WHEN GENERALIZED SUMSETS ARE DIFFERENCE DOMINATED

VIRGINIA HOGAN AND STEVEN J. MILLER

Abstract. We study the relationship between the number of minus signs in a generalized sumset, $A+\cdots+A-\cdots-A$, and its cardinality; without loss of generality we may assume there are at least as many positive signs as negative signs. As addition is commutative and subtraction is not, we expect that for most A a combination with more minus signs has more elements than one with fewer; however, recently Iyer, Lazarev, Miller and Zhang [ILMZ] proved that a positive percentage of the time the combination with fewer minus signs can have more elements. Their analysis involves choosing sets A uniformly at random from $\{0,\dots,N\}$; this is equivalent to independently choosing each element of $\{0,\dots,N\}$ to be in A with probability $1/2$. We investigate what happens when instead each element is chosen with probability $p(N)$, with $\lim_{N\to\infty} p(N)=0$. We prove that the set with more minus signs is larger with probability 1 as $N\to\infty$ if $p(N)=cN^{-\delta}$ for $\delta\ge\frac{h-1}{h}$, where h is the number of total summands in $A+\cdots+A-\cdots-A$, and explicitly quantify their relative sizes. The results generalize earlier work of Hegarty and Miller [HM], and we see a phase transition in the behavior of the cardinalities when $\delta=\frac{h-1}{h}$.

Contents

Date: January 24, 2013.

2010 *Mathematics Subject Classification.* 11P99 (primary), 11K99 (secondary).

Key words and phrases. Sum dominated sets, more sum than difference sets, strong concentration, phase transitions.

The first named author was partially supported by NSF Grant DMS0850577, and second named author was partially supported by NSF Grant DMS0970067. We thank the participants of the 2012 SMALL REU program, especially Kevin Vissuet, as well as Kevin O'Bryant and Dmitrii Zhelezov for helpful discussions.

1. Introduction

1.1. **Previous Results.** Let A be a subset of the integers. We define the **sumset** $A+A$ and the **difference set** $A-A$ by

$$A+A \;=\; \{a_1+a_2 : a_i \in A\}, \quad A-A \;=\; \{a_1-a_2 : a_i \in A\}. \tag{1.1}$$

Many important problems in number theory are related to these sets and their generalizations. For example, if P denotes the set of primes and K the set of k^{th} powers of positive integers, then the Goldbach conjecture is equivalent to $P+P$ contains all even numbers, the twin prime conjecture is $P-P$ contains 2 infinitely often, Fermat's Last Theorem is $(K+K)\cap K$ is empty if $k \geq 3$, and Waring's problem is that for each k there is an s such that $K+\cdots+K$ (s times) contains all positive integers.

Note the last problem involves more than one binary operation; the main goal of this paper is to explore what happens to generalized sumsets in different models. Before stating our results, we review some previous work. 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 A has a larger difference set than sumset. If this is the case then we say A is **difference dominated**, while if the two sets have the same size we say the set is **balanced**, and if the sumset is larger then A is **sum dominated** (also called a **more sums than differences (MSTD) set**). It was conjectured that if A is chosen uniformly at random from $\{0,\dots,N\}$ then as $N\to\infty$ almost all sets are difference dominated. In 2007, however, Martin and O'Bryant [MO] disproved this conjecture by showing a positive percentage of sets are sum dominated. The percentage is small, around $4.5\cdot 10^{-4}$ [Zh].

While these results imply that sum dominated sets are not too rare, this is a consequence of how the sets are chosen. An equivalent formulation is that each element of $I_N := \{0,\dots,N\}$ is chosen to be in A with probability $1/2$. With high probability a randomly chosen subset A has approximately $N/2$ 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. Almost all possible sums and differences are realized; the expected number of missing differences is 6, while the expected number of missing sums is 10. Thus, a typical set needs just a small nudge to become sum dominated. This can be accomplished by appropriately choosing the fringe elements of A (the elements near 0 and N), as almost surely changes at the fringes do not affect whether or not most possible sums and differences are realized.

This observation suggests that instead of taking each element of I_N with probability $1/2$ (or any fixed, non-zero probability), we should instead explore what happens when all of these elements are chosen independently with probability $p(N)$, where p is some function tending to zero; this is a binomial model with parameter $p(N)$. Such an analysis was done by Hegarty and Miller [HM] in 2009. They showed that if $p(N)=cN^{-\delta}$ for some $\delta\in(0,1)$, then almost surely A is difference dominated. The analysis breaks into three cases based on the probability for choosing elements in A. The authors study **fast decay** ($\delta>1/2$), **critical decay** ($\delta=1/2$), and **slow decay** ($\delta<1/2$). There is a phase transition at $\delta=1/2$, leading to the name critical decay.

Before stating their results we first introduce some definitions, notation, conventions, and standard facts that we use in our results as well.

We start with notation for sizes. By $f(x) = O(g(x))$ we mean that there exist constants x_0 and C such that for all $x \geq x_0$, $|f(x)| \leq Cg(x)$. We write $f(x) = \Theta(g(x))$ if both $f(x) = O(g(x))$ and $g(x) = O(f(x))$. If $\lim_{x\to\infty} f(x)/g(x) = 0$ then we write $f(x) = o(g(x))$, which is equivalent to $f(x) \ll g(x)$.

As the fundamental objects of study are sizes of sets, we need a way to denote asymptotic behavior. Let X be a real-valued random variable depending on some positive integer parameter N, and let $f(N)$ be some real-valued function. By "$X \sim f(N)$" we mean that, for any $\epsilon_1, \epsilon_2 > 0$, there exists $N_{\epsilon_1,\epsilon_2} > 0$ such that, for all $N > N_{\epsilon_1,\epsilon_2}$,

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

We now state the main past result, which we will generalize.

Theorem 1.1 (Hegarty-Miller [HM])**.** *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

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

Notice there is a phase transition at $\delta = 1/2$, where $|A-A|$ goes from almost surely having twice as many elements as $|A+A|$ (when $\delta > 1/2$) to having the same number of elements to first order (when $\delta < 1/2$); further, an explicit, tractable formula is obtained for the relative sizes when $\delta = 1/2$ as a simple function of c.

The goal of this paper is to generalize this theorem to arbitrary combinations of sums and differences.

1.2. **Results.** Before stating our results, we need some combinatorial results. We use the extended definition of the binomial coefficient, setting $\binom{a}{b} = 0$ for integers $0 \le a < b$. A central result, which we use again and again, is the stars and bars (or cookie) problem: for any pair of positive integers n, k, the number of distinct k-tuples of non-negative integers that sum to n is $\binom{n+k-1}{k-1}$. Note this is equivalent to counting the number of solutions in non-negative integers to $x_1 + \cdots + x_k = n$. This is readily found. If we choose $k-1$ objects from $n+k-1$ (there are $\binom{n+k-1}{k-1}$ ways to do so), we partition the remaining n objects into k sets, and there is a one-to-one correspondence between these partitions and our desired solutions.

In our investigations below we always choose elements for our set A from $I_N := \{0, \dots, N\}$ independently with probability $p(N) = cN^{-\delta}$ for fixed $\delta \in (0,1)$ and $c > 0$.

- Given a set A we define its generalized sumset $A_{s,d}$ with s sums and d differences to be $A + \cdots + A - \cdots - A$; as we are only interested in cardinalities we may always assume $d \le s$.
- We write $|A_{s,d}|$ for its size. We always use h for the number of summands, so $h = s+d$.
- An $h_{(s,d)}$-tuple is a set of $h = s + d$ integers, $\{a_1, \dots, a_s, a_{s+1}, \dots, a_h\}$.
- If the associated sum $\sum_{i=1}^{s} a_i - \sum_{j=1}^{d} a_{s+j}$ equals λ then we say the tuple generates λ. Note that the generalized sumset is the set of all numbers generated by $h_{(s,d)}$-tuples of elements of A.
- Related to this is $R(n,s,d)$, which we define to be the number of ways to generate n through $h_{(s,d)}$-tuples of integers drawn from $\{0, \dots, N\}$. As $R(n,s,d)$ counts all permutations equally, order matters; for example, if $a_1 + a_2 - a_3 = n$, then $R(n,2,1)$ counts (a_1, a_2, a_3) and (a_2, a_3, a_1) as two different entities. As N is fixed throughout our calculations and then sent to infinity only at the end, to simplify notation we write $R(n,s,d)$, though really it should be $R_N(n,s,d)$ to emphasize this dependence.

In the course of our investigations we encounter the following constants and functions. For k a positive integer and $j \in (0, h/2)$, set

$$b_{h,k} := \frac{1}{k!(h-1)!^k} \cdot 2 \sum_{j \le h/2} j^{(h-1)k} \int_0^1 \left(\sum_{i=0}^{j} (-1)^i \binom{h}{i} \left(1 - \frac{(i-t)}{j}\right)^{h-1} \right)^k dt. \tag{1.8}$$

These constants emerge in our phase transition function

$$g(x; s, d) := \sum_{k=1}^{\infty} (-1)^{k-1} \frac{b_{h,k}}{(s!d!)^k} x^{(s+d)k}, \tag{1.9}$$

which for $h = s + d \ge 2$ converges for all x.

Our main result is the following.

Theorem 1.2. *Let h be a positive integer, $c > 0$ a real number, and choose pairs of integers (s_i, d_i) with $s_i \ge d_i$ and $s_i + d_i = h$; for definiteness let $d_1 > d_2$. Consider subsets $A \subset I_N$ where each element of I_N is independently chosen to be in A with probability $p(N) = cN^{-\delta}$.*

- *For $\delta > \frac{h-1}{h}$, the set A_{s_i,d_i} with the larger d_i is larger almost surely. In particular, as $N \to \infty$ with probability one we have $|A_{s_1,d_1}|/|A_{s_2,d_2}| = (s_2!d_2!)/(s_1!d_1!) + o(1)$.*
- *If $\delta = \frac{h-1}{h}$ then almost surely $|A_{s_i,d_i}| \sim N g(c; s_i, d_i)$ (with g defined in (1.9)), and thus with probability one $|A_{s_1,d_1}|/|A_{s_2,d_2}|$ is $g(c; s_1, d_1)/g(c; s_2, d_2) + o(1)$.*

Thus for two sets with the total number of summands fixed, the set with more minus signs is larger almost surely when $\delta > \frac{h-1}{h}$, so there are more distinct elements in the generalized sumset with more minus signs. There is a phase transition in the behavior when δ passes from being greater than $\frac{h-1}{h}$ to equaling $\frac{h-1}{h}$.

The proof is similar to that in [HM], which does the $h = 2$ case. The idea is to bound the number of times distinct $h_{(s,d)}$-tuples generate the same element. This allows us to discount the number of repeated elements in the generalized sumset. If we already know that most elements are distinct, then simple combinatorics allows us to compare their sizes; however, as δ gets smaller, we choose more and more elements for A, which leads to more repeated elements in the generalized sumset. The analysis is significantly easier when there are fewer repeated generalized sums, as then the sizes of the two generalized sumsets are well separated. Specifically, in the case of fast decay, the analysis follows from Chebyshev's inequality. The case of critical decay is significantly more challenging and requires recent strong concentration results. We first must show that we can estimate the number of $h_{(s,d)}$-tuples with a constant sum by the number of $h_{(s,d)}$-tuples with h distinct elements. We then show that if we partition our $h_{(s,d)}$-tuples into equivalence classes based on the number of other $h_{(s,d)}$-tuples with the same sum, the class of singletons is the largest, so most $h_{(s,d)}$-tuples generate a unique integer. We then define our function $g(x; s, d)$ in terms of $|A_{s,d}|$.

In Section 2.1 we define $R(n, s, d)$ to count possible values for $A_{s,d}$. In Section 2.2 we bound the expected number of repeated elements, and in Section 2.3 we show that the number of repeated elements is close to its expected value. In Section 3.1 we study the case of fast decay. In Section 3.2 we study the case of critical decay. We end with a discussion of future work.

2. Strong Concentration

In this section, we first derive formulas for quantities related to the number of $h_{(s,d)}$-tuples generating a given number. These results are key ingredients in the strong concentration analysis.

2.1. Determining $R(n, s, d)$.

Our first step is to determine a tractable formula for $R(n, s, d)$, the number of $h_{(s,d)}$-tuples of integers drawn from $\{0, \dots, N\}$ that generate n.

Lemma 2.1. *Let $n' := n + dN$. We have*

$$R(n, s, d) \;=\; \sum_{i=0}^{\lfloor \frac{n'}{N} \rfloor - 1} (-1)^i \binom{h}{i} \binom{n' - i(N+1) + h - 1}{h-1}. \tag{2.1}$$

Proof. We first assume $d = 0$, so all signs are positive and $n' = n$. By the stars and bars / cookie problem, the number of ways to write n as a sum of h non-negative integers is $\binom{n+h-1}{h-1}$. As it will be important later, it is worth noting that this treats $4 + 3 + 1$ and $3 + 4 + 1$ as two different representations. Also, note this is equivalent to solving $x_1 + \cdots + x_h = n$ with each x_i a non-negative integer. We desire each summand to be in $I_N := \{0, \dots, N\}$, and thus $\binom{n+h-1}{h-1}$ may overcount. We remedy this by using inclusion-exclusion to remove representations with summands exceeding N.

We first remove all representations where at least one summand exceeds N. There are $\binom{h}{1}$ ways to choose which summand this is. We write that summand as $x_j = y_j + N + 1$,

and write $x_j = y_j$ for the remaining summands. Thus the number of representations where summand j exceeds N and the other summands are at least zero is the number of solutions to $y_1 + \cdots + y_h = n - (N+1)$, which is just $\binom{n-(N+1)+h-1}{h-1}$. If instead i summands are greater than N, we would get $y_1 + \cdots + y_h = n - i(N+1)$, for $\binom{n-i(N+1)+h-1}{h-1}$ solutions. The claim now follows by inclusion-exclusion.

We only need trivial modifications if $d > 0$. For the d elements occurring with a minus sign, $a_{s+1}, \ldots, a_{s+d}$, write $a'_j = N - a_j$. Then

$$a_1 + \cdots + a_s - a_{s+1} - \cdots - a_{s_j} \;=\; n \tag{2.2}$$

becomes

$$a_1 + \cdots + a_s + a'_{s+1} + \cdots + a'_{s+d} \;=\; n + dN, \tag{2.3}$$

reducing us to the first case. □

In our strong concentration applications we need not $R(n, s, d)$, but the closely related quantity $R_{\text{distinct}}(n, s, d)$, which counts the number of representations of n by $h = s + d$ distinct elements. The next lemma shows that these two quantities differ in a lower order term (relative to N).

Lemma 2.2. *The number of $h_{(s,d)}-$tuples which generate n using h distinct elements is of a higher order than repeated elements. In particular, if $n' = n + dN$ then*

$$R(n, s, d) \;=\; R_{\text{distinct}}(n, s, d) + O(N^{h-2}). \tag{2.4}$$

Remark 2.3. *If $(n')^{h-2} = o(N)$, the error term in Lemma 2.2 exceeds the main term. While a more careful analysis gives a better error estimate, the bound above suffices for our applications as the main term is summed over a large enough regime that its contribution exceeds that of the error.*

Proof. If there is at least one repeated element, there are at most $h - 2$ free choices for the summands (we lose one choice for the repetition, and one choice as the sum must equal n). Thus the contribution from representations of n with a repeated element is at most $O(N^{h-2})$. □

2.2. **Generalizing Hegarty-Miller's Random Variables.** Hegarty and Miller [HM] introduce some useful random variables to prove their strong concentration results. We begin with a generalization of these quantities, and then derive useful bounds which give the needed asymptotic relations.

For a set A, define

$$A_k \;:=\; \Bigg\{ \left\{\{a_1, ..., a_h\}, \ldots, \left\{a_{(k-1)h+1}, ..., a_{kh}\right\}\right\} : \textstyle\sum_{i=1}^{s} a_i - \sum_{i=s+1}^{h} a_i \;=\; \cdots$$
$$\cdots \;=\; \textstyle\sum_{i=(k-1)h+1}^{kh-d} a_i - \sum_{i=kh-d}^{kh} a_i \Bigg\}, \tag{2.5}$$

and let $X_k = |A_k|$. Note that now the ordering of elements *within* the h-tuples matters (because subtraction is not commutative), so we are looking at unordered k-tuples of *ordered* elements. The dependence on the ordering is, however, weak. Given any one of these k-tuples, we can permute the first s elements or permute the last d elements without changing the number it generates, and thus such a permutation is the same element of the k-tuple.

If all the elements of an h-tuple are distinct (actually, all we need are no repeats among the first s and no repeats among the final d), then there are $s!d!$ ways to reorder the tuple *without* changing the number it generates, and thus all of these correspond to the same *set* (remember, the only way the ordering matters in the set of h elements is which are the first s elements and which are the last d). Thus, if all elements are distinct, there is overcounting by a factor of $s!d!$; we must take this into account later.

We want to study these k-tuples because they shed light on how many repeated elements are in the generalized sumset. We have k-tuples of $h_{(s,d)}$-tuples, so each k-tuple has a total of hk integers. We place $h_{(s,d)}$-tuples in the same k-tuple if they all generate the same number. Intuitively, because we need to subtract out repeated elements, all $h_{(s,d)}$-tuples within the same k-tuple only count once in our generalized sumset, so counting these k-tuples is equivalent to counting $h_{(s,d)}$-tuples. To make this more concrete, we present a short example. If $h = 3$, $s = 3$, and $d = 0$, then $\{3, 4, 7\}$, $\{5, 6, 3\}$, $\{1, 11, 2\}$, and $\{1, 5, 8\}$ would all be in the same k-tuple because they all sum to 14. If these four $h_{(s,d)}$-tuples were the only $h_{(s,d)}$-tuples that generated 14, then we would have $A_4 = \{\{3, 4, 7\}, \{5, 6, 3\}, \{1, 11, 2\}, \{1, 5, 8\}\}$. X_k counts the number of k-tuples, so the number of times there are exactly k $h_{(s,d)}$-tuples generating the same number. For example, if we also only had 4 $h_{(s,d)}$-tuples that generated 5, and 5 and 14 were the only two numbers generated, then $X_4 = 2$ for the two different numbers generated by exactly 4 $h_{(s,d)}$-tuples.

The reason why A_1 is so important is that $h_{(s,d)}$-tuples are only in A_1 if no other $h_{(s,d)}$-tuples generate that number. We want the number of single $h_{(s,d)}$-tuples in A_1 (counted by X_1) to be the largest because then we know that most $h_{(s,d)}$-tuples generate a unique sum. The larger k is, the more constant sums we must have. We have k-tuples of $h_{(s,d)}$-tuples, and within each k-tuple, all $h_{(s,d)}$-tuples contained inside generate the same number. If there were another $h_{(s,d)}$-tuple that generated the same number, then the two $h_{(s,d)}$-tuples would be in A_2. Therefore, X_1 counts the number of *distinct* sums among our $h_{(s,d)}$-tuples, which is important because we will show that this is the higher order than X_k for any $k > 1$, so X_1 becomes critically important in measuring the size of the generalized sumset. So, X_1 counts the number of $h_{(s,d)}$-tuples that generate a distinct sum, because if an $h_{(s,d)}$-tuple is in A_1, then there are no other $h_{(s,d)}$-tuples that generate the same number. Similarly, X_2 counts how many $h_{(s,d)}$-tuples generate the same sum as exactly one other $h_{(s,d)}$-tuple. Therefore, if we know that $X_2 = o(X_1)$, then we know there are significantly more $h_{(s,d)}$-tuples with a unique sum than those with any number of repeated sums (because any k-tuples in A_k for $k > \ell$ are also in A_ℓ).

The goal is to generalize Theorem 1.1 and Lemma 2.1 of [HM]. To do this we must bound the number of repeated elements in $A_{s,d}$. *If* we knew that our generalized sumset contains mostly distinct sums (so most $h_{(s,d)}$-tuples generate a distinct integer), then a simple combinatorial argument and Chebyshev's theorem would suffice to prove Theorem 1.2. In the case of fast decay, $\delta > \frac{h-1}{h}$, the number of repeated elements is a lower order than the number of distinct elements. The case of critical decay, $\delta = \frac{h-1}{h}$, is more difficult because now the number of our repeated elements is of the same order as the number of distinct elements. Intuitively, the smaller δ is, the more elements from $\{0, \dots, N\}$ are in A, so the more likely it is that two $h_{(s,d)}$-tuples generate the same element. Thus a more sophisticated argument is needed to find the relevant cardinalities.

We first introduce some terminology.

Definition 2.1. *By **Type** 0 we mean the k-tuples with hk distinct elements of I_N, while **Type** i refers to k-tuples with i repeated elements.*

By repeated elements, we mean total number of elements that would need to be removed for all elements to be distinct. For example, in the 7-tuple $\{1,1,1,2,2,3,4\}$, we say there are three repeated elements because we would need to remove $\{1,1,2\}$ for all remaining elements to be distinct. For a fixed k-tuple α, since we draw our A from a binomial model with parameter $p(N) = cN^{-\delta}$, we know

$$\text{Prob}(\alpha \text{ is of Type } t) \;=\; \binom{hk}{t}\, c^{kh-t} N^{-\delta(kh-t)}. \tag{2.6}$$

Equation (2.6) holds because the probability of choosing any element is independent of the probability of choosing any other element. We need a binomial coefficient because we have to choose t of the k-tuple's total hk elements to repeat. Note that (2.6) is for a fixed k-tuple, but we do not know the locations of the repeated elements, so the binomial coefficient is necessary for all possible combinations of repeats.

Let $\xi_{i,k}(N)$ be the number of k-tuples of type i. Note that we have k-tuples of h-element sets; in those h element sets, the ordering of elements within matters a bit, though we may permute the first s or permute the last d without changing the number it generates. As in equation (2.4) of [HM],

$$\mathbb{E}(X_k) \;=\; \sum_{i=0}^{hk-1} \xi_{i,k}(N) p(N)^{(k-i)h}. \tag{2.7}$$

This holds because we are summing over all possible types of k-tuples times the probability of choosing a k-tuple of that type, so we get the expected number of k-tuples.

Similar to [HM], for $\delta \geq \frac{h-1}{h}$, the only contribution to (2.7) that matters is from the first term. This is equivalent to estimating the number of k-tuples by only considering the number of k-tuples with no repeated elements. We first estimate the contribution from this term, and then bound the contribution from the remaining ones.

Lemma 2.4. *We have*

$$\xi_{0,k}(N) \;\sim\; \frac{b_{h,k}}{(s!d!)^k} N^{(h-1)k+1}. \tag{2.8}$$

The error above is $O(N^{(h-2)(k-1)+1})$, and $b_{h,k}$ is defined in (1.8).

Proof. Because we have to sum over all n in the interval to count how many times a k-tuple can generate the same number, the number of k-tuples of Type 0 is

$$\xi_{0,k}(N) \;=\; \sum_{n=-dN}^{sN} \binom{R(n,s,d)/s!d! + O(N^{h-2})}{k}, \tag{2.9}$$

where $R(n,s,d)$ is the number of $h_{(s,d)}$-tuples elements in $\{0,\dots,N\}$ that generate n. The error is because $\xi_{0,k}(N)$ counts *distinct* tuples, while $R(n,s,d)$ allows repeats; however, our earlier analysis showed that the number of tuples with repeated elements is lower order (this is because h is fixed and N tends to infinity). From the Binomial Theorem and standard bounds on approximating binomial coefficients with the largest term (specifically, $\binom{f(n)}{m} =$

$\frac{f(n)^m}{m!} + O(f(n)^{m-1})$), we find

$$\xi_{0,k}(N) \;=\; \sum_{n=-dN}^{sN} \binom{R(n,s,d)/s!d!}{k} + O\left(\sum_{n=-dN}^{sN} \binom{N^{h-2}}{k}\right) \;\sim\; \sum_{n=-dN}^{sN} \binom{R(n,s,d)/s!d!}{k}. \tag{2.10}$$

Letting $n' = n + dN$ as before, define

$$S_j(N) \;:=\; \sum_{n'=jN}^{(j+1)N} \binom{R(n',s,d)/s!d!}{k}. \tag{2.11}$$

We use the notation $S_j(N)$ to sum over all possible n in $R(n,s,d)$ in one of h intervals of length n. In $S_j(N)$, j gives the index of the interval. From (2.10), we see that it is useful to break $\xi_{0,k}(N)$ into these intervals in order to compute the total sum. To distinguish between $R(n,s,d)$ and $S_j(N)$, recall that $R(n,s,d)$ is for a fixed n, while $S_j(N)$ is for a fixed interval of length N.

Assume h is even (the case of h odd is similar). We first approximate $R(n,s,d)$. Let $j = \lfloor \frac{n'}{N} \rfloor - 1$. Assume $j > 0$; the case of $j = 0$ follows similarly, and we mostly omit the details. We have

$$\begin{aligned} R(n',s,d) &= \sum_{i=0}^{\lfloor \frac{n'}{N} \rfloor - 1} (-1)^i \binom{h}{i} \binom{n' - i(N+1) + h - 1}{h-1} \\ &= \sum_{i=0}^{j} (-1)^i \binom{h}{i} \frac{(n'-iN)^{h-1}}{(h-1)!} + O(N^{h-2}); \end{aligned} \tag{2.12}$$

this follows from standard approximation for the binomial coefficients and the Binomial Theorem.

For the rest of this subsection, in all the analysis below the error term in the asymptotic relations denoted by $\sim$ are at least one order smaller in N.

We now have

$$\begin{aligned} S_j(N) &= \sum_{n'=jN}^{(j+1)N} \binom{\frac{1}{s!d!} \sum_{i=0}^{j} (-1)^i \binom{h}{i} \frac{(n'-iN)^{h-1}}{(h-1)!}}{k} + O\left(\sum_{n'=jN}^{(j+1)N}\right) (n - iN)^{h-2} \\ &\sim \sum_{n'=jN}^{(j+1)N} \frac{\left(\frac{1}{s!d!}\right)^k \left(\sum_{i=0}^{j} (-1)^i \binom{h}{i} \left(\frac{(n'-iN)^{h-1}}{(h-1)!}\right)^k\right)}{k!} \end{aligned} \tag{2.13}$$

We pull out all terms that do not depend on n' to get

$$S_j(N) \;\sim\; \frac{1}{(s!d!)^k k! (h-1)!^k} \sum_{n'=jN}^{(j+1)N} \left(\sum_{i=0}^{j} (-1)^i \binom{h}{i} (n' - iN)^{h-1}\right)^k. \tag{2.14}$$

Our goal is to find the dependence on s, d and N. To do this, we first approximate (2.14) with an integral to get

$$S_j(N) \sim \frac{1}{(s!d!)^k k!(h-1)!^k} \int_{jN}^{(j+1)N} \left(\sum_{i=0}^{j} (-1)^i \binom{h}{i} (x - iN)^{h-1} \right)^k dx; \tag{2.15}$$

the cost of the approximation is one order lower in N as we have sums of polynomials.

We change variables by taking $x = (j+t)N$ with t ranging from 0 to 1. Thus $dx = Ndt$ and as $j > 0$ (if $j = 0$ we cannot pull out the power of j)

$$S_j(N) \sim \frac{1}{(s!d!)^k k!(h-1)!^k} j^{(h-1)k} N^{k(h-1)+1} \int_0^1 \left(\sum_{i=0}^{j} (-1)^i \binom{h}{i} \left(1 - \frac{(i-t)}{j}\right)^{h-1} \right)^k dt. \tag{2.16}$$

From our definition of $b_{h,k}$ (see (1.8)) and the fact that by symmetry it suffices to sum n' up to $hN/2$, summing over j completes the proof. □

Lemma 2.5. *We have*

$$\mathbb{E}(X_k) \sim \xi_{0,k}(N) p(N)^{-hk\delta} = \frac{b_{h,k} c^{hk}}{(s!d!)^k} N^{(h-1)k+1-hk\delta}, \tag{2.17}$$

with $b_{h,k}$ defined in (1.8).

Proof. By Lemma 2.4, it suffices to show $\mathbb{E}(X_k) \sim \xi_{0,k}(N) p(N)^{-hk\delta}$. To show that we can estimate $\mathbb{E}(X_k)$ by $\xi_{0,k}(N) p(N)^{hk}$, it suffices to prove that for each $\ell > 0$ we have

$$\xi_{\ell,k}(N) p(N)^{hk-\ell} = o(\xi_{0,k}(N) p(N)^{hk}). \tag{2.18}$$

From (2.6), for $\ell > 0$ the probability of being Type ℓ is $p(N)^{hk-\ell}$ and the number of such k-tuples is $\xi_{\ell,k}(N)$. The repeated elements can either be in the same $h_{(s,d)}$-tuple or in different $h_{(s,d)}$-tuples. In both cases we have the same order, though. We have k sets of h-tuples. That would give us hk independent variables; however, each of the h-tuples must sum to N (so we lose k degrees of freedom), and then we lose another ℓ by assumption (if $\ell = 0$ we have no repeated elements, which is the main term). Thus for a fixed n the number of solutions is at most on the order of $N^{hk-k-\ell}$; summing over n gives at most order N, for a total contribution of at most order $N^{(h-1)k-\ell+1}$.

We now multiply by the probability $p(N)^{hk-\ell}$ and get

$$\xi_{\ell,k}(N) p(N)^{hk-\ell} = O\left(N^{(h-1)k-\ell+1-(hk-\ell)\delta}\right). \tag{2.19}$$

Because $\delta < 1$ and $\ell > 0$, we know that

$$\xi_{\ell,k}(N) p(N)^{hk-\ell} = O\left(N^{(h-1)k+1-hk\delta-\ell(1-\delta)}\right) = O\left(\xi_{0,k}(N) N^{-\delta hk} \cdot N^{-\ell(1-\delta)}\right), \tag{2.20}$$

so the probability of choosing k-tuples with ℓ repeats is of a lower order than the probability of choosing a k-tuple with no repeats, completing the proof. □

2.3. **Strong Concentration Results.** We need to show X_k is strongly concentrated about its expected value as $N \to \infty$ to conclude that the actual number of distinct elements in the generalized sumset approaches the expectation. We know from Lemma 2.5 that the *expected* number of distinct elements is of a higher order than the *expected* number of repeated elements, but if we do not know that the *actual* number of distinct elements is close to its expectation, then Lemma 2.5 is of little use. Here we show that the actual number does indeed approach its mean. This is similar to equations (2.9) and (2.10) of [HM].

Lemma 2.6. *For $\delta \geq \frac{h-1}{h}$, X_k becomes strongly concentrated about its expected value as $N \to \infty$.*

Proof. We employ a second moment method to show that $N^{-\left(\frac{(h-1)k+1}{kh}\right)} = o(p(N))$ implies X_k is highly concentrated about its mean.

Let $\triangle = \sum_{a \sim b} P(Y_\alpha \cap Y_\beta)$ where $\alpha \sim \beta$ if k-tuples α, β have at least one number in common, and Y_α is an indicator variable for each unordered k-tuple having a constant sum. As $N \to \infty$, from Lemma 2.5 we know that $\mathbb{E}(X_k) \to \infty$ so, as in equation (2.9) of [HM], it suffices to show that

$$\triangle \;=\; o(\mathbb{E}(X_k)^2) \;=\; o_k\left(\left(N^{2(h-1)k+2}\right)\left(c^{2hk}N^{-2hk\delta}\right)\right). \tag{2.21}$$

The main contribution is from pairs with hk distinct elements and exactly 1 element in common. From Proposition 2.4, we have $O(N^{(h-1)k+1})$ choices for α. There are hk choices for common element with β, $O(N^{(h-1)k})$ choices for the rest of β, and $2kh-1$ elements in $\alpha \cup \beta$, so

$$P(Y_\alpha \cap Y_\beta) \;=\; O(p(N)^{2kh-1}). \tag{2.22}$$

Generalizing equation (2.10) in [HM],

$$\begin{aligned} \triangle &= \sum_{a \sim b} P(Y_\alpha \cap Y_\beta) \\ &= \sum_{a \sim b} p(N)^{2kh-1} \\ &= O(N^{(h-1)k+1+(h-1)k}) c^{2kh-1} N^{-(2kh-1)\delta}) \\ &= O(N^{2k(h-1)+1-(2kh-1)\delta}). \end{aligned} \tag{2.23}$$

Because $\delta < 1$,

$$\triangle \;=\; o_k(N^{2k(h-1)+2-2kh\delta}). \tag{2.24}$$

which proves our lemma. □

3. Phase Transition

3.1. **Fast Decay.** Here we prove the first claim of Theorem 1.2. We can do this using Chebyshev's inequality and Lemmas 2.5 and 2.6. This is equations 2.11-2.12 of [HM]. For $\delta > \frac{h-1}{h}$,

$$\begin{aligned} X_1 &\sim E(X_1) \sim \left(b_{h,1}N^h\right)\left(c^h N^{-h\delta}\right) \\ X_2 &\sim \left(b_{h,2}N^{2(h-1)+1}\right)\left(c^{2h}N^{-2h\delta}\right). \end{aligned} \tag{3.1}$$

We get the above equations from plugging $k = 1, 2$ into Lemma 2.5. Because $\delta > \frac{h-1}{h}$ and $X_1 = \Theta(N^{h-h\delta})$,

$$X_2 \;=\; O(N^{2(h-1)+1-2h\delta}) \;=\; O(X_1) + O(N^{-h+1+h\delta}). \tag{3.2}$$

The error term is lower order, so as $N \to \infty$, all but a vanishing proportion of $h_{(s,d)}$-tuples will generate distinct sums.

3.2. **Critical Decay.** We are now ready to prove the second claim in Theorem 1.2. The key result in our earlier approximation of X_k is that when we plug in $\delta = \frac{h-1}{h}$, all the exponents on N sum to 1, so we are left with a term on the order of N.

We first claim

$$\left| |A_{s,d}| - \sum_{k=1}^{m} (-1)^{k-1} X_k \right| \;\leq\; X_m. \tag{3.3}$$

We omit the proof because of its similarity to [HM] (see equation (2.16) there).

We now want to show

$$|A_{s,d}| \;\sim\; \sum_{k=1}^{m} (-1)^{k-1} X_k. \tag{3.4}$$

To do this, we need to show that the coefficients on X_m go to 0 as $k \to \infty$. The proof of this is a rote bound; we omit the details. By our concentration result in Lemma 2.6,

$$X_m \;\sim\; E(X_m) \;\sim\; b_{h,k} N^{(h-1)k+1-hk\delta} c^{hk} \;\sim\; b_{h,k} c^{hk} N. \tag{3.5}$$

Therefore, because $\delta = \frac{h-1}{h}$,

$$X_m \;\sim\; b_{h,k} c^{hk} N. \tag{3.6}$$

Following equation (2.18) of [HM] and using equation (3.3):

$$|A_{s,d}| \;\sim\; \sum_{k=1}^{m} (-1)^{k-1} X_k \;\sim\; N \sum_{k=1}^{m} (-1)^{k-1} b_{h,k} c^{hk}. \tag{3.7}$$

We conclude that

$$S_d^s \;\sim\; N g(c; s, d). \tag{3.8}$$

We define $g(c; s, d)$ to capture the N-dependency of the size of our generalized sumset $A_{s,d,}$. Unlike in the case of $A + A$ versus $A - A$, this function no longer has a nice closed form. The function we have defined arises in the generalization of Hegarty-Miller's random variables, and the purpose of $g(c; s, d)$ is to identify and pull out the N to determine how the size of the generalized sumset depends on N.

We want to compare the sizes of two sets A_{s_1,d_1} and A_{s_2,d_2} for $s_1 + d_1 = s_2 + d_2 = h$. The k, h, N factors are all the same and cancel, so $|A_{s_1,d_1}|/|A_{s_2,d_2}|$ depends only on s_1, s_2, d_1, d_2. Therefore,

$$\frac{|A_{s_1,d_1}|}{|A_{s_2,d_2}|} \;=\; \frac{s_2! d_2!}{s_1! d_1!}. \tag{3.9}$$

Because $d \leq s$, the maximum value of $1/(s!d!)$ is achieved at the minimum value of $s!d!$, which occurs when $s = d$. Thus, we conclude that as $N \to \infty$, with a probability of choosing elements decaying in N, the set with the most minus signs is almost surely larger. This proves the second claim of Theorem 1.2.

3.3. **Future Work: Slow Decay.** We are left with the case when $\delta < \frac{h-1}{h}$. This was done in the third case of Theorem 1.1 in [HM] for two summands, but in the general case of slow decay it is considerably more difficult for a number of reasons. The crucial difference in the analysis of the case of critical decay and the case of slow decay is that the case of slow decay focuses on the number of elements missing from $A_{s,d}$, while the case of critical decay focuses on the number of elements present in $A_{s,d}$. In the previous sections, we approximated $|A_{s,d}|$ by focusing on the middle of the interval $[-dN, sN]$ because it was here that elements were most likely to be present. However, to measure the number of sums missing from $A_{s,d}$, we instead need to look at the fringes of the interval, so the analysis shifts completely. Following [HM], we would need to estimate the expectation of the number of elements missing from the generalized sumset. In [HM], they let $\mathscr{E}_n$ denote the event that $n \notin A + A$. They can then find the expected number of missing sums,

$$\mathbb{E}[\mathscr{S}^c] \;=\; \sum_{n=0}^{2N} \mathbb{P}(\mathscr{E}_n); \tag{3.10}$$

however, to find $P(\mathscr{E}_n)$, they use that all ways of representing any integer n are independent of one another. This leads to the following nice equation in [HM]:

$$\mathbb{P}(\mathscr{E}_n) \;=\; \begin{cases} (1-p^2)^{n/2}(1-p) & \text{if } n \text{ is even} \\ (1-p^2)^{(n+1)/2} & \text{if } n \text{ is odd.} \end{cases} \tag{3.11}$$

In the general case, this formula is significantly less tractable as now the various ways to summing to n all depend on one another. The probability must be conditioned on each previous element chosen to be in the $h_{s,d}$-tuple, and that is the major difficulty in finding this formula in the general case. In the next equation of [HM], they sum over the probabilities of each n in the interval:

$$\mathbb{E}[\mathscr{S}^c] \;\sim\; 4 \cdot \sum_{m=0}^{\lfloor N/2 \rfloor} (1-p^2)^m \;\sim\; \frac{4}{p^2}; \tag{3.12}$$

however, in the $h = 2$ case, this summation takes advantage of nice geometric series properties which are not have available in the general case, and are thus left for future work.

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*, Random Structures and Algorithms **35** (2009), no. 1, 118–136.

[ILMZ] G. Iyer, O. Lazarev, S. J. Miller and L. Zhang, *Generalized More Sums Than Differences Sets*, Journal of Number Theory **132** (2012), no. 5, 1054–1073.

[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.

[LMO] O. Lazarev, S. J. Miller and K. O'Bryant, *Distribution of Missing Sums in Sumsets*, to appear in Experimental Mathematics. `http://arxiv.org/abs/1109.4700`.

[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* Additive Number Theory: Festschrift In Honor of the Sixtieth Birthday of Melvyn B. Nathanson, David Chudnovsky and Gregory Chudnovsky (Editors), Springer-Verlag, 2010. 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.

[Zh] Y. Zhao, *Sets Characterized by Missing Sums and Differences*, Journal of Number Theory **131** (2011), 2107–2134.

E-mail address: ginny6@stanford.edu

Department of Mathematics, Stanford University, Stanford, CA 94305

E-mail address: sjm1@williams.edu, Steven.Miller.MC.96@aya.yale.edu

Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267