Constructing MSTD Sets Using Bidirectional Ballot Sequences

Yufei Zhao

Department of Mathematics, Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, MA 02139, USA
yufeiz@mit.edu

November 14, 2009

Abstract

A more sums than differences (MSTD) set is a finite subset S of the integers such that $|S+S| > |S-S|$. We construct a new dense family of MSTD subsets of $\{0, 1, 2, \ldots, n-1\}$. Our construction gives $\Theta(2^n/n)$ MSTD sets, improving the previous best construction with $\Omega(2^n/n^4)$ MSTD sets by Miller, Orosz, and Scheinerman.

Keywords: MSTD, sumset, difference set, bidirectional ballot sequence
2010 Mathematics Subject Classification: 11P99, 05A16

1 Introduction

A more sums than differences (MSTD) set is a finite set S of integers with $|S+S| > |S-S|$, where the sum set $S+S$ and the difference set $S-S$ are defined as

$$S + S = \{s_1 + s_2 : s_1, s_2 \in S\}$$
$$S - S = \{s_1 - s_2 : s_1, s_2 \in S\}.$$

Since addition is commutative while subtraction is not, two distinct integers s_1 and s_2 generate one sum but two differences. This suggests that $S+S$ should "usually" be smaller than $S-S$. Thus we expect MSTD sets to be rare.

The first example of an MSTD was found by Conway in the 1960's: $\{0, 2, 3, 4, 7, 11, 12, 14\}$. The name MSTD was later given by Nathanson [8]. MSTD sets have recently become a popular research topic [1, 2, 5, 6, 7, 8, 16, 17]. For older papers see [3, 4, 9, 11, 12, 13, 14]. We refer the reader to [7, 8] for the history of the problem.

Let $\rho_{n-1}2^n$ be the number of MSTD subsets of $\{0, 1, 2, \ldots, n-1\}$. We refer to ρ_n informally as the *density* of the family of MSTD sets. This quantity was first studied by Martin and O'Bryant [5], who showed that $\rho_n \geq 2 \times 10^{-7}$ for $n \geq 14$. However, this bound is far from optimal. Recently, the author [17] showed that ρ_n converges to a limit, and computed a lower bound of 4×10^{-4} for this limit. From Monte Carlo experiments, we expect limiting density to be about 4.5×10^{-4} [5].

The proofs of the lower bounds on ρ_n are non-constructive. On the other hand, infinite families of MSTD sets were constructed by Hegarty [1], Nathanson [8], and Miller, Orosz, and Scheinerman [6]. In particular, Miller et al. gave the densest construction in terms of the number of subsets of $\{0, 1, \ldots, n-1\}$; their construction has density $\Omega(1/n^4)$.

In this paper, we offer a new construction of an infinite family of MSTD sets. Our construction, described in Section 2, has density $\Theta(1/n)$, improving the previous result of Miller et al. [6] In

Section 3 we prove that our family of MSTD sets has the claimed size. In the process we introduce a new combinatorial object called a bidirectional ballot sequence, whose additional properties are discussed in Section 4.

2 Construction of MSTD sets

We use $[a, b]$ to denote the set $\{a, a+1, \dots, b\}$. In this section we describe our construction of a new family of MSTD subsets of $[0, n-1]$.

The first idea used in our construction is similar to the techniques used in both [5] and [6]; namely we look for sets of the form

$$S = L \cup M \cup R,$$

where

$$\begin{aligned} L &= S \cap [0, \ell - 1], \\ M &= S \cap [\ell, n - r - 1], \\ R &= S \cap [n - r, n - 1]. \end{aligned}$$

We will fix L and R to be sets with certain desirable properties and let M vary.

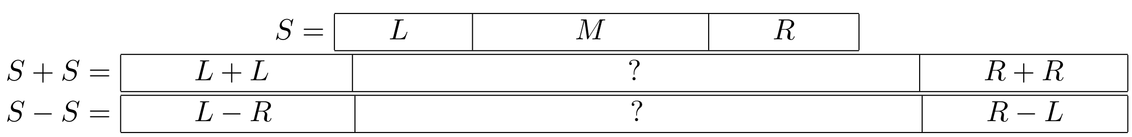

Figure 1: Illustration of the construction of S.

For instance, adapting the construction from [5] and taking $\ell = r = 11$ and

$$L = \{0, 2, 3, 7, 8, 9, 10\}, \tag{1}$$

$$R = \{n-11, n-10, n-9, n-8, n-6, n-3, n-2, n-1\}, \tag{2}$$

we have

$$L + L = [0, 20] \setminus \{1\}, \qquad R + R = [2n - 22, 2n - 2].$$

On the other hand, $S - S$ is missing at least two differences, namely $\pm(n-7)$, so $|S - S| \le 2n - 3$. If we can get $S + S$ to contain $[21, 2n - 23]$ (i.e., all the middle sums not yet covered by $L + L$ or $R + R$), then $S + S$ is only missing the sum 1, and thus $|S + S| = 2n - 2$, thereby making S an MSTD set.

So our goal is to choose M so that $S + S$ is not missing any sums in the middle segment, i.e., $[21, 2n - 23]$. From the probabilistic argument of [5], we know that the set of all M's with this property occupies a positive lower density of all subsets of $[11, n - 12]$. However, that proof is non-constructive.

Note that if $M + M$ is not missing any sums (i.e., $M + M = [2 \cdot 11, 2(n - 12)]$), then S has the desired properties. This condition forces $11, n - 12 \in M$, so that $21, 2n - 23 \in S + S$ as well. Let us temporarily do some re-indexing so that the problem becomes finding subsets M of $[1, m]$ such that $M + M = [2, 2m]$. Note that the probabilistic argument of [5] also shows that the set of such M's has at least positive constant density.

The construction of [6] is as follows: let M contain all k elements on each of its two ends (i.e., $[1, k] \cup [m - k + 1, n] \subset M$), and furthermore let M have the property that it does not have a run of more than k consecutive missing elements, which can be satisfied by dividing into blocks of size $k/2$ and choosing at least one element from each block. Here k is allowed to vary. This construction gives a density of $\Omega(1/n^4)$.

We use a different approach to construct M. The property of M that we seek is the following: for every prefix and suffix of $[1, m]$, more than half of the elements are in M. The following lemma proves that this constraint is sufficient for our purposes.

Lemma 2.1. *If $M \subset [1, m]$ satisfies*

$$|M \cap [1, k]| > \frac{k}{2}, \qquad \text{and} \qquad |M \cap [m - k + 1, m]| > \frac{k}{2}$$

for every $0 < k \leq m$, then $M + M = [2, 2m]$.

Proof. Let $2 \leq x \leq 2m$. If $x \leq m$, then since M contains more than half of the elements in $[1, x-1]$, by the pigeonhole principle, there is some y so that $y, x - y \in M$, so that $x \in M + M$. Similarly, if $x > m$, then since M contains more than half of the elements in $[x - m, m]$, we can find some $x - y, y \in M$ so that $m \in M + M$ as well. □

The construction of this new family of MSTD sets is summarized in the theorem below.

Theorem 2.2. *Let $n \geq 24$. Moreover, let M be a subset of $[11, n-12]$ with the property that every prefix and every suffix of the interval $[11, n - 12]$ has more than half of its elements in M. Then $S = L \cup M \cup R$ is an MSTD set, where L and R are given in* (1) *and* (2). *The number of MSTD sets of $\{0, 1, \ldots, n - 1\}$ in this family is $\Theta(2^n/n)$.*

To prove the last assertion in the theorem, we need to count the number of sets in our family. This is done in the next section.

Remark. It seems that our method can be extended to construct sets A satisfying $|A + A + A| > |A + A - A|$ in the spirit of [6, Sec. 4]. However, we do not pursue this direction here since no new ideas are introduced and the details can get somewhat messy.

3 Bidirectional ballot sequence

In order to study the sizes of our new families of MSTD sets, we introduce the following combinatorial construction.

Definition 3.1. A 0-1 sequence of length n is a *bidirectional ballot sequence* if every prefix and suffix contains strictly more 1's than 0's. The number of bidirectional ballot sequences of length n is denoted B_n.

Recall that a classical *ballot sequence* is a 0-1 sequence where we only require that every prefix has more 1's than 0's. The name comes from the interpretation where there is an election with two candidates and we want one candidate to always stay strictly ahead of the other candidate through the vote count. Then, a bidirectional ballot sequence is a ballot sequence whose reverse is also a ballot sequence. Table 1 gives some values of B_n. This construction appears to be new. The sequence B_n was previously not found on the Sloane On-Line Encyclopedia of Integer Sequences [15], so we added it as a new entry A167510.

It is easy to see that the possibilities for the set M in the construction in Theorem 2.2 correspond bijectively with bidirectional ballot sequences of length $n - 22$. Then, the proof of the final assertion in the theorem is equivalent to the following result about the number of bidirectional ballot sequences of a given length.

Table 1: Number of bidirectional ballot sequences of length n.

n	1	2	3	4	5	6	7	8	9	10	11	12
B_n	1	1	1	1	2	3	5	9	15	28	49	91
n	13	14	15	16	17	18	19	20	21	22	23	24
B_n	166	307	574	1065	2016	3769	7176	13532	25842	49113	93995	179775

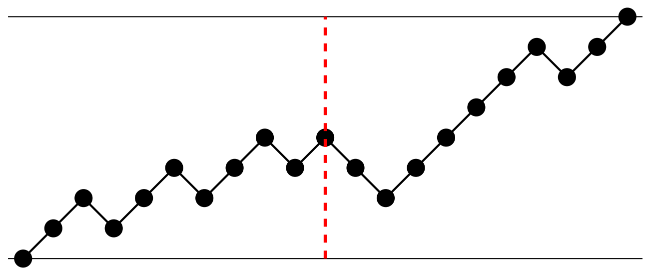

Figure 2: A bidirectional ballot walk corresponding to the sequence 11011011010011111011. The middle dashed line divides the walk into two halves.

Proposition 3.2. *The number of bidirectional ballot sequences satisfies* $B_n = \Theta\left(2^n/n\right)$.

This rest of this section contains a proof of Proposition 3.2.

We can interpret 0-1 sequences in terms of lattice walks, where we start at the origin and take steps of the form $(1,1)$ and $(1,-1)$, corresponding to the terms 1 and 0 in the sequence, respectively. Let a *ballot walk* (resp. *bidirectional ballot walk*) be such a lattice walk corresponding to a ballot sequence (resp. bidirectional ballot sequence). So, a ballot walk is a lattice walk with the property that the starting point is the unique lowest point, and a bidirectional ballot walk has the additional property that the ending point is the unique highest point. See Figure 2 for an example.

The key idea in the proof of Proposition 3.2 is to divide a bidirectional ballot walk into two halves, as in Figure 2. The second half should be "reversed," i.e., viewed with a 180° rotation. For the upper bound, we notice that each half is necessarily a ballot walk. For the lower bound, we need some sufficient condition on the two halves so that neither "overshoots" the other when the two halves are glued together.

Let us recall the following classic theorem about ballot sequences (e.g., see [10]).

Theorem 3.3 (Ballot Theorem)**.** *Let $p > q$. The number of ballot sequences with p 1's and q 0's, or equivalently the number of ballot walks with p steps of the form $(1,1)$ and q steps of the form $(1,-1)$, is equal to*

$$\binom{p+q-1}{p-1} - \binom{p+q-1}{p} = \frac{p-q}{p+q}\binom{p+q}{p}.$$

The Ballot Theorem can be proven by counting the number of "bad" walks, i.e., those that dip below the x-axis. This is usually done using a method is known as the *reflection principle*. As a special case, when $p = q+1$, we obtain the Catalan numbers $\frac{1}{2p+1}\binom{2p+1}{p} = \frac{1}{p+1}\binom{2p}{p}$.

The following Corollary of the Ballot Theorem will be useful in our proof of Proposition 3.2.

Corollary 3.4. *Let $0 \le a < b$ be real numbers. The number of ballot walks with n steps and whose final height is inclusively between a and b is*

$$\binom{n-1}{\left\lceil \frac{1}{2}(a+n)\right\rceil - 1} - \binom{n-1}{\left\lfloor \frac{1}{2}(b+n)\right\rfloor}.$$

Proof. We use the Ballot Theorem and sum over all (p, q) with $p + q = n$ and $a \leq 2p - n \leq b$ to find that the desired quantity is

$$\sum_{a \leq 2p-n \leq b} \left(\binom{n-1}{p-1} - \binom{n-1}{p} \right) = \binom{n-1}{\lceil \frac{1}{2}(a+n) \rceil - 1} - \binom{n-1}{\lfloor \frac{1}{2}(b+n) \rfloor}. \qquad \square$$

We will also use the following well-known fact about the normal approximation of binomial coefficients. It can be proved using either Stirling's formula or the Central Limit Theorem.

Proposition 3.5. *For any real number t,*

$$\lim_{n \to \infty} \frac{\sqrt{n}}{2^n} \binom{n}{\frac{1}{2}(n + t\sqrt{n})} = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}t^2}. \tag{3}$$

3.1 Upper Bound

Lemma 3.6. *The number of ballot walks with n steps is $\binom{n-1}{\lceil n/2 \rceil - 1} \sim \frac{2^n}{\sqrt{2\pi n}}$.*

Proof. This follows directly from Corollary 3.4 and Proposition 3.5. $\square$

Let $n_0 = \lfloor n/2 \rfloor$ and $n_1 = \lceil n/2 \rceil$. A bidirectional ballot walk is necessarily a ballot walk of length n_0 followed by the reverse of a ballot walk of length n_1. Therefore, the number of bidirectional ballot walks with n steps is at most

$$O\left(\frac{2^{n_0}}{\sqrt{n_0}}\right) O\left(\frac{2^{n_1}}{\sqrt{n_1}}\right) = O\left(\frac{2^n}{n}\right).$$

Thus we have proven the following upper bound on B_n.

Proposition 3.7. $B_n = O(2^n/n)$.

3.2 Lower Bound

We know that the first half and the reverse of the second half of a bidirectional ballot walk are both ballot walks, but this alone is not enough to guarantee that the overall walk is a bidirectional ballot walk. So we place additional constraints on each half of the walk.

Definition 3.8. Let b be a positive integer. A b-bounded walk is a ballot walk that never goes into the region $y > 2b$ and ends in the region $y > b$.

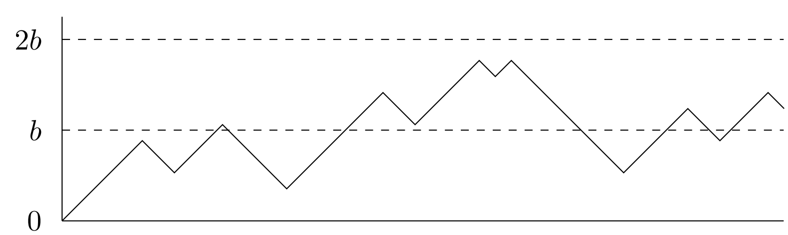

Figure 3: An example of a b-bounded walk.

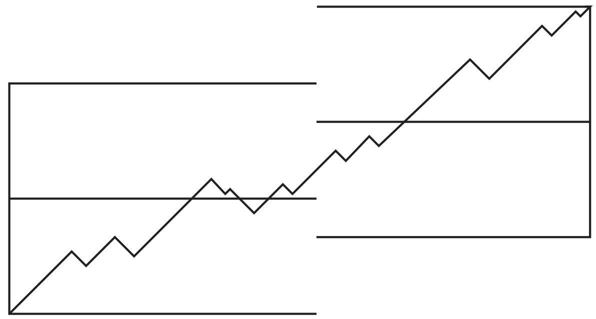

Figure 4: "Proof by picture" of Lemma 3.9.

Lemma 3.9. *The concatenation of a b-bounded walk followed by the reverse of another b-bounded walk is necessarily a bidirectional ballot walk.*

Figure 4 is a "proof by picture" of the lemma. The b-boundedness ensures that neither half overshoots the other.

Lemma 3.10. *The number of $\lfloor\sqrt{n}\rfloor$-bounded walks of n steps is $\Omega(2^n/\sqrt{n})$.*

Proof. We see that b-bounded walks of n steps are precisely ballot walks that end in the region $b+1 \leq y \leq 2b$ and never go into the region $y > 2b$. Using Corollary 3.4, we see that the number of ballot walks with n steps that end in $b+1 \leq y \leq 2b$ is equal to

$$\binom{n-1}{\lceil\frac{1}{2}(n+b-1)\rceil} - \binom{n-1}{\lfloor n/2\rfloor + b}.$$

Now we need to consider those ballot walks that end in $b < y \leq 2b$ but go into $y > 2b$ at some point in the walk. Let $(t, 2b+1)$ be the last point in walk that is in the region $y > 2b$. We can reflect the portion of the walk after that point to get a ballot walk that ends in $y > 2b+1$. See Figure 5 for an illustration. This map is injective since we can always get back to the original walk, but it is not necessarily onto. Then, we know that the number of ballot walks that end in $b < y \leq 2b$ but go into $y > 2b$ at some point is at most the number of ballot walks that end in $y \geq 2b+2$. By Corollary 3.4, the number of ballot walks that end in $y \geq 2b+2$ is equal to $\binom{n-1}{\lceil n/2\rceil + b}$.

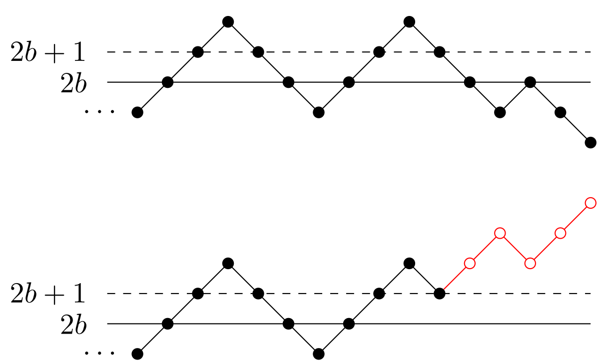

Figure 5: Reflecting the last segment of a walk.

Therefore, the number of b-bounded walks is at least

$$\binom{n-1}{\lceil\frac{1}{2}(n+b-1)\rceil} - \binom{n-1}{\lfloor n/2\rfloor + b} - \binom{n-1}{\lceil n/2\rceil + b}.$$

Let $b = \lfloor \sqrt{n} \rfloor$. Using Proposition 3.5, we have

$$\lim_{n\to\infty} \frac{\sqrt{n}}{2^n} \left(\binom{n-1}{\lceil \frac{1}{2}(n+b-1) \rceil} - \binom{n-1}{\lfloor n/2 \rfloor + b} - \binom{n-1}{\lceil n/2 \rceil + b} \right) = \frac{1}{\sqrt{2\pi}}(e^{-1/2} - 2e^{-2}) > 0.$$

It follows that the number of $\lfloor \sqrt{n} \rfloor$-bounded walks is $\Omega(2^n/\sqrt{n})$. □

As before, we can form bidirectional ballot walks by concatenating two b-bounded walks, where the second half is reversed. Let $n_0 = \lfloor n/2 \rfloor$ and $n_1 = \lceil n/2 \rceil$. Then, the number of bidirectional ballot walks is at least

$$\Omega\left(\frac{2^{n_0}}{\sqrt{n_0}}\right) \Omega\left(\frac{2^{n_1}}{\sqrt{n_1}}\right) = \Omega(2^n/n).$$

Thus we have proven the following.

Proposition 3.11. $B_n = \Omega(2^n/n)$.

Propositions 3.7 and 3.11 together complete the proof of Proposition 3.2 and hence also Theorem 2.2.

4 Further remarks

We believe that there is more potential to bidirectional ballot sequences than what is presented here. Knowing that $B_n = \Theta(2^n/n)$, we can ask whether the ratio $nB_n/2^n$ approaches a limit. Table 2 contains some values computed from an exact formula for B_n. The data suggest that $nB_n/2^{n-2} \to 1$. This is indeed true. We have a proof of this fact, but our proof is rather long and technical, so we do not present it here. The proof involves first finding an exact formula for B_n using repeated applications of the reflection principle, and then some analysis to estimate the sum. The data in Table 2 also suggest the asymptotic expansion

$$\frac{B_n}{2^n} = \frac{1}{4n} + \frac{1}{6n^2} + O\left(\frac{1}{n^3}\right),$$

which we pose as a conjecture.

Table 2: Some values of $nB_n/2^{n-2}$.

n	$nB_n/2^{n-2}$
100	1.0067268...
1000	1.00066729...
10000	1.0000666729...

Bidirectional ballot sequences look superficially similar to Dyck paths and Catalan numbers. However, the former lack the nice enumerative properties enjoyed by the latter two. There does not seem to be any simple recursive structure in bidirectional ballot sequences, and we were unable to find any useful recurrence relations or generating functions for B_n. This is what makes the enumeration of bidirectional ballot sequences particularly difficult.

We can interpret bidirectional ballot sequences in terms of random walks. Suppose we take a random walk of n steps in $\mathbb{Z}$ where each step independently moves one unit to the left or the right, each with 1/2 probability. Let p_n denote the probability that, among all the points visited by the walk, the starting point is minimum and the ending point is maximum. Then $p_n = B_{n+2}/2^n \sim 1/n$ as $n \to \infty$.

Were it the case that $p_n \sim c/n$ for any other constant c, then perhaps the result might be much less interesting[1]. However, as it stands, we feel that $p_n \sim 1/n$ is not merely a coincidence, and we believe that it deserves a better explanation then the calculation-heavy proof that we have. There should be some natural, combinatorial explanation, perhaps along the lines of grouping all possible walks into orbits of size mostly n under some symmetry, so that almost every orbit contains exactly one walk with the desired property. So far, we do not know of any such explanation.

We are also currently investigating higher dimensional analogues of this type of random walk problems. We have some experimental data that suggest the prevalence of the $1/n$ asymptotics for analogous walks in higher dimensions. We currently have no proof or explanation of this phenomenon.

The asymptotics related to bidirectional ballot sequences are very intriguing, and we hope to generate more interest in these objects.

Acknowledgments

This research was carried out at the University of Minnesota Duluth under the supervision of Joseph Gallian with the financial support of the National Science Foundation and the Department of Defense (grant number DMS 0754106), the National Security Agency (grant number H98230-06-1-0013), and the MIT Department of Mathematics. The author would like to thank Joseph Gallian for his encouragement and support. The author would also like to thank Nathan Kaplan and Ricky Liu for reading the paper and making valuable suggestions.

References

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

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

[3] F. Hennecart, G. Robert, and A. Yudin, *On the number of sums and differences*, Astérisque (1999), no. 258, xiii, 173–178, Structure theory of set addition.

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

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

[6] S. J. Miller, B. Orosz, and D. Scheinerman, *Explicit constructions of infinite families of MSTD sets*, to appear.

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

[8] ———, *Sets with more sums than differences*, Integers **7** (2007), A5, 24 pp. (electronic).

[1] Indeed, if we only require the starting point to be minimum, then it is easy to show that $p_n \sim \sqrt{\frac{2}{\pi n}}$; the constants here are not nearly as nice.

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

[10] M. Renault, *Four proofs of the ballot theorem*, Math. Mag. **80** (2007), no. 5, 345–352.

[11] F. Roesler, *A mean value density theorem of additive number theory*, Acta Arith. **96** (2000), no. 2, 121–138.

[12] I. Z. Ruzsa, *On the cardinality of $A + A$ and $A - A$*, Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II, Colloq. Math. Soc. János Bolyai, vol. 18, North-Holland, Amsterdam, 1978, pp. 933–938.

[13] ———, *Sets of sums and differences*, Seminar on number theory, Paris 1982–83 (Paris, 1982/1983), Progr. Math., vol. 51, Birkhäuser Boston, Boston, MA, 1984, pp. 267–273.

[14] ———, *On the number of sums and differences*, Acta Math. Hungar. **59** (1992), no. 3-4, 439–447.

[15] N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, published electronically at `www.research.att.com/~njas/sequences/`, 2009.

[16] Y. Zhao, *Counting MSTD sets in finite abelian groups*, preprint at `http://arxiv.org/abs/0911.2288`.

[17] ———, *Subsets characterized by the number of missing sums and differences*, preprint at `http://arxiv.org/abs/0911.2292`.