with $\mathcal{Q}(m,d,N)\cap\mathcal{B}_i\ne\varnothing$, then either we have $t\le d$ and then

$$\begin{aligned}
\sum_{j=1}^{t}|\mathcal{Q}(m,d,N)\cap\mathcal{B}_{i_j}| &\le \sum_{j=1}^{t}\left(\left[\frac{k-i_j}{d}\right]+1\right)\\
&\le \sum_{j=0}^{t-1}\left(\left[\frac{k-j}{d}\right]+1\right)\le\sum_{j-0}^{d-1}\left[\frac{k-j}{d}\right]+d
\end{aligned}$$

or we have $t = d + 1$, $i_2 = i_1 + 1$, $i_3 = i_1 + 2$, $\ldots$, $i_t = i_{d+1} = i_1 + d$ and then, using also (5.6),

$$\begin{aligned}\sum_{j=1}^{t} |\mathcal{Q}(m,d,N) \cap \mathcal{B}_{i_j}| &= (|\mathcal{Q}(m,d,N) \cap \mathcal{B}_{i_1}| + |\mathcal{Q}(m,d,N) \cap \mathcal{B}_{i_t}|) \\ &\qquad + \sum_{j=2}^{t-1} |\mathcal{Q}(m,d,N) \cap \mathcal{B}_{i_j}| \\ &\leq \left(\left[\frac{k-i_1}{d}\right] + 2\right) + \sum_{j=2}^{t-1}\left(\left[\frac{k-i_j}{d}\right] + 1\right) \\ &= \sum_{j=1}^{t-1}\left(\left[\frac{k-i_j}{d}\right] + 1\right) + 1 \leq \sum_{j=0}^{t-2}\left(\left[\frac{k-j}{d}\right] + 1\right) + 1 \\ &= \sum_{j=0}^{d-1}\left[\frac{k-j}{d}\right] + (d+1).\end{aligned}$$

In both cases we have

$$\begin{aligned}|\mathcal{Q}(m,d,N) \cap k\mathcal{A}| &\leq \sum_{j=1}^{t} |\mathcal{Q}(m,d,N) \cap \mathcal{B}_{i_j}| \\ &\leq \sum_{j=0}^{d-1}\left[\frac{k-j}{d}\right] + (d+1).\end{aligned}$$

Define the integers q, r by $k = qd + r$, $0 \leq r < d$. Then, using (5.3), we have

$$\begin{aligned}\sum_{j=0}^{d-1}\left[\frac{k-j}{d}\right] + (d+1) &= \sum_{j=0}^{r}\left[\frac{k-j}{d}\right] + \sum_{j=r+1}^{d-1}\left[\frac{k-j}{d}\right] + (d+1) \\ &= (r+1)q + (d-1-r)(q-1) + (d+1) = qd + r + 2 = k + 2\end{aligned}$$

which proves that

$$E(N,k,\mathcal{A}) \leq k + 2. \tag{5.7}$$

To see that also

$$E(N,k,\mathcal{A}) \geq k + 2 \tag{5.8}$$

holds, observe that by (5.1) we have

$$\{k, k+1, \ldots, 2k, N+k-1\} \subset \{(k-1)+1, (k-1)+2, \ldots, (k-1)+N\} \cap k\mathcal{A}.$$

(5.2) follows from (5.7) and (5.8), and this completes the proof of the theorem. □

6. In this section it will be proved:

Theorem 9. — *If $N \in \mathbb{N}$, $i \in \mathbb{N}$, $k \in \mathbb{N}$,*

$$N > 4ki, \tag{6.1}$$

and we write $\mathcal{A} = \{1, 2, \dots, i, N-i+1, N-i+2, \dots, N\}$, then we have

$$E(N, k, \mathcal{A}) \leq ki + i.$$

For $N > 4ki$ this implies

$$E(N, k, \mathcal{A}) \leq ki + i = \left(\min(1, \frac{k+1}{2k} \cdot k\frac{|\mathcal{A}|}{N})\right) N$$

so that

Corollary 3. — *For all $k \in \mathbb{N}$, $k \geq 2$ we have*

$$C_\infty(k) \leq \frac{k+1}{2k}$$

whence

$$C_\infty \leq \frac{1}{2}.$$

Proof of Theorem 9. — Clearly we have

$$\begin{aligned} k\mathcal{A} &= k\{1, 2, \dots, i, N-i+1, N-i+2, \dots, N\} \\ &= k(\{0, N-i\} + \{1, 2, \dots, i\}) = k\{0, N-i\} + k\{1, 2, \dots, i\} \\ &= \{0, N-i, \dots, k(N-i)\} + \{k, k+1, \dots, ki\} \\ &= \bigcup_{j=0}^{k} \{j(N-i)+k, j(N-i)+k+1, \dots, j(N-i)+ki\} = \bigcup_{j=0}^{k} \mathcal{B}_j, \end{aligned}$$

where

$$\mathcal{B}_j = \{j(N-i)+k, j(N-i)+k+1, \dots, j(N-i)+ki\}.$$

Consider now an arithmetic progression $\mathcal{Q}(m, d, N) = \{m+d, m+2d, \dots, m+Nd\}$ with $m \in \mathbb{Z}$, $d \in \mathbb{N}$. We have to distinguish two cases. Assume first that

$$d \geq k+1. \tag{6.2}$$

Then we have

$$\begin{aligned} |\mathcal{Q}(m, d, N) \cap k\mathcal{A}| &= |\mathcal{Q}(m, d, N) \cap \bigcup_{j=0}^{k} \mathcal{B}_j| \\ &\leq \sum_{j=0}^{k} |\mathcal{Q}(m, d, N) \cap \mathcal{B}_j|. \end{aligned} \tag{6.3}$$

Here clearly we have

$$
\begin{aligned}
&|\mathcal{Q}(m,d,N)\cap\mathcal{B}_j| \\
(6.4)\qquad &\quad \leq |\{n : n\equiv m \pmod d,\quad j(N-i)k\leq n\leq j(N-i)+ki\}| \\
&\quad \leq \left[\frac{ki-k}{d}\right]+1 \text{ for } j=0,1,\dots,k.
\end{aligned}
$$

It follows from (6.2), (6.3) and (6.4) that

$$
\begin{aligned}
&|\mathcal{Q}(m,d,N)\cap k\mathcal{A}|\leq\sum_{j=0}^{k}\left(\left[\frac{ki-k}{d}\right]+1\right) \\
(6.5)\qquad &\quad =(k+1)\left(\left[\frac{ki-k}{d}\right]+1\right)\leq(k+1)\left(\frac{ki-k}{k+1}+1\right) \\
&\quad = ki+1 \text{ (for } d\geq k+1).
\end{aligned}
$$

Assume now that

$$(6.6)\qquad d\leq k.$$

Note that the assumption (6.2) was not used in the proof of (6.4) so that (6.4) holds also in this case.

Assume that $0\leq u<v\leq k$ and both $\mathcal{B}_u$ and $\mathcal{B}_v$ meet $\mathcal{Q}(m,d,N)$. Then the difference between the smallest element of $\mathcal{B}_v$ and the greatest element of $\mathcal{B}_u$ cannot exceed the difference between the greatest and smallest elements of $\mathcal{Q}(m,d,N)$:

$$(v(N-i)+k)-(u(N-i)+ki)\leq(N-1)d$$

whence, by (6.1) and (6.6),

$$v-u\leq d+\frac{(i-1)d+(ki-k)}{N-i}<d+\frac{2(ki-k)}{N/2}<d+1.$$

Moreover, if $v-u=d$, then denote the greatest element of $\mathcal{Q}(m,d,N)\cap\mathcal{B}_{u+d}$ (where $u+d=v$) by x. Then $y<x-d(N-1)$ implies that $y\notin\mathcal{Q}(m,d,N)$ since $x\in\mathcal{Q}(m,d,N), x-y>d(N-1)$, and the greatest difference between two elements of

$\mathcal{Q}(m,d,N)$ is $d(N-1)$. Thus we have

$$\begin{aligned}
&|\mathcal{Q}(m,d,N)\cap\mathcal{B}_u| + |\mathcal{Q}(m,d,N)\cap\mathcal{B}_{u+d}| \\
&\le |\{n : n\equiv m \pmod d,\quad x-(N-1)d\le n\le u(N-i)+ki\}| \\
&\qquad + |\{n : n\equiv m \pmod d,\quad (u+d)(N-i)+k\le n\le x\}| \\
&\le \left(\left[\frac{(u(N-i)+ki)-(x-(N-1)d)}{d}\right]+1\right) \\
&\qquad + \left(\left[\frac{x-((u+d)(N-i)+k)}{d}\right]+1\right) \qquad (6.7)\\
&\le \frac{u(N-i)+ki-x+(N-1)d}{d} \\
&\qquad + \frac{x-u(N-i)-d(N-i)-k}{d}+2 \\
&= i+1+\frac{ki-k}{d}.
\end{aligned}$$

It follows from (6.4), (6.6) and the discussion above that if $j_1<j_2<\cdots<j_t$ denote the integers j with

$$\mathcal{Q}(m,d,N)\cap\mathcal{B}_j\neq\varnothing,$$

then either we have $t\le d$ and then

$$\begin{aligned}
\sum_{\ell=1}^{t}|\mathcal{Q}(m,d,N)\cap\mathcal{B}_{j_\ell}| &\le \sum_{\ell=1}^{t}\left(\left[\frac{ki-k}{d}\right]+1\right) \\
&\le d\left(\left[\frac{ki-k}{d}\right]+1\right)\le ki-k+d\le ki,
\end{aligned}$$

or we have $t=d+1$, $j_2=j_1+1$, $j_3=j_1+2,\dots,j_t=j_{d+1}=j_1+d$ and then, using also (6.7),

$$\begin{aligned}
\sum_{\ell=1}^{t}|\mathcal{Q}(m,d,N)\cap\mathcal{B}_{j_\ell}| &= (|\mathcal{Q}(m,d,N)\cap\mathcal{B}_{j_1}|+|\mathcal{Q}(m,d,N)\cap\mathcal{B}_{j_t}|) \\
&\qquad + \sum_{\ell=2}^{t-1}|\mathcal{Q}(m,d,N)\cap\mathcal{B}_{j_\ell}| \\
&\le i+1+\frac{ki-k}{d}+(t-2)\left(\left[\frac{ki-k}{d}\right]+1\right) \\
&\le i+t-1+(t-1)\frac{ki-k}{d}=i+d+ki-k\le ki+i.
\end{aligned}$$

In both cases we have

$$|\mathcal{Q}(m,d,N)\cap k\mathcal{A}|\le\sum_{\ell=1}^{t}|\mathcal{Q}(m,d,N)\cap\mathcal{B}_{j_\ell}|\le ki+i$$

which completes the proof of the theorem. □

7. One might like to make a guess on the values of the constants C and C_∞. Suggested by the results above, I would risk two conjectures:

(i) we have
$$C < C_\infty$$
(this is, perhaps, not quite hopeless);

(ii) we have
$$C_\infty = \frac{1}{2};$$
this seems to the closest finite analog of the $\alpha + \beta$ conjecture but probably it will not be easy to prove it.

On the other hand, I have no idea whether Corollary 2 gives the best possible upper bound for C, i.e., we have $C = 1/3$; it is quite possible that (perhaps, using computers) one can find a set $\mathcal{A}$ whose study leads to an upper bound smaller than 1/3.

8. In the rest of this paper, I will study another extension of the classical Schnirelmann-Khintchin-Mann-Kneser theory of addition theorems. Namely, in this theory as well as in the finite case studied above, our basic problem is the following: we start out from a set $\mathcal{A}$ whose density in a certain sense is $\geq \delta(> 0)$ and then our goal is to give a lower bound for the density of $k\mathcal{A}$ in terms of k and δ. (This lower bound is usually $k\delta$ or, at least, $ck\delta$.) In particular, how large k is needed to be to ensure that the density of $k\mathcal{A}$ should be 1 ? (Khintchin's theorem (3.6) and my Theorem 3 above are typical results of this type). This problem can be generalized in the following way:

Suppose we start out from a set $\mathcal{B}$ known to be a basis, like the set of the primes or k-th powers. What happens if we take a subset $\mathcal{A}$ of $\mathcal{B}$ whose "density relative to $\mathcal{B}$" is $\geq 1/k$ (where $k \in \mathbb{N}$, $k \geq 2$), i.e., we take $\geq 100/k$ percent of the elements of $\mathcal{B}$ as $\mathcal{A}$? What additional condition is needed to ensure that $\mathcal{A}$ should form a basis, and if such a condition holds, then what upper bound can be given for the order of the basis $\mathcal{A}$ in terms of k and the order of the basis $\mathcal{B}$? The difficulty is that usually one needs a coprimality condition concerning the set $\mathcal{A}$. The most interesting problem of this type is when $\mathcal{B}$ consists of the primes, namely, then no coprimality condition is needed. Thus here we shall restrict ourselves to this special case. In other words, the problem is the following:

Assume that $k \in \mathbb{N}$, $k \geq 2$ and $\mathcal{P}$ is an infinite set of primes with the property that
$$\liminf_{n\to+\infty} \frac{P(n)}{\pi(n)} \geq \frac{1}{k}. \tag{8.1}$$
Then by Schnirelman's method, it can be shown that $\{0\} \cup \mathcal{P}$ is an asymptotic basis of finite order. Let $H = H(k)$ denote the smallest integer h such that for every set $\mathcal{P}$ of primes satisfying (8.1), $\{0\} \cup \mathcal{P}$ is an asymptotic basis of order $\leq h$ (i.e., H is the smallest integer such that for every $\mathcal{P}$ satisfying (8.1), every large integer can be represented as the sum of at most H elements of $\mathcal{P}$). The problem is to estimate H in terms of k. It will be proved that

Theorem 10. — *For all $k \in \mathbb{N}$ we have*

$$c_1 k \log\log(k+2) < H(k) < c_2 k^4. \tag{8.2}$$

Probably the lower bound gives the right order of magnitude of $H(k)$; unfortunately, I have not been able to prove this. Moreover, we remark that a finite analog of Theorem 10 (a theorem covering finite sets $\mathcal{P}$ of primes) could be proved as well but it would be much more complicated; thus we restrict ourselves to the much simpler infinite case.

Proof. — First we will prove the lower bound in (8.2). We will show that if c_3 is a small positive constant to be fixed later, then for every $k \in \mathbb{N}$ there is a positive integer m such that

$$m > c_3 k \log\log(k+2) \tag{8.3}$$

and

$$\varphi(m) \le k. \tag{8.4}$$

Indeed, denote the i-th prime by q_i, and define t by

$$q_1 q_2 \dots q_t \le c_3 k \log\log(k+2) < q_1 q_2 \dots q_{t+1}. \tag{8.5}$$

(If $c_3 k \log\log(k+2) < 2$, then (8.3) and (8.4) hold with $m = 2$.) By the prime number theorem, it follows from (8.5) that

$$q_t = (1 + o(1)) \log k \qquad (\text{for } k \to +\infty). \tag{8.6}$$

Define u by

$$u = \left[\frac{c_3 k \log\log(k+2)}{q_1 q_2 \dots q_{t-1}}\right]$$

and let

$$m = q_1 q_2 \dots q_{t-1}(u+1). \tag{8.7}$$

Then (8.3) holds trivially. Moreover, for $k \to +\infty$ clearly we have

$$m = (1 + o(1)) c_3 k \log\log(k+2). \tag{8.8}$$

By Mertens' formula, it follows from (8.6), (8.7) and (8.8) that for $k > k_o$ (where k_o may depend also on c_3)

$$\begin{aligned}\varphi(m) = m \prod_{p|m}(1 - \frac{1}{p}) &\le m \prod_{i=1}^{t-1}(1 - \frac{1}{q_i}) \\ &< c_4 \frac{m}{\log q_{t-1}} < 2c_4 \frac{m}{\log\log(k+2)} \\ &< 3c_4 \frac{c_3 k \log\log(k+2)}{\log\log(k+2)} = 3c_3 c_4 k\end{aligned}$$

so that (8.4) holds if we choose $c_3 = 1/3c_4$ and $k > k_o$. Finally, if $k \le k_o$, then (8.3) and (8.4) hold with $m = 1$ at the expense of replacing the constant c_3 computed above by another smaller constant (small enough in terms of k_o) and this proves the existence of a number m satisfying (8.3) and (8.4).

Now define $\mathcal{P}$ by

$$\mathcal{P} = \{p : p \text{ prime}, p \equiv 1 \pmod m\}.$$

Then by the prime number theorem for the arithmetic progressions of small moduli, it follows from (8.4) that for $n \to +\infty$ we have

$$P(n) = (1 + o(1))\frac{\pi(n)}{\varphi(m)} > (1 + o(1))\frac{\pi(n)}{k}$$

which proves (8.1). Moreover, if $v < m$, then $v(\{0\} \cap \mathcal{P})$ does not contain the positive multiples of m, thus if the order of the asymptotic basis $\{0\} \cup \mathcal{P}$ is h, then, by (8.3), we have

$$h \geq m > c_3 k \log\log(k + 2)$$

which proves the lower bound in (8.2).

To prove the upper bound, we need two lemmas.

Lemma 1. — *There is an absolute constant c_5 such that if $\mathcal{P}$ is a set of primes satisfying (8.1) and N is large enough depending on $\mathcal{P}$, then we have*

$$S(\mathcal{P}, N) > c_5 N k^{-4} \tag{8.9}$$

where $S(\mathcal{P}, N)$ denotes the counting function of the set $S(\mathcal{P}) = \mathcal{P} + \mathcal{P}$.

Proof of Lemma 1. — Let $R(n)$ denote the number of pairs (p, q) of primes with

$$p + q = n$$

so that, by Brun's sieve (see, *e.g.*, [15, p. 80]), for $n \in \mathbb{N}$, $n > 1$ we have

$$R(n) < c_6 \prod_{p|n}(1 + \frac{1}{p})\frac{n}{(\log n)^2}. \tag{8.10}$$

Moreover, denote the number of solutions of

$$p + q = n, \quad p \in \mathcal{P}, \ q \in \mathcal{P}$$

by $r(\mathcal{P}, n)$.

By (8.1) and the prime number theorem, for sufficiently large N we have

$$\begin{aligned}\sum_{n=1}^{N} r(\mathcal{P}, n) &= \sum_{n=1}^{N} |\{(p, q) : p + q \leq N, \ p, q \in \mathcal{P}\}| \\ &\geq \sum_{n=1}^{N} |\{(p, q) : p, q \leq N/2, \ p, q \in \mathcal{P}\}| = (P([N/2]))^2 \\ &\geq (1 + o(1))\frac{1}{k^2}(\pi([N/2]))^2 > \frac{1}{5k^2}\frac{N^2}{(\log N)^2}.\end{aligned}$$

Thus by Cauchy's inequality we have

$$\tag{8.11}\begin{aligned}\sum_{n=1}^{N} r^2(\mathcal{P},n) &\geq \left(\sum_{n=1}^{N} r(\mathcal{P},n)\right)^2 |\{n : n \in S(\mathcal{P}),\ n \leq N\}|^{-1}\\ &> \frac{1}{25k^4}\frac{N^4}{(\log N)^4}\frac{1}{S(\mathcal{P},N)}.\end{aligned}$$

On the other hand, by (8.10) we have

$$\tag{8.12}\begin{aligned}\sum_{n=1}^{N} r^2(\mathcal{P},n) \leq \sum_{n=1}^{N} R^2(n) &< c_7\frac{N^2}{(\log N)^4}\sum_{n=1}^{N}\prod_{p|n}(1+\frac{1}{p})^2\\ &< c_8\frac{N^2}{(\log N)^4}\sum_{n=1}^{N}\prod_{p|n}(1+\frac{2}{p})\\ &= c_8\frac{N^2}{(\log N)^4}\sum_{n=1}^{N}\sum_{d|n,|\mu(d)|=1}\frac{2^{\omega(d)}}{d}\\ &= c_8\frac{N^2}{(\log N)^4}\sum_{d\leq N,|\mu(d)|=1}\frac{2^{\omega(d)}}{d}[\frac{N}{d}]\\ &\leq c_8\frac{N^3}{(\log N)^4}\sum_{|\mu(d)|=1}\frac{2^{\omega(d)}}{d^2}\\ &= c_8\frac{N^3}{(\log N)^4}\prod_{p}(1+\frac{2}{p^2}) < c_9\frac{N^3}{(\log N)^4}.\end{aligned}$$

(8.9) follows from (8.11) and (8.12), and this completes the proof of Lemma 1. □

Lemma 2. — *If $\ell \in \mathbb{N}$, and $\mathcal{A}$ is an infinite set of non-negative integers such that $0 \in \mathcal{A}$ and*

$$\underline{d}(\ell\mathcal{A}) < \ell\underline{d}(\mathcal{A}),$$

then there is a set $\mathcal{E}$ and a number g such that

$$\tag{8.14}\mathcal{E} \subset \ell\mathcal{A},$$

$$\tag{8.15}0 \in \mathcal{E},$$

there is a number n_o such that

$$\tag{8.16}e \in \mathcal{E},\ e' \equiv e \pmod g,\ e' \geq n_o \text{ imply } e' \in \mathcal{E}$$

(so that $[n_o, +\infty) \cap \mathcal{E}$ is the union of the intersection of certain modulo g residue classes, including the 0 residue class, with $[n_o, +\infty)$) and

$$\tag{8.17}d(\mathcal{E}) \geq \ell\underline{d}(\mathcal{A}) - \frac{\ell}{g}.$$

Proof of lemma 2. — This follows from Kneser's theorem [18] and, indeed, it is a special case of [16, p. 57, Theorem 19].

To complete the proof of the upper bound in (8.2), first we use Lemma 2 with $\ell = [\frac{2}{c_5}k^4]+1$ (where c_5 is the constant in Lemma 1) and with $S(\{0\}\cup\mathcal{P}) = 2(\{0\}\cup\mathcal{P})$ in place of $\mathcal{A}$. By Lemma 1 we have

$$\underline{d}(S(\{0\}\cup\mathcal{P})) \geq \liminf_{N\to+\infty} \frac{S(\mathcal{P},N)}{N} \geq c_5 k^{-4}.$$

Thus we have

$$\ell\underline{d}(\{0\}\cup S(\mathcal{P})) \geq \left(\left[\frac{2}{c_5}k^4\right]+1\right)c_5 k^{-4} > 2 \tag{8.18}$$

so that (8.13) certainly holds thus, indeed, Lemma 2 can be applied. By (8.17) and (8.18), we have

$$1 \geq d(\mathcal{E}) \geq \ell\underline{d}(S(\{0\}\cup\mathcal{P})) - \frac{\ell}{g} > 2 - \frac{\ell}{g}$$

whence

$$g < \ell. \tag{8.19}$$

Now it will be proved that every large integer n can be represented in the form

$$p_1 + p_2 + \cdots + p_u = n \text{ with } p_1, p_2, \ldots p_u \in \mathcal{P},\ u \leq 3l-2. \tag{8.20}$$

Indeed, let p' denote the smallest prime with

$$p' > g, \quad p' \in \mathcal{P}, \tag{8.21}$$

and assume that

$$n \geq n_o + (g-1)p' \tag{8.22}$$

where n_o is defined by (8.16). By (8.21) we have $(p', g) = 1$, thus there is an integer i such that

$$n - ip' \equiv 0 \pmod{g} \tag{8.23}$$

and

$$0 \leq i \leq g-1. \tag{8.24}$$

By (8.22) and (8.24) we have

$$n - ip' \geq (n_o + (g-1)p') - (g-1)p' = n_o. \tag{8.25}$$

It follows from (8.14), (8.15), (8.16), (8.23) and (8.25) that

$$n - ip' \in \mathcal{E} \subset \ell\mathcal{A} = \ell S(\{0\}\cup\mathcal{P}) = (2\ell)(\{0\}\cup\mathcal{P})$$

so that there are primes $p_1, p_2, \ldots, p_v$ with

$$n - ip' = p_1 + p_2 + \cdots + p_v, \quad p_1, p_2, \ldots p_v \in \mathcal{P} \tag{8.26}$$

and

$$v \leq 2\ell. \tag{8.27}$$

(8.26) can be rewritten in the form

$$p_1 + p_2 + \cdots + p_v + ip' = n.$$

This is a representation of the form (8.20) where, by (8.19), (8.24) and (8.27), the number of the terms on the left hand side is

$$u = v + i \le 2\ell + g - 1 \le 3\ell - 2.$$

Thus every integer n satisfying (8.22) has a representation in form (8.20). It follows that $\{0\} \cup \mathcal{P}$ is an asymptotic basis of order

$$h \le 3\ell - 2 = 3\left(\left[\frac{2}{c_5}k^4\right] + 1\right) - 2 < c_{10}k^4$$

which proves the upper bound in (8.2). □

References

[1] N. Alon, *Subset sums*, J. Number Theory, **27**, 1987, 196–205.

[2] N. Alon and G. Freiman, *On sums of subsets of a set of integers*, Combinatorica, **8**, 1988, 297–306.

[3] F. Dyson, *A theorem on the densities of sets of integers*, J. London Math. Soc., **20**, 1945, 8–14.

[4] P. Erdős, *On the representation of large integers as sums of distinct summands taken from a fixed set*, Acta. Arith., **7**, 1961/62, 345–354.

[5] P. Erdős and G. Freiman, *On two additive problems*, J. Number Theory, **34**, 1990, 1–12.

[6] P. Erdős, J.-L. Nicolas and A. Sárközy, *On the number of partitions of n without a given subsum*, II, Analytic Number Theory, Proceedings of a Conference in Honor of P. T. Bateman, B. C. Berndt et al. eds., Birkhäuser, Boston-Basel-Berlin, 1990, 205–234.

[7] P. Erdős, J.-L. Nicolas and A. Sárközy, *On the number of pairs of partitions of n without common subsums*, Colloquium Math., **63**, 1992, 61–83.

[8] P. Erdős and A. Sárközy, *On a problem of Straus*, Disorder in Physical Systems (a volume in Honour of John M. Hammersley), G. R. Grimmett and D. J. Welsh eds., Clarendon Press, Oxford, 1990, 55–66.

[9] P. Erdős and A. Sárközy, *Arithmetic progression in subset sums*, Discrete Mathematics, **102**, 1992, 249–264.

[10] P. Erdős, A. Sárközy and C. L. Stewart, *On prime factors of subset sums*, J. London Math. Soc., **49**, 1994, 209–218.

[11] J. Folkman, *On the representation of integers as sums of distinct terms from a fixed sequence*, Canadian J. Math, **18**, 1966, 643–655.

[12] G. A. Freiman, *Foundations of a Structural Theory of Set Additions*, Translations of Mathematical Monographs, **37**, Amer. Math. Soc., Providence, RI.

[13] G. A. Freiman, *New analytical results in subset-sum problem*, Discrete Mathematics, **114**, 1993, 205–218.

[14] G. A. Freiman, *Sumsets and powers of 2*, Coll. Math. Soc. J. Bolyai, **60**, 1992, 279–286.

[15] H. Halberstam and H.-E. Richert, *Sieve Methods*, Academic Press, 1974.

[16] H. Halberstam and K.F. Roth, *Sequences*, Springer Verlag, Berlin, 1983.

[17] A. Khintchin, *Zur additiven Zahlentheorie*, Math. Sb. N.S., **39**, 1932, 27–34.

[18] M. Kneser, *Abschätzungen der asymptotischen Dichte von Summenmengen*, Math. Z, **58**, 1953, 459–484.

[19] H. B. Mann, *A proof of the fundamental theorem on the density of sums of sets of positive integers*, Ann. Math., **43**, 1942, 523–527.

[20] M. B. Nathanson and A. Sárkőzy, *Sumsets containing long arithmetic progressions and powers of 2*, Acta Arith., **54**, 1989, 147–154.

[21] A. Sárkőzy, *Finite addition theorems, I*, J. Number Theory, **32**, 1989, 114–130.

[22] A. Sárkőzy, *Finite addition theorems, II*, J. Number Theory, **48**, 1994, 197–218.

[23] A. Sárkőzy, *Finite addition theorems, III*, Publ. Math. d'Orsay, **92–01**, 105–122

[24] L. G. Schnirelmann *Über additive Eigenschaften von Zahlen*, Annals Inst. Polyt. Novocherkassk, **14**, (1930), 3–28; Math. Annalen, **107**, 1933, 649–90.

A. Sárkőzy, Department of Algebra and Number Theory, Eőtvős Loránd University, Rákóczi út 5, H-1088 Budapest, Hungary • *E-mail :* sarkozy@cs.elte.hu

JOHN STEINIG

On Freiman's theorems concerning the sum of two finite sets of integers

Astérisque, tome 258 (1999), p. 129-140

<http://www.numdam.org/item?id=AST_1999__258__129_0>

Astérisque **258**, 1999, p. 129–140

ON FREIMAN'S THEOREMS CONCERNING THE SUM OF TWO FINITE SETS OF INTEGERS

by

John Steinig

Abstract. — Details are provided for a proof of Freiman's theorems [1] which bound $|M+N|$ from below, where M and N are finite subsets of $\mathbb{Z}$.

1. Introduction

If M and N are subsets of $\mathbb{Z}$, their sum $M+N$ is the set

$$M+N := \{x \in \mathbb{Z} : x = b+c,\ b \in M,\ c \in N\}.$$

If a set $E \subset \mathbb{Z}$ is finite and non-empty, its cardinality will be denoted by $|E|$, and its largest and smallest element by $\max(E)$ and $\min(E)$, respectively. If A is some collection of integers, say $a_1, \dots, a_k$, not all zero, their greatest common divisor will be denoted by $(a_1, \dots, a_k)$, or by $\gcd(A)$.

Now let M and N be finite sets of non-negative integers, such that $0 \in M \cap N$, say

$$M = \{b_0, \dots, b_{m-1}\} \quad \text{with} \quad b_0 = 0 \quad \text{and} \quad b_i < b_{i+1} \quad (\text{all } i) \tag{1.1}$$

and

$$N = \{c_0, \dots, c_{n-1}\} \quad \text{with} \quad c_0 = 0 \quad \text{and} \quad c_i < c_{i+1} \quad (\text{all } i). \tag{1.2}$$

It is easily seen that

$$|M+N| \geq |M| + |N| - 1 \tag{1.3}$$

(consider $b_0, \dots, b_{m-1}$, $b_{m-1}+c_1, \dots, b_{m-1}+c_{n-1}$) .

The following two theorems of Freiman's [**1**] give a better lower bound for $|M+N|$, when additional conditions are imposed on M and N.

Theorem X. *Let M and N be finite sets of non-negative integers with $0 \in M \cap N$, as in (1.1) and (1.2). If*

$$c_{n-1} \leq b_{m-1} \leq m+n-3 \tag{1.4}$$

or

$$c_{n-1} < b_{m-1} = m+n-2, \tag{1.5}$$

1991 ***Mathematics Subject Classification.*** — 11 B 13.
Key words and phrases. — Inverse theorems, sumsets of integers.

then

$$|M+N| \geq b_{m-1} + n\,. \tag{1.6}$$

If

$$c_{n-1} = b_{m-1} \leq m+n-3\,, \tag{1.7}$$

then

$$|M+N| \geq b_{m-1} + \max(m,n)\,. \tag{1.8}$$

Theorem XI. *Let M and N be finite sets of non-negative integers with $0 \in M \cap N$, as in (1.1) and (1.2). If*

$$\max(b_{m-1}, c_{n-1}) \geq m+n-2 \tag{1.9}$$

and

$$(b_1, \ldots, b_{m-1}, c_1, \ldots, c_{n-1}) = 1\,, \tag{1.10}$$

then

$$|M+N| \geq m+n-3+\min(m,n)\,. \tag{1.11}$$

We remark here that if $\min(m,n) \geq 2$, then any sets M and N which satisfy (1.4) or (1.5) also satisfy (1.10). In fact, either of these conditions implies that $\gcd(M) = 1$ or $\gcd(N) = 1$. For if $\gcd(M) > 1$, then M contains neither 1, nor any pair of consecutive positive integers; that is, $b_\nu - b_{\nu-1} \geq 2$ for $\nu = 1, \ldots, m-1$. Hence, by summing up, $b_{m-1} \geq 2m-2$. Similarly, $c_{n-1} \geq 2n-2$ if $\gcd(N) > 1$. And these two lower bounds are incompatible if (1.4) or (1.5) holds.

Interesting applications of these two theorems to the study of sum-free sets of positive integers are given in **[2]** and **[3]**.

The proof of Theorem XI in **[1]** is presented very succinctly, but divides the argument into many cases and is in fact quite long once the necessary details are provided. The aim of this paper is to give a detailed proof, separated into fewer cases than in **[1]**. As in **[1]**, one proceeds by induction on $m+n$ and distinguishes two situations (called here, and there, Cases (I) and (II)), essentially according to the size of $\max(b_{m-2}, c_{n-2})$.

Inequality (2.11) and Theorem 2.1 (below) are essential tools, here and in **[1]**. Case (I) requires fewer subcases here than in **[1]**, and uses an argument which is applied again at the end of Case (II). Case (II) has been simplified by avoiding consideration of the sign of $b_p - c_p$ (cf. **[1]**, after (26)), and of $m - p_1 - p_1^*$ (**[1]**, after (29)).

For completeness, Theorem X is also proved, since it is used to prove Theorem XI. We follow **[1]** here, but the formulation of Theorem X given above differs from Freiman's in including (1.5) and (1.7), which in **[1]** are embodied in the proof of Theorem XI.

I am grateful to Felix Albrecht, who helped me by translating **[1]** into English.

2. Preliminaries

We now introduce some more notation and three auxiliary results.

Part of the proof of Theorem XI exploits a certain symmetry between M and N and the sets

$$M^* := \{b_{m-1} - b_\nu\}_{\nu=0}^{m-1}, \tag{2.1}$$

and

$$N^* := \{c_{n-1} - c_\nu\}_{\nu=0}^{n-1}, \tag{2.2}$$

which we also write as

$$M^* = \{x_0, x_1, \dots, x_{m-1}\}, \quad \text{with} \quad x_\nu = b_{m-1} - b_{m-1-\nu}, \tag{2.3}$$

and

$$N^* = \{y_0, y_1, \dots, y_{n-1}\}, \quad \text{with} \quad y_\nu = c_{n-1} - c_{n-1-\nu} \tag{2.4}$$

($x_0 = 0$, $x_{m-1} = b_{m-1}$ and $x_i < x_{i+1}$ for all i; $y_0 = 0$, $y_{n-1} = c_{n-1}$ and $y_i < y_{i+1}$ for all i).

The hypotheses of Theorem XI are met by M^* and N^* if they are by M and N, because

$$(b_{m-1} - b_{m-2}, \dots, b_{m-1} - b_1, b_{m-1}) = (b_1, \dots, b_{m-1}), \tag{2.5}$$

$|M^*| = |M|$, $|N^*| = |N|$ and $\max(x_{m-1}, y_{n-1}) = \max(b_{m-1}, c_{n-1})$. And the theorem's conclusion holds for $|M+N|$ if it does for $|M^* + N^*|$, since the two are equal.

For any r and s with $0 \le r \le m$ and $0 \le s \le n$, let

$$M'_r := \{b_i \in M : i \le r-1\}, \; N'_s := \{c_i \in N : i \le s-1\}, \tag{2.6}$$

and

$$(M^*)'_r := \{x_i \in M^* : i \le r-1\}, \; (N^*)'_s := \{y_i \in N^* : i \le s-1\}.$$

Theorem XI is proved by induction. Typically, one writes $M = M'_r \cup (M \backslash M'_r)$, then subtracts from each element of $M \backslash M'_r$ its smallest element, b_r, in order to obtain a set with the same cardinality, which contains 0. This set is, for $0 \le r \le m-1$,

$$M''_{m-r} := \{0, b_{r+1} - b_r, \dots, b_{m-1} - b_r\} = \{b_\nu - b_r\}_{\nu=r}^{m-1}, \tag{2.7}$$

and the corresponding set for $N \backslash N'_s$ is

$$N''_{n-s} := \{0, c_{s+1} - c_s, \dots, c_{n-1} - c_s\} = \{c_\nu - c_s\}_{\nu=s}^{n-1}. \tag{2.8}$$

For any r and s with $0 \le r < m$ and $0 \le s < n$, we have

$$|M''_{m-r}| = m - r \qquad \text{and} \qquad |N''_{n-s}| = n - s. \tag{2.9}$$

Many of the estimates involving these sets will be combined with the following elementary inequality: if E_1 and E_2 are subsets of the finite set E, then

$$|E| \ge |E_1| + |E_2| - |E_1 \cap E_2|. \tag{2.10}$$

We shall use the following form of (2.10): if $k \le r \le m-1$ and $\ell \le s \le n-1$, then

$$|M+N| \ge |M'_r + N'_s| + |M''_{m-k} + N''_{n-\ell}| - |(M'_r + N'_s) \cap ((M \backslash M'_k) + (N \backslash N'_\ell))|. \tag{2.11}$$

To obtain (2.11), set $E = M + N$, $E_1 = M'_r + N'_s$ and $E_2 = (M \backslash M'_k) + (N \backslash N'_\ell)$ in (2.10), and observe that

$$M''_{m-k} + N''_{n-\ell} = \{x \in \mathbb{Z} : x = b_u + c_v - (b_k + c_\ell),\; k \le u \le m-1,\; \ell \le v \le n-1\},$$

so that if x runs through the elements of $M''_{m-k} + N''_{n-\ell}$, then $x + (b_k + c_\ell)$ runs through those of E_2; consequently

$$|M''_{m-k} + N''_{n-\ell}| = |\{x \in \mathbb{Z} : x = b_u + c_v,\ k \le u \le m-1\,,\ \ell \le v \le n-1\}|\,. \quad (2.12)$$

From (2.10) and (2.12) we get (2.11).

The following property of the counting functions

$$B(s) := |\{b_i \in M : 1 \le b_i \le s\}|\,,\ C(s) := |\{c_i \in N : 1 \le c_i \le s\}| \quad (2.13)$$

follows from Mann's inequality ([**4**], Chap. I.4; [**5**]); we will apply it to choose the parameters in (2.11).

Theorem 2.1. *If* $B(s) + C(s) \ge s$ *for* $s = 1, \dots, k$, *then* $\{0, 1, \dots, k\} \subset M + N$.

We will use the following proposition in establishing Case (II) of Theorem XI. Its proof is suggested by an argument of Freiman's ([**1**], p. 152). There is an arithmetical hypothesis, different from (1.10), but no condition on the size of $\max(M \cup N)$. The conclusion is stronger than (1.11).

Proposition 2.2. *If* M *and* N *are finite subsets of* $\mathbb{Z}$, *such that* $0 \in M \cap N$, $|M| \ge 2$, $|N| \ge 2$ *and* $\gcd(N) \nmid \gcd(M)$, *then*

$$|M + N| \ge |M| + 2|N| - 2\,. \quad (2.14)$$

Proof. — Set $d := \gcd(N)$, and $N_0 := N\backslash\{0\}$. Since $0 \in M$ and $d \nmid \gcd(M)$, some, but not all elements of M are divisible by d. Let b_r and b_s be the largest integers in M such that, respectively, $b_r \equiv 0$ and $b_s \not\equiv 0 \pmod d$. Then M, $\{b_r\} + N_0$ and $\{b_s\} + N_0$ are pairwise disjoint subsets of $M + N$ (for instance, $b = b_r + c$ for some $b \in M$ and $c \in N_0$ would imply both $b \equiv 0 \pmod d$ and $b \ge b_r + 1$). This proves (2.14).

Corollary 2.3. *Let* M *and* N *be as in (1.1) and (1.2), and such that (1.10) holds. Assume also that* $\min(m, n) \ge 3$. *Then (1.11) is true, if any one of the following conditions is satisfied*:

$$\gcd(M) > 1\,, \quad (2.15)$$

$$\gcd(M'_{m-1}) > 1\,, \quad (2.16)$$

$$\gcd((M^*)'_{m-1}) > 1\,. \quad (2.17)$$

Proof. — Because of (1.10), $\gcd(M) \nmid \gcd(N)$ if $\gcd(M) > 1$; and then $|M + N| \ge m + n - 2 + \min(m, n)$, by (2.14). Thus (1.11) follows from (1.10) and (2.15).

Now suppose that (2.16) is verified. We may assume that $\gcd(N) = 1$, for if not, (1.11) is true (exchange M and N in Proposition 2.2 and argue as above). Then, $\gcd(M'_{m-1}) \nmid \gcd(N)$ and by Proposition 2.2,

$$|M'_{m-1} + N| \ge 2\,(m-1) + n - 2 \ge m + n - 4 + \min(m, n)\,.$$

This implies (1.11), since $b_{m-1} + c_{n-1} \notin M'_{m-1} + N$.

Finally, (1.10) and (2.5) imply that $(x_1, \dots, x_{m-1}, y_1, \dots, y_{n-1}) = 1$. The preceding arguments then show that (2.17) implies (1.11) for M^* and N^*, hence also for M and N.

3. Freiman's Theorems

3.1. Proof of Theorem X. — Consider the sets

$$A := \{b_0, \ldots, b_{m-1}, b_{m-1} + c_1, \ldots, b_{m-1} + c_{n-1}\}$$

and

$$B := \{g \in \mathbb{Z} : 1 \leq g < b_{m-1}, g \notin M\}.$$

Since $A \subset (M + N)$ and $|A| + |B| = b_{m-1} + n$, (1.6) is true if $B = \phi$. If $B \neq \phi$, (1.6) is proved by constructing an injective mapping, say f, of B into $(M + N)\backslash A$, as follows. Let $g \in B$.

If $g \in N$, then $g \in M + N$; $g \notin A$, since $A \cap B = \phi$. In this case, set $f(g) = g$.

If $g \notin N$, if $c_{n-1} < b_{m-1}$ and $c_{n-1} < g < b_{m-1}$, then the n integers

$$g - c_0, \ g - c_1, \ldots, \ g - c_{n-1} \tag{3.1}$$

are in the interval $[1, b_{m-1})$. Since $|B| = b_{m-1} - (m - 1) \leq n - 1$, some integer in (3.1) belongs to M, say $g - c_s = b_r$, whence $g = b_r + c_s \in M + N$. As before, $g \notin A$. Here also, set $f(g) = g$.

If $g \notin N$ and $g < c_{n-1}$, let $i \, (0 \leq i \leq n - 2)$ be such that $c_i < g < c_{i+1}$. The $n - 1$ integers

$$g + b_{m-1} - c_\nu \ (\nu = i + 1, \ldots, n - 2), \ g - c_\nu \ (\nu = 0, \ldots, i) \tag{3.2}$$

are distinct $(g + b_{m-1} - c_{n-2} > g = g - c_0)$, and in $[1, b_{m-1})$. If $b_{m-1} - (m-1) \leq n-2$, as in (1.4), one of them must belong to M. If $b_{m-1} - (m-1) = n-1$ and $c_{n-1} < b_{m-1}$ as in (1.5), we may include $g + b_{m-1} - c_{n-1}$ in (3.2) since $g + b_{m-1} - c_{n-1} > g$ in this case, and reach the same conclusion. Hence g or $g + b_{m-1}$ is in $M + N$. Neither is in A; $g \notin A$ as before, and $g + b_{m-1} \notin A$ since $g + b_{m-1} > b_{m-1}$ and $g \notin N$. We set $f(g) = g$, or $f(g) = g + b_{m-1}$, so as to have $f(g) \in M + N$.

This f is injective. Indeed, $f(g) = g$ or $f(g) = g + b_{m-1}$ for each $g \in B$; and if $g < g' < b_{m-1}$ then $g < g' < g + b_{m-1} < g' + b_{m-1}$.

This concludes the proof of (1.6). And (1.8) now follows on observing that if $b_{m-1} = c_{n-1}$ in (1.4), the roles of M and N may be exchanged.

3.2. Proof of Theorem XI. — The proof proceeds by induction on $m + n$. Since (1.3) implies (1.11) if $\min(m, n) \leq 2$, we may assume that $\min(m, n) \geq 3$. We shall show that (1.11) is true for M and N, if it is true for all finite sets A and B of non-negative integers which are such that

$$|A| + |B| < m + n, \tag{3.3}$$

$$0 \in A \cap B, \tag{3.4}$$

$$\gcd(A \cup B) = 1, \tag{3.5}$$

and

$$\max(A \cup B) \geq |A| + |B| - 2. \tag{3.6}$$

We consider separately the two cases

(I) $\max(b_{m-2}, \ c_{n-2}) < m + n - 4,$ (3.7)

(II) $\max(b_{m-2}, \ c_{n-2}) \geq m + n - 4.$ (3.8)

We first deal with

Case (I). Clearly, (3.7) implies that $M \cap N \neq \{0\}$. We proceed to make this remark more precise.

Let B and C be the counting functions defined in (2.13). Because of (3.7), we have

$$B(m+n-4) + C(m+n-4) \geq m+n-4 \tag{3.9}$$

and

$$B(m+n-5) + C(m+n-5) > m+n-5\,. \tag{3.10}$$

It follows from Theorem 2.1 that (1.11) is true, if also

$$B(s) + C(s) \geq s \quad \text{for} \quad s = 1, \ldots, m+n-6\ . \tag{3.11}$$

Indeed, Theorem 2.1 and (3.9) through (3.11) ensure that $\{0, 1, \ldots, m+n-4\} \subset M+N$. And if $b_{m-1} \geq c_{n-1}$, then the n integers $b_{m-1} + c_\nu$ ($\nu = 0, \ldots, n-1$) are in the set $(M+N)\backslash\{0, 1, \ldots, m+n-4\}$, because of (1.9); if $c_{n-1} > b_{m-1}$ we can find m integers in this set. Hence, $|M+N| \geq (m+n-3) + \min(m,n)$ if (3.7) and (3.11) are true.

It therefore suffices to consider the possibility that (3.11) fails to hold, say that

$$B(s_o) + C(s_o) < s_o \tag{3.12}$$

for some s_o, $1 \leq s_o \leq m+n-6$. Then,

$$B(s_o+1) + C(s_o+1) \leq s_o + 1\ . \tag{3.13}$$

It follows from (3.10), (3.12) and (3.13) that there is an integer i, with $s_o + 2 \leq i \leq m+n-5$, such that

$$B(s) + C(s) \leq s \quad \text{for} \quad s_o \leq s \leq i-1 \tag{3.14}$$

and $B(i) + C(i) > i$.

Then,

$$B(i-1) + C(i-1) = i-1 \tag{3.15}$$

and

$$B(i) + C(i) = i+1, \tag{3.16}$$

whence $i \in M \cap N$. And $i - 2 \geq s_o$ by definition, hence from (3.14),

$$B(i-2) + C(i-2) \leq i-2\,. \tag{3.17}$$

With (3.15), this implies that $i - 1 \in M \cup N$.

We now define q_1 and q_2 ($1 \leq q_1 \leq m-2$ and $1 \leq q_2 \leq n-2$) by setting

$$b_{q_1} = i = c_{q_2}\ ; \tag{3.18}$$

then $\max(b_{q_1-1}, c_{q_2-1}) = i-1\,.$

From (3.16) and (3.18) we have

$$i = q_1 + q_2 - 1\,; \tag{3.19}$$

hence $q_1 + q_2 \geq 4$, since $i \geq 3$. And from (3.18) and (3.19),

$$b_{q_1} = c_{q_2} = q_1 + q_2 - 1\ . \tag{3.20}$$

We may invoke the induction hypothesis to obtain the following estimates: if $b_{q_1-1} = i-1$, then

$$|M''_{m-q_1+1} + N''_{n-q_2}| \geq m+n-(q_1+q_2)-2+\min(m-q_1+1\,,\, n-q_2); \tag{3.21}$$

if $c_{q_2-1} = i-1$, then

$$|M''_{m-q_1} + N''_{n-q_2+1}| \geq m+n-(q_1+q_2)-2+\min(m-q_1,\, n-q_2+1)\,; \tag{3.22}$$

and in both cases,

$$|M''_{m-q_1+1} + N''_{n-q_2+1}| \geq m+n-(q_1+q_2)+\min(m-q_1,\, n-q_2)\,. \tag{3.23}$$

Indeed, (3.3) is verified each time because of (2.9) and since $q_1+q_2 \geq 4$. Condition (3.4) is met, since $0 \in M''_{m-r} \cap N''_{n-s}$ by (2.7) and (2.8). Condition (3.5) is satisfied because by (3.18) we have $1 = b_{q_1} - b_{q_1-1} \in M''_{m-q_1+1}$ if $b_{q_1-1} = i-1$, and $1 \in N''_{n-q_2+1}$ if $c_{q_2-1} = i-1$. To verify (3.6) we observe that by (2.7) and (1.9),

$$\begin{aligned}\max(M''_{m-r} \cup N''_{n-s}) &= \max(b_{m-1}-b_r,\, c_{n-1}-c_s)\\ &\geq (m+n-2)-\max(b_r, c_s),\end{aligned}$$

from which (3.6) follows in each case.

We shall also need two consequences of Theorem X, namely

$$|M'_{q_1+1} + N'_{q_2+1}| \geq q_1+q_2+\max(q_1,\, q_2\,) \tag{3.24}$$

and

$$|M'_{q_1} + N'_{q_2+1}| \geq 2q_1+q_2-1\,. \tag{3.25}$$

To obtain (3.24) we observe that because of (3.20) the sets M'_{q_1+1} and N'_{q_2+1} satisfy (1.7) since

$$|M'_{q_1+1}| + |N'_{q_2+1}| - 3 = q_1+q_2-1\,;$$

(3.24) is (1.8) for these sets.

For (3.25), we note that M'_{q_1} and N'_{q_2+1} verify (1.5) since by (1.1) and (3.20),

$$b_{q_1-1} < c_{q_2} = q_1+q_2-1 = |M'_{q_1}| + |N'_{q_2+1}| - 2.$$

By (1.6) then,

$$|M'_{q_1} + N'_{q_2+1}| \geq c_{q_2} + q_1\,,$$

and this is (3.25).

We proceed to apply (3.21) through (3.25). The argument in Case (I) is now separated into two subcases,

(Ia) $\qquad b_{q_1-1} = c_{q_2-1}\,, \qquad$ (3.26)

(Ib) $\qquad b_{q_1-1} \neq c_{q_2-1}\,.$

Case (Ia). In this case,

$$|M+N| \geq |M'_{q_1+1} + N'_{q_2+1}| + |M''_{m-q_1+1} + N''_{n-q_2+1}| - 3\,. \tag{3.27}$$

To prove (3.27) we use (2.11) with $r = q_1+1$, $s = q_2+1$, $k = q_1-1$, $\ell = q_2-1$. For simplicity of notation, set $M_1 = M'_{q_1+1}$, $N_1 = N'_{q_2+1}$, $M_2 = M \backslash M'_{q_1-1}$ and $N_2 = N \backslash N'_{q_2-1}$. We must show that $|(M_1+N_1) \cap (M_2+N_2)| = 3$ in order to get (3.27) from (2.11). Indeed, $b_{q_1-1}+c_{q_2-1}$, $b_{q_1}+c_{q_2-1}$, $b_{q_1-1}+c_{q_2}$ and $b_{q_1}+c_{q_2}$ are in

$(M_1+N_1)\cap(M_2+N_2)$, and $b_{q_1}+c_{q_2-1}=b_{q_1-1}+c_{q_2}$ by (3.18) and (3.26). These are the only elements of $(M_1+N_1)\cap(M_2+N_2)$. For consider some $x\in M_1+N_1$, say $x=b_u+c_v$, with $u<q_1-1$ or $v<q_2-1$; then $x<b_{q_1-1}+c_{q_2}$, hence $x\in M_2+N_2$ only if $x=b_{q_1-1}+c_{q_2-1}$.

Return now to (3.27). On combining (3.27), (3.23) and (3.24) we have

$$|M+N|\geq m+n-3+\max(q_1,q_2)+\min(m-q_1,n-q_2),$$

and this implies (1.11). This concludes the proof in Case (Ia).

Case (Ib). The argument when $b_{q_1-1}<c_{q_2-1}$ is typical. Then, we have

$$|M+N|\geq|M'_{q_1+1}+N'_{q_2+1}|+|M''_{m-q_1}+N''_{n-q_2+1}|-2 \tag{3.28}$$

and

$$|M+N|\geq|M'_{q_1}+N'_{q_2+1}|+|M''_{m-q_1}+N''_{n-q_2+1}|\,. \tag{3.29}$$

To verify (3.28), set $r=q_1+1$, $s=q_2+1$, $k=q_1$, $\ell=q_2-1$ in (2.11) and observe that if $u\leq q_1-1$ and $v\leq q_2$, then $b_u+c_v\in M'_{q_1+1}+N'_{q_2+1}$ but $b_u+c_v\leq b_{q_1-1}+c_{q_2}<b_{q_1}+c_{q_2-1}=\min(M\backslash M'_{q_1})+(N\backslash N'_{q_2-1})$. Hence $b_{q_1}+c_{q_2-1}$ and $b_{q_1}+c_{q_2}$ are the only elements of $(M'_{q_1+1}+N'_{q_2+1})\cap((M\backslash M'_{q_1})+(N\backslash N'_{q_2-1}))$. And (3.29) follows from (2.11) with $r=q_1$, $s=q_2+1$, $k=q_1$, $\ell=q_2-1$, since $b_{q_1-1}+c_{q_2}<b_{q_1}+c_{q_2-1}$ that is, $\max(M'_{q_1}+N'_{q_2+1})<\min((M\backslash M'_{q_1})+(N\backslash N'_{q_2-1}))$.

From (3.28), (3.22) and (3.24),

$$|M+N|\geq m+n-4+\max(q_1,q_2)+\min(m-q_1\,,\,n-q_2+1),$$

from which (1.11) follows if $q_2>q_1$.

If $q_1\geq q_2$ we use (3.29), (3.22) and (3.25) which together yield

$$|M+N|\geq m+n-3+q_1+\min(m-q_1\,,\,n-q_2+1),$$

and (1.11) follows.

This settles Case (Ib) when $b_{q_1-1}<c_{q_2-1}$. If $b_{q_1-1}>c_{q_2-1}$ the argument goes through as above on replacing (3.22) by (3.21) and similarly interchanging the roles of M and N in (3.25), (3.28) and (3.29).

This disposes of Case (I).

Case (II). This case is determined by condition (3.8). We may also assume that

$$\max(b_{m-1}-b_1,\ c_{n-1}-c_1)\geq m+n-4\,, \tag{3.30}$$

for otherwise, by Case (I), the conclusion of Theorem XI holds for M^* and N^*, since $b_{m-1}-b_1=x_{m-2}$ and $c_{n-1}-c_1=y_{n-2}$.

Because of Corollary 2.3, it suffices to consider sets M and N such that

$$\gcd(M)=\gcd(N)=1\,, \tag{3.31}$$

$$\gcd((M^*)'_{m-1})=1\,, \tag{3.32}$$

and

$$\gcd(M'_{m-1})=1\,. \tag{3.33}$$

In Case (II), we may further assume that

$$b_1=c_1=1 \tag{3.34}$$

and that

$$b_{m-1} - b_{m-2} = c_{n-1} - c_{n-2} = 1\,, \tag{3.35}$$

as we proceed to show. Consider (3.34) first. If $b_1 \neq c_1$ then 0, b_1, c_1 are distinct elements of $M + N$, not in $M_0 + N_0$ (in the notation of Proposition 2.2). Hence if $b_1 \neq c_1$,

$$|M + N| \geq |M_0 + N_0| + 3 = |(M^*)'_{m-1} + (N^*)'_{n-1}| + 3 \tag{3.36}$$

($b_{m-1} + c_{n-1} - x$ runs through $(M^*)'_{m-1} + (N^*)'_{n-1}$, if x runs through $M_0 + N_0$).

Inequality (3.36) also holds if $b_1 = c_1 \geq 2$. For if $b_1 = c_1 \geq 2$, let b_u and c_v be the smallest integers in M and N, respectively, such that $b_1 \nmid b_u$ and $b_1 \nmid c_v$ (they are well-defined, because of (3.31)). Then $u \geq 2$ and $v \geq 2$, whence

$$b_0 + c_0 < b_1 + c_0 < \min(b_u,\, c_v)\,. \tag{3.37}$$

And $\min(b_u, c_v) \notin M_0 + N_0$. Indeed, say $b_u \leq c_v$, and suppose that $b_u = b_k + c_\ell$ for some $k \geq 1$ and $\ell \geq 1$. Then $b_u > b_k$ and $c_v \geq b_u > c_\ell$, whence $b_k \equiv c_\ell \equiv 0 \pmod{b_1}$. This is impossible since $b_1 \nmid b_u$. Hence with (3.37), we have (3.36) again.

Now the induction hypothesis applies to $(M^*)'_{m-1}$ and $(N^*)'_{n-1}$ because of (3.30) and (3.32). With it, (3.36) yields (1.11). This justifies assumption (3.34).

To justify (3.35), we use M^* and N^*; note that (3.35) is equivalent to $x_1 = y_1 = 1$. By (2.5) and (3.31), $\gcd(M^*) = \gcd(N^*) = 1$. By reasoning as for (3.34) we see that

$$|M^* + N^*| \geq |M'_{m-1} + N'_{n-1}| + 3\,, \tag{3.38}$$

except perhaps if $x_1 = y_1 = 1$. And because of (3.8) and (3.33), we may apply the induction hypothesis to M'_{m-1} and N'_{n-1}; (1.11) then follows from (3.38).

Another restriction is possible in Case (II): we may assume that $m = n$. Indeed, suppose $m < n$. The induction hypothesis applies to M and N'_{n-1}: (3.5) is satisfied because of (3.31); so is (3.6) since by (1.9) and (3.35),

$$\max(M \cup N'_{n-1}) = \max(b_{m-1},\, c_{n-1} - 1) \geq m + n - 3 = |M| + |N'_{n-1}| - 2\ .$$

From the induction hypothesis we get

$$|M + N'_{n-1}| \geq m + (n-1) - 3 + \min(m, n-1) = m + n - 4 + \min(m, n),$$

and (1.11) follows. If $m > n$ we can reason in the same manner with M'_{m-1} and N.

Finally, since Theorem XI is symmetric in M and N, and since we have made no assumptions distinguishing M from N, we may assume that $b_{m-1} \geq c_{n-1}$.

We again consider the function $B(s) + C(s) - s$, where B and C are as in (2.13). It is ultimately negative, since M and N are finite. In fact, since now $b_{m-1} \geq c_{n-1}$ and consequently $b_{m-1} \geq m + n - 2$,

$$B(s) + C(s) < s \quad \text{for} \quad s > b_{m-1}\ . \tag{3.39}$$

On the other hand, because of (3.34), we have $B(1) + C(1) > 1$, and $B(2) + C(2) \geq 2$. Hence there is an integer j, with $2 \leq j \leq b_{m-1}$, such that $B(s) + C(s) \geq s$ for $1 \leq s \leq j$ and $B(j+1) + C(j+1) < j+1$. Then $B(j) + C(j) = j = B(j+1) + C(j+1)$, whence $j + 1 \notin M \cup N$. And by Theorem 2.1,

$$\{0, 1, \ldots, j\} \subset M + N\ . \tag{3.40}$$

If $j \geq m+n-4$ then (1.11) is true, by the argument developed after (3.11). We may therefore assume that $j \leq m+n-5$; then, $j+1 < b_{m-1}$ by (1.9). With this assumption, let p_1 be such that $b_{p_1-1} < j+1 < b_{p_1}$. By (3.34) and (3.35), $2 \leq p_1 \leq m-2$. Then, either $c_{n-1} < j+1 < b_{p_1}$ or $j+1 < c_{n-1}$.

If $c_{n-1} < j+1 < b_{p_1}$ then $B(j+1)+C(j+1)=j$ yields

$$j = n + p_1 - 2\ . \tag{3.41}$$

The integers in (3.40), the b_i with $p_1 \leq i \leq m-1$ and the $b_{m-1}+c_k$ with $1 \leq k \leq n-1$ are distinct, and in $M+N$. By (3.41) they are $(j+1)+(m-p_1)+(n-1) = m+2n-2$ in number; this implies (1.11).

If $j+1 < c_{n-1}$, let $p_2\,(2 \leq p_2 \leq n-2)$ be such that $c_{p_2-1} < j+1 < c_{p_2}$. Then (3.41) is replaced by

$$j = p_1 + p_2 - 2\ . \tag{3.42}$$

We now distinguish three subcases, according to the sign of $p_1 - p_2$. Suppose first that $p_1 = p_2 = p$, say. Then by arguing as for (3.27), we have

$$|M+N| \geq |M'_{p+1} + N'_{p+1}| + |M''_{m-p+1} + N''_{n-p+1}| - a\,, \tag{3.43}$$

where

$$a = \begin{cases} 4 & \text{if} \quad b_{p-1}+c_p \neq b_p + c_{p-1} \qquad (3.44) \\ 3 & \text{else.} \qquad (3.45) \end{cases}$$

For the first member on the right side of (3.43), we have

$$|M'_{p+1}+N'_{p+1}| \geq \begin{cases} 3p+1 & \text{if} \quad b_{p-1}+c_p \neq b_p+c_{p-1} \qquad (3.46) \\ 3p & \text{else.} \qquad (3.47) \end{cases}$$

Indeed, $\{0,1,\ldots,j\} \subset M'_{p+1} + N'_{p+1}$ because of (3.40) and since

$$b_u + c_v > \min(b_p,\, c_p) > j$$

if $u > p$ or $v > p$. And if $b_p + c_{p-1} < b_{p-1} + c_p$, then the $p+2$ integers $b_p + c_\nu$ ($\nu = 0,1,\ldots,p$) and $b_{p-1}+c_p$ are distinct, in $M'_{p+1} + N'_{p+1}$, and larger than j. This proves (3.46), since $(j+1)+p+2 = 3p+1$. (If $b_p + c_{p-1} > b_{p-1}+c_p$, use the $b_\nu + c_p$ with $0 \leq \nu \leq p$, and $b_p + c_{p-1}$.) To prove (3.47), use the same integers as for (3.46), except $b_{p-1}+c_p$ (or $b_p + c_{p-1}$, as the case may be).

For the second member on the right side of (3.43), we have

$$|M''_{m-p+1} + N''_{n-p+1}| \geq 3\,(m-p+1) - 3 \tag{3.48}$$

by the induction hypothesis: condition (3.5) is verified since $b_{m-1} - b_{p-1}$ and $b_{m-2} - b_{p-1}$ are consecutive integers, by (3.35); and (3.6) is met, since

$$\begin{aligned} &\max(b_{m-1} - b_{p-1}\,,\, c_{n-1} - c_{p-1}) \\ &\quad \geq \max(b_{m-1},\, c_{n-1}) - \max(b_{p-1},\, c_{p-1}) \\ &\quad \geq (m+n-2) - j = (m-p+1) + (n-p+1) - 2\,. \end{aligned}$$

Now (3.43) through (3.48) imply (1.11). This settles the subcase in which $p_1 = p_2$.

Suppose now that $p_1 > p_2$ in (3.42). Because of (3.40) and since $c_{p_2} > j$,

$$|M+N| \geq (j+1) + |M + \{c_{p_2},\, c_{p_2+1}, \ldots,\, c_{n-1}\}|, \tag{3.49}$$

whence with (2.12),

$$|M+N| \geq (j+1) + |M + N''_{n-p_2}|\,. \tag{3.50}$$

The induction hypothesis applies to M and N''_{n-p_2}, by (3.31) and (1.9), and since $b_{m-1} > c_{n-1} - c_{p_2}$ and $p_2 \geq 2$. With it and (3.42), (3.50) yields

$$|M+N| \geq (p_1 + p_2 - 1) + m + 2(m - p_2) - 3 = 3m - 4 + (p_1 - p_2)\,,$$

whence $|M+N| \geq 3m-3$.

We must still treat the subcase in which

$$p_1 < p_2\,. \tag{3.51}$$

Arguing as for (3.50), we see that (3.40) and $b_{p_1} > j$ imply that

$$|M+N| \geq (j+1) + |M''_{m-p_1} + N|\,. \tag{3.52}$$

If max $(M''_{m-p_1} \cup N) \geq |M''_{m-p_1}| + |N| - 2$, that is, if

$$\max(b_{m-1} - b_{p_1}\,,\, c_{n-1}) \geq 2m - p_1 - 2\,, \tag{3.53}$$

then by the induction hypothesis,

$$|M''_{m-p_1} + N| \geq 3(m-1) - 2p_1\,. \tag{3.54}$$

With (3.54), (1.11) follows from (3.52), (3.42) and (3.51).

In order to conclude the proof of Theorem XI, we must consider subcase (3.51) when, instead of (3.53),

$$\max\,(b_{m-1} - b_{p_1}, c_{n-1}) \leq 2m - p_1 - 3\,. \tag{3.55}$$

For this we use the sets M^* and N^*, as defined in (2.3) and (2.4). In analogy to (2.13), let B^* and C^* denote the counting functions of the positive elements of M^* and N^*, respectively. By (1.9) and (3.35) there is an integer j^* with $2 \leq j^* \leq b_{m-1}$, such that $B^*(s) + C^*(s) \geq s$ for $1 \leq s \leq j^*$ and $B^*(j^*+1) + C^*(j^*+1) < j^*+1$. Then $j^*+1 \notin M^* \cup N^*$, $j^* = B^*(j^*+1) + C^*(j^*+1)$, and by Theorem 2.1,

$$\{0, 1, \ldots, j^*\} \subset M^* + N^*\,. \tag{3.56}$$

By a previous assumption, $y_{n-1} := c_{n-1} \leq b_{m-1} =: x_{m-1}$. By the argument applied after (3.40), we may assume that $j^*+1 < x_{m-1}$. Then define p_1^* $(1 < p_1^* < m)$ by

$$x_{p_1^*-1} < j^* + 1 < x_{p_1^*}\,. \tag{3.57}$$

If $y_{n-1} < j^*+1 < x_{p_1^*}$, we can prove (1.11) by reasoning as when $c_{n-1} < j+1 < b_{p_1}$ (use (3.56), and replace (3.41) by $j^* = n + p_1^* - 2$). Accordingly, let us assume that

$$j^* + 1 < c_{n-1}\,. \tag{3.58}$$

Because of (3.55), and since $b_{m-1} - b_{p_1} = x_{m-p_1-1}$ and $c_{n-1} = y_{m-1}$, we have

$$B^*(2m - p_1 - 3) + C^*(2m - p_1 - 3) \geq (m - p_1 - 1) + (m-1) > 2m - p_1 - 3\,.$$

And $2m - p_1 - 3 \geq c_{n-1} > j^* + 1$ by (3.55) and (3.58). Thus, if (3.55) and (3.58) hold, then

$$B^*(s) + C^*(s) > s \qquad \text{for some} \quad s > j^* + 1\,.$$

Now

$$B^*(j^*+1) + C^*(j^*+1) < j^* + 1\,. \tag{3.59}$$

Hence (3.55) and (3.58) imply the existence of an integer g such that

$$B^*(s) + C^*(s) \leq s \qquad \text{for} \qquad j^* + 1 \leq s \leq g - 1 \tag{3.60}$$

and

$$B^*(g) + C^*(g) > g\,.$$

Then,

$$B^*(g-1) + C^*(g-1) = g - 1, \tag{3.61}$$

$$B^*(g) + C^*(g) = g + 1\,, \tag{3.62}$$

and therefore $g \in M^* \cap N^*$. Furthermore, $g \geq j^* + 2$ by definition, and $g = j^* + 2$ is excluded by comparing (3.59) and (3.61). Thus $g - 2 \geq j^* + 1$, and from (3.60),

$$B^*(g-2) + C^*(g-2) \leq g - 2\,; \tag{3.63}$$

with (3.61) this implies that $g - 1 \in M^* \cup N^*$.

Now define r_1 and r_2 by setting

$$x_{r_1} = g = y_{r_2}\,; \tag{3.64}$$

then $x_{r_1 - 1} = g - 1$ or $y_{r_2 - 1} = g - 1$. And from (3.62) and (3.64),

$$g = r_1 + r_2 - 1\,. \tag{3.65}$$

We now have a situation entirely similar to the one encountered in Case (I): compare (3.61) through (3.65) with (3.15) through (3.19).

To complete the proof of (1.11) when (3.51) holds, it suffices to proceed as in Case (I). On replacing there M and N by M^* and N^*, respectively, q_i by r_i $(i = 1, 2)$, each b by x and each c by y, and remembering that $|M^* + N^*| = |M + N|$, we dispose of this last subcase.

This concludes the proof of Theorem XI.

References

[1] Freiman G.A., *Inverse Problems in Additive Number Theory*, VI. On the Addition of Finite Sets, III (in Russian). Izv. Vyss. Ucebn. Zaved. Matematika, **3 (28)**, 1962, 151–157.

[2] Freiman G.A., *On the Structure and the Number of Sum-Free Sets*, Journées Arithmétiques de Genève 1991, Astérisque **209**, 195–203.

[3] Deshouillers J.-M., Sós V., Freiman G.A. and Temkin M., *On the Structure of Sum-Free Sets, 2*, this volume.

[4] Halberstam H. and Roth K.F., *Sequences* (second edition), Springer-Verlag, New York, Berlin, 1983.

[5] Mann H.B., *A Proof of the Fundamental Theorem on the Density of Sums of Sets of Positive Integers*, Ann. Math., **43**, 1942, 523–527.

J. Steinig, Section de Mathématiques, Université de Genève, C.P. 240, 1211 Genève 24, Suisse
E-mail : `Peterman@ibm.unige.ch (qui transmettra)`

JEAN-MARC DESHOUILLERS
GREGORY A. FREIMAN

On an additive problem of Erdős and Straus, 2

Astérisque, tome 258 (1999), p. 141-148
<http://www.numdam.org/item?id=AST_1999__258__141_0>

Astérisque **258**, 1999, p. 141–148

ON AN ADDITIVE PROBLEM OF ERDŐS AND STRAUS, 2

by

Jean-Marc Deshouillers & Gregory A. Freiman

Abstract. — We denote by $s^\wedge A$ the set of integers which can be written as a sum of s pairwise distinct elements from A. The set A is called admissible if and only if $s \neq t$ implies that $s^\wedge A$ and $t^\wedge A$ have no element in common.

P. Erdős conjectured that an admissible set included in $[1, N]$ has a maximal cardinality when A consists of consecutive integers located at the upper end of the interval $[1, N]$. The object of this paper is to give a proof of Erdős' conjecture, for sufficiently large N.

Let $\mathcal{A}$ be a set of positive integers having the property that each time an integer n can be written as a sum of distinct elements of $\mathcal{A}$, the number of summands is well defined, in that the integer n cannot be written as a sum of distinct elements of $\mathcal{A}$ with a different number of summands. This notion has been introduced by P. Erdős in 1962 (cf. [**2**]) and called **admissibility** by E.G. Straus in 1966 (cf. [**5**]). In other words, if we denote by $s^\wedge\mathcal{A}$ the set of integers which can be written as a sum of s pairwise distinct elements from $\mathcal{A}$ then $\mathcal{A}$ is **admissible** if and only if $s \neq t$ implies that $s^\wedge\mathcal{A}$ and $t^\wedge\mathcal{A}$ have no element in common.

Erdős conjectured that an admissible subset $\mathcal{A}$ included in $[1, N]$ has a cardinality which is maximal when $\mathcal{A}$ consists of consecutive integers located at the upper end of the interval $[1, N]$. As it was computed by E.G. Straus, the set

$$\{N - k + 1, N - k + 2, \dots, N\}$$

is admissible if and only if $k \leq 2\sqrt{N + 1/4} - 1$.

Straus himself proved that $\sqrt{N}$ is the right order of magnitude for the cardinality of a maximal admissible subset from $[1, N]$. More precisely, he proved the inequality $|\mathcal{A}| \leq (4/\sqrt{3} + o(1))\sqrt{N}$. The constant involved has been slightly reduced by P. Erdős, J-L. Nicolas and A. Sárkőzy (cf. [**3**]) and we proved (cf. [**1**]) the inequality

1991 ***Mathematics Subject Classification.*** — 11 P99, 05 D05.

Key words and phrases. — Admissible sets, arithmetic progressions.

J.-M. D.: Cette recherche a bénéficié du soutien du CNRS (UMR 9936, Université Bordeaux 1) et de l'Université Victor Segalen Bordeaux 2.

$|\mathcal{A}| \le (2+o(1))\sqrt{N}$. The object of this paper is to give a proof of Erdős conjecture, at least when N is sufficiently large.

Theorem 1. — *There exists an integer N_0, effectively computable, such that for any integer $N \ge N_0$ and any admissible subset $\mathcal{A} \subset [1,N]$ we have*

$$\operatorname{Card}\mathcal{A} \le 2\sqrt{N+1/4}-1.$$

The proof is based on the description of the structure of large admissible sets we obtained previously, namely :

***Theorem 2* (J-M. Deshouillers, G.A. Freiman [1]).** — *Let $\mathcal{A}$ be an admissible set included in $[1,N]$, such that* $\operatorname{Card}\mathcal{A} > 1.96\sqrt{N}$. *If N is large enough, there exist $\mathcal{C} \subset \mathcal{A}$ and an integer q having the following properties :*
(i) $\operatorname{Card}\mathcal{C} \le 10^5 N^{5/12}$,
(ii) for some t the set $t^\wedge\mathcal{C}$ contains at least $3N^{5/6}$ terms in an arithmetic progression modulo q,
(iii) $\mathcal{A}\setminus\mathcal{C}$ is included in an arithmetic progression modulo q containing at most $N^{7/12}$ terms.

Although we do not develop this point, it will be clear from the proof that our arguments may be used to describe the structure of maximal admissible subsets of $[1,N]$, leading for example to the fact that when N has the shape n^2 or n^2+n (and n sufficiently large), the Erdős - Straus example is the only maximal subset of $[1,N]$.

1. We first establish a lemma expressing the fact that if a set of integers $\mathcal{D}$ is part of a finite arithmetic progression with few missing elements, then the same is locally true for $s^\wedge\mathcal{D}$.

Proposition 1. — *Let us consider integers r,s,t and a,q such that $t \ge 2s-q$, $s \ge 4r+3+q$ and $0 \le a < q$.*
Let further $\mathcal{D} = \{d_1 < d_2 < \cdots < d_t\}$ be a set of t distinct integers congruent to a modulo q such that $d_t - d_1 = (t-1+r)q$,and denote by m (resp. M) the smallest (resp. largest) element in $s^\wedge\mathcal{D}$. Then, among $2r+1$ consecutive integers congruent to sa modulo q and laying in the interval $[m,M]$, at least $r+1$ belong to $s^\wedge\mathcal{D}$.

Proof. — We treat the special case when $a=0, q=1$ and $\mathcal{D}$ is included in $[1,t]$. We notice that the general case reduces to this one by writing
$d_l = d_1 + q(\delta_l - 1)$ and considering the set $\{\delta_1,\dots,\delta_t\}$.

Let x be an integer in $s^\wedge\mathcal{D} \cap [m,(m+M)/2]$. We first show that the interval $[x, x+3r]$ contains at least $2r+1$ elements from $s^\wedge\mathcal{D}$. Since x is in $s^\wedge\mathcal{D}$, we can find $d(1) < \cdots < d(s)$, elements in $\mathcal{D}$, the sum of which is x.

Let us show that $d(1)$ is less than $t-s-3r$. On the one hand we have

$$m+M \le (r+1)+\cdots+(r+s)+(t+r-s+1)+\cdots+(t+r) = \frac{s}{2}(2t+4r+2),$$

and on the other hand we have

$$x \ge d(1)+(d(1)+1)+\cdots+(d(1)+s-1) = \frac{s}{2}(2d(1)+s-1).$$

The inequality $x \leq (m+M)/2$ implies that we have

$$2d(1)+s-1 \leq t+2r+1,$$

whence

$$2d(1) \leq 2(t-s-3r)-(t-s-4r-2),$$

and we notice that $t-s-4r-2$ is positive, by the assumptions of Proposition 1.

Since $d(1)$ is less than $t-s-3r$, the interval $[d(1), t+r]$ contains at least $s+4r+1$ integers. We denote by $i_1 < \cdots < i_l$ the indexes of those $d's$ such that $d(i_k+1)-d(i_k) \geq 2$, with the convention that $d(i_l+1)=3Dt+r+1$ in the case when $d(s)<t+r$. The set

$$\bigcup_{k=1}^{l}]d(i_k)+1, d(i_k+1)-1[$$

contains at least $4r+1$ integers. We now suppress from those intervals those which contain no element from $\mathcal{D}$, and we rewrite the remaining ones as

$$]d(j_1)+1, d(j_1+1)-1[, \ldots,]d(j_h)+1, d(j_h+1)-1[.$$

They contain at least $3r+1$ integers, among which at most r are not in $\mathcal{D}$.

Let us define u_1 to be the largest integer such that $d(j_1)+u_1$ is in $\mathcal{D}$ and is less than $d(j_1+1)$, and let us define $u_2, \ldots, u_h$ in a similar way. We consider the integers

$$x = y + d(j_1) + \cdots + d(j_h) \quad \text{(which defines } y\text{)},$$

$$x+1 = y + d(j_1) + 1 + d(j_2) + \cdots + d(j_h),$$

$$\ldots$$

$$x+u_1 = y + d(j_1) + u_1 + d(j_2) + \cdots + d(j_h),$$

$$\ldots$$

$$x+u_1+\cdots+u_h = y + d(j_1) + u_1 + d(j_2) + u_2 + \cdots + d(j_h) + u_h.$$

One readily deduces from this construction that the interval

$$[x, x+\min(3r, u_1+\cdots+u_h)]$$

contains at most r elements which are not in $s^\wedge\mathcal{D}$.

What we have proven so far easily implies that any interval $[z-r, z]$ with $m \leq z \leq (M+m)/2$ contains at least one element in $s^\wedge\mathcal{D}$. Let us consider an interval $[y, y+2r]$ with $m \leq y \leq (M+m)/2$. By what we have just said, the interval $[y-r, y]$ contains an element in $s^\wedge\mathcal{D}$, let us call it x. As we have shown the interval $[x, x+3r]$ contains at most r integers not in $s^\wedge\mathcal{D}$, so that $[y, y+2r]$ contains at most r integers not in $s^\wedge\mathcal{D}$, which is equivalent to say that it contains at least $r+1$ elements from $s^\wedge\mathcal{D}$.

A similar argument taking into account decreasing sequences and starting with M shows that any interval $[y-2r, y]$ with $(m+M)/2 \leq y \leq M$ contains at least $r+1$ elements from $s^\wedge\mathcal{D}$.

2. We now prove the following result concerning the structure of a large admissible finite set.

Theorem 3. — *Let $\mathcal{A} = \{a_1 < \cdots < a_A\}$ be an admissible subset of $[1, N]$ with cardinality $A = 2N^{1/2} + O(N^{5/12})$, and let us define q to be the largest integer such that $\mathcal{A}$ is contained in an arithmetic progression modulo q. We have $q = O(N^{5/12})$ and there exists an integer u in $[N^{11/24}, 2N^{11/24}]$ such that*

$$a_{A-u} - a_{u+1} = q(2N^{1/2} + O(N^{11/24})).$$

Proof. — The proof is based on the structure result we quoted in the introduction as Theorem 2. We keep its notation and first show that an integer q satisfying (ii) and (iii) is indeed the largest integer such that $\mathcal{A}$ is contained in an arithmetic progression modulo q. We let $\mathcal{B}$ denote $\mathcal{A} \setminus \mathcal{C}$.

A simple counting argument will show that $\mathcal{A}$ is included in the same arithmetic progression as $\mathcal{B}$. Otherwise, let us consider an element $a \in \mathcal{A}$ which is not in the same arithmetic progression as $\mathcal{B}$ modulo q. The set $s^\wedge\mathcal{A}$ contains the disjoint sets $s^\wedge\mathcal{B}$ and $a + (s-1)^\wedge\mathcal{B}$. We thus have $|s^\wedge\mathcal{A}| \geq |s^\wedge\mathcal{B}| + |(s-1)^\wedge\mathcal{B}|$. It is well-known (cf. [4] for example) that $|s^\wedge\mathcal{B}| \geq s(|\mathcal{B}| - s)$ for $s \leq |\mathcal{B}|$, and since $\mathcal{A} \subset [1, N]$ is admissible we have

$$\begin{array}{rcl} N(|\mathcal{B}|+1) & \geq & \text{Card } (\bigcup_s (s^\wedge\mathcal{B} \cup (a + (s-1)^\wedge\mathcal{B})) \\ & \geq & 2\sum_s |s^\wedge\mathcal{B}| \geq 2\sum_s s = 20(|\mathcal{B}| - s) = \frac{1}{3}|\mathcal{B}|^3 + O(N), \end{array}$$

which implies $|\mathcal{B}| \leq (\sqrt{3} + o(1))\sqrt{N}$, so that we have $|\mathcal{A}| = |\mathcal{B}| + |\mathcal{C}| \leq (\sqrt{3} + o(1))\sqrt{N}$, a contradiction.

We have so far proven that q divides $g := gcd(a_2 - a_1, \ldots, a_A - a_1)$. Property (ii) implies that q is a multiple of g, so that we have $q = g$, as we wished to show.

The second step in the proof consists in showing that for $0 < k \leq |\mathcal{B}| - q$, any element in $k^\wedge\mathcal{B}$ is less than any element in $(k+q)^\wedge\mathcal{B}$. Let us call J the $3N^{5/6}$ consecutive terms of the arithmetic progression modulo q, the existence of which is asserted in (ii). Since $\mathcal{B}$ is included in an arithmetic progression modulo q with less that $3N^{5/6}$ terms, the sets $k^\wedge\mathcal{B} + J$ and $(k+q)^\wedge\mathcal{B} + J$ consists of consecutive terms of arithmetic progressions modulo q, and moreover, they are in the *same* class modulo q. Since $\mathcal{A}$ is admissible, the sets $k^\wedge\mathcal{B} + J$ (included in $(k+t)^\wedge\mathcal{A}$) and $(k+q)^\wedge\mathcal{B} + J$ (included in $(k+q+t)^\wedge\mathcal{A}$) do not intersect. To prove that any element of $k^\wedge\mathcal{B}$ is less that any element of $(k+q)^\wedge\mathcal{B}$, it is now sufficient to notice that $k^\wedge\mathcal{B}$ contains an element (we can consider the smallest element of $k^\wedge\mathcal{B}$), which is smaller than some element of $(k+q)^\wedge\mathcal{B}$.

We now prove that $q = O(N^{5/12})$. The cardinality of $\mathcal{A}$ and Theorem 2 imply that $|\mathcal{B}| = 2N^{1/2} + O(N^{5/12})$. We choose k so that $2k + q$ is $|\mathcal{B}|$ or $|\mathcal{B}| - 1$. (We notice that this is always possible since $\mathcal{A}$ contains at least $N^{1/2}$ integers from $[1, N]$ in an arithmetic progression modulo q, so that $q \leq N^{1/2}$). By the second step, the largest element in $k^\wedge\mathcal{B}$ is smaller than the largest element in $(k+q)^\wedge\mathcal{B}$. Let z be $(k+q)$-th element from $\mathcal{B}$, in the increasing order. We have

$$z \leq N - (k-1)q$$

and

$$(z+q)+\cdots+(z+qk)\leq z+(z-q)+\cdots+(z-(k+q-1)q) \quad ;$$

by an easy computation, we get

$$(q+2k)^2 \leq 2N+2k^2+3q,$$

but $2k+q=|\mathcal{B}|+O(1)=|\mathcal{A}|+O(N^{5/12})$, which implies

$$2k^2 \geq 2N(1+O(N^{-1/12})),$$

so that we have

$$k=N^{1/2}+O(N^{5/12}).$$

We now use again the same argument, being more precise. Let us write $\mathcal{B}=\{b_1<\cdots<b_{k+q}<b_{k+q+1}<\cdots<b_{2k+q}\leq b_B\}$. We have

$$b_{k+q+1}+\cdots+b_{2k+q}<b_1\cdots+b_k+b_{k+1}+b_{k+q}.$$

Let t be any integer in $[1,k]$. We have

$$b_{k+1}+\cdots+b_{k+q}>(b_{2k+q}-b_1)+\cdots+(b_{2k+q-t+1}-b_t)+\cdots+(b_{k+q+1}-b_k).$$

We clearly have the inequalities

$$\begin{aligned}
&b_{k+q+1}-b_k \geq (q+1)q,\\
&b_{k+q+2}-b_{k-1} \geq (q+3)q,\\
&\cdots\\
&b_{2k+q-t-1}-b_{t+2} \geq (q+1+2(k-t-2))q,\\
&b_{2k+q-t}-b_{t+1} \geq b_{2k+q-t}-b_{t+1},\\
&b_{2k+q-t+1}-b_t \geq (b_{2k+q-t}-b_{t+1})+2q,\\
&\cdots\\
&(b_{2k+q}-b_1) \geq (b_{2k+q-t}-b_{t+1})+2tq.
\end{aligned}$$

We thus obtain

$$\begin{aligned}
b_{k+1}+\cdots+b_{k+q} &> (t+1)(b_{2k+q-t}-b_{t+1})\\
&+q\textstyle\sum_{l=0}^{k-t-2}(q+1+2l)+q\sum_{h=0}^{t}2h.
\end{aligned}$$

Taking into account that $b_{k+q}\leq N-kq$, a dull computation leads to

$$(t+1)(b_{2k+q-t}-b_{t+1})\leq q(N-k^2+2kt+O(N^{11/12})),$$

when $t=O(N^{11/24})$. This in turn leads to

$$b_{2k+q-t}-b_{t+1}\leq q(2k+O(N^{11/24})),$$

when $t=\frac{3}{2}N^{11/24}+O(1)$.
Let C the cardinality of $\mathcal{C}$. Since $\mathcal{A}=\mathcal{B}\cup\mathcal{C}$, we have

$$b_{t+1}\leq a_{t+C}\leq a_{A+t+C-2k-q+1}\leq a_{2k+q-C-t}\leq b_{2k+q-t} \quad ;$$

we choose $u=A+t+C-2k-q$ and recall that $A-2k-q\leq C+1=O(N^{5/12})$, so that Theorem 3 is proven.

3. We now embark on the proof of Theorem 1 which will follow from Theorem 3 and Proposition 1. Let $\mathcal{A}$ be an admissible subset of $[1, N]$ with maximal cardinality. By [**1**], we know that $A = 2\sqrt{N} + O(N^{5/12})$, so we can apply Theorem 3 : there exists integers u and r such that

$$a_{A-u} - a_{u+1} = q(A - 2u + r),$$

with $u \in [N^{11/24}, 2N^{11/24}]$ and $r = O(N^{11/24})$.

We let

$$\mathcal{D} := \mathcal{A} \cap [a_{u+1}, a_{A-u}], \quad t := A - 2u, \quad \sigma := [(t-q)/2],$$

and we shall apply Proposition 1 with $s = \sigma$ and $s = \sigma + q$ (one readily checks that the conditions of application of Proposition 1 are fulfilled). Let us further denote by $m(s)$ (resp. $M(s)$) the smallest (resp. largest) element in $s^\wedge\mathcal{D}$.

As a first step, we show that $a_1 + a_2 + \cdots + a_q$ cannot be too small. We have

$$\begin{aligned} M(\sigma) - m(\sigma) &\geq (a_{A-u-\sigma+1} - a_{u+\sigma}) + \cdots + (a_{A-u} - a_{u+1}) \\ &\geq q(2 + 4 + \cdots + 2(\sigma - 1) = q\sigma(\sigma - 1) \\ &= qN + O(qN^{23/24}). \end{aligned}$$

If $\alpha_q := a_1 + \cdots + a_q$ were less than $M(\sigma) - m\sigma - (2r+1)q$, the intersection of $[m(\sigma), M(\sigma)]$ and $[m(\sigma) + \alpha_q, M(\sigma) + \alpha_q]$ would be an interval containing at least $(2r+1)$ integers in each class modulo q. By the property of $\sigma^\wedge\mathcal{D}$ established in Proposition 1, property obviously shared by $\alpha_q + \sigma^\wedge\mathcal{D}$, the pigeon-hole principle would imply that $\sigma^\wedge\mathcal{D}$ and $\alpha_q + \sigma^\wedge\mathcal{D}$ have an element in common, and this would contradict the admissibility of $\mathcal{A}$. (We may notice that this implies that a_1 itself is not too small, but we shall not use this fact).

By using the same pigeon-hole argument, we see that the admissibility of $\mathcal{A}$ implies

$$M(\sigma) + a_{A-u+1} + \cdots + a_A \leq m(\sigma + q) + a_1 + \cdots + a_u + (2r-1)q,$$

that is to say

$$a_{A-u-\sigma+1} + \cdots + a_{A-u} + \cdots a_A \leq a_1 + \cdots + a_u + a_{u+1} + \cdots + a_{u+\sigma+q} + (2r-1)q,$$

whence we deduce

$$\begin{aligned} (a_A - a_1) + (a_{A-1} - a_2) \quad &+ \cdots + (a_{A-u-\sigma+1} - a_{u+\sigma}) \leq \\ &a_{u+\sigma+1} + \cdots + a_{u+\sigma+q} + (2r-1)q. \end{aligned}$$

We have $a_{A-u-\sigma+1} - a_{u+\sigma} \geq q(A - u - \sigma + 1 - u - \sigma) = q(A - 2u - 2\sigma + 1)$ and, by the definition of σ, we can write

$$A - 2u - 2\sigma = q + \theta,$$

where $\theta = 0$ if $A - q$ is even and $\theta = 1$ if $A - q$ is odd. We thus have

$$\begin{aligned} urq + q(1 + q + \theta) + q(3 + q + \theta) \quad &+ \cdots + q(2(u + \sigma) - 1 + q + \theta) \leq \\ &a_{u+\sigma+1} + \cdots + a_{u+\sigma+q} + (2r-1)q. \end{aligned}$$

Since $u \geq 2$ and $r \geq 0$, we have

$$\begin{aligned} q(u+\sigma)(u+\sigma+q+\theta) &\leq a_{u+\sigma+1} + \cdots + a_{u+\sigma+q} - q \\ &\leq N - (A-u-\sigma-1)q + \cdots + \\ &\quad N - (A-u-\sigma-q)q - q \\ &\leq Nq - Aq^2 + uq^2 + \sigma q^2 + \tfrac{q^2(q+1)}{2} - q. \end{aligned}$$

We now replace $u+\sigma$ by $\frac{A-q-\theta}{2}$, which leads to

$$q\left(\frac{A-q-\theta}{2}\right)\left(\frac{A+q+\theta}{2}\right) \leq Nq - q^2\left(\frac{A+q+\theta}{2}\right) + \frac{q^2(q-1)}{2}.$$

If $A-q$ is even, we get

$$A^2 - q^2 \leq 4N - 2Aq - 2q^2 + 2q^2 - 2q,$$

whence

$$A^2 + 2Aq + q^2 \leq 4N + 2q^2 - 2q,$$

or

$$(A+q)^2 \leq 4N + 2q^2 - 2q.$$

if $q = 1$, this is $(A+1)^2 \leq 4N$;
if $q \geq 2$, we have

$$\begin{aligned} (A+1)^2 &\leq (A+q)^2 - (A+q)^2 + (A+1)^2 \\ &\leq 4N + 2q^2 - 2q - A^2 - 2Aq - q^2 + A^2 + 2A + 1 \\ &\leq 4N + 2A(1-q) + (q-1)^2 \\ &\leq 4N - (q-1)(2A-q+1) \leq 4N. \end{aligned}$$

If $A-q$ is odd, we get

$$A^2 - (1+q)^2 \leq 4N - 2Aq - 2q^2 - 2q + 2q^2 - 2q.$$

if $q = 1$, this is $A^2 + 2A + 1 \leq 4N + 1$;
if $q \geq 2$, we have

$$(A+1)^2 \leq A^2 - (1+q)^2 + 2q + q^2 + 2 + 2A$$

$$\leq 4N - 2A(q-1) + q^2 - 2q + 2$$

$$\leq 4N - (q-1)(2A-q+1) + 1$$

$$\leq 4N + 1.$$

In all cases, we thus have $(A+1)^2 \leq 4N+1$, which ends the proof of our main result.

References

[1] Deshouillers J-M. and Freiman G.A., *On an additive problem of Erdős and Straus, 1*, Israël J. Math., **92**, 1995, 33–43.

[2] Erdős P., *Some remarks on number theory*, III. Mat. Lapok **13**, 1962, 28–38.

[3] Erdős P., Nicolas J-L., Sarkőzy A., *Sommes de sous-ensembles*, Sem. Th. Nb. Bordeaux **3**, 1991, 55–72.

[4] Freiman G.A., *The addition of finite sets, 1*,(Russian), Izv. Vyss. Ucebn. Zaved. Matematika, **6(13)**, 1959, 202–213.

[5] Straus E.G., *On a problem in combinatorial number theory*, J. Math. Sci., **1**, 1966, 77–80.

J.-M. Deshouillers, Mathématiques Stochastiques, Université Victor Segalen Bordeaux 2, 33076 Bordeaux, France • *E-mail :* `j-m.deshouillers@u-bordeaux2.fr`

G.A. Freiman, School of Mathematical Sciences, Department of Mathematics, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, 69978 Tel Aviv, Israel
E-mail : `grisha@math.tau.ac.il`

Astérisque

JEAN-MARC DESHOUILLERS
GREGORY A. FREIMAN
VERA SÓS
MIKHAIL TEMKIN

On the structure of sum-free sets, 2

Astérisque, tome 258 (1999), p. 149-161
<http://www.numdam.org/item?id=AST_1999__258__149_0>

Astérisque **258**, 1999, p. 149–161

ON THE STRUCTURE OF SUM-FREE SETS, 2

by

Jean-Marc Deshouillers, Gregory A. Freiman, Vera Sós & Mikhail Temkin

Abstract. — A finite set of positive integers is called sum-free if $\mathbb{A} \cap (\mathbb{A} + \mathbb{A})$ is empty, where $\mathbb{A} + \mathbb{A}$ denotes the set of sums of pairs of non necessarily distinct elements from $\mathbb{A}$. Improving upon a previous result by G.A. Freiman, a precise description of the structure of sum-free sets included in $[1, M]$ with cardinality larger than $0.4M - x$ for $M \geq M_0(x)$ (where x is an arbitrary given number) is given.

1. Introduction

A finite set of positive integers $\mathcal{A}$ is called **sum-free** if $\mathcal{A} \cap (\mathcal{A} + \mathcal{A})$ is empty, where $\mathcal{A} + \mathcal{A}$ denotes the set of sums of pairs of elements from $\mathcal{A}$.

Such sum-free sets have been considered by Cameron and Erdős (cf. [1]), and the first result concerning their structure has been obtained by Freiman (cf. [3]). It is clear that for odd n, the sets $\{1, 3, 5, \ldots, n\}$ and $\{\frac{n+1}{2}, \frac{n+3}{2}, \ldots, n\}$ are sum-free. Freiman showed that when $\mathcal{A}$ is included in $[1, n]$ and its cardinality is at least $5n/12 + 2$, then $\mathcal{A}$ is essentially a subset of the ones we just described. In an unpublished paper, Deshouillers, Freiman and Sós showed the following improvement.

Theorem 1.1. — *Let $\mathcal{A}$ be a sum-free set with minimal element m and maximal element M. Under the assumption that $A = \operatorname{Card} \mathcal{A} > 0.4M + 0.8$, we have either*

(i) : *all the elements of $\mathcal{A}$ are odd,*

(ii) : *the minimal element of $\mathcal{A}$ is at least A, and we have*

$$\operatorname{Card}(\mathcal{A} \cap [1, M/2]) \leq (M - 2A + 3)/4.$$

1991 ***Mathematics Subject Classification.*** — 05 B10, 11 B13.

Key words and phrases. — Sum-free sets, additive number theory, combinatorial number theory, arithmetic progressions.

J.-M. D.: Cette recherche a bénéficié du soutien du CNRS (UMR 9936, Université Bordeaux 1) et de l'Université Victor Segalen Bordeaux 2.

Examples have been produced to show that all the bounds in the theorem are sharp. We are not going to discuss the bound in (ii), but show what may happen if the condition on A is relaxed: let s be a positive integer, and consider

$$\mathcal{A}_1 = \{s, s+1, \ldots, 2s-1\} \cap \{4s-1, \ldots, 5s-2\},$$

as well as

$$\mathcal{A}_2 = \{2, 3, 7, 8, 12, 13, \ldots, 5k-3, 5k-2, \ldots, 5s-3, 5s-2\}$$

it is easy to see that $\mathcal{A}_1$ and $\mathcal{A}_2$ are sum-free, that their cardinality, $2s$, is precisely equal to $0.4(5s-2)+0.8$, and that they are very far from satisfying properties (i) or (ii) from Theorem 1.1. A further example, with $A = 0.4M + 0.4$ is

$$\mathcal{A}_3 = \{1, 4, 6, 9, \ldots, 5k-4, 5k-1, \ldots, 5s-4, 5s-1\}$$

Our aim is to show that when A is not much less than $0.4M$, then the structure of a sum-free set is described by Theorem 1.1, or close to one of the previous examples. More precisely, we have the following

Theorem 1.2. — *Let x be a positive real numbers; there exist real number $M_0(x)$ and $C(x)$ such that for every sum-free set $\mathcal{A}$ with largest element $M \geq M_0(x)$ and cardinality $A \geq 0.4M - x$, at least one of the following properties holds true*

(i) : *all the elements of $\mathcal{A}$ are odd,*
(ii) : *all the elements of $\mathcal{A}$ are congruent to 1 or 4 modulo 5,*
(iii) : *all the elements of $\mathcal{A}$ are congruent to 2 or 3 modulo 5,*
(iv) : *the smallest elements of $\mathcal{A}$ is at least equal to A and we have $|\mathcal{A} \cap [1, M/2]| \leq (M - 2A + 3)/4$*
(v) : *$\mathcal{A}$ is included in $[\frac{M}{5} - C(x), \frac{2M}{5} + C(x)] \cup [\frac{4M}{5} - C(x), M]$.*

The constants $C(x)$ and $M_0(x)$ may be computed explicitly from our proof. However, they are not good enough to lead us to the structure of $\mathcal{A}$ when A is about $0.375M$, where new structures appear.

We may reduce the proof of Theorem 1.2 to the case when $\mathcal{A}$ contains at least one even element. From now on, we take this assumption for granted. The proof will be conducted according to the location of the smallest element m of $\mathcal{A}$: section 4 and 5 are devoted to show that m is around 1 or $M/5$, or that it is at least equal to A; the structure of $\mathcal{A}$ will be deduced from this location in section 6 and 7. Section 3 aims at filling the gap between the content of [3] and a proof of Theorem 1.1, as well as presenting in a simple frame some of the ideas that will be developed later on. In the next section, we present our notation as well as general results.

2. Notation - General results

Letters $\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$ with or without indices or other diacritical symbols denote finite sets of integers. Their cardinality is represented by $|\mathcal{A}|, |\mathcal{B}|, |\mathcal{C}|, \ldots$ or $A, B, C, \ldots$ with the same diacritical symbols. For a non empty set $\mathcal{B}$, we further let

$M(\mathcal{B})$: be its maximal element,
$m(\mathcal{B})$: be its minimal element,

$l(\mathcal{B})$: be its length, i.e. $M(\mathcal{B}) - m(\mathcal{B}) + 1$,
$d(\mathcal{B})$: be the gcd of all the differences $(b_i - b_j)$ between pairs of elements of $\mathcal{B}$,
$\mathcal{B}_+$: $= \mathcal{B} \cap [1, +\infty[$.

The letter $\mathcal{A}$ is restricted to denote a non empty sum-free set of positive integers, and we let

$$\mathcal{A}_0 = \mathcal{A} \cap 2\mathbb{Z},\ \mathcal{A}_1 = \mathcal{A} \cap (2\mathbb{Z}+1),$$

$$\mathcal{A}^- = \mathcal{A} \cap [1, M/2],\ \mathcal{A}^+ = \mathcal{A} \cap [M/2, M],$$

M (resp. m, resp. $M_0, \dots$) denote $M(\mathcal{A})$ (resp. $m(\mathcal{A})$, resp. $M(\mathcal{A}_0) \cdots$).

By x we denote a real number larger than -1. All the constants $C_1, C_2, \dots$ depend on x at most, and their value may change from one section to the other. Further, when we say that a property holds for M sufficiently large, we understand that there exist $M_0(x)$ depending on x at most, such that the considered property holds for M at least equal to $M_0(x)$.

We turn now our attention towards general results that will be used systematically, beginning with section 4.

Definition 2.1. — *A set $\mathcal{A}$ of positive integers is said to satisfy the general assumptions if it is a sum-free set that contains at least one even element and has cardinality $A = 0.4M - x$.*

Proposition 2.1. — *If $\mathcal{A}$ satisfies the general assumptions and M is large enough, we have the following properties*

(i) : *$\mathcal{A}$ contains an odd number,*
(ii) : $d(\mathcal{A}) = 1$,
(iii) : *$\mathcal{A} \cap (\mathcal{A} - \mathcal{A})$ is empty,*
(iv) : $M - m \geq 2A - 2 \Longrightarrow |(\mathcal{A} - \mathcal{A})_+| \geq \frac{3}{2}A - 2$,
(v) : $M - m \leq 2A - 3 \Longrightarrow |(\mathcal{A} - \mathcal{A})_+| \geq (M - m + A - 1)/2$,
(vi) : *for any integers u and v : $|\mathcal{A} \cap [u, u+v]| \leq (v+m)/2$,*
(vii) : *for any integer u : $|\mathcal{A} \cap [u, u+2m]| \leq m$*

Proof

(i) If $\mathcal{A}$ contains only even numbers, then the set $\mathcal{A}/2 = \{a/2 | a \in \mathcal{A}\}$ is a sum-free set that is contained in $[1, M/2]$, and so its cardinality is at most $M/4 + 1$ as can be directly seen (cf. also [4]). But $|\mathcal{A}/2| = |\mathcal{A}| = 0.4M - x$ which is larger than $M/4 + 1$ when M is large enough.

(ii) The number $d(\mathcal{A})$ is defined in such a way that $\mathcal{A}$ is included in an arithmetic progression modulo $d(\mathcal{A})$. Since $\mathcal{A}$ contains an even number (by our general assumption) as well as an odd number (by (i)), we have $d(\mathcal{A}) \neq 2$. On the other hand, we cannot have $d(\mathcal{A}) \geq 3$, otherwise $\mathcal{A}$ would have at most $M/3 + 1$ elements, which would contradict our general assumptions. Thus, $d(\mathcal{A}) = 1$.

(iii) Let $b \in \mathcal{A} \cap (\mathcal{A} - \mathcal{A})$. We can find a_1, a_2, a_3 in $\mathcal{A}$ such that $b = a_1 = a_2 - a_3$. This implies $a_2 = a_1 + a_3$, which is impossible. Thus $\mathcal{A} \cap (\mathcal{A} - \mathcal{A})$ is empty, and our argument shows even that last condition implies that $\mathcal{A}$ is sum-free.

(iv) and (v) are straightforward application of the following result ([2] and [5]):

Lemma 2.1. — *Let $\mathcal{B}$ and $\mathcal{C}$ be to finite sets of integers with $m(\mathcal{B}) = m(\mathcal{C}) = 0$, and let $M(\mathcal{B},\mathcal{C})$ be $max(M(\mathcal{B}), M(\mathcal{C}))$.*
If $M(\mathcal{B},\mathcal{C}) \leq |\mathcal{B}| + |\mathcal{C}| - 3$, then we have $|\mathcal{B} + \mathcal{C}| \geq M(\mathcal{B}) + |\mathcal{C}|$.
If $M(\mathcal{B},\mathcal{C}) \geq |\mathcal{B}| + |\mathcal{C}| - 2$ and $d(\mathcal{B} \cup \mathcal{C}) = 1$, then we have $|\mathcal{B}+\mathcal{C}| \geq M(\mathcal{B}) + |\mathcal{C}| - 3 + min(|\mathcal{B}|, |\mathcal{C}|)$.

(vi) The result is obvious when $v \leq m$, so we way assume $v > m$. We let $\mathcal{B} = \mathcal{A} \cap [u, u+v-m]$ and $\mathcal{C} = \mathcal{A} \cap [u+m, u+v]$. Since $\mathcal{A}$ is sum-free and m is in $\mathcal{A}$, we have $|\mathcal{B}| + |\mathcal{C}| \leq v - m$. Combined with the trivial upper bound $|\mathcal{A}| \leq |\mathcal{B}| + m$ and $|\mathcal{A}| \leq |\mathcal{C}| + m$, this inequality leads us to (vi).
(vii) We apply the same argument as above, leading to $|\mathcal{B}| + |\mathcal{C}| \leq v - m = m$, and further notice that $\mathcal{A} \cap]u, u+2m]$ is the union of $\mathcal{B}$ and $\mathcal{C}$.

The next results are fairly simple.

Lemma 2.2. — *Let $\mathcal{B}$ be a finite set of integers such that $2|\mathcal{B}| > l(\mathcal{B})$. Then $\mathcal{B} - \mathcal{B}$ contains $[1, 2|\mathcal{B}| - l(\mathcal{B}) - 1]$.*

Proof. — We consider a positive integer y which is not the difference of two elements of $\mathcal{B}$. We way assume $\mathcal{B} \subset [1, l(\mathcal{B})]$ and let

$$\begin{array}{ll} \mathcal{B}_1 = \mathcal{B} \cap [1, y] & , \mathcal{B}_2 = \mathcal{B} \cap [y+1, l(\mathcal{B})], \\ \mathcal{B}_3 = \mathcal{B} \cap [1, l(\mathcal{B}) - y] & , \mathcal{B}_4 = \mathcal{B} \cap [l(\mathcal{B}) - y + 1, l(\mathcal{B})]. \end{array}$$

Since y is not difference of two elements of $\mathcal{B}$, the sets $\mathcal{B}_2$ and $\mathcal{B}_3 + y$ are disjoint so that we have

$$|\mathcal{B}_2| + |\mathcal{B}_3| \leq l(\mathcal{B}) - y.$$

This easily leads to

$$2|\mathcal{B}| = |\mathcal{B}_1| + |\mathcal{B}_2| + |\mathcal{B}_3| + |\mathcal{B}_4| \leq y + l(\mathcal{B}) - y + y = l(\mathcal{B}) + y,$$

whence the inequality $y \geq 2|\mathcal{B}| - l(\mathcal{B})$.

Lemma 2.3. — *Let $\mathcal{B} = \{b_1 < b_2 < \cdots < b_B\}$ and $\mathcal{D} = \{d_1 < \cdots < d_D\}$ be to sets of integers such that we have $b_{i+1} - b_i < l(\mathcal{D})$ for $1 \leq i \leq B-1$, and $card\mathcal{D} \geq l(\mathcal{D}) - C$. We have $|\mathcal{B} + \mathcal{D}| \geq (l(\mathcal{B}) + l(\mathcal{D}) + 1)(1 - 3C/l(\mathcal{D}))$*

Proof. — Let $l(\mathcal{D}) = d_D - d_1 + 1$. We show that for any integer $u \in [b_1 + d_1, b_B + d_1[$, the interval $[u, u + l(\mathcal{D})]$ contains at most $2C$ integers which are not in $\mathcal{B} + \mathcal{D}$. We define the integer i such that $b_i + d_1 \leq u < b_{i+1} + d_1$. Since $b_{i+1} - b_i$ is less than $l(\mathcal{D})$, the interval $[u, u + l(\mathcal{D})]$ is included in $[b_i + d_1, b_{i+1} + d_D]$, which contains only elements in $\{b_i, b_{i+1}\} + \mathcal{D}$, with at most $2C$ exception. Since $[b_1 + d_1, b_B + d_D]$ can be covered with at most $(b_B + d_D + b_1 + d_1 + 1)/l(\mathcal{D}) + 1$ intervals of length $l(\mathcal{D})$, we have

$$\begin{array}{rcl} |\mathcal{B} + \mathcal{D}| & \geq & l(\mathcal{B}) + l(\mathcal{D}) + 1 - ((l(\mathcal{B}) + l(\mathcal{D}) + 1)/l(\mathcal{D}) + 1)2C \\ & \geq & (l(\mathcal{B}) + l(\mathcal{D}) + 1)(1 - 3C/l(\mathcal{D})). \end{array}$$

3. Contribution to the proof of Theorem 1.1

Combined with the result of [3], the following proposition leads to a proof of Theorem 1.1.

Proposition 3.1. — *Let $\mathcal{A}$ be a sum-free set of positive elements containing at least one even and one odd numbers, such that*

$$0.4M + 1 \leq A.$$

Then m is either smaller than $0.2M + 1$ or at least equal to $0.25M$.

Proof. — We assume on the contrary that we have

$$0.2M + 1 \leq m \leq 0.25M.$$

This condition implies the chain of inequalities

$$0 < m < M - 3m < (M - m)/2 < 2m < M - 2m < M - m < M.$$

Let $m - \eta$ denote $|\mathcal{A} \cap]M - m, M]|$. Since the interval $]M - m, M]$ is shifted from $]M - 2m, M - m]$ by m which belongs to $\mathcal{A}$, the number of elements in $]M - 2m, M - m]$ is at most η.

The two intervals $]m, M - 3m]$, $]2m, M - 2m]$ being shifted by m, there are at most $M - 4m + 1$ elements from $\mathcal{A}$ in their union.

The interval $]M - 3m, \frac{M-2m}{2}]$ contains $(M - m)/2 - M + 3m$ integers, and so at most $(M - m)/2 - M + 3m$ elements from $\mathcal{A}$.

Let now $\mathcal{B} = \mathcal{A} \cap]\frac{M-m}{2}, 2m[$; then

$$\mathcal{B} + \mathcal{B} \subset (\mathcal{A} + \mathcal{A}) \cap]M - m, 4m[\subset (\mathcal{A} + \mathcal{A}) \cap [M - m, M].$$

Since $\mathcal{A}$ is sum-free, there are at most η elements of $\mathcal{A} + \mathcal{A}$ in $]M - m, M]$, which implies that $|\mathcal{B} + \mathcal{B}|$ is at most η and so $|\mathcal{B}|$ is at most $(\eta + 1)/2$.

Putting all those upper bounds together, we obtain

$$\begin{aligned} A &\leq m - \eta + \eta + M - 4m + 1 + (M - m)/2 - M + 3m + (\eta + 1)/2 \\ &\leq (M - m + 3 + \eta)/2. \end{aligned}$$

Our last step is to obtain an upper bound for η. By Proposition 2.1, we have

$$|(\mathcal{A} - \mathcal{A})_+| > (A - 1 + M - m)/2,$$

and since $\mathcal{A}$ is sum-free, the intersection $\mathcal{A} \cap (\mathcal{A} - \mathcal{A})_+$ is empty. This implies that we have

$$|\mathcal{A}| + |(\mathcal{A} - \mathcal{A})_+| > (3A - 1 + M - m)/2;$$

the total number of elements in $[1, M]$ which are not in $\mathcal{A} \cup (\mathcal{A} - \mathcal{A})_+$ is thus less than

$$M - (3A - 1 + M - m)/2 = (M + m + 1 - 3A)/2,$$

and this is also an upper bound for $|(\mathcal{A} \cup (\mathcal{A} - \mathcal{A})_+) \cap]M - m, M]|$. Since $]M - m, M]$ contains no elements from $(\mathcal{A} - \mathcal{A})_+$, we have $\eta < (M + m + 1 - 3A)/2$, which implies

$$\begin{aligned} A &< (M-m+3)/2+(M+m+1-3A)/4A \\ &\leq (M-m+3)/2+(M+m+1-1.2M-3)/4 \\ &\leq 0.45M-0.25m+1 \\ &\leq 0.4M+0.75 < A, \end{aligned}$$

a contradiction which proves the proposition.

4. On the location of m in $[1, M/5]$

Proposition 4.1. — *Under our general assumptions, there exists C such that $m \notin [C, M/5-C]$, when M is large enough.*

Proof. — We assume that $m \in [C, M/5-C]$, and that C has been chosen sufficiently large. We have

$$M-m \geq 4M/5+C \geq 2(2M/5-x)-2 = 2A-2,$$

so that properties (iii) and (iv) from Proposition 2.1 imply

$$|\mathcal{A} \cup (\mathcal{A}-\mathcal{A})_+| = |\mathcal{A}| + |(\mathcal{A}-\mathcal{A})_+| \geq 5A/2-2 = M-C_2.$$

Since $]M-m, M] \cap (\mathcal{A}-\mathcal{A})_+$ is empty, we have

(4.1) $$|]M-m, M] \cap \mathcal{A}| \geq m - C_2,$$

which in turn implies

(4.2) $$|[1, m[\cap(\mathcal{A}-\mathcal{A})_+| \geq m-C_2.$$

On the other hand, we have

$$M \geq |\mathcal{A} \cup (\mathcal{A}-\mathcal{A}_+)| = |\mathcal{A}| + |(\mathcal{A}-\mathcal{A})_+| = 2M/5 - x + (\mathcal{A}-\mathcal{A})_+|$$

so that we get

(4.3) $$m-C_2 \leq |(\mathcal{A}-\mathcal{A})_+| \leq 3M/5+x.$$

which will be used later on.

We define the integer k and the sequence $a^{(1)} < \cdots < a^{(k)} \leq M-2m$ to be the set of elements a in $\mathcal{A}$ such that $]a-m, a[$ contains no element from $\mathcal{A}$. We further let $a^{(k+1)} = M-2m+1$, and

$$\mathcal{A}^{(i)} = [a^{(i)}, a^{(i+1)}[\cap \mathcal{A},$$

and $l^{(i)} = M(\mathcal{A}^{(i)}) - m(\mathcal{A}^{(i)}) + 1$, for $i = 1, \ldots, k$. We use (vi) (resp. (vii)) in Proposition 2.1 to get an upper bound for $A^{(i)}$ (resp. $]M-2m, M] \cap \mathcal{A}$), which leads us to

(4.4) $$2M/5 - x = A \leq \sum_{i=1}^{k} (l^{(i)}+m)/2 + m.$$

Let us consider the set $\mathcal{D} =]M-m, M] \cap \mathcal{A}$. We already noticed in (4.1) that its cardinality is at least $m-C_2$, and Lemma 2.3 leads us to

(4.5) $$|\mathcal{D} - \mathcal{A}^{(i)}| \geq (1-\frac{4C_2}{m})(l^{(i)}+m-1)\text{for} i = 1, \ldots, k.$$

We easily notice that the sets $(\mathcal{D}-\mathcal{A}^{(i)})$ are pairwise disjoint, and disjoint from $(\mathcal{A}-\mathcal{A})_+\cap[1,m[$. Relation (4.3) in conjunction with (4.2) and (4.5) implies

$$m-C_2+\sum_{i=1}^{k}(1-\frac{4C_2}{m})(l^{(i)}+m-1)\leq 3M/5+x,$$

and the use of (4.4) leads to

$$m-C_2+(1-\frac{4C_2}{m})(2A-2m)-k\leq 3M/5+x.$$

Since the $a^{(i)}$ are separated by intervals of length m, we have $km<M$, and we are led to a quadratic inequality

$$m^2-m(M/5-C_4)+C_5M>0$$

which cannot be fulfilled if $m\in[C,M/5-C]$, for C sufficiently large.

5. On the location of m in $[M/5,A]$

Proposition 5.1. — *Under our general assumptions, there exists C such that $m\notin[M/5+C,A[$, when M is large enough.*

We first assume that $m\in]M/3,A[$; in this case, we have $\mathcal{A}\subset]M-2m,M]$, and relation (vii) in Proposition (2.1) implies $A=|\mathcal{A}\cap]M-2m,M]|\leq m$, a contradiction.

We now assume that $m\in[M/5+C,M/3]$, for some sufficiently large C. We then have $M-m\leq 4M/5-C\leq 2A-3$, so that relations (iii) and (v) in Proposition (2.1) imply

$$|\mathcal{A}\cup(\mathcal{A}-\mathcal{A})_+|\geq(3M-m+A-1)/2=M-(5m-M)/10-C_1.$$

In the same way as we obtained (4.1), we get

$$|]M-m,M]\cap\mathcal{A}|\geq m-(5m-M)/10-C_1. \tag{5.1}$$

This relation will be used to get an upper bound for the cardinality of $\mathcal{B}=\mathcal{A}\cap](M-m)/2,M/2]$; we have $\mathcal{B}+\mathcal{B}\subset]M-m,M]$, so that (4.1) implies $|\mathcal{B}+\mathcal{B}|\leq(5m-M)/10+C_1$, and so we get

$$|\mathcal{B}|\leq(5m-M)/20+C_2.$$

When we combine this inequality with an easy consequence of relation (vii) in Proposition 2.1, we get

$$|\mathcal{A}\cap(]\frac{M-n}{2},\frac{M}{2}]\cup]M-2M,M])|\leq(25m-M)/20+C_2. \tag{5.2}$$

We consider finally two subcases, according as m is larger than $M/4$ or smaller. If $m\in]M/4,M/3]$, we have the chain of inequalities

$$m\leq(M-m)/2\leq M-2m\leq M/2\leq M-m\leq M,$$

so that (4.2) and a trivial upper bound $[m,(M-m)/2]$ leads to

$$\begin{aligned}A &\leq (25m-M)/20+C_2+(M-m)/2-m+1\\ &\leq \tfrac{9M}{20}-\tfrac{m}{4}+C_3\leq\tfrac{22M}{60}+C_3,\end{aligned}$$

which is less than $\frac{24M}{60} - x = A$, when M is large enough.

We are thus left to consider the case when $m \in]M/5 + C, M/4]$, in which we have the chain of inequalities

$$m \leq (M - m)/2 \leq M/2 \leq M - 2m \leq M.$$

We easily see that the interval $[m + m, (M - m)/2 + m]$ covers the interval $[M/2, M - 2m]$, so that the number of elements in $\mathcal{A}$ that lie in $[m, (M - m)/2] \cup [M/2, M - 2m]$ is at most the number of integers that lie in $[m, (M - m)/2]$. This means that we get as above

$$\begin{array}{rcl} A & \leq & \frac{9M}{20} - \frac{m}{4} + C_3 \\ & \leq & (2M)/5 + C_3 - C/4 \end{array}$$

which is again a contradiction, when C is large enough.

6. The structure of $\mathcal{A}$ when its minimal value is close to $M/5$

We prove in this section that if the minimal element m of $\mathcal{A}$ is close to $M/5$, in the sense that there exists C such that $M/5 - C < m < M/5 + C$, and $\mathcal{A}$ satisfy our general assumptions, then we are in the case (v) of Theorem 1.2.

Our first step is to show that there exist C_1 and C_2 such hat all elements from $\mathcal{A}$, with at most C_2 exception, lie in $[M/5 - C_1, 2M/5 + C_1] \cup [4M/5 - C_1, M]$. The argument is very similar to that of the previous section, so we just present a sketch of it. We have the chain of inequalities

$$m \leq (M - m)/2 \leq M/2 \leq M - 2m \leq M - m \leq M,$$

and m is about $M/5$, $(M - m)/2$ is about $2M/5$, $M - 2m$ is about $3M/5$ and $M - m$ is about $4M/5$.

We may apply (iv) or (v) from Proposition 2.1, getting $|(\mathcal{A} - \mathcal{A})_+| \geq 3M/5 - C_3$. This implies that $|\mathcal{A} \cap]M - m, M]| \geq m - C_4$ so that $|\mathcal{A} \cap]M - 2m, M - m]| \leq C_4$, as well as $|\mathcal{A} \cap](M - m)/2, M/2]| \leq C_5$ by using respectively the translation by m and the doubling argument. It remains to take care of $]M/2, M - 2m]$. Summing up what we have up to now, we know that at least $M/5 - C_6$ elements of $\mathcal{A}$ are located in $[m, M - 3m] \cup]M/2, M - 2m]$. By translating by m, we know that there are at most $M/10 + C_7$ elements of $\mathcal{A}$ in $]\frac{M}{2} - m, M - 3m] \cup]M/2, M - 2m]$, so that there remain at least $M/10 + C_8$ elements of $\mathcal{A}$ in $[m, M/2 - m]$. This implies that $\mathcal{A} + \mathcal{A}$ almost covers $[2m, M - 2m]$, so that it almost covers $]M/2, M - 2m]$, whence there are at most C_9 elements of $\mathcal{A}$ in $]M/2, M - 2m]$, which ends the proof of the first step.

In the second and last step, we show that there is no element of $\mathcal{A}$ in $I =]2M/5 + 2C_2 + 2 - 2C_1, 4M/5 + C_1 - 2C_2 - 2[$. Let indeed y be an element in this set. Since we have $M/5 - C_1 + y < M - 2C_2 + 2$, and $2M/5 + C_1 + y > 4M/5 - C_1 + 2C_2 + 2$ the two intervals $[M/5 - C_1 + y, 2M/5 + C_1 + y]$ and $[4M/5 - C_1, M]$ have at least $2C_2 + 1$ integers in common. Thanks to the first step and the pigeon-hole principle, we know that there exist a_1 and a_2 in $\mathcal{A}$ such that $a_1 + y = a_2$, thus y cannot belong to $\mathcal{A}$, and $\mathcal{A}$ is concentrated in $[m, 2M/5 + C_{10}] \cup [4M/5 - C_{10}, M]$ as we wished to show.

7. Some properties of $\mathcal{A}$ when m is small

We recall our notation, namely

$$\mathcal{A}_0 = \mathcal{A} \cap 2\mathbb{Z},\ \mathcal{A}_1 = \mathcal{A} \cap (2\mathbb{Z}+1),$$
$$\mathcal{A}^- = \mathcal{A} \cap [1, M/2],\ \mathcal{A}^+ = \mathcal{A} \cap [M/2, M],$$

$m = m^- = \min(\mathcal{A})$, $M = M^+ = \max(\mathcal{A})$, $m_0 = \min(\mathcal{A}_0)$.

In his section we prove the following.

Proposition 7.1. — *Let $\mathcal{A}$ satisfy our general assumptions, and be such that $m < M/20$. There exists C such that, when M is large enough, we have*

$$||\mathcal{A}^-| - M/5| \leq C, \tag{7.1}$$

$$m_0 \leq C. \tag{7.2}$$

The proof will be led in three steps, where we prove that (7.1) holds, then that we have the following inequality

$$|\mathcal{A}_0| \geq M/5 - C, \tag{7.3}$$

and finally that (7.2) holds.

7.1. The set $\mathcal{A}$ is balanced between small and large elements. — We first show that $\mathcal{A}^-$ cannot be too large. Indeed, if $|\mathcal{A}^-| > M/5+2$, we may apply Theorem 1.1, and, since $m(\mathcal{A}^-) = m < M/20$, the set $\mathcal{A}^-$ consists only of odd elements, so that $m_0 \geq M/2$. There are at most $m_0/4$ elements from $\mathcal{A}$ in $[1, m_0[$, since they are odd, m_0 is in $\mathcal{A}$, and at most $(M - m_0 + M/20)/2$ elements from $\mathcal{A}$ in $]m_0, M]$, so that $2M/5 - x \leq m_0/4 + (M - m_0 + M/20)/2 + 1$, which implies $M/2 \leq m_0 \leq 9M/20 + C_1$, a contradiction.

We now show that the two simultaneous relations $|\mathcal{A}^+| > M/5 + C_2$ and $d(\mathcal{A}^+) > 1$ lead to a contradiction. We first notice that $\mathcal{A}^+$ contains at least $M/5$ elements, so that $d(\mathcal{A}^+) > 1$ is equivalent to $d(\mathcal{A}^+) = 2$, i.e. $\mathcal{A}^+$ consists only of odd integers, or of even integers. In either case, Lemma 2.2 implies that $(\mathcal{A}^+ - \mathcal{A}^+)$ contains all the non-negative even integers at most equal to $4|\mathcal{A}^+| - M/2$. So we have $m_0 > |\mathcal{A}^+| - M/2 \geq 3M/10$.

Assume that we have $|\mathcal{A}^+| > M/5 + C_2$ and that $\mathcal{A}^+$ consists only of even numbers. We have already shown that all the elements in $\mathcal{A} \cap [1, M/4]$ are odd, so there are at most $M/8$ of them. Thus, $\mathcal{A} \cap]M/4, M/2]$ has at least $|\mathcal{A}^-| - M/8$ elements. By doubling them, we obtain $|\mathcal{A}^-| - M/8$ even numbers in $]M/2, M] \setminus \mathcal{A}$. The total number of even number in $]M/2, M]$, which is about $M/4$, must be at least $|\mathcal{A}^+| + |\mathcal{A}^-| - M/8$, leading to a contradiction with $|\mathcal{A}| = |\mathcal{A}^+| + |\mathcal{A}^-| = 2M/5 - x$.

We now assume that $|\mathcal{A}^+| > M/5 + C_2$ and that $\mathcal{A}^+$ consists of odd numbers, so that $M_0 \leq M/2$. Let u be the number of odd elements less that M_0 which are in $\mathcal{A}$. There are at most $M_0/2 - u$ odd elements in $\mathcal{A} \cap]M_0, 2M_0]$, since $(2a+1) + M_0$ is odd and cannot be in $\mathcal{A}$ when $2a+1$ is in $\mathcal{A}$. In $\mathcal{A}$, there are thus at most $M_0/2 + (M - 2M_0)/2$ odd elements, and the number of even elements in $\mathcal{A}$ is at least $2M/5 - x - (M - M_0)/2$. The largest even element in $\mathcal{A}$ is at least $m_0 + 2(2M/5 -$

$x - (M - M_0)/2) = m_0 + M_0 - M/5 - 2x$, which implies $m_0 \leq M/5 + 2x$, which contradicts the inequality $m_0 \geq 3M/10$ we already obtained.

It remains to show that when $d(\mathcal{A}^+) = 1$, the set $\mathcal{A}^+$ cannot be too large. We apply Lemma 2.1 with $\mathcal{B} = \mathcal{A}^+ - \{m^+\}$ and $\mathcal{C} = \{M^+\} - \mathcal{A}^+$. By Proposition 2.1, we have $m < C_3$, and this implies that $M(\mathcal{B})$ is larger than $2|\mathcal{B}| - C_4$. Either case of Lemma 2.1 leads to $|\mathcal{B}+\mathcal{C}| \geq 3|\mathcal{A}^+| - C_5$, so that we have $|(\mathcal{A}^+ - \mathcal{A}^+)_+| \geq 3|\mathcal{A}^+|/2 - C_6$. But the set $(\mathcal{A}^+ - \mathcal{A}^+)_+$ is included in $[1, M/2]$ and disjoint from $\mathcal{A}^-$, so that we have $3|\mathcal{A}^+|/2 + |\mathcal{A}^-| \leq M/2 + C_7$, or $|\mathcal{A}^+|/2 \leq M/2 + C_7 - |\mathcal{A}| = M/2 + C_7 - 2M/5 + x$, which implies $|\mathcal{A}^+| \leq M/5 + C_8$, or $|\mathcal{A}^-| \geq M/5 - C_9$.

We have so far proved that (7.1) holds.

7.2. The set $\mathcal{A}$ contains many even numbers. — We assume in this subsection that (7.3) does not hold, so that we have $|\mathcal{A}_1| > |\mathcal{A}_0|$.

Since m_0 in the least even element in $\mathcal{A}$, we have $|\mathcal{A} \cap [1, m_0]| \leq m_0/4$, and because $m < C_2$, we have $|\mathcal{A} \cap [m_0, M]| \leq (M - m_0)/2 + C_3$. We thus have

$$2M/5 - x = |\mathcal{A}| \leq m_0/4 + (M - m_0)/2 + C_3, \text{ whence } m_0 \leq 2M/5 + C_4.$$

We may apply he same reasoning to $\mathcal{A}^-$, since we now know that $m_0 < M/2$; using (7.1), we get $m_0 \leq M/5 + C_5$.

By repeating the argument used in previous section, as well as the previous one, we may show, that, up to a constant, $|\mathcal{A} \cap [1, M/4]|$ is about $M/10$, so that $m_0 \leq M/10 + C_6$. We may reduce further the bound on m_0 by the same type of idea, but this would lead us only to $m_0 \leq \varepsilon M$ for any positive ε which is not as strong an inequality as the one we need.

We wish to apply Lemma 2.1 with

$$\mathcal{B} = \{(a_0 - m_0)/2,\ a_0 \in \mathcal{A}_0\} \text{ and } \mathcal{C} = \{(M_1 - a_1)/2, a_1 \in \mathcal{A}_1\}.$$

We have $|\mathcal{B}| = |\mathcal{A}_0|$, $|\mathcal{C}| = |\mathcal{A}_1|$ and

$$\max(M(\mathcal{B}), M(\mathcal{C})) \geq (M - M/10 - C_6)/2 > |\mathcal{A}| - 3 = |\mathcal{B}| + |\mathcal{C}| - 3.$$

Since $\mathcal{B} \cup \mathcal{C}$ contains more than $M/6$ elements and is included in $[0, M/2[$, we have $d(\mathcal{B} \cup \mathcal{C}) = 1$ or 2. We first show that when $d(\mathcal{B} \cup \mathcal{C}) = 2$ then $\mathcal{A}_0$ is large.

When $d(\mathcal{B} \cup \mathcal{C}) = 2$, even elements of $\mathcal{A}$ are either all congruent to 0 modulo 4 or all congruent to 2 modulo 4, and in the same way, odd elements of $\mathcal{A}$ are either all congruent to 1 modulo 4, or all congruent to 3 modulo 4.

If m_0 is congruent to 0 modulo 4, then the set $\{m_0\} + \mathcal{A}_1$ and $\mathcal{A}_1$ are disjoint, in the same class modulo 4 and included in $[1, M + M/10 + C_6]$, so that $2|\mathcal{A}_1| \leq 11M/40 + C_6$, in contradiction to $|\mathcal{A}_1| > M/5 - C_1$.

If all the even elements are congruent to 2 modulo 4, we are going to use the fact that the sum of two elements in $\mathcal{A}_1$ is also congruent to 2 modulo 4. We first notice that the number of elements in $\mathcal{A}_1$ is at most $M/4$, so that $|\mathcal{A}_0|$ is least $2M/5 - x - M/4 = 3M/20 - x$, which implies that M_0 is at least equal to $3M/5 - x$. The number of odd elements is at most $M_0/8 + (M - M_0)/4 + 1$, which is less than $7M/40 + C$, contradicting our assumption that $|\mathcal{A}_1| > 8M/40 - C_1$.

We now know that $d(\mathcal{B}\cup\mathcal{C})=1$, and Lemma 2.1 leads to

$$|(\mathcal{A}_0-\{m_0\})+(\{l_1\}-\mathcal{A}_1)|\geq|\mathcal{A}_1|+2|\mathcal{A}_0|-3.$$

Since $(\mathcal{A}_0-\mathcal{A}_1)_+\cap\mathcal{A}_1=\varnothing$, we get $|\mathcal{A}_1|+|(\mathcal{A}_0+\mathcal{A}_1)_+|\leq M/2$, and in the same way $|\mathcal{A}_1|+|(\mathcal{A}_1+\mathcal{A}_0)_+|\leq M/2$. This leads to

$$M\geq 2|\mathcal{A}_1|+|\mathcal{A}_1+\mathcal{A}_0|\geq 3|\mathcal{A}_1|+2|\mathcal{A}_0|-3\geq|\mathcal{A}_1|+4M/5-2x-3,$$

which implies that $|\mathcal{A}_1|\leq M/5+2x+3$, whence (7.3) holds.

7.3. The set $\mathcal{A}$ contains a small even number. — Since we have (7.3), we may apply to the set $\{a_0/2, a_0\in\mathcal{A}_0\}$ the result we have obtained so far. One of the following cases holds

(i) : $\mathcal{A}_0\subset 4\mathbb{Z}+2$,
(ii) : $m_0>2M/5+C_1$,
(iii) : $\mathcal{A}_0\subset[M/5-C_1,2M/5+C_1]\cup[4M/5-C_1,M]$,
(iv) : $m_0<C$,

so that we just have to rule out the first three cases in order to complete the proof of Proposition 7.1.

Case (i) cannot hold because the sets $\{2a_1,a_1\in\mathcal{A}_1\}$ and $\mathcal{A}_0$ are disjoint, included in $[1,M]\cap(4\mathbb{Z}+2)$, and the cardinality of their union is $\mathcal{A}$ which is larger than $M/4+1$.

Cases (ii) and (iii) cannot hold, because the argument we used at the beginning of (7.2) implies that m_0 is less than $M/10$, up to a constant.

8. End of the proof of Theorem 1.2

Let $\mathcal{A}$ be a sum-free set satisfying our general assumptions. We know that $m\in[1,C]\cup[M/5-C,M/5+C]\cup[A,M]$. We have already shown that $m\in[M/5-C,M/5+C]$ leads to case (v) in Theorem 1.2. The argument given in [3] for the second case in Theorem 1.1 implies that $m\in[A,M]$ leads to case (iv). It remains to show that $m\leq C$ leads to case (ii) or (iii). We shall make use of Proposition 7.1 and retain in the sequel the notation C for constant implied in (7.1) and (7.2). We let $C_1=38C+60$.

Our first task is to show that we can find a_1 and a_2 in $\mathcal{A}\cap[M/2-C_1,M/2]$ such that $a_2=a_1+m_0/2$. We assume that it is not the case; by this assumption and the fact that m_0 is in $\mathcal{A}$, any interval of length $3m_0/2$ in $[M/2-C_1,M/2]$ contains at most $m_0/2$ elements from $\mathcal{A}$. We thus have $|\mathcal{A}\cap[M/2-C_1,M/2]|\leq(C_1/3)+(3m_0/2)$; this implies that $|\mathcal{A}\cap[1,M/2-C_1]|>2(M/2-C_1)/5+4$, and we now have a contradiction with Theorem 1.1 applied to $\mathcal{A}\cap[1,M/2-C_1]$ and the fact that this set contains m_0, a small even integer. So there exist a_1 and a_2 with the prescribed properties.

Let us now define, for $i=1$ and 2,

$$f_i(t)=\begin{cases}t+a_i & when\quad t\leq M/2\\ t-a_i & when\quad t>M/2\end{cases}$$

There exists C_2 such that $||f_i(\mathcal{B})\cap[1,M]|-|\mathcal{B}|| \leq C_2$ for any $\mathcal{B}$ in $[1,M]$, and $i \in \{1,2\}$. Since we clearly have

(i) : $f_1(\mathcal{A}) \cap \mathcal{A} = f_2(\mathcal{A}) \cap \mathcal{A} = \varnothing$,
(ii) : we must have
(iii) : $f_1(\mathcal{A}) \cap f_2(\mathcal{A}) > M/5 - C_3$.

We can find C_4 so that the relations

$$\begin{cases} a'_1 \in [C_4, M/2 - C_4] \cup [M/2 + C_4, M - C_4], \\ f_1(a'_1) = f_2(a'_2), \end{cases}$$

imply $|a'_1 - a'_2| = m_0/2$. Moreover, for any number t, at most one of $t + m_0/2$ and $t - m_0/2$ belongs to $\mathcal{A}$. All that imply that, with at most C_5 exceptions, all elements in $\mathcal{A}$ can be organized in pairs with common difference $m_0/2$.

The largest such pair is larger than $M - C_6$, for a suitable C_6. Otherwise, there are no more than C_5 elements in $\mathcal{A} \cap [M - C_6, M]$, and so there are more than $2M/5 - x - C_5$, i.e. more than $2(M - C_6)/5 + 2$ elements in $\mathcal{A} \cap [1, M - C_6]$, which is in contradiction with Theorem 1.1 and the fact that $\mathcal{A}$ contains a small even element. Let us call $(M - C_7, M - C_7 + m_0/2)$ the largest pair of elements in $\mathcal{A}$.

To each pair $(a, a + m_0/2)$, we associate a triple $(M - C_7 - a - m_0/2, M - C_7 - a, M - C_7 - a + m_0/2)$ of integers that do not belong to $\mathcal{A}$. Since two pairs have no element in common and difference between first and second elements of different pairs is not equal to m, two such triples have no element in common neither. The set $[1, M]\backslash\mathcal{A}$ contains $3M/5$ elements, up to a constant, and there are, again up to a constant, $M/5$ triples, so up a constant number of exceptions, $[1, M]\backslash\mathcal{A}$ is a union of triples.

Let us consider any arithmetic progression modulo $m_0/2$. Up to a constant number of exceptions, the progression is covered by $(M/5)/(m_0/2)$ structures of five consecutive points, the first two belonging to $\mathcal{A}$, and the last three not belonging to $\mathcal{A}$. We consider the set $\mathcal{B}$ of the first element in each pentuple. If $b_1 < b_2$ are two elements in $\mathcal{B}$, then $b_2 - b_1$ is the midpoint of a triple of elements that do not belong to $\mathcal{A}$. Since there are $M/5 - C_8$ such triples, we have $|(\mathcal{B} - \mathcal{B})_+| = |\mathcal{B}| + C_9$, the reasoning as in [5] imply that $\mathcal{B}$ is located in an arithmetic progression of length at most $|\mathcal{B}| + C_{10}$. Due to the cardinality of $\mathcal{B}$, this progression is modulo $5m_0/2$.

We consider the set $\mathcal{S}$ of the residues modulo $5m_0/2$ of the first and second elements of the pentuples associated to each arithmetic progression modulo $m_0/2$. The set $\mathcal{S}$ consist of m_0 elements. Since $\mathcal{A}$ is sum-free and is equal, up to a constant number of terms, to the numbers in $[1, M]$ which are above the elements of $\mathcal{S}$, the set $\mathcal{S}$ must be sum-free. We are thus left with the characterization of sum-free subsets of $\mathbb{Z}/5L\mathbb{Z}$ which satisfy the following property: each subset $\{u, u + L, u + 2L, u + 3L, u + 4L\}$ contains exactly two elements, and those two elements are consecutive; by this we mean that those two elements are $\{u + iL, u + jL\}$ where (i, j) is one of the pair (0,1),(1,2),(2,3),(3,4),(4,0), and we shall call the first of those two elements the one that corresponds to the first element in the associated pair (i, j). We call $\mathcal{S}_1$ the set of the first elements, and $\mathcal{S}_2$ the set of the second ones. We have $|\mathcal{S}_1| = |\mathcal{S}_2| = L$. Let s and s' be elements in $\mathcal{S}_1$; then $s + L$ and $s' + L$ are in $\mathcal{S}_2$, so that $s + s'$, $s + s' +$

L, $s+s'+2L$ are not in $\mathcal{S}$ which is sum-free; this implies that $s+s'-2L$ is in $\mathcal{S}_1$. This implies that $|\mathcal{S}_1+\mathcal{S}_1|=|\mathcal{S}_1|$, and Kneser's theorem (cf. [4]) implies that $\mathcal{S}_1$ is a coset associated to a subgroup $\mathcal{H}$ of $\mathbb{Z}/5L\mathbb{Z}$ with cardinality L. This implies that $\mathcal{S}_1$, as well as $\mathcal{S}_2$, is the image in $\mathbb{Z}/5L\mathbb{Z}$ of an arithmetic progression modulo 5.

We have thus proved that, up to a constant number of terms, $\mathcal{A}$ is the union of two arithmetic progressions modulo 5. Since $\mathcal{A}$ is sum-free it is easily seen that those two arithmetic progression are either $5\mathbb{Z}+1$ and $5\mathbb{Z}+4$, or $5\mathbb{Z}+2$ and $5\mathbb{Z}+3$, and that $\mathcal{A}$ is indeed included in the two arithmetic progressions, so that either (ii) or (iii) holds in Theorem 1.2.

References

[1] Cameron P.J. and Erdős P., *On the Number of Sets of Integers With Various Properties*, Proceeding of the first Conference of the CNTA, R.Molin ed., Alberta, April 17-27, 1988.

[2] Freiman G. A., *Inverse problems in additive number theory, VI.* On addition of finite sets, III. The method of trigonometric sums (Russian), Izvestiya Vuzov, Mathem., **3(28)**, 1962, 151–157.

[3] Freiman G.A., *On the structure and the number of sum-free sets*, Astérisque, **209**, 1992, 195–201.

[4] Mann H. B., *Addition Theorems*, Wiley, New-York, 1965, xi+114 p.

[5] Steinig J., *On Freiman's theorems concerning the sum of finite sets of integers*, this volume.

J.-M. Deshouillers, Mathématiques Stochastiques, Université Victor Segalen Bordeaux 2, 33076 Bordeaux, France • *E-mail :* j-m.deshouillers@u-bordeaux2.fr

G.A. Freiman, School of Mathematical Science, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, 69978 Tel Aviv, Israel • *E-mail :* grisha@math.tau.ac.il

V. Sós, Math. Institute,, Hungarian Academy of Sciences,, Realtanoda u.13-15,, Budapest, Hungary *E-mail :* sos@math-inst.hu

M. Temkin, School of Mathematical Science, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, 69978 Tel Aviv, Israel

Astérisque

GREGORY A. FREIMAN

LEWIS LOW

JANE PITMAN

Sumsets with distinct summands and the Erdős-Heilbronn conjecture on sums of residues

Astérisque, tome 258 (1999), p. 163-172

<http://www.numdam.org/item?id=AST_1999__258__163_0>

Astérisque **258**, 1999, p. 163–172

SUMSETS WITH DISTINCT SUMMANDS AND THE ERDŐS-HEILBRONN CONJECTURE ON SUMS OF RESIDUES

by

Gregory A. Freiman, Lewis Low & Jane Pitman

Abstract. — Let S be a set of integers or of residue classes modulo a prime p, with cardinality $|S| = k$, and let T be the set of all sums of two distinct elements of S. For the integer case, it is shown that if $|T|$ is less than approximately $2.5k$ then S is contained in an arithmetic progression with relatively small cardinality. For the residue class case a result of this type is derived provided that $k > 60$ and $p > 50k$. As an application, it is shown that $|T| \geq 2k-3$ under these conditions. Earlier results of Freiman play an essential role in the proofs.

1. Introduction

1.1. Let Z be the set of all integers and let F_p be the finite field of residue classes modulo p, where p is a prime number. If A is a subset of Z or F_p (written $A \subset Z$ or $A \subset F_p$) we denote the cardinality of A by $|A|$. For a finite subset A of Z or F_p we shall consider $2A$, the set of all sums of two elements of A, and also $2^\wedge A$, the set of all sums of two *distinct* elements of A, that is,

$$\begin{aligned} 2A &= \{a+b : a,b \in A\}, \\ 2^\wedge A &= \{a+b : a,b \in A,\ a \neq b\}. \end{aligned}$$

1.2. Sums of elements from a set of integers. — First we consider the sumset $2A$ for $A \subset Z$. We write

$$A = \{a_0, a_1, \ldots, a_{k-1}\}, \quad k = |A|,$$

where

$$a_0 < a_1 < \cdots < a_{k-1}\ .$$

Since the $k-1$ sums $a_i + a_{i+1}$ and the k sums $a_i + a_i = 2a_i$ are all distinct we have

$$|2A| \geq 2k-1, \tag{1}$$

1991 ***Mathematics Subject Classification.*** — 11 B75, 11 B13.
Key words and phrases. — sumset, set addition, sums of residues.

and it is easily seen that equality holds if and only if A is an arithmetic progression (that is, the differences $a_{i+1} - a_i$ are all equal). Freiman [6, page 11] has proved the following more precise result (which will be used in Section 2 below).

Theorem A **(Freiman).** — *Let $D \subset Z$. If $|2D| \leq 2|D| - 1 + C_1$, where $C_1 \leq |D| - 3$, then $D \subset L$, where L is an arithmetic progression such that*

$$|L| \leq |D| + C_1$$

(so that D is obtained by deleting at most C_1 terms from the arithmetic progression L).

1.3. Sums of elements from a subset of F_p. — Next we look at $2A$ for $A \subset F_p$ such that $|A| = k$. By analogy with (1) we have the following special case of the well-known Cauchy-Davenport theorem:

$$|2A| \geq \min(p, 2k - 1). \tag{2}$$

More detailed results have been obtained by various authors and Freiman [6, p.46] has used the above theorem on $2D$ for $D \subset Z$ to obtain the following result in the same vein for $A \subset F_p$.

Theorem B **(Freiman).** — *Let $A \subset F_p$ such that*

$$|A| = k < p/35.$$

Suppose that $|2A| = 2k - 1 + b$, $b < 0.4k - 2$. Then $A \subset L$, where L is an arithmetic progression in F_p such that $|L| = k + b$.

1.4. Sums with distinct summands and the Erdős-Heilbronn conjecture

For $A \subset Z$ as in 1.2 above, the $k - 1$ sums $a_i + a_{i+1}$ and the $k - 2$ sums $a_i + a_{i+2}$ are all distinct and belong to $2^\wedge A$. Thus

$$|2^\wedge A| \geq 2k - 3, \tag{3}$$

and it can be checked that for $k \geq 5$ equality holds if and only if A is an arithmetic progression.

By analogy with the Cauchy-Davenport theorem (2), Erdős and Heilbronn conjectured in the 1960's (see Erdős and Graham [5]) that for $A \subset F_p$ such that $|A| = k$ we must have

$$|2^\wedge A| \geq \min(p, 2k - 3). \tag{4}$$

Although there is a short elementary proof of (2) (see, for example, Davenport [3]), the corresponding result for distinct summands seems to be more difficult. As discussed further below, the full conjecture (4) has been proved since 1993. The main published contribution prior to 1993 seems to be that of Mansfield [7], who proved the following theorem.

Theorem **(Mansfield).** — *Let $A \subset F_p$ such that $|A| = k$. Then the Erdős-Heilbronn conjecture (4) is true if*

$$\text{either} \quad k \leq 11 \quad \text{or} \quad 2^{k-1} \leq p.$$

Our aim in this paper was to develop analogues for $2^\wedge A$ of Freiman's results on $2D$, both for $D \subset Z$ and for $A \subset F_p$, which would be strong enough for the purposes of proving (4) for a wider range, as well as being of independent interest.

1.5. Results obtained. — In Section 2 we use simple combinatorial arguments together with Freiman's theorem on $2D$ for $D \subset Z$ to prove the following theorem.

Theorem 1. — *Let D be a set of k integers for which*

$$|2^\wedge D| \leq 2k - 3 + C,$$

where

$$0 \leq C \leq \frac{1}{2}(k-5).$$

Then D is contained in an arithmetic progression L such that

$$|L| \leq k + 2C + 2.$$

In Section 3 we use Theorem 1 and arguments based on trigonometric sums to prove the main result of the paper, which is as follows.

Theorem 2. — *Let $A \subset F_p$ such that*

$$|A| = k < \frac{p}{50}, \quad k > 60.$$

Suppose that

$$|2^\wedge A| \leq 2k - 3 + C,$$

where $C < 0.06k$. Then $A \subset L$, where L is an arithmetic progression in F_p such that

$$|L| \leq k + 2C + 2.$$

As a corollary, we will show that for $A \subset F_p$ such that $|A| = k$ the Erdős-Heilbronn conjecture (4) is true if

$$k < \frac{p}{50}, \quad k > 60. \tag{5}$$

Pybus [10] told us that he had obtained a proof of a version of the Erdős-Heilbronn conjecture based on different ideas. More recent work by others, including proofs of the full conjecture, will be discussed in Section 4 at the end of the paper.

1.6. Isomorphisms. — We note that the sumsets $2A$ and $2^\wedge A$ can be considered for any set A with addition. If A and B are two sets, each with an addition, and $\phi : A \to B$ is a bijection, we call ϕ an isomorphism if and only if

$$\phi(a) + \phi(b) = \phi(c) + \phi(d) \Leftrightarrow a + b = c + d.$$

We call A and B as above *isomorphic* if such an isomorphism exists, in which case we have

$$|2A| = |2B| \quad \text{and} \quad |2^\wedge A| = |2^\wedge B|.$$

We shall use the fact that affine transformations of Z or F_p are isomorphisms.

2. Sums of distinct elements from a set of integers

2.1. In this section we consider a set of integers A such that $|A| = k$ and use notation as at the beginning of section 1.2, together with some further vocabulary as follows. We note that $2^\wedge A$ is isomorphic to the set

$$\{\tfrac{1}{2}(a_i + a_j) : 1 \leq i < j \leq n\}$$

and it is helpful to think geometrically in terms of the points a_i and the mid-points of pairs a_i, a_j $(i < j)$. We shall say that a_i is *representable* if and only if a_i coincides with one of the mid-points, that is

$$2a_i \in 2^\wedge A.$$

We shall call a sum $a_i + a_{i+s}$ with $s \geq 1$ an *s-step sum*, and we recall that the 1-step and 2-step sums are all distinct. For $s \geq 1$, an s-step sum will be called *new* if and only if it is not equal to any j-step sum with $1 \leq j < s$. All 1-step and 2-step sums are new, but for $s \geq 3$ an s-step sum is not necessarily new. We shall use the notations

$$\begin{aligned} k_1 = k_1(A) &= \text{ total number of } \textit{new } s\text{-step sums with } s \geq 3, \\ k_2 = k_2(A) &= \text{ number of } a_j\text{'s which are } \textit{representable.} \end{aligned}$$

If an s-step sum a_i+a_{i+s} is not new, then for some j, k such that $i < j < j+k < i+s$ we must have

$$a_i + a_{i+s} = a_j + a_{j+k}$$

and hence

$$0 < a_j - a_i = a_{i+s} - a_{j+k}\ .$$

We therefore consider the associated *difference set*

$$\mathcal{D}(a_i, a_{i+s}) = (a_{i+1} - a_i, a_{i+2} - a_{i+1}, \ldots, a_{i+s} - a_{i+s-1}).$$

Our proof of Theorem 1 will be based on the following lemma.

Lemma. — *For $A \subset Z$ such that $|A| = k$, $k \geq 5$, let k_1, k_2 be the number of new s-step sums with $s \geq 3$ and the number of representable elements of A as defined above. Then*

$$k_1 + k_2 \geq k - 4. \tag{6}$$

Proof. — Consider a particular subscript i such that $0 \leq i \leq k - 5$. If $a_i + a_{i+3}$ is not new we must have

$$\mathcal{D}(a_i, a_{i+3}) = (x, y, x)$$

for some $x, y > 0$, and so

$$\mathcal{D}(a_i, a_{i+4}) = (x, y, x, z)$$

for some z. If $z = x$ or $z = x + y$ then a_{i+3} is a mid-point and so is representable, while if $z \neq x$ and $z \neq x + y$ then $a_i + a_{i+4}$ is new. Thus at least one of the following three statements holds:

(i) $a_i + a_{i+3}$ is new; (ii) $a_i + a_{i+4}$ is new; (iii) a_{i+3} is representable.

This is true for $i = 0, 1, \ldots, k-5$; the new sums arising from (i) and (ii) for different i's are distinct, and the representable elements arising from (iii) are also distinct. Hence at least one element counted in $k_1 + k_2$ arises in this way from each of the $k-4$ possible values of the subscript i and so (6) follows.

We note that the above argument involves only 3-step and 4-step sums. By more detailed arguments using s-step sums with $s \geq 5$ it can be shown that in fact

$$k_1 + k_2 \geq k - 2 \quad \text{for} \quad k \geq 8. \tag{7}$$

2.2. Proof of Theorem 1. — We now consider $D \subset Z$ such that $|D| = k$. Let $k_1 = k_1(D)$, $k_2 = k_2(D)$ and suppose that D satisfies the hypotheses of Theorem 1, so that

$$|2^\wedge D| \leq 2k - 3 + C \tag{8}$$

$$0 \leq C \leq \tfrac{1}{2}(k-5). \tag{9}$$

Since

$$(2D)\backslash(2^\wedge D) = \{2d \mid d \in D, \ \ d \text{ is not representable}\}, \tag{10}$$

we have

$$|2D| = |2^\wedge D| + k - k_2 \,. \tag{11}$$

Using (8) and the above lemma we obtain

$$|2D| \leq 2k - 3 + C + k - k_2 = 3k - 3 + C - k_2 = 3k - 3 + C + k_1 - (k_1 + k_2)$$

$$\leq 3k - 3 + C + k_1 - (k-4) = 2k + 1 + C + k_1 \,. \tag{12}$$

The number of 1-step sums is equal to $k-1$, the number of 2-step sums is equal to $k-2$, and the number of new sums (different from these) is equal to k_1. Thus, we have

$$|2^\wedge D| = 2k - 3 + k_1 \,,$$

and hence by (8),

$$k_1 \leq C.$$

Applying this inequality in (12), we get

$$|2D| \leq 2|D| - 1 + 2C + 2. \tag{13}$$

It now follows from (9) and (13), by Theorem A in Section 1.2, that $D \subset L$, where L is an arithmetic progression such that $|L| \leq |D| + 2C + 2$, as required.

3. Sums of distinct summands from a subset of F_p

3.1. The proof of our main result, Theorem 2, will depend on the use of trigonometric sums. We view the elements of F_p as residue classes modulo p, and note that for $a \in Z$ and $x \in F_p$, $e^{2\pi i\, a\, x/p}$ is defined uniquely by

$$e^{2\pi i\, a\, x/p} = e^{2\pi i\, a\, x_0/p} \,,$$

where x_0 is any representative residue belonging to the residue class x.

For finite sets $A \subset F_p$ we shall consider trigonometric sums of the form

$$T = \sum_{x \in A} e^{2\pi i\, a\, x/p}\ . \tag{14}$$

We note that for such sums it is easily checked that if $k = |A|$ then

$$\sum_{a=1}^{p-1} \Big| \sum_{x \in A} e^{2\pi i\, a\, x/p} \Big|^2 \leq pk - k^2\ . \tag{15}$$

3.2. We shall need the following lemma of Freiman [6].

Lemma. — *Let A be a subset of F_p such that $|A| = k$, and let $a \in Z$ such that $a \not\equiv 0$ (mod p) and let T be the corresponding trigonometric sum defined by (14). Suppose that $|T| > C_0 k$, where $0 < C_0 < 1$. Then, for some u and v in F_p such that $v \neq 0$, at least*

$$\frac{1}{2}\,(C_0 + 1)k$$

distinct elements of A belong to the arithmetic progression

$$\{u + sv : 0 \leq s \leq \tfrac{p-1}{2}\}.$$

Proof. — See Freiman [6], Section 1 of Chapter II, Corollary to Lemma on pages 46-47 and discussion on page 50.

3.3. Proof of Theorem 2. — We now turn to the proof of Theorem 2. We therefore consider $A \subset F_p$ such that

$$|A| = k < p/50, \tag{16}$$

$$|2^\wedge A| \leq 2k - 3 + C, \quad C < 0.06k, \quad k > 60. \tag{17}$$

Consider the sum

$$S = \sum_{a=0}^{p-1} \sum_{x_1, x_2 \in A} \sum_{x_3 \in 2^\wedge A} e^{2\pi i (a/p)(x_1 + x_2 - x_3)}\ .$$

We divide the sum S into two parts,

$$S = \sum_{a=0}^{p-1} \sum_{x_1, x_2 \in A} \sum_{x_3 \in 2A} - \sum_{a=0}^{p-1} \sum_{x_1, x_2 \in A} \sum_{x_3 \in (2A) \setminus (2^\wedge A)} = S_1 - S_2\ , \tag{18}$$

say. Since each pair x_1, x_2 of elements of A yields exactly one x_3 in $2A$ such that $x_1 + x_2 = x_3$, we have

$$S_1 = k^2 p \tag{19}$$

(as in Freiman [6], p.48 (2.3.2)).

Denote by B the set of all elements of A which are *not* representable. Then, in view of (10) we have

$$S_2 = \sum_{a=0}^{p-1} \sum_{x_1, x_2 \in A} \sum_{a_j \in B} e^{2\pi i (a/p)(x_1 + x_2 - 2a_j)}\ .$$

For a_j in B, the equation $x_1+x_2-2a_j=0$ holds only if $x_1=x_2=a_j$ and therefore

$$S_2 = p|B|. \tag{20}$$

It follows from (18), (19) and (20) that

$$S \geq p(k^2-k). \tag{21}$$

Then from (21) and the definition of S, we obtain

$$p(k^2-k) \leq \sum_{a=0}^{p-1} \left| \sum_{x_1,x_2\in A} \sum_{x_3\in 2^\wedge A} e^{2\pi i(a/p)(x_1+x_2-x_3)} \right|$$

$$= \sum_{a=0}^{p-1} \left| \sum_{x\in A} e^{2\pi i(a/p)x} \right|^2 \cdot \left| \sum_{x\in 2^\wedge A} e^{2\pi i(a/p)x} \right|$$

$$= k^2|2^\wedge A| + \sum_{a=1}^{p-1} \left| \sum_{x\in A} e^{2\pi i(a/p)x} \right|^2 \cdot \left| \sum_{x\in 2^\wedge A} e^{2\pi i(a/p)x} \right|$$

$$\leq k^2|2^\wedge A| + \max_{a\not\equiv 0 \pmod p} \left| \sum_{x\in A} e^{2\pi i(a/p)x} \right| \cdot \sum_{a=1}^{p-1} \left| \sum_{x\in A} e^{2\pi i(a/p)x} \right| \cdot \left| \sum_{x\in 2^\wedge A} e^{2\pi i(a/p)x} \right|.$$

By using Cauchy's inequality and applying (15) to A and $2^\wedge A$, we see that this expression is

$$\leq k^2|2^\wedge A| + \max_{a\not\equiv 0 \pmod p} \left| \sum_{x\in A} e^{2\pi i(a/p)x} \right| \cdot \sqrt{pk-k^2} \cdot \sqrt{p|2^\wedge A| - |2^\wedge A|^2}\,.$$

Dividing by pk^2 and solving the inequality for

$$U = \max_{a\not\equiv 0 \pmod p} \left| \sum_{x\in A} e^{2\pi i(a/p)x} \right|$$

we obtain

$$\frac{U}{k} \geq \frac{1-\alpha\beta-\gamma}{\sqrt{(\alpha(1-\beta)(1-\alpha\beta))}} = f(\alpha,\beta,\gamma), \text{ say,}$$

where

$$\alpha = \frac{|2^\wedge A|}{k}\,, \quad \beta = \frac{k}{p}\,, \quad \gamma = \frac{1}{k}\,,$$

and so, by (16) and (17)

$$0 < \alpha < 2.06 - 3\gamma, \quad 0 < \beta < \frac{1}{50}\,, \quad 0 < \gamma < \frac{1}{60} < \frac{1}{50}\,.$$

By consideration of partial derivatives in the relevant range it can be checked that

$$f(\alpha,\beta,\gamma) \geq f(2.06-3\gamma,\beta,\gamma) \geq f\left(2.01\,,\ \frac{1}{50}\,,\ \frac{1}{60}\right),$$

and hence

$$U > 0.6859k\,. \tag{22}$$

By applying the lemma in Section 3.2 above to the sum

$$T = \sum_{x \in A} e^{2\pi i (a/p) x}$$

and using (22), we see that there exist u, v in F_p with $v \neq 0$ and a subset A_1 of A such that

$$A_1 \subset \left\{ u + vs : 0 \leq s \leq \tfrac{1}{2}(p-1) \right\}$$

and $|A_1| = m_1$, say, satisfies

$$m_1 = |A_1| \geq 0.8429k. \tag{23}$$

We consider the set

$$B_1 \subset \{0, 1, \ldots, \tfrac{1}{2}(p-1)\} \subset Z$$

defined by

$$B_1 = \{s : 0 \leq s \leq \tfrac{1}{2}(p-1),\ u + vs \in A_1\}. \tag{24}$$

By changing u and v if necessary we can assume that the first element of B_1 is 0 and that the greatest common divisor of the differences between successive elements of B_1 is 1.

Since the mapping ϕ given by $\phi(u + vs) = s$ gives an isomorphism of A_1 onto B_1 under addition mod p on A_1 and addition in Z on B_1, it follows that A_1 is isomorphic to B_1 as a subset of Z, so that (using (3) on B_1)

$$|B_1| = |A_1| = m_1\,, \quad |2^\wedge A_1| = |2^\wedge B_1| \geq 2m_1 - 3. \tag{25}$$

Suppose now that

$$|2^\wedge A_1| \geq 2|A_1| + C_1 - 3, \quad C_1 = \frac{|A_1| - 5}{2}\,.$$

Then from (23) it follows that

$$|2^\wedge A| \geq |2^\wedge A_1| \geq 2.5|A_1| - 5.5 \geq 2.107k - 5.5$$

and further, remembering that $k > 60$, we get

$$|2^\wedge A| \geq 2.06k - 3.$$

contradicting (17). Thus we can assume that

$$|2^\wedge A_1| < 2|A_1| + C_1 - 3,$$

and hence

$$|2^\wedge B_1| < 2|B_1| + C_1 - 3.$$

Then from Theorem 1 we get that B_1 is contained in an arithmetic progression $L \subset Z$ such that

$$|L| \leq |B_1| + 2C_1 + 2 = 2|B_1| - 3 \leq 2k - 3.$$

By our assumptions on B_1 (following (24)) it follows that

$$B_1 \subset L \subset \{0, 1, 2, \ldots, 2k - 4\}. \tag{26}$$

All elements of F_p, and in particular those of A can be written in the form

$$a = u + vs, \quad 0 \leq s \leq p - 1,$$

for u, v as above. If A contained an element a with

$$6k < s < p - 4k \tag{27}$$

then in view of (24) and (26) and the fact that $p > 10k$ the sets $2^\wedge A_1$ and $A_1 + a$ would be disjoint and so, by (25) we would have

$$|2^\wedge A| \geq |A_1 + a| + |2^\wedge A_1| = m_1 + |2^\wedge A_1| \geq 3m_1 - 3$$

and hence by (23)

$$|2^\wedge A| > 2.06k - 3,$$

a contradiction to (17). Hence (27) does not hold for elements of A and it is easily seen that all elements of A can be written in the form $a = u + vs$ with

$$-4k \leq s \leq 6k.$$

As $p > 20k$, addition mod p on s in the above range $[-4k, 6k]$ coincides with ordinary addition. Thus A is isomorphic to the set $B \subset Z$ (with addition in Z) given by

$$B = \{s : u + vs \in A, -\tfrac{1}{2}(p-1) \leq s \leq \tfrac{1}{2}(p-1)\} \subset [-4k, 6k],$$

so that by (17)

$$|2^\wedge B| = |2^\wedge A| \leq 2k - 3 + C, \quad C < 0.06k.$$

By Theorem 1 it follows that B is contained in an arithmetic progression L' with

$$|L'| \leq k + 2C + 2,$$

where $C < 0.06k$, and so A is contained in the arithmetic progression

$$L = \{u + vs : s \in L'\}$$

with $|L| = |L'|$. This completes the proof of Theorem 2.

3.4. Application to Erdős Conjecture. — We now obtain the following corollary on the Erdős conjecture.

Corollary to Theorem 2. — *Let $A \subset F_p$ such that*

$$|A| = k < \frac{p}{50}, \quad k > 60.$$

Then

$$|2^\wedge A| \geq 2k - 3.$$

Proof. — If $|2^\wedge A| \geq 2k - 2$ there is nothing to prove, so suppose that $|2^\wedge A| \leq 2k - 3$. Then by Theorem 2 (with $C = 0$) we have $A \subset L$, where L is an arithmetic progression in F_p with $|L| \leq k + 2$. Since $p > 2k + 5$ and $|A| = k$, it follows that A is isomorphic to a set B of integers (under addition in Z) such that

$$|B| = k, \quad B \subseteq \{1, 2, \ldots, k + 2\}.$$

Hence, using (3), we have

$$|2^\wedge A| = |2^\wedge B| \geq 2k - 3.$$

4. Postscript on the Erdős-Heilbronn conjecture

Rődseth [11], also using results of Freiman, has proved (4) for $p > ck$, for some positive constant c. More detailed arguments along the lines of the present paper and based on (7) can be used to obtain (4) for $p \geq 8k$, but some such restriction is essential to this approach.

Recently, two independent proofs have been given of the full Erdős-Heilbronn conjecture (4), without any restriction at all. For the first, see Dias da Silver and Hamidoune [4] and Nathanson [8]. The second, which uses only simple properties of polynomials over finite fields, is due to Alon, Nathanson and Ruzsa, [1],[2]. We are grateful to these authors for information about this work and to Professor Nathanson for the opportunity to see a preliminary version of his expository account of this topic in Nathanson [9].

References

[1] Alon N., Nathanson M. B. and Ruzsa I. Z., *Adding distinct congruence classes modulo a prime*, Amer. Math. Monthly, **102**, 1995, 250–255.

[2] Alon N., Nathanson M. B., and Ruzsa I. Z., *The polynomial method and restricted sums of congruence classes*, J. Number Theory, **56**, 1996, 404–417.

[3] Davenport H., *On the addition of residue classes*, J. London Math. Soc., **10**, 1935, 30–32.

[4] Dias da Silva J. A. and Hamidoune Y. O., *Cyclic spaces for Grassman derivatives and additive theory*, Bull. London Math. Soc., **26**, 1994, 140–146.

[5] Erdős P. and Graham R. L., *Old and new results in combinatorial number theory*, Monographie 28 de L'Enseignement Math. Gen., 1980.

[6] Freiman G. A., *Foundations of a Structural Theory of Set Addition*, Translations of Mathematical Monographs, vol.37, Amer. Math. Soc., Providence, R.I., 1973.

[7] Mansfield R., *How many slopes in a polygon*, Israel J. Math., **39**, 1981, 265–272.

[8] Nathanson M. B., *Ballot numbers, alternating products and the Erdős-Heilbronn conjecture*, in Graham, R.L., and Nestril, J., (editors), *The Mathematics of Paul Erdős*, Springer, Heidelberg, 1994.

[9] Nathanson M. B., *Additive Number Theory 2: Inverse theorems and the geometry of sumsets*, Graduate Texts in Mathematics, Springer, New York, 1996.

[10] Pyber L., *Personal communication*, 1991.

[11] Rődset O. J., *Sums of distinct residues mod p*, Acta Arith., **65**, 1993, 181–184.

G.A. Freiman, School of Mathematical Sciences, Department of Mathematics, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, 69978 Tel Aviv, Israel
E-mail : grisha@math.tau.ac.il

L. Low, Department of Pure Mathematics, University of Adelaide, Adelaide,, SA 5005, Australia
E-mail : llow@maths.adelaide.edu.au

J. Pitman, Department of Pure Mathematics, University of Adelaide, Adelaide,, S.A. 5001, Australia
E-mail : jpitman@maths.adelaide.edu.au

Astérisque

François Hennecart
Gilles Robert
Alexander Yudin

On the number of sums and differences

Astérisque, tome 258 (1999), p. 173-178
<http://www.numdam.org/item?id=AST_1999__258__173_0>

Astérisque **258**, 1999, p. 173–178

ON THE NUMBER OF SUMS AND DIFFERENCES

by

François Hennecart, Gilles Robert & Alexander Yudin

Abstract. — It is proved that $\inf_{A\subset\mathbb{Z}} \ln|A+A|/\ln|A-A|$ is less than .7865, improving a previous result due to G. Freiman and W. Pigarev.

1. Introduction

Let A be a set of integers. Write

$$\begin{array}{rcl}\mathbf{2}A = A+A & := & \{x+y \mid x,y\in A\}\\ \mathbf{D}A = A-A & := & \{x-y \mid x,y\in A\},\end{array}$$

and

$$\alpha(A) = \frac{\ln|\mathbf{2}A|}{\ln|\mathbf{D}A|}.$$

If $|A| = n$, then we have

$$\begin{array}{rcl}2n-1 & \leq & |\mathbf{2}A| \leq \dfrac{n^2+n}{2}\\ 2n-1 & \leq & |\mathbf{D}A| \leq n^2-n+1,\end{array}$$

where equality on the left side occurs for arithmetical progressions, and on the right side for "generic" sets, in which there is no nontrivial coincidence between sums and differences. Denote for any $n \geq 1$

$$\alpha_n = \inf_{A\subset N, |A|=n} \frac{\ln|\mathbf{2}A|}{\ln|\mathbf{D}A|}.$$

The lower bound $\alpha_n \geq 3/4$ follows from the inequality (see Freiman and Pigarev [1] or Ruzsa [6])

$$|\mathbf{D}A|^{3/4} \leq |\mathbf{2}A|. \tag{1}$$

1991 ***Mathematics Subject Classification.*** — 11P99, 11B75, 05B10.

Key words and phrases. — Additive number theory, combinatorial number theory, difference sets, sumsets.

In fact, Ruzsa proved a sharper result, namely

$$|A| \cdot |\mathbf{D}A| \leq |\mathbf{2}A|^2. \tag{2}$$

By squaring (2) and using $|\mathbf{D}A| \leq |A|^2$, we obtain (1).

Conversely, Freiman and Pigarev have shown that

$$\liminf_{n \to +\infty} \alpha_n < 0.89. \tag{3}$$

Here we shall improve this result by showing

Theorem. — *The sequence $(\alpha_n)_{n \geq 1}$ converges to $\alpha := \inf_A \alpha(A)$ and we have*

$$3/4 \leq \alpha \leq \ln 2/\ln(1+\sqrt{2}) < .7865.$$

Let us notice that inequality (2) implies that there is no set A such that $\alpha(A) = 3/4$. Furthermore a set A such that $\alpha(A) \leq 3/4 + \epsilon$, where $\epsilon > 0$, satisfies $|\mathbf{D}A| \geq |\mathbf{2}A|^{4/3-2\epsilon}$, and then again from (2) we deduce $|A|^{3/2} \geq |\mathbf{2}A| \geq |A|^{3/2-5\epsilon}$ and $|A|^2 \geq |\mathbf{D}A| \geq |A|^{2-10\epsilon}$.

This shows that if the value of α is 3/4, then there exists a set A with arbitrary large cardinality, which is almost a generic set to insure that $\ln |\mathbf{D}A|$ is close to $2 \ln |A|$, even when in the same time $\ln |\mathbf{2}A|$, which should be close to $1.5 \ln |A|$, does not at all correspond to a generic set A. In [5], Ruzsa was interested by such sets, and proved that there exist $c > 0$ and arbitrary large sets A such that $|\mathbf{D}A| = |A|^2(1+o(1))$ and $|\mathbf{2}A| \leq |A|^{2-c}$.

2. The convergence of α_n

In this section, we study the convergence of α_n.

In the table below, we give the value of α_n for small n: to compute them, we have looked for all the sequences $s = (s_k)_{k \in \mathbb{Z}}$ and $d = (d_k)_{k \in \mathbb{Z}}$ of nonnegative integers such that

$$\sum_{k \in \mathbb{Z}} s_k = n^2, \quad \sum_{k \in \mathbb{Z}} d_k = n^2 \quad \text{and} \quad \sum_{k \in \mathbb{Z}} s_k^2 = \sum_{k \in \mathbb{Z}} d_k^2. \tag{4}$$

For a finite set of integers A with $|A| = n$, these three conditions are satisfied by the sequences $(s_k(A))_{k \in \mathbb{Z}}$ and $(d_k(A))_{k \in \mathbb{Z}}$ where $s_k(A)$ (resp. $d_k(A)$) is the number of representations of k as a sum (resp. a difference) of two elements of A. For any pairs of solutions $s = (s_k)$ and $d = (d_k)$ of (4), we denote by N_s (resp. N_d) the number of integers k such that $s_k \neq 0$ (resp. $d_k \neq 0$).

From the inclusion

$$\{\alpha(A) \ : \ |A| = n\} \subset E_n = \{\ln N_s / \ln N_d \ : \ (s, d) \text{ solution of } (4)\},$$

we obtain α_n by exhibiting a set A of cardinality n which achieves the minimum of the finite set E_n.

For $1 \leq n \leq 7$, the infimum of $\alpha(A)$ is reached for set A for which both $|\mathbf{2}A|$ and $|\mathbf{D}A|$ are maximal. It is no more the case when $n = 8$.

n	1	2	3	4	5	6	7
α_n	1	1	.9208	.8977	.8895	.8866	.8859

Observe that α_n seems to be decreasing. In fact we are only able to prove that α_n has a limit when n tends to infinity.

The definition of $\alpha(A)$ still has a sense if $A \subset \mathbb{Z}^m$, and even if A is a subset of a lattice Λ generated by a basis $\{\omega_j\}_{1\leq j\leq m}$ in $\mathbb{R}^m$. Now we map A into $\mathbb{Z}$ by

$$x_1\omega_1 + \cdots + x_m\omega_m \longmapsto x_1 + qx_2 + \cdots + q^{m-1}x_m.$$

If q is sufficiently large, then this mapping conserves the number of sums and differences of our set. Thus we have

$$\alpha_n = \inf\left\{ \frac{\ln|\mathbf{2}A|}{\ln|\mathbf{D}A|} \ : \ A \subset \Lambda \text{ lattice in } \mathbb{R}^m \text{ for some } m \text{ and } |A| = n \right\}.$$

For any set A of integers, and any $k \geq 1$, we denote by A^k the cartesian product

$$A^k = \{(a_1, \ldots, a_k) \ : \ a_j \in A\}.$$

This set satisfies

$$\begin{aligned} |\mathbf{2}A^k| &= |\mathbf{2}A|^k \\ |\mathbf{D}A^k| &= |\mathbf{D}A|^k, \end{aligned}$$

whence

(5) $$\alpha(A^k) = \alpha(A).$$

We are now in position to prove that (α_n) converges.
Let $\varepsilon > 0$. There exist an integer n and a set A with $|A| = n$ such that

(6) $$\alpha < \alpha(A) < \alpha + \varepsilon/2.$$

Let q be an integer. We can write $n^k \leq q < n^{k+1}$ for some integer k. We define B_q as being any subset of A^{k+1} of cardinality q, and containing $A^k \times \{a\}$ for some a in A. This is possible because $|B_q| \geq |A^k| = n^k$. Then we have

$$A^k \times \{a\} - A^k \times \{a\} \subset B_q - B_q,$$

and

$$B_q + B_q \subset A^{k+1} + A^{k+1},$$

whence by (5)

$$\alpha(B_q) - \alpha(A) \leq \frac{\alpha(A)}{k}.$$

In view of (6) and using $\alpha_q \geq \alpha$, we obtain that for any sufficiently large q,

$$|\alpha_q - \alpha| < \varepsilon.$$

Thus α_q converges to α.

3. The upper bound

From the previous section, we deduce that $\alpha \leq \alpha(A)$ for any set A. To show the upper bound in the theorem, we shall construct a sequence of finite sets A_ℓ of integers such that $\lim_{\ell \to +\infty} \alpha(A_\ell) = \ln 2 / \ln(1+\sqrt{2})$.

Analyzing the proof in [1] of the upper bound (3), we see that as a set with comparatively small number of sums and large number of differences, an isomorphic (in the sense of G. Freiman [2]) image of vertices of a simplex in $\mathbb{R}^6$ was taken. It corresponds to the result of C.A. Rogers and G.G. Shephard [3], that for each convex set K of $\mathbb{R}^m$ we have

$$\operatorname{mes}(\mathbf{D}K) \leq \binom{2m}{m} \operatorname{mes}(K) \tag{7}$$

The equality in (7) is achieved only in the case when K is a simplex. Therefore, in order to get an estimate for α, it is natural to take a set of points of some lattice in a simplex.

Let $\{e_1, \dots, e_m\}$ be an orthonormal basis of $\mathbb{R}^m$ and let Λ be the lattice it generates. Let $A(m, L)$ be the subset of Λ, consisting in points $x = (x_1, \dots, x_m)$, such that $\forall j,\ x_j \geq 0$ and $\sum_{j=1}^{m} x_j \leq L$, where $L \in \mathbb{Z}^+$.
In addition, let $\operatorname{Sim}(m, L) = \operatorname{card} A(m, L)$, i.e., $\operatorname{Sim}(m, L)$ is a number of points of the lattice Λ in a rectilinear closed simplex with an edge of length L.
We shall use two lemmas.

Lemma 1. — *We have*

$$\operatorname{Sim}(m, L) = \binom{m+L}{L}.$$

This result is standard and its proof is left to the reader.

Lemma 2. — *We have*

$$|\mathbf{D}A(m, L)| = \sum_{k=0}^{\min(m,L)} \binom{m}{k}\binom{L}{k}\binom{L+m-k}{m-k} = \sum_{k=0}^{\min(m,L)} \binom{m}{k}^2 \binom{L+m-k}{m}.$$

Proof. — For any set of integers $P \subset \{1, 2 \dots, m\}$, we define the sets $\mathcal{S}_>(P, L)$

$$= \{\ (x_1, \dots, x_m) \in \mathbb{Z}^m\ :\ x_j > 0 \text{ for } j \in P,\ x_j = 0 \text{ if } j \notin P \text{ and } \sum_{j \in P} x_j \leq L\ \},$$

and $\mathcal{S}_\leq(P, L)$

$$= \{\ (x_1, \dots, x_m) \in Z^m\ :\ x_j \leq 0 \text{ for } j \in P,\ x_j = 0 \text{ if } j \notin P \text{ and } \sum_{j \in P} x_j \geq -L\ \}.$$

If we denote by $\overline{P}$ the complementary set of P in $\{1, 2, \dots, m\}$, then we have the following decomposition into disjoint sets

$$\mathbf{D}A(m, l) = \bigcup_{P \subset \{1, \dots, m\}} \mathcal{S}_>(P, L) \oplus \mathcal{S}_\leq(\overline{P}, L).$$

Since $\mathrm{card}(S_{>}(P,L)) = \mathrm{Sim}(k, L-k)$ and $\mathrm{card}(S_{\leq}(P,L)) = \mathrm{Sim}(k,L)$ if $\mathrm{card}P = k$, and counting all different subsets of $\{1,2,\dots,m\}$ with cardinal k, we obtain the lemma. □

Proof of the Theorem. — Let us write $M_{m,L} = \max_k u(m,L,k)$ where

$$u(m,L,k) = \binom{m}{k}^2 \binom{L+m-k}{m}.$$

We obviously have

$$M_{m,L} \leq |\mathbf{D}A(m,L)| \leq (m+1)M_{m,L}. \tag{8}$$

Thus using Stirling's Formula, we obtain

$$\frac{1}{2m} \ln |\mathbf{D}A(2m,m)| \sim 2\ln(1+\sqrt{2}),$$

and since $\mathbf{2}A(m,L) = A(m,2L)$,

$$\frac{1}{2m} \ln |\mathbf{2}A(2m,m)| \sim 2\ln 2,$$

when m tends to infinity. Thus

$$\alpha\Big(A(2m,m)\Big) = \frac{\ln 2}{\ln(1+\sqrt{2})}(1+o(1)),$$

when m tends to infinity. This gives the upper bound in the Theorem. □

We end by noticing that our set $A = A(2m,m)$ satisfies

$$\ln|\mathbf{2}A| / \ln|A| \sim 4\ln 2/(3\ln 3 - 2\ln 2) = 1.4519...$$

and

$$\ln|\mathbf{D}A| / \ln|A| \sim 4\ln(1+\sqrt{2})/(3\ln 3 - 2\ln 2) = 1.8462...,$$

when m tend to infinity.

References

[1] Freiman G. A. and Pigarev W. P., *The relation between the invariants R and T* (Russian), Kalinin., Gos., Univ., Moscow, 1973, 172–174.

[2] Freiman G. A., *Foundations of structural theory of set addition*, Providence, AMS, 1973.

[3] Rogers C. A. and Shephard G. C., *The difference body of a convex body*, Arch. Math. **8**, 1957, 220–233.

[4] Ruzsa I. Z., *Sets of sums and differences*, in: Sêminaire de théorie des nombres de Paris, 1982–83, Birkhauser Boston, Inc.

[5] Ruzsa I. Z., *On the number of sums and differences*, Acta Math. Hung., **59(3-4)**, 1992, 439–447.

[6] Ruzsa I. Z., *On the cardinality of $A+A$ and $A-A$*, in: Coll. math. Soc. Bolyai, **18**, Combinatorics (Keszthely, 1976), Akadêmiai Kiadó (Budapest, 1979), 933–938.

F. Hennecart, Algorithmique Arithmétique Expérimentale, Université Bordeaux 1, 351, cours de la libération, 33405 Talence, France • *E-mail :* hennec@math.u-bordeaux.fr

G. Robert, Laboratoire de Mathématiques Pures, Université Bordeaux 1, 351, cours de la libération, 33405 Talence, France • *E-mail :* robert@math.u-bordeaux.fr

A. Yudin, Department of Mathematics, Vladimir Pedagogical University, 11, pr. Stroiteley, Vladimir, Russia • *E-mail :* aayudin@vgpu.elcom.ru

Astérisque

VSEVOLOD F. LEV

The structure of multisets with a small number of subset sums

Astérisque, tome 258 (1999), p. 179-186

<http://www.numdam.org/item?id=AST_1999__258__179_0>

Astérisque **258**, 1999, p. 179–186

THE STRUCTURE OF MULTISETS WITH A SMALL NUMBER OF SUBSET SUMS

by

Vsevolod F. Lev

Abstract. — We investigate multisets of natural numbers with relatively few subset sums. Namely, let A be a multiset such that the number of distinct subset sums of A is bounded by a fixed multiple of the cardinality of A (that is, $|P(A)| \ll |A|$). We show that the set $P(A)$ of subset sums is then a union of a small number of arithmetic progressions sharing a common difference.

Similar problems were considered by G. Freiman (see [1]) and M. Chaimovich (see [2]). Unlike those papers, our conditions are stated in terms of the cardinality of the subset sums set $P(A)$ only and not on the largest element of the original multiset A.

The result obtained is nearly best possible.

1. Notation and definitions

By a *multiset* we mean a finite collection of natural numbers with repetitions allowed: $A = \{a_1, \dots, a_k\}$, where $a_1 \le \cdots \le a_k$ are the elements of A. The number of appearances of an element will be called its *multiplicity.*

As with "normal" sets, $|A| = k$ is called the *cardinality* of A. The sum of all elements of the multiset is $\sigma(A) = a_1 + \cdots + a_k$, and its *subset sums set* is

$$P(A) = \{\varepsilon_1 a_1 + \cdots + \varepsilon_k a_k \colon 0 \le \varepsilon_1, \dots, \varepsilon_k \le 1\}.$$

Notice that 0 and $\sigma(A)$ are both included in $P(A)$; generally, e belongs to $P(A)$ if and only if $\sigma(A) - e$ does.

Another useful notation:

$$A = \{a_1 \times k_1, \dots, a_s \times k_s\},$$

meaning that $a_1 < \cdots < a_s$ are *distinct* elements of A with multiplicities $k_1, \dots, k_s \ge 1$. In these terms, the cardinality of A is $|A| = k_1 + \cdots + k_s$, the sum of its elements is $\sigma(A) = k_1 a_1 + \cdots + k_s a_s$, and its subset sums set is

$$P(A) = \{\kappa_1 a_1 + \cdots + \kappa_s a_s \colon 0 \le \kappa_1 \le k_1, \dots, 0 \le \kappa_s \le k_s\}.$$

1991 ***Mathematics Subject Classification.*** — 11P99, 11B75.

Key words and phrases. — Subset sums, small doubling, multisets.

2. The main result

The following theorem is our main result.

Theorem 1. — *Let A satisfy*

$$|P(A)| \leq C|A| - 4C^3, \tag{1}$$

where C is a natural number, and suppose that the cardinality of A is sufficiently large: $|A| \geq 8C^3$. Then $P(A)$ is a union of at most $C-1$ arithmetic progressions with the same common difference.

Theorem 1 (the proof of which will be given in Section 5) is somewhat unusual in describing the structure of the subset sums set $P(A)$ rather then the structure of the multiset A itself. As the reader will notice, this reflects the essence of the problem: one can change A substantially without affecting $P(A)$, and thus it seems impossible to describe the structure of A under any reasonable condition on $P(A)$.

I conjecture that (1) can be replaced by the weaker restriction

$$|P(A)| \leq C|A| - (C-1)^2. \tag{2}$$

The following examples show that inequality (2) cannot be further relaxed.

Example 1. — *Let $A = \{1 \times (k - C + 1), b \times (C-1)\}$, where $k = |A|$ and b are sufficiently large. Then $P(A)$ is the union of C progressions*

$$\begin{gathered} 0, 1, \ldots, k - C + 1, \\ b, b+1, \ldots, b + (k - C + 1), \\ \cdot \quad \cdot \quad \cdot \\ (C-1)b, (C-1)b + 1, \ldots, (C-1)b + (k - C + 1), \end{gathered}$$

so that $|P(A)| = C(k - C + 2) = Ck - (C-1)^2 + 1$. However, $P(A)$ cannot be represented as a union of at most $C-1$ arithmetic progressions with a common difference.

Example 2. — *Let $A = \{1 \times (C-1), b \times (k - C + 1)\}$, where $k = |A|$ and b are sufficiently large. Then $P(A)$ is the union of C progressions*

$$\begin{gathered} 0, b, \ldots, (k - C + 1)b, \\ 1, 1 + b, \ldots, 1 + (k - C + 1)b, \\ \cdot \quad \cdot \quad \cdot \\ C - 1, C - 1 + b, \ldots, C - 1 + (k - C + 1)b, \end{gathered}$$

so that $|P(A)| = Ck - (C-1)^2 + 1$, and again $P(A)$ cannot be represented as a union of at most $C-1$ arithmetic progressions with a common difference.

Note that in view of Lemma 2 below, the inequality $|P(A)| \geq |A| + 1$ is always true. Hence, the conditions of Theorem 1 are never satisfied for $C = 1$, and from now on we assume $C \geq 2$.

3. Small values of C

For A satisfying (1) (or even (2)) with small values of C ($C = 2, 3$) the structure of $P(A)$, as well as the structure of A itself, can be completely described.

We begin with some basic properties of subset sums set. First, we estimate by how much $|P(A)|$ increases if one adds an element to A.

Lemma 1. — *Let $A = \{a_1 \times k_1, \ldots, a_s \times k_s\}$, $A^+ = A \cup \{a\}$, and suppose that A contains at least $i-1$ different elements less then a (that is, $a > a_{i-1}$ unless $i = 1$). Then*

$$|P(A^+)| \geq |P(A)| + i.$$

Proof. — $P(A^+)$ contains all the elements of $P(A)$, as well as the i additional elements

$$\sigma(A) + a, \sigma(A) + a - a_1, \ldots, \sigma(A) + a - a_{i-1}.$$

□

As a direct corollary, we obtain a lower-bound estimate for $|P(A)|$.

Lemma 2. — *The cardinality of the subset sums set $P(A)$ of the multiset*

$$A = \{a_1 \times k_1, \ldots, a_s \times k_s\}$$

satisfies

$$|P(A)| \geq 1 + k_1 + 2k_2 + \cdots + sk_s.$$

In particular, $|P(A)| \geq 1 + |A|$.

Proof. — The assertion is obviously true for $|A| = 1$, and we use induction on $|A|$. Denote by A^- the multiset obtained by removing from A its largest element a_s. Applying Lemma 1, we obtain then

$$\begin{aligned} |P(A)| &\geq |P(A^-)| + s \geq (1 + k_1 + 2k_2 + \cdots + s(k_s - 1)) + s \\ &= 1 + k_1 + 2k_2 + \cdots + sk_s. \end{aligned}$$

□

It follows from Lemma 2 that a multiset A with relatively small value of $|P(A)|$ has at least one element with large multiplicity.

Lemma 3. — *Let $A = \{a_1 \times k_1, \ldots, a_s \times k_s\}$, and let $k_0 = \max\limits_{1 \leq i \leq s} k_i$ be the maximal multiplicity of an element of A. Then*

$$k_0 > \frac{k^2}{2|P(A)|}.$$

Proof. — For $1 \leq i \leq s$ we have:

$$\begin{aligned} |P(A)| &\geq 1 + k_1 + 2k_2 + \cdots + ik_i + (i+1)(k_{i+1} + \cdots + k_s) \\ &> (i+1)k - (k_i + 2k_{i-1} + \cdots + ik_1) \\ &\geq (i+1)k - \frac{1}{2}i(i+1)k_0. \end{aligned}$$

The resulting estimate

$$|P(A)| > (i+1)k - \frac{1}{2}i(i+1)k_0$$

also holds for $i > s$, as in this case the expression in the right-hand side, considered as a function of real i, has a negative derivative:

$$k - \frac{1}{2}(2i+1)k_0 < k - sk_0 \leq 0.$$

Hence,

$$k_0 > \frac{2}{i}\left(k - \frac{|P(A)|}{i+1}\right)$$

for every $i = 1, 2, \dots$ We choose i under the condition

$$2\frac{|P(A)|}{k} - 1 \leq i < 2\frac{|P(A)|}{k}.$$

Then

$$\frac{2}{i} > \frac{k}{|P(A)|}, \quad \frac{|P(A)|}{i+1} \leq \frac{k}{2},$$

and so

$$k_0 > \frac{k}{|P(A)|} \cdot \frac{k}{2} = \frac{k^2}{2|P(A)|}.$$

□

We now construct multisets whose subset sums sets have a particularly simple structure.

Example 3. — *Let $A = \{a_1, \dots, a_k\}$ be a multiset such that*

i) $a_2, \dots, a_k \equiv 0 \pmod{a_1}$;

ii) $a_{i+1} \leq a_1 + \dots + a_i$ *for* $i = 1, \dots, k-1$.

Then $P(A)$ is an arithmetic progression: $P(A) = \{0, a_1, 2a_1, \dots, \sigma(A)\}$.

This easily follows by induction on k: if $A^- = \{a_1, \dots, a_{k-1}\}$, then

$$\begin{aligned} P(A) &= P(A^-) \cup (a_k + P(A^-)) \\ &= \{0, a_1, \dots, \sigma(A^-)\} \cup \{a_k, a_k + a_1, \dots, a_k + \sigma(A^-)\} \\ &= \{0, a_1, \dots, \sigma(A)\}, \end{aligned}$$

since $a_k \leq \sigma(A^-)$ and $a_k + \sigma(A^-) = \sigma(A)$.

Proposition 1. — *Any multiset A, satisfying $|P(A)| \leq 2|A| - 1$ (that is satisfying (2) with $C = 2$) has the structure, described in Example 3.*

Proof. — Suppose, on the contrary, that there exists an index $2 \leq i \leq k$ for which either $a_i \not\equiv 0 \pmod{a_1}$ or $a_i > a_1 + \dots + a_{i-1}$; we assume, moreover, that i is the *minimum* index with this property. Then, writing $A_j = \{a_1, \dots, a_j\}$ $(j = 1, \dots, k)$ and applying Lemma 2, we obtain

$$|P(A_i)| = 2|P(A_{i-1})| \geq 2i$$

(since $P(A_i) = P(A_{i-1}) \cup (a_i + P(A_{i-1}))$, and $P(A_{i-1})$ is disjoint with $a_i + P(A_{i-1})$), therefore

$$|P(A)| = |P(A_k)| \geq |P(A_{k-1})| + 2 \geq \cdots \geq |P(A_i)| + 2(k-i) \geq 2k.$$

□

The following example describes the construction of multisets whose subset sums set consists of exactly *two* arithmetic progressions.

Example 4. — *Let* $A = \{a_1, \ldots, a_m\} \cup \{b_1, b_2\} \cup \{c_1, \ldots, c_n\}$, *where*

- $a_m < b_1 \leq c_1$;
- $A_m = \{a_1, \ldots, a_m\}$ *satisfies conditions (i) and (ii) of Example 3 with* $a_1 = 2$;
- $b_1, b_2 \not\equiv 0 \pmod 2$;
- $b_1 + b_2 \leq \sigma(A_m) + 2$;
- $c_{i+1} \leq \sigma(C_i) - 2b_1 + 3 \quad (0 \leq i \leq n-1)$, *where* $C_i = \{a_1, \ldots, a_m\} \cup \{b_1, b_2\} \cup \{c_1, \ldots, c_i\}$.

Then $P(A)$ *is a union of two progressions with the common difference* 2: *if* $\sigma(A)$ *is even, then*

$$P(A) = \{0, 2, \ldots, \sigma(A) - 2, \sigma(A)\} \cup \{b_1, b_1 + 2, \ldots, \sigma(A) - b_1\},$$

and if $\sigma(A)$ *is odd, then*

$$P(A) = \{0, 2, \ldots, \sigma(A) - b_1\} \cup \{b_1, b_1 + 2, \ldots, \sigma(A)\}.$$

In either case, $|P(A)| = \sigma(A) - b_1 + 2$.

The verification is left to the interested reader.

Proposition 2. — *Any multiset* A *with co-prime elements satisfying* $|P(A)| \leq 3|A| - 4$ *(that is satisfying (2) with* $C = 3$*) has either the structure described in Example 3, or the structure described in Example 4.*

This proposition will not be used in the sequel and is given just for completeness. Its proof (which is rather long and tedious) is available from the author.

4. More lemmas and properties of $P(A)$

In this section, we prepare for the proof of Theorem 1. To this end, we first determine the value of $|P(A)|$ for multisets A with only two different elements. Without loss of generality we can restrict ourselves to the case when these two elements are co-prime.

Lemma 4. — *Let* $A = \{a_1 \times k_1, a_2 \times k_2\}$, *where* $(a_1, a_2) = 1$. *Then*

i) *if* $k_1 \leq a_2 - 1$ *or* $k_2 \leq a_1 - 1$, *then*

$$|P(A)| = (k_1 + 1)(k_2 + 1);$$

ii) *if* $k_1 \geq a_2 - 1$ *and* $k_2 \geq a_1 - 1$, *then*

$$|P(A)| = a_1 k_1 + a_2 k_2 - (a_1 - 1)(a_2 - 1) + 1.$$

Proof. — In Case i), the assertion follows from the fact that all values of the linear form

$$a_1x + a_2y; \quad 0 \le x \le k_1,\ 0 \le y \le k_2$$

are pairwise distinct: if, for instance, $k_1 \le a_2 - 1$, and $a_1x + a_2y = a_1x' + a_2y'$, then $x \equiv x' \pmod{a_2}$; therefore (in view of $0 \le x, x' \le k_1 < a_2$) we have $x = x'$, whence $y = y'$.

In Case ii) we use induction on k_2. For $k_2 = a_1 - 1$ we apply the above proved:

$$\begin{aligned} |P(A)| &= (k_1+1)(k_2+1) = a_1k_1 + a_1 \\ &= a_1k_1 + a_2k_2 - a_2k_2 + a_1 \\ &= a_1k_1 + a_2k_2 - a_1a_2 + a_1 + a_2 \\ &= a_1k_1 + a_2k_2 - (a_1-1)(a_2-1) + 1. \end{aligned}$$

Suppose now that $k_2 > a_1 - 1$. Write $A^- = \{a_1 \times k_1, a_2 \times (k_2 - 1)\}$, so that

$$|P(A^-)| = a_1k_1 + a_2(k_2-1) - (a_1-1)(a_2-1) + 1.$$

We have to prove, therefore, that $|P(A)| = |P(A^-)| + a_2$. Obviously, the difference $|P(A)| - |P(A^-)|$ counts the numbers of the form

$$xa_1 + k_2a_2; \ 0 \le x \le k_1, \tag{3}$$

which cannot be represented in the form

$$xa_1 + ya_2; \quad 0 \le x \le k_1,\ 0 \le y \le k_2 - 1.$$

We show that this particular subset of (3) is obtained when $k_1 - a_2 + 1 \le x \le k_1$; that is, there exist exactly a_2 such numbers. Indeed, if $x < k_1 - a_2 + 1$ then the number $e = xa_1 + k_2a_2$ possesses the representation $e = (x + a_2)a_1 + (k_2 - a_1)a_2$. On the other hand, for $x \ge k_1 - a_2 + 1$ the equality $xa_1 + k_2a_2 = x'a_1 + ya_2$ is impossible: otherwise $x' \equiv x \pmod{a_2}$, meaning that $x' \le x$, and then $xa_1 + k_2a_2 > x'a_1 + ya_2$, a contradiction. □

The following lemma shows that under certain conditions, a multiset can be slightly modified in such a way that the number of its elements will increase while its subset sums set will not change. Once again, we start with multisets with exactly two distinct elements.

Lemma 5. — *Let $A = \{a_1 \times k_1, a_2 \times k_2\}$, where*

$$a_1 < a_2,\ k_1 \ge a_2 - 1,\ k_2 \ge 2a_1 - 1.$$

Then there exist k_1', k_2' such that the multiset $A' = \{a_1 \times k_1', a_2 \times k_2'\}$ satisfies

$$P(A') = P(A), \quad |A'| > |A|.$$

Proof. — We set $k_1' = k_1 + a_2$, $k_2' = k_2 - a_1$. Since $k_1' + k_2' = k_1 + k_2 + (a_2 - a_1) > k_1 + k_2$, we have only to prove that $P(A') = P(A)$.

1) Suppose that $e = xa_1 + ya_2 \in P(A)$, where $0 \le x \le k_1$, $0 \le y \le k_2$, and show that $e \in P(A')$. Indeed, this is trivial if $y \le k_2 - a_1$, and otherwise it follows from $e = (x + a_2)a_1 + (y - a_1)a_2$.

2) Suppose that $e = xa_1 + ya_2 \in P(A')$, where $0 \le x \le k_1'$, $0 \le y \le k_2'$, and show that $e \in P(A)$. Indeed, this is trivial if $x \le k_1$, and otherwise this follows from $e = (x - a_2)a_1 + (y + a_1)a_2$. □

We now wish to bring the assumptions of Lemma 5 to a more convenient form, as well as to extend this lemma for the case of multisets with arbitrarily many distinct elements.

Lemma 5'. — *Let $A = \{a_1 \times k_1, \dots, a_s \times k_s\}$, and suppose that some two multiplicities k_i, k_j $(1 \le i < j \le s)$ satisfy $k_ik_j \ge 2(|P(A)| - k)$. Then there exists a multiset A' such that*

$$P(A') = P(A), \quad |A'| > |A|.$$

Proof. — Write $A_0 = \{a_i \times k_i, a_j \times k_j\}$ and $A_1 = A \setminus A_0$ (so that $A = A_0 \cup A_1$). We denote $d = (a_i, a_j)$ and set $a_i' = a_i/d$, $a_j' = a_j/d$. Clearly,

$$|P(A)| \ge |P(A_0)| + |P(A_1)| - 1 \ge |P(A_0)| + |A_1|,$$
$$|P(A_0)| \le (\tfrac{1}{2}k_ik_j + k) - (k - k_i - k_j) = \tfrac{1}{2}k_ik_j + k_i + k_j < (k_i + 1)(k_j + 1),$$

whence, in view of Lemma 4, $k_i \ge a_j'$ and $k_j \ge a_i'$. Moreover, applying Lemma 4 once more (this time part (ii)) we obtain:

$$\begin{aligned} |P(A_0)| &= k_ia_i' + k_ja_j' - (a_i' - 1)(a_j' - 1) + 1 \\ &> k_ia_i' + k_j + k_j(a_j' - 1) - (a_i' - 1)(a_j' - 1) \\ &\ge k_ia_i' + k_j, \end{aligned}$$

which implies

$$\tfrac{1}{2}k_ik_j + k_i + k_j > k_ia_i' + k_j,$$
$$k_j > 2a_i' - 2.$$

This allows us to apply Lemma 5 to A_0 (more precisely, to the multiset $\{a_i' \times k_i, a_j' \times k_j\}$) to find A_0' with $P(A_0') = P(A_0)$, $|A_0'| > |A_0|$. Then the multiset $A' = A_0' \cup A_1$ will obviously satisfy the required conditions $P(A') = P(A)$, $|A'| > |A|$. □

5. Proof of the main theorem

Two multisets A and A' will be called *equivalent*, if $P(A) = P(A')$. Without loss of generality we can assume that A is a multiset of the maximum possible cardinality of all equivalent multisets. We write A in the form

$$A = \{a_0 \times k_0\} \cup B, \; B = \{b_1 \times k_1, \dots, b_s \times k_s\},$$

where $k_1, \dots, k_s \le k_0$, and $b_1, \dots, b_s \ne a_0$.

By a *chain* we will mean a sequence $E = \{e_1, \dots, e_t\}$ of the elements of $P(B)$, satisfying the two following conditions:

i) $0 < e_{i+1} - e_i \le k_0a_0$; $\quad i = 1, \dots, t - 1$;

ii) $e_1 \equiv \dots \equiv e_t \pmod{a_0}$.

The chain E will be referred to as *maximal* if no more elements of $P(B)$ can be added to E without violating either (i) or (ii).

Let $S = P(\{a_0 \times k_0\}) = \{0, a_0, \dots, k_0 a_0\}$. It is obvious that:

- if E is a chain, then the sum $S + E$ is an arithmetic progression with the difference a_0;
- if E_1 and E_2 are two distinct maximal chains, then the progressions $S + E_1$ and $S + E_2$ are disjoint.

Clearly, there is exactly one way to decompose $P(B)$ into maximal chains, and we denote the number of these chains by N. We assume $N \geq C$ (since otherwise $P(A)$ consists of at most $C - 1$ progressions with the difference a_0) and show that this assumption leads to a contradiction.

Since obviously $|P(A)| - |P(B)| \geq N k_0$, we obtain

$$|P(B)| \leq |P(A)| - Nk_0 < C(|A| - k_0) = C|B|$$

(in fact, one can easily prove that B satisfies (2)). Therefore, by Lemma 3,

$$\max_{1 \leq i \leq s} k_i \geq \frac{|B|^2}{2|P(B)|} > \frac{|B|}{2C}.$$

By Lemma 5′ and in view of the maximality of A,

$$k_0 \cdot \tfrac{|B|}{2C} < 2(|P(A)| - k) < 2(C-1)k,$$
$$k_0(k - k_0) < 4C(C-1)k.$$

The left-hand side of the last inequality is a quadratic polynomial of k_0 with zeroes at 0 and k, maximum at $k/2$, and attaining both at $4C^2$ and $k - 4C^2$ the same common value

$$4C^2(k - 4C^2) \geq 4C(C-1)k.$$

Therefore, either $k_0 < 4C^2$ or $k_0 > k - 4C^2$ holds true.

The first is actually impossible, since by Lemma 3, $k_0 > \frac{k^2}{2Ck} \geq 4C^2$. Hence $k_0 > k - 4C^2$, and it follows that

$$|P(A)| \geq N(k_0 + 1) > C(k - 4C^2) = Ck - 4C^3,$$

a contradiction with (1). (Notice that this is the only place where we use (1) instead of the weaker (2).) This completes the proof of Theorem 1.

References

[1] Freiman G.A., *Subset-sum problem with different summands*, Congressus Numerantium, **70**, 1990, 207–215.

[2] Chaimovich M., *Solving a value-independent knapsack problem with use of methods of additive number theory*, Congressus Numerantium, **72**, 1990, 115–123.

V.F. Lev, Department of Mathematics, University of Georgia, Athens, GA 30602
E-mail : seva@math.tau.ac.il • *Url :* http://www.math.uga.edu:80/~seva/

EDITH LIPKIN

Subset sums of sets of residues

Astérisque, tome 258 (1999), p. 187-193

<http://www.numdam.org/item?id=AST_1999__258__187_0>

Astérisque **258**, 1999, p. 187–193

SUBSET SUMS OF SETS OF RESIDUES

by

Edith Lipkin

Dedicated to Grisha Freiman, with respect and affection

Abstract. — The number m is called the critical number of a finite abelian group G, if it is the minimal natural number with the property:
for every subset A of G with $|A| \geq m, 0 \notin A$, the set of subset sums A^* of A is equal to G. In this paper, we prove the conjecture of G. Diderrich about the value of the critical number of the group G, in the case $G = \mathbb{Z}_q$, for sufficiently large q.

Let G be a finite Abelian group, $A \subset G$ such that $0 \notin A$. Let $A = \{a_1, a_2, \dots, a_{|A|}\}$, where $|A| = \text{card}A$.

Let

$$A^* := \{x \,|\, x = a_1\varepsilon_1 + a_2\varepsilon_2 + \cdots + \varepsilon_{|A|}a_{|A|},\ \varepsilon_j \in \{0,1\},\ 1 \leq j \leq |A|,\ \sum_{j=1}^{|A|} \varepsilon_j > 0\}$$

and

$$X := \{m \in \mathbf{N} \,|\, \forall A \subset G, |A| \geq m \Rightarrow A^* = G\}.$$

Since $|G| - 1 \in X$, then $X \neq \varnothing$ if $|G| > 2$. The number

$$c(G) = \min_{m \in X} m$$

was introduced by George T. Diderrich in [**1**] and called the critical number of the group G.

In this note we study the magnitude of $c(G)$ in the case $G = \mathbb{Z}_q$, where $\mathbb{Z}_q$ is a group of residue classes modulo q. We set $c(q) := c(\mathbb{Z}_q)$. A survey of the problem was given by G.T. Diderrich and H.B. Mann in [**2**].

In the case when q is a prime number John Olson [**3**] proved that

$$c(q) \leq \sqrt{4q-3} + 1.$$

1991 ***Mathematics Subject Classification.*** — 11 P99, 05 D99.
Key words and phrases. — Subset sum, residue.

Recently J.A. Dias da Silva and Y.O. Hamidoune [4] have found the exact value of $c(q)$ for which an estimate

$$2q^{1/2} - 2 < c(q) < 2q^{1/2}$$

is valid.

If $q = p_1 p_2$, $p_1 \geq p_2$, p_1, p_2 – prime numbers, then

$$p_1 + p_2 - 2 \leq c(G) \leq p_1 + p_2 - 1$$

as was proved by Diderrich [1].

It was proved in [2] that for $q = 2\ell$, $\ell > 1$

$$c(G) = \ell \text{ if } \ell \geq 5 \text{ or } q = 8$$

$$c(G) = \ell + 1 \text{ in all other cases.}$$

Thus, to give thorough solution for $G = \mathbb{Z}_q$ we have to find $c(q)$ when q is a product of no less than three prime odd numbers.

G. Diderrich in [1] has formulated the following conjecture:

Let G be an Abelian group of odd order $|G| = ph$ where p is the least prime divisor of $|G|$ and h is a composite number. Then

$$c(G) = p + h - 2.$$

We prove here this conjecture for the case $G = \mathbb{Z}_q$ for sufficiently large q.

Theorem 1. — *There exists a positive integer q_0 that if $q > q_0$ and $q = ph$, $p > 2$, where p is the least prime divisor of q and h is a composite number, we have*

$$c(q) = p + h - 2.$$

To prove Theorem 1 we need the following results.

Lemma 1. — *Let $A = \{a_1, a_2, \ldots, a_{|A|}\} \subset N, N = \{1, 2, \ldots, \ell\}, S(A) = \sum_{i=1}^{|A|} a_i$,*

$A(g) = \{x \in A | x \equiv 0 (\text{mod } g)\}$, $B(A) = \frac{1}{2}\left(\sum_{i=1}^{|A|} a_i^2\right)^{1/2}$. Suppose that for some $\varepsilon > 0$ and $\ell > \ell_1(\varepsilon)$ we have $|A| \geq \ell^{2/3+\varepsilon}$ and

$$|A(g)| \leq |A| - \ell^{\frac{2}{3}+\frac{\varepsilon}{2}}, \tag{1}$$

for every $g \geq 2$. Then for every M for which

$$|M - \frac{1}{2}S(A)| \leq B(A)$$

we have $M \subset A^*$.

Lemma 2. — *Let ε be a constant, $0 < \varepsilon \leq 1/3$. There exists $\ell_0 = \ell_0(\varepsilon)$ such that for every $\ell \geq \ell_0$ and every set of integers $A \subset [1, \ell]$, for which*

$$|A| \geq \ell^{\frac{2}{3}+\varepsilon}, \tag{2}$$

the set A^* contains an arithmetic progression of ℓ elements and difference d satisfying the condition

$$d < \frac{2\ell}{|A|}. \tag{3}$$

We cited as Lemma 1 the Proposition 1.3 on page 298 of [**5**].

Proof of Lemma 2. — Let us first assume that A fulfills the condition (1) in Lemma 1. Since we have

$$B(A) \geq \frac{1}{2}\sqrt{\sum_{i=1}^{|A|} i^2} > \frac{1}{2}\sqrt{\frac{|A|^3}{3}} > \frac{1}{2\sqrt{3}}\ell^{1+\frac{3}{2}\varepsilon}$$

and every M from the interval $(\frac{1}{2}S(A) - B(A), \frac{1}{2}S(A) + B(A))$ belong to A^*, there exists an arithmetic progression in A^* of the length $2B(A) > \ell$, if $\ell > \ell_0 = \ell_1(\varepsilon)$.

Now we study the case when A does not satisfy (1). We can then find an integer $g_1 \geq 2$ such that $B_1 \subset A = A_0$ and B_1 contains those elements of A_0 which are divisible by g_1 and for the set $A_1 = \{x/g_1 | x \in B_1 \text{ and } x \equiv 0 (\text{mod } g_1)\}$ we have

$$|A_1| > |A_0| - \ell^{\frac{2}{3}+\frac{\varepsilon}{2}} .$$

Suppose that this process was repeated s times and numbers $g_1, g_2, \ldots, g_s$ were found and sets $A_1, A_2, \ldots, A_s$ defined inductively, B_j being a subset of A_{j-1} containing those elements of A_{j-1} which are divisible by g_j and

$$A_j = \{x/g_j | x \in B_j \text{ and } x \equiv 0 (\text{mod } g_j)\}$$

so that we have

$$|A_j| > |A_{j-1}| - \ell^{\frac{2}{3}+\frac{\varepsilon}{2}}, \quad j = 1, 2, \ldots, s .$$

From

$$|A_s| \geq |A_{s-1}| - \ell^{\frac{2}{3}+\frac{\varepsilon}{2}} > |A| - s\ell^{\frac{2}{3}+\frac{\varepsilon}{2}}$$

and

$$\ell_s = \left[\frac{\ell_{s-1}}{q_s}\right] \leq \frac{\ell}{2^s}$$

it follows that

$$|A_s| \geq \frac{1}{2}|A| \geq \frac{1}{2}\ell^{\frac{2}{3}+\frac{\varepsilon}{2}} > \ell_s^{\frac{2}{3}+\varepsilon}. \tag{4}$$

The condition (2) of Lemma 2 for A_s is verified, for some sufficiently large s the condition (3) is fulfilled and thus A_s^* contains an interval

$$\left(\frac{1}{2}S(A_s) - B(A_s), \frac{1}{2}S(A_s) + B(A_s)\right).$$

We have, in view of (4),

$$\begin{aligned} B(A_s) &\geq \frac{1}{2}\sqrt{\sum_{i=1}^{|A_s|} i^2} > \frac{1}{2}\sqrt{\frac{|A_s|^3}{3}} \\ &\geq \frac{1}{4\sqrt{6}}\ell^{1+\frac{3}{2}\varepsilon} > \ell. \end{aligned} \tag{5}$$

We have shown that A_s^* contains an arithmetic progression of length ℓ and difference $d = g_1 g_2 \cdots g_s$, and thus A^* has the same property.

We now prove (2). From

$$\ell_s = \left[\frac{\ell}{d}\right], \quad \ell_s \geq |A_s| \geq \frac{1}{2}|A|$$

we have

$$\left[\frac{\ell}{d}\right] \geq \frac{1}{2}|A|$$

or

$$d \leq \frac{2\ell}{|A|}.$$

Lemma 2 is proved.

***Lemma 3* (M. Chaimovich [6]).** — *Let $B = \{b_i\}$ be a multiset, $B \subset \mathbb{Z}_q$. Suppose that for every $s \geq 2$, s dividing q, we have*

$$|B \backslash B(s)| \geq s - 1. \tag{6}$$

There exists $F \subset B$ for which

$$|F| \leq q - 1,$$
$$F^* = \mathbb{Z}_q.$$

Proof of Theorem 1. — Let $q = p_1 p_2 \cdots p_k$, $k \geq 4$, $p = p_1 \leq p_2 \leq \cdots \leq p_k$. We have

$$p^k \leq q \Rightarrow p \leq q^{1/4}\ . \tag{7}$$

Let $A \subset \mathbb{Z}_q$ be such that $0 \notin A$ and

$$|A| \geq \frac{q}{p} + p - 2; \tag{8}$$

we have to prove that $A^* = \mathbb{Z}_p$.

From (7) and (8) we get

$$|A| > \frac{q}{p} \geq q^{3/4}\ . \tag{9}$$

Let us consider some divisor d of q, and denote by A_d a multiset A viewed as a multiset of residues mod d. Let us show that for every δ dividing d the number of residues in A_d which are not divisible by δ satisfies the condition of Lemma 3.

The number of residues in $\mathbb{Z}_q$ which are divisible by δ is equal to q/δ. Therefore the number of such residues in A (which are all different) is not larger than $q/\delta - 1$, because $0 \notin A$.

From this reasoning and from (7) we get the estimate

$$\begin{aligned} &|A_d \backslash A(\delta)| \geq |A| - \left(\frac{q}{\delta} - 1\right) \geq \\ &\frac{q}{p} + p - 2 - \frac{q}{\delta} + 1 = \frac{q}{p} + p - \left(\frac{q}{\delta} + \delta\right) + \delta - 1\ . \end{aligned} \tag{10}$$

The function $x + q/x$ is decreasing on the segment $[1, \sqrt{q}]$.

The least divisor of q is equal to p, and the maximal one to q/p. Therefore

$$p \leq \delta \leq \frac{q}{p} .$$

If $p \leq \delta \leq \sqrt{q}$, we have

$$\frac{q}{p} + p \geq \frac{q}{\delta} + \delta . \tag{11}$$

In the case $\sqrt{q} \leq \delta \leq \frac{q}{p}$, let $\rho = \frac{q}{\delta}$. Then $\delta = \frac{a}{\rho}$, $\sqrt{q} \leq \frac{q}{\rho} \leq \frac{q}{p}$ and $p \leq \rho \leq \sqrt{q}$ and we have

$$\frac{q}{p} + p \geq \frac{q}{\rho} + \rho = \delta + \frac{q}{\delta} . \tag{12}$$

From (11) and (12) it follows from (10) that we have

$$|A_d \backslash A(\delta)| \geq \delta - 1 . \tag{13}$$

Let us apply the Lemma 3 to A_d. Condition (13) is condition (6) of Lemma 3. Therefore there exists $F_d \subset A_d$ such that $|F_d| \leq d - 1$ and $F_d^* = \mathbb{Z}_d$.

Viewing F_d as a set of residues mod q, let

$$A' = \bigcup_{\substack{d/q \\ p \leq d < q^{1/3}}} F_d .$$

It is well known that the number of divisors $d(q) = O(q^\varepsilon)$ for every $\varepsilon > 0$ so that

$$|A'| < q^{\frac{1}{3}+\varepsilon}$$

for sufficiently large q.

Take now $A'' = A \backslash A'$. Take the least positive integer from each class of residues of the set A'' and denote this set by $\widehat{A''}$. We have $\widehat{A''} \subset [1, q-1]$. We set $\ell = q$ and see that all conditions of Lemma 1 are valid for $\widehat{A''}$. Thus, $(\widehat{A''})^*$ contains an arithmetic progression $\mathcal{L}$ with a length q and a difference Δ such that

$$\Delta < \frac{2q}{q^{\frac{3}{4}}} = 2q^{1/4} . \tag{14}$$

If $(\Delta, q) = 1$ then $(A'')^* = \mathbb{Z}_q$. Suppose that $D = (\Delta, q) > 1$. Then $\mathcal{L}$ (and therefore $(\widehat{A''})^*$ which contains $\mathcal{L}$) contains the residues of $\mathbb{Z}_q$ which are divisible by D. If $\mathbb{Z}_D$ is a system of residues mod q representing a system of all residues mod D/q, then $(\widehat{A''})^* + \mathbb{Z}_D = \mathbb{Z}_q$. But $F_D \subset A'$ and $F_D^* = \mathbb{Z}_D$. Thus

$$A^* \supset (\widehat{A''})^* + (A')^* = \mathbb{Z}_q .$$

Theorem 1 is proved in the case $k \geq 4$.

Now we have to study the case when q is a product of three primes. Let $q = p_1 p_2 p_3$, $p = p_1 \leq p_2 \leq p_3$. Suppose that for some positive ε we have $p < p^{\frac{1}{3+\varepsilon}}$. The proof may be completed in a similar way to what was done.

In the general case we can use a stronger result than Lemma 2. Namely, the formulation of Lemma 2 is valid if in (2) we replace the number 2/3 in the exponent by 1/2 (see G. Freiman [7] and A. Sárkőzy [8]). So, in the case of q being a product of three primes, we can use this stronger version and prove Theorem 1.

As we have seen, the version of Lemma 1 with the exponent 2/3 was sufficient in the majority of cases. It is preferable to use this version, for its proof is much simpler than the case 1/2. Secondly, in the case 2/3 estimates of error terms have been obtained explicitly by M. Chaimovich. It provides us with the possibility to get an explicit range of validity for Theorem 1.

Lemma 4. — *Define a function of ℓ in the following manner:*

$$m_0(\ell) = \left(\frac{12}{\pi^2}\right)^{1/3} \ell^{2/3}(\log \ell + 1/6)^{1/3}\left(2 - \frac{4\gamma}{3}\right)^{1/3} \tag{15}$$

where $\gamma = \left(\frac{12}{\pi^2}\ \frac{\log \ell + 1/6}{\ell}\right)^{1/3}$.

Then for $\ell > 155$ a subset sum of each subset $A \subset \{1, 2, \dots, \ell\}$ with $|A| = m > m_0(\ell)$ contains an arithmetic progression of cardinality ℓ.

Simplifying (15) we can take

$$m_0(\ell) = 1.3\,\ell^{2/3}(\log \ell + 1/6)^{1/3}.$$

In the case of four or more primes in a representation of q we have to verify an inequality

$$\ell^{3/4} > 1.3\,\ell^{2/3}(\log \ell + 1/6)^{1/3} \tag{16}$$

which is fulfilled for

$$\ell \geq 3000.$$

In some special cases we can give better estimates. For example, if $p = 3$ we have $m > q/3$ and instead of (16) we have

$$\ell/3 > 1.3\,\ell^{2/3}(\log \ell + 1/6)^{1/3},$$

$$\ell > 64(\log \ell + 1/6)$$

which is valid for

$$\ell \geq 500.$$

References

[1] Diderrich G. T., *An addition theorem for Abelian groups of order pq*, Journal of Number Theory, **7**, 1975, 33–48.

[2] Diderrich G. T., Mann H. B., *Combinatorial problems in finite Abelian groups*, J.N.Srivastava et al., eds., A Survey of Combinatorial Theory. North-Holland Publishing Company, 1973, 95–100.

[3] Olson J. E., *An addition theorem modulo p*, Journal of Combinatorial Theory, **5**, 1968, 45–52.

[4] Dias da Silva J.A., Hamidoune Y.O., *Cyclic spaces for Grassman derivatives and additive theory*, Bull London. Math. Soc., **26**, 1994, 140–146.

[5] Alon N., Freiman G. A., *On sums of subsets of a set of integers*, Combinatorica, **8**(4), 1988, 297–306.

[6] Chaimovich M., *Solving a value-independent knapsack problem with the use of methods of additive number theory*, Congressus Numerantium, **72**, 1990, 115–123.

[7] Freiman G.A., *New analytical results in subset-sum problem*, Discrete Mathematics, **114**, 1993, 205–218.

[8] Sárkőzy A., *Finite addition theorems*, II. J. Number Theory, **48(2)**, 1994, 197–218.

E. LIPKIN, School of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel-Aviv, Israel

Astérisque

MELVYN B. NATHANSON
GÉRALD TENENBAUM

Inverse theorems and the number of sums and products

Astérisque, tome 258 (1999), p. 195-204
<http://www.numdam.org/item?id=AST_1999__258__195_0>

Astérisque **258**, 1999, p. 195–204

INVERSE THEOREMS AND THE NUMBER OF SUMS AND PRODUCTS

by

Melvyn B. Nathanson & Gérald Tenenbaum

Abstract. — Let $\epsilon > 0$. Erdős and Szemerédi conjectured that if A is a set of k positive integers which large k, there must be at least $k^{2-\epsilon}$ integers that can be written as the sum or product of two elements of A. We shall prove this conjecture in the special case that the number of sums is very small.

1. A conjecture of Erdős and Szemerédi

Let A be a nonempty, finite set of positive integers, and let $|A|$ denote the cardinality of the set A. Let

$$2A = \{a + a' : a, a' \in A\}$$

denote the 2-fold *sumset* of A, and let

$$A^2 = \{aa' : a, a' \in A\}$$

denote the 2-fold *product set* of A. We let

$$E_2(A) = 2A \cup A^2$$

denote the set of all integers that can be written as the sum or product of two elements of A. If $|A| = k$, then

$$|2A| \leqslant \binom{k+1}{2}$$

and

$$|A^2| \leqslant \binom{k+1}{2},$$

1991 ***Mathematics Subject Classification.*** — Primary 11B05, 11B13, 11B75, 11P99, 05A17.

Key words and phrases. — Additive number theory, sumsets, product sets, inverse theorems, Freiman's theorem, sums and products of integers, divisors.

M.B. N.: This work was supported in part by grants from the PSC-CUNY Research Award Program and the National Security Agency Mathematical Sciences Program.

and so the number of sums and products of two elements of A is

$$|E_2(A)| \leqslant k^2 + k.$$

Erdős and Szemerédi [**3**, p. 60] made the beautiful conjecture that a finite set of positive integers cannot have simultaneously few sums and few products. More precisely, they conjectured that for every $\varepsilon > 0$ there exists an integer $k_0(\varepsilon)$ such that, if A is a finite set of positive integers and

$$|A| = k \geqslant k_0(\varepsilon),$$

then

$$|E_2(A)| \gg_\varepsilon k^{2-\varepsilon}.$$

Very little is known about this question. Erdős and Szemerédi [**4**] have shown that there exists a real number $\delta > 0$ such that

$$|E_2(A)| \gg k^{1+\delta},$$

and Nathanson [**11**] proved that

$$|E_2(A)| \geqslant ck^{32/31},$$

where $c = 0.00028\ldots$.

Erdős and Szemerédi [**4**] also remarked that, in the special case that $|2A| \leqslant ck$, "perhaps there are more than $k^2/(\log k)^\varepsilon$ elements in A^2 ". This cannot be true for arbitrary finite sets of positive integers and arbitrarily small $\varepsilon > 0$. For example, if A is the set of all integers from 1 to k, then Tenenbaum [**16, 17**], improving a result of Erdős [**2**], proved that

$$\frac{k^2}{(\log k)^{\varepsilon_0}} \mathrm{e}^{-c\sqrt{\log_2 k \log_3 k}} \ll |A^2| \ll \frac{k^2}{(\log k)^{\varepsilon_0}\sqrt{\log_2 k}}, \tag{1}$$

where $\log_r$ denotes the r-fold iterated logarithm, and

$$\varepsilon_0 = 1 - \left(\frac{1+\log_2 2}{\log 2}\right) \geqslant 0.08607 \tag{2}$$

(cf. Hall and Tenenbaum [**8**, Theorem 23]).

Using an inverse theorem of Freiman, we shall prove that if A is a set of k positive integers such that $|2A| \leqslant 3k - 4$, then

$$|A^2| \gg (k/\log k)^2.$$

We obtain a similar result for the sumset and product set of two possibly different sets of integers. Let A_1 and A_2 be nonempty, finite sets of positive integers, and let

$$A_1 + A_2 = \{a_1 + a_2 : a_1 \in A_1, a_2 \in A_2\}$$

and

$$A_1A_2 = \{a_1a_2 : a_1 \in A_1, a_2 \in A_2\}.$$

Let $|A_1| = |A_2| = k$. We prove that whenever $|A_1 + A_2| \leqslant 3k - 4$, then we have $|A_1A_2| \gg (k/\log k)^2$.

2. Product sets of arithmetic progressions

A set Q of positive integers is an *arithmetic progression* of length ℓ and difference q if there exist positive integers r, q, and ℓ such that

$$Q = \{r + uq : 0 \leqslant u < \ell\}.$$

We shall always assume that

$$\ell \geqslant 2.$$

For any sets A and B of positive integers, let $\varrho_{A,B}(m)$ denote the number of representations of m in the form $m = ab$, where $a \in A$ and $b \in B$. Let $\varrho_A(m) = \varrho_{A,A}(m)$. Let $\tau(m)$ denote the number of positive divisors of m. Clearly, for every integer m,

$$\varrho_{A,B}(m) \leqslant \tau(m).$$

If $A_1 \subseteq Q_1$ and $A_2 \subseteq Q_2$, then $\varrho_{A_1,A_2}(m) \leqslant \varrho_{Q_1,Q_2}(m)$.

***Lemma 1* (Shiu).** — *Let $0 < \alpha < 1/2$ and let $0 < \beta < 1/2$. Let x and y be real numbers and let s and q be integers such that*

$$0 < s \leqslant q \text{ and } (s,q) = 1, \tag{3}$$

$$q < y^{1-\alpha}, \tag{4}$$

and

$$x^{\beta} < y \leqslant x. \tag{5}$$

Then

$$\sum_{\substack{w \equiv s \, (\mathrm{mod}\, q) \\ x-y<w\leqslant x}} \tau(w) \ll_{\alpha,\beta} \frac{\varphi(q) y \log x}{q^2}.$$

Proof. This is a special case of Theorem 2 in Shiu **[14]** (see also Vinogradov and Linnik **[18]** and Barban and Vehov **[1]**).

Lemma 2. — *Let s, q, h, and ℓ be integers such that $h \geqslant 0$, $\ell \geqslant 2$, $0 < s \leqslant q$, and $(s,q) = 1$. Let Q be the arithmetic progression*

$$Q = \{s + vq : h \leqslant v < h + \ell\}.$$

If $(h+1)q < \ell^5$, then

$$\sum_{w \in Q} \tau(w) \ll \ell \log \ell.$$

Proof. We apply Lemma 1 with $\alpha = \beta = 1/6$, $x = (h+\ell)q$, and $y = \ell q$. The integers s and q satisfy (3). Since $q \leqslant (h+1)q < \ell^5$, we have $q^{1/6} < \ell^{5/6}$, and so

$$q = q^{1/6} q^{5/6} < (\ell q)^{5/6} = y^{1-\alpha}.$$

This shows that (4) is satisfied.

To obtain (5), we consider two cases. If $h \leqslant \ell$, then, since $2 \leqslant \ell \leqslant \ell q$, we have

$$x^{\beta} = ((h+\ell)q)^{\beta} \leqslant (2\ell q)^{\beta} \leqslant (\ell q)^{2\beta} = (\ell q)^{1/3} < \ell q = y \leqslant x.$$

If $h > \ell$, then, since $hq < \ell^5$, we have

$$x^{\beta} = \{(h+\ell)q\}^{\beta} < (\ell h q)^{\beta} < \ell^{6\beta} = \ell \leqslant \ell q = y \leqslant x.$$

This shows that (5) holds.

Applying Lemma 1, we obtain

$$\begin{aligned}\sum_{w\in Q} \tau(w) &= \sum_{\substack{w\equiv s \,(\mathrm{mod}\, q)\\ hq<w\leqslant(h+\ell)q}} \tau(w) \ll \frac{\varphi(q)(\ell q)\log((h+\ell)q)}{q^2} \\ &\ll \ell\log(\ell(h+1)q) \ll \ell\log\ell^6 \ll \ell\log\ell.\end{aligned}$$

This completes the proof.

Lemma 3. — *Let Q_1 and Q_2 be two arithmetic progressions of length $\ell \geqslant 2$, and let $m \in Q_1Q_2$. Then*

$$\varrho_{Q_1,Q_2}(m) \ll_{\varepsilon} \ell^{\varepsilon} \tag{6}$$

for every $\varepsilon > 0$, and

$$\sum_{m\in Q_1Q_2} \varrho_{Q_1,Q_2}(m)^2 \ll (\ell\log\ell)^2. \tag{7}$$

Proof. Let $Q_i = \{r_i + uq_i : 0 \leqslant u < \ell\}$ for $i = 1, 2$. We may assume without loss of generality that $(r_i, q_i) = 1$. We write $r_i = s_i + h_iq_i$, where $0 < s_i \leqslant q_i$ and $h_i \geqslant 0$. Then

$$Q_i = \{s_i + vq_i : h_i \leqslant v < h_i + \ell\}.$$

If $w_1 \in Q_1$ and $w_2 \in Q_2$, then, for suitable $v_1 \in [h_1, h_1+\ell[$, $v_2 \in [h_2, h_2+\ell[$, we have

$$h_1q_1 < w_1 = s_1 + v_1q_1 \leqslant (h_1+\ell)q_1 \leqslant \ell(h_1+1)q_1 \tag{8}$$

and

$$h_2q_2 < w_2 = s_2 + v_2q_2 \leqslant (h_2+\ell)q_2 \leqslant \ell(h_2+1)q_2. \tag{9}$$

We can assume that

$$(h_2+1)q_2 \leqslant (h_1+1)q_1.$$

There are two cases. In the first case,

$$(h_1+1)q_1 < \ell^5.$$

By (8) and (9), we deduce that

$$w_1 \leqslant \ell(h_1+1)q_1 < \ell^6, \quad \text{and} \quad w_2 \leqslant \ell(h_2+1)q_2 \leqslant \ell(h_1+1)q_1 < \ell^6.$$

If $m \in Q_1Q_2$, then m is of the form $m = w_1w_2$, and so $m < \ell^{12}$. Since, by a classical estimate, $\tau(m) \ll_{\varepsilon} m^{\varepsilon/12}$, it follows that

$$\varrho_{Q_1,Q_2}(m) \leqslant \tau(m) \ll_{\varepsilon} m^{\varepsilon/12} \ll_{\varepsilon} \ell^{\varepsilon}.$$

This proves (6).

To prove (7), we use the submultiplicativity of the divisor function, that is, $\tau(uv) \leqslant \tau(u)\tau(v)$ for all positive integers u, v. Then

$$\begin{aligned}\sum_{m\in Q_1Q_2} \varrho_{Q_1,Q_2}(m)^2 &= \sum_{w_1\in Q_1}\sum_{w_2\in Q_2} \varrho_{Q_1,Q_2}(w_1w_2)\\ &\leqslant \sum_{w_1\in Q_1}\sum_{w_2\in Q_2} \tau(w_1w_2)\\ &\leqslant \sum_{w_1\in Q_1} \tau(w_1) \sum_{w_2\in Q_2} \tau(w_2) \ll \ell^2(\log \ell)^2,\end{aligned}$$

where the last upper bound follows from Lemma 2.

Consider now the second case

$$(h_1+1)q_1 \geqslant \ell^5.$$

We shall prove that

$$\varrho_{Q_1,Q_2}(m) \leqslant 3 \tag{10}$$

for all $m \geqslant 1$. Suppose that $w_1 = r_1 + uq_1 \in Q_1$ and $w_1' = r_1 + u'q_1 \in Q_1$ are distinct divisors of m, and that $w_1 < w_1'$. Then $(r_1, q_1) = 1$ implies that $(w_1, q_1) = (w_1', q_1) = 1$, and so $((w_1, w_1'), q_1) = 1$. Since (w_1, w_1') divides

$$w_1' - w_1 = (u' - u)q_1,$$

it follows that (w_1, w_1') divides $u' - u$, and so

$$1 \leqslant (w_1, w_1') \leqslant u' - u < \ell.$$

Suppose that $\varrho_{Q_1,Q_2}(m) \geqslant 4$. Then m has at least four distinct representations in the form $m = w_1w_2$ with $w_1 \in Q_1$ and $w_2 \in Q_2$, and so m has at least four different divisors in Q_1, that is, at least four divisors of the form

$$r_1 + uq_1 = s_1 + (h_1 + u)q_1$$

with $0 \leqslant u < \ell$. At most one of these divisors is $s_1 + h_1q_1$, and so m has at least three different divisors, which we shall denote by w_1, w_1', and w_1'', such that

$$\min\{w_1, w_1', w_1''\} \geqslant s_1 + (h_1+1)q_1 > (h_1+1)q_1 \geqslant \ell^5.$$

Let $[w_1, w_1', w_1'']$ denote the least common multiple of w_1, w_1', and w_1''. Since each of these three numbers is a divisor of m, we have

$$\begin{aligned} m &\geqslant [w_1, w_1', w_1''] \geqslant \frac{w_1w_1'w_1''}{(w_1,w_1')(w_1,w_1'')(w_1',w_1'')}\\ &> \left(\frac{(h_1+1)q_1}{\ell}\right)^3 = \frac{(h_1+1)q_1}{\ell^3}(h_1+1)^2q_1^2\\ &\geqslant \ell^2\Big((h_1+1)q_1\Big)^2 \geqslant \ell(h_1+1)q_1 \cdot \ell(h_2+1)q_2 \geqslant w_1w_2 = m,\end{aligned}$$

which is impossible. This proves (10), and inequalities (6) and (7) follow immediately.

Lemma 4. — *Let Q be an arithmetic progression of length $\ell \geqslant 2$, and let $m \in Q^2$. Then*

$$\varrho_Q(m) \ll_\varepsilon \ell^\varepsilon \tag{11}$$

for every $\varepsilon > 0$, and

$$\sum_{m \in Q^2} \varrho_Q(m)^2 \ll (\ell \log \ell)^2. \tag{12}$$

Proof. This follows immediately from Lemma 3 with $Q_1 = Q_2 = Q$.

Lemma 5. — *Let Q_1 and Q_2 be arithmetic progressions of length $\ell \geqslant 2$. Then*

$$|Q_1Q_2| \gg \Big(\frac{\ell}{\log \ell}\Big)^2.$$

Proof. Let $\varrho_{Q_1,Q_2}(m)$ denote the number of representations of m in the form $m = q_1q_2$, where $q_1 \in Q_1$ and $q_2 \in Q_2$. By the Cauchy-Schwarz inequality and inequality (7) of Lemma 3,

$$\begin{aligned} \ell^2 &= \sum_{m \in Q_1Q_2} \varrho_{Q_1,Q_2}(m) \leqslant |Q_1Q_2|^{1/2}\Big(\sum_{m \in Q_1Q_2} \varrho_{Q_1,Q_2}(m)^2\Big)^{1/2} \\ &\ll |Q_1Q_2|^{1/2}\ell \log \ell. \end{aligned}$$

Therefore,

$$|Q_1Q_2| \gg \Big(\frac{\ell}{\log \ell}\Big)^2.$$

This completes the proof.

Lemma 6. — *Let Q be an arithmetic progression of length $\ell \geqslant 2$. Then*

$$|Q^2| \gg \Big(\frac{\ell}{\log \ell}\Big)^2.$$

Proof. This follows immediately from Lemma 5 with $Q_1 = Q_2 = Q$.

3. Application of some inverse theorems

We shall use the following two inverse theorems of Freiman.

***Lemma 7* (Freiman).** — *Let A be a nonempty set of k positive integers. If*

$$|2A| \leqslant 3k - 4,$$

then A is a subset of an arithmetic progression of length $\ell < 2k$.

Proof. See [**5, 7, 10, 12**].

***Lemma 8* (Freiman).** — *Let A_1 and A_2 be nonempty finite sets of positive integers, and let $|A_i| = k_i$ for $i = 1, 2$. If*

$$|A_1 + A_2| \leqslant k_1 + k_2 + \min\{k_1, k_2\} - 4,$$

then A_1 and A_2 are subsets of arithmetic progressions Q_1 and Q_2, respectively, where Q_1 and Q_2 have the same difference and the same length $\ell < k_1 + k_2$.

Proof. See [**6, 9, 12, 15**].

Theorem 1. — *Let A be a finite set of positive integers, and let $|A| = k \geqslant 2$. If*

$$|2A| \leqslant 3k - 4,$$

then

$$|A^2| \gg \left(\frac{k}{\log k}\right)^2.$$

Proof. By Lemma 7, if $|2A| \leqslant 3k - 4$, then there exists an arithmetic progression Q of length $\ell < 2k$ such that $A \subseteq Q$. Since

$$\varrho_A(m) \leqslant \varrho_Q(m),$$

it follows from (12) that

$$\begin{aligned} k^2 &= \sum_{m \in A^2} \varrho_A(m) \leqslant |A^2|^{1/2} \Big(\sum_{m \in A^2} \varrho_A(m)^2 \Big)^{1/2} \\ &\leqslant |A^2|^{1/2} \Big(\sum_{m \in Q^2} \varrho_Q(m)^2 \Big)^{1/2} \\ &\ll |A^2|^{1/2} \ell \log \ell \ll |A^2|^{1/2} k \log k. \end{aligned}$$

Therefore,

$$|A^2| \gg \left(\frac{k}{\log k}\right)^2. \tag{13}$$

This completes the proof.

Theorem 2. — *Let $\lambda \geqslant 1$. Let A_1 and A_2 be finite sets of positive integers such that $|A_i| = k_i \geqslant 2$ for $i = 1, 2$ and*

$$\frac{1}{\lambda} \leqslant \frac{k_2}{k_1} \leqslant \lambda. \tag{14}$$

If

$$|A_1 + A_2| \leqslant k_1 + k_2 + \min\{k_1, k_2\} - 4,$$

then

$$|A_1 A_2| \gg_\lambda \frac{k_1 k_2}{\Big(\log(k_1 k_2)\Big)^2}.$$

Proof. It follows from (14) that

$$(k_1+k_2)^2 \leqslant (1+\lambda)^2 k_1^2 = (1+\lambda)^2\lambda k_1(k_1/\lambda) \leqslant (1+\lambda)^2\lambda k_1 k_2,$$

and so

$$k_1+k_2 \ll_\lambda (k_1k_2)^{1/2}.$$

By Lemma 8, if $|A_1+A_2| \leqslant k_1+k_2+\min\{k_1,k_2\}-4$, there exist arithmetic progressions Q_1 and Q_2, each of length $\ell < k_1+k_2$, such that $A_1 \subseteq Q_1$ and $A_2 \subseteq Q_2$. Since

$$\varrho_{A_1,A_2}(m) \leqslant \varrho_{Q_1,Q_2}(m),$$

it follows from (7) that

$$\begin{aligned} k_1k_2 &= \sum_{m\in A_1A_2} \varrho_{A_1,A_2}(m) \\ &\leqslant |A_1A_2|^{1/2}\Big(\sum_{m\in A_1A_2} \varrho_{A_1,A_2}(m)^2\Big)^{1/2} \\ &\leqslant |A_1A_2|^{1/2}\Big(\sum_{m\in Q_1Q_2} \varrho_{Q_1,Q_2}(m)^2\Big)^{1/2} \\ &\ll |A_1A_2|^{1/2}\ell\log\ell \ll |A_1A_2|^{1/2}(k_1+k_2)\log(k_1+k_2) \\ &\ll_\lambda |A_1A_2|^{1/2}(k_1k_2)^{1/2}\log(k_1k_2). \end{aligned}$$

Therefore,

$$|A_1A_2| \gg_\lambda \frac{k_1k_2}{\Big(\log(k_1k_2)\Big)^2}. \tag{15}$$

This completes the proof.

Theorem 3. — *Let A_1 and A_2 be finite sets of positive integers such that $|A_1| = |A_2| = k \geqslant 2$. If*

$$|A_1+A_2| \leqslant 3k-4,$$

then

$$|A_1A_2| \gg \Big(\frac{k}{\log k}\Big)^2.$$

Proof. This follows immediately from Theorem 2 with $k_1 = k_2 = k$ and $\lambda = 1$.

4. Open problems

By Theorem 1, if $|A| = k$ and $|2A| \leqslant 3k-4$, then $|A^2| \gg k^{2-\varepsilon}$. This gives the first general case in which we know that the conjecture of Erdős and Szemerédi is true. It would be nice to prove that if $c \geqslant 3$ and if A is a finite set of k positive integers such that

$$|2A| \leqslant ck, \tag{16}$$

then

$$|A^2| \gg_{c,\varepsilon} k^{2-\varepsilon}.$$

By a general inverse theorem of Freiman [**7, 12, 13**], a finite set of integers whose sumset satisfies inequality (16) is a "large" subset of what is called an n-dimensional arithmetic progression. This is a set Q with the following structure: For $n \geqslant 1$, there exist positive integers $r, q_1, \ldots, q_n, \ell_1, \ldots, \ell_n$ such that

$$Q = \{r + u_1 q_1 + \cdots + u_n q_n : 0 \leqslant u_i < \ell_i \text{ for } i = 1, \ldots, n\}. \tag{17}$$

The *length* of Q is defined as $\ell(Q) = \ell_1 \cdots \ell_n$. Clearly,

$$|Q| \leqslant \ell(Q)$$

for every n-dimensional arithmetic progression. Freiman's inverse theorem should be applicable to the Erdős-Szemerédi conjecture for sets satisfying the additive condition (16).

Let Q be an n-dimensional arithmetic progression of the form (17). If j is such that $\ell_j = \max\{\ell_i : i = 1, \ldots, n\}$ in (17), then

$$Q \supseteq Q_j = \{r + u_j q_j : 0 \leqslant u_j < \ell_j\}.$$

It follows from Lemma 6 that

$$|Q^2| \geqslant |Q_j^2| \gg \left(\frac{\ell_j}{\log \ell_j}\right)^2. \tag{18}$$

The following example shows that this inequality is almost best possible. Fix $n \geqslant 2$. For $\ell \geqslant 2$, consider the n-dimensional arithmetic progression Q with $r = 1$, $q_i = i$ and $\ell_i = \ell$ for $i = 1, \ldots, n$. Then

$$Q = \left\{1 + \textstyle\sum_{i=1}^n i u_i : 0 \leqslant u_i < \ell\right\} \subseteq \left[1, 1 + \tfrac{1}{2}n(n+1)(\ell - 1)\right] \subseteq [1, n^2\ell].$$

We apply the lower bound (18) with $\ell = \max\{\ell_i : i = 1, \ldots, n\}$, and we apply the upper bound (1) with $k = n^2\ell$. For sufficiently large ℓ we obtain

$$\left(\frac{\ell}{\log \ell}\right)^2 \ll |Q^2| \ll \frac{k^2}{(\log k)^{\varepsilon_0}} \ll_n \frac{\ell^2}{(\log \ell)^{\varepsilon_0}},$$

where ε_0 is defined by (2). Since $\ell(Q) = \ell^n$, it is clearly not true that

$$|Q^2| \gg_{n,\varepsilon} \ell(Q)^{2-\varepsilon}.$$

It would be interesting to obtain sufficient conditions for an n-dimensional arithmetic progression Q to satisfy

$$|Q^2| \gg_{n,\varepsilon} |Q|^{2-\varepsilon}.$$

Let A be a set of k positive integers. For $h \geqslant 3$, let $E_h(A)$ denote the set of all numbers that can be written as the sum or product of h elements of A. Erdős and Szemerédi [**4**] also conjectured that

$$|E_h(A)| \gg_\varepsilon k^{h-\varepsilon}$$

for all $\varepsilon > 0$. Nothing is known about this.

References

[1] Barban M. B. and Vehov P. P., *Summation of multiplicative functions of polynomials* (in Russian), Mat. Zametki, 5(6):669–680, 1969. English translation: Math. Notes Acad. Sci. USSR, **5**, 1969, 400–407.

[2] Erdős P., *An asymptotic inequality in the theory of numbers* (in Russian), Vestnik Leningrad Univ., Serija Mat. Mekh. i Astr., **15(13)**, 1960, 41–49.

[3] Erdős P., *Problems and results on combinatorial number theory III*, In M. B. Nathanson, editor, Number Theory Day, New York 1976, volume 626 of Lecture Notes in Mathematics, Berlin, Springer-Verlag, 1977, 43–72.

[4] Erdős P. and Szemerédi E., *On sums and products of integers*, In P. Erdős, L. Alpár, G. Halász, and A. Sárközy, editors, Studies in Pure Mathematics, To the Memory of Paul Turán, Birkhäuser Verlag, Basel, 1983, 213–218.

[5] Freiman G. A., *On the addition of finite sets. I*, Izv. Vysh. Zaved. Matematika, **13(6)**, 1959, 202–213.

[6] Freiman G. A., *Inverse problems of additive number theory. VI on the addition of finite sets. III* Izv. Vysh. Ucheb. Zaved. Matematika, **28(3)**, 1962, 151–157.

[7] Freiman G. A., *Foundations of a Structural Theory of Set Addition*, volume 37 of Translations of Mathematical Monographs, American Mathematical Society, Providence, 1973.

[8] Hall R. R. and Tenenbaum G., *Divisors*, Number 90 in Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1988.

[9] Lev V. F. and Smeliansky P. Y., *On addition of two different sets of integers*, Preprint, 1994.

[10] Nathanson M. B., *The simplest inverse problems in additive number theory*, In A. Pollington and W. Moran, editors, *Number Theory with an Emphasis on the Markoff Spectrum*, Marcel Dekker, 1993, 191–206.

[11] Nathanson M. B., *On sums and products of integers*, submitted, 1994.

[12] Nathanson M. B., *Additive Number Theory: 2. Inverse Theorems and the Geometry of Sumsets*, Graduate Texts in Mathematics. Springer-Verlag, New York, 1995, to appear.

[13] Ruzsa I. Z., *Generalized arithmetic progressions and sumsets*, to appear.

[14] Shiu P., *A Brun-Titchmarsh theorem for multiplicative functions*, J. reine angew. Math., **313**, 1980, 161–170.

[15] Steinig J., *On G. A. Freiman's theorems concerning the sum of two finite sets of integers*, In *Conference on the Structure Theory of Set Addition*, CIRM, Marseille, 1993, 173–186.

[16] Tenenbaum G., *Sur la probabilité qu'un entier possède un diviseur dans un intervalle donné*, In Séminaire de Théorie des Nombres, Paris 1981-1982, volume **38** of Progress in Math., Birkhäuser, Boston, 1983, 303–312.

[17] Tenenbaum G., *Sur la probabilité qu'un entier possède un diviseur dans un intervalle donné*, Compositio Math., **51**, 1984, 243–263.

[18] Vinogradov A. I. and Linnik Yu. V., *Estimate of the sum of the number of divisors in a short segment of an arithmetic progression*, Uspekhi Mat. Nauk (N.S.), **12**, 1957, 277–280.

M.B. Nathanson, Department of Mathematics, Lehman College (CUNY), Bronx, New York 10468, USA • *E-mail :* `nathansn@dimacs.rutgers.edu`

G. Tenenbaum, Institut Élie Cartan, Université Henri-Poincaré–Nancy 1, 54506 Vandœuvre lès Nancy Cedex, France • *E-mail :* `tenenb@ciril.fr`

Astérisque

JEAN-LOUIS NICOLAS

Stratified sets

Astérisque, tome 258 (1999), p. 205-215

<http://www.numdam.org/item?id=AST_1999__258__205_0>

Astérisque **258**, 1999, p. 205–215

STRATIFIED SETS

by

Jean-Louis Nicolas

Abstract. — A set $\mathcal{A}$ of integers is said "stratified" if, for all t, $0 \le t < \operatorname{Card} \mathcal{A}$, the sum of any t distinct elements of $\mathcal{A}$ is smaller than the sum of any $t+1$ distinct elements of $\mathcal{A}$. That implies that all elements of $\mathcal{A}$ should be positive. It is proved that the number of stratified sets with maximal element equal to N is exactly the number $p(N)$ of partitions of N.

1. Introduction

Let $\mathbb{N} = \{1, 2, 3, \dots\}$ denote the set of positive integers. After Erdős and Straus (see [**3**] and [**7**]), a set $\mathcal{A} \subset \mathbb{N}$ is said admissible if for any pairs $\mathcal{A}_1, \mathcal{A}_2$ of subsets of $\mathcal{A}$, one has

$$\left(\sum_{a\in\mathcal{A}_1} a = \sum_{a\in\mathcal{A}_2} a\right) \Rightarrow \mid \mathcal{A}_1 \mid = \mid \mathcal{A}_2 \mid .$$

Here $\mid \mathcal{A} \mid$ will denote the number of elements of $\mathcal{A}$.

Straus has observed that, if $k = \lfloor 2\sqrt{N + 1/4} - 1 \rfloor$, then the set $\mathcal{A} = \{N, N-1, \dots, N-k+1\}$ is admissible, On the other hand, he proved (cf. [**7**]) that if $N = \max_{a\in\mathcal{A}} a$, and $\mathcal{A}$ is admissible, then $\mid \mathcal{A} \mid \le \left(\frac{4}{\sqrt{3}} + o(1)\right)\sqrt{N}$. The constant $4/\sqrt{3}$ has been improved in [**4**], and in [**1**], J.M. Deshouillers and G.A. Freiman have replaced it by the best possible constant 2. In [**2**], they prove that for N large enough, the above example of Straus is the greatest possible admissible set with maximal element N.

Definition 1. — *A set $\mathcal{A} \subset \mathbb{Z}$ is stratified, if for $0 \le t < t'$ the sum of any t distinct elements of $\mathcal{A}$ is strictly smaller than the sum of any t' distinct elements of $\mathcal{A}$.*

1991 ***Mathematics Subject Classification.*** — 11P81, 11B75.

Key words and phrases. — Partition, admissible, stratified, Durfee square.

Research partially supported by C.N.R.S., Institut Girard Desargues, UPRES-A 5028.

Note that from the above definition all the elements of $\mathcal{A}$ should be positive (choose $t = 0$ and $t' = 1$), and so $\mathcal{A}$ is included in $\mathbb{N}$.

Clearly a stratified set is admissible. The above example of Straus is stratified, and in the table at the end of **[4]**, it can be seen that most of large admissible sets are stratified.

In this paper, stratified sets will be described in terms of partitions (Theorem 1). Further, we shall reformulate some of the conjectures about admissible sets given in **[4]** in terms of stratified sets. Finally, we shall show that the number of stratified sets with maximal element N is equal to the number of partitions of N (Theorem 2) and a one to one correspondence associating such a stratified set to a partition of N is explicited. As a corollary, the lower bound given in **[4]** for the total number of admissible sets with maximal element N will be improved.

It is possible to extend the notion of stratified set to subsets in arithmetic progression and in this way to describe some other classes of admissible sets. For instance a subset of odd numbers $\mathcal{A}$ which satisfies that the sum of any t distinct elements is smaller than the sum of any $t + 2$ distinct elements will certainly be admissible, since the sum of t elements and the sum of $(t + 1)$ elements are of different parity and therefore are unequal. I hope to return to this question in an other paper.

At the end of this article, a table of the numbers $p(N)$ of stratified sets and $a(N)$ of admissible sets with largest element N is given. The table of $a(N)$ given in **[4]** is erroneous.

This work has started in September 1991, when G. Freiman was visiting me in Lyon. At that time he was trying to understand the structure of a large admissible set (it was before getting the result of **[1]**), and he wrote on the blackboard many equations like (11) or (12) below. So an important part of this paper is due to G. Freiman, and I thank him strongly.

I thank also very much Marc Deléglise for calculating the values of $a(N)$, and for listing stratified sets which drove me to Theorem 2. I thank also Paul Erdős, András Sárközy, Etienne Fouvry and Jean-Marc Deshouillers for fruitful discussions on this subject.

Notation. — $t^\wedge\mathcal{A}$ will denote the set of the sums of t distinct elements of $\mathcal{A}$.

2. Description of a stratified set

First it will be proved:

Proposition 1. — *Let $\mathcal{A} = \{a_1 < a_2 < \cdots < a_k = N\}$ be a set of positive integers, and $t_0 = \lfloor (k-1)/2 \rfloor$. Then $\mathcal{A}$ is stratified if and only if*

$$\max t_0{}^\wedge\mathcal{A} < \min(t_0 + 1)^\wedge\mathcal{A}. \tag{1}$$

Proof. — From the definition, $\mathcal{A}$ is stratified if for all $t, 1 \le t \le k - 1$,

$$\max t^\wedge\mathcal{A} < \min(t + 1)^\wedge\mathcal{A}. \tag{2}$$

Let us first prove (2) for $t \le t_0$. From (1), one has:

$$a_k + a_{k-1} + \cdots + a_{k-t_0+1} < a_1 + a_2 + \cdots + a_{t_0+1}$$

which implies

$$(3) \qquad a_k + a_{k-1} + \cdots + a_{k-t+1} < a_1 + \cdots + a_{t+1} + \sum_{i=1}^{t_0-t} (a_{t+1+i} - a_{k-t+1-i})\,.$$

But for $1 \le i \le t_0 - t$, we have $t+1+i < k-t+1-i$ since $2i \le 2(t_0 - t) \le k-1-2t < k-2t$; thus, the last sum in (3) is non-positive and (3) yields (2). Let us now suppose that $t > t_0$ and set $S = \sum_{i=1}^{k} a_i$ and $t' = k - t - 1$. We have $k/2 - 1 \le t_0 \le (k-1)/2$, so that

$$t' = k - t - 1 < k - t_0 - 1 \le k - (k/2 - 1) - 1 = k/2 \le t_0 + 1,$$

and so, $t' \le t_0$. From the above proof, one gets

$$(4) \qquad \max(t')^\wedge\mathcal{A} < \min(t'+1)^\wedge\mathcal{A},$$

and from the definition of t',

$$(5) \qquad \max t^\wedge\mathcal{A} = S - \min(t'+1)^\wedge\mathcal{A}$$

and

$$(6) \qquad \min(t+1)^\wedge\mathcal{A} = S - \max(t')^\wedge\mathcal{A}.$$

(4), (5) and (6) prove (2), and this completes the proof of Proposition 1.

Theorem 1

(a) Let k be even. There is a one to one correspondence between the stratified sets $\mathcal{A} \subset \mathbb{Z}$ with $\max\mathcal{A} = N$ and $|\mathcal{A}| = k$ and the solutions of the inequality

$$(7) \qquad x_1 + 2(x_2 + x_{k-1}) + 3(x_3 + x_{k-2}) + \cdots + \frac{k}{2}(x_{k/2} + x_{k/2+1}) \le N - \frac{k^2}{4} - \frac{k}{2}.$$

where the x_i's are non negative integers.

(b) Let k be odd. There is a one to one correspondence between the stratified sets $\mathcal{A} \subset \mathbb{Z}$ with $\max\mathcal{A} = N$ and $|\mathcal{A}| = k$ and the solutions of the inequality:

$$(8) \qquad x_1 + 2(x_2 + x_{k-1}) + \cdots + \frac{k-1}{2}(x_{(k-1)/2} + x_{(k+3)/2}) + \frac{k+1}{2}\, x_{(k+1)/2} \le N - \frac{(k+1)^2}{4}.$$

Proof. — Let $\mathcal{A} = \{a_1, a_2, \dots, a_k\} \subset \mathbb{Z}$ with

$$(9) \qquad a_1 < a_2 < \cdots < a_{k-1} < a_k = N.$$

be a stratified set. Let us introduce the new variables

$$x_i = a_{i+1} - a_i - 1, \quad 1 \le i \le k-1.$$

From (9), one has

$$x_i \geq 0, \quad 1 \leq i \leq k-1. \tag{10}$$

Conversely, (9) is clearly equivalent to $a_k = N$, and (10). Now we have to express (1) in terms of the $x_i's$, and this explains the role played by the parity of k.

Let us suppose k is even. From the definition of x_i, one has

$$x_1 + 2x_2 + \cdots + tx_t = -\frac{t(t+1)}{2} - a_1 - a_2 - \cdots - a_t + ta_{t+1} \tag{11}$$

$$\begin{aligned} 2x_{k-1} + 3x_{k-2} + \cdots + (u+1)x_{k-u} &= N + a_k + a_{k-1} + \cdots + a_{k-u+1} \\ &\quad -(u+1)a_{k-u} \\ &\quad -\frac{(u+1)(u+2)}{2} + 1. \end{aligned} \tag{12}$$

One chooses $t = t_0 + 1 = k/2$ in (11) and $u = t_0 = \frac{k}{2} - 1$ in (12) and then (11) and (12) give

$$\begin{aligned} \max t_0{}^\wedge\mathcal{A} - \min(t_0+1)^\wedge\mathcal{A} &= a_k + a_{k-1} + \cdots + a_{k-t_0+1} \\ &\quad -a_1 - a_2 - \cdots - a_{t_0+1} \\ &= x_1 + 2(x_2 + x_{k-1}) + \cdots \\ &\quad +\frac{k}{2}(x_{k/2} + x_{k/2+1}) \\ &\quad -N + \frac{k^2}{4} + \frac{k}{2} - 1. \end{aligned}$$

The last term -1 allows us to transform the strict inequality (2) in inequality (7) with $\leq$ sign.

The proof of (8) when k is odd is similar.

Corollary 1. — *Let us denote the number of stratified sets with k elements, and maximal element N by $S_k(N)$. The generating functions are:*
for k even

$$\sum_{N=0}^{\infty} S_k(N)x^N = x^{k^2/4+k/2} \prod_{i=1}^{k/2} \frac{1}{(1-x^i)^2}, \tag{13}$$

for k odd:

$$\sum_{N=0}^{\infty} S_k(N)x^N = \frac{x^{(k+1)^2/4}}{1-x^{(k+1)/2}} \prod_{i=1}^{(k-1)/2} \frac{1}{(1-x^i)^2}. \tag{14}$$

Proof. — It follows easily from the theorem, by the classical method of generating series. For k even, the generating series of the number of solutions of

$$x_1 + 2(x_2 + x_{k-1}) + \cdots + \frac{k}{2}(x_{k/2} + x_{(k/2+1)}) = n \tag{15}$$

is

$$\frac{1}{1-x}\prod_{i=2}^{k/2}\frac{1}{(1-x^i)^2},$$

and if in (15) " $= n$ " is replaced by " $\leq n$ ", the generating function must be multiplied by $1/(1-x)$. At last, in (7), the right hand side is $N - k^2/4 - k/2$, which explains the factor $x^{k^2/4+k/2}$ in (13).
For k odd, the proof is similar.

Corollary 2. — *Let $p(n)$ denote the classical partition function, i.e. the number of ways of writing $n = n_1 + n_2 + \cdots + n_k, n_1 \geq n_2 \cdots \geq n_k \geq 1$, and let us define $P(n) = \sum_{i=0}^{n} p(i)p(n-i)$. So, the generating function of $P(n)$ is*

$$\sum_{n=0}^{\infty} P(n)x^n = \prod_{i=1}^{\infty}\frac{1}{(1-x^i)^2}. \tag{16}$$

The number of stratified sets $\mathcal{A}$ with largest element N and with a maximal number of elements is given by

$$P(N-m^2) \; if \; m^2 \leq N < m(m+1)$$

and by

$$P(N-m^2-m) \; if \; m(m+1) \leq N < (m+1)^2.$$

Proof. — Let us suppose first that $m^2 \leq N < m(m+1)$. For $k = 2m-1$, one has

$$N - \frac{(k+1)^2}{4} = N - m^2 \leq m(m+1) - 1 - m^2 = m - 1 = \frac{k-1}{2}, \tag{17}$$

But by (14), (16) and (17), the number of stratified sets, $S_k(N)$, is equal to $P(N-m^2)$. For $k = 2m$, one has $N - \frac{k^2}{4} - \frac{k}{2} = N - m(m+1) < 0$, and from (13) there is no stratified sets with k elements.

The proof of the second case, $m(m+1) \leq N < (m+1)^2$ is similar.

Remark. — It follows from theorem 1 and the above proof, that the maximal number of elements of a stratified set $\mathcal{A}$ with maximal element N is $\lfloor 2\sqrt{N+1/4} - 1 \rfloor$, that is $2m-1$ if $m^2 \leq N < m(m+1)$ and $2m$ if $m(m+1) \leq N < (m+1)^2$.

Table of $P(N)$:

N	=	0	1	2	3	4	5	6	7	8	9	10	11
$P(N)$	=	1	2	5	10	20	36	65	110	185	300	481	752

This table has to be compared with the column $p(N)$ in the table of [4].

3. A conjecture about admissible sets with maximal size

Let m be an integer, $N = m^2 + m - 2, k = 2m - 1$, and let us consider the set $\{N, N-1, \ldots, N-m+2, N-m, N-m-1, \ldots, N-2m+1\}$. If the elements of this set are denoted by $a_1 < a_2 < \cdots < a_k$, and if we set $x_i = a_{i+1} - a_i - 1$, we have $x_i = 0$ for all i but

$$x_{m-1} = x_{(k-1)/2} = 1.$$

So, (8) writes:

$$m - 1 = \frac{k-1}{2} \leq N - \frac{(k+1)^2}{4} = m^2 + m - 2 - m^2 = m - 2$$

which does not hold. Therefore the set is not stratified. It is easy to see that $t_0 = \lfloor (k-1)/2 \rfloor = m - 1$,

$$\max t_0{}^\wedge\mathcal{A} = \min(t_0+1)^\wedge\mathcal{A} + 1$$

but the second largest term of $t_0{}^\wedge\mathcal{A}$ is smaller than all elements of $(t_0+1)^\wedge\mathcal{A}$, and the set is admissible.

A similar counter example admissible but not stratified does exist for $N = m^2 + m - 1, k = 2m - 1$, omitting $N - 2m$ instead of $N - 2m + 1$.
These two counterexamples will be said quasistratified.
Now, conjectures 1 to 4 of [**4**] can be reformulated in the following terms:
Conjecture 1 of [**4**], that the maximal number of elements of an admissible set with greatest element N is $\lfloor 2\sqrt{N + 1/4} - 1 \rfloor$, has been proved by J.M. Deshouillers and G.A. Freiman in [**2**], for N large enough.
Conjecture 2 is replaced by: For $N \geq 20$, the admissible sets of maximal size and largest element N are either stratified, or one of the sets made of odd elements described in conjecture 3 of [**4**] (whenever N is of the form $m^2 - 1$ or $m^2 + m - 1$), or a quasistratified set described above (whenever N is of the form $m^2 + m - 1$ or $m^2 + m - 2$).
Conjecture 4 of [**4**] then becomes an easy consequence of our new conjecture 2.
This new conjecture 2 fits the table of [**4**] for $20 \leq N \leq 50$. This table has been extended up to $N = 60$, and the conjecture is verified for $20 \leq N \leq 60$.

4. How many stratified sets are there ?

Theorem 2. — *The set of stratified sets with largest element N and the set of partitions of N have same cardinal. Moreover an explicit one to one correspondence between these two sets is given.*

Proof. — Let m be an integer, and, as above, let us denote $S_k(N)$ the number of stratified sets with largest element N and with k elements. From (13) and (14) the generating function of $S_{2m-1}(N) + S_{2m}(N)$ is:

$$\begin{aligned}\sum_{N=0}^{\infty} \left(S_{2m-1}(N) + S_{2m}(N)\right) x^N &= \frac{x^{m^2}}{1-x^m} \prod_{i=1}^{m-1} \frac{1}{(1-x^i)^2} + x^{m^2+m} \prod_{i=1}^{m} \frac{1}{(1-x_i)^2} \\ &= x^{m^2} \prod_{i=1}^{m} \frac{1}{(1-x_i)^2}\end{aligned}$$

Now, the generating function of $S(N)$, the total number of stratified sets with largest elements N will be:

$$\sum_{N=0}^{\infty} S(N) x^N = \sum_{m=1}^{\infty} x^{m^2} \prod_{i=1}^{m} \frac{1}{(1-x_i)^2}.$$

But, by an identity due to Euler (cf. [**5**], p. 280):

$$\sum_{m=1}^{\infty} x^{m^2} \prod_{i=1}^{m} \frac{1}{(1-x^i)^2} = \prod_{i=1}^{\infty} \frac{1}{1-x^i} = \sum_{N=0}^{\infty} p(N) x^N, \tag{18}$$

and so, $S(N) = p(N)$.

To the partition $n = n_1 + n_2 + \cdots + n_k$, with $n_1 \geq n_2 \geq \cdots \geq n_k$, let us associate the so-called Ferrers diagram, that is the array of dots made with n_1 dots on the first line, n_2 dots on the second line, ..., and so on n_k dots on the k^{th} line. For instance, to $10 = 4 + 3 + 1 + 1 + 1$ corresponds the array:

● ● ● ●
● ● ●
●
●
●

This graphical representation contains a square in the upper left corner, and the largest such square is called "Durfee square" in [5, p. 281].

In the combinatorial proof of Euler's identity (18), it is observed in [**5**] that

$$x^{m^2} \prod_{i=1}^{m} \frac{1}{(1-x^i)^2}$$

is the generating function of the number of partitions such that the Durfee squares have an edge of length exactly m. To find the wanted one to one correspondence we just have to use the combinatorial proof of (18) in [5, p. 281].

From the above proof, one can see that $S_{2m}(N)$ is equal to the number of partitions of N such that the Durfee square has an edge of length m, and moreover such that the corresponding array contains the rectangle of length $m+1$ and height m. Similarly, $S_{2m-1}(N)$ is equal to the number of partitions of N such that the Durfee square has an edge of length m, but such that the array does not contain the above mentioned rectangle.

Let us suppose first that the Ferrers diagram does not contain the rectangle $(m+1) \times m$ (that means that $n_m = m$) and choose $k = 2m-1$. The Ferrers diagram consists of three parts: The Durfee square and two tails. Let us denote by $\mathcal{U}$ the upper right tail and by $\mathcal{V}$ the lower left tail, so that

$$n = m^2 + |\mathcal{U}| + |\mathcal{V}|.$$

Now, $\mathcal{U}$ can be interpreted as the Ferrers diagram (in column) of a partition of $|\mathcal{U}|$, the parts of which are $\leq m-1$. Let us denote by x_i the number of columns of $\mathcal{U}$ with height i ; then this partition writes

$$(19) \qquad x_1 + 2x_2 + \cdots + (m-1)x_{m-1} = |\mathcal{U}|.$$

Similarly $\mathcal{V}$ can be interpreted as the Ferrers diagram (in row) of a partition of $|\mathcal{V}|$, the parts of which are $\leq m$. Let us denote by y_i the number of rows of $\mathcal{V}$ with length i ; then this partition writes

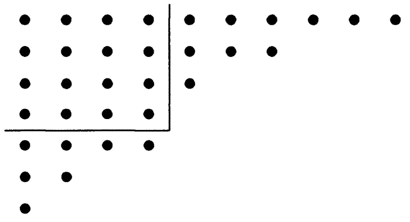
$$(20) \qquad y_1 + 2y_2 + \cdots + my_m = |\mathcal{V}|.$$

By introducing a new variable x_k, let us transform the inequality (8) in the equality:

$$(x_1 + x_k) + 2(x_2 + x_{k-1}) + \cdots + (m-1)(x_{m-1} + x_{m+1}) + mx_m = N - m^2 = |\mathcal{U}| + |\mathcal{V}|.$$

The values of $x_1, \ldots, x_{m-1}$ are given by (19), the values of $x_k = y_1, x_{k-1} = y_2, \ldots, x_m = y_m$ are given by (20), and the stratified set $(a_1, a_2, \ldots, a_k = N)$ can be obtained by $a_k = N$, and $a_i = a_{i+1} - 1 - x_i$.

Example $N = 33 = 10 + 7 + 5 + 4 + 4 + 2 + 1$



$m = 4, k = 7$
$|\mathcal{U}| = 10, x_3 = 1, x_2 = 2, x_1 = 3.$
$|\mathcal{V}| = 7, y_4 = 1, y_3 = 0, y_2 = 1, y_1 = 1.$
The stratified set corresponding to this partition is (19, 23, 26, 28, 30, 31, 33).

Whenever the Ferrers diagram does contain the rectangle $(m+1) \times m$ (that means that $n_{m-1} \geq m+1$) one chooses $k = 2m$, and before defining the tails $\mathcal{U}$ and $\mathcal{V}$ we have to take off the rectangle, so that

$$n = m(m+1) + |\mathcal{U}| + |\mathcal{V}|.$$

The parts of the partition represented by $\mathcal{U}$ are allowed to be equal to m, so that (19) becomes

$$x_1 + 2x_2 + \cdots + mx_m = |\mathcal{U}|,$$

while (20) does not change. (7) becomes:

$$(x_1 + x_k) + 2(x_2 + x_{k-1}) + \cdots + m(x_m + x_{m+1}) = N - m(m+1) = |\mathcal{U}| + |\mathcal{V}|$$

and the end of the calculation of the $a_i's$ is similar.

For instance the stratified set associated to the partition of $10 = 4+3+1+1+1$, the array of which is displayed above, is $(6, 8, 9, 10)$.

Corollary 3. — *The number $a(N)$ of admissible sets with largest element N is greater than $p(N)$, the number of partitions of N.*

Proof. — It follows immediately from Theorem 2, since any stratified set is admissible.
From the result of Hardy and Ramanujan, it is known that (cf. [**6**], formula 1.41):

$$p(n) \sim \frac{1}{4n\sqrt{3}} \exp\left(\pi\sqrt{\frac{2n}{3}}\right).$$

So the above corollary improves the lower bound given in [**4**]:

$$a(N) \geq 2^{2\sqrt{N-2}} - 1.$$

For the moment, I am not able to improve the upper bound of [**4**]:

$$a(N) \leq \exp(c\sqrt{N}\log N),$$

but I conjecture that $a(N)$ is not much greater than $p(N)$ and satisfies

$$a(N) = \exp\left((\pi\sqrt{2/3} + o(1))\sqrt{N}\right),$$

and, may be (see the table) that $a(N) \sim p(N)$.

N	$p(N)$	$a(N)$	$a(N)/p(N)$	N	$p(N)$	$a(N)$	$a(N)/p(N)$
1	1	1	1.00	41	44583	235189	5.28
2	2	2	1.00	42	53174	273087	5.14
3	3	3	1.00	43	63261	335262	5.30
4	5	6	1.20	44	75175	394565	5.25
5	7	9	1.29	45	89134	465548	5.22
6	11	15	1.36	46	105558	551586	5.23
7	15	23	1.53	47	124754	659344	5.29
8	22	39	1.77	48	147273	750256	5.09
9	30	54	1.80	49	173525	912459	5.26
10	42	87	2.07	50	204226	1051209	5.15
11	56	121	2.16	51	239943	1230129	5.13
12	77	178	2.31	52	281589	1433643	5.09
13	101	249	2.47	53	329931	1705477	5.17
14	135	362	2.68	54	386155	1900438	4.92
15	176	484	2.75	55	451276	2308752	5.12
16	231	708	3.06	56	526823	2604726	4.94
17	297	928	3.12	57	614154	3041041	4.95
18	385	1265	3.29	58	715220	3483815	4.87
19	490	1685	3.44	59	831820	4132473	4.97
20	627	2306	3.68	60	966467	4527898	4.69
21	792	2886	3.64	61	1121505	5491786	4.90
22	1002	3918	3.91	62	1300156	6101289	4.69
23	1255	4987	3.97	63	1505499	7090459	4.71
24	1575	6418	4.07	64	1741630	8019859	4.60
25	1958	8265	4.22	65	2012558	9504818	4.72
26	2436	10601	4.35	66	2323520	10230396	4.40
27	3010	13104	4.35	67	2679689	12413471	4.63
28	3718	16947	4.56	68	3087735	13595124	4.40
29	4565	21069	4.62	69	3554345	15791911	4.44
30	5604	26088	4.66	70	4087968	17584116	4.30
31	6842	32804	4.79	71	4697205	20860378	4.44
32	8349	40935	4.90	72	5392783	22095088	4.10
33	10143	49360	4.87	73	6185689	26904818	4.35
34	12310	61712	5.01	74	7089500	29025643	4.09
35	14883	75338	5.06	75	8118264	33687817	4.15
36	17977	90456	5.03	76	9289091	37071664	3.99
37	21637	111771	5.17	77	10619863	44046119	4.15
38	26015	134685	5.18	78	12132164	45918783	3.78
39	31185	160353	5.14	79	13848650	56109976	4.05
40	37338	195993	5.25	80	15796476	59689468	3.78

References

[1] Deshouillers J-M. and Freiman G.A., *On an additive problem of Erdős and Straus I*, Israel J. of Maths, **92**, 1995, 33–43.

[2] Deshouillers J-M. and Freiman G.A., *On an additive problem of Erdős and Straus, 2*, this volume.

[3] Erdős P., *Some remarks on number theory III*, Math. Lapok, **13**, 1962, 28–38.

[4] Erdős P., Nicolas J-L., Sárkőzy A., *Sommes de sous-ensembles*, Sem. Th. Nb. Bordeaux. **3**, 1991, 55–72.

[5] Hardy G. H. and Wright E. M., *An introduction to the theory of numbers*, 4th edition, Oxford at the Clarendon Press, 1960.

[6] Hardy G. H. and Ramanujan S., *Asymptotic formulae in combinatory analysis*, Proc. London Math. Soc. 2, **17**, 1918, 75–115. Collected Papers of Ramanujan, 276–309.

[7] Straus E. G., *On a problem in combinatorial number theory*, J. Math. Sci., **1**, 1966, 77–80.

J-L. NICOLAS, Mathématiques,, Université Claude Bernard (Lyon 1), F-69622 VILLEURBANNE Cedex • *E-mail :* jlnicola@in2p3.fr

Astérisque

YONUTZ V. STANCHESCU

On the structure of sets of lattice points in the plane with a small doubling property

Astérisque, tome 258 (1999), p. 217-240

<http://www.numdam.org/item?id=AST_1999__258__217_0>

Astérisque **258**, 1999, p. 217–240

ON THE STRUCTURE OF SETS OF LATTICE POINTS IN THE PLANE WITH A SMALL DOUBLING PROPERTY

by

Yonutz V. Stanchescu

Abstract. — We describe the structure of sets of lattice points in the plane, having a small doubling property. Let $\mathbb{K}$ be a finite subset of $\mathbb{Z}^2$ such that

$$|\mathbb{K}+\mathbb{K}| < 3.5|\mathbb{K}| - 7.$$

If $\mathbb{K}$ lies on three parallel lines, then the convex hull of $\mathbb{K}$ is contained in three compatible arithmetic progressions with the same common difference, having together no more than

$$|\mathbb{K}| + \frac{3}{4}\Big(|\mathbb{K}+\mathbb{K}| - \frac{10}{3}|\mathbb{K}| + 5\Big)$$

terms. This upper bound is best possible.

Notation

We write $[m,n] = \{x \in \mathbb{Z} | \ m \leq x \leq n\}$. For any nonempty finite set $K \subseteq \mathbb{R}$, $K = \{u_1 < u_2 < \cdots < u_k\}$ we denote by $k = |K|$ the *cardinality* of K and by $\ell(K)$ the *length* of K, that is the difference between its maximal and minimal elements. If $K \subseteq \mathbb{Z}$ and $k \geq 2$, by $d(K)$ we denote the *greatest common divisor* of $u_i - u_1$, $1 \leq i \leq k$. If $k = 1$, we put $d(K) = 0$. Let $h(K) = \ell(K) - |K| + 1$ denote the number of *holes* in K, that is $h(K) = | \ [u_1, u_k] \ \setminus \ K|$.

Let A and B be two subsets of an abelian group $(G, +)$. As usual, their *sum* is defined by $A + B = \{x \in G \mid x = a + b, \ a \in A, \ b \in B\}$ and we put $2A = A + A$. The *convex hull* of a set $\mathbb{S} \subseteq \mathbb{R}^2$ is denoted by $\mathrm{conv}(\mathbb{S})$. Vectors will be written in the form $u = (u_1, u_2)$, where u_1 and u_2 are the coordinates with respect to the canonical basis $e_1 = (1,0)$, $e_2 = (0,1)$.

1991 ***Mathematics Subject Classification.*** — 05D05, 11B75, 11P99.
Key words and phrases. — Two-dimensional lattice points, small doubling property.

1. Introduction

In additive number theory we usually ask what may be said about $M + M$, for a given set M. As a counterbalance to this direct approach, consider now the inverse problem: we study the properties of M, when some characteristic of $M + M$ is given, for example, the cardinality of the sum set $M + M$. It was noticed by Freiman **[F1]** that the assumption that $|2M|$ is small compared to $|M|$, implies strong restrictions on the structure of the set M. If $|2M| = 2|M| - 1$ and $M \subseteq \mathbb{Z}$, then M is an arithmetic progression. If we choose bigger values for $|2M|$, the problem ceases to be trivial. The fundamental theorem of G.A. Freiman **[F2]** gives the structure of finite sets of integers with small doubling property: $|2M| < c_0|M|$, where c_0 is any given positive number. This theorem was proved using geometric methods of number theory and a modification of the method of trigonometric sums. Y. Bilu recently studied in **[B]** a case when c_0 is a slowly growing function of $|M|$. The generalization to the case of different summands $M+N$, with a new proof, is to be found in the paper of I.Z. Ruzsa **[R]**.

However, in the case of small values of the constant c_0, elementary methods yield sharper results. Let $\mathbb{K} \subseteq \mathbb{Z}^2$ be a finite set of lattice points. Two cases have been studied by G. A. Freiman **[F1]**, pp.11, 28.

Theorem A. — *If* $|\mathbb{K} + \mathbb{K}| < 3|\mathbb{K}| - 3$, *then*
(1) $\mathbb{K}$ *lies on a straight line.*
(2) $\mathbb{K}$ *is contained in an arithmetic progression of no more than* $v = |\mathbb{K}+\mathbb{K}| - |\mathbb{K}| + 1$ *terms.* □

Theorem B. — *If* $|\mathbb{K} + \mathbb{K}| < \frac{10}{3}|\mathbb{K}| - 5$, $|\mathbb{K}| \geq 11$ *and* $\mathbb{K}$ *is not contained in a line, then*
(1) $\mathbb{K}$ *lies on two parallel straight lines.*
(2) $\mathbb{K}$ *is contained in two arithmetic progressions with the same common difference having together no more than* $v = |\mathbb{K} + \mathbb{K}| - 2|\mathbb{K}| + 3$ *terms.* □

The generalization of Theorems A(1) and B(1), to s lines, $s \geq 3$, was obtained in **[S2]**:

Theorem C. — *If* $|\mathbb{K} + \mathbb{K}| < \left(4 - \frac{2}{s+1}\right)|\mathbb{K}| - (2s + 1)$ *and* $|\mathbb{K}| \geq 16s(s + 1)(2s + 1)$, *then there exist s parallel lines which cover the set* $\mathbb{K}$. □

A result which generalizes Theorems A(2) and B(2) was obtained in **[S3]**.

Theorems A(1), B(1) and C cannot be sharpened by increasing the upper bound for $|2\mathbb{K}|$. (see Example A in **[S2]**.) Assertion (2) of Theorems A and B gives the precise structure theorem for $s = 1$ and $s = 2$. In **[S2]** we obtained a sharpening of Theorem B(2) by giving the best possible value of the upper bound for $|2\mathbb{K}|$, under the additional assumption that $\mathbb{K}$ lies on $s = 2$ parallel lines. We proved that Theorem B(2) is true, even we replace $|2\mathbb{K}| < \frac{10}{3}|\mathbb{K}| - 5$ by $|2\mathbb{K}| < 4|\mathbb{K}| - 6$. More precisely:

Theorem S. — *Let* $\mathbb{K} \subseteq \mathbb{Z}^2$ *be a finite set, which lies on the lines* $x_2 = 0$ *and* $x_2 = 1$. *Let the set of abscissae for* $x_2 = 0$ *and* $x_2 = 1$, *respectively be equal to* A *and* B.

(1) If $\ell(A)+\ell(B)\le 2|\mathbb{K}|-5$, *then* $(d(A),d(B))=1$ *and*

$$|2\mathbb{K}|\ge(3|\mathbb{K}|-3)+h(A)+h(B)=(2|\mathbb{K}|-1)+\ell(A)+\ell(B).$$

(2) If $\ell(A)+\ell(B)\ge 2|\mathbb{K}|-4$ *and* $(d(A),d(B))=1$, *then* $|2\mathbb{K}|\ge 4|\mathbb{K}|-6$. □

It is not difficult to give examples to show that Theorems A(2), B(2) and Theorem S cannot be sharpened by reducing the quantity v or by increasing the upper bound for $|2\mathbb{K}|$. (see Examples B1 and B2 of Section 3, **[S2]**)

The present paper is devoted to the generalization of Theorem A(2) and S to the case of $s=3$ parallel lines. Instead of condition $|2\mathbb{K}|<3k-3$, of Theorem A and condition $|2\mathbb{K}|<\frac{10}{3}k-5$ of Theorem B, we study now a set $\mathbb{K}$ of integer points on a plane, with the following small doubling property

$$|2\mathbb{K}|<3.5|\mathbb{K}|-7.$$

Take a lattice $\mathcal{L}$ generated by $\mathbb{K}$. We wish to obtain an estimate for the number of points of $\mathcal{L}$ that lie in $\mathrm{conv}(\mathbb{K})$; we are interested in an upper bound of $|\mathcal{L}\cap\mathrm{conv}(\mathbb{K})|$. Some estimate of this number was obtained in [S2, Theorem C]. In this paper we shall give the best possible estimate for $|\mathcal{L}\cap\mathrm{conv}(\mathbb{K})|$. The result implies an affirmative answer to a question of G.A. Freiman **[F3]** and generalizes previous results of **[F1]** and **[S2]**.

2. Main Result

An *arithmetic progression* in $\mathbb{Z}^2$ is a set of the form

$$P=P(a,\Delta)=\{a,a+\Delta,a+2\Delta,\dots,a+(p-1)\Delta\},$$

where $a,\ \Delta\in\mathbb{Z}^2$ and $p=|P|\ge 1$. The vector Δ is called the *common difference* of the progression and a is the *initial term* . We say that $P_i=P_i(a_i,\ \Delta_i)$, $i=1,2,3$ are *compatible* arithmetic progressions, if $\Delta_1=\Delta_2=\Delta_3=\Delta$ and $a_1+a_3\equiv 2a_2(\mathrm{mod}\,\Delta)$.

Now we are ready to formulate our main result.

Theorem 1. — *Let* $\mathbb{L}\subseteq\mathbb{Z}^2$ *be a finite set of lattice points with small doubling property:*

$$|\mathbb{L}+\mathbb{L}|<3.5|\mathbb{L}|-7. \tag{2.1}$$

(1) If $|\mathbb{L}|\ge 1344$, *then the set* $\mathbb{L}$ *lies on no more than three parallel lines.*
(2) If $\mathbb{L}$ *is not contained in any two parallel lines, then* $\mathrm{conv}(\mathbb{L})\cap\mathbb{Z}^2$ *is included in three compatible arithmetic progressions having together no more than*

$$v=|\mathbb{L}|+\frac{3}{4}\Big(|\mathbb{L}+\mathbb{L}|-\frac{10}{3}|\mathbb{L}|+5\Big)=\frac{3}{4}\Big(|\mathbb{L}+\mathbb{L}|-2|\mathbb{L}|+5\Big) \tag{2.2}$$

terms.

Assertion (1) of Theorem 1 is a partial case of Theorem C, for $s=3$. We shall reformulate our main result and prove that the new formulation implies assertion (2) of Theorem 1. We need some definitions. Let $\mathbb{K} \subseteq \mathbb{Z}^2$ be a finite set of lattice points that lies on three parallel lines:

$$\begin{aligned}&\mathbb{K} = \mathbb{K}_1 \cup \mathbb{K}_2 \cup \mathbb{K}_3,\\ &\mathbb{K}_1 \subseteq (x_2 = 0),\ \mathbb{K}_2 \subseteq (x_2 = 1),\ \mathbb{K}_3 \subseteq (x_2 = h),\ h \geq 2.\end{aligned} \tag{2.3}$$

Let the set of abscissae of $\mathbb{K}_i$ be respectively equal to K_i and denote $d_i = d(K_i)$.

Put

$$\mathbb{K}^* = \operatorname{conv}(\mathbb{K}) \cap \mathbb{Z}^2,\ k = |K|,\ k^* = |\mathbb{K}^*| \tag{2.4}$$

and

$$d(\mathbb{K}) = \gcd(d_1, d_2, d_3). \tag{2.5}$$

Such a finite set of $\mathbb{Z}^2$ is called a *reduced set of lattice points,* if $h = 2$ and $d(\mathbb{K}) = 1$.

We would like to note at this point that this definition may be formulated in an obvious way, for sets that lie on $s \geq 2$ parallel lines. In this paper, however, a reduced set of lattice points will always be a set that lies on three parallel lines.

Theorem 2. — *Let $\mathbb{K} \subseteq \mathbb{Z}^2$ be a reduced set of lattice points. If $|2\mathbb{K}| < 3.5|\mathbb{K}| - 7$, then*

$$k^* := |\operatorname{conv}(\mathbb{K}) \cap \mathbb{Z}^2| \leq |\mathbb{K}| + \frac{3}{4}\Big(|2\mathbb{K}| - \frac{10}{3}|\mathbb{K}| + 5\Big) = \frac{3}{4}\Big(|2\mathbb{K}| - 2|\mathbb{K}| + 5\Big).$$

Proof of case (2) of Theorem 1, assuming Theorem 2. — Since $\mathbb{L}$ lies on three parallel lines, there is an affine isomorphism of the plane which maps $\mathbb{L}$ onto a set $\mathbb{K}$ such that

(i) $\mathbb{K}$ lies on $(x_2 = 0),\ (x_2 = 1),\ (x_3 = h),\ h \geq 2$,

(ii) $m_1 = m_2 = 0$, where we put $m_i = \min(K_i)$, for $i = 1, 2, 3$.

Since the function $|2\mathbb{L}|$ is an affine invariant of the set $\mathbb{L}$, we see that

$$|2\mathbb{K}| = |2\mathbb{L}| < 3.5|\mathbb{L}| - 7 = 3.5|\mathbb{K}| - 7. \tag{2.6}$$

Denote $d = d(\mathbb{K})$. Remark that, thanks to the small doubling property (2.6) one has

$$h = 2 \text{ and } m_1 + m_3 \equiv 2m_2 (\operatorname{mod} d). \tag{2.7}$$

Indeed, if $h > 2$, then $(\mathbb{K}_1 + \mathbb{K}_3) \cap 2\mathbb{K}_2 = \varnothing$ and thus

$$\begin{aligned}|2\mathbb{K}| &\geq |2K_1| + |K_1 + K_2| + |K_1 + K_3| + |2K_2| + |K_2 + K_3| + |2K_3|\\ &\geq (2k_1 - 1) + (k_1 + k_2 - 1) + (k_1 + k_3 - 1)\\ &\quad + (2k_2 - 1) + (k_2 + k_3 - 1) + (2k_3 - 1)\\ &= 4k - 6 \geq 3.5k - 7.\end{aligned} \tag{2.8}$$

In the same way, if $m_1 + m_3 \not\equiv 2m_2 (\operatorname{mod} d)$, then for $x \in K_1; y', y'' \in K_2$, $z \in K_3$ we have $y' + y'' \equiv 2m_2 \not\equiv m_1 + m_3 = x + z (\operatorname{mod} d)$. Thus, $(\mathbb{K}_1 + \mathbb{K}_3) \cap 2\mathbb{K}_2 = \varnothing$ is valid and (2.8) follows again.

Consequently, $\mathbb{K}$ and $\mathbb{L}$ are contained each in three equidistant compatible arithmetic progressions.

Equation (2.7) and (ii) ensure that $m_3 \equiv 2m_2 - m_1 = 0 (\operatorname{mod} d)$. This yields $w \equiv 0 (\operatorname{mod} d)$ for every $w \in K_1 \cup K_2 \cup K_3$. We can now easily check that the linear isomorphism $(x, y) \to (x/d, y)$, maps $\mathbb{K}$ onto a reduced set $\mathbb{K}'$ of lattice points. Assertion (2) of Theorem 1 follows now easily, because of the inequality

$$v = |\mathbb{K}'^*| \le \frac{3}{4}(|2\mathbb{K}'^*| - 2k'^* + 5) = \frac{3}{4}(|2\mathbb{K}| - 2k + 5) = \frac{3}{4}(|2\mathbb{L}| - 2|\mathbb{L}| + 5),$$

due to Theorem 2 applied to the set $\mathbb{K}'^*$. □

As usual, the solution of an inverse problem allows us to obtain nontrivial lower bounds for $|\mathbb{K} + \mathbb{K}|$, thus solving at the same time a direct additive problem. By $L^* = L(\mathbb{K}^*) = \sum_{i=1}^{3} \ell_i^*$, we denote the *length* of $\mathbb{K}^*$, where $\ell_1^* = \ell_1 = \ell(K_1)$, $\ell_3^* = \ell_3 = \ell(K_3)$ and $\ell_2^* = \max(\mathrm{conv}(\mathbb{K}) \cap (x_2 = 1)) - \min(\mathrm{conv}(\mathbb{K}) \cap (x_2 = 1)) \ge \ell_2 = \ell(K_2)$. The assertion of Theorem 1 and 2 may be reworded as follows:

Theorem 3. — *Let $\mathbb{K} \subseteq \mathbb{Z}^2$ be a finite set of lattice points which lies on three parallel lines $x_2 = 0$, $x_2 = 1$, $x_2 = 2$.*
(1) If $L^* \le \frac{9}{8}(|\mathbb{K}| - 4)$, then $d(\mathbb{K}) = 1$ and $|2\mathbb{K}| \ge (2|\mathbb{K}| - 1) + \frac{4}{3}L^*$.
(2) If $L^* \ge \frac{9}{8}(|\mathbb{K}| - 4)$ and $d(\mathbb{K}) = 1$, then $|2\mathbb{K}| \ge 3.5|\mathbb{K}| - 7$.

We conjecture that inequality $|2\mathbb{K}| < 3.5k - 7$ of Theorem 2 may be actually replaced by $|2\mathbb{K}| \le 4k - 7$.

Conjecture. — *Let $\mathbb{K} \subseteq \mathbb{Z}^2$ be a reduced set of lattice points that lies on three parallel lines. If $|2\mathbb{K}| \le 4|\mathbb{K}| - 7$, then*

$$k^* := |\mathrm{conv}(\mathbb{K}) \cap \mathbb{Z}^2| \le |\mathbb{K}| + \frac{3}{4}\Big(|2\mathbb{K}| - \frac{10}{3}|\mathbb{K}| + 5\Big) = \frac{3}{4}\Big(|2\mathbb{K}| - 2|\mathbb{K}| + 5\Big). \qquad \square$$

We construct an example $\mathbb{K} \subseteq \mathbb{Z}^2$ such that
(i) $\mathbb{K}$ satisfies the small doubling property $|2\mathbb{K}| < 3.5k - 7$ or $|2\mathbb{K}| \le 4k - 7$.
(ii) The number of lattice points in conv($\mathbb{K}$) is exactly $k^* = \frac{3}{4}(|2\mathbb{K}| - 2k + 5)$.
This means that the upper bound (2.2) is best possible. Thus, Theorems 1 and 2 cannot be sharpened by reducing the quantity $v = k^*$.

Example. — Choose $a \ge b$ two natural numbers and define $\mathbb{K} \subseteq \mathbb{Z}^2$ by :

$$K_1 = \{0, 1, 2, \ldots, 2a+b\} \cup \{2a+2b\},\ K_2 = \{0, 1, 2, \ldots, a\} \cup \{a+b\},\ K_3 = \{0\}. \quad (2.9)$$

Then $k_1 = 2a + b + 2$, $k_2 = a + 2$, $k_3 = 1$, $k = 3a + b + 5$, $k^* = L^* + 3 = L + 3 = 3a + 3b + 3, 4k - 7 = 12a + 4b + 13$. Note that $2\mathbb{K}_2 = \mathbb{K}_1 + \mathbb{K}_3$ and therefore

$$\begin{aligned}|2\mathbb{K}| &= |2K_1| + |K_1 + K_2| + |K_1 + K_3| + |K_2 + K_3| + |2K_3| \\ &= (4a + 3b + 2) + (3a + 2b + 2) + (2a + b + 2) + (a + 2) + 1 \\ &= 10a + 6b + 9 = (2k - 1) + \frac{4}{3}L^* = (2k - 1) + \frac{4}{3}(k^* - 3).\end{aligned}$$

This proves (ii), that is $k^* = \frac{3}{4}(|2\mathbb{K}| - 2k + 5)$. Moreover, assertion (i) is also true because, if $a \ge b - 2$, then $|2\mathbb{K}| \le 4|\mathbb{K}| - 7$ and if $a > 5b - 3$, then $|2\mathbb{K}| < 3.5|\mathbb{K}| - 7$. □

We shall need a generalization of Theorem A(2) to the case of distinct summands. The first result in this direction, due to G.A. Freiman ([**F4**]), was sharpened recently by Lev & Smeliansky [**L-S**] and Stanchescu [**S1**].

Let $A = \{0 = a_1 < a_2 < \cdots < a_k\}$, $B = \{0 = b_1 < b_2 < \cdots < b_\ell\}$ be two sets of integers. Define $\varepsilon = \varepsilon(A, B) \in \{0, 1\}$ by $\varepsilon = 1$, if $\ell(A) = \ell(B)$, and $\varepsilon = 0$, if $\ell(A) \neq \ell(B)$.

Theorem D
(1) If $\max(\ell(A), \ell(B)) \leq |A| + |B| - 2 - \varepsilon$, *then* $|A + B| \geq (|A| + |B| - 1) + \max(h(A), h(B))$.

(2) If $\ell(A) \geq |A| + |B| - 1 - \varepsilon$, $\ell(A) \geq \ell(B)$ *and* $d(A) = 1$, *then* $|A + B| \geq |A| + 2|B| - 2 - \varepsilon$. *If* $\ell(A) \geq |A| + |B| - 2$, $\ell(A) \geq \ell(B)$ *and* $d(A \cup B) = 1$, *then* $|A + B| \geq |A| + |B| - 3 + \min(|A|, |B|)$.

(3) If $d = d(A) > 1$ *and B intersects exactly s residue classes modulo d, then* $|A+B| \geq |B| + s(|A| - 1)$. *If* $d(A \cup B) = 1$, *then* $|A + B| \geq |B| + 2(|A| - 1)$. □

The proof of D(1) is to be found in [**S1**] and of D(2) in [**L-S**]. We shall use this theorem for $A = K_i$ and $B = K_j$, $1 \leq i, j, \leq 3$. In this case we put $\varepsilon_{ij} = \varepsilon(K_i, K_j)$.

Denote by $k_i = |\mathbb{K}_i| = |K_i|$, $m_i = \min(K_i)$, $M_i = \max(K_i)$, $\ell_i = \ell(K_i)$, $d_i = d(K_i)$, $h_i = h(K_i)$, for every $1 \leq i \leq 3$. Denote by $H = h_1 + h_2 + h_3$, the number of *interior holes* of $\mathbb{K}$ and by $H^* = |K^*| - |K| = k^* - k$, the *total number of holes* of $\mathbb{K}$. By $L = L(\mathbb{K}) = \ell_1 + \ell_2 + \ell_3 = H + k - 3$, we denote the *length* of $\mathbb{K}$. For every pair $1 \leq i < j \leq 3$, we let $\mathbb{K}_{ij} = \mathbb{K}_i \cup \mathbb{K}_j$ and $d_{ij} = (d_i, d_j)$ the greatest common divisor of d_i and d_j.

In the remaining sections the set $\mathbb{K} \subseteq \mathbb{Z}^2$ denotes a reduced set of lattice points (on three parallel lines). We note at this point two inequalities, which will be used in the paper:

$$\begin{aligned}|2\mathbb{K}| &= |2K_1| + |K_1 + K_2| + \max(|2K_2|, |K_1 + K_3|) + |K_2 + K_3| + |2K_3| \\ &\geq (2k_1 - 1) + (k_1 + k_2 - 1) + \max(2k_2 - 1, k_1 + k_3 - 1) \\ &\qquad\qquad + (k_2 + k_3 - 1) + (2k_3 - 1),\end{aligned}$$

which leads to

$$|2\mathbb{K}| \geq 3k_1 + 4k_2 + 3k_3 - 5, \tag{2.10}$$

$$|2\mathbb{K}| \geq 4k_1 + 2k_2 + 4k_3 - 5. \tag{2.11}$$

3. Some Lemmas

Lemma 3.1. — *Suppose* $k_2 = 1$. *Then* $|2\mathbb{K}| \geq 4|\mathbb{K}| - 7$.

Proof. — Since $k_2 = 1$, inequality (2.11) yields $|2\mathbb{K}| \geq 4k_1 + 2k_2 + 4k_3 - 5 = 4k - 7$. □

Lemma 3.2. — *Suppose that* K_1 *and* K_3 *lie each in one residue class modulo d,* $d > 1$. *Then* $|2\mathbb{K}| \geq 4|\mathbb{K}| - 7$.

Proof. — Since $\mathbb{K}$ is a reduced set, it follows that in K_2 there are at least two elements in-congruent modulo d. We estimate $|K_1 + K_2|$ and $|K_2 + K_3|$ by Theorem D(3) and obtain

$$|2\mathbb{K}| \geq (2k_1 - 1) + (k_2 + 2k_1 - 2) + (2k_2 - 1) + (2k_3 + k_2 - 2) + (2k_3 - 1) \geq 4k - 7. \quad \square$$

Lemma 3.3. — *Suppose* $k_1 = \max(k_1, k_2, k_3)$, $d_1 = d(K_1) > 1$. *Then* $|2\mathbb{K}| \geq 4|\mathbb{K}| - 7$.

Proof. — If $k_2 = 1$, we use Lemma 3.1. If $k_3 = 1$, then K_1 and K_3 lie each in only one residue class modulo d_1 and we apply Lemma 3.2. Therefore, we assume

$$\min(k_1, k_2, k_3) \geq 2. \tag{3.1}$$

We distinguish three cases:

(a) Suppose that $(d_1, d_2) = (d_1, d_3) = 1$. Theorem D(3) gives

$$\begin{aligned}|2\mathbb{K}| &\geq |2K_1| + |K_1 + K_2| + |K_1 + K_3| + |K_2 + K_3| + |2K_3| \\ &\geq (2k_1 - 1) + (k_2 + 2k_1 - 2) + (2k_1 + k_3 - 2) + (k_2 + k_3 - 1) + (2k_3 - 1) \\ &= 6k_1 + 2k_2 + 4k_3 - 7 = 4k - 6 + 2(k_1 - k_2) - 1 \geq 4k - 7.\end{aligned}$$

(b) Suppose $d = (d_1, d_2) > 1$. It follows that $(d, d_3) = 1$, because $\mathbb{K}$ is reduced. Theorem D(3) yields

$$\begin{aligned}|2\mathbb{K}| &\geq (2k_1 - 1) + (k_1 + k_2 - 1) + (2k_1 + k_3 - 2) + (2k_2 + k_3 - 2) + (2k_3 - 1) \\ &= 4k - 7 + (k_1 - k_2) \geq 4k - 7.\end{aligned}$$

(c) Suppose $(d_1, d_3) > 1$. We apply Lemma 3.2. □

Lemma 3.4. — *Suppose* $k_2 = \max(k_1, k_2, k_3)$ *and* $d_2 = d(K_2) > 1$. *Then* $|2\mathbb{K}| \geq 4|\mathbb{K}| - 7$.

Proof

(a) Suppose $1 = k_3 = k_1$. Then we apply Lemma 3.2.

(b) Suppose $1 = k_3 < k_1 \leq k_2$. It is clear that $(d_1, d_2) = 1$, because $\mathbb{K}$ is reduced. Theorem D(3) implies

$$|2\mathbb{K}| \geq (2k_1 - 1) + (2k_2 + k_1 - 2) + (2k_2 - 1) + k_2 + 1 = 3k_1 + 5k_2 - 3 \geq 4k - 7.$$

We may suppose now that $k_1 \geq k_3 \geq 2$.

(c) Suppose $(d_2, d_1) = (d_2, d_3) = 1$. Using Theorem D(3) we get

$$\begin{aligned}|2\mathbb{K}| &\geq (2k_1 - 1) + (2k_2 + k_1 - 2) + (2k_2 - 1) + (2k_2 + k_3 - 2) + (2k_3 - 1) \\ &\geq (4k - 6) + (k_2 - k_1) + (k_2 - k_3) - 1 \geq 4k - 7.\end{aligned} \tag{3.2}$$

(d) Suppose $d = (d_2, d_1) > 1$ (the case $(d_2, d_3) > 1$ is similar). It is clear that $(d, d_3) = 1$ and therefore $|K_2 + K_3| \geq 2k_2 + k_3 - 2$. Moreover $|(\mathbb{K}_1 + \mathbb{K}_3) \backslash 2\mathbb{K}_2| \geq k_1$. Indeed, K_1 and $2K_2$ each lie in only one residue class modulo d and in K_3 there are at least two elements, say $x < y$, non-congruent modulo d. Thus, $(x + K_1) \cap 2K_2 = \varnothing$ or $(y + K_1) \cap 2K_2 = \varnothing$, which yields $|2\mathbb{K}_2 \cup (\mathbb{K}_1 + \mathbb{K}_3)| \geq 2k_2 - 1 + k_1$. In conclusion,

$$\begin{aligned}|2\mathbb{K}| &\geq (2k_1 - 1) + (k_1 + k_2 - 1) + (2k_2 - 1 + k_1) + (2k_2 + k_3 - 2) + (2k_3 - 1) \\ &= 4k_1 + 5k_2 + 3k_3 - 6 \geq 4k - 6.\end{aligned}$$

□

Lemma 3.5. — *Suppose that K_2 and K_3 lie each in only one residue class modulo d, with $d > 1$. Then $|2\mathbb{K}| > 3.5|\mathbb{K}| - 7$.*

Proof. — If $k_2 = 1$, we use Lemma 3.1. Suppose $k_2 \geq 2$. K_1 intersects at least two residue classes modulo d, because $\mathbb{K}$ is a reduced set. Thanks to Theorem D(3), we get

$$\begin{aligned}|2\mathbb{K}| &\geq (2k_1 - 1) + (k_1 + 2k_2 - 2) + (k_1 + 2k_3 - 2) + (k_2 + k_3 - 1) + (2k_3 - 1)\\ &= (4k_1 + 3k_2 + 5k_3 - 7). \end{aligned} \tag{3.3}$$

Take the arithmetic mean between inequalities (3.3) and (2.10). We get

$$|2\mathbb{K}| \geq 3.5k_1 + 3.5k_2 + 4k_3 - 6 > 3.5k - 7.$$

□

Next we discuss what happens if $d_2 > 1$. By the previous Lemma it is enough to study the case $k_3 \geq 2, k_1 \geq 2, d_{23} = d_{21} = 1$.

Lemma 3.6. — *Suppose that $d(K_2) > 1, (d_2, d_1) = (d_2, d_3) = 1$. Then $|2\mathbb{K}| \geq 4|\mathbb{K}| - 7$.*

Proof. — In view of Theorem D(3), we get

$$\begin{aligned}|2\mathbb{K}| \geq |2K_1| + |K_1 + K_2| + |K_1 + K_3| + |K_2 + K_3| + |2K_3| \geq (2k_1 - 1) +\\ + (k_1 + 2k_2 - 2) + (k_1 + k_3 - 1) + (k_3 + 2k_2 - 2) + (2k_3 - 1) = 4k - 7.\end{aligned}$$

□

Lemma 3.7. — *If $d_2 = 1, d_1 > 1, d_3 > 1$, then $|2\mathbb{K}| \geq 4|\mathbb{K}| - 7$.*

Proof. — We apply Theorem D(3) and we get

$$\begin{aligned}|2\mathbb{K}| &\geq |2K_1| + |K_1 + K_2| + |2K_2| + |K_2 + K_3| + |2K_3|\\ &\geq (2k_1 - 1) + (k_2 + 2k_1 - 2) + (2k_2 - 1) + (k_2 + 2k_3 - 2) + (2k_3 - 1) \geq 4k - 7.\end{aligned}$$

□

Conclusion. — Lemmas 3.1-3.7 and inequality $|2\mathbb{K}| < 3.5|\mathbb{K}| - 7$ ensure that $k_2 \geq 2$, $d_2 = d(K_2) = 1$. Indeed, if $d_2 > 1$, then Lemma 3.6 yields $(d_2, d_3) > 1$ or $(d_2, d_1) > 1$ and this leads to a contradiction, in view of Lemma 3.5. We obtained that $d_2 = 1$. By Lemma 3.7, d_1 and d_3 cannot be simultaneously greater than one. Suppose that $d_2 = d_3 = 1$, $d_1 > 1$. Lemma 3.3 shows that $k_1 \neq \max(k_1, k_2, k_3)$. Similarly, one has $k_3 \neq \max(k_1, k_2, k_3)$, if $d_2 = d_1 = 1$, $d_3 > 1$. In consequence, one of the following situations holds

$$(\alpha) \quad d_1 = d_2 = d_3 = 1, \ k_2 \geq 2, \tag{3.4}$$

$$(\beta) \quad d_2 = d_3 = 1, \ d_1 > 1, \ k_1 \neq \max(k_1, k_2, k_3), \ k_2 \geq 2, \tag{3.5}$$

$$(\gamma) \quad d_2 = d_1 = 1, \ d_3 > 1, \ k_3 \neq \max(k_1, k_2, k_3), \ k_2 \geq 2. \tag{3.6}$$

We end Section 3, by proving a lemma which will be used several times in the sequel.

Lemma 3.8. — *Suppose* $\max(h_1,h_2,h_3) \le \min(k_1,k_2,k_3) - 2$. *If*

$$|2\mathbb{K}| \le 4|\mathbb{K}| - 7, \tag{3.7}$$

then

(a) $|2\mathbb{K}| \ge (2|\mathbb{K}| - 1) + 2\ell_1 + 2\ell_3$ *and* $|2\mathbb{K}| \ge (2|\mathbb{K}| - 1) + \ell_1 + 2\ell_2 + \ell_3$,

(b) $|2\mathbb{K}| \ge (\frac{10}{3}|\mathbb{K}| - 5) + \frac{5}{3}H = \frac{5}{3}(|\mathbb{K}| + L)$,

(c) $|2\mathbb{K}| \ge (2|\mathbb{K}| - 1) + \frac{4}{3}L^*$.

Proof. — It is clear that $\ell_i \le 2k_i - 3, \max(\ell_i,\ell_j) \le k_i + k_j - 3$, for every $1 \le i,j \le 3$. Applying Theorem D(1) we obtain $|K_i + K_j| \ge k_i + k_j - 1 + \max(h_i,h_j)$. First, we estimate $|2\mathbb{K}|$ by using $2K_1,\ K_1+K_2,\ K_1+K_3,\ K_2+K_3,\ 2K_3$. We can write

$$\begin{aligned}
|2\mathbb{K}| &\ge (2k_1 - 1 + h_1) + (k_1 + k_2 - 1 + \max(h_1,h_2)) + (k_1 + k_3 - 1 + \max(h_1,h_3)) \\
&\qquad + (k_2 + k_3 - 1 + \max(h_2,h_3)) + (2k_3 - 1 + h_3) \\
&= 4k_1 + 2k_2 + 4k_3 - 5 + h_1 + h_3 \\
&\qquad + \max(h_1,h_2) + \max(h_2,h_3) + \max(h_3,h_1) \\
&\ge \begin{cases} 4k_1 + 2k_2 + 4k_3 - 5 + H + 2\max(h_1,h_2,h_3), & \text{if } h_2 \ne \max(h_1,h_2,h_3). \\ 4k_1 + 2k_2 + 4k_3 - 5 + 2H - \min(h_1,h_2,h_3), & \text{if } h_2 = \max(h_1,h_2,h_3). \end{cases} \\
&\ge 4k_1 + 2k_2 + 4k_3 - 5 + \frac{5}{3}H.
\end{aligned} \tag{3.8}$$

Thus, $|2\mathbb{K}| \ge (2|\mathbb{K}|-1)+2\ell_1+2\ell_3$. Moreover, inequality (b) is also true, if $k \ge 3k_2$. Second, using $2K_2$ instead of $K_1 + K_3$ we get

$$\begin{aligned}
|2\mathbb{K}| &\ge (2k_1 - 1 + h_1) + (k_1 + k_2 - 1 + \max(h_1,h_2)) + (2k_2 - 1 + h_2) \\
&\qquad + (k_2 + k_3 - 1 + \max(h_2,h_3)) + (2k_3 - 1 + h_3) \\
&= 3k_1 + 4k_2 + 3k_3 - 5 + h_1 + h_2 + h_3 + \max(h_1,h_2) + \max(h_2,h_3) \\
&\ge \begin{cases} 3k_1 + 4k_2 + 3k_3 - 5 + H + 2\max(h_1,h_2,h_3), & \text{if } h_2 = \max(h_1,h_2,h_3). \\ 3k_1 + 4k_2 + 3k_3 - 5 + 2H - \min(h_1,h_2,h_3), & \text{if } h_2 \ne \max(h_1,h_2,h_3). \end{cases} \\
&\ge 3k_1 + 4k_2 + 3k_3 - 5 + \frac{5}{3}H.
\end{aligned}$$

Thus, $|2\mathbb{K}| \ge (2|\mathbb{K}|-1)+\ell_1+2\ell_2+\ell_3$. Moreover, inequality (b) is also true, if $k \le 3k_2$. We prove now inequality (c).

First Case. — Suppose $[m_2, M_2] \supseteq \left[\frac{1}{2}(m_1 + m_3), \frac{1}{2}(M_1 + M_3)\right]$.

It is clear that $L^* = L = \ell_1 + \ell_2 + \ell_3$ and in this case inequality (c) follows from (a), in view of $2\ell_2 \ge \ell_1 + \ell_3$. We could have used (b). Indeed,

$$|2\mathbb{K}| \ge (\frac{10}{3}|\mathbb{K}| - 5) + \frac{5}{3}H \ge (\frac{10}{3}|\mathbb{K}| - 5) + \frac{4}{3}H = (2|\mathbb{K}| - 1) + \frac{4}{3}L = (2|\mathbb{K}| - 1) + \frac{4}{3}L^*.$$

Second Case. — Suppose $[m_2, M_2] \subseteq \left[\frac{1}{2}(m_1+m_3), \frac{1}{2}(M_1+M_3)\right]$.

It is clear that

$$L^* = \ell_1 + \frac{1}{2}(\ell_1+\ell_3) + \ell_3 = \frac{3}{2}(\ell_1+\ell_3),\ \frac{4}{3}L^* = 2\ell_1 + 2\ell_3.$$

Inequality (c) follows from (a). Actually, (3.8) shows that a sharper inequality is true:

$$\begin{aligned}|2\mathbb{K}| &\geq 4k_1 + 2k_2 + 4k_3 - 5 + 2h_1 + 2h_3 + \max(h_1,h_2,h_3)\\ &= (2k-1) + 2\ell_1 + 2\ell_3 + \max(h_1,h_2,h_3)\\ &\geq (2k-1) + 2\ell_1 + 2\ell_3 = (2k-1) + \frac{4}{3}L^*.\end{aligned}$$

Third Case. — Suppose $m_2 < \frac{1}{2}(m_1+m_3) \leq M_2 < \frac{1}{2}(M_1+M_3)$. Put

$$\delta = \frac{m_1+m_3}{2} - m_2.$$

Define

$$K_2^- = K_2 \cap \left[m_2, \frac{m_1+m_3}{2}\right),\quad k_2^- = |K_2^-|. \tag{3.9}$$

We improve inequality (3.8) by taking into account

$$|2\mathbb{K}_2 \backslash (\mathbb{K}_1 + \mathbb{K}_3)| \geq |2K_2^-| \geq (2k_2^- - 1).$$

One has $k_2^- \geq \delta - h_2$ and therefore (3.8) shows that

$$\begin{aligned}|2\mathbb{K}| &\geq \Big(|2\mathbb{K}_{13}| + |K_2 + K_1| + |K_2 + K_3|\Big) + |2\mathbb{K}_2 \backslash (\mathbb{K}_1 + \mathbb{K}_3)|\\ &\geq \Big(4k_1 + 2k_2 + 4k_3 - 5 + h_1 + h_3 + \max(h_1,h_3) + 2h_2\Big) + (2k_2^- - 1)\\ &\geq 4k_1 + 2k_2 + 4k_3 - 6 + \frac{3}{2}(h_1+h_3) + 2\delta. \end{aligned}\tag{3.10}$$

If $\frac{2}{3}\delta \geq \frac{h_1+h_3}{2} + 1$, then inequality (c) is proved, because (3.10) ensures

$$\begin{aligned}|2\mathbb{K}| &\geq (4k_1 + 2k_2 + 4k_3 - 5) + 2h_1 + 2h_3 + \frac{4}{3}\delta\\ &= (2k-1) + 2\ell_1 + 2\ell_3 + \frac{4}{3}\delta\\ &= (2k-1) + \frac{4}{3}\Big(\ell_1 + \frac{1}{2}(\ell_1+\ell_3) + \delta + \ell_3\Big)\\ &= (2k-1) + \frac{4}{3}L^*.\end{aligned}$$

Now we may suppose that $\frac{2}{3}\delta < \frac{h_1+h_3}{2} + 1$. First of all, note that

$$2\delta - 1 \leq k_2 - 1 \leq \ell_2. \tag{3.11}$$

Indeed, in view of (3.10) one has

$$|2\mathbb{K}| > (4k_1 + 2k_2 + 4k_3 - 6) + 3(\frac{2}{3}\delta - 1) + 2\delta = (4k_1 + 2k_2 + 4k_3 - 9) + 4\delta \tag{3.12}$$

and if $2\delta \geq k_2 + 1$ we would obtain $|2\mathbb{K}| > 4k - 7$, in contradiction with (3.7).

We estimate now $|2\mathbb{K}_2\backslash(\mathbb{K}_1+\mathbb{K}_3)|$. It is clear that $m_2+(K_2\cap[m_2,m_1+m_3-m_2))$ is included in $2\mathbb{K}_2\backslash(\mathbb{K}_1+\mathbb{K}_3)$. The length of $[m_2,m_1+m_3-m_2)$ is exactly 2δ and in view of inequality (3.11) we obtain $|K_2\cap[m_2,m_1+m_3-m_2)|\geq 2\delta-h_2$. Therefore

$$|2\mathbb{K}_2\backslash(\mathbb{K}_1+\mathbb{K}_3)|\geq 2\delta-h_2. \tag{3.13}$$

As in (3.10), we improve inequality (3.8) by taking into account $2\mathbb{K}_2\backslash(\mathbb{K}_1+\mathbb{K}_3)$. One has

$$\begin{aligned}|2\mathbb{K}| &\geq |2\mathbb{K}_{13}|+|K_2+K_1|+|K_2+K_3|+|2\mathbb{K}_2\backslash(\mathbb{K}_1+\mathbb{K}_3)|\\ &\geq (4k_1+2k_2+4k_3-5+2h_1+2h_3+h_2)+(2\delta-h_2)\\ &\geq (2k-1)+2\ell_1+2\ell_3+\frac{4}{3}\delta=(2k-1)+\frac{4}{3}L^*.\end{aligned}$$

□

In the remaining part of the proof, we shall distinguish *three main cases* according to $k_2=\min(k_1,k_2,k_3)$, $k_2=\max(k_1,k_2,k_3)$, $k_3<k_2<k_1$.

4. First Case : $k_2=\min(k_1,k_2,k_3)$

Theorem 4.1. — *Suppose $k_2\leq k_3\leq k_1$. If $|\mathbb{K}+\mathbb{K}|<3.5|\mathbb{K}|-7$, then $k_2\geq 2$ and*

(i) $d_1=d_2=d_3=1$ *and* $\max(h_1,h_2,h_3)\leq k_2-2$.

(ii) $|2\mathbb{K}|\geq\left(\frac{10}{3}|\mathbb{K}|-5\right)+\frac{5}{3}H=\frac{5}{3}(|\mathbb{K}|+L)$.

(iii) $|2\mathbb{K}|\geq(2|\mathbb{K}|-1)+\frac{4}{3}L^*$.

Proof. — In view of $k_2\leq k_3\leq k_1$, it is enough to prove only (i), because assertions (ii) and (iii) are direct consequences of Lemma 3.8 and inequality $\max(h_1,h_2,h_3)\leq k_2-2$. Using $k_1=\max(k_1,k_2,k_3)$, equations (3.4), (3.5), (3.6) and the small doubling hypothesis we deduce that $k_2\geq 2$, $d_1=1$, $d_2=1, d_3\geq 1$.

1. *We show that* $d_3=1$.

Suppose that $d_3>1$. By Theorem D(3) we have $|2\mathbb{K}|\geq(2k_1-1)+(k_1+k_2-1)+(k_1+2k_3-2)+(k_2+2k_3-2)+(2k_3-1)=4k-6+2(k_3-k_2)-1\geq 4k-7$.

2. *We show that* $\max(h_1,h_2,h_3)\leq k_2-2$.

Suppose that $\max(h_2,h_i)\geq k_2-1$, for $i=1$ or $i=3$. Using Theorem D, one has $|K_2+K_i|\geq\min\left(k_i+2k_2-3,k_i+k_2-1+\max(h_i,h_2)\right)\geq(k_i+k_2-1)+(k_2-2)$. In consequence, we improve (2.11) to $|2\mathbb{K}|\geq 4k_1+3k_2+4k_3-7$. Take the arithmetic mean between (2.10) and the previous inequality. We obtain $|2\mathbb{K}|\geq 3.5k-6>3.5k-7$, which contradicts (4.1). □

5. Second Case : $k_2=\max(k_1,k_2,k_3)$

Theorem 5.1. — *Suppose $k_3\leq k_1\leq k_2$ and $|2\mathbb{K}|<3.5|\mathbb{K}|-7$. Then, $k_3\geq 2$ and*

(a) $\max(h_1,h_2,h_3)\leq k_3-2$.

(b) $|2\mathbb{K}|\geq(\frac{10}{3}|\mathbb{K}|-5)+\frac{5}{3}H=\frac{5}{3}(|\mathbb{K}|+L)$.

(c) $|2\mathbb{K}|\geq(2|\mathbb{K}|-1)+\frac{4}{3}L^*$.

Proof. — Inequalities (b) and (c) follow from Lemma 3.8 and assertion (a). Therefore, we need to prove only $\max(h_1, h_2, h_3) \le k_3 - 2$. Lemma 3.4 implies that $d_2 = 1$.

1. *We show that* $d_1 = 1$.

Suppose $d_1 \ge 2$ and prove that $|2\mathbb{K}| > 3.5k - 7$, which contradicts our small doubling hypothesis. If $k_3 = 1$, then K_1 and K_3 lie each in only one residue class modulo d_1 and we use Lemma 3.2. If $k_3 \ge 2$ and $d_3 \ge 2$, we apply Lemma 3.7. We may assume now that $k_3 \ge 2$, $d_2 = d_3 = 1$ and estimate $|K_1 + K_2|$ by Theorem D(3):

$$\begin{aligned}|2\mathbb{K}| &\ge |2K_1| + |K_1 + K_2| + |2K_2| + |K_2 + K_3| + |2K_3| \\ &\ge (2k_1 - 1) + (k_2 + 2k_1 - 2) + (2k_2 - 1) + (k_2 + k_3 - 1) + (2k_3 - 1) \\ &\ge 4k_1 + 4k_2 + 3k_3 - 6 = (4k - 6) - k_3 \ge \frac{11}{3}k - 6 > 3.5k - 7.\end{aligned}$$

2. *We show that* $\max(h_1, h_2, h_3) \le k_3 - 2$.

Suppose that $\max(h_1, h_2, h_3) \ge k_3 - 1$. We use this inequality in order to improve (2.10) by $k_3 - 2$ and thus obtain

$$|2\mathbb{K}| \ge 3k_1 + 4k_2 + 4k_3 - 7 = (4k - 7) - k_1 \ge 3.5k - 7 + \frac{k_3}{2} > 3.5k - 7, \qquad (5.1)$$

in contradiction with the small doubling hypothesis.

If $k_3 = 1$ holds, then clearly (5.1) is true. Suppose $k_3 \ge 2$.

(i) If $h_2 \ge k_3 - 1$, then $|K_2 + K_3| \ge (k_2 + k_3 - 1) + (k_3 - 2)$, by using Theorem D(2).

(ii) If $h_1 \ge k_3 - 1$, then $|2K_1| \ge (2k_1 - 1) + \min(h_1, k_1 - 2) \ge (2k_1 - 1) + (k_3 - 2)$, thanks to Theorem D.

(iii) If $h_3 \ge k_3 - 1, d_3 > 1$, then $|K_2 + K_3| \ge (k_2 + k_3 - 1) + (k_3 - 1)$, by Theorem D(3). Finally, if $h_3 \ge k_3 - 1$, $d_3 = 1$, then $|2K_3| \ge (2k_3 - 1) + (k_3 - 2)$, due to Theorem D(2). The proof of Theorem 5.1 is now complete. □

6. Third Case : $k_3 \le k_2 \le k_1$

Theorem 6.1. — *Suppose* $k_3 \le k_2 \le k_1$ *and* $|2\mathbb{K}| < 3.5|\mathbb{K}| - 7$. *Then,*

$$|2\mathbb{K}| \ge (2|\mathbb{K}| - 1) + \frac{4}{3}L^*.$$

This theorem is a consequence of lemmas 6.1, 6.2 and 6.3.

Lemma 6.1. — *Suppose* $k_3 \le k_2 \le k_1$, $\max(h_1, h_2) \ge k_2 - 1$. *Then* $|2\mathbb{K}| \ge 3.5|\mathbb{K}| - 7$.

Proof. — Using Lemma 3.3, we deduce that $d_1 = 1$. We estimate $|K_1 + K_2|$ by Theorem D(1),(2). One has

$$|K_1 + K_2| \ge \min\Big(k_1 + k_2 - 1 + \max(h_1, h_2), k_1 + 2k_2 - 3\Big) \ge (k_1 + k_2 - 1) + (k_2 - 3).$$

Inequality (2.11) becomes $|2\mathbb{K}| \ge 4k_1 + 3k_2 + 4k_3 - 7$. Taking the arithmetic mean between this inequality and (2.10) we get $|2\mathbb{K}| \ge 3.5k - 6 > 3.5k - 7$. □

Lemma 6.2. — *Suppose* $k_3 \le k_2 \le k_1$, $\max(h_1, h_2) \le k_2 - 2$ *and* $\ell_2 \ge \ell_1$ *or* $\ell_3 \ge \ell_2$. *If* $|2\mathbb{K}| < 3.5|\mathbb{K}| - 7$, *then* $\max(h_1, h_2, h_3) \le k_3 - 2$ *and* $|2\mathbb{K}| \ge (2k - 1) + \frac{4}{3}L^*$.

Proof. — We shall show that $|2\mathbb{K}| \geq 4k_1 + 3k_2 + 3k_3 - 5 + H$, which yields $|2\mathbb{K}| \geq 3.5k - 7 + H + 2 - \frac{1}{2}k_3$. This proves $\max(h_1, h_2, h_3) \leq H < \frac{1}{2}k_3 - 2$ and Lemma 6.2 follows thanks to Lemma 3.8.

(i) Assume first that $\ell_2 \geq \ell_1$. We get $|K_1 + K_2| \geq (2k_1 - 1)$ and thus

$$\begin{aligned} |2(\mathbb{K}_2 \cup \mathbb{K}_3)| &\leq |2\mathbb{K}| - (|2K_1| + |K_1 + K_2|) \\ &\leq |2\mathbb{K}| - (2k_1 - 1) - (2k_1 - 1) \\ &= |2\mathbb{K}| - (4k_1 - 2) \\ &\leq 4k_2 + 4k_3 - 7, \end{aligned}$$

in view of the small doubling property of $\mathbb{K}$. Since $d_2 = 1$, Theorem S yields

$$|2(\mathbb{K}_2 \cup \mathbb{K}_3)| \geq 3k_2 + 3k_3 - 3 + h_2 + h_3,$$

and thus

$$\begin{aligned} |2\mathbb{K}| &\geq (2k_1 - 1 + h_1) + (k_1 + k_2 - 1 + \max(h_1, h_2)) + (3k_2 + 3k_3 - 3 + h_2 + h_3) \\ &\geq (3k_1 + 4k_2 + 3k_3 - 5) + h_1 + 2h_2 + h_3 \\ &= (3k - 4) + H + \ell_2 \\ &\geq (4k_1 + 3k_2 + 3k_3 - 5) + H. \end{aligned}$$

(ii) Assume $\ell_3 \geq \ell_2$. We get $|K_2 + K_3| \geq 2k_2 - 1$ and thus

$$\begin{aligned} |2(\mathbb{K}_1 \cup \mathbb{K}_3)| &\leq |2\mathbb{K}| - (|K_1 + K_2| + |K_2 + K_3|) \\ &\leq |2\mathbb{K}| - (k_1 + k_2 - 1) - (2k_2 - 1) \\ &\leq |2\mathbb{K}| - (4k_2 - 2) \\ &\leq 4k_1 + 4k_3 - 7, \end{aligned}$$

by the small doubling property of $\mathbb{K}$. Since $d_1 = 1$, Theorem S gives $|2(\mathbb{K}_1 \cup \mathbb{K}_3)| \geq 3k_1 + 3k_3 - 3 + h_1 + h_3$ and thus

$$\begin{aligned} |2\mathbb{K}| &\geq |2(\mathbb{K}_1 \cup \mathbb{K}_3)| + |K_1 + K_2| + |K_2 + K_3| \geq \\ &\geq (3k_1 + 3k_3 - 3 + h_1 + h_3) + (k_1 + k_2 - 1 + \max(h_1, h_2)) + (2k_2 - 1) \\ &\geq (4k_1 + 3k_2 + 3k_3 - 5) + H. \end{aligned}$$

□

Lemma 6.3. — *Suppose* $k_3 \leq k_2 \leq k_1$, $\max(h_1, h_2) \leq k_2 - 2$ *and* $\ell_3 \leq \ell_2 \leq \ell_1$. *If* $|2\mathbb{K}| < 3.5|\mathbb{K}| - 7$, *then* $|2\mathbb{K}| \geq (2k - 1) + \frac{4}{3}L^*$.

Proof. — **(I)** We begin the proof by obtaining an upper bound for ℓ_3, see (6.6), and by showing in (6.7), (6.8) that we may estimate $|2(\mathbb{K}_2 \cup \mathbb{K}_3)|$ and $|2(\mathbb{K}_1 \cup \mathbb{K}_3)|$ by using Theorem S(1). In the same time, we shall obtain (6.12) below, an inequality which will be used several times in the proof.

The hypothesis $\max(h_1, h_2) \leq k_2 - 2$ ensures that

$$\ell_1 \leq k_1 + k_2 - 3 \leq 2k_1 - 3, \ d_1 = 1, \tag{6.1}$$

$$\ell_2 \leq 2k_2 - 3 \leq k_1 + k_2 - 3, \ d_2 = 1. \tag{6.2}$$

Note that $\ell_3 \leq \ell_i$, for $i = 1$ and $i = 2$ give

$$\begin{aligned}|K_i + K_3| &\geq |m_3 + K_i| + |M_3 + \Big(K_i \cap (M_i - \ell_3, M_i]\Big)| \\ &\geq k_i + \ell_3 - h_i = (k_i + k_3 - 1) + h_3 - h_i. \qquad (6.3)\end{aligned}$$

Applying (6.3) and Theorem D(1) for $|2K_1|, |K_1 + K_2|$, inequalities (6.1), (6.2) yield

$$\begin{aligned}|2\mathbb{K}| &\geq |2K_1| + |K_1 + K_2| + |K_1 + K_3| + |K_2 + K_3| + |2K_3| \\ &\geq (2k_1 - 1 + h_1) + (k_1 + k_2 - 1 + \max(h_1, h_2)) + ((k_1 + k_3 - 1) \\ &\quad + \max(0, h_3 - h_1)) + ((k_2 + k_3 - 1) + \max(0, h_3 - h_2)) + (2k_3 - 1),\end{aligned}$$

and thus

$$|2\mathbb{K}| \geq (4k_1 + 2k_2 + 4k_3 - 5) + 2h_3, \qquad (6.4)$$

$$|2\mathbb{K}| \geq (4k_1 + 2k_2 + 4k_3 - 5) + \max(h_1, h_2) + h_3. \qquad (6.5)$$

We claim that

$$h_3 \leq k_2 - 2, \ \ell_3 \leq k_2 + k_3 - 3. \qquad (6.6)$$

On the contrary, suppose that $h_3 \geq k_2 - 1$. Inequality (6.4) gives $|2\mathbb{K}| \geq 4k - 7 > 3.5k - 7$, a contradiction. By a similar argument, the small doubling property and inequality (6.5) lead to

$$h_2 + h_3 \leq k_2 + k_3 - 3, \ h_1 + h_3 \leq k_1 + k_3 - 3, \qquad (6.7)$$

which shows that

$$|2(\mathbb{K}_2 \cup \mathbb{K}_3)| \geq 3k_2 + 3k_3 - 3 + h_2 + h_3, \quad |2(\mathbb{K}_1 \cup \mathbb{K}_3)| \geq 3k_1 + 3k_3 - 3 + h_1 + h_3, \ (6.8)$$

in view of $d_1 = d_2 = 1$ and Theorem S(1). We are now able to deduce

$$\begin{aligned}|2\mathbb{K}| &\geq |2K_1| + |K_1 + K_2| + |2(\mathbb{K}_2 \cup \mathbb{K}_3)| \\ &\geq (2k_1 - 1 + h_1) + (k_1 + k_2 - 1 + \max(h_1, h_2)) + (3k_2 + 3k_3 - 3 + h_2 + h_3) \\ &= (3k_1 + 4k_2 + 3k_3 - 5) + h_1 + \max(h_1, h_2) + h_2 + h_3. \qquad (6.9)\end{aligned}$$

If $\max(h_1, h_2, h_3) \leq k_3 - 2$, Lemma 6.3 follows from Lemma 3.8. Therefore, we have to examine only the case

$$\max(h_1, h_2, h_3) \geq k_3 - 1. \qquad (6.10)$$

(II) *We prove (6.12), inequality which will be repeatedly used.*

In order to obtain (6.12), we need one more lower bound for $|2\mathbb{K}|$ (see (6.11) below). We use (6.10) and consider two cases :

(a) On the one hand, if $\max(h_2, h_3) \geq k_3 - 1$, then $|K_2 + K_3| \geq k_2 + 2k_3 - 2$. Indeed, if $\ell_2 \geq k_2 + k_3 - 2$, then $\varepsilon_{23} = 0$ thanks to (6.6); using Theorem D(1),(2) we get $|K_2 + K_3| \geq \min(k_2 + 2k_3 - 2, k_2 + k_3 - 1 + \max(h_2, h_3)) \geq k_2 + 2k_3 - 2$. If $\ell_2 \leq k_2 + k_3 - 3$, then $|K_2 + K_3| \geq k_2 + k_3 - 1 + \max(h_2, h_3) \geq k_2 + 2k_3 - 2$, by Theorem D(1). Therefore, in each of these two cases, $k_2 + 2k_3 - 2$ is a lower bound for $|K_2 + K_3|$. We may estimate $|K_2 + K_1|$ by Theorem D(1), because of (6.1) and (6.2). Finally, in view of (6.8) one has

$$\begin{aligned}|2\mathbb{K}| &\geq |2(\mathbb{K}_1 \cup \mathbb{K}_3)| + |K_2 + K_1| + |K_2 + K_3| \\ &\geq (3k_1 + 3k_3 - 3 + h_1 + h_3) + (k_1 + k_2 - 1 + \max(h_1, h_2)) + (k_2 + 2k_3 - 2) \\ &\geq (4k_1 + 2k_2 + 4k_3 - 6) + h_1 + \max(h_1, h_2) + h_3 + k_3.\end{aligned}$$

(b) On the other hand, if $\max(h_2, h_3) \leq k_3 - 2$ and $h_1 \geq k_3 - 1$, then we estimate $|2K_1|, |K_1 + K_2|, |K_2 + K_3|, |2K_3|$ by Theorem D(1) and $|K_1 + K_3|$ by Theorem D(2), for $\varepsilon_{13} = 0$. We get

$$\begin{aligned}|2\mathbb{K}| &\geq |2K_1| + |K_1 + K_2| + |K_1 + K_3| + |K_2 + K_3| + |2K_3| \\ &\geq (2k_1 - 1 + h_1) + (k_1 + k_2 - 1 + \max(h_1, h_2)) + (k_1 + 2k_3 - 2) + \\ &\qquad + (k_2 + k_3 - 1 + \max(h_2, h_3)) + (2k_3 - 1 + h_3) \geq \\ &\geq (4k_1 + 2k_2 + 4k_3 - 6) + h_1 + \max(h_1, h_2) + \max(h_2, h_3) + h_3 + k_3.\end{aligned}$$

Thus, in both cases (a) and (b), we obtain

$$|2\mathbb{K}| \geq (4k_1 + 2k_2 + 4k_3 - 6) + h_1 + \max(h_1, h_2) + h_3 + k_3. \tag{6.11}$$

Taking the arithmetic mean between (6.9) and (6.11) we obtain

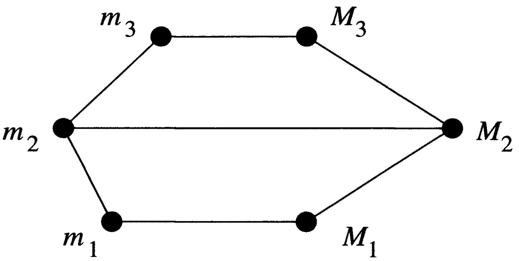
$$|2\mathbb{K}| \geq (3.5k - 7) + \frac{1}{2}\Big(2h_1 + 2\max(h_1, h_2) + h_2 + 2h_3 + k_3 + 3 - k_2\Big).$$

Applying the small doubling property, we deduce immediately that

$$2h_1 + 2\max(h_1, h_2) + h_2 + 2h_3 + k_3 + 4 \leq k_2. \tag{6.12}$$

As in Lemma 3.8, we shall distinguish at this point three situations, depending on the relative position of $[m_2, M_2]$ and $[\frac{m_1+m_3}{2}, \frac{M_1+M_3}{2}]$.

First Case. — $m_2 \leq \frac{m_1+m_3}{2} \leq \frac{M_1+M_3}{2} \leq M_2$.

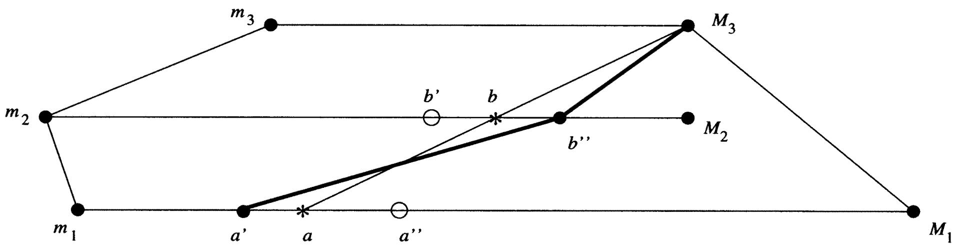


We already proved in (6.9) that $|2\mathbb{K}| \geq (2k - 1) + \ell_1 + 2\ell_2 + \ell_3$, and this implies $|2\mathbb{K}| \geq (2k - 1) + \frac{4}{3}L^*$, in view of

$$\ell_1 + 2\ell_2 + \ell_3 - \frac{4}{3}L^* = (\ell_1 + 2\ell_2 + \ell_3) - \frac{4}{3}(\ell_1 + \ell_2 + \ell_3) = \frac{2}{3}\Big(\ell_2 - \frac{\ell_1 + \ell_3}{2}\Big) \geq 0.$$

□

Second Case. — $m_2 \le \frac{m_1+m_3}{2} \le M_2 \le \frac{M_1+M_3}{2}$.



As usual, put $\delta = \frac{m_1+m_3}{2} - m_2 \ge 0$.

(i) *We split* $\mathbb{K}$ *into two subsets,* $\mathbb{K}'$ *and* $\mathbb{K}''$ *and get a lower estimate of* $|2\mathbb{K}|$ *by adding* $|2\mathbb{K}'|$ *and* $|2\mathbb{K}''|$.

Let us take a line l which intersects $(x_2 = 0)$ at $(0, a)$, $(x_2 = 1)$ at $(1, b)$, $(x_2 = 2)$ at $(2, M_3)$. In the sequel we shall prove that we may choose l such that

$$m_1 \le a \le M_1 \text{ and } \frac{1}{2}(m_1 + m_3) + \frac{\ell_3}{2} \le b \le M_2. \tag{6.13}$$

Take $b' \le b''$ two consecutive elements of K_2 such that $b' \le b \le b''$. Take $a' \le a''$ two consecutive elements of K_1 such that $a' < a \le a''$. Define

$$K_1' = K_1 \cap [m_1, a'], \quad K_2' = K_2 \cap [m_2, b''], \quad K_3' = K_3, \tag{6.14}$$
$$K_1'' = K_1 \cap [a', M_1], \quad K_2'' = K_2 \cap [b'', M_2], \quad K_3'' = \{M_3\}. \tag{6.15}$$

It will be shown in step (v) bellow, that we may choose a and b such that

$$\ell_1' \le 2k_1' - 3, \ \max(\ell_1', \ell_2') \le k_1' + k_2' - 3, \ \ell_2' + \ell_3' \le 2k_2' + 2k_3' - 5, \tag{6.16}$$
$$\ell_1'' \le 2k_1'' - 3, \ \max(\ell_1'', \ell_2'') \le k_1'' + k_2'' - 3. \tag{6.17}$$

Now Lemma 6.3 follows easily. Indeed, we estimate $|2\mathbb{K}|$ by adding $|2\mathbb{K}'|$ and $|2\mathbb{K}''|$ and paying attention to the points counted twice:

$$2\mathbb{K}_2' \cap (\mathbb{K}_1'' + \mathbb{K}_3'') \quad \text{and} \quad 2a', \ a' + b'', \ b'' + M_3, \ 2M_3.$$

Thus, if we denote by $x = |2\mathbb{K}_2' \cap (\mathbb{K}_1'' + \mathbb{K}_3'')|$, then

$$\begin{aligned}|2\mathbb{K}| &\geq |2\mathbb{K}'| + |2\mathbb{K}''| - \Big(4 + |2\mathbb{K}_2' \cap (\mathbb{K}_1'' + \mathbb{K}_3'')\Big)| \\ &\geq |2K_1'| + |K_1' + K_2'| + |2(\mathbb{K}_2' \cup \mathbb{K}_3')| \\ &\qquad + |2K_1''| + |K_1'' + K_2''| + |K_1'' + K_3''| + |K_2'' + K_3''| + |2K_3''| - (4 + x) \\ &\geq \Big[(\ell_1' + k_1') + (\ell_2' + k_1') + (2k_2' + 2k_3' - 1 + \ell_2' + \ell_3')\Big] \\ &\qquad + \Big[(\ell_1'' + k_1'') + (\ell_1'' + k_2'') + k_1'' + k_2'' + 1\Big] - 4 - x \\ &= \Big[2(k_1' + k_2' + k_3') - 1 + \ell_1' + 2\ell_2' + \ell_3'\Big] \\ &\qquad + \Big[2(k_1'' + k_2'' + k_3'') - 1 + 2\ell_1''\Big] - 4 - x \\ &= 2(k_1' + k_1'' + k_2' + k_2'' + k_3' + k_3'') - 2 + \ell_1' + 2\ell_2' + \ell_3' + 2\ell_1'' - 4 - x \\ &= 2k + \ell_1' + 2\ell_2' + \ell_3' + 2\ell_1'' - x. \end{aligned} \tag{6.18}$$

In the last equality we used $k_i' + k_i'' = k_i + 1$, for $1 \leq i \leq 3$. It is clear that

$$x = |2\mathbb{K}_2' \cap (\mathbb{K}_1'' + \mathbb{K}_3'')| \leq 1 + |[2b, 2b'']| = 2 + 2(b'' - b). \tag{6.19}$$

In view of the collinearity of a and b we have $(a - m_1) + \ell_3 = 2b - (m_1 + m_3)$ and thus

$$\begin{aligned}\ell_1' + 2\ell_2' + \ell_3' + 2\ell_1'' &= \ell_1' + 2\Big((b'' - b) + (b - \frac{m_1 + m_3}{2}) + (\frac{m_1 + m_3}{2} - m_2)\Big) \\ &\qquad + \ell_3' + 2\ell_1'' \\ &= \ell_1' + 2(b'' - b) + 2\Big(b - \frac{m_1 + m_3}{2}\Big) + 2\Big(\frac{m_1 + m_3}{2} - m_2\Big) \\ &\qquad + \ell_3' + 2\ell_1'' \\ &= \ell_1' + 2(b'' - b) + \Big((a - m_1) + \ell_3\Big) + 2\delta + \ell_3' + 2\ell_1'' \\ &= \ell_1' + 2(b'' - b) + \Big(\ell_1' + (a - a') + \ell_3\Big) + 2\delta + \ell_3' + 2\ell_1'' \\ &= 2(\ell_1' + \ell_1'') + (\ell_3 + \ell_3') + 2\delta + 2(b'' - b) + (a - a') \\ &= 2\ell_1 + 2\ell_3 + 2\delta + 2(b'' - b) + (a - a'). \end{aligned} \tag{6.20}$$

Thus, using (6.19), (6.20) in (6.18), we conclude that

$$|2\mathbb{K}| \geq 2k - 2 + 2\ell_1 + 2\ell_3 + 2\delta + (a - a') \geq (2k - 1) + 2\ell_1 + 2\ell_3 + \frac{4}{3}\delta.$$

(ii) *We put inequalities (6.16), (6.17) in a slightly different form:*

$$h_1' \leq k_1' - 2,\ h_1' \leq k_2' - 2,\ h_2' \leq k_1' - 2,\ h_2' + h_3' \leq k_2' + k_3' - 3 = k_2' + k_3 - 3,$$
$$h_1'' \leq k_1'' - 2,\ h_1'' \leq k_2'' - 2,\ h_2'' \leq k_1'' - 2.$$

Consequently, it is enough to choose a and b such that (6.13) and the following four inequalities are true:

$$k_1'' \geq \max(h_1, h_2) + 2, \ k_2'' \geq h_1 + 2, \tag{6.21}$$

$$k_1' \geq \max(h_1, h_2) + 2, \ k_2' \geq \max(h_1 + 2, h_2 + h_3 - k_3 + 3). \tag{6.22}$$

(iii) *We define now the line l.*

To define l, we need only to choose b as

$$b = \min\left\{M_2 - (h_1 + h_2 + 2), \frac{M_1 + M_3}{2} - \frac{\max(h_1, h_2) + h_1 + 1}{2}\right\}. \tag{6.23}$$

Using the collinearity condition a is defined by

$$(a - m_1) + \ell_3 = 2b - (m_1 + m_3). \tag{6.24}$$

We shall show in step **(v)** that this choice ensures (6.13), (6.21) and (6.22). But first, we need some more estimates.

(iv) *We estimate δ and compare h_1, h_2, h_3 to k_1.*

(a) We prove that

$$2\delta + 2h_1 + 2\max(h_1, h_2) + 2h_3 + k_3 + 3 \leq k_2 + h_2. \tag{6.25}$$

Improve (6.11) by taking into account $|2\mathbb{K}_2 \backslash (\mathbb{K}_1 + \mathbb{K}_3)| \geq 2(\delta - h_2) - 1$. We get

$$|2\mathbb{K}| \geq (4k_1 + 2k_2 + 4k_3 - 7) + h_1 + \max(h_1, h_2) + h_3 + k_3 + 2\delta - 2h_2. \tag{6.26}$$

We take the arithmetic mean between (6.26) and (6.9) and obtain

$$|2\mathbb{K}| \geq (3.5k - 7) - \frac{1}{2}(k_2 + h_2) + \frac{1}{2}\Big(2h_1 + 2\max(h_1, h_2) + 2h_3 + k_3 + 2\delta + 2\Big).$$

In view of the small doubling property, we deduce that (6.25) holds.

(b) We prove that

$$3h_1 + 4\max(h_1, h_2) + 2h_2 + 3h_3 + k_3 + 8 \leq k_1. \tag{6.27}$$

Remark that in the Second Case, one has $\frac{\ell_1 + \ell_3}{2} > \ell_2 - \delta$. This gives $2\delta > 2\ell_2 - (\ell_1 + \ell_3)$. Thanks to (6.25), the last inequality implies

$$\Big(2\ell_2 - (\ell_1 + \ell_3)\Big) + 2h_1 + 2\max(h_1, h_2) + 2h_3 + k_3 + 3 < k_2 + h_2,$$

$$(2k_2 - k_1 - k_3) + (2h_2 - h_1 - h_3) + 2h_1 + 2\max(h_1, h_2) + 2h_3 + k_3 + 3 < k_2 + h_2,$$

$$k_2 + h_2 + h_1 + 2\max(h_1, h_2) + h_3 + 4 \leq k_1. \tag{6.28}$$

Combining inequalities (6.12) and (6.28), we obtain the desired result (6.27).

(v) *We prove (6.13), (6.21) and (6.22).*

We begin with (6.13). In view of the collinearity condition (6.24), we shall prove now $\frac{m_1 + m_3}{2} + \frac{\ell_3}{2} \leq b \leq M_2$, which ensures that $m_1 \leq a \leq M_1$. Since b is defined by

(6.23), we have actually to check the two inequalities stated below.

$$\begin{aligned}(1)\quad \Big(M_2-(h_1+h_2+2)\Big)-\Big(\frac{m_1+m_3}{2}+\frac{\ell_3}{2}\Big)\\ &=(M_2-\frac{m_1+m_3}{2})-(h_1+h_2+2+\frac{\ell_3}{2})\\ &=(\ell_2-\delta)-(h_1+h_2+\frac{k_3}{2}+\frac{h_3}{2}+1.5)\\ &=(k_2+h_2)-(\delta+h_1+h_2+\frac{k_3}{2}+\frac{h_3}{2}+2.5)\geq 0,\end{aligned}$$

in view of (6.25);

$$\begin{aligned}(2)\quad \left(\frac{M_1+M_3}{2}-\frac{\max(h_1,h_2)+h_1+1}{2}\right)-\Big(\frac{m_1+m_3}{2}+\frac{\ell_3}{2}\Big)\\ &=\frac{1}{2}\Big((M_1+M_3)-(m_1+m_3)-(\max(h_1,h_2)+h_1+1)-\ell_3\Big)\\ &=\frac{1}{2}\Big(\ell_1+\ell_3-(\max(h_1,h_2)+h_1+1)-\ell_3\Big)\\ &=\frac{1}{2}(k_1-\max(h_1,h_2)-2)\geq 0,\end{aligned}$$

by hypothesis $k_1\geq k_2$ and inequality (6.12).

We estimate k_1', k_2', k_1'', k_2''. First of all, we verify (6.21) :

$$\begin{aligned}k_1''&=|K_1\cap[a',M_1]|\geq M_1-a'+1-h_1\geq M_1-a+1-h_1=\\ &=2\Big(\frac{M_1+M_3}{2}-b\Big)+1-h_1\geq 2\frac{\max(h_1,h_2)+h_1+1}{2}+1-h_1=\\ &=\max(h_1,h_2)+2.\\ k_2''&=|K_2\cap[b'',M_2]|=|K_2\cap(b,M_2]|\geq M_2-[b]-h_2\geq M_2-b-h_2\geq\\ &\geq M_2-(M_2-h_1-h_2-2)-h_2=h_1+2.\end{aligned}$$

Further, using (6.24) one has

$$\begin{aligned}k_1'&=|K_1\cap[m_1,a']|=|K_1\cap[m_1,a)|\geq[a]-m_1+1-h_1\geq a-m_1-h_1=\\ &=2\Big(b-\frac{m_1+m_3}{2}\Big)-\ell_3-h_1\geq\max(h_1,h_2)+2,\end{aligned}$$

in view of the following two inequalities

$$
\begin{aligned}
(1)\quad &\left(2\left(M_2 - h_1 - h_2 - 2 - \frac{m_1+m_3}{2}\right) - \ell_3 - h_1\right) - \left(\max(h_1,h_2)+2\right)\\
&= 2\left(M_2 - \frac{m_1+m_3}{2}\right) - (3h_1 + 2h_2 + \ell_3 + \max(h_1,h_2) + 6)\\
&= 2(\ell_2 - \delta) - (3h_1 + 2h_2 + \ell_3 + \max(h_1,h_2) + 6)\\
&= 2k_2 - (2\delta + 3h_1 + \ell_3 + \max(h_1,h_2) + 8)\\
&= 2k_2 - (2\delta + 3h_1 + h_3 + k_3 + \max(h_1,h_2) + 7)\\
&\geq 2k_2 - (2\delta + 2h_1 + 2\max(h_1,h_2) + h_3 + k_3 + 7)\\
&= \Big((k_2+h_2) - (2\delta + 2h_1 + 2\max(h_1,h_2) + 2h_3 + k_3 + 3)\Big) + \Big(k_2 + h_3 - h_2 - 4\Big)\\
&\quad \geq k_2 + h_3 - h_2 - 4 \geq 0, \quad \text{because of (6.25) and (6.12),}
\end{aligned}
$$

$$
\begin{aligned}
(2)\quad &\left(2\left(\frac{M_1+M_3}{2} - \frac{\max(h_1,h_2)+h_1+1}{2} - \frac{m_1+m_3}{2}\right) - \ell_3 - h_1\right)\\
&\qquad\qquad - \left(\max(h_1,h_2)+2\right)\\
&= (M_1 - m_1 + M_3 - m_3) - (2\max(h_1,h_2) + 2h_1 + \ell_3 + 3)\\
&= (\ell_1 + \ell_3) - (2\max(h_1,h_2) + 2h_1 + \ell_3 + 3)\\
&= k_1 - (h_1 + 2\max(h_1,h_2) + 4) \geq 0,
\end{aligned}
$$

due to (6.12) and $k_1 \geq k_2$.

It only remains to estimate k_2'. Note that

$$
\begin{aligned}
k_2' &= |K_2 \cap [m_2, b'']| \geq b'' - m_2 + 1 - h_2\\
&\geq b - m_2 + 1 - h_2 \geq \max(h_1+2, h_2+h_3-k_3+3),
\end{aligned}
$$

because we may write the following four inequalities

$$
\begin{aligned}
(1)\quad &\Big((M_2 - h_1 - h_2 - 2) - m_2 + 1 - h_2\Big) - (h_1+2)\\
&= (M_2 - m_2) - (2h_1 + 2h_2 + 3)\\
&= \ell_2 - (2h_1 + 2h_2 + 3)\\
&= k_2 - (2h_1 + h_2 + 4)\\
&\geq 0, \quad \text{because of (6.12),}
\end{aligned}
$$

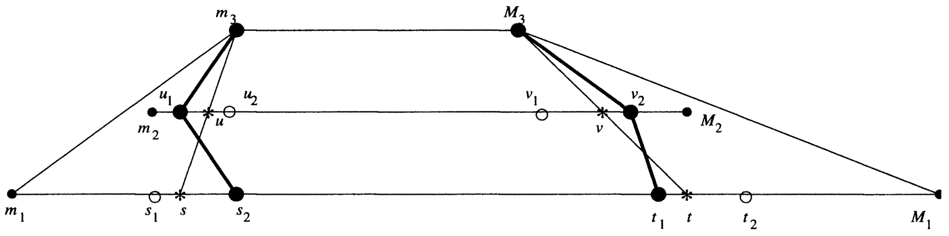
$$\begin{aligned}(2)\quad &\Big((M_2 - h_1 - h_2 - 2) - m_2 + 1 - h_2\Big) - (h_2 + h_3 - k_3 + 3)\\ &= (M_2 - m_2) + k_3 - (h_1 + 3h_2 + h_3 + 4)\\ &= \ell_2 + k_3 - (h_1 + 3h_2 + h_3 + 4)\\ &= (k_2 + k_3) - (h_1 + 2h_2 + h_3 + 5)\\ &\geq 0, \quad \text{due to (6.12),}\end{aligned}$$

$$\begin{aligned}(3)\quad &\Big(\frac{M_1 + M_3}{2} - \frac{\max(h_1, h_2) + h_1 + 1}{2} - m_2 + 1 - h_2\Big) - (h_1 + 2)\\ &= \Big(\frac{M_1 + M_3}{2} - m_2\Big) - \frac{1}{2}(\max(h_1, h_2) + 3h_1 + 2h_2 + 3)\\ &= (\delta + \frac{\ell_1 + \ell_3}{2}) - \frac{1}{2}(\max(h_1, h_2) + 3h_1 + 2h_2 + 3)\\ &= \frac{1}{2}\Big((k_1 + k_3 + h_3) - (\max(h_1, h_2) + 2h_1 + 2h_2 + 5)\Big) + \delta\\ &\geq 0, \quad \text{thanks to (6.12),}\end{aligned}$$

$$\begin{aligned}(4)\quad &\Big(\frac{M_1 + M_3}{2} - \frac{\max(h_1, h_2) + h_1 + 1}{2} - m_2 + 1 - h_2\Big) - (h_2 + h_3 - k_3 + 3)\\ &= k_3 + \Big(\frac{M_1 + M_3}{2} - m_2\Big) - \frac{\max(h_1, h_2) + h_1 + 4h_2 + 2h_3 + 5}{2}\\ &= k_3 + (\delta + \frac{\ell_1 + \ell_3}{2}) - \frac{\max(h_1, h_2) + h_1 + 4h_2 + 2h_3 + 5}{2}\\ &= k_3 + \delta + \frac{1}{2}\Big((k_1 + k_3) - (\max(h_1, h_2) + 4h_2 + h_3 + 7)\Big)\\ &\geq 0, \quad \text{because of (6.27).}\end{aligned}$$

The proof of case 2 is now complete. □

Third Case. — $\frac{m_1+m_3}{2} \leq m_2 \leq M_2 \leq \frac{M_1+M_3}{2}$.



We split $\mathbb{K}$ into three sets $\mathbb{K}', \mathbb{K}'', \mathbb{K}'''$. Choose $s \leq t$ between m_1 and M_1 and $u \leq v$ between m_2 and M_2 such that the points $(0, s), (1, u), (2, m_3)$ and $(0, t), (1, v), (2, M_3)$ are collinear.

Take s_1, s_2, t_1, t_2 in K_1 such that $s_1 \leq s < s_2$, $t_1 < t \leq t_2$ and there is no point of K_1 in the intervals $(s_1, s), (s, s_2)$, $(t_1, t), (t, t_2)$.

Take u_1, u_2, v_1, v_2 in K_2 such that $u_1 \leq u \leq u_2$, $v_1 \leq v \leq v_2$ and there is no point of K_2 in the intervals $(u_1, u), (u, u_2)$, $(v_1, v), (v, v_2)$. Define

$$K_1' = K_1 \cap [m_1, s_2],\ K_1'' = K_1 \cap [s_2, t_1],\ K_1''' = K_1 \cap [t_1, M_1], \tag{6.29}$$

$$K_2' = K_2 \cap [m_2, u_1],\ K_2'' = K_2 \cap [u_1, v_2],\ K_2''' = K_2 \cap [v_2, M_2], \tag{6.30}$$

$$K_3' = \{m_3\},\ K_3'' = K_3,\ K_3''' = \{M_3\}. \tag{6.31}$$

It will be shown that we may choose the points s, t, u, v such that

$$\ell_1' \leq 2k_1' - 3, \quad \max(\ell_1', \ell_2') \leq k_1' + k_2' - 3, \tag{6.32}$$

$$\ell_1'' \leq 2k_1'' - 3, \quad \max(\ell_1'', \ell_2'') \leq k_1'' + k_2'' - 3,\ \ell_2'' + \ell_3'' \leq 2k_2'' + 2k_3'' - 5, \tag{6.33}$$

$$\ell_1''' \leq 2k_1''' - 3, \quad \max(\ell_1''', \ell_2''') \leq k_1''' + k_2''' - 3. \tag{6.34}$$

We can easily deduce Lemma 6.3 from (6.32-34). Denote $x = |2\mathbb{K}_2'' \cap (\mathbb{K}_1' + \mathbb{K}_3')| + |2\mathbb{K}_2'' \cap (\mathbb{K}_1''' + \mathbb{K}_3''')|$. It is clear that

$$\begin{aligned}|2\mathbb{K}| \geq{} & |2\mathbb{K}_1'| + |\mathbb{K}_1' + \mathbb{K}_2'| + |\mathbb{K}_1' + \mathbb{K}_3'| + |\mathbb{K}_2' + \mathbb{K}_3'| + |2\mathbb{K}_3'| \\ & + |2\mathbb{K}_1''| + |\mathbb{K}_1'' + \mathbb{K}_2''| + |2(\mathbb{K}_2'' \cup \mathbb{K}_3'')| \\ & + |2\mathbb{K}_1'''| + |\mathbb{K}_1''' + \mathbb{K}_2'''| + |\mathbb{K}_1''' + \mathbb{K}_3'''| + |\mathbb{K}_2''' + \mathbb{K}_3'''| + |2\mathbb{K}_3'''| - 8 - x.\end{aligned} \tag{6.35}$$

Indeed, the above inequality is true, because the points $2s_2, 2t_1$, $u_1 + s_2, v_2 + t_1, u_1 + m_3, M_3 + v_2, 2m_3, 2M_3, 2\mathbb{K}_2'' \cap (\mathbb{K}_1' + \mathbb{K}_3'), 2\mathbb{K}_2'' \cap (\mathbb{K}_1''' \cup \mathbb{K}_3''')$ are counted twice. By Theorem D we get

$$\begin{aligned}|2\mathbb{K}| \geq{} & (\ell_1' + k_1') + (\ell_1' + k_2') + k_1' + k_2' + 1 \\ & + (\ell_1'' + k_1'') + (\ell_2'' + k_1'') + (2k_2'' + 2k_3'' - 1 + \ell_2'' + \ell_3'') \\ & + (\ell_1''' + k_1''') + (\ell_1''' + k_2''') + k_1''' + k_2''' + 1 - (8 + x) \\ ={} & (2k_1' + 2k_2' + 1 + 2\ell_1') + (2k_1'' + 2k_2'' + 2k_3'' - 1 + \ell_1'' + 2\ell_2'' + \ell_3'') \\ & + (2k_1''' + 2k_2''' + 1 + 2\ell_1''') - (8 + x) \\ ={} & 2(k_1' + k_1'' + k_1''') + 2(k_2' + k_2'' + k_2''') + 2k_3'' - 7 \\ & + (2\ell_1 - \ell_1'') + 2\ell_2'' + \ell_3'' - x \\ ={} & 2k + 1 + 2\ell_1 + 2\ell_3 + [2\ell_2'' - (\ell_1'' + \ell_3)] - x.\end{aligned} \tag{6.36}$$

We have used here $k_3'' = k_3$, $\ell_3'' = \ell_3$ and $k_i' + k_i'' + k_i''' = k_i + 2$, for $i = 1, 2$. Note that the collinearity condition gives $2(v - u) = (t - s) + \ell_3$ and thus we have

$$\begin{aligned}2\ell_2'' - (\ell_1'' + \ell_3) &= 2[(v - u) + (u - u_1) + (v_2 - v)] \\ &\quad - [(t - s) - (t - t_1) - (s_2 - s) + \ell_3] \\ &= 2(u - u_1) + 2(v_2 - v) + (t - t_1) + (s_2 - s).\end{aligned} \tag{6.37}$$

It is clear that

$$|(\mathbb{K}_1' + \mathbb{K}_3') \cap 2\mathbb{K}_2''| \leq 1 + |[2u_1, 2u]| = 2 + 2(u - u_1),$$

$$|2\mathbb{K}_2'' \cap (\mathbb{K}_1''' + K_3''')| \leq 1 + |[2v, 2v_2]| = 2 + 2(v_2 - v).$$

Therefore, $x \leq 4 + 2(u - u_1) + 2(v_2 - v)$; applying this inequality and (6.37) in (6.36), we obtain the desired lower bound:

$$|2\mathbb{K}| \geq 2k - 3 + 2\ell_1 + 2\ell_3 + (t - t_1) + (s_2 - s) \geq (2k - 1) + 2\ell_1 + 2\ell_3. \qquad (6.38)$$

The last step in the proof is to choose u, s, v, t such that (6.32-6.34) are valid.

First of all, we rewrite these inequalities in the form

$$\begin{aligned}
&h_1' \leq k_1' - 2,\ h_1' \leq k_2' - 2,\ h_2' \leq k_1' - 2,\\
&h_1'' \leq k_1'' - 2,\ h_1'' \leq k_2'' - 2,\ h_2'' \leq k_1'' - 2,\\
&h_2'' + h_3'' \leq k_2'' + k_3'' - 3 = k_2'' + k_3 - 3,\\
&h_1''' \leq k_1''' - 2,\ h_1''' \leq k_2''' - 2,\ h_2''' \leq k_1''' - 2.
\end{aligned}$$

Consequently, it will be enough to find u, s, t, v such that

$$k_2' \geq h_1 + 2,\ k_2''' \geq h_1 + 2, k_2'' \geq \max(h_1 + 2, h_2 + h_3 - k_3 + 3), \qquad (6.39)$$

$$k_1' \geq \max(h_1, h_2) + 2,\ k_1'' \geq \max(h_1, h_2) + 2,\ k_1''' \geq \max(h_1, h_2) + 2. \qquad (6.40)$$

Define u, v between m_2 and M_2 by

$$u = m_2 + (h_1 + h_2 + 2), \qquad v = M_2 - (h_1 + h_2 + 2). \qquad (6.41)$$

Take s, t between m_1 and M_1 so that $(0, s), (1, u), (2, m_3)$ and $(0, t), (1, v), (2, M_3)$ are collinear. We obtain

$$v - u = \ell_2 - 2(h_1 + h_2 + 2) \text{ and } t - s = 2(v - u) - \ell_3. \qquad (6.42)$$

In order to prove (6.39-40), it will suffice to estimate k_2', k_2'', k_2''', k_1', k_1'', k_1'''. We begin by establishing (6.39).

$$\begin{aligned}
k_2' &= |K_2 \cap [m_2, u_1]| = |K_2 \cap [m_2, u)| \geq (u - m_2) - h_2 = h_1 + 2.\\
k_2''' &= |K_2 \cap [v_2, M_2]| = |K_2 \cap (v_1, M_2]| \geq (M_2 - v) - h_2 = h_1 + 2.\\
k_2'' &= |K_2 \cap [u_1, v_2]| = 2 + (v - u) - h_2 = 2 + \Big(\ell_2 - 2(h_1 + h_2 + 2)\Big) - h_2\\
&= k_2 - 2h_1 - 2h_2 - 3 \geq \max(h_1 + 2, h_2 + h_3 - k_3 + 3).
\end{aligned}$$

Indeed, $3h_1 + 2h_2 + 5 \leq k_2$ and $2h_1 + 3h_2 + h_3 + 6 \leq k_2 + k_3$ follow from (6.12). Remark that $v - u = \ell_2 - 2(h_1 + h_2 + 2) \geq \ell_3$, because it is equivalent to $2h_1 + h_2 + k_3 + h_3 + 4 \leq k_2$, which follows again from (6.12). Therefore, we may choose $s \leq t$ between m_1 and M_1 such that the points $(2, m_3)$, $(1, u)$, $(0, s)$ and $(2, M_3), (1, v), (0, t)$ are collinear. Define $s_1 \leq s < s_2$ and $t_1 < t \leq t_2$ such that s_1, s_2 and t_1, t_2 are consecutive points in K_1.

We check inequalities (6.40). Note that $s - m_1 \geq 2(u - m_2)$ and $M_1 - t \geq 2(M_2 - v)$. Using (6.42) we have

$$\begin{aligned}
k_1' &= |K_1 \cap [m_1, s_2]| \geq s_2 - m_1 + 1 - h_1 \geq (s - m_1) + 1 - h_1\\
&\geq 2(u - m_2) + 1 - h_1 = 2(h_1 + h_2 + 2) + 1 - h_1 = h_1 + 2h_2 + 5\\
&\geq \max(h_1, h_2) + 2,
\end{aligned}$$

$$\begin{aligned} k_1''' &= |K_1 \cap [t_1, M_1]| \geq (M_1 - t_1 + 1) - h_1 \geq (M_1 - t) + 1 - h_1 \\ &\geq 2(M_2 - v) + 1 - h_1 = 2(h_1 + h_2 + 2) + 1 - h_1 = h_1 + 2h_2 + 5 \\ &\geq \max(h_1, h_2) + 2, \end{aligned}$$

$$\begin{aligned} k_1'' &= |K_1 \cap [s_2, t_1]| = |K_1 \cap (s, t)| \geq [t] - [s] - h_1 \geq (t - s) - 1 - h_1 \\ &= \Big(2(v - u) - \ell_3\Big) - 1 - h_1 = 2\Big((\ell_2 - 2h_1 - 2h_2 - 4) - \ell_3\Big) - 1 - h_1 \\ &= 2(k_2 - 2h_1 - h_2 - 5) - k_3 - h_3 - h_1 = 2k_2 - (5h_1 + 2h_2 + h_3 + k_3 + 10) \\ &\geq \max(h_1, h_2) + 2, \quad \text{because of (6.12).} \end{aligned}$$

□

References

[B] Bilu Y., *Structure of sets with small sumsets,* Mathématiques Stochastiques, Univ. Bordeaux 2, Preprint **94-10**, 1994.

[F1] Freiman G.A., *Foundations of a structural theory of set addition* ; Translation of Mathematical Monographs, **37**, Amer. Math. Soc., Providence, R.I., USA, 1973.

[F2] Freiman G.A., *What is the structure of K if K + K is small?* ; Lecture Notes in Mathematics **1240**, 1987, Springer-Verlag, New-York, 109–134.

[F3] Freiman G.A., *Private Communication.*

[F4] Freiman G.A., *Inverse problems of additive number theory, VI. On addition of finite sets, III,* Izvest. Vuz. Mathem., **3 (28)**, 1962, 151–157.

[L-S] Lev V.F., Smeliansky P.Y., *On addition of two distinct sets of integers,* Acta Arithmetica, **LXX.1**, 1995, 85–91.

[R] Ruzsa I.Z., *Generalized arithmetic progressions and sumsets,* Acta Math. Hungar., **65(4)**, 1994, 379–388.

[S1] Stanchescu Y., *On addition of two distinct sets of integers,* Acta Arithmetica, **LXXV.2**, 1996, 191–194.

[S2] Stanchescu Y., *On the structure of sets with small doubling property on the plane (i),* Acta Arithmetica, **LXXXIII.2**, 1998, 127–141.

[S3] Stanchescu Y., *On the structure of sets with small doubling property on the plane (ii),* preprint.

Y. Stanchescu, School of Mathematical Sciences, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, Ramat Aviv, Israel 69978 • *E-mail :* ionut@math.tau.ac.il

Astérisque

YAKOV BERKOVICH

Non-solvable groups with a large fraction of involutions

Astérisque, tome 258 (1999), p. 241-248

<http://www.numdam.org/item?id=AST_1999__258__241_0>

Astérisque **258**, 1999, p. 241–248

NON-SOLVABLE GROUPS WITH A LARGE FRACTION OF INVOLUTIONS

by

Yakov Berkovich

Abstract. — In this note we classify the non-solvable finite groups G such that the class number of G is at least $|G|/16$. Some consequences are derived as well.

C.T.C. Wall classified all finite groups in which the fraction of involutions exceeds 1/2 (see [**1**], Theorem 11.24). In this paper we classify all non-solvable finite groups in which the fraction of involutions is not less than 1/4.

We recall some notation.

Let $k(G)$ be the class number of G. Let $i(G)$ denote the number of all involutions of G, $T(G) = \sum \chi(1)$ where χ runs over the set $\mathrm{Irr}(G)$. Now

$$\mathrm{mc}(G) = k(G)/|G|, \ f(G) = T(G)/|G|, \ i_o(G) = i(G)/|G|.$$

It is well-known (see [**1**], chapter 11) that

$$i(G) < T(G), \ i_o(G) < f(G), \ f(G)^2 \leq \mathrm{mc}(G)$$

(with equality if and only if G is abelian).

In this note we prove the following three theorems.

Theorem 1. — *Let G be a non-solvable group.*

If $mc(G) \geq 1/16$ then $G = G'Z(G)$, where G' is the commutator subgroup of G, $Z(G)$ is the centre of G, $G' \in \{PSL(2,5), SL(2,5)\}$.

Theorem 2. — *Let G be a non-solvable group.*

If $f(G) \geq 1/4$ then $G = G'Z(G)$ and $G' \in \{PSL(2,5), SL(2,5)\}$.

Theorem 3. — *Let G be a non-solvable group.*

Then $i_o(G) \geq 1/4$ if and only if $G = PSL(2,5) \times E$ with $\exp E \leq 2$.

Lemma 1 contains some well-known results.

Lemma 1

(a) If G is simple and a non-linear $\chi \in Irr(G)$ is such that $\chi(1) < 4$, then $\chi(1) = 3$ and $G \in \{PSL(2,5), PSL(2,7)\}$; see [2].

1991 ***Mathematics Subject Classification.*** — Primary 20C15, 20D05, 20E34.

Key words and phrases. — Non-solvable finite groups, class number.

(b) *(Isaacs; see* **[1]***, Theorem 14.19). If G is non-solvable, then $|cdG| \geq 4$; here $cdG = \{\chi(1) | \chi \in Irr(G)\}$.*

(c) *(see, for example,* **[1]***, Chapter 11). If G is non-abelian then*

$$mc(G) \leq 5/8, \ f(G) \leq 3/4.$$

***Lemma* 2.** — *Let $G = G' > 1$, $d \in \{4, 5, 6\}$. If $mc(G) \geq (1/d)^2$ then there exists a non-linear $\chi \in Irr(G)$ such that $\chi(1) < d$.*

Proof. — Suppose that G is a counterexample. Then by virtue of Lemma 1(b) one has

$$\begin{aligned} |G| &= \textstyle\sum_\chi \chi(1)^2 \geq 1 + d^2(k(G) - 3) + (d+1)^2 + (d+2)^2 \\ &\geq 1 + d^2(\tfrac{|G|}{d^2} - 3) + 2d^2 + 6d + 5 = |G| - d^2 + 6d + 6 > |G| \end{aligned}$$

since $d \in \{4, 5, 6\}$, — a contradiction (here χ runs over the set $\mathrm{Irr}(G)$).

Lemma 3 contains the complete classification of all groups G satisfying $i_o(G) = 1/4$.

***Lemma* 3.** — *If $i_o(G) = 1/4$ then one and only one of the following assertions holds:*

(a) *$G \cong A_4$, the alternating group of degree* 4.

(b) *$G \cong PSL(2, 5)$.*

(c) *G is a Frobenius group with kernel of index* 4.

(d) *G is a non-cyclic abelian group of order* 12.

(e) *G contains a normal subgroup R of order* 3 *such that $G/R \cong S_3 \times S_3$; if x is an involution in G then $|C_G(x)| = 12$ (here S_3 is the symmetric group of degree* 3*).*

Proof. — By the assumption $|G|$ is even. $i(G)$ is therefore odd by the Sylow Theorem and $|G| = 4i(G)$, $P \in \mathrm{Syl}_2(G)$ has order 4.

(i) Suppose that G has no a normal 2-complement. Then P is abelian of type $(2, 2)$ and by the Frobenius normal p-complement Theorem G contains a minimal non-nilpotent subgroup $F = C(3^a) \cdot P$ (here $C(m)$ is a cyclic group of order m and $A \cdot B$ is a semi-direct product of A and B with kernel B). Since all involutions are conjugate in F, all involutions are conjugate in G. Hence $C_G(x) = P$ for $x \in P^\# = P - \{1\}$, $a = 1$. If G is simple then by the Brauer-Suzuki-Wall Theorem (see **[1]**, Theorem 5.20) one has

$$|G| = (2^2 - 1)2^2(2^2 + 1) = 60.$$

Now we assume that G is not simple. Take H, a non-trivial normal subgroup of G. If $|G : H|$ is odd, then

$$\begin{aligned} i(G) = i(H), \ i_o(H) = i(H)/|H| = i(G)/|H| = \\ |G|i_o(G)/|H| = |G : H|i_o(G) = |G : H|/4. \end{aligned}$$

Therefore $|G : H| = 3$ and $i_o(H) = 3/4$. Now $f(H) > i_o(H)$, hence H is abelian (Lemma 1(c)) and $f(H) = 1$. It is easy to see that H is an elementary abelian 2-group, $H = P$. Now $|P| = 4$ implies $|G| = 12$, $F = G \cong A_4$.

Now suppose that H has even index. Since G is not 2-nilpotent (= has no a normal 2-complement) then $|H|$ is odd. In view of $|C_G(x)| = 4$ for $x \in P^\#$ one obtains that PH is a Frobenius group with kernel H, P is cyclic — a contradiction.

(ii) G has a normal 2-complement K.

First assume that P is cyclic. Then all involutions are conjugate in G, and for the involution $x \in P$ one has $C_G(x) = P$. Then G is a Frobenius group with kernel K of index 4.

Assume that $P = \langle\alpha\rangle \times \langle\beta\rangle$ is not cyclic. We have $P = \{1, \alpha, \beta, \alpha\beta\}$, and all elements from $P^\#$ are not pairwise conjugate in G. Thus

$$|G : C_G(\alpha)| + |G : C_G(\beta)| + |G : C_G(\alpha\beta)| = i(G) = |G : P|.$$

Note that $C_G(\alpha) = P \cdot C_K(\alpha)$, and similarly for β and $\alpha\beta$. Therefore

$$(1) \qquad |C_K(\alpha)|^{-1} + |C_K(\beta)|^{-1} + |C_K(\alpha\beta)|^{-1} = 1.$$

Since $|K| > 1$ is odd then (1) implies

$$(2) \qquad |C_K(\alpha)| = |C_K(\beta)| = |C_K(\alpha\beta)| = 3.$$

By the Brauer Formula (see **[1]**, Theorem 15.47) one has

$$(3) \qquad |K||C_K(P)|^2 = |C_K(\alpha)||C_K(\beta)||C_K(\alpha\beta)| = 3^3.$$

If $C_K(P) > 1$ then (3) implies $|K| = 3$ and $G = P \times K$ is an abelian non-cyclic group of order 12.

Assume $C_K(P) = 1$. Then $|K| = 3^3$. Now (2) implies that K is not cyclic. By analogy, (2) implies that $\exp K = 3$. From $\exp P = 2$ follows that G is supersolvable. Therefore R, a minimal normal subgroup of G, has order 3. Applying the Brauer Formula to G/R, one obtains $G/R \cong S_3 \times S_3$, and we obtain group (e).

Proof of Theorem 1. — Denote by $S = S(G)$ the maximal normal solvable subgroup of G.

(i) If G is non-abelian simple then $G \cong \mathrm{PSL}(2,5)$.

Proof. — Take $d = 4$ in Lemma 2. Then there exists $\chi \in \mathrm{Irr}(G)$ with $\chi(1) = 3$. Now Lemma 1(a) implies $G \in \{\mathrm{PSL}(2,5), \mathrm{PSL}(2,7)\}$. Since

$$\mathrm{mc}(\mathrm{PSL}(2,7)) = 1/28 < 1/16$$

then $G \cong \mathrm{PSL}(2,5)$ (note that $\mathrm{mc}(\mathrm{PSL}(2,5)) = 1/12$).

(ii) If G is semi-simple then $G \cong \mathrm{PSL}(2,5)$.

Proof. — Take in G a minimal normal subgroup D. Then $D = D_1 \times \cdots \times D_s$ where the D_i's are isomorphic non-abelian simple groups. Since (see **[1]**, Chapter 11) $\mathrm{mc}(D_1) \geq \mathrm{mc}(G) \geq 1/16$, $D \cong \mathrm{PSL}(2,5)$ by (i) and so $\mathrm{mc}(D_1) = 1/12$. Now

$$\mathrm{mc}(D) = \mathrm{mc}(D_1)^s = (1/12)^s \geq 1/16$$

implies that $s = 1$. Therefore $D \cong \mathrm{PSL}(2,5)$. Since $G/C_G(D)$ is isomorphic to a subgroup of $\mathrm{Aut}D \cong S_5$, $\mathrm{mc}(S_5) = 7/120 < 1/16$, then $G/C_G(D) \cong \mathrm{PSL}(2,5)$. Because $D \cap C_G(D) = 1$, $G = D \times C_G(D)$. Now

$$1/16 \leq \mathrm{mc}(G) = \mathrm{mc}(C_G(D))\mathrm{mc}(D) = (1/12)\mathrm{mc}(C_G(D))$$

implies that $\mathrm{mc}(C_G(D)) \geq 3/4 > 5/8$, $C_G(D)$ is abelian (Lemma 1(c)), $C_G(D) = 1$ (since G is semi-simple), and $G \cong \mathrm{PSL}(2,5)$.

(iii) $G/S \cong \mathrm{PSL}(2,5)$.

This follows from $\mathrm{mc}(G/S) \geq \mathrm{mc}(G)$ (P.Gallagher; see **[1]**, Theorem 7.46) and (ii).

(iv) If $G = G'$ then $G \in \mathrm{PSL}(2,5), \mathrm{SL}(2,5)\}$.

Proof. — By virtue of (iii) we may assume that $S > 1$.

Suppose that (iv) is true for all proper epimorphic images of G. Take in S a minimal normal subgroup R of G, and put $|R| = p^n$. Then by the Gallagher Theorem and induction one has $G/R \in \{PSL(2,5), \mathrm{SL}(2,5)\}$.

(1iv) $G/R \cong \mathrm{PSL}(2,5)$, i.e. $R = S$.

If $Z(G) > 1$ then $R = Z(G)$ is isomorphic to a subgroup of the Schur multiplier of G/R so $|R| = 2$ and $G \cong \mathrm{SL}(2,5)$ (Schur). In the sequel we suppose that $Z(G) = 1$.

Then $C_G(R) = R$, so $n > 1$. If $x \in R^\#$ then $|G : C_G(x)| \geq 5$, since index of any proper subgroup of $\mathrm{PSL}(2,5)$ is at least 5. Let $k_G(M)$ denote the number of conjugacy classes of G ($= G$-classes), containing elements from M. Then

$$k_G(R) \leq 1 + |R^\#|/5 = (p^n + 4)/5.$$

If $x \in G - R$ then $Z(G) = 1$, and the structure of G/R imply $|G : C_G(x)| \geq 12p$ (indeed, x does not centralize R and $|G/R : C_{G/R}(xR)| \geq 12$). Hence

$$k_G(G - R) = k(G) - k_G(R) = |G|\mathrm{mc}(G) - k_G(R) \geq$$
$$60p^n/16 - (p^n + 4)/5 = (71p^n - 16)/20.$$

Now

(1) $$|G - R| = 59p^n \geq 12pk_G(G - R) \geq 12p(71p^n - 16)/20,$$

(2) $$5 \times 59p^{n-1} = 295p^{n-1} \geq 213p^n - 48 \geq 426p^{n-1} - 48 \Rightarrow 131p^{n-1} \leq 48,$$

a contradiction.

(2iv) $G/R \cong \mathrm{SL}(2,5)$.

Proof. — Suppose that $R_1 \neq R$ is a minimal normal subgroup of G. Then (by induction)

$$RR_1 = R \times R_1 = S, \ |R_1| = 2, \ G/R_1 \cong \mathrm{SL}(2,5)$$

and $G' < G$, since the multiplier of $\mathrm{SL}(2,5)$ is trivial, a contradiction. Therefore R is a unique minimal normal subgroup of G. Similarly, one obtains $Z(G) = 1$.

Let $p > 2$. Then $C_G(R) = R$. In this case $Z(S) < R$, so $Z(S) = 1$ and S is a Frobenius group with kernel R of index 2. As in (1iv) one has

$$k_G(S) = k_G(S - R) + k_G(R) \leq 1 + (p^n + 4)/5 = (p^n + 9)/5.$$

If $x \in G - S$ then $|G : C_G(x)| \geq 12p$ and

$$k_G(G - S) = k(G) - k_G(S) = |G|\mathrm{mc}(G) - k_G(S) \geq$$
$$120p^n/16 - (p^n + 9)/5 = (73p^n - 18)/10,$$
$$|G - S| = 118p^n \geq 12pk_G(G - S) \geq 6p(73p^n - 18)/5,$$
$$295p^{n-1} \geq 219p^n - 54 \geq 657p^{n-1} - 54,$$
$$54 \geq 362p^{n-1},$$

a contradiction.

Let $p = 2$. Since R is the only minimal normal subgroup of G and $Z(G) = 1$ then,

$$\begin{aligned} k_G(S) &\leq 1 + (2^{n+1} - 1)/5 = (2^{n+1} + 4)/5, \\ k_G(G - S) &\geq 120.2^n/16 - (2^{n+1} + 4)/5 = (71.2^n - 8)/10, \\ 59.2^{n+1} = |G - S| &\geq 24k_G(G - S) \geq 24(71.2^n - 8)/10, \\ 295.2^n &\geq 426.2^n - 48, \\ 48 &\geq 131.2^n, \end{aligned}$$

a contradiction.

(v) If D is the last term of the derived series of G then $D \in \{\mathrm{PSL}(2,5), \mathrm{SL}(2,5)\}$.

Proof. — Since $D = D'$ and $\mathrm{mc}(D) \geq \mathrm{mc}(G) \geq 1/16$ the result follows from (iv).

(vi) The subgroup D from (v) coincides with G'.

Proof. — We have $D \in \{\mathrm{PSL}(2,5), \mathrm{SL}(2,5)\}$ by (v). Since $Z(G) < D$ we may, by virtue of the Gallagher Theorem [**1**], Theorem 7.46, assume that $Z(D) = 1$. Then $D \cong \mathrm{PSL}(2,5)$. Since

$$\mathrm{Aut}D \cong S_5, \ \mathrm{mc}(S_5) = 7/120 < 1/16$$

then

$$G/C_G(D) \cong \mathrm{PSL}(2,5), \ G = D \times C_G(G),$$

and $C_G(D)$ is abelian (see (ii)). So $D = G'$.

(vii) $G = SG'$.

This follows from (iii) and (vi).

(viii) $|S'| \leq 2$. In particular, S is nilpotent and all its Sylow subgroups of odd orders are abelian.

Proof. — In fact, $S' \leq S \cap G' \leq Z(G')$.

(ix) $G = S * G'$, a central product.

Proof. — Take an element x of order 5 in G'. Since $G' \cap S \leq Z(G)$, then

$$G/G' \cap S = G'/G' \cap S \times S/S \cap G'$$

implies that $\langle x, S\rangle$ is nilpotent. Hence $\langle S, x\rangle = P \times A$ where $P \in \mathrm{Syl}_2(S)$ and A is abelian. As $x \in A$ then $x \in C_G(S)$. Since $G' = \langle x \in G' | x^5 = 1\rangle$ it follows that $G = SG' = S * G'$.

(x) S is abelian.

Proof. — We have $G = (S \times G')/Z$ where $|Z| \leq 2$. For $G' \cong \mathrm{PSL}(2,5)$ our assertion is evident. Now let $G' \cong \mathrm{SL}(2,5)$. Then $|Z| = 2$, $Z \geq S'$. Suppose that S is non-abelian. Then $Z = S'$.

Take $\chi \in \mathrm{Irr}(G)$. We consider χ as a character of $G' \times S$ such that $Z \leq \ker\chi$. Then $\chi = \tau\vartheta$ where $\tau \in \mathrm{Irr}(G')$, $\vartheta \in \mathrm{Irr}(S)$ and $\chi_Z = \chi(1)1_Z = \tau(1)\vartheta(1)1_Z$. Now $\tau_Z = \tau(1)\lambda$, $\vartheta_Z = \vartheta(1)\mu$ where $\lambda, \mu \in \mathrm{Irr}(Z)$, $\lambda\mu = 1_Z$. Noting that $|Z| = 2$, one has

$\lambda = \mu$ and $\tau_Z = \tau(1)\lambda, \vartheta_Z = \vartheta(1)\lambda$. Since S is non-abelian then $\mathrm{cd}S = \{1, m\}$ where $m^2 = |S : Z(S)|$.

Suppose that $\lambda = 1_Z$. $\mathrm{Irr}(G')$ has exactly 5 characters containing Z in their kernels, so for τ we have exactly 5 possibilities. Since $Z \leq \ker\vartheta$ then $\vartheta \in \mathrm{Lin}(S)$, and for ϑ we have exactly $|\mathrm{Lin}(S)| = |S|/2$ possibilities. Hence for χ we have exactly $5|S|/2$ possibilities if $\lambda = 1_Z$.

Suppose that $\lambda \neq 1_Z$. Then Z is not contained in $\ker\tau$, so for τ we have exactly $|\mathrm{Irr}(G')| - |\mathrm{Irr}(G'/Z)| = 9 - 5 = 4$ possibilities. Since $S' = Z$ is not contained in $\ker\vartheta$, then ϑ is not linear, and for ϑ we have exactly $(|S| - |S/S'|)/m^2 = |S|/2m^2$ possibilities. For χ we have, in this case, exactly $4|S|/2m^2 = 2|S|/m^2$ possibilities.

Finally,

$$k(G) = 5|S|/2 + 2|S|/m^2$$

and

$$\mathrm{mc}(G) = k(G)/|G| = k(G)/60|S| = 1/24 + 1/30m^2.$$

Since $m > 1$ then

$$\mathrm{mc}(G) \leq 1/24 + 1/120 = 1/20 < 1/16,$$

a contradiction. Therefore S is abelian, $S = Z(G)$ and $G = G'Z(G)$. In this case $\mathrm{mc}(G) \in \{1/12, 3/40\}$. The theorem is proved.

Let now $f(G) \geq 1/4$. Then $\mathrm{mc}(G) > f(G)^2 \geq 1/16$, and Theorem 2 is a corollary of Theorem 1. It is easy to see that in this case $f(G) = f(G') \in \{4/15, 1/4\}$.

Proof of Theorem 3. — In view of Lemma 3 we may assume that $i_o(G) > 1/4$. Since

$$\mathrm{mc}(G) \geq f(G)^2 > i_o(G)^2 > 1/16$$

we may apply Theorem 1. By this theorem $G = G'Z(G)$ where

$$G' \in \{\mathrm{PSL}(2,5), \mathrm{SL}(2,5)\}.$$

If $G' = G$ then $G \cong \mathrm{PSL}(2,5)$ since $i_0(\mathrm{SL}(2,5)) = 1/120 < 1/4$. Now let $G' < G$.

Suppose that $\exp(G/G') > 2$. Let M/G' be the subgroup generated by all involutions of G/G'. Then $i(M) = i(G)$,

$$i_o(M) = i(M)/|M| = |G : M|i(G)/|G| =$$
$$|G : M|i_o(G) \geq |G : M|/4 \geq 1/2,$$

and M is solvable by **[1]** Theorem 11.24 (since $f(M) > i_o(M) \geq 1/2$), a contradiction. Thus $\exp(G/G') = 2$.

If $G' = \mathrm{PSL}(2,5)$ then $G = G' \times Z(G)$. If $\exp Z > 2$ and $M = G' \times \Omega_1(Z(G))$ then

$$i(G) = i(M),\ i_o(M) = |G : M|i_o(G) > |G : M|/4 \geq 1/2,$$

and M is solvable (see **[1]**, Theorem 11.24) — a contradiction. Hence if $G' \cong \mathrm{PSL}(2,5)$ then $G = \mathrm{PSL}(2,5) \times E$ with $\exp E \leq 2$.

Now suppose that $G = G'Z(G)$, $G' \cong \mathrm{SL}(2,5)$ and $Z(G)$ is a 2-subgroup. Set $\langle z \rangle = Z(G')$.

If $\exp Z(G) = 2$ then $Z(G) = \langle z \rangle \times E$, $G = G' \times E$, and $i_o(G) < 1/4$. Assume that $\exp Z(G) = 4$. Then

$$G' \cap Z(G) = \langle z \rangle = \Phi(G)$$

where $\Phi(G)$ is the Frattini subgroup of G.

Let s be an element of order 4 in $Z(G)$. Then $Z(G) = \langle s \rangle \times E$ and

$$G = (G'\langle s \rangle) \times E, \exp E \leq 2.$$

Let us calculate $i_o(H)$ where

$$H = G'\langle s \rangle,\ Z(H) = \langle s \rangle,\ o(s) = 4.$$

Take $P \in \mathrm{Syl}_2(G')$. Then $P \cong Q(8)$ contains exactly three distinct cyclic subgroups $\langle a \rangle$, $\langle b \rangle$, $\langle c \rangle$ of order 4, and $a^2 = b^2 = c^2 = s^2 = z$. Hence

$$(as)^2 = (bs)^2 = (cs)^2 = 1$$

and it is easy to see that $i_o(\langle P, s \rangle) = 7$. Now

$$\langle P, s \rangle \in \mathrm{Syl}_2(H),\ |H : N_H(\langle P, s \rangle)| = 5,$$
$$\langle P, s \rangle \cap \langle P, s \rangle^x = \langle s \rangle$$

for all $x \in H - N_H(\langle P, s \rangle)$. Thus

$$i_o(H) = |H : N_H(\langle P, s \rangle)| i_o(\langle P, s \rangle) -$$
$$(|H : N_H(\langle P, s \rangle)| - 1) i_o(\langle s \rangle) = 5 \times 7 - 4 = 31.$$

Since

$$G = H \times E,\ |E| = 2^\alpha,\ \exp E \leq 2,$$

then

$$i(G) = i(H)|E| + |E| - 1 = 31.2^\alpha + 2^\alpha - 1 = 32.2^\alpha - 1,$$
$$i_o(G) = i(G)/|G| = (32.2^\alpha - 1)/240.2^\alpha < 2/15 < 1/4,$$

a contradiction. Therefore $G' \not\cong \mathrm{SL}(2,5)$ and the theorem is proved.

Question. — Find all non-solvable groups G with $i_o(G) = 2^{-n}$, $n > 2$.

There exist four multiplication tables for two-element subsets of group elements (see [**3**]). These multiplication tables afford the following 2×2 squares:

$$\begin{matrix} A & B \\ B & A \end{matrix} \qquad \begin{matrix} A & B \\ B & C \end{matrix} \qquad \begin{matrix} A & B \\ C & A \end{matrix} \qquad \begin{matrix} A & B \\ C & D \end{matrix}$$

Here distinct letters denote distinct elements of a group.

Let us calculate the number $P(1)$ of the squares of the first type in a finite group G. If a pair $\{a, b\}$ of elements of G affords a square of the first type, then $a^2 = b^2$, $ab = ba$. Then $(a^{-1}b)^2 = 1$, so $i = a^{-1}b$ is the involution commuting with a and b. If $i \in \mathrm{Inv}(G)$ (the set of all involutions of G), $x \in C_G(i)$, then the pair (x, xi) affords the square of the first type. Therefore $i \in \mathrm{Inv}(G)$ affords exactly $|C_G(i)|$ squares of the first type. Let

$$\mathrm{Inv}(G) = K(1) \cup \cdots \cup K(r),$$

where $K(1), \dots, K(r)$ are distinct conjugacy classes of G. Then

$$P(1) = \sum_{i \in \mathrm{Inv}(G)} |C_G(i)| = \sum_{j=1}^{r} \sum_{i \in K(j)} |C_G(i)| = r|G|.$$

Thus $P(1) = r|G|$, where r is the number of conjugacy classes of involutions in G.

By analogy, we may prove that the number $P(1,2)$ of commutative squares in the multiplicative table of G is equal to $k(G)|G|$. The number $P(2)$ of squares of the second type in the multiplicative table of G is therefore equal to $P(2) = P(1,2) - P(1) = (k(G) - r)|G|$. If $p(n)$ is the fraction of squares of the n-th type in the multiplicative table of G then

$$p(1) = r/|G|, \ p(2) = (k(G) - r)/|G| = \mathrm{mc}(G) - p(1).$$

It is easy to see that the number $P(1) + P(3)$ of squares of the first and the third type in the multiplicative table of G is equal to $|G|s$ where s is the number of real classes (a class K of G is said to be real if $x \in K \Rightarrow x^{-1} \in K$). Thus

$$P(4) \equiv 0 \pmod{|G|}.$$

References

[1] Berkovich Ya. G., Zhmud' E. M., *Characters of finite groups*, Part 1, Amer. Math. Soc., Providence, Rhode Island, 1998.

[2] Blichfeldt H. F., *Finite collineation groups*, Chicago, 1917.

[3] Freiman G. A., *On two- and three-element subsets of groups*, Æquat. Math., **22**, 1981, 140–152.

Y. Berkovich, Research Institute of Afula, Department of Mathematics and Computer Science, University of Haifa, 31905 Haifa, Israel • *E-mail :* berkov@mathcs2.haifa.ac.il

YAKOV BERKOVICH

Questions on set squaring in groups

Astérisque, tome 258 (1999), p. 249-253

<http://www.numdam.org/item?id=AST_1999__258__249_0>

Astérisque **258**, 1999, p. 249–253

QUESTIONS ON SET SQUARING IN GROUPS

by

Yakov Berkovich

Abstract. — Some questions on small subsets in groups are posed and discussed.

Let M be a subset of a group G. Define

$$M^2 = \{x \mid x = ab, a, b \in M\},$$

the square of M. M is a set with a large square if $a, b, c, d \in M$ and $ab = cd$ implies $a = c, b =$ d. If M is a finite set note that $\mid M^2 \mid = \mid M \mid^2$. In the opposite case M is said to be a set with small square.

It is natural to consider two group subsets M, N as equivalent if they have equal multiplication tables. To be more precise, we give the following

Definition. — *Let $M \subseteq G, N \subseteq H$ where G, H are groups. A bijection φ from M onto N is said to be an S -isomorphism if for $a, b, c, d \in M$ the equality $ab = cd$ implies $\varphi(a)\varphi(b) = \varphi(c)\varphi(d)$, and conversely.*

The group isomorphism is an S-isomorphism, but the converse assertion is not true. Moreover, if G is a group with non-trivial centre then there exists an S-isomorphism from G onto G which is not a group isomorphism. The automorphism group AUT(M) of a group subset M is defined as usual. If M is a finite group set with great squaring then AUT(M) $\cong S_n$ where $n = \mid M \mid$.

Question 1. — *Find all group n-sets M with* $\mathrm{AUT}(M) \cong \{1\}$.

Question 2. — *Is there for any group H a group set M such that* AUT(M) $\cong H$ *?*

Question 3. — *Find all group n-sets M such that* AUT(M) $\cong S_n$.

Question 4. — *Find all group n-sets M such that* AUT(M) $\cong A_n$ *(may be the set of all such M for $n > 3$ is empty).*

1991 ***Mathematics Subject Classification.*** — 20E34.
Key words and phrases. — Deficient squares groups, Squaring bounds in groups.

The classification of all group n-sets is a very difficult problem. Let Set(n) be the number of all pairwise non-isomorphic group n-sets. Then Set(2) $= 4$, Set(3) $= 54$ (G. Freiman); see[4,6].

Question 5. — *Find Set(4).*

Consider the easiest case $n = 2$. As we saw there are four distinct group 2-sets with the following squares (i.e., their multiplication tables):

$$\begin{matrix} AB & AB & AB & AB \\ BA & BC & CA & CD \end{matrix}$$

We note that the number $Gr(n)$ of pairwise non-isomorphic groups of order n is not a monotone function. In the same time Set(n) is a monotone function.

Let us continue to consider the case $n = 2$. Suppose that G is finite. Denote by $P_G(i)$ the number of 2-sets of type i in G. The following result is due to Freiman: If $P_G(4) = 0$ then G is abelian or a dedekindian 2-group, and conversely [6]. Now $P_G(1) = 0$ if and only if G is of odd order, and $P_G(2) = 0$ if and only if G is an elementary abelian 2-group. Lastly $P_G(3) = 0$ if and only if a Sylow 2-subgroup S is normal in G and $P_S(3) = 0$ [4]. As A. Mann showed, a 2-group S has no squares of third type if and only if $x^2 = y^2 \Leftrightarrow (xy^{-1})^2 = 1$ for $x, y \in$ S. Next
$P_G(1) + P_G(2) =\mid G \mid k(G)$ where $k(G)$ is the class number of G(A.Mann);
$P_G(1) + P_G(3) =\mid G \mid r(G)$ where $r(G)$ is the number of real G-classes (a class K is real if $x \in K \Leftrightarrow x^{-1} \in K$);
$P_G(1) =\mid G \mid k_i(G)$ where $k_i(G)$ is the number of G-classes containing involutions. Note that $P_G(1) + P_G(3) =\mid \{(x,y) \in G \times G \mid x^2 = y^2\} \mid$.

In particular $P_G(i) \equiv 0(\text{mod} \mid G \mid)$.

Now we see that a fraction of commutative 2×2-squares in G is equal to

$$mc(G) =\mid G \mid k(G)/ \mid G \mid^2 = k(G)/ \mid G \mid,$$

the measure of commutativity of G. Note that $k(G) =$|Irr(G)|, the number of ordinary irreducible characters of G. Therefore we may study $mc(G)$ by means of representation theory. This function has a number of nice properties. For example, if $H \leq G$ then $mc(H) \geq mc(G)$; if H is normal in G then $mc(G) \leq mc(H)mc(G/H)$; see [1], §§7.8,7.11,11.3.

Question 6. — *Is it true that the number of n-subsets of given type in G is divisible by $\mid G \mid$ for small values n (for example, for $n = 3$) and large $\mid G \mid$?*

A number of authors have classified all groups without 3×3-squares with 9 distinct elements; see [2].

Question 7. — *Classify all groups without 4×4-squares with 16 distinct elements.*

P. Neumann showed that if all n-subsets of G have small squares, then G contains a finite normal subgroup H such that G/H is an extension of an abelian group by a finite group. Herzog, Longobardi and Maj classified all such groups; see [7].

B. Neumann showed that $\mid G : Z(G) \mid$ is finite if and only if any infinite subset of G contains a pair of commuting elements.

Question 8. — *What we may say about a structure of G if any of its infinite subset contains a small square?*

If for some $k > 1$ there is a connection between $\mid M^2 \mid$ and k for all k-subsets M of G then in some cases we may make strong assertions on G. We note the following characterization of abelian groups:

Theorem **(L. Brailovsky)**. — *Let $k > 2$ be a positive integer such that $(k^2-3)(k-2) <\mid G \mid /15$ if G is finite. If*

$$\mid M^2 \mid\leq (k^2 + 2k - 3)/2$$

is true for any $M \subset G$ with $\mid M \mid= k$ then G is abelian.

We note that if G is abelian then $\mid M^2 \mid\leq k(k+1)/2$ for all such M.

It is interesting to consider a group generated by a set with small square or cube. Some results in this direction are contained in the following theorem

Theorem **(S. Brodsky)**

(a) If $\mid \{a,b\}^3 \mid< 7$ then the subgroup $\langle a,b\rangle$ is solvable.

(b) If $\mid \{a,b\}^4 \mid< 11$ then the subgroup $\langle a,b\rangle$ is solvable.

(c) The author completely described groups $G = \langle a,b\rangle$ for which

$$\mid \{a,b\}^3 \mid> 6 \text{ and } \mid \{a,b\}^4 \mid< 14.$$

This theorem was proved by means of a computer; see [5].

The set Q of $n \times n$-squares is said to be minimal, corresponding to a $n \times n$-square q, if it satisfies the following conditions:

(a) $q \in$ Q.

(b) Let $q_1 \in Q - \{q\}$ and T be the set of all groups containing $Q - \{q_1\}$. Then there exists a square $q_0 \notin Q - \{q_1\}$ which is contained in all groups of the set T.

Question 9. — *Find all one element minimal sets of $n \times n$-squares.*

Question 10. — *For $n = 3$ find for any square q all minimal sets containing q.*

We consider in the remaining part of the lecture "large subsets".

Theorem **(G. Freiman)**. — *Let M be a finite subset of a group G such that $\langle M\rangle = G$. Suppose that $\mid M^2 \mid< 1.5 \mid M \mid$. Then one of the following assertions is true;*

(a) M^2 is a subgroup of G.

(b) $M^2 = xH$ where H is a normal subgroup of G.

Question 11. — *Change in this theorem 1.5 to 2, i.e., consider the case when $\mid M^2 \mid\leq 2 \mid M \mid -1$.*

L. Brailovsky and G. Freiman described for torsion free groups the case when $\mid K^2 \mid= 2 \mid K \mid -1$. If K, M are finite subsets of a torsion free group then $\mid KM \mid\geq\mid K \mid + \mid M \mid -1$(*Kemperman*). L. Brailovsky and G. Freiman showed that if $\mid KM \mid=\mid K \mid + \mid M \mid -1$ then K and M are geometric progressions with the same factor.

Question 12. — *Let $M \subset G$ and for any $c \in G$ one has*

$$\mid (M \cup c)^2 - M^2 \mid\leq 1. \tag{$*$}$$

Describe the position of M in G.

It is easy to show that if in Question 12 $\mid M \mid> 1$ then $G = \langle M\rangle$. Now if $\mid M \mid= 2$ then G is solvable of derived length 2. But in the case $\mid M \mid= 3$ this question is very complicated.

Many results on squares of large subsets are contained in the lecture of M. Herzog.

Question 12 is, in some sense, a development of the idea of special elements, which was studied by Brailovsky, Freiman, and Herzog. An element $a \in G$ is said to be (m, n)-special if for any $b \in G$ one has $\mid \{a, b\}^m \mid\leq$ n.

Let $S_{m,n}(G)$ be the set of all (m, n)-special elements of a group G. The same three authors proved that $S_{2,3}(G)$ and $S_{3,5}(G)$ are characteristic subgroups of G. We may consider the sets $S_{m,n}(G)$ as natural generalizations of the centre $Z(G)$ of G(we note that $Z(G) \subset S_{2,3}(G)$). However Brailovsky showed that, in general, $S_{3,6}(G)$ is not a subgroup of G(he found that among 2,328 groups of order 2^7 only two cases for which this subset is not a subgroup).

A group G is said to be a $P(m, n)$-group if for any subset $\{a_1, \ldots, a_m\}$ of G one has

$$\mid \{a_{\sigma(1)} \ldots a_{\sigma(m)} \mid \sigma \in S_m\} \mid\leq \text{ n.}$$

G. Freiman and B. Schein showed that G is a $P(3, 2)$-group if and only if $\mid G' \mid\leq 2$. They also proved the analogous result for $P(3, 3)$-groups. They obtained many important results on $P(3, 4)$-groups.

I hope that the approach inspired by Freiman's number theoretical investigations on finite subsets with small doubling will considerably increase the subjects of group theory and will lead to new interesting results. For further information and references, see Freiman's survey.

References

[1] Berkovich Y. G., Zhmud' E. M., *Characters of finite groups*, Part 1,2, Amer. Math. Soc., Providence, Rhode Island, 1998.

[2] Berkovich Y. G., Freiman G. A. and Praeger C. E., *Small squaring and cubing properties of finite groups*, Bull. Austral. Math. Soc., **44**, 1991, 429–450.

[3] Brailovsky L., Freiman G. A. and Herzog M., *Special elements in groups* , Suppl. Rend. Circ. Mat. Palermo, **2 (23)**, 1990, 33–42.

[4] Brailovsky L., Freiman G. A., *On two-elements subsets in groups*, Ann. N.Y. Acad. Sci., **373**, 1981, 183–190.

[5] Brodsky S., this volume.

[6] Freiman G. A., *On two- and three-element subsets of groups*, Æquat. Math., **22**, 1981, 140–152.

[7] Herzog M., Longobardi P. and Maj M., *On a combinatorial problem in group theory*, Israel J. Math., **82**, 1993, 329–340.

Y. Berkovich, Department of Mathematics and Computer Science, University of Haifa, 31905 Haifa, Israel • *E-mail :* `berkov@mathcs2.haifa.ac.il`

Astérisque

SERGEI BRODSKY

On groups generated by a pair of elements with small third or fourth power

Astérisque, tome 258 (1999), p. 255-279

<http://www.numdam.org/item?id=AST_1999__258__255_0>

Astérisque **258**, 1999, p. 255–279

ON GROUPS GENERATED BY A PAIR OF ELEMENTS WITH SMALL THIRD OR FOURTH POWER

by

Sergei Brodsky

Abstract. — The paper is devoted to an investigation of two-generated groups such that the $m-$th power of the generating pair contains less than 2^m elements . It is proved, in particular, that if the cube of the generating pair contains less than 7 elements or its fourth power contains less than 11 elements, then the group is solvable. Otherwise, it is not necessarily solvable. The proofs use computer calculations.

1. Introduction

Let G be a group. A finite subset M of G is called *a set with small $m-$th power* (m is some integer) if $|M^m| < |M|^m$ (here $M^m = \{a_1 \dots a_m | a_1, \dots, a_m \in M\}$ and $|\,.\,|$ denotes the cardinality of the set). The structure of the groups in which each $p-$element subset has a small $m-$th power (for some small p and m), as well as the structure of the set of all special elements [1], was investigated in papers [1-5,7], among others. Notice that the notion of identification pattern, which is introduced in the present paper, is close to the notion "type of square" which was introduced in [3], but we will not discuss the relationship between these concepts.

In this paper we are interested in the structure of groups generated by a two-element set $M = \{a, b\}$ with a small third and fourth power. The proofs are based on pure combinatorial considerations, and are ultimately reduced to enumerating a list of very concrete groups, unfortunately; the total number of cases which appear here is so large that we need to use a computer. All computer calculations were developed by the author on an IBM PC using self-made programs which were written in the frame-work of the mathematical package MATLAB-386 [2]. These programs provide a simplification of finite group presentations using Tietze transformations, a calculation

1991 ***Mathematics Subject Classification.*** — 20F05, 20F16.
Key words and phrases. — Corepresentations of groups, solvable groups, small subsets in groups.

Supported in part by the Israel Ministry of Absorption and the Rashi Foundation.

[1] The element $a \in G$ is called *special* if the set $\{a, b\}$ has the small $m-$th power for some fixed integer m and each $b \in G$.

[2] ©The MathWorks, Inc., 1984-1991, version 3.5k.

of a commutator subgroups in the case of a finite index, and also recognition of groups of some types. The methods of programming are in some interest. Since their description would lead us too far from the topic of the present paper, the topic could be a subject of a separate publication. The results of the mentioned calculations are given in the Appendix.

Acknowledgment. — The author would like to thank Prof. Ya. Berkovich for the introduction into the subject of the investigations, as well as for useful discussions.

Let us formulate a general combinatorial assertion which will be needed below. Let A be a finite set, θ an equivalence relation on A, and $R \subseteq A \times A$. We say that the equivalence relation θ *is generated by* R, and write $\theta = \mathrm{eq}(R)$ if θ is the least equivalence relation contaning R. The relation θ will be called *independent* if θ is the minimal generating relation for its closure $\mathrm{eq}(R)$. The following lemma can be easily proved using induction on $|R|$.

Lemma 1. — *Let θ be an equivalence relation on the set A generated by a relation $R \subseteq A \times A$. Then $|A/\theta| \geq |A| - |R|$. If, in addition, R is independent, then $|A/\theta| = |A| - |R|$.*

2. Identification graphs and their properties

Let G be a group generated by two elements a and b: $G = \mathrm{gp}(a, b)$. We fix a and b as signature constants and regard the group G as the quotient-group of the free group $F = \langle a, b \rangle$. The natural epimorphism $\Phi_G : F \twoheadrightarrow G$ defines an equivalence relation on the group F which will be denoted by the symbol θ_G. We define $H(G)$ as the normal closure of the element ab^{-1} in G: $H(G) = (ab^{-1})^G$, and set $u_i = a^i b a^{-i-1}$ for each $i \in \mathbb{Z}$, so $H = \mathrm{gp}(u_i | i \in \mathbb{Z})$. For each element, or a subset P of $H(G)$, we let $P^{(s)}$ denote the element (the subset) $a^s P a^{-s}$; it is clear that $P^{(s)}$ can be obtained from P by adding s to all indices of the u-symbols. We also apply the same notation to elements and subsets of the Cartesian square $H_G \times H_G$: $(P, Q)^{(s)} = (P^{(s)}, Q^{(s)})$. Since $|\{a, b\}^m| = |\{a, b\}^m a^{-m}|$, the condition $|\{a, b\}^m| = n \leq 2^m$ $(m \geq 2)$ is equivalent to the condition $|H_m(G)| = n$ where $H_m(G) = \{a, b\} a^{-m}$. One can see that $H_m(G)$ consists of values in G of all strictly increasing positive words in symbols $u_0, \dots, u_{m-1}$:

$$H_m(G) = \{u_{i_1} \dots u_{i_k} \mid 0 \leq i_1 < \dots < i_k \leq m - 1,\ 0 \leq k < m\} \subseteq H(G).$$

We denote by U_m the set of all strictly increasing positive words in symbols $u_0, \dots,$ u_{m-1} itself, so that $H_m(F) = \mathrm{gp}(U_m)$ and $H_m(G) = \mathrm{gp}(\Phi_G(U_m))$.

For $S, T \in U_m$ we say that the pair (S, T) is an *irreducible m-pair* if exactly one of the words S, T begins with u_0 and exactly one of them ends with u_{m-1}. If the irreducible m-pair e has the form $(u_0 P, Q u_{m-1})$ we say that it is *positive*, otherwise e has the form $(u_0 P u_{m-1}, Q)$ and in this case we say that e is *negative*. In both cases we define $i(e) = P$ and $t(e) = Q$. The set of all positive irreducible m-pairs is denoted by I_m^+ and the set of all negative irreducible pairs is denoted as I_m^-.

For given $R \in U_m$, let $\tilde{R}$ be the word in symbols a and b which freely equals R; it is clear that $\tilde{R}$ is a positive word of length m. We say that an irreducible m-pair

(S,T) is *degenerate* if there exists some irreducible $(m-1)$-pair $(P,Q) \in \theta_G$ such that one of the words $\tilde{P}, \tilde{Q}$ is a subword of one of the words $\tilde{S}, \tilde{T}$. The following lemma is obvious.

Lemma 2. — *Let* $\theta_0 = \theta_G \cap (U_{m-1} \times U_{m-1})$ *and let* (S,T) *be a degenerate irreducible m-pair. Then* $(S,T) \in \theta$ *if and only if* $(S,T) \in \mathrm{eq}(\theta_0 \cup \theta_0^{(1)} \cup \theta_0 u_{m-1} \cup u_0\theta_0^{(1)})$.

Let us now define the *positive identification m-graph* $\Gamma_m^+(G)$ of G as the oriented graph with the set of vertices $H_{m-2}^{(1)}$ and the set of edges $E_m^+(G) = (\Phi_G \times \Phi_G)(I_m^+ \cap \theta_G)$, and the *negative identification m-graph* $\Gamma_m^-(G)$ of G as the graph with the same set of vertices and the set of edges $E_m^-(G) = (\Phi_G \times \Phi_G)(I_m^- \cap \theta_G)$. The incidence relations in both these graphs are given by the following rule: if $e \in E_m^+ \cup E_m^-$ and $e = (\Phi_G \times \Phi_G)(e_0)$, where e_0 is some irreducible m-pair, then the initial vertex of e is $\Phi_G(i(e_0))$ and the terminal vertex of e is $\Phi_G(t(e_0))$.

The correctness of the last definition, as well as the validity of the following lemma, can be easily verified.

Lemma 3. — *Let* $G = \mathrm{gp}(a,b)$ *and* $m \geq 2$. *Then each vertex of the positive m-identification graph* $\Gamma_m^+(G)$, *and each vertex of the negative m-identification graph* $\Gamma_m^-(G)$, *has at most one incoming edge and at most one outgoing edge.*

For $e \in E_m^+(G) \cup E_m^-(G)$, we call e a *degenerate* edge if and only if the set $(\Phi_G \times \Phi_G)^{-1}(e)$ contains some degenerate irreducible pair. Lastly, let $\mathrm{def}_m(G)$ denote the total number of nondegenerate edges in the set $E_m^+(G) \cup E_m^-(G)$.

Lemma 4. — *Let* $G = \mathrm{gp}(a,b)$ *and* $m \geq 2$. *Then*

$$\mathrm{def}_m(G) \geq -2^m - |H_m(G)| + 4|H_{m-1}(G)|.$$

Proof. — Let $d = 2^{m-1} - |H_{m-1}(G)|$. Then, by Lemma 1, the trace θ_0 of the equivalence relation θ_G on the set U_{m-1} is generated by some relation R_0 of cardinality d. Since $U_m \times U_m = (U_{m-1} \times U_{m-1}) \cup (U_{m-1}^{(1)} \times U_{m-1}^{(1)}) \cup (U_{m-1}u_{m-1} \times U_{m-1}u_{m-1}) \cup (u_0U_{m-1}^{(1)} \times u_0U_{m-1}^{(1)})$, the trace θ of the equivalence relation θ_G on the set U_m can be represented as the union of their traces $\theta_0, \theta_1, \theta_2, \theta_3$ on the sets $U_{m-1}, U_{m-1}^{(1)}, U_{m-1}u_{m-1}, u_0U_{m-1}^{(1)}$, respectively, and the relation $(I_G^+ \cup I_G^-) \cap \theta_G$. Each of the equivalence relations θ_k $(k = 1,2,3,4)$ is generated by a d-element relation $(R_0, R_0^{(1)}, R_0u_{m-1}, u_0R_0^{(1)}$, respectively). The union R of last the four relations contains no more than $4d$ elements. By Lemma 2, the difference $(I_G^+ \cup I_G^-) \cap \theta_G \setminus \mathrm{eq}(R)$ is contained in the set of all nondegenerate irreducible m-pairs from θ. Now let us define R_1 as the set which contains one $\Phi_G \times \Phi_G$ pre-image of each nondegenerate edge from $E_m^+(G) \cup E_m^-(G)$. Then $\theta_0 = \mathrm{eq}(R \cup R_1)$, and it only remains to apply Lemma 1.

The inequality which was obtained in Lemma 4 provides us with good necessary conditions for a group to be generated by a pair with a small power. However, we need a more detailed version of this result which also includes some sufficient conditions.

Lemma 5. — *Let $G = \mathrm{gp}(a, b)$ and $H_{m-1}(G) \geq 2^{m-1} - 1$ $(m \geq 2)$. Then*

$$\mathrm{def}_m(G) = -2^m - |H_m(G)| + 4|H_{m-1}(G)|.$$

Proof. — Let $H_{m-1}(G) = 2^{m-1}$. Preserving the notations which were introduced in the Proof of Lemma 4, we have here that $R = \varnothing$ and R_1 coinsides with $E_m^+(G) \cup E_m^-(G)$. Lemma 3 asures us that the last relation is independent. By Lemma 1, the inequality of Lemma 4 becomes an exact equality.

Let now $H_{m-1}(G) = 2^{m-1} - 1$. In this case R consists of four pairs, and one can verify that it is independent. Repeating the previous argument, and bearing in mind that the definition of a nondegenerate edge provides the independence of the united relation R_1 we again have an exact equality - instead of the inequality - in Lemma 4.

The fact that the quotient group $G/H(G)$ is cyclic reduces the investigation of the group $G(\Gamma)$ to an investigation of the group $H(G)$. The following lemma shows that in nontrivial situations this group is finitely generated.

Lemma 6. — *Let $|H_m(G)| < 2^m$. Then $H(G) = \mathrm{gp}(u_0, \ldots, u_{m-2})$.*

Proof. — If $m = 1$ then $u_0 = 1$ and $H = 1$. Hence, we may assume that $m \geq 2$. Without loss of generality, we may also assume that $|H_{m-1}(G)| = 2^{m-1}$. By Lemma 4, $\mathrm{def}_m(G) \geq 1$, and thus there exists an irreducible m-pair (S, T) such that G satisfies the equality $S = T$ - implying that G also satisfies the equality $S^{(i)} = T^{(i)}$ for each $i \in \mathbb{Z}$. Therefore, for each $i \in \mathbb{Z}$, $u_i \in \mathrm{gp}(u_{i-m+1}, \ldots, u_{i-1})$ and $u_i \in \mathrm{gp}(u_{i+1}, \ldots, u_{i+m-1})$. Now, using induction on i, one can prove that for each $i \in \mathbb{Z}$, $u_i \in \mathrm{gp}(u_0, \ldots, u_{m-2})$.

It should be noted that in the case $m = 2$ Lemma 6 asserts that the group H is cyclic. (In fact, this assertion is obvious and well known).

3. Identification patterns and their universal groups

Let us consider a finite sequence $\Gamma = \langle E_2^+, E_2^-, \ldots, E_m^+, E_m^- \rangle$ such that the set E_k^+ of its *positive k-edges* and the set of E_k^- of its *negative k-edges* consist of positive and negative irreducible k-pairs, respectively $(2 \leq k \leq m)$. For each $e \in E_k^+ \cup E_k^-$, we define the *initial* vertex of e as $i(e)$ and the *terminal* vertex of e as $t(e)$; so for each $2 \leq k \leq m$ we obtain two oriented graphs with the set of vertices U_{k-2}: the *positive k-graph of* Γ which will be denoted by $(\Gamma)_k^+$, and the *negative k-graph of* Γ which will be denoted by $(\Gamma)_k^-$. We write $e = (w_1, w_2)_k^+$ (or $e = (w_1, w_2)_k^-$) if e is a positive (or a negative) k-edge with the initial vertex w_1 and the terminal vertex w_2. If we need to describe any such sequence in a concrete situation, we do this by enumerating of its edges. Further, we consider the sequence of groups $\{H_k(\Gamma) | 2 \leq k \leq m\}$ which are defined in the set of generators $\{u_i | i \in \mathbb{Z}\}$ by the sets of relations $\bigcup\{\mathcal{R}_k(\Gamma)^{(s)} | s \in \mathbb{Z}\}$, where $\mathcal{R}_k(\Gamma) = \{u_0 i(e) = t(e) u_{p-1}^{\varepsilon(e)} | e \in E_p^+ \cup E_p^-,\ 2 \leq p \leq k\}$, $\varepsilon(e) = 1$ for $e \in E_p^+$ and $\varepsilon(e) = -1$ for $e \in E_p^-$. For each of these groups, the natural epimorphism $\Phi_{\Gamma,k} : U_k \to H_k$ defines the equivalence relation on the group U_k which is denoted by the symbol $\theta_{\Gamma,k}$. Let us denote the quotient-graphs $(\Gamma)^{+}_{k}/\theta_{\Gamma,k-2}$ and $(\Gamma)^{i}_{k}/\theta_{\Gamma,k-2}$ by the symbols $[\Gamma]^{+}_{k}$ and $[\Gamma]^{-}_{k}$, respectively.

As above, we say that an irreducible k-pair (S,T) is *degenerate* (in respect to Γ) if there exists some irreducible $(k-1)$-pair $(P,Q) \in \theta_{\Gamma,k}$ such that one of the words $\tilde{P},\tilde{Q}$ is a subword of one of the words $\tilde{S},\tilde{T}$.

Finally, we call the sequence Γ to be an *identification pattern* if, for each $3 \leq k \leq m$, the set $E^{+}_{k} \cup E^{-}_{k}$ consists of nondegenerate pairs, and each of the graphs $[\Gamma]^{+}_{k}, [\Gamma]^{-}_{k}$ has the property that each of its vertices has at most one incoming and at most one outgoing edge.

For a given identification pattern Γ, we let the symbol $\mathcal{G}(\Gamma)$ denote the class of all groups $G = \mathrm{gp}(a,b)$, such that for each $2 \leq k \leq m$, $E^{+}_{k}(G) \supseteq (\Phi_G \times \Phi_G)(E^{+}_{k})$ and $E^{-}_{k}(G) \supseteq (\Phi_G \times \Phi_G)(E^{-}_{k})$. Let us now define the *universal group* $G(\Gamma)$ of the identification pattern Γ as the infinite cyclic extension of the group $H(\Gamma) = H_m(\Gamma)$ with the naturally defined extending automorphism: $G(\Gamma) = \langle a\rangle \leftthreetimes H(\Gamma)$, $au_ia^{-1} = u_{i+1}$. It is easy to see that for each identification pattern $\Gamma, G(\Gamma) \in \mathcal{G}(\Gamma)$ and $\mathcal{G}(\Gamma)$ consists of all quotient-groups of $G(\Gamma)$. The group $H(\Gamma)$ itself we call the *universal kernel* of Γ.

Example 1. — Let $\Gamma = \langle(1,1)^{+}_{3}, (u_1u_2, u_1u_2)^{+}_{4}\rangle$. Then the universal kernel of Γ has the following presentation: $H(\Gamma) = \langle u_0, u_1 \mid u_0u_1u_0 = u_1u_0u_1\rangle$; and the inner automorphism, afforded by a, acts in the following way: $au_0a^{-1} = u_1, au_1a^{-1} = u_0$. Using the Reidemeister-Schreier method (see, for instance, [8,9]), we see that the group H is the infinite cyclic extension of the free group $K = \langle v_0, v_1\rangle$ with the extending automorphism defined by the equalities $u_1v_0u_1^{-1} = v_1$, $u_1v_1u_1^{-1} = v_0^{-1}v_1$ $(u_0 = v_0u_1)$. Direct calculations show that $|H_3(G(\Gamma))| = 7$ and $|H_4(G(\Gamma))| = 11$.

Example 2. — Let $\Gamma = \langle(1,1)^{+}_{4}, (u_1u_2, u_1u_2)^{+}_{4}\rangle$. Then

$$H(\Gamma) = \langle u_0, u_1, u_2 \mid u_0u_1u_2 = u_1u_2u_0 = u_2u_0u_1\rangle,$$

$au_0a^{-1} = u^1, au_1a^{-1} = u_2$ and $au_2a^{-1} = u_0$. Using Tietze transformations (see, for instance, [8,9]), we have $H(\Gamma) = \langle v_0, v_1, v_2 \mid v_0v_2 = v_2v_0, v_1v_2 = v_2v_1\rangle$, where $u_0 = v_0, u_1 = v_0^{-1}v_1, u_2 = v_1^{-1}v_2$. That is, $H(\Gamma)$ is a direct product of the free group $\langle v_0, v_1\rangle$ of rank two and the infinite cyclic group $\langle v_2\rangle$. In this case we have $|H_3(G(\Gamma))| = 8$ and $|H_4(G(\Gamma))| = 14$.

In an informal way the above examples show that there exist arbitrarily large groups generated by a pair of elements with small third and fourth powers. In precise terms we have the following two theorems:

Theorem 1. — *For each countable (finite) group L, there exists a (finite) group $G = \mathrm{gp}(a,b)$ such that $|\{a,b\}^3| = 7$, $|\{a,b\}^4| = 11$, and the group L is the homomorphic image of a subgroup of G.*

Theorem 2. — *For each countable (finite) group L, there exists a (finite) group $G = \mathrm{gp}(a,b)$ such that $|\{a,b\}^3| = 8$, $|\{a,b\}^4| = 14$, and the group L is the homomorphic image of a subgroup of G.*

Proof. — In order to prove the infinite versions of these theorems, it is enough to note that each of the groups $G(\Gamma)$ in the above examples contains the free group of rank two, and also to realize that each countable group can be embedded into a two-generated group ([6]).

The proofs of the finite versions can be obtained by the method of the proof of Theorem 1 in [10],which asserts that a semidirect product of a residually finite group and a finitely generated residually finite group is residually finite.

Let us prove in details Theorem 1. Let $G_0 = G(\Gamma), H = H(\Gamma)$, and K be the groups from Example 1. First we embed the group L into some two-generated finite group P (we can take, for instance, $P = S_n$ for the relevant permutation group S_n). Consider the two-generator free group $\tilde{F}$ of the variety generated by P and the relevant verbal subgroup M of K, so that $\tilde{F} = K/M$ and $|\tilde{F}| < \infty$ (see, for instance, [12]). The subgroup M is a normal divisor of the group H, the quotient-group H/M is the semidirect product of an infinite cyclic group, and the group $\tilde{F}$: $H/M = \langle u_1 \rangle \lambda \tilde{F}$. Since the group $\tilde{F}$ is finite, the extending automorphism of this semidirect product has finite order, say l, and hence we may consider the semidirect product $P_1 = \langle u_1 \mid u_1^l = 1 \rangle \lambda \tilde{F}$ which is a finite group. We then obtain the following chain of epimorphisms and embeddings:

$$H \twoheadrightarrow P_1 \hookleftarrow \tilde{F} \twoheadrightarrow P \hookleftarrow L.$$

Repeating these considerations, with the usage of P_1 instead of P, H instead of K, and G_0 instead of H, we can extend the above chain to the chain

$$G_0 \twoheadrightarrow P_2 \hookleftarrow \tilde{F}_1 \twoheadrightarrow P_1 \hookleftarrow \tilde{F} \twoheadrightarrow P \hookleftarrow L,$$

where all groups besides G are finite. Now, using Theorem 1 from [10], we may assert that the group G_0 is residually finite. Therefore, it is possible to insert into the last chain the finite group G which satisfies the conditions of Theorem 1:

$$G_0 \twoheadrightarrow G \twoheadrightarrow P_2 \hookleftarrow \tilde{F}_1 \twoheadrightarrow P_1 \hookleftarrow \tilde{F} \twoheadrightarrow P \hookleftarrow L.$$

The proof of the finite version of Theorem 2 is obtained by similar considerations.

Theorems 1 and 2 show that if we want to obtain any definite information about the groups generated by a pair with a small third or fourth power, we need to impose stronger restrictions on the cardinalities of these powers than those used in the above theorems. Noting that in the case where $|H_2(G)| < 4$ the group H is cyclic, we have to investigate only the following situations:

(a) $|H_2(G)| = 4$ and $|H_3(G)| \leq 7$;
(b) $|H_3(G)| = 7$ and $|H_4(G)| \leq 11$;
(c) $|H_3(G)| = 8$ and $|H_4(G)| \leq 14$.

Using lemmas 2,3 and 5, one can easily verify the following three lemmas.

Lemma 7. — *Let $G = \mathrm{gp}(a, b)$ and $|H_2(G)| = 4$. Then $|H_3(G)| < 7$ if and only if G is a quotient of the universal group $G(\Gamma)$ for some identification pattern Γ with two 3-edges (and no other edges).*

Lemma 8. — *Let $G = \mathrm{gp}(a, b)$ and $|H_3(G)| = 7$. Then $|H_4(G)| < 11$ if and only if G is a quotient of the universal group $G(\Gamma)$ for some identification pattern Γ with one 3-edge and two 4-edges (and no other edges).*

Lemma 9. — *Let $G = \mathrm{gp}(a, b)$ and $|H_3(G)| = 8$. Then $|H_4(G)| \leq 16 - k$ if and only if G is a quotient of the universal group $G(\Gamma)$ for some identification pattern Γ with k 4-edges (and no other edges).*

The conditions of lemmas 7 - 9 provide the diagonality of the relations $\theta_{1,\Gamma}, \theta_{2,\Gamma}$, and therefore the graphs $[\Gamma]_3^+$, $[\Gamma]_3^-$, $[\Gamma]_4^+$, $[\Gamma]_4^-$ coinside with the graphs $(\Gamma)_3^+$, $(\Gamma)_3^-$, $(\Gamma)_4^+$, $(\Gamma)_4^-$ respectively. We see now that the problem of describing groups which satisfy the conditions (a)-(c) above is reduced to enumerating the relevant graphs with the sets of vetices $U_1^{(1)}$ and $U_2^{(1)}$ and calculating the relevant universal kernels. The major part of this enumeration can be eliminated by using the considerations below.

For a word $P \in U_k$, we define the *k-complementary* word $\alpha_k(P)$ as the word from U_k such that the set of all u–symbols which occur in $\alpha_k(P)$ is the complement in $\{u_0, \dots u_{k-1}\}$ of the set of all u-symbols which occur in P. If $P = u_{i_1}u_{i_2}\dots u_{i_l}$, we define the *k-opposite* word $\beta_k(P) = u_{k-1-i_l}\dots u_{k-1-i_2}u_{k-1-i_1}$. Extending these mappings componentwise onto the Cartesian square $U_k \times U_k$, we obtain two sign preserving involutions on the set of all irreducible k-pairs which we denote by the same symbols α_k and β_k. It follows from the definitions that these involutions commute, and hence they define an action of the Klein four-group $\mathcal{K}$ on the set of all irreducible k-pairs. Furthermore, for $g \in \mathcal{K}$ and any identification pattern $\Gamma = \langle E_2^+, E_2^-, \dots, E_m^+, E_m^-\rangle$, we define $g(\Gamma) = \langle g(E_2^+), g(E_2^-), \dots, g(E_m^+), g(E_m^-)\rangle$ and so we obtain the action of $\mathcal{K}$ on the set of all identification patterns. We say that two identification patterns are $\mathcal{K}$ – *equivalent* if they belong to the same orbit of this action.

Lemma 10. — *If identification patterns Γ_1 and Γ_2 are $\mathcal{K}$-equivalent, then $H(\Gamma_1) \cong H(\Gamma_2)$ and $G(\Gamma_1) \cong G(\Gamma_2)$.*

Proof. — In order to prove this lemma it is enough to note that the map α is the restriction of the automorphism of the free group $F = \langle a, b\rangle$ defined by the rule $a \mapsto b, b \mapsto a$, and the map β is the restriction of the composition of the automorphism of F defined by the rule $a \mapsto a^{-1}, b \mapsto b^{-1}$ and the group inversion $g \mapsto g^{-1}$.

4. Main results

Now we turn directly to the problem of calculating the universal kernel for a given identification pattern Γ. By Lemma 6, $H(\Gamma)$ is finitely generated, but it is not necessarily finitely presented. Let us denote by the symbol $H^{[n]}(\Gamma)$ the group which is defined in the set of generators $\{u_i | 0 \leq i \leq n-1\}$ by the set of all relations from the union $\bigcup\{\mathcal{R}_m(\Gamma)^{(s)} | s \in \mathbb{Z}\}$ which contain only the symbols $u_0, \dots, u_{n-1}$; we call this group the *n-particular kernel* of Γ. The group $H(\Gamma)$ is the direct limit of the family of groups $\{H^{[n]}(\Gamma) | n > 0\}$; if we have, for some n, $H^{[n]}(\Gamma) \cong H^{[n+1]}(\Gamma)$, and the group $H^{[n]}(\Gamma)$ is hopfian, then we may conclude that $H(\Gamma) = H^{[n]}(\Gamma)$. The lists

1 and 2 of the universal kernels in the Appendix are obtained using this argument: for each type of identification patterns which appears in lemmas 7-9, we enumerate up to $\mathcal{K}$-equivalence identification patterns of the given type, and calculate $H^{[5]}(\Gamma)$ and $H^{[6]}(\Gamma)$ (taking into account only the patterns for which the order of $H^{[n]}(\Gamma)$ is large enough). It is shown by the calculations that the first of above conditions holds for each identification pattern of those types. On the other hand, all of the groups in these lists are finite, except the first one in List 1 and the second one in List 2; yet these two groups are finite extensions of residaully finite groups and so they are residually finite themselves. Therefore, all groups in the lists 1 and 2 are hopfian, so these lists present the exact description of the needed universal kernels. List 3 is obtained in the similar way using $H^{[7]}(\Gamma)$ and $H^{[8]}(\Gamma)$. For a few identificational patterns of this type it turns out that $H^{[7]}(\Gamma) \neq H^{[8]}(\Gamma)$. In this case we also calculate $H^{[9]}(\Gamma)$, and have $H^{[8]}(\Gamma) = H^{[9]}(\Gamma)$. Again, all of these groups are residually finite, and therefore they are hopfian. In order to prove this assertion, we can apply the same line of argument, or, in some cases, Malčev's theorem, which is mentioned in the proof of Theorem 1.

Summarizing the information which is contained in the mentioned lists, and bearing in mind lemmas 7-9, we obtain the following theorems:

Theorem 3. — *Let $G = gp(a,b)$, $|\{a,b\}^2| = 4$ and $|\{a,b\}^3| < 7$. Then the normal subgroup $H = (ab^{-1})^G$ of the group G, generated by the element ab^{-1}, is isomorphic to one of the following groups:*

a) cyclic group of order 5;

b) direct product of two cyclic groups of the same order p $(2 \leq p \leq \infty)$;

c) dihedral group of order greater than 2;

d) quaternion group.

All these possibilities are realizable.

Proof. — By Lemma 7, the group G satisfies the conditions of the Theorem if and only if its subgroup H is a homomorphic image of some group in List 1 in the Appendix. Taking into account that groups number 3,7 and 9 are all isomorphic to the quaternion group, we see that all homomorphic images of the groups 1,3,4,6,7,8 and 9, which have at least four elements, is one of the groups described in items *a),c),d)*. The groups 2 and 5 are free abelian of rank two, and it is easy to verify that their free generators are conjugated by the element a. Thus images of these generators are conjugated in each quotient-group of the universal group $G(\Gamma)$. Therefore, these quotient-groups satisfy the condition *b)* for H. A similar situation holds also for group 1 in List 1: it is the free product of two groups of order two which are conjugated by a. It follows that each normal subgroup P of this group is a-invariant (that is P is normal in the group $G(\Gamma)$) and hence for each dihedral group $\hat{H}$ there exists homomorphic image $\hat{G}$ of $G(\Gamma)$ with $H(\hat{G}) \cong \hat{H}$.

Since the condition of $|H| < 4$ implies the cyclicity of H (Lemma 6) for all groups satisfying *b),c),d)* we have that $|H| = 4$. Group 4 satisfies the condition *a)* which may be checked directly.

Corollary 1. — *Let $G = gp(a,b)$, $|\{a,b\}^3| < 7$. Then the group G is solvable of derived length not greater than three.*

Theorem 4. — *Let $G = gp(a,b)$, $|\{a,b\}^3| = 7$ and $|\{a,b\}^4| < 11$. Then the normal subgroup $H = (ab^{-1})^G$ of the group G, generated by the element ab^{-1}, is isomorphic to one of the following groups:*
a) cyclic group of order 7;
b) direct product of two cyclic groups of the same order p $(3 \leq p \leq \infty)$;
c) direct product of cyclic groups of orders 2 and 4;
d) quaternion group;
e) nonabelian semidirect product of cyclic group of even finite or infinite order with a cyclic group of order 3;
f) nonabelian semidirect product of cyclic group of order 3 with a Klein four-group;
g) group defined by presentation $\langle x, y \mid x^2 = y^2, (xy)^2 = 1\rangle$ (extension of cyclic group of order 4 by group of order 2);
h) special linear group $\mathbf{SL}(2,3)$;
All these possibilities are realizable.

Proof. — At first let us remark that there exist identification patterns with one 3-edge and two 4-edges such that $H(\Gamma)$ is the cyclic group of order 7 (we may take as the example $\Gamma = \langle (u_1,1)_3^-, (1,u_1)_4^-, (u_1,u_2)_4^- \rangle$) and so there exists a group G satisfying the conditions of the Theorem such that H satisfies the condition *a)*. If H is not isomorphic to the cyclic group of order 7 then, by Lemma 8, it must by a homomorphic image of some group in List 2. Taking into account that the group 2 has the presentation $\langle v, w \mid w^3 = 1, v^{-1}wv = w^{-1}\rangle$, that the groups 12 and 19 are isomorphic to the quaternion group and that the group 5 is isomorphic to $\mathbf{SL}(2,3)$ (the last isomorphism can be defined by the rule $x \mapsto \left[\begin{smallmatrix}-1 & 1\\ 0 & -1\end{smallmatrix}\right], y \mapsto \left[\begin{smallmatrix}-1 & 0\\ 1 & -1\end{smallmatrix}\right]$), we see that each group in List 2 satisfies one of the conditions *b)-h)*. In order to complete the proof, it remains to make the following observations: it is possible to apply to groups 2,14 and 18 considerations similar to those which were applied to the groups 1,2 and 5 in the proof of Theorem 3; the unique quotient-group of $\mathbf{SL}(2,3)$ which has order greater than 7 satisfies condition *f)*; and the unique quotient-group of the group 1 which has order greater than 7, satisfies condition *c)*.

Corollary 2. — *Let $G = gp(a,b)$, $|\{a,b\}^3| = 7$ and $|\{a,b\}^4| < 11$. Then the group G is solvable of derived length not greater than four.*

Theorem 5. — *Let $G = gp(a,b)$, $|\{a,b\}^3| = 8$ and $|\{a,b\}^4| < 14$. Then either the normal subgroup $H = (ab^{-1})^G$ of the group G generated by the element ab^{-1} is solvable of derived length not greater than three, or it is a central extension of a cyclic group of order not greater two by the alternating group A_5.*

Proof. — By Lemma 9, our group is a homomorphic image of some group in List 3. All these groups except group 17 are solvable of derived length not greater than three. Group 17 is presented in this list in the following way: $H = \langle x, y \mid xyx = yxy, xyx^{-1}yx = y^2\rangle$, but in generators $v = xy$ and $w = xyx$ it has the presentation

$\langle v, w \mid v^3 = w^2, (vw^{-1})^5 w^2 = 1\rangle$. The quotient-group of this group by the central cyclic subgroup generated by the element w^2 is isomorphic (as was proved in ([11]) to the alternating group A_5. Moreover, let us define homomorphism $\varphi : H \to A_5$ by the rule $v \mapsto (135)$ and $w \mapsto (12)(34)$. A computation using the Reidemeister-Schreier method shows that φ is an epimorphism, and its kernel is isomorphic to a cyclic group of order two.

Remark. — Using List 3, one can make a full classification of groups which satisfy the conditions of Theorem 5 as it has done in theorems 3 and 4, but it seems to be too extensive for our liking.

Corollary 3. — *Let $G = gp(a, b)$, $|\{a, b\}^3| = 8$ and $|\{a, b\}^4| < 14$. Then either the group G is solvable of derived length not greater than four, or it has an invariant series $1 \trianglelefteq N \triangleleft H \trianglelefteq G$ such that N is a cyclic central subgroup of H, $H/N \cong A_5$ and G/H is cyclic of order not greater two.*

Theorem 6. — *Let $G = gp(a, b)$, $|\{a, b\}^3| = 8$ and $|\{a, b\}^4| < 13$. Then the normal subgroup $H = (ab^{-1})^G$ of the group G generated by the element ab^{-1} is solvable of derived length not greater than three.*

Proof. — If G satisfies the conditions of the Theorem, but H does not satisfy its conclusion, then, by Lemma 9 and properties of groups in lists 3,4, it must be a quotient-group of some identification pattern with four 3-edges such that all of them are contained in the union of $\mathcal{K}$-orbit of the identification pattern number 17 in List 3. This union consists of edges $(1,1)_4^+$, $(u_1, u_1u_2)_4^+$, $(u_1u_2, u_2)_4^+$, $(1, u_2)_4^+$, $(u_1, 1)_4^+$, $(u_1u_2, u_1u_2)_4^+$ and it is easy to see that it is impossible to construct any four-element identification pattern of these edges.

Corollary 4. — *Let $G = gp(a, b)$, $|\{a, b\}^3| = 8$ and $|\{a, b\}^4| < 13$. Then the group G is solvable of derived length not greater than four.*

Corollary 5. — *Let $G = gp(a, b)$ and $|\{a, b\}^4| < 11$. Then the group G is solvable of derived length not greater than four.*

5. Appendix

Below are given results of mechanical computations of the universal kernels for the identification patterns which appear in the proofs of the theorems of the last section. These tables use the notation abel$(m_1, \ldots, m_k)$ for the direct product of k cyclic groups of orders $m_1, \ldots, m_k$ $(2 \leq m_i \leq \infty)$.

List 1
Universal kernels of the identification patterns with two 3-edges, only for $|H(\Gamma)| > 4$.

N	Γ	$H(\Gamma)$
1	$\langle (1,1)_3^+, (1,1)_3^- \rangle$	$H = \langle x, y \mid x^2, y^2 \rangle, H' = \text{abel}(1);$
2	$\langle (1,1)_3^+, (u_1,u_1)_3^+ \rangle$	$H = \text{abel}(\infty,\infty);$
3	$\langle (1,1)_3^+, (u_1,u_1)_3^- \rangle$	$H = \langle x, y \mid yxyx^{-1}, xyxy^{-1} \rangle,$ $H/H' = \text{abel}(2,2), H' = \text{abel}(2);$
4	$\langle (1,1)_3^-, (1,u_1)_3^+ \rangle$	$H = \text{abel}(5);$
5	$\langle (1,1)_3^-, (u_1,u_1)_3^- \rangle$	$H = \text{abel}(\infty,\infty);$
6	$\langle (1,u_1)_3^+, (1,u_1)_3^- \rangle$	$H = \text{abel}(2,2);$
7	$\langle (1,u_1)_3^+, (u_1,1)_3^+ \rangle$	$H = \langle x, y \mid y^2x^{-2}, yxyx^{-1}, yxy^{-1}x \rangle,$ $H/H' = \text{abel}(2,2), H' = \text{abel}(2);$
8	$\langle (1,u_1)_3^+, (u_1,1)_3^- \rangle$	$H = \text{abel}(2,2);$
9	$\langle (1,u_1)_3^-, (u_1,1)_3^- \rangle$	$H = \langle x, y \mid yxyx^{-1}, y^2x^2, xy^{-1}xy \rangle,$ $H/H' = \text{abel}(2,2), H' = \text{abel}(2)$

List 2
Universal kernels of the identification patterns with one 3-edge and two 4-edges, only for $|H(\Gamma)| > 7$.

N	Γ	$H(\Gamma)$
1	$\langle (1,1)_3^+, (u_2,u_1)_4^+,$ $(u_1u_2,1)_4^- \rangle$	$H = \langle x, y \mid y^2x^{-2}, yxyx \rangle,$ $H/H' = \text{abel}(2,4); H' = \text{abel}(2);$
2	$\langle (1,1)_3^+, (u_2,u_1)_4^+,$ $(u_1u_2,u_1u_2)_4^+ \rangle$	$H = \langle x, y \mid y^2x^{-2}, yxyx^{-1}y^{-1}x^{-1} \rangle,$ $H/H' = \text{abel}(\infty), H' = \text{abel}(3);$
3	$\langle (1,1)_3^+, (u_2,u_1)_4^+,$ $(u_1u_2,u_1u_2)_4^- \rangle$	$H = \langle x, y \mid y^2x^{-2}, yxyxy^{-1}x^{-1}, y^2x^2 \rangle,$ $H/H' = \text{abel}(4); H' = \text{abel}(3);$
4	$\langle (1,1)_3^+, (u_2,u_1u_2)_4^+,$ $(u_1u_2,1)_4^- \rangle$	$H = \langle x, y \mid xyxy^{-1}xy^{-1}, y^2, x^3 \rangle,$ $H/H' = \text{abel}(3); H' = \text{abel}(2,2);$

List 2
(continuation)

N	Γ	$H(\Gamma)$
5	$\langle (1,1)_3^+, (u_2, u_1u_2)_4^+,$ $(u_1u_2, u_1)_4^+ \rangle$	$H = \langle x, y \mid xyxy^{-2}, yxyx^{-2} \rangle, H/H' = \text{abel}(3);$ $H'/H'' = \text{abel}(2,2); H'' = \text{abel}(2);$
6	$\langle (1,1)_3^+, (u_1u_2, 1)_4^-,$ $(u_1u_2, u_1)_4^+ \rangle$	$H = \langle x, y \mid xyxy^{-1}xy^{-1}, y^2, x^3 \rangle,$ $H/H' = \text{abel}(3); H' = \text{abel}(2,2);$
7	$\langle (1,1)_3^-, (1, u_1u_2)_4^-,$ $(u_1, u_2)_4^+ \rangle$	$H = \text{abel}(2,4);$
8	$\langle (1,u_1)_3^+, (u_1, 1)_4^-,$ $(u_2, 1)_4^+ \rangle$	$H = \text{abel}(3,3);$
9	$\langle (1,u_1)_3^+, (u_1, 1)_4^-,$ $(u_1u_2, u_1)_4^+ \rangle$	$H = \text{abel}(3,3);$
10	$\langle (1,u_1)_3^+, (u_1, 1)_4^-,$ $(u_1u_2, u_2)_4^- \rangle$	$H = \text{abel}(3,3);$
11	$\langle (1,u_1)_3^+, (u_2, 1)_4^+,$ $(u_1u_2, u_1)_4^+ \rangle$	$H = \text{abel}(3,3);$
12	$\langle (1,u_1)_3^-, (1, 1)_4^+,$ $(u_1u_2, u_1u_2)_4^+ \rangle$	$H = \langle x, y \mid xyx^{-1}y, yxy^{-1}x, x^{-2}y^2 \rangle,$ $H/H' = \text{abel}(2,2); H' = \text{abel}(2);$
13	$\langle (1,u_1)_3^-, (1, 1)_4^-,$ $(u_1, u_2)_4^+ \rangle$	$H = \text{abel}(3,3);$
14	$\langle (1,u_1)_3^-, (1, 1)_4^-,$ $(u_1u_2, u_1u_2)_4^- \rangle$	$H = \text{abel}(\infty,\infty);$
15	$\langle (1,u_1)_3^-, (u_1, u_2)_4^+,$ $(u_1u_2, 1)_4^- \rangle$	$H = \langle x, y \mid y^3, x^3, xyxy \rangle,$ $H/H' = \text{abel}(3); H' = \text{abel}(2,2);$
16	$\langle (u_1,1)_3^-, (1, u_1u_2)_4^-,$ $(u_1, u_1)_4^+ \rangle$	$H = \text{abel}(3,3);$
17	$\langle (u_1,1)_3^-, (1, u_1u_2)_4^-,$ $(u_2, u_1)_4^+ \rangle$	$H = \langle x, y \mid y^3, x^3, xy^{-1}xy^{-1} \rangle,$ $H/H' = \text{abel}(3); H' = \text{abel}(2,2);$
18	$\langle (u_1,1)_3^-, (u_1, u_1)_4^+,$ $(u_2, u_2)_4^+ \rangle$	$H = \text{abel}(\infty,\infty);$
19	$\langle (u_1,1)_3^-, (u_1, u_1)_4^-,$ $(u_2, u_2)_4^- \rangle$	$H = \langle x, y \mid xyxy^{-1}, x^{-1}yxy \rangle,$ $H/H' = \text{abel}(2,2); H' = \text{abel}(2)$

List 3
Universal kernels of the identification patterns with three 4-edges,
only for $|H(\Gamma)| > 7$.

N	Γ	$H(\Gamma)$
1	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(1,1)_4^- \rangle$	$H = \text{abel}(2,2,2);$
2	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(1,u_1u_2)_4^- \rangle$	$H = \text{abel}(3,\infty);$
3	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_1,1)_4^- \rangle$	$H = \text{abel}(9);$
4	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_1,u_1)_4^- \rangle$	$H = \text{abel}(2,2,2);$
5	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_1,u_2)_4^- \rangle$	$H = \text{abel}(14);$
6	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_2,1)_4^- \rangle$	$H = \text{abel}(9);$
7	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_2,u_1)_4^- \rangle$	$H = \text{abel}(14);$
8	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_2,u_2)_4^+ \rangle$	$H = \text{abel}(\infty,\infty,\infty);$
9	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_2,u_2)_4^- \rangle$	$H = \text{abel}(2,2,2);$
10	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_1u_2,u_1)_4^- \rangle$	$H = \text{abel}(9);$
11	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_1u_2,u_2)_4^- \rangle$	$H = \text{abel}(9);$
12	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_1u_2,u_1u_2)_4^+ \rangle$	$H = \text{abel}(\infty,\infty,\infty);$
13	$\langle (1,1)_4^+, (u_1,u_1)_4^+,$ $(u_1u_2,u_1u_2)_4^- \rangle$	$H = \text{abel}(2,2,2);$
14	$\langle (1,1)_4^+, (u_1,u_2)_4^+,$ $(1,u_1u_2)_4^- \rangle$	$H = \text{abel}(\infty);$
15	$\langle (1,1)_4^+, (u_1,u_2)_4^+,$ $(u_2,u_1)_4^+ \rangle$	$H = \langle x,y \mid y^3, x^{-1}yxy \rangle,$ $H/H' = \text{abel}(\infty), H' = \text{abel}(3);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
16	$\langle(1,1)_4^+,(u_1,u_2)_4^+,$ $(u_1u_2,u_1u_2)_4^+\rangle$	$H=\text{abel}(\infty);$
17	$\langle(1,1)_4^+,(u_1,u_1u_2)_4^+,$ $(u_1u_2,u_2)_4^+\rangle$	$H=\langle x,y\mid xyxy^{-1}x{-}1y{-}1,xyx^{-1}yxy^{-2}\rangle,$ $H/H'=1;$
18	$\langle(1,1)_4^+,(u_2,u_1)_4^+,$ $(1,u_1u_2)_4^-\rangle$	$H=\langle x,y\mid yx^{-1}yx^{-1},x^2yx^{-1}y^{-2}\rangle,$ $H/H'=\text{abel}(\infty),H'=\text{abel}(2,2);$
19	$\langle(1,1)_4^+,(u_2,u_1)_4^+,$ $(u_1u_2,u_1u_2)_4^+\rangle$	$H=\langle x,y\mid yxyx^{-1}y^{-1}x^{-1},y^3x^{-1}y^{-3}x\rangle,$ $H/H'=\text{abel}(\infty),$ $H'/H''=\text{abel}(2,2),H''=\text{abel}(2);$
20	$\langle(1,1)_4^+,(u_1u_2,u_1u_2)_4^+,$ $(1,1)_4^-\rangle$	$H=\langle x,y,z\mid x^2,y^2,z^2,zxyzyx,xyzxzy,$ $yzxyxz\rangle,\ H/H'=\text{abel}(2,2,2),H'=\text{abel}(2);$
21	$\langle(1,1)_4^+,(u_1u_2,u_1u_2)_4^+,$ $(1,u_1u_2)_4^-\rangle$	$H=\langle x,y\mid y^2x^{-1}y^2x^{-1}y^{-1}x^{-1},x^3y^{-3}\rangle,$ $H/H'=\text{abel}(3,\infty),H'=\text{abel}(\infty,\infty);$
22	$\langle(1,1)_4^+,(u_1u_2,u_1u_2)_4^+,$ $(u_1,1)_4^-\rangle$	$H=\text{abel}(9);$
23	$\langle(1,1)_4^+,(u_1u_2,u_1u_2)_4^+,$ $(u_1,u_1)_4^-\rangle$	$H=\langle x,y,z\mid zxzx^{-1},yzy^{-1}z,xyxy^{-1},$ $zxz^{-1}x,yzyz^{-1},xyx^{-1}y\rangle,$ $H/H'=\text{abel}(2,2,2),H'=\text{abel}(2);$
24	$\langle(1,1)_4^+,(u_1u_2,u_1u_2)_4^+,$ $(u_1,u_2)_4^-\rangle$	$H=\text{abel}(14);$
25	$\langle(1,u_1)_4^+,(u_1,1)_4^+,$ $(1,u_1)_4^-\rangle$	$H=\text{abel}(2,2,2);$
26	$\langle(1,u_1)_4^+,(u_1,1)_4^+,$ $(u_1,1)_4^-\rangle$	$H=\text{abel}(2,2,2);$
27	$\langle(1,u_1)_4^+,(u_1,1)_4^+,$ $(u_2,u_1u_2)_4^+\rangle$	$H=\text{abel}(2,2,2);$
28	$\langle(1,u_1)_4^+,(u_1,1)_4^+,$ $(u_2,u_1u_2)_4^-\rangle$	$H=\text{abel}(2,2,2);$
29	$\langle(1,u_1)_4^+,(u_1,1)_4^+,$ $(u_1u_2,u_2)_4^+\rangle$	$H=\text{abel}(2,2,2);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
30	$\langle (1,u_1)_4^+,(u_1,1)_4^+,$ $(u_1u_2,u_2)_4^-\rangle$	$H=\text{abel}(2,2,2);$
31	$\langle (1,u_1)_4^+,(u_1,u_2)_4^+,$ $(1,1)_4^-\rangle$	$H=\text{abel}(9);$
32	$\langle (1,u_1)_4^+,(u_1,u_2)_4^+,$ $(u_1,u_1)_4^-\rangle$	$H=\text{abel}(9);$
33	$\langle (1,u_1)_4^+,(u_1,u_2)_4^+,$ $(u_2,1)_4^+\rangle$	$H=\langle x,y \mid yx^{-2}yx, y^{-2}xyx\rangle,$ $H/H'=\text{abel}(3),$ $H'/H''=\text{abel}(2,2), H''=\text{abel}(2);$
34	$\langle (1,u_1)_4^+,(u_1,u_2)_4^+,$ $(u_2,u_2)_4^-\rangle$	$H=\text{abel}(9);$
35	$\langle (1,u_1)_4^+,(u_1,u_2)_4^+,$ $(u_2,u_1u_2)_4^+\rangle$	$H=\text{abel}(9);$
36	$\langle (1,u_1)_4^+,(u_1,u_2)_4^+,$ $(u_1u_2,1)_4^-\rangle$	$H=\langle x,y \mid yx^{-1}yxyx^{-1}, x^2, y^3\rangle,$ $H/H'=\text{abel}(3), H'=\text{abel}(2,2);$
37	$\langle (1,u_1)_4^+,(u_1,u_2)_4^+,$ $(u_1u_2,u_1u_2)_4^-\rangle$	$H=\text{abel}(9);$
38	$\langle (1,u_1)_4^+,(u_2,1)_4^+,$ $(u_1u_2,1)_4^-\rangle$	$H=\langle x,y \mid x^3, yxyx, y^3\rangle,$ $H/H'=\text{abel}(3), H'=\text{abel}(2,2);$
39	$\langle (1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(1,1)_4^-\rangle$	$H=\text{abel}(9);$
40	$\langle (1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(1,u_1)_4^-\rangle$	$H=\text{abel}(2,2,2);$
41	$\langle (1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(1,u_2)_4^-\rangle$	$H=\text{abel}(11);$
42	$\langle (1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(u_1,1)_4^-\rangle$	$H=\text{abel}(2,2,2);$
43	$\langle (1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(u_1,u_1)_4^-\rangle$	$H=\text{abel}(9);$
44	$\langle (1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(u_1,u_2)_4^-\rangle$	$H=\text{abel}(17);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
45	$\langle(1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(u_2,1)_4^-\rangle$	$H=\mathrm{abel}(3,3);$
46	$\langle(1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(u_2,u_1)_4^-\rangle$	$H=\mathrm{abel}(13);$
47	$\langle(1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(u_1u_2,1)_4^+\rangle$	$H=\mathrm{abel}(11);$
48	$\langle(1,u_1)_4^+,(u_2,u_1u_2)_4^+,$ $(u_1u_2,u_2)_4^+\rangle$	$H=\mathrm{abel}(2,2,2);$
49	$\langle(1,u_1)_4^+,(u_1u_2,1)_4^+,$ $(1,u_2)_4^-\rangle$	$H=\mathrm{abel}(11);$
50	$\langle(1,u_1)_4^+,(u_1u_2,1)_4^+,$ $(u_1,u_1u_2)_4^-\rangle$	$H=\mathrm{abel}(11);$
51	$\langle(1,u_1)_4^+,(u_1u_2,u_2)_4^+,$ $(1,u_1)_4^-\rangle$	$H=\mathrm{abel}(2,2,2);$
52	$\langle(1,u_1)_4^+,(u_1u_2,u_2)_4^+,$ $(u_1,1)_4^-\rangle$	$H=\mathrm{abel}(2,2,2);$
53	$\langle(1,u_1)_4^+,(u_1u_2,u_2)_4^+,$ $(u_2,u_1u_2)_4^-\rangle$	$H=\mathrm{abel}(2,2,2);$
54	$\langle(1,u_1)_4^+,(u_1u_2,u_2)_4^+,$	$H=\mathrm{abel}(2,2,2);$
55	$\langle(1,u_2)_4^+,(u_1,u_1u_2)_4^+,$ $(1,u_2)_4^-\rangle$	$H=\mathrm{abel}(2,2,2);$
56	$\langle(1,u_2)_4^+,(u_1,u_1u_2)_4^+,$ $(u_1,1)_4^-\rangle$	$H=\mathrm{abel}(11);$
57	$\langle(1,u_2)_4^+,(u_1,u_1u_2)_4^+,$ $(u_1,u_2)_4^-\rangle$	$H=\mathrm{abel}(19);$
58	$\langle(1,u_2)_4^+,(u_1,u_1u_2)_4^+,$ $(u_2,1)_4^+\rangle$	$H=\mathrm{abel}(2,2,2);$
59	$\langle(1,u_2)_4^+,(u_1,u_1u_2)_4^+,$ $(u_2,1)_4^-\rangle$	$H=\mathrm{abel}(2,2,2);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
60	$\langle(1,u_2)_4^+,(u_1,u_1u_2)_4^+,$ $(u_2,u_1)_4^-\rangle$	$H=\text{abel}(11);$
61	$\langle(1,u_1u_2)_4^+,(u_1,u_1)_4^+,$ $(1,u_1u_2)_4^-\rangle$	$H=\langle x,y \mid x^2,yxyx,y^5\rangle,$ $H/H'=\text{abel}(2),H'=\text{abel}(5);$
62	$\langle(1,u_1u_2)_4^+,(u_1,u_1)_4^+,$ $(u_1,u_1)_4^-\rangle$	$H=\langle x,y \mid y^2,x^2,xyxyxyxyxy\rangle,$ $H/H'=\text{abel}(2),H'=\text{abel}(5);$
63	$\langle(1,u_1u_2)_4^+,(u_1,u_1)_4^+,$ $(u_1u_2,1)_4^+\rangle$	$H=\langle x,y \mid y^2,yx^{-1}yx^{-1},x^4yx^{-1}y\rangle,$ $H/H'=\text{abel}(2),H'=\text{abel}(5);$
64	$\langle(1,u_1u_2)_4^+,(u_1,u_2)_4^+,$ $(1,u_1u_2)_4^-\rangle$	$H=\text{abel}(2,2,2);$
65	$\langle(1,u_1u_2)_4^+,(u_1,u_2)_4^+,$ $(u_1,u_2)_4^-\rangle$	$H=\text{abel}(2,2,2);$
66	$\langle(1,u_1u_2)_4^+,(u_1,u_2)_4^+,$ $(u_2,u_1)_4^+\rangle$	$H=\text{abel}(2,2,2);$
67	$\langle(1,u_1u_2)_4^+,(u_1,u_2)_4^+,$ $(u_2,u_1)_4^-\rangle$	$H=\text{abel}(2,2,2);$
68	$\langle(1,u_1u_2)_4^+,(u_1,u_2)_4^+,$ $(u_1u_2,1)_4^+\rangle$	$H=\langle x,y,z \mid zxzx^{-1},yzyz^{-1},x^2y^{-2},$ $yxyx^{-1},yxy^{-1}x,z^2y^{-2},zxz^{-1}x,yzy^{-1}z\rangle,$ $H/H'=\text{abel}(2,2,2),H'=\text{abel}(2);$
69	$\langle(1,u_1u_2)_4^+,(u_1,u_2)_4^+,$ $(u_1u_2,1)_4^-\rangle$	$H=\text{abel}(2,2,2);$
70	$\langle(1,u_1u_2)_4^+,(u_2,u_1)_4^+,$ $(1,u_1u_2)_4^-\rangle$	$H=\text{abel}(2,2,2);$
71	$\langle(1,u_1u_2)_4^+,(u_2,u_1)_4^+,$ $(u_1,u_2)_4^-\rangle$	$H=\text{abel}(2,2,2);$
72	$\langle(1,u_1u_2)_4^+,(u_2,u_1)_4^+,$ $(u_2,u_1)_4^-\rangle$	$H=\text{abel}(2,2,2);$
73	$\langle(1,u_1u_2)_4^+,(u_2,u_1)_4^+,$ $(u_1u_2,1)_4^+\rangle$	$H=\langle x,y,z \mid x^2z^2,z^2y^2,z^{-1}yzy,$ $z^{-1}xzx,yx^{-1}yx,x^{-1}zxz,xzy^{-1}xzy^{-1}\rangle,$ $H/H'=\text{abel}(2,2,2),H'=\text{abel}(2);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
74	$\langle(1,u_1u_2)_4^+,(u_2,u_1)_4^+,$ $(u_1u_2,1)_4^-\rangle$	$H=\text{abel}(2,2,2);$
75	$\langle(1,u_1u_2)_4^+,(u_1u_2,1)_4^+,$ $(1,u_1u_2)_4^-\rangle$	$H=\langle x,y,z\mid x^2,z^2,yzyz,yzxzyzxz,yxyx\rangle,$ $H/H'=\text{abel}(2,2,2),H'=\text{abel}(\infty,\infty);$
76	$\langle(1,u_1u_2)_4^+,(u_1u_2,1)_4^+,$ $(u_1,u_1)_4^-\rangle$	$H=\langle x,y\mid x^2,y^2,xyxyxyxyxy\rangle,$ $H/H'=\text{abel}(2),H'=\text{abel}(5);$
77	$\langle(1,u_1u_2)_4^+,(u_1u_2,1)_4^+,$ $(u_1,u_2)_4^-\rangle$	$H=\text{abel}(2,2,4);$
78	$\langle(1,u_1u_2)_4^+,(u_1u_2,1)_4^+,$ $(u_1u_2,1)_4^-\rangle$	$H=\langle x,y,z\mid x^2,y^2,xzxz,zyzy\rangle,$ $H/H'=\text{abel}(2,2,2),H'=\text{abel}(\infty,\infty);$
79	$\langle(u_1,u_1)_4^+,(u_2,u_2)_4^+,$ $(1,1)_4^-\rangle$	$H=\langle x,y,z\mid zx^{-1}zx,y^{-1}xyx,$ $z^{-1}yzy,zy^{-1}zy,z^{-1}xzx,xyx^{-1}y\rangle,$ $H/H'=\text{abel}(2,2,2),H'=\text{abel}(2);$
80	$\langle(u_1,u_1)_4^+,(u_2,u_2)_4^+,$ $(1,u_1u_2)_4^-\rangle$	$H=\text{abel}(3,\infty);$
81	$\langle(u_1,u_1)_4^+,(u_2,u_2)_4^+,$ $(u_1,1)_4^-\rangle$	$H=\text{abel}(9);$
82	$\langle(u_1,u_1)_4^+,(u_2,u_2)_4^+,$ $(u_1,u_1)_4^-\rangle$	$H=\langle x,y,z\mid x^2,y^2,xzxz,zxyxz^{-1}xyx,$ $yxzyxz^{-1},zxyz^{-1}xy,z^{-1}yzy\rangle,$ $H/H'=\text{abel}(2,2,2),H'=\text{abel}(2);$
83	$\langle(u_1,u_1)_4^+,(u_2,u_2)_4^+,$ $(u_1,u_2)_4^-\rangle$	$H=\text{abel}(14);$
84	$\langle(u_1,u_2)_4^+,(u_2,u_1)_4^+,$ $(1,1)_4^-\rangle$	$H=\langle x,y\mid yx^{-1}yx,x^2y^{-1}x^2y,y^3\rangle,$ $H/H'=\text{abel}(4),H'=\text{abel}(3);$
85	$\langle(u_1,u_2)_4^+,(u_2,u_1)_4^+,$ $(1,u_1u_2)_4^-\rangle$	$H=\text{abel}(2,2,\infty);$
86	$\langle(u_1,u_2)_4^+,(u_2,u_1)_4^+,$ $(u_1,u_2)_4^-\rangle$	$H=\langle x,y,z\mid z^2,xz^{-1}xz^{-1},xzxz,y^2,xyxy^{-1}\rangle,$ $H/H'=\text{abel}(2,2,2),H'=\text{abel}(\infty,\infty);$
87	$\langle(u_1,u_2)_4^+,(u_2,u_1)_4^+,$ $(u_1u_2,1)_4^-\rangle$	$H=\langle x,y,z\mid y^2,zyxz^{-1}yx^{-1},xyzxyz,$ $z^2yx^{-1}yx^{-1},yzyx^{-2}z,z^{-1}x^{-1}zx\rangle,$ $H/H'=\text{abel}(2,2,4),H'=\text{abel}(2);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
88	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(1,1)_4^+\rangle$	$H = \text{abel}(2,2,2);$
89	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(1,u_1)_4^+\rangle$	$H = \text{abel}(9);$
90	$\langle (1,1)_4^-, (u_1,u_1)_4^-, \rangle$ $(1,u_1u_2)_4^+$	$H = \text{abel}(2,7);$
91	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_1,u_1)_4^+\rangle$	$H = \text{abel}(2,2,2);$
92	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_1,u_2)_4^+\rangle$	$H = \text{abel}(3,\infty);$
93	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_2,1)_4^+\rangle$	$H = \text{abel}(9);$
94	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_2,u_2)_4^+\rangle$	$H = \text{abel}(2,2,2);$
95	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_2,u_2)_4^-\rangle$	$H = \text{abel}(\infty,\infty,\infty);$
96	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_2,u_1u_2)_4^+\rangle$	$H = \text{abel}(9);$
97	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_1u_2,1)_4^+\rangle$	$H = \text{abel}(14);$
98	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_1u_2,u_1)_4^+\rangle$	$H = \text{abel}(9);$
99	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_1u_2,u_1u_2)_4^+\rangle$	$H = \text{abel}(2,2,2);$
100	$\langle (1,1)_4^-, (u_1,u_1)_4^-,$ $(u_1u_2,u_1u_2)_4^-\rangle$	$H = \text{abel}(\infty,\infty,\infty);$
101	$\langle (1,1)_4^-, (u_1u_2,u_1u_2)_4^-,$ $(1,1)_4^+\rangle$	$H = \langle x,y,z \mid x^2,y^2,z^2,zxyzyx,$ $xyzxzy,yzxyxz,zxyzyx\rangle,$ $H/H' = \text{abel}(2,2,2), H' = \text{abel}(2);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
102	$\langle (1,1)_4^-, (u_1u_2, u_1u_2)_4^-,$ $(1, u_1)_4^+ \rangle$	$H = \text{abel}(9);$
103	$\langle (1,1)_4^-, (u_1u_2, u_1u_2)_4^-,$ $(1, u_1u_2)_4^+ \rangle$	$H = \text{abel}(14);$
104	$\langle (1,1)_4^-, (u_1u_2, u_1u_2)_4^-,$ $(u_1, u_1)_4^+ \rangle$	$H = \langle x, y, z \mid y^{-1}zyz, y^{-1}xyx,$ $xz^{-1}xz, z^{-1}yzy, yxyx^{-1}, zx^{-1}zx \rangle,$ $H/H' = \text{abel}(2,2,2), H' = \text{abel}(2);$
105	$\langle (1,1)_4^-, (u_1u_2, u_1u_2)_4^-,$ $(u_1, u_2)_4^+ \rangle$	$H = \text{abel}(3, \infty);$
106	$\langle (1,u_1)_4^-, (u_1, 1)_4^-,$ $(1, u_1)_4^+ \rangle$	$H = \text{abel}(2,2,2);$
107	$\langle (1,u_1)_4^-, (u_1, 1)_4^-,$ $(u_1, 1)_4^+ \rangle$	$H = \text{abel}(2,2,2);$
108	$\langle (1,u_1)_4^-, (u_1, 1)_4^-,$ $(u_2, u_1u_2)_4^+ \rangle$	$H = \text{abel}(2,2,2);$
109	$\langle (1,u_1)_4^-, (u_1, 1)_4^-,$ $(u_2, u_1u_2)_4^- \rangle$	$H = \text{abel}(2,2,2);$
110	$\langle (1,u_1)_4^-, (u_1, 1)_4^-,$ $(u_1u_2, u_2)_4^+ \rangle$	$H = \text{abel}(2,2,2);$
111	$\langle (1,u_1)_4^-, (u_1, 1)_4^-,$ $(u_1u_2, u_2)_4^- \rangle$	$H = \text{abel}(2,2,2);$
112	$\langle (1,u_1)_4^-, (u_2, u_1u_2)_4^-,$ $(1, u_1)_4^+ \rangle$	$H = \text{abel}(2,2,2);$
113	$\langle (1,u_1)_4^-, (u_2, u_1u_2)_4^-,$ $(1, u_1u_2)_4^+ \rangle$	$H = \text{abel}(11);$
114	$\langle (1,u_1)_4^-, (u_2, u_1u_2)_4^-,$ $(u_1, 1)_4^+ \rangle$	$H = \text{abel}(2,2,2);$
115	$\langle (1,u_1)_4^-, (u_2, u_1u_2)_4^-,$ $(u_2, 1)_4^+ \rangle$	$H = \text{abel}(11);$
116	$\langle (1,u_1)_4^-, (u_2, u_1u_2)_4^-,$ $(u_1u_2, 1)_4^+ \rangle$	$H = \text{abel}(19);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
117	$\langle (1,u_1)_4^-, (u_1u_2,u_2)_4^-,$ $(1,u_1)_4^+\rangle$	$H = \text{abel}(2,2,2);$
118	$\langle (1,u_1)_4^-, (u_1u_2,u_2)_4^-,$ $(u_1,1)_4^+\rangle$	$H = \text{abel}(2,2,2);$
119	$\langle (1,u_1)_4^-, (u_1u_2,u_2)_4^-,$ $(u_2,u_1u_2)_4^+\rangle$	$H = \text{abel}(2,2,2);$
120	$\langle (1,u_1)_4^-, (u_1u_2,u_2)_4^-,$ $(u_1u_2,u_2)_4^+\rangle$	$H = \text{abel}(2,2,2);$
121	$\langle (1,u_1u_2)_4^-,$ $(u_1,1)_4^-, (1,1)_4^+\rangle$	$H = \text{abel}(9);$
122	$\langle (1,u_1u_2)_4^-, (u_1,1)_4^-,$ $(u_1,u_1)_4^+\rangle$	$H = \text{abel}(9);$
123	$\langle (1,u_1u_2)_4^-, (u_1,1)_4^-,$ $(u_2,u_2)_4^+\rangle$	$H = \langle x,y \mid y^{-2}xyxy^{-1}x, y^2x^2y^{-1}x\rangle,$ $H/H' = \text{abel}(9), H' = \text{abel}(7);$
124	$\langle (1,u_1u_2)_4^-, (u_1,1)_4^-,$ $(u_1u_2,u_1)_4^-\rangle$	$H = \langle x,y \mid x^3, yxyxy^{-1}x^{-1}, y^3\rangle,$ $H/H' = \text{abel}(3), H' = \text{abel}(7);$
125	$\langle (1,u_1u_2)_4^-, (u_1,1)_4^-,$ $(u_1u_2,u_2)_4^-\rangle$	$H = \text{abel}(9);$
126	$\langle (1,u_1u_2)_4^-, (u_1,1)_4^-,$ $(u_1u_2,u_1u_2)_4^+\rangle$	$H = \text{abel}(9);$
127	$\langle (1,u_1u_2)_4^-, (u_1,u_1)_4^-,$ $(1,u_1u_2)_4^+\rangle$	$H = \langle x,y \mid y^2, yx^{-1}yx^{-1}, xyxy, x^5\rangle,$ $H/H' = \text{abel}(2), H' = \text{abel}(5);$
128	$\langle (1,u_1u_2)_4^-, (u_1,u_1)_4^-,$ $(u_1,u_1)_4^+\rangle$	$H = \langle x,y \mid xyxy, x^2, y^3x^{-1}y^{-2}x, xy^{-1}xy^{-1}\rangle,$ $H/H' = \text{abel}(2), H' = \text{abel}(5);$
129	$\langle (1,u_1u_2)_4^-, (u_1,u_1)_4^-,$ $(u_1,u_2)_4^+\rangle$	$H = \text{abel}(\infty);$
130	$\langle (1,u_1u_2)_4^-, (u_1,u_1)_4^-,$ $(u_2,u_2)_4^-\rangle$	$H = \text{abel}(\infty);$
131	$\langle (1,u_1u_2)_4^-, (u_1,u_1)_4^-,$ $(u_1u_2,1)_4^-\rangle$	$H = \langle x,y \mid x^{-1}yxy, y^5\rangle,$ $H/H' = \text{abel}(\infty), H' = \text{abel}(5);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
132	$\langle(1,u_1u_2)_4^-,(u_1,u_2)_4^-,$ $(1,u_1u_2)_4^+\rangle$	$H=\mathrm{abel}(2,2,2);$
133	$\langle(1,u_1u_2)_4^-,(u_1,u_2)_4^-,$ $(u_1,u_2)_4^+\rangle$	$H=\mathrm{abel}(2,2,2);$
134	$\langle(1,u_1u_2)_4^-,(u_1,u_2)_4^-,$ $(u_2,u_1)_4^+\rangle$	$H=\mathrm{abel}(2,2,2);$
135	$\langle(1,u_1u_2)_4^-,(u_1,u_2)_4^-,$ $(u_2,u_1)_4^-\rangle$	$H=\langle x,y,z \mid z^2x^2,zx^{-1}zx,z^{-2}y^2,$ $zyz^{-1}y,z^{-1}yxz^{-1}yx^{-1},y^{-1}zyz,y^{-1}xyx^{-1}\rangle,$ $H/H'=\mathrm{abel}(2,2,2),H'=\mathrm{abel}(2)$
136	$\langle(1,u_1u_2)_4^-,(u_1,u_2)_4^-,$ $(u_1u_2,1)_4^+\rangle$	$H=\mathrm{abel}(2,2,2);$
137	$\langle(1,u_1u_2)_4^-,(u_1,u_2)_4^-,$ $(u_1u_2,1)_4^-\rangle$	$H=\langle x,y,z \mid y^2x^2,y^{-1}zyz,z^2x^2,$ $xy^{-1}xy,xyx^{-1}y,xzx^{-1}z,yz^{-1}yz,xzxz^{-1}\rangle,$ $H/H'=\mathrm{abel}(2,2,2),H'=\mathrm{abel}(2);$
138	$\langle(1,u_1u_2)_4^-,(u_1u_2,1)_4^-,$ $(1,u_1u_2)_4^+\rangle$	$H=\langle x,y,z \mid x^2,xyxy,zyz^{-1}y,z^2\rangle,$ $H/H'=\mathrm{abel}(2,2,2),H'=\mathrm{abel}(\infty,\infty);$
139	$\langle(1,u_1u_2)_4^-,(u_1u_2,1)_4^-,$ $(u_1,u_2)_4^+\rangle$	$H=\mathrm{abel}(2,2,\infty);$
140	$\langle(1,u_1u_2)_4^-,(u_1u_2,1)_4^-,$ $(u_2,u_1)_4^+\rangle$	$H=\langle x,y,z \mid z^2,xy^{-1}xy^{-1},y^2x^{-1}zx^{-1}z,$ $yzyzx^{-2},xz^{-1}xyzy,x^2y^2\rangle,$ $H/H'=\mathrm{abel}(2,2,4),H'=\mathrm{abel}(2);$
141	$\langle(u_1,1)_4^-,(u_2,u_1)_4^-,$ $(1,u_2)_4^+\rangle$	$H=\mathrm{abel}(11);$
142	$\langle(u_1,1)_4^-,(u_2,u_1)_4^-,$ $(u_1,u_1u_2)_4^+\rangle$	$H=\mathrm{abel}(11);$
143	$\langle(u_1,1)_4^-,(u_2,u_1)_4^-,$ $(u_1u_2,u_2)_4^-\rangle$	$H=\mathrm{abel}(11);$
144	$\langle(u_1,1)_4^-,(u_1u_2,u_1)_4^-,$ $(u_2,u_2)_4^+\rangle$	$H=\langle x,y \mid x^3,y^3,xyxy^{-1}x^{-1}y\rangle,$ $H/H'=\mathrm{abel}(3),H'=\mathrm{abel}(7);$
145	$\langle(u_1,1)_4^-,(u_1u_2,u_2)_4^-,$ $(1,1)_4^+\rangle$	$H=\mathrm{abel}(9);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
146	$\langle (u_1,1)_4^-, (u_1u_2,u_2)_4^-,$ $(1,u_1)_4^+\rangle$	$H = \mathrm{abel}(2,2,2);$
147	$\langle (u_1,1)_4^-, (u_1u_2,u_2)_4^-,$ $(1,u_2)_4^+\rangle$	$H = \mathrm{abel}(11);$
148	$\langle (u_1,1)_4^-, (u_1u_2,u_2)_4^-,$ $(1,u_1u_2)_4^+\rangle$	$H = \mathrm{abel}(17);$
149	$\langle (u_1,1)_4^-, (u_1u_2,u_2)_4^-,$ $(u_1,1)_4^+\rangle$	$H = \mathrm{abel}(2,2,2);$
150	$\langle (u_1,1)_4^-, (u_1u_2,u_2)_4^-,$ $(u_1,u_1)_4^+\rangle$	$H = \mathrm{abel}(9);$
151	$\langle (u_1,1)_4^-, (u_1u_2,u_2)_4^-,$ $(u_2,1)_4^+\rangle$	$H = \mathrm{abel}(3,3);$
152	$\langle (u_1,1)_4^-, (u_1u_2,u_2)_4^-,$ $(u_1u_2,1)_4^+\rangle$	$H = \mathrm{abel}(13);$
153	$\langle (u_1,u_1)_4^-, (u_2,u_2)_4^-,$ $(1,1)_4^+\rangle$	$H = \langle x,y,z \mid zxzx^{-1}, xyxy^{-1}, yzyz^{-1},$ $zyzy^{-1}, xzxz^{-1}, yxyx^{-1}\rangle,$ $H/H' = \mathrm{abel}(2,2,2), H' = \mathrm{abel}(2);$
154	$\langle (u_1,u_1)_4^-, (u_2,u_2)_4^-,$ $(1,u_1)_4^+\rangle$	$H = \mathrm{abel}(9);$
155	$\langle (u_1,u_1)_4^-, (u_2,u_2)_4^-,$ $(1,u_1u_2)_4^+\rangle$	$H = \mathrm{abel}(14);$
156	$\langle (u_1,u_1)_4^-, (u_2,u_2)_4^-,$ $(u_1,u_1)_4^+\rangle$	$H = \langle x,y,z \mid x^2, z^2, xy^{-1}xy^{-1},$ $y^{-1}z^{-1}yz, xy^{-1}z^{-1}x^{-1}yz\rangle,$ $H/H' = \mathrm{abel}(2,2,2), H' = \mathrm{abel}(2);$
157	$\langle (u_1,u_1)_4^-, (u_2,u_2)_4^-,$ $(u_1,u_2)_4^+\rangle$	$H = \langle x,y \mid xy^{-2}xyx^{-2}y, y^{-3}x^3\rangle,$ $H/H' = \mathrm{abel}(3,\infty), H' = \mathrm{abel}(\infty,\infty);$
158	$\langle (u_1,u_1)_4^-, (u_2,u_2)_4^-,$ $(u_1u_2,1)_4^-\rangle$	$H = \langle x,y \mid x^2y^{-1}x^{-2}y, y^{-2}xyx^{-1}yx\rangle,$ $H/H' = \mathrm{abel}(\infty),$ $H'/H'' = \mathrm{abel}(2,6), H'' = \mathrm{abel}(2);$
159	$\langle (u_1,u_1)_4^-, (u_1u_2,1)_4^-,$ $(1,u_1u_2)_4^+\rangle$	$H = \langle x,y \mid x^2, y^2, xyxyxyxyxy\rangle,$ $H/H' = \mathrm{abel}(2); H' = \mathrm{abel}(5);$

List 3
(continuation)

N	Γ	$H(\Gamma)$
160	$\langle (u_1,u_1)_4^-, (u_1u_2,1)_4^-,$ $(u_1,u_1)_4^+\rangle$	$H = \langle x,y \mid x^2, y^2, xyxyxyxyxy\rangle,$ $H/H' = \text{abel}(2); H' = \text{abel}(5);$
161	$\langle (u_1,u_1)_4^-, (u_1u_2,1)_4^-,$ $(u_1,u_2)_4^+\rangle$	$H = \langle x,y \mid y^2, x^2yx^{-1}yx^{-1}y\rangle,$ $H/H' = \text{abel}(\infty), H' = \text{abel}(2,2);$
162	$\langle (u_1,u_2)_4^-, (u_2,u_1)_4^-,$ $(1,u_1u_2)_4^+\rangle$	$H = \text{abel}(2,2,4);$
163	$\langle (u_1,u_2)_4^-, (u_2,u_1)_4^-,$ $(u_1,u_2)_4^+\rangle$	$H = \langle x,y,z \mid x^2, y^2, zyzy, zxzx\rangle,$ $H/H' = \text{abel}(2,2,2), H' = \text{abel}(\infty,\infty);$
164	$\langle (u_1,u_2)_4^-, (u_2,u_1)_4^-,$ $(u_2,u_1)_4^+\rangle$	$H = \langle x,y,z \mid x^2, y^2, zxzx, zyzy\rangle,$ $H/H' = \text{abel}(2,2,2), H' = \text{abel}(\infty,\infty);$
165	$\langle (u_1,u_2)_4^-, (u_2,u_1)_4^-,$ $(u_1u_2,1)_4^-\rangle$	$H = \langle x,y,z \mid z^2x^2, y^2z^{-2}, xz^{-1}xz, xyx^{-1}y,$ $xy^{-1}xy, yzy^{-1}z, yz^{-1}yz, x^{-1}zxz\rangle,$ $H/H' = \text{abel}(2,2,2), H' = \text{abel}(2);$
166	$\langle (u_1,u_2)_4^-, (u_1u_2,1)_4^-,$ $(1,u_1u_2)_4^+\rangle$	$H = \text{abel}(2,2,2);$
167	$\langle (u_1,u_2)_4^-, (u_1u_2,1)_4^-,$ $(u_1,u_2)_4^+\rangle$	$H = \text{abel}(2,2,2);$
168	$\langle (u_1,u_2)_4^-, (u_1u_2,1)_4^-,$ $(u_2,u_1)_4^+\rangle$	$H = \text{abel}(2,2,2);$
169	$\langle (u_1,u_2)_4^-, (u_1u_2,1)_4^-,$ $(u_1u_2,1)_4^+\rangle$	$H = \text{abel}(2,2,2);$

References

[1] Berkovich J.G, Freiman G. A and Praeger C. E., *Small squaring and cubing properties of finite groups*, Bull. Austral. Math. Soc., **44**, 1991, 429–450.

[2] Brailovsky L. V., *On (3,m)-special elements in groups*, Manuscript.

[3] Brailovsky L. V. and Freiman G. A., *On two-element subsets in groups*, Ann. N.Y. Acad. Sci., **373**, 1981, 183–190.

[4] Brailovsky L. V. and Freiman G. A., *Groups with small cardinality of the cubes of their two-element subsets*, Ann. N.Y. Acad. Sci., **410**, 1983, 75–82.

[5] Freiman G. A. and Schein B. M., *Interconnections between the structure theory of set addition and rewritability in groups*, Proc. Amer. Math. Soc., **113**, 1991, n. 4, 899–910.

[6] Higman G., Neumann B. H., and Neumann H., *Embedding theorems for groups*, J.London Math. Soc., **29**, 1949, 247-254.

[7] Longobardi P. and Maj M., *The classification of groups with small squaring property on 3-sets*, Bull. Austral. Math. Soc., **46** (1992), 263–269.

[8] Lyndon C. and Schupp E., *Combinatorial Group Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1977.

[9] Magnus V., Karrass A. and Solitar D., *Combinatorial Group Theory: Presentations of groups in terms of generators and relations*, John Wiley and Sons, Inc., New York, London, Sydney, 1966.

[10] Malčev A. I., *On homomorphisms onto the finite groups*, Uchenye Zapiski Ivanovskogo pedinstituta, (Russian),**18**, (1958), n.5, 49–60.

[11] Marston C., *Three-relator quotients of the Modular Group*, Quart. J. Math. Oxford (2), **38**, 1987, 127–447.

[12] Neumann H., *Varieties of Groups*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 37, Springer-Verlag, Berlin-Heidelberg-N.Y., 1967.

S. Brodsky, Eventus Logistics Ltd., 8 Dov Friedman st., P.O.B. 1930, Ramat-Gan 52118, Israel
E-mail : `sergei@eventus.co.il`

Astérisque

YAHYA OULD HAMIDOUNE

On small subset product in a group

Astérisque, tome 258 (1999), p. 281-308

<http://www.numdam.org/item?id=AST_1999__258__281_0>

Astérisque **258**, 1999, p. 281–308

ON SMALL SUBSET PRODUCT IN A GROUP

by

Yahya Ould Hamidoune

Abstract. — We generalise some known addition theorems to non abelian groups and to the most general case of relations having a transitive group of automorphisms.

The classical proofs of addition theorems use local transformations due to Davenport, Dyson and Kempermann. We present a completely different method based on the study of some blocks of imprimitivity with respect to the automorphism group of a relation.

Several addition theorems including the finite $\alpha + \beta$-Theorem of Mann and a formula proved by Davenport and Lewis will be generalised to relations having a transitive group of automorphisms.

We study the critical pair theory in the case of finite groups. We generalise Vosper Theorem to finite not necessarily abelian groups.

Chowla, Mann and Straus obtained in 1959 a lower bound for the size of the image of a diagonal form on a prime field. This result was generalised by Tietäväienen to finite fields with odd characteristics. We use our results on the critical pair theory to generalise this lower bound to an arbitrary division ring.

Our results apply to the superconnectivity problems in networks. In particular we show that a loopless Cayley graph with optimal connectivity has only trivial minimum cuts when the degree and the order are coprime.

1. Introduction

Let p be a prime number, and let A and B be two subsets of $\mathbf{Z}_p$, such that $|A|$, $|B| \geq 2$. The Cauchy-Davenport Theorem states that

$$|A + B| \geq \min(p, |A| + |B| - 1),$$

cf. [2,5]. Vosper Theorem states that

$$|A + B| \geq \min(p - 1, |A| + |B|),$$

unless A and B form arithmetic progressions, cf. [31,32]. Freiman obtained a structure theorem for all $A \subset \mathbf{Z}_p$ such that $|2A| < 12|A|/5 - 3$, cf. [26].

1991 ***Mathematics Subject Classification.*** — Primary: 20D60, Secondary: 20K01, 11B13, 11B75, 05C25.

Key words and phrases. — Addition theorems, blocks of imprimitivity, network reliability.

Let A and B be finite subsets of an abelian group G. We shall say that B a Cauchy subset if for every finite non-empty subset X,

$$|X + B| \geq \min(|G|, |X| + |B| - 1).$$

Mann proved in [24] that B is a Cauchy subset if and only if for every finite subgroup H, $|H+B| \geq \min(|G|, |H|+|B|-1)$. Kneser Theorem states that $|A+B| < |A|+|B|-1$ only if there is a finite non-null subgroup H such that $A + H + B = A + B$. Some progress toward the determination of all pairs A, B such that $|A + B| \leq |A| + |B| - 1$ is obtained by Kempermann in [20]. In [14], we could classify all the pairs, $\{A, B\}$ with $|A + B| = |A| + |B| - 1$, if B is a Cauchy subset.

Less results are know in the non-abelian case. The classical basic tools in this case are two nice results proved by Kempermann in [19]. No generalisation of Kneser Theorem is known in the non-abelian case. The natural one is false in general, cf. [28,33]. Diderrich obtained in [7] a generalisation of Kneser Theorem in the case where the elements of B commute. But this result is an easy corollary of Kneser Theorem as showed in [13]. Brailowski and Freiman obtained a Vosper Theorem in free torsion groups, cf. [1]. It was observed recently that some results involving the connectivity of Cayley graphs are strongly related to addition theorems. This connection will be explained below.

A natural question consists of asking how addition Theorems generalise to a group acting on a set. The connectivity of Cayley graphs belongs to this kind of problems. The connectivity of a reflexive relation $\Gamma = (V, E)$ is

$$\kappa(\Gamma) = \min\{|\Gamma(F)| - |F| : 1 \leq |\Gamma(F)| < |V|\}.$$

Let B be a finite subset of a group G containing 1 and let Γ be the Cayley relation $x^{-1}y \in B$. In this case, $\kappa(\Gamma)$ is the best possible lower bound for $|AB^{-1}| - |A|$, where $AB \neq G$. The Cauchy-Davenport Theorem may be expressed using this language as $\kappa(\Gamma) = |B| - 1$, for $|G|$ prime. Under this formulation, this result was rediscovered in [9]. The method used in [9] is based on the study of blocks of imprimitivity with respect to the group of automorphisms. The same method is used in [12] to prove a local generalisation of Mann Theorem for finite groups. Zemor used the same method in [33] to obtain a global one. More complicated blocks are studied in [14] to calculate the critical pairs in Mann Theorem in the abelian case.

The connection between connectivity problems and addition theorems were observed only recently.

The results obtained in [14] are strongly based on the well known fact that an abelian Cayley relation is isomorphic to its reverse. We generalise some of the results to the non abelian case. The organisation of the paper is as follows. In section 2, we study the connectivity of relations. We give also lemmas allowing to translate connectivity bounds into addition theorems. We improve some results contained in [9,10,11,12,14]. In section 3, we generalise several basic additive inequalities. In particular, we give a generalisation of Mann Theorem to non-abelian groups and to relations with a transitive group of automorphisms. We generalise also a formula proved by Davenport and Lewis for finite fields to division rings and to arc-transitive relations. We generalise also a result proved by Olson [27] to point transitive relations. This generalisation in the finite case was proved in [10, Proposition 3.4]. In section 4, we study the superatoms. They form the main tool for the critical pair problem in our approach. The main result of section 5 is the following result which characterizes the equality cases in Mann Theorem. We state it below.

Let B be a subset of a finite group G such that $1 \in B$. Then the following conditions are equivalent.

(i) For all $A \subset G$ such that $2 \leq |A|$,

$$|AB| \geq \min(|G|-1, |A|+|B|).$$

(ii) For every subgroup H of G and for every $a \in G$ such that $|H \cup Ha| \geq 2$,

$$\min(|B(H \cup aH)|, |(H \cup Ha)B|) \geq \min(|G|-1, |H \cup Ha|+|B|).$$

The main result of section 6 is a critical pair theorem which generalises Vosper Theorem. We state it below.

Let G be a finite group and let B be a Cauchy subset of G such that $(|G|, |B|-1) = 1$. Let $A \subset G$ such that

$$|AB| = |A|+|B|-1 \leq |G|-1.$$

Then one of the following conditions holds.

(i) $|A| = 1$ or $A = G \setminus aB^{-1}$, for some $a \in G$.

(ii) There are $a, b, r \in G$, $k, s \in \mathbf{N}$ such that

$$A = \{a, ar, ar^2, \ldots, ar^{k-1}\} \quad \textit{and} \quad B = (G \setminus \langle r \rangle b) \cup \{b, rb, r^2b, \ldots, r^{s-1}b\}.$$

(iii) There are $a, b, r \in G$, $k, s \in \mathbf{N}$ such that

$$A = \{ab^{-1}, arb^{-1}, ar^2b^{-1}, \ldots, ar^{k-1}b^{-1}\} \quad \textit{and} \quad B = (G \backslash b\langle r \rangle) \cup \{b, rb, r^2b, \ldots, r^{s-1}b\}.$$

One of the classical applications of the critical pair theory is the estimation of the range of a diagonal form. Using Vosper's Theorem, Chowla, Mann and Straus obtained in [4] an estimation of the range of a diagonal form over $\mathbf{Z}_p$. Tietäväinen obtained in [30] the same bound in the case of finite fields with odd characteristics. We gave in [14] a proof for all finite fields based on the method of superatoms. We generalise this bound to all division rings in this paper as follows.

Let R be a division ring and let P be a finite subset of R such that $0 \in P$ and $P \setminus 0$ is multiplicative subgroup. Let R_0 be the additive subgroup generated by P. Suppose that $|P| \geq 4$ and let $a_1, a_2, \ldots, a_n$ be non-zero elements of R. Then

$$|a_1P + a_2P + \cdots + a_nP| \geq \min(|R_0|, (2n-1)(|P|-1)+1).$$

In section 8, we apply our results to solve some problems raised in network Theory. We also explain the connections between Cayley graphs reliability and Additive group Theory. In particular we show that a loopless Cayley graph with optimal connectivity has only trivial minimum cuts when the degree and the order are coprime.

2. The connectivity of a relation

In this section we study subsets with a small image with respect to a given relation. Restricted to Cayley relations defined on a group, this problem becomes the study of subsets with a small product. The results obtained in this section improve slightly our previous results obtained in [9,10,11,12,14].

The cardinality of a finite set V will be denoted by $|V|$. For an infinite set V, we write $|V| = \infty$. By a *relation* we mean an ordered pair $\Gamma = (V, E)$, where V is a set and E is a subset of $V \times V$. A permutation σ of V is said to be an automorphism of Γ if $E = \{(\sigma(x), \sigma(y)) : (x, y) \in E\}$. The group of automorphisms of Γ will be denoted by $\mathrm{Aut}(\Gamma)$. A relation will be called *point transitive* if its group of automorphisms acts transitively on V. Let $A \subset V$. The *subrelation induced* on A is $\Gamma[A] = (A, E \cap (A \times A))$.

We introduce some notations. Let $\Gamma = (V, E)$ be a relation and let F be a subset of V. The *image* of F will be denoted by $\Gamma(F)$. We recall that

$$\Gamma(F) = \{y \in V : \text{ there is } x \in F \text{ such that } (x, y) \in E\}.$$

We write $\partial_\Gamma(F) = \Gamma(F) \setminus F$ and $\delta_\Gamma(F) = V \setminus (F \cup \Gamma(F))$. The reference to Γ will be omitted when the meaning is clear from the context. In particular we shall write $\partial_{\Gamma^-}(F) = \partial^-(F)$ and $\delta_{\Gamma^-}(F) = \delta^-(F)$. The *degree* of a point $x \in V$ is by definition $d_\Gamma(x) = |\Gamma(x)|$. A relation Γ is said to be *locally finite* if both Γ and Γ^- have only finite degrees. A relation $\Gamma = (V, E)$ is said to be *regular* if Γ is locally finite and if all $x, y \in V$, $|\Gamma(x)| = |\Gamma(y)|$ and $|\Gamma^-(x)| = |\Gamma^-(y)|$. Let Γ be a regular relation. The degree of every point with respect to Γ will be called the degree of Γ and denoted by $d(\Gamma)$.

A relation Γ on a set V is said to be *connected* if $\Gamma(A) \not\subset A$ for every finite proper subset A of V. A subset C of V is said to be *connected* if $\Gamma[C]$ is connected. A *block* of Γ is a subset B of V such that for every automorphism f of Γ, either $f(B) = B$ or $f(B) \cap B = \varnothing$.

The following remark is easy to show and well known.

Remark 2.a. — If Γ is regular and if V is finite then $d(\Gamma) = d(\Gamma^-)$.

Let Γ be a reflexive relation on V. The *connectivity* of Γ is by definition:

$$\kappa(\Gamma) = \min\{|\partial(F)| : 1 \leq |\Gamma(F)| < |V| \text{ or } |F| = 1\}.$$

The inequality $1 \leq |\Gamma(F)| < |V|$ is never satisfied if $V \times V = \Gamma$. In the other cases,

$$\kappa(\Gamma) = \min\{|\partial(F)| : 1 \leq |\Gamma(F)| < |V|\}.$$

Remark 2.b. — The connectivity of a relation coincides with the connectivity of its reflexive closure. For this reason we restrict ourselves to reflexive relations. This choice simplifies the proofs and the notations. In some previous papers [9,10,11,12] we adopted the opposite choice, where a relation is assumed to be disjoint from its diagonal. These two choices are essentially equivalent.

Lemma 2.1. — *Let Γ be a locally finite reflexive relation. Then $\kappa(\Gamma)$ is the maximal k such that for every non-empty finite subset A, $|\Gamma(A)| \geq \min(|V|, |A| + k)$.*

Proof. — The Lemma follows easily from the definitions. □

One see easily that a locally finite reflexive relation Γ on a set V is connected if and only if $\kappa(\Gamma) \geq 1$.

Let Γ be a locally finite reflexive relation on V and let F be a subset of V. We say that F is a *fragment* of Γ if the following conditions are satisfied.

(i) $\kappa(\Gamma) = |\partial(F)|$ and $\Gamma(F) \neq V$.

(ii) F is finite or $V \setminus F$ is finite.

A fragment with minimal cardinality will be called an *atom.* The cardinality of an atom of Γ will be denoted by $\mu(\Gamma)$. We see easily that an atom is always finite. We write $\rho(\Gamma) = \min(d(x); x \in V)$. Observe that $\rho(\Gamma)$ is the minimal degree.

***Lemma 2.2* ([9]).** — *Let Γ be a locally finite reflexive relation on a set V and let A be an atom of Γ. Then $\Gamma[A]$ is connected. Moreover $\kappa(\Gamma) \leq \rho(\Gamma) - 1$ and equality holds if and only if $\mu(\Gamma) = 1$ or $\Gamma = V \times V$.*

Proof. — Lemma 2.2 follows easily from the definitions. □

***Lemma 2.3* ([9]).** — *Let Γ be a reflexive relation on a finite set V. Then $\kappa(\Gamma) = \kappa(\Gamma^-)$.*

Proof. — The equality is obvious if $\Gamma = V \times V$. Suppose the contrary and let F be a fragment of Γ. We have clearly $\partial^-(\delta(F)) = \partial^-(V \setminus \Gamma(F)) \subset \partial(F)$. By the definition we have

$$\kappa(\Gamma^-) \leq |\partial^-(\delta(F))| \leq |\partial(F)| = \kappa(\Gamma).$$

The other inequality follows by duality. □

We use the following result.

***Lemma 2.4* ([9]).** — *Let Γ be a locally finite reflexive relation on a set V such that $\kappa(\Gamma) = \kappa(\Gamma^-)$. Let F and M be two fragments of Γ. Then*

(i) $\partial^-(\delta(F)) = \partial(F)$ and $\delta^-(\delta(F)) = F$.

(ii) $\delta(F)$ is a fragment of Γ^-.

(iii) $F \subset M$ if and only if $\delta(F) \supset \delta(M)$.

Proof. — We have clearly

$$\partial^-(\delta(F)) = \partial^-(V \setminus \Gamma(F)) \subset \partial(F).$$

If $\delta(F)$ is finite, then

$$|\partial^-(\delta(F))| \geq \kappa(\Gamma^-) = \kappa(\Gamma) = |\partial(F)|.$$

Therefore $\partial^-(\delta(F)) = \partial(F)$. Assume now that $\delta(F)$ is infinite. Hence F is finite. We shall prove that $\partial(F) \subset \partial^-(\delta(F))$. Suppose on the contrary that there exists $x \in \partial(F) \setminus \partial^-(\delta(F))$. It follows that $\Gamma(F \cup \{x\}) \subset \Gamma(F)$. Therefore

$$|\partial(F \cup \{x\})| \leq |\partial(F)| - 1 = \kappa(\Gamma) - 1.$$

It follows by the definition of κ that $|\Gamma(F \cup \{x\})| = \infty$, a contradiction. The first equality in (i) is proved. It follows that

$$\delta^-(\delta(F)) = V \setminus (\partial^-(\delta(F)) \cup \delta(F)) = V \setminus (\partial(F) \cup \delta(F)) = F.$$

Hence (i) holds. It follows that

$$|\partial^-(\delta(F))| = |\partial(F)| = \kappa(\Gamma) = \kappa(\Gamma^-).$$

Since $F \cap \Gamma^-(\delta(F)) = \varnothing$, $\delta(F)$ is a fragment of Γ^-. Thus (ii) is proved.

Suppose that $F \subset M$. We have $\delta(F) = V \setminus \Gamma(F) \supset V \setminus \Gamma(M) = \delta(M)$.

Suppose that $\delta(F) \supset \delta(M)$. We see as above that $\delta^-(\delta(F)) \subset \delta^-(\delta(M))$. Using (i) we obtain, $F \subset M$. □

We shall use the following lemma.

***Lemma* 2.5.** — *Let Γ be reflexive relation on a set V. Let M be a finite fragment and let F be a fragment of Γ such that $M \cap F \neq \varnothing$. Then $|\delta(F) \setminus \delta(M)| \leq |M \setminus F|$.*

Suppose that $\kappa(\Gamma) = \kappa(\Gamma^-)$ or that F is finite. Then one of the following conditions holds.

(i) $\delta(F) \cap \delta(M) = \varnothing$.

(ii) $F \cap M$ is a fragment of Γ and $\Gamma(M \cap F) = \Gamma(M) \cap \Gamma(F)$.

Proof. — We have clearly

$$\Gamma(M \cap F) \subset \Gamma(M) \cap \Gamma(F) = M \cup \partial(M) \cap (F \cup \partial(F)).$$

Therefore

$$\partial(F \cap M) \subset (\partial(M) \setminus \delta(F)) \cup (M \cap \partial(F)). \tag{1}$$

Clearly $\Gamma(M \cap F) \neq V$. By the definition of the connectivity and since $|\partial(M)| = \kappa(\Gamma)$, we have $|\partial(M)| \leq |\partial(M \cap F)|$. Using (1), we have

$$|\delta(F) \cap \partial(M)| \leq |\partial(F) \cap M|. \tag{2}$$

It follows that

$$|\delta(F) \setminus \delta(M)| = |\delta(F) \cap M| + |\delta(F) \cap \partial(M)| \leq |\delta(F) \cap M| + |M \cap \partial(F)| = |M \setminus F|.$$

This proves the first part of the lemma.

We have clearly $\partial(M \cup F) \subset \partial(M) \cup \partial(F)$. Therefore

$$\partial(F \cup M) \subset (\partial(F) \setminus M) \cup (\delta(F) \cap \partial(M)). \tag{3}$$

Assume now $\delta(F) \cap \delta(M) \neq \varnothing$. It follows that $\Gamma(M \cup F) \neq V$. We shall show the following inequality.

$$|\partial(F) \cap M| \leq |\delta(F) \cap \partial(M)|. \tag{4}$$

Consider first the case where F is finite. By the definition of the connectivity and since $\Gamma(M \cup F) \neq V$ and $|\partial(F)| = \kappa(\Gamma)$, we have $|\partial(F)| \leq |\partial(M \cup F)|$. Using (3), we obtain (4). Assume now that $\kappa(\Gamma) = \kappa(\Gamma^-)$ and that F is infinite. By the definitions $\delta(F)$ is finite. By Lemma 2.4, $\delta(F)$ and $\delta(M)$ are fragments of Γ. By applying (2) to Γ^-, with M replaced by $\delta(F)$ and F replaced by $\delta(M)$, we obtain

$$|\delta^-\delta(M) \cap \partial^-(\delta(F))| \leq |\partial^-(\delta(M)) \cap \delta(F)|.$$

(4) follows now using Lemma 2.4.

By (2) and (4) we have

$$|M \cap \partial(F)| = |\partial(M) \cap \delta(F)|. \tag{5}$$

Since $\Gamma(F \cap M) \neq \varnothing$ and $\Gamma(F \cap M) \neq V$, we have $|\partial(F \cap M)| \geq \kappa(\Gamma)$. By (1) and (5) we have

$$\kappa(\Gamma) \leq |\partial(F \cap M)| \leq |(\partial(F) \setminus \delta(M)) \cup (K \cap \partial(M))| \leq |\partial(M)| = \kappa(\Gamma).$$

It follows that $F \cap M$ is a fragment of Γ. It follows also that

$$\partial(F \cap M) = (\partial(F) \setminus \delta(M)) \cup (F \cap \partial(M)).$$

Therefore

$$\Gamma(F \cap M) = (F \cap M) \cup (\partial(F) \setminus \delta(M)) \cup (F \cap \partial(M)) = \Gamma(F) \cap \Gamma(M).$$

□

Remark 2.c. — If A and B be two finite fragments such that $|A| = |B|$ then $|\delta(A)| = |\delta(B)|$.

Clearly we have $|\delta(B)| = |V| - (|B| + \kappa(\Gamma)) = |\delta(A)|$. □

The fundamental property of atoms is the following.

Proposition 2.6. — *Let A and B be two distinct atoms of a locally finite reflexive relation Γ and let F be a fragment of Γ. Suppose that $\kappa(\Gamma) = \kappa(\Gamma^{-})$ or that F is finite.*

(i) Assume that $|A| \leq |\delta(F)|$. Then either $A \subset F$ or $A \cap F = \varnothing$.
(ii) Assume that $|A| \leq |\delta(A)|$. Then $A \cap B = \varnothing$.
(iii) Assume that $|A| \geq |\delta(A)| + 1$. Then $\delta(A) \cap \delta(B) = \varnothing$.

Proof. — Assume that $|A| \leq |\delta(F)|$ and suppose that $A \cap F \neq \varnothing$. By Lemma 2.5, we have

$$|\delta(F) \setminus \delta(A)| \leq |A \setminus F| < |A|.$$

Hence $\delta(F) \cap \delta(A) \neq \varnothing$. By Lemma 2.5, $A \cap F$ is a fragment of Γ. By the minimality of $|A|$, we have $A \cap F = A$. Therefore $A \subset F$. This proves (i).

Assume that $|A| \leq |\delta(A)|$ and that $A \cap B \neq \varnothing$. By Remark 2.c and (i), we have $A \cap B = \varnothing$, a contradiction. Hence (ii) is proved.

Assume that $|A| \geq |\delta(A)| + 1$ and that $\delta(A) \cap \delta(B) \neq \varnothing$. Clearly $|V|$ is finite.

By Lemma 2.5, we have

$$|A \setminus B| = |B \setminus A| \leq |\delta(A) \setminus \delta(B)|.$$

Therefore $A \cap B \neq \varnothing$. By Lemma 2.5, $A \cap B$ is a fragment. Hence $A = B$, a contradiction. □

Corollary 2.7. — *Let Γ be a locally finite reflexive relation such that either V is infinite or $\mu(\Gamma) \leq \mu(\Gamma^{-})$. Let A be an atom of Γ and let F be a finite fragment of Γ. Then either $A \subset F$ or $A \cap F = \varnothing$.*

Proof. — The inequality $\mu(\Gamma^-) \leq |\delta(F)|$ holds clearly if V is infinite and follows in the finite case by Lemmas 2.3 and 2.4. Therefore $|A| = \mu(\Gamma) \leq \mu(\Gamma^-) \leq |\delta(F)|$. The result follows now using Proposition 2.6. □

The above result was proved for finite symmetric relations by Mader [23, sätz 1] and generalised to arbitrary finite relations in [9, proposition 1]. A basic property of atoms is the following.

Corollary 2.8. — *Let $\Gamma = (V, E)$ be a locally finite reflexive relation and let A be an atom of Γ.*

(i) Assume that $|A| \leq |\delta(A)|$. Then A is a block.

(ii) Assume that $|A| \geq |\delta(A)| + 1$. Then $\delta(A)$ is a block.

Proof. — Let f be an automorphism of Γ. Clearly $f(A)$ is an atom. We have also

$$f(\delta(A)) = f(V \setminus \Gamma(A)) = V \setminus \Gamma(f(A)) = \delta(f(A)).$$

The results follows now easily using Proposition 2.6. □

Let A and B be subsets of a group G. We write

$$AB = \{xy \ : \ x \in A \text{ and } y \in B\}.$$

Let $a \in G$, the *left translation* $\gamma_a : G \longrightarrow G$ is defined by the equality $\gamma_a(x) = ax$. As usual the image of a subgroup H by a left translation will be called a *left coset* of H.

Let S be a subset of G. The relation $x^{-1}y \in S$ is called a *Cayley relation.* It will be denoted by $\Lambda(G, S)$. Let $\Gamma = \Lambda(G, S)$ and let $F \subset G$. Clearly $\Gamma(F) = FS$.

The following result is easy to show and well known.

Lemma 2.9. — *Let G be a group and let S be a finite subset of G. Then $(\Lambda(G, S))^- = \Lambda(G, S^{-1})$.*

For every $a \in G$, $\gamma_a \in \mathrm{Aut}(\Lambda(G, S))$. In particular $\Lambda(G, S)$ is point transitive.

Cayley relations defined above form an important class of the relations with a transitive group of automorphisms.

We use the following result which is implicit in [12].

Lemma 2.10. — *Let G be a group containing a subset S and let B be a finite non-empty block of $\Lambda(G, S)$. Then B is a left coset of some subgroup of G.*

Proof. — Choose $b \in B^{-1}$ and set $H = bB$. Let $x \in H$. By Lemma 2.9, H is a block. Clearly $1 \in H$. Therefore $x \in H \cap \gamma_x(H)$, and hence $H = xH$. Therefore $HH = H$. Since H is finite, H is a subgroup. □

Theorem 2.11. — *Let G be a group and let S be a finite subset of G such that $1 \in S$. Let A be an atom of $\Lambda(G, S)$ containing an element a and let $b \in \delta(A)$.*

(i) If $|A| \leq |\delta(A)|$ then $a^{-1}A$ is a subgroup.

(ii) If $|A| \geq |\delta(A)| + 1$ then $b^{-1}\delta(A)$ is a subgroup.

Proof. — The result follows from Corollary 2.8 and Lemma 2.10. □

Corollary 2.12 **([12, proposition 1]).** — *Let G be a group and let S be a finite subset of G such that $1 \in S$ and let A be an atom of $\Lambda(G,S)$ containing 1. Suppose that $\mu(\Lambda(G,S)) \le \mu(\Lambda(G,S^{-1}))$. Then A is a subgroup. Moreover for every finite fragment F of $\Lambda(G,S)$, $FA = F$.*

Proof. — As shown in the proof of Corollary 2.7, we have $|A| = \mu(\Gamma) \le \mu(\Gamma^{-}) \le |\delta(F)|$. By Theorem 2.11, A is a subgroup. Since $1 \in A$, we have $F \subset FA$. Let $x \in F$. By Lemma 2.10, xA is an atom. By Corollary 2.7, $xA \subset F$. Hence $FA \subset F$. □

We shall now describe a method allowing to apply connectivity bounds for connected relations in the non connected case. This happens in Cayley relations when B generates a proper finite subgroup. In this case, one could decompose A as union of cosets modulo $\langle B \rangle$. We shall generalise this decomposition in the case of relations with a transitive group of automorphisms. Let us begin with an easy lemma

Lemma 2.13. — *Let $\Gamma = (V,E)$ be a point transitive relation and let C be a block. Let f be an automorphism of Γ. Then $\Gamma[C]$ and $\Gamma[f(C)]$ are isomorphic point transitive relations.*

Proof. — Clearly $f/C : C \longrightarrow f(C)$ defines an isomorphism from $\Gamma[C]$ onto $\Gamma[f(C)]$.

Let x, $y \in C$. Since Γ is point transitive, there is $g \in \mathrm{Aut}(\Gamma)$ such that $g(x) = g(y)$. By the definition of a block $g(C) = C$.

Now $g/C : C \longrightarrow C$ defines an automorphism of $\Gamma[C]$. □

Let $\Gamma = (V,E)$ be a reflexive relation. A subset C of V will be called a *component* of Γ if $\Gamma[C]$ is connected and if C is maximal with respect to this property. It follows easily from Zorn Lemma that every connected subset is contained in a component. It is easy also to check that two distinct components are disjoint. In particular the connected components form a partition. In follows also that a component is a block. The following remark follows easily from the definitions.

Remark 2.d. — Let $\Gamma = (V,E)$ be a reflexive relation and let $\{C_i; i \in I\}$ be a family of components of Γ and let $A \subset \bigcup_{i\in I} C_i$, be such that $\Gamma(A) \cap (\bigcup_{i\in I} C_i) = A$. There is $J \subset I$ such that $A = \bigcup_{j\in J} C_j$.

Remark 2.e. — Let $\Gamma = (V,E)$ be a reflexive relation. Then Γ has at most one infinite component.

By Remark 2.d, the union of two infinite components is connected. Hence any two infinite components must coincide.

We mention that the path connectedness, considered in section 8, leads to distinct notion of components in the infinite case. The following lemma contains all we need on components.

Lemma 2.14. — *Let $\Gamma = (V,E)$ be a reflexive point transitive relation and let C and D be components of Γ. Then the following conditions hold.*

(i) $\Gamma[C]$ and $\Gamma[D]$ are isomorphic point transitive relations.

(ii) $C = V$ or C is finite.

(iii) $\Gamma(A \cap C) = (\Gamma(A)) \cap C$ *and* $d(\Gamma[C]) = d(\Gamma)$.

Proof. — The validity of (i) follows easily from Lemma 2.14. The validity of (ii) follows from (i) and Remark 2.e.

Assume that (iii) does not hold. There are distinct connected components C_1 and C_2 such that $\Gamma(C_1) \cap C_2 \neq \varnothing$. Using the transitivity of $\mathrm{Aut}(\Gamma)$, we may construct a sequence of connected components $\{C_i; i \geq 1\}$ such that, $C_i \neq C_{i+1}$ and $\Gamma(C_i) \cap C_{i+1} \neq \varnothing$, for all $i \geq 1$.

For all i, $j \geq 1$, we have

$$C_i \neq C_{j+i}. \tag{1}$$

Assume the contrary and choose j minimal with respect to this property. By the definition $\bigcup_{0 \leq k \leq j} C_{i+k}$ is non-connected.

Hence there exists $A \subset \bigcup_{0 \leq k \leq j} C_{i+k}$ such that $\Gamma(A) \cap (\bigcup_{0 \leq k \leq j} C_{i+k}) = A$ and $A \neq \varnothing$. By Remark 2.d, there is $J \subset [i, i+j]$ such that $A = \bigcup_{i \in J} C_i$. By the construction of C_i, one should have $J = [i, j]$. In particular $\bigcup_{i \geq 1} C_i$ is connected, a contradiction.

Let A be a finite non-empty subset of $\bigcup_{i \geq 1} C_i$ By (1), there exists clearly j such that $A \cap C_j \neq \varnothing$ and $A \cap C_{j+1} = \varnothing$. In particular $\Gamma(A) \not\subset \bigcup_{i \geq 1} C_i$. It follows that $\bigcup_{i \geq 1} C_i$ is connected, contradicting the maximality of C_1. □

Lemma 2.15. — *Let* $\Gamma = (V, E)$ *be a locally finite reflexive point symmetric relation and let* C *be a component of* Γ.

Then for every non-empty finite subset A, *either* $\Gamma(\Gamma(A)) = \Gamma(A)$ *or* $|\Gamma(A)| \geq |A| + \kappa(\Gamma[C])$.

Proof. — Assume first C infinite. By Lemma 2.14, $V = C$ and the result holds trivially by Lemma 2.2.

Therefore we may assume C finite. By Lemma 2.14, all the connected components generate isomorphic relations. In particular $\kappa(\Gamma[C]) = \kappa(\Gamma[D])$.

Suppose $\Gamma(\Gamma(A)) \neq \Gamma(A)$. By Lemma 2.14.iii, there is a connected component D such that $\Gamma(\Gamma(A \cap D)) \neq \Gamma(A \cap D)$. In particular we have using Lemma 2.14.iii, $\Gamma(A \cap D) \neq D$.

By Lemma 2.2,

$$|\Gamma(A) \cap D| \geq |A \cap D| + \kappa(\Gamma[D]).$$

By Lemma 2.14.iii,

$$|\Gamma(A)| = |\Gamma(A \cap D)| + |\Gamma(A \setminus D)| \geq |A \cap D| + \kappa(\Gamma[C]) + |A \setminus D| = |A| + \kappa(\Gamma[C]).$$

□

3. Some basic additive inequalities generalised to relations

We begin by a generalisation of Mann Theorem to non-abelian groups and to relations with a transitive group of automorphisms.

A reflexive locally finite relation $\Gamma = (V, E)$ will be called a *Cauchy relation* if

$$\kappa(\Gamma) = \rho(\Gamma) - 1.$$

Lemma 3.1. — *Let $\Gamma = (V, E)$ be a reflexive locally finite relation. Then Γ is a Cauchy relation if and only if for every finite non-empty subset A of V,*

$$|\Gamma(A)| \geq \min(|V|, |A| + \rho(\Gamma) - 1).$$

If Γ is finite and regular, then Γ is a Cauchy relation if and only if Γ^- is a Cauchy relation.

Proof. — The first part follows easily from Lemma 2.1 and Lemma 2.2. The second part follows from Remark 2.a and Lemma 2.3. □

Lemma 3.2. — *Let B be a finite subset of a group G such that $1 \in B$. Then B is a Cauchy subset if and only if for every finite non-empty subset A of G,*

$$|AB| \geq \min(|G|, |A| + |B| - 1).$$

Proof. — Set $\Gamma = \Lambda(G, B)$. For every subset $F \subset G$, $\Gamma(F) = FB$. The result follows now easily by Lemma 3.1. □

According to Lemma 3.2, the Cauchy-Davenport inequality is satisfied for every non-empty subset A of G if B is a Cauchy subset. The Cayley graphs of such subsets are used in network models and said to be optimally connected.

Theorem 3.3. — *Let $\Gamma = (V, E)$ be a reflexive locally finite point transitive relation and let $v \in V$. Then Γ is a Cauchy relation if and only if one of the following conditions holds.*

(i) V is infinite and for every finite block B of Γ containing v,

$$|\Gamma(B)| \geq \min(|V|, |B| + d(\Gamma) - 1).$$

(ii) V is finite and for every block B of Γ containing v,

$$\min(|\Gamma(B)|, |\Gamma^-(B)|) \geq \min(|V|, |B| + d(\Gamma) - 1).$$

Proof. — By Lemma 3.1, the theorem is invariant by interchanging Γ and Γ^- in the finite case. We may assume without lost of generality V is infinite or $\mu(\Gamma) \leq \mu(\Gamma^-)$. The necessity follows by Lemma 3.1. Suppose that (ii) holds. We may assume that $\Gamma \neq V \times V$, since otherwise the result is obvious. By the transitivity of $\mathrm{Aut}(\Gamma)$, there exists an atom A of Γ such that $v \in A$.

By Corollary 2.8, A is a block.

It follows using the definition of an atom and (ii) that

$$\kappa(\Gamma) = |\Gamma(A)| - |A| \geq d(\Gamma) - 1.$$

By Lemma 2.2, we have $\kappa(\Gamma) = d(\Gamma) - 1$. Hence Γ is a Cauchy relation. □

Corollary 3.4. — *Let S be a finite subset of a group G such that $1 \in S$. Then S is a Cauchy subset if and only if one of the following conditions holds.*

(i) G is infinite and for every finite subgroup H of G,

$$|HS| \geq \min(|G|, |H| + |S| - 1).$$

(ii) G is finite and for every subgroup H of G,

$$\min(|SH|, |HS|) \geq \min(|G|, |H| + |S| - 1).$$

Proof. — Set $\Gamma = \Lambda(G, S)$. By Lemma 2.10 every finite block of Γ containing 1 is a subgroup. Observe that for every subgroup H, $|\Gamma^-(H)| = |HS^{-1}| = |SH|$. The result follows now using Theorem 3.3. □

The second part of Corollary 3.4 follows from [33, Theorem 1.2]. Zemor studied the same problem where B is not assumed to contain 1. A example contained in [33] shows that for a finite group G, the condition

$$\min(|SH|, |HS|) \geq \min(|G|, |H| + |S| - 1)$$

can not be replaced by the weaker one

$$|HS| \geq \min(|G|, |H| + |S| - 1).$$

Corollary 3.5 **([24,25]).** — *Let B be a finite non-empty subset of an abelian group G. Then the following conditions are equivalent.*

(i) For every finite non-empty subset A of G,

$$|A + B| \geq \min(|G|, |A| + |B| - 1).$$

(ii) For every finite subgroup H of G,

$$|H + B| \geq \min(|G|, |H| + |B| - 1).$$

Proof. — Choose $b \in B$ and set $S = B - b$. Using Lemma 3.2, one see easily that (ii) holds if and only if S is a Cauchy subset. The result follows now using Corollary 3.4. □

The following result generalises a result proved in [10] for finite relations.

Theorem 3.6. — *Let Γ be a locally finite connected reflexive point transitive relation such that $d(\Gamma) \geq d(\Gamma^-)$. Then $\kappa(\Gamma) \geq d(\Gamma)/2$.*

Proof. — According to Lemma 2.3 and Remark 2.a, the statement is invariant if we replace Γ by Γ^- if V is finite. Hence we may assume without lost of generality $\mu(\Gamma) \leq \mu(\Gamma^-)$, in the finite case.

Let M be an atom of Γ. By Corollary 2.7, M is a block. It follows easily that $\Gamma[M]$ is point transitive and that any other atom T generates a relation isomorphic to $\Gamma[M]$. Since M is finite, we have by Remark 2.a, $d(\Gamma[M]) = d((\Gamma[M])^-)$. Set $t = d((\Gamma[M]))$. Set $d^+ = d(\Gamma)$ and $d^- = d(\Gamma^-)$. Let X be the graph obtained from Γ by deleting all the arcs interior to the atoms. As for every block, the atoms partition V. It follows that $d(X) = d^+ - t$ and $d(X^-) = d^- - t$. The number of edges of R originating in M is not greater than the number of edges terminating in $\partial(M)$. Therefore $\sum_{x \in M}(d^+ - t) \leq \sum_{x \in \partial(M)}(d^- - t)$.

Therefore $|M|(d^+ - t) \leq \kappa(\Gamma)(d^- - t)$. Hence $|M|(d^+ - t) \leq \kappa(\Gamma)(d^+ - t)$. Observe that $d^+ - t \neq 0$, since otherwise $\kappa(\Gamma) = 0$, contradicting the assumption that Γ is connected. It follows that $|M| \leq \kappa(\Gamma)$.

Let $x \in M$, we have

$$d(\Gamma) = |\Gamma(x)| = |\Gamma(x) \cap M| + |\Gamma(x) \cap \partial(M)|.$$

It follows that

$$d(\Gamma) \leq |M| + \kappa(\Gamma) \leq 2\kappa(\Gamma).$$

This proves the theorem. □

Corollary 3.7. — *Let $\Gamma = (V, E)$ be a locally finite reflexive point transitive relation such that $d(\Gamma) \geq d(\Gamma^-)$.*

Then for every non-empty finite subset A, either $\Gamma(\Gamma(A)) = \Gamma(A)$ or $|\Gamma(A)| \geq |A| + d(\Gamma)/2$.

Proof. — Let C be a component of Γ such that $A \cap C \neq \varnothing$. By Lemma 2.14, $d(\Gamma) = d(\Gamma(C))$. By Theorem 3.6, $\kappa(\Gamma[C]) \geq d(\Gamma)/2$. The result follows now from Lemma 2.15. □

Corollary 3.8. — *(Olson [27]) Let A and B be finite nonempty subsets of a group G such that $1 \in B$. Then $|AB| \geq \min(|A\langle B\rangle|, |A| + |B|/2)$.*

Proof. — Let $\Gamma = \Lambda(G, B)$. Clearly $d(\Gamma) = |B| = |B^{-1}| = d(\Gamma^-)$. By Corollary 3.7, either $ABB = AB$ or $|AB \geq |A| + |B|/2$.

The result is now obvious since the two conditions $ABB = AB$ and $A\langle B\rangle = AB$ are equivalent (observe that A and B are finite). The second one implies the first by multiplication with B. Assume the first one holds. Hence $AB^j = AB$, for all $j \geq 1$. Since A, B are finite, this last condition implies easily that $A\langle B\rangle = AB$. □

As we have seen, Theorem 3.6 generalises Corollary 3.8 (Olson [27]) to point transitive relations. In the finite case, this generalisation was proved in [10, Proposition 3.4] before the result of Olson.

A relation $\Gamma = (V, E)$ is said to be *arc-transitive* if for all x, y, v, $w \in V$, such that $(x, y) \in E$ and $(v, w) \in E$ and $x \neq y$ and $v \neq w$, there is $f \in \mathrm{Aut}(\Gamma)$ such that $v = f(x)$ and $w = f(y)$. Observe that a connected arc-transitive relation is point transitive also.

A basic example of arc transitive relation is the following one.

Let R be a division ring and U be a finite multiplicative subgroup of $R \setminus \{0\}$. Set $\Omega = \Lambda(R, U \cup \{0\})$. The relation Ω is clearly point transitive. Let us prove that it is arc transitive. Consider two arcs (a, b) and (c, d) such that $b \neq a$ and $d \neq c$. Therefore $b - a$, $d - c \in U$. Consider the application $f(x) = (d-c)(b-a)^{-1}(x-a)+c$. Clearly $f(a) = c$ and $f(b) = d$. It remains to show that $f \in \mathrm{Aut}(\Omega)$. Now f is the composition of a translation and an application of the form $g(x) = ux$, where $u \in U$. The translation is an automorphism by Lemma 2.9. It remains to show that $g \in \mathrm{Aut}(\Omega)$. This follows from the following obvious equivalence.

$$x - y \in U \quad \text{if and only if} \quad ux - uy \in U.$$

The following result is proved in [9] in the finite case.

Theorem 3.9. — *Let $\Gamma = (V, E)$ be a locally finite connected reflexive arc-transitive relation.*

Then Γ is a Cauchy relation. In particular $\kappa(\Gamma) = d(\Gamma) - 1$.

Proof. — According to Lemma 2.3 and Remark 2.a, the statement is invariant if we replace Γ by Γ^- if V is finite. Hence we may assume without lost of generality $\mu(\Gamma) \leq \mu(\Gamma^-)$, in the finite case.

Let M be an atom of Γ. We shall prove that $|M| = 1$. Suppose the contrary. By Lemma 2.2, $\Gamma[M]$ is connected. In particular, there are $x, y \in M$ with $x \neq y$ and $(x, y) \in E$. Since Γ is connected, $\kappa(\Gamma) \geq 1$. In particular there is $v \in M$ and $w \in M$, such that $(v, w) \in E$. By the transitivity of the group of automorphisms on the arcs, there is $f \in \mathrm{Aut}(\Gamma)$ such that $f(x) = v$ and $f(y) = w$. It follows that $f(M) \neq M$ and $f(M) \cap M \neq \varnothing$, contradicting Corollary 2.8.

It follows that $|M| = 1$. Hence $\kappa(\Gamma) = d(\Gamma) - 1$, by Lemma 2.2. □

Corollary 3.10. — *Let G be a group containing a finite subset B such that $1 \in B$. Assume that $\Lambda(\langle B\rangle, B)$ is arc-transitive.*

Then for every finite subset $A \subset G$, $|AB| \geq \min(|A\langle B\rangle|, |A| + |B| - 1)$.

Proof. — The proof is similar to the proof of Corollary 3.8. □

Corollary 3.11. — *Let R be a division ring and let P be a finite subset of R such that $0 \in P$ and $P \setminus \{0\}$ is multiplicative subgroup. Then P is a Cauchy subset of the additive subgroup generated by P.*

Proof. — The result follows easily by Corollary 3.10. □

Corollary 3.11 generalises an inequality proved by Davenport-Lewis in [6] in the case of finite fields. We shall improve this result in section 7.

The notion of a base can be generalised easily to relations as follows.

A subset A of a group G is said to be a base with order h if h is the smallest integer such that $G = A^h$.

Let $\Gamma = (V, E)$ be a point transitive reflexive relation. The *diameter* of Γ is the smallest integer k such that $\Gamma^k = V \times V$, where Γ^k is the composition of Γ with itself k times.

Clearly if $1 \in A$, then A is a base of order h if and only if $\Lambda(G, A)$ has diameter h.

Lemma 3.12. — *Let $X = (V, E)$ be a finite connected reflexive point transitive relation with diameter h. Then*

$$h \leq \max(2, 3 + \frac{|V| - 2d(\Gamma)}{\kappa(\Gamma)}).$$

Proof. — The result holds if $h \leq 2$. Assume the contrary. Choose $v \in V$ Let X be a nonempty subset of G. By the definition of κ, we have

$$|\Gamma(\Gamma)| \geq \min(|V|, |X| + \kappa(\Gamma)).$$

It follows that

$$|\Gamma^{h-2}(v)| \geq \min(|V|, d(\Gamma) + (h-3)\kappa(\Gamma)).$$

Since h is the exact diameter of X, there is $y \in V$ such that $\Gamma^-(y) \cap \Gamma^{h-2}(v) = \varnothing$. Hence

$$|V| \geq |d^-(y)| + d(v) + (h-3)\kappa(\Gamma) = 2d(\Gamma) + (h-3)\kappa(\Gamma).$$

□

Theorem 3.13 **([11]).** — *Let $\Gamma = (V, E)$ be a finite connected vertex transitive reflexive relation with diameter h. Then $h \le \max\left(2, \frac{2|V|}{d(\Gamma)} - 1\right)$.*

Proof. — By Lemma 3.12,

$$h \le \max(2, 3 + \frac{|V| - 2d(\Gamma)}{\kappa(\Gamma)}).$$

By Theorem 3.6,

$$h \le \max(2, 3 + \frac{2(|V| - 2d(\Gamma))}{d(\Gamma)}).$$

Therefore

$$h \le \max(2, \frac{2|V|}{|d(\Gamma)|} - 1).$$

□

There was an omission in the statement in [11]. We did add max(2, . . .), to correct the statement.

Theorem 3.13, applied to bases of finite groups, is proved independently by Rödseth [29]. This last result is used in [15] to generalise results of Cherly and Deshouillers [3], Jia and Nathanson [17] to arbitrary σ-finite groups.

4. The critical inequalities

We introduce in this section some new objects. The properties of these objects will be used later to solve the critical pair problem.

Let Γ be a relation on a set V. A fragment F of Γ is said to be a *strict fragment* if $\mu(\Gamma) + 1 \le |F|$ and $\mu(\Gamma^-) + 1 \le |\delta(F)|$.

The relation Γ is said to be *degenerate* if Γ has a finite strict fragment. Let G be a group containing a subset B such that $1 \in B$. We shall say that B is *degenerate* if $\Lambda(G, B)$ is degenerate.

Remark 4.a. — Let Γ be a relation on a finite set V and let $F \subset V$. The following conditions are equivalent.

(i) F is a strict fragment of Γ.

(ii) $\delta(F)$ is a strict fragment of Γ^-.

Proof. — Suppose that (i) holds. By Lemmas 2.3 and 2.4, $\delta(F)$ is a fragment of Γ^-. We have also $\delta^-(\delta(F)) = F$. Therefore (ii) holds. The other implication holds by duality using Lemmas 2.3 and 2.4. □

Lemma 4.1

(i) Let Γ be reflexive Cauchy relation on a set V. Then Γ is non-degenerate if and only if for all $A \subset V$ such that $2 \le |A| < \infty$,

$$|\Gamma(A)| \ge \min(|V| - 1, |A| + \kappa(\Gamma) + 1).$$

(ii) Let B be a finite Cauchy subset of a group G such that $1 \in B$. Then B is non-degenerate if and only if for all $A \subset G$ such that $2 \leq |A| < \infty$,

$$|AB| \geq \min(|G| - 1, |A| + |B|).$$

Proof. — The result holds trivially if V is infinite. Observe that a fragment with cardinality $\neq 1$ is a strict fragment in this case. Assume the contrary. By Lemma 3.1, Γ^- is also a Cauchy relation.

Suppose that there is $A \subset V$ such that $2 \leq |A|$ and

$$|\Gamma(A)| < \min(|V| - 1, |A| + \kappa(\Gamma) + 1).$$

By Lemma 2.1, we have $|\Gamma(A)| = |A| + \kappa(\Gamma)$. Hence A is a fragment. We have $|\delta(A)| = |V| - |\Gamma(A)| \geq 2$. It follows that Γ is degenerate. Similarly one see easily that if Γ is degenerate then every strict fragment A verifies the inequality

$$|\Gamma(A)| < \min(|V| - 1, |A| + \kappa(\Gamma) + 1).$$

Clearly (ii) is a particular case of (1). □

Lemma 4.2. — *Let Γ be a reflexive regular Cauchy relation on a finite set V and let F be a fragment of Γ. The following conditions are equivalent.*

(i) $F = V \setminus \Gamma^-(x)$, for some $x \in V$.

(ii) $|F| = |V| - d(\Gamma)$.

(iii) $|\delta(F)| = 1$.

Proof. — Suppose that $F = V \setminus \Gamma^-(x)$, for some $x \in V$. We have

$$|F| = |V| - |\Gamma^-(x)| = |V| - d(\Gamma^-) = |V| - d(\Gamma).$$

Hence (ii) holds. Suppose now that (ii) holds. We have

$$|\delta(F)| = |V \setminus \Gamma(F)| = |V| - |F| - \kappa(\Gamma) = 1.$$

Hence (iii) holds. Suppose now that (iii) holds and take $\delta(F) = \{y\}$. By Lemmas 2.3 and 2.4,

$$F = \delta^-\delta(F) = V \setminus \Gamma^-(y).$$

□

The minimal cardinality of a strict fragment of a degenerate relation Γ will be denoted by $\omega(\Gamma)$. Clearly $\omega(\Gamma)$ is finite. A strict fragment of Γ with cardinality $\omega(\Gamma)$ will be called a *superatom* of Γ.

Lemma 4.3. — *Let Γ be a reflexive relation on a finite set V. Then Γ is degenerate if and only if it is Γ^- is degenerate.*

Proof. — Using Remark 4.a, a fragment F is strict with respect to Γ if and only if $\delta(F)$ is strict with respect to Γ^-. Using Lemma 2.4, we see easily that Γ is degenerate if and only if Γ^- is degenerate. □

Let Γ be a degenerate Cauchy relation and let K be a superatom of Γ. We shall say that Γ is *singular* if $|\delta(K)| \leq |K| - 1$.

Notice that Γ is singular if and only if $2\omega(\Gamma) + \kappa(\Gamma) - 1 \geq |V|$. In particular every singular relation is finite.

Proposition 4.4. — *Let Γ be a reflexive regular degenerate Cauchy relation. Assume that Γ is singular and let K and M be superatoms of Γ. Then either $\delta(K) = \delta(M)$ or $\delta(K) \cap \delta(M) = \varnothing$.*

Proof. — Suppose on the contrary that $\delta(K) \neq \delta(M)$ and $\delta(K) \cap \delta(M) \neq \varnothing$. By Lemma 2.4 $K \neq M$, $\delta^-(\delta(K)) = K$ and $\delta^-(\delta(M)) = M$. Using Lemma 2.5 applied to Γ^-, we have

$$|K \setminus M| = |M \setminus K| \leq |\delta(K) \setminus \delta(M)|.$$

Since $|K| > |\delta(K)|$ we have $K \cap M \neq \varnothing$. By Lemma 2.5, $K \cap M$ is a fragment of Γ. Since K is a superatom of Γ we have $|K \cap M| = 1$. Clearly

$$|K| = |M| = 1 + |M \setminus K| \leq |\delta(K)|,$$

a contradiction. □

Proposition 4.5. — *Let Γ be a reflexive degenerate Cauchy relation on a set V and let M be a superatom of Γ. Let F be a finite strict fragment of Γ such that $M \not\subset F$, $M \cap F \neq \varnothing$ and $|\delta(F)| \geq |M|$. Then*

(i) $|M \cap F| = 1$

(ii) $\Gamma(M \cap F) = \Gamma(M) \cap \Gamma(F)$.

Proof. — By Lemma 2.5, $|\delta(F) \setminus \delta(M)| < |M|$. Therefore $\delta(F) \cap \delta(M) \neq \varnothing$.

By Lemma 2.5, (ii) is satisfied and $M \cap F$ is a fragment of Γ. By the definition of a superatom and since $M \cap F \neq M$, we have $|M \cap F| = 1$. □

The above proposition generalises a result proved in the case of finite symmetric relations by Jung. Our finite symmetric relations are equivalent to undirected graphs considered by Jung, cf. [18, sätz 2]. The notion of a superatom coincides in this case with the notion of a 2-atom of Jung.

Corollary 4.6. — *Let Γ be a reflexive degenerate Cauchy relation. Assume that Γ is non-singular and let K and M be two distinct superatoms of Γ such that $K \cap M \neq \varnothing$. Then $|K \cap M| = 1$ and $\Gamma(K \cap M) = \Gamma(K) \cap \Gamma(M)$.*

Proof. — By the definition we have $|\delta(K)| \geq |K| = |M|$. By Proposition 4.5, we have $|K \cap M| = 1$ and $\Gamma(K \cap M) = \Gamma(K) \cap \Gamma(M)$. □

Corollary 4.7. — *Let Γ be a reflexive degenerate Cauchy relation such that $\omega(\Gamma) \leq \omega(\Gamma^-)$. Let M be a superatom of Γ and let F be a finite strict fragment of Γ. Then either $M \subset F$ or $|M \cap F| \leq 1$.*

Proof. — The inequality $\omega(\Gamma^-) \leq |\delta(F)|$ holds clearly if V is infinite and follows in the finite case by Remark 4.a. Therefore $|M| = \omega(\Gamma) \leq \omega(\Gamma^-) \leq |\delta(F)|$. The corollary follows now by Proposition 4.5. □

Proposition 4.8. — *Let Γ be a reflexive point transitive relation on a set V such that both Γ and Γ^- are non-singular degenerate Cauchy relations and $d(\Gamma) = d(\Gamma^-)$. Assume that*

$$3 \leq \min(\omega(\Gamma), \omega(\Gamma^-)).$$

Then one of the following conditions holds.

(i) Any three distinct superatoms of Γ *have an empty intersection.*

(ii) Any three distinct superatoms of Γ^- *have an empty intersection.*

Proof. — The statement is invariant by interchanging Γ and Γ^-. We may assume without lost of generality that $\omega(\Gamma) \leq \omega(\Gamma^-)$.

Suppose on the contrary that both (i) and (ii) are not satisfied. Choose two distinct superatoms A , B of Γ and an element x such that $x \in A \cap B$.

Since $A \cap B \neq \varnothing$, we have by Lemma 2.5

$$|\delta A \setminus \delta B| \leq |B \setminus A| < |B| = \omega(\Gamma). \tag{1}$$

As in the proof of Corollary 4.7, we have $\omega(\Gamma) \leq |\delta(A)|$. Therefore we have using (1), $\delta(A) \setminus \delta(B) \neq \varnothing$. Choose $y \in \delta(A) \setminus \delta(B)$.

Let K, L and M be distinct superatoms of Γ^- such that $y \in K \cap L \cap M$. Such superatoms exist by the transitivity of the group of automorphisms and by the hypothesis that (ii) is not satisfied. By Corollary 4.6,

$$K \cap L = K \cap M = L \cap M = \{y\}. \tag{2}$$

Suppose that there are F_1 and $F_2 \in \{K, L, M\}$ such that $F_1 \cup F_2 \subset \delta(A)$ and $F_1 \neq F_2$. By (2),

$$|F_1 \cap F_2| = 1. \tag{3}$$

Let $i \in \{1, 2\}$. By Lemma 2.4, $\delta(B)$ is a fragment of Γ^-. By Lemma 2.4, $A = \delta^-(\delta(A))$ and $A \subset \delta^-(F_1) \cap \delta^-(F_2)$. Hence $x \in \delta^-(F_1) \cap \delta^-(F_2)$. Now we have $y \in F_i \setminus \delta(B)$ and $x \in B \cap \delta^-(F_i) = \delta^-(\delta(B)) \cap \delta^-(F_i)$ (using Lemma 2.4). By Lemma 2.5, applied to Γ^- with $M = F_i$ and $F = \delta(B)$, we have $F_i \cap \delta(B) = \varnothing$ or $F_i \cap \delta(B)$ is a fragment of Γ^-. By the definition of a superatom we have $|F_i \cap \delta(B)| \leq 1$. Therefore

$$|(F_1 \cup F_2) \cap \delta(B)| \leq 2. \tag{4}$$

By (4) we have

$$|(F_1 \cup F_2) \cap (\delta(A) \setminus \delta(B))| \geq |F_1 \cup F_2| - 2. \tag{5}$$

By (3) and (5), we have

$$|\delta(A) \setminus \delta(B)| \geq |F_1| + |F_2| - 3 \geq \omega(\Gamma^-) \geq \omega(\Gamma).$$

This inequality contradicts (1). It follows that at most one superatom $F \in \{K, L, M\}$ is contained in $\delta(A)$. We may assume without loss of generality $K \not\subset \delta(A)$ and $L \not\subset \delta(A)$. By Lemma 2.4, $A \not\subset \delta^-(K)$ and $A \not\subset \delta^-(L)$.

By Corollary 4.7, $|A \cap (\delta^-(K) \cup \delta^-(L))| \leq 2$. Therefore

$$A \cap (\Gamma^-(K) \cap \Gamma^-(L)) = A \setminus (\delta^-(K) \cup \delta^-(L)) \neq \varnothing.$$

By (2) and Corollary 4.6,

$$\Gamma^-(y) = \Gamma^-(K \cap L) = \Gamma^-(L) \cap \Gamma^-(L).$$

Therefore $\Gamma^-(y) \cap A \neq \varnothing$, contradicting the assumption $y \in \delta(A)$. This contradiction proves the result. □

Proposition 4.8 generalises a result proved in [14].

5. The Vosper inequality

We apply in this section the results obtained in section 4 to the case of a finite group. Let G be a group and let $r \in G$. The subgroup of G generated by r will be denoted by $\langle r \rangle$. We recall the following elementary fact.

Remark 5.a. — Let S be a finite subset of G and let $r \in G$. The following conditions are equivalent.

(i) S is a union of right $\langle r \rangle$-cosets.
(ii) $\langle r \rangle S = S$.
(iii) $rS = S$.

In this section we study the inequality $|AB| \geq \min(|G|-1, |A|+|B|)$, where A and B are subsets of a finite group G.

Theorem 5.1. — *Let B be a degenerate Cauchy subset of a finite group G such that $1 \in B$. Set $\Gamma = \Lambda(G, B)$. Let L be a superatom of Γ and let M be a superatom of Γ^- such that $1 \in L \cap M$.*

(i) If B is singular, then $x^{-1}\delta(L)$ is a subgroup for every $x \in \delta(L)$.

(ii) If B and B^{-1} are non-singular, then there are a subgroup H and $a \in G$ such that $L = H \cup Ha$ or $M = H \cup Ha$.

Proof. — Suppose first that Γ is singular. Choose $y \in \delta(K)$ and set $M = y^{-1}\delta(K)$. We have clearly $1 \in M$. Let $x \in M$. We have clearly $xM \cap M \neq \varnothing$. But

$$xM = x(G \setminus y^{-1}K) = \delta(xy^{-1}K).$$

By Proposition 4.4, $M = xM$. Hence $MM = M$ and therefore M is a subgroup. This proves (i).

Assume now that Γ and Γ^- are non-singular. By Lemma 4.4, Γ^- is degenerate. The result holds clearly if $\omega(\Gamma) = 2$ or $\omega(\Gamma^-) = 2$. Assume $\omega(\Gamma) \geq 3$ and $\omega(\Gamma^-) \geq 3$. By Lemma 3.1, B^{-1} is a Cauchy subset. By Proposition 4.8, there exists $\Psi \in \{\Gamma, \Gamma^-\}$ such that any three distinct superatoms of Ψ are disjoint.

Set $K = L$ if $\Gamma = \Psi$ and $K = M$ if $\Gamma^- = \Psi$. By Lemma 2.8, for any $x \in G$, xK is a superatom of Ψ. This observation will be used without reference.

Take $H = \{x \mid xK = K\}$. Clearly H is a subgroup contained in K. Let $Q = K \setminus H$. If $Q = \varnothing$, the result holds with $a = 1$. Assume $Q \neq \varnothing$ and let $a \in Q$.

Let $x \in Q$, we have $1 \in K \cap a^{-1}K \cap x^{-1}K$. By Proposition 4.8, two of these superatoms coincide. Since $a, x \in Q$, we have $a^{-1}K \neq K$ and $x^{-1}K \neq K$. Therefore $a^{-1}K = x^{-1}K$. Hence $x \in Ha$. Hence $Q \subset Ha$. Since $|K| \geq 3$ and $K = H \cup Q$, we have $|H| \geq 2$.

Let $x \in H$. We have $|xK \cap K| \geq |H| \geq 2$. By Corollary 4.6, $xK = K$. Therefore $HK = K$. Hence K is a union of right cosets of H. Hence $|Q| \geq |H|$ and therefore $|Q| = |H|$. It follows that $K = H \cup Ha$. □

Corollary 5.2. — *Let B be a degenerate Cauchy subset of a finite group G such that $1 \in B$. There are $S \in \{B, B^{-1}\}$, a subgroup H and $a \in G$ such that $H \cup Ha$ is a strict fragment of $\Lambda(G, S)$.*

Proof. — This result follows immediately from Theorem 5.1. □

Theorem 5.3. — *Let B be a subset of a finite group G such that $1 \in B$. Then the following conditions are equivalent.*

(i) For all $A \subset G$ such that $2 \leq |A|$,

$$|AB| \geq \min(|G| - 1, |A| + |B|).$$

(ii) For every subgroup H of G and for every $a \in G$ such that $|H \cup Ha| \geq 2$,

$$\min(|B(H \cup aH)|, |(H \cup Ha)B|) \geq \min(|G| - 1, |H \cup Ha| + |B|).$$

Proof. — Suppose that (i) holds. It follows that for every non-empty $A \subset G$,

$$|AB| \geq \min(|G|, |A| + |B| - 1).$$

By Lemma 3.2, B is a Cauchy subset.

By Lemma 4.1, B is non-degenerate. By Lemma 4.3, B^{-1} is non-degenerate. Hence for all $A \subset G$ such that $2 \leq |A|$,

$$|AB^{-1}| \geq \min(|G| - 1, |A| + |B|).$$

Therefore (ii) holds. Suppose that (i) is not satisfied. Hence there exists A such that $2 \leq |A|$ and

$$|AB| \leq \min(|G| - 2, |A| + |B| - 1).$$

Case 1. B is a Cauchy subset. — By Lemma 4.1, B is degenerate. By Corollary 5.2, there are $a \in G$ and a subgroup H such that $H \cup Ha$ is a fragment of $\Lambda(G, S)$ or a fragment of $\Lambda(G, S^{-1})$. In this case (ii) is not satisfied.

5.0.1. *Case 2. B is not a Cauchy subset.* — By Corollary 3.5, there exists a subgroup H of G such that $\min(|BH|, |HB|) \leq \min(|G| - 1, |H| + |B| - 2)$.

Clearly $|H| \geq 2$. Since $|H|$ divides $|G|$, $|BH|$ and $|HB|$, we have

$$\min(|BH|, |HB|) \leq \min(|G| - |H|, |H| + |B| - 1).$$

Therefore

$$\min(|BH|, |HB|) \leq \min(|G| - 2, |H| + |B| - 1).$$

It follows that (ii) is not satisfied (with $a = 1$). □

6. The critical pair theory

Let G be a group and let $r \in G \setminus \{1\}$. A subset $B \subset G$ will be called a *right progression with ratio r*, if there are $b \in G$ and a number k such that $1 \leq k < |\langle r \rangle|$ such that $B = \{b, rb, r^2b, \ldots, r^{k-1}b\}$.

A subset $B \subset G$ will be called a *right coprogression with ratio r*, if $G \setminus B$ is a right progression with ratio r.

A subset $B \subset G$ will be called a *left progression with ratio r*, if there are $b \in G$ and a number k such that $1 \leq k < |\langle r \rangle|$ such that $B = \{b, br, br^2, \ldots, br^{k-1}\}$.

A subset $B \subset G$ is a *left coprogression with ratio r*, if $G \setminus B$ is a left progression with ratio r.

We say that a subset $B \subset G$ is a *right semi-progression with ratio* r, if there are $b \in G$ and a number k such that $1 \le k < |\langle r\rangle|$ satisfying the following properties.

(1) $B \supset \{b, rb, r^2b, \ldots, r^{k-1}b\}$

(2) $B \setminus \{b, rb, r^2b, \ldots, r^{k-1}b\}$ is a union (possibly void) of right $\langle r\rangle$-cosets.

A right semi-progression with $k = 1$ will be called a *right almost-periodic.*

A right semi-progression with $B \supset G \setminus \langle r\rangle b$ is a *right coprogression.* A subset B is said to be a *left semi-progression* if B^{-1} is a right semi-progression.

We introduce the following notion. Let $r \in G \setminus \{1\}$ and let $A \subset \langle r\rangle$. We say that $\{r^i, r^{i+1}, \ldots, r^j\}$ is an *r-string* of A if $\{r^i, r^{i+1}, \ldots, r^j\} \subset A$ and $\{r^{i-1}, r^{j+1}\} \cap A = \varnothing$.

Lemma 6.1. — *Let B be a finite subset of a group G and let $r \in G \setminus \{1\}$.*

If $|\{1,r\}B| = |B| + 1$, then B is a right semi-progression with ratio r.

Proof. — Take $B = B_1 \cup B_2 \cup \cdots \cup B_k$, where B_i is the intersection of B with an $\langle r\rangle$-right coset. We assume also $B_i \neq \varnothing$, for all $1 \le i \le k$.

We have

$$|\{1,r\}B| = |\{1,r\}B_1| + \cdots + |\{1,r\}B_k| = |B_1| + \cdots + |B_k| + 1.$$

It follows that there is j, $1 \le j \le k$, such that

(i) $|\{1,r\}B_j| = |B_j| + 1$.

(ii) $|\{1,r\}B_i| = |B_i|$, for all $i \neq j$.

By (ii), we have $rB_i = B_i$, for all $i \neq j$. It follows using Remark 5.a that B_i is an $\langle r\rangle$-right coset, for all $i \neq j$.

It remains to show that B_j is a right progression with ratio r. Take $x \in G$ such that $B_j \subset \langle r\rangle x^{-1}$ and let $C = B_j x$. It would be enough to show that C is an r-string.

We have clearly $|\{1,r\}C| = |C| + 1$ and $C \subset \langle r\rangle$. We decompose C into $\langle r\rangle$-strings. Clearly every string $\{r^i, r^{i+1}, \ldots, r^j\}$ of C determines uniquely an element r^{j+1} of $\{1,r\}C \setminus C$. Hence there is exactly one string. □

Lemma 6.2. — *Let G be a finite group and let B be a right semi-progression. If B is a Cauchy subset then one of the following conditions holds.*

(i) B is a right almost-periodic subset.

(ii) B is a right coprogression.

Proof. — We have $|\langle r\rangle B| = |B \setminus (\langle r\rangle b)| + |\langle r\rangle|$.

If $\langle r\rangle B = G$, then clearly B is right coprogression. Assume $\langle r\rangle B \neq G$.

By the definition of κ, we have

$$|B| - 1 = \kappa \le |\partial\langle r\rangle| = |B \setminus (\langle r\rangle b)| = |B| - |B \cap b\langle r\rangle|.$$

It follows that $|B \cap (b\langle r\rangle)| = 1$. Thus B is right almost-periodic. □

Proposition 6.3. — *Let G be a finite group and B be a Cauchy subset of G such that $(|G|, |B| - 1) = 1$. Then B is degenerate if and only if B a right coprogression or a left coprogression.*

Proof. — Set $\Gamma = \Lambda(G,B)$. By Corollary 5.2, there are $S \in \{B, B^{-1}\}$, a finite subgroup H and $r \in G$ such that $K = H \cup Hr$ is a strict fragment $\Lambda(G,S)$. We have

$$|B| - 1 = |S| - 1 = \kappa(\Lambda(G,S)) = |(H \cup Ha)S| - |H \cup Ha|.$$

Hence $|H|$ divides $|B| - 1$. Therefore $|H| = 1$ and $r \neq 1$. Hence $K = \{1, r\}$. By Lemma 6.1, S is a right semi-progression with ratio r. The subset S can not be almost-periodic since otherwise $|\langle r\rangle|$ would divide $|S| - 1$.

By Lemma 6.2, S is a right coprogression. Clearly B is a right coprogression if $S = B$. It follows easily that B is a left coprogression if $S = B^{-1}$. □

We need the following lemma.

Lemma 6.4. — *Let A be a finite subset of a group G and let $r \in G \setminus \{1\}$. Let B be a finite right coprogression with ratio r such that $|B| \geq 2$ and $|AB| = |A| + |B| - 1 \leq |G| - 1$. Then A is a left progression with ratio r.*

Proof. — Take $B = G \setminus \langle r\rangle b \cup \{b, rb, r^2b, \dots, r^{k-1}b\}$ and take $a \in A$. Let $C = a^{-1}A$ and let $D = Bb^{-1}$. Clearly $|CD| = |C| + |D| - 1 \leq |G| - 1$.

We shall prove first that $C \subset \langle r\rangle$. Assume there is $x \in C\setminus\langle r\rangle$. Since $D \supset G\setminus\langle r\rangle$, we have $x^{-1}\langle r\rangle \subset D$. It follows that $\langle r\rangle \subset CD$. Since $1 \in C$, we have $G\setminus\langle r\rangle \subset D \subset CD$. Therefore $CD = G$, a contradiction.

This shows that $C \subset \langle r\rangle$. The argument used in the last part of the proof of Lemma 6.1, shows that A is a left progression of $\langle r\rangle$. □

Lemma 6.5. — *Let A be a finite subset of a group G with cardinality m and let $r \in G \setminus \{1\}$. Let $B = (G \setminus b\langle r\rangle) \cup \{b, br, br^2, \dots, br^{k-1}\}$ be a finite left coprogression with ratio r such that $|B| \geq 2$ and $|AB| = |A| + |B| - 1 \leq |G| - 1$. Then there are $a \in G$ such that $A = a\{1, r, \dots, r^{m-1}\}b^{-1}$.*

Proof. — Take $a \in A$ and set $C = a^{-1}A$. We shall see that

$$b^{-1}Cb \subset \langle r\rangle \tag{1}$$

Assume there is $c \in C$ such that $b^{-1}cb \notin \langle r\rangle$. It follows that $c^{-1}b \notin b\langle r\rangle$. Since $B \supset G \setminus b\langle r\rangle$, we have $c^{-1}b\langle r\rangle \subset B$. It follows that $b\langle r\rangle \subset CB$. Since $1 \in C$, we have $G \setminus b\langle r\rangle \subset B \subset CB$. Therefore $CB = G$. It follows that $AB = G$, a contradiction.

Set now $B_1 = B \setminus b\langle r\rangle$ and $B_2 = B \cap (\langle r\rangle)$. Let us prove that

$$CB_1 \cap CB_2 = \varnothing. \tag{2}$$

We have clearly $|CB_1| \leq |CB| < |G|$. Since CB_1 is a union of left cosets we have $|CB_1| \leq |G| - |\langle r\rangle|$. On the other side $B_1 \subset CB_1$. Therefore $CB_1 = B_1$. Now (2) follows easily from (1).

It follows that

$$|C| + |B_1| + |B_2| - 1 = |CB| = |CB_1| + |CB_2| = |B_1| + |CB_2|.$$

Therefore

$$|CB_2| = |C| + |B_2| - 1. \tag{3}$$

Therefore $|(b^{-1}Cb)(b^{-1}B_2)| = |b^{-1}Cb| + |b^{-1}B_2| - 1$. Now $b^{-1}Cb$ and $b^{-1}B_2$ are subsets of $\langle r\rangle$. By Lemma 6.4, $b^{-1}Cb$ progression with ratio r. The result follows now easily. □

We prove now the main result of this section. It implies a generalisation of Vosper Theorem to any Cauchy subset of a finite group, where we replace the condition "$|G|$ is prime" in Vosper Theorem by the weaker one "$(|G|, |B|-1) = 1$".

Theorem 6.6. — *Let G be a finite group and let B be a Cauchy subset of G such that $(|G|, |B|-1) = 1$.*

Let $A \subset G$ such that $|AB| = |A| + |B| - 1 \leq |G| - 1$. Then one of the following conditions holds.

(i) $|A| = 1$ or $A = G \setminus aB^{-1}$, for some $a \in G$.

(ii) There are $a, b, r \in G$, $k, s \in \mathbf{N}$ such that

$$A = \{a, ar, ar^2, \dots, ar^{k-1}\} \quad \textit{and} \quad B = (G \setminus \langle r\rangle b) \cup \{b, rb, r^2b, \dots, r^{s-1}b\}.$$

(iii) There are $a, b, r \in G$, $k, s \in \mathbf{N}$ such that

$$A = \{ab^{-1}, arb^{-1}, ar^2b^{-1}, \dots, ar^{k-1}b^{-1}\} \quad \textit{and} \quad B = (G\backslash b\langle r\rangle)\cup\{b, br, br^2, \dots, br^{s-1}\}.$$

Proof. — Assume now that (i) does not hold. Then $|A| \geq 2$. By Lemma 4.2, $|\delta(A)| \geq 2$. It follows that A is a strict fragment and hence by Lemma 4.1, $\Lambda(G, B)$ is degenerate.

By Proposition 6.3, there exists $r \in G\backslash\{1\}$ such that B is a right coprogression or a left coprogression with ratio r. Consider first the case where B is a right coprogression. Choose $b \in G$ and $s \in \mathbf{N}$ such that $B = G \setminus \{b, rb, r^2b, \dots, r^{s-1}b\}$

By Lemma 6.4, A is a left progression with ratio r. Choose $a \in G$ and $k \in \mathbf{N}$ such that $A = \{a, ar, ar^2, \dots, ar^{k-1}\}$. Therefore (ii) holds. A similar argument using Lemma 6.5 shows that (iii) holds if B is a left coprogression. □

***Corollary 6.7* (Vosper Theorem).** — *Let p be a prime number, and let A and B be two non-empty subsets of Z_p such that*

$$|A + B| = |A| + |B| - 1 \leq p - 1.$$

Then one of the following conditions holds.

(i) $|A| = 1$ or $|B| = 1$

(ii) $A = Z_p \setminus (a - B)$, for some $a \in Z_p$.

(iii) A and B are arithmetic progressions with the same difference

Proof. — Vosper Theorem may be reduced without lost of generality to subsets B such that $0 \in B$ and $|B| \geq 2$. Using the Cauchy-Davenport Theorem, B is a Cauchy subset.

The result is now an obvious consequence of Theorem 6.6. □

***Corollary 6.8* ([14]).** — *Let B be a Cauchy subset of an abelian group G. Then B is degenerate if and only if one of the following conditions holds.*

and (i) B is a progression or B is a coprogression.

(ii) There exists a finite subgroup H such that

$$|H| \geq 2 \quad \textit{and} \quad |G| > |H+B| = |H| + |B| - 1.$$

The proof of this result follows along the lines of Theorem 5.3. One should use the fact that an abelian Cayley relation is isomorphic to its inverse to show that the inverse relation is also degenerate.

7. Diagonal forms over a division ring

The estimation of the range of a diagonal form is one of the classical applications of the critical pair theory, cf. [4, 29, 14]. Let us show that our methods imply the validity the estimation given in [4, 29, 14] for finite fields in the case of an arbitrary division ring.

Let us begin by a general lemma.

Lemma 7.1. — *Let G be a group and let B be a finite subset of G such that $1 \in B$. Assume that $\Lambda(\langle B\rangle, B)$ is a a nondegenerate Cauchy subset of G.*

Then for every finite subset A such that $|A| \geq 2$,

$$|AB| \geq \min(|A\langle B\rangle| - 1, |A| + |B|).$$

The proof is similar to the proof of Lemma 2.15.

Lemma 7.2. — *Let R be a division ring and let P be a finite subset of R such that $0 \in P$ and $P \setminus \{0\}$ is multiplicative subgroup. If $|R| > |P| \geq 4$, then P is neither an arithmetic progression nor a coprogression.*

Proof. — This result is proved in [25] in the case of primes fields. The argument given there is not easy to generalise to our case. But we shall deduce this result using the fact that $\Gamma = \Lambda(\langle P\rangle, P)$ is arc-transitive.

Consider the case of an arithmetic progression. The case of a coprogression works in the same way. Assume that P is an arithmetic progression. Set

$$P = \{a, a+r, a+2r, \ldots, a+(k-1)r\}.$$

We may assume without loss of generality that $r, 2r \in P$. Therefore one $P = \{b, b+1, b+2, \ldots, b+(k-1)\}$, where $b = ar^{-1}$. It follows that

$$|\Gamma(0) \cap \Gamma(1)| = k - 1 > |\Gamma(0) \cap \Gamma(2)|,$$

contradicting the arc-transitivity of Γ. □

Proposition 7.3. — *Let R be a division ring and let P be a finite subset of R such that $0 \in P$ and $P \setminus \{0\}$ is multiplicative subgroup. Let R_0 be the additive subgroup generated by P. Then P is a non-degenerate Cauchy subset of R_0.*

Proof. — By Corollary 3.11, P is a Cauchy subset of R_0. Suppose that P is degenerate. By Lemma 7.2, P can not be a progression or a coprogression. By Corollary 6.8, there is a finite non-trivial subgroup $H \subset R_0$ such that $|R_0| > |H+B| = |H|+|P|-1$. Let p be the characteristic of R. Clearly p divides the order of $|H|$. It follows that p

divides $|P| - 1$. Since $P \setminus \{0\}$ is a subgroup, it follows that $u \in P \setminus \{0,1\}$ such that $u^p = 1$. Hence $(u-1)^p = 0$, a contradiction. □

Theorem 7.4. — *Let R be a division ring and let P be a finite subset of R such that $0 \in P$ and $P \setminus \{0\}$ is multiplicative subgroup. Let R_0 be the additive subgroup generated by P.*

Suppose that $|P| \geq 4$ and let $a_1, a_2, \ldots, a_n$ be non-zero elements of R. Then

$$|a_1P + a_2P + \cdots + a_nP| \geq \min(|R_0|, (2n-1)(|P|-1)+1).$$

Proof. — The proof is by induction. The statement is obvious for $n = 1$. Suppose it true for n. We may assume clearly $a_{n+1} = 1$. By Lemma 7.2, Proposition 7.3 and the induction hypothesis, we have

$$|b_1P + b_2P + \cdots + b_nP + P| \geq \min(|R_0| - 1, 2n(|P|-1)+2).$$

Set $U = P \setminus \{0\}$. Since

$$((a_1P + a_2P + \cdots + a_nP + P) \setminus \{0\})U = (a_1P + a_2P + \cdots + a_nP + P) \setminus \{0\}.$$

It follows that $|U|$ divides $|a_1P + a_2P + \cdots + a_nP + P| - 1$. It follows that

$$|a_1P + a_2P + \cdots + a_nP + P| \geq \min(|R_0| - 1, (2n+1)(|P|-1)+1).$$

It follows easily from this equality that

$$|a_1P + a_2P + \cdots + a_nP + P| \geq \min(|R_0|, (2n+1)(|P|-1)+1).$$

□

Theorem 7.4 was first proved in the case of $\mathbf{Z}_p$, by Chowla, Mann and Straus in [4]. Tietäväinen proved in [29] the above Theorem 7.4 in the case of finite fields with odd characteristics. We gave in [14] a proof for all finite fields based on the method of superatoms.

8. An application to networks

In this section, we identify a relation and its graph. We assume the loops coloured with white and the other edges coloured black.

A network will be modelled by a reflexive graph. The usual models are graphs without loops. Basically the two models are equivalent. The first one is more appropriate in our approach. In particular all the results and notions contained in this paper apply immediately. The second model requires some easy transformations. The reader could consider the black part as the network model and the white part as introduced for theoretical reasons. A point will be called a node or a vertex and an edge will be called a link (directed one).

Let $\Gamma = (V, E)$ be a reflexive graph. A *sink* of Γ is a proper finite subset of V such that $\Gamma(A) = A$. Clearly Γ is connected if and only if Γ has no sinks. We shall say that Γ is strongly connected if for all x,$y \in V$, there is a directed path from x to y. It is easy to show that a finite graph is connected if and only if it is strongly connected. This is not the case for infinite graphs. The Cayley graph $\Lambda(\mathbf{Z}, \{1\})$ is clearly connected and not strongly connected.

From now on, all the graphs considered will be assumed for simplicity finite.

Let $\Gamma = (V, E)$ be a finite reflexive regular graph. A set of vertices will be called a *cutset* if its deletion and its incident edges disconnects the graph. A cutset with smallest cardinality is called a *minimum cutset.* It is easy to see that the cardinality of a minimum cutset is $\kappa(\Gamma)$. Let us mention that a fragment is just a sink in the subgraph obtained by the deletion of a minimum cutset.

In a good network, the connectivity should be maximised. By Lemma 2.2, the maximal possible value of the connectivity is $d(\Gamma) - 1$. If In particular, if $\kappa(\Gamma) = d(\Gamma) - 1$, then after the failure of $d(\Gamma) - 2$ nodes, the remaining nodes remain connected. This property shows that Γ must be a Cauchy graph. The next property studied in network models is the superconnectdness. Let $x \in V$, clearly $\Gamma(x) \setminus \{x\}$ creates the sink $\{x\}$, it is thus a cutset with cardinality $d(\Gamma) - 1$. A similar remark holds for $\Gamma^-(x) \setminus \{x\}$. A graph is said to be superconnected if it has no other cutsets with cardinality $d(\Gamma) - 1$. It follows easily from the lemmas proved in section 3 that a graph is vosperian if and only if all its fragments are trivial, where a trivial fragment is either $\{x\}$ or $V \setminus \Gamma^-(x)$, for some $x \in V$.

Most of the models are Cayley graphs on cyclic groups, called usually *loop networks.* Several attempts were made to characterise superconnected loop networks, cf. [16] and the references mentioned there. A first solution to this problem, based on Kempermann critical pair theory, is contained in [16]. There is also a characterisation of vosperian abelian Cayley graphs in [16]. Easier characterisations, based on the properties of superatoms, are obtained later in [14].

Proposition 6.3 has the following implication.

Corollary 8.1. — *Let G be a finite group and let B be a Cauchy subset of G such that $(|G|, |B|) = 1$. Assume that B is neither a left coprogression nor a right coprogression. Then $\Lambda(G, B)$ is superconnected.* □

We conclude this section by explaining the characterisation of vosperian graphs in network reliability. This characterisation is contained in an unpublished manuscript of the present author.

Consider a reflexive regular graph Γ. Set $d(\Gamma) = d$. The following property will be denoted by $\mathbf{P}_k$:

$$\forall\, A, B \subset V,\ |A| = |B| = k,\ \exists\, k \text{ disjoint paths from } A \text{ into } B.$$

Clearly $\mathbf{P}_d$ can not hold, since every d paths starting from $\Gamma(x)$ contains two intersecting paths. Clearly the path starting in x must use an other vertex of $\Gamma(x)$. It is an easy consequence of Menger Lemma that Γ satisfies $\mathbf{P}_{d-1}$ if and only if Γ is a Cauchy graph. The Vosper property is in some sense the critical situation of this problem. In particular we have the following characterisation.

Γ is vosperian if and only if for all $A \notin \{\Gamma(x) : x \in V\}$, $B \notin \{\Gamma^-(x) : x \in V\}$, with $|A| = |B| = d$, there exist d disjoint paths from A into B.

References

[1] Brailowski L.V. and Freiman G. A., *On a product of finite subsets in a torsion free group*, J. Algebra, **130**, 1990, 462–476.

[2] Cauchy A., *Recherches sur les nombres*, J. Ecole Polytechnique, **9**, 1813, 99–116.

[3] Cherly J. and Deshouillers J-M., *Un théorème d'addition dans $F_q[X]$*, J. Number Theory, **34**, 128–131.

[4] Chowla S., Mann H.B. and Strauss L.G., *Some applications of the Cauchy-Davenport theorem*, Norske Vid. Selsk. Forh. (Trondheim), **32**, 1959, 74–80.

[5] Davenport H., *On the addition of residue classes*, J. London Math. Soc., **10**, 1935, 30–32.

[6] Davenport H. and Lewis D.J., *Notes on congruences (III)*, Quart. Math. Oxford, **(2) 17**, 1966, 339–344.

[7] Diderrich G.T., *On Kneser's addition theorem in groups*, Proc. Amer. Math. Soc., 1973, 443–451.

[8] Halberstam H. and Roth K.F., *sequences*, Springer-Verlag, 1982.

[9] Hamidoune Y.O., *Sur les atomes d'un graphe orienté*, C.R. Acad. Sc. Paris A, **284**, 1977, 1253–1256.

[10] Hamidoune Y.O., *Quelques problèmes de connexité dans les graphes orientés*, J. Comb. Theory B, **30**, 1981, 1–10.

[11] Hamidoune Y.O., *An application of connectivity Theory in graphs to factorizations of elements in groups*, Europ. J. Combinatorics, **2**, 1981, 349–355.

[12] Hamidoune Y.O., *On the connectivity of Cayley digraphs*, Europ. J. Combinatorics, **5**, 1984, 309–312.

[13] Hamidoune Y.O., *On a subgroup contained in words with a bounded length*, Discrete Math., **103**, 1992, 171–176.

[14] Hamidoune Y.O., *On subsets with a small sum in abelian groups*, Europ. J. of Combinatorics, **18**, 1997, 541–566.

[15] Hamidoune Y.O. and Rödseth Ö.J., *On bases for σ-finite groups*, Math. Scand., 1994, 246–254.

[16] Hamidoune Y.O., Llàdo A.S. and Serra O., *Vosperian and superconnected abelian Cayley digraphs*, Graphs and Combinatorics, **7**, 1991, 143–152.

[17] Jia X.B. and Nathanson M.B., *Additions theorems for σ-finite groups*, In Proc. Rademacher Centenary conference, contemporary mathematics, Amer math. Soc. 1194.

[18] Jung H.A., *Über den Zusammenhang von Graphen, mit Anwendungen auf symmetricher Graphen*, Math. Ann., **202**, 1973, 307–320.

[19] Kempermann J.H.B., *On complexes in a semigroup*, Indag. Math., **18**, 1956, 247–254.

[20] Kempermann J.H.B., *On small sumsets in abelian groups*, Acta Math., **103**, 1960, 66–88.

[21] Kneser M., *Anwendung eines satzes von Mann auf die Geometrie von Zahlen*, Proc. Int. Cong. Math. Amsterdam, **2**, 1954, 32.

[22] Kneser M., *Eine Satz über abelesche gruppen mit Anwendungen auf die Geometrie der Zahlen*, Math. Z., 1955, 429–434.

[23] Mader W., *Eine Eigenschaft der Atome endlicher Graphen*, Arch. Math., **22**, 1971, 333–336.

[24] Mann H.B., *An addition theorem for sets of elements of an abelian group*, Proc. Amer. Math Soc, **4**, 1953, 423.

[25] Mann H.B., *Addition theorems: The addition theorems of group theory and number theory*, Interscience, New York, 1965.

[26] Nathanson M.B., *Additive number theory: Inverse problems and the Geometry of sumsets*, Springer-Verlag, 1994, to appear.

[27] Olson J.E., *On the sum of of two sets in a group*, J. Number Theory, **18**, 1984, 110–120.

[28] Olson J.E., *On the symmetric difference of two sets in a group*, Europ. J. Combinatorics, 1986, 43–54.

[29] Rödseth Ö.J., *Two remarks on linear forms in non-negative integers*, Math. Scand., **51**, 1982, 193–198.

[30] Tietäväinen A., *On diagonal forms over finite fields*, Ann. Univ. Turku Ser. A, 1968, 1–10.

[31] Vosper G., *The critical pairs of subsets of a group of prime order*, J. London Math. Soc., **31**, 1956, 200–205.

[32] Vosper G., *Addendum to "The critical pairs of subsets of a group of prime order"*, J. London Math. Soc., **31**, 1956, 280–282.

[33] Zemor G., *A generalisation to noncommutative groups of a theorem of Mann*, Discrete Math., **126**, 1994, 365–372.

Y.O. Hamidoune, Équipe de Combinatoire, Case 189, UFR 921, Université P. et M. Curie, Place Jussieu, 75230 Paris, France • *E-mail :* `yha@ccr.jussieu.fr`

Astérisque

MARCEL HERZOG

New results on subset multiplication in groups

Astérisque, tome 258 (1999), p. 309-315

<http://www.numdam.org/item?id=AST_1999__258__309_0>

Astérisque **258**, 1999, p. 309–315

NEW RESULTS ON SUBSET MULTIPLICATION IN GROUPS

by

Marcel Herzog

Abstract. — This paper presents results and open problems related to the following topics: group with deficient multiplication sub-tables, product bases in finite groups.

In this paper, I would like to discuss several topics which deal with subset multiplication in groups. The topics are:

(1) Deficient squares groups;
(2) Squaring bounds in groups;
(3) Deficient products in groups;
(4) Product bases in finite groups.

The paper will be concluded by a list of some related open problems.

The letter G will always denote a group and the center of G will be denoted by $Z(G)$.

1. Deficient squares groups

Let m be an integer and let M be an *m-subset* of G, i.e. $M \subseteq G$ and $|M| = m$. We say that M has the *deficient square property* if

$$|M^2| := |\{xy|x, y \in M\}| < |M|^2 = m^2 \ . \tag{1}$$

A group G has the *deficient squares property for m* ($G \in DS(m)$ in short) if (1) holds for all m-subsets M of G. A group G has the *deficient squares property* ($G \in DS$ in short) if $G \in DS(m)$ for some integer m. If G is a finite group, then of course $G \in DS$.

The first mathematician to consider the $DS(m)$ property was Gregory Freiman, who classified in [8] the $DS(2)$-groups and who collaborated with others in the classification of the $DS(3)$-groups (see [2] and [19]). It was Peter Neumann who raised the problem of classifying the DS-groups. During his visit to Australia in 1989 Peter Neumann proved that DS-groups belong to the family of finite-by-abelian-by-finite

1991 ***Mathematics Subject Classification.*** — 20F99, 20E34.

Key words and phrases. — Deficient squares groups, squaring bounds in groups, deficient products in groups, product bases in finite groups.

groups [22]. In a recent paper, Patrizia Longobardi, Mercede Maj and myself completely characterized the DS-groups. We proved

Theorem 1.1 **(cf. [9]).** — *A group $G \in DS$ if and only if either G is nearly-dihedral or $|G^{(2)}|$ is finite.*

Here a group G is called *nearly-dihedral* if it contains an abelian normal subgroup H of finite index, such that each element of G acts on H by conjugation either as the identity automorphism or as the inverting automorphism. By $G^{(2)}$ we mean $\langle g^2 | g \in G \rangle$. Instead of requiring $|G^{(2)}|$ to be finite, we could have required the finiteness of $|\{g^2 | g \in G\}|$. Our proof relies on the above mentioned result of Peter Neumann, the proof of which was included in our paper by his permission.

A group G is called *central-by-finite* or an *FIZ-group* if the center of G is of finite index in G. Clearly $G \in FIZ$ implies that G is a nearly-dihedral group and it follows by Theorem 1.1 that DS-groups are a generalization of FIZ-groups. In 1976, B.H.Neumann proved the following beautiful theorem:

Theorem 1.2 **(cf. [21]).** — *The group $G \in FIZ$ if and only if G does not contain an infinite independent subset.*

A subset M of G is called *independent* if $xy = yx$ for $x, y \in M$ implies $x = y$. If $G \in FIZ$, say $|G : Z(G)| = n$, then clearly the size of an independent subset of G is bounded by n. The difficulty in Theorem 1.2 lies in proving the other direction of the theorem.

Recently, Carlo Scoppola and myself characterized the DS-groups in the spirit of the B.H.Neumann's result. Call a subset M of G *fully-independent* if $uv = yz$ for $u, v, y, z \in M$ implies $u = y$ and $v = z$. We proved

Theorem 1.3 **(cf. [11]).** — *The group $G \in DS$ if and only if G does not contain an infinite fully-independent subset.*

Again, one direction of the theorem is trivial, since the existence of an infinite fully-independent subset in G clearly implies that $G \notin DS$. In our proof of the opposite direction, the following result of Babai-Sós [1,Proposition 8.1] was very useful:

Theorem 1.4 **(cf. [1]).** — *If U is an infinite subset of the group G, then U contains an infinite subset V such that: if $u, v, y, z \in V$ and $|\{u, v, y, z\}| \geq 3$, then $uv \neq yz$.*

The only non-trivial relations allowed in V by Theorem 1.4 are $xy = yx$ and $x^2 = y^2$. Thus, if $G \notin DS$, in order to construct an infinite fully-independent subset of G it suffices to construct an infinite subset U of G satisfying: $xy \neq yx$ and $x^2 \neq y^2$ for $x, y \in U$, $x \neq y$. By Theorem 1.4 U contains an infinite fully-independent subset of G.

2. Squaring bounds in groups

Of course, we can require from G more than the DS-property, i.e. not only $|M^2| < |M|^2$ for all m-subsets, but some stronger inequality. Such questions were considered by Leonid Brailovsky in his Ph.D. thesis, written under the supervision of G. Freiman and myself. L. Brailovsky proved, among other results, the following

***Theorem 2.1* (cf. [6]).** — *The group $G \in FIZ$ if and only if there exists a positive integer k, such that*

$$|K^2| \leq k^2 - k$$

for each k-subset K of G.

I want to prove one direction of Theorem 2.1. The other direction is easy too, but a bit more technical.

I'll prove: If k is an integer and $G \notin FIZ$ then $|K^2| > k^2 - k$ for some k-subset K of G.

By Theorem 1.2, there exists an infinite independent subset U of G and by Theorem 1.4, U contains an infinite subset V such that $uv \neq yz$ for $u, v, y, z \in V$ with $|\{u, v, y, z\}| \geq 3$. Thus, if K is a k-subset of V, then the only non-trivial equalities among the elements of K^2 are of the type $x^2 = y^2$, thus yielding

$$|K^2| \geq k^2 - (k-1) > k^2 - k\ .$$

The proof is complete.

Suppose now that G is an abelian group. Then clearly

$$(2) \qquad |K^2| \leq \frac{1}{2}k(k+1) \quad \text{for } k\text{-subsets } K \text{ of } G.$$

Does this property characterize the abelian groups? Generally speaking, the answer is NO. For $k = 1$, the inequality (2) always holds and for $k = 2$, the groups $G = Q_8 \times E$ satisfy (2), where Q_8 is the quaternion group of order 8 and E denotes an elementary abelian 2-group, finite or infinite. Moreover, if G is finite and $\frac{1}{2}k(k+1) \geq |G|$, then again (2) is trivially satisfied. But for the majority of cases, the answer is YES. More precisely, Leonid Brailovsky proved in his thesis

***Theorem 2.2* (cf. [4]).** — *If $k > 2$ is an integer and G is an infinite group, then (2) implies that G is abelian. In the finite case the same is true provided that $k^3 - k < \frac{1}{2}|G|$.*

Theorem 2.2 also holds if the bound $\frac{1}{2}k(k+1)$ in (2) is increased to $\frac{1}{2}k(k+1) + \frac{1}{2}(k-3)$, but then in the finite case we must require that $(k^2-3)(k-1) < \frac{1}{15}|G|$ (see [5]).

In the infinite case much more can be proved. We define the integral valued function of an integral variable

$$f(n) = \left\lceil \frac{5n^2 - 3n - 2}{6} \right\rceil$$

where $\lceil x \rceil$ for a real x denotes the smallest integer m such that $x \leq m$. In his thesis, L.Brailovsky proved:

***Theorem 2.3* (cf. [6]).** — *Let $k \geq 2$ be an integer. Then:*

1 : *If $|K^2| \leq f(k)$ for all k-subsets K of an infinite group G, then G is abelian.*

2 : *There exists a non-abelian infinite group G such that $|K^2| \leq f(k)+1$ for all k-subsets K of G.*

So $f(n)$ is the best possible squaring bound for infinite abelian groups. Moreover, there is a gap between $\frac{1}{2}k(k+1)$ and $\lceil\frac{5k^2-3k-2}{6}\rceil$. Each infinite abelian group satisfies $|K^2| \leq \frac{1}{2}k(k+1)$ for all k-subsets, whereas for infinite non-abelian groups the bound for $|K^2|$ on all k-subsets is larger than $\lceil\frac{5k^2-3k-2}{6}\rceil$.

3. Deficient products in groups

Let n be a positive integer. We say that G has the *deficient products property for n* ($G \in DP(n)$ in short) if for all couples of n-sets X and Y in G the following inequality holds:

$$|XY \cup YX| < 2n^2 \ . \tag{3}$$

More generally, if k is an integer with $k \geq 2$, we say that $G \in DP(n,k)$ if all k-tuples $X_1, X_2, \ldots, X_k$ of n-sets in G satisfy

$$UP(X_1,\ldots,X_k) =_{def} |\cup\{X_iX_j | 1 \leq i,j \leq k,\ i \neq j\}| < (k^2-k)n^2 \ . \tag{4}$$

Thus $DP(n) = DP(n,2)$. Finally, we say that $G \in DP$ if $G \in DP(n,k)$ for some positive integers n, k, $k \geq 2$.

In a recent paper, Federico Menegazzo from Padova and myself proved the following results concerning groups satisfying the various conditions which were introduced above.

Theorem 3.1 **(cf. [10]).** — *Let G be an infinite group. Then $G \in DP(n)$ if and only if G is abelian.*

This theorem follows easily from the following characterization of infinite non-abelian groups. First a definition: two subsets A and B of G are *product-independent* if whenever $a, a' \in A$ and $b, b' \in B$, then $ab \neq b'a'$ and $ab = a'b'$ or $ba = b'a'$ only if $a = a'$ and $b = b'$.

Theorem 3.2 **(cf. [10]).** — *Let G be an infinite group. Then G is non-abelian if and only if it contains two infinite product-independent subsets.*

Theorem 3.1 generalizes Theorem B of [17]. We proved also the following characterization of FIZ-groups.

Theorem 3.3 **(cf. [10]).** — *Let G be an infinite group. Then G contains $\aleph_0$ mutually product-independent infinite subsets if and only if $G \notin FIZ$.*

The characterization of infinite DP-groups is an easy consequence of Theorem 3.3.

Theorem 3.4 **(cf. [10]).** — *Let G be an infinite group. Then $G \in DP$ if and only if $G \in FIZ$.*

Consider now related but different conditions. Let $(n) = (n_1, n_2, \dots)$ be an infinite sequence of positive integers. We say that $G \in P^*_{(n)}$ ($G \in P^{**}_{(n)}$) if every infinite sequence $X_1, X_2, \dots$ of distinct subsets of G of sizes $|X_i| = n_i$ for all i contains a pair X, Y of distinct members satisfying $XY = YX$ ($|XY \cup YX| < 2|X||Y|$). If $n_i = n$ for all i write P^*_n for $P^*_{(n)}$. Theorem 1.2 states that $G \in P^*_1$ if and only if $G \in FIZ$. In [20] F. Menegazzo proved that an infinite group G satisfies P^*_n for $n \geq 2$ if and only if G is abelian. In [10] we proved:

Theorem 3.5. — *Let G be an infinite group. Then $G \in P^*_{(n)}$ with $n_i \geq 2$ for all i if and only if G is abelian.*

It is easy to see that Theorem 3.3 implies:

Theorem 3.6. — *Let G be an infinite group. Then $G \in P^{**}_{(n)}$ if and only if $G \in FIZ$.*

4. Product bases in finite groups

A subset A of a finite group G is called a *basis (2-basis)* of G if $A^2 =_{def} \{ab | a, b \in A\} = G$. The problem of finding bases for G of size $c|G|^{\frac{1}{2}}$ for families of finite groups, where c denotes a fixed real number, was first posed by H. Rohrbach in 1937 in [23]. Such bases were found for certain families by Rohrbach himself [23], by Bertram and Herzog [3] and by Jia [12,13]. Recently, two graduate students in the Tel-Aviv University Gadi Kozma and Arie Lev proved that such bases exist for the family of all finite groups. They proved:

***Theorem 4.1* (cf. [15]).** — *If G is a finite group then there exists $A \subset G$ such that $A^2 = G$ and $|A| \leq \frac{4}{\sqrt{3}}|G|^{\frac{1}{2}} \approx 2.3094|G|^{\frac{1}{2}}$.*

The proof of Theorem 4.1 was based on the following strengthening of the Brauer-Fowler theorem:

***Theorem 4.2* (cf. [18]).** — *If G is a finite group of a non-prime order then there exists a proper subgroup H of G with $|H| \geq |G|^{\frac{1}{2}}$.*

Brauer and Fowler proved only that $|H| \geq |G|^{\frac{1}{3}}$ for groups G of even order. However, the proof of Theorem 4.2 uses the classification of the finite simple groups. We were recently informed that results similar to Theorems 4.1 and 4.2 appeared in the computer-science oriented papers [14] and [7].

Finally, if h is a positive integer, a subset A of a finite group is called an *h-basis* if $A^h = G$. Kozma and Lev proved the following theorem about h-bases in finite solvable groups:

***Theorem 4.3* (cf. [16]).** — *Let G be a finite solvable group. Then G contains an h-basis A such that $|A| \leq (2h-1)|G|^{\frac{1}{h}}$.*

A similar theorem is probably true for all finite groups.

5. Some open problems

I am going to list now some open problems, which are related to the results mentioned in this lecture.

1. Let $G \in DS(m)$. Prove that there exists an integer $N = f(m)$ such that either $|G^{(2)}| \leq N$ or G is nearly-dihedral with $|G : H| \leq N$ (see Theorem 1.1).

2. Does there exist a purely graph-theoretical proof of Theorem 1.3? In other words, can one prove directly that the existence of fully-independent subsets of G of size m for all integers m implies the existence of an infinite fully-independent subset of G?

3. Let n and m denote positive integers. A group G has the *deficient n-powers property for m* ($G \in DNP(m)$ in short) if $|M^n| < |M|^n$ for all m-subsets M of G. A group $G \in DNP$ if $G \in DNP(m)$ for some integer m. A subset M of G is *n-fully-independent* if $x_1x_2 \ldots x_n = y_1y_2 \ldots y_n$ for $x_i, y_j \in M$ implies $x_i = y_i$ for $i = 1, 2, \ldots, n$. Is it true that: $G \in DNP$ if and only if G does not contain an infinite n-fully-independent subset? (see Theorem 1.3)

4. Does there exist a constant c such that if $k > 2$ is an integer and G is a finite group satisfying condition (2), then G is abelian, provided that $k^2 < c|G|$? (see Theorem 2.2)

5. Does there exist a constant c such that if G is a finite group and $|G| \neq p, p^2, pq$, where p and q are arbitrary distinct primes, then there exists a proper subgroup H of G satisfying $|H| \geq c|G|^{\frac{2}{3}}$? (see Theorem 4.2)

6. Prove: Let h be a positive integer. Then there exists a function $f(x)$ such that if G is a finite group, then G has an h-basis A satisfying $|A| \leq f(h)|G|^{\frac{1}{h}}$? (see Theorem 4.3)

References

[1] Babai L. and Sós V.T., *Sidon sets in groups and induced subgraphs of Cayley graphs*, Europ. J. Combinatorics, **6**, 1985, 101-114

[2] Berkovich Ja.G., Freiman G.A. and Praeger C.E., *Small squaring and cubing properties for finite groups*, Bull. Austral. Math. Soc., **44**, 1991, 429-450

[3] Bertram E.A. and Herzog M., *On medium-size subgroups and bases of finite groups*, J. of Combin. Theory, Series A, **57**, 1991, 1-14

[4] Brailovsky L., *A characterization of abelian groups*, Proc. Amer. Math. Soc., **117**, 1993, no. 3, 627–629.

[5] Brailovsky L., *On the small squaring and commutativity*, Bull. London Math. Soc., **25**, 1993, 330–336.

[6] Brailovsky L., *Combinatorial conditions forcing commutativity of an infinite group*, J. of Algebra, **165**, 1994, no. 2, 394–400.

[7] Finkelstein L., Kleitman D. and Leighton T., *Applying the classification theorem for finite simple groups to minimize pin count in uniform permutation architectures*, VLSI algorithms and architectures, J.H. Reif(ed), Springer, 1989, 247-256

[8] Freiman G.A., *On two and three-element subsets of groups* Aeq. Math.,**22**, 1981, 140-152

[9] Herzog M., Longobardi P. and Maj M., *On a combinatorial problem in group theory*, Israel J. Math., **82**, 1993, no. 1-3, 329–340.

[10] Herzog M. and Menegazzo F., *On deficient products in infinite groups*, Rend. Sem. Mat. Univ. Padova, **93**, 1995, 1–6.

[11] Herzog M. and Scoppola C.M., *On deficient squares groups and fully-independent subsets* Bull. London Math. Soc., **27**, 1995, no. 1, 65–70.

[12] Jia X-D., *Thin bases for abelian groups*, J. Number Theory, **36**, 1990, 254-256

[13] Jia X-D., *Thin bases for finite nilpotent groups*, J. Number Theory, **41**, 1992, 303-313

[14] Kilian J., Kipnis S. and Leierson Ch.E., *The organization of permutation architectures with bussed interconnections*, Proc. 1987 IEEE Conf. On The Foundation Of Comp. Science, IEEE, 1987, 305-315

[15] Kozma G. and Lev A., *Bases and decomposition numbers of finite groups*, Arch. Math., **58**, 1992, 417-424

[16] Kozma G. and Lev A., *On h-bases and h-decompositions of the finite solvable and alternating groups*, J. Number Theory, **49**, 1994, no. 3, 385–391

[17] Lennox J.C., Hassanabadi A.M. and Wiegold J., *Some commutativity criteria*, Rend. Sem. Mat. Univ. Padova, **84**, 1990, 135-141

[18] Lev A., *On large subgroups of finite groups*, J. of Algebra, **152**, 1992, 434-438

[19] Longobardi P. and Maj M., *The classification of groups with the small-squaring property on 3-sets*, Bull. Austral Math. Soc., **46**, 1992, 263-269

[20] Menegazzo F., *A property equivalent to commutativity for infinite groups* Rend. Sem. Mat. Univ. Padova, **87**, 1992, 299-301

[21] Neumann B.H., *A problem of Paul Erdős on groups*, J. Austral. Math. Soc., **21**, 1976, 467-472, *Selected works of B.H.Neumann and Hanna Neumann*, **5**, 1003-1008

[22] Neumann P.M., *A combinatorial problem in group theory*, Private communication.

[23] Rohrbach H., *Anwendung eines Satzes der additiven Zahlentheorie auf eine gruppentheoretishe Frage*, Math. Z., **42**, 1937, 538-542

M. Herzog, School of Mathematical Sciences, Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv, Israel

Astérisque

LEV F. VSEVOLOD

On small sumsets in abelian groups

Astérisque, tome 258 (1999), p. 317-321

<http://www.numdam.org/item?id=AST_1999__258__317_0>

Astérisque **258**, 1999, p. 317–321

ON SMALL SUMSETS IN ABELIAN GROUPS

by

Vsevolod F. Lev

Abstract. — In this paper we investigate the structure of those pairs of finite subsets of an abelian group whose sums have relatively few elements: $|A+B| < |A| + |B|$. In 1960, J. H. B. Kemperman gave an exhaustive but rather sophisticated description of recursive nature. Using intermediate results of Kemperman, we obtain below a description of another type. Though not (generally speaking) sufficient, our description is intuitive and transparent and can be easily used in applications.

1. Introduction

By G we denote an abelian group. A finite non-empty subset $S \subseteq G$ is said to be *an arithmetic progression with difference d* if S is of the form

$$S = \{a + id \colon i = 1, \dots, |S|\} \quad (a, d \in G).$$

If, in addition, the order of the group element d satisfies ord $d \geq |S| + 2$, then we say that S is a *true* arithmetic progression.

Let A and B be finite subsets of G. We write

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

and consider the following condition:

$$|A + B| \leq |A| + |B| - 1. \qquad (*)$$

The aim of this paper is to prove the following

Main Theorem. — *Let A and B satisfy $(*)$, and suppose that* $\max\{|A|, |B|\} > 1$. *Then there exist a finite subgroup $H \subseteq G$ and two finite subsets $S_1, S_2 \subseteq G$ such that $A \subseteq S_1 + H$, $B \subseteq S_2 + H$, and one of the following holds:*

i) $|S_1| = |S_2| = 1$, *and* $|A + B| \geq \frac{1}{2}|H| + 1$*;*

ii) $|S_1| = 1$, $|S_2| > 1$, *and* $|A + B| \geq (|S_2| - 1)|H| + 1$*;*

iii) $|S_1| > 1$, $|S_2| = 1$, *and* $|A + B| \geq (|S_1| - 1)|H| + 1$*;*

1991 ***Mathematics Subject Classification.*** — 11P99, 11B75.
Key words and phrases. — Sumsets, small doubling.

iv) $\min\{|S_1|, |S_2|\} > 1$, *and* $|A+B| \geq (|S_1|+|S_2|-2)|H|+1$; *moreover,* S_1 *and* S_2 *are true arithmetic progressions with common difference* d *of order at least* $\operatorname{ord} d \geq |S_1|+|S_2|+1$.

It can be easily verified that the conclusion of Main Theorem implies

$$|A+B+H| - |A+B| \leq |H|-1$$

in cases ii)–iv), and

$$|A+B+H| - |A+B| \leq \frac{1}{2}|H|-1$$

in case i): just observe that

$$|A+B+H| \leq |S_1+S_2+H| \leq |S_1+S_2||H|.$$

Thus, $A+B$ "almost" fills in a system of H-cosets, while both $(A+H)/H$ and $(B+H)/H$ are in arithmetic progressions — unless some of them consists of just one element.

The Main Theorem will be proved in Section 3. Now, we give two definitions.

We say that the subgroup $H \subseteq G$, $|H| \geq 2$ is *a period* of the finite subset $C \subseteq G$ if C is a union of one or more H-cosets, that is if $C+H=C$. In this case C is called *periodic* and we write $H = P(C)$.

We say that the subgroup $H \subseteq G$, $|H| \geq 2$ is a *quasi-period* of the finite subset $C \subseteq G$, if C is a union of one or more H-cosets and possibly a subset of yet another H-coset. In this case C is called *quasi-periodic* and we write $H = Q(C)$.

If $H = P(C)$, we also say that H is a *true* period of C, as opposed to $H = Q(C)$, when C is a *quasi*-period. Obviously, if $H = P(C)$ or $H = Q(C)$ then $|H| < \infty$. Notice that according to the above definitions each periodic set is also quasi-periodic.

2. Auxiliary results

The following deep result due to Kemperman (see [**1**]) plays the central role in our proof.

***Theorem 1* (Kemperman).** — *Let* A *and* B *be finite subsets of* G *such that* $(*)$ *holds and* $\min\{|A|, |B|\} > 1$. *Then either* $A+B$ *is an arithmetic progression or* $A+B$ *is quasi-periodic.*

Corollary 1. — *Under the assumptions of Theorem 1, one of the following holds:*

i) $A+B$ *is in true arithmetic progression;*

ii) $A+B = c+H \setminus \{0\}$ *where* $H \subseteq G$ *is a subgroup, and* $c \in G$ *— an element of* G*;*

iii) $A+B$ *is quasi-periodic.*

The next lemma also originates in [**1**].

***Lemma 1* (Kemperman).** — *Suppose that* $(*)$ *holds and that* $A+B$ *is in true arithmetic progression of difference* d. *Then also* A *and* B *are in true arithmetic progressions with the same difference* d. *Moreover, in* $(*)$ *equality holds, and therefore* $\operatorname{ord} d \geq |A|+|B|+1$.

We need three more lemmas.

***Lemma* 2.** — *Let A and B be finite non-empty subsets of G, and let $H \subseteq G$ be a finite non-zero subgroup of G, satisfying*

$$(|A+H|-|A|)+(|B+H|-|B|)<|H|.$$

Then $H = P(A+B)$.

Proof. — We choose $c = a+b \in A+B$ and $h \in H$ and we prove that $c+h \in A+B$. We have:

$$|(a+H)\cap\overline{A}|+|(b+H)\cap\overline{B}| \leq |(A+H)\cap\overline{A}|+|(B+H)\cap\overline{B}|<|H|,$$

hence

$$|(a+H)\cap A|+|(b+H)\cap B|>|H|,$$
$$|H\cap(A-a)|+|h-H\cap(B-b)|>|H|,$$

and therefore there exist $h_a, h_b \in H$ such that

$$h_a = h-h_b,\ h_a = a'-a,\ h_b = b'-b \quad (a' \in A,\ b' \in B).$$

But then $c+h = a+b+h_a+h_b = a'+b' \in A+B$ which was to be proved. □

***Lemma* 3.** — *Let $A, B \subseteq G$ satisfy $(*)$. Suppose that $A+B$ is quasi-periodic, and write $H = Q(A+B)$. Denote by σ the canonical homomorphism $\sigma\colon G \to G/H$, and set $A_1 = \sigma A$, $B_1 = \sigma B$. Then*

i) $|A_1+B_1| \leq |A_1|+|B_1|-1$;
ii) $|A_1+B_1| < |A+B|$;
iii) $|A+B|-1 \geq (|A_1+B_1|-1)|H|$.

Proof. — i) Suppose first that $H = P(A+B)$. Obviously, $|A+B| \leq |A+H|+|B+H|-1$. But the left-hand side, as well as $|A+H|$ and $|B+H|$, divides by $|H|$, so we also have $|A+B| \leq |A+H|+|B+H|-|H|$. Eventually, $|A+H| = |A_1||H|$, $|B+H| = |B_1||H|$ and $|A+B| = |A_1+B_1||H|$.

Now consider the situation, when H is a quasi-period, but not a *true* period of $A+B$. Then by Lemma 2,

$$|A+B|+1 \leq |A|+|B| \leq |A+H|+|B+H|-|H|,$$

hence (since the right-hand side divides by $|H|$) we also have $|A+B+H| \leq |A+H|+|B+H|-|H|$, and the proof finishes as in the case $H = P(A+B)$.

ii) Follows from iii).

iii) If $H = P(A+B)$, then

$$|A+B|-1 = |A_1+B_1||H|-1 > (|A_1+B_1|-1)|H|.$$

If H is not a true period of $A+B$, then $A+B$ contains $|A_1+B_1|-1$ full H-cosets, and at least one element in yet another H-coset, therefore $|A+B| \geq (|A_1+B_1|-1)|H|+1$.

□

Lemma 4. — *Let $A+B=c+H\setminus\{0\}$ and suppose that* $\min\{|A|,|B|\}\geq 2$, *where $A,B\subseteq G$ are subsets, $H\subseteq G$ a subgroup, and $c\in G$ an element of G. Then $|H|\geq 4$.*

Proof. — We have: $|H|-1=|A+B|\geq|A|\geq 2$, hence $|H|\geq 3$. Suppose $|H|=3$, and so $|A|=|B|=|A+B|=2$. Let $A=a+\{0,d_1\}$, $B=b+\{0,d_2\}$. Then $A+B=a+b+\{0,d_1,d_2,d_1+d_2\}$, hence $d_2=d_1$, $d_1+d_2=0$, and $H=\{0\}\cup\{a+b-c,a+b+d-c\}$, where $d=d_1=d_2$, $2d=0$. Therefore $d=(a+b+d-c)-(a+b-c)\in H$, which contradicts to $|H|=3$, $2d=0$. □

3. Proof of the Main Theorem

Denote $G_0=G$, $A_0=A$, $B_0=B$ and consider the following conditions:

1) $|A|=|B|=1$;
2) $|A|=1$, $|B|>1$;
3) $|A|>1$, $|B|=1$;
4) $A+B=c+\widetilde{H}\setminus\{0\}$, where $\widetilde{H}$ is a subgroup, and $c\in G$ — an element of G;
5) $A+B$ is in true arithmetic progression.

If all these conditions fail, then by Corollary 1 the sum A_0+B_0 is quasi-periodic, and we put $H_1=Q(A_0+B_0)$, $G_1=G_0/H_1$, denote by σ_1 the canonical homomorphism $\sigma_1\colon G_0\to G_1$ and set $A_1=\sigma_1A_0$, $B_1=\sigma_1B_0$, so that A_1,B_1 satisfy $(*)$ by Lemma 3, i). Now check, whether some of the conditions 1)–5) is met with G_1,A_1,B_1 substituted for G,A,B. If not, we continue the process by defining

$$H_2=Q(A_1+B_1),\ G_2=G_1/H_2,$$
$$\sigma_2\colon G_1\to G_2,\ A_2=\sigma_2A_1,\ B_2=\sigma_2B_1$$

and so on. At each step we obtain a pair of subsets $A_i,B_i\subseteq G_i$, satisfying $(*)$ and also $|A_i+B_i|<|A_{i-1}+B_{i-1}|$ (by Lemma 3, ii)). Eventually we obtain a pair $A_k,B_k\subseteq G_k$ ($k\geq 0$), which meets at least one of the conditions 1)–5). We write $\sigma=\sigma_k\cdots\sigma_1\colon G\to G_k$ (or $\sigma=\mathrm{id}_G$ in the case $k=0$) so that $A_k=\sigma A$, $B_k=\sigma B$, and we write $H=\sigma^{-1}\widetilde{H}$ if the first condition met is 4), or $H=\ker\sigma$ otherwise. We distinguish 5 cases according to the first condition satisfied.

1) Here $k>0$ and $A_{k-1}+B_{k-1}=c+H_k$, where $c\in G_{k-1}$ (since H_k is a quasi-period of $A_{k-1}+B_{k-1}$), therefore $A_{k-1}\subseteq a+H_k$, $B_{k-1}\subseteq b+H_k$ ($a,b\in G_{k-1}$), whence $A\subseteq a'+H$, $B\subseteq b'+H$ ($a',b'\in G$). We choose now $S_1=\{a'\}$, $S_2=\{b'\}$ and observe, that by Lemma 3, iii)

$$\begin{aligned}|A+B|-1 &\geq (|A_1+B_1|-1)|H_1|\geq\cdots\geq\\ &\geq (|A_{k-1}+B_{k-1}|-1)|H_{k-1}|\cdots|H_1|=\\ &= (|H_k|-1)|H_{k-1}|\cdots|H_1|\geq\\ &\geq \frac{1}{2}|H_k||H_{k-1}|\cdots|H_1|=\frac{1}{2}|H|.\end{aligned}$$

2) Also here we may assume $k>0$, since otherwise the result is trivial if we choose $S_1=A$, $S_2=B$, $H=\{0\}$. Furthermore, as in 1) we have $A\subseteq a+H$. We choose $S_1=\{a\}$, and for S_2 we choose the system of arbitrary representatives of all

H-cosets, containing at least one element of B, so that $A \subseteq S_1 + H,\ B \subseteq S_2 + H$ and $|S_2| = |B_k|$. Then

$$|A+B| - 1 \geq \cdots \geq (|A_k + B_k| - 1)|H_k| \cdots |H_1| = (|S_2| - 1)|H|.$$

3) This case is analogous to the previous one in view of the symmetry between A and B.

4) In this case there exist $a, b \in G$ such that $A \subseteq a + H,\ B \subseteq b + H$ and we choose $S_1 = \{a\},\ S_2 = \{b\}$. Then

$$\begin{aligned} |A+B| - 1 &\geq \cdots \geq (|A_k + B_k| - 1)|H_k| \cdots |H_1| = \\ &= (|\widetilde{H}| - 2)|H_k| \cdots |H_1| \geq \frac{1}{2}|\widetilde{H}||H_k| \cdots |H_1| = \frac{1}{2}|H| \end{aligned}$$

(since $|\widetilde{H}| \geq 4$ by Lemma 4).

5) In this case, by Lemma 1, A_k and B_k are in true arithmetic progressions with common difference d of order ord $d \geq |A_k| + |B_k| + 1$, and $|A_k + B_k| = |A_k| + |B_k| - 1$. It is easily seen that we can choose two true arithmetic progressions $S_1, S_2 \subseteq G$ with a common difference d' in such a way, that $A_k = \sigma S_1,\ B_k = \sigma S_2$ and $|S_1| = |A_k|,\ |S_2| = |B_k|$, ord $d' \geq$ ord d. Then

$$A \subseteq S_1 + H,\ B \subseteq S_2 + H,\ \text{ord } d' \geq |S_1| + |S_2| + 1$$

and

$$|A+B| - 1 \geq \cdots \geq (|A_k + B_k| - 1)|H_k| \cdots |H_1| = (|S_1| + |S_2| - 2)|H|.$$

This completes the proof. □

References

[1] Kemperman J.H.B., *On small sumsets in an abelian group*, Acta Math., **103**, 1960, 63–88.

V.F. Lev, Inst. of Mathematics, Hebrew University, Jerusalem, Israel 91904
E-mail : seva@math.huji.ac.il • *Url :* http://www.ma.huji.ac.il/~seva/

Imre Z. Ruzsa

An analog of Freiman's theorem in groups

Astérisque, tome 258 (1999), p. 323-326

<http://www.numdam.org/item?id=AST_1999__258__323_0>

Astérisque **258**, 1999, p. 323–326

AN ANALOG OF FREIMAN'S THEOREM IN GROUPS

by

Imre Z. Ruzsa

Abstract. — It is proved that in a commutative group G, where the order of elements is bounded by an integer r, any set A having n elements and at most αn sums is contained in a subgroup of size Cn with $C = f(r,\alpha)$ depending on r and α but not on n. This is an analog of a theorem of G. Freiman which describes the structure of such sets in the group of integers.

Let A be a set of integers, $|A| = n$, and suppose that $|A + A| \le cn$. A famous theorem of Freiman [1, 2] provides a certain structural description of these sets; in one of the possible formulations, it says that A can be covered by a generalized arithmetic progression

$$\{a + q_1x_1 + q_2x_2 + \cdots + q_dx_d : 0 \le x_i \le l_i - 1\},$$

where $d < c$ and $\prod l_i \le Cn$ with C depending on c.

One can ask for a description of sets with few sums in every Abelian group. In this paper we consider groups which are in a sense very far from $\mathbb{N}$.

Theorem. — *Let $r \ge 2$ be an integer, and let G be a commutative group in which the order of every element is at most r. Let $A \subset G$ be a finite set, $|A| = n$. If there is another $B \subset G$ such that $|B| = n$ and $|A + B| \le \alpha n$ (in particular, if $|A + A| \le \alpha n$ or $|A - A| \le \alpha n$), then A is contained in a subgroup H of G such that*

$$|H| \le f(r,\alpha)n,$$

where

$$f(r,\alpha) = \alpha^2 r^{\alpha^4}.$$

1991 ***Mathematics Subject Classification.*** — 11B75, 11P99.
Key words and phrases. — Addition of sets, sumsets, inverse theorems.

This paper was finished while author was visiting DIMACS, Rutgers University. The author gratefully acknowledges support by DIMACS.

The proof goes along similar lines to my proof of Freiman's theorem [3, 4], but is considerably simpler.

For a nonnegative integer k and a set $A \subset G$ we introduce the notation

$$kA = A + \cdots + A, \quad k \text{ summands},$$
$$0A = \{0\}, \quad 1A = A.$$

Lemma. — *If $A, B \subset G$, $|B| = n$ and $|A + B| \le \alpha n$, then for arbitrary nonnegative integers k, l we have*

$$|kA - lA| \le \alpha^{k+l} n.$$

See [3], Lemma 3.3. Observe the asymmetric role of A and B. No a priori bound is assumed for $|A|$; an alternative formulation (like in the Theorem) would be "if A is such that the union of n suitable translations has at most αn elements, then A is so small that even the sets $kA - lA$ are small".

Proof of the Theorem. Let $b_1, b_2, \ldots, b_k$ be a maximal collection of elements such that $b_i \in 2A - A$ and the sets $b_i - A$ are all disjoint. We have

$$b_i - A \subset 2A - 2A,$$

hence

$$\left| \bigcup (b_i - A) \right| = kn \le |2A - 2A| \le \alpha^4 n$$

(the last inequality follows from the Lemma). This implies $k \le \alpha^4$.

Take an arbitrary $x \in 2A - A$. Since the collection $b_1, \ldots, b_k$ was maximal, there must be an i such that

$$(x - A) \cap (b_i - A) \ne \varnothing,$$

that is, $x - a_1 = b_i - a_2$ with some $a_1, a_2 \in A$, which means

$$x = b_i + a_1 - a_2 \in b_i + (A - A).$$

Hence

$$2A - A \subset \bigcup (b_i + (A - A)) = B + A - A, \tag{1}$$

where $B = \{b_1, \ldots, b_k\}$.

Now we prove

$$jA - A \subset (j-1)B + A - A \quad (j \ge 2) \tag{2}$$

by induction on j. By (1), this holds for $j = 2$. Now we have

$$\begin{aligned}(j+1)A - A &= (2A - A) + (j-1)A \\ &\subset B + A - A + (j-1)A \text{ by (1)} \\ &= B + (jA - A) \\ &\subset B + (j-1)B + A - A \\ &= jB + A - A,\end{aligned}$$

which provides the inductive step.

Let H and I be the subgroups generated by A and B, respectively. By (2) we have

$$jA - A \subset I + (A - A) \tag{3}$$

for every j. We have also

$$\bigcup (jA - A) = H, \tag{4}$$

which easily follows from the fact that the order of elements of G is bounded. Relations (3) and (4) imply that

$$H \subset I + (A - A).$$

Since I is generated by k elements of order $\le r$ each, we have

$$|I| \le r^k \le r^{\alpha^4},$$

consequently

$$|H| \le |I||A - A| \le \alpha^2 r^{\alpha^4} n$$

(the estimate for $|A - A|$ follows from the Lemma). QED

Remarks. — Take a group of the form $G = Z_r^m$, where Z_r is a cyclic group of order r, and a set $A \subset G$ of the form

$$A = (a_1 + G') \cup \cdots \cup (a_k + G')$$

with a subgroup G'. Here $|A| = n = k|G'|$, and if all the sums $a_i + a_j$ lie in different cosets of G', then

$$|A + A| = \frac{k(k+1)}{2}|G'| = \alpha n, \quad \alpha = \frac{k+1}{2}.$$

The subgroup generated by A can have as many as $r^k|G'|$ elements, hence our function

$$f(r, \alpha) = \alpha^2 r^{\alpha^4}$$

cannot be replaced by anything smaller than

$$r^k = r^{2\alpha - 1}.$$

Conjecture. — *The Theorem holds with $f(r, \alpha) = r^{C\alpha}$ with a suitable constant C.*

The following conjecture of Katalin Marton would yield a more efficient covering in a slightly different form.

Conjecture. — *If $|A| = n$, $|A + A| \le \alpha n$, then there is a subgroup H of G such that $|H| \le n$ and A is contained in the union of α^c cosets of H, where the constant c may depend on r but not on n or α.*

This also suggests that perhaps in Freiman's original problem a better result can be formulated in terms of covering by a small number of generalized arithmetical progressions than just one.

References

[1] Freiman G. A., *Foundations of a structural theory of set addition*, Translation of Math. Monographs vol. **37**, Amer. Math. Soc., Providence, R. I., USA, 1973.

[2] Freiman G. A., *What is the structure of K if $K + K$ is small?*, in: *Lecture Notes in Mathematics 1240*, Springer-Verlag, New York – Berlin, 1987, 109–134.

[3] Ruzsa I. Z., *Arithmetical progressions and the number of sums*, Periodica Math. Hung., **25**, 1992, 104–111.

[4] Ruzsa I. Z., *Generalized arithmetical progressions and sumsets*, Acta Math. Hungar., **65**, 1994, 379–388.

I. Ruzsa, Mathematical Institute, of the Hungarian Academy of Science, Budapest, Pf. 127, H-1364 Hungary • *E-mail :* ruzsa@math-inst.hu

Astérisque

GILLES COHEN

GÉRARD ZÉMOR

Subset sums and coding theory

Astérisque, tome 258 (1999), p. 327-339

<http://www.numdam.org/item?id=AST_1999__258__327_0>

Astérisque **258**, 1999, p. 327–339

SUBSET SUMS AND CODING THEORY

by

Gérard Cohen & Gilles Zémor

Abstract. — We study some additive problems in the group $(\mathbb{Z}/_{2\mathbb{Z}})^r$. Our purpose is to show how those problems are closely related to coding theory. We present some relevant classical coding techniques and make use of them to obtain some original contributions.

1. Introduction

Let G denote the group $\mathbf{F}^r$ where $\mathbf{F} = \{0, 1\}$ stands for the additive group with two elements. Let S be a generating set of G. For any positive integer i, denote by S^i the set of sums of i distinct elements of S. Set $S^0 = \{0\}$ and for any set I of non-negative integers, let $S^I = \cup_{i \in I} S^i$. Let us denote by $\rho(S)$ the smallest integer t such that any element of G can be expressed as a sum of t or less elements of S, i.e. such that

$$G = S^{[0,t]}.$$

Let us denote by $d(S)$ the smallest integer i such that 0 can be expressed as a sum of i distinct elements of S, i.e. let $d(S) - 1$ be the largest t such that

$$0 \notin S^{[1,t]}.$$

We wish to focus on the following 'additive' problems.

Problem 1. — *For given r and t, find the smallest s such that $|S| \geq s$ implies $\rho(S) \leq t$.*

Problem 2. — *For given r and t, find the largest s such that $|S| \leq s$ implies $\rho(S) \geq t$.*

Problem 3. — *For given r and d, find the smallest s such that $|S| \geq s$ implies $d(S) \leq d$.*

1991 ***Mathematics Subject Classification.*** — 94B05, 94B75, 05C25, 05C50, 11P99.
Key words and phrases. — Coding theory, additive theory.

Those three problems can be expressed as problems in coding theory. Indeed, problems 2 and 3 are classical coding problems of which we shall give a short self-contained presentation for the non specialist. Problem 1, although less known to coding theorists, is also amenable to coding techniques, and we shall present original contributions to it and also to the following generalisation of problem 3.

Problem 4. — *Given r and an arbitrary set of integers I, find the smallest s such that $|S| \geq s$ implies $0 \in S^I$.*

2. Coding-theoretic formulation of problems 1-4

What coding theorists call a (binary) *linear code* of length n is simply a subspace of the vector space $\mathbf{F}^n$. Let S be a generating set of $\mathbf{F}^r$ with $|S| = n$. There is an important linear code $C(S)$ associated to S whose coding-theoretic properties reflect the additive properties of S. To obtain it let $s_1, \ldots, s_n$ be any ordering of its elements that we shall write as column vectors. Consider the $r \times n$ matrix $\mathbf{H} = [s_1 \ldots s_n]$ and the associated function

$$\begin{array}{rcl} \sigma : \mathbf{F}^n & \to & G = \mathbf{F}^r \\ \mathbf{x} = (x_1 \ldots x_n) & \mapsto & \sigma(\mathbf{x}) = \mathbf{H}\,{}^t\mathbf{x} \end{array}$$

Define $C(S)$ to be the set of vectors $\mathbf{x}$ of $\mathbf{F}^n$ such that $\sigma(\mathbf{x}) = 0$. When defining such a code $C(S)$ associated to a set S we shall usually not specify which ordering $s_1, \ldots, s_n$ we are choosing because the properties of $C(S)$ that interest us are independent of it. To help distinguish between the two structures $G = \mathbf{F}^r$ and $\mathbf{F}^n$, we shall use plain letters to denote elements of G and bold letters to denote vectors of $\mathbf{F}^n$: furthermore, since the vector space structure of $\mathbf{F}^n$ will be used rather more heavily than that of G, we shall systematically refer to elements of $\mathbf{F}^n$ as *vectors*. $C(S)$ (or simply C when there is little ambiguity) is a subspace of $\mathbf{F}^n$ of dimension $k = n - r$. Its elements are referred to as *codewords*. $\mathbf{H}$ is called a *parity-check matrix* of C, and for any vector $\mathbf{x} \in \mathbf{F}^n$, $\sigma(\mathbf{x})$ is called the *syndrome* of $\mathbf{x}$. Two vectors $\mathbf{x} = (x_1 \ldots x_n)$ and $\mathbf{y} = (y_1 \ldots y_n)$ of $\mathbf{F}^n$ are said to be *orthogonal* if

$$\sum_{i=1}^{n} x_i y_i = 0$$

where computations are performed in $\mathbf{F}$. If C is a linear code of $\mathbf{F}^n$ of dimension k, then the set $C^\perp$ of vectors orthogonal to C is a linear code of dimension $n - k$. Any matrix $\mathbf{H}$ whose rows are independent vectors orthogonal to C make up a parity-check matrix of C.

Remark. — *Not every code C need be a code $C(S)$ for some set S. This is because not every code has a parity-check matrix with distinct columns.*

Coding theorists regard $\mathbf{F}^n$ as a metric space, i.e. endowed with the *Hamming distance* $d(\cdot,\cdot)$:

$$\begin{array}{rcl} \mathbf{F}^n \times \mathbf{F}^n & \to & [0, n] \\ (\mathbf{x}, \mathbf{y}) & \mapsto & d(\mathbf{x}, \mathbf{y}) \end{array}$$

where $d(\mathbf{x},\mathbf{y})$ is defined as the number of coordinates where $\mathbf{x}$ and $\mathbf{y}$ differ. The *minimum distance* $d(C)$ of a code C is the smallest distance between a pair of distinct codewords,

$$d(C) = \min_{\substack{\mathbf{x},\mathbf{y}\in C\\ \mathbf{x}\neq\mathbf{y}}} d(\mathbf{x},\mathbf{y}).$$

Note that $d(C)$ is also the minimum distance $d(\mathbf{x},\mathbf{0})$ between the $\mathbf{0}$ vector and any non-zero codeword $\mathbf{x}$: this is because $d(\cdot,\cdot)$ is invariant by translation and C is an additive subgroup. The integer $d(\mathbf{x},\mathbf{0})$ is called the *weight* of $\mathbf{x}$ and denoted by $w(\mathbf{x})$. The classical parameters of a linear code C are usually denoted by $[n,k,d]$ and refer respectively to its length, dimension and minimum distance.

Another classical parameter of a code C is its *covering radius* $\rho(C)$: it is the maximum distance between a vector of $\mathbf{F}^n$ and the code C, i.e.

$$\rho(C) = \max_{\mathbf{x}\in\mathbf{F}^n} d(\mathbf{x},C)$$

where $d(\mathbf{x},C) = \min_{\mathbf{c}\in C} d(\mathbf{x},\mathbf{c})$.

Given a vector $\mathbf{x} = (x_1\dots x_n)$ of $\mathbf{F}^n$, it is common to define its *support* by $supp(\mathbf{x}) = \{i, x_i = 1\}$. The syndrome of $\mathbf{x}$ can therefore be written as

$$\sigma(\mathbf{x}) = \sum_{i\in supp(\mathbf{x})} s_i$$

where the sum is computed in $\mathbf{F}^r$. It is now clear that the minimum distance of C equals the minimum cardinality of a subset I of S such that $\sum_{i\in I} s_i = 0$. In particular we have :

Remark. — *For any code C, there exists a set S not containing* 0 *such that $C = C(S)$ if and only if $d(C) \geq 3$.*

Similarly, it is readily checked that the covering radius of C is the smallest number of additions necessary to generate every non-zero element of $\mathbf{F}^r$ with elements of S. Summarizing,

Proposition 2.1. — *The correspondence $S \to C(S)$ is such that*

$$\begin{aligned} d(S) &= d(C(S)) \\ \rho(S) &= \rho(C(S)). \end{aligned}$$

The above correspondence transforms problems of an additive nature into *packing* and *covering* problems in a metric space. In particular, we see that problem 3 is equivalent to the fundamental problem of coding theory, namely determine the largest possible minimum distance of a linear code of length n and dimension k. There are several classical bounds relating n, k and d. Let us mention two simple bounds that we shall make use of later on.

***Proposition 2.2* (Hamming bound).** — *Any $[n, n-r, d]$ code satisfies*

$$\sum_{i=0}^{\lfloor (d-1)/2 \rfloor} \binom{n}{i} \leq 2^r.$$

Proof. — Since any vector $\mathbf{x} \in \mathbf{F}^n$ of weight $\leq d-1$ satisfies $\sigma(\mathbf{x}) \neq 0$, then all vectors with weight at most $\lfloor (d-1)/2 \rfloor$ must have distinct syndromes.

Using classical estimates for binomial coefficients, the Hamming bound states, asymptotically, that any $[n, nR, n\delta]$ code satisfies

$$R \leq 1 - h(\delta/2) + o(1) \tag{1}$$

where $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$ denotes the binary entropy function.

***Proposition 2.3* (Varshamov-Gilbert bound).** — *Let n and r be given. There exists an $[n, n-r, d]$ code whenever*

$$\sum_{i=0}^{d-1} \binom{n-1}{i} < 2^r.$$

Proof. — We construct inductively a parity-check matrix of such a code. Suppose constructed an $r \times i$ matrix $\mathbf{H}_i$ such that any $d-1$ columns are linearly independent. They are at most N_i distinct linear combinations of columns involving at most $d-2$ terms, with

$$N_i = \sum_{j=1}^{d-2} \binom{n}{j}.$$

If $N_i < 2^r - 1$, then a nonzero element of $G = \mathbf{F}^r$ can be added to the set of columns of $\mathbf{H}_i$ to yield an $r \times (i+1)$ matrix $\mathbf{H}_{i+1}$ with the property that any $d-1$ of its columns are linearly independent ; equivalently $\mathbf{H}_{i+1}$ is the parity-check matrix of a code of minimal distance $\geq d$.

Asymptotically, the Varshamov-Gilbert bound reads: there exist $[n, nR, n\delta]$ codes with

$$R \geq 1 - h(\delta) + o(1). \tag{2}$$

There is no known better asymptotic lower bound on $R = k/n$. Let us just mention the most powerful upper bound on R due to McEliece, Rodemich, Rumsey, and Welch (see e.g. [**10**]) for a proof):

Proposition 2.4. — *Any $[n, nR, n\delta]$ code satisfies*

$$R \leq h\left(\frac{1}{2} - \sqrt{\delta(1-\delta)}\right) + o(1). \tag{3}$$

Note that the Varshamov-Gilbert bound is not really constructive (the complexity of constructing a parity-check matrix for such codes is exponential in the length n). There are no known constructions of codes achieving the Varshamov-Gilbert bound for growing n and fixed R, $0 < R < 1$. There are, however, good constructions of codes with fixed d and growing n. We give a very short presentation of such codes, to which we shall refer later on.

Cyclic and BCH codes. — The one-to-one mapping

$$\mathbf{v} = (v_0, v_1, \dots, v_{n-1}) \leftrightarrow v(X) = v_0 + v_1 X + \cdots + v_{n-1}X^{n-1}$$

gives us an identification of the binary vector space $\mathbf{F}^n$ with the additive structure of the algebra $\mathcal{A} = \mathbf{F}[X]/(X^n - 1)$. If a subspace of $\mathcal{A}$ has the additional property of being an ideal of the ring $\mathcal{A}$, it is called a *cyclic* code. Every ideal of $\mathcal{A}$ is principal and generated by a polynomial $g(X)$ which divides $X^n - 1$. Take n of the form $n = 2^m - 1$ so that $g(X)$ can be considered to have all its roots in the finite field $\mathbf{F}_{2^m}$ on 2^m elements. Let α be a primitive element of $\mathbf{F}_{2^m}$ and define the cyclic code C_e as the set of polynomials (modulo $X^n - 1$) whose roots contain $\alpha, \alpha^3, \dots, \alpha^{2e-1}$ (since all these elements are roots of $X^n - 1$, this definition makes sense). It is a vector space over $\mathbf{F}$ of dimension at least $n - em$. Since these polynomials have their coefficients in $\mathbf{F}$, the set of their roots must be stable by the Frobenius homomorphism $x \mapsto x^2$, so that polynomials of C_e also have $\alpha^2, \alpha^4, \dots, \alpha^{2e}$ as roots. Note that C_e can also be described as those vectors $\mathbf{v} = (v_0, \dots, v_{n-1}) \in \mathbf{F}^n$ that are orthogonal to the rows of the matrix

$$\mathbf{H} = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{2^m-2} \\ 1 & \alpha^2 & \alpha^4 & \dots & \alpha^{2(2^m-2)} \\ 1 & \alpha^3 & \alpha^6 & \dots & \alpha^{3(2^m-2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \alpha^{2e} & \alpha^{4e} & \dots & \alpha^{2e(2^m-2)} \end{bmatrix}.$$

Now any $2e \times 2e$ submatrix of $\mathbf{H}$ is a van der Monde matrix and hence full-rank. Therefore any vector of $\mathbf{F}^n$ that is orthogonal to all the rows of $\mathbf{H}$ must have weight not less than $2e + 1$. The set C_e is called a BCH code: we have just proved that its parameters are

$$[n = 2^m - 1, k \geq 2^m - 1 - em, d \geq 2e + 1].$$

Cyclic and BCH codes have been extensively studied: see e.g. **[10]** and references therein. For fixed $d = 2e + 1$ and growing n they are, except for sporadic counter-examples, the best known constructions. Their dimension $k \geq n - e\log_2 n$ meets the asymptotic Hamming bound $k \leq n - e\log_2 n + 0(1)$.

Problem 2 is also classical, and can be reworded as: 'determine the smallest possible covering radius of a linear code of length n and dimension k'. We do not wish to dwell further on those two classical problems but rather refer to **[10]** for problem 3 and general background on coding theory, and to **[9]** for problem 2. Problem 4 is a generalisation of problem 3 that we shall discuss in section 4.

In the next section, we focus on problem 1 which is a truly additive problem in the sense that we are asking for those sets S, and their cardinalities, such that $S^{[1,t]} = S \cup S^2 \cup \cdots \cup S^t$ grows as slowly as possible.

3. Problem 1

Denote by $s_G(t)$ the smallest integer s such that, for any generating set S of $G = \mathbf{F}^r$, $\rho(S) \leq t$ whenever $|S| \geq s$. In other words $s_G(t) - 1$ is the largest cardinality of a generating set S of G such that $\rho(S) > t$. Problem 1 asks for the determination of $s_G(t)$. Proposition 2.1 tells us that this can be understood as asking for the largest possible covering radius of a linear code $C(S)$ of given length. Because $\rho(S) = \rho(S \setminus \{0\})$, and in view of the remark preceding proposition 2.1, we are really asking for the largest covering radius of a linear code of given length and minimum distance $d \geq 3$.

3.1. A lower bound on $s_G(t)$. — It is natural to consider the following sets.

Definition 3.1. — *Call a τ-*cylinder *of $\mathbf{F}^r$, a subset isomorphic to $S = B_{\tau,1}(0) \times \mathbf{F}^{r-\tau}$, where $B_{\tau,1}(0)$ denotes the ball centered on 0 and of radius 1 in $\mathbf{F}^\tau$. In other words, for some properly chosen basis of $\mathbf{F}^r$, S is the subset of vectors of $\mathbf{F}^r$ whose first τ coordinates make up a vector of weight at most one. If S is a τ-cylinder of $\mathbf{F}^r$, then $\rho(S) = \tau$ and $|S| = (\tau+1)2^{r-\tau}$.*

Since $s_G(t)$ must be larger than the cardinality of a $(t+1)$-cylinder, we have:

Proposition 3.1. — *Let* $\log_2 |G| \geq t+1$. *Whenever* $\log_2 |G| \geq t+1$,

$$s_G(t) > \frac{t+2}{2^{t+1}}|G|.$$

3.2. Upperbounding $s_G(t)$. — It is possible to prove that the above lower bound is the best possible for some values of r by a coding argument. The idea is to say, broadly speaking, that a code can't have too large a covering radius, otherwise, without changing the minimum distance, one would use it to construct a code with an impossibly large dimension.

Denote by $k(n,d)$ the maximum dimension of a linear code of length n and minimum distance at least d. We have, ([**9**])

Proposition 3.2. — *Let C be an $[n,k,d]$ linear code, and let ρ be its covering radius. We have:*

$$k + k(\rho, d) \leq k(n,d).$$

Proof. — Let $\mathbf{z}$ be a vector such that $d(\mathbf{z}, C) = \rho$ and of weight ρ. Assume, without loss of generality that its support is $supp(\mathbf{z}) = \{1,2,\ldots,\rho\}$. Let C' be a code of length ρ and dimension $k(\rho,d)$. Let $(C'|\mathbf{0})$ be the code of length n obtained from C' by appending $\mathbf{0} \in \mathbf{F}^{n-\rho}$ to all words in C'. It is not difficult to check that the sum $C + (C'|\mathbf{0})$ is a code with minimal distance at least d and dimension $k + k(\rho,d)$.

One has, besides:

Lemma 3.1. — $k(n,3) = n - 1 - \lfloor \log_2 n \rfloor$.

Proof. — To prove this, one just needs to find the smallest r such that there exists an $r \times n$ matrix with distinct non-zero columns.

Applying Proposition 3.2 and Lemma 3.1 we obtain the following upper bound on $s_G(t)$.

Proposition 3.3. — $s_G(t) \leq |G|/2^{t-\lfloor \log_2(t+1) \rfloor} + 1$.

Remarkably, the two bounds 3.1 and 3.3 coincide for t of the form $2^m - 2$, so that we have:

Corollary. — *For $m \geq 2$, $\log_2 |G| \geq 2^m - 1$, the following equality holds.*

$$s_G(2^m - 2) = |G|/2^{2^m - m - 1} + 1.$$

For the remaining values of t the upper and lower bounds of Propositions 3.1 and 3.3 leave a gap. For $t = 3$, the first value for which some uncertainty remains, we can obtain an improvement over Proposition 3.3. We make use of a theorem proved in **[13]** with a more traditional "additive" approach. It says that the subsets S of $\mathbf{F}^r$ such that $|S + S|$ is "small" tend to cluster around subgroups: the result is sharper than what can be said for general abelian groups. The precise statement is :

Theorem 3.1. — *Let S be a subset of $G = \mathbf{F}^r$. Let k be a nonnegative integer. One of the following holds.*

i. *There is a subgroup $H \neq \{0\}$ of G such that*

$$|S + H| - |S| < |H| + k$$

ii. *For any subset T of G such that $k \leq |T|^2 - 2$ and $2 \leq |G| - |S + T|$ we have*

$$|S + T| \geq |S| + |T| + k$$

Applying Theorem 3.1 with $k = 0$ yields:

Proposition 3.4. — $s_G(3) \leq |G|/3 + 1$.

Proof. — We prove that if S generates $G = \mathbf{F}^r$ and $|S| > |G|/3$ then $S + S + S = G$. We argue as follows. First check the result by hand for $\log_2 |G| \leq 4$. Then proceed by induction. If S satisfies the above, then:

1. Either $|S + S| \geq 2|S|$, and then $|S| + |S + S| > |G|$ so that the pigeon-hole principle implies $S + S + S = G$.
2. Or $|S + S| < 2|S|$, in which case Theorem 3.1 implies the existence of a non-trivial subgroup H of G such that $|S + H| - |S| \leq |H| - 1$. Consider now the partition $S = \cup S_i$ induced by the partition of G into cosets modulo H. Expressing $S + S$ as a union of sums $S_i + S_j$, we obtain by repeated application of the pigeon-hole principle that $S + S = S + S + H$. We finish by applying the induction hypothesis in G/H to the set of those cosets modulo H that intersect S, so as to obtain that $S + S + S$ intersects all cosets modulo H of G.

4. Constrained distances

In this section we consider problem 4, i.e. studying large sets S such that $0 \notin S^I$ for arbitrary $I \subset [1, |S|]$. Let us restate the problem in coding terms. We shall use the notation of **[5]**. For a code C, let

$$D(C) = \{w(\mathbf{c}) \mid \mathbf{c} \in C, \mathbf{c} \neq \mathbf{0}\}$$

$$l(n, D) = \max\{\dim C \mid D(C) \subset D\}.$$

Denote by $\overline{D} = [1, n] \setminus D$ the complement of D.

If $D \subset D(C)$, C is sometimes called a *D-clique*. The classical coding case is $D = [d, n]$, but the function $l(n, D)$ can vary very much with the nature of the set

D. For instance D-cliques with $D = [0, d]$, in other words sets with *maximal distance* d, have been considered under the name of *anticodes* **[6]**. These anticodes have been used to construct good codes, see ch. 17 §6 of **[10]**. More recently the problem of forbidding one distance, i.e. studying $l(n, \overline{\{d\}})$, has been considered. A variety of approaches to the problem have been put forward, among which additive techniques and more traditional coding approaches. By way of illustration, let us mention the problem of determining $l(4t, \overline{\{2t\}})$. It was conjectured by Ito that $l(4t, \overline{\{2t\}}) = 2t$. An elegant proof was found by Alon in the case $4t = 2^m$, using the following theorem of Olson. For an abelian group G, denote by $s(G)$ the smallest positive integer such than any sequence $g_1 \dots g_s$ of (not necessarily distinct) non-zero elements of G contains a subsequence summing to zero. Olson's Theorem **[12]** states.

Theorem 4.1. — *Consider the finite abelian p-group* $G = \mathbb{Z}/p^{e_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p^{e_k}\mathbb{Z}$ *; then* $s(G) = 1 + \sum_{i=1}^{k}(p^{e_i} - 1)$.

Sketch of Alon's proof. — Let C be a linear code of length $4t$ and dimension $2t + 1$. Consider the columns of a $(2t - 1) \times 4t$ parity-check matrix of C and add to each of them an extra coordinate consisting of the 1 element of the group $\mathbb{Z}/2^{m+1}\mathbb{Z}$. Thus we are dealing with $4t$ elements of the group $G = (\mathbb{Z}/2\mathbb{Z})^{2t-1} \times \mathbb{Z}/2^{m+1}\mathbb{Z}$. Olson's Theorem implies $s(G) = 4t - 1$, hence the existence of a proper subset of those elements that sum to zero: because of the last coordinate this subset must consist of exactly $2t$ elements and therefore correspond to a word of C of weight $2t$.

Ito's conjecture was finally proved in **[5]** for all t.

4.1. General results. — Most of the results of this section carry over to non linear codes: we shall not concern ourselves with these generalisations, however, since they would take us too far from our additive motivation.

Let us start by a general result of Delsarte **[4]**.

Theorem 4.2

$$2^{l(n,D)} \leq \sum_{i=0}^{|D|} \binom{n}{i}.$$

We present a concise proof of this classical result which should give some flavour of the methods of coding theory.

Proof of Theorem 4.2. — Let C be a code with parity-check matrix $\mathbf{H}$. Let S be the set of the columns of $\mathbf{H}$. Let us associate to C the Cayley graph $\mathcal{C}$ defined as having $G = \mathbf{F}^r$ as vertex set, and edge set $\{(g, g + s) \mid g \in G, s \in S\}$. Let $\mathbf{A} = (a_{uv})$ be the adjacency matrix of $\mathcal{C}$, i.e. the matrix whose rows and columns are indexed by G and such that

$$a_{uv} = \begin{cases} 1 \text{ if } v = u + s, & s \in S \\ 0 \text{ otherwise} \end{cases}$$

Notice that the quantity $\rho(C) = \rho(S)$ is exactly the diameter Δ of $\mathcal{C}$.

Let $Spe(\mathcal{C})$ be the set of eigenvalues of $\mathbf{A}$. The following lemma is classical in graph theory.

Lemma 4.1. — *Suppose the graph $\mathcal{C}$ has diameter Δ. Then*

$$\Delta + 1 \leq |Spe(\mathcal{C})|.$$

Proof of lemma 4.1. — Recall that the entry in position (u, v) of the matrix $\mathbf{A}^i$ equals the number of walks $u = u_0, u_1 \dots, u_i = v$ of length i from vertex u to vertex v. Consider the algebra $\mathbf{C}[\mathbf{A}]$ of polynomials in $\mathbf{A}$. On the one hand, it is standard linear algebra that $\dim \mathbf{C}[\mathbf{A}] = |Spe(\mathcal{C})|$. On the other hand, whenever $i \leq \Delta$, there is a walk of length i from some vertex u to some vertex v such that no walk of length $< i$ exists between u and v. This means that $(\mathbf{A}^i)_{uv} \neq 0$ while $(\mathbf{A}^j)_{uv} = 0$ for $j < i$. Therefore $\{\mathbf{I}, \mathbf{A}, \dots, \mathbf{A}^\Delta\}$ is a linearly independent set in $\mathbf{C}[\mathbf{A}]$.

Now in our particular case, it is straightforward to check that for any character χ of G, $[\chi(v)]_{v \in G}$ is an eigenvector of $\mathbf{A}$ associated to the eigenvalue

$$\lambda_\chi = \sum_{s \in S} \chi(s).$$

Every character of the group $G = \mathbf{F}^r$ is of the form

$$\chi_u : v \mapsto (-1)^{(u|v)}$$

for some $u \in G$, where $(u|v)$ denotes the scalar product in $\mathbf{F}^r$. So we see that $\lambda_{\chi_u} = n - 2w({}^t u.\mathbf{H})$, so that the number of distinct eigenvalues of $\mathcal{C}$ is exactly the number of distinct weights in the subspace generated by the rows of $\mathbf{H}$, i.e. the dual code $C^\perp$ of C. Summarizing, we have:

***Theorem 4.3* (Delsarte)**. — $\rho(C) \leq |D(C^\perp)|$.

Note now that, from the definition of $\rho(S) = \rho(C)$, one has the inequality

$$\sum_{i=0}^{\rho(C)} \binom{n}{i} \geq 2^r. \tag{4}$$

Relation (4) together with Theorem 4.3 prove Theorem 4.2.

Let us state the following result from **[5]**.

Proposition 4.1. — *For $n \geq 4t$,*

$$\begin{aligned} l(n, \overline{\{2t\}}) &\leq n - 2t \\ l(n, \overline{\{2t, 2t+1\}}) &\leq n - 2t - 1. \end{aligned}$$

We shall now derive a variation on the so-called "Elias-Bassalygo lemma" **[1]**.

Denote by $A(n, D)$ the maximal size of a (not necessarily linear) subset of $\mathbf{F}^n$ such that any two of its elements have distance in D.

Denote by $A(n, D, w)$ the maximal size of a subset of $\mathbf{F}^n$ such that any two of its elements have weight w and distance in D.

Proposition 4.2

$$A(n,D) \leq \frac{2^n}{\binom{n}{w}} A(n,D,w).$$

Proof. — Let C be a code (simply a set of vectors in the non-linear case) realizing $A(n,D)$. Consider its 2^n translates $C+\tau, \tau \in \mathbf{F}^n$. Each vector of $\mathbf{F}^n$, and in particular those of weight w, appear $A(n,D)$ times in the union of the translates $C+\tau$. Thus one of the translates, in itself a D-clique because $d(\cdot,\cdot)$ is invariant by translation, must contain at least $\binom{n}{w}A(n,D)2^{-n}$ vectors of weigh w. Hence

$$\binom{n}{w} A(n,D)2^{-n} \leq A(n,D,w).$$

4.2. Forbidding one distance. — We shall need the following result [**7**].

Proposition 4.3. — *If $\mathcal{F}$ is a family of w-subsets of an n-set no two of which intersect in exactly e elements, then*

$$|\mathcal{F}| \leq c_w n^{\max\{e,w-e-1\}}$$

where c_w is a constant depending only on w.

Set $w = d = 2e$, then clearly any two members of a family achieving $A(n,\overline{2e},2e)$ do not intersect in e elements. Thus Proposition 4.3 yields

$$A(n,\overline{2e},2e) \leq c_{2e}n^e$$

and by Proposition 4.2 we get, fixing e and letting n go to infinity,

$$A(n,\overline{2e}) = O\left(\frac{2^n}{n^e}\right).$$

Hence,

$$l(n,\overline{2e}) \leq n - e\log_2 n + O(1). \tag{5}$$

In other words, for fixed e, it is asymptotically just as costly to forbid the distance $2e$ between codewords as to forbid all distances d, $1 \leq d \leq 2e$, since BCH codes meet (5). We have:

Proposition 4.4. — $l(n,\overline{2e}) = n - e\log_2 n + O(1)$.

We now consider the case when the forbidden distance d increases linearly with n. In other words, we fix λ and study $l(n,\overline{\lambda n})$ by which we mean $l(n,\overline{\{\lfloor\lambda n\rfloor\}})$. Some caution is in order when dealing with the asymptotical behaviour of $n^{-1}l(n,\overline{\lambda n})$, since this function of n does not converge: indeed, the $[n,n-1,2]$ even weight subcode of $\mathbf{F}^n$ shows that $l(n,\overline{2e+1}) = n-1$, hence $\limsup n^{-1}l(n,\overline{\lambda n}) = 1$. We suspect that the sequence $n^{-1}l(n,\overline{\lambda n})$ actually has many accumulation points. We shall now derive a result on $\liminf n^{-1}l(n,\overline{\lambda n})$.

We shall need the following from [**8**]:

Proposition 4.5. — *Let q be a prime power. Let $\mathcal{F}$ be a set of w-subsets of the n-set $\{1,2,\dots,n\}$. Suppose that for any distinct $F,F' \in \mathcal{F}$ we have*

$$|F \cap F'| \not\equiv w \bmod q$$

then

$$|\mathcal{F}| \le \binom{n}{q-1}.$$

We now obtain:

Proposition 4.6

$$\liminf_{n\to\infty} n^{-1}l(n,\overline{\lambda n}) \le 1 - h(\lambda) + h(\lambda/2) + o(1).$$

Proof. — Suppose d equals twice the power of a prime $d = 2q$. Let $w = 2q-1$. Any code of constant weight w and such that no two codewords are at distance d from each other yields a set $\mathcal{F}$ such that $|F\cap F'| \not\equiv -1 \bmod q$ for distinct $F,F' \in \mathcal{F}$. Hence

$$A(n,\overline{2q},2q-1) \le \binom{n}{q-1} \le 2^{n(h(\lambda/2)+o(1))}.$$

Apply Proposition 4.2 to conclude the proof.

Note that for $\lambda < 0.27$, this improves on Proposition 4.1.

4.3. Forbidding multiples of a given distance. — More generally, if q is a prime power and $\lambda n = 2iq$, considering constant weight codes of weight $w = (i+1)q-1$, one obtains

Proposition 4.7

$$n^{-1}l(n,\overline{\{2q,4q,\dots,2iq\}}) \le 1 - h\left(\frac{i+1}{2i}\lambda\right) + h\left(\frac{\lambda}{2i}\right) + o(1).$$

Remark. — *For growing i, the right hand side of this last inequality tends to $1 - h(\lambda/2)$, so that it can be considered as a refinement of the Hamming bound (1)*

$$n^{-1}l(n,\overline{[1,\dots,\lambda n]}) \le 1 - h(\lambda/2)$$

in the sense that one need not forbid every distance in $[1,\dots,\lambda n]$.

4.4. A construction. — We have the lower bound:

Proposition 4.8. — *For $\lambda \le 1/3$,*

$$n^{-1}l(n,\overline{\lambda n}) \ge 1 - (1-\lambda)h\left(\frac{\lambda}{1-\lambda}\right) + o(1).$$

Proof. — Consider the generating matrix

$$\mathbf{G} = \begin{bmatrix} \mathbf{I}_{\lambda n-1} & 0 \\ 0 & \mathbf{G}_0 \end{bmatrix}$$

where $\mathbf{G}_0$ is a generator matrix of an optimal code C_0 of length $n - \lambda n + 1$ and distance $\lambda n + 1$. Obviously every combination of rows of $\mathbf{G}$ has weight at most $\lambda n - 1$ - if it does not use rows of $\mathbf{G}_0$ - or at least $\lambda n + 1$ if it does.

Take for C_0 a code lying on the Varshamov-Gilbert bound (2) to get the asymptotical result.

Large gaps remain between upper and lower bounds.

Open problem. — *It would be particularly interesting to known what is the most "persistent" distance in linear codes, in other words, what is the value of λ that minimizes* $\liminf n^{-1} l(n, \overline{\lambda n})$ *?*

5. Intersecting codes

We would like to conclude by another intriguing problem with an additive flavour. Let us say that a subset $S = \{s_1, \ldots, s_n\}$ of an abelian group G has the *intersecting property* if there do not exist two disjoint subsets I and J of $[1, n]$ such that both

$$\sum_{i \in I} s_i = 0 \qquad \text{and} \qquad \sum_{j \in J} s_j = 0.$$

An *intersecting code C* is a linear code with the property that any two non-zero codewords have intersecting supports. Equivalently, it is a code C such that the set of columns of any parity-check matrix of C has the intersecting property in $G = \mathbf{F}^r$.

Problem 5. — *Given r, what is the maximal size $\iota(r)$ of $S \subset \mathbf{F}^r$ with the intersecting property ?*

This problem was first investigated by Miklós [**11**], and has since proved to lead to a variety of applications, see [**2**]. A lower bound on $\iota(r)$ can be derived by random arguments [11,2] Asymptotically it reads:

$$\iota(r) \geq \frac{2r}{\log_2 3} \approx 1.26r.$$

To obtain an upper bound, notice that an intersecting code must have $d \geq k$. Otherwise choose a minimum weight codeword $\mathbf{c}$: among the 2^k codewords there must be two, $\mathbf{c}'$ and $\mathbf{c}''$, that coincide on the d coordinates of the support of $\mathbf{c}$. Therefore $\mathbf{c}$ and $\mathbf{c}' + \mathbf{c}''$ have nonintersecting supports, a contradiction. This argument, namely $\delta \geq R$, together with the bound (3) gives

$$\iota(r) \leq 1.40r.$$

References

[1] Bassalygo L.A., *New bounds for error-correcting codes.*, Problemy Peredachi Informatsii, **1**, 1965, 41–45.

[2] Cohen G. and Zémor G., *Intersecting codes and independent families*, IEEE Trans. on Inf. Theory, **40**, 1994, 1872–1881.

[3] Cohen G.D., Karpovsky M., Mattson H.F. Jr. and Schatz J., *Covering radius - survey and recent results*, IEEE Trans. Inf. Theory, **31**, 1985, 328–344.

[4] Delsarte P., *Four fundamental parameters of a code and their combinatorial significance*, Info. and control, **23**, 1973, 407–438.

[5] Enomoto H., Frankl P., Ito N. and Nomura K., *Codes with given distances*, Graphs and Combinatorics, **3**, 1987, 25–38.

[6] Farrell P.G., *Linear binary anticodes*, Electronics Letters, **6**, 1970, 419–421.

[7] Frankl P. and Füredi Z., *Forbidding just one intersection*, J.C.T. A, **39**, 1985, 160–176.

[8] Frankl P. and Wilson R.M., *Intersection theorems with geometric consequences*, Combinatorica, **1**, 1981, 357–368.

[9] Godlewski P., *WOM-codes construits à partir des codes de Hamming*, Discrete Math., **65**, 1987, 237–243.

[10] MacWilliams F.J. and Sloane N.J.A., *The theory of error-correcting codes*, North-Holland, 1977.

[11] Miklós D., *Linear binary codes with intersection properties*, Discrete Applied Math., **9**, 1984, 187–196.

[12] Olson J.E., *A combinatorial problem on finite abelian groups*, J. Number Theory, **1**, 1969, 195–199.

[13] Zémor G., *Subset sums in binary spaces*, Europ. J. of Combinatorics, **13**, 1992, 221–230.

GÉRARD COHEN, Ecole Nationale Supérieure des Télécommunications, 46 rue Barrault, 75 634 Paris Cedex 13, France • *E-mail :* cohen@inf.enst.fr

GILLES ZÉMOR, Ecole Nationale Supérieure des Télécommunications, 46 rue Barrault, 75 634 Paris Cedex 13, France • *E-mail :* zemor@res.enst.fr

Astérisque

MARK CHAIMOVICH

New structural approach to integer programming : a survey

Astérisque, tome 258 (1999), p. 341-362

<http://www.numdam.org/item?id=AST_1999__258__341_0>

Astérisque **258**, 1999, p. 341–362

NEW STRUCTURAL APPROACH TO INTEGER PROGRAMMING: A SURVEY

by

Mark Chaimovich

Abstract. — The survey discusses a new approach to Integer Programming which is based on the structural characterization of problems using methods of additive number theory. This structural characterization allows one to design algorithms which are applicable in a narrower, yet still wide, domain of problems, and substantially improve the time boundary of existing algorithms. The new algorithms are polynomial for the class of problems in which they are applicable, and even linear ($O(m)$) for a wide class of the Subset-Sum and Value-Independent Knapsack problems. Previously known polynomial time algorithms for the same classes of problems are at least two orders of magnitude slower.

1. Introduction

This survey considers a recently developed approach to Integer Programming (IP) which is based on the application of analytical methods of Additive Number Theory. Elaborated by G. Freiman in the early 1980's, this new approach was developed by N. Alon, P. Buzytsky, M. Chaimovich, P. Erdős, G. Freiman, Z. Galil, E. Lipkin and O. Margalit (in alphabetical order).

In general, the number of Integer Programming models is vast and they have numerous applications; only a few of them – Subset-Sum (one and multi-dimensional), Value-Independent Knapsack and k-Partition problems – were investigated using the new structural approach. Theorems from analytical number theory allow one to characterize the structure of the domain of solutions for a wide class of problems and to design efficient algorithms for these problems. These new algorithms substantially improve the time boundary of existing algorithms. They are polynomial for the class of problems in which they are applicable, and even linear ($O(m)$) for certain classes of the Subset-Sum and Value-Independent Knapsack problems. That is at least two orders of magnitude faster than previously known polynomial time algorithms for the

1991 ***Mathematics Subject Classification.*** — Primary: 90C02 Alternate: 05A17, 11B25, 68Q25.
Key words and phrases. — Analytical Number Theory, Integer Programming, Subset Sum Problem.

same classes of problems. This fact allows one to solve problems with a much larger number of variables.

This article is organized into several parts. In section 2 the general idea for development of an analytical approach to Integer Programming is considered. Sections 3 and 4 deal with the Subset-Sum Problem (SSP). The first of them provides a detailed, structural analysis of the problem including an example of the analytical theorem while the second describes algorithms for solving SSP based on this structural analysis. Proofs of the validity of the algorithms are not provided in this survey, however, they may be found in the references. Section 5 describes the application of the structural approach to multi-dimensional Subset-Sum, Value-Independent Knapsack and k-Partition problems. (Only the main theorems and outlines of the algorithms are presented.) In the conclusion possible directions for future research are discussed.

2. General idea of the application of the structural approach to IP

In this section the main idea of the structural approach is described. We begin with a simple example that illustrates the approach. Further, the concept of density is discussed, this explains how the structural characterization of the problem may be obtained. We conclude the section with a short history of the research in the field of structural characterization.

2.1. A simple illustration of the structural approach. — In order to understand a structural approach to IP, consider the problem of feasibility of a single boolean equation. Given an integer m, an integral vector $(a_1, a_2, \dots, a_m)$ and an integer N, does equation

$$a_1x_1 + a_2x_2 + \cdots + a_mx_m = N \tag{1}$$

have any solutions for $x_i \in \{0, 1\}$ for all i? To illustrate the approach, we use the following concrete equation

$$7x_1 + 8x_2 + 14x_3 + 15x_4 + 22x_5 + 28x_6 + 56x_7 = 75, \tag{2}$$

i.e. $m = 7$, $(a_1, \dots, a_7) = (7, 8, 14, 15, 22, 28, 56)$ and $N = 75$.

Dynamic programming approach

Denoting $S_0 = \{0\}$ and $S_k = \{b | b = \sum_{j=1}^{k} a_jx_j, x_j \in \{0, 1\}\}$ for $1 \leq k \leq 7$, we have $S_k = S_{k-1} + \{0, a_k\} = \{b | b \in S_{k-1}$ *or* $b - a_k \in S_{k-1}\}$. Thus, having S_7 – the set of all possible values of the linear form in the left-hand side of (2),– it remains only to check if $N = 75 \in S_7$. In fact,

$$\begin{aligned} S_1 &= \{0, 7\}, \\ S_2 &= \{0, 7, 8, 15\}, \\ S_3 &= \{0, 7, 8, 14, 15, 21, 22, 29\}, \\ &\dots \end{aligned}$$

and so on. Finally,

$$\begin{aligned} S_7 &= \{0,7,8,14,15,21,22,23,28,29,30,35,36,37,42,43,44,45,49,50, \\ &\quad 51,52,56,57,58,59,63,64,65,66,70,71,72,73,77,78,79,80,\ldots\}, \end{aligned}$$

i.e., $75 \notin S_7$ and equation (2) does not have a solution.

Structural approach

We characterize the structure of S_7 without explicitly enumerating it. Observe, that some of the coefficients of the equation are divisible by 7: $a_1 \equiv a_3 \equiv a_6 \equiv a_7 \equiv 0(\mathrm{mod}\, 7)$. Then, for $b \in S_7$ we have $b \equiv 8x_2 + 15x_4 + 22x_5 \equiv x_2 + x_4 + x_5(\mathrm{mod}\, 7)$, i.e.,

$$b \equiv 0,1,2,3(\mathrm{mod}\, 7). \tag{3}$$

However, $75 \equiv 5(\mathrm{mod}\, 7)$, so, the equation does not have a solution.

Condition (3) determines a necessary condition for solvability equation (2). In order to obtain a sufficient condition let us analyze the same equation with another right-hand side:

$$7x_1 + 8x_2 + 14x_3 + 15x_4 + 22x_5 + 28x_6 + 56x_7 = 79.$$

Clearly, $79 \equiv 2(\mathrm{mod}\, 7)$, so it can belong to S_7 according to (3). To confirm that it really belongs to S_7, consider a linear form

$$L = 7x_1 + 14x_3 + 28x_6 + 56x_7 = 7(x_1 + 2x_3 + 4x_6 + 8x_7).$$

The linear form $L' = x_1 + 2x_3 + 4x_6 + 8x_7$ can take all values from 0 to 15, thus, the linear form L can, correspondingly, take values of the form $7t$, where $0 \leq t \leq 15$. When we combine these values with the other coefficients $(8, 15, 22)$, we have

$$\begin{aligned} S_7 = \{b \mid\ & b \equiv 0(\mathrm{mod}\, 7), 0 \leq b \leq 7 \cdot 15, \ \textit{or} \\ & b \equiv 1(\mathrm{mod}\, 7), 8 \leq b \leq 22 + 7 \cdot 15, \ \textit{or} \\ & b \equiv 2(\mathrm{mod}\, 7), 23 \leq b \leq 37 + 7 \cdot 15, \ \textit{or} \\ & b \equiv 3(\mathrm{mod}\, 7), 45 \leq b \leq 45 + 7 \cdot 15\}. \end{aligned} \tag{4}$$

Here 8, 23, 45 are the smallest numbers with residues 1, 2, 3 modulo 7 that can be represented by the linear form in the left-hand side of the equation. Since $79 \equiv 2(\mathrm{mod}\, 7)$ and $79 = 23 + 7 \cdot 8$, the answer is that the equation has at least one solution.

Observe that the above consideration determines the structure of the set of possible values of a linear form on the left-hand side of an equation as a collection of arithmetic progressions with a common difference. This fact allows one to solve the problem immediately for each right-hand side. One can suppose that this example was especially selected to illustrate the approach and that would be true. However the situation obtained can be generalized: for a wide class of problems we can always determine the structure.

To obtain a general structural characterization of the IP problem (in the same way that (4) was obtained for a concrete equation), a specific analytical theorem must be proven. Of course, certain conditions have to be imposed on the coefficients in order to obtain such a characterization. These conditions follow directly from the analytical theorem. Once we have the conditions, it is possible to go to the next step – to design algorithms to verify these conditions and to obtain the structure.

Indeed, the structure obtained and the conditions of its existence provide an understanding of why some problems are easy and others are very hard for various enumerative algorithms. To confirm this statement consider the following problem which was investigated by R. Jeroslow (1974) **[19]**: maximize x_1 satisfying $2x_1 + 2x_2 + \cdots + 2x_n = n$ where n is odd. Although this problem is by nature trivial, it requires almost complete enumeration using different enumerative techniques. (Branch and Bound, for example, is one of them.) The secret is the fact that the constraint has no solutions, however, we must verify all possibilities to confirm this fact. The structural approach allows one to obtain an answer for this problem in no time.

2.2. Concept of density and its use in structural characterization. — In order to apply analytical methods to solve an IP problem, it is necessary for the problem to have a *high density.* To explain the notion of "density" and its importance in the application of the analytical approach to IP, let us consider again the feasibility of equation (1).

Let $\ell = \max_{1 \le i \le m} a_i$. The linear function on the left in (1) has a domain of size 2^m and a range of size $m\ell$. Since the domain size represents the overall number of "solutions" for all possible values of the right-hand side, the ratio $\frac{2^m}{m\ell}$ represents the average number of "solutions" for a value from the range. We say that this ratio characterizes the density of the problem. The density of other IP problems can be defined similarly.

In the case of equation (1), the density condition means that $\ell = o(\frac{2^m}{m})$ or $\frac{2^m}{m\ell} \to \infty$. Currently, algorithms are still not capable of handling this density. The only situation that has been investigated is $\ell = O(\frac{m^2}{\log m})$. The conjecture of G. Freiman is that the new approach can be refined to handle the case $\ell = O(m^c)$ for any positive constant c.

To highlight basic features of the approach, we present some non-strict considerations resulting from probability theory. In view of

$$\int_0^1 e^{2\pi i\alpha b} d\alpha = \begin{cases} 0 & \text{for } b \in \mathbb{Z}, b \neq 0, \\ 1 & \text{for } b = 0, \end{cases}$$

it is easy to verify that the number of solutions of (1) can be expressed by the integral

$$(5) \quad J(N) = \int_0^1 \prod_{j=1}^m (1 + e^{2\pi i\alpha a_j}) e^{-2\pi i\alpha N} d\alpha = 2^m \int_0^1 \prod_{j=1}^m (\tfrac{1}{2} + \tfrac{1}{2} e^{2\pi i\alpha a_j}) e^{-2\pi i\alpha N} d\alpha.$$

One may look at $\frac{1}{2} + \frac{1}{2}e^{2\pi i\alpha a_j}$ as the characteristic function of a random variable ξ_j taking values 0 and a_j with probabilities equal to $\frac{1}{2}$. Then the value of integral (5) is equal to the probability $P(\zeta = N)$, where $\zeta = \xi_1 + \cdots + \xi_m$ is a random variable with mathematical expectation $M = \frac{1}{2}\sum_{j=1}^m a_j$ and dispersion $\sigma^2 = \frac{1}{4}\sum_{j=1}^m a_j^2$. Assuming that the local limit theorem can be applied, the variable ζ has asymptotically normal distribution; we therefore have

$$J(N) \sim \frac{2^m}{\sqrt{2\pi\sigma^2}} e^{-\frac{(M-N)^2}{2\sigma^2}}, \tag{6}$$

which implies the existence of solutions for equation (1) for right-hand sides N in a wide interval of the mathematical expectation M.

As a rule, a local limit theorem is not always available. In spite of this difficulty, the precise analysis of integral (5) (see **[11]**, **[20]**, **[1]**, for example) confirms the validity of the asymptotic formula (6) for sufficiently dense equations whose coefficients satisfy some distributive properties. These distributive properties require that not "too many" coefficients have one common divisor. Non-compliance with this requirement provides a special structure for the range of the linear form on the left in (1).

Note that application of the analytical approach to a specific IP model requires one to prove the model's own structural theorem. Alternatively, one can reduce the new problem to another problem for which the structural theorem is already proved.

2.3. Historical background. — The possibility of using analytical methods for solving IP problems was shown for the first time by G. Freiman in 1980 **[12]** (see also P. Buzytsky and G. Freiman **[3]**). However, at that time, his concepts did not provide an explicit structural characterization of the problem. Only recently has determination of a precise structure for some IP problems become possible on the basis of methods proposed by G. Freiman and P. Erdős in **[11]**.

The first works investigating structural characterization of IP using analytical methods were concentrated on the following Subset-Sum Problem (SSP) with *different summands*: Given a set A of positive integers and a number N, find (a) $z = \max\{S_B = \sum_{a \in B} \mid S_B \leq N, B \subseteq A\}$ and (b) subset $B \subset A$ such that $S_B = z$.

The authors proved the analytical theorems, showing that a set of subset-sums around the middle sum may be characterized as a collection of long arithmetic progressions with a common difference. P. Erdős and G. Freiman **[11]** assumed very dense inputs ($m > \frac{\ell}{3}$) and a very small interval. E. Lipkin **[20]** improved the density ($m > \ell^{4/5+\varepsilon}$) and enlarged the interval size. N. Alon and G. Freiman **[1]** further improved the density ($m > \ell^{2/3+\varepsilon}$) but used a small interval (like in **[11]**). Later, G. Freiman **[17]** proved the same result for sets with density $m > c(\ell \log \ell)^{1/2}$. All of these characterization theorems used analytic number theory and hold true for sufficiently large values of ℓ. M. Chaimovich (**[5]** and **[9]**) shows the existence of an arithmetic progression in subset sums for sets with density $m > g(\ell)\ell^{2/3} \log^{1/3} \ell$, $\ell \geq 155$, where $g(\ell)$ is some function depending on ℓ, $1.9 < g(\ell) < 2.5$. The proof is done with exact computation of all constants, which allows one to use the result in practical algorithms.

A. Sárkőzy **[22]** has independently obtained an arithmetic progression for sets with the same density as **[17]**; he used algebra and combinatorial methods. However, his proof is not constructive and therefore it may not be applicable to algorithmic design. Z. Galil and O. Margalit **[18]**, using elementary number theoretic facts only (in contrast to A. Sárkőzy's approach), have explicitly constructed a long arithmetic progression in subset-sums. They achieved density $m = O(\ell^{1/2} \log \ell)$ (slightly weaker than **[17]** and **[22]**) and matched the interval size of **[20]**.

To complete the discussion of the results related to SSP with different summands, we mention that G. Freiman has shown in **[17]** that the density $m = \Omega(\ell^{1/2})$ is the lowest density for which the characterization of the structure of subset-sums is an arithmetic progression. For sets of different summands with $m < \ell^{1/2}$ the structure is more complicated. G. Freiman conjectures that this structure is multi-dimensional in the sense that it is formed by a few, relatively short, arithmetic progressions, and that each of these arithmetic progressions can be viewed as a "dimension" of the structure.

The first algorithms solving the SSP were derived by G. Freiman **[13]**, **[14]** and M. Chaimovich **[5]**. They solve in linear time problem (a) which finds the maximal sum but not the subset. In comparison, dynamic programming solves the same problem in $O(m^2\ell)$ time which is two orders of magnitude slower. Solving problem (b) with this approach (see **[10]**) takes $O(\ell^2 \log \ell)$ time. The algorithm of Z. Galil and O. Margalit **[18]** solves both problems (a) and (b) – finding the maximal sum and the subset. It reaches $O(\ell \log \ell)$ time improving **[10]** by one order of magnitude.

The SSP with repeated summands (relaxing the restriction that the summands must be distinct) was considered by M. Chaimovich in **[4]**, **[6]**. The existence of a long arithmetic progression in a set of subset-sums was proved for $m > 6\ell \log \ell$. This estimate is the best possible apart from a logarithmic factor and a constant.

Investigation of the multi-dimensional SSP, where vectors take place of the integers, was begun by G. Freiman **[16]**. He has shown that for two-dimensional problems an integral lattice takes the place of an arithmetic progression in the structural characterization. M. Chaimovich **[8]** extended this result for an arbitrary number of dimensions. For multi-dimensional problems the time boundary of the new algorithm is more impressive than the one dimensional one: for n-dimensional SSP it reaches $O(m^2)$ time instead of $O(m^{n+2})$ in dynamic programming.

Another problem investigated recently by using the structural approach was a k-partition problem (KPP): Given a set A of positive integers and k positive target numbers $N_1 \leq N_2 \leq \cdots \leq N_k$ such that $\sum_{i=1}^{k} N_i = S_A$, find a partition of A into k subsets $(B_1, \ldots, B_k), \bigcup_{j=1}^{k} B_j = A$, whose sums are closest to the target numbers in the sense that they minimize (or maximize) an appropriate objective function z.

This problem is especially hard to solve using traditional methods. It was solved by dynamic programming (see **[21]**) in $O(m^{2k})$ time. Applying the structural approach to it (M. Chaimovich **[7]**) gives $O(m^{1+1/(k-1)})$ time for sufficiently dense input sets. The gain is considerable for a fixed k as well as for k increasing with ℓ (its value is bounded by $\ell^{1/4}$).

3. Analytical method for structural analysis of the Subset-Sum Problem.

In this section we provide a detailed explanation of the structural characterization of the set of subset-sums. First we determine sufficient conditions of the existence of a long interval in a set of subset-sums (Theorem 3.1). Next we elaborate the structure of the set of subset-sums in the case where sufficient conditions are not fulfilled. Our consideration is based on the proofs from **[1]** and **[9]**.

3.1. Existence of a long interval in a set of subset-sums – sufficient conditions.

— For a set X define

$$m_X = |X|, \quad \ell_X = \max\{x \in X\}, \quad S_X = \sum_{x\in X} x, \quad \sigma_X = \tfrac{1}{2}(\sum_{x\in X} x^2)^{1/2},$$
$$X^* = \{z \mid \exists Y \subseteq X, S_Y = z\}, \quad X(s,q) = \{x \in X \mid x \equiv s (\operatorname{mod} q)\}.$$

Note: In the following theorem and further on $c_0, c_1, c_2, \dots,$ always denote absolute positive constants. We will also omit the subscript identifying the set if it is clear from the context which set is being discussed.

Theorem 3.1. — *Let A be a set of positive integers, such that*

$$m > c_1 \ell^{2/3} \log^{1/3} \ell > c_0. \tag{7}$$

Suppose that for all integers q the inequality

$$|A(0,q)| \le m - \ell^{2/3} \log^{1/3} \ell \tag{8}$$

is true. Then all integers N for which

$$\left|N - \tfrac{1}{2} S_A\right| \le c_2 \sigma_A \tag{9}$$

belong to the set of subset-sums of A, i.e., $\left[\frac{1}{2}S_A - c_2\sigma_A, \frac{1}{2}S_A + c_2\sigma_A\right] \subseteq A^*$.

General idea of the proof. — The fact that an integer N belongs to the set of subset-sums is equivalent to the existence of a solution of a linear equation (1).

For $1 \le j \le m$ define $\varphi_j(\alpha) = \frac{1}{2}\left(1 + e^{2\pi i \alpha a_j}\right)$ and $\varphi(\alpha) = \prod_{j=1}^{m} \varphi_j(\alpha)$. As mentioned on page 344, the number of solutions to equation (1) can be expressed by the integral $J(N) = 2^m \displaystyle\int_0^1 \varphi(\alpha) e^{-2\pi i \alpha N} d\alpha$, thus, it is necessary to show that $J(N) \ge 1$ whenever N satisfies (9). In order to do this we can prove the asymptotic formula

$$J(N) = (1 + o(1)) \frac{2^m}{\sqrt{2\pi\sigma^2}} e^{-\frac{(M-N)^2}{2\sigma^2}}, \tag{10}$$

for the number of solutions of equation (1).

Let us analyze the nature of conditions of the theorem. A restriction (9) on number N ensures that the exponent in (10) is not too small. This restriction is necessary to obtain an asymptotic formula, but not to prove the existence of a solution and/or a structure, therefore, we will relax this restriction below.

Condition (7) represents the density of a problem in the sense that the number of combinations of unknowns is large with respect to a range of possible values of the linear form. This condition can be strengthened to $m > c\ell^{1/2} \log^{1/2} \ell$ (see **[17]** and **[22]**) , but then the proof becomes quite complicated. In any case we need a condition of density to ensure the existence of the structure.

Finally, condition (8) is a condition of distribution. Its validity is necessary to obtain an asymptotic formula, but it is not necessary to obtain a structural characterization. The influence of a distribution of summands on a structure will be studied in the next paragraph.

Let us define $F_N(\alpha) = \varphi(\alpha)e^{-2\pi i\alpha N}$ and $L = 2\ell$. Observing that $F_N(\alpha)$ is a periodic function with a period equal to 1, one can write

$$(11)\quad J(N) = 2^m \int_{-\frac{1}{L}}^{1-\frac{1}{L}} F_N(\alpha)d\alpha = 2^m \left(\int_{-\frac{1}{L}}^{\frac{1}{L}} + \int_{\frac{1}{L}}^{1-\frac{1}{L}}\right) \geq 2^m \left(\left|\int_{-\frac{1}{L}}^{\frac{1}{L}}\right| - \left|\int_{\frac{1}{L}}^{1-\frac{1}{L}}\right|\right).$$

Note that for a sufficiently high density, the first integral on the right side of the equation in (11) provides the major part of the asymptotic formula for $J(N)$; the second integral forms the error term. The proof estimates these two integrals separately. It shows that

$$\int_{-\frac{1}{L}}^{\frac{1}{L}} F_N(\alpha)d\alpha = (1+o(1))\frac{2^m}{\sqrt{2\pi\sigma_A^2}}e^{-\frac{(M-N)^2}{2\sigma_A^2}},$$

and that

$$\left|\int_{\frac{1}{L}}^{1-\frac{1}{L}} F_N(\alpha)d\alpha\right| = o(\frac{1}{\sigma_A}).$$

In this survey we omit the detailed explanation of the integrals' evaluation.

Enlarging the interval. — According to Theorem 3.1, the interval $[\frac{1}{2}S_A - c_2\sigma_A, \frac{1}{2}S_A + c_2\sigma_A]$ belongs to the set of subset-sums. The length of the interval may be easily estimated to be at least $\sigma_A = \Omega(\ell \log^{1/2}\ell) \gg \ell$. However, this length is "small" relative to the range of subset sums which is $S_A > \frac{1}{2}m^2 \gg \ell^{4/3}$ (for our density).

Now we are going to show that this interval may be larger without considerably enlarging density.

Take the set A with $m_A = c_3 m_0$ where $m_0 = c_1\ell^{1/2}\log^{1/2}\ell$ (the density required by the theorem) and $c_3 \geq 2$. Let $1 \leq a_1 < a_2 < \cdots < a_m$ where $a_i \in A$ and denote $A' = \{a_j\}_{j=1}^{m_0}$. Suppose also that A' satisfies a condition similar to condition (8) of Theorem 3.1 so that Theorem 3.1 can be applied to set A'. According to the theorem, interval $I = [\frac{1}{2}S_{A'} - c_2\sigma_{A'}, \frac{1}{2}S_{A'} + c_2\sigma_{A'}]$, which is longer than ℓ, belongs to the set of subset sums of A'.

Return now to the original set A. Denote $N_i = N - \sum_{j=1}^{i} a_{m_0+j}$ for $1 \leq i \leq m - m_0$. Clearly, $0 < N_i - N_{i+1} \leq \ell$ and, whenever $N \in [\frac{1}{2}S_{A'}, S_A - \frac{1}{2}S_{A'}]$, for some i_0 we will have $N_{i_0} \in I$, which means that the entire interval $[\frac{1}{2}S_{A'}, S_A - \frac{1}{2}S_{A'}]$ belongs to the set of subset sums of A.

To estimate the length of this long interval we recall that set A' consists of m_0 smallest elements of A such that $S_{A'} \leq \frac{1}{c_3}S_A$. Thus, the length of the interval in the set of subset sums is at least $(1 - \frac{1}{c_3})S_A = O(S_A)$.

3.2. Elaborating on the structure of the set of subset sums. — To elaborate on the structure of the set of subset sums, we consider the case where condition (8) is not satisfied for the set A. We will show that in this situation we have an arithmetic progression (instead of an interval) belonging to the set of subset sums.

The situation when an arithmetic progression is obtained is unusual situation. It is characterized by the fact that many elements of A are divisible by the same integer Q.

To refine the structure of the set of subset sums we manipulate it with those elements of A which are not divisible by this integer Q, i.e., have non zero residues modulo Q.

The results from **[15]** and **[9]** are used in the presentation of the section.

Arithmetic progression. — If condition (8) is true for all q's, then an arithmetic progression with the difference $Q = 1$ beginning before $s = \frac{1}{2}S_A - \sigma_A$ and having length more than $h = 2\sigma_A \gg \ell$ belongs to the set of subset sums.

Assuming that (8) fails for some integer q, we construct a sequence of sets $A_0, \dots, A_p$ and a sequence of integers $q_0, \dots, q_p$ in the following way:

Assign $A_0 = A, q_0 = 1$, and assume that set A_i has already been found. Introduce also $q_i' = \prod_{j=0}^{i} q_j$. The integer q_{i+1} will be an integer such that

$$|A_i \setminus A_i(0, q_{i+1})| \leq \ell_{A_i}^{2/3} \log^{1/3} \ell_{A_i}. \tag{12}$$

If such an integer q_{i+1} exists, construct $A_{i+1} = \frac{A_i(0,q_{i+1})}{q_{i+1}} = \frac{A(0,q'_{i+1})}{q'_{i+1}}$; if such an integer q_{i+1} does not exist, set $p = i$ and $Q = q'_p$. This Q is not large, it is less than $\frac{3\ell}{2m}$ (see **[15]**).

Consider the set A_p obtained at the end of the process. It may be shown that for the set A_p, all the conditions of Theorem 3.1 are true. Therefore, apply Theorem 3.1 to A_p in order to arrive at

$$\left[\tfrac{1}{2}S_{A_p} - c_2\sigma_{A_p}, \tfrac{1}{2}S_{A_p} + c_2\sigma_{A_p}\right] \subseteq A_p^*.$$

Recalling that $A(0, Q) = \{aQ \mid a \in A_p\}$, we obtain a long segment of a progression with difference Q being contained in $(A(0, Q))^*$.

Refining the structure using residues. — In the previous paragraph it was shown that an arithmetic progression with a small difference ($Q \leq \frac{3\ell}{2m}$) belongs to the set of subset sums. Furthermore, these subset sums (elements of the arithmetic progression) may be constructed using only the elements of A that have zero residue modulo Q (the difference of the arithmetic progression). The next step is to try to use the remaining elements of A (with non-zero residues modulo Q) to refine the structure by "filling" the "holes" in the progression. To do this we need some properties of subset sums. In this survey we will only list these properties; the proofs may be found in **[15]**, **[9]**.

Properties of subset sums modulo integer q. — Consider ring $\mathbb{Z}_q$ of residues mod q. For $d \in \mathbb{Z}_q$, $d \mid q$, define $H_d = \{0, d, 2d, \dots, (\frac{q}{d} - 1)d\}$, and for $r \in \mathbb{Z}_q$ define $H_d(r) = r + H_d$.

(a) Let C be a set of elements of the ring $\mathbb{Z}_q$. If an element $b \in \mathbb{Z}_q$ is such that $C = C + \{0, b\}$, then the set C has the following structure: for each $r \in C$, we have $H_d(r) \subseteq C$, where $d = \gcd(b, q)$; i.e., $C = \bigcup_{r \in C} H_d(r)$.

(b) Let set $C \subset \mathbb{Z}_q$ have the following structure: $C = \bigcup_{r \in C} H_d(r)$ for some d, $d|q$. Then for any $b \in \mathbb{Z}_q$, the set $C + \{0, b\}$ has the same structure.

Refining the structure. — We continue from the following point: an integer $Q \leq 3\ell_A/2m_A$ is found, such that a long arithmetic progression with the difference Q belongs to $(A(0,Q))^*$. Let $A \setminus A(0,Q) = \{b_1, \dots, b_w\}$, and define a sequence of numbers $d_0, \dots, d_w$ in the following way.

Let $B_i = \{b_1, \dots, b_i\}$ and C_i be the set of the smallest non-negative residues modulo Q of $B_i^* \cup \{0\}$, i.e., $C_0 = \{0\}$ and $C_i = C_{i-1} + \{0, b_i\} (\operatorname{mod} Q)$. Let $d_0 = Q$. If the numbers $d_0, \dots, d_{i-1}$ have already been determined, take $d_i = d_{i-1}$ when $|C_i| > |C_{i-1}|$ and $d_i = \gcd(d_{i-1}, b_i)$ when $|C_i| = |C_{i-1}|$. In this way the numbers d_i and sets C_i possess property (a), i.e., for any $c \in C_i$ we have $H_{d_i}(c) \in C_i$. At the end of the process we obtain the set C of all non-zero residues modulo Q which may be represented by subset sums A^*. This set has the following structure: $C = \cup_{c \in C} H_{d_w}(c)$ where $d_w \leq Q$. Combining the set C with the previously obtained arithmetic progression we conclude that the structure of the set of subset sums may be characterized as *a collection of long arithmetic progressions with a common difference.*

Relaxing the condition of distribution. — Working with residues allows not only the refining of the structure as was shown above, it also provides the way to relax the condition (8) of distribution in Theorem 3.1. Indeed, looking on the structure of set C above, one can see that the result of the previous paragraph is a collection of arithmetic progressions with a difference $d_w \leq Q$. It might be shown that $d_w \neq 1$ only if $|A(0, d_w)| > m - d_w$. This means that condition (8) might be replaced by condition

$$|A(0,q)| \leq m - q \tag{13}$$

and we would still get a long interval belonging to A^*.

3.3. Reducing density. — There are a few ways to achieve an arithmetic progression in a set of subset sums for lower density, namely, for $m > c(\ell \log \ell)^{1/2}$.

Analytical approach

G. Freiman in **[17]** proves that the asymptotic formula (10) is still valid for the lower density if the elements of A are "well distributed". "Bad distribution" in his consideration means one of the following:
(1) there are too many small elements in A.
(2) there are too many elements in A divisible by one number q.
(3) there are too many elements in A belonging to a two-dimensional structure.
All these situations, where the asymptotic formula is not valid, are investigated separately and an arithmetic progression is constructed for each of them. The third case is of special interest because its analysis shows the possibility of future improvements. Let us outline main points of this thought.

Build injection $A \overset{\rho}{\to} \mathbb{Z}^2$. This map ρ transforms our one-dimensional problem into a two-dimensional one. (Recall that dense two-dimensional problems were solved in **[16]**.) G. Freiman shows that, in the case that the asymptotic formula does not work, there is a rectangle $H \subset \mathbb{Z}^2$ which contains images of most elements of A such that for this rectangle, the density condition of **[16]** holds. Thus, subset sums of these elements represent all integer points of a lattice – a two-dimensional analogue of an arithmetic progression. Now, transforming back to one dimension, we get a collection of short arithmetic progressions, the union of which forms a long one.

Finite addition approach

A. Sárkőzy arrives at an arithmetic progression for the same density ($m > c(\ell \log \ell)^{1/2}$) using a different approach. He proves a sequence of theorems that leads to the existence of an arithmetic progression (we formulate his result using the notation of this survey).

Theorem 3.2. — *Let $\ell > 2500$ and $|A| = m > 200(\ell \log \ell)^{1/2}$. Then there are integers d, y, z that*

$$1 \leq d < 10^4 \frac{\ell}{m}, \quad z > 7^{-1} \cdot 10^{-4} m^2, \quad y < 7 \cdot 10^4 \frac{\ell}{m^2} z,$$

and

$$\{M : M \equiv 0 (\operatorname{mod} d), yd \leq M \leq zd\} \subseteq A^*.$$

A. Sárkőzy (**[22]**) shows that this theorem is the best possible apart of the constants and a logarithmic factor in the density constraint. However, the proof of this theorem does not lead to an explicit way of calculating a difference d of an arithmetic progression for a specific instance of a set A.

Algorithmic approach

Z. Galil and O. Margalit (**[18]**) obtain almost the same density ($m > c\ell^{1/2} \log \ell$) while explicitly constructing a progression. We will discuss this approach in the next section whilst explaining their algorithm.

4. Algorithms for the Subset-Sum Problem based on the structural characterization

This section is dedicated to algorithms for solving SSP. The first algorithm using the structural approach is due to G. Freiman **[14]**. Using structural characterization from **[1]** (density $m > \ell^{2/3+\varepsilon}$) this algorithm solves SSP (finding the maximal sum but not the subset) in $O(\frac{\ell^{5/3}}{m} + m \log^2 m)$. In **[15]** G. Freiman improved this algorithm obtaining a linear time algorithm for the same density of problems. This algorithm also works perfectly for lower density (up to $m > c(\ell \log \ell)^{1/2})$) but then it is not linear. Its time grows and becomes $O(m^2 / \log m)$ for the lowest density. This algorithm was improved by M. Chaimovich (see **[5]**, **[9]**) using the same idea but more complicated technique for verifying the divisibility of the summands.

Z. Galil and O. Margalit **[18]** use another technique. Their algorithm finds both the maximal sum S_B and the subset B. Its time is $O(m)$ for the high density ($m > c\ell^{3/4} \log \ell$) and $O(\ell \log \ell) = O(m^2 / \log m)$ for the lower one ($m \sim \ell^{1/2} \log \ell$). Moreover, this algorithm provides an elementary proof of the structural characterization theorem by explicit construction of the desired structure.

In this survey two algorithms are presented. The first of them ([**15**]) is based on the analytical theorem. We discuss also the methods that were used in [**5**] and [**9**] in order to improve the algorithm. The second algorithm is created by Z. Galil and O. Margalit [**18**]. We will only present descriptions of the algorithms and their estimated complexities (for detailed proofs the reader may refer to cited articles).

4.1. Algorithm for finding the maximal subset sum S_B. — The main idea of the first algorithm is to find the difference Q of the arithmetic progression in the set of subset sums. Based on the analytical theorem (as in Theorem 3.1), finding this difference requires verification of a condition similar to condition (12). Verification is done in the same way as was explained on page 349. In Algorithm 1 condition (14) is used. It may be shown that only prime numbers $q < 3\ell/2m$ must be verified. Once Q is found, elements of A with non-zero residues modulo this Q allow one to "fulfill" the "holes" in the progression and to complete the construction of the structure.

The algorithm does not require that all summands are different but that the amount of the different ones is large enough (see [**5**]). In the algorithm below, the number of different elements in a multi-set X is denoted by $\overline{m}_X$. Recall also that N is the target number and z is the maximal subset sum that does not exceed N.

Algorithm 1

```
1. Finding $Q$.
    (a) Initialization. $q_0 \leftarrow 1, A_0 \leftarrow A, t_0 \leftarrow \lfloor \frac{3\ell_A}{2\overline{m}_A} \rfloor, i \leftarrow 0$.
    (b) Find the smallest prime number $q_{i+1}$ such that $2 \leq q_{i+1} \leq t_i$ and

(14)                $|A_i \setminus A_i(0, q_{i+1})| < t_i$.

        If such a number $q_{i+1}$ exists, compute $A_{i+1} \leftarrow \frac{A_i(0,q_{i+1})}{q_{i+1}}$, $t_{i+1} \leftarrow \lfloor \frac{3\ell_{A_{i+1}}}{2\overline{m}_{A_{i+1}}} \rfloor$
        and continue to next $i (i \leftarrow i+1)$.
    (c) If such a number $q_{i+1}$ does not exist, set $p \leftarrow i$ and compute $Q \leftarrow \prod_{j=0}^{p} q_j$.
        If $Q = 1$, then go to step 3.
2. Finding $C$. Let $G = \{b_1, \dots, b_k\} = A \backslash A(0, Q)$.
    (a) Initialization. $d_0 \leftarrow Q, C_0 \leftarrow \{0\}, i \leftarrow 0$.
    (b) Computing $C_{i+1}$ and $d_{i+1}$.
        If $b_{i+1}$ is divisible by $d_i$, then $C_{i+1} \leftarrow C_i$ and $d_{i+1} \leftarrow d_i$.
        Otherwise, compute $C_{i+1}$ explicitly
        $(C_{i+1} = C_i + \{0, b_{i+1}\} (\mathrm{mod}\, Q) = \{s \mid s \in C_i \text{ or } s - b_{i+1} (\mathrm{mod}\, Q) \in C_i\})$.
        If $|C_{i+1}| = |C_i|$ then $d_{i+1} \leftarrow \gcd(d_i, b_{i+1})$; otherwise, set $d_{i+1} \leftarrow d_i$.
        If $|C_{i+1}| = Q$ or $i = k$ go to step 3; otherwise, continue to next $i$ $(i \leftarrow i+1)$.
3. If $Q = 1$ or $|C| = Q$ then $z = \lfloor N \rfloor$, otherwise compute $r = N - \lfloor \frac{N}{Q} \rfloor \cdot Q$ and
   $z = N - r + \max\{r_i \in C | r_i \leq r\}$.
```

The complexity of Algorithm 1 is $O((\frac{\ell}{\overline{m}})^2 + m \log^2 m)$ which is $O(m \log^2 m)$ for $\ell = O(\overline{m}^{3/2} \log \overline{m})$.

To improve the time boundary of Algorithm 1 to $O(m \log m)$ M. Chaimovich uses condition $|A_i \setminus A_i(0, q_{i+1})| < \left(\prod_{q \leq p \leq (11\ell)/(8m), p \textit{ is prime}} \frac{p}{p-1}\right) \cdot \frac{11\ell}{8m}$ (see **[5]**) and condition $|A_i \setminus A_i(0, q_{i+1})| < \frac{t_i^{5/3}}{q_{i+1}^{2/3}}$ (see **[9]**) instead of (14). In both versions gain (comparing with Algorithm 1) is achieved owing to the fact that as soon as prime number q is verified, there is no need to return and to check it again.

To obtain linear complexity (for slightly higher density) the following modification may be used:

Algorithm 1A

1. Let A' be a multi-set consisting of the $\overline{m}_{A'} = \frac{\sqrt{8\overline{m}}}{\log^2 \overline{m}}$ first different elements of A. Find a number Q applying the process from step 1 of Algorithm 1 to set A'.
2. Execute steps 2 and 3 of Algorithm 1.

4.2. Algorithm for finding the optimal subset (Z. Galil and O. Margalit)

Z. Galil and O. Margalit **[18]** solve the SSP by constructing a long arithmetic progression belonging to the set of subset sums. To do this they partition the input set A into three parts: $A = A_1 \cup A_2 \cup A_3$. First, they construct A_1 – a small set satisfying $(A_1)^* (\operatorname{mod} d) = A^* (\operatorname{mod} d)$ for every small enough integer d. Set A_2 consists of a number of the smallest elements of $A \setminus A_1$. These elements are used to construct the segment of the progression longer than ℓ. A_3 contains the remaining elements of A. They are used to extend the progression.

The algorithm is based on two main processes. The first of them reduces the problem to the case where $A^* (\operatorname{mod} d) = [0, d)$ for every small enough integer d. This constitutes Step 1 of Algorithm 2. Logically this process is similar to the first step of Algorithm 1 and results in the number d_0 such that the set $A' = \frac{A(0,d_0)}{d_0}$ possesses the above mentioned property. The technique used in this algorithm is different than the one used in Algorithm 1. It allows us to obtain the linear time boundary. The same method is employed again in Step 3 when we apply it to set $A'_{(m/4)}$ – the set of $\frac{m}{4}$ smallest elements of A' – in order to construct the subset A'_1.

The second process, used in Algorithm 2, provides a way to construct an arithmetic progression belonging to the set of subset sums. This constitutes Step 5 of Algorithm 2. This process is based on the following simple consideration: Given a set A of μ distinct integers in an interval of length $\lambda \leq \ell$, consider the sets P_i of pairs with difference i, i.e., $P_i = \{(a, b) \in A \times A | a - b = i\}$. There are $\Omega(\mu^2)$ pairs (a, b) and thus (by pigeon-hole argument) there are many pairs with the same difference. We first take many P_i's that contain large enough number of pairs. Taking $k - j$ pairs from P_ρ and j pairs from P_σ gives a sequence of subsets $D_j \subset A$ with $S_{D_j} = k\rho + j(\sigma - \rho)$ - an arithmetic progression. (Observe that the pairs in each P_i are disjointed, but the pairs in different P_i's may intersect. So, only some of the pairs from P_i's may be used in our construction, some pairs have to be "deleted" in order to restore the disjointness property.)

The arithmetic progressions generated this way are still too short, but it is possible to generate many of them and then combine them in order to create a longer arithmetic progression. Starting with the progression of minimal difference, $(i+1)$-st progression is inductively combined with the previously obtained arithmetic progression of the first i progressions. An element of a combined progression is the sum of an appropriate element from each of the two progressions.

As the full description of the process is quite complicated we will omit it in this survey. Thus, we are ready to outline steps of the algorithm.

Algorithm 2

1. Let $t = \lceil \frac{103\ell}{m} \log_2 \ell \rceil$. $A_{(i)}$ denotes the set of i smallest elements of A and q stands for a power of prime. We find an integer $d_0 < t$ such that $|A \setminus A(0, d_0)| < d_0$ and $|A' \setminus A'(0, d)| > d$ for each $d < t$ where $A' = A(0, d_0)/d_0$.
   (a) Compute $G_1 = \{q : 1 < q < t, |A_{(2t)} \setminus A_{(2t)}(0, q)| < t\}$ by verifying all prime powers from 1 to t.
   (b) Compute $G_2 = \{q : q \in G_1, |A_{(b)} \setminus A_{(b)}(0, q)| < t\}$, where $b = 3t \log_2 \ell$ by verifying all elements of G_1.
   (c) Compute $G_3 = \{q : q \in G_2, |A \setminus A(0, q)| < t\}$ by computing $|A(i, \mathrm{lcm}(G_2))|$ for all $i \in [0, \mathrm{lcm}(G_2))$ and using elements of G_2 as candidates for G_3.
   (d) Compute $G_4 = \{d : 1 < d < t, d \equiv 0(\mathrm{mod}(\mathrm{lcm}(G_3))), |A \setminus A(0, d)| < d\}$ using elements of G_3 as candidates for G_4.
   (e) Compute $d_0 = \max(G_4)$.
2. Use dynamic programming modulo d_0 to compute $A^*(\bmod d_0)$. In computing the set $A^*(\bmod d_0)$ keep a subset $C_i \subseteq A \setminus A(0, d_0)$ for each $i \in A^*(\bmod d_0)$ such that $S_{C_i} \equiv i(\bmod d_0)$ and $S_{C_i} < \ell d_0$. Also compute the function $f_{d_0}(i) = \max\{j | 0 \le j \le i \; and \; j \in A^*(\bmod d_0)\}$. (The use of this function will be clarified in step 9.)
3. Reduce the problem to another one by taking $A' = \frac{A(0, d_0)}{d_0}$ instead of A. Apply sub-steps (a)-(d) of Step 1 of the algorithm to $A'_{(m/4)}$ (the first smallest $\frac{m}{4}$ elements of A') and construct $A'_1 = A'_{(m/4)} \cup (\cup_{d \in G'_4} C'_d)$ where G'_4 is the set obtained in sub-step (d) of the second application of Step 1 and C'_d are d elements from $A' \setminus A'(0, d)$.
4. Defining $\lambda = \lceil 64 \ell^{1/2} \log_2 \ell \frac{4S_{A'}}{m^2} \rceil$ and $\mu = \lceil 15 \ell^{1/2} \log_2 \ell \rceil$ choose $A'_2 \subseteq A' \setminus A'_1$ which contains μ elements where each one is less than $\frac{4S_{A'}}{m}$ and lies in a sub-interval of length λ. This is done by taking elements of $A' \setminus A'_1$ smaller than $\frac{4S_{A'}}{m}$, splitting them into sub-intervals of length λ and choosing the most dense sub-interval.
5. Using the elements of A'_2 obtain the sequence of subsets $\{B'_i\}_{i=0}^{2\ell-1}$ such that their sums form an arithmetic progression with a small difference – $S_{B'_i} = s_0 + i g_r$, $g_r < t$. (A detailed description of the process may be found in section 4 of **[18]**.)
6. Using dynamic programming, build a sequence $\{E'_i\}_{i=0}^{g_r - 1}$ of subsets of A'_1 such that $S_{E'_i} \equiv i(\bmod g_r)$ and $S_{E'_i} < \ell' g_r$.
7. Construct sets $F'_i = E'_{i(\bmod g_r)} + B'_j$, where $j = \ell' + (i - S_{E'_{i(\bmod g_r)}})/g_r$, for $0 \le i < \ell g_r$. (Note that $S_{F'_i} = s_0 + \ell' g_r + i$.)
8. Compute all the prefix sums of the set $A'_3 = A' \setminus (A'_1 \cup A'_2)$.
9. Given a target number N, the following sub-steps are executed.

(a) Denoting $r_0 = f_{d_0}(N(\operatorname{mod} d_0))$ (for definition of $f_{d_0}(i)$ see step 2), compute $S_B = d_0\lfloor N/d_0 \rfloor + r_0$.
(b) Compute $N' = (S_B - S_{C_{r_0}})/d_0 - s_0 - \ell' g_r$.
(c) Using a binary search on the set of prefix sums of A'_3 find

$$n = \max\{i | S_{A'_{3(i)}} \leq N'\}.$$

(d) The desired subset is $B_? = C_{r_0} \cup d_0 A'_{3(n)} \cup d_0 F'_{N' - S_{A'_{3(n)}}}$.

Observe that the first eight steps of the algorithm are preprocessing steps that may be performed only once in the case that SSP is solved many times for the same set A and different target numbers.

As mentioned before, Algorithm 2 finds the maximal sum and the optimal subset in $O(m + (\frac{\ell}{m}\log\ell)^2 + \frac{S_A \ell^{1/2}}{m^2}\log^2\ell)$ time. (Note that the last term of the expression is required for finding the optimal subset only.) This gives $O(m)$ time for $m > c\ell^{3/4}\log\ell$ and $O(\ell\log\ell)$ time for $m \sim c\ell^{1/2}\log\ell$.

5. Application of an analytical structural approach to other IP models

5.1. Value-Independent Knapsack. — The Value-Independent Knapsack Problem (VIKP) is IP problem of the form: maximize $z = \sum_{i=1}^{r} a_i x_i$ subject to $\sum_{i=1}^{r} a_i x_i \leq N$, where $0 \leq x_i \leq n_i$, $x_i \in \mathbb{Z}, i = 1, 2, \ldots, r$, and all coefficients are integers (see [**2**]). One can reformulate the VIKP as the SSP with a multi-set A, containing element a_i exactly n_i times (for each i, $1 \leq i \leq r$), and a target number N.

In view of the fact that structural analysis of the SSP was done assuming that the elements are distinct ([**15**], [**18**]) or assuming that the number of distinct elements is sufficiently large ([**5**]), the VIKP requires its own structural analysis. This analysis was done in [**4**] and [**6**] proving the structural characterization of the VIKP for $\ell = O(\frac{m^{3/4}}{\log m})$ and $\ell = O(\frac{m}{\log m})$ respectively. In this survey we formulate the structural characterization as it was done in [**6**] and give a short sketch of the algorithm presented there.

Structural Characterization. — For convenience, we will view A as at a set of pairs of positive integers such that the elements of the pair are an integer and the number of its appearances in A respectively. Thus, we write $A = \{(a'_i, n_i) \mid 1 \leq i \leq r\}$, where $\{a'_1, \ldots, a'_r\}$ is the set of distinct elements of A and n_i is the numbers of appearances of a'_i in A. Define also $t = \max\{n_i \mid 1 \leq i \leq r\}$.

Using this notation, the existence of an arithmetic progression in the set of subset sums was proved in [**6**] for $m \geq \min\{6\ell\log\ell, 9(\ell t)^{2/3}\log^{1/3}(\ell t)\}$.

Indeed, the estimation $m > 6\ell\log\ell$ is the best possible apart from a logarithmic factor and a constant: Let $A = \{(\ell + 1 - i, t) \mid 1 \leq i \leq r\}$ for some integers ℓ, t, r. Clearly, $m = |A| = rt$. A^* consists of rt disjoint intervals (each of which is not longer than $\frac{r^2 t}{2}$) whenever $\frac{r^2 t}{2} \leq \ell - r$, i.e., $m^2 + m \leq \ell t$. We therefore do not have a long arithmetic progression for $m \asymp (\ell t)^{1/2}$ and for $m \asymp \ell$ when $t \asymp \ell$.

Sketch of the algorithm. — One can see two important parts in each of the algorithms (see Algorithm 1) based on the new approach. The first part finds the difference Q of an arithmetic progression in A^*, and the second explicitly constructs subset sums with non-zero residues mod Q, i.e., $A^*(\bmod Q)$. The same is true for VIKP. In this survey only the main considerations of the algorithm will be presented. Details can be found in [**6**] or in [**9**].

The step that finds the difference of the progression employs two ideas in order to reduce the number of operations of the algorithm. First, the elements of A are grouped in order to present A as a list of pairs (a', n) - an element and the number of its appearances in A. These pairs are sorted such that the most frequent elements appear first.

Second, the difference Q of the progression is found using three different methods depending on the number r of distinct summands in A. If this number is large enough, the method similar to Step 1 of Algorithm 1 is used. Otherwise, Q is determined as the greatest common divisor of the k most frequent elements of A (elements that appear more than $\frac{2\ell}{k} + 1$ times each).

Construction of the subset sums with non-zero residues modulo Q is done using the same technique as in Algorithm 1.

Precise analysis of the steps of the algorithm shows that it carries out the solution in $O(Qr_Q + m)$ time, where r_Q is the number of different residues mod Q of A. In the worst scenario, the first term of the expression dominates and, taking into account that the number of different residues mod Q is limited by the number of different elements of A, we have a $O(\ell^{3/2} \log^{1/2} \ell + m)$ time algorithm.

However, the algorithm becomes linear if (a)$r > c(\ell \log \ell)^{1/2}$ or $r = O(\frac{m}{\ell})$; (b) $k \asymp r$, where k is the number used for calculation of Q (this condition means that there are not many elements with very small number of appearances implying $Q = O(\ell^{1/2})$). Therefore, linear time is not achieved for the following special case: The number of different elements of A is neither large nor very small and all elements with a large number of appearances belongs to one arithmetic progression with a sufficiently large difference. The number of these elements is extremely small in relation to the number of elements with a small number of appearances.

5.2. Multi-dimensional Subset-Sum Problem. — This paragraph is concerned with the multi-dimensional Subset-Sum problem which is a particular case of the multi-dimensional Knapsack Problem. Recall its definition ([**8**]): Let $\mathcal{A}$ be a set of n-dimensional non-zero integral vectors taken from the convex hull $\mathcal{D}$, i.e.,

$$\mathcal{A} = \{\overline{a}_i = (a_{1i}, \dots, a_{ni})^t\}_{i=1}^m \subseteq ((\mathbb{Z}^n \cap \mathcal{D}) \setminus \{\overline{0}\}).$$

(The notation $(\cdot)^t$ means the transpose of a vector $(\cdot)$, i.e., $\overline{a}_i$'s are viewed as column-vectors.) The problem is: for the given target vector $\overline{b} \in \mathbb{N}^n$, find the vector $\overline{z} \in \mathcal{A}^*$ satisfying $\overline{z} \leq \overline{b}$ and having maximal length, where a partial order on n-dimensional vector space is defined in any appropriate way.

The two-dimensional SSP was investigated by G. Freiman in [**16**]. It has been found that in this case a lattice becomes a basic element of the structure and takes the place of an arithmetic progression – a basic element of the structure in the one-dimensional case. Further, this result was extended to n dimensions by M. Chaimovich [**8**].

Two-dimensional SSP. — Let $\Gamma_U = \{\overline{v} \mid \overline{v} = k_1\overline{u}_1 + k_2\overline{u}_2, k_j \in \mathbb{Z}, \overline{u}_j \in U\}$ denote the lattice generated by the set $U = \{\overline{u}_1, \overline{u}_2\} \subseteq \mathbb{Z}^2$ of linearly independent integral vectors. Hereafter the subscript is omitted whenever it is clear from the context which lattice is being considered and let V_Γ denote the number of integer points in the fundamental parallelogram of Γ.

Two vectors $\overline{v}_1, \overline{v}_2$ are congruent modulo Γ (written as $\overline{v}_1 \equiv \overline{v}_2 (\operatorname{mod} \Gamma)$) if $\overline{v}_1 - \overline{v}_2 \in \Gamma$. Two sets are congruent modulo Γ (written $\mathcal{A}_1 \equiv \mathcal{A}_2 (\operatorname{mod} \Gamma)$) if for each vector $\overline{v}_1 \in \mathcal{A}_1$ there is a vector $\overline{v}_2 \in \mathcal{A}_2$ congruent to $\overline{v}_1$ and inversely for each vector $\overline{v}_2 \in \mathcal{A}_2$ there is a vector $\overline{v}_1 \in \mathcal{A}_1$ congruent to $\overline{v}_2$. In addition, $\mathcal{A}(\Gamma) = \mathcal{A} \cap \Gamma$, and $\overline{b} \in \mathcal{A}(\operatorname{mod} \Gamma)$ means that there is $\overline{v} \in \mathcal{A}$ congruent to $\overline{b}$.

For a given $\mathcal{A}$ define $B_j^2 = \frac{1}{4}\sum_{i=1}^m a_{ji}^2$, $j = 1,2$; $B_{12} = \frac{1}{4}\sum_{i=1}^m a_{1i}a_{2i}$.

Using this notation the following theorem (Theorem 2 [**16**]) gives a structural characterization for a two-dimensional case.

Theorem 5.1. — *Let $\mathcal{A} \subseteq D \cap \mathbb{Z}^2$ be a set of two-dimensional integral vectors where D is a convex set with $|D \cap \mathbb{Z}^2| = \ell$, $|\mathcal{A}| \geq c_1\ell^{2/3}\log^{1/3}\ell$, $\ell > \ell_0$. Suppose that for each line "a" containing zero*

$$|\mathcal{A} \cap a| \leq \tfrac{1}{2}|\mathcal{A}|. \tag{15}$$

Then (i) there is the lattice Γ_0 with $V_{\Gamma_0} = O(\frac{\ell}{m})$ and the subset $H \subseteq \mathcal{A}$ such that $|H| \leq V_{\Gamma_0}$ and $\mathcal{A}^* \equiv H^*(\operatorname{mod} \Gamma_0)$ and (ii) for convex hull $\mathcal{F}$ defined by

$$\mathcal{F} = \{\overline{v} : |(\tfrac{1}{2}S_{\mathcal{A}} - \overline{v})^t \begin{pmatrix} B_1^2 & B_{12} \\ B_{12} & B_2^2 \end{pmatrix}^{-1} (\tfrac{1}{2}S_{\mathcal{A}} - \overline{v})| \leq c_2\},$$

vector $\overline{b}$ belongs to $\mathcal{A}^* \cap \mathcal{F}$ if and only if $\overline{b} \in \mathcal{F}$ and $\overline{b} \equiv H^*(\operatorname{mod} \Gamma_0)$.

According to this theorem, the structural characterization of the set of two-dimensional subset sums is quite simple: a collection of all points from certain classes of residues modulo lattice (including zero residue class) within a two-dimensional convex hull in the wide vicinity of the mid-point $\frac{1}{2}S_{\mathcal{A}}$.

The proof of this result is too complicated to be presented in this survey. First of all, the case where all vectors are taken from the rectangle with edges parallel to axes is investigated and for this case the asymptotic formula for the number of representations by the set of two-dimensional subset sums is obtained. In this step the condition for validity of the asymptotic formula is determined: not too many vectors may belong to a one integer lattice.

Further, this result is extended replacing a rectangle by an arbitrary convex set $\mathcal{D}$. This is done by applying to the set $\mathcal{D}$ a certain transformation which is invariant with regard to the integer lattice. In addition, the image of $\mathcal{D}$ is contained in the rectangle and the number of integer points in this rectangle is of the same order as in $\mathcal{D}$.

Finally, the case where the asymptotic formula is not true or, in other words, where most of the vectors belong to some lattice Γ is considered. This is done in the similar way as in the one-dimensional case.

Observe that condition (15) is crucial for obtaining two-dimensional structural characterization. If this condition is not satisfied, the problem is actually one-dimensional because most of its vectors lie on one line.

The last step of the proof provides a simple algorithmic way to construct the structure and to find the solution to the problem. Precise analysis of the algorithm (not presented here) shows that its time boundary is $O(m^2 \log m)$. For very dense problems ($\ell = O(m)$), the time boundary of the new algorithm is more impressive. It is $O(m \log^2 m)$ – almost linear.

n-dimensional SSP. — Analysis of the n-dimensional SSP is quite similar to the two-dimensional case. The difficulty in the generalization lies in the complexity of the geometry of an arbitrary number of dimensions compared to the geometry of two dimensions. However, the structural characterization of the set of n-dimensional subset sums, explicitly determined by the algorithm, seems to be quite simple: it consists of a collection of all points with certain classes of residues modulo lattice within an n-dimensional convex body.

The density condition for the n-dimensional case is $m \geq (n\ell^{n-1} \log \ell)^{1/n}$, $n > 2$, and the time boundary of the algorithm becomes $O(m^{2+1/(n-1)} \log^n m)$ or even $O(m \log^n m)$ for very dense problems ($\ell = O(m)$).

5.3. The k-Partition Problem. — A structural approach for solving the k-partition problem (KPP) was studied in [7] (see page 346 for problem definition). Although the proposed method works for a wide spectrum of objective functions, the author chooses as an objective function the function $z = \max_j \frac{S_{B_j}}{N_j}$. Under this objective function the problem can also be viewed as a problem of scheduling independent tasks on uniform machines so as to minimize an end (make-span) time (see [21] for scheduling problem definition).

The solution is based on the reduction of the k-partition problem to a sequence of dense SSP and on the structural characterization of SSP by a collection of arithmetic progressions. As a result, the proposed algorithm solves the problem in $O(k\ell \log \ell)$ time which is considerably faster than previously known polynomial algorithms (dynamic programming, [21]) that achieved $O(m^{2k-1}\ell)$ time only.

In this survey general concepts of the reduction process and of the algorithm are presented.

General concepts. — Let $z^* = \min z$ be a value of the objective function for the optimal partition $(B'_1, \ldots, B'_k)$, i.e., $z^* = \max_j \frac{S_{B'_j}}{N_j}$. If $z^* = 1$ we say that KPP is *exactly solvable.* This is equivalent to the existence of a partition $(B_1, \ldots, B_k)$ with $S_{B_j} = N_j$ for all j, $1 \leq j \leq k$.

Suppose that a KPP $(A, N_1, \dots, N_k)$ has an exact solution. Consider the sequence of the following $(k-1)$ Subset-Sum problems:

$$(A, N_1, S_A - N_1), (A \backslash B_1, N_2, S_{A\backslash B_1} - N_2), \dots,$$
$$(A \backslash \bigcup_{i=1}^{k-2} B_i, N_{k-1}, S_{A\backslash \cup_{i=1}^{k-2} B_i} - N_{k-1}),$$

where B_i is some solution of the i-th SSP. Assuming that the first SSP is already solved and $S_{B_1} = N_1$, it is not necessarily true that the remaining SSPs are still exactly solvable. This is because elements which are necessary to find an exact solution for the second SSP could already have been used in B_1 – the solution for the first problem. In other words, the solution B_1, which was chosen from the set of all possible exact solutions of SSP, can be "bad"; thus the rest of SSP will not be exactly solvable.

To overcome this difficulty, a certain subset $C \subset A$, for which SSP $(C, N_1, S_C - N_1)$ has an exact solution, is defined. This subset is created such that selection of any one of these solutions ensures the existence of exact solutions for all subsequent problems. In that way, KPP can be replaced by solving a sequence of SSPs.

Some conditions must be imposed on multi-set A in order to ensure successful application of this method. Recall that an exact solution of SSP in a wide interval of target numbers N is ensured by condition (8), i.e., we have "many" non-zero residues for each modulo q. To solve KPP, it is natural to strengthen condition (8), requiring as many non-zero residues for each modulo q as we need for exact solvability of all $(k-1)$ SSP: the condition (8) becomes

$$|A \setminus A(q)| \geq (k-1)^2 \frac{4\ell_A}{\overline{m}_A} \log_2 2\ell_A, \tag{16}$$

where $\overline{m}_A$ again stands for the number of different elements of A.

Indeed, multi-set C, mentioned above, and from which subset B_1 is chosen, includes the amount of non-zero residues needed to solve one problem only, leaving the rest to be used when solving subsequent problems.

In fact, in addition to condition (16) the density relation

$$\overline{m} > c_1(k(k-1)\ell \log \ell)^{2/3}, \tag{17}$$

must be imposed on A in order to ensure the possibility of the reduction. Condition (17) and the trivial inequality $\ell \geq \overline{m}$ restrict the values of k for which the class of KPP, solved using the above method, is not empty. Namely, $k < \frac{\ell^{1/4}}{c_1^{3/2} \log^{1/2} \ell}$.

The situation where condition (16) fails for some integer q, can be viewed in a similar way to the way it was handled in the case of SSP. It can be shown that there exists an integer q_0 such that multi-set $A' = \frac{A(0,q_0)}{q_0}$ satisfies conditions similar to conditions (17) and (16), and that KPP $(A(0, q_0), N_1', \dots, N_k')$, where $q_0 | N_i'$, has an exact solution.

Introduce the set Q_{q_0} of k-tuples $(s_1, \dots, s_k)$ of residues modulo q_0 which can be represented by $A \backslash A(0, q_0)$: for each $(s_1, \dots, s_k) \in Q_{q_0}$ there is a partition $(G_1, \dots, G_k)$ of $A \backslash A(0, q_0)$ such that $s_j \equiv S_{G_j} (\operatorname{mod} q_0)$, $0 \leq s_j < q_0$.

Combining the solution of the above mentioned KPP $(A(0,q_0),N'_1,\dots,N'_k)$ and k-tuples from Q_{q_0}, it may be concluded that the original KPP is exactly solvable if and only if there is a k-tuple $(n_1,\dots,n_k) \in Q_{q_0}$ such that $n_j \equiv N_j (\mathrm{mod}\, q_0), 0 \le n_j < q_0, 1 \le j \le k$.

Thus, to determine if KPP is exactly solvable, it is sufficient to find q_0, and also to verify that the k-tuple of residues of target numbers modulo q_0 can be represented by a partition of $A \setminus A(0,q_0)$.

Finally, it is necessary to describe how to find the solution of a KPP, that is not exactly solvable. For each partition $(B_1,\dots,B_k)$ of A, we have $S_{B_j} = N_j + d_j$ and $z = \max_j \frac{S_{B_j}}{N_j} = \max_j (1 + \frac{d_j}{N_j})$. The goal is to find a certain set of deviations $\{d^*_j\}$ which will minimize the objective function z.

For $(s_1,\dots,s_k) \in Q_{q_0}$ define

$$(18)\quad z(s_1,\dots,s_k) = \min\left\{ \max_j (1+\frac{d'_j}{N_j}) \,\middle|\, \sum_{j=1}^{k} d'_j = 0, d'_j + N_j \equiv s_j (\mathrm{mod}\, q_0), 1 \le j \le k, \right\}$$

where the minimum of the function is taken over all possible sets $\{d'_j\}$. It is shown in [**7**] that an optimal set of deviations $\{d^*_j\}$ is the set which minimizes this function $z^* = \min\{z(s_1,\dots,s_k) \mid (s_1,\dots,s_k) \in Q_{q_0}\}$.

Once we have this set of deviations, we obtain the KPP $(A, N_1 + d^*_1,\dots,N_k + d^*_k)$ which has an exact solution. To solve this problem, we construct a new problem $(A', N'_1,\dots,N'_k)$ where $A' = \frac{A(0,q_0)}{q_0}$ and $N'_j = \frac{N_j + d^*_j - S_{G_j}}{q_0}$ and solve it by solving $k-1$ subsequent SSPs. (Algorithm 2 from page 354 may be used to solve each SSP.)

Let $(G'_1,\dots,G'_k)$ be a solution of this KPP. Then $(G_1 \cup q_0 G'_1, \dots, G_k \cup q_0 G'_k)$ is a solution of the original KPP, since $S_{G_j \cup q_0 G'_j} = S_{G_j} + q_0 N'_j = N_j + d^*_j$.

The complexity of the algorithm. — The complexity of the algorithm is evaluated (details can be found in [**7**]) as

$$(19)\qquad O(k^3 \frac{\ell}{\overline{m}} q_0^{k-1} \log \ell + k\ell \log \ell).$$

Indeed, for $q_0 = O(\frac{\ell}{\overline{m}})$ and $\ell = O(\frac{m^{3/2}}{k^2 \log \overline{m}})$ (see (17)), the first term in (19) dominates the second one and the time is $o(\overline{m}^{k/2})$. However, in the case where $q_0 < \left(\frac{\overline{m}}{k^2}\right)^{1/(k-1)}$, the dominant term in (19) is the second term and we obtain an almost linear time algorithm with $O(k\ell \log \ell)$ time. This time remains polynomial, even if k increases with ℓ. Observe also that $q_0 < \left(\frac{\overline{m}}{k^2}\right)^{1/(k-1)}$ is always satisfied for a dense 3-partition problem ($k = 3$) and for problems with sufficiently high density, namely, for $\overline{m} > k^{2/k} \ell^{1-1/k}$.

6. Conclusion

There are several other directions which deserve to be explored in order to proceed with this new approach.

One of them is to study structural characterization of subset sums for problems with lower density. G. Freiman conjectures that analytical techniques can be refined to handle the case $\ell = O(m^c)$ for any constant c.

This characterization will allow us to derive new algorithms. Recall that the structural approach, contrary to classical methods, works for problems with a large number of variables. There is, therefore, a gap between an upper boundary of classical algorithms and a lower boundary of the existing algorithms based on the new approach. The purpose of an algorithmic design, from the operational point of view, is to overlap this gap. From the theoretical point of view, the future algorithms will allow us to verify the conjecture: is it true that certain IP problems that are NP-hard have a less than exponential time solution for dense instances?

The other direction of the development is to analyze additional IP problems and to extend new methods to them. These efforts can proceed in two ways. One is to work directly on other specific problems and try to characterize their structure. The other is to reduce a problem to the SSP (one or multi-dimensional) as was done for the k-partition problem. In order to do this we need density-preserving reductions that yield instances of the SSP that are sufficiently dense.

References

[1] Alon N. and Freiman G.A., *On Sums of Subsets of a Set of Integers*, Combinatorica, **8**, 1988, 305–314.

[2] Balas E. and Zemel E., *An Algorithm for Large Zero-One Knapsack Problems*, Operations Research, **28**, 1980, 1130–1154.

[3] Buzytsky P. and Freiman G.A., *Analytical Methods in Integer Programming*, Moscow, ZEMJ., (Russian), 48 pp., 1980.

[4] Chaimovich M., *An Efficient Algorithm for the Subset-Sum Problem*, a manuscript, 1988.

[5] Chaimovich M., *Subset-Sum Problems with Different Summands: Computation*, Discrete Applied Mathematics, **27**, 1990, 277–282.

[6] Chaimovich M., *Solving a Value-Independent Knapsack Problem with the Use of Methods of Additive Number Theory*, Congressus Numerantium, **72**, 1990, 115–123.

[7] Chaimovich M., *Fast Exact and Approximate Algorithms for k-Partition and Scheduling Independent Tasks*, Discrete Mathematics, **114**, 1993, 87–103.

[8] Chaimovich M., *On Solving Dense n-Dimensional Subset-Sum Problem*, Congressus Numerantium, **84**, 1992, 41–50.

[9] Chaimovich M., *Analytical Methods of Number Theory in Integer Programming*, Ph. D. Thesis, Tel-Aviv University, Israel, 1991.

[10] Chaimovich M., Freiman G. and Galil Z., *Solving Dense Subset-Sum Problem by Using Analytical Number Theory*, J. of Complexity, **5**, 1989, 271–282.

[11] Erdős P. and Freiman G., *On Two Additive Problems*, J. Number Theory, **34**, 1990, 1–12.

[12] Freiman G.A., *An Analytical Method of Analysis of Linear Boolean Equations*, Ann. New York Acad. Sci., **337**, 1980, 97–102.

[13] Freiman G.A., *What is the Structure of K if $K + K$ is Small?*, in Lecture Notes in Mathematics, **1240**, 1987, 109–134.

[14] Freiman G.A., *On Extremal Additive Problems of Paul Erdős*, ARS Combinatoria, **26B**, 1988, 93–114.

[15] Freiman G.A., *Subset-Sum Problem with Different Summands*, Congressus Numerantium, **70**, 1990, 207–215.

[16] Freiman G.A., *On Solvability of a System of Two Boolean Linear Equations*, The Proceedings of the Number Theory Conference, New York, 1989.

[17] Freiman G.A., *New Analytical Results in Subset-Sum Problem*, Discrete Mathematics, **114**, 1993, 205–218.

[18] Galil Z. and Margalit O., *An Almost Linear-Time Algorithm for the Dense Subset-Sum Problem*, SIAM J. of Computing, **20**, 1991, 1157–1189.

[19] Jeroslow R.G., *Trivial Integer Programs Unsolvable by Branch and Bound*, Mathematical Programming, **6**, 1974, 105–109.

[20] Lipkin E., *On Representation of r-Powers by Subset-Sums*, Acta Arithmetica, **LII**, 1989, 353–366.

[21] Sahni S.K., *Algorithms for Scheduling Independent Tasks*, J. ACM, **23**, 1976, 116–127.

[22] Sárkőzy A., *Finite Addition Theorems II*, J. Number Theory, **48**, 1994, 197–218.

M. Chaimovich, 7041 Wolftree Lane, Rockville MD 20852, USA
E-mail : mark.chaimovich@bellatlantic.COM

Astérisque

MARK CHAIMOVICH

New algorithm for dense subset-sum problem

Astérisque, tome 258 (1999), p. 363-373

<http://www.numdam.org/item?id=AST_1999__258__363_0>

Astérisque **258**, 1999, p. 363–373

NEW ALGORITHM FOR DENSE SUBSET-SUM PROBLEM

by

Mark Chaimovich

Abstract. — A new algorithm for the dense subset-sum problem is derived by using the structural characterization of the set of subset-sums obtained by analytical methods of additive number theory. The algorithm works for a large number of summands (m) with values that are bounded from above. The boundary (ℓ) moderately depends on m. The new algorithm has $O(m^{7/4}/\log^{3/4} m)$ time boundary that is faster than the previously known algorithms the best of which yields $O(m^2/\log^2 m)$.

1. Introduction

Consider the following subset-sum problem (see [**13**]). Let $A = \{a_1, \ldots, a_m\}$, $a_i \in I\!N$. For $B \subseteq A$, let $S_B = \sum_{a_i \in B} a_i$ and let $A^* = \{S_B \mid B \subseteq A\}$. The problem is to find the maximal subset-sum $S^* \in A^*$ satisfying $S^* \leq M$ for a given target number $M \in I\!N$.

Although the problem is NP-hard (the partition problem is easily reduced to the SSP), its restriction can be solved in polynomial time. Denote $\ell = \max\{a_i \mid a_i \in A\}$. Introducing restriction $\ell \leq m^\alpha$ where α is some positive real number (or equivalently $m \geq \ell^{1/\alpha}$), one can easily solve problems from this restricted class in $O(m^2\ell)$ time using dynamic programming.

This work belongs to the school of thought that applies analytical methods of number theory to integer programming (see [**8**], [**2**]). It continues the application of a new approach, the main idea of which is as follows: analytical methods enable us to effectively characterize the set A^* of subset-sums as a collection of arithmetic progressions with a common difference (see [**7**], [**12**], [**1**], [**10**]). Once this characterization is obtained, it is quite easy to find the largest element of A^* that is not greater than the given M.

Efficient algorithms have recently been derived using the new approach. In almost linear time (with respect to the number m of summands) they solve the following class

1991 ***Mathematics Subject Classification.*** — Primary: 90C10 Alternate: 05A17, 11B25, 68Q25.
Key words and phrases. — Analytical Number Theory, Integer Programming, Subset Sum Problem.

of SSP: the target number M is within a wide range of the mid-point of the interval $[0, S_A]$ and $m > c\ell^{2/3}\log^{1/3}\ell$, $\ell > \ell_0$ when A is a set of distinct summands ([**9**], [**4**], [**6**], [**11**]) or $m > 6\ell\log\ell$ when A is an arbitrary multi-set without any limitation on the number of distinct summands ([**5**]). Here and further on $\ell_0, c, c_1, c_2, \ldots$ denote some absolute positive constants.

The latest analytical result ([**10**]) allows one to apply the algorithm from [**9**] to problems with density $m > c_1(\ell\log\ell)^{1/2}$. The algorithm from [**11**] works for density $m > c_2\ell^{1/2}\log\ell$ which is almost the same as in [**10**]. For $m < \ell^{2/3}$, the time boundary for both algorithms is estimated as $O((\frac{\ell}{m})^2)$, i.e., $O(\frac{m^2}{\log^2 m})$ for the lowest density $(m \sim (\ell\log\ell)^{1/2})$.

This work refines the structural characterization of the set of subset-sums which allows us to use more efficient conditions in the process of determining the structure. These refinements are discussed in Section 2. They lead to the development of a new algorithm which is described in Section 3. It works in $O(m\log m + \min\{\frac{\ell^{5/4}\log^{1/2}\ell}{m^{3/4}}, (\frac{\ell}{m})^2\})$ time which improves [**9**] and [**11**] for $m \leq \frac{\ell^{3/5}}{\log^{2/5}\ell}$ and yields $O(m^{7/4}/\log^{3/4} m)$ time for $m \sim (\ell\log\ell)^{1/2}$.

2. Refinement of the structural characterization of the set A^* of subset-sums

The following Theorem 2.1 [**10**] determines the structure of the set A^* of subset-sums for $m > c_1(\ell\log\ell)^{1/2}$ as a long segment of an arithmetic progression.

Theorem 2.1 **(G. Freiman).** — *Let $A = \{a_1, \ldots, a_m\}$ be a set of m integers taken from the segment $[1, \ell]$. Assume that $m > c_1(\ell\log\ell)^{1/2}$ and $\ell > \ell_0$.*
(i) There is an integer d, $1 \leq d \leq \frac{3\ell}{m}$, such that

$$|A(0,d)| > m - d \tag{1}$$

and

$$\{M : M \equiv 0 (\bmod d), |M - \tfrac{1}{2}S_{A(0,d)}| \leq c_2 dm^2\} \subseteq A^*(0,d),$$

where $A(s,t) = \{a : a \equiv s (\bmod t), a \in A\}$.
(ii) If for all prime numbers p, $2 \leq p \leq \frac{3\ell}{m}$,

$$|A(0,p)| \leq m - \frac{3\ell}{m}, \tag{2}$$

then the assertion (i) of the Theorem holds true with $d = 1$.

Simple consideration shows that verification of condition (2) is crucial for the structural characterization of a set A^* of subset-sums. Algorithms from [**9**] and [**11**] use this condition directly ([**9**]) or indirectly ([**11**]). Our intention is to replace condition (2) by a condition (or a set of conditions), verification of which is easier in the sense that the number of required operations is smaller. To do this we introduce the notion of *d-full set*. We say that set A is *d*-full if A^* contains all classes of residues modulo d, i.e., in other words, $A^*(\bmod d) = \{0, 1, \ldots, d-1\}$.

Let us study some properties of d-full sets.

Define $S_{r(\bmod d)} = \min\{s \in A^*, s \equiv r(\bmod d)\}$.

Lemma 2.2. — *Let A be a set of integers taken from the segment $[1,\ell]$. Suppose that A is d-full. Then for each r, $0 < r < d$,*

$$S_{r(\bmod d)} \leq d\ell. \tag{3}$$

Proof. — Assume that for some r condition (3) is not true, i.e., $S_{r(\bmod d)} > d\ell$. This means that $S_{r(\bmod d)} = a_{i_1} + a_{i_2} + \cdots + a_{i_k}$ for some $k > d$. Consider the sequence of subset-sums $T_s = \sum_{j=1}^{s} a_{i_j}$, $1 \leq s \leq k$. Obviously, at least two of these sums (assume T_s and T_q, $s < q$) belong to the same residue class modulo d (since $k > d$). Then $T_q - T_s \equiv 0(\bmod d)$ and subset-sum $T_k - (T_q - T_s) = a_{i_1} + \cdots + a_{i_s} + a_{i_{q+1}} + \cdots + a_{i_k} \equiv r(\bmod d)$ and this subset-sum is smaller than $S_{r(\bmod d)}$. This fact contradicts the minimality of $S_{r(\bmod d)}$. □

Lemma 2.3. — *Suppose that the set A is d-full. Then there is a d-full subset of A with cardinality less than d.*

Proof. — Let us assume that contrary to the Lemma the smallest d-full subset of A has more than $d-1$ elements. Denote this subset by $A' = \{a_1, \ldots, a_k\}$. In fact, $d \not| a_i$ for all i's.

Let B be the multi-set of non-zero residues modulo d in A', that is B is composed with $|A'(i,d)|$ times i for any $1 \leq i < d$. Naturally one has $B^* = (A')^*(\bmod d)$. Then, as a multi-set, $|B| = \sum_{i=1}^{d-1} |A'(i,d)| \geq d$, by the assumption.

Define a sequence of multi-sets $B_0, B_1, \ldots, B_k$ as follows: B_0 is an empty set and $B_i = \{b_1, \ldots, b_i\}$ for $i > 0$. Note that $0 \in B_i^*$ (since it is the sum of an empty subset), and that

$$B_i^* = B_{i-1}^* + \{0, b_i\} = B_{i-1}^* \cup (B_{i-1}^* + b_i), 1 \leq i \leq k. \tag{4}$$

Thus, obviously, $|B_{i-1}^*| \leq |B_i^*|$.

Taking into account that $|B_0^*| = 1$ and that $|B| = k \geq d$, for some i we have $|B_{i-1}^*| = |B_i^*|$ implying that residue b_i (and element a_i respectively) does not add new residue classes, i.e., $(B \setminus b_i)^* = B^*$. Therefore, $A' \setminus a_i$ is d-full as well as A'. This fact contradicts the assumption that A' is the smallest d-full subset of A and proves the Lemma. □

The next lemma refines the second assertion (ii) of Theorem 2.1.

Lemma 2.4. — *Let A be a set of integers taken from the segment $[1,\ell]$. Assume that $|A| = m > c_1(\ell \log \ell)^{1/2}$, $\ell > \ell_0$, and suppose that A is q-full for each q, $2 \leq q \leq \frac{3\ell}{m}$. Then the assertion (i) of Theorem 2.1 holds with $d = 1$.*

Proof. — Assume that $d > 1$ in Theorem 2.1. By the theorem, a long segment of an arithmetic progression belongs to $A^*(0,d)$. On the other hand, A is d-full (since $d \leq \frac{3\ell}{m}$) and subset-sum $S_{r(\bmod d)}$ exists for each r, $1 \leq r < d$. Combine a long segment of an arithmetic progression (with difference d) in interval

$$[\tfrac{1}{2}S_{A(0,d)} - c_2 dm^2, \tfrac{1}{2}S_{A(0,d)} + c_2 dm^2]$$

(belonging to $A^*(0,d)$) with subset-sums $S_{1(\bmod d)}, S_{2(\bmod d)}, \ldots, S_{d-1(\bmod d)}$ (these subset-sums are obtained without using elements of $A(0,d)$). Thus we obtain an interval

$$[\tfrac{1}{2}S_{A(0,d)} - c_2 dm^2 + \max\{S_{r(\bmod d)} : 1 \leq r < d\}, \tfrac{1}{2}S_{A(0,d)} + c_2 dm^2],$$

all integers of which belong to A^*. In fact, if the length of this new interval is sufficiently large ($O(m^2)$, for example), we will obtain the result of Theorem 2.1 with $d' = 1$. Actually, since we are interested only in the case $d > 1$ and since $\max\{S_{r(\bmod d)} : 1 \leq r < d\} < d\ell = O(dm^2/\log m)$, the length of the obtained interval is

$$O(dm^2 - \max\{S_{r(\bmod d)} : 1 \leq r < d\}) = O(dm^2 - \frac{dm^2}{\log m}) = O(dm^2)$$

which completes the proof. □

The latest property (Lemma 2.4) shows that in order to obtain a structural characterization of A^*, it is sufficient to verify that set A is q-full for all q's, $2 \leq q \leq \frac{3\ell}{m}$. Clearly, the new condition is weaker than (2): A can be q-full even if $|A(0,q)| > m - \frac{3\ell}{m}$. However, from an algorithmic point of view this new condition is difficult to verify. To correct this we have to use some lemmas which determine different sufficient conditions implying that set A is q-full. We will also show that it is sufficient to verify the prime numbers only.

Lemma 2.5 ([**3**]). — *If p is prime and*

$$\sum_{i=1}^{p-1} |A(i,p)| \geq p - 1 \tag{5}$$

then A is p-full.

The proof of this lemma is presented here because of the difficulty in accessing of reference [**3**].

Proof. — Using the fact that all elements of $A(i,p), i \neq 0$, are relatively prime to p, introduce ring $\mathbb{Z}_p$ of residues mod p. In the following reasoning it is implied that all arithmetic operations, including the operations for computing subset-sums, are operations modulo p in $\mathbb{Z}_p$.

Put, as in the proof of Lemma 2.3, $B = \{b_1, b_2, \ldots, b_k\}$ for the multi-set of non-zero residues modulo p in A and define the sequence of multi-sets $B_0, B_1, \ldots, B_k$ where B_0 is an empty set and $B_i = \{b_1, \ldots, b_i\}$ for $i > 0$.

By the hypothesis, $|B| = \sum_{i=1}^{p-1} |A(i,p)| \geq p - 1$. If for all $i \leq p-1, |B_{i-1}^*| < |B_i^*|$, then $|B_i^*| \geq |B_{i-1}^*| + 1 \geq |B_0^*| + i = i + 1$, i.e., $|B_{p-1}^*| \geq p$, which concludes the proof, since we are dealing with residues modulo p.

Otherwise, the fact that $|B_{i-1}^*| = |B_i^*|$ for some $i < p - 1$ implies that for any $c \in B_{i-1}^*$, $c + b_i$ also belongs to B_{i-1}^*. Continuing this reasoning we obtain $c + rb_i \in B_{i-1}^* \subseteq B^*$ for any r. Recalling that all operations are modulo p and that $\gcd(b_i, p) = 1$, one obtains that all residues modulo p are in B^*, i.e., A is p-full. □

***Lemma 2.6* (Olson [14]).** — *If p is prime and*

$$|\{i : |A(i,p)| \neq 0, 1 \leq i < p\}| > 2p^{1/2} \tag{6}$$

then A is p-full.

***Lemma 2.7* (Theorem 7, Sárközy [15]).** — *If p is prime and*

$$(\sum_{i=1}^{p-1} |A(i,p)|)^3 \geq c_5 p \log p \sum_{i=1}^{p-1} |A(i,p)|^2 \tag{7}$$

where $c_5 = 4 \cdot 10^6$, then A is p-full.

Note that condition (7) implies $\sum_{i=1}^{p-1} |A(i,p)| \geq (c_5 p \log p)^{1/2}$ in view of

$$\sum_{i=1}^{p-1} |A(i,p)| \leq \sum_{i=1}^{p-1} |A(i,p)|^2.$$

The next two lemmas show that it is sufficient to verify the prime numbers only.

Lemma 2.8. — *If for prime numbers p, $2 \leq p \leq Q^{1/2}$,*

$$|A(0,p)| \leq m - Q, \tag{8}$$

and for prime numbers p, $Q^{1/2} < p \leq Q$, the set A is p-full, then the set A is t-full for all integers t, $2 \leq t \leq Q$.

Proof. — The proof employs induction for the total number of prime divisors of t.

1. t is prime. Condition (8) ensures that Lemma 2.5 can be applied to all prime numbers $t \leq Q^{1/2}$. For prime numbers $t > Q^{1/2}$, the set A is t-full by definition.
2. For $n > 1$, assume that the Lemma is true for each number whose total number of prime divisors is less than n. Now we are going to prove the Lemma for any integer t having n prime divisors.

   Let $t = p_1 \cdots p_n$ where $p_1 \leq p_2 \leq \cdots \leq p_n$ are the prime divisors of t. One has $p_1 \leq t^{1/2} \leq Q^{1/2}$ and, in view of (8), $|B| = |A \setminus A(0,t)| \geq |A \setminus A(0,p_1)| \geq Q \geq t$.

   Denote $s = t/p_1$. This integer s has $n-1$ prime divisors. By the induction hypothesis, A is s-full. Thus, according to Lemma 2.3, there is $A' \subseteq A$ such that A' is s-full and $|A'| < s$. Put, as in the proof of Lemma 2.5, $B = \{b_1, b_2, \ldots, b_k\}$ for the multi-set of non-zero residues modulo t in A and define $B_i = \{b_1, \ldots, b_i\}$. Without losing generality, assume that the first residues in B corresponds to elements of A'. Thus, $B^*_{|A'|}$ contains all classes of residue modulo s implying $|B^*_{|A'|}| \geq s$. Continue with the same reasoning as in Lemma 2.5.

   Again, if for all $i, |A'| < i \leq t-1, |B^*_{i-1}| < |B^*_i|$, then $|B^*_i| \geq |B^*_{i-1}| + 1 \geq |B^*_{|A'|}| + (i - |A'|) \geq i + 1$, i.e., $|B^*_{t-1}| \geq t$, which concludes the proof, since we are dealing with residues modulo t.

   Otherwise, the fact that $|B^*_{i-1}| = |B^*_i|$ for some i, $|A'| < i \leq t-1$ implies that for any $c \in B^*_{i-1}$, $c + b_i \in B^*_{i-1}$. Continuing this reasoning we obtain $c + rb_i \in B^*_{i-1} \subseteq B^*$ for any r. Recalling that $B^*_{|A'|}$ contains $c_1, \ldots, c_s$ - different residues modulo s - we generate s disjoint sequences $c_j + rb_i$. Since each sequence has $r = \frac{t}{s}$ elements modulo t, all sequences together cover the entire set of residues modulo t, i.e., A is t-full.

This concludes the proof that the set A is t-full for all $t \leq Q$. □

Now we can formulate a sufficient condition for a long interval to exist in the set A^* of subset-sums:

Corollary 2.9. — *Let A be a set of integers taken from the segment $[1, \ell]$. Assume that $|A| = m > c_1(\ell \log \ell)^{1/2}$, $\ell > \ell_0$, and suppose that for all primes p, $2 \leq p \leq (\frac{3\ell}{m})^{1/2}$, condition (2) holds and for all primes p, $(\frac{3\ell}{m})^{1/2} < p \leq \frac{3\ell}{m}$, at least one of the conditions (5), (6) or (7) is satisfied. Then A^* contains a long interval: a segment of an arithmetic progression with difference 1 and length $O(m^2)$.*

Proof. — The corollary follows from previously mentioned Lemmas 2.4, 2.5, 2.6, 2.7 and 2.8. □

3. Algorithm

In the previous section we determined a sufficient condition, ensuring the existence of a long interval contained in A^*. In the case where this condition is not satisfied, namely, if for some p_1 either condition (2) (if p_1 is small) or conditions (5), (6) and (7) (if p_1 is large) fail, the process similar to the process described in **[9]** may be applied. This process finds a number d such that an arithmetic progression with difference d belongs to the set of subset-sums. It is implemented in the first step of the algorithm. The second step of the algorithm finds all non-zero residues modulo this d in A^* by using a modification of dynamic programming approach modulo d.

Now we are ready to describe the algorithm.

Notation. — $n_p(i)$, $0 \leq i < p$: the counter of summands belonging to residue class i $\bmod p$ (when all summands of A are verified $n_p(i) = |A(i,p)|$);
$r_p = |\{i \mid 1 \leq i < p, n_p(i) \neq 0\}|$: the counter of different non-zero residues modulo p;
$R_p = \sum_{i=1}^{p-1} n_p(i)$; $R'_p = R_p + n_p(0)$; $S_p = \sum_{i=1}^{p-1} n_p^2(i)$;
$\frac{A(0,p)}{p} = \{a \mid ap \in A(0,p)\}$;
prevpr(x): the prime number preceding x;
nextpr(x): the prime number following x;

In this notation conditions (5), (6) and (7) will take form $R_p \geq p - 1$, $r_p > 2p^{1/2}$ and $R_p^3 \geq (c_5 p \log p) S_p$, respectively.

Algorithm 1.

```
1. Finding $d$
   (a) Initialization: $d \leftarrow 1$, $p \leftarrow 2$, $Q \leftarrow \lfloor \frac{3\ell}{m} \rfloor$.
   (b) $R_p \leftarrow 0$.
       For each $a \in A$ where $a \equiv 0 (\bmod d)$, compute $s = \frac{a}{d} - \lfloor \frac{a}{dp} \rfloor p$ and if $s \neq 0$
       then advance the counter $R_p \leftarrow R_p + 1$;
       Continue this process until $R_p \geq Q$ or all elements are processed.
```

```
        If $R_p \geq Q$ then set $p \leftarrow nextpr(p)$;
        otherwise set $d \leftarrow dp$, $Q \leftarrow \lfloor \frac{3\ell}{d|A(0,d)|} \rfloor$ and $p \leftarrow 2$.
        If $p \leq Q^{1/2}$ return to 1(b);
        otherwise set $p \leftarrow prevpr(Q)$ and go to 1(c).
    (c) $n_p(i) \leftarrow 0\ (0 \leq i < p), R_p \leftarrow 0, S_p \leftarrow 0, R'_p \leftarrow 0, r_p \leftarrow 0$.
        For each $a \in A$ for which $a \equiv 0 (\bmod d)$ compute $s = \frac{a}{d} - \lfloor \frac{a}{dp} \rfloor p$ and
        advance the counters:
        $n_p(s) \leftarrow n_p(s) + 1$, $R'_p \leftarrow R'_p + 1$;
        if $s \neq 0$ then ($R_p \leftarrow R_p + 1$, $S_p \leftarrow S_p + 2n_p(s) - 1$;
                          if $n_p(s) = 1$ then $r_p \leftarrow r_p + 1$);
        Continue this process until one of the following inequalities is true:
(9)             $r_p > 2p^{1/2}, \quad R_p \geq p - 1, \quad R_p^3 \geq (c_5 p \log p) S_p,$
        or all elements are processed.
        If all elements are processed ($n_p(0) > |A(0,d)| - p$) then $d \leftarrow dp$.
        If $R'_p \geq (\frac{16 c_5 r_p \ell \log \ell}{p})^{1/2}$ then $p \leftarrow prevpr(\min\{p-1, \frac{4 r_p \ell}{p R'_p}\})$;
        otherwise $p \leftarrow prevpr(p-1)$.
        If $p \geq Q^{1/2}$ return to 1(c); otherwise go to 1(d).
    (d) Find $n_d(i)$, $1 \leq i < d$, and $r_d$ for the set $A$.
2. Finding C – the set of all non-zero residues modulo $d$ in $A^*$.
     Define the sequence of sets $C_0, C_1, \ldots, C_{d-1}$ in the following way: $C_0 = \{0\}$
   and, for $i > 0$, $C_i = C_{i-1} + \{0, i, \ldots, n_d(i) i\} (\bmod d)$ if $n_d(i) \neq 0$ or $C_i = C_{i-1}$
   if $n_d(i) = 0$. Clearly, $C_{d-1} = C$.
     Let $v$ be a vector with $d$ coordinates (numbered from 0 to $d-1$) which
   represents $C_i$ in the way that if $j \in C_i$ then $v(j) = i$ and if $j \notin C_i$ then
   $v(j) = -1$.
    (a) Initialization: $v \leftarrow (0, -1, \ldots, -1)$.
    (b) For all $i$, $1 \leq i < d$, for which $n_d(i) \neq 0$ do
            for all $j$, $1 \leq j < d$, for which $0 \leq v(j) < i$ do
                $v(j) \leftarrow i$ and
                for $s$ running from 1 to $n_d(i)$ while $v(j + si (\bmod d)) = -1$
                    $v(j + si (\bmod d)) \leftarrow i$.
3. Finding $S^*$. Define $s \equiv M (\bmod d)$, $0 \leq s < d$.
   Find $S^* = M - s + s_0$, where $s_0 = \max\{s_i \mid s_i \in C, s_i \leq s\}$.
```

To prove the validity of the algorithm we need to ensure that its step 1 finds a proper number d such that a set $\frac{A(0,d)}{d}$ satisfies all the conditions of Corollary 2.9. Indeed, sub-steps 1(b) and 1(c) use the conditions of the corollary. Therefore, the only thing that needs to be proved is the validity of the condition in sub-step 1(c) $\left(R'_p \geq \left(\frac{16 c_5 r_p \ell \log \ell}{p} \right)^{1/2} \right)$ which allows us to skip verification of some p's.

Recall that R'_p is the counter of elements of the set that have been checked for divisibility by p and that we stop the verification process for a particular prime number p once one of the conditions in (9) is satisfied. Therefore, the number of elements that have been checked for a particular p may be small (if many different non-zero residues are found in the beginning of the process) but this value may also be quite large. However, the fact that many elements have been checked for some $p' > Q^{1/2}$ ensures that A is p-full for many p's, namely, for $p > \frac{4r_{p'}\ell}{p'R'_{p'}}$. This is proved in the following lemma.

Lemma 3.1. — *Let B be a set of integers taken from the segment $[1,\ell]$. Assume that there is a prime $p' < \ell^{1/2}$ which satisfies the inequality*

$$|B| \geq \left(\frac{16c_5 r_{p'}\ell\log\ell}{p'}\right)^{1/2}, \tag{10}$$

where $r_{p'} = |\{i : |B(i,p')| \neq 0, 0 \leq i < p'\}|$ and c_5 is the constant from Lemma 2.7. Then, for prime numbers p, $\frac{4r_{p'}\ell}{p'|B|} < p < \ell^{1/2}$, $p \neq p'$, the set B is p-full.

Proof. — We are going to show that condition (7) of Lemma 2.7 is satisfied for all p's from the required interval. From this point on, for convenience we will use r without a subscript to denote $r_{p'}$.

Let $\{b_1,\dots,b_r\}$ be the set of all classes of residues modulo p' of the set B and let t_i, $1 \leq i \leq r$, be the number of occurrences of residues from class b_i in the set B. Without losing generality, assume that $t_1 \geq t_2 \geq \cdots \geq t_r$. Among the t_i elements which are in the class of b_i modulo p', only $\lceil\frac{\ell}{pp'}\rceil < \frac{2\ell}{pp'}$ elements can belong to the same class of residues modulo p, $p \neq p'$. Therefore, these t_i elements of B belong to at least $\lceil\frac{t_i pp'}{2\ell}\rceil$ different classes of residues modulo p.

To estimate from above the value of $\sum_{i=1}^{p-1}|B(i,p)|^2$ in the left-hand side in (7) we have taken the worst case scenario where the number of different classes of residues modulo p is the smallest possible. For a given $|B|$, this case occurs when each class of residues contains the maximum possible number of elements. Thus, the number of classes is at least $\lceil\frac{t_1 pp'}{2\ell}\rceil$ and each class can include the following number of elements of B: less than $\frac{2\ell r}{pp'}$ elements in $\lceil\frac{t_r pp'}{2\ell}\rceil$ classes, $\frac{2\ell(r-1)}{pp'}$ elements in $\lceil\frac{t_{r-1}pp'}{2\ell}\rceil - \lceil\frac{t_r pp'}{2\ell}\rceil$ classes, $\dots$, and $\frac{2\ell}{pp'}$ elements in $\lceil\frac{t_1 pp'}{2\ell}\rceil - \lceil\frac{t_2 pp'}{2\ell}\rceil$ classes. (Recall that $|B| = \sum_{i=1}^{r} t_i$ is being given.) Using these values we can estimate

$$\begin{aligned}
\sum_{i=1}^{p-1}|B(i,p)|^2 &\leq \left(\frac{2\ell r}{pp'}\right)^2\left\lceil\frac{t_r pp'}{2\ell}\right\rceil + \left(\frac{2\ell(r-1)}{pp'}\right)^2\left(\left\lceil\frac{t_{r-1}pp'}{2\ell}\right\rceil - \left\lceil\frac{t_r pp'}{2\ell}\right\rceil\right)\\
&\quad + \cdots + \left(\frac{2\ell}{pp'}\right)^2\left(\left\lceil\frac{t_1 pp'}{2\ell}\right\rceil - \left\lceil\frac{t_2 pp'}{2\ell}\right\rceil\right) - |B(0,p)|^2\\
&= \left(\frac{2\ell}{pp'}\right)^2\left(\left\lceil\frac{t_r pp'}{2\ell}\right\rceil(2r-1) + \left\lceil\frac{t_{r-1}pp'}{2\ell}\right\rceil(2r-3)\right.\\
&\quad \left. + \cdots + \left\lceil\frac{t_1 pp'}{2\ell}\right\rceil\cdot 1\right) - |B(0,p)|^2
\end{aligned}$$

$$\begin{aligned}
&\leq \left(\frac{2\ell}{pp'}\right)^2 \left(\frac{t_r pp'}{2\ell}(2r-1) + \frac{t_{r-1}pp'}{2\ell}(2r-3)\right. \\
&\qquad \left. + \cdots + \frac{t_1 pp'}{2\ell} + r^2\right) - |B(0,p)|^2 \\
&\leq \left(\frac{2\ell r}{pp'}\right)^2 \cdot \frac{|B|}{r} \cdot \frac{pp'}{2\ell} + \left(\frac{2\ell r}{pp'}\right)^2 - |B(0,p)|^2 \\
&= \frac{2\ell r|B|}{pp'}\left(1 + \frac{2\ell r}{|B|pp'} - \frac{pp'|B(0,p)|^2}{2\ell r|B|}\right)
\end{aligned}$$

and, taking into account (10) and that $|B| > \frac{4r\ell}{pp'}$, we continue

$$\begin{aligned}
\sum_{i=1}^{p-1}|B(i,p)|^2 &\leq \frac{|B|^3}{8c_5 p\log\ell}\left(1 + \frac{1}{2} - \frac{2|B(0,p)|^2}{|B|^2}\right) \\
&= \frac{(\sum_{i=1}^{p-1}|B(i,p)|)^3}{8c_5 p\log\ell} \cdot \frac{\frac{3}{2} - 2\alpha^2}{(1-\alpha)^3},
\end{aligned}$$

where $\alpha = \frac{|B(0,p)|}{|B|}$. To prove now the validity of (7) for p it is sufficient to show that $\frac{\frac{3}{2}-2\alpha^2}{(1-\alpha)^3} \leq 8$. It is easy to see that the function in the left-hand side of this inequality increases with α for $\alpha < \frac{2}{3}$ and, therefore, the inequality holds true for $\alpha \leq \frac{1}{2}$. Indeed, since the number of elements in one class of residues modulo p cannot exceed $\frac{2\ell r}{pp'}$ and $|B| > \frac{4\ell r}{pp'}$, $\alpha = \frac{|B(0,p)|}{|B|} \leq \frac{1}{2}$ that concludes the proof. □

The complexity. — Step 1 checks the divisibility of elements a_i by different prime numbers p. Since $a_i \leq \ell$, the number of prime divisors of a_i cannot be more than $\log_2 \ell$. Therefore, the overall number of occurrences where some p divides some element of A is $O(m\log m)$. In order to estimate the number of occurrences where some p does not divide some element of A we need to investigate each part of Step 1 separately.

In Step 1(b), in the worst case, we may find Q elements not divisible by p while verifying this number p. Since this part of Step 1 deals with prime numbers less than $Q^{1/2}$, the number of operations in Step 1(b) where some p does not divide some element of A is $O(Q^{3/2}) = O((\frac{\ell}{m})^{3/2})$. (Recall that $Q \sim \frac{\ell}{m}$.)

In step 1(c), again, no more than p elements not divisible by p may be found. Thus, the number of operations in Step 1(c) where some p does not divide some element of A is limited by $O(Q^2) = O((\frac{\ell}{m})^2)$. In fact, for $m \leq \frac{\ell^{3/5}}{\log^{2/5}\ell}$ this estimate can be improved.

If the number of verified elements is sufficiently large ($R'_p \geq (\frac{16c_5 r_p \ell\log\ell}{p})^{1/2}$) for some p, we are able to skip verification of some numbers according to Lemma 3.1. (The above "skipping" condition supersedes condition $R'_p > \frac{4r_p\ell}{p^2}$ for $p > \ell^{2/5}$ which ensures that the next number to be verified is less than p.)

Let us analyze this situation. The worst scenario (from a complexity point of view) occurs when we do not reach the "skipping" condition during verification. Thus, the number of operations in Step 1(c) where some p does not divide some element of A

is limited by

$$\sum_{p=\lceil Q^{1/2}\rceil}^{\lfloor \ell^{2/5}\rfloor} p + \sum_{p=\lfloor \ell^{2/5}\rfloor+1}^{\lfloor Q\rfloor} \left(\frac{16c_5 r_p \ell \log \ell}{p}\right)^{1/2} = O\left(\int_{Q^{1/2}}^{\ell^{2/5}} x dx + \int_{\ell^{2/5}}^{Q} \frac{(\ell \log \ell)^{1/2}}{x^{1/4}} dx\right).$$

Here we took into consideration the first condition in (9) which implies $r_p \le 2p^{1/2}$. By keeping after integration only the most significant term in each integral, we obtain complexity

$$O(\ell^{1/2} Q^{3/4} \log^{1/2} \ell) = O\left(\frac{\ell^{5/4} \log^{1/2} \ell}{m^{3/4}}\right). \tag{11}$$

This estimate is obtained assuming $p > \ell^{2/5}$. Observe that p can be greater than $\ell^{2/5}$ only for $m \le \ell^{3/5}$ since $p \le Q \sim \frac{\ell}{m}$. Comparing (11) with the first estimate – $O((\frac{\ell}{m})^2)$ – one can see that (11) improves it for $m \le \frac{\ell^{3/5}}{\log^{2/5} \ell}$.

Combining the results for sub-steps 1(b) and 1(c), one can get the overall complexity of the process that verifies divisibility of elements of A:

$$O\left(m \log m + \min\left\{\left(\frac{\ell}{m}\right)^2, \frac{\ell^{5/4} \log^{1/2} \ell}{m^{3/4}}\right\}\right). \tag{12}$$

This estimate also holds true for the overall complexity of the algorithm, since in the worst scenario both steps 1(d) and 2 have complexity $O(m)$.

In conclusion, the only thing that remains is to analyze the above expression (12). The second term dominates for $m \le \ell^{2/3} \log^{1/3} \ell$. It is equal to $O(\frac{\ell^{5/4} \log^{1/2} \ell}{m^{3/4}})$ for $m \le \frac{\ell^{3/5}}{\log^{2/5} \ell}$ and $O((\frac{\ell}{m})^2)$ otherwise. This improves the algorithms from **[9]** and **[11]** for low density $\left(m \le \frac{\ell^{3/5}}{\log^{2/5} \ell}\right)$. In the worst case ($m \sim (\ell \log \ell)^{1/2}$) time is $O(m^{7/4}/\log^{3/4} m)$.

References

[1] Alon N., and Freiman G. A., *On Sums of Subsets of a Set of Integers*, Combinatorica, **8**, 1988, 305–314.

[2] Buzytsky P., and Freiman G.A., *Analytical Methods in Integer Programming*, Moscow, ZEMJ., (Russian), 1980, 48 pp.

[3] Chaimovich M., *An Efficient Algorithm for the Subset-Sum Problem*, a manuscript, 1988.

[4] Chaimovich M., *Subset-Sum Problems with Different Summands: Computation*, Discrete Applied Mathematics, **27**, 1990, 277–282.

[5] Chaimovich M., *Solving a Value-Independent Knapsack Problem with the Use of Methods of Additive Number Theory*, Congressus Numerantium, **72**, 1990, 115–123.

[6] Chaimovich M., Freiman G.A., and Galil Z., *Solving Dense Subset-Sum Problem by Using Analytical Number Theory*, J. of Complexity, **5**, 1989, 271–282.

[7] Erdős P., and Freiman G., *On Two Additive Problems*, J. Number Theory, **34**, 1990, 1–12.

[8] Freiman G.A., *An Analytical Method of Analysis of Linear Boolean Equations*, Ann. New York Acad. Sci., **337**, 1980, 97–102.

[9] Freiman G.A., *Subset-Sum Problem with Different Summands*, Congressus Numerantium, **70**, 1990, 207–215.

[10] Freiman G.A., *New Analytical Results in Subset-Sum Problem*, Discrete Mathematics, **114**, 1993, 205–218.

[11] Galil Z., and Margalit O., *An Almost Linear-Time Algorithm for the Dense Subset-Sum Problem*, SIAM J. of Computing, **20**, 1991, 1157–1189.

[12] Lipkin E., *On Representation of r-Powers by Subset-Sums*, Acta Arithmetica, **LII**, 1989, 353–366.

[13] Martello S. and Toth T., *The 0-1 Knapsack Problem*, in Combinatorial Optimization, ed: N. Christofides, A.Mingozzi, P. Toth, C.Sandi, Wiley, 1979, 237–279.

[14] Olson J., *An Addition Theorem Modulo p*, J. of Combinatorial Theory, **5**, 1968, 45–52.

[15] Sárkőzy A., *Finite Addition Theorems II*, J. Number Theory, **48**, 1994, 197–218.

M. Chaimovich, 7041 Wolftree Lane, Rockville MD 20852, USA
E-mail : `mark.chaimovich@bellatlantic.COM`

Astérisque

ALAIN PLAGNE

On the two-dimensional subset sum problem

Astérisque, tome 258 (1999), p. 375-409

<http://www.numdam.org/item?id=AST_1999__258__375_0>

Astérisque **258**, 1999, p. 375–409

ON THE TWO-DIMENSIONAL SUBSET SUM PROBLEM

by

Alain Plagne

Abstract. — We consider a system of two linear boolean equations. Using methods from analytic number theory, we obtain sufficient conditions ensuring the solvability of the system. This completes Freiman's work on the subject.

1. Introduction

In this paper, we are interested in considering the system of two linear equations

$$a_1x_1 + \cdots + a_mx_m = b, \tag{1}$$

where $a_i = (a_{i,1}, a_{i,2})$ and $b = (b_1, b_2)$ are in $\mathbb{Z}^2$ and the x_i's, the unknowns, restricted to be either 0 or 1: that is, we are only interested in the boolean system induced by (1). Our intention is to give sufficient conditions for the set of coefficients $A = \{a_1, \ldots, a_m\}$ and b to ensure the solvability of (1). Probabilistic considerations show that, if the a_i's are "well distributed" and if their number is large enough, we should have solutions for all b in the neighbourhood of $\sum_{i=1}^{m} a_i/2$ and, more precisely, that the distribution of the number of solutions must be Gaussian: in fact, we are expecting a central limit theorem. So that here we investigate conditions ensuring a "good" distribution and then deduce the general case, that is, we describe the structure of A^*, the set of all sums $a_1x_1 + \cdots + a_mx_m$ with boolean unknowns.

The corresponding one-dimensional problem has been much studied in the past recent years from this point of view (see for example **[F80, AF88, EF90, F93]** and **[C91b]** for a complete bibliography). It has been shown that A^* is a collection of arithmetical progressions with the same difference. Each of these papers uses methods coming from analytic number theory, in the vein introduced in the 80's by Freiman (in the first quoted paper), essentially the principle of the circle method.

1991 ***Mathematics Subject Classification.*** — 11P99, 11P55, 11H06.

Key words and phrases. — additive number theory, structure theorem, subset sum, two-dimensional subset sum problem, geometry of numbers, integer points, Farey dissection, convex set.

Freiman began to generalize these results in two dimensions [**F96**] but some details remained obscure (computations on page 143 for example). A little later, Chaimovich [**C91a**] tried to generalize this in higher dimensions but some algorithmical problems arose in these cases (see, for example, our counterexample in section 2.3 to the extension of Proposition 4 stated in [**C91a**]). Our goal here is to make clear the situation. We complete, correct and improve in some places Freiman's [**F96**]. In addition, the results given here are in an explicit form, because of the opportunity they offer to design algorithms. However the constants for which we prove the theorems are still far from being the best one could expect.

For the sake of completeness, the present paper is self-contained except for very classical tools (as, for example, Farey dissection) for which we refer as usual to [**HW**].

In this paper we shall use the following notation: if u is in $\mathbb{R}^2$, we denote by u_1 and u_2 its coordinates with respect to the canonical basis (ϵ_1, ϵ_2) and by O the origin point. The e function is, as usual, defined by $e(t) = \exp(2\pi i t)$. For a real t, $||t||$ will denote the distance between t and $\mathbb{Z}$ and $[t]$ its integer part. The usual Euclidean scalar product is denoted simply with a point and the Lebesgue measure is denoted by μ. Finally, the volume of a fundamental parallelogram of any lattice Γ is denoted Vol Γ.

When $k, l \geq 1$ (in order to deal with really two-dimensional problems), we denote $P_{k,l}$ the integer rectangle

$$P_{k,l} = ([-k,k] \times [-l,l]) \cap \mathbb{Z}^2$$

and v its "volume", $v = (2k+1)(2l+1)$, that is, the number of integer points of $P_{k,l}$. In the sequel, A will denote a set of $m = |A|$ different integer points, $A = \{a_1, \dots, a_m\}$ and $J(b)$ the number of solutions of (1). We write $M = \sum_{i=1}^m a_i/2$ and

$$V = \begin{pmatrix} V_1^2 & V_{12} \\ V_{12} & V_2^2 \end{pmatrix},$$

where we have put $V_{12} = \sum_{j=1}^m a_{j,1}a_{j,2}$ and $V_i^2 = \sum_{j=1}^m a_{j,i}^2$ for $i = 1, 2$.

We denote by q_V the quadratic form naturally associated to this matrix $q_V(x) = \sum_{j=1}^m (a_j.x)^2$ $(x \in \mathbb{R}^2)$, and by $q_{V^{-1}}$ that one associated to V^{-1} that is $q_{V^{-1}}(x) = \frac{1}{\det V} \sum_{j=1}^m \det^2(a_j, x)$. Finally, we define the constants

$$k_1 = 25, \qquad k_2 = 6, \qquad k_4 = 189912,$$
$$k_5 = 100k_1 = 2500, \quad k_6 = 100k_2 = 600,$$
$$k_8 = \max(10k_6, k_5) = 6000, \qquad k_9 = \frac{9}{20},$$

and $k_3 = k_7$ being any constant $< 1/2$.

Our aim is to prove the following three Theorems:

Theorem 1. — *Let $A \subset P_{l_1,l_2}$ and $v = (2l_1+1)(2l_2+1)$. Assume*

$$|A| \geq k_1 v^{2/3} \log^{1/3} v \tag{2}$$

and that for each integer lattice Γ different from $\mathbb{Z}^2$ we have

$$|A \setminus A \cap \Gamma| \geq k_2 v^{2/3} \log^{1/3} v, \tag{3}$$

then we have the following asymptotic equivalent (when $v \to +\infty$)

$$J(b) \sim \frac{2^{m+1}}{\pi\sqrt{\det V}} \exp\{-2q_{V^{-1}}(M-b)\}, \tag{4}$$

provided that $q_{V^{-1}}(M-b) \leq k_3 \log\log v - 4$.

Notice first that the density hypothesis (2) implies

$$\frac{v}{\log v} \geq k_1^3,$$

that implies

$$v \geq k_4. \tag{5}$$

The previous Theorem is slightly better than Freiman's Theorem 1 of **[F96]**, the main difference being that the size of domain of validity of (4) is increased by a factor $\log\log v$ tending to infinity with v. This result is the heart of this work, but this is not entirely satisfying because dealing only with rectangle cases. That is why it is generalized in the following form.

Theorem 2. — *Let C be a compact convex set in $\mathbb{R}^2$ containing O, E be its integer points, and A be a subset of E. Assume*

$$|A| \geq k_5|E|^{2/3}\log^{1/3}|E|$$

and that for each integer lattice Γ different from $\mathbb{Z}^2$, we have

$$|A \setminus A \cap \Gamma| \geq k_6|E|^{2/3}\log^{1/3}|E|, \tag{6}$$

then we have the following asymptotic equivalent (when $|E| \to +\infty$)

$$J(b) \sim \frac{2^{m+1}}{\pi\sqrt{\det V}} \exp\{-2q_{V^{-1}}(M-b)\},$$

provided that $q_{V^{-1}}(M-b) \leq k_7 \log\log|E| - 4$.

Once again, it is not completely satisfying because it deals only with "good" cases: those where the elements of A are "well distributed". The conclusion of this paper will be the following general result.

Theorem 3. — *Let C be a compact convex set in $\mathbb{R}^2$ containing O, E be its integer points, and A be a subset of E. Assume $|A| \geq k_8|E|^{2/3}\log^{1/3}|E|$ and that for each line D such that $O \in D$, one has*

$$|A \cap D| < k_9|A|. \tag{7}$$

Then there exists a lattice Λ_0 such that, if A' stands for $A \setminus A \cap \Lambda_0$, one has $|A'| \leq |A \cap \Lambda_0|$ and

$$A'^* + (\Lambda_0 \cap F) \subset A^*,$$

where $F = \{x \in \mathbb{Z}^2, q_{W^{-1}}(M'-x) \leq k_7 \log\log(|A|/2) - 4\}$ and $W(x) = \sum_{a \in A'}(a.x)^2$, $M' = \sum_{a \in A'} a/2$.

This is a structural theorem because it describes how the set A^* is made, at least locally. It is a powerful result in order to design algorithms, as it has already been done in the one-dimensional case (see for example **[CFG89]**).

We notice that hypothesis (7) is in fact not very restrictive: it ensures that our set A is an essentially two-dimensional set. If that condition is not fulfilled, we have the possibility to treat our problem as a one-dimensional one, by forgetting some points and this is even much simpler.

Acknowledgments. — Many thanks go to G.A. Freiman for his kind help during the preparation of this paper. I would like also to thank J.-M. Deshouillers for his advice.

2. Preliminary lemmas

We begin this section by quoting some inequalities (whose validity can easily be seen by using, for instance, some Taylor-Lagrange's inequalities). For any real t, if $0 \leq |t| \leq 1/2$, we have

$$|1 + e(t)| \leq 2\exp(-\pi^2 t^2/2), \tag{8}$$

and if $|t| \leq \pi/2$,

$$0 \leq 1 - \exp(t^2/2)\cos t \leq (2t/\pi)^4. \tag{9}$$

Finally, for reals $(\epsilon_i)_{1\leq i \leq n}$ between 0 and 1, we have, with a trivial induction argument,

$$\prod_{i=1}^{n}(1-\epsilon_i) \geq 1 - \sum_{i=1}^{n}\epsilon_i. \tag{10}$$

Now we present several propositions that we shall need in the sequel.

2.1. Arithmetical lemmas. — Here, we give two results concerning the number of solutions of a Diophantine inequality.

Lemma 1. — *Let a, b, ϵ be real numbers and k, n be integers such that $0 < |a|k < 1$ and $\epsilon < (1 - k|a|)/2$. Then we have*

$$|\{x \in \mathbb{N}, n \leq x \leq n+k : ||ax+b|| \leq \epsilon\}| \leq 1 + [2\epsilon/|a|].$$

Proof. — Without loss of generality we may assume $a > 0$ and write $u_s = as+b$. This is a strictly increasing sequence. Let s_1 be the smallest integer, with $n \leq s_1 \leq n+k$, such that $||u_{s_1}|| \leq \epsilon$ (if s_1 does not exist, then the cardinality studied is zero); we thus have $|u_{s_1} - e| \leq \epsilon$ for some integer e. Let s_2 be the largest integer satisfying $|u_{s_2} - e| \leq \epsilon$. We claim that $s_2 < t \leq n+k$ implies $||u_t|| > \epsilon$; indeed $|u_t - e| > \epsilon$ is clear by definition of s_2 and

$$u_t = u_{s_1} + (t - s_1)a \leq u_{s_1} + ka \leq e + \epsilon + ka < e + 1 - \epsilon.$$

Since $s_2 - s_1 = (u_{s_2} - u_{s_1})/a \leq 2\epsilon/a$, we get, for the cardinality studied, the desired upper bound. □

Now, we prove a result due to Freiman. We write here a complete proof, in view of the lack of details in Freiman's paper **[F96]**.

Proposition 1. — *Let n, k, P be integers, $0 \leq P \leq k$ and a, b be two reals. Assume $a = p/q + z$ with $(p, q) = 1$, $q \leq P$, $1/2qk \leq |z| \leq 1/qP$. We have*

$$|\{x \in \mathbb{N}, n \leq x \leq n+k : ||ax+b|| \leq P^{-1}\}| \leq 3(4kP^{-1}+1).$$

Proof. — By just changing the value of b, the problem reduces to the case where $n = 0$. For $P \leq 12$, the result is clear, so we assume from now on $P > 12$ and without loss of generality $z > 0$. The solutions of $||ax + b|| \leq P^{-1}$ are clearly in bijection with those of the following problem, that we shall denote (P), consisting in finding $(x, x_0, t) \in \{0, \dots, k\} \times \{0, \dots, q-1\} \times \mathbb{Z}$ satisfying

$$\begin{cases} px \equiv x_0 \bmod q, \\ \left|\dfrac{x_0}{q} + zx + b - t\right| \leq P^{-1}. \end{cases}$$

One can easily bound from above the cardinality, J, of the set of solutions of (P) as follows

$$\begin{aligned} J \leq\ & |\{x_0 \mid \exists (x,t), (x, x_0, t) \text{ solution of (P)}\}| \\ & \times \max_{x_0} |\{t \mid \exists x, (x, x_0, t) \text{ solution of (P)}\}| \times \max_{x_0, t} |\{x \mid (x, x_0, t) \text{ solution of (P)}\}|. \end{aligned}$$

Now, write $|x_0/q + zx + b - t| \leq P^{-1}$ in the following form

$$(11) \qquad -qP^{-1} - zxq - bq + tq \leq x_0 \leq qP^{-1} - zxq - bq + tq.$$

It implies, because $x \geq 0$, that x_0 belongs to $[-qP^{-1} - zxq - bq + tq, qP^{-1} - bq + tq]$. But t is an integer, $0 \leq x \leq k$ and x_0 stays in $\{0, \dots, q-1\}$, so x_0 belongs to $[-qP^{-1} - zkq - bq, qP^{-1} - bq] \bmod q$, which has length $2qP^{-1} + zkq$, this yields

$$|\{x_0 | \exists (x,t), (x, x_0, t) \text{ solution of (P)}\}| \leq \inf([2qP^{-1} + zkq] + 1, q) \leq \inf([zkq] + 3, q).$$

Now, the value of x_0 being given, equation (11) can be rewritten

$$-P^{-1} + x_0/q + zx + b \leq t \leq P^{-1} + x_0/q + zx + b,$$

and, as $0 \leq x \leq k$, one has

$$-P^{-1} + x_0/q + b \leq t \leq P^{-1} + x_0/q + zk + b,$$

thus t belongs to an interval of length $2P^{-1} + zk$, which implies

$$\max_{x_0} |\{t | \exists x, (x, x_0, t) \text{ solution of (P)}\}| \leq [2P^{-1} + zk] + 1.$$

In the same vein, we can get

$$(12) \qquad \max_{x_0, t} |\{x | (x, x_0, t) \text{ solution of (P)}\}| \leq [2/qPz] + 1.$$

Indeed, x_0 and t being given, we have

$$(-P^{-1} - x_0q - b + t)/z \leq x \leq (P^{-1} - x_0q - b + t)/z,$$

so the x's which are possible solutions are consecutive integers in an interval of length $2/Pz$. The condition $px \equiv x_0 \bmod q$ implies moreover that on a complete set of residues modulo q only one x can be solution. This proves (12).

We have finally, and in any case,

$$J \leq \inf([|z|kq]+3, q)([2P^{-1}+|z|k]+1)([2/qP|z|]+1). \tag{13}$$

At this point, we have to distinguish two cases.

If $(1+2P^{-1})/2k \leq |z| \leq 1/qP$, equation (13) gives

$$J \leq q(2P^{-1}+1+|z|k)(1+2/|z|qP),$$

but, in view of the hypothesis, this is $\leq q(3|z|k)(3/|z|qP) = 9kP^{-1}$.

Now, if $|z| \leq (1+2P^{-1})/2k$, one has $2P^{-1}+|z|k < 2P^{-1}+(1+2P^{-1})/2 = 1/2+3P^{-1} < 1$, because $P > 12$, thus (13) implies

$$J \leq ([|z|kq]+3)([2/qP|z|]+1). \tag{14}$$

There are now three sub-cases according to the position of $|z|kq$.

If $|z|kq \geq 3/2$, (14) leads to

$$\begin{aligned} J &\leq (3+|z|kq)(1+2/qP|z|) \\ &\leq (3|z|q)(3/qP|z|) = 9kP^{-1}, \end{aligned}$$

because one has, in every case $1 \leq 1/qP|z|$.

If now $1 \leq |z|kq \leq 3/2$, equation (14) yields

$$\begin{aligned} J &\leq ([3/2]+3)(1+2/qP|z|) \\ &\leq 4(1+2kP^{-1}) \leq 12kP^{-1}+3, \end{aligned}$$

because $kP^{-1} \geq 1$.

Finally, if $1/2 \leq |z|kq < 1$, in virtue of (14),

$$\begin{aligned} J &\leq ([|z|kq]+3)(1+2/qP|z|) \\ &\leq 3(1+4kP^{-1}) \leq 12kP^{-1}+3. \end{aligned}$$

This concludes the proof of Proposition 1. □

2.2. A two-dimensional "reverse-Cauchy-Schwarz" inequality. — This section is devoted to the proof of an inequality used in **[F96]** without explanation. Our aim is to find a good lower bound for the ratio

$$\left(\sum_{j=1}^{m}(a_j.\alpha)^2\right)^2 \Big/ \sum_{j=1}^{m}(a_j.\alpha)^4 \tag{15}$$

which is naturally ≥ 1 and $\leq m$ (by the Cauchy-Schwarz inequality). We would like to "reverse" the Cauchy-Schwarz inequality, that is to find, for (15), a better lower bound than 1 (a power of m or $\log m$ for example). This is generally not possible, but here the a_j's have special properties which allow to get the desired result.

Let us consider the one-dimensional corresponding problem. Since α^4 can be factorized, the problem becomes to minimize

$$\left(\sum_{j=1}^{m} a_j^2\right)^2 \Big/ \sum_{j=1}^{m} a_j^4 \tag{16}$$

for distinct integers a_j's satisfying $1 \leq a_j \leq l$. It is easily seen that this ratio is

$$\geq \frac{\left(\sum_{j=1}^{m} a_j^2\right)^2}{l^2 \sum_{j=1}^{m} a_j^2} = \frac{\sum_{j=1}^{m} a_j^2}{l^2} \geq \frac{\sum_{j=1}^{m} j^2}{l^2} \sim \frac{m^3}{3l^2},$$

which is better than $O(1)$ as soon as $l^{2/3} = o(m)$.

This can be guessed in another way: if one tries to choose the a_j's such that (16) is near to 1, a natural idea (see below) is to take $a_j = j$ for $1 \leq j \leq m-1$ and $a_m = l$. This choice yields

$$\left(\sum_{j=1}^{m} a_j^2\right)^2 \Big/ \sum_{j=1}^{m} a_j^4 \asymp \frac{m^6 + l^4}{m^5 + l^4},$$

and this won't be $O(1)$ as soon as $l^4 = o(m^6)$, that is to say $l^{2/3} = o(m)$.

In dimension 2, the situation is not so clear but we will show that an analogous phenomenon happens. We begin by proving a preliminary lemma, which corresponds to a generalized one-dimensional case, for which we present two proofs: the first one will be direct while the second one corresponds to what we called the "natural idea" above. Although this second approach is much more intricate, we believe that the method could be efficient in some other contexts where the first one would not work.

The notation

$$\{a_1^{(e_1)}, \ldots, a_n^{(e_n)}\}$$

is for the multi-set (that is the set "with repetition") composed with e_1 times a_1, e_2 times a_2, and so on.

Lemma 2. — *Let r, s be integers ≥ 1, $A \subset E = \{1^{(r)}, \ldots, s^{(r)}\}$. Assume that*

$$|A| \geq c|E|^{2/3} \log^{1/3} |E| \tag{17}$$

for some constant c, then we have, for $k_{10} = 1/10$,

$$\left(\sum_{a\in A} a^2\right)^2 \geq k_{10} c^3 \log |E| \left(\sum_{a\in A} a^4\right). \tag{18}$$

First proof of Lemma 2. — We use the fact that

$$F(A) = \left(\sum_{a\in A} a^2\right)^2 \Big/ \left(\sum_{a\in A} a^4\right) \geq \left(\sum_{a\in A} a^2\right) \Big/ s^2. \tag{19}$$

Now, suppose first that $|A| \geq \sqrt{10}r$, then

$$F(A) \geq \frac{r(1^2 + 2^2 + \cdots + [|A|/r]^2)}{s^2} \geq \frac{r}{3s^2}\left[\frac{|A|}{r}\right]^3$$

which can be bounded from below by

$$\frac{r}{3s^2}\left(\frac{|A|}{r} - 1\right)^3 \geq \frac{1}{3}\left(1 - \frac{1}{\sqrt{10}}\right)^3 \frac{|A|^3}{r^2s^2} \geq \frac{|A|^3}{10r^2s^2}.$$

Since $rs = |E|$, we get the lower bound $\frac{c^3}{10}\log|E|$.

Suppose now $|A| < \sqrt{10}r$, then

$$\sqrt{10}r > |A| \geq c(rs)^{2/3}\log^{1/3}|E|,$$

which furnishes

$$\frac{r}{s^2} > \frac{c^3}{10\sqrt{10}}\log|E|.$$

But (19) implies

$$F(A) \geq \frac{|A|}{s^2} \geq c\left(\frac{r}{s^2}\right)^{2/3}\log^{1/3}|E|$$

and thus

$$F(A) \geq \frac{c^3\log|E|}{10}.$$

□

Now, we present the second method for obtaining a proof of Lemma 2 (in fact, the proof given here does not yield the same value for k_{10} but we did not try to optimize it). It begins with some definitions and a lemma.

Let E be a multi-set. If A is a sub-multi-set of E, a an element of A and b an element of $E \setminus A$, we denote by

$$A_a(b) = (A \setminus \{a\}) \cup \{b\},$$

the set obtained by replacing a by b in A.

Suppose E is a multi-set of reals and $\mathcal{F}$ a sub-family of the family of all sub-multi-sets of E. If A is a sub-multi-set of E belonging to $\mathcal{F}$ and a an element of A, we say that A is a-minimal relatively to $\mathcal{F}$ if for any b in $E \setminus A$ such that $b < a$, one has

$$A_a(b) \notin \mathcal{F}.$$

In the same way, we define A to be a-maximal relatively to $\mathcal{F}$ if for any b in $E \setminus A$ such that $b > a$, one has

$$A_a(b) \notin \mathcal{F},$$

and A is said to be a-extremal relatively to $\mathcal{F}$ if it is a-minimal or a-maximal relatively to $\mathcal{F}$. Finally, we say that A is $\mathcal{F}$-extremal if for any a in A, A is a-extremal. This can be restated in the following way: A is $\mathcal{F}$-extremal if for any $a \in A$ and $b, c \in E \setminus A$ such that $b < a < c$ then at least one of the sets $A_a(b), A_a(c)$ is not in $\mathcal{F}$. For example, if E is finite and $\mathcal{F} = \mathcal{P}(E)$, the $\mathcal{F}$-extremal sub-multi-sets are those in which the elements are accumulated on the extremities, with no "hole".

Lemma 3. — *Let t be any real, E be a finite multi-set of positive reals and $\mathcal{F}$ be any sub-family of the family of all sub-multi-sets of E. Assume the sub-multi-set B of E minimizes, on $\mathcal{F}$, the function D defined, for any $A \in \mathcal{F}$, by the formula*

$$D(A) = \left(\sum_{a\in A} a^2\right)^2 - t\left(\sum_{a\in A} a^4\right),$$

then B is $\mathcal{F}$-extremal.

Proof. — As E is finite, there is a minimum (on the sets belonging to $\mathcal{F}$) for D: so B always exists. Assume that B is not $\mathcal{F}$-extremal: it contains an element β such that there exists α, γ not in B and such that $0 \le \alpha < \beta < \gamma$ holds. Denoting simply $B(\alpha)$ and $B(\gamma)$ the sets obtained by replacing β in B, respectively by α and γ, it follows, by hypothesis, that $B(\alpha)$ and $B(\gamma)$ belong to $\mathcal{F}$. As B is minimal for D on $\mathcal{F}$, one has

$$D(B(\alpha)) - D(B) \ge 0 \qquad \text{and} \qquad D(B(\gamma)) - D(B) \ge 0.$$

Denoting S the sum of squares of $B \setminus \{\beta\}$, this can be rewritten,

$$(S+\alpha^2)^2 - (S+\beta^2)^2 - t(\alpha^4 - \beta^4) \ge 0, \tag{20}$$

$$(S+\gamma^2)^2 - (S+\beta^2)^2 - t(\gamma^4 - \beta^4) \ge 0. \tag{21}$$

Introduce now the following notations:

$$X = \alpha^4, \qquad Y = \beta^4, \qquad Z = \gamma^4,$$

and

$$F(u) = (S + u^{1/2})^2.$$

We have

$$F''(u) = -S/2u^{3/2},$$

which is strictly negative: therefore F is strictly concave. But equations (20) and (21) show

$$\frac{F(Y)-F(X)}{Y-X} \le t \le \frac{F(Z)-F(Y)}{Z-Y},$$

and that is not possible for a strictly concave function in view of $X < Y < Z$. □

Second proof of Lemma 2. — Let $\mathcal{F}$ be the set of all sub-multi-sets of E satisfying (17). Assume that (18) is proven for every $\mathcal{F}$-extremal set, then by Lemma 3, (18) is proven for every sub-multi-set of E belonging to $\mathcal{F}$ and we are done. Thus we only have to check that (18) holds for $\mathcal{F}$-extremal sets. That is what we do now, after having noticed that conditions (17) and $|E| \ge |A|$ imply

$$|A| \ge c^3 \log |E|.$$

Thus it is enough to get a lower bound with $k_{10}|A| \left(\sum_{a\in A} a^4\right)$ in the right-hand side of equation (18).

As above, we define the ratio

$$F(A) = \left(\sum_{a\in A} a^2\right)^2 \Big/ \sum_{a\in A} a^4.$$

We have to investigate the cases where A is of the following form

$$A = \{1^{(r)}, \dots, (a-1)^{(r)}, a^{(x_a)}, (s-b)^{(x_b)}, (s-b+1)^{(r)}, \dots, s^{(r)}\},$$

with $a \geq 1$, $0 \leq x_a, x_b \leq r$, b being possibly zero. We have (using elementary tools)

$$\begin{aligned} F(A) &= \frac{(r(1^2 + \cdots + (a-1)^2) + x_a a^2 + x_b(s-b)^2 + r(s^2 + \cdots + (s-b+1)^2))^2}{(r(1^4 + \cdots + (a-1)^4) + x_a a^4 + x_b(s-b)^4 + r(s^4 + \cdots + (s-b+1)^4))} \\ (22) \qquad &\geq \frac{(r(a-1)^3/3 + x_a a^2 + x_b(s-b)^2 + rbs^2/3)^2}{r(a-1)^5 + x_a a^4 + x_b(s-b)^4 + rbs^4}. \end{aligned}$$

Furthermore, the cardinality $|A|$ verifies:

$$|A| = r(a+b-1) + x_a + x_b. \tag{23}$$

Consider now the following sub-cases.

(1) If $a = 1$, equation (22) shows that

$$F(A) \geq \frac{(x_a + x_b(s-b)^2 + rbs^2/3)^2}{x_a + x_b(s-b)^4 + rbs^4}. \tag{24}$$

(1a) If $b = 0$ then equation (24) produces

$$F(A) \geq \frac{(x_a + x_b s^2)^2}{x_a + x_b s^4}.$$

(1a1) If $x_a \geq x_b s^4$, we get

$$2F(A) \geq x_a + \frac{x_b^2 s^4}{x_a} \geq x_a + \frac{x_b^2}{x_a} \geq \sup(x_a, x_b) \geq \frac{|A|}{2}$$

because of (23). Thus $F(A) \geq |A|/4$.

(1a2) If $x_a \leq x_b s^4$, we get

$$F(A) \geq \frac{x_a^2 + x_b^2 s^4}{2x_b s^4} \geq \begin{cases} x_b/2 \\ \sqrt{\dfrac{x_a^2 x_b^2 s^4}{x_b^2 s^8}} = \dfrac{x_a}{s^2}, \end{cases}$$

the second lower bound following from the arithmetico-geometric inequality.

If $x_b \geq |A|/2$, one has $F(A) \geq |A|/4$. Otherwise, as in (1a1), equation (23) implies $|A| = x_a + x_b \leq 2r$. Using (17), we have

$$2r \geq |A| \geq c|E|^{2/3} \log^{1/3}|E| \geq c(sr)^{2/3} \log^{1/3}|E|,$$

from which we deduce $8r \geq c^3 s^2 \log|E|$. Now, writing s^2 as $(s^2 s)^{2/3}$, we get

$$\begin{aligned} s^2 &\leq \left(\frac{8rs}{c^3 \log|E|}\right)^{2/3} \\ &\leq \frac{4|E|^{2/3}}{c^2 \log^{2/3}|E|}. \end{aligned}$$

Finally, we have (as $x_a \geq |A|/2$)

$$F(A) \geq \left(\frac{|A|}{2}\right)\left(\frac{c^2 \log^{2/3}|E|}{4|E|^{2/3}}\right) \geq \frac{c^3}{8}\log|E|.$$

(1b) Now, $b \geq 1$. Equation (24) gives

$$F(A) \geq \frac{(rbs^2/3)^2}{x_a + r(b+1)s^4} \geq \frac{(rbs^2/3)^2}{r(1+(b+1)s^4)} \geq \frac{b}{9(b+2)}rb \geq \frac{rb}{27},$$

in the same manner as above. Once again, using (23), we deduce $|A| \leq r(b+2) \leq 3rb$. Finally, in this case we have $F(A) \geq |A|/81$.

(2) If $a \geq 2$, we have

$$F(A) \geq \frac{(r(a-1)^3/3 + x_b(s-b)^2 + rbs^2/3)^2}{ra^5 + x_b(s-b)^4 + rbs^4}. \tag{25}$$

(2a) If $b = 0$, we get

$$F(A) \geq \frac{(r/3)^2(a-1)^6 + x_b^2 s^4}{ra^5 + x_b s^4}. \tag{26}$$

(2a1) If $ra^5 \geq x_b s^4$ then equation (26) implies

$$F(A) \geq \frac{(r/3)^2(a-1)^6}{2ra^5} = \frac{(a-1)r}{18}\left(\frac{a-1}{a}\right)^5 \geq \frac{(a-1)r}{576}.$$

But (23) implies $3(a-1)r \geq |A|$, so that finally $F(A) \geq |A|/1728$.

(2a2) If $ra^5 \leq x_b s^4$ then equation (26) yields

$$F(A) \geq \frac{(r(a-1)^3/3)^2 + x_b^2 s^4}{2x_b s^4}.$$

Applying once again the arithmetico-geometric inequality, we deduce $F(A) \geq r(a-1)^3/3s^2$. As $3r(a-1) \geq |A|$, using condition (17) we deduce the lower bound

$$(3(a-1)r)^3 \geq |A|^3 \geq (c(rs)^{2/3}\log^{1/3}|E|)^3 = c^3(rs)^2\log|E|,$$

thus

$$r(a-1)^3/s^2 \geq (c^3\log|E|)/27.$$

Finally, we have in this case $F(A) \geq (c^3\log|E|)/81$.

(2b) If $b \geq 1$, we have

$$F(A) \geq \frac{(r(a-1)^3/3 + rbs^2/3)^2}{ra^5 + r(b+1)s^4} = \frac{((a-1)^3/3 + bs^2/3)^2 r}{a^5 + (b+1)s^4}. \tag{27}$$

(2b1) If $a \geq b$, the cardinality equation (23) shows that $|A| \leq r(2a+1) \leq 5(a-1)r$.

(2b11) If $a^5 \geq (b+1)s^4$, we have successively

$$F(A) \geq \frac{r(a-1)^6/9}{2a^5} = \frac{(a-1)r}{18}\left(\frac{a-1}{a}\right)^5 \geq \frac{(a-1)r}{576} \geq \frac{|A|}{2880}.$$

(2b12) If $a^5 \leq (b+1)s^4$, we have

$$F(A) \geq r\frac{(a-1)^6/9 + b^2s^4/9}{2(b+1)s^4},$$

and after applying the arithmetico-geometric inequality

$$F(A) \geq \frac{b}{9(b+1)}\left(\frac{r(a-1)^3}{s^2}\right) \geq \frac{r(a-1)^3}{18s^2}.$$

Now, proceeding as in (2a2), we get that ($|A| \leq r(2a-1)+2r \leq r(2a+1) \leq 5r(a-1)$)

$$r(a-1)^3/s^2 \geq (c^3 \log|E|)/125,$$

and finally $F(A) \geq (c^3 \log|E|)/2250$.

(2b2) If $a \leq b$, using (27), we get

$$\begin{aligned} F(A) &\geq rb^2s^4/(9\{(b+1)s^4 + a^5\}) \\ &\geq rb^2/18(b+1) \geq rb/36. \end{aligned}$$

Once again (23) yields $3rb \geq |A|$, whence $F(A) \geq |A|/108$.

This completes this proof of Lemma 2. □

Before going a step further, it is interesting to notice that hypothesis (3) implies trivially the following:

(28) For each line D containing O, $|A \setminus A \cap D| \geq k_2 v^{2/3} \log^{1/3} v$,

since $D \cap \mathbb{Z}^2$ can be completed in some integral lattice different from $\mathbb{Z}^2$.

We are now able to deduce the following

Proposition 2. — *If $A \subset P_{l_1,l_2}$ satisfies $|A| \geq k_1 v^{2/3} \log^{1/3} v$ and hypothesis (3), then, for every $\alpha \in \mathbb{R}^2$, we have (recall $v = (2l_1+1)(2l_2+1)$)*

$$\left(\sum_{a\in A}(a.\alpha)^2\right)^2 \geq k_{11}\left(\sum_{a\in A}(a.\alpha)^4\right)\log v,$$

where

$$k_{11} = 1.12\ 10^{-3}.$$

as a consequence of

Proposition 3. — *If $A \subset P_{l_1,l_2}$ satisfies $|A| \geq k_1 v^{2/3} \log^{1/3} v$ and hypothesis (28), then, for every $\alpha \in \mathbb{R}^2$, we have*

$$\left(\sum_{a\in A}(a.\alpha)^2\right)^2 \geq k_{11}\left(\sum_{a\in A}(a.\alpha)^4\right)\log v.$$

Proof. — Let us first notice some facts. The formula is homogeneous and continuous with respect to α and symmetrical (as P_{l_1,l_2} is). Thus it suffices to prove it for every $\alpha = (p,q)$ with p,q positive integers sufficiently large and $\gcd(p,q) = 1$ (during this proof l_1 and l_2 are assumed to be fixed). Indeed the fractions q/p subject to these conditions are dense in $\mathbb{R}^+$.

In all this proof, we write $N = pl_1 + ql_2$ and assume, with no loss of generality, that

$$ql_2 \geq pl_1, \tag{29}$$

and

$$p \geq 2l_2 + 1, \tag{30}$$

$$q \geq 2l_1 + 1. \tag{31}$$

Then

$$\mathcal{S} = \{|a_j.\alpha|, a_j \in A\} \subset \{0^{(1)}, 1^{(2)}, \dots, N^{(2)}\} \subset \{0^{(2)}, 1^{(2)}, \dots, N^{(2)}\};$$

indeed $x.\alpha = t \in \mathbb{Z}$ is the equation of a line which can have at most one point in P_{l_1,l_2} because if

$$\begin{cases} px_1 + qx_2 = t, \\ py_1 + qy_2 = t, \end{cases}$$

then $q|y_1 - x_1$. This implies $x_1 = y_1$ as a consequence of $|x_1|, |y_1| \leq l_1 < q/2$ and then $x_2 = y_2$.

We now examine the value of $|n.\alpha|$ when $n \in P_{l_1,l_2}$. Take first $u, v \in \mathbb{Z}$ by Bezout Theorem such that $pu + qv = 1$. If $n.\alpha = t$, then there exists an integer e such that

$$n = t(u,v) + e(q, -p).$$

So, $n \in A$ implies that $|tu + eq| \leq l_1$ for some integer e, that is to say

$$\left\|t\frac{u}{q}\right\| \leq \frac{l_1}{q}. \tag{32}$$

We now distinguish two cases.

First case. — We assume that $(2l_2+1)^2 \geq 2l_1 + 1$ or that $q/p \leq 2v^{1/3}/3$. In the case where $(2l_2+1)^2 \geq 2l_1+1$, we get

$$(2l_1 + 1) \leq (2l_1+1)^{2/3}(2l_2+1)^{2/3} = v^{2/3},$$

and in the case where $q/p \leq 2v^{1/3}/3$, we get (using relation (29))

$$(2l_1+1) \leq (2l_1+1)^{1/2}\left(\frac{3q}{2p}(2l_2+1)\right)^{1/2} \leq v^{2/3}.$$

Here we have used $2l_1 + 1 \leq 3l_1 \leq 3ql_2/p \leq \frac{3q}{2p}(2l_2+1)$.

Let $k = [q/2] + 1$ and $P = [q/2l_1] < k$, we approximate u/q by an element α/β of the Farey dissection of order P:

$$\frac{u}{q} = \frac{\alpha}{\beta} + z,$$

with

$$(2k)^{-1} \leq q^{-1} \leq \beta|z| \leq P^{-1},$$

the lower bound being due to the fact that $u/q \neq \alpha/\beta$ because $\beta \leq P < q$ and $\gcd(u,q) = 1$. We can apply Proposition 1 that yields

$$\left|\left\{-k \leq t \leq k : \left\|t\frac{u}{q}\right\| \leq P^{-1}\right\}\right| \leq 3(8kP^{-1}+1) \leq 13(2l_1+1),$$

if q is large enough. Almost similarly (we have to consider separately the cases $t > 0$ and $t < 0$ but we get the upper bound $6(4kP^{-1}+1)$ and finally the same result), for any integer $0 \leq w \leq M = [N/k]$, we infer

$$\left|\left\{t \in \mathbb{Z}, wk \leq |t| \leq (w+1)k : \left\|t\frac{u}{q}\right\| \leq P^{-1}\right\}\right| \leq 13(2l_1+1);$$

thus, by putting $b_j = |a_j.\alpha|/k$ and

$$\mathcal{A}_w = \{j : w \leq b_j < w+1\},$$

for $w = 0, 1, \dots, M$, one has $|\mathcal{A}_w| \leq 13(2l_1+1)$. Now

$$\frac{\left(\sum_{j=1}^{m}(a_j.\alpha)^2\right)^2}{\left(\sum_{j=1}^{m}(a_j.\alpha)^4\right)} = \frac{\left(\sum_{j=1}^{m} b_j^2\right)^2}{\sum_{j=1}^{m} b_j^4} = \frac{\left(\sum_{w=0}^{M}\left(\sum_{j\in\mathcal{A}_w} b_j^2\right)\right)^2}{\sum_{j\in\mathcal{A}_0} b_j^4 + \sum_{w=1}^{M}\sum_{j\in\mathcal{A}_w} b_j^4}.$$

But, $\sum_{j\in\mathcal{A}_0} b_j^4 \leq |\mathcal{A}_0|$, thus

$$\sum_{w=1}^{M}\sum_{j\in\mathcal{A}_w} b_j^4 \geq \left|\bigcup_{w=1}^{M}\mathcal{A}_w\right| \geq |A|/2 \geq |\mathcal{A}_0| \geq \sum_{j\in\mathcal{A}_0} b_j^4,$$

because $|\mathcal{A}_0| \leq 13(2l_1+1) \leq 13v^{2/3} \leq |A|/2$ (this is due to the fact that $13 \leq (k_1/2)\log^{1/3} v$ for $v \geq k_4$). Therefore we obtain

$$\frac{\left(\sum_{j=1}^{m}(a_j.\alpha)^2\right)^2}{\left(\sum_{j=1}^{m}(a_j.\alpha)^4\right)} \geq \frac{\left(\sum_{w=1}^{M}|\mathcal{A}_w|w^2\right)^2}{2\sum_{w=1}^{M}(w+1)^4|\mathcal{A}_w|} \geq \frac{\left(\sum_{w=1}^{M}|\mathcal{A}_w|w^2\right)^2}{2^5\sum_{w=1}^{M}w^4|\mathcal{A}_w|} = \frac{F(C)}{2^5},$$

in the notations of the proof of Lemma 2 with $C \subset E = \{1^{(r)}, \dots, s^{(r)}\}$ and $r = 13(2l_1+1)$ and $s = M = [N/k]$. We have (q large), using inequality (29),

$$\begin{aligned} |E| = rs &\leq 13(2l_1+1)N/k \leq 52(2l_1+1)l_2 \leq 26v, \\ |E| &\geq 13(2l_1+1)(N/k-1) \geq 13(2l_1+1)(3l_2/2-1) \geq 2v. \end{aligned}$$

Thus $|E| \geq 2k_4$ that implies

$$\begin{aligned} |C| = |A| - |\mathcal{A}_0| \geq \frac{|A|}{2} \geq \frac{k_1}{2}v^{2/3}\log^{1/3} v &\geq \frac{k_1}{2}(|E|/26)^{2/3}\log^{1/3}(|E|/26) \\ &\geq \frac{k_1}{4(26)^{2/3}}|E|^{2/3}\log^{1/3}|E|. \end{aligned}$$

Consequently, thanks to Lemma 2, we get the lower bound

$$\frac{F(C)}{2^5} \geq \frac{k_{10}k_1^3}{1384448}\log|E| \geq \frac{k_{10}k_1^3}{1384448}\log v.$$

Second case. — We now consider degenerate cases, namely when $2l_1+1 \geq (2l_2+1)^2$ and $q/p \geq 2v^{1/3}/3$. This corresponds to cases where P_{l_1,l_2} is "thin" and α "almost orthogonal" to P_{l_1,l_2}. It requires a particular treatment.

We examine the case where

$$q/p \leq 2l_1 \tag{33}$$

and show that what has been done in the previous lines holds. We put $\epsilon = l_1/q$ and

$$k = [q(2l_2+1)^{2/3}/(2l_1+1)^{1/3}] \geq 1,$$

for large enough q. By using the Bezout relation, we see that

$$t\frac{u}{q} = t\frac{-v}{p} + \frac{t}{pq}.$$

We obtain

$$\begin{aligned}\left|\left\{-k \leq t \leq k : \left\|t\frac{u}{q}\right\| \leq \epsilon\right\}\right| &= \left|\left\{0 \leq t \leq 2k : \left\|\frac{tv}{p} + \frac{ku}{q} - \frac{t}{pq}\right\| \leq \epsilon\right\}\right| \\ &\leq \sum_{w=0}^{[2k/p]} \left|\left\{wp \leq t < (w+1)p : \left\|\frac{tv}{p} + \frac{ku}{q} - \frac{t}{pq}\right\| \leq \epsilon\right\}\right| \\ &\leq \sum_{w=0}^{[2k/p]} \left|\left\{wp \leq t < (w+1)p : \left\|\frac{tv}{p} + \frac{ku-w}{q}\right\| \leq \eta\right\}\right|\end{aligned}$$

where

$$\eta = \epsilon + q^{-1}.$$

But, as v is invertible modulo p, the number of solutions to $\left\|\frac{tv}{p} - c\right\| \leq \eta$ in a residue class modulo p is $\leq 2\eta p + 1$. Thus

$$\left|\left\{-k \leq t \leq k : \left\|t\frac{u}{q}\right\| \leq \epsilon\right\}\right| \leq (2\eta p + 1)(1 + 2k/p). \tag{34}$$

For q large enough, one has

$$\frac{k}{p} = \frac{\left[\frac{q(2l_2+1)^{2/3}}{(2l_1+1)^{1/3}}\right]}{p} \sim \frac{q(2l_2+1)^{2/3}}{p(2l_1+1)^{1/3}}$$

and this is

$$\geq (2v^{1/3}/3)\frac{(2l_2+1)^{2/3}}{(2l_1+1)^{1/3}} = 2(2l_2+1)/3 \geq 2.$$

Therefore for q large, $k/p \geq 1$. Concerning ηp, using the supplementary hypothesis (33), we have

$$2\eta p = 2\left(\frac{l_1+1}{q}\right)p \geq \frac{2l_1 p}{q} \geq 1,$$

thus (34) leads to

$$\left|\left\{-k \leq t \leq k : \left\|t\frac{u}{q}\right\| \leq \epsilon\right\}\right| \leq 4\eta p(3k/p) = 12\eta k$$
$$\leq 12\left(\frac{l_1+1}{q}\right)\left(\frac{(2l_2+1)^2}{2l_1+1}\right)^{1/3} q \leq 8v^{2/3}.$$

As above, we get the same result in the general case:

$$\left|\left\{wk \leq |t| < (w+1)k : \left\|t\frac{u}{q}\right\| \leq \epsilon\right\}\right| \leq 8v^{2/3},$$

and, with the same notation as before, we have

$$|E| \leq 8v^{2/3}(1+N/k) \leq 8v^{2/3}(1+2v^{1/3}) \leq 24v,$$

therefore we can conclude as previously.

Case $\alpha = (0,1)$. — To complete the result, we first establish it in the case where $\alpha = (0,1)$. In this case, we have

$$\frac{\left(\sum_{j=1}^{m}(a_j.\alpha)^2\right)^2}{\left(\sum_{j=1}^{m}(a_j.\alpha)^4\right)} = \frac{\left(\sum_{j\in J} a_{j,2}^2\right)^2}{\left(\sum_{j\in J} a_{j,2}^4\right)},$$

with $J = \{j : |a_{j,2}| \neq 0\}$. If $J_1 = \{|a_{j,2}|, j \in J\} \subset E = \{1^{(4l_1+2)}, \dots, l_2^{(4l_1+2)}\}$, hypothesis (28) applied to the line $\mathbb{R}\epsilon_1$ implies (since $v \geq |E|$)

$$|J_1| \geq k_2 v^{2/3}\log^{1/3} v \geq k_2|E|^{2/3}\log^{1/3}|E|,$$

which permits us to apply Lemma 2 with $c = k_2$ and to conclude that the fraction is

$$\geq k_{10}k_2^3 \log|E| \geq \frac{k_{10}k_2^3}{2}\log v,$$

because $|E| = (4l_1+2)l_2 \geq 2v/3$.

Extension of the formula. — Now, we extend the formula by continuity in the neighbourhood of $\alpha = (0,1)$ to fill the gap, namely we have to show that for every $0 \leq \theta \leq 1/2l_1$, the relation holds for the vector $(\theta, 1)$. But for any $a_{j,1}$, $|a_{j,1}\theta| \leq 1/2$, thus if we denote by K the set of j's such that $a_{j,2} = 0$ and by J its complementary (on which $|a_{j,2}|/2 \le |a_{j,1}\theta + a_{j,2}| \le 3|a_{j,2}|/2$), we obtain

$$\begin{aligned} F = \frac{\left(\sum_{j=1}^{m}(a_{j,1}\theta + a_{j,2})^2\right)^2}{\sum_{j=1}^{m}(a_{j,1}\theta + a_{j,2})^4} &\ge \frac{\left(\sum_{j\in K} a_{j,1}^2\theta^2 + \frac{1}{4}\sum_{j\in J} a_{j,2}^2\right)^2}{\sum_{j\in K} a_{j,1}^4\theta^4 + \left(\frac{3}{2}\right)^4 \sum_{j\in J} a_{j,2}^4} \\ &\ge \frac{S_{K,2}^2\theta^4 + S_{J,2}^2}{81(S_{K,4}\theta^4 + S_{J,4})} = g(\theta^4)/81, \end{aligned}$$

where $S_{K,2} = \sum_{j\in K} a_{j,1}^2, S_{J,2} = \sum_{j\in J} a_{j,2}^2, S_{K,4} = \sum_{j\in K} a_{j,1}^4, S_{J,4} = \sum_{j\in J} a_{j,2}^4$. This is a monotonic function g of θ^4. Thus

$$81F \ge \inf(g(0), g((1/2l_1)^4)).$$

We just estimated $g(0)$ (cf. Case $\alpha = (0,1)$), there remains to calculate $g((1/2l_1)^4)$,

$$g((1/2l_1)^4) \ge \frac{S'^2_{K,2} + S'^2_{J,2}}{S'_{K,4} + S'_{J,4}},$$

where $S'_{K,2} = \sum_{j\in K}(a_{j,1}/2l_1)^2, S'_{J,2} = S_{J,2}, S'_{K,4} = \sum_{j\in K}(a_{j,1}/2l_1)^4, S'_{J,4} = S_{J,4}$. We have now to consider two different cases.

If $S'_{K,4} \ge S'_{J,4}$ (this implies $|K| \ge |J|$ and consequently $|K| \ge |A|/2$), then writing

$$g((1/2l_1)^4) \ge \frac{S'^2_{K,2}}{2S'^2_{K,4}} = \frac{S^2_{K,2}}{2S_{K,4}},$$

we can apply the result of Lemma 2, since the cardinality of K', the set of j's in K such that $a_{j,1} \ne 0$ verifies

$$|K'| \ge |K| - 1 \ge |A|/3 \ge \frac{k_1}{3} v^{2/3}\log^{1/3} v$$

and

$$\{|a_{j,1}|, j \in K'\} \subset \{1^{(2)}, \dots, l_1^{(2)}\} \subset \{1^{(4l_2+2)}, \dots, l_1^{(4l_2+2)}\} = E.$$

We have $2v/3 \le (4l_2 + 2)l_1 = |E| \le v$, thus we obtain

$$g((1/2l_1)^4) \ge \frac{k_{10}(k_1/3)^3}{2}\log(2v/3) \ge \frac{k_{10}k_1^3}{100}\log v.$$

If $S'_{K,4} \le S'_{J,4}$ then

$$g((1/2l_1)^4) \ge \frac{S'^2_{J,2}}{2S'_{J,4}} = \frac{S^2_{J,2}}{2S_{J,4}}.$$

But $|J| \ge k_2 v^{2/3}\log^{1/3} v$ because of hypothesis (28) applied to the line $\mathbb{R}\epsilon_1$. As above $\{|a_{j,2}|, j \in J\} \subset E = \{1^{(4l_2+2)}, \dots, l_1^{(4l_2+2)}\}$ and $2v/3 \le |E| \le v$ that allows to obtain the lower bound

$$g((1/2l_1)^4) \ge \frac{k_{10}k_2^3}{2}\log|E| \ge \frac{k_{10}k_2^3}{4}\log v.$$

All this computation show that, in fact, we can take any

$$k_{11} \leq \inf\left\{\frac{k_1^3}{1384448}, \frac{k_2^3}{324}\right\} k_{10}$$

and ends the proof. □

2.3. Geometrical lemmas . — This section is devoted to the geometrical aspects of the problem. We complete Freiman's proof **[F96]** by studying every cases and improve some interesting intermediate results.

Through all this section we refer to **[C]** for extra information.

For C a compact convex body of $\mathbb{R}^2$, let us denote E its integer points, $E = C \cap \mathbb{Z}^2$. If Λ is a sub-lattice of $\mathbb{Z}^2$, we consider here $E \cap \Lambda$, which we assume to be two-dimensional, that is, not included in a line (this implies $|E| \geq 3$) and introduce some vocabulary and notation. If Δ is a line maximizing the cardinality $|\Delta \cap E \cap \Lambda|$, we write $\Delta \cap \mathbb{Z}^2 = \mathbb{Z}e_1$ for some e_1. Now, $\Delta \cap E \cap \Lambda = \{A_0 + k\alpha e_1, 0 \leq k \leq n\}$ for some point A_0, and α, n positive integers, because of the convexity of C. In the sequel, without loss of generality, we assume $A_0 = O$ and write $A = n\alpha e_1$. Next we choose e_2' completing e_1 in a $\mathbb{Z}^2$-basis, this is always possible. Now take β the unique (in view of $|\alpha\beta| =$ Vol Λ, the volume of a fundamental parallelogram of Λ) positive integer and γ' in $\mathbb{Z}$ such that

$$\Lambda = \alpha\mathbb{Z}e_1 + \mathbb{Z}(\beta e_2' + \gamma' e_1).$$

Define $u = \inf\{t \in \mathbb{Z}, \beta e_2' + te_1 \in E \cap \Lambda\}$ and $e_2 = e_2' + [u/\beta]e_1$. Then (e_1, e_2) is, as well, a $\mathbb{Z}^2$-basis and

$$\Lambda = \alpha\mathbb{Z}e_1 + \mathbb{Z}(\beta e_2 + \gamma e_1),$$

for some $\gamma \in \mathbb{Z}$. By definition of u, one can easily see that if $\beta e_2 + te_1 \in E \cap \Lambda$ then $t \geq \beta(u/\beta - [u/\beta]) \geq 0$. This remark will be needed in the sequel. Points of Λ are of the shape $(k\alpha + l\gamma)e_1 + l\beta e_2$ with $k, l \in \mathbb{Z}$. We note

$$d^+ = \max\ \{l | (k\alpha + l\gamma)e_1 + l\beta e_2 \in E \cap \Lambda \text{ for some } k\} \geq 0,$$

$$d^- = -\min\ \{l | (k\alpha + l\gamma)e_1 + l\beta e_2 \in E \cap \Lambda \text{ for some } k\} \geq 0.$$

Changing, if needed, e_2 in $-e_2$, one can assume that

$$d = \max\{d^+, d^-\} = d^+,$$

and since $E \cap \Lambda$ is not one-dimensional, $d \geq 1$. Clearly

$$|E \cap \Lambda| \leq (d^+ + d^- + 1)(n+1) \leq (2d+1)(n+1). \tag{35}$$

Finally, we note Δ_s the line $se_2 + \mathbb{R}e_1$ and define

$$c_s = |\Delta_s \cap E|,$$
$$c_s' = |\Delta_s \cap E \cap \Lambda|.$$

First we have to prove some preparatory lemmas.

***Lemma* 4.** — *With the preceding conditions and notation, we have*

(i) $\Delta_{(n+1)\beta} \cap E \cap \Lambda$ *is empty,*

(ii) *If* $n \geq 2$: $d \leq n+3$.

(iii) *If* $n = 1$: $d \leq 3$.

Proof. — We first prove (i). Assume that $\Delta_{(n+1)\beta} \cap E \cap \Lambda$ is not empty. We can find

$$P = (k\alpha + (n+1)\gamma)e_1 + (n+1)\beta e_2,$$

a point in that intersection. Performing the Euclidean division of k by $n+1$, we find an integer r between 0 and n such that $k = (n+1)q + r$ for some integer q. Now, the convexity of C shows, on the one hand, that $r\alpha e_1$ belongs to E, because it is located between O and $n\alpha e_1$ and, on the other hand, that any integer point on the segment joining

$$P = (n+1)(\beta e_2 + (q\alpha + \gamma)e_1) + r\alpha e_1$$

and $r\alpha e_1$ is in E too; in particular, for each integer s belonging to $\{0, \ldots, n+1\}$, the point $s(\beta e_2 + (q\alpha + \gamma)e_1)) + r\alpha e_1$ belongs to E. But now, these points are on a same line and their cardinality is $n+2$, which is impossible.

Let us now turn to (ii). Assume $n \geq 2$ and that there exists a positive integer j such that $\Delta_{(n+4+j)\beta} \cap E \cap \Lambda$ contains some point M. Convexity of C implies that the "full" triangle (AOM) is entirely in C. Let us denote L the length of the intersection of that triangle and $\Delta_{(n+1)\beta}$ (which is a segment). Application of Thales's Theorem gives

$$\frac{L}{n\alpha|e_1|} = \frac{3+j}{n+4+j},$$

that is

$$L = \frac{(3+j)n}{n+4+j}\alpha|e_1|,$$

an increasing expression with respect to n and j, which is therefore minimal when $n = 2$ and $j = 0$. It implies that $L \geq \alpha|e_1|$. But then $\Delta_{(n+1)\beta} \cap E \cap \Lambda$ contains at least one point and is subsequently not empty, contrarily to (i). Because each point of Λ is on some line $\Delta_{s\beta}$, one has $d \leq n+3$.

Now, let us see (iii): n is 1. If $d = 1$ or 2, there is nothing to prove. Assume we have $d \geq 3$ and choose M a point in $\Delta_{d\beta} \cap E \cap \Lambda$. Part (i) of the Lemma shows that $\Delta_{2\beta}$ does not intersect $E \cap \Lambda$, and then that the diameter of the intersection $\Delta_{2\beta} \cap C$ is less than $\alpha|e_1|$. Whence there exists a unique couple (S, T) of points of Λ verifying the properties

1. Non-void segment $\Delta_{2\beta} \cap C$ is included in the segment $[S, T]$,
2. $T = S + \alpha e_1$.

These points are of the following shape

$$S = (k\alpha + 2\gamma)e_1 + 2\beta e_2,$$
$$T = ((k+1)\alpha + 2\gamma)e_1 + 2\beta e_2 = S + \alpha e_1.$$

We now show that k is odd. In fact, if k were even, say $= 2l$, the point $S' = \frac{1}{2}S = (l\alpha + \gamma)e_1 + \beta e_2$ would be in Λ. Moreover, αe_1 and S' form a basis of Λ. But the "full" triangle (OAM) is contained in C, which forces the point M to be in the open strip (if not so, S or T would be in C) between the lines (OS) and (AT). Its coordinates are then of the form:

$$M = xS + y\alpha e_1,$$

with $x > 1$ and $0 < y < 1$. The point M can not belong to Λ, its coordinates not being entire (in the basis $(\alpha e_1, S')$). A contradiction.

The integer k is consequently odd, say $= 2l + 1$. If $T' = \frac{1}{2}T$, as before, we get that T' and αe_1 generate Λ. Writing

$$M = xS + y\alpha e_1,$$

with $x > 1$ and $0 < y < 1$, we get $M = 2xT' + (y - x)\alpha e_1$. But, because M is in Λ, $2x$ and $y - x$ are integers. The only possibility is $y = 1/2$ and $x = 1/2 + u$ with $u \in \mathbb{Z}$. Looking at the coordinate on e_2, we get $d = 1 + 2u$. But now, the convexity of C forces the integer points belonging to the line joining the middle of the segment $[O, A]$ to M to belong to E, in particular T'. As $M = T' + u(2T' - \alpha e_1)$ and since there is no line containing more than 2 points of $E \cap \Lambda$, one has $u \leq 1$ i.e. $d \leq 3$. □

Lemma 5. — *Let i, j, k be integers and t a real, $0 < t < 1$, such that $k = ti + (1-t)j$. Then the following holds*

(i) *if $\Delta_i \cap C, \Delta_j \cap C \neq \varnothing$, then one has $c_k \geq tc_i + (1-t)c_j - 2$,*

(ii) *if $\Delta_{\beta i} \cap C, \Delta_{\beta j} \cap C \neq \varnothing$, then one has $c'_{\beta k} \geq tc'_{\beta i} + (1-t)c'_{\beta j} - 2$,*

(iii) $c'_{\beta k} \leq 1 + \dfrac{c_{\beta k} - 1}{\alpha}$.

Proof. — Because of convexity, $\Delta_i \cap C$,$\Delta_j \cap C$ and $\Delta_k \cap C$ are non-empty segments. If l_i, l_j and l_k denote their respective length, one has, once again by convexity, $tl_i + (1-t)l_j \leq l_k$, but one has $l_i - |e_1| \leq c_i|e_1| \leq l_i + |e_1|$ (and the same for j and k), so we get

$$t(c_i - 1) + (1-t)(c_j - 1) \leq c_k + 1,$$

that is the first inequality.

The second one is similar (that is just a question of scale). And the third one is the consequence of an easy counting argument. □

Lemma 6. — *One has*

$$d \leq \frac{2}{\sqrt{\alpha}}|E|^{1/2} + 3 + \frac{4}{3\alpha}.$$

Proof. — We first notice that, because of $|E| \geq 3$, the formula is easily verified in the following cases: $\alpha = 1$ and $d \leq 6$, $\alpha = 2$ or 3 and $d \leq 4$ and $\alpha \geq 4$ and $d \leq 3$ and that except in those cases, which from now on we do not consider anylonger, one has $\alpha(d-3) - 3 \geq 1$. This remark is useful to make easier the forthcoming estimations. If $d \geq 4$, one can write

$$|E| \geq \sum_{k=0}^{\beta d} c_k \geq c_0 + \sum_{k=1}^{[\beta d/2]} (c_0/2 - 2),$$

because of (i) in Lemma 5. This way, we obtain the lower bound

$$|E| \geq c_0 + (c_0/2 - 2)[\beta d/2] \geq n\alpha + 1 + [\beta d/2](n\alpha - 3)/2,$$

with the inequality $c_0 = |\Delta_0 \cap E| \geq n\alpha + 1$; it implies, by virtue of Lemma 4, (ii), that

$$|E| \geq \alpha(d-3) + 1 + [\beta d/2]\frac{\alpha(d-3)-3}{2}.$$

But the preliminary remark of this Lemma ensures that the last fraction is positive. Since $\beta \geq 1$, we get the lower bound for $|E|$:

$$|E| \geq \alpha(d-3) + 1 + \left(\frac{d-1}{2}\right)\left(\frac{\alpha(d-3)-3}{2}\right),$$

which can be rewritten as follows:

$$4|E| \geq \alpha d^2 - 3d + 7 - 9\alpha.$$

It is easy to see that it implies

$$\begin{aligned} d &\leq \frac{3}{2\alpha} + \sqrt{\frac{4|E|}{\alpha} + 9 - \frac{7}{\alpha} + \frac{9}{4\alpha^2}} \\ &\leq \frac{3}{2\alpha} + \frac{2\sqrt{|E|}}{\sqrt{\alpha}} + \sqrt{9 - \frac{7}{\alpha} + \frac{9}{4\alpha^2}} \\ &\leq \frac{3}{2\alpha} + \frac{2\sqrt{|E|}}{\sqrt{\alpha}} + 3 - \frac{7}{6\alpha} + \frac{\sqrt{9/4 - 49/36}}{\alpha}, \end{aligned}$$

which implies the announced result. □

Lemma 7. — *We have*

$$\left(\sum_{i=-d^-+1}^{d^+-1} c_{i\beta}\right) - 4(d^+ + d^-) \geq -39.$$

Proof. — Suppose first that $d^- \geq 1$ (and thus $d^+ \geq 1$). For any positive integer i, we have, in view of Lemma 5, (i),

$$c_{i\beta} \geq \left(1 - \frac{i}{d^+}\right) c_0 + \frac{i}{d^+} c_{\beta d^+} - 2.$$

The same inequality holds, symmetrically, for the $c_{-i\beta}$'s (changing d^+ in d^-). By summing these inequalities, we get

$$\begin{aligned}\sum_{i=-d^-+1}^{d^+-1} c_{i\beta} &\geq c_0+\sum_{i=1}^{d^+-1}\left(\left(1-\frac{i}{d^+}\right)c_0+\frac{i}{d^+}c_{\beta d^+}-2\right)\\ &\quad +\sum_{i=1}^{d^--1}\left(\left(1-\frac{i}{d^-}\right)c_0+\frac{i}{d^-}c_{-\beta d^-}-2\right)\\ &= c_0+\left(\frac{c_0+c_{d+\beta}}{2}-2\right)(d^+-1)+\left(\frac{c_0+c_{d-\beta}}{2}-2\right)(d^--1)\\ &\geq c_0+\left(\frac{c_0+1}{2}-2\right)(d^++d^--2),\end{aligned}$$

where we have used the fact that $c_{d+\beta}, c_{d-\beta} \geq 1$ because of the non-emptiness of $\Delta_{d+\beta}\cap E$ and $\Delta_{d-\beta}\cap E$. Finally we get

$$\left(\sum_{i=-d^-+1}^{d^+-1} c_{i\beta}\right)-4(d^++d^-)\geq 3+\frac{c_0-11}{2}(d^++d^-).$$

Now, we consider two cases. If $c_0 \geq 11$, this is greater than 3. Or else, in view of $c_0 \geq n\alpha+1 \geq n+1$, and Lemma 4, (ii), this is $\geq 3+(n-10)(n+3)$ if $2 \leq n \leq 9$. This expression is minimal for $n=3$ or 4 and is in these cases equal to -39.

Assume now $d^-=0$ (recall $d=d^+$), the same inequalities as for the case $d^-\neq 0$ show that our expression is

$$\begin{aligned}\geq \sum_{i=1}^{d-1} c_{i\beta}-4d &\geq \left(\frac{c_0+1}{2}-2\right)(d-1)-4d\\ &= \frac{n-10}{2}(d-1)-4,\end{aligned}$$

which is ≥ -4 if $n \geq 10$ and if $n \leq 9$, this is $\geq (n+2)(n-10)/2-4 \geq -22$. This completes the proof. □

We are now ready to prove the first proposition of this section. It will be useful for Theorem 3 and algorithmical aspects of our problem but we think that it is an interesting result in itself.

Proposition 4. — *Let C be a compact convex body in $\mathbb{R}^2$ and E denote the set of its integer points. Assume E is not included in a line. Then, for each integer lattice Λ different from $\mathbb{Z}^2$, one has either $E\cap\Lambda$ included in a line, or $|E\cap\Lambda| \leq \frac{2}{3}|E|+39$.*

Freiman [**F96**] obtained a non-effective version of this result with a "reduction" factor 3/4 in place of our 2/3 which is the best possible, as one can see by considering the family depending on an integer parameter n:

$$E_n = C_n\cap\mathbb{Z}^2 = \{(i,j), 0\leq i\leq n-1, 0\leq j\leq 2\}\cup\{(n,0),(-1,0)\}$$

(where C_n denotes the convex hull of E_n) and the lattice

$$\Lambda = \mathbb{Z}e_1 + 2\mathbb{Z}e_2,$$

because then $|E_n| = 3n+2$, $|E_n \cap \Lambda| = 2n+2$ and consequently $|E_n \cap \Lambda| = \frac{2}{3}|E_n| + \frac{2}{3}$.

The constant 39 appearing in Proposition 4 seems to be larger than the one one might expect. Once again, the reason is that our computations are rough. Indeed it seems that one could expect a constant very near from 1. This problem of minimizing that constant seems to be open.

Note that the higher dimensional analogue to Proposition 4 is false, contrarily to what is announced in [**C91a**, Lemma 2], as can be seen by considering the following example. In $\mathbb{R}^3$ consider the points $a = (1,0,0), b = (0,-n,0), c = (0,n,0)$ and $d = (-1,0,2)$ for an integer parameter n. Let C'_n be the convex hull of these points

$$E'_n = \mathbb{Z}^3 \cap C'_n = \{(1,0,0),(0,0,1),(-1,0,2),(0,j,0), -n \leq j \leq n\}$$

and $\Lambda = \mathbb{Z}e_1 + \mathbb{Z}e_2 + 2\mathbb{Z}e_3$. One has $E'_n \cap \Lambda = E'_n \setminus \{(0,0,1)\}$ and thus

$$\frac{|E'_n \cap \Lambda|}{|E'_n|} = 1 - \frac{1}{|E'_n|}$$

that tends to 1 as n tends to infinity. At the same time, $E'_n \cap \Lambda$ is 3-dimensional. This shows that no strictly less than 1 analogue to the constant 2/3 exists in dimension 3.

Proof of Proposition 4. — If Λ is not $\mathbb{Z}^2$ then α or β is different from 1, that is at least 2.

First we consider the case where $\alpha \geq 2$. We can write, using Lemma 5 (iii),

$$\begin{aligned} |E \cap \Lambda| = \sum_{k=-d^-}^{d^+} c'_{\beta k} &\leq \sum_{k=-d^-}^{d^+} \left(1 + \frac{c_{\beta k} - 1}{\alpha}\right) \\ &= \left(1 - \frac{1}{\alpha}\right)(1 + d^- + d^+) + \frac{1}{\alpha} \sum_{k=-d^-}^{d^+} c_{\beta k} \\ &\leq \frac{|E|}{\alpha} + (2d+1). \qquad (36) \end{aligned}$$

In view of Lemma 6, equation (36) can be rewritten, because of $\alpha \geq 2$,

$$\begin{aligned} |E \cap \Lambda| &\leq \frac{|E|}{2} + 2\left(\sqrt{2}|E|^{1/2} + \frac{11}{3}\right) + 1 \\ &\leq \frac{|E|}{2} + 2\sqrt{2}|E|^{1/2} + \frac{25}{3}, \end{aligned}$$

which is bounded by $\frac{2}{3}|E| + 21$, as one can easily check.

Now we consider the case where $\alpha = 1$. Then one has $\beta \geq 2$ and $c'_{\beta k} = c_{\beta k}$. We have the trivial lower bound

$$|E| \geq \sum_{k=-\beta d^-}^{\beta d^+} c_k. \qquad (37)$$

We put for $0 \leq i \leq d^+ - 1$,

$$S_i^+ = \sum_{k=i\beta}^{(i+1)\beta-1} c_k,$$

and, likewise, for $0 \leq i \leq d^- - 1$,

$$S_i^- = \sum_{k=-i\beta}^{-(i+1)\beta+1} c_k;$$

then equation (37) becomes

$$\text{(38)} \qquad |E| + c_0 \geq (S_0^+ + \cdots + S_{d^+-1}^+ + c_{d^+\beta}) + (S_0^- + \cdots + S_{d^--1}^- + c_{-d^-\beta}).$$

But, application of Lemma 5, after summation, yields

$$S_i^+ = c_{i\beta} + \sum_{k=i\beta+1}^{(i+1)\beta-1} c_k \geq \frac{\beta+1}{2} c_{i\beta} + \frac{\beta-1}{2} c_{(i+1)\beta} - 2(\beta-1),$$

and, symmetrically,

$$S_i^- \geq \frac{\beta+1}{2} c_{-i\beta} + \frac{\beta-1}{2} c_{-(i+1)\beta} - 2(\beta-1),$$

so equation (38) implies

$$\begin{aligned} |E| + c_0 \quad \geq \quad & \left(\frac{\beta+1}{2} \sum_{i=0}^{d^+-1} c_{i\beta} + \frac{\beta-1}{2} \sum_{i=1}^{d^+} c_{i\beta} - 2(\beta-1)d^+ + c_{d^+\beta} \right) \\ & + \left(\underbrace{\frac{\beta+1}{2} \sum_{i=0}^{d^--1} c_{-i\beta} + \frac{\beta-1}{2} \sum_{i=1}^{d^-} c_{-i\beta}}_{=0 \text{ if } d^-=0} -2(\beta-1)d^- + c_{d^-\beta} \right), \end{aligned}$$

which takes the following simplified form

$$|E| \geq \frac{\beta+1}{2} \sum_{k=-d^-}^{d^+} c_{k\beta} + \frac{\beta-1}{2} \left(\sum_{k=-d^-+1}^{d^+-1} c_{k\beta} - 4(d^+ + d^-) \right).$$

But the first sum is $|E \cap \Lambda|$ and the second ≥ -39 in view of Lemma 7, so we have

$$|E| \geq \frac{\beta+1}{2} |E \cap \Lambda| - 39 \frac{\beta-1}{2}.$$

Thus,

$$|E \cap \Lambda| \leq \frac{2}{\beta+1} |E| + 39 \frac{\beta-1}{\beta+1} \leq \frac{2}{\beta+1} |E| + 39 \leq \frac{2}{3} |E| + 39.$$

□

From now on, we are only interested in E itself (which corresponds to $\Lambda = \mathbb{Z}^2$ or equivalently to $\alpha = \beta = 1$) for which we need two more lemmas. We keep the same notation as above but, for avoiding confusion, we put an index E so that d, n have nothing to do with d_E and n_E.

Lemma 8. — *Let k be an integer such that $1 \leq k \leq d_E$, then*

$$ke_2 + te_1 \in E \textit{ implies } -k+1 \leq t \leq n_E + k - 1.$$

If $-d_E^- \leq k \leq -1$, then

$$ke_2 + te_1 \in E \textit{ implies } k \leq t \leq 2n_E - k + 1.$$

Proof. — Remember that, by construction, $O, n_E e_1, e_2$ belong to E while $-e_1$, $e_2 - e_1$, $e_2 + (n_E + 1)e_1$ do not belong to E.

Let $k \geq 1$.

If $ke_2 + te_1 \in E$ then, as $O \in E$ too, one would have, by convexity

$$e_2 + \frac{t}{k}e_1 = \frac{1}{k}(ke_2 + te_1) + \frac{k-1}{k}O \in C.$$

Suppose $t \leq -k \leq 0$, then

$$e_2 - e_1 = -\frac{k}{t}\left(e_2 + \frac{t}{k}e_1\right) + \left(1 + \frac{k}{t}\right)e_2 \in E,$$

which is not true. Thus $t \geq -k + 1$.

On the other hand, if $ke_2 + te_1 \in E$, one would have also

$$e_2 + \left(\frac{t + (k-1)n_E}{k}\right)e_1 = \frac{1}{k}(ke_2 + te_1) + \left(\frac{k-1}{k}\right)n_E e_1 \in C.$$

Suppose $t \geq n_E + k$, then

$$e_2 + (n_E + 1)e_1 =$$
$$\left(\frac{(n_E + 1)k}{t + (k-1)n_E}\right)\left(e_2 + \left(\frac{t + (k-1)n_E}{k}\right)e_1\right) + \left(\frac{t - n_E - k}{t + (k-1)n_E}\right)e_2 \in E,$$

which is not true. Thus $t \leq n_E + k - 1$.

Now, let $k \leq -1$.

If $ke_2 + te_1 \in E$, then, since $e_2 \in E$,

$$\left(\frac{t}{1-k}\right)e_1 = \left(\frac{1}{1-k}\right)(ke_2 + te_1) + \left(\frac{-k}{1-k}\right)e_2 \in C.$$

Suppose $t \leq k - 1$, then

$$-e_1 = \left(\frac{k-1}{t}\right)\left(\frac{t}{1-k}e_1\right) + \left(\frac{t+1-k}{t}\right)O \in E,$$

which is false. Thus $t \geq k$.

Suppose $ke_2 + te_1 \in E$. It is known, by construction, that there is some point $d_E e_2 + xe_1 \in E$ and that, as previously seen, $x \geq 1 - d_E$. But then

$$\left(\frac{d_E t - xk}{d_E - k}\right)e_1 = \left(\frac{d_E}{d_E - k}\right)(ke_2 + te_1) + \left(\frac{-k}{d_E - k}\right)(d_E e_2 + xe_1) \in C.$$

Suppose $t \geq 2n_E + 2 - k$, since

$$\frac{d_E t - xk}{d_E - k} \geq \frac{d_E(2n_E + 2 - k) - k(1 - d_E)}{d_E - k} = \frac{2n_E d_E + 2d_E - k}{d_E - k} \geq n_E + 1,$$

the point $(n_E + 1)e_1$ is located on the segment joining O to $\left(\frac{d_E t - xk}{d_E - k}\right) e_1$ and is consequently in E, which is not true. Thus $t \leq 2n_E + 1 - k$. □

Lemma 9. — *One has*

$$|E| \geq n_E d_E/3.$$

Proof. — As in the previous proofs, one has:

$$\begin{aligned} |E| = c_0 + c_{d_E} + \sum_{i=1}^{d_E - 1} c_i &\geq (c_0 + c_{d_E})(d_E + 1)/2 - 2(d_E - 1) \\ &\geq (n_E + 2)(d_E + 1)/2 - 2(d_E - 1). \end{aligned}$$

Then $(|E| - n_E d_E/3) \geq 0$ follows from $(n_E - 6)d_E + 3n_E + 18 \geq 0$. For $n_E \geq 6$ this is trivially true and one checks that for $n_E \leq 6$, using Lemma 4,

$$(n_E - 6)d_E + 3n_E + 18 \geq n_E^2 \geq 0.$$

□

Now, we prove the second geometric lemma, which will be the key result for obtaining Theorem 2. We have here to remember that A_0 can be different from O.

Proposition 5. — *Let C be a compact convex body containing O, and E denote $C \cap \mathbb{Z}^2$. Then there exists a unimodular linear application ϕ and two integers l, m such that*

$$\phi(E) \subset P_{l,m},$$

with $v = (2l + 1)(2m + 1) \leq 345|E|$.

Proof. — Using the construction described before with $\alpha = \beta = 1$, we find a point A_0 and two integer vectors e_1, e_2 such that

$$\begin{aligned} C &\subset \{A_0 + \{-d_E, \dots, 2n_E + d_E + 1)\}e_1 + \{-d_E, \dots, d_E\}e_2\} \\ &\subset \{A_0 + \{-(n_E + 3), \dots, 3n_E + 4)\}e_1 + \{-d_E, \dots, d_E\}e_2\} \end{aligned}$$

in view of Lemma 8 and $d_E \leq n_E + 3$ (Lemma 4).

Recall that (ϵ_1, ϵ_2) is the canonical basis. Let ϕ be the linear transformation sending the $\mathbb{Z}^2$-basis $(e_i)_{i=1,2}$ onto the $\mathbb{Z}^2$-basis $(\epsilon_i)_{i=1,2}$. We have $\det \phi = \pm 1$ (ϕ is unimodular) and

$$\phi(C) \subset \{\phi(A_0) + \{-(n_E + 3), \dots, (3n_E + 4)\}\epsilon_1 + \{-d_E, \dots, d_E\}\epsilon_2\}.$$

But $\phi(O) = O \in C$ implies the existence of some r, s such that

$$O = \phi(A_0) + r\epsilon_1 + s\epsilon_2,$$

with $-(n_E + 3) \leq r \leq 3n_E + 4, |s| \leq d_E$. This fact shows that $\phi(C) \subset P_{4n_E+7, 2d_E}$, whose volume is $v = (8n_E + 15)(4d_E + 1) \leq 115 n_E d_E$. But, by Lemma 9, $|E| \geq n_E d_E/3 \geq v/345$, which ends the proof. □

3. Proof of Theorem 1

We assume without loss of generality that

$$2l_1+1 \geq v^{1/2} \geq 2l_2+1.$$

Let us remember that $v \geq k_4$. Define $m=|A|$ and write $m_0 = k_1 v^{2/3}\log^{1/3} v$.

The trivial orthogonality relations for e type functions easily yield

$$J(b) = 2^m \int_{[0,1]^2} \phi(\alpha)e(-b.\alpha)d\alpha = 2^m \int_{[-1/2,1/2]^2} \phi(\alpha)e(-b.\alpha)d\alpha,$$

with

$$\phi_j(\alpha) = \frac{1+e(a_j.\alpha)}{2}$$

and $\phi = \prod_{j=1}^m \phi_j$. In fact, we investigate $I = I(b) = J(b)/2^m$, by splitting the domain of integration into two parts, the major and minor arcs, corresponding respectively to the domains

$$K_0 = [-1/4l_1, 1/4l_1] \times [-1/4l_2, 1/4l_2]$$

and $K_1 = [-1/2,1/2]^2 \setminus K_0$. The corresponding integrals are denoted respectively I_0 and I_1.

3.1. The error term. — We have $|I_1| \leq \int_{K_1} |\phi(\alpha)|d\alpha \leq \mu(K_1)\sup_{\alpha\in K_1}|\phi(\alpha)|$. Applying (8) we get

$$|I_1| \leq \mu(K_1)\exp\left(-\frac{\pi^2}{2}\inf_{\alpha\in K_1}\sum_{j=1}^m ||\alpha.a_j||^2\right).$$

Our main aim in this section is to find a uniform lower bound on K_1 for the sum $\sum_{j=1}^m ||\alpha.a_j||^2$ appearing in the exponential. For this, let us use a Farey dissection of order Q of $[-1/2,1/2]\setminus[-1/4l_1,1/4l_1]$: we write (modulo 1) each α_1 of this interval as $p/q+z$ with

$$\gcd(p,q)=1, 0<q\leq Q \text{ and } |z| \leq 1/qQ.$$

Here we choose $Q=[k_{12}v/m]$, with

$$k_{12} = 69.$$

We notice that

$$\frac{2l_1+1}{Q} \geq \frac{v^{1/2}}{k_{12}v/m} \geq \frac{m_0}{k_{12}v^{1/2}} \geq \frac{k_1}{k_{12}} v^{1/6}\log^{1/3} v > 2, \tag{39}$$

because $v \geq k_4$.

Now, we distinguish different cases.

3.1.1. $|z| \geq 1/4ql_1$. — Using Proposition 1 with $P = Q$, $k = 2l_1$ which is $\geq Q$, $a = \alpha_1$, $b = -\alpha_1 l_1 + \alpha_2 n_2$ and $n = -l_1$, we get for each n_2 subject to $-l_2 \leq n_2 \leq l_2$:

$$|\{-l_1 \leq n_1 \leq l_1 : ||\alpha.n|| \leq Q^{-1}\}| \leq 3(8l_1 Q^{-1} + 1).$$

Thus we get, after summation on n_2,

$$|\{n \in P_{l_1,l_2} : ||\alpha.n|| \leq Q^{-1}\}| \leq 3(8l_1 Q^{-1} + 1)(2l_2 + 1),$$

and this is

$$\leq 3\left(\frac{8l_1 + 4}{Q} + \frac{2l_1 + 1}{2Q}\right)(2l_2 + 1) = 13.5vQ^{-1},$$

in view of (39).

3.1.2. $|z| < 1/4ql_1$. — Write $\alpha_2 = (h+\theta)/q$ where $h \in \mathbb{Z}, 0 \leq \theta < 1$ and $\theta = p'/q' + z'$ by using a Farey dissection of order $4l_2$ (therefore $q' \leq 4l_2$ and $|z'| \leq 1/4q'l_2$). We have

$$\begin{aligned}\alpha.n &= \left(\frac{p}{q} + z\right)n_1 + \left(\frac{h}{q} + \frac{p'}{qq'} + \frac{z'}{q}\right)n_2 \\ &= \frac{q'(pn_1 + hn_2) + p'n_2}{qq'} + zn_1 + \frac{z'n_2}{q}.\end{aligned}$$

If $D_{a,b}$ denotes the set $\{aq, aq+1, \ldots, (a+1)q - 1\} \times \{bq', bq'+1, \ldots, (b+1)q' - 1\}$, define Ψ the application

$$\begin{array}{rccc}\Psi : & D_{a,b} & \longrightarrow & \mathbb{Z}/qq'\mathbb{Z} \\ & (x,y) & \longmapsto & q'(px + hy) + p'y.\end{array}$$

It is easy to check that Ψ is bijective (injectivity is just a trivial consequence of $\gcd(p', q') = 1$).

Finally, equation $||\alpha.n|| \leq Q^{-1}$ implies

$$\left|\left|\frac{\Psi(n)}{qq'}\right|\right| \leq Q^{-1} + |z|n_1 + \frac{|z'|n_2}{q} \leq Q^{-1} + |z|l_1 + \frac{1}{4qq'}, \tag{40}$$

in view of $|zn_1| \leq |z|l_1$ and $|z'n_2/q| \leq 1/4qq'$.

3.1.2.1. *Case 1:* $Q^{-1} \geq |z|l_1, 1/4qq'$. — We have

$$Q^{-1} + |z|l_1 + \frac{1}{4qq'} \leq 3Q^{-1},$$

and since Ψ bijective,

$$\begin{aligned}|\{n \in P_{l_1,l_2} : ||\alpha.n|| \leq Q^{-1}\}| &\leq |\{n \in P_{l_1,l_2} : ||\Psi(n)/qq'|| \leq 3Q^{-1}\}| \\ &\leq (1 + 6Q^{-1}qq')([2l_1/q] + 1)([2l_2/q'] + 1),\end{aligned}$$

because $[2l_1/q] + 1$ is the maximal number of integers between $-l_1$ and l_1 having same residue modulo q.

Now, $1 + 6Q^{-1}qq' \leq 10Q^{-1}qq'$ by hypothesis of case 1. One has

$$[2l_1/q] + 1 \leq \frac{2l_1 + 1}{q} + 1 \leq \frac{3}{2}\left(\frac{2l_1 + 1}{q}\right),$$

in view of (39). And $[2l_2/q'] + 1 \leq 3(2l_2+1)/q'$ because $q' \leq 4l_2$. Finally,

$$|\{n \in P_{l_1,l_2} : ||\alpha.n|| \leq Q^{-1}\}| \leq 10Q^{-1}qq'(3(2l_1+1)/2q)(3(2l_2+1)/q') \leq 45vQ^{-1}.$$

3.1.2.2. *Case 2:* $Q^{-1}, |z|l_1 < 1/4qq'$. — In this case, a solution of $||\alpha.n|| \leq Q^{-1}$ verifies $||\Psi(n)/qq'|| < 1/qq'$, but this implies $\Psi(n) = 0 \bmod qq'$, that is

$$q'(pn_1 + hn_2) + p'n_2 \equiv 0 \bmod qq'.$$

This implies that $q'|n_2$, so that, when $q' \neq 1$, n belongs to some lattice different from $\mathbb{Z}^2$ (namely $\mathbb{Z}\epsilon_1 + \mathbb{Z}q'\epsilon_2$).

If $q' = 1$, one has $pn_1 + (h+p')n_2 \equiv 0 \bmod q$, which is the equation of a lattice different from $\mathbb{Z}^2$ as soon as $q \neq 1$, but that is the case because if $q = q' = 1$, one has $p = p' = 0$ (and then $h = 0$), and $|z| < 1/4l_1, |z'| \leq 1/4l_2$. Therefore $|\alpha_i| < 1/4l_i$ for $i = 1, 2$, which shows that, in this case, α is not in K_1.

Consequently, we can bound the number of solutions in this case, using hypothesis (3):

$$|\{n \in P_{l_1,l_2} : ||\alpha.n|| \leq Q^{-1}\}| \leq m - k_2v^{2/3}\log^{1/3} v.$$

3.1.2.3. *Case 3:* $|z|l_1 \geq Q^{-1}, 1/4qq'$. — If the integer vector n satisfies

$$||\alpha.n|| \leq Q^{-1}, \tag{41}$$

it satisfies a fortiori

$$||\Psi(n)/qq'|| \leq 3|z|l_1 \tag{42}$$

and so the number of couples of residues (x_0, y_0), modulo q and q' respectively, solutions to (42), is less than $1 + 6|z|l_1qq'$.

Let us now give an upper bound for the number of solutions of (41) with n_1 restricted to be equal to some x_0 modulo q and n_2 fixed. The equation becomes

$$\left|\left|\frac{pn_1}{q} + z(x_0 + qt) + \alpha_2 n_2\right|\right| \leq Q^{-1};$$

this is of the form

$$||\eta + zqt|| \leq Q^{-1},$$

for which we can apply Lemma 1 with $a = zq, b = \eta, \epsilon = Q^{-1} \leq k_{12}^{-1} < 1/6, k = [2l_1/q] + 1$. We have $k|a| \leq |z|q(1 + 2l_1/q) = |z|q + 2l_1|z| < Q^{-1} + 1/2q \leq Q^{-1} + 1/2$, by hypothesis of section 3.1.2 and thus $(1 - k|a|)/2 > \epsilon$ is verified. We get the upper bound $1 + [2/Q|z|q]$. Consequently, if now, n_1 and n_2 are restricted to be constant modulo q and q' respectively, the number of solutions of (41) is

$$\leq \left(1 + \left[\frac{2l_2}{q'}\right]\right)\left(1 + \frac{2}{Q|z|q}\right).$$

Finally, the total number of possible solutions to (41) is bounded above by

$$(1 + 6|z|l_1qq')\left(1 + \left[\frac{2l_2}{q'}\right]\right)\left(1 + \frac{2}{Q|z|q}\right) \leq 45vQ^{-1},$$

by using $|z|l_1qq' \geq 1/4$, $|z| \leq 1/qQ$ and $q' \leq 4l_2$.

3.1.3. *Conclusion.* — In the cases of sections 3.1.1 and 3.1.2 cases 1 and 3, the total number of solution to (41) is bounded by

$$45vQ^{-1} \leq \frac{45}{k_{12}-1}m.$$

Consequently,

$$|\{n \in A, ||\alpha.n|| \geq Q^{-1}\}| \geq \frac{k_{12}-46}{k_{12}-1}m.$$

Thus we get

$$\sum_{j=1}^{m} ||\alpha.a_j||^2 \geq \frac{k_{12}-46}{k_{12}-1}mQ^{-2} \geq \frac{k_{12}-46}{(k_{12}-1)k_{12}^2}\frac{m^3}{v^2} \geq \frac{(k_{12}-46)k_1^3}{(k_{12}-1)k_{12}^2}\log v \geq 0.75\log v.$$

In the case of section 3.1.2, case 2, we have

$$|\{n \in A, ||\alpha.n|| \geq Q^{-1}\}| \geq k_2 v^{2/3}\log v,$$

thus

$$\sum_{j=1}^{m} ||\alpha.a_j||^2 \geq \left(\frac{m}{k_{12}v}\right)^2 k_2 v^{2/3}\log^{1/3} v \geq \frac{k_1^2 k_2}{k_{12}^2}\log v \geq 0.75\log v.$$

Finally,

$$|I_1| \leq \mu(K_1)\exp\left(-\frac{\pi^2}{2}\inf_{\alpha\in K_1}\sum_{j=1}^{m}||\alpha.a_j||^2\right) \leq \mu(K_1)/v^3.$$

3.2. The major part. — Here, we have to investigate $I_0 = \int_{K_0}\phi(\alpha)e(-b.\alpha)d\alpha$. Let us denote

$$K = \{\alpha \in K_0 : V(\alpha) \leq 0.75\log v\}$$

and $K' = K_0 \setminus K$. The contribution of K' can be evaluated as follows

$$\begin{aligned}\left|\int_{K'}\phi(\alpha)e(-b.\alpha)d\alpha\right| &\leq \int_{K'}\prod_{j=1}^{m}|\phi_j(\alpha)|d\alpha \leq \mu(K')\exp\left(-\pi^2\sum_{j=1}^{m}||\alpha.a_j||^2/2\right)\\ &\leq \mu(K')\exp\left(-\pi^2 V(\alpha)/2\right) \leq \mu(K')/v^{3\pi^2/8},\end{aligned}$$

in view of the definition of V and because on K_0, $|\alpha_i a_{j,i}| \leq 1/4$ so that $||\alpha.a_j|| = |\alpha.a_j| \leq 1/2$.

Now, rewrite $\phi(\alpha)e(-b.\alpha)$ as follows

$$\phi(\alpha)e(-b.\alpha) = e((M-b).\alpha)\prod_{j=1}^{m}\cos(\pi a_j.\alpha),$$

but for $|\pi a_j.\alpha| \leq \pi/2$, one can write, in view of (9),

$$\cos(\pi a_j.\alpha) = \exp(-\pi^2(a_j.\alpha)^2/2)(1-g(a_j.\alpha)),$$

with

$$0 \leq g(a_j.\alpha) \leq (2\pi a_j.\alpha/\pi)^4 \leq 1.$$

Finally the major part is

$$\int_K \phi(\alpha)e(-b.\alpha)d\alpha = \int_K \exp\left(-\frac{\pi^2}{2}\sum_{j=1}^m (a_j.\alpha)^2 + 2i\pi(M-b).\alpha\right)(1-R(\alpha))d\alpha,$$

where $R(\alpha) = 1 - \prod_{j=1}^m (1 - g(a_j.\alpha))$. One can write this integral $A_0 - A_1 - A_2$ by splitting it into three parts

$$\begin{aligned} A_0 &= \int_{\mathbb{R}^2} \exp\left(-\frac{\pi^2}{2}V(\alpha) + 2i\pi(M-b).\alpha\right)d\alpha, \\ A_1 &= \int_{\mathbb{R}^2\setminus K} \exp\left(-\frac{\pi^2}{2}V(\alpha) + 2i\pi(M-b).\alpha\right)d\alpha, \\ A_2 &= \int_K R(\alpha)\exp\left(-\frac{\pi^2}{2}V(\alpha) + 2i\pi(M-b).\alpha\right)d\alpha. \end{aligned}$$

Let us write $d_i = M_i - b_i$ and investigate these three integrals. By the change of variables

$$x = \frac{\pi}{V_1}(V_1^2\alpha_1 + V_{12}\alpha_2), y = \pi\frac{\sqrt{\det V}}{V_1}\alpha_2, \tag{43}$$

we get

$$\begin{aligned} A_0 &= \frac{1}{\pi^2\sqrt{\det V}}\int_{\mathbb{R}}\exp\left(-\frac{x^2}{2}+i\frac{2d_1x}{V_1}\right)dx\int_{\mathbb{R}}\exp\left(-\frac{y^2}{2}+i\frac{2(d_2V_1^2-d_1V_{12})y}{V_1\sqrt{\det V}}\right)dy \\ &= \frac{1}{\pi^2\sqrt{\det V}}\sqrt{2\pi}\exp\left(-\frac{1}{2}\left(2\frac{d_1}{V_1}\right)^2\right)\sqrt{2\pi}\exp\left(-\frac{1}{2}\left(2\frac{d_2V_1^2-d_1V_{12}}{V_1\sqrt{\det V}}\right)^2\right) \\ &= \frac{2\exp(-2q_{V^{-1}}(M-b))}{\pi\sqrt{\det V}}. \end{aligned}$$

An upper bound for $|A_1|$ can be achieved by noticing that if $\alpha \notin K$ then $V(\alpha) \geq 0.75\log v$. Suppose $\alpha \notin \mathbb{Z}^2 + K_0$. We have

$$V(\alpha) = \sum_{i=1}^m (\alpha.a_j)^2 \geq \sum_{j=1}^m ||\alpha.a_j||^2.$$

Since the last sum is not changed by an integral translation of α, the problem is reduced to the study of this sum for $\alpha \in K_1$. This has been done in the preceding section and thus we get the lower bound $0.75\log v$. If now $\alpha \in (\mathbb{Z}^2 \setminus \{(0,0)\}) + K_0$, it can be written as $\alpha = h + \epsilon$ with $h \in \mathbb{Z}^2 \setminus \{(0,0)\}$ and $\epsilon \in K_0$. Since $|\epsilon_i| \leq 1/4l_i$, $|\epsilon.a_j| \leq 1/2$ and in view of hypothesis (3) at least $k_2v^{2/3}\log^{1/3}v$ elements a_j of A verify $h.a_j \neq 0$ (cf. hypothesis (28)) thus $(\alpha.a_j)^2 \geq 1/4$ for these values and we obtain

$$V(\alpha) \geq \frac{k_2}{4}v^{2/3}\log^{1/3}v \geq 0.75\log v.$$

This ends the proof of the lower bound of $V(\alpha)$ on the complementary of K. Now, the change of variables (43) and a polar change of variables produce

$$|A_1| \leq \frac{1}{\pi^2\sqrt{\det V}} \int_{\pi\sqrt{3\log v/4}}^{+\infty} r\exp(-r^2/2)dr \int_0^{2\pi} d\theta,$$

and finally

$$|A_1| \leq 2/(\pi\sqrt{\det V} v^{3\pi^2/8}).$$

Now, we consider A_2. We have, in view of inequality (10),

$$|R(\alpha)| = |1 - \prod_{j=1}^{m}(1 - g(a_j.\alpha))| \leq \sum_{j=1}^{m} g(a_j.\alpha) \leq 16\sum_{j=1}^{m}(a_j.\alpha)^4,$$

the last inequality being due to (9). Here is the place where we need Proposition 2. We get

$$\begin{aligned}|A_2| &\leq 16\int_K \left(\sum_{j=1}^{m}(a_j.\alpha)^4\right)\exp\left(-\frac{\pi^2}{2}V(\alpha)\right)d\alpha \\ &\leq \frac{16}{k_{11}\log v}\int_{\mathbb{R}^2} V(\alpha)^2\exp\left(-\frac{\pi^2}{2}V(\alpha)\right)d\alpha.\end{aligned}$$

In the same way as above we obtain finally

$$|A_2| \leq \frac{256}{k_{11}\pi^5\sqrt{\det V}\log v} \leq \frac{750}{\log v\sqrt{\det V}}.$$

3.3. Conclusion. — The dominant term is

$$A_0 = \frac{2}{\pi\sqrt{\det V}}\exp(-2q_{V^{-1}}(M-b)).$$

The error term is bounded from above by

$$\frac{\mu(K_1)}{v^3} + \frac{\mu(K')}{v^{3\pi^2/8}} + \frac{2}{\pi\sqrt{\det V}v^{3\pi^2/8}} + \frac{750}{\log v\sqrt{\det V}} \leq \frac{800}{\log v\sqrt{\det V}},$$

using $|\det V| \leq m^2v^2/4 \leq v^4/4$. It is readily seen that if $q_{V^{-1}}(M-b) \leq k_3\log\log v - 4$ then the main term A_0 is $\geq 1800/\sqrt{\det V}\log^{2k_3} v$. At the same time the error term is

$$\leq \frac{800}{\sqrt{\det V}\log v} = o(A_0),$$

(with a constant 1) thanks to our choice for k_3. This concludes the proof of Theorem 1.

4. Proof of Theorems 2 and 3

4.1. Theorem 2. — By Proposition 5, we can find a linear application ϕ sending C onto P_{l_1,l_2} for some l_1, l_2 with $v \leq 345|E|$. We can assume that $v \geq |E|$ with no loss of generality. Consequently,

$$\begin{aligned}|\phi(A)| = |A| \geq k_5|E|^{2/3}\log^{1/3}|E| &\geq k_5\left(\frac{v}{345}\right)^{2/3}\log^{1/3}\left(\frac{v}{345}\right)\\ &\geq \frac{k_5}{100}v^{2/3}\log^{1/3}v = k_1 v^{2/3}\log^{1/3}v.\end{aligned}$$

Since the linear transformation ϕ sends lattices onto lattices we have

$$|\phi(A) \setminus (\phi(A) \cap \Gamma)| = |A \setminus (A \cap \phi^{-1}(\Gamma))| \geq k_6|E|^{2/3}\log^{1/3}|E|,$$

by hypothesis (6). As above this is

$$\geq \frac{k_6}{100}v^{2/3}\log^{1/3}v = k_2 v^{2/3}\log^{1/3}v.$$

Applying Theorem 1 to the set $\phi(A)$, the asymptotic formula (4) is changed in (J' stands for the number of solutions to the boolean equation induced by $\phi(A)$)

$$J(b) = J'(\phi(b)) \sim \frac{2^{m+1}}{\pi\sqrt{\det W}}\exp\{-2q_{W^{-1}}(\phi(M) - \phi(b))\},$$

for any b such that $q_{W^{-1}}(\phi(M) - \phi(b)) \leq k_3 \log\log v - 4$, where W is the matrix obtained with the $\phi(a_j)$'s instead of the a_j's that is to say $W = \phi V \phi^t$. Thus $\det W = \det^2\phi\ \det V = \det V$, $q_{W^{-1}}(\phi(M) - \phi(b)) = q_{V^{-1}}(M - b)$ and we can take $k_7 = k_3$ due to $\log\log|E| \leq \log\log v$. □

From Theorem 2 we deduce the following result.

Corollary 1. — *Let C be a compact convex set in $\mathbb{R}^2$ containing O, Λ be an integer lattice and $E = C \cap \Lambda$. Let A be a subset of E. Assume*

$$|A| \geq k_5|E|^{2/3}\log^{1/3}|E|$$

and that for each Γ sub-lattice of Λ different from $\mathbb{Z}^2$, we have

$$|A \setminus A \cap \Gamma| \geq k_6|E|^{2/3}\log^{1/3}|E|. \tag{44}$$

Then we have the following asymptotic equivalent (when $|E| \to +\infty$)

$$J(b) \sim \frac{2^{m+1}\ \mathit{Vol}\,\Lambda}{\pi\sqrt{\det V}}\exp\{-2q_{V^{-1}}(M - b)\}, \tag{45}$$

provided that $q_{V^{-1}}(M - b) \leq k_7 \log\log|E| - 4$.

Proof of the Corollary. — Take a basis of Λ and Ψ a linear application sending this basis onto the canonical basis of $\mathbb{Z}^2$. If $A' = \Psi(A), E' = \Psi(C) \cap \mathbb{Z}^2 = \Psi(E)$ and since the sub-lattices of Λ are sent onto integer lattices, we can apply Theorem 2 to $\Psi(C)$ and $\Psi(A)$. With the same computation as above, we get the asymptotic equivalent (45), the factor Vol Λ being due to the formula of change of basis for quadratic forms and $|\det\Psi| = \text{Vol }\Lambda$. □

4.2. Theorem 3. — Either condition (6) is fulfilled by A and we are done by applying Theorem 2 (notice $k_8 \geq k_5$) or there is an integral lattice Γ_1 such that

$$|A \cap \Gamma_1| \geq |A| - k_6 |E|^{2/3} \log^{1/3} |E|. \tag{46}$$

Write $\Gamma_0 = \mathbb{Z}^2$, $A_i = A \cap \Gamma_i$ and $E_i = E \cap \Gamma_i$. Since A_1 is not contained in a line in view of hypothesis (7) ($|A_1| \geq \left(1 - \frac{k_6}{k_8}\right) |A|$ and $k_9 < 1 - k_6/k_8$) we have, by Proposition 4, $|E_1| \leq \frac{2}{3}|E| + 39 \leq 0.7|E_0|$ (the smallest possible value of $|E|$ allows to write this). Therefore we get

$$|A_1| \geq k_8 |E_1|^{2/3} \log^{1/3} |E_1| \tag{47}$$

in view of equation (46) and $(0.7)^{2/3} k_8 \leq k_8 - k_6$.

Now either A_1 verifies condition (44) of Corollary 1 and we stop here the process, or there exists a lattice $\Gamma_2 \subset \Gamma_1$ violating (44).

Let us show that, more generally, we can construct a decreasing finite sequence of lattices $(\Gamma_i)_{1 \leq i \leq p}$: assume we have already built $\Gamma_1, \ldots, \Gamma_i$ and $A_1, \ldots, A_i$. If condition (44) is fulfilled for $A_i \subset E_i$ then we stop the process else we find a lattice $\Gamma_{i+1} \subset \Gamma_i$ violating (44).

Lemma 10. — *We have for each* i

$$|A_i| \geq k_8 |E_i|^{2/3} \log^{1/3} |E_i| \text{ and } |E_i| \geq |A_i| > |A|/2.$$

Proof. — For $i = 1$, equation (47), $k_8 > 2k_6$ and (46) prove the result. Assume that the result is true for $1, 2, \ldots, i$ and that we have built Γ_{i+1}, E_{i+1} and A_{i+1}. One has

$$|A_{i+1}| \geq |A_i| - k_6 |E_i|^{2/3} \log^{1/3} |E_i| \geq \\ (k_8 - k_6)|E_i|^{2/3} \log^{1/3} |E_i| \geq k_8 |E_{i+1}|^{2/3} \log^{1/3} |E_{i+1}|$$

in view of Proposition 4 (here we used the fact that $|A_{i+1}| \geq (1 - k_6/k_8)|A_i| \geq \frac{1}{2}(1 - k_6/k_8)|A| \geq k_9 |A|$ which implies first that E_{i+1} is not included in a line (in view of (7)) and, second, that $|E_i|$ is large enough). Now

$$\begin{aligned} |A_{i+1}| &\geq |A| - k_6 \sum_{j=0}^{i} |E_j|^{2/3} \log^{1/3} |E_j| \\ &\geq |A| - k_6 \sum_{j=0}^{\infty} (0.7^j |E|)^{2/3} \log^{1/3} |E| \\ &\geq |A| - 5k_6 |E|^{2/3} \log^{1/3} |E| > |A|/2 \end{aligned}$$

due to the definition of k_6 and k_8. □

This Lemma shows that the process is well defined. As it is clearly finite (since $(|E_i|)_{1 \leq i \leq p}$ is a strictly decreasing sequence and E_i is never included in a line in view of hypothesis (7) and Lemma 10) and at the end we have a lattice $\Lambda_0 = \Gamma_p$, $A_p = A \cap \Gamma_p$ and $E_p = E \cap \Gamma_p$ such that condition (44) and the cardinality condition are fulfilled. Thus we can apply the Corollary to Theorem 2 which gives the result.

References

[AF88] Alon N. and Freiman G. A., *On sums of subsets of a set of integers*, Combinatorica, **8 (4)**, 1988, 297–306.

[C] Cassels J. W. S., *An introduction to the geometry of numbers*, Springer Verlag, 1971.

[C91a] Chaimovich M., *On solving dense n-dimensional subset sum problems*, Congressus Numerantium, **84**, 1991, 41–49.

[C91b] Chaimovich M., *Analytical methods of number theory in integer programming*, PhD, University of Tel-Aviv, 1991.

[CFG89] Chaimovich M., Freiman G. A. and Galil Z., *Solving dense subset sum problem by using analytical number theory*, J. of Complexity, **5**, 1989, 271–282.

[EF90] Erdős P. and Freiman G. A., *On two additive problems*, J. Number Theory, **34**, 1990, 1–12.

[F80] Freiman G. A., *An analytical method of analysis of linear boolean equations*, Ann. New-York Acad. Sci., **337**, 1980, 97–102.

[F93] Freiman G. A., *New analytical results in subset sum problem*, Discrete Math., **114**, 1993, 205–217. For erratum, see Discrete Math., **126**, 1994, 447.

[F96] Freiman G. A., *On solvability of a system of two boolean linear equations*, Number Theory: New-York Seminar 1991-1995, Springer-Verlag, 1996, 135–150.

[HW] Hardy G. W. and Wright E. M., *An introduction to the theory of numbers*, 5th ed., Oxford University Press, 1979.

A. Plagne, Algorithmique Arithmétique Expérimentale, CNRS UMR 9936,
Université Bordeaux I, 351 cours de la Libération, 33405 Talence Cedex, FRANCE
E-mail : plagne@math.u-bordeaux.fr

Astérisque

JEAN-MARC DESHOUILLERS
GREGORY A. FREIMAN
WILLIAM MORAN

On series of discrete random variables, 1 : real trinomial distributions with fixed probabilities

Astérisque, tome 258 (1999), p. 411-423
<http://www.numdam.org/item?id=AST_1999__258__411_0>

Astérisque **258**, 1999, p. 411–423

ON SERIES OF DISCRETE RANDOM VARIABLES, 1: REAL TRINOMIAL DISTRIBUTIONS WITH FIXED PROBABILITIES

by

Jean-Marc Deshouillers, Gregory A. Freiman & William Moran

Abstract. — This paper begins the study of the local limit behaviour of triangular arrays of independent random variables $(\zeta_{n,k})_{1\leq k\leq n}$ where the law of $\zeta_{n,k}$ depends on on n. We consider the case when $\zeta_{n,1}$ takes three integral values $0 < a_1(n) < a_2(n)$ with respective probabilities p_0, p_1, p_2 which do not depend on n. We show three types of limit behaviours for the sequence of r. v. $\eta_n = \zeta_{n,1} + \cdots + \zeta_{n,n}$, according as $a_2(n)/\gcd(a_1(n), a_2(n))$ tends to infinity slower, quicker or at the same speed as $\sqrt{n}$.

These notes are a first step in the description of the local behaviour of *series* of discrete random variables, that is to say sequences $(\eta_1, \ldots, \eta_n, \ldots)$ of random variables such that η_n is the sum of n independent discrete random variables $(\xi_{n,k})_{1\leq k\leq n}$ following a same law that may depend on n. We are restricting ourselves to the case when the $\xi_{n,k}$'s take three integer values $a_0 = 0 < a_1(n) < a_2(n)$, where a_1 and a_2 are coprime, with fixed positive probabilities p_0, p_1 and p_2 respectively.

When the values $a_1(n)$ and $a_2(n)$ do not depend on n, it follows from a result of Gnedenko that we have a *local limit* result, namely

$$P\{\eta_n = N\} = \frac{1}{\sigma\sqrt{2\pi n}}\left(\exp\left(-\frac{(n\mu - N)^2}{2n\sigma^2}\right) + o(1)\right) \qquad \text{as } n \to \infty,$$

uniformly in N, where μ and σ^2 are the expectation and the variance of the $\xi_{n,k}$'s.

Our aim is to give a complete description of the case when a_1 and a_2 depend on n, showing that there exist three different behaviors according as $a_2(n)$ is bounded or tends to infinity slower than $\sqrt{n}$, tends to infinity at the same speed as $\sqrt{n}$, or tends to infinity quicker than $\sqrt{n}$. In the first case, we get a *local limit* result similar to

1991 ***Mathematics Subject Classification.*** — 60F05, 60E10, 11P55, 11Z05.

Key words and phrases. — Sums of discrete random variables, local limit theorems, characteristic function, number theoretic methods, circle method.

J.-M. D: Cette recherche a bénéficié du soutien du CNRS (UMR 9936, Université Bordeaux 1) et de l'Université Victor Segalen Bordeaux 2.

the one we just quoted. The second case leads to a result of a similar structure with a limiting law which is no more normal. The third case can be seen as *isomorphic* to a two-dimensional series with a fixed law; this notion, which may be of future importance, will be presented in the last section. This notion will also be useful to explain what happens when the coprimality condition of the a_i's is removed, without having to rewrite the statement of the Theorem in a heavier form where a_1 and a_2 would be replaced by $a_1/\gcd(a_1,a_2)$ and $a_2/\gcd(a_1,a_2)$, and $\{\eta_n = N\}$ by $\{\eta_n = N\gcd(a_1,a_2)\}\ldots$

We now state our main result.

Theorem. — *Let p_0, p_1, p_2 be three positive numbers with sum 1, and $a_0 = 0 < a_1(n) < a_2(n)$ be three coprime integers. Let further*

$$\begin{aligned}\mu &= \mu_n = p_1a_1(n) + p_2a_2(n),\\ \sigma^2 &= \sigma_n^2 = p_1a_1^2(n) + p_2a_2^2(n) - \mu_n^2.\end{aligned}$$

We consider n independent random variables $(\xi_{n,k})_{1\le k\le n}$, each of which takes the values $a_0, a_1(n), a_2(n)$ with probabilities p_0, p_1, p_2 respectively, and we denote by η_n the sum $\xi_{n,1} + \cdots + \xi_{n,n}$.

When $a_2(n) = o(\sqrt{n})$ as n tends to infinity, we have, uniformly with respect to the integer N

$$P\{\eta_n = N\} = \frac{1}{\sigma_n\sqrt{2\pi n}}\left(\exp\left(-\frac{(n\mu_n - N)^2}{2n\sigma_n^2}\right) + o(1)\right) \quad \textit{as } n \to \infty.$$

When $a_2(n)/\sqrt{n}$ tends to infinity with n, we have, uniformly with respect to the integer N

$$P\{\eta_n = N\} = \frac{n!}{k_0!k_1!k_2!}p_0^{k_0}p_1^{k_1}p_2^{k_2} + o(\frac{1}{n}) \quad \textit{as } n \to \infty,$$

where the integral triple (k_0, k_1, k_2) is defined by

$$-a_2(n)/2 < k_1 - np_1 \le a_2(n)/2,\ k_1a_1(n) + k_2a_2(n) = N,\ k_0 + k_1 + k_2 = n.$$

When $a_2(n)/\sqrt{n}$ tends to a finite positive limit c when n tends to infinity, we have, uniformly with respect to c and to the integer N

$$P\{\eta_n = N\} = \frac{1}{2\pi n\sqrt{p_0p_1p_2}}\sum_{(k_0,k_1,k_2)} \exp(Q(k_0, k_1, k_2)) + o(\frac{1}{n}) \quad \textit{as } n \to \infty,$$

where the sum is extended to integral triples satisfying $k_0 + k_1 + k_2 = n$, $a_1(n)k_1 + a_2(n)k_2 = N$, and the quadratic form Q is defined by

$$Q(k_0, k_1, k_2) = -\frac{1}{2}\sum_{i=1}^{3}\frac{1}{np_i}(k_i - np_i)^2.$$

1. The case when $a_2(n) = o(\sqrt{n})$

Due to the arithmetical nature of our problem (the a_i's are integers), we shall use the Fourier kernel $\exp(2\pi i \cdot x)$ and define by

$$\Psi(t) = \Psi_n(t) = p_0 + p_1 \exp(2\pi i t a_1) + p_2 \exp(2\pi i t a_2)$$

the characteristic function of $\xi_{n,k}$ so that we have

$$P\{\eta_n = N\} = \int \Psi^n(t) \exp(-2\pi i t N)\, dt\ ,$$

where the integral is performed over any interval of length 1.

We shall divide the range of summation into a *major arc*, when t is close to 0, and a *minor arc* when t is far from 0. Let ε be a positive real number (that will be specified later to be $1/(a_2(n)n^{2/5})$), and let

$$\mathfrak{M} = [-\varepsilon, \varepsilon] \qquad \text{and} \qquad \mathfrak{m} =]\varepsilon; 1-\varepsilon].$$

1.1. Contribution of the minor arc.

— The following lemma will be used to get an upper bound for Ψ on the minor arc, playing either with the term $\exp(2\pi i a_1 t)$ or with $\exp(2\pi i a_2 t)$.

Lemma 1. — *Let* $C = 8 \min\left(\frac{p_0 p_1}{p_0+p_1}, \frac{p_0 p_2}{p_0+p_2}\right)$. *We have*

$$|p_0 + p_1 \exp(2\pi i u_1) + p_2 \exp(2\pi i u_2)| \le \exp(-C \max(||u_1||^2, ||u_2||^2))\ ,$$

where $||u||$ *denotes the distance from* u *to the nearest integer.*

Proof. — It is of course enough to prove the inequality

$$|p_0 + p_1 \exp(2\pi i u_1) + p_2 \exp(2\pi i u_2)| \le \exp\left(\frac{-8 p_0 p_1}{p_0 + p_1} ||u_1||^2\right).$$

We have

$$\begin{aligned} |p_0 + p_1 \exp(2\pi i u_1)|^2 &= p_0^2 + p_1^2 + 2 p_0 p_1 \cos 2\pi u_1 \\ &= (p_0 + p_1)^2 - 4 p_0 p_1 \sin^2 \pi u_1 \\ &\le (p_0 + p_1)^2 - 16 p_0 p_1 ||u_1||^2 \\ &\le (p_0 + p_1)^2 (1 - \frac{8 p_0 p_1}{(p_0 + p_1)^2} ||u_1||^2)^2\ . \end{aligned}$$

This implies

$$\begin{aligned} |p_0 + p_1 \exp(2\pi i u_1) + p_2(2\pi i u_2)| &\le p_0 + p_1 + p_2 - \frac{8 p_0 p_1}{p_0 + p_1} ||u_1||^2 \\ &\le 1 - \frac{8 p_0 p_1}{p_0 + p_1} ||u_1||^2 \\ &\le \exp\left(-\frac{8 p_0 p_1}{p_0 + p_1} ||u_1||^2\right)\ , \end{aligned}$$

which is the inequality we looked for. □

We now present the dissection of the minor arc. For any integer r, we shall denote by $\overline{r}$ the integer in $[0, a_2/2]$ such that r is congruent either to $\overline{r}$ or to $-\overline{r}$ modulo a_2. The reader will easily check that $\mathfrak{m}$ is the disjoint union of the following intervals:

$$\begin{aligned}
\mathfrak{m}_1(r) &= \left]\frac{r}{a_2} - \frac{\overline{ra_1}}{2a_2^2}, \frac{r}{a_2} + \frac{\overline{ra_1}}{2a_2^2}\right[, \quad \text{for } r = 1, 2, \dots, a_1 - 1,\\
\mathfrak{m}_2(0) &= \left]\varepsilon, \frac{1}{a_2} - \frac{\overline{1a_1}}{2a_2^2}\right],\\
\mathfrak{m}_2(a_2-1) &= \left[\frac{a_2-1}{a_2} + \frac{\overline{(a_2-1)a_1}}{2a_2^2}, 1-\varepsilon\right[\ ,\\
\mathfrak{m}_2(r) &= \left[\frac{r}{a_2} + \frac{\overline{ra_1}}{2a_2^2}, \frac{r+1}{a_2} - \frac{\overline{(r+1)a_2}}{2a_2^2}\right], \quad \text{for } r = 1, 2, \dots, a_2 - 2\ .
\end{aligned}$$

The intervals of type $\mathfrak{m}_2$ stay away from rationals with denominator a_2, so that $||a_2 t||$ is rather large when t is in such an interval. More precisely, if we consider $\mathfrak{m}_2^-(r) = [\frac{r}{a_2} + \frac{\overline{ra_1}}{2a_2^2}, \frac{2r+1}{2a_2}]$, in the case when $1 \le r \le a_2/2$, we have $t \in \mathfrak{m}_2^-(r) \Rightarrow ||ta_2|| = ||ta_2 - r|| = (ta_2 - r)$, so that, by Lemma 1, we have

$$\begin{aligned}
\int\limits_{\mathfrak{m}_2^-(r)} |\Psi(t)|^n dt &\le \int\limits_{\mathfrak{m}_2^-(r)} \exp(-Cn(ta_2 - r)^2) dt\\
&\le \frac{1}{a_2\sqrt{n}} \int\limits_{\frac{\overline{ra_1}}{2a_2}\sqrt{n}}^{\infty} \exp(-Cu^2) du\ .
\end{aligned}$$

In a similar way, the contribution of each of $]\varepsilon, \frac{1}{2a_2}]$ and $[\frac{2a_2-1}{a_2}, 1-\varepsilon[$ is at most

$$\frac{1}{a_2\sqrt{n}} \int\limits_{\varepsilon a_2\sqrt{n}}^{\infty} \exp(-Cu^2) du.$$

The coprimality of a_1 and a_2 implies that $\overline{ra_1}$ is different from 0 for $r = 1, 2, \dots, a_1 - 1$, and that any integer s in $[1, a_2/2]$ is equal to some $\overline{ra_1}$ for at most two values of r. If we denote by $\mathfrak{m}_2$ the union of the $\mathfrak{m}_2(r)$ for $r = 0, 1, \dots, a_2 - 1$, we get

$$\int\limits_{\mathfrak{m}_2} |\Psi(t)|^n dt \le \frac{2}{a_2\sqrt{n}} \int\limits_{\varepsilon a_2\sqrt{n}}^{\infty} \exp(-Cu^2) du + \frac{4}{a_2\sqrt{n}} \sum_{s=1}^{\infty} \int\limits_{\frac{s\sqrt{n}}{2a_2}}^{\infty} \exp(-Cu^2) du\ ,$$

so that the condition $a_2 = o(\sqrt{n})$ implies

$$\int\limits_{\mathfrak{m}_2}^{\infty} |\Psi(t)|^n dt \le \frac{2}{a_2\sqrt{n}} \int\limits_{\varepsilon a_2\sqrt{n}}^{\infty} \exp(-Cu^2) du + o(\frac{1}{a_2\sqrt{n}}). \tag{1}$$

We now turn our attention to the $\mathfrak{m}_1(r)$'s, the union of which is denoted by $\mathfrak{m}_1$. The length of $\mathfrak{m}_1(r)$ is $\frac{\overline{ra_1}}{2a_2^2}$, and for t in $\mathfrak{m}_1(r)$ we have

$$||a_1t|| \geq ||\frac{a_1r}{a_2}| - \frac{a_1\overline{ra_1}}{2a_2^2} \geq \frac{\overline{ra_1}}{2a_2},$$

so that we get

$$\begin{aligned}\int\limits_{\mathfrak{m}_1} |\Psi(t)|^n dt &\leq \frac{2}{2a_2^2}\sum_{s=1}^{\infty}\exp(-Cn(\frac{s}{2a_2^2})^2)\\ &\leq \frac{2}{a_2^2}\exp(-\frac{Cn}{4a_2^2})\ ,\end{aligned}$$

as soon as $\exp(-Cn/(4a_2^2))$ is less than $1/2$, which is the case when n is large enough. We thus have

$$\int\limits_{\mathfrak{m}_2}^{\infty} |\Psi(t)|^n dt \leq \frac{2}{a_2\sqrt{n}}\cdot\frac{\sqrt{n}}{a_2}\exp(-\frac{Cn}{4a_2^2}) = o(\frac{1}{a_2\sqrt{n}}).$$

Combining the contributions of the different parts of the minor arc, we get the following

Proposition 1. — *We have, for any* $\varepsilon > 0$

$$\int\limits_{\varepsilon}^{1-\varepsilon} |\Psi(t)|^n dt \leq \frac{2}{a_2\sqrt{n}}\int\limits_{\varepsilon a_2\sqrt{n}}^{\infty}\exp(-Cu^2)du + o(\frac{2}{a_2\sqrt{n}}).$$

1.2. Contribution of the major arc

Proposition 2. — *Let* $\varepsilon = O(1/(a_2n^{1/3}))$. *We have*

$$\int\limits_{-\varepsilon}^{\varepsilon}\Psi^n(t)\exp(-2\pi iNt)dt = \frac{1}{\sigma\sqrt{2\pi n}}\exp\Big(-\frac{(\mu n - N)^2}{2n\sigma^2}\Big)$$

$$+O(a_2^3n\varepsilon^4) + O(\frac{1}{a_2\sqrt{n}}\int\limits_{\varepsilon a_2\sqrt{n}}^{\infty}\exp(-Cu^2)du),$$

uniformly in N, *where* n *tends to infinity.*

This proposition is an easy consequence of the two following lemmata

Lemma 2. — *Let* $\varepsilon = O(1/(a_2n^{1/3}))$. *We have*

$$\int\limits_{-\varepsilon}^{\varepsilon}\Psi^n(t)\exp(-2\pi iNt)dt = \int\limits_{-\varepsilon}^{\varepsilon}\exp(-2\pi it(n\mu - N) - 2\pi^2n\sigma^2t^2)dt + O(a_2^3n\varepsilon^4),$$

uniformly in N, *when* n *tends to infinity.*

Proof. — On the interval $[-\varepsilon, +\varepsilon]$ we use the second order Taylor expansion of $\exp(2\pi i a_1 t)$ and $\exp(2\pi i a_2 t)$, noticing that $a_1 t$ and $a_2 t$ are bounded terms. We get

$$\begin{aligned} p_0 &+ p_1 \exp(2\pi i a_1 t) + p_2 \exp(2\pi i a_2 t) \\ &= 1 + 2\pi i t(a_1 p_1 + a_2 p_2) - 2\pi^2(a_1^2 p_1 + a_2^2 p_2)t^2 + O(a_2^3|t|^3) \\ &= \exp(2\pi i \mu t - 2\pi^2\sigma^2 t^2 + O(a_2^3|t|^3)) \ . \end{aligned}$$

The integrand in the LHS of the formula in Lemma 2 can thus be written as

$$\exp(2\pi i(n\mu - N)t - 2\pi^2 n\sigma^2 t^2 + O(na_2^3|t|^3)).$$

which is also equal to

$$\exp(2\pi i(n\mu - N)t - 2\pi^2 n\sigma^2 t^2) + O(na_2^3\varepsilon^3),$$

since $na_2^3\varepsilon^3$ is bounded. The lemma follows by integrating over $[-\varepsilon, \varepsilon]$. □

Lemma 3. — *We have*

$$\int\limits_{|t|>\varepsilon} |\exp(2\pi t i(n\mu - N) - 2\pi^2 n\sigma^2 t^2)|dt \leq \frac{2}{a_2\sqrt{n}} \int\limits_{\varepsilon a_2\sqrt{n}}^{\infty} \exp(-Cu^2)du.$$

Proof. — We first notice that $\sigma^2 \geq a_2^2 \frac{p_0 p_2}{p_0 + p_2}$. We have

$$\begin{aligned} \sigma^2 &= p_1 a_1^2 + p_2 a_2^2 - (p_1 a_1 + p_2 a_2)^2 \\ &= p_1 q_1 (a_1^2 - 2\frac{p_2}{q_1} a_1 a_2 + \frac{p_2 q_2}{p_1 q_1} a_2^2) \\ &= p_1 q_1 ((a_1 - \frac{p_2}{q_1} a_2)^2 + (\frac{p_2 q_2}{p_1 q_1} - \frac{p_2^2}{q_1^2}) a_2^2) \\ &\geq a_2^2 \left(\frac{p_1 q_2 q_1 - p_1 p_2^2}{q_1} \right) \\ &= a_2^2 \frac{(q_2 q_1 - p_1 p_2) p_2}{q_1} \\ &= a_2^2 \frac{p_0 p_2}{p_0 + p_2} \ . \end{aligned}$$

This inequality on σ^2 implies $2\pi^2\sigma^2 \geq Ca_2^2$.
We thus obtain

$$\begin{aligned} \int\limits_{|t|>\varepsilon} |\exp(2\pi i t(n\mu - N) &- 2\pi^2 n\sigma^2 t^2|dt = \int\limits_{|t|>\varepsilon} \exp(-2\pi^2 n\sigma^2 t^2)dt \\ &\leq 2\int\limits_{\varepsilon}^{\infty} \exp(-Cna_2^2 t^2)dt \\ &= \frac{2}{a_2\sqrt{n}} \int\limits_{\varepsilon a_2\sqrt{n}}^{\infty} \exp(-Cu^2)du \ . \end{aligned}$$

□

The proof of Proposition 2 follows from Lemma 2, Lemma 3, and the well-known relation

$$\int_{-\infty}^{\infty} \exp(2\pi it(n\mu - N) - 2\pi^2 n\sigma^2 t^2)dt = \frac{1}{\sigma\sqrt{2\pi n}}\exp(-\frac{(\mu n - N)^2}{2n\sigma^2}).$$

1.3. Proof of the Theorem when $a_2(n) = o(\sqrt{n})$. — According to Proposition 1 and Proposition 2, we merely have to notice that σ and a_2 have the same order of magnitude (we already showed that $\sigma^2 \geq a_2^2\frac{p_0p_2}{p_0+p_2}$, and the relation $\sigma^2 \leq a_2^2$ is trivial), and to show that one can find a function ε such that:

(i) $\varepsilon a_2\sqrt{n} \to \infty$,
(ii) $a_2^3 n\varepsilon^4 = o(\frac{1}{a_2\sqrt{n}})$,
(iii) $a_2 n^{1/3}\varepsilon = O(1)$.

As we already mentioned it, the quantity $\varepsilon = \frac{1}{a_2(n)n^{2/5}}$ satisfies these three conditions.

2. The case when $a_2(n)/\sqrt{n}$ tends to infinity.

For non-negative integers k_0, k_1, k_2 we define

$$P(k_0, k_1, k_2) = \frac{(k_0 + k_1 + k_2)!}{k_0!k_1!k_2!}p_0^{k_0}p_1^{k_1}p_2^{k_2},$$

and we extend this definition by letting the RHS be 0 when at least one of the k's is negative.

The second case of the theorem easily follows from the following lemma.

Lemma 4. — *Let $a_1(n)$ and $a_2(n)$ be positive coprime integers. For an integer N, we define*

$$\mathcal{E}(N) = \{(k_0, k_1, k_2), k_0 + k_1 + k_2 = n \quad \text{and} \quad k_1a_1(n) + k_2a_2(n) = N\}.$$

Let φ be a function that tends to infinity with its argument. Uniformly in N, $a_1(n)$ and $a_2(n)$ we have

$$\sum_{\substack{(k_0,k_1,k_2)\in\mathcal{E}(N)\\ |k_1-p_1n|>\varphi(n)\sqrt{n}}} P(k_0, k_1, k_2) = o\left(\frac{1}{n}\right).$$

Proof. — We consider the fundamental triple (k_0^*, k_1^*, k_2^*) in $\mathcal{E}(N)$ satisfying $-a_2/2 < k_1^* - p_1n \leq a_2/2$. We have

$$\mathcal{E}(N) = \{(k_0^* - s(a_2 - a_1), k_1^* + sa_2^*, k_2^* - sa_1), s \in \mathbb{Z}\}.$$

We remember for later use that the second components of two triples in $\mathcal{E}(N)$ differ by a multiple of a_2, and that for any given k_1, there exists in $\mathcal{E}(N)$ at most one triple the second element of which is k_1.

Let now k_1 be a given non-negative integers less than n. We define

$$\tilde{p_0} = \frac{p_0}{p_0 + p_2}, \quad \tilde{p_2} = \frac{p_2}{p_0 + p_2}, \quad \tilde{k_0} = \lfloor(n - k_1)\tilde{p_0}\rfloor, \quad \tilde{k_2} = \lfloor(n - k_1)\tilde{p_2}\rfloor.$$

By a simple property of the binomial distribution $\mathcal{B}(n-k_1, \widetilde{p_0})$, we have

$$\frac{(n-k_1)!}{k_0!k_2!}\widetilde{p_0}^{\,k_0}\widetilde{p_2}^{k_2} = O\left(\frac{n-k_1)!}{\widetilde{k_0}!\widetilde{k_2}!}\widetilde{p_0}^{\widetilde{k_0}}\widetilde{p_2}^{\widetilde{k_2}}\right) = O\left(\frac{1}{\sqrt{n-k_1}}\right).$$

From this follows that for (k_0, k_1, k_2) in $\mathcal{E}(N)$ we have

$$\begin{aligned}\frac{n!}{k_0!k_1!k_2!}p_0^{k_0}p_1^{k_1}p_2^{k_2} &= \frac{n!}{k_1!(n-k_1)!}p_1^{k_1}\frac{(n-k_1)!}{k_0!k_2!}\widetilde{p_0}^{k_0}\widetilde{p_2}^{k_2}\left(\frac{p_0^{k_0}p_2^{k_2}}{\widetilde{p_0}^{k_0}\widetilde{p_2}^{k_2}}\right)\\ &= O\left(\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}\frac{p_0^{k_0}p_2^{k_2}}{\widetilde{p_0}^{k_0}\widetilde{p_2}^{k_2}}\frac{1}{\sqrt{n-k_1}}\right)\\ &= O\left(\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}(p_0+p_2)^{k_0+k_2}\frac{1}{\sqrt{n-k_1}}\right)\\ &= O\left(\frac{1}{\sqrt{n-k_1}}\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}q_1^{n-k_1}\right).\end{aligned}$$

Using the properties of the terms k_1 in the triple (k_0, k_1, k_2) belonging to $\mathcal{E}(N)$, the series

$$\sum_{\substack{(k_0,k_1,k_2)\in\mathcal{E}(N)\\ |k_1-p_1n|>\varphi(n)\sqrt{n}}} P(k_0, k_1k_2)$$

is – up to a constant factor – at most equal to its largest term, which is $o(\frac{1}{n})$. □

3. The case when $a_2(n)/\sqrt{n}$ tends to a positive limit.

We retain the definition of $P(k_0, k_1, k_2)$ and $\mathcal{E}(N)$ introduced in the previous section, and notice that Lemma 4 may still be applied in our present case, as well as in the description of $\mathcal{E}(N)$ in terms of the fundamental triple (k_0^*, k_1^*, k_2^*) satisfying $-a_2/2 < k_1^* - np_1 \le a_2/2$. We further write $\mathbf{k} = (k_0, k_1, k_2)$. Let $\varphi(n) = n^{1/10}$; we have

$$\begin{aligned}&\left|\sum_{\mathbf{k}\in\mathcal{E}(N)} P(\mathbf{k}) - \sum_{\mathbf{k}\in\mathcal{E}(N)} \frac{\exp(Q(\mathbf{k}))}{2\pi n\sqrt{p_0p_1p_2}}\right|\\ \le &\sum_{\substack{\mathbf{k}\in\mathcal{E}(N)\\ |k_1-np_1|\le\varphi(n)\sqrt{n}\\ |k_2-np_2|\le\varphi(n)\sqrt{n}}} \left|P(\mathbf{k}) - \frac{\exp(Q(\mathbf{k}))}{2\pi n\sqrt{p_0p_1p_2}}\right|\\ + &\sum_{\substack{\mathbf{k}\in\mathcal{E}(N)\\ |k_1-np_1|>\varphi(n)\sqrt{n}}} P(\mathbf{k}) + \sum_{\substack{\mathbf{k}\in\mathcal{E}(N)\\ |k_2-np_2|\le\varphi(n)\sqrt{n}}} P(\mathbf{k})\\ + &O\left(\frac{1}{n}\sum_{\substack{\mathbf{k}\in\mathcal{E}(N)\\ |k_1-np_1|>\varphi(n)\sqrt{n}}} \exp(Q(\mathbf{k})\right) + O\left(\frac{1}{n}\sum_{\substack{\mathbf{k}\in\mathcal{E}(N)\\ |k_1-np_1|\le\varphi(n)\sqrt{n}\\ |k_2-np_2|>\varphi(n)\sqrt{n}}} \exp(Q(\mathbf{k}))\right).\end{aligned}$$

By Lemma 4, the second and the third terms in the RHS are $o(1/n)$. The fourth term is easily dealt with since we have

$$\sum_{\substack{\mathbf{k}\in\mathcal{E}(N)\\ |k_1-np_1|>\varphi(n)\sqrt{n}}} \exp(Q(\mathbf{k})) \leq \sum_{\substack{\mathbf{k}\in\mathcal{E}(N)\\ |k_1-np_1|>\varphi(n)\sqrt{n}}} \exp\Big(-\frac{1}{2np_1}(k_1-np_1)^2\Big)$$

$$\leq 2 \sum_{\substack{s\in N\\ s>(\varphi(n)-1)(\sqrt{n}/a_2)}} \exp\Big(-\frac{1}{2p_1}(s-1)^2\frac{a_2^2}{n}\Big) = o(1).$$

Concerning the fifth term in the RHS, we first notice that the growth condition on a_2 implies that there are $O(\varphi(n))$ elements $\mathbf{k}$ in $\mathcal{E}(N)$ such that $|k_1-np_1|\leq\varphi(n)\sqrt{n}$. We thus have

$$\sum_{\substack{\mathbf{k}\in\mathcal{E}(N)\\ |k_1-np_1|\leq\varphi(n)\sqrt{n}\\ |k_2-np_2|>\varphi(n)\sqrt{n}}} \exp(Q(\mathbf{k})) = O\Big(\varphi(n)\exp\Big(-\frac{1}{2np_2}\varphi^2(n)n\Big)\Big) = o(1).$$

We now turn our attention toward the first and main term. We are going to prove that the conditions $\mathbf{k}\in\mathcal{E}(N)$, $|k_1-np_1|\leq\varphi(n)\sqrt{n}$, $|k_2-np_2|\leq\varphi(n)\sqrt{n}$ imply

$$P(\mathbf{k}) = \frac{1}{2\pi n\sqrt{p_0p_1p_2}}\exp(Q(\mathbf{k})) + o(\frac{1}{n\varphi(n)}),$$

which is sufficient to induce

$$\sum_{\mathbf{k}\in\mathcal{E}(N)} P(\mathbf{k}) = \sum_{\mathbf{k}\in\mathcal{E}(N)} \frac{\exp(Q(\mathbf{k}))}{2\pi n\sqrt{p_0p_1p_2}} + o(\frac{1}{n}),$$

which is the last part of the Theorem. We notice that the conditions we have stated imply $|k_i-np_i|\leq 2\varphi(n)\sqrt{n}$ for $i=0,1$ and 2.

We use Stirling's formula in the shape

$$s! = \sqrt{2\pi s}\Big(\frac{s}{e}\Big)^s e^{\Theta_s},\quad \text{with } |\Theta_s|\leq\frac{1}{12s}.$$

We have

$$P(\mathbf{k}) = \frac{1}{2\pi}\sqrt{\frac{n}{k_0k_1k_2}}(\frac{np_0}{k_0})^{k_0}(\frac{np_1}{k_1})^{k_1}(\frac{np_2}{k_2})^{k_2}\exp(O(\frac{1}{n}))$$

$$= \frac{1}{2\pi n\sqrt{p_0p_1p_2}}(\frac{np_0}{k_0})^{k_0+1/2}(\frac{np_1}{k_1})^{k_1+1/2}(\frac{np_2}{k_2})^{k_2+1/2}\exp(O(\frac{1}{n})).$$

A second order expansion easily leads to

$$\begin{aligned}\log(\frac{np}{k})^{k+1/2} &= (k+1/2)\log(1+\frac{np-k}{k})\\ &= np-k-\frac{(np-k)^2}{2np}+O(\frac{\varphi(n)^3}{\sqrt{n}})\ ;\end{aligned}$$

since the sum $(np_1-k_1)+(np_2-k_2)+(np_3-k_3)$ is 0, we have

$$P(\mathbf{k}) = \frac{1}{2\pi n\sqrt{p_0p_1p_2}}\exp(Q(\mathbf{k})) + O(\frac{\varphi(n)^3}{n^{3/2}})$$

which is stronger that what we need as soon as $\varphi(n) = o(n^{1/8})$. The Theorem is now completely proved. We can even notice that the error term in the last case can be further reduced.

4. Isomorphism between series of discrete random variables

In the case when $a_2(n)/\sqrt{n}$ tends to infinity with n, the Theorem tells us that at most one term (the fundamental one, if any) is meaningful in the series

$$P\{\eta_n = N\} = \sum_{\mathbf{k}\in\mathcal{E}(N)} P(\mathbf{k}).$$

This situation is similar to a two-dimensional one: let $(\zeta_{n,k})$ be a family of independent bivariate random variables taking the values $(0,0), (0,1)$ and $(1,0)$ with respective probabilities p_0, p_1 and p_2. Letting $\varphi_n = \zeta_{n,1} + \cdots + \zeta_{n,n}$, we have

$$P\{\varphi_n = (k_1, k_2)\} = P(n - k_1 - k_2, k_1, k_2).$$

We thus would like to consider that the sequence of linear maps $\mathbb{R}^2 \xrightarrow{\mathbf{a}(n)} \mathbb{R}$ defined by $\mathbf{a}(n).\mathbf{x} = a_1(n)x_1 + a_2(n)x_2$ induces an isomorphism between the series $(\xi_{n,k})$ and $(\zeta_{n,k})$; this notion, for which we now suggest a precise definition, bears some similarity with the one that has been introduced by the second named author when dealing with additive problems (cf. [**1**]). Let us first recall this notion.

Let A and B be two finite subsets of two monoids $(E,+)$ and $(F,+)$, respectively, and let s be a positive integer. We denote by sum_s the map from E^s to E defined by $\mathrm{sum}_s(n_1, \ldots, n_s) = n_1 + \cdots + n_s$, and, since there is no fear of confusion, we use the same symbol to denote the sum of s elements of F. We let sA denote the set $\mathrm{sum}_s(A^S)$. The sets A and B are said to be *s-isomorphic* if there exist a bijection $\varphi : A \to B$ and a bijection $\varphi^{(s)} : sA \to sB$ such that

$$\forall x \in sA : \mathrm{sum}_s \circ \varphi^{\otimes s} \circ \mathrm{sum}_s^{-1}(\{x\}) = \{\varphi^{(s)}(x)\} ,$$
$$\forall y \in sB : \mathrm{sum}_s \circ (\varphi^{-1})^{\otimes s} \circ \mathrm{sum}_s^{-1}(\{y\}) = \{(\varphi^{(s)})^{-1}(y)\} .$$

A straightforward way to extend this definition to the s-isomorphism of probability measures P and Q, supported respectively by A and B, is to request that A and B are s-isomorphic and that the image of P by φ is Q and the image of P^{*s} by $\varphi^{(s)}$ is Q^{*s}.

For practical purposes, this definition is however too strong, as we shall show after stating a convenient definition of a weaker concept. We may use to our benefit the fact that we measure sets, by allowing the loss of some points in sA and sB with a total measure not exceeding some given δ, and allowing also the preservation of the measure by $\varphi^{(s)}$ to be approximate, up to some given ε. This leads to the following definition.

Definition. — *Let P and Q be two probability measures with finite supports A and B in monoids $(E,+)$ and $(F,+)$, respectively. Let s be a positive integer and δ and*

ε be two non-negative real numbers. The measures P and Q are said to be (s,δ,ε)-isomorphic if there exists a bijection $\varphi : A \to B$, two subsets $A^{(s)} \subset sA$, $B^{(s)} \subset sB$ and a bijection $\varphi^{(s)} : A^{(s)} \to B^{(s)}$ such that:

(i) $P^{*s}(A^{(s)}) \geq 1-\delta, \quad Q^{*s}(B^{(s)}) \geq 1-\delta$,

(ii) $\forall x \in A^{(s)} : \mathrm{sum}_s \circ \varphi^{\otimes s} \circ \mathrm{sum}_s^{-1}(\{x\}) = \{\varphi^{(s)}(x)\}, \quad \forall y \in B^{(s)} : \mathrm{sum}_s \circ (\varphi^{-s})^{\otimes s} \circ \mathrm{sum}_s^{-1}(\{y\}) = \{(\varphi^{(s)})^{-1}(y)\}$,

(iii) $\forall x \in A^{(s)} : |P^{*s}(\{x\}) - Q^{*s}(\{\varphi^{(s)}(x)\})| \leq \varepsilon$.

We first notice that when (ii) is satisfied, (iii) is equivalent to

(iii′) $\forall y \in B^{(s)} : |Q^{*s}(\{y\}) - P^{*s}(\{(\varphi^{(s)})^{-1}(y)\})| \leq \varepsilon$,
which implies that when P and Q are (s,δ,ε)-isomorphic, so are Q and P.

When A and B are exactly the support of P and Q, the $(s,0,0)$-isomorphism is exactly the strong isomorphism we introduced at first.

As an example, we consider the case where $E = \mathbb{R}$, $F = \mathbb{R}^2$, $A = \{0,1,n\}, B = \{(0,0),(0,1),(1,0)\}$ with uniform measures, which we denote by P and Q, respectively. The map $\varphi : A \to B$ defined by $\varphi(0) = (0,0), \varphi(1) = (0,1), \varphi(n) = (1,0)$ naturally extends to sA with $2 \leq s < n$ and defines an $(s,0,0)$-isomorphism from A to B. For $s = n$, one difficulty comes from the fact that n has $n+1$ representations in nA (namely, $1+1+\cdots+1$ and the n permutations of $n+0+\cdots+0$), which correspond to the unique representation of $(0,n)$ in nB and the n representations of (1,0) in nB. Of course, from the beginning we knew that there is non $(n,0,0)$-isomorphism between A and B since nA and nB do not have the same cardinality; but we were not far from succeeding.

There are indeed ways to build satisfactory approximate isomorphisms between A and B. One way is to consider $A^{(n)} = nA \backslash \{n\}$ and $B^{(n)} = nB \backslash \{(0,n),(1,0)\}$; in this way, we get a $\left(n,(n+1)3^{-n},0\right)$-isomorphism.

Another way to proceed is to keep the n representations of (1,0) in nB, and take $B^{(n)} = nB \backslash \{(0,n)\}$; this implies that we have to keep n in nA, with its $(n+1)$ representations; we thus take $A^{(n)} = nA$. We are losing one representation on the ε-side (local side), but winning n of them on the δ-side (global side), and finally obtain a $\left(n,3^{-n},3^{-n}\right)$-isomorphism.

In general, the problem will be to find a trade between the δ and ε-sides, which play in opposite directions. For example, when dealing with sequences of probability measures (P_n) and (Q_n), a natural choice will be to obtain $(n,\delta_n,\varepsilon_n)$-isomorphisms such that $\delta_n \to 0$ and $\varepsilon_n = o\Big(\max_x \sup\left(P_n^{*n}(\{x\}),\ Q_n^{*n}(\{x\})\right)\Big)$.

We now turn our attention to our main result.

First, this notion of isomorphism helps us to understand that the conditions $a_1 = 0$ and (a_1,a_2) coprime induce no real restriction, since two random variables ξ and ζ, taking respectively the integral values

$$(a_0,a_1,a_2) \quad \text{and} \quad \Big(0, \frac{a_1-a_0}{gcd(a_1-a_0,a_2-a_0)}, \frac{a_2-a_0}{gcd(a_1-a_0,a_2-a_0)}\Big)$$

with the same weights (p_0, p_1, p_2), have laws which are $(h, 0, 0)$-isomorphic for any $h \geq 1$.

Second, in the case when $a_2(n)/\sqrt{n}$ tends to infinity with n, we can indeed show that $\xi_{n,1}$ is $(n, \delta_n, \varepsilon_n)$-isomorphic to the two dimensional integral valued random variable η taking the *fixed* values (0,0), (1,0), (0,1) with respective probabilities p_0, p_1, p_2, for some $\delta_n = o(1)$ and $\varepsilon_n = o(\frac{1}{n})$. This requires some explanation. We let

$$\begin{aligned}
&A = \{(0,0),(1,0),(0,1)\}, \quad B = \{0, a_1(n), a_2(n)\}\ ,\\
&P_n \text{ be the law of } \xi_{n,1}, \quad Q \text{ be the law of } \eta\ ,\\
&\varphi : A \to B \text{ such that } \varphi(0,0) = 0, \quad \varphi(1,0) = a_1, \quad \varphi(0,1) = a_2\ ,\\
&A^{(n)} = \{(k_1, k_2) : -a_2/2 < k_1 - p_1 n \leq a_2/2, \quad k_1 + k_2 \leq n\}\ ,\\
&\varphi^{(n)} : A^{(n)} \to nB \text{ such that } \varphi(k_1, k_2) = k_1 a_1 + k_2 a_2,\\
&\qquad B^{(n)} = \varphi^{(n)}(A^{(n)})\ ,
\end{aligned}$$

and we check that we have an isomorphism.

(i) The set $A^{(n)}$ contains

$$\{(k_1, k_2)/ - a_2/2 < k_1 - p_1 n \leq a_2/2, \ -a_2/2 < k_2 - p_2 n \leq a_2/2\}\ ;$$

since $a_2(n)/\sqrt{n}$ tends to infinity, we have $P^{*n}(A^{(n)}) \to 1$ when n tends to infinity. Now let N be in $B^{(n)}$. By the definition of $B^{(n)}$, there exists a triple (k_0, k_1, k_2) such that (k_1, k_2) is in $A^{(n)}$, $k_1 a_1 + k_2 a_2 = N$ and $k_0 + k_1 + k_2 = n$. The measure $Q^{*n}(\{N\})$ is the sum over all the triples (k_0, k_1, k_2) such that $k_1 a_1 + k_2 a_2 = N$ and $k_0 + k_1 + k_2 = n$ of the expression $\frac{n!}{k_0! k_1! k_2!} p_0^{k_0} p_1^{k_1} p_2^{k_2}$; but this last expression is $P^{*n}(\{(k_1, k_2)\})$. We thus have

$$\begin{aligned}
Q^{*n}(B^{(n)}) = \textstyle\sum_{N \in B^{(n)}} Q^{*n}(\{N\}) \ &\geq \textstyle\sum_{(k_1,k_2) \in A^{(n)}} P^{*n}(\{(k_1, k_2)\})\\
&= P^{*n}(A^{(n)})\ ,
\end{aligned}$$

and so we have $Q^{*n}(B^{(n)}) \to 1$ when $n \to \infty$.

(ii) Let (k_1, k_2) be in $A^{(n)}$. The only way to write (k_1, k_2) as the sum of n elements of A is to take k_1-times (1,0), k_2-times (0,1) and $(n - k_1 - k_2)$-times (0,0). This leads to $\mathrm{sum}_n \circ \varphi^{\otimes s} \circ \mathrm{sum}_n^{-1}(\{(k_1, k_2)\}) = \{k_1 a_1 + k_2 a_2\}$.

Now let N be in $B^{(n)}$. By the definition of $B^{(n)}$, there exists (k_1, k_2) in $A^{(n)}$ such that $N = k_1 a_1 + k_2 a_2$. But there exists at most one pair (k_1, k_2) with $N = k_1 a_1 + k_2 a_2$ and $-a_2/2 < k_1 - p_1 n \leq a_2/2$ (since a_1, a_2 are coprime). This implies that $\mathrm{sum}_n \circ (\varphi^{-1})^{\otimes s} \circ \mathrm{sum}_n^{-1}(\{N\})$ consists of a single pair (k_1, k_2).

(iii) The result of the theorem states precisely that for N in $B^{(n)}$, written as $k_1 a_1 + k_2 a_2$ with $(k_1, k_2) \in A^{(n)}$, we have

$$Q^{*n}(\{N\}) = P^{*n}(\{(k_1, k_2)\}) + o\Big(\frac{1}{n}\Big)\ .$$

We may further notice, which is not asked in the definition, that when N is not in $B^{(n)}$ we have $Q^{*n}(\{N\}) = o(\frac{1}{n})$ and when (k_1, k_2) is not in $A^{(n)}$, we have $P^{*n}(\{(k_1, k_2)\}) = o(\frac{1}{n})$, so that $nA \backslash A^{(n)}$ as well as $nB \backslash B^{(n)}$ is not only globally small (condition (i)) but also locally small.

References

[1] Freiman G. A., *Foundations of a structural theory of set addition.* Translations of mathematical monographs, bf 37, Amer. Math. Soc., Providence, R.I., 1973.

J.-M. Deshouillers, Mathématiques stochastiques, Université Bordeaux 2, BP 26, 33076 Bordeaux, France • *E-mail :* j-m.deshouillers@u-bordeaux2.fr

G.A. Freiman, School of Mathematical Sciences, Department of Mathematics, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, 69978 Tel Aviv, Israel *E-mail :* grisha@math.tau.ac.il

W. Moran, Discipline of Mathematics, Flinders University, POB 2100, Adelaide SA 5001, Australia *E-mail :* bill@ist.flinders.edu.au

Astérisque

JEAN-MARC DESHOUILLERS
GREGORY A. FREIMAN
ALEXANDER A. YUDIN

On bounds for the concentration function. 1

Astérisque, tome 258 (1999), p. 425-436
<http://www.numdam.org/item?id=AST_1999__258__425_0>

Astérisque **258**, 1999, p. 425–436

ON BOUNDS FOR THE CONCENTRATION FUNCTION. 1

by

Jean-Marc Deshouillers, Gregory A. Freiman & Alexander A. Yudin

Abstract. — We give an upper bound for the concentration function of a sum of independent identically distributed integral valued random variables in terms of a lower bound for their tail, under the necessary extra condition that the random variables are not essentially supported in a proper arithmetic progression.

1. Introduction

Let $X_1, \dots, X_k, \dots$ be independent real random variables and $S_n = \sum_{k=1}^{n} X_k$. It is well known that, in general, the distribution of S_n spreads out as n grows. When all the X_k's are square-integrable, the relation $\sigma^2(S_n) = \sum_{k=1}^{n} \sigma^2(X_k)$ is a way to express this fact. In the general case, Doeblin and Lévy [**2**] were the first to measure this phenomenon in terms of concentration functions. The concentration function of a real random variable X is defined by

$$Q(X; \lambda) = \sup_t P\{t < X \leq t + \lambda\} \text{ for } \lambda \geq 0 \; .$$

The results of Doeblin and Lévy have been successively improved by Kolmogorov [**6**], Rogozin [**12**] and Kesten [**5**]. Let us quote a corollary to Kesten's result, for the case when the X_k's are identically distributed.

***Theorem* (Kesten [5], Corollary 1, p. 134)).** — *There exists an absolute constant C such that for any set of independent identically distributed random variables $X_1, \dots, X_n$*

1991 ***Mathematics Subject Classification.*** — 60 E15, 60 E10, 11 B05, 11 Z05.

Key words and phrases. — Sums of discrete random variables; concentration function, DLKRK-inequalities, characteristic function, additive number theory.

J.-M. D: Cette recherche a benificié du soutien du CNRS (UMR 9936, Université Bordeaux 1) et de l'Université Victor Segalen Bordeaux 2.

A. Y: This work was supported in part by RFFI GRANT N 93-01-00260.

and any $0 < \lambda \leq 2L$ *we have*

$$Q(S_n;L) \leq C\frac{L}{\lambda}\frac{Q(X_1;L)}{\sqrt{n(1-Q(X_1;\lambda))}} \ . \tag{1.1}$$

Let us consider the case when the X_k's follow a Cauchy law $\mathcal{C}(1)$,where the Cauchy law $\mathcal{C}(a)$ with parameter $a>0$ has density $a/\big(\pi(t^2+a^2)\big)$. One readily sees that for $L=1$ and $0<\lambda\leq 2$, the right hand side of (1.1) has order of magnitude $(\lambda\sqrt{n})^{-1}$ and is never $o(1/\sqrt{n})$. However, the random variable S_n follows the law $\mathcal{C}(n)$, and so

$$Q(S_n;1) = \frac{2}{\pi}\arctan\left(\frac{1}{2n}\right) = \frac{1}{n\pi}\big(1+o(1)\big) \ .$$

The dispersion (in the standard sense) of S_n is due to the dispersion of the X_k's themselves; but the dispersion of the X_k's is not reflected in a small concentration $Q(X_1;\lambda)$ for *small* λ's, but indeed for *large* λ's: the law of X_1 has a large tail, as can be seen from the fact that X_1 is not integrable.

A connection between the moments of the X_k's and the concentration of their sums has been provided by Esséen **[3]**, who proves that the integrability of $|X_1|^r$ for some $0<r\leq 2$ implies the *lower bound*

$$Q(S_n;L) \geq K(r)L\big(L+(n\mu_r)^{1/r}\big)^{-1} \ ,$$

where $\mu_r = \inf_a E(|X_1-a|^r)$ and $K(r)$ is an explicitly given expression that only depends on r.

We aim at giving an *upper bound* for the concentration function of S_n in terms of the *tail* of the distribution of the X_k's. There is however a difficulty that will be better seen on discrete random variables. Let us consider an integer $q>1$ and two integral valued random variables X_1 and X_1' such that

$$P\{X_1=0\} = P\{X_1'=0\} = 1/2 \ ,$$
$$P\{X_1'=\ell\} = \begin{cases} P\{X_1=\ell/q\} \neq 0 & \text{when } q \text{ divides } \ell, \\ 0 & \text{otherwise .}\end{cases}$$

We clearly have $Q(X_1;1)=Q(X_1';1)=1/2$ and the tail of the distribution of X_1' is heavier than that of X_1. However, if we consider two sets $X_1,\ldots,X_n$ and $X_1',\ldots,X_n'$ of n independent identically distributed random variables, their sums S_n and S_n' are such that $Q(S_n;1)=Q(S_n';1)$; we have indeed $P\{S_n=N\}=P\{S_n'=qN\}$ and so

$$Q(S_n;1) = \max_N P\{S_n=N\} = \max_M P\{S_n'=M\} = Q(S_n';1) \ .$$

We give in this paper an upper bound for the concentration function of a sum of independent identically distributed integral valued random variables in terms of the measure of their tail, under the assumption that the support of the random variables is not essentially contained in a proper arithmetic progression.

Theorem 1. — *Let* $\frac{\log 4}{\log 3} < \sigma < 2$, $\varepsilon > 0$, $A \geq 1$ *and* $a>0$ *be given real numbers. Let* n *be a positive integer and* $X_1,\ldots,X_n$ *a set of independent identically distributed* *integral valued random variables such that*

$$\max_{q\geq 2}\ \max_{s \bmod q}\ \sum_{\ell\equiv s(\bmod q)} P\{X_1=\ell\} \leq 1-\varepsilon\ , \tag{1.2}$$

$$\forall L \geq A : Q(X_1;L) \leq 1-aL^{-\sigma}\ . \tag{1.3}$$

Then we have

$$Q(S_n;1) \leq cn^{-1/\sigma}\ , \tag{1.4}$$

where c depends on σ, ε, A and a at most.

The main aim of this paper being to illustrate the use of inverse additive results to probability theory, we kept the statement and proof of our main result as simple as possible. We have thus restricted our attention to integral valued random variables, have not considered the general case when $0 < \sigma < 2$, and have not made explicit the dependence of c on the parameters ε, A and a. Let us simply notice here that Theorem 1 is valid under the condition $1 < \sigma < 2$: this depends on the fact that, under iterated applications of Lemma 3, the constant 3^k that arise may be improved to $(4-\epsilon)^k$, an observation which is basically due to Lev. However, when $\sigma < 1$, new phenomena enter the matter (generalized arithmetic progressions); we shall soon return to this topic.

The statement of Theorem 1 becomes false if condition (1.2) is suppressed. Of course, if the constant c in (1.4) is allowed to depend on the law of X_1, then condition (1.2) is no longer necessary.

The proof of this theorem may be summarized as follows. The concentration $Q(S_n;1)$ is majorized by the mean value of the modulus of the characteristic function of S_n; this latter is the n-th power of that of X_1, which we call φ, so that the problem reduces to the study of the large values of φ. Here we use two ideas that have been introduced by Freiman, Moskvin and Yudin in [**4**] in the context of local limit theorems. The first one, which can be seen as a consequence of Bochner's theorem, is that $\varphi(t_1+t_2)$ is large as soon as both $\varphi(t_1)$ and $\varphi(t_2)$ are large. The second one comes from the structure theory of set addition: either the set E of the arguments of the large values of φ is small, or it has a structure. In the first case, φ cannot be too large, and so we get (1.4). It remains to exclude the second case; were it to occur, then, as we shall see, either E would contain the vertices of a regular polygon, which would violate (1.2), or it would contain a large interval around 0, which would contradict (1.3).

Problems of estimating the measure of the set of large values of the characteristic function have also been studied by Arak and Zaitsev [**1**]. This gave them the possibility to solve a famous problem of Kolmogorov on the estimation of the approximation of the n-th convolution of any probability distribution by that of an infinitely divisible law.

As a warm up, and in order to introduce some tools and techniques, we devote the second paragraph to prove a special case of the Doeblin-Lévy-Kolmogorov-Rogozin-Kesten (DLKRK) inequality which stems from the same ideas and follows [**10**], [**11**].

The interested reader will find questions of a similar flavour in the classical monographs by Petrov [**9**] and the more recent one by Ledoux and Talagrand [**7**].

2. A DLKRK inequality for discrete random variables

***Theorem 2* (DLKRK).** — *Let $X_1, \ldots, X_n$ be independent identically distributed integral valued random variables, and let S_n be their sum and $p = \max_N P\{X_1 = N\}$. For every integer N, we have*

$$P\{S_n = N\} \leq 40 \frac{p}{\sqrt{n(1-p)}} .$$

Let us start by giving some notation that will be used in this paragraph and the next. We let

$$\begin{aligned}
&p_\ell = P\{X_1 = \ell\} \text{ for any } \ell \in \mathbb{Z} ,\\
&\varphi(t) = \sum_{\ell \in \mathbb{Z}} p_\ell \exp(2\pi i t \ell) \text{ for } t \in \mathbb{T} = \mathbb{R}/\mathbb{Z} ,\\
&E(\theta) = \{t \in \mathbb{T} : |\varphi(t)| \geq \cos\theta\} \text{ for } 0 \leq \theta \leq \pi/2 ,\\
&\theta^* \text{ be such that } \cos\theta^* = \min|\varphi(t)| \text{ and } 0 \leq \theta^* \leq \pi/2 .
\end{aligned}$$

The proof of Theorem 2 will be based on the following two results, for the first of which we give a sketch of a proof.

***Lemma 1* (cf. [4]).** — *For $\theta_1 \geq 0$, $\theta_2 \geq 0$ and $\theta_1 + \theta_2 \leq \frac{\pi}{2}$, we have*

$$E(\theta_1) + E(\theta_2) \subset E(\theta_1 + \theta_2) .$$

Proof. — For $j = 1, 2$, we consider t_j in $E(\theta_j)$, and let $\alpha_j = \arg^t \varphi(t_j)$ and $\lambda_j = \sqrt{1 - |\varphi(t_{3-j})|^2} e^{-i\alpha_j}$. We use the Cauchy inequality to get an upper bound on

$$|\lambda_1 \varphi(t_1) + \lambda_2 \varphi(-t_2)|^2 = \left| \sum_\ell \sqrt{p_\ell} \left(\lambda_1 \sqrt{p_\ell} e^{2\pi i \ell t_1} + \lambda_2 \sqrt{p_\ell} e^{-2\pi i \ell t_2} \right) \right|^2 .$$

□

***Lemma 2* (Macbeath-Kneser Theorem, cf. [8], p. 13-14).** — *Let E_1 and E_2 be two non-empty closed sets in $\mathbb{T}$. We have*

$$|E_1 + E_2| \geq \min(1, |E_1| + |E_2|) ,$$

where $|A|$ represents the Haar measure of A in $\mathbb{T}$.

Proof of Theorem 2. — We may of course assume that p is strictly less that 1 and so θ^* is strictly positive. Our first task is to show that

$$|E(\theta)| \leq 12 \frac{\theta p}{\sqrt{1-p}}, \text{ for } \theta \in]0, \theta^*/2[, \tag{2.1}$$

and

$$|E(\theta)| \leq p/\cos^2\theta, \text{ for } \theta \in]0, \frac{\pi}{2}] . \tag{2.2}$$

By the definition of θ^*, we have

$$\cos^2\theta^* \le \int_{\mathbb{T}} |\varphi(t)|^2 dt = \sum_\ell p_\ell^2 \le p \sum_\ell p_\ell = p \ ,$$

whence

$$\theta^* > |\sin\theta^*| = \sqrt{1-\cos^2\theta^*} \ge \sqrt{1-p} \ . \tag{2.3}$$

Now let θ be in $]0, \theta^*/2[$, and let $M = [\theta^*/\theta]$; we have $M \ge 2$. We may write

$$\begin{aligned} p &\ge \int_{\mathbb{T}} |\varphi(t)|^2 dt \\ &\ge \sum_{m=1}^{M} \int_{\cos(m\theta)\le|\varphi(t)|<\cos((m-1)\theta)} |\varphi(t)|^2 dt \\ &\ge \sum_{m=1}^{[M/2]} |E(m\theta)\backslash E\big((m-1)\theta\big)| \cos^2 m\theta \ , \end{aligned}$$

and by appealing to Lemma 2, we get

$$\begin{aligned} p &\ge |E(\theta)| \sum_{m=1}^{[M/2]} \cos^2 m\theta \\ &\ge |E(\theta)| \cdot [M/2] \cdot \frac{1}{2} \\ &\ge \frac{M}{6}|E(\theta)| \\ &\ge \frac{\theta^*}{12\theta}|E(\theta)| \ . \end{aligned}$$

This last inequality and (2.3) imply (2.1). Inequality (2.2) simply follows from

$$p \ge \int_{\mathbb{T}} |\varphi(t)|^2 dt \ge |E(\theta)| \cdot \cos^2\theta \ .$$

For every integer N, we have

$$\begin{aligned} P\{S_n = N\} &= \int_{\mathbb{T}} \varphi^n(t) \exp(-2\pi i N t) dt \\ &\le \int_{\mathbb{T}} |\varphi(t)|^n dt \ . \end{aligned}$$

By change of variable, and then integration by parts, we have

$$\begin{aligned}\int\limits_{\mathbb{T}} |\varphi(t)|^n dt &= \int\limits_0^{\frac{\pi}{2}} \cos^n \theta d|E(\theta)| \\ &= \int\limits_0^{\frac{\pi}{2}} n\cos^{n-1}\theta \sin\theta \cdot |E(\theta)| d\theta \ .\end{aligned}$$

We now use (2.1), getting

$$\begin{aligned}&\int\limits_0^{\theta^*/2} n\cos^{n-1}\theta\sin\theta|E(\theta)|d\theta \\ &\qquad\qquad \le \frac{12p}{\sqrt{1-p}}\int\limits_0^{\frac{\pi}{2}} \cos^n\theta d\theta \\ &\qquad\qquad \le \frac{12p\sqrt{\pi/2}}{\sqrt{(1-p)}\sqrt{n}} \ .\end{aligned}$$

On the remaining interval $[\theta^*/2, \pi/2]$, we use (2.2), which leads, for $n \geq 3$, to

$$\begin{aligned}\int\limits_{\theta^*/2}^{\pi/2} n\cos^{n-1}\theta\sin\theta\frac{p}{\cos^2\theta}d\theta &\le \frac{n}{n-2}p\int\limits_{\theta^*/2}^{\pi/2}(n-2)\cos^{n-3}\theta\sin\theta d\theta \\ &= \frac{np}{n-2}\cos^{n-2}(\frac{\theta^*}{2}) \\ &\le \frac{np}{n-2}\cos^{n-2}\left(\frac{\sqrt{(1-p)}}{2}\right) .\end{aligned}$$

From the inequality $x\cos^n x \le \frac{1}{\sqrt{n}}$,we get

$$\begin{aligned}&\int\limits_{\theta^*/2}^{\pi/2} n\cos^{n-1}\theta\sin\theta|E(\theta)|d\theta \\ &\qquad\qquad \le \frac{2np}{(n-2)\sqrt{(n-2)(1-p)}} \\ &\qquad\qquad \le \frac{12p}{\sqrt{n(1-p)}}.\end{aligned}$$

We have completed the proof of Theorem 2 when $n \geq 3$. For $n \leq 2$, we simply have to notice that $P\{S_n = N\} \leq p \leq \frac{40p}{\sqrt{n(1-p)}}$. □

3. Proof of Theorem 1

The letter c (with or without indices) always represents a constant that depends on σ, ε, A and a at most. The value may change from one occurrence to the next.

The key ingredient which will play the rôle of Lemma 2 is the following result, which will be applied with $k = 2^{2/\sigma}$.

***Lemma 3* (Freiman - Moskvin - Yudin [4]).** — *Let $2 < k < 3$; there exists a positive real number $\mu = \mu(k)$ such that for any closed set F in $\mathbb{T}$, symmetric with respect to the origin and satisfying:*

$$|F| \leq \mu \quad \text{and} \quad |2F| < k|F| \ ,$$

there exists a positive integer q such that

$$\bigcup_{r=0}^{q-1} \left[\frac{r}{q} - \frac{|F|}{q}, \frac{r}{q} + \frac{|F|}{q}\right] \subset 3F \ .$$

We give two further results connecting the values of the characteristic function with the concentration of the associated distribution, the second one being of an arithmetical nature.

Lemma 4. — *Let α and L^{-1} be two positive real numbers. If $|\varphi(t)| \geq 1 - \alpha$ for all t in $\left[-\frac{1}{2L}, \frac{1}{2L}\right]$, then $Q(X_1; L) \geq 1 - 6\alpha$.*

Proof. — Let us write $Q = Q(L)$, and define x_0 the supremum of

$$\{x : P\{X_1 \leq x\} \leq (1 - Q)/2\}.$$

Since $x \mapsto P\{X_1 \leq x\}$ is continuous from the right, we have $P\{X_1 \leq x_0\} \geq (1-Q)/2$. On the other hand, for $x < x_0$, we have

$$P\{X_1 \geq x + L\} \geq 1 - P\{x < X_1 \leq x + L\} - P\{X_1 \leq x\} \geq (1 - Q)/2 \ ,$$

and, since $P\{X_1 \geq x + L\}$ is continuous from the left, we have $P\{X_1 \geq x_0 + L\} \geq (1 - Q)/2$.

Let us define

$$\varphi_-(t) = \sum_{k \leq x_0} p_k \exp(2\pi i k t), \quad \varphi_+(t) = \sum_{\ell \geq x_0 + L} p_\ell \exp(2\pi i \ell t)$$

and $\varphi_0(t) = \sum\limits_{x_0 < m < x_0 + L} p_m \exp(2\pi i m t)$.

Let us denote by I the interval $\left[-\frac{1}{2L}, \frac{1}{2L}\right]$. By the Schwarz inequality, we get

$$\begin{aligned}\left(L\int_I |\varphi_-(t) &+ \varphi_+(t)|dt\right)^2 \leq L\int_I |\varphi_-(t)+\varphi_+(t)|^2 dt \\ &\leq \varphi_-^2(0)+\varphi_+^2(0)+2L\sum_{k\leq x_0}\sum_{\ell\geq x_0+L} p_k p_\ell \left|\int_I \cos\big(2\pi(k-\ell)t\big)dt\right| \\ &\leq \varphi_-^2(0)+\varphi_+^2(0)+\frac{2}{\pi}\varphi_-(0)\varphi_+(0) \\ &\leq \Big(\big(\varphi_-(0)+\varphi_+(0)\big)^2-\Big(2-\frac{2}{\pi}\Big)\varphi_-(0)\varphi_+(0)\Big) \\ &= \big(\varphi_-(0)+\varphi_+(0)\big)^2\left(1-\Big(2-\frac{2}{\pi}\Big)\frac{\varphi_-(0)\varphi_+(0)}{\big(\varphi_-(0)+\varphi_+(0)\big)^2}\right) \\ &\leq \big(\varphi_-(0)+\varphi_+(0)\big)^2\left(1-\Big(1-\frac{1}{\pi}\Big)\frac{\varphi_-(0)\varphi_+(0)}{\big(\varphi_-(0)+\varphi_+(0)\big)^2}\right)^2 ,\end{aligned}$$

and so we get

$$L\int_I |\varphi_-(t)+\varphi_+(t)|dt \leq \varphi_-(0)+\varphi_+(0)-\frac{2}{3}\,\frac{\varphi_-(0)\varphi_+(0)}{\varphi_-(0)+\varphi_+(0)} .$$

This leads to

$$1-\alpha \leq L\int_I |\varphi(t)|dt \leq \varphi_0(0)+\varphi_-(0)+\varphi_+(0)-\frac{2}{3}\,\frac{\varphi_-(0)\varphi_+(0)}{\varphi_-(0)+\varphi_+(0)} .$$

Since the minimum of $\frac{xy}{x+y}$ under the conditions $\frac{1-Q}{2} \leq x,\ y \leq \frac{1+Q}{2}$ and $1-Q \leq x+y \leq 1$ is $\frac{1-Q}{4}$, we have

$$1-\alpha \leq 1-\frac{2}{3}\left(\frac{1-Q}{4}\right)$$

which implies

$$1-Q \leq 6\alpha.$$

□

Lemma 5. — *Let $q \geq 1$ be an integer. We have*

$$\frac{1}{q}\sum_{r=0}^{q-1}|\varphi(\frac{r}{q})|^2 \leq \max_s \sum_{\ell\equiv s (\mathrm{mod}\, q)} p_\ell .$$

Proof. — We have

$$\varphi\Big(\frac{r}{q}\Big) = \sum_\ell p_\ell \exp\Big(2\pi i\ell\frac{r}{q}\Big) = \sum_{s=0}^{q-1}\exp\Big(2\pi i s\frac{r}{q}\Big)P_s ,$$

where $P_s = \sum_{\ell \equiv s \bmod q} p_\ell$, so that

$$\begin{aligned}
\frac{1}{q}\sum_{r=0}^{q-1}|\varphi(\tfrac{r}{q})|^2 &= \frac{1}{q}\sum_{r=0}^{q-1}\sum_{s=0}^{q-1}\sum_{t=0}^{q-1}\exp\left(2\pi i\frac{r(s-t)}{q}\right)P_sP_t\\
&= \sum_{s=0}^{q-1}\sum_{t=0}^{q-1}\left\{\frac{1}{q}\sum_{r=0}^{q-1}\exp\left(2\pi i\frac{r(s-t)}{q}\right)\right\}P_sP_t\\
&= \sum_{s=0}^{q-1}P_sP_s\\
&\le \left(\max_{s \bmod q} P_s\right)\sum_{s=0}^{q-1}P_s \ ,
\end{aligned}$$

but

$$\sum_{s=0}^{q-1}P_s = \sum_{s=0}^{q-1}\sum_{\ell\equiv s(\bmod q)} p_\ell = \sum_\ell p_\ell = 1 \ ,$$

and Lemma 5 is proved. □

We keep the notation that we introduced at the beginning of section 2, and further let

$$\theta_1 = \min\left(\theta^*/4, \frac{\mu\sqrt{1-p}}{12p}, \sqrt{\varepsilon}/2, \frac{\sqrt{1-p}}{12pA}\right) \ , \tag{3.1}$$

so that (2.1) implies that $|E(\theta)| \le \mu$, for $\theta \le \theta_1$.

Let us assume that there exists $\theta_0 \le \theta_1$ such that $|2E(\theta_0)| < k|E(\theta_0)|$. According to Lemma 3, one of the following possibilities holds true:

$$\exists q \ge 2 : \left\{\tfrac{0}{q}, \tfrac{1}{q}, \dots, \tfrac{q-1}{q}\right\} \subset E(3\theta_0) \tag{3.2}$$

$$\Big[-|E(\theta_0)|, |E(\theta_0)|\Big] \subset E(3\theta_0) \ . \tag{3.3}$$

Let us first assume that (3.2) is satisfied. We use Lemma 5, which leads to

$$\max_s \sum_{\ell\equiv s(\bmod q)} p_\ell \ge \cos^2 3\theta_0 \ ,$$

and by condition (1.2) of Theorem 1, we obtain

$$\cos^2 3\theta_0 \le 1-\varepsilon \ ,$$

whence

$$\theta_0 \ge \sqrt{\varepsilon/3} \ ,$$

in contradiction with the inequalities $\theta_0 \le \theta_1 \le \sqrt{\varepsilon}/2$.

Hence, condition (3.3) is satisfied. Lemma 4 then leads us to

$$Q(X_1; |E(\theta_0)|^{-1}) \ge 1 - 6\big(1-\cos(3\theta_0)\big)$$

and by condition (1.3) of Theorem 1,

$$1 - 6(1 - \cos 3\theta_0) \leq 1 - a|E(\theta_0)|^\sigma \ ,$$

whence

$$a|E(\theta_0)|^\sigma \leq 6\big(1 - \cos(3\theta_0)\big)$$

or

$$|E(\theta_0)| \leq c\theta_0^{2/\sigma} \ . \tag{3.4}$$

Let us summarize what we have proved so far, remembering that Lemma 1 implies $2E(\theta) \subset E(2\theta)$. For $\theta \leq \theta_1$, we have

$$\text{either} \quad k|E(\theta)| \leq |E(2\theta)| \quad \text{or} \quad |E(\theta)| \leq c\theta^{2/\sigma} \ .$$

Our next step is to prove that for $\theta \leq \theta_1$, we have

$$|E(\theta)| \leq k \max\left(c, \frac{|E(\theta_1)|}{\theta_1^{2/\sigma}}\right) \theta^{2/\sigma} \ . \tag{3.5}$$

Indeed, by induction, we have for any $\ell \geq 0$:

$$|E(\theta_1/2^\ell)| \leq \max\left(c, \frac{|E(\theta_1)|}{\theta_1^{2/\sigma}}\right) (\theta_1/2^\ell)^{2/\sigma} \ ,$$

either by using

$$|E(\theta_1/2^{\ell+1})| \leq \frac{1}{k}|E(\theta_1/2^\ell)| = \frac{1}{2^{2/\sigma}}|E(\theta_1/2^\ell)|,$$

or directly

$$|E(\theta_1/2^{\ell+1})| \leq c(\theta_1/2^{\ell+1})^{2/\sigma}$$

when the previous inequality does not hold. Now, for $\theta \leq \theta_1$, we choose ℓ such that $\theta_1/2^{\ell+1} < \theta \leq \theta_1/2^\ell$, and notice that $E(\theta) \subset E(\theta_1/2^\ell)$, so that we have

$$\begin{aligned} |E(\theta)| &\leq \max\left(c, \frac{|E(\theta_1)|}{\theta_1^{2/\sigma}}\right) (\theta_1/2^\ell)^{2/\sigma} \\ &\leq k \max\left(c, \frac{|E(\theta_1)|}{\theta_1^{2/\sigma}}\right) (\theta_1/2^{\ell+1})^{2/\sigma} \ , \end{aligned}$$

whence (3.5) follows.

As in section 2, we write

$$\begin{aligned} P\{S_n = N\} &\leq \int_{\mathbb{T}} |\varphi(t)|^n dt \\ &\leq \int_0^{\frac{\pi}{2}} n \cos^{n-1}\theta \sin\theta |E(\theta)| d\theta \ , \end{aligned}$$

and we majorize $|E(\theta)|$ by (3.5) on the set $[0, \theta_1]$, and by (2.2) on $\left[\theta_1, \frac{\pi}{2}\right]$. We get, for $n \geq 3$

$$\begin{aligned} P\{S_n = N\} \quad &\leq \quad k \max\left(c, \frac{|E(\theta_1)|}{\theta_1^{2/\sigma}}\right) \int_0^{\theta_1} n \cos^{n-1}\theta \sin\theta \cdot \theta^{2/\sigma} d\theta \\ &\quad + \frac{n}{n-2} \cos^{n-2}\theta_1 \ . \end{aligned}$$

We now find upper bounds for the terms containing θ_1, as well as for the integral. We have $\theta_1 \leq \pi/8$ so that $\theta_1^{2/\sigma} \leq (\pi/8)^{2/\sigma}$, and

$$|E(\theta_1)| \leq \frac{1}{\cos^2\theta_1} \leq \frac{1}{\cos^2 \pi/8} \ .$$

By (2.3), we have $\theta^* \geq \sqrt{1-p} \geq \sqrt{\varepsilon}$, so that θ_1, defined in (3.1), depends at most on μ (and so on σ), ε, A, a, so that $\cos^{n-2}\theta_1 \leq c/n^2$.

We finally have, for $1 \leq \ell \leq 2$

$$\begin{aligned} \int_0^{\theta_1} n \cos^{n-1}\theta \sin\theta \cdot \theta^{\ell} d\theta \quad &\leq \quad 2 \int_0^{\frac{\pi}{8}} \theta^{\ell-1} \cos^n \theta d\theta \\ &\leq \quad 2 \int_0^{\frac{\pi}{8}} \theta^{\ell-1} \exp\left(-\frac{\theta^2 n}{2}\right) d\theta \\ &\leq \quad 2n^{-\ell/2} \int_0^{\infty} (\theta\sqrt{n})^{\ell-1} \exp\left(-\frac{\theta^2 n}{2}\right) d(\theta\sqrt{n}) \\ &\leq \quad cn^{-\ell/2} \ , \end{aligned}$$

where c is an absolute constant.

We have thus obtained

$$P\{S_n = N\} \leq cn^{-1/\sigma} \ ,$$

where c depends at most on σ, ε, A and a. □

We are thankful to Ruth Lawrence for her careful reading of the paper.

References

[1] Arak T.V. and Zaitsev A.Yu., *Uniform limit theorems for sums of independent random variables*, Proceedings of the Steklov Institute of Mathematics, issue **1**, 1988, 41–61.

[2] Doeblin W. and Lévy P., *Sur les sommes de variables aléatoires indépendantes à dispersions bornées inférieurement*, C.R. Acad. Sci. Paris **202**, 1936, 2027–2029.

[3] Esséen C.G., *On the Kolmorogov-Rogozin inequality for the concentration function*, Z. Wahrscheinlichkeitstheorie und verw. Gebiete, **5**, 1966, 210–216.

[4] Freiman G.A., Moskvin D.A. and A.A. Yudin A.A., *Inverse problems of additive number theory and local limit theorems for lattice random variables*, (in Russian), Number theoretic studies in the Markov spectrum and in the structure theory of set addition, Kalinin Gos. Univ., Moscow, 1973, 148–162.

[5] Kesten H., *A sharper form of the Doeblin-Lévy-Kolmogorov-Rogozin inequality for concentration functions*, Math. Scand. **25**, 1969, 133–144.

[6] Kolmogorov A.N., *Sur les propriétés des functions de concentration de M.P. Lévy*, Ann. Inst. H. Poincaré **16**, 1958-1960, 27–34.

[7] Ledoux M. and Talagrand M. *Probability in Banach spaces. Isoperimetry and processes*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), **23**, Springer-Verlag, Berlin, 1991, xii+480 pp. ISBN: 3-540-52013-9.

[8] Mann H.B., *The Addition Theorems of Groups Theory and Number Theory*, Interscience, **18**, J. Wiley, New York, 1965.

[9] Petrov V.V., *Limit theorems for sums of independent random variables*, (Russian), Probability Theory and Mathematical Statistics, "Nauka", Moscow, 1987, 318 pp.

[10] Postnikova L.P. and Yudin A.A., *On the concentration function*, Th. Probability Appl., **22**, 1977, 302–305.

[11] Postnikova L.P. and Yudin A.A., *An analytic method for estimates of the concentration function*, Proc. Steklov Inst. Math., 1980, 153–161.

[12] Rogozin B.A., *On the increase of dispersion of sums of independent random variables*, Th. Probability Appl., **6**, 1961, 97–99.

J.-M. Deshouillers, Mathématiques stochastiques, Université Bordeaux 2, BP 26, 33076 Bordeaux, France • *E-mail :* `j-m.deshouillers@u-bordeaux2.fr`

G.A. Freiman, School of Mathematical Sciences, Department of Mathematics, Raymond and Beverly Sackler, Faculty of Exact Sciences, Tel Aviv University, 69978 Tel Aviv, Israel *E-mail :* `grisha@math.tau.ac.il`

A. Yudin, Department of Mathematics, Vladimir Pedagogical University, 11, pr. Stroiteley, Vladimir, Russia • *E-mail :* `aayudin@vgpu.elcom.ru`

AST

Structure theory of set addition - Pages préliminaires

Astérisque, tome 258 (1999), p. I-XXII

<http://www.numdam.org/item?id=AST_1999__258__R1_0>

ASTÉRISQUE 258

STRUCTURE THEORY OF SET ADDITION

edited by

Jean-Marc Deshouillers

Bernard Landreau

Alexander A. Yudin

Société Mathématique de France 1999

Jean-Marc Deshouillers

Mathématiques stochastiques, Université Bordeaux 2, BP 26,
33076 Bordeaux, France.

E-mail : `j-m.deshouillers@u-bordeaux2.fr`

Bernard Landreau

Laboratoire A2X, Université Bordeaux 1, 33405 Talence Cedex, France.

E-mail : `landreau@math.u-bordeaux.fr`

Alexander A. Yudin

Department of Mathematics, Vladimir Pedagogical University,
11, pr. Stroiteley, Vladimir, Russia.

E-mail : `aayudin@vgpu.elcom.ru`

1991 *Mathematics Subject Classification.* — 05-XX, 11Bxx, 11Hxx, 11Lxx, 11Pxx, 20Cxx, 20Dxx, 20Exx, 20Fxx, 60Exx, 60Fxx, 68Qxx, 90Cxx, 94Bxx.

Key words and phrases. — additive number theory, combinatorial number theory, additive finite groups, structure theory of set addition, inverse additive problems, sums of discrete random variables, subset sums, integer programming, coding theory.

STRUCTURE THEORY OF SET ADDITION

edited by Jean-Marc Deshouillers, Bernard Landreau, Alexander A. Yudin

Abstract. — For a long time, additive number theory, motivated by conjectures such as that of Goldbach or Waring, has been concerned by the study of additive properties of *special* sequences. In the 1930's it was noticed that the consideration of the additive properties of *general* sequences turned out, not only to be a beautiful subject for its own sake, but was able to lead to improvements in the study of special sequences: thus, in the paper founding this philosophy, Schnirel'man introduced a density on sets of integers, gave a general lower bound for the density of the sum of two sets, and applied it to the special sequence of primes to show that every integer can be written as a sum of a uniformly bounded number of primes. Additive number theory evolved towards the definition of invariants for sets of (non-necessarily commutative) monoids and the study of the invariants for the "sum" of different sets in terms of the invariants of those sets.

A new trend appeared in the 1950's, with authors like M. Kneser and G. A. Freiman, which is sometimes described as *inverse* additive theory: knowing that the relation between the invariants of a family of sets and the invariant of their sum is extremal (or close to), what can be said on the *structure* of the sets themselves ?

In the recent years, there has been a renewed interest for this approach which turns out to have applications to different others fields. It seemed appropriate to gather in a single volume 24 contemporary original research papers and 3 survey articles dealing with *the structure theory of set addition* and its applications to elementary or combinatorial number theory, group theory, integer programming and probability theory.

***Résumé* (Problèmes additifs inverses).** — La théorie additive des nombres, motivée par des conjectures telles que celles de Goldbach ou Waring, s'est longtemps consacrée à l'étude des propriétés additives de suites *particulières*. Dans les années 1930, on a remarqué que la considération des propriétés additives de suites *générales*, non seulement constituait un magnifique sujet en lui-même, mais en outre permettait des améliorations dans l'étude de suites particulières : ainsi, dans l'article fondateur de cette problématique, Schnirel'man a introduit une notion de densité sur les suites d'entiers, donné une minoration de la densité de la somme de deux suites et l'a appliquée à l'ensemble des nombres premiers montrant que tout entier peut être

représenté comme une somme de nombres premiers, avec un nombre de termes uniformément borné. La théorie additive des nombres a évolué vers la définition d'invariants pour des parties de monoïdes (non nécessairement commutatifs) et l'étude des invariants de la somme d'ensembles en fonction des invariants liés à ces ensembles.

Une nouvelle tendance est apparue dans les années 1950, avec les travaux de M. Kneser et G.A. Freiman, que l'on désigne parfois sous le vocable de théorie additive *inverse* : sachant que le rapport entre les invariants d'une famille d'ensembles et l'invariant de leur somme est extrêmal (ou presque extrêmal), que peut-on dire de la *structure* des ensembles eux-mêmes ?

Cet abord a connu récemment un regain d'intérêt qui se trouve porter ses fruits dans d'autres domaines. Il a semblé judicieux de regrouper en un unique volume 24 articles de recherches originaux et 3 synthèses ayant trait à cette théorie de la structure des sommes d'ensembles et ses applications à la théorie des nombres élémentaire ou combinatoire, à la théorie des groupes, à la programmation entière et à la théorie des probabilités.

Contents

Introduction

Additive Number Theory

Combinatorial Number Theory

Algebra

Coding Theory

Integer Programming

Probability

RÉSUMÉS DES ARTICLES

Structure Theory of Set Addition
Gregory A. Freiman 1

Nous présentons une synthèse des résultats fondamentaux de la théorie connue sous le nom de “structure theory of set addition” et de leurs applications à d'autres domaines.

Sets of integers with large trigonometric sums
Amnon Besser 35

Nous cherchons à optimiser, pour un entier k et un réel u fixés, sur tous les ensembles $K = \{a_1 < a_2 < \cdots < a_k\} \subset \mathbb{Z}$, la mesure de l'ensemble des $\alpha \in [0,1]$ tels que la valeur absolue de la somme trigonométrique $S_K(\alpha) = \sum_{j=1}^{k} e^{2\pi i \alpha a_j}$ soit supérieure à $k - u$. Lorsque u est suffisamment petit par rapport à k, nous sommes en mesure de construire un ensemble K_{ex} qui est presque optimal. Cet ensemble est une union finie de progressions arithmétiques. Nous montrons que tout ensemble plus performant, s'il existe, a une structure similaire à celle de K_{ex}. On obtient également des bornes inférieures et supérieures précises pour la mesure maximale.

Structure of sets with small sumset
Yuri Bilu 77

Freiman a démontré qu'un ensemble fini d'entiers K satisfaisant $|K+K| \leq \sigma|K|$ est nécessairement un sous-ensemble d'une petite progression arithmétique généralisée de rang m avec $m \leq \lfloor \sigma - 1 \rfloor$. Nous donnons une preuve complète de ce résultat accompagnée de quelques améliorations ainsi que du calcul explicite des constantes impliquées.

On finite addition theorems
András Sárkőzy 109

Si un ensemble fini A d'entiers inclus dans $\{1, \ldots, N\}$ a plus de N/k éléments, on peut s'attendre à ce que l'ensemble ℓA des sommes de ℓ éléments de

A, contienne, quand ℓ est comparable à k, une progression arithmétique (homogène ou non) assez longue. Après la présentation de l'état des lieux, nous montrons que certains de ces résultats ne peuvent pas être améliorés autant que la considération du cas infini pourrait le laisser prévoir. L'article s'achève sur un résultat fournissant des majorations et minorations de l'ordre, en tant que base asymptotique, des sous-suites, de densité relative positive, des nombres premiers.

On Freiman's Theorems concerning the sum of two finite sets of integers
JOHN STEINIG 129

En suivant les indications données par Freiman [1], cet article fournit une preuve détaillée de ses deux théorèmes minorant $|M+N|$, où M et N sont des sous-ensembles finis de $\mathbb{Z}$.

On an additive problem of Erdős and Straus, 2
JEAN-MARC DESHOUILLERS & GREGORY A. FREIMAN 141

On désigne par $s^\wedge A$ l'ensemble des entiers qui peuvent s'écrire comme somme de s éléments distincts de A. L'ensemble A est dit admissible si et seulement si $s \neq t$ implique que $s^\wedge A$ et $t^\wedge A$ n'ont aucun élément en commun.

P. Erdős a conjecturé qu'un ensemble admissible inclus dans $[1, N]$ a un cardinal maximal lorsque A est constitué d'entiers consécutifs situés à l'extrémité supérieure de l'intervalle $[1, N]$. L'objet de cet article est de donner une preuve de la conjecture d'Erdős, pour N suffisamment grand.

On the structure of sum-free sets, 2
JEAN-MARC DESHOUILLERS, GREGORY A. FREIMAN, VERA SÓS & MIKHAIL TEMKIN 149

On dit qu'un ensemble d'entiers positifs est additivement libre si l'ensemble $\mathbb{A} \cap (\mathbb{A}+\mathbb{A})$ est vide, où $\mathbb{A}+\mathbb{A}$ désigne l'ensemble des sommes de deux éléments de $\mathbb{A}$ non nécessairement distincts. Améliorant un résultat précédent de G.A. Freiman, on donne une description précise de la structure des ensembles additivement libres inclus dans $[1, M]$ de cardinalité au moins $0.4M - x$ pour $M \geq M_0(x)$ (où x est un entier arbitraire).

Sumsets with distinct summands and the Erdős-Heilbronn conjecture on sums of residues
GREGORY A. FREIMAN, LEWIS LOW & JANE PITMAN 163

Soit S un ensemble d'entiers ou de classes de résidus modulo un nombre premier p, de cardinalité $|S| = k$, et soit T l'ensemble de toutes les sommes de deux éléments distincts de S. Dans le cas des entiers, on démontre que, si $|T|$ est plus petit qu'un nombre proche de $2.5k$, alors S est contenu dans une progression arithmétique de cardinal relativement petit. Dans le cas des résidus, un résultat du même genre est obtenu, pourvu que $k > 60$ et $p > 50k$. Comme application, on prouve que $|T| \geq 2k - 3$ sous ces conditions. Des résultats antérieurs de Freiman jouent un rôle essentiel dans les démonstrations.

On the number of sums and differences
FRANÇOIS HENNECART, GILLES ROBERT & ALEXANDER YUDIN 173

Dans cet article, nous montrons que $\inf_{A \subset \mathbb{Z}} \ln |A+A| / \ln |A-A|$ est inférieur à $0,7865$, améliorant en cela un résultat antérieur dû à G. Freiman et W. Pigarev.

The structure of multisets with a small number of subset sums
VSEVOLOD F. LEV 179

On recherche ici des ensembles d'entiers naturels $A = \{a_1, \ldots, a_k\}$ (avec répétitions possibles) tels que l'ensemble des sommes $P(A) = \{\varepsilon_1 a_1 + \cdots + \varepsilon_k a_k \colon 0 \leq \varepsilon_1, \ldots, \varepsilon_k \leq 1\}$ est petit. Précisément, soit A un tel ensemble pour lequel le cardinal de $P(A)$ est borné par un multiple fixe du cardinal de A (i.e. $|P(A)| \ll |A|$), nous montrons que l'ensemble $P(A)$ est alors la réunion d'un petit nombre de progressions arithmétiques de même raison.

Des problèmes similaires ont déjà été considérés par G. Freiman [1] et M. Chaimovich [2]. À la différence de ces articles, nos conditions s'expriment seulement à l'aide du cardinal de $P(A)$ sans faire appel au plus grand élément de A.

Subset sums of sets of residues
EDITH LIPKIN 187

On appelle nombre critique d'un groupe abélien G, le plus petit entier naturel m vérifiant la propriété suivante :
pour toute partie A de G avec $|A| \geq m, 0 \notin A$, l'ensemble A^* des sommes partielles de A est égal à G. Dans cet article, on démontre la conjecture de G. Diderrich concernant la valeur du nombre critique du groupe G, lorsque $G = \mathbb{Z}_q$, pour q suffisamment grand.

Inverse theorems and the number of sums and products
MELVYN B. NATHANSON & GÉRALD TENENBAUM 195

Soit $\epsilon > 0$. Erdős et Szemerédi ont conjecturé que, si A est un ensemble de k nombres entiers positifs avec k assez grand, le nombre des entiers qui sont représentables comme somme ou produit de deux éléments de A est au moins égal à $k^{2-\epsilon}$. Nous confirmons cette conjecture dans le cas particulier où le nombre des sommes est très petit.

Stratified Sets
JEAN-LOUIS NICOLAS 205

On dit qu'un ensemble $\mathcal{A}$ de nombres entiers est "stratifié" si, pour tout t, $0 \leq t < \operatorname{Card} \mathcal{A}$, la somme de t éléments distincts de $\mathcal{A}$ est toujours strictement inférieure à la somme de $t+1$ éléments distincts de $\mathcal{A}$. Cela implique que les éléments de $\mathcal{A}$ sont positifs. On démontre que le nombre d'ensembles stratifiés de plus grand élément N est exactement égal au nombre $p(N)$ de partitions de N.

On the structure of sets of lattice points in the plane with a small doubling property
Yonutz V. Stanchescu 217

On décrit la structure des ensembles $\mathbb{K}$ de points d'un réseau plan tels que $|\mathbb{K}+\mathbb{K}|$ est petit comparé à $|\mathbb{K}|$. Soit $\mathbb{K}$ un sous-ensemble fini de $\mathbb{Z}^2$ tel que

$$|\mathbb{K}+\mathbb{K}| < 3.5|\mathbb{K}| - 7.$$

Si $\mathbb{K}$ est porté par trois droites parallèles, alors l'enveloppe convexe de $\mathbb{K}$ est contenu dans trois progressions arithmétiques compatibles de même raison ayant en totalité au plus

$$|\mathbb{K}| + \frac{3}{4}\Big(|\mathbb{K}+\mathbb{K}| - \frac{10}{3}|\mathbb{K}| + 5\Big)$$

termes. Cette majoration est optimale.

Non-solvable groups with a large fraction of involutions
Yakov Berkovich 241

Dans cet article, on classifie les groupes finis G non résolubles tels que le nombre de classes de G est au moins $|G|/16$. On en déduit certaines conséquences.

Questions on set squaring in groups
Yakov Berkovich 249

Quelques questions sur des petits sous-ensembles de groupes sont posées et discutées.

On groups generated by a pair of elements with small third or fourth power
Sergei Brodsky 255

Cet article se propose d'étudier les groupes bi-générés, tels que la puissance m-ème de la paire génératrice contienne moins de $2m$ éléments. Nous prouvons en particulier, que si le cube de la paire génératrice contient moins de 7 éléments ou si la puissance quatrième contient moins de 11 éléments, alors le groupe est résoluble. Sinon, il n'est pas nécessairement résoluble. Les démonstrations sont effectuées à l'aide de calculs par ordinateurs.

On small subset product in a group
Yahya Ould Hamidoune 281

Nous généralisons des théorèmes d'addition connus pour le cas des groupes non abéliens.

Les preuves classiques des théorèmes d'addition utilisent des transformations locales dues à Davenport, Dyson et Kempermann.

Notre approche est basée sur l'étude de certains blocs d'imprimitivité du groupe d'automorphismes d'une relation.

Cet article présente des résultats et des problèmes ouverts sur les sujets suivants : groupes avec sous-tables de multiplication déficientes, bases multiplicatives des groupes finis.

On étudie dans cet article la structure des paires de parties finies A, B d'un groupe abélien pour lesquelles les sommes sont peu nombreuses : $|A + B| < |A| + |B|$. En 1960, J. H. B. Kemperman en a donné une description complète de nature récursive mais relativement compliquée. En utilisant des résultats intermédiaires de Kemperman, on obtient ici une description d'une autre nature. Bien qu'elle ne soit pas suffisante d'un point de vue général, notre description a l'avantage d'être claire et intuitive, et peut être utilisée pour des applications.

On montre que pour un groupe abélien G, tel que l'ordre des éléments est majoré par un entier r, tout ensemble ayant n éléments et au plus αn sommes est contenu dans un sous-groupe de taille Cn avec $C = f(r, \alpha)$ dépendant de r et α mais non de n. C'est un résultat analogue au Théorème de G. Freiman qui décrit la sructure de tels ensembles dans le groupe des entiers.

Nous nous intéressons à quelques problèmes additifs dans le groupe $(\mathbb{Z}/2\mathbb{Z})^r$. Notre propos est de montrer comment ces problèmes sont étroitement liés à la théorie des codes correcteurs. Nous présentons des techniques classiques de codage que nous utilisons pour obtenir quelques contributions originales.

Cet article de synthèse présente un nouvel abord de la programmation entière basée sur la caractérisation de configurations extrêmes en théorie additive des nombres. La structure de ces configurations extrêmes nous permet d'élaborer des algorithmes applicables à des familles suffisamment larges de problèmes; ces algorithmes améliorent notablement les bornes actuellement connues. Là où ils sont applicables, ces algorithmes sont polynômiaux voire linéaires; c'est en particulier le cas pour les problèmes de type sac à dos. Pour cette classe de problèmes, l'amélioration sur les algorithmes antérieurs est d'au moins de deux ordres de grandeur.

New Algorithm for Dense Subset-Sum Problem
MARK CHAIMOVICH 363

On présente un nouvel algorithme pour le problème des sommes partielles (subset-sum problem) dans le cas dense. Il est basé sur une caractérisation de la famille des sommes partielles obtenue par des méthodes analytiques de la théorie additive des nombres. L'algorithme fonctionne pour un grand nombre de sommants (m) avec des valeurs qui sont majorées. La borne (ℓ) dépend modérément de m. Le temps requis par ce nouvel algorithme est en $O(m^{7/4} / \log^{3/4} m)$, ce qui est plus rapide que les précédents algorithmes connus, le meilleur d'entre eux prenant un temps en $O(m^2 / \log^2 m)$.

On the Two-Dimensional Subset Sum Problem
ALAIN PLAGNE 375

Dans cet article, on considère un système de deux équations booléennes linéaires. Grâce à des méthodes de théorie analytique des nombres, on montre que, sous certaines conditions, le système admet toujours des solutions. Cela complète le travail de Freiman sur ce sujet.

On series of discrete random variables, 1: real trinomial distributions with fixed probabilities
JEAN-MARC DESHOUILLERS, GREGORY A. FREIMAN & WILLIAM MORAN .. 411

Cet article démarre l'étude du comportement limite local d'un système triangulaire de variables aléatoires indépendantes $(\zeta_{n,k})_{1\le k\le n}$, où la loi de $\zeta_{n,k}$ dépend de n. Nous considédrons le cas où $\zeta_{n,1}$ prend trois valeurs entières $0 < a_1(n) < a_2(n)$ avec des probabilités respectives p_0, p_1, p_2 qui ne dépendent pas de n. Nous montrons qu'il y a trois types de comportement limite pour la suite des variables aléatoires $\eta_n = \zeta_{n,1} + \cdots + \zeta_{n,n}$, selon que $a_2(n)/\mathrm{pgcd}(a_1(n), a_2(n))$ tend vers l'infini plus lentement, plus vite ou à la même vitesse que $\sqrt{n}$.

On Bounds for the Concentration Function. 1
JEAN-MARC DESHOUILLERS, GREGORY A. FREIMAN & ALEXANDER A. YUDIN
.... 425

Nous donnons une majoration de la fonction de concentration d'une somme de variables aléatoires entières indépendantes et équidistribuées, en fonction d'une minoration de leur queue de distribution, sous l'hypothèse supplémentaire nécessaire que le support de ces variables aléatoires n'est pas essentiellement contenu dans une progression arithmétique non triviale.

ABSTRACTS

Structure Theory of Set Addition
Gregory A. Freiman 1

We review fundamental results in the so-called structure theory of set addition as well as their applications to other fields.

Sets of integers with large trigonometric sums
Amnon Besser 35

We investigate the problem of optimizing, for a fixed integer k and real u and on all sets $K = \{a_1 < a_2 < \cdots < a_k\} \subset \mathbb{Z}$, the measure of the set of $\alpha \in [0,1]$ where the absolute value of the trigonometric sum $S_K(\alpha) = \sum_{j=1}^{k} e^{2\pi i\alpha a_j}$ is greater than $k - u$. When u is sufficiently small with respect to k we are able to construct a set K_{ex} which is very close to optimal. This set is a union of a finite number of arithmetic progressions. We are able to show that any more optimal set, if one exists, has a similar structure to that of K_{ex}. We also get tight upper and lower bounds on the maximal measure.

Structure of sets with small sumset
Yuri Bilu 77

Freiman proved that a finite set of integers K satisfying $|K+K| \leq \sigma|K|$ is a subset of a "small" m-dimensional arithmetical progression, where $m \leq \lfloor \sigma - 1 \rfloor$. We give a complete self-contained exposition of this result, together with some refinements, and explicitly compute the constants involved.

On finite addition theorems
András Sárkőzy 109

If a finite set A of integers included in $\{1, \ldots, N\}$ has more than N/k elements, one may expect that the set ℓA of sums of ℓ elements of A, contains, when ℓ is comparable to k, a rather long arithmetic progression (which can be required to be homogeneous or not). After presenting the state of the art, we show that some of the results cannot be improved as far as it would be thought possible in view of the known results in the infinite case. The paper ends with lower and upper bounds for the order, as asymptotic bases, of the subsequences of the primes which have a positive relative density.

On Freiman's Theorems concerning the sum of two finite sets of integers
JOHN STEINIG 129

Details are provided for a proof of Freiman's theorems [1] which bound $|M + N|$ from below, where M and N are finite subsets of $\mathbb{Z}$.

On an additive problem of Erdős and Straus, 2
JEAN-MARC DESHOUILLERS & GREGORY A. FREIMAN 141

We denote by $s^\wedge A$ the set of integers which can be written as a sum of s pairwise distinct elements from A. The set A is called admissible if and only if $s \neq t$ implies that $s^\wedge A$ and $t^\wedge A$ have no element in common.

P. Erdős conjectured that an admissible set included in $[1, N]$ has a maximal cardinality when A consists of consecutive integers located at the upper end of the interval $[1, N]$. The object of this paper is to give a proof of Erdős' conjecture, for sufficiently large N.

On the structure of sum-free sets, 2
JEAN-MARC DESHOUILLERS, GREGORY A. FREIMAN, VERA SÓS & MIKHAIL TEMKIN 149

A finite set of positive integers is called sum-free if $\mathbb{A} \cap (\mathbb{A} + \mathbb{A})$ is empty, where $\mathbb{A} + \mathbb{A}$ denotes the set of sums of pairs of non necessarily distinct elements from $\mathbb{A}$. Improving upon a previous result by G.A. Freiman, a precise description of the structure of sum-free sets included in $[1, M]$ with cardinality larger than $0.4M - x$ for $M \geq M_0(x)$ (where x is an arbitrary given number) is given.

Sumsets with distinct summands and the Erdős-Heilbronn conjecture on sums of residues
GREGORY A. FREIMAN, LEWIS LOW & JANE PITMAN 163

Let S be a set of integers or of residue classes modulo a prime p, with cardinality $|S| = k$, and let T be the set of all sums of two distinct elements of S. For the integer case, it is shown that if $|T|$ is less than approximately $2.5k$ then S is contained in an arithmetic progression with relatively small cardinality. For the residue class case a result of this type is derived provided that $k > 60$ and $p > 50k$. As an application, it is shown that $|T| \geq 2k - 3$ under these conditions. Earlier results of Freiman play an essential role in the proofs.

On the number of sums and differences
FRANÇOIS HENNECART, GILLES ROBERT & ALEXANDER YUDIN 173

It is proved that $\inf_{A \subset \mathbb{Z}} \ln |A + A| / \ln |A - A|$ is less than .7865, improving a previous result due to G. Freiman and W. Pigarev.

We investigate multisets of natural numbers with relatively few subset sums. Namely, let A be a multiset such that the number of distinct subset sums of A is bounded by a fixed multiple of the cardinality of A (that is, $|P(A)| \ll |A|$). We show that the set $P(A)$ of subset sums is then a union of a small number of arithmetic progressions sharing a common difference.

Similar problems were considered by G. Freiman (see [1]) and M. Chaimovich (see [2]). Unlike those papers, our conditions are stated in terms of the cardinality of the subset sums set $P(A)$ only and not on the largest element of the original multiset A.

The result obtained is nearly best possible.

The number m is called the critical number of a finite abelian group G, if it is the minimal natural number with the property:
for every subset A of G with $|A| \geq m, 0 \notin A$, the set of subset sums A^* of A is equal to G. In this paper, we prove the conjecture of G. Diderrich about the value of the critical number of the group G, in the case $G = \mathbb{Z}_q$, for sufficiently large q.

Let $\epsilon > 0$. Erdős and Szemerédi conjectured that if A is a set of k positive integers which large k, there must be at least $k^{2-\varepsilon}$ integers that can be written as the sum or product of two elements of A. We shall prove this conjecture in the special case that the number of sums is very small.

A set $\mathcal{A}$ of integers is said "stratified" if, for all t, $0 \leq t < \operatorname{Card} \mathcal{A}$, the sum of any t distinct elements of $\mathcal{A}$ is smaller than the sum of any $t+1$ distinct elements of $\mathcal{A}$. That implies that all elements of $\mathcal{A}$ should be positive. It is proved that the number of stratified sets with maximal element equal to N is exactly the number $p(N)$ of partitions of N.

We describe the structure of sets of lattice points in the plane, having a small doubling property. Let $\mathbb{K}$ be a finite subset of $\mathbb{Z}^2$ such that

$$|\mathbb{K} + \mathbb{K}| < 3.5|\mathbb{K}| - 7.$$

If $\mathbb{K}$ lies on three parallel lines, then the convex hull of $\mathbb{K}$ is contained in three compatible arithmetic progressions with the same common difference, having together no more than

$$|\mathbb{K}| + \frac{3}{4}\Big(|\mathbb{K} + \mathbb{K}| - \frac{10}{3}|\mathbb{K}| + 5\Big)$$

terms. This upper bound is best possible.

Non-solvable groups with a large fraction of involutions
YAKOV BERKOVICH 241

In this note we classify the non-solvable finite groups G such that the class number of G is at least $|G|/16$. Some consequences are derived as well.

Questions on set squaring in groups
YAKOV BERKOVICH 249

Some questions on small subsets in groups are posed and discussed.

On groups generated by a pair of elements with small third or fourth power
SERGEI BRODSKY 255

The paper is devoted to an investigation of two-generated groups such that the $m-$th power of the generating pair contains less than 2^m elements . It is proved, in particular, that if the cube of the generating pair contains less than 7 elements or its fourth power contains less than 11 elements, then the group is solvable. Otherwise, it is not necessarily solvable. The proofs use computer calculations.

On small subset product in a group
YAHYA OULD HAMIDOUNE 281

We generalise some known addition theorems to non abelian groups and to the most general case of relations having a transitive group of automorphisms.

The classical proofs of addition theorems use local transformations due to Davenport, Dyson and Kempermann. We present a completely different method based on the study of some blocks of imprimitivity with respect to the automorphism group of a relation.

Several addition theorems including the finite $\alpha+\beta$-Theorem of Mann and a formula proved by Davenport and Lewis will be generalised to relations having a transitive group of automorphisms.

We study the critical pair theory in the case of finite groups. We generalise Vosper Theorem to finite not necessarily abelian groups.

Chowla, Mann and Straus obtained in 1959 a lower bound for the size of the image of a diagonal form on a prime field. This result was generalised by Tietäväienen to finite fields with odd characteristics. We use our results on the critical pair theory to generalise this lower bound to an arbitrary division ring.

Our results apply to the superconnectivity problems in networks. In particular we show that a loopless Cayley graph with optimal connectivity has only trivial minimum cuts when the degree and the order are coprime.

This paper presents results and open problems related to the following topics: group with deficient multiplication sub-tables, product bases in finite groups.

In this paper we investigate the structure of those pairs of finite subsets of an abelian group whose sums have relatively few elements: $|A+B| < |A|+|B|$. In 1960, J. H. B. Kemperman gave an exhaustive but rather sophisticated description of recursive nature. Using intermediate results of Kemperman, we obtain below a description of another type. Though not (generally speaking) sufficient, our description is intuitive and transparent and can be easily used in applications.

It is proved that in a commutative group G, where the order of elements is bounded by an integer r, any set A having n elements and at most αn sums is contained in a subgroup of size Cn with $C = f(r,\alpha)$ depending on r and α but not on n. This is an analog of a theorem of G. Freiman which describes the structure of such sets in the group of integers.

We study some additive problems in the group $(\mathbb{Z}/2\mathbb{Z})^r$. Our purpose is to show how those problems are closely related to coding theory. We present some relevant classical coding techniques and make use of them to obtain some original contributions.

The survey discusses a new approach to Integer Programming which is based on the structural characterization of problems using methods of additive number theory. This structural characterization allows one to design algorithms which are applicable in a narrower, yet still wide, domain of problems, and substantially improve the time boundary of existing algorithms. The new algorithms are polynomial for the class of problems in which they are applicable, and even linear ($O(m)$) for a wide class of the Subset-Sum and

Value-Independent Knapsack problems. Previously known polynomial time algorithms for the same classes of problems are at least two orders of magnitude slower.

New Algorithm for Dense Subset-Sum Problem
MARK CHAIMOVICH 363

A new algorithm for the dense subset-sum problem is derived by using the structural characterization of the set of subset-sums obtained by analytical methods of additive number theory. The algorithm works for a large number of summands (m) with values that are bounded from above. The boundary (ℓ) moderately depends on m. The new algorithm has $O(m^{7/4}/\log^{3/4} m)$ time boundary that is faster than the previously known algorithms the best of which yields $O(m^2/\log^2 m)$.

On the Two-Dimensional Subset Sum Problem
ALAIN PLAGNE 375

We consider a system of two linear boolean equations. Using methods from analytic number theory, we obtain sufficient conditions ensuring the solvability of the system. This completes Freiman's work on the subject.

On series of discrete random variables, 1: real trinomial distributions with fixed probabilities
JEAN-MARC DESHOUILLERS, GREGORY A. FREIMAN & WILLIAM MORAN .. 411

This paper begins the study of the local limit behaviour of triangular arrays of independent random variables $(\zeta_{n,k})_{1\leq k\leq n}$ where the law of $\zeta_{n,k}$ depends on on n. We consider the case when $\zeta_{n,1}$ takes three integral values $0 < a_1(n) < a_2(n)$ with respective probabilities p_0, p_1, p_2 which do not depend on n. We show three types of limit behaviours for the sequence of r. v. $\eta_n = \zeta_{n,1} + \cdots + \zeta_{n,n}$, according as $a_2(n)/\mathrm{gcd}(a_1(n), a_2(n))$ tends to infinity slower, quicker or at the same speed as $\sqrt{n}$.

On Bounds for the Concentration Function. 1
JEAN-MARC DESHOUILLERS, GREGORY A. FREIMAN & ALEXANDER A. YUDIN
.... 425

We give an upper bound for the concentration function of a sum of independent identically distributed integral valued random variables in terms of a lower bound for their tail, under the necessary extra condition that the random variables are not essentially supported in a proper arithmetic progression.