Explicit constructions of infinite families of MSTD sets

Steven J. Miller and Daniel Scheinerman

Abstract Given a finite set of integers A, we may consider its sumset $A+A$ and its difference set $A-A$. As addition is commutative and subtraction is not, it was initially believed that as $r \to \infty$ almost all of the 2^r subsets of $\{1,\dots,r\}$ would have $|A-A| > |A+A|$; if $|A+A| > |A-A|$ we say A is an MSTD (more sums than differences) set. While Martin and O'Bryant [MO06] disproved this conjecture by showing that a small but positive percentage of such sets are MSTD, previous explicit constructions only found families of size $f(r)2^{r/2}$ for some polynomial $f(r)$.

Below we present a new construction that yields a family of MSTD sets in $\{1,\dots,r\}$ of size $C2^r/r^4$ for a fixed, non-zero constant C; thus our family is significantly denser than previous constructions. Our method has been generalized further with Brooke Orosz to handle certain ternary combinations; the details below are adapted from our paper [MOS09].

We conclude with an appendix on a special case of a result of Hegarty and Miller [HM07] which supports the intuition behind the false conjecture. Specifically, if $p(r)$ is a monotonically decreasing function tending to 0, and for each r every element in $\{1,\dots,r\}$ is in a subset A with probability $p(r)$, then as $r \to \infty$ almost no subsets (with respect to this probability) are MSTD.

Keywords: more sum than difference sets.
MSC 2010: 11P99.

Steven J. Miller
Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267 e-mail: Steven.J.Miller@williams.edu

Daniel Scheinerman
Department of Mathematics, Brown University, Providence, RI 02912 e-mail: Daniel_Scheinerman@brown.edu

1 Introduction

Given a finite set of integers A, we define its sumset $A+A$ and difference set $A-A$ by

$$\begin{aligned} A+A &= \{a_i + a_j : a_i, a_j \in A\} \\ A-A &= \{a_i - a_j : a_i, a_j \in A\}, \end{aligned} \tag{1}$$

and let $|X|$ denote the cardinality of X. If $|A+A| > |A-A|$, then, following Nathanson, we call A an MSTD (more sums than differences) set. As addition is commutative while subtraction is not, we expect that for a 'generic' set A we have $|A-A| > |A+A|$, as a typical pair (x,y) contributes one sum and two differences; thus we expect MSTD sets to be rare.

Martin and O'Bryant [MO06] proved that, in some sense, this intuition is wrong. They considered the uniform model[1] for choosing a subset A of $\{1,\dots,n\}$, and showed that there is a positive probability that a random subset A is an MSTD set (though, not surprisingly, the probability is quite small). However, the answer is very different for other ways of choosing subsets randomly, and if we decrease slightly the probability an element is chosen then our intuition is correct. Specifically, consider the binomial model with parameter $p(n)$, with $\lim_{n\to\infty} p(n) = 0$ and $n^{-1} = o(p(n))$ (so $p(n)$ doesn't tend to zero so rapidly that the sets are too sparse).[2] Hegarty and Miller [HM07] recently proved that, in the limit as $n \to 0$, the percentage of subsets of $\{1,\dots,n\}$ that are MSTD sets tends to zero in this model. See Appendix 2 for full statements and a self-contained proof when $p(n) = o(n^{-1/2})$.

Though MSTD sets are rare, they do exist (and, in the uniform model, are somewhat abundant by the work of Martin and O'Bryant). Examples go back to the 1960s. Conway is said to have discovered $\{0,2,3,4,7,11,12,14\}$, while Marica [Ma69] gave $\{0,1,2,4,7,8,12,14,15\}$ in 1969 and Freiman and Pigarev [FP73] found $\{0,1,2,4,5, 9,12,13, 14,16,17, 21,24,25,26,28,29\}$ in 1973. Recent work includes infinite families constructed by Hegarty [He07] and Nathanson [Na07], as well as existence proofs by Ruzsa [Ru76, Ru84, Ru92].

Most of the previous constructions[3] of infinite families of MSTD sets start with a symmetric set which is then 'perturbed' slightly through the careful addition of a few elements that increase the number of sums more than the number of differences; see [He07, Na07] for a description of some previous constructions and methods. In

[1] This means each of the 2^n subsets of $\{1,\dots,n\}$ are equally likely to be chosen, or, equivalently, that the probability any $k \in \{1,\dots,n\}$ is in A is just $1/2$.

[2] This model means that the probability $k \in \{1,\dots,n\}$ is in A is $p(n)$.

[3] An alternate method constructs an infinite family from a given MSTD set A by considering $A_t = \{\sum_{i=1}^{t} a_i m^{i-1} : a_i \in A\}$. For m sufficiently large, these will be MSTD sets; this is called the base expansion method. Note, however, that these will be very sparse. See [He07] for more details.

many cases, these symmetric sets are arithmetic progressions; such sets are natural starting points because if A is an arithmetic progression, then $|A+A| = |A-A|$.[4]

In this work we present a new method which takes an MSTD set satisfying certain conditions and constructs an infinite family of MSTD sets. While these families are not dense enough to prove a positive percentage of subsets of $\{1,\dots,r\}$ are MSTD sets, we are able to elementarily show that the percentage is at least C/r^4 for some constant C. Thus our families are far denser than those in [He07, Na07]; trivial counting[5] shows all of their infinite families give at most $f(r)2^{r/2}$ of the subsets of $\{1,\dots,r\}$ (for some polynomial $f(r)$) are MSTD sets, implying a percentage of at most $f(r)/2^{r/2}$.

We first introduce some notation:

- We let $[a,b]$ denote all integers from a to b; thus $[a,b] = \{n \in \mathbf{Z} : a \le n \le b\}$.
- We say a set of integers A has the property P_n (or is a P_n-set) if both its sumset and its difference set contain all but the first and last n possible elements (and of course it may or may not contain some of these fringe elements).[6] Explicitly, let $a = \min A$ and $b = \max A$. Then A is a P_n-set if

$$[2a+n,\ 2b-n] \ \subset\ A+A \tag{2}$$

and

$$[-(b-a)+n,\ (b-a)-n] \ \subset\ A-A. \tag{3}$$

We can now state our construction and main result.

[4] As $|A+A|$ and $|A-A|$ are not changed by mapping each $x \in A$ to $\alpha x + \beta$ for any fixed α and β, we may assume our arithmetic progression is just $\{0,\dots,n\}$, and thus the cardinality of each set is $2n+1$.

[5] For example, consider the following construction of MSTD sets from [Na07]: let $m,d,k \in \mathbf{N}$ with $m \ge 4$, $1 \le d \le m-1$, $d \ne m/2$, $k \ge 3$ if $d < m/2$ else $k \ge 4$. Set $B = [0,m-1]\backslash\{d\}$, $L = \{m-d, 2m-d, \dots, km-d\}$, $a^* = (k+1)m - 2d$ and $A = B \cup L \cup (a^* - B) \cup \{m\}$. Then A is an MSTD set. The width of such a set is of the order km. Thus, if we look at all triples (m,d,k) with $km \le r$ satisfying the above conditions, these generate on the order of at most $\sum_{k \le r} \sum_{m \le r/k} \sum_{d \le m} 1 \ll r^2$, and there are of the order 2^r possible subsets of $\{0,\dots,r\}$; thus this construction generates a negligible number of MSTD sets. Though we write $f(r)/2^{r/2}$ to bound the percentage from other methods, a more careful analysis shows it is significantly less; we prefer this easier bound as it is already significantly less than our method. See for example Theorem 2 of [He07] for a denser example.

[6] It is not hard to show that for fixed $0 < \alpha \le 1$ a random set drawn from $[1,n]$ in the uniform model is a $P_{\lfloor \alpha n \rfloor}$-set with probability approaching 1 as $n \to \infty$.

Theorem 1. *Let $A = L \cup R$ be a P_n, MSTD set where $L \subset [1,n]$, $R \subset [n+1,2n]$, and $1,2n \in A$;*[7] *see Remark 2 for an example of such an A. Fix a $k \geq n$ and let m be arbitrary. Let M be any subset of $[n+k+1, n+k+m]$ with the property that it does not have a run of more than k missing elements (i.e., for all $\ell \in [n+k+1, n+m+1]$ there is a $j \in [\ell, \ell+k-1]$ such that $j \in M$). Assume further that $n+k+1 \notin M$ and set $A(M;k) = L \cup O_1 \cup M \cup O_2 \cup R'$, where $O_1 = [n+1, n+k]$, $O_2 = [n+k+m+1, n+2k+m]$ (thus the O_i's are just sets of k consecutive integers), and $R' = R + 2k + m$. Then*

(1) $A(M;k)$ is an MSTD set, and thus we obtain an infinite family of distinct MSTD sets as M varies;

(2) there is a constant $C > 0$ such that as $r \to \infty$ the percentage of subsets of $\{1, \dots, r\}$ that are in this family (and thus are MSTD sets) is at least C/r^4.

Remark 1. We quickly highlight the main idea of the construction, referring to §2 for details. The idea is to take an MSTD set A and augment it to a new set A' such that the number of sums added ($|A'+A'| - |A+A|$) equals the number of differences added ($|A'-A'| - |A-A|$). This is accomplished by having the two blocks O_1, O_2 of consecutive elements and then making sure that we always take at least one out of every k elements between O_1 and O_2. Counting arguments then show that every possible new difference and new sum is included.

Remark 2. In order to show that our theorem is not trivial, we must of course exhibit at least one P_n, MSTD set A satisfying all our requirements (else our family is empty!). We may take the set[8] $A = \{1,2,3,5,8,9,13,15,16\}$; it is an MSTD set as

$$\begin{aligned} A+A &= \{2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,\\ &\qquad 22,23,24,25,26,28,29,30,31,32\} \\ A-A &= \{-15,-14,-13,-12,-11,-10,-8,-7,-6,-5,-4,-3,-2,-1,\\ &\qquad 0,1,2,3,4,5,6,7,8,10,11,12,13,14,15\} \end{aligned} \tag{4}$$

(so $|A+A| = 30 > 29 = |A-A|$). A is also a P_n-set, as (2) is satisfied since $[10,24] \subset A+A$ and (3) is satisfied since $[-7,7] \subset A-A$.

For the uniform model, a subset of $[1,2n]$ is a P_n-set with high probability as $n \to \infty$, and thus examples of this nature are plentiful. For example, of the 1748 MSTD sets with minimum 1 and maximum 24, 1008 are P_n-sets.

Unlike other estimates on the percentage of MSTD sets, our arguments are not probabilistic, and rely on explicitly constructing large families of MSTD sets. Our arguments share some similarities with the methods in [He07] (see for example Case

[7] Requiring $1,2n \in A$ is quite mild; we do this so that we know the first and last elements of A.

[8] This A is trivially modified from [Ma69] by adding 1 to each element, as we start our sets with 1 while other authors start with 0. We chose this set as our example as it has several additional nice properties that were needed in earlier versions of our construction which required us to assume slightly more about A.

I of Theorem 8) and [MO06]. There the fringe elements of the set were also chosen first. A random set was then added in the middle, and the authors argued that with high probability the resulting set is an MSTD set. We can almost add a random set in the middle; the reason we do not obtain a positive percentage is that we have the restriction that there can be no consecutive block of size k of numbers in the middle that are not chosen to be in $A(M;k)$. This is easily satisfied by requiring us to choose at least one number in consecutive blocks of size $k/2$, and this is what leads to the loss of a positive percentage (though we do obtain sets that are known to be MSTD sets, and not just highly likely to be MSTD sets).

The paper is organized as follows. We describe our construction in §2, and prove our claimed lower bounds for the percentage of sets that are MSTD sets in §3. We end with some concluding remarks and suggestions for future research in §4.

On a personal note, the first named author would like to thank Mel for introducing him to much of additive number theory, ranging from his accessible books to numerous conversations over the years. This paper (as well as the paper by Hegarty and Miller [HM07]) is an outgrowth of a talk Mel gave at Brown in 2007 on MSTD sets as well as conversations at CANT 2007, and it is a pleasure to thank him for an introduction to such a fascinating subject.

2 Construction of infinite families of MSTD sets

Let $A \subset [1,2n]$. We can write this set as $A = L \cup R$ where $L \subset [1,n]$ and $R \subset [n+1,2n]$. We have

$$A+A \ = \ [L+L] \cup [L+R] \cup [R+R] \tag{5}$$

where $L+L \subset [2,2n]$, $L+R \subset [n+2,3n]$ and $R+R \subset [2n+2,4n]$, and

$$A-A \ = \ [L-R] \cup [L-L] \cup [R-R] \cup [R-L] \tag{6}$$

where $L-R \subset [-1,-2n+1]$, $L-L \subset [-(n-1),n-1]$, $R-R \subset [-(n-1),n-1]$ and $R-L \subset [1,2n-1]$.

A typical subset A of $\{1,\dots,2n\}$ (chosen from the uniform model, see Footnote 1) will be a P_n-set (see Footnote 6). It is thus the interaction of the "fringe" elements that largely determines whether a given set is an MSTD set. Our construction begins with a set A that is both an MSTD set and a P_n-set. We construct a family of P_n, MSTD sets by inserting elements into the middle in such a way that the new set is a P_n-set, and the number of added sums is equal to the number of added differences. Thus the new set is also an MSTD set.

In creating MSTD sets, it is very useful to know that we have a P_n-set. The reason is that we have all but the "fringe" possible sums and differences, and are thus reduced to studying the extreme sums and differences. The following lemma shows that if A is a P_n, MSTD set and a certain extension of A is a P_n-set, then this extension is also an MSTD set. The difficult step in our construction is determining a large class of extensions which lead to P_n-sets; we will do this in Lemma 2.

Lemma 1. *Let $A = L \cup R$ be a P_n-set where $L \subset [1,n]$ and $R \subset [n+1,2n]$. Form $A' = L \cup M \cup R'$ where $M \subset [n+1,n+m]$ and $R' = R+m$. If A' is a P_n-set then $|A'+A'| - |A+A| = |A'-A'| - |A-A| = 2m$ (i.e., the number of added sums is equal to the number of added differences). In particular, if A is an MSTD set then so is A'.*

Proof. We first count the number of added sums. In the interval $[2,n+1]$ both $A+A$ and $A'+A'$ are identical, as any sum can come only from terms in $L+L$. Similarly, we can pair the sums of $A+A$ in the region $[3n+1,4n]$ with the sums of $A'+A'$ in the region $[3n+2m+1,4n+2m]$, as these can come only from $R+R$ and $(R+m)+(R+m)$ respectively. Since we have accounted for the n smallest and largest terms in both $A+A$ and $A'+A'$, and as both are P_n-sets, the number of added sums is just $(3n+2m+1)-(3n+1) = 2m$.

Similarly, differences in the interval $[1-2n,-n]$ that come from $L-R$ can be paired with the corresponding terms from $L-(R+m)$, and differences in the interval $[n,2n-1]$ from $R-L$ can be paired with differences coming from $(R+m)-L$. Thus the size of the middle grows from the interval $[-n+1,n-1]$ to the interval $[-n-m+1,n+m-1]$. Thus we have added $(2n+2m+3)-(2n+3) = 2m$ sums. Thus $|A'+A'| - |A+A| = |A'-A'| - |A-A| = 2m$ as desired.

The above lemma is not surprising, as in it we assume A' is a P_n-set; the difficulty in our construction is showing that our new set $A(M;k)$ is also a P_n-set for suitably chosen M. This requirement forces us to introduce the sets O_i (which are blocks of k consecutive integers), as well as requiring M to have at least one of every k consecutive integers.

We are now ready to prove the first part of Theorem 1 by constructing an infinite family of distinct P_n, MSTD sets. We take a P_n, MSTD set and insert a set in such a way that it remains a P_n-set; thus by Lemma 1 we see that this new set is an MSTD set.

Lemma 2. *Let $A = L \cup R$ be a P_n-set where $L \subset [1,n]$, $R \subset [n+1,2n]$, and $1, 2n \in A$. Fix a $k \geq n$ and let m be arbitrary. Choose any $M \subset [n+k+1,n+k+m]$ with the property that M does not have a run of more than k missing elements, and form $A(M;k) = L \cup O_1 \cup M \cup O_2 \cup R'$ where $O_1 = [n+1,n+k]$, $O_2 = [n+k+m+1,n+2k+m]$, and $R' = R+2k+m$. Then $A(M;k)$ is a P_n-set.*

Proof. For notational convenience, denote $A(M;k)$ by A'. Note $A'+A' \subset [2,4n+4k+2m]$. We begin by showing that there are no missing sums from $n+2$ to $3n+4k+2m$; proving an analogous statement for $A'-A'$ shows A' is a P_n-set. By symmetry[9] we only have to show that there are no missing sums in $[n+2,2n+2k+m]$. We consider various ranges in turn.

[9] Apply the arguments below to the set $2n+2k+m-A'$, noting that $1, 2n+2k+m \in A'$.

We observe that $[n+2,n+k+1] \subset A'+A'$ because we have $1 \in L$ and these sums result from $1+O_1$. Additionally, $O_1+O_1 = [2n+2,2n+2k] \subset A'+A'$. Since $n \leq k$ we have $n+k+1 \geq 2n+1$, these two regions are contiguous and overlap and thus $[n+2,2n+2k] \subset A'+A'$.

Now consider O_1+M. Since M does not have a run of more than k missing elements, the worst case scenario for us for elements in the sumset is that the smallest element of M is $n+2k$ and that the largest element is $n+m+1$ (and, of course, we still have at least one out of every k consecutive integers is in M). If this is the case then we still have $O_1+M \supset [(n+1)+(n+2k),(n+k)+(n+m+1)] = [2n+2k+1,2n+k+m+1]$. We had already shown that $A'+A'$ has all sums up to $2n+2k$; this extends the sumset to all sums up to $2n+k+m+1$.

All that remains is to show we have all sums in $[2n+k+m+2,2n+2k+m]$. This follows immediately from $O_1+O_2 = [2n+k+m+2,2n+3k+m] \subset A'+A'$. This extends our sumset to include all sums up to $2n+3k+m$, which is well past our halfway mark of $2n+2k+m$; the remaining sums follow from a similar argument. Thus we have shown that $A'+A' \supset [n+2,3n+4k+2m+1]$.

We now do a similar calculation for the difference set, which is contained in $[-(2n+2k+m)+1,(2n+2k+m)-1]$. As we have already analyzed the sumset, all that remains to prove A is a P_n-set is to show that $A'-A' \supset [-n-2k-m+1,n+2k+m-1]$. As all difference sets are symmetric about and contain 0, it suffices to show the positive elements are present, i.e., that $A'-A' \supset [1,n+2k+m-1]$.

We easily see $[1,k-1] \subset A'-A'$ as $[0,k-1] \subset O_1-O_1$. Now consider $M-O_1$. Again the worst case scenario for us is that the least element of M is $n+2k$ and the greatest is $n+m+1$. With this in mind we see that $M-O_1 \supset [(n+2k)-(n+k),(n+m+1)-(n+1)] = [k,m]$. Now $O_2-O_1 \supset [(n+k+m+1)-(n+k),(n+2k+m)-(n+1)] = [m+1,2k+m-1]$, and we therefore have all differences up to $2k+m-1$.

Since $2n \in A$ we have $2n+2k+m \in A'$. Consider $(2n+2k+m)-O_1 = [n+k+m,n+2k+m-1]$. Since $k \geq n$ we see that $n+k+m \leq 2k+m$; this implies that we have all differences up to $n+2k+m-1$ (this is because we already have all differences up to $2k+m-1$, and $n+k+m$ is either less than $2k+m-1$, or at most one larger).

Proof of Theorem 1(1). The proof of the first part of Theorem 1 follows immediately. By Lemma 2 our new sets $A(M;k)$ are P_n-sets, and by Lemma 1 they are also MSTD. All that remains is to show that the sets are distinct; this is done by requiring $n+k+1$ is not in our set (for a fixed k, these sets have elements $n+1,\dots,n+k$ but not $n+k+1$; thus different k yield distinct sets).

3 Lower bounds for the percentage of MSTDs

To finish the proof of Theorem 1, for a fixed n we need to count how many sets $\widetilde{M}$ of the form $O_1 \cup M \cup O_2$ (see Theorem 1 for a description of these sets) of width

$r = 2k+m$ can be inserted into a P_n, MSTD set A of width $2n$. As O_1 and O_2 are just intervals of k consecutive ones, the flexibility in choosing them comes solely from the freedom to choose their length k (so long as $k \geq n$). There is far more freedom to choose M.

There are two issues we must address. First, we must determine how many ways there are there to fill the elements of M such that there are no runs of k missing elements. Second, we must show that the sets generated by this method are distinct. We saw in the proof of Theorem 1(1) that the latter is easily handled by giving $A(M;k)$ (through our choice of M) slightly more structure. Assume that the element $n+k+1$ is *not* in M (and thus not in A). Then for a fixed width $r = 2k+m$ each value of k gives rise to necessarily distinct sets, since the set contains $[n+1, n+k]$ but not $n+k+1$. In our arguments below, we assume our initial P_n, MSTD set A is fixed; we could easily increase the number of generated MSTD sets by varying A over certain MSTD sets of size $2n$. We choose not to do this as n is fixed, and thus varying over such A will only change the percentages by a constant independent of k and m.

Fix n and let r tend to infinity. We count how many $\widetilde{M}$'s there are of width r such that in M there is at least one element chosen in any consecutive block of k integers. One way to ensure this is to divide M into consecutive, non-overlapping blocks of size $k/2$, and choose at least one element in each block. There are $2^{k/2}$ subsets of a block of size $k/2$, and all but one have at least one element. Thus there are $2^{k/2} - 1 = 2^{k/2}(1 - 2^{-k/2})$ valid choices for each block of size $k/2$. As the width of M is $r-2k$, there are $\lceil \frac{r-2k}{k/2} \rceil \leq \frac{r}{k/2} - 3$ blocks (the last block may have length less than $k/2$, in which case any configuration will suffice to ensure there is not a consecutive string of k omitted elements in M because there will be at least one element chosen in the previous block). We see that the number of valid M's of width $r-2k$ is at least $2^{r-2k}\left(1-2^{-k/2}\right)^{\frac{r}{k/2}-3}$. As O_1 and O_2 are two sets of k consecutive 1's, there is only one way to choose either.

We therefore see that, for a fixed k, of the $2^r = 2^{m+2k}$ possible subsets of r consecutive integers, we have at least $2^{r-2k}\left(1-2^{-k/2}\right)^{\frac{r}{k/2}-3}$ are permissible to insert into A. To ensure that all of the sets are distinct, we require $n+k+1 \not\in M$; the effect of this is to eliminate one degree of freedom in choosing an element in the first block of M, and this will only change the proportionality constants in the percentage calculation (and *not* the r or k dependencies). Thus if we vary k from n to $r/4$ (we could go a little higher, but once k is as large as a constant times r the number of generated sets of width r is negligible) we have at least some fixed constant times $2^r \sum_{k=n}^{r/4} \frac{1}{2^{2k}} \left(1-2^{-k/2}\right)^{\frac{r}{k/2}-3}$ MSTD sets; equivalently, the percentage of sets $O_1 \cup M \cup O_2$ with O_i of width $k \in \{n, \dots, r/4\}$ and M of width $r-2k$ that we may add is at least this divided by 2^r, or some universal constant times

$$\sum_{k=n}^{r/4} \frac{1}{2^{2k}} \left(1 - \frac{1}{2^{k/2}}\right)^{\frac{r}{k/2}} \tag{7}$$

(as $k \geq n$ and n is fixed, we may remove the -3 in the exponent by changing the universal constant).

We now determine the asymptotic behavior of this sum. More generally, we can consider sums of the form

$$S(a,b,c;r) \ = \ \sum_{k=n}^{r/4} \frac{1}{2^{ak}} \left(1 - \frac{1}{2^{bk}}\right)^{r/ck}. \tag{8}$$

For our purposes we take $a = 2$ and $b = c = 1/2$; we consider this more general sum so that any improvements in our method can readily be translated into improvements in counting MSTD sets. While we know (from the work of Martin and O'Bryant [MO06]) that a positive percentage of such subsets are MSTD sets, our analysis of this sum yields slightly weaker results. The approach in [MO06] is probabilistic, obtained by fixing the fringes of our subsets to ensure certain sums and differences are in (or not in) the sum- and difference sets. While our approach also fixes the fringes, we have far more possible fringe choices than in [MO06] (though we do not exploit this). While we cannot prove a positive percentage of subsets are MSTD sets, our arguments are far more elementary.

The proof of Theorem 1(2) is clearly reduced to proving the following lemma, and then setting $a = 2$ and $b = c = 1/2$.

Lemma 3. *Let*

$$S(a,b,c;r) \ = \ \sum_{k=n}^{r/4} \frac{1}{2^{ak}} \left(1 - \frac{1}{2^{bk}}\right)^{r/ck}. \tag{9}$$

Then for any $\varepsilon > 0$ we have

$$\frac{1}{r^{a/b}} \ \ll \ S(a,b,c;r) \ \ll \ \frac{(\log r)^{2a+\varepsilon}}{r^{a/b}}. \tag{10}$$

Proof. We constantly use $(1-1/x)^x$ is an increasing function in x. We first prove the lower bound. For $k \geq (\log_2 r)/b$ and r large, we have

$$\left(1 - \frac{1}{2^{bk}}\right)^{r/ck} \ = \ \left(1 - \frac{1}{2^{bk}}\right)^{2^{bk} \frac{r}{ck2^{bk}}} \ \geq \ \left(1 - \frac{1}{r}\right)^{r \cdot \frac{b}{c \log_2 r}} \ \geq \ \frac{1}{2} \tag{11}$$

(in fact, for r large the last bound is almost exactly 1). Thus we trivially have

$$S(a,b,c;r) \ \geq \ \sum_{k=(\log_2 r)/b}^{r/4} \frac{1}{2^{ak}} \cdot \frac{1}{2} \ \gg \ \frac{1}{r^{a/b}}. \tag{12}$$

For the upper bound, we divide the k-sum into two ranges: (1) $bn \leq bk \leq \log_2 r - \log_2 (\log r)^{\delta}$; (2) $\log_2 r - \log_2 (\log r)^{\delta} \leq bk \leq br/4$. In the first range, we have

$$\begin{aligned}\left(1-\frac{1}{2^{bk}}\right)^{r/ck} &\leq \left(1-\frac{(\log r)^{\delta}}{r}\right)^{r/ck} \\ &\ll \exp\left(-\frac{b(\log r)^{\delta}}{c\log_2 r}\right) \\ &\leq \exp\left(-\frac{b\log 2}{c}\cdot(\log r)^{\delta-1}\right). \qquad (13)\end{aligned}$$

If $\delta > 2$ then this factor is dominated by $r^{-\frac{b\log 2}{c}\cdot(\log r)^{\delta-2}} \ll r^{-A}$ for any A for r sufficiently large. Thus there is negligible contribution from k in range (1) if we take $\delta = 2+\varepsilon/a$ for any $\varepsilon > 0$.

For k in the second range, we trivially bound the factors $\left(1-1/2^{bk}\right)^{r/ck}$ by 1. We are left with

$$\sum_{k\geq\frac{\log_2 r}{b}-\frac{\log_2(\log r)^{\delta}}{b}} \frac{1}{2^{ak}}\cdot 1 \leq \frac{(\log r)^{a\delta}}{r^{a/b}}\sum_{\ell=0}^{\infty}\frac{1}{2^{a\ell}} \ll \frac{(\log r)^{a\delta}}{r^{a/b}}. \qquad (14)$$

Combining the bounds for the two ranges with $\delta = 2+\varepsilon/a$ completes the proof.

Remark 3. The upper and lower bounds in Lemma 3 are quite close, differing by a few powers of $\log r$. The true value will be at least $\left(\frac{\log r}{r}\right)^{a/b}$; we sketch the proof in Appendix 1.

Remark 4. We could attempt to increase our lower bound for the percentage of subsets that are MSTD sets by summing r from R_0 to R (as we have fixed r above, we are only counting MSTD sets of width $2n+r$ where 1 and $2n+r$ are in the set. Unfortunately, at best we can change the universal constant; our bound will still be of the order $1/R^4$. To see this, note the number of such MSTD sets is at least a constant times $\sum_{r=R_0}^{R} 2^r/r^4$ (to get the percentage, we divide this by 2^R). If $r \leq R/2$ then there are exponentially few sets. If $r \geq R/2$ then $r^{-4} \in [1/R^4, 16/R^4]$. Thus the percentage of such subsets is still only at least of order $1/R^4$.

4 Concluding remarks and future research

We observed earlier (Footnote 6) that for a constant $0 < \alpha \leq 1$, a set randomly chosen from $[1,2n]$ is a $P_{\lfloor \alpha n\rfloor}$-set with probability approaching 1 as $n \to \infty$. MSTD sets are of course not random, but it seems logical to suppose that this pattern continues.

Conjecture 1. Fix a constant $0 < \alpha \leq 1/2$. Then as $n \to \infty$ the probability that a randomly chosen MSTD set in $[1,2n]$ containing 1 and $2n$ is a $P_{\lfloor \alpha n\rfloor}$-set goes to 1.

In our construction and that of [MO06], a collection of MSTD sets is formed by fixing the fringe elements and letting the middle vary. The intuition behind both is that the fringe elements matter most and the middle elements least. Motivated by this it is interesting to look at all MSTD sets in $[1,n]$ and ask with what frequency a given element is in these sets. That is, what is

$$\gamma(k;n) \;=\; \frac{\#\{A : k \in A \text{ and } A \text{ is an MSTD set}\}}{\#\{A :\ A \text{ is an MSTD set}\}} \tag{15}$$

as $n \longrightarrow \infty$? We can get a sense of what these probabilities might be from Figure 1.

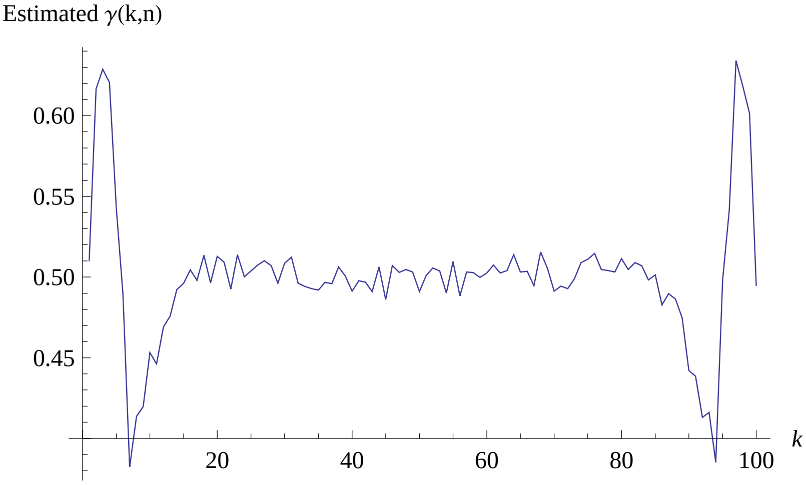

Fig. 1 Estimation of $\gamma(k,100)$ as k varies from 1 to 100 from a random sample of 4458 MSTD sets. The sample was obtained by choosing sets from the uniform model (i.e., for each $A \subset \{1,\dots,n\}$ the probability $k \in A$ is 1/2).

Note that, as the graph suggests, γ is symmetric about $\frac{n+1}{2}$, i.e. $\gamma(k,n) = \gamma(n+1-k,n)$. This follows from the fact that the cardinalities of the sumset and difference set are unaffected by sending $x \longrightarrow \alpha x + \beta$ for any α, β. Thus for each MSTD set A we get a distinct MSTD set $n+1-A$ showing that our function γ is symmetric. These sets are distinct since if $A = n+1-A$ then A is sum-difference balanced.[10]

From [MO06] we know that a positive percentage of sets are MSTD sets. By the central limit theorem we then get that the average size of an MSTD set chosen from $[1,n]$ is about $n/2$. This tells us that on average $\gamma(k,n)$ is about $1/2$. The graph above suggests that the frequency goes to $1/2$ in the center. This leads us to the following conjecture:

[10] The following proof is standard (see, for instance, [Na07]). If $A = n+1-A$ then

$$|A+A| \;=\; |A+(n+1-A)| \;=\; |n+1+(A-A)| \;=\; |A-A|. \tag{16}$$

Conjecture 2. Fix a constant $0 < \alpha < 1/2$. Then $\lim_{n\to\infty} \gamma(k,n) = 1/2$ for $\lfloor \alpha n \rfloor \leq k \leq n - \lfloor \alpha n \rfloor$.

Remark 5. More generally, we could ask which non-decreasing functions $f(n)$ have $f(n) \to \infty$, $n - f(n) \to \infty$ and $\lim_{n\to\infty} \gamma(k,n) = 1/2$ for all k such that $\lfloor f(n) \rfloor \leq k \leq n - \lfloor f(n) \rfloor$.

Appendix 1: Size of $S(a,b,c;r)$

We sketch the proof that the sum

$$S(a,b,c;r) \;=\; \sum_{k=n}^{r/4} \frac{1}{2^{ak}} \left(1 - \frac{1}{2^{bk}}\right)^{r/ck} \tag{17}$$

is at least $\left(\frac{\log r}{r}\right)^{a/b}$. We determine the maximum value of the summands

$$f(a,b,c;k,r) \;=\; \frac{1}{2^{ak}} \left(1 - \frac{1}{2^{bk}}\right)^{r/ck}. \tag{18}$$

Clearly $f(a,b,c;k,r)$ is very small if k is small due to the second factor; similarly it is small if k is large because of the first factor. Thus the maximum value of $f(a,b,c;k,r)$ will arise not from an endpoint but from a critical point.

It is convenient to change variables to simplify the differentiation. Let $u = 2^k$ (so $k = \log u / \log 2$). Then

$$g(a,b,c;u,r) \;=\; f(a,b,c;k,r) \;=\; u^{-a} \left(1 - \frac{1}{u^b}\right)^{u^b \cdot \frac{m \log 2}{c u^b \log u}}. \tag{19}$$

Thus

$$g(a,b,c;u,r) \;\approx\; u^{-a} \exp\left(-\frac{r \log 2}{c u^b \log u}\right). \tag{20}$$

Maximizing this is the same as minimizing $h(a,b,c;u,r) = 1/g(a,b,c;u,r)$. After some algebra we find

$$h'(a,b,c;u,r) \;=\; \frac{h(a,b,c;u,r)}{c u \log^2 u} \Big(a c u^b \log^2 u - r \log 2 \cdot (b \log u + 1)\Big). \tag{21}$$

Setting the derivative equal to zero yields

$$a c u^b \log^2 u \;=\; r \log 2 \cdot (b \log u + 1). \tag{22}$$

As we know u must be large, looking at just the main term from the right hand side yields

$$acu^b \log u \approx rb \log 2, \tag{23}$$

or

$$u^b \log u \approx Cr, \quad C = \frac{b \log 2}{ac}. \tag{24}$$

To first order, we see the solution is

$$u_{\max} = \left(\frac{(Cr)}{\frac{\log(Cr)}{b}} \right)^{\frac{1}{b}} \approx C' \left(\frac{r}{\log r} \right)^{\frac{1}{b}}. \tag{25}$$

Straightforward algebra shows that the maximum value of our summands is approximately $(C'e^{1/b})^{-a} \left(\frac{\log r}{r} \right)^{a/b}$.

Appendix 2: When almost all sets are not MSTD sets

Peter Hegarty and Steven J. Miller

In [Na06], Nathanson remarked: *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* $|A-A| > |A+A|$. While we now know (thanks to the work of Martin and O'Bryant [MO06]) that a positive percentage of all subsets of $\{1,\dots,N\}$ are MSTD sets, the answer is markedly different when we consider instead a binomial model with parameter decreasing to zero as $N \to \infty$. In [HM07] it is shown that Nathanson's intuition is correct for such a model.

Theorem 2. *Let* $p : \mathbf{N} \to (0,1)$ *be any function such that*

$$N^{-1} = o(p(N)) \quad \text{and} \quad p(N) = o(1). \tag{26}$$

For each $N \in \mathbf{N}$ *let A be a random subset of* $\{1,\dots,N\}$ *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*

$$\mathscr{S} \sim \frac{(N \cdot p(N))^2}{2} \quad \text{and} \quad \mathscr{D} \sim 2\mathscr{S} \sim (N \cdot p(N))^2. \tag{27}$$

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

$$g(x) \; := \; 2\left(\frac{e^{-x} - (1-x)}{x}\right). \tag{28}$$

Then

$$\mathscr{S} \; \sim \; g\left(\frac{c^2}{2}\right)N \;\; \text{and} \;\; \mathscr{D} \; \sim \; g(c^2)N. \tag{29}$$

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

$$\mathscr{S}^c \; \sim \; 2 \cdot \mathscr{D}^c \; \sim \; \frac{4}{p(N)^2}. \tag{30}$$

Parts (i) and (ii) of the theorem can be proven by elementary means; a standard second moment analysis (Chebyshev's inequality applied to a sum of indicator random variables) suffices to prove strong concentration of the variables $\mathscr{S}$ and $\mathscr{D}$, while in part (ii) an additional inclusion-exclusion type argument is used to obtain the correct form of the function g. Our proof of part (iii) requires different and more sophisticated concentration machinery recently developed by Kim and Vu [KV00, Vu00, Vu02]. For the benefit of the reader not familiar with probabilistic techniques, we present below an entirely self-contained proof of a more explicit form of the simplest case of our theorem, namely part (i). See [HM07] for proofs of the other cases, as well as generalizations to comparing arbitrary binary forms.[11]

We prove the following special case of Theorem 2.

Theorem 3. *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 $P_1(N) := \frac{4}{c}N^{-(1-\delta)}$ 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 $\delta \in (1/2,1)$.*

We first establish some notation, and then prove a sequence of lemmas from which Theorem 3 immediately follows. Our goal is to provide explicit bounds which decay like N to a power.

Let $I_N = \{1,\dots,N\}$ and let $X_{n;N}$ denote the binary indicator variable for n being in a subset (it is thus 1 with probability $cN^{-\delta}$ and 0 otherwise), and let X be the random variable denoting the cardinality of a subset (thus $X = \sum_n X_{n;N}$). For two pairs of

[11] Let u_1,u_2,v_1,v_2 be fixed integers, and define two binary forms $f(x,y) = u_1x + v_1y$ and $g(x,y) = u_2x + v_2y$. By $f(A)$ we mean $\{f(a_1,a_2) : a_i \in A\}$ (and similarly for $g(A)$). Theorem 2 can be generalized to analyze how often $|f(A)| > |g(A)|$ when A is drawn from $\{1,\dots,N\}$ from a binomial model with parameter $p(N)$.

ordered elements (m,n) and (m',n') in $I_N \times I_N$ ($m < n$, $m' < n'$), let $Y_{m,n,m',n'} = 1$ if $n - m = n' - m'$, and 0 otherwise.

Lemma 4. *With probability at least* $1 - P_1(N)$*,*

$$X \in \left[\frac{1}{2}\, cN^{1-\delta},\ \frac{3}{2}\, cN^{1-\delta}\right]. \tag{31}$$

Let $\mathscr{O}$ denote the number of ordered pairs (m,n) (with $m < n$) in a subset of I_N chosen with respect to the binomial model. Then with probability at least[12] $1 - P_1(N)$ *we have*

$$\frac{\frac{1}{2}cN^{1-\delta}\left(\frac{1}{2}cN^{1-\delta} - 1\right)}{2} \leq \mathscr{O} \leq \frac{\frac{3}{2}cN^{1-\delta}\left(\frac{3}{2}cN^{1-\delta} - 1\right)}{2}. \tag{32}$$

Proof. We have $\mathbf{E}[X] = \sum_n \mathbf{E}[X_{n;N}] = cN^{1-\delta}$. As the $X_{n;N}$ are independent,

$$\sigma_X^2 = \sum_n \sigma_{X_{n;N}}^2 = N\left(cN^{-\delta} - c^2N^{-2\delta}\right). \tag{33}$$

Thus

$$\sigma_X \leq \sqrt{c} \cdot N^{\frac{1-\delta}{2}}. \tag{34}$$

By Chebyshev's inequality,

$$\text{Prob}(|X - cN^{1-\delta}| \leq k\sigma_X) \geq 1 - \frac{1}{k^2}. \tag{35}$$

For $X \in \left[\frac{1}{2}\, cN^{1-\delta},\ \frac{3}{2}\, cN^{1-\delta}\right]$ we choose k so that

$$k\sigma_X = \frac{1}{2}\, cN^{1-\delta} \leq k\sqrt{c}N^{\frac{1-\delta}{2}}. \tag{36}$$

Thus $k \geq \frac{1}{2}\, \sqrt{c}N^{(1-\delta)/2}$, and the probability that X lies in the stated interval is at least $1 - (cN^{1-\delta}/4)^{-1}$. The second claim follows from the fact that there are $\binom{r}{2}$ ways to choose two distinct objects from r objects.

Proof of Theorem 3. By Lemma 4, (32) holds with probability at least $1 - P_1(N)$. The main contribution to the cardinalities of the sumset and the difference set is from ordered pairs (m,n) with $m < n$. With probability at least $1 - P_1(N)$ there are on the order $N^{2-2\delta}$ such pairs, which is much larger than the order $N^{1-\delta}$ pairs with $m = n$. The proof is completed by showing that almost all of the ordered pairs yield distinct sums (and differences). Explicitly, we shall show that for a subset chosen from I_N with respect to the binomial model, if $\mathscr{O}$ is the number of ordered pairs (which is of size $N^{2-2\delta}$ with high probability), then with high probability the

[12] By using the Central Limit Theorem instead of Chebyshev's inequality we could obtain a better estimate on the probability of X lying in the desired interval.

cardinality of its difference set is $2\mathscr{O} + O_C(N^{3-4\delta})$ while the cardinality of its sumset is $\mathscr{O} + O_C(N^{3-4\delta})$. This argument crucially uses $\delta > 1/2$ (if $\delta = 1/2$ then the error term is the same size as the main term, and a more delicate argument is needed). We shall show that almost all of the ordered pairs generate distinct differences; the argument for the sums follows similarly.

Each ordered pair (m,n) yields two differences: $m-n$ and $n-m$. The problem is that two different ordered pairs could generate the same differences. To calculate the size of the difference set, we need to control how often two different pairs give the same differences. Consider two distinct ordered pairs (m,n) and (m',n') with $m<n$ and $m'<n'$ (as the $N^{1-\delta} \ll N^{2-2\delta}$ 'diagonal' pairs (n,n) yield the same difference, namely 0, it suffices to study the case of ordered pairs with distinct elements). Without loss of generality we may assume $m \le m'$. If $n-m = n'-m'$ then these two pairs contribute the same differences. There are two possibilities: (1) all four indices are distinct; (2) $n = m'$.

We calculate the expected number of pairs of non-diagonal ordered pairs with the same difference by using our binary indicator random variables $Y_{m,n,m',n'}$. Set

$$Y \;=\; \sum_{1\le m\le m'\le N} \; \sum_{m'<n'\le N} \; \sum_{\substack{m<n\le N \\ n'-m'=n-m}} Y_{m,n,m',n'}. \tag{37}$$

If the four indices are distinct then $\mathbf{E}[Y_{m,n,m',n'}] = c^4 N^{-4\delta}$; if $n=m'$ then $\mathbf{E}[Y_{m,n,m',n'}] = c^3 N^{-3\delta}$.

The number of tuples (m,n,m',n') of distinct integers satisfying our conditions is bounded by N^3 (once m, n and m' are chosen there is at most one choice for $n' \in \{m+1,\dots,N\}$ with $n'-m' = n-m$)[13]. If instead $n=m'$ then there are at most N^2 tuples satisfying our conditions (once m and n are chosen, m' and n' are uniquely determined, though they may not satisfy our conditions). Therefore

$$\mathbf{E}[Y] \;\le\; N^3 \cdot c^4 N^{-4\delta} + N^2 \cdot c^2 N^{-3\delta} \;\le\; 2C^4 N^{3-4\delta} \tag{38}$$

as $\delta \in (1/2, 1)$.

As $N^{3-4\delta}$ is much smaller than $N^{2-2\delta}$ for $\delta > 1/2$, most of the differences are distinct. To complete the proof, we need some control on the variance of Y. In Lemma 5 we show that

$$\sigma_Y \;\le\; 7C^4 N^{r(\delta)}, \tag{39}$$

where

$$2r(\delta) \;=\; \max\{3-4\delta, 5-7\delta\}. \tag{40}$$

While we cannot use the Central Limit Theorem (the $Y_{m,n,m',n'}$ are not independent), we may use Chebyshev's inequality to bound the probability that Y is close to its

[13] Although we do not need the actual value, simple algebra yields the number of tuples is $N^3/6 + O(N^2)$.

mean (recall the mean is at most $2C^4N^{3-4\delta}$). We have

$$\text{Prob}(|Y - \mathbf{E}[Y]| \ \leq \ k\sigma_Y) \geq 1 - \frac{1}{k^2}. \tag{41}$$

Simple algebra shows that if we take $k = N^{2-2\delta-r(\delta)-g(\delta)}$ then with probability at least $1 - N^{-(f(\delta)-g(\delta))}$ we have $Y \leq 9C^4N^{2-2\delta-g(\delta)}$, which is a positive power of N less than $N^{2-2\delta}$. Thus an at most negligible amount of the differences are repeated.

The argument for two ordered pairs yielding the same sum proceeds similarly: if $\mu + \nu = \mu' + \nu'$ then $\nu - \mu' = \nu' - \mu$.

For our ratio to be $2 + O_C(N^{-g(\delta)})$, two events must happen. As the probability the first does not occur is at most $P_1(N)$ and the probability the second does not occur is at most $P_2(N)$, the probability that the two desired events happen is at least $1 - P_1(N) - P_2(N)$.

Except for the claimed estimate on σ_Y, the above completes the proof of Theorem 3. We now prove our bound for σ_Y.

Lemma 5. *Let the notation be as in Theorem 3 and (A.10). We have*

$$\sigma_Y \ \leq \ 7C^4N^{r(\delta)}. \tag{42}$$

Proof. If U and V are two random variables, then

$$\text{Var}(U+V) \ = \ \text{Var}(U) + \text{Var}(V) + 2\text{CoVar}(U,V). \tag{43}$$

By the Cauchy-Schwartz inequality, $\text{CoVar}(U,V) \leq \sqrt{\text{Var}(U)\text{Var}(V)}$. Thus

$$\text{Var}(U+V) \ \leq \ 3\text{Var}(U) + 3\text{Var}(V). \tag{44}$$

We may therefore write

$$\sum Y_{m,n,m',n'} \ = \ \sum U_{m,n,m',n'} + \sum V_{m,n,n'} \ = \ U + V, \tag{45}$$

where in the U-sum all four indices are distinct (with $1 \leq m < m' \leq N$, $m < n \leq N$, $m' < n' \leq N$ and $n - m = n' - m'$) and in the V-sum all three indices are distinct (with $1 \leq m < n < n' \leq N$ and and $n - m = n' - n$). As $\text{Var}(Y) \leq 3\text{Var}(U) + 3\text{Var}(V)$, we are reduced to bounding the variances of U and V.

We first bound $\text{Var}(U)$. Standard algebra yields

$$\begin{aligned} \text{Var}(U) \ &= \ \text{Var}\left(\sum U_{m,n,m',n'}\right) \\ &= \ \sum \text{Var}(U_{m,n,m',n'}) + 2 \sum_{(m,n,m',n') \neq (\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}')} \text{CoVar}(U_{m,n,m',n'}, U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'}). \end{aligned} \tag{46}$$

As $\mathrm{Var}(U_{m,n,m',n'}) = c^4N^{-4\delta} - c^8N^{-8\delta}$ and there are at most N^3 ordered tuples (m,n,m',n') of distinct integers with $n-m = m'-n'$, the $\mathrm{Var}(U_{m,n,m',n'})$ term is bounded by $c^4N^{3-4\delta}$.

For the covariance piece, if all eight indices $(m,n,m',n',\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}')$ are distinct, then $U_{m,n,m',n'}$ and $U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'}$ are independent and thus the covariance is zero. There are four cases; in each case there are always at most N^3 choices for the tuple (m,n,m',n'), but often there will be significantly fewer choices for the tuple $(\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}')$. We only provide complete details for the first and third cases, as the other cases follow similarly.

- Seven distinct indices: There are at most N^2 choices for $(\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}')$. The covariance of each such term is bounded by $c^7N^{-7\delta}$. To see this, note

$$\begin{aligned}&\mathrm{CoVar}(U_{m,n,m',n'},U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'})\\ &\quad = \mathbf{E}[U_{m,n,m',n'}U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'}] - \mathbf{E}[U_{m,n,m',n'}]\mathbf{E}[U_{\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'}]. \qquad (47)\end{aligned}$$

  The product of the expected values is $c^8N^{-8\delta}$, while the expected value of the product is $c^7N^{-7\delta}$. Thus the covariances of these terms contribute at most $c^7N^{5-7\delta}$.
- Six distinct indices: The covariances of these terms contribute at most $c^6N^{4-6\delta}$.
- Five distinct indices: The covariances of these terms contribute at most $c^5N^{3-5\delta}$ (once three of the $\widetilde{m},\widetilde{n},\widetilde{m}',\widetilde{n}'$ have been determined, the fourth is uniquely determined; thus there are at most N^3 choices for the first tuple and at most 1 choice for the second).
- Four distinct indices: The covariances of these terms contribute at most $c^4N^{3-4\delta}$.

The N-dependence from the case of seven distinct indices is greater than the N-dependence of the other cases (except for the case of four distinct indices if $\delta > 2/3$). We also only increase the contributions if we replace c with $C = \max(c,1)$. We therefore find

$$\begin{aligned}\mathrm{Var}(U) &\le C^4N^{3-4\delta} + 2\left(C^7N^{5-7\delta} + C^6N^{4-6\delta} + C^5N^{3-5\delta} + C^4N^{3-4\delta}\right)\\ &= 3C^4N^{3-4\delta} + 6C^7N^{5-7\delta}. \qquad (48)\end{aligned}$$

Similarly we have

$$\begin{aligned}\mathrm{Var}(V) &= \mathrm{Var}(\textstyle\sum V_{m,n,n'})\\ &= \sum \mathrm{Var}(V_{m,n,n'}) + 2\sum_{(m,n,n')\neq(\widetilde{m},\widetilde{n},\widetilde{n}')} \mathrm{CoVar}(V_{m,n,n'},V_{\widetilde{m},\widetilde{n},\widetilde{n}'}). \qquad (49)\end{aligned}$$

The $\mathrm{Var}(V_{m,n,n'})$ piece is bounded by $N^2 \cdot c^3N^{-3\delta}$ (as there are at most N^2 tuples with $n'-n = n-m$). The covariance terms vanish if the six indices are distinct. A similar argument as before yields bounds of $c^5N^{3-5\delta}$ for five distinct indices, $c^4N^{2-4\delta}$ for four distinct indices, and $c^3N^{2-3\delta}$ for three distinct indices. The largest N-dependence is from the $c^3N^{2-3\delta}$ term (as $\delta > 1/2$). Arguing as before and replacing c with C yields

$$\operatorname{Var}(V) \ \le \ C^3N^{2-3\delta} + 2\cdot 3C^3N^{2-3\delta} \ \le \ 7C^3N^{2-3\delta}. \tag{50}$$

As $\delta < 1, 2-3\delta < 3-4\delta$. Therefore

$$\begin{aligned}\operatorname{Var}(Y) \ &\le \ 3\cdot\left(3C^4N^{3-4\delta} + 6C^7N^{5-7\delta}\right) + 3\cdot 7C^3N^{2-3\delta} \\ &\le \ 30C^4N^{3-4\delta} + 18C^7N^{5-7\delta} \ \le \ 49C^8N^{2r(\delta)},\end{aligned} \tag{51}$$

which yields

$$\sigma_Y \ \le \ 7C^4N^{r(\delta)}. \tag{52}$$

Remark 6. An extreme choice of g would be to choose $g(\delta) = \varepsilon$, for some small positive constant ε. Since $f(\delta) \ge 1/4$ for all $\delta \in (1/2,1)$, we then obtain a bound of $2+O_C(N^{-\varepsilon})$ for the ratio of the cardinality of the difference set to the sumset with probability $1-O_C(N^{-\min\{1-\delta,\frac{1}{4}-\varepsilon\}})$.

References

[FP73] Freiman, G.A. and Pigarev, V.P.: The relation between the invariants R and T. In: Number theoretic studies in the Markov spectrum and in the structural theory of set addition (Russian), pp. 172–174. Kalinin. Gos. Univ., Moscow (1973)

[He07] Hegarty, P.V.: Some explicit constructions of sets with more sums than differences. Acta Arithmetica **130**, no. 1, 61–77 (2007).

[HM07] Hegarty, P.V. and Miller, S.J.: When almost all sets are difference dominated. Random Structures and Algorithms **35**, no. 1, 118–136 (2009).

[KV00] Kim, J.H. and Vu, V.H.: Concentration of multivariate polynomials and its applications. Combinatorica **20** 417–434 (2000).

[Ma69] Marica, J.: On a conjecture of Conway. Canad. Math. Bull. **12**, 233–234 (1969).

[MO06] Martin, G. and O'Bryant, K.: Many sets have more sums than differences, in: Additive Combinatorics, in: CRM Proc. Lecture Notes, vol. 43, Amer. Math. Soc., Providence, RI, 2007, pp. 287305.

[MOS09] Miller, S.J., Orosz, B. and Scheinerman, D.: Explicit constructions of infinite families of MSTD sets. To appear in: Journal of Number Theory (2010), doi:10.1016/j.jnt.2009.09.003.

[Na06] Nathanson, M.B.: Problems in additive number theory, I, in: Additive Combinatorics, in: CRM Proc. Lecture Notes, vol. 43, Amer. Math. Soc., Providence, RI, 2007, pp. 263270.

[Na07] Nathanson, M.B.: Sets with more sums than differences. Integers : Electronic Journal of Combinatorial Number Theory **7**, Paper A5 (24pp) (2007).

[Ru76] Ruzsa, I.Z.: On the cardinality of $A+A$ and $A-A$. In: Combinatorics year (Keszthely, 1976), vol. 18, Coll. Math. Soc. J. Bolyai, North-Holland-Bolyai Tàrsulat, pp. 933–938 (1978).

[Ru84] Ruzsa, I.Z.: Sets of sums and differences. In: Séminaire de Théorie des Nombres de Paris 1982-1983, pp. 267–273. Birkhäuser, Boston (1984).

[Ru92] Ruzsa, I.Z.: On the number of sums and differences. Acta Math. Sci. Hungar. **59**, 439–447 (1992).

[Vu00] Vu, V.H.: New bounds on nearly perfect matchings of hypergraphs: Higher codegrees do help. Random Structures and Algorithms **17**, 29–63 (2000).

[Vu02] Vu, V.H.: Concentration of non-Lipschitz functions and Applications. Random Structures and Algorithms **20**, no. 3, 262-316 (2002).