Limiting Behavior in Missing Sums of Sumsets

Aditya Jambhale (adijambhale@gmail.com), Rauan Kaldybayev (rk19@williams.edu), Chris Yao (chris.yao@yale.edu);
Advisor: Steven J. Miller (sjm1@williams.edu)
Number Theory and Probability Group - SMALL 2023 - Williams College

1. Enter sumsets

1.1 Sum and difference sets in number theory

Many of the most important problems in number theory concern the sum or difference set of a given set of integers. For example, Goldbach's conjecture states that if P is the set of all primes, $P+P$ includes all even numbers starting from four; the Twin Prime Conjecture states that two is represented infinitely often in $P-P$; and Fermat's Last Theorem states that if S_k is the set of k-th powers of positive integers, the intersection of $S_k + S_k$ with S_k is empty whenever $k > 2$.

1.2 The problem we investigate

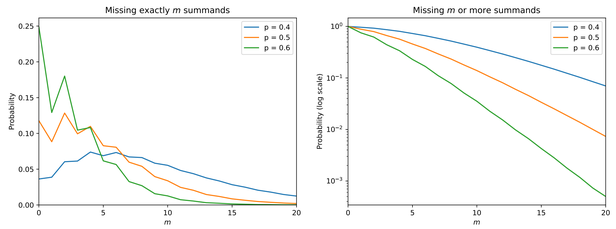
Pick a number $0 < p < 1$. A set of nonnegative integers $A \subseteq \mathbb{Z}_{\geq 0}$ is chosen randomly such that for every n, the probability of $n \in A$ is equal to p:

$$\mathbb{P}(n \in A) = p \quad \forall n \in \mathbb{Z}_{\geq 0}.$$

We investigate the sum set $A + A$,

$$A + A \coloneqq \{x + y : x, y \in A\}.$$

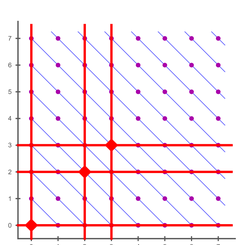
Our main object of study is $\mathbb{P}(|A+A|^c = m)$, the probability of missing exactly m summands as a function of m. The figure below provides numerically computed $\mathbb{P}(|A+A|^c = m)$, as well as the "cumulative distribution function" $\mathbb{P}(|A+A|^c \geq m)$, for three different values of p.



The most interesting theorems in mathematics are those that are "out there," in plain sight, waiting to be "discovered," not "created." Our problem boasts a number of interesting patterns. Missing $0, 2, 4, \ldots$ summands appears to be more likely than missing $1, 3, 5, \ldots$ summands; is there a reason why? The plot of $\log(\mathbb{P}(|A+A|^c \geq m))$ against m is practically linear; why is that, and what is its slope? What is the likelihood of missing zero summads, that is, of having $A + A = \mathbb{Z}_{\geq 0}$? We calculate the mean and variance of the number of missing summands and provide bounds for other relevant quantities.

1.3 Missing sums and annihilation

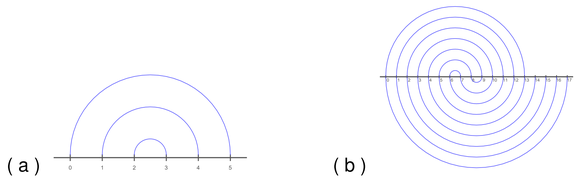
Imagine the quarter-infinite grid $\mathbb{Z}_0^2$. Whenever a number $k \in \mathbb{Z}_0$ is not in A, a fictional "orbital canon" fires at the point (k, k) and annihilates the corresponding column and row. For any integer n, if all of the squares $(n, 0), (n-1, 1), \ldots, (0, n)$ are annihilated, n is missing from $A + A$. In the figure below, the canon fires at $0, 2, 3$, and one can see that $0, 1, 3, 4$ are missing from $A + A$.



In order for a number n to be missing from $A + A$, $n + 1$ squares must be annihilated. As n gets large, the probability $\mathbb{P}(n \notin A + A)$ decays exponentially. So even though A in general misses infinitely many integers, $A + A$ will on average only miss finitely many.

2. Our methods and results

2.1 Circle diagrams



In order for an integer n to be missing from $A + A$, for every pair $(0, n), (1, n-1), \ldots, (n, 0)$, at least one of the two numbers in the pair must be missing from A. Figure (a) depicts the condition for $5 \notin A + A$, and figure (b) depicts the condition for $13, 17 \notin A + A$. Figure (a) splits into three semicircles, and figure (b) splits into two spirals. In general, a circle diagram for one or two numbers is a combination of one or more "chains."

2.2 The likelihood of missing one or two given integers

Using circle diagrams, we were able to compute exact formulas for the probabilities of missing one or two given summands. These formulas have a piecewise character. The probability of missing any given integer n is

$$\mathbb{P}(n \notin A + A) = F_2^{\frac{n+1}{2}} \begin{cases} 1 & n \text{ odd} \\ \sqrt{\frac{1-p}{1+p}} & n \text{ even} \end{cases}, \tag{1}$$

and the probability of missing any two given integers m, n with $m > n$ is

$$\mathbb{P}(m, n \notin A + A) = \omega(m, n, l)\kappa(m, n, l). \tag{2}$$

Here, $l = \lfloor \frac{m+1}{m-n} \rfloor$, $s = (m, n, l) \mod 2$ encodes the parities of m, n, l,

$$\kappa(m, n, l) = \begin{cases} 1 & s = 110 \text{ or } 111 \\ (1-p)F_{l-1}/\sqrt{F_{2l}} & s = 100 \text{ or } 011 \\ (1-p)F_l/\sqrt{F_{2l+2}} & s = 101 \text{ or } 010 \\ (1-p)^2 F_{l-1}F_l/\sqrt{F_{2l}F_{2l+2}} & s = 000 \text{ or } 001 \end{cases}$$

is a factor close to one, $\omega(m, n, l) = F_{2l}^{l^{\frac{m-n}{2}} - \frac{n+1}{2}} F_{2l+2}^{\frac{m+1}{2} - l^{\frac{m-n}{2}}}$ is a piece decaying exponentially with m and n, and (F_l) are a generalization of the Fibonaccis:

$$\begin{aligned} F_0 &= F_1 = 1, \\ F_{l+2} &= (1-p)F_{l+1} + p(1-p)F_l. \end{aligned} \tag{3}$$

When $p = 1/2$, F_l is the $(l+1)$-st Fibonacci number divided by 2^l. The F_l's are strictly between 0 and 1, and they decay exponentially. Qualitatively, F_l is the probability that if a string of zeros and ones of length l is chosen at random with $\mathbb{P}(1) = p$, there will be no two consecutive ones; this quantity arises because circle diagrams always split into disjoint "chains."

2.3 Mean and variance

Let $Y = |(A + A)^c|$ be the random variable for the number of nonnegative integers missing from $A + A$. Summing equation 1 yields

$$\mathbb{E}(Y) = \frac{2}{p^2} - \frac{1}{p} - 1. \tag{4}$$

From equation 2, we compute the second moment to be

$$\mathbb{E}\left(Y^2\right) = -\left(\frac{2}{p^2} - \frac{1}{p} - 1\right) + 2\sum_{l=1}^{\infty} \frac{F_{2l} + (1-p)F_{l-1} + (1-p)F_lF_{2l} + (1-p)^2F_{l-1}F_l}{(1-F_{2l})(1-F_{2l+2})}. \tag{5}$$

The sum, though hard to analyze, converges very quickly, since exponentially its terms decay. Once the first two moments are known, the variance is calculated as $\mathbb{E}\left(Y^2\right) - \mathbb{E}(Y)^2$. For $p = 1/2$, the expectation value of Y is 5, and the variance is 17.9829.

2.4 Establishing the decay rate of Y

Using equation 1 and observing that $\mathbb{P}(U \cap V) \leq \mathbb{P}(V)$ for any events U and V, one can obtain a crude but surprisingly nice upper bound on the k-th moment of Y:

$$\begin{aligned} \mathbb{E}\left(Y^k\right) &= \sum \mathbb{P}(n_1, \ldots, n_k \notin A + A) \leq \\ &\leq \sum \mathbb{P}(\max\{n_1, \ldots, n_k\} \notin A + A) \leq \left(1 + \frac{\alpha}{\sqrt{2\pi}}\right) \frac{k!}{\alpha^k}. \end{aligned} \tag{6}$$

(Here, $\alpha \coloneqq \log \frac{1}{\sqrt{1-p^2}}$.) Applying Chernoff's inequality, we obtain an exponential upper bound on the probability of missing more than m summands in $A + A$:

$$\mathbb{P}(Y \geq m) \leq \left(1 + \frac{\alpha}{\sqrt{2\pi}}\right) m\, e^{-\alpha m + 1} = O\left(m\, e^{-(\log \frac{1}{\sqrt{1-p^2}})m}\right). \tag{7}$$

(This bound is valid whenever $m > 1/\alpha$. For $p = 1/2$, $\alpha = 0.1438$ and $1/\alpha = 6.9521$.) One cannot bound $\mathbb{P}(Y \geq m)$ tighter than exponential. If $0, \ldots, m/2$ are missing from A, then $0, \ldots, m$ are missing from $A + A$. Therefore, for even m,

$$\mathbb{P}(Y \geq m) \geq \mathbb{P}(0, \ldots, m \notin A + A) \geq \mathbb{P}(0, \ldots, m/2 \notin A) = (1-p)^{m/2} = \Omega\left(e^{-(\log \frac{1}{\sqrt{1-p}})m}\right). \tag{8}$$

Equations 7 and 8 together establish an approximate decay rate for $\mathbb{P}(Y \geq m)$. Bounded above and below by two exponential functions, $\mathbb{P}(Y \geq m)$ must itself be "approximately exponential."

3. Future work

3.1 Tightening the upper bound

Monte Carlo simulations allow us to compute $\mathbb{P}(Y \geq m)$ as a function of m. Plotting $\log(\mathbb{P}(Y \geq m))$ against m yields an approximately linear graph whose slope we call the *decay rate*. For $p = 1/2$, the decay rate was estimated to be around -0.3. This is very close to the theoretical lower bound $\log\sqrt{1-p} = -0.3466$ and quite far from the upper bound $\log\sqrt{1-p^2} = -0.1438$. Therefore, tightening equation 7 is an important direction of future work. One immediate improvement can be obtained by using equation 2 instead of 1.

3.2 Calculating the even-odd disparity

The difference in probabilities of missing an even/odd numbers of summands can be quantified using Euler's identity:

$$\mathbb{P}(Y \text{ even}) - \mathbb{P}(Y \text{ odd}) = \mathbb{E}\left((-1)^Y\right) = \mathbb{E}\left(e^{i\pi Y}\right) = \mathbb{E}(\cos \pi Y).$$

Since the moment generating function $M(t) = \mathbb{E}\left(e^{tY}\right)$ is analytic, the tools of complex analysis, including Cauchy's integral formula and series approximations, could potentially be used to evaluate $\mathbb{P}(Y \text{ even}) - \mathbb{P}(Y \text{ odd})$.

3.3 Towards a closed-form expression for variance

The difficulty of getting a closed-form expression for the variance of the number of missing summands in $A + A$ boils down to understanding the function

$$H(x) = \sum_{n=1}^{\infty} \frac{x^n}{(1-F_{2n})(1-F_{2n+2})}.$$

It is not entirely unimaginable that this sum has a closed-form expression.

3.4 The k-th additive power of A

Define A^{+k} to be the k-fold sum of A with itself, $A + A + \cdots + A$ with k summands. What is the probability distribution for the number of missing summands in A^{+k}? We have investigated this for $k = 2$. Another possible object of study is $A^{+\infty} = \{0\} \cup A \cup A^{+2} \cup \ldots$, the set of all numbers expressible as an arbitrary sum of elements of A.

4. Acknowledgements

Financial support was provided by Williams College and the U.S. National Science Foundation, grant numbers DMS2241623 and DMS1947438.