

# SUM AND DIFFERENCE SETS IN GENERALIZED DIHEDRAL GROUPS

RUBEN ASCOLI, JUSTIN CHEIGH, GUILHERME ZEUS DANTAS E MOURA, RYAN JEONG, ANDREW KEISLING, ASTRID LILLY, STEVEN J. MILLER, PRAKOD NGAMLAMAI, MATTHEW PHANG

**ABSTRACT.** Given a group  $G$ , we say that a set  $A \subseteq G$  has more sums than differences (MSTD) if  $|A + A| > |A - A|$ , has more differences than sums (MDTS) if  $|A + A| < |A - A|$ , or is sum-difference balanced if  $|A + A| = |A - A|$ . A problem of recent interest has been to understand the frequencies of these type of subsets.

The seventh author and Vissuet studied the problem for arbitrary finite groups  $G$  and proved that almost all subsets  $A \subseteq G$  are sum-difference balanced as  $|G| \rightarrow \infty$ . For the dihedral group  $D_{2n}$ , they conjectured that of the remaining sets, most are MSTD, i.e., there are more MSTD sets than MDTS sets. Some progress on this conjecture was made by Haviland et al. in 2020, when they introduced the idea of partitioning the subsets by size: if, for each  $m$ , there are more MSTD subsets of  $D_{2n}$  of size  $m$  than MDTS subsets of size  $m$ , then the conjecture follows.

We extend the conjecture to generalized dihedral groups  $D = \mathbb{Z}_2 \ltimes G$ , where  $G$  is an abelian group of size  $n$  and the nonidentity element of  $\mathbb{Z}_2$  acts by inversion. We make further progress on the conjecture by considering subsets with a fixed number of rotations and reflections. By bounding the expected number of overlapping sums, we show that the collection  $\mathcal{S}_{D,m}$  of subsets of the generalized dihedral group  $D$  of size  $m$  has more MSTD sets than MDTS sets when  $6 \leq m \leq c_j \sqrt{n}$  for  $c_j = 1.3229/\sqrt{111 + 5j}$ , where  $j$  is the number of elements in  $G$  with order at most 2. We also analyze the expectation for  $|A + A|$  and  $|A - A|$  for  $A \subseteq D_{2n}$ , proving an explicit formula for  $|A - A|$  when  $n$  is prime.

## 1. INTRODUCTION AND MAIN RESULTS

Given a set of  $A$  integers, the sumset and difference set of  $A$  are defined as

$$A + A = \{a_1 + a_2 : a_1, a_2 \in A\} \quad \text{and} \quad A - A = \{a_1 - a_2 : a_1, a_2 \in A\}. \quad (1)$$

These elementary operations are fundamental in additive number theory. A natural problem of recent interest has been to understand the relative sizes of the sum and difference sets of sets  $A$ .

**Definition 1.1.** We say that a set  $A$  has **more sums than differences (MSTD)** if  $|A + A| > |A - A|$ ; has **more differences than sums (MDTS)** if  $|A + A| < |A - A|$ ; or is **sum-difference balanced** if  $|A + A| = |A - A|$ .

We intuitively expect most sets to be MDTS since addition is commutative and subtraction is not. Nevertheless, MSTD subsets of integers exist. Nathanson detailed in [nathanson2007problems] the history of the problem, and attributed to John Conway the first recorded example of an MSTD subset of integers,  $\{0, 2, 3, 4, 7, 11, 12, 14\}$ . Martin and O’Bryant proved in [martin2006many] that the proportion of the  $2^n$  subsets  $A$  of  $\{0, 1, \dots, n - 1\}$  which are MSTD is bounded below by a positive value for all  $n \geq 15$ . They proved this by controlling the “fringe” elements of  $A$ , those close to 0 and  $n - 1$ , which have the most influence over whether elements are missing from the sum and difference sets. In [zhao2011characterized], Zhao gave a deterministic algorithm to compute the limit of the ratio of MSTD subsets of  $\{0, 1, \dots, n - 1\}$  as  $n$  goes to infinity and found that

---

*Date:* October 4, 2022.

*2020 Mathematics Subject Classification.* 11P99, 05B10.

*Key words and phrases.* More Sums Than Differences, Dihedral Group, Generalized Dihedral Group.

This work was supported by NSF grant DMS1947438, Williams College, and Harvey Mudd College.

this ratio is at least  $4.28 \times 10^{-4}$ . For more on the problem of MSTD sets in the integers, see also [hegarty2007explicit] and [nathanson2006sets] for constructive examples of infinite families of MSTD sets, [miller2010explicit] and [zhao2010ballot] for non-constructive proofs of existence of infinite families of MSTD sets, and [hegarty2009almost] and [hogan2013generalized] for an analysis of sets with each integer from 0 to  $n - 1$  included with probability  $cn^{-\delta}$ .

More recently, several authors have examined analogous problems for groups  $G$  other than the integers. For example, Do, Kulkarni, Moon, Wellens, Wilcox, and the seventh author studied in [do2015polytopes] the analogous problem for higher-dimensional integer lattices.

For finite groups, although the usual notation for the operation of the group is multiplication, we match the notation from previous work and define, for a subset  $A \subseteq G$ , its sumset and difference set as

$$A + A = \{a_1 a_2 : a_1, a_2 \in A\} \quad \text{and} \quad A - A = \{a_1 a_2^{-1} : a_1, a_2 \in A\}. \quad (2)$$

Definition 1.1 of MSTD, MDTs, and sum-difference balanced sets apply in this context.

The approaches used to study MSTD subsets of integers do not generalize for MSTD subsets of finite groups due to the lack of fringes. Zhao proved asymptotics for numbers of MSTD subsets of finite abelian groups as the size of the group goes to infinity in [zhao2010counting]. The seventh author and Vissuet examined the problem for arbitrary finite groups  $G$ , also with the size of the group going to infinity, and proved Theorem 1.2.

**Theorem 1.2** ([miller2014most]). *Let  $\{G_n\}$  be a sequence of finite groups, not necessarily abelian, with  $|G_n| \rightarrow \infty$ . Let  $A_n$  be a uniformly chosen random subset of  $G_n$ . Then  $\mathbb{P}[A_n + A_n = A_n - A_n = G_n] \rightarrow 1$  as  $n \rightarrow \infty$ . In other words, as the size of the finite groups increases without bound, almost all subsets are balanced (with sumset and difference set equalling the entire group).*

Furthermore, for the case of dihedral groups  $D_{2n}$ , they proposed Conjecture 1.3.

**Conjecture 1.3** ([miller2014most]). *Let  $n \geq 3$  be an integer. There are more MSTD subsets of  $D_{2n}$  than MDTs subsets of  $D_{2n}$ .*

Given a set  $A \subseteq D_{2n} = \langle r, s \mid r^n, s^2, rsrs \rangle$ , define  $R$  (resp.  $F$ ) as the set of elements of  $A$  of the form  $r^i$  (resp.  $r^i s$ ), called rotation elements (resp. flip elements). Hence,  $A = R \cup F$ . Then, we can write

$$\begin{aligned} A + A &= (R + R) \cup (F + F) \cup (R + F) \cup (-R + F), \\ A - A &= (R - R) \cup (F + F) \cup (R + F). \end{aligned} \quad (3)$$

Intuition for Conjecture 1.3 comes from noting that  $F + F$  and  $R + F$  contribute to both  $A + A$  and  $A - A$ ;  $R + R$  and  $-R + F$  contribute only to  $A + A$ ; and  $R - R$  contributes only to  $A - A$ .

In 2020, Haviland, Kim, L  m, Lentfer, Trejos Su  res, and the seventh author made progress towards Conjecture 1.3 by partitioning subsets of  $D_{2n}$  by size. They proposed Conjecture 1.4 as a means of proving Conjecture 1.3.

**Conjecture 1.4** ([haviland2020more]). *Let  $n \geq 3$  be an integer, and let  $\mathcal{S}_{2n,m}$  denote the collection of subsets of  $D_{2n}$  of size  $m$ . For any  $m \leq 2n$ ,  $\mathcal{S}_{2n,m}$  has at least as many MSTD sets as MDTs sets.*

They showed that Conjecture 1.4 holds for  $m = 2$ , which we reproduce in this paper, and we also extend their approach to  $m = 3$ . They also showed that Conjecture 1.4 holds for  $m > n$  by showing that all sets in  $\mathcal{S}_{2n,m}$  are sum-difference balanced. We prove this result in this paper, using Lemma 1.5. These results are proved in Section 2.

**Lemma 1.5.** *Let  $n \geq 3$  be an integer, and let  $A \subseteq D_{2n}$ . Let  $R$  (resp.  $F$ ) be the subset of rotations (resp. flips) in  $A$ . Suppose that  $|F| > \frac{n}{2}$  or  $|R| > \frac{n}{2}$ . Then,  $A$  cannot be MDTs.*

Furthermore, we extend Conjecture 1.3 as follows. A **generalized dihedral group** is given by  $D = \mathbb{Z}_2 \rtimes G$ , where  $G$  is any abelian group and where the nonidentity element of  $\mathbb{Z}_2$  acts on  $G$  by inversion.

**Conjecture 1.6.** *Let  $G$  be an abelian group with at least one element of order 3 or greater, and let  $D = \mathbb{Z}_2 \rtimes G$  be the corresponding generalized dihedral group. Then, there are more MSTD subsets of  $D$  than MDTs subsets of  $D$ .*

Conjecture 1.3 is a special case of Conjecture 1.6, with  $G = \mathbb{Z}_n$ . We also state Conjecture 1.7, analogous to Conjecture 1.4.

**Conjecture 1.7.** *Let  $D$  be a generalized dihedral group of size  $2n$ , and let  $\mathcal{S}_{D,m}$  denote the collection of subsets of  $D$  of size  $m$ . For any  $m \leq 2n$ ,  $\mathcal{S}_{D,m}$  has at least as many MSTD sets as MDTs sets.*

The version of Lemma 1.5 that we prove in Section 2 deals with the generalized dihedral group.

In Section 3, we prove our main theorem, verifying Conjecture 1.7 for the case of  $m \leq c_j \sqrt{n}$ , where  $c_j$  is a constant (independent of  $n$ ) depending only on the quantity  $j$ , the number of elements of order at most 2 in the abelian group  $G$ . More explicitly, we show the following.

**Theorem 1.8.** *Let  $D = \mathbb{Z}_2 \rtimes G$  be a generalized dihedral group of size  $2n$ . Let  $\mathcal{S}_{D,m}$  denote the collection of subsets of  $D$  of size  $m$ , and let  $j$  denote the number of elements in  $G$  with order at most 2. If  $6 \leq m \leq c_j \sqrt{n}$ , where  $c_j = 1.3229/\sqrt{111 + 5j}$ , then there are more MSTD sets than MDTs sets in  $\mathcal{S}_{D,m}$ .*

See Section 3 the proof of this result and two related theorems. We also extend these results to the dihedral group on finitely generated abelian groups in Section 3.3.

Next, in Section 4, we discuss the following result about the expected size of  $|A - A|$  when  $A$  is a randomly chosen set in  $\mathcal{S}_{2n,m}$ , the collection of subsets of  $D_{2n}$  of size  $m$ .

**Theorem 1.9.** *If  $n$  is prime, and  $A$  is chosen uniformly at random from  $\mathcal{S}_{2n,m}$ , then*

$$\mathbb{E}[|A - A|] = 2n - \frac{nm2^m \binom{n}{m} + 2n(n-1) \binom{n-m-1}{m-1}}{m \binom{2n}{m}} - \frac{n^2(n-1)}{\binom{2n}{m}} \sum_{k=1}^{m-1} \frac{\binom{n+k-m-1}{m-k-1} \binom{n-k-1}{k-1}}{k(m-k)}. \quad (4)$$

Finally, in Section 5, we discuss directions for further research.

## 2. DIRECT ANALYSIS

**2.1. Small Subsets.** For the case of the usual dihedral group  $D_{2n}$ , we have the following two results.

**Lemma 2.1** ([haviland2020more]). *Let  $n \geq 3$ , and let  $\mathcal{S}_{2n,2}$  denote the collection of subsets of  $D_{2n}$  of size 2. Then,  $\mathcal{S}_{2n,2}$  has strictly more MSTD sets than MDTs sets.*

**Lemma 2.2.** *Let  $n \geq 3$ , and let  $\mathcal{S}_{2n,3}$  denote the collection of subsets of  $D_{2n}$  of size 3. Then,  $\mathcal{S}_{2n,3}$  has strictly more MSTD sets than MDTs sets.*

The proofs for both these lemmas use basic and somewhat tedious casework; they can be found in Appendix A.

Similar results for the generalized dihedral group likely follow from similar arguments.

**2.2. Large Subsets.** We consider what happens when  $m$  gets close to  $n$ . Here we can prove a result for any generalized dihedral group. For the rest of this section, let  $G$  be a finite abelian group of size  $n$ . Recall that the generalized dihedral group is given by  $D = \mathbb{Z}_2 \ltimes G$ , where the nonidentity element of  $\mathbb{Z}_2$  acts on  $G$  by inversion. Note that  $|D| = 2n$ . Writing an element of the group  $D$  as  $(z, g)$  where  $z \in \{0, 1\}$  and  $g \in G$ , we write  $R_D$  to mean the subset of  $D$  consisting of elements with  $z = 0$  and  $F_D$  to mean the subset of  $D$  consisting of elements with  $z = 1$ . Note that  $|R_D| = |F_D| = n$ . For the case where  $D = D_{2n} = \mathbb{Z}_2 \ltimes \mathbb{Z}_n$  is the usual dihedral group,  $R_D$  and  $F_D$  are the sets of rotations and flips, respectively; out of convenience, we will use these terms for the general case as well.

It turns out that having  $m > n$  ensures that  $A$  is balanced.

**Lemma 2.3.** *Let  $D$  be a generalized dihedral group of size  $2n$ . Let  $A \subseteq D$ , and let  $R = A \cap R_D$  and  $F = A \cap F_D$ . If  $\max(|R|, |F|) > n/2$ , then  $R_D \subseteq A + A$  and  $R_D \subseteq A - A$ .*

*Proof.* Let  $L$  be the larger of  $R$  and  $F$ , and define  $L_D = R_D$  if  $L = R$  and  $L_D = F_D$  if  $L = F$ .

For each rotation  $r \in R_D$ , define  $rL^{-1} = \{r\ell^{-1} \mid \ell \in L\}$  and  $rL = \{r\ell \mid \ell \in L\}$ . Note that  $|rL^{-1}| = |rL| = |L| > n/2$ , and  $rL^{-1}, rL, L$  are subsets of the set  $L_D$  which has size  $n$ . Hence, by the inclusion–exclusion principle,  $L \cap rL^{-1}$  and  $L \cap rL$  are nonempty. Thus,  $r \in L + L \subseteq A + A$  and  $r \in L - L \subseteq A - A$ .

Therefore, as desired,  $R_D \subseteq A + A$  and  $R_D \subseteq A - A$ .  $\square$

*Remark 1.* Note Lemma 2.3 implies if  $\max(|R|, |F|) > n/2$ , then  $A$  is not MDTs. This follows from the discussion after the statement of Conjecture 1.3 (which we explicitly extend to the general dihedral group case in Section 3). For a set to be MDTs,  $R - R$  must contribute rotations to  $A - A$  that the set  $A + A$  does not have. But here we have shown that if  $\max(|R|, |F|) > n/2$ , then  $A + A$  has all the rotations.

**Lemma 2.4.** *Let  $D$  be a generalized dihedral group of size  $2n$ , and let  $A \subseteq D$  with  $|A| = m$ . If  $m > n$ , then  $A + A = A - A = D$ .*

*Proof.* Let  $R$  (resp.  $F$ ) be the subset of rotations (resp. flips) in  $A$ ; hence  $A = R \cup F$ . Let  $L, S$  be the larger and smaller of  $R$  and  $F$ , respectively. Define  $|L| = n_1, |S| = n_2$  for  $n_1 + n_2 = m > n$ . Thus,  $n_1 > n/2$ .

By Lemma 2.3, we have that  $R_D \subseteq A + A$  and  $R_D \subseteq A - A$ .

For each flip  $f \in F_D$ , define  $fL^{-1} = \{f\ell^{-1} \mid \ell \in L\}$  and  $fL = \{f\ell \mid \ell \in L\}$ .

Note that  $|fL^{-1}| = |fL| = |L| = n_1, |S| = n_2$ , and  $fL^{-1}, fL, S$  are subsets of  $S_D$ , which has size  $n < n_1 + n_2$ . Hence, by the inclusion–exclusion principle,  $fL^{-1} \cup S$  and  $fL \cup S$  are nonempty. Thus,  $f \in S + L \subseteq A + A$  and  $f \in S - L \subseteq A - A$ . Therefore,  $F_D \subseteq A + A$  and  $F_D \subseteq A - A$ .

Thus, we have  $R_D, F_D \subseteq A + A$  and  $R_D, F_D \subseteq A - A$ , which imply  $A + A = A - A = D$ .  $\square$

### 3. COLLISION ANALYSIS

Let  $G$  be a finite abelian group of size  $n$ , written multiplicatively. Recall that the generalized dihedral group is given by  $D = \mathbb{Z}_2 \ltimes G$ , where the nonidentity element of  $\mathbb{Z}_2$  acts on  $G$  by inversion.

This section is dedicated to proving the following.

**Theorem 1.8.** *Let  $D = \mathbb{Z}_2 \ltimes G$  be a generalized dihedral group of size  $2n$ . Let  $\mathcal{S}_{D,m}$  denote the collection of subsets of  $D$  of size  $m$ , and let  $j$  denote the number of elements in  $G$  with order at most 2. If  $6 \leq m \leq c_j \sqrt{n}$ , where  $c_j = 1.3229/\sqrt{111 + 5j}$ , then there are more MSTD sets than MDTs sets in  $\mathcal{S}_{D,m}$ .*

Note that the theorem is only useful when  $c_j \sqrt{n} \geq 6$ , or  $n \geq (6/c_j)^2$ .

If  $n$  is arbitrarily large and  $j$  is a constant compared to  $n$ , we can make a stronger statement: we can replace  $c_j$  in the above theorem with a constant arbitrarily close to  $\sqrt{2/7} \approx 0.5345$ . Specifically, we have the following.

**Theorem 3.1.** *For fixed  $j$  and  $\epsilon > 0$ , there exists  $n_{j,\epsilon}$  with the following property. Let  $G$  be an abelian group of size  $n \geq n_{j,\epsilon}$  with at most  $j$  elements of order 2 or 1. Then with  $D$  and  $\mathcal{S}_{D,m}$  defined as in Theorem 1.8, we have that if  $6 \leq m \leq (\sqrt{2/7} - \epsilon)\sqrt{n}$ , then there are more MSTD sets than MDTs sets in  $\mathcal{S}_{D,m}$ .*

We can give a stronger, more general statement on the proportion of MSTD sets in  $\mathcal{S}_{D,m}$  if  $m$  is large and also bounded above by a (smaller) constant times  $\sqrt{n}$ .

**Theorem 3.2.** *Let  $D, j$ , and  $\mathcal{S}_{D,m}$  be defined as in Theorem 1.8. For any  $\epsilon > 0$ , there exist  $m_\epsilon$  and  $c_{\epsilon,j}$  such that if  $m_\epsilon \leq m \leq c_{\epsilon,j}\sqrt{n}$ , the proportion of MSTD sets in  $\mathcal{S}_{D,m}$  is at least  $1 - \epsilon$ .*

Here  $m_\epsilon$  and  $c_{\epsilon,j}$  are independent of  $n$ , but similarly to before, this theorem is only useful when  $n \geq (m_\epsilon/c_{\epsilon,j})^2$ .

*Remark 2.* In practice, Theorems 1.8 and 3.2 are most useful if  $j$  is essentially a constant compared to  $n$ . This is indeed the case for the original dihedral group  $D_{2n} = \mathbb{Z}_2 \ltimes \mathbb{Z}_n$ , where  $j = 1$  when  $n$  is odd and  $j = 2$  when  $n$  is even, yielding  $c_j \geq 0.12$ . The family of original dihedral groups is also a good example of how to use Theorem 3.1: we can apply that theorem with  $j = 2$  and  $\epsilon$  arbitrarily small to get that for large enough dihedral groups, we can get the coefficient of the  $\sqrt{n}$  in the theorem to be very close to  $\sqrt{2/7}$ , which is a significant improvement over 0.12.

However, these theorems are not useful when, for example,  $G = \mathbb{Z}_2 \times \dots \times \mathbb{Z}_2 \times \mathbb{Z}_3$ . Here  $G$  does have an element of order at least 3, so Conjecture 1.6 applies, but we have  $j = n/3$ , which is too large for Theorems 1.8 and 3.2 to apply to any values of  $m$ . In fact, if  $j \geq n/100$ , then  $n \leq (6/c_j)^2$ , and Theorem 1.8 does not apply to any values of  $m$ .

We first prove Theorem 1.8 here. Then, in Subsection 3.1 we demonstrate Theorems 3.1 and 3.2.

Recall the definitions of  $R_D$  and  $F_D$  from Section 2.2. Note that any element in  $F_D$  has order 2 in  $D$ . Furthermore, any element in  $R_D$  has order at most 2 in  $D$  if and only if it has order at most 2 in  $G$ .

We begin with a set  $A$  with size  $m$  and count the number of elements in  $A + A$  and  $A - A$ . In this count, we will make a naive assumption: there are no overlaps between sums and differences that we do not expect to overlap. Decompose  $A$  into the union of the set of rotations  $R = A \cap R_D$  of and the set of flips  $F = A \cap F_D$ , and define  $k = |F|$ . We have:

$$\begin{aligned} A + A &= (R + F) \cup (F + R) \cup (R + R) \cup (F + F); \\ A - A &= (R - F) \cup (F - R) \cup (R - R) \cup (F - F) \end{aligned} \tag{5}$$

Consider first the flips in  $A + A$  and  $A - A$ . In  $A + A$ , these are in  $R + F$  and  $F + R$ . Note that we do not expect a lot of overlap in general; for a rotation  $(0, g_1) \in R$  and a flip  $(1, g_2) \in F$ ,  $(0, g_1) \cdot (1, g_2) = (1, g_1 g_2)$  does not equal  $(1, g_2) \cdot (0, g_1) = (1, g_2 g_1^{-1})$  unless  $g_1$  has order 1 or 2 in  $G$ . On the other hand, for the flips in  $A - A$ , we have  $R - F = F - R$ . This is because  $(0, g_1) \cdot (1, g_2)^{-1} = (0, g_1) \cdot (1, g_2) = (1, g_1 g_2)$ , and  $(1, g_2) \cdot (0, g_1)^{-1} = (1, g_2) \cdot (0, g_1^{-1}) = (1, g_1 g_2)$ , so these two are the same. There are  $m - k$  rotations and  $k$  flips in  $A$ , so we thus expect the flips to contribute  $2(m - k)k$  to  $A + A$  but only  $(m - k)k$  to  $A - A$ .

Next, consider the rotations in  $A + A$  and  $A - A$ . Begin with  $F + F$  and  $F - F$ . Since all flips have order 2, these are in fact the same set and thus always contribute equally to  $A + A$  and to  $A - A$ . Also note that if  $k \neq 0$  (we will treat the  $k = 0$  case later), the identity 1 is contained in  $F + F$  and  $F - F$ .

Next consider  $R + R$  and  $R - R$ . Adding rotations is commutative, so we expect  $R + R$  to contribute  $\binom{m-k}{2} + (m-k)$  to the size of  $A + A$ , where the  $m-k$  term comes from the sum of each rotation in  $R$  with itself. On the other hand, in general  $g_1 g_2^{-1} \neq g_2 g_1^{-1}$ , so  $R - R$  is expected to contribute  $2\binom{m-k}{2}$ . Here there is no additional  $m-k$  term since when  $g_1 = g_2$ , we have  $g_1 g_2^{-1} = 1$ , and 1 was already counted in  $A - A$  from  $F - F$ .

We now put this all together. For  $|A + A| > |A - A|$ , we need

$$\begin{aligned} 2(m-k)k + \binom{m-k}{2} + (m-k) &> (m-k)k + 2\binom{m-k}{2}, \\ \iff k &> m/3 - 1 \text{ and } m \neq k, \\ \iff m/3 &\leq k < m. \end{aligned} \tag{6}$$

Note that when  $k = m$  the set is necessarily balanced as  $F + F = F - F$ . Further, one can now see why we may assume  $k \neq 0$ : when  $k$  is smaller than  $m/3$ , we expect the set to be MDTs, and indeed we will assume this is the case.

To use this naive estimate to prove our theorem, we first formalize our assumption that we have minimal overlaps within the sumset.

**Definition 3.3.** Let  $A \subseteq D$  such that  $|A| = m$ , and let  $q = (a, b, c, d) \in A^4$ . We say that  $q$  represents a **collision** if  $ab = cd$ .

Every collision that occurs in  $A$  has the potential to make  $|A + A|$  smaller relative to our naive estimate, unless the quadruple  $q$  is of a form that we already took into account. For example, if  $q = (a, b, a, b)$ , then this is not a collision we need to count as  $ab = ab$  trivially. Similarly, collisions of the form  $q = (a, b, b, a)$  with  $a$  and  $b$  both rotations do not subtract from  $|A + A|$  in Equation (6) as we already accounted for commutativity of addition for rotations. And finally, if  $q = (a, b, c, d)$  is a collision where  $a, b, c$ , and  $d$  are all flips, then  $q$  does not impact Equation (6) as  $F + F = F - F$  do not affect the relative sizes of the sum and difference sets. We refer to these three kinds of quadruples as **redundant**.

Every non-redundant collision  $(a, b, c, d)$  of  $A$ , together with  $(c, d, a, b)$ , decreases the size of  $A + A$  by at most 1 from our naive estimate. Let  $X_A$  be half the total number of non-redundant collisions of  $A$ . Then combining the above analysis with Equation (6), we are guaranteed to have that  $A$  is MSTD when

$$\begin{aligned} 2(m-k)k + \binom{m-k}{2} + (m-k) - X_A &> (m-k)k + 2\binom{m-k}{2}, \\ \iff m/3 + \frac{2X_A}{3(m-k)} &\leq k < m, \\ \iff 3k^2 - 4mk + m^2 + 2X_A &\leq 0 \text{ (and } k < m). \end{aligned} \tag{7}$$

We use the quadratic equation to solve for when the above quantity equals 0 and obtain  $k = (1/6)(4m \pm \sqrt{16m^2 - 12(m^2 + 2X_A)}) = (1/3)(2m \pm \sqrt{m^2 - 6X_A})$ . Thus Equation (7) is satisfied when:

$$\frac{2m - \sqrt{m^2 - 6X_A}}{3} \leq k \leq \frac{2m + \sqrt{m^2 - 6X_A}}{3} \quad (\text{and } k < m). \tag{8}$$

*Remark 3.* We are assuming that no “collisions” of the form  $ab^{-1} = cd^{-1}$  happen to lower the size of  $A - A$ . Because our objective is to guarantee  $|A + A| > |A - A|$  for a large proportion of  $A$ , this assumption still gives a sufficient condition on  $X_A$  and  $k$ .

One therefore sees that for most values of  $k$ , when the number of collisions is not too large,  $A$  is MSTD. More formally, suppose that  $A$  is chosen randomly out of the subsets of  $D$  with size  $m$ . Suppose that the expectation value of  $X_A$  is bounded above by  $c_1 m^2$ , where  $c_1 = 7/1152 \approx 0.006076$ . Then the actual value of  $X_A$  exceeds  $12c_1 m^2$  at most  $1/12$  of the time by Markov's inequality. Of sets  $A$  with  $5m/12 \leq k \leq 11m/12$ , the actual value of  $X_A$  exceeds  $12c_1 m^2$  at most  $2/12 = 1/6$  of the time. Thus when  $5m/12 \leq k \leq 11m/12$ , Equation (8) is true at least  $5/6$  of the time since when  $X_A = 12c_1 m^2$ , the equation reads

$$\begin{aligned} \frac{2m - \sqrt{m^2 - 72c_1 m^2}}{3} &\leq k \leq \frac{2m + \sqrt{m^2 - 72c_1 m^2}}{3}, \\ \Leftrightarrow \frac{2m - \sqrt{9/16 \cdot m^2}}{3} &\leq k \leq \frac{2m + \sqrt{9/16 \cdot m^2}}{3}, \\ \Leftrightarrow \frac{5m}{12} &\leq k \leq \frac{11m}{12}. \end{aligned} \quad (9)$$

Now, we need to make sure that for our values of  $m$ , more than  $(1/2)/(5/6) = 3/5$  proportion of sets in  $\mathcal{S}_{D,m}$  have  $5m/12 \leq k \leq 11m/12$ . This will ensure that a proportion greater than  $1/2$  of sets in  $\mathcal{S}_{D,m}$  satisfy Equation (8) and are therefore MSTD.

More formally speaking, we require

$$\lim_{n \rightarrow \infty} \frac{\sum_{k=\lceil 5m/12 \rceil}^{\lfloor 11m/12 \rfloor} \binom{n}{k} \binom{n}{m-k}}{\binom{2n}{m}} > 3/5 \quad (10)$$

$$\Leftrightarrow \frac{m!}{2^m} \sum_{k=\lceil 5m/12 \rceil}^{\lfloor 11m/12 \rfloor} \frac{1}{k!(m-k)!} > 3/5, \quad (11)$$

where we may take the limit as  $n \rightarrow \infty$  because the left hand side of Equation (10) decreases with increasing  $n$ . Equation (11) can be verified numerically to be true for  $m \geq 6$ .<sup>1</sup>

The problem has therefore been reduced to placing an upper bound on the values of  $m$  such that the expected value of  $X_A$  is at most  $c_1 m^2$  when  $A$  is chosen uniformly at random from  $\mathcal{S}_{D,m}$ . The following lemma gives us the result we need.

**Lemma 3.4.** *When  $A$  is chosen uniformly at random from  $\mathcal{S}_{D,m}$ , we have*

$$\mathbb{E}[X_A] \leq \left( \frac{7}{32} + \frac{1}{m} + \frac{5j}{8m^2} \right) \frac{m^4}{n}. \quad (12)$$

Much of the machinery of this proof lies in Lemma 3.4; its proof is rather technical and can be found in Subsection 3.2.

When  $m \geq 6$ , Lemma 3.4 implies that under the hypothesis of the lemma,

$$\mathbb{E}[X_A] \leq c_2 \frac{m^4}{n}, \quad \text{where } c_2 = \frac{7}{32} + \frac{1}{6} + \frac{5j}{288} = \frac{111 + 5j}{288}. \quad (13)$$

We are ready to complete the proof of Theorem 1.8. Recall that we wanted to have  $\mathbb{E}[X_A] \leq c_1 m^2$  to ensure most subsets of size  $m$  would be MSTD. Thus the requisite upper bound on  $m$  is determined as

$$c_2 \frac{m^4}{n} \leq c_1 m^2,$$

<sup>1</sup>In fact, when  $m$  is large, we expect almost all of the sets in  $\mathcal{S}_{D,m}$  to have  $5m/12 \leq k \leq 11m/12$ ; see Subsection 3.1 for further discussion on this fact. For smaller values of  $m \geq 6$ , one can verify Equations (10) and (11) using the following Desmos link: <https://www.desmos.com/calculator/e4zqwbmmcr>.

$$\iff m \leq \sqrt{\frac{c_1 n}{c_2}} = c_j \sqrt{n}, \quad (14)$$

where  $c_j = \sqrt{c_1/c_2} = \sqrt{7/(4(111 + 5j))} \approx 1.3229/\sqrt{5j + 111}$ .

This concludes the proof of Theorem 1.8.

**3.1. Proof of Theorems 3.1 and 3.2.** In this section we prove Theorem 3.1, and in the process we outline the steps needed to prove Theorem 3.2. We now assume that  $n > n_{j,\epsilon}$ , where  $j$  is a constant upper bound on the number of elements of order at most 2 in the group  $G$  and  $n_{j,\epsilon}$  is sufficiently large.

Having proved Theorem 1.8 for  $6 \leq m \leq (1.3229/\sqrt{5j + 111})\sqrt{n}$ , we may now assume  $m \geq (1.3229/\sqrt{5j + 111})\sqrt{n}$ ; that is,  $m$  is now large. We will follow similar steps to the previous proof, but using this assumption, we will increase  $c_1$  to be arbitrarily close to  $1/16$ , and we will decrease  $c_2$  to be arbitrarily close to  $7/32$ . Then we will have that the coefficient of the  $\sqrt{n}$ , which is  $\sqrt{c_1/c_2}$  (as discussed in the previous proof), is arbitrarily close to  $\sqrt{2/7}$ .

Take small  $\epsilon_1 > 0$ . Note that inside of  $\mathcal{S}_{D,m}$ , the distribution of values of  $k$  is a hypergeometric distribution. This is because one can construct a random set in  $\mathcal{S}_{D,m}$  by taking  $m$  random elements of the group  $D$  without replacement, one at a time; to begin with there is a  $1/2$  chance each time that we choose a flip. Thus since  $n$  is very large and  $j$  is fixed, having  $m \geq (1.3229/\sqrt{5j + 111})\sqrt{n}$  is sufficient for a proportion at least  $1 - \epsilon_1$  of sets in  $\mathcal{S}_{D,m}$  to have  $k \in [(1/2 - \epsilon_1)m, (1/2 + \epsilon_1)m]$ .

Going back to Equation (8), we thus see that we just need  $\sqrt{m^2 - 6X_A}$  to be at least  $m/2 + \epsilon_1$ , or  $6X_A \leq m^2 - (m/2 + \epsilon_1)^2$ , slightly more than half the time when  $k$  is in the relevant interval. More specifically, we need

$$\mathbb{P} \left[ 6X_A \leq \frac{3m^2}{4} - m\epsilon_1 - \epsilon_1^2 \mid (1/2 - \epsilon_1)m \leq k \leq (5/6 + \epsilon_1)m \right] \geq \frac{1}{2} \cdot \frac{1}{(1 - \epsilon_1)^2}. \quad (15)$$

Then, among sets with  $(1/2 - \epsilon_1)m \leq k \leq (5/6 + \epsilon_1)m$  (which, recall, form a proportion of at least  $1 - \epsilon_1$  of sets in  $\mathcal{S}_{D,m}$ ), at least a proportion of  $(1/2)/(1 - \epsilon_1)^2$  satisfy Equation (8). This means that a proportion of at least  $(1/2)/(1 - \epsilon_1) > 1/2$  of sets in  $\mathcal{S}_{D,m}$  are MSTD.

For Equation (15) to hold, we claim that it suffices to have the following probability bound, not conditioned on the size of  $k$ :

$$\mathbb{P} \left[ X_A \leq \frac{m^2}{8} - \frac{m\epsilon_1}{6} - \frac{\epsilon_1^2}{6} \right] \geq \frac{1}{2} \cdot \frac{1}{1 - \epsilon_1} + \epsilon_1. \quad (16)$$

To see why Equation (16) implies Equation (15), call  $B$  the event that  $(1/2 - \epsilon_1)m \leq k \leq (5/6 + \epsilon_1)m$  and  $C$  the event that  $X_A \leq \frac{m^2}{8} - \frac{m\epsilon_1}{6} - \frac{\epsilon_1^2}{6}$ . Then, we manipulate conditional probabilities as follows.

$$\begin{aligned} \mathbb{P}[C] &= \mathbb{P}[C \mid B] \mathbb{P}[B] + \mathbb{P}[C \mid \neg B] \mathbb{P}[\neg B] \leq \mathbb{P}[C \mid B] \mathbb{P}[B] + \mathbb{P}[\neg B], \\ \iff \mathbb{P}[C \mid B] &\geq \frac{\mathbb{P}[C] - \mathbb{P}[\neg B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[C] - (1 - \mathbb{P}[B])}{\mathbb{P}[B]} = 1 - \frac{1 - \mathbb{P}[C]}{\mathbb{P}[B]}. \end{aligned} \quad (17)$$

Since  $\mathbb{P}[B] \geq 1 - \epsilon_1$  and Equation (16) says that  $\mathbb{P}[C] \geq (1/2)/(1 - \epsilon_1) + \epsilon_1$ , we have that if Equation (16) is true, then

$$\mathbb{P}[C \mid B] \geq 1 - \frac{1 - ((1/2)/(1 - \epsilon_1) + \epsilon_1)}{1 - \epsilon_1} = 1 - \frac{(1 - \epsilon_1) - (1/2)/(1 - \epsilon_1)}{1 - \epsilon_1} = \frac{1/2}{(1 - \epsilon_1)^2}, \quad (18)$$

and the claim is shown.

To ensure that Equation (16) is true, we require

$$\mathbb{E}[X_A] \leq \left(1 - \epsilon_1 - \frac{1/2}{1 - \epsilon_1}\right) \left(\frac{m^2}{8} - \frac{m\epsilon_1}{6} - \frac{\epsilon_1^2}{6}\right). \quad (19)$$

Then by Markov's inequality, the probability that  $X_A$  exceeds  $\frac{m^2}{8} - \frac{m\epsilon_1}{6} - \frac{\epsilon_1^2}{6}$  is at most  $1 - \epsilon_1 - \frac{1/2}{1 - \epsilon_1}$ , which is equivalent to Equation (16).

We may now choose a small value  $\epsilon_2$  such that Equation (19) is true if

$$\mathbb{E}[X_A] \leq \left(\frac{1}{16} - \epsilon_2\right) m^2. \quad (20)$$

Notice that in the limit  $\epsilon_1 \rightarrow 0$ , Equation (19) boils down to the statement that  $\mathbb{E}[X_A] \leq m^2/16$ , so we can make  $\epsilon_2$  be as small as desired by making  $\epsilon_1$  be small.

We now revisit Lemma 3.4. We are now assuming that  $m$  is large and  $j$  is a constant, so only the first term dominates:

$$\mathbb{E}[X_A] \leq \left(\frac{7}{32} + \epsilon_3\right) \frac{m^4}{n}. \quad (21)$$

We wanted to have  $\mathbb{E}[X_A] \leq (1/16 - \epsilon_2)m^2$ . Thus the upper bound on  $m$  is now determined as

$$\begin{aligned} \left(\frac{7}{32} + \epsilon_3\right) \frac{m^4}{n} &\leq \left(\frac{1}{16} - \epsilon_2\right) m^2 \\ \iff m &\leq c\sqrt{n}, \end{aligned} \quad (22)$$

where

$$c = \sqrt{\frac{1/16 - \epsilon_2'}{7/32 + \epsilon_3}}. \quad (23)$$

If  $n$  is arbitrarily large and  $j$  is constant compared to  $n$ , we can choose  $\epsilon_2$  and  $\epsilon_3$  to be very small, so that  $c$  is arbitrarily close to  $\sqrt{(1/16)/(7/32)} \approx 0.5345$ .

This completes the proof of Theorem 3.1. To prove Theorem 3.2, we follow a very similar method. We choose  $m_\epsilon$  large enough that almost all of the sets in  $\mathcal{S}_{D,m}$  have  $(0.5 - \epsilon_1)m \leq k \leq (0.5 + \epsilon_1)m$ . The only difference is that now, in Equation (15) we replace the right-hand side with  $(1 - \epsilon_1)$ , so that the proportion of MSTD sets is at least  $(1 - \epsilon_1)^2$ . (We may choose  $\epsilon_1$  so that  $(1 - \epsilon_1)^2$  equals the desired  $(1 - \epsilon)$ ). Then in Equation (16) we replace the right-hand side with  $1 - (1 - \epsilon_1)\epsilon_1$ , leading to the analog of Equation (20) being that  $\mathbb{E}[X_A] \leq \epsilon_2 m^2$  for some small  $\epsilon_2$  depending on  $\epsilon_1$ . The rest of the proof continues as before, leading to a coefficient  $c_{\epsilon,j}$  proportional to  $\sqrt{\epsilon_2}$ . This completes the proof.

**3.2. Proof of Lemma 3.4.** We now prove Lemma 3.4. Recall that  $X_A$  is defined to be half the number of non-redundant collisions in the set  $A$ , and we are interested in bounding above the expectation value of  $X_A$  when  $A$  is chosen uniformly at random from  $\mathcal{S}_{D,m}$ .

To more easily count the collisions in  $A$ , we make the following definition.

**Definition 3.5.** A **redundant triple** is a triple  $(a, b, c) \in D^3$  such that the quadruple  $(a, b, c, c^{-1}ab)$  is redundant. That is, a triple  $(a, b, c)$  is redundant if  $a = c$ , or if  $a, b$ , and  $c$  are all flips, or if  $b = c$  and  $a$  and  $b$  are both rotations. Denote  $T \subseteq D^3$  to be the set of non-redundant triples.

Define the function  $\chi : \mathcal{S}_{D,m} \times T \rightarrow \{0, 1\}$  by  $\chi(A, t) = 1$  if for the non-redundant triple  $t = (a, b, c)$ , the element  $c^{-1}ab$  is in  $A$ , and  $\chi(A, t) = 0$  otherwise.

For a fixed set  $A$ , the set  $A^3 \cap T$  is the set of non-redundant triples with all three elements contained in  $A$ . Notice that we have

$$X_A = \frac{1}{2} \sum_{t \in A^3 \cap T} \chi(A, t). \quad (24)$$

That is, the number of non-redundant collisions in  $A$  is the same as the number of non-redundant triples  $(a, b, c) \in A^3 \cap T$  such that the element  $d = c^{-1}ab$  is in  $A$ , forming a quadruple  $(a, b, c, d)$  representing a collision  $ab = cd$ .

By definition of expectation value, we write

$$\mathbb{E}[X_A] = \sum_{A \in \mathcal{S}_{D,m}} \mathbb{P}[A] X_A. \quad (25)$$

Since  $A$  is chosen uniformly at random from the  $\binom{2n}{m}$  sets in  $\mathcal{S}_{D,m}$ , we have  $\mathbb{P}[A] = 1/\binom{2n}{m}$ . Thus,

$$\mathbb{E}[X_A] = \frac{1}{\binom{2n}{m}} \sum_{A \in \mathcal{S}_{D,m}} \frac{1}{2} \sum_{t \in A^3 \cap T} \chi(A, t). \quad (26)$$

We swap the order of the sums.

$$\mathbb{E}[X_A] = \frac{1}{2} \frac{1}{\binom{2n}{m}} \sum_{t \in T} \sum_{\substack{A \in \mathcal{S}_{D,m} \\ A^3 \ni t}} \chi(A, t). \quad (27)$$

To compute the inner sum, we must simply count the number of sets  $A \in \mathcal{S}_{D,m}$  with  $t = (a, b, c) \in A^3$  such that  $A$  contains  $c^{-1}ab$ . That is,

$$\mathbb{E}[X_A] = \frac{1}{2} \frac{1}{\binom{2n}{m}} \sum_{(a,b,c) \in T} |\{A \in \mathcal{S}_{D,m} \mid a, b, c, c^{-1}ab \in A\}|. \quad (28)$$

We now break this sum into seven pieces for different kinds of triples  $(a, b, c) \in T$ . These are:

- $T_1 = \{(a, b, c) \in T \mid a, b, c, c^{-1}ab \text{ are distinct}\}$
- $T_2 = \{(a, b, c) \in T \mid a = b; a, c, c^{-1}a^2 \text{ are distinct}\}$
- $T_3 = \{(a, b, c) \in T \mid b = c; a, c, c^{-1}ac \text{ are distinct}\}$
- $T_4 = \{(a, b, c) \in T \mid c^{-1}ab = a; a, b, c \text{ are distinct}\}$
- $T_5 = \{(a, b, c) \in T \mid c^{-1}ab = c; a, b, c \text{ are distinct}\}$
- $T_6 = \{(a, b, c) \in T \mid b = c; a = c^{-1}ac; a, c \text{ are distinct}\}$
- $T_7 = \{(a, b, c) \in T \mid a = b; c = c^{-1}a^2; a, c \text{ are distinct}\}$

We have  $T = \bigcup_{i=1}^7 T_i$ , for these seven cases cover all the cases of possible equalities between the four elements except for those where  $a = c$ , or equivalently,  $c^{-1}ab = b$ , since those cases are redundant triples. Furthermore, this union is disjoint.

Note that for triples  $(a, b, c) \in T_1$ , the quantity  $|\{A \in \mathcal{S}_{D,m} \mid a, b, c, c^{-1}ab \in A\}|$  is given by  $\binom{2n-4}{m-4}$  since we are requiring four distinct elements to be in  $A$ , and we have  $2n-4$  choices for the remaining  $m-4$  elements. For triples in  $T_2, T_3, T_4$ , and  $T_5$ , we are requiring three distinct elements to be in  $A$ , so we have  $|\{A \in \mathcal{S}_{D,m} \mid a, b, c, c^{-1}ab \in A\}| = \binom{2n-3}{m-3}$ . Finally, for triples in  $T_6$  and  $T_7$  we have  $|\{A \in \mathcal{S}_{D,m} \mid a, b, c, c^{-1}ab \in A\}| = \binom{2n-2}{m-2}$ .

Therefore, from Equation (28), we may write:

$$\begin{aligned} \mathbb{E}[X_A] &= \frac{1}{2} \frac{1}{\binom{2n}{m}} \left[ \binom{2n-4}{m-4} |T_1| + \binom{2n-3}{m-3} (|T_2| + |T_3| + |T_4| + |T_5|) + \binom{2n-2}{m-2} (|T_6| + |T_7|) \right] \\ &\leq \frac{1}{2} \left[ \left( \frac{m}{2n} \right)^4 |T_1| + \left( \frac{m}{2n} \right)^3 (|T_2| + |T_3| + |T_4| + |T_5|) + \left( \frac{m}{2n} \right)^2 (|T_6| + |T_7|) \right], \end{aligned} \quad (29)$$

where in the second line we used the fact that

$$\frac{\binom{2n-4}{m-4}}{\binom{2n}{m}} = \frac{m(m-1)(m-2)(m-3)}{2n(2n-1)(2n-2)(2n-3)} \leq \left(\frac{m}{2n}\right)^4, \quad (30)$$

and similarly for the other two terms.

Now, to use Equation (29) to find an upper bound on  $\mathbb{E}[X_A]$ , all that remains to be done is find an upper bound on each of the  $|T_i|$ 's. We do so next.

We bound  $|T_1|$  using the trivial inequality  $|T_1| \leq |T|$ . There are  $(2n)^3$  total triples in  $D^3$ , but we may subtract the redundant triples, including the  $n^3$  triples consisting of three flips. Thus we obtain

$$|T_1| \leq 7n^3. \quad (31)$$

We bound  $|T_2|$  and  $|T_3|$  next. We have  $2n$  choices for  $b$ . In  $T_2$ ,  $a$  must equal  $b$ , and in  $T_3$ ,  $c$  must equal  $b$ , and in both, we have at most  $2n$  choices for the remaining value of  $a$  or  $c$ . The extra condition that  $c^{-1}ab$  is distinct from the others only lowers  $|T_2|$  and  $|T_3|$ , so we do not have to take it into account to obtain an upper bound. Thus,

$$|T_2|, |T_3| \leq 4n^2. \quad (32)$$

For  $|T_4|$  and  $|T_5|$ , we have  $2n$  choices for  $a$  and  $2n$  choices for  $c$ , but  $b$  must be the element  $a^{-1}ca$  for  $T_4$  or  $a^{-1}c^2$  for  $T_5$ . Thus we have

$$|T_4|, |T_5| \leq 4n^2. \quad (33)$$

When considering  $T_6$ , we note that since  $b = c$ , the triple  $(a, b, c)$  is redundant if  $a$  and  $c$  are both rotations, or if both are flips. So, one must be a rotation and the other must be a flip, and we require that  $a = c^{-1}ac$ , or  $ca = ac$ . Thus to bound  $|T_6|$  we must count the number of pairs of elements with  $(0, g_1) \cdot (1, g_2) = (1, g_2) \cdot (0, g_1)$ , that is,  $(1, g_1g_2) = (1, g_2g_1^{-1})$ . This happens if and only if  $g_1^{-1} = g_1$ , or  $g_1^2 = 1$ . Recalling that there are  $j$  elements in  $G$  with order 2 or less, there are therefore  $j$  choices for the rotation element, and  $n$  choices for the flip element. We multiply by 2 since  $a$  can be either the flip or the rotation, and  $c$  is the other of the two. Thus,

$$|T_6| \leq 2nj. \quad (34)$$

Finally, for  $T_7$ , we again split into two cases: firstly where  $a$  is a rotation, and secondly where  $a$  is a flip, so  $c$  is a rotation to avoid redundancy. Here we require  $c = c^{-1}a^2$ , or  $c^2 = a^2$ . For the first case, we first consider the number of pairs with  $a^2 = c^2 = 1$ . Since  $a$  must be a rotation, there are only  $j$  choices for  $a^2 = 1$ ;  $c$  can be a flip or a rotation, so there are  $n + j$  ways to have  $c^2 = 1$ . So, there are at most  $(n + j)j$  pairs of this kind. If  $a$  is a rotation and  $a^2 \neq 1$ , then  $c$  must also be a rotation since otherwise  $c^2 = 1 \neq a^2$ . Thus  $a$  and  $c$  commute, so  $a^2 = c^2 \iff (ac^{-1})^2 = 1$ . There are  $n - j$  choices of  $a$  with  $a^2 \neq 1$ , and for each, there are  $j$  choices of  $c$  which have  $(ac^{-1})^2 = 1$ . Thus there are at most  $(n - j)j$  pairs of this kind.

Next we consider the case where  $a$  is a flip and  $c$  is a rotation. Here  $a^2 = 1$ , so there are only  $j$  choices for  $c$  that have  $c^2 = a^2$ . Thus there are  $nj$  pairs of this kind. In total,

$$|T_7| \leq (n + j)j + (n - j)j + nj = 3nj. \quad (35)$$

We now substitute these bounds on each  $|T_i|$  into Equation (29).

$$\begin{aligned} \mathbb{E}[X_A] &\leq \frac{1}{2} \left[ \left(\frac{m}{2n}\right)^4 (7n^3) + \left(\frac{m}{2n}\right)^3 (4n^2 + 4n^2 + 4n^2 + 4n^2) + \left(\frac{m}{2n}\right)^2 (2nj + 3nj) \right] \\ &= \left( \frac{7}{32} + \frac{1}{m} + \frac{5j}{8m^2} \right) \frac{m^4}{n}. \end{aligned} \quad (36)$$

This concludes the proof of Lemma 3.4.

**3.3. Finitely Generated Abelian Groups.** We transition to a discussion of finitely generated abelian groups  $G$ . When  $|G| < \infty$  Theorems 1.8, 3.1, and 3.2 hold, so the remaining case is when  $G$  is an infinite group. However, we must make some restriction as to ensure taking subsets uniformly at random is well defined. By the fundamental theorem of finitely generated abelian groups,

$$G \cong \mathbb{Z}^{r_0} \oplus \mathbb{Z}_{q_1}^{r_1} \oplus \cdots \oplus \mathbb{Z}_{q_k}^{r_k}, \quad (37)$$

where  $q_i$  are powers of (not necessarily distinct) primes. We denote elements of  $G$  as a tuple  $(g_{0,1}, g_{0,2}, \dots, g_{0,r_0}, g_{1,1}, \dots, g_{k,r_k}) \in G$ , where  $g_{0,b} \in \mathbb{Z}$  and  $g_{a,b} \in \mathbb{Z}_{q_a}$  for  $a > 0$ . Since this section will occasionally require us to deal with multiple groups simultaneously, we update our notation of  $D$  as the generalized dihedral group of  $G$  to the standard  $\text{Dih}(G)$ . We still denote elements of  $\text{Dih}(G)$  as  $(z, g)$ , where  $z \in \{0, 1\}, g \in G$ .

For some fixed  $\alpha \in \mathbb{N}$ , we will consider taking subsets uniformly at random from the finite

$$\text{Dih}(G_\alpha) = \{(z, (g_{0,1}, \dots, g_{0,r_0}, \dots, g_{k,r_k})) \in \text{Dih}(G) \mid 0 \leq g_{0,b} < \alpha\}. \quad (38)$$

Our goal is to leverage Theorems 1.8, 3.1, and 3.2, and, to do so, we will refer to  $\text{Dih}(G')$ , where

$$G' = \mathbb{Z}_\alpha^{r_0} \oplus \mathbb{Z}_{q_1}^{r_1} \oplus \cdots \oplus \mathbb{Z}_{q_k}^{r_k}. \quad (39)$$

To adhere to prior notation, let  $j$  be the number of elements in  $G'$  that are at most order 2 and let  $\mathcal{S}_m$  denote the collection of subsets of  $\text{Dih}(G_\alpha)$  that have size  $m$ . Then we get the following corollaries of Theorems 1.8, 3.1, and 3.2:

**Corollary 3.6.** *If  $6 \leq m \leq c_j \sqrt{n}$ , then there are more MSTD than MDTs in  $\mathcal{S}_m$ .*

**Corollary 3.7.** *If  $n \geq n_{j,\epsilon}$ , then if  $6 \leq m \leq (\sqrt{2/7} - \epsilon) \sqrt{n}$ , then there are more MSTD than MDTs sets in  $\mathcal{S}_m$ .*

**Corollary 3.8.** *For any  $\epsilon > 0$ , there exist  $m_\epsilon$  and  $c_{\epsilon,j}$  such that if  $m_\epsilon \leq m \leq c_{\epsilon,j} \sqrt{n}$ , the proportion of MSTD sets in  $\mathcal{S}_m$  is at least  $1 - \epsilon$ .*

*Proof.* We will establish a bijection  $\phi : \text{Dih}(G_\alpha) \rightarrow \text{Dih}(G')$  such that, if  $(a, b, c, d) \in (\text{Dih}(G_\alpha))^4$  is a collision, then  $(\phi(a), \phi(b), \phi(c), \phi(d)) \in (\text{Dih}(G'))^4$  is also a collision. Since we are still working with generalized dihedral groups, this immediately implies Corollaries 3.6, 3.7, and 3.8, as the number of non-degenerate collisions in  $\text{Dih}(G_\alpha)$  is bounded above by the number of non-degenerate collisions in  $\text{Dih}(G')$ .

Let  $\phi : \text{Dih}(G_\alpha) \rightarrow \text{Dih}(G')$  defined by  $\phi((z, (g_{0,1}, \dots, g_{k,r_k}))) = (z, (g_{0,1}, \dots, g_{k,r_k}))$ . As defined, this is clearly a bijection. Let  $(x_1, x_2, x_3, x_4) \in (\text{Dih}(G_\alpha))^4$  be a collision. By definition,  $x_1 x_2 = x_3 x_4$ . Let  $x_j = (z_j, (g_{0,1,j}, \dots, g_{k,r_k,j}))$ . Define the binary operation  $\star_t : \mathbb{Z}_t \times \mathbb{Z}_t \rightarrow \mathbb{Z}_t$  by

$$g_1 \star g_2 = \begin{cases} g_1 + g_2 \pmod t, & z_1 = 0, \\ g_1 - g_2 \pmod t, & z_1 = 1. \end{cases} \quad (40)$$

Then we get

$$x_1 x_2 = (z_1 + z_2 \pmod 2, (g_{0,1,1} \star_{q_0} g_{0,1,2}, \dots, g_{k,r_k,1} \star_{q_k} g_{k,r_k,2})). \quad (41)$$

In other words, we are given the following system of equations:

$$\begin{aligned} z_1 + z_2 &\equiv z_3 + z_4 \pmod 2, \\ g_{0,1,1} + g_{0,1,2} &= g_{0,1,3} + g_{0,1,4}, \\ &\vdots \\ g_{k,r_k,1} \star_{q_k} g_{k,r_k,2} &= g_{k,r_k,3} \star_{q_k} g_{k,r_k,4}. \end{aligned}$$

Consider  $q = (\phi(x_1), \phi(x_2), \phi(x_3), \phi(x_4))$ . For  $q$  to be a collision in  $\text{Dih}(G')$ , we require the following system of equations:

$$\begin{aligned} z_1 + z_2 &\equiv z_3 + z_4 \pmod{2}, \\ g_{0,1,1} \star_{\alpha} g_{0,1,2} &= g_{0,1,3} \star_{\alpha} g_{0,1,4}, \\ &\vdots \\ g_{k,r_k,1} \star_{q_k} g_{k,r_k,2} &= g_{k,r_k,3} \star_{q_k} g_{k,r_k,4}, \end{aligned}$$

which is implied by our given system. Therefore the result follows.  $\square$

We present an example of the number of collisions in another dihedral group  $\mathbb{Z}_2 \ltimes \mathbb{Z}_n^2$ . We show that the dihedral group  $\mathbb{Z}_2 \ltimes \mathbb{Z}_n^2$  possesses exactly the same number of possible collisions as the dihedral group  $\mathbb{Z}_2 \ltimes \mathbb{Z}_{n^2}$  if and only if  $n$  is odd and there are more collisions in  $\mathbb{Z}_2 \ltimes \mathbb{Z}_n^2$  otherwise. This result only provides some intuition on the the expected number of collisions depending on  $j$ , the number of elements of order 2 within the particular abelian group.

**Lemma 3.9.** *The number of possible collisions within the two groups are equal if and only if  $n$  is odd and there are more collisions in  $\mathbb{Z}_2 \ltimes \mathbb{Z}_n^2$  otherwise.*

To do so, we make use of the following useful results:

**Lemma 3.10.** *The number of pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 - b_1 \equiv x \pmod{n}$  and  $a_2 - b_2 \equiv y \pmod{n}$  and the number of pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) - (b_1n + b_2) \equiv xn + y \pmod{n^2}$  are both equal to  $n^2$*

*Proof.* For any given  $(b_1, b_2)$  and  $b_1n + b_2$ , there exist only a single  $(a_1, a_2)$  and  $a_1n + a_2$  which satisfies the conditions respectively. Thus, there are  $n^2$  such pairs for both cases.  $\square$

**Lemma 3.11.** *The number of pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 + b_1 \equiv x \pmod{n}$  and  $a_2 + b_2 \equiv y \pmod{n}$  is as follows:*

- when  $n$  is odd, there are  $\frac{n^2+1}{2}$  such pairs
- when  $n$  is even and  $y$  is odd or  $x$  is odd, there are  $\frac{n^2}{2}$  such pairs
- when  $n, y$ , and  $x$  are all even, there are  $\frac{n^2+4}{2}$  such pairs

*Proof.* When  $n$  is odd, each choice of  $a_1$  is paired with a single possible  $b_1$ , with one such pair being  $a_1 = b_1$ . Similarly, the same holds for  $a_2$  and  $b_2$ . This gives  $n^2$  as the number of pairs. However, we over-counted since swapping  $a_1$  with  $b_1$  and  $a_2$  with  $b_2$  does not yield a distinct pair. Thus, we divide by 2 except for the single pair where  $a_1 = b_1$  and  $a_2 = b_2$  which we did not over-count to get  $\frac{n^2-1}{2} + 1 = \frac{n^2+1}{2}$  pairs of  $(a_1, a_2), (b_1, b_2)$ .

When  $n$  is even but  $y$  is odd, we choose pairs  $(a_2, b_2)$  first which gives  $n/2$  distinct pairs where  $a_2$  is never equal to  $b_2$ . For each of these pairs, we can choose any choice of  $a_1$  which forces  $b_1$  without over-counting. Thus, we get  $\frac{n^2}{2}$  pairs. Similar arguments hold for when  $x$  is odd.

When all of  $n, y$ , and  $x$  are even, we get that there are two pairs of identical number and  $\frac{n-2}{2}$  pairs of different numbers which sum to  $x$  and similar for  $y$  in  $\mathbb{Z}_n$ . We choose  $a_2, b_2$  first. If  $a_2 = b_2$ , we have  $\frac{n-2}{2} + 2 = \frac{n+2}{2}$  choices for choosing  $a_1, b_1$ . If  $a_2 \neq b_2$ , we have that for any  $a_1$ , we uniquely determines  $b_1$  without over-counting and thus there are  $n$  choices. In total, we have  $2 * \frac{n+2}{2} + n * \frac{n-2}{2} = \frac{n^2+4}{2}$  pairs of  $(a_1, a_2), (b_1, b_2)$ .  $\square$

**Lemma 3.12.** *The number of pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) + (b_1n + b_2) \equiv xn + y \pmod{n^2}$  is as follows:*

- when  $n$  is odd, there are  $\frac{n^2+1}{2}$  such pairs
- when  $n$  is even but  $y$  is odd, there are  $\frac{n^2}{2}$  such pairs
- when both  $n$  and  $y$  are even, there are  $\frac{n^2+2}{2}$  such pairs

*Proof.* When  $n$  is odd, we have that each of  $a_1n + a_2$  is paired with another  $b_1n + b_2$ , with exactly one being paired with itself. Thus, there are  $\frac{n^2-1}{2} + 1 = \frac{n^2+1}{2}$  pairs of  $a_1n + a_2$  and  $b_1n + b_2$ .

When  $n$  is even but  $y$  is odd, we have that each of  $a_1n + a_2$  is paired with another  $b_1n + b_2$ , necessarily distinct. Thus, there are  $\frac{n^2}{2}$  pairs of  $a_1n + a_2$  and  $b_1n + b_2$ .

When  $n$  is even and  $y$  is even, we have that each of  $a_1n + a_2$  is paired with another  $b_1n + b_2$ , with exactly two being paired with themselves. Thus, there are  $\frac{n^2-2}{2} + 2 = \frac{n^2+2}{2}$  pairs of  $a_1n + a_2$  and  $b_1n + b_2$ .  $\square$

Combining these results over the casework where elements of our pairs may be a rotation or a flip yield the desired result, with the details of the proof in Appendix B.

#### 4. EXPECTED SIZE OF SUM AND DIFFERENCE SETS

In this section, we only consider the classical dihedral groups

$$D_{2n} = \mathbb{Z}_2 \ltimes \mathbb{Z}_n = \langle r, s \mid r^n, s^2, rsrs \rangle \quad (42)$$

Throughout this section, we use  $\mathcal{S}_{2n,m}$  to denote the set of subsets of size  $m$  in  $D_{2n}$ .

The method of collision analysis will likely not be sufficient to prove that  $\mathcal{S}_{2n,m}$  has more MSTD sets than MDTs sets for values of  $m$  greater in order of magnitude than  $\sqrt{n}$ . The intuition for this comes from the fact that the sum and difference sets for  $A \subseteq D_{2n}$  should very roughly have size on the same order of magnitude as  $A^2$ . Hence, one would expect to usually have  $A+A = A-A = D_{2n}$  when  $m$  is much greater than  $\sqrt{n}$ . The analysis for relative numbers of MSTD and MDTs sets in  $\mathcal{S}_{2n,m}$  for these larger values of  $m$  should therefore be based on counting the number of *missed* sums and differences in  $D_{2n}$ , in direct analogy with the case of slow decay for the integers in [hegarty2009almost].

We take the first steps toward such an analysis by proving the following special case.

**Theorem 1.9.** *If  $n$  is prime, and  $A$  is chosen uniformly at random from  $\mathcal{S}_{2n,m}$ , then*

$$\mathbb{E}[|A - A|] = 2n - \frac{nm2^m \binom{n}{m} + 2n(n-1) \binom{n-m-1}{m-1}}{m \binom{2n}{m}} - \frac{n^2(n-1)}{\binom{2n}{m}} \sum_{k=1}^{m-1} \frac{\binom{n+k-m-1}{m-k-1} \binom{n-k-1}{k-1}}{k(m-k)}. \quad (4)$$

This follows from the following straightforward yet useful lemma, which reduces the problem of computing the probability of missing a sum or difference to an analogous problem in  $\mathbb{Z}_n$ .

**Lemma 4.1.** *Let  $A$  be a subset in  $\mathcal{S}_{2n,m}$  chosen uniformly at random. Then if  $r^i$  is a rotation in  $D_{2n}$ , we have*

$$\mathbb{P}[r^i \notin A + A] = \sum_{k=0}^m \frac{\mathbb{P}[A \text{ has } k \text{ flips}]}{\mathbb{P}[i \notin S + S \mid S \subseteq \mathbb{Z}_n, |S| = m-k]} \cdot \mathbb{P}[i \notin S - S \mid S \subseteq \mathbb{Z}_n, |S| = k], \quad (43)$$

and

$$\mathbb{P}[r^i \notin A - A] = \sum_{k=0}^m \frac{\mathbb{P}[A \text{ has } k \text{ flips}]}{\mathbb{P}[i \notin S - S \mid S \subseteq \mathbb{Z}_n, |S| = m-k]} \cdot \mathbb{P}[i \notin S + S \mid S \subseteq \mathbb{Z}_n, |S| = k]. \quad (44)$$

If  $r^i$  is a flip in  $D_{2n}$ , we have

$$\mathbb{P}[r^i \notin A + A] = \sum_{k=0}^m \frac{\mathbb{P}[A \text{ has } k \text{ flips}]}{\mathbb{P}[i \notin S_1 + S_2 \wedge i \notin S_2 - S_1 \mid S_1, S_2 \subseteq \mathbb{Z}_n, |S_1| = m-k, |S_2| = k]}, \quad (45)$$

and

$$\mathbb{P}[r^i s \notin A - A] = \sum_{k=0}^m \mathbb{P}[A \text{ has } k \text{ flips}] \cdot \mathbb{P}[i \notin S_1 + S_2 | S_1, S_2 \subseteq \mathbb{Z}_n, |S_1| = m - k, |S_2| = k]. \quad (46)$$

*Proof.* Partition  $A$  into its set of rotations  $R$  and flips  $F$ , and suppose that  $|F| = k$  and  $|R| = m - k$ . The rotation element  $r^i$  can appear in  $A + A$  precisely as  $r^j r^\ell = r^{j+\ell}$  for  $r^j, r^\ell \in R$  or as  $(r^j s)(r^\ell s) = r^{j+\ell}$  for  $r^j s, r^\ell s \in F$ . Taking the probability of the negations of these events respectively give second and third probabilities appearing in the sum of Equation (43). Equation (44) follows similarly.

The flip element  $r^i s$  can appear in  $A + A$  precisely as  $r^j(r^\ell s) = r^{j+\ell} s$  or as  $(r^\ell s)r^j = r^{\ell-j} s$  for  $r^j \in R, r^\ell s \in F$ , but unlike in the previous cases, these events are no longer independent so Equation (45) cannot be broken up into a product of simpler probabilities. Finally,  $r^i s$  can appear in  $A - A$  precisely  $r^j(r^\ell s)^{-1} = r^{j+\ell} s = (r^\ell s)(r^j)^{-1}$ , from which Equation (46) follows.  $\square$

A number of these probabilities can be expressed explicitly in terms of  $n$ ,  $m$ , and  $k$ . To prove Theorem 1.9, we need to compute all probabilities appearing in Equation (44) and Equation (46), but we will also compute the probabilities appearing in Equation (43) for completeness.

**Lemma 4.2.** *Let  $S$  be a subset of  $\mathbb{Z}_n$  of size  $m - k$  chosen uniformly at random, and let  $i$  be any element of  $\mathbb{Z}_n$ . Then*

$$\mathbb{P}[i \notin S + S] = \begin{cases} \frac{2^{m-k} \binom{\frac{n}{2}-1}{k}}{\binom{n}{m-k}}, & n \text{ and } i \text{ both even,} \\ \frac{2^{m-k} \binom{\frac{n}{2}}{m-k}}{\binom{n}{m-k}}, & n \text{ even and } i \text{ odd,} \\ \frac{2^{m-k} \binom{\frac{n-1}{2}}{m-k}}{\binom{n}{m-k}}, & n \text{ odd.} \end{cases} \quad (47)$$

*Proof.* If  $n$  and  $i$  are both even, then there exist exactly 2 elements of  $\mathbb{Z}_n$  that give  $i$  when added to themselves. The remaining elements of  $\mathbb{Z}_n$  partition into pairs of distinct elements adding to  $i$ , and any  $S$  such that  $i \notin S + S$  is obtained by selecting  $m - k$  of these  $\frac{n}{2} - 1$  pairs and one element from each pair, hence the result. The case where  $n$  is even and  $i$  is odd is identical except that all  $n$  of the elements of  $\mathbb{Z}_n$  are now partitioned into pairs as there are no elements that give  $i$  when added to themselves. Finally, if  $n$  is odd, then there is always a unique element that gives  $i$  when added to itself, and the remaining elements partition into  $\frac{n-1}{2}$  pairs, after which the same analysis can be applied.  $\square$

**Lemma 4.3.** *Let  $S$  be a subset of  $\mathbb{Z}_n$  of size  $k$  chosen uniformly at random, and let  $i$  be any nonzero element of  $\mathbb{Z}_n$  of order  $n/d$ . Then*

$$\mathbb{P}[i \notin S - S] = \frac{1}{\binom{n}{k}} \sum_{\substack{(k_1, \dots, k_d) \in \mathbb{Z}_{\geq 0}^d \\ k_1 + \dots + k_d = k}} \prod_{t=1}^d g\left(\frac{n}{d}, k_t\right), \quad (48)$$

where

$$g\left(\frac{n}{d}, k_t\right) = \begin{cases} 1, & k_t = 0, \\ \frac{\frac{n}{d} \binom{\frac{n}{d}-1-k_t}{k_t-1}}{k_t}, & k_t > 0. \end{cases} \quad (49)$$

*Remark 4.* When  $n$  is prime, we must have  $d = 1$  always and Equation (48) simplifies significantly to

$$\mathbb{P}[i \notin S - S] = \frac{g(n, k)}{\binom{n}{k}} = \frac{1}{\binom{n}{k}} \cdot \begin{cases} 1, & k = 0, \\ \frac{n \binom{n-1-k}{k-1}}{k}, & k > 0. \end{cases} \quad (50)$$

This is the reason why we restrict to such  $n$  in Theorem 1.9.

*Proof.* Partition  $\mathbb{Z}_n$  into the  $d$  additive cosets of the subgroup  $i\mathbb{Z}_n$ . Each of these cosets has size  $n/d$  and has elements that can be cyclically ordered such that the difference between any element and its predecessor is  $i$ . Choosing a set  $S$  of size  $k$  such that  $i \notin S - S$  is then equivalent to partitioning  $k$  into  $k_1 + \dots + k_d$  with each  $k_t \geq 0$ , and choosing  $k_t$  non-adjacent elements from the  $t^{\text{th}}$  coset to include in  $S$ . This is precisely what is counted by Equation (48) if  $g(n/d, k_t)$  is equal to the number of ways to choose  $k_t$  non-adjacent elements from a cyclically ordered set of size  $n/d$ .

Indeed, this is clear for  $k_t = 0$ , so assume  $k_t > 0$ . There are  $n/d$  choices for the first element to be included, and each selection of  $k_t$  elements may be made in  $k_t$  different ways by designating different elements to be this first choice. Once the first element has been selected, the number of ways to choose the remaining elements is equal to the number of ways to choose  $k_t - 1$  non-adjacent elements from a linearly ordered set of size  $n/d - 1$  without choosing the extremal elements. This is the classic stars and bars problem, which yields precisely the binomial coefficient in the definition of  $g(n/d, k_t)$ .  $\square$

**Lemma 4.4.** *Let  $S_1$  and  $S_2$  be subsets of  $\mathbb{Z}_m$  of sizes  $m - k$  and  $k$ , respectively, chosen uniformly at random, and let  $i$  be any element of  $\mathbb{Z}_n$ . Then*

$$\mathbb{P}[i \notin S_1 + S_2] = \frac{\binom{n-k}{m-k}}{\binom{n}{m-k}}. \quad (51)$$

*Proof.* For any fixed choice of  $S_2$ , we have  $\binom{n}{m-k}$  total choices for  $S_1$ . For each of these  $S_1$ , we have  $i \notin S_1 + S_2$  if and only if  $i - j \notin S_1$  for all  $j \in S_2$ . Because  $|S_2| = k$ , this leaves  $\binom{n-k}{m-k}$  choices for  $S_1$  such that  $i \notin S_1 + S_2$ .  $\square$

We now have all the necessary parts to prove Theorem 1.9. Simple combinatorics yields

$$\mathbb{P}[A \text{ has } k \text{ flips}] = \frac{\binom{n}{k} \binom{n}{m-k}}{\binom{2n}{m}}. \quad (52)$$

Combining this with Lemmas 4.1, 4.3, and 4.4 yields the result after some simplification. See Appendix C for complete details.

By plotting  $\mathbb{E}[|A - A|]$  against  $m$  for certain large values of  $n$ , we can see evidence for our intuition from the beginning of this section that  $\mathbb{E}[|A - A|]$  quickly becomes close to  $2n$  for  $m$  on the order of magnitude of  $\sqrt{n}$  (Figure 1). Numerical evidence from many large primes  $n$  suggests that the value of  $m$  such that  $\mathbb{E}[|A - A|] = n$  is about  $1.3875\sqrt{n}$ .<sup>2</sup>

## 5. FUTURE WORK

An immediate future direction of research is to extend the bounds on  $m$  to show Conjecture 1.7 for all  $m$  and  $n$ . However, it is unlikely that the methods we have used in this paper will be useful for larger values of  $m$ . This is because our approach showed that for values of  $m$  that we considered, the majority of subsets of the generalized dihedral group with size  $m$  were MSTD. But this is simply not the case for larger  $m$ : numerical evidence shows that for any  $m \gg \sqrt{n}$ , the vast majority of the sets are balanced. A new approach is required to show that out of the sets that are not balanced, more are MSTD.

To this end, it would also be productive to more carefully analyze how many elements of the group  $D$  are not in  $A + A$  (or not in  $A - A$ ) for  $|A|$  closer to  $n$ . This follows the approach of [hegarty2009almost] for the “slow decay” case. Such an analysis could be used to find explicit

<sup>2</sup>Some code written for this project can be found at <https://github.com/ZeusDM/MSTD-experiments>.

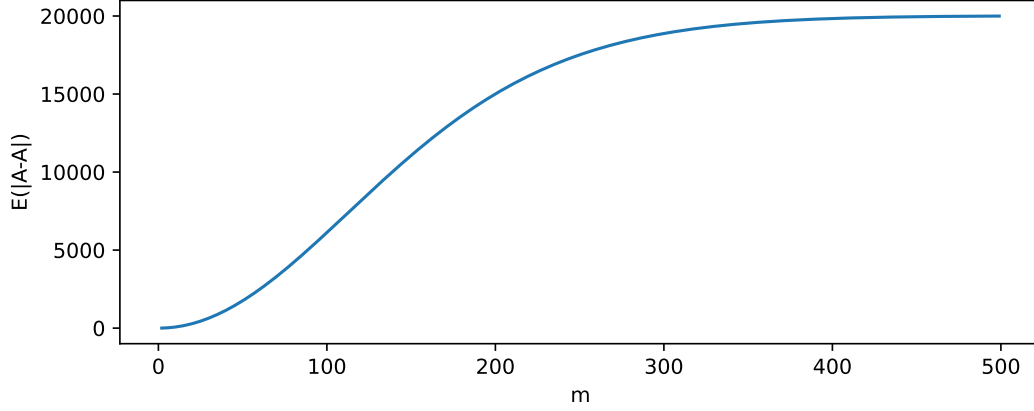


FIGURE 1. A plot of  $\mathbb{E}[|A - A|]$  versus  $m$  for  $A \in \mathcal{S}_{2n,m}$  with  $n = 10007$ . Note that  $\mathbb{E}[|A - A|]$  rapidly approaches  $2n = 20014$  as  $m$  approaches a small multiple of  $\sqrt{n} \approx 100$ .

formulas for the expectation values of  $|A + A|$  and  $|A - A|$ . In this paper we found a formula for  $\mathbb{E}[|A - A|]$ , but only when  $n$  is prime. Further, to use these results, we would also need to bound the variance of  $|A + A|$  and  $|A - A|$ . Depending on the results, this could be enough to prove Conjecture 1.6 if we can deduce an upper bound  $M_\ell$  on the number of MDTs sets and a lower bound  $M_s$  on the number of MSTD sets such that  $M_\ell < M_s$ .

Yet another possible approach to prove Conjecture 1.6 is to construct an injective map from MDTs sets to MSTD sets in the group. Such an approach has proven to be difficult, but has the potential advantage of working for both large and small values of  $m$ .

#### APPENDIX A. PROOF OF LEMMAS 2.1 AND 2.2

**Lemma 2.1** ([haviland2020more]). *Let  $n \geq 3$ , and let  $\mathcal{S}_{2n,2}$  denote the collection of subsets of  $D_{2n}$  of size 2. Then,  $\mathcal{S}_{2n,2}$  has strictly more MSTD sets than MDTs sets.*

*Proof.* There are only 3 possible cases to consider for  $A$ .

- $A$  contains two flip elements: in this case, adding and subtracting flips is an identical operation as each flip element has order 2. Thus,  $A$  will necessarily be sum-difference balanced.
- $A$  contains one flip and one rotation element: let  $r^i, r^j s \in A$ . Here,  $A + A = \{1, r^{2i}, r^{i+j} s, r^{j-i} s\}$ . However,  $A - A = \{1, r^{i+j} s\}$ . When  $r^{2i} \neq e$  or  $r^{j-i} s \neq r^{i+j} s$ ,  $A$  will be MSTD. In the case where both of these are actually equalities ( $A + A = \{e, r^{i+j} s\}$ ),  $A$  will simply be balanced.
- $A$  contains two rotation elements: let  $r^a, r^b \in A$ . Here,  $A + A = \{r^{2i}, r^{i+j}, r^{2j}\}$ . However,  $A - A = \{1, r^{i-j}, r^{j-i}\}$ . Note that  $i \neq j$ . Both the sum set and the difference set have 3 elements except in one special case. Suppose that  $r^{2i} = r^{2j}$ . Then, we have that  $r^i r^{-j} = r^j r^{-i} \implies r^{i-j} = r^{j-i}$ . Thus, when  $A$  contains two rotation elements only,  $A$  is always balanced.

Therefore,  $A$  has strictly more MSTD sets in  $\mathcal{S}_{2n,2}$  than MDTs sets.  $\square$

**Lemma 2.2.** *Let  $n \geq 3$ , and let  $\mathcal{S}_{2n,3}$  denote the collection of subsets of  $D_{2n}$  of size 3. Then,  $\mathcal{S}_{2n,3}$  has strictly more MSTD sets than MDTs sets.*

*Proof.* There are 4 possible cases to consider for  $A$ .

- $A$  contains three flip elements: just as in the  $m = 2$  case, since addition and subtraction of two flips are equivalent,  $A$  is balanced.

- $A$  contains two flip elements and one rotation element: let  $A = \{r^i, r^j s, r^k s\}$  for  $j \neq k$ . Then the sumset is

$$A + A = \{r^{2i}, r^{i+j} s, r^{i+k} s, r^{j-k}, r^{k-j}, 1, r^{j-i} s, r^{k-i} s\}, \quad (53)$$

while the difference set is

$$A - A = \{r^{i+j} s, r^{i+k} s, r^{j-k}, r^{k-j}, 1\}, \quad (54)$$

so  $A - A \subseteq A + A$ . Moreover,  $A$  is MSTD precisely when at least one of  $r^{j-i} s$  and  $r^{k-i} s$  is distinct from each of the flips in  $A + A$  that lie in  $A - A$ . We enumerate all the ways that this could fail to occur:

We have  $r^{i+j} s = r^{j-i} s$  and  $r^{i+k} s = r^{k-i} s$  if and only if  $i = -i \pmod{n}$ . This happens if and only if  $i = 0$  or if  $n$  is even and  $i = n/2$ , and in each of these cases we have  $\binom{n}{2} = \frac{1}{2}(n^2 - n)$  choices for  $j$  and  $k$ . On the other hand, we have  $r^{i+j} s = r^{k-i} s$  and  $r^{i+k} s = r^{j-i} s$  if and only if  $2i = k - j = j - k \pmod{n}$ , which implies that  $2j = 2k \pmod{n}$ . We also require  $2i = 1 \pmod{n}$ ; because  $j \neq k$ , these equations can occur if and only if  $n$  is divisible by 4,  $i = n/4$  or  $3n/4$ , and  $k - j = n/2 \pmod{n}$ . We therefore have 2 choices for  $i$ , and  $\frac{1}{2}n$  choices for  $j$  and one corresponding choice of  $k$  for each  $i$ . Finally, note that we cannot have  $r^{j-i} s = r^{k-i} s$  because  $k \neq i$ , so this completes the enumeration of all such sets  $A$  that fail to be MSTD.

There are  $n\binom{n}{2} = \frac{1}{2}(n^3 - n^2)$  total sets  $A$  containing exactly 2 flips. Subtracting the balanced sets that we just enumerated, we obtain the following numbers of MSTD sets:

$$\begin{aligned} & \frac{1}{2}(n^3 - n^2) \text{ if } n \equiv 1 \pmod{2}, \\ & \frac{1}{2}(n^3 - n^2) - n(n-1) \text{ if } n \equiv 2 \pmod{4}, \\ & \frac{1}{2}(n^3 - n^2) - n(n-1) - n \text{ if } n \equiv 0 \pmod{4}. \end{aligned} \quad (55)$$

- $A$  contains one flip element and two rotation element: let  $A = \{r^i, r^j, r^k s\}$  for  $i \neq j$ . Then the sumset is

$$A + A = \{r^{i+j}, r^{2i}, r^{2j}, r^{i+k} s, r^{j+k} s, r^{k-i} s, r^{k-j} s, 1\}, \quad (56)$$

while the difference set is

$$A - A = \{r^{i-j}, r^{j-i}, r^{i+k} s, r^{j+k} s, 1\}, \quad (57)$$

We show that  $A$  cannot be MDTs by separately comparing the flips and the rotations in  $A + A$  and  $A - A$ . The flips in  $A - A$  are contained in  $A + A$ , so this comparison is trivial. The rotations in  $A - A$  are  $r^{i-j}, r^{j-i}$ , and 1; we will show that  $A + A$  has at least as many rotations. Consider the rotations in  $A + A$ : note that  $r^{i+j}$  is different from  $r^{2i}$  and from  $r^{2j}$  since  $i \neq j$ , so there are at least 2 distinct rotations in  $A + A$ . If  $r^{2i} \neq r^{2j}$ , then in fact  $A + A$  has at least three distinct rotations, and we have  $|A + A| \geq |A - A|$ . On the other hand, if  $r^{2i} = r^{2j}$ , then  $r^{i-j} = r^{j-i}$ , so there are only at most 2 distinct rotations in  $A - A$ , so again  $|A + A| \geq |A - A|$ .

- $A$  contains three rotation elements: for this case, there are  $\binom{n}{3}$  such subsets. We compare directly to Equation (55), the case that yields the least possible number of MSTD sets in  $\mathcal{S}_{2n,3}$ . Here we have

$$\frac{1}{2}(n^3 - n^2) - n(n-1) - n > \frac{1}{6}n(n-1)(n-2) \quad (58)$$

or equivalently,

$$n^3 - 3n^2 - n > 0, \quad (59)$$

which is true if  $n \geq 4$ . When  $n = 3$ , since  $n \equiv 1 \pmod{2}$ , we compare  $\binom{n}{3}$  to  $\frac{1}{2}(n^3 - n^2)$  instead, and the result is verified once again.

Therefore, even if all of the subsets with three rotation elements are MDTs, which is never true to begin with, we still have strictly more MSTd subsets of size 3 than MDTs subsets of size 3.  $\square$

## APPENDIX B. PROOF OF LEMMA 3.9

For any given  $(x, y, F) \in \mathbb{Z}_2 \times \mathbb{Z}_n^2$ , we calculate the number of pairs  $a, b \in \mathbb{Z}_2 \times \mathbb{Z}_n^2$  where  $a - b = (x, y, F)$ . Similarly, we also calculate the number of pairs  $a, b \in \mathbb{Z}_2 \times \mathbb{Z}_n^2$  where  $a + b = (x, y, F)$ . Note that these can be counted through counting the number of pairs in  $\mathbb{Z}_n^2$  which results in  $(x, y)$  as follows:

Case 1,  $F = 0$

- Case 1.1,  $a - b = (x, y, 0)$ 
  - Case 1.1.1  $a$  and  $b$  are not flips.  
We get that this is equivalent to counting pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 - b_1 \equiv x \pmod{n}$  and  $a_2 - b_2 \equiv y \pmod{n}$
  - Case 1.1.2  $a$  and  $b$  are both flips.  
We get that this is equivalent to counting pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 - b_1 \equiv x \pmod{n}$  and  $a_2 - b_2 \equiv y \pmod{n}$
- Case 1.2,  $a + b = (x, y, 0)$ 
  - Case 1.2.1  $a$  and  $b$  are not flips.  
We get that this is equivalent to counting pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 + b_1 \equiv x \pmod{n}$  and  $a_2 + b_2 \equiv y \pmod{n}$
  - Case 1.2.2  $a$  and  $b$  are both flips.  
We get that this is equivalent to counting pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 - b_1 \equiv x \pmod{n}$  and  $a_2 - b_2 \equiv y \pmod{n}$

Case 2,  $F = 1$

- Case 2.1  $a - b = (x, y, 1)$ 
  - Case 2.1.1  $a$  is a flip.  
We get that this is equivalent to counting pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 + b_1 \equiv x \pmod{n}$  and  $a_2 + b_2 \equiv y \pmod{n}$
  - Case 2.1.2  $b$  is a flip.  
We get that this is equivalent to counting pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 + b_1 \equiv x \pmod{n}$  and  $a_2 + b_2 \equiv y \pmod{n}$
- Case 2.2  $a + b = (x, y, 1)$ 
  - Case 2.2.1  $a$  is a flip.  
We get that this is equivalent to counting pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 - b_1 \equiv x \pmod{n}$  and  $a_2 - b_2 \equiv y \pmod{n}$
  - Case 2.2.2  $b$  is a flip.  
We get that this is equivalent to counting pairs  $(a_1, a_2), (b_1, b_2) \in \mathbb{Z}_n^2$  where  $a_1 + b_1 \equiv x \pmod{n}$  and  $a_2 + b_2 \equiv y \pmod{n}$

We get a similar result with  $a, b \in \mathbb{Z}_2 \times \mathbb{Z}_{n^2}$  where  $a - b = (xn + y, F)$  or  $a + b = (xn + y, F)$  as follows:

Case 1,  $F = 0$

- Case 1.1,  $a - b = (xn + y, 0)$

- Case 1.1.1  $a$  and  $b$  are not flips.

We get that this is equivalent to counting pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) - (b_1n + b_2) \equiv xn + y \pmod{n^2}$

- Case 1.1.2  $a$  and  $b$  are both flips.

We get that this is equivalent to counting pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) - (b_1n + b_2) \equiv xn + y \pmod{n^2}$

- Case 1.2,  $a + b = (xn + y, 0)$

- Case 1.2.1  $a$  and  $b$  are not flips.

We get that this is equivalent to counting pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) + (b_1n + b_2) \equiv xn + y \pmod{n^2}$

- Case 1.2.2  $a$  and  $b$  are both flips.

We get that this is equivalent to counting pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) - (b_1n + b_2) \equiv xn + y \pmod{n^2}$

Case 2,  $F = 1$

- Case 2.1  $a - b = (xn + y, 1)$

- Case 2.1.1  $a$  is a flip.

We get that this is equivalent to counting pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) + (b_1n + b_2) \equiv xn + y \pmod{n^2}$

- Case 2.1.2  $b$  is a flip.

We get that this is equivalent to counting pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) + (b_1n + b_2) \equiv xn + y \pmod{n^2}$

- Case 2.2  $a + b = (xn + y, 1)$

- Case 2.2.1  $a$  is a flip.

We get that this is equivalent to counting pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) - (b_1n + b_2) \equiv xn + y \pmod{n^2}$

- Case 2.2.2  $b$  is a flip.

We get that this is equivalent to counting pairs  $a_1n + a_2, b_1n + b_2 \in \mathbb{Z}_{n^2}$  where  $(a_1n + a_2) + (b_1n + b_2) \equiv xn + y \pmod{n^2}$

Note that the signs are the same between corresponding cases in  $\mathbb{Z}_2 \times \mathbb{Z}_{n^2}$  and  $\mathbb{Z}_2 \times \mathbb{Z}_n^2$ . And thus there are only four distinct cases we need to count, and it suffices to compare the number of 4-tuple collisions over all  $(x, y) \in \mathbb{Z}_n^2$  with the number of 4-tuple collisions over all  $xn + y \in \mathbb{Z}_{n^2}$

We will prove the main result by counting the collisions of each of the 3 types. When  $n$  is odd, it is simple to see from Lemma 3.10, Lemma 3.11, and Lemma 3.12 that the number of collisions for each of the  $(x, y)$  in  $\mathbb{Z}_n^2$  is the same for the corresponding  $xn + y \in \mathbb{Z}_{n^2}$  and thus there are an equal total number of collisions. We now assume that  $n$  is even.

- $a + b = c + d$

Suppose  $y$  is odd, from Lemma 3.11 we have  $(\frac{n^2}{2})^2$  pairs for each  $(x, y)$  for a total of  $\frac{n^5}{4}$  collisions. From Lemma 3.12, we also get  $(\frac{n^2}{2})^2$  pairs for each  $xn + y$  for an equal total  $\frac{n^5}{4}$  collisions.

If  $y$  is even, we have a half of  $(x, y)$  yielding  $(\frac{n^2}{2})^2$  pairs and another half yielding  $(\frac{n^2+4}{2})^2$  pairs for a total of  $\frac{n}{2} * \frac{n^4+n^4+8n^2+16}{4} = \frac{n^5+4n^3+8n}{4}$ . Meanwhile in  $\mathbb{Z}_{n^2}$  we have each  $xn + y$  yielding  $(\frac{n^2+2}{2})^2$  for an equal total of  $\frac{n^5+4n^3+4n}{4}$ . Thus, there are strictly more collisions in  $\mathbb{Z}_n^2$  in this case.

- $a + b = c - d$

Suppose  $y$  is odd, we have  $\frac{n^2}{2}$  choices for  $a + b$  and  $n^2$  choices for  $c - d$  for a total of  $\frac{n^5}{2}$ . We also get the same results for  $\mathbb{Z}_{n^2}$ .

If  $y$  is even, we have half of  $(x, y)$  yielding  $\frac{n^2}{2} * n^2$  pairs and another half yielding  $\frac{n^2+4}{2} * n^2$  pairs for a total of  $\frac{n}{2} * \frac{n^4+n^4+4n^2}{2} = \frac{n^5+2n^3}{2}$ . For  $\mathbb{Z}_{n^2}$ , we have each  $(x, y)$  yielding  $\frac{n^2+2}{2} * n^2$  collisions for a total of  $\frac{n^5+2n^3}{2}$

- $a - b = c - d$

Both groups have the same number of collisions for each  $(x, y)$  and the corresponding  $xn+y$  at  $n^4$ . Thus, the total number of collisions in this case is  $n^5$ .

Thus, if we look over all the cases where  $n$  is even, we found that either there are strictly more collisions in  $\mathbb{Z}_n^2$  or there are equal amount of collisions. Thus, we can conclude that there are strictly more collisions in  $\mathbb{Z}_2 \ltimes \mathbb{Z}_n^2$  when  $n$  is even.

### APPENDIX C. PROOF OF THEOREM 1.9

We have

$$\begin{aligned}
\mathbb{E}[|A - A|] &= |D_{2n}| - \sum_{g \in D_{2n}} \mathbb{P}[g \notin A - A] \\
&= 2n - \mathbb{P}[1 \notin A - A] - \left[ \sum_{i=1}^{n-1} \mathbb{P}[r^i \notin A - A] \right] - \left[ \sum_{i=0}^{n-1} \mathbb{P}[r^i s \notin A - A] \right] \\
&= 2n - 0 - \left[ \sum_{i=1}^{n-1} \sum_{k=0}^m \frac{\mathbb{P}[A \text{ has } k \text{ flips}]}{\mathbb{P}[i \notin S - S | S \subseteq \mathbb{Z}_n, |S| = m - k]} \right] \\
&\quad - \left[ \sum_{i=0}^{n-1} \sum_{k=0}^m \frac{\mathbb{P}[A \text{ has } k \text{ flips}]}{\mathbb{P}[i \notin S_1 + S_2 | S_1, S_2 \subseteq \mathbb{Z}_n, |S_1| = m - k, |S_2| = k]} \right] \\
&= 2n - \left[ \sum_{i=1}^{n-1} \sum_{k=0}^m \frac{\binom{n}{k} \binom{n}{m-k}}{\binom{2n}{m}} \cdot \frac{g(n, k)}{\binom{n}{k}} \cdot \frac{g(n, m-k)}{\binom{n}{m-k}} \right] \\
&\quad - \left[ \sum_{i=0}^{n-1} \sum_{k=0}^m \frac{\binom{n}{m-k}}{\binom{2n}{m}} \cdot \frac{\binom{n-k}{m-k}}{\binom{n}{m-k}} \right] \\
&= 2n - \left[ (n-1) \sum_{k=0}^m \frac{g(n, k) \cdot g(n, m-k)}{\binom{2n}{m}} \right] - \left[ n \sum_{k=0}^m \frac{\binom{n}{k} \binom{n-k}{m-k}}{\binom{2n}{m}} \right] \\
&= 2n - \left[ 2(n-1) \frac{g(n, 0) \cdot g(n, m)}{\binom{2n}{m}} + (n-1) \sum_{k=1}^{m-1} \frac{\frac{n \binom{n-1-k}{k-1}}{k} \cdot \frac{n \binom{n-1-m+k}{m-k}}{m-k}}{\binom{2n}{m}} \right] \\
&\quad - \left[ \frac{n}{\binom{2n}{m}} \sum_{k=0}^m \binom{n}{k} \binom{n-k}{m-k} \right] \\
&= 2n - \frac{nm2^m \binom{n}{m} + 2n(n-1) \binom{n-m-1}{m-1}}{m \binom{2n}{m}} - \frac{n^2(n-1)}{\binom{2n}{m}} \sum_{k=1}^{m-1} \frac{\binom{n+k-m-1}{m-k-1} \binom{n-k-1}{k-1}}{k(m-k)}, \quad (60)
\end{aligned}$$

by the combinatorial identity

$$\sum_{k=0}^m \binom{n}{k} \binom{n-k}{m-k} = 2^m \binom{n}{m}. \quad (61)$$