Generating and generalizing MSTD sets through Markov processes

Frank He (fsh30@cam.ac.uk) **Karol Daniewski (kkd2@williams.edu)**
Azmera Gebre, Mateo Palomares
Mentor: Steven J Miller

Supported by NSF, Williams College,
Winston Churchill Foundation, Finnerty Fund

Affiliated with University of Cambridge, Williams College, University of Michigan

Yale
July 21, 2026

Williams SMALL
Probability and Number Theory

Sum and difference sets

Let A be a subset of the integers. We define its sum set and difference set as follows:

- $A + A = \{a_1 + a_2 : a_1, a_2 \in A\}$
- $A - A = \{a_1 - a_2 : a_1, a_2 \in A\}$

We may compare their sizes, and call A

- *Sum dominant* if $|A + A| > |A - A|$
- *Difference dominant* if $|A + A| < |A - A|$
- *Balanced* if $|A + A| = |A - A|$

Introduction to MSTD sets

More Sums Than Differences (MSTD) problems studies finite sets $A \subset \{0, 1, \ldots, n\}$ for which A is *sum dominant*.

Addition commutes, so expect most sets to be *difference dominant*,

$$a + b = b + a$$

so sums repeated. However, subtraction is not commutative

$$a - b \neq b - a$$

so more differences produced. This is true, but some sets are still sum dominant [MO06].

Differences dominate on average

Martin and O'Bryant chose set $A \subseteq \{0, \ldots, n\}$ uniformly, and discovered that in the limit $n \to \infty$, the average $|A - A| - |A + A|$ over all A is 4.

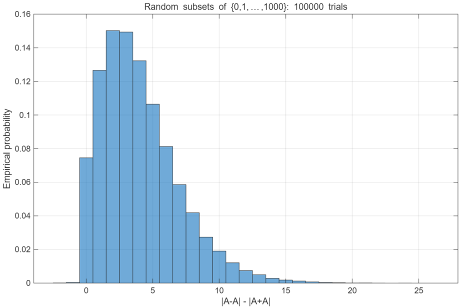


Figure 1: Simulating expected differences

MSTD sets are still likely

Martin and O'Bryant in [MO06] proved a positive proportion of sets are MSTD. Hegarty, in [Heg07], gave explicit constructions of MSTD sets. Zhao in [Zha11] showed the limiting proportions exist.

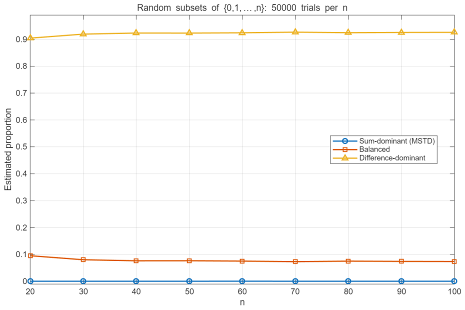


Figure 2: Convergence of proportions

Extending to further probability models

In [Do+15], authors considered subsets A and B of $\{0, \ldots, n\}$, where B chosen with probability dependent on A.

They considered $A + B$ and $A - B$, and found that for fixed, non-degenerate probabilities, MSTD sets occur with positive limiting probabilities.

The original problem is recovered as a special case of this setup.

We explore the problem of the choice of elements in A being dependent on the set itself.

Define a Markov chain

We generate A with a Markov chain.

A Markov chain is a sequence of random variable $\{X_i\}_{i\geq 0}$ such that

$$\mathbb{P}(X_{i+1} = x_{i+1} \mid X_i = x_i) = \mathbb{P}(X_{i+1} = x_{i+1} \mid X_i = x_i, \ldots X_0 = x_0)$$

Introduction to the Markov-chain

Here is our Markov chain.

Let $(X_i)_{i=0}^n$ be a stationary two-state Markov chain with transition matrix

$$P = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}, \text{ where } p, q \in (0,1)$$

and define

$$A = \{i \in 0, \ldots, n : X_i = 1\}.$$

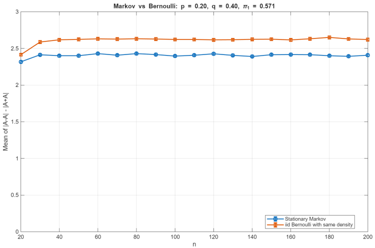
We note that $p = q = 1/2$ recovers the original problem, and $p = 1 - q$ gives an independent probability q for each element to be included in A.

Observations

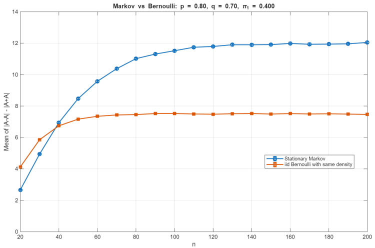
A simple computation gives the following stationary distribution

$$\pi = \left(\frac{1-q}{2-p-q}, \frac{1-p}{2-p-q} \right)$$

However, the behavior of the chain (in blue) is very different from the Bernoulli(π_1) case (in orange), as we can see in the following simulations:



(a) States frequently alternate



(b) States frequently cluster

Many Markov Sets Have More Sums

We extend Zhao's Theorem for MSTD sets to the Markov case, using the following Theorem:

Theorem 1

Let $A_n \subseteq S := \{0, \ldots, n\}$ be generated by Markov process $\{X_i\}$. Let $\rho_+^{(n)}, \rho_-^{(n)}$, and $\rho_0^{(n)}$ be the probabilities that A_n is Sum-dominant, Difference-dominant, and Balanced respectively. As $n \to \infty$, $\rho_+^{(n)}, \rho_-^{(n)}$, and $\rho_0^{(n)}$ tend to strictly positive limits.

We first need to prove that the limits exist, then we need to show that they are positive.

Limiting proportions exist

Define left and right fringes of A_n as

$$L = A_n \cap [0, K] \qquad\qquad R = (n - A_n) \cap [0, K]$$

And define Fringe Score

$$\Delta(L, R) = |(L + L) \cap [0, K]| + |(R + R) \cap [0, K]| - 2|(L + R) \cap [0, K]|$$

Limiting proportions exist

Define left and right fringes of A_n as

$$L = A_n \cap [0, K] \qquad\qquad R = (n - A_n) \cap [0, K]$$

And define Fringe Score

$$\Delta(L, R) = |(L + L) \cap [0, K]| + |(R + R) \cap [0, K]| - 2|(L + R) \cap [0, K]|$$

Lemma 2 (adapted from [Zha11])

Let the middles of $A_n + A_n$ and $A_n - A_n$ be full, i.e.
$[K + 1, 2n - K - 1] \subseteq A_n + A_n$ and $[K + 1 - n, n - K - 1] \subseteq A_n - A_n$, then

$$\Delta = |A_n + A_n| - |A_n - A_n|$$

Proof of Lemma 2 (sketch)

1. **Look at left and right fringes.** By assumption, we only need to consider the size $K+1$ fringes of $A_n + A_n$ and $A_n - A_n$.
2. **Write in terms of L and R.**

$$|(A_n + A_n) \cap [0, K]| = |(L + L) \cap [0, K]|$$
$$|(A_n + A_n) \cap [2n - K, K]| = |(R + R) \cap [0, K]|$$

3. **Do the same with $A_n - A_n$.** Subtracting the two expressions for fringe sizes gives the result.

Limiting proportions exist

We extend the idea from [Zha11] that most sets have full middles.

Limiting proportions exist

We extend the idea from [Zha11] that most sets have full middles.

Lemma 3

Let $F_{K,n}$ be the event that $[K+1, 2n-K-1] \subseteq A_n + A_n$ and $[K+1-n, n-K-1] \subseteq A_n - A_n$, since we know their middles are full. There exists constants $C > 0$, $0 < \theta < 1$, depending only on p and q, such that for every $K < n/2$,

$$\mathbb{P}(F_{K,n}^c) \leq C\theta^K$$

Limiting proportions exist

We extend the idea from [Zha11] that most sets have full middles.

Lemma 3

Let $F_{K,n}$ be the event that $[K+1, 2n-K-1] \subseteq A_n + A_n$ and $[K+1-n, n-K-1] \subseteq A_n - A_n$, since we know their middles are full. There exists constants $C > 0$, $0 < \theta < 1$, depending only on p and q, such that for every $K < n/2$,

$$\mathbb{P}(F_{K,n}^c) \leq C\theta^K$$

Use the fact that there is some minimum probability $\delta > 0$ such that any two states X_i, X_j are 1.

Limiting proportions exist

Proving that limiting proportions exist is a direct application of the previous Lemma. We take note:

$$\begin{aligned}\left|\mathbb{P}(\Delta > 0) - \mathbb{P}(\text{MSTD})\right| &\leq \mathbb{P}\big(\text{Middle not full}\big)\\ &\leq C\theta^K\end{aligned}$$

Limiting probability as $n \to \infty$ for each fringe to occur by asymptotic independence. Thus, there is a limiting probability that $\Delta > 0$.

Therefore, the probability that a set is MSTD also has a limit, and the same argument works for MDTS and balanced cases.

MSTD sets have positive limiting probability

Here is a fringe that has positive limiting probability and produces $\Delta > 0$. For $K \geq 12$,

$$L_K^+ = \{0, 2, 3, 7, 8, 9, 10\} \cup [12, K] \quad R_K^+ = \{1, 2, 3, 6, 8, 9, 10, 11\} \cup [12, K]$$

produces $\Delta = 1$, which is an MSTD construction.

We can prove that Lemma 3 works even when conditioning on a specific fringe, which occurs with positive probability.

Hence, MSTD sets must occur with positive limiting probability.

We apply other fringes to prove the MDTS and balanced cases.

Expected difference between sum sets and difference sets

In the sets generated by Markov chain we show that differences dominate sums.

Theorem 4

Let $A_n \subseteq S\{0, \ldots, n\}$ be generated by Markov process $\{X_i\}$. As $n \to \infty$

$$\mathbb{E}\big[|A_n - A_n| - |A_n + A_n|\big] \to \Lambda(p, q) = \frac{2(1-q)(1+p-p^2-pq)}{(1-p)^2(p+q)(2-p-q)}$$

First simplify this problem to missing sums and then use Markov properties to finalize it. Note $p = q = 1/2$ recovers the original case.

Expected difference between sum sets and difference sets (proof sketch)

Define $\mathbb{E}\big[|A_n - A_n| - |A_n + A_n|\big] = E_n$. Firstly, let's translate this problem to using missing sums and differences. Notice that

$$|A_n - A_n| - |A_n + A_n| = |(A_n + A_n)^c| - |(A_n - A_n)^c|$$

because $A_n + A_n$ and $A_n - A_n$ live in $[0, 2n]$ and $[-n, n]$ respectively, and $|[-n, n]| = |[0, 2n]|$.

Expected difference between sum sets and difference sets (proof sketch)

Define $\mathbb{E}\big[|A_n - A_n| - |A_n + A_n|\big] = E_n$. Firstly, let's translate this problem to using missing sums and differences. Notice that

$$|A_n - A_n| - |A_n + A_n| = |(A_n + A_n)^c| - |(A_n - A_n)^c|$$

because $A_n + A_n$ and $A_n - A_n$ live in $[0, 2n]$ and $[-n, n]$ respectively, and $|[-n, n]| = |[0, 2n]|$. By linearity of expectation, we obtain

$$E_n = \mathbb{E}[\textbf{missing sums}] - \mathbb{E}[\textbf{missing differences}]$$

Expected difference between sum sets and difference sets (proof sketch)

Define probability of missing k in $A + A$ as

$$u_k = \mathbb{P}(k \notin A_n + A_n),$$

probability does not depend on n. Furthermore,

$$u_{2n-k} = \mathbb{P}(2n - k \notin A_n + A_n) = u_k$$

Expected difference between sum sets and difference sets (proof sketch)

Define probability of missing k in $A + A$ as

$$u_k = \mathbb{P}(k \notin A_n + A_n),$$

probability does not depend on n. Furthermore,

$$u_{2n-k} = \mathbb{P}(2n - k \notin A_n + A_n) = u_k$$

Define the probability of $n - k$ not included in $A - A$ as

$$v_{n,k} = \mathbb{P}(n - k \notin A_n - A_n)$$

The difference $n - k$ is missing exactly when there is no $i \in [0, k]$ such that $X_i = X_{n-k+i} = 1$

Expected difference between sum sets and difference sets (proof sketch)

Bringing everything together we get

$$\mathbb{E}_n = 2\sum_{i=0}^{n-1}(u_k - v_{n,k}) + u_n - \pi_0 p^n$$

Note $\mathbb{P}(0 \notin A_n - A_n) = \mathbb{P}(A_n = \emptyset) = \pi_0 p^n$.

Expected difference between sum sets and difference sets (proof sketch)

$$\mathbb{E}_n = 2\sum_{i=0}^{n-1}(u_k - v_{n,k}) + u_n - \pi_0 p^n$$

By the Convergence Theorem, for fixed k, blocks $X_0, \ldots, X_k$ and $X_n, X_{n-1}, \ldots, X_{n-k}$ are asymptotically independent as $n \to \infty$. Defining w_k as the probability that k is missing from the sum of two independent Markov chains will give us

$$v_{n,k} \to w_k, \text{ as } n \to \infty$$

Therefore,

$$\lim_{n\to\infty} \mathbb{E}_n = 2\sum_{k=0}^{\infty}(u_k - w_k)$$

Expected difference between sum sets and difference sets (proof sketch)

We calculate u_{2m+1} and u_{2m} separately. For $k = 2m$, we start the Markov chain in (m, m) and let it run to positions $(m-1, m+1), (m-2, m+2), \dots$ and compute their probabilities. We repeat for u_{2m+1} and find the infinite sum.

Computation for w_k is a direct combination of two identical Markov chains in parallel.

Matrix algebra yields

$$\lim_{n\to\infty} E_n = \frac{2(1-q)(1+p-p^2-pq)}{(1-p)^2(p+q)(2-p-q)}$$

Expected difference converges very quickly

When we ran simulations of the Markov chain, we found that, even for relatively small n (such as 1000), our computational results closely matched what our theorems predicted. In fact, this convergence rate is expected.

Expected difference converges very quickly

When we ran simulations of the Markov chain, we found that, even for relatively small n (such as 1000), our computational results closely matched what our theorems predicted. In fact, this convergence rate is expected.

Theorem 5

There exists a constant $\lambda < 1$ dependent only on p and q such that

$$E_n = \Lambda(p, q) + O(n\lambda^n)$$

Convergence rate (proof sketch)

We recall that

$$E_n = 2\sum_{k=0}^{n-1}(u_k - v_{n,k}) + u_n - \pi_0 p^n$$

$$\Lambda(p,q) = 2\sum_{k=0}^{\infty}(u_k - w_k)$$

By the triangle inequality,

$$|E_n - \Lambda| \leq u_n + \pi_0 p^n + 2\sum_{k\geq n}(u_k + w_k) + 2\sum_{k=0}^{n-1}|v_{n,k} - w_k|$$

It remains to show that each term is bounded.

Convergence rate (proof sketch)

1. **First three terms.** Observe all Markov transition probabilities are positive, so there is a minimum probability $0 < \delta < 1$ that each summand or difference pair producing k is included in A_n. Hence, we may find some $0 < \alpha < 1$ such that
$$u_k + w_k = O(\alpha^k)$$
Hence, if we WLOG choose $\alpha \leq p$,
$$u_n + \pi_0 p^n + 2 \sum_{k \geq n} (u_k + w_k) = O(\alpha^n)$$

It remains to bound the final term.

Convergence rate (proof sketch)

2. **Final term.** We need to bound $\sum_{k=0}^{n-1} |v_{n,k} - w_k|$, which we may do by using the geometric convergence to stationarity of our finite-state chain, which states that there is some $0 < \beta < 1$ such that

$$|(P^j)_{ab} - \pi_b| = O(\beta^j)$$

   a. **Case $k \leq n/3$:** Endpoint blocks $X_0, \ldots, X_k$ and $X_{n-k}, \ldots, X_n$ have gap $n/3$, so asymptotically independent.

      The difference $|v_{n,k} - w_k|$ measures the difference the upper block has from stationarity, hence,

$$|v_{n,k} - w_k| = O(\beta^{n/3})$$

Convergence rate (proof sketch)

b. **Case $k > n/3$:** Cannot use the gap argument, but we can bound $v_{n,k}$ and w_k separately.
   There are $k+1$ pairs that can produce a difference of $n-k$, of which at least $(k+1)/2$ pairs must be distinct. $k > n/3$, so

$$v_{n,k} \leq (1-\delta)^{n/6}$$

   All pairs considered in the process of w_k are disjoint, so simply

$$w_k \leq (1-\delta)^{n/3}$$

c. **Sum over all possible k:** There are only $n+1$ possible values for k, which produces the factor of n in the convergence.

Hence, $O(n\lambda^n)$ is the rate of convergence.

References

[Do+15] Thao Do et al. “Sums and Differences of Correlated Random Sets”. In: *Journal of Number Theory* 147 (2015), pp. 44–68. DOI: 10.1016/j.jnt.2014.06.022.

[Heg07] Peter V. Hegarty. “Some Explicit Constructions of Sets with More Sums than Differences”. In: *Acta Arithmetica* 130.1 (2007), pp. 61–77. DOI: 10.4064/aa130-1-4.

[MO06] Greg Martin and Kevin O’Bryant. “Many Sets Have More Sums Than Differences”. In: *Additive Combinatorics* (2006).

[Zha11] Yufei Zhao. “Sets Characterized by Missing Sums and Differences”. In: *Journal of Number Theory* 131.11 (2011), pp. 2107–2134. DOI: 10.1016/j.jnt.2011.05.003.

Thank You!

Frank He (fsh30@cam.ac.uk) Karol Daniewski (kkd2@williams.edu)
Azmera Gebre, Mateo Palomares
Mentor: Steven J Miller

Contact: fsh30@cam.ac.uk, kkd2@williams.edu

Yale
July 21, 2026

Williams SMALL

Probability and Number Theory