WHEN ALMOST ALL SETS ARE DIFFERENCE DOMINATED

PETER HEGARTY AND STEVEN J. MILLER

ABSTRACT. We investigate the relationship between the sizes of the sum and difference sets attached to a subset of $\{0, 1, ..., N\}$, chosen randomly according to a binomial model with parameter $p(N)$, with $N^{-1} = o(p(N))$. We show that the random subset is almost surely difference dominated, as $N \to \infty$, for any choice of $p(N)$ tending to zero, thus confirming a conjecture of Martin and O'Bryant. The proofs use recent strong concentration results.

Furthermore, we exhibit a threshold phenomenon regarding the ratio of the size of the difference- to the sumset. If $p(N) = o(N^{-1/2})$ then almost all sums and differences in the random subset are almost surely distinct, and in particular the difference set is almost surely about twice as large as the sumset. If $N^{-1/2} = o(p(N))$ then both the sum and difference sets almost surely have size $(2N + 1) - O(p(N)^{-2})$, and so the ratio in question is almost surely very close to one. If $p(N) = c \cdot N^{-1/2}$ then as c increases from zero to infinity (i.e., as the threshold is crossed), the same ratio almost surely decreases continuously from two to one according to an explicitly given function of c.

We also extend our results to the comparison of the generalized difference sets attached to an arbitrary pair of binary linear forms. For certain pairs of forms f and g, we show that there in fact exists a sharp threshold at $c_{f,g} \cdot N^{-1/2}$, for some computable constant $c_{f,g}$, such that one form almost surely dominates below the threshold, and the other almost surely above it.

The heart of our approach involves using different tools to obtain strong concentration of the sizes of the sum and difference sets about their mean values, for various ranges of the parameter p.

1. INTRODUCTION

To know whether a random variable is strongly concentrated is an issue of fundamental importance in many areas of mathematics and statistics. In this paper we apply recent results of Kim and Vu [KiVu, Vu1, Vu2] to completely solve a combinatorial number theory question on the size of difference- and sumsets of integers. A classical strong concentration result (due to Chernoff) states that if $Y = \sum_{i=1}^{n} t_i$ with the t_i i.i.d. binary random variables, then for any $\lambda > 0$ we have $\text{Prob}(|Y - \mathbb{E}[Y]| \geq \sqrt{\lambda n}) \leq 2e^{-\lambda/2}$. Within number theory, this result was used by Erdős (see [AS], Chapter 8) to prove the existence of so-called 'thin' bases of $\mathbb{N}$ of order 2. The general requirement for many applications is to obtain Chernoff-like exponential deviation bounds in situations when the atom variables t_i are not independent. For modern surveys of strong concentration

Date: October 9, 2008.

2000 *Mathematics Subject Classification.* 11P99 (primary), 11K99 (secondary).

Key words and phrases. Sum dominated sets, Binary linear forms, Strong concentration, Thresholds.

inequalities see, for example, [Ta] and [Vu2]; the latter, in particular, contains a fine selection of applications in random graph theory, combinatorial number theory and finite geometry.

The specific result we shall utilise is a martingale inequality which appears as Lemma 3.1 in [Vu2]. It is an extension of the classical Azuma inequality ([AS], Chapter 7) to functions whose Lipschitz coefficients are small 'on average'. As remarked in [Vu2], this type of inequality is very general and robust, and is expected to be applicable in numerous situations; this is definitely true for our problem.

Let S be a subset of the integers. We define the sumset $S+S$ and difference set $S-S$ by

$$\begin{aligned} S+S &= \{s_1+s_2 : s_i \in S\} \\ S-S &= \{s_1-s_2 : s_i \in S\}, \end{aligned} \tag{1.1}$$

and denote the cardinality of a set A by $|A|$. As addition is commutative and subtraction is not, a typical pair of integers generates two differences but only one sum. It is therefore reasonable to expect a generic finite set S will have a larger difference set than sumset. We say a set is *sum dominated* (such sets are also called *more sums than differences*, or MSTD, sets) if the cardinality of its sumset exceeds that of its difference set. If the two cardinalities are equal we say the set is *balanced*, otherwise *difference dominated*. Sum dominated sets exist: consider for example $\{0, 2, 3, 4, 7, 11, 12, 14\}$ (see [He, MS, Na2] for additional examples). In [Na1], Nathanson wrote *"Even though there exist sets A that have more sums than differences, such sets should be rare, and it must be true with the right way of counting that the vast majority of sets satisfies* $|A-A| > |A+A|$*."*

Recently Martin and O'Bryant [MO] showed there are many sum dominated sets. Specifically, let $I_N = \{0, \dots, N\}$. They prove the existence of a universal constant $\kappa_{\rm SD} > 0$ such that, for any $N \geq 14$, at least $\kappa_{\rm SD} \cdot 2^{N+1}$ subsets of I_N are sum dominated (there are no sum dominated sets in I_{13}). Their proof is based on choosing a subset of I_N by picking each $n \in I_N$ independently with probability $1/2$. The argument can be generalized to independently picking each $n \in I_N$ with any probability $p \in (0, 1)$, and yields the existence of a constant $\kappa_{{\rm SD},p} > 0$ such that, as $N \to \infty$, a randomly chosen (with

respect to this model) subset is sum dominated with probability at least $\kappa_{{\rm SD},p}$. Similarly one can prove there are positive constants $\kappa_{{\rm DD},p}$ and $\kappa_{{\rm B},p}$ for the probability of having a difference dominated or balanced set.

While the authors remark that, perhaps contrary to intuition, sum dominated sets are ubiquitous, their result is a consequence of how they choose a probability distribution on the space of subsets of I_N. Suppose $p = 1/2$, as in their paper. With high probability a randomly chosen subset will have $N/2$ elements (with errors of size $\sqrt{N}$). Thus the density of a generic subset to the underlying set I_N is quite high, typically about $1/2$. Because it is so high, when we look at the sumset (resp., difference set) of a typical A there are many ways of expressing elements as a sum (resp., difference) of two elements of A. For example (see [MO]), if $k \in A+A$ then there are roughly $N/4-|N-k|/4$ ways of writing k as a sum of two elements in A (similarly, if $k \in A-A$ there are roughly $N/4 - |k|/4$ ways of writing k as a difference of two elements of A). This enormous redundancy means almost all numbers which can be in the sumset or difference set are.

In fact, using uniform density on the subsets of I_N (i.e., taking $p = 1/2$), Martin and O'Bryant show that the average value of $|A+A|$ is $2N-11$ and that of $|A-A|$ is $2N-7$ (note each set has at most $2N-1$ elements). In particular, it is only for k near extremes that we have high probability of not having k in an $A+A$ or an $A-A$. In [MO] they prove a positive percentage of subsets of I_N (with respect to the uniform density) are sum dominated sets by specifying the fringe elements of A. Similar conclusions apply for any value of $p > 0$.

At the end of their paper, Martin and O'Bryant conjecture that if, on the other hand, the parameter p is a function of N tending to zero arbitrarily slowly, then as $N \to \infty$ the probability that a randomly chosen subset of I_N is sum dominated should also tend to zero. In this paper we will, among other things, prove this conjecture.

We shall find it convenient to adopt the following (fairly standard) shorthand notations. Let X be a real-valued random variable depending on some positive integer parameter N, and let $f(N)$ be some real-valued function. We write '$X \sim f(N)$' to denote the fact that, for any $\epsilon_1, \epsilon_2 > 0$, there exists $N_{\epsilon_1,\epsilon_2} > 0$ such that, for all $N > N_{\epsilon_1,\epsilon_2}$,

$$\mathbb{P}\left(X \notin [(1-\epsilon_1)f(N), (1+\epsilon_1)f(N)]\right) < \epsilon_2. \tag{1.2}$$

In particular we shall use this notation when X is just a function of N (hence not 'random'). In practice X will in this case be the expectation of some other random variable.

By $f(x) = O(g(x))$ we mean that there exist constants x_0 and C such that for all $x \geq x_0$, $|f(x)| \leq Cg(x)$. By $f(x) = \Theta(g(x))$ we mean that both $f(x) = O(g(x))$ and $g(x) = O(f(x))$ hold. Finally, if $\lim_{x\to\infty} f(x)/g(x) = 0$ then we write $f(x) = o(g(x))$.

Our main findings can be summed up in the following theorem.

Theorem 1.1. *Let $p : \mathbb{N} \to (0,1)$ be any function such that*

$$N^{-1} = o(p(N)) \quad \text{and} \quad p(N) = o(1). \tag{1.3}$$

For each $N \in \mathbb{N}$ let A be a random subset of I_N chosen according to a binomial distribution with parameter $p(N)$. Then, as $N \to \infty$, the probability that A is difference dominated tends to one.

More precisely, let $\mathscr{S}, \mathscr{D}$ denote respectively the random variables $|A+A|$ and $|A-A|$. Then the following three situations arise :

(i) $p(N) = o(N^{-1/2})$: Then

$$\mathscr{S} \sim \frac{(N \cdot p(N))^2}{2} \quad \text{and} \quad \mathscr{D} \sim 2\mathscr{S} \sim (N \cdot p(N))^2. \tag{1.4}$$

(ii) $p(N) = c \cdot N^{-1/2}$ for some $c \in (0,\infty)$: Define the function $g : (0,\infty) \to (0,2)$ by

$$g(x) := 2\left(\frac{e^{-x} - (1-x)}{x}\right). \tag{1.5}$$

Then

$$\mathscr{S} \sim g\left(\frac{c^2}{2}\right) N \quad \text{and} \quad \mathscr{D} \sim g(c^2)N. \tag{1.6}$$

(iii) $N^{-1/2} = o(p(N))$ *: Let* $\mathscr{S}^c := (2N+1) - \mathscr{S}$, $\mathscr{D}^c := (2N+1) - \mathscr{D}$. *Then*

$$\mathscr{S}^c \sim 2 \cdot \mathscr{D}^c \sim \frac{4}{p(N)^2}. \tag{1.7}$$

Remark 1.2. Obviously, not all functions $p : \mathbb{N} \to (0,1)$ satisfying (1.3) conform to the requirements of (i), (ii) or (iii) above, but these are the natural functions to investigate in the current context. Similar remarks apply to Theorem 3.1 and Conjecture 4.2 below.

Theorem 1.1 proves the conjecture in [MO] and re-establishes the validity of Nathanson's claim in a broad setting. It also identifies the function $N^{-1/2}$ as a *threshold function*, in the sense of [JŁR], for the ratio of the size of the difference- to the sumset for a random set $A \subseteq I_N$. Below the threshold, this ratio is almost surely $2 + o(1)$, above it almost surely $1 + o(1)$. Part (ii) tells us that the ratio decreases continuously (a.s.) as the threshold is crossed. Below the threshold, part (i) says that most sets are 'nearly Sidon sets', that is, most pairs of elements generate distinct sums and differences. Above the threshold, most numbers which can be in the sumset (resp., difference set) usually are, and in fact most of these in turn have many different representations as a sum (resp., a difference). However the sumset is usually missing about twice as many elements as the difference set. Thus if we replace 'sums' (resp., 'differences') by 'missing sums' (resp., 'missing differences'), then there is still a symmetry between what happens on both sides of the threshold.

We prove Theorem 1.1 in the next section. Our strategy will consist of first establishing an estimate for the expectation of the random variable $\mathscr{S}$ or $\mathscr{D}$, followed by establishing sufficiently strong concentration of these variables about their mean values. For the second part of this strategy we will use different approaches for $p = O(N^{-1/2})$ and $N^{-1/2} = o(p(N))$. In the former range a fairly straightforward second moment argument works. In the latter range, however, we will employ a specialization of the Kim-Vu martingale lemma (Lemma 3.1 in [Vu2], Lemma 2.2 below).

In Section 3, we extend our result to arbitrary binary linear forms. The paper [NOORS] provides motivation for studying these objects. By a *binary linear form* we mean a function $f(x,y) = ux + vy$ where $u, v \in \mathbb{Z}_{\neq 0}$, $u \geq |v|$ and $\text{GCD}(u,v) = 1$. For a set A of integers we let

$$f(A) := \{ua_1 + va_2 : a_1, a_2 \in A\}. \tag{1.8}$$

Except in the special case $u = v = 1$ we always have that $f(x,y) \neq f(y,x)$ whenever $x \neq y$. Thus we refer to f as a *difference form* and a set $f(A)$ as a *generalized difference set*, whenever $u > |v|$. Theorem 3.1 allows us to compare the sizes of $f(A)$ and $g(A)$ for random sets A, and arbitrary difference forms f and g when $N^{-3/5} = o(p(N))$.

Two situations arise :

(a) for some pairs of forms, the same one a.s. dominates the other for all parameters $p = p(N)$ in this range. In fact every other difference form dominates $x - y$, and hence also $x + y$.

(b) for certain pairs f and g, something very nice happens. Namely, there is now a *sharp* *threshold*, in the sense of [JŁR], at $c_{f,g}N^{-1/2}$, for some computable constant $c_{f,g} > 0$, depending on f and g. One form dominates a.s. below the threshold, and the other one a.s. above it. This fact may be considered a partial generalization of the main result of [NOORS] to random sets, partial in the sense that is only applies to certain pairs of forms. Namely, they proved that for any two forms f and g (including $x + y$), there exist finite sets A_1 and A_2 such that $|f(A_1)| > |g(A_1)|$ whereas $|f(A_2)| < |g(A_2)|$.

We leave it to future work to investigate what happens as the threshold is crossed in this situation.

In Section 4 we make a brief summary of this and other remaining questions and make suggestions for other problems to study. In particular, we suggest looking at other probabilistic models for choosing random sets. This is partly motivated by the fact that our results in Section 3 only apply when $N^{-3/5} = o(p(N))$. The reason is that, for faster decaying $p(N)$, as we shall see, the variance in the size of the random set A itself swamps all other error terms, and it is meaningless to compare $|f(A)|$ and $|g(A)|$; in other words, the model itself becomes useless. This may be considered a problem when $N^{-3/4} = o(p(N))$. For $p(N) = o(N^{-3/4})$, the results of [GJLR] imply that all pairs (x, y) in a random set a.s. generate different values $f(x, y)$, for any f, so that $|f(A)| = |g(A)|$ a.s. for any f and g.

2. Proof of Theorem 1.1

Our strategy for proving the various assertions in Theorem 1.1 is the following. Let $\mathscr{X}$ be one of the random variables $\mathscr{S}, \mathscr{D}, \mathscr{S}^c, \mathscr{D}^c$, as appropriate. We then carry out the following two steps :

Step 1 : Prove that $\mathbb{E}[\mathscr{X}]$ behaves asymptotically as asserted in the theorem.
Step 2 : Prove that $\mathscr{X}$ is strongly concentrated about its mean.

As already mentioned, the calculations required to perform these two steps differ according as to whether $p(N) = O(N^{-1/2})$ or $N^{-1/2} = o(p(N))$. In particular, in the former case, *Step 2* is achieved by a fairly straightforward second moment argument, whereas a more sophisticated concentration inequality is used in the latter case. We thus divide the proof of the theorem into two separate cases, depending on the parameter function p.

Throughout the paper we often abuse notation to save space, writing p for $p(N)$. As we never consider the case where $p(N)$ is constant (as this case has been analyzed in [MO]), this should not cause any confusion.

Case I : $p(N) = O(N^{-1/2})$.

We first concentrate on the sumset and prove the various assertions in parts (i) and (ii) of the theorem. The proofs for the difference set will be similar. For any finite set $A \subseteq \mathbb{N}_0$ and any integer $k \geq 1$, let

$$A_k := \{\{\{a_1, a_2\}, \dots, \{a_{2k-1}, a_{2k}\}\} : a_1 + a_2 = \cdots = a_{2k-1} + a_{2k}\}. \tag{2.1}$$

In words, A_k consists of all unordered k-tuples of unordered pairs of elements of A having the same sum. Let $X_k := |A_k|$. So if A is a random set, then each X_k is a non-negative integer valued random variable. The crucial observation for our work is that, in the model we are considering, the random variables X_k are all highly concentrated :

Lemma 2.1. *For $p(N) = O(N^{-1/2})$ we have for every k that*

$$\mathbb{E}[X_k] \sim \frac{2}{(k+1)!}\left(\frac{p(N)^2}{2}\right)^k N^{k+1} \tag{2.2}$$

and, more significantly, $X_k \sim \mathbb{E}[X_k]$ whenever $N^{-\left(\frac{k+1}{2k}\right)} = o(p(N))$.

Proof. We write p for $p(N)$. By the central limit theorem it is clear that

$$|A| \sim Np. \tag{2.3}$$

Each X_k can be written as a sum of indicator variables Y_α, one for each unordered k-tuple α. There are two types of k-tuples : those consisting of $2k$ distinct elements of I_N and those in which one element is repeated twice in one of the k pairs, and the sum of each of the k pairs is even. The probability of any k-tuple of the former type occurring in A_k is p^{2k}, whereas for k-tuples of the latter type this probability is p^{2k-1}. Let there be a total of $\xi_{1,k}(N)$ k-tuples of the former type and $\xi_{2,k}(N)$ of the latter type. Then, by linearity of expectation,

$$\mathbb{E}[X_k] = \xi_{1,k}(N)\cdot p^{2k} + \xi_{2,k}(N)\cdot p^{2k-1}. \tag{2.4}$$

We have

$$\xi_{1,k}(N) = \sum_{n=0}^{2N}\binom{R(n)}{k}, \tag{2.5}$$

where $R(n)$ is the number of representations of n as a sum of two distinct elements of I_N, and hence we easily estimate

$$\begin{aligned}\xi_{1,k}(N) &= \sum_{n=2k}^{2N-2k}\binom{\min\{\lfloor\frac{n}{2}\rfloor, \lfloor\frac{2N-n}{2}\rfloor\}}{k} \sim 2\cdot\sum_{n=2k}^{N}\binom{\lfloor\frac{n}{2}\rfloor}{k}\\ &\sim 2\cdot 2\cdot\sum_{n=1}^{\lfloor N/2\rfloor}\binom{n}{k} \sim 4\binom{\lceil N/2\rceil}{k+1} \sim 4\frac{(N/2)^{k+1}}{(k+1)!}.\end{aligned} \tag{2.6}$$

Thus

$$\xi_{1,k}(N)\cdot p^{2k} \sim \frac{2}{(k+1)!}\left(\frac{p^2}{2}\right)^k N^{k+1}. \tag{2.7}$$

A similar calculation shows that $\xi_{2,k}(N) = O_k(N^k)$, hence $\xi_{2,k}(N)\cdot p^{2k-1} = O_k(N^k p^{2k-1})$. Since $N^{-1} = o(p)$ it follows that

$$\mathbb{E}[X_k] \sim \frac{2}{(k+1)!}\left(\frac{p^2}{2}\right)^k N^{k+1}, \tag{2.8}$$

in accordance with the lemma. To complete the proof of the lemma, we need to show that, whenever $N^{-\left(\frac{k+1}{2k}\right)} = o(p)$, the random variable X_k becomes highly concentrated about its mean as $N \to \infty$. We apply a standard second moment method. In the notation of [AS], Chapter 4, since we already know in this case that $\mathbb{E}[X_k] \to \infty$, it suffices to show that $\Delta = o(\mathbb{E}[X_k]^2) = o_k(N^{2k+2}p^{4k})$, where

$$\Delta \; = \; \sum_{\alpha \sim \beta} \mathbb{P}(Y_\alpha \wedge Y_\beta), \tag{2.9}$$

the sum being over pairs of k-tuples which have at least one number in common. It is easy to see that, since $N^{-1} = o(p)$, the main contribution to Δ comes from pairs $\{\alpha, \beta\}$ of k-tuples, each of which consist of $2k$ distinct elements of I_N, and which have exactly one element in common. The number of such pairs is $O_k(N^{2k+1})$ since there are

- $O_k(N^{k+1})$ choices for α,
- $2k$ choices for the common element with β,
- $O(N)$ choices for the sum of each pair in β,
- $O_k(N^{k-1})$ choices for the remaining elements in β.

Since a total of $4k-1$ elements of I_N occur in total in $\alpha \cup \beta$, we have $\mathbb{P}(Y_\alpha \wedge Y_\beta) = p^{4k-1}$. Thus

$$\Delta \; = \; O_{c,k}(N^{2k+1}p^{4k-1}) \; = \; o_{c,k}(N^{2k+2}p^{4k}), \quad \text{since } N^{-1} = o(p). \tag{2.10}$$

This completes the proof of Lemma 2.1. □

We can now prove parts (i) and (ii) of the theorem. First suppose $p = o(N^{-1/2})$. By (2.2) we have $X_1 \sim \frac{1}{2}N^2p^2$, whereas $X_2 \sim \frac{1}{12}N^3p^4$ for $N^{-3/4} = o(p)$ and $\mathbb{E}[X_2] = O(1)$ otherwise.

Since $p = o(N^{-1/2})$ we have $\max(1, N^3p^4) = o(N^2p^2)$ and thus $X_2 = o(X_1)$ almost surely. In other words, as $N \to \infty$, all but a vanishing proportion of pairs of element of A will have distinct sums. It follows immediately that

$$\mathscr{S} \; \sim \; X_1 \; \sim \; \frac{(Np)^2}{2}, \tag{2.11}$$

as claimed.

Now suppose $p = cN^{-1/2}$ for some fixed $c > 0$. This time we will need to consider all the X_k together. Let $\mathscr{P}$ be the partition on A_1 whereby $\{a_1, a_2\}$ and $\{a_3, a_4\}$ are in the same part if and only if $a_1 + a_2 = a_3 + a_4$. For each $i > 0$ let τ_i denote the number of parts of size i (as a random variable). Observe that

$$\mathscr{S} \; = \; \sum_{i=0}^{\infty} \tau_i \tag{2.12}$$

and, for each $k \geq 1$, that

$$\sum_{i=1}^{\infty} \binom{i}{k} \tau_i \; = \; X_k. \tag{2.13}$$

(2.13) is a system of infinitely many equations in the variables τ_i, which together determine $\mathscr{S}$. For any $m \geq 1$, the general solution of the subsystem formed by the first m equations (i.e.: $k = 1, \dots, m$) is readily checked to be

$$\mathscr{S} \; = \; \sum_{k=1}^{m} (-1)^{k-1} X_k + \sum_{k=m+1}^{\infty} \left\{ \sum_{i=0}^{m} (-1)^i \binom{k}{i} \right\} \tau_k. \tag{2.14}$$

Regarding the second sum on the right of (2.14) we have that

$$\left|\sum_{k=m+1}^{\infty}\left\{\sum_{i=0}^{m}(-1)^i\binom{k}{i}\right\}\tau_k\right| \;\le\; \sum_{k=m+1}^{\infty}\binom{k}{m}\tau_k \;=\; X_m-\tau_m\le X_m. \tag{2.15}$$

Hence it follows that, for any $m\ge 1$,

$$\left|X-\sum_{k=1}^{m}(-1)^{k-1}X_k\right| \;\le\; X_m. \tag{2.16}$$

Now for $p=cN^{-1/2}$, Lemma 2.1 says that

$$X_m \;\sim\; 2\frac{\left(\frac{c^2}{2}\right)^m}{(m+1)!}N, \tag{2.17}$$

and since $\frac{\left(c^2/2\right)^m}{(m+1)!}\to 0$ as $m\to\infty$, another application of Lemma 2.1 implies that

$$\mathscr{S} \;\sim\; \sum_{k=1}^{\infty}(-1)^{k-1}X_k \;\sim\; 2\cdot\left(\sum_{k=1}^{\infty}\frac{(-1)^{k-1}\left(\frac{c^2}{2}\right)^k}{(k+1)!}\right)\cdot N. \tag{2.18}$$

So to prove (1.6) it just remains to verify that

$$g(x) \;=\; 2\cdot\sum_{k=1}^{\infty}\frac{(-1)^{k-1}x^k}{(k+1)!}, \tag{2.19}$$

which is an easy exercise.

This proves parts (i) and (ii) of Theorem 1.1 for the sumset. For the difference set one reasons in an entirely parallel manner. One now defines, for each $k\ge 1$,

$$A_k' := \{\{(a_1,a_2),\dots,(a_{2k-1},a_{2k})\} : a_1-a_2=\cdots=a_{2k-1}-a_{2k}\neq 0\}. \tag{2.20}$$

In words, A_k' consists of all k-tuples of ordered pairs of elements of A which have the same non-zero difference. We let $X_k' := |A_k|$ and in a completely analogous manner to Lemma 2.1 prove that

$$\mathbb{E}[X_k'] \;\sim\; \frac{2}{(k+1)!}p^{2k}N^{k+1}, \tag{2.21}$$

and that $X_k'\sim\mathbb{E}[X_k']$ whenever $N^{-\left(\frac{k+1}{2k}\right)}=o(p)$. We define the partition $\mathscr{P}'$ of A_1' in the obvious way and let τ_i' denote the number of parts of size i, for each $i\ge 1$. Since $\mathscr{D}=1+\sum_{i=1}^{\infty}\tau_i$ we can follow exactly the same analysis as above to deduce (1.4) and (1.6). This completes the proofs of parts (i) and (ii) of the theorem.

Case II : $N^{-1/2}=o(p(N))$.

Recall $p(N)=o(1)$. Set $p=p(N)$ and $P=1/p$; thus $P=o(N^{1/2})$ (as $p=o(1)$ we have $\lim_{N\to\infty}P=\infty$). Again we begin with the sumset. Recall the two steps to be accomplished :

Step 1 : We prove that $\mathbb{E}[\mathscr{S}^c]\sim 4P^2$.

Step 2 : We prove that the random variable $\mathscr{S}^c$ is strongly concentrated about its mean.

We begin with the simpler *Step 1*. For each $n \in I_{2N}$, let $\mathscr{E}_n$ denote the event that $n \notin A + A$. Thus

$$\mathbb{E}[\mathscr{S}^c] = \sum_{n=0}^{2N} \mathbb{P}(\mathscr{E}_n). \tag{2.22}$$

Observe that $\mathbb{P}(\mathscr{E}_n) = \mathbb{P}(\mathscr{E}_{2N-n})$. Since all the ways of representing any given n as a sum of two elements of I_N are independent of one another, we have, for $n \in I_N$,

$$\mathbb{P}(\mathscr{E}_n) \ = \ \begin{cases} (1-p^2)^{n/2}(1-p) & \text{if } n \text{ is even} \\ (1-p^2)^{(n+1)/2} & \text{if } n \text{ is odd.} \end{cases} \tag{2.23}$$

Since $p = o(1)$ we have $1 - p \sim 1$, and since $N^{-1/2} = o(p)$ we have $(1-p^2)^N = o(1)$. Thus it is easy to see that

$$\mathbb{E}[\mathscr{S}^c] \ \sim \ 4 \cdot \sum_{m=0}^{\lfloor N/2 \rfloor} (1-p^2)^m \ \sim \ \frac{4}{p^2} \ = \ 4P^2, \tag{2.24}$$

as claimed. This completes *Step 1*.

For *Step 2* we need the martingale machinery of Kim and Vu.

We use notation consistent with [Vu2]. Consider a fixed N, which shall tend to infinity in our estimates. Let $\Omega := \{0,1\}^{N+1}$. Thus every subset A of I_N can be identified with an element of Ω. We are working in the probability space (Ω, μ) where μ is the product measure with parameter p. For each $A \in \Omega$, $n \in I_N$ and $x \in \{0,1\}$, define

$$C_n(x, A) \ := \ \left|\mathbb{E}\left[\mathscr{S}^c | a_0, ..., a_{n-1}, a_n = x\right] - \mathbb{E}\left[\mathscr{S}^c | a_0, ..., a_{n-1}\right]\right|; \tag{2.25}$$

by $\mathbb{E}\left[\mathscr{S}^c | a_0, ..., a_m\right]$ we mean the expected value of the random variable $\mathscr{S}^c$, given that for $k \in \{0, \dots, m\}$ the element k is always (resp., never) in the subset if $a_k = 1$ (resp., $a_k = -1$). Let

$$C(A) \ := \ \max_{n,x} C_n(x, A). \tag{2.26}$$

Further put

$$V_n(A) \ := \ \int_0^1 C_n^2(x, A) d^n\mu = pC_n^2(0, A) + (1-p)C_n^2(1, A) \tag{2.27}$$

and

$$V(A) \ := \ \sum_{n=0}^{N} V_n(A). \tag{2.28}$$

For two arbitrary positive numbers $\mathbf{V}$ and $\mathbf{C}$, define the event

$$\mathbb{B}_{\mathbf{V},\mathbf{C}} \ := \ \{A : C(A) \geq \mathbf{C} \text{ or } V(A) \ \geq \ \mathbf{V}\}. \tag{2.29}$$

Then the following is a specialization of a result appearing in [Vu2] :

Lemma 2.2. *For any positive numbers $\lambda, \mathbf{V}, \mathbf{C}$ such that $\lambda \leq 4\mathbf{V}/\mathbf{C}^2$ we have*

$$\mathbb{P}\left(|\mathscr{S}^c - \mathbb{E}[\mathscr{S}^c]| \geq \sqrt{\lambda \mathbf{V}}\right) \ \leq \ 2e^{-\lambda/4} + \mathbb{P}(\mathbb{B}_{\mathbf{V},\mathbf{C}}). \tag{2.30}$$

We quickly sketch how Lemma 2.2 completes the proof of assertion (iii) of Theorem 1.1. We shall take

$$\lambda := \kappa_0 \log P, \quad \mathbf{V} := \kappa_1 (P \log P)^3, \quad \mathbf{C} := \kappa_2 P \log P. \tag{2.31}$$

We show that for appropriately chosen κ_1, κ_2 we have

$$\mathbb{P}(\mathbb{B}_{\mathbf{V},\mathbf{C}}) \ = \ o(1). \tag{2.32}$$

From (2.30) and $\lim_{N\to\infty} P = \infty$, for sufficiently small κ_0 we will then be able to conclude that

$$|\mathscr{S}^c - \mathbb{E}[\mathscr{S}^c]| \ = \ O(P^{3/2}\log^2 P) \ \text{ a.s. as } N \to \infty. \tag{2.33}$$

As $\mathbb{E}[\mathscr{S}^c] \sim 4P^2$ (see (2.24)), assertion (iii) in Theorem 1.1 follows immediately. Thus we are reduced to proving (2.32), which we now proceed to do.

First we simplify things a little. For any $n \in I_N$ and $A \in \Omega$, we introduce the shorthand

$$\mathscr{U}_{n,A} \ := \ \mathscr{S}^c | a_0, ..., a_{n-1}. \tag{2.34}$$

Let

$$\Delta_n(A) \ := \ \mathbb{E}[\mathscr{U}_{n,A} | a_n = 0] - \mathbb{E}[\mathscr{U}_{n,A} | a_n = 1]. \tag{2.35}$$

As $\mathbb{E}[\mathscr{U}_{n,A} | a_n = 0] \geq \mathbb{E}[\mathscr{U}_{n,A} | a_n = 1]$, we see $\Delta_n(A) \geq 0$. For $x \in \{0,1\}$,

$$C_n(x, A) \ = \ |\mathbb{E}[\mathscr{U}_{n,A} | a_n = x] - \mathbb{E}[\mathscr{U}_{n,A}]|. \tag{2.36}$$

Since

$$\begin{aligned} \mathbb{E}[\mathscr{U}_{n,A}] \ &= \ p\mathbb{E}[\mathscr{U}_{n,A} | a_n = 1] + (1-p)\mathbb{E}[\mathscr{U}_{n,A} | a_n = 0] \\ &= \ \mathbb{E}[\mathscr{U}_{n,A} | a_n = 1] + (1-p)\Delta_n(A), \end{aligned} \tag{2.37}$$

$C_n(x, A)$ from (2.25) simplifies to

$$C_n(1, A) \ = \ (1-p)\Delta_n(A), \quad C_n(0, A) \ = \ p\Delta_n(A). \tag{2.38}$$

Since $p < 1 - p$ for all sufficiently large N, we have then

$$C(A) = (1-p) \max_{0 \leq n \leq N} \Delta_n(A). \tag{2.39}$$

Further, (2.27), (2.28) and (2.38) yield

$$V(A) = p(1-p)\sum_{n=0}^{N} \Delta_n^2(A). \tag{2.40}$$

This completes our simplifications.

Recall that in order to use (2.30) from Lemma 2.2 we need to prove (2.32) (namely that $\mathbb{P}(\mathbb{B}_{\mathbf{V},\mathbf{C}}) = o(1)$). The heart of the proof of (2.32) is to show that for an appropriate choice of $\kappa_1, \kappa_2, \kappa_3 > 0$, with probability $1 - o(1)$ all three of the following events occur:

$$\sum_{|N-n| > \kappa_3 P^2 \log P} \Delta_n(A) = o(1), \tag{2.41}$$

$$\max_{0\leq|N-n|\leq\kappa_3P^2\log P}\Delta_n(A)\leq\kappa_2P\log P, \tag{2.42}$$

$$V(A)\leq\kappa_1(P\log P)^3. \tag{2.43}$$

We claim that (2.41)-(2.43) imply (2.32). This follows immediately from applying the trivial bound $\mathbb{P}(\mathbb{B}_{\mathbf{V},\mathbf{C}})\leq\mathbb{P}(V\geq\mathbf{V})+\mathbb{P}(C\geq\mathbf{C})$ and using (2.41)-(2.43) to show these two probabilities are both $o(1)$.

To summarize, the proof is completed by verifying (2.41)-(2.43). Notice also that (2.41) and (2.42), together with (2.40), imply (2.43) for any choice of $\kappa_1>\kappa_2^2\kappa_3$, so it just remains to prove the former two. As in the arguments that follow there is a symmetry between n and $N-n$, we consider n with $0\leq n\leq N/2$; the remaining n follow similarly.

First, consider (2.41). Note that, depending on the parameter P, this sum could be empty. This will not affect the argument to follow. The proof is by an averaging argument, i.e.: for each n we first consider $\mathbb{E}_{a_0,\dots,a_{n-1}}[\Delta_n(A)]$. This quantity has a very natural interpretation : in words, it is the expected increase in the size of the sumset $A+A$ brought about by the addition of the number n to A. For every $z\in\{n,...,n+N\}$, adding n to A will add z to $A+A$ if and only if $z-n\in A$ and, for any other numbers n_1,n_2 such that $n_1+n_2=z$, either $n_1\notin A$ or $n_2\notin A$. Let $\mathscr{E}_z$ be the event that z gets added to $A+A$ by the addition of n to A. Then using (2.23) we can explicitly estimate

$$\mathbb{E}_{a_0,\dots,a_{n-1}}[\Delta_n(A)]\;=\;\sum_{z=n}^{n+N}\mathbb{P}[\mathscr{E}_z]\;\sim\;p\sum_{z=n}^{n+N}(1-p^2)^{\min\{\lfloor\frac{z}{2}\rfloor,\lfloor\frac{2N-z}{2}\rfloor\}}. \tag{2.44}$$

Since $n\leq N/2$, the last sum is asympotically no more than

$$2p\sum_{r=\lfloor n/2\rfloor}^{\lfloor N/2\rfloor}(1-p^2)^r\;=\;(2+o(1))P(1-p^2)^{n/2}. \tag{2.45}$$

By Markov's inequality, we deduce that for any n,

$$\mathbb{P}(\Delta_n(A)\geq2P(1-p^2)^{n/4})\;\leq\;(1+o(1))(1-p^2)^{n/4}. \tag{2.46}$$

Then, just using a trivial union bound

$$\mathbb{P}\left(\bigvee\mathscr{E}_n\right)\;\leq\;\sum\mathbb{P}(\mathscr{E}_n), \tag{2.47}$$

it follows that, with probability at least

$$1-(1+o(1))\cdot\sum_{n=\lceil\kappa_3P^2\log P\rceil}^{\lfloor N/2\rfloor}(1-p^2)^{n/4}, \tag{2.48}$$

we have

$$\Delta_n(A)\;\leq\;2P(1-p^2)^{n/4}\;\text{ for all } n \text{ such that } \kappa_3P^2\log P\leq n\leq N/2. \tag{2.49}$$

Then (2.41) clearly follows provided

$$P\cdot\sum_{n=\lceil\kappa_3P^2\log P\rceil}^{\lfloor N/2\rfloor}(1-p^2)^{n/4}\;=\;o(1), \tag{2.50}$$

which is clearly the case for sufficiently large κ_3, since $1 = o(P)$.

We now turn to (2.42). Firstly, a similar argument to the one just given shows that, even if $n \leq \kappa_3 P^2 \log P$, adding n to a random set A is very probably not going to add any elements at all to $A + A$ which are larger than $\kappa_3 P^2 \log P$. Secondly, among the numbers in $I_{\kappa_3 P^2 \log P}$, the addition to A of one number cannot add to $A + A$ more numbers than were in A already, plus maybe one more. But Chernoff's inequality ([AS], Corollary A.14) implies that, with probability $1 - e^{-c_1 \kappa_3 P \log P}$, where c_1 is some universal positive constant, $|A \cap I_{\kappa_3 P^2 \log P}| \leq 2\kappa_3 P \log P$. Then (2.42) follows from a simple union bound, as long as $\kappa_2 > 2\kappa_3$ for example.

This completes the proof of the assertion of Theorem 1.1(iii) as regards the sumset.

For the difference set, we proceed in two identical steps. First consider the estimate of $\mathbb{E}[\mathscr{D}^c]$. Let $\mathscr{E}_n$ now denote instead the event that $n \notin A - A$ for each $n \in \pm I_N$. Clearly,

$$\mathbb{E}[\mathscr{D}^c] \;=\; 2 \cdot \sum_{n=1}^{N} \mathbb{P}(\mathscr{E}_n) + o(1). \tag{2.51}$$

For each $n > 0$ we have

$$\mathscr{E}_n \;=\; \bigwedge_{m=0}^{N-n} \overline{\mathscr{B}}_{m,n}, \tag{2.52}$$

where $\mathscr{B}_{m,n}$ is the (bad) event that both m and $m + n$ lie in A and $\overline{\mathscr{B}}_{m,n}$ is the complementary event. These events are not independent, but the dependencies will not affect our estimates. To see this rigorously, one can for example use Janson's inequality (see [AS], Chapter 8, though this is certainly overkill!)

$$M \;\leq\; \mathbb{P}\left(\bigwedge_m \overline{\mathscr{B}}_{m,n}\right) \;\leq\; M \exp\left(\frac{\Delta}{1-\epsilon}\right), \tag{2.53}$$

where all $\mathbb{P}(\mathscr{B}_{m,n}) \leq \epsilon$, $M = \prod_m \mathbb{P}(\overline{\mathscr{B}}_{m,n})$ and

$$\Delta = \sum_{m \sim m'} \mathbb{P}(\mathscr{B}_{m,n} \wedge \mathscr{B}_{m',n}), \tag{2.54}$$

the sum being over dependent pairs $\{m, m'\}$, i.e.: pairs such that $m' = m + n$.

Note that we can take $\epsilon = p^2$, we have $M = (1 - p^2)^{N-n+1}$ and

$$\Delta \;=\; \begin{cases} (N - 2n + 1)p^3 & \text{if } n \leq N/2 \\ 0 & \text{if } n > N/2, \end{cases} \tag{2.55}$$

since there is a 1-1 correspondence between dependent pairs and 3-term arithmetic progressions in I_N of common difference n. It is then easy to see that this correction term can be ignored when we make the estimate

$$\mathbb{E}[\mathscr{D}^c] \;\sim\; 2 \cdot \sum_{n=1}^{N} (1 - p^2)^{N-n+1} \;\sim\; \frac{2}{p^2} \;\sim\; \frac{1}{2}\mathbb{E}[\mathscr{S}^c], \tag{2.56}$$

as desired.

The concentration of $\mathscr{D}^c$ about its mean can be established in the same way as we did with $\mathscr{S}^c$ above. A little more care is required in estimating quantities analogous to $\mathbb{E}_{a_0,\dots,a_{n-1}}[\Delta_n(A)]$, because of the dependencies between different representations of the same difference, but Janson's inequality can again be used to see rigorously that this will not affect our estimates. We omit further details and simply note that we will again obtain the result that

$$|\mathscr{D}^c - \mathbb{E}[\mathscr{D}^c]| \;=\; O(P^{3/2}\log^2 P) \;\text{ a.s. as } N \to \infty. \tag{2.57}$$

This completes the proof of Theorem 1.1. □

3. General Binary Linear Forms

We have the following generalization of Theorem 1.1 :

Theorem 3.1. *Let* $p : \mathbb{N} \to (0,1)$ *be a function satisfying* (1.3). *Let* u, v *be non-zero integers with* $u \geq |v|$, *GCD*$(u,v) = 1$ *and* $(u,v) \neq (1,1)$. *Put* $f(x,y) := ux + vy$. *For a positive integer* N, *let* A *be a random subset of* I_N *obtained by choosing each* $n \in I_N$ *independently with probability* $p(N)$. *Let* $\mathscr{D}_f$ *denote the random variable* $|f(A)|$. *Then the following three situations arise :*

(i) $p(N) = o(N^{-1/2})$ *: Then*

$$\mathscr{D}_f \sim (N \cdot p(N))^2. \tag{3.1}$$

(ii) $p(N) = c \cdot N^{-1/2}$ *for some* $c \in (0,\infty)$ *: Define the function* $g_{u,v} : (0,\infty) \to (0, u+|v|)$ *by*

$$g_{u,v}(x) \;:=\; (u+|v|) - 2|v|\left(\frac{1-e^{-x}}{x}\right) - (u-|v|)e^{-x}. \tag{3.2}$$

Then

$$\mathscr{D}_f \;\sim\; g_{u,v}\left(\frac{c^2}{u}\right) N. \tag{3.3}$$

(iii) $N^{-1/2} = o(p(N))$ *: Let* $\mathscr{D}_f^c := (u+|v|)N - \mathscr{D}_f$. *Then*

$$\mathscr{D}_f^c \;\sim\; \frac{2u|v|}{p(N)^2}. \tag{3.4}$$

Proof. One follows exactly the method of proof of Theorem 1.1, so we only give a sketch here.

Case I : $p(N) = O(N^{-1/2})$.

We again write p for $p(N)$. For any finite set $A \subseteq \mathbb{N}_0$ and any integer $k \geq 1$, let

$$A'_{k,f} \;:=\; \{\{(a_1,a_2),\dots,(a_{2k-1},a_{2k})\} : f(a_1,a_2) = \cdots = f(a_{2k-1},a_{2k})\}. \tag{3.5}$$

Let $X'_{k,f} := |A_{k,f}|$. Then (2.21) has the following generalization :

$$\mathbb{E}[X'_{k,f}] \;\sim\; \left(\frac{2|v|}{(k+1)!} + \frac{u-|v|}{k!}\right) p^{2k} N^{k+1}, \tag{3.6}$$

and $X'_{k,f} \;\sim\; \mathbb{E}[X'_{k,f}]$ whenever $N^{-\left(\frac{k+1}{2k}\right)} = o(p)$.

We shall just sketch the proof that $\mathbb{E}[X'_{k,f}]$ behaves like the right-hand side of (3.6) in the case when $v > 0$. The proof for $v < 0$ is similar, and the concentration of X_k about its mean when $N^{-\left(\frac{k+1}{2k}\right)} = o(p)$ is established by the same kind of second moment argument as in Section 2.

If $v > 0$ then for any $A \subseteq I_N$ we have $f(A) \subseteq I_{(u+v)N}$. Then

$$\mathbb{E}[X'_{k,f}] \ \sim \ \xi_{k,f}(N) \cdot p^{2k}, \tag{3.7}$$

where

$$\xi_{k,f}(N) \ = \ \sum_{n=0}^{(u+v)N} \binom{R(n)}{k} \tag{3.8}$$

and $R(n)$ denotes the number of solutions to the equation $ux + vy = n$ satisfying $(x, y) \in I_N \times I_N$. For any integer n, the general integer solution to $ux + vy = n$ is of course

$$x = nx_0 - vt, \quad y = ny_0 + ut, \quad t \in \mathbb{Z}, \tag{3.9}$$

where $ux_0 + vy_0 = 1$. If $n > 0$ then there are $\lfloor n/uv \rfloor + O(1)$ solutions in non-negative integers, and for all such solutions, $(x, y) \in I_{\lfloor n/u \rfloor} \times I_{\lfloor n/v \rfloor}$. For $n \in I_{(u+v)N}$ the following three situations then arise :

(I) $n \in I_{vN}$: then all non-negative solutions satisfy $(x, y) \in I_N \times I_N$, so $R(n) = \lfloor n/uv \rfloor + O(1)$ in this case.
(II) $vN < n \leq uN$: we have $R(n) = \lfloor N/u \rfloor + O(1)$ for any such n.
(III) $uN < n \leq (u+v)N$: we have $R(n) = \lfloor \frac{1}{uv} \left[(u+v)N - n \right] \rfloor + O(1)$ for these n.

Thus it follows that

$$\xi_{k,f} \sim \sum_{n=0}^{vN} \binom{\lfloor n/uv \rfloor}{k} + \sum_{n=vN}^{uN} \binom{\lfloor N/v \rfloor}{k} + \sum_{n=uN}^{(u+v)N} \binom{\lfloor \frac{1}{uv} \left[(u+v)N - n \right] \rfloor}{k} \tag{3.10}$$

$$\sim 2 \cdot uv \sum_{n=0}^{\lfloor N/u \rfloor} \binom{n}{k} + (u-v)N \binom{\lfloor N/u \rfloor}{k} \tag{3.11}$$

$$\sim 2uv \frac{\left(\frac{N}{u}\right)^{k+1}}{(k+1)!} + (u-v)N \frac{\left(\frac{N}{u}\right)^{k}}{k!} \tag{3.12}$$

which, together with (3.7), verifies our claim that $\mathbb{E}[X'_{k,f}]$ behaves like the right-hand side of (3.6).

Once we have (3.6) then, in a similar manner to Section 2, we can prove part (i) of Theorem 3.1 by noting that $X'_{2,f} = o(X'_{1,f})$ almost surely when $p = o(N^{-1/2})$, and part (ii) by showing that

$$\mathscr{D}_f \ \sim \ \sum_{k=1}^{\infty} (-1)^{k-1} X'_{k,f} \tag{3.13}$$

when $p = cN^{-1/2}$. It's a simple exercise to check that (3.13) and (3.6) yield (3.3).

Case II : $N^{-1/2} = o(p(N))$.

We give a sketch of the estimate for $\mathbb{E}[\mathscr{D}_f^c]$, the details of the concentration estimate being completely analogous to what has gone before. Let us continue to assume $v > 0$, the proof for $v < 0$ being similar. As in the proof of Theorem 1.1(iii) one may check that various dependencies do not affect our estimates which, using observations (I),(II),(III) above, lead to

$$\mathbb{E}[\mathscr{D}_f^c] \sim \sum_{n=0}^{vN} (1-p^2)^{\lfloor n/uv \rfloor} + \sum_{n=vN}^{uN} (1-p^2)^{N/u} + \sum_{n=uN}^{(u+v)N} (1-p^2)^{\lfloor \frac{1}{uv}[(u+v)N-n] \rfloor} \tag{3.14}$$

$$\sim 2 \cdot uv \sum_{n=0}^{\lfloor N/u \rfloor} (1-p^2)^n + (u-v)N(1-p^2)^{N/u}. \tag{3.15}$$

The sum is $\sim 1/p^2$ and the second term is negligible since $1 = o(Np^2)$, so $\mathbb{E}[\mathscr{D}_f^c] \sim 2uv/p^2$, as claimed.

This completes the proof of Theorem 3.1. □

As mentioned earlier, the main result of [NOORS] was that, for any two binary forms f and g, including the case when $g(x,y) = x + y$, there exist finite sets A_1, A_2 of integers such that $|f(A_1)| > |g(A_1)|$ and $|f(A_2)| < |g(A_2)|$. Theorem 3.1 has a number of consequences on the matter of comparing $|f(A)|$ and $|g(A)|$ for given f and g and random subsets A of I_N for large N. We now reserve the notations $f(x,y) := u_1 x + v_1 y$ and $g(x,y) := u_2 x + v_2 y$ for two forms being compared. Unless otherwise stated, we assume neither f nor g is the form $x + y$. A generic form $ux + vy$ will be denoted $h(x,y)$.

It is convenient to formalize a piece of terminology which we used informally in the introduction :

Definition 3.2. *Let f, g be two binary linear forms as above. Let $p : \mathbb{N} \to (0,1)$ satisfy* (1.3). *Then we say that f* **dominates** *g for the parameter p if, as $N \to \infty$, $|f(A)| > |g(A)|$ almost surely when A is a random subset of I_N obtained by choosing each $n \in I_N$ independently with probability $p(N)$.*

We now consider three different regimes (depending on how rapidly $p(N)$ decays). In the arguments below we shall write p for $p(N)$. The most interesting behavior will be isolated afterwards as Theorem 3.3.

Regime 1 : $N^{-1/2} = o(p(N))$.

Then part (iii) of Theorem 3.1 implies, in particular, that $\mathscr{D}_f \sim (u_1 + |v_1|)N$ and $\mathscr{D}_g \sim (u_2 + |v_2|)N$. Hence f dominates g when $u_1 + |v_1| > u_2 + |v_2|$. In particular, this is the case if $g(x,y) = x - y$. If $u_1 + |v_1| = u_2 + |v_2|$ then the theorem says that f dominates g if and only if $u_1|v_1| < u_2|v_2|$, which is the case if and only if $u_1 > u_2$.

Regime 2 : $p = o(N^{-1/2})$.

Part (i) of Theorem 3.1 says that $\mathscr{D}_f \sim \mathscr{D}_g \sim (Np)^2$ for any f and g. For every $k \geq 1$ we have

$$X'_{k,h} = \Theta_{k,u,v}(N^{k+1}p^{2k}). \tag{3.16}$$

Thus

$$X'_{k+1,h} \;=\; O(Np^2 X'_{k,h}) = o(X'_{k,h}) \text{ almost surely.} \tag{3.17}$$

The second moment method gives standard deviations

$$\sigma(X'_{k,h}) = \Theta_{k,u,v}\left(\sqrt{\mathbb{E}[X'_{k,h}] + \Delta}\right) \;=\; \Theta_{k,u,v}\left(\max\{N^{\frac{k+1}{2}}p^k, N^{k+\frac{1}{2}}p^{2k-\frac{1}{2}}\}\right). \tag{3.18}$$

In particular we have $\sigma(X'_{1,h}) = \Omega\left([Np]^{3/2}\right)$ and $X'_{2,h} = \Theta(N^3p^4)$. First of all, then, if $(N^3p^4) = O((Np)^{3/2})$, i.e.: if $p = O(N^{-3/5})$, then the uncertainty in the size of the random set A itself swamps everything else, and our model is worthless.

If $N^{-3/5} = o(p)$ then, by (3.17), it is in the first instance the $X'_{2,h}$-term which will be decisive. By (3.6) we have

$$X'_{2,h} \;\sim\; \frac{1}{u^2}\left(\frac{|v|}{3} + \frac{u-|v|}{2}\right)N^3p^4. \tag{3.19}$$

Hence f dominates g in this range of δ if $\alpha(u_1, v_1) < \alpha(u_2, v_2)$ where

$$\alpha(u,v) \;:=\; \frac{1}{u^2}\left(\frac{|v|}{3} + \frac{u-|v|}{2}\right) \;=\; \frac{3u-|v|}{6u^2}. \tag{3.20}$$

Since it is easy to see that $\alpha(u_1v_1) = \alpha(u_2, v_2)$ if and only if $u_1 = u_2, v_1 = \pm v_2$, this allows us to compare any pair of forms in the range $N^{-3/5} = o(p)$ and $p = o(N^{-1/2})$, except a pair $ux \pm vy$. But for such a pair, our methods are entirely worthless anyway, since all the estimates in this section depend only on $|v|$. Note in particular that $\alpha(u,v) < \alpha(1,-1)$ for any $(u,v) \neq (1,-1)$ so that any other form dominates $x - y$.

Regime 3 : $p = cN^{-1/2}$.

By part (iii) of Theorem 3.1, for a given value of the parameter $c \in (0,\infty)$, f dominates g if

$$g_{u_1,v_1}\left(\frac{c^2}{u_1}\right) \;>\; g_{u_2,v_2}\left(\frac{c^2}{u_2}\right). \tag{3.21}$$

Since $g_{u,v}(x) \to u + |v|$ as $x \to \infty$, f will dominate g for sufficiently large values of c, provided $u_1 + |v_1| > u_2 + |v_2|$. This is as expected from Regime 1. On the other hand, the Taylor expansion of $g_{u,v}$, as a function of c, around $c = 0$, reads

$$g_{u,v}(c) \;=\; c^2 - \alpha(u,v)c^4 + O_{u,v}(c^6). \tag{3.22}$$

Thus f dominates g for sufficiently small values of c provided $\alpha(u_1, v_1) < \alpha(u_2, v_2)$. Again this is as expected, this time from Regime 2. Note that the injectivity of α allows us to even compare forms with the same value of $u + |v|$, namely : for a fixed value of $u + |v|$, $\alpha(u,v)$ is clearly a decreasing function of u. Hence if $u_1 + |v_1| = u_2 + |v_2|$ then f dominates g for all values of $c \in (0,\infty)$ if and only if $u_1 > u_2$. Note that this is the same condition as in Regime 1. More generally, we have that f dominates g for all values of c whenever $\alpha(u_1, v_1) < \alpha(u_2, v_2)$ and $u_1 + |v_1| \leq u_2 + |v_2|$. In particular this is the case for $g(x,y) = x - y$ and any other f.

The most interesting phenomenon arises when we compare two forms such that

$$u_1 + |v_1| \;>\; u_2 + |v_2| \quad\text{and}\quad \alpha(u_1, v_1) \;>\; \alpha(u_2, v_2). \tag{3.23}$$

Then the combined observations of Regimes 1, 2 and 3 imply that there exists some $c_{f,g} > 0$ such that

$$\begin{aligned} &g \text{ dominates } f \text{ whenever } N^{-3/5} = o(p) \text{ and } p = o(N^{-1/2}) \\ &\text{or } p = cN^{-1/2} \text{ for any } 0 < c < c_{f,g}, \end{aligned} \tag{3.24}$$

whereas

$$f \text{ dominates } g \text{ whenever } p = cN^{-1/2} \text{ for any } c > c_{f,g} \text{ or } N^{-1/2} = o(p). \tag{3.25}$$

This observation may be considered a partial generalization of the main result of [NOORS] to random sets, partial in the sense that it only applies to pairs of forms satisfying (3.23). Equations (3.24) and (3.25) say that we have a sharp threshold, below which g dominates f and above which f dominates g. We leave it to future work to determine what happens as one crosses this sharp threshold.

We close this section by summarizing the most important observations above in a theorem.

Theorem 3.3. *Let $f(x, y) = u_1x + u_2y$ and $g(x, y) = u_2x + g_2y$, where $u_i \geq |v_i| > 0$, GCD$(u_i, v_i) = 1$ and $(u_2, v_2) \neq (u_1, \pm v_1)$. Let $\alpha : \mathbb{Z}^2_{\neq 0} \rightarrow \mathbb{Q}$ be the function given by* (3.20). *The following two situations can be distinguished :*

(i) $u_1 + |v_1| \geq u_2 + |v_2|$ and $\alpha(u_1, v_1) < \alpha(u_2, v_2)$.

Then f dominates g for all p such that $N^{-3/5} = o(p)$ and $p = o(1)$. In particular, every other difference form dominates the form $x - y$ in this range.

(ii) $u_1 + |v_1| > u_2 + |v_2|$ and $\alpha(u_1, v_1) > \alpha(u_2, v_2)$.

Then there exists $c_{f,g} \in \mathbb{R}^+$ such that (3.24) *and* (3.25) *hold. Specifically, $c_{f,g}$ is the unique positive root of the equation*

$$g_{u_1,v_1}\left(\frac{c^2}{u_1}\right) = g_{u_2,v_2}\left(\frac{c^2}{u_2}\right), \tag{3.26}$$

where $g_{u,v}(x) : \mathbb{R}^+ \rightarrow (0, u + |v|)$ is given by (3.2).

4. Open Problems

Here is a sample of issues which could be the subject of further investigations :

1. One unresolved matter is the comparison of arbitrary difference forms in the range where $N^{-3/4} = O(p)$ and $p = O(N^{-3/5})$. Here the problem is that the binomial model itself does not prove of any use. This provides, more generally, motivation for looking at other models. Obviously one could look at the so-called *uniform* model on subsets (see [JŁR]), but this seems a more awkward model to handle. Note that the property of one binary form dominating another is not monotone, or even convex.

2. Secondly, a very tantalizing problem is to investigate what happens while crossing a sharp threshold, whenever it arises under the conditions of Theorem 3.3(ii).

3. Thirdly, one can ask if the various concentration estimates in Theorem 1.1 can be improved. When $p = o(N^{-1/2})$ we have only used an ordinary second moment argument, and it is possible to provide explicit estimates. Explicitly, the following follows from Chebyshev's Theorem (see the appendix to [HM] or [MS] for a proof).

Theorem 4.1. *Let $p(N) := cN^{-\delta}$ for some $c > 0$, $\delta \in (1/2, 1)$. Set $C := \max(1, c)$, $f(\delta) := \min\{\frac{1}{2}, \frac{3\delta-1}{2}\}$ and let $g(\delta)$ be any function such that $0 < g(\delta) < f(\delta)$ for all $\delta \in (1/2, 1)$. Set $P_1(N) := (4/c)N^{-(1-\delta)}$ and $P_2(N) := N^{-(f(\delta)-g(\delta))}$. For any subset chosen with respect to the binomial model with parameter $p = p(N)$, with probability at least $1 - P_1(N) - P_2(N)$ the ratio of the cardinality of its difference set to the cardinality of its sumset is $2 + O_C(N^{-g(\delta)})$. Thus the probability a subset chosen with respect to the binomial model is not difference dominated is at most $P_1(N) + P_2(N)$, which tends to zero rapidly with N for $\delta \in (1/2, 1)$.*

The range $N^{-1/2} = o(p(N))$ seems more interesting, however. Here we proved that the random variable $\mathscr{S}^c$ has expectation of order $P(N)^2$, where $P(N) = 1/p(N)$, and is concentrated within $P(N)^{3/2}\log^2 P(N)$ of its mean. Now one can ask whether the constant $3/2$ can be improved, or at the very least can one get rid of the logarithm?

4. Finally, it is natural to ask for extensions of our results to $\mathbb{Z}$-linear forms in more than two variables. Let

$$f(x_1, ..., x_k) \;=\; u_1x_1 + \cdots + u_kx_k, \quad u_i \in \mathbb{Z}_{\neq 0}, \tag{4.1}$$

be such a form. We conjecture the following generalization of Theorem 3.1 :

Conjecture 4.2. *Let $p : \mathbb{N} \to (0, 1)$ be a function satisfying (1.3). For a positive integer N, let A be a random subset of I_N obtained by choosing each $n \in I_N$ independently with probability $p(N)$. Let f be as in (4.1) and assume that $GCD(u_1, ..., u_n) = 1$. Set*

$$\theta_f \;:=\; \#\{\sigma \in S_k : (u_{\sigma(1)}, ..., u_{\sigma(k)}) = (u_1, ..., u_k)\}. \tag{4.2}$$

Let $\mathscr{D}_f$ denote the random variable $|f(A)|$. Then the following three situations arise :

(i) $p(N) = o(N^{-1/k})$: Then

$$\mathscr{D}_f \;\sim\; \frac{1}{\theta_f}(N \cdot p(N))^k. \tag{4.3}$$

(ii) $p(N) = c \cdot N^{-1/k}$ for some $c \in (0, \infty)$: There is a rational function $R(x_0, ..., x_k)$ in $k + 1$ variables, which is increasing in x_0, and an increasing function $g_{u_1,...,u_k} : (0, \infty) \to (0, \sum_{i=1}^{k} |u_i|)$ such that

$$\mathscr{D}_f \;\sim\; g_{u_1,...,u_k}\left(R(c, u_1, ..., u_k)\right) \cdot N. \tag{4.4}$$

(iii) $N^{-1/k} = o(p(N))$: Let $\mathscr{D}_f^c := \left(\sum_{i=1}^{k} |u_i|\right) N - \mathscr{D}_f$. Then

$$\mathscr{D}_f^c \;\sim\; \frac{2\theta_f \prod_{i=1}^{k} |u_i|}{p(N)^k}. \tag{4.5}$$

ACKNOWLEDGEMENT

We thank Devdatt Dubhashi for pointing us in the direction of Vu's paper, the participants of CANT 2007 for interesting conversations, and the referees for comments on an earlier draft. The second named author was partly supported by NSF grant DMS0600848.

REFERENCES

[AS] N. Alon and J. H. Spencer, *The Probabilistic Method*, Wiley, 1992.

[GJLR] A. P. Godbole, S. Janson, N. W. Locantore Jr. and R. Rapoport, *Random Sidon sequences*, J. Number Theory **75** (1999), no. 1, 7–22.

[He] P. V. Hegarty, *Some explicit constructions of sets with more sums than differences*, Acta Arith. **130** (2007), no. 1, 61–77.

[HM] P. V. Hegarty and S. J. Miller, *When almost all sets are difference dominated*, preprint. http://www.arxiv.org/abs/0707.3417

[JŁR] S. Janson, T. Łuczak and A. Ruciński, *Random Graphs*, Wiley, 2000.

[KiVu] J. H. Kim and V. H. Vu, *Concentration of multivariate polynomials and its applications*, Combinatorica **20** (2000), 417–434.

[MO] G. Martin and K. O'Bryant, *Many sets have more sums than differences*, Additive combinatorics, 287–305, CRM Proc. Lecture Notes **43**, Amer. Math. Soc., Providence, RI, 2007.

[MS] S. J. Miller and D. Scheinerman, *Explicit constructions of infinite families of MSTD sets*, preprint. http://arxiv.org/abs/0809.4621

[Na1] M. B. Nathanson, *Problems in additive number theory, 1*, Additive combinatorics, 263–270, CRM Proc. Lecture Notes **43**, Amer. Math. Soc., Providence, RI, 2007.

[Na2] M. B. Nathanson, *Sets with more sums than differences*, Integers : Electronic Journal of Combinatorial Number Theory **7** (2007), Paper A5 (24pp).

[NOORS] M. B. Nathanson, K. O'Bryant, B. Orosz, I. Ruzsa and M. Silva, *Binary linear forms over finite sets of integers*, Acta Arith. **129** (2007), no. 4, 341–361.

[Ta] M. Talagrand, *A new look at indepedence*, Ann. Prob **24** (1996), 1–34.

[Vu1] V. H. Vu, *New bounds on nearly perfect matchings of hypergraphs: Higher codegrees do help*, Random Structures and Algorithms **17** (2000), 29–63.

[Vu2] V. H. Vu, *Concentration of non-Lipschitz functions and Applications*, Random Structures and Algorithms **20** (2002), no. 3, 262-316.

E-mail address: hegarty@math.chalmers.se

MATHEMATICAL SCIENCES, CHALMERS UNIVERSITY OF TECHNOLOGY AND GÖTEBORG UNIVERSITY, GÖTEBORG, SWEDEN

E-mail address: Steven.J.Miller@williams.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267