Sum Constructions

Jacob Richey

February 19, 2023

The question we investigate in this note is the following:

Question 0.1. *Let $S \subset [0,1]$ be any set of reals. Describe the set of $k-$sums, or countable sums, namely*

$$kS = \{s_1 + s_2 + \cdots + s_k : s_1, s_2, \ldots, s_k \in S\} \tag{0.1}$$

or

$$\mathrm{sum}(S) = \left\{ \sum_{s \in T} s : T \text{ a countable subset of } S \right\}. \tag{0.2}$$

In particular, assuming the sumsets are measurable, what is the measure of kS? Of sum(S)?

(In general, the sumset need not be measurable. If S is countable, then any of the sumsets is also measurable.)

One motivation is the following: suppose you have access to supply of independent samples X of some discrete distribution F, say Poisson(1).

Question 0.2. *How many different samples do you need to simulate a Bernoulli(p) event for some $p \in (0,1)$?*

Alternatively:

Question 0.3. *For which p can you simulate a Bernoulli(p) with just one sample from F? With two? With k?*

Without loss of generality, we can assume F takes the form

$$F = \sum_{n \in \mathbb{N}} s_n \delta_n. \tag{0.3}$$

Simulating a Bernoulli(p) event is equivalent to finding an event A, measurable with respect to X, such that $\mathbb{P}(A) = p$. Since F is discrete, this is the same as finding a subset $I \subset \mathbb{N}$ such that

$$\sum_{i \in I} s_i = p. \tag{0.4}$$

There is some literature on the *Bernoulli factory problem*: how do you use an infinite supply of Bernoulli(p)'s to generate a Bernoulli($f(p)$), where f is known but p is unknown? There may be some work on other distributions, but I'm not sure.

1 Generating Bernoullis from a single sample

The following will turn out to be a relevant condition for discrete probability measures. Let $S = \{s_n\}_{n\in\mathbb{N}}$ be a countable probability distribution, i.e. $s_n \geq 0$ and

$$\sum_n s_n = 1 \tag{1.1}$$

By convention we always write our countable sets S in decreasing order, so that $s_1 \geq s_2 \geq \ldots$.

Definition 1.1. *S has the* small tail sum property at level n *if*

$$s_n > \sum_{m>n} s_m. \tag{1.2}$$

If the above holds for all n, then we simply say that S has the small tail sum property*.*

Note that if S has the small tail sum property at any level n, then

$$\sum_m s_m < \left(\sum_{m\leq n} s_m\right) + s_n < \infty, \tag{1.3}$$

so WLOG we may assume that 1.1 holds by scaling.

(Example) The geometric series $s_n = (1-p)p^n$ has the small tail sum property if and only if $p < 1/2$. Indeed,

$$(1-p)p^n \geq \sum_{m>n}(1-p)p^m = p^{n+1} \iff 1-p \geq p \iff 1/2 \geq p. \tag{1.4}$$

Equality holds for every n exactly when $p = 1/2$; in that case, $\text{sum}(\{1/2, 1/4, 1/8, \ldots\}) = [0,1]$, which is equivalent to the fact that every real number in $[0,1]$ has a binary decomposition.

Let λ denote Lebesgue measure. For countable S and $k \in \mathbb{N}$, kS is countable and thus $\lambda(kS) = 0$, so it is natural to consider $\text{sum}(S)$ in the context of Lebesgue measure. We have the following characterization of $\text{sum}(S)$ in this case.

Theorem 1.2. *Let $S \subset [0,1]$ be a countable probability distribution with the small tail sum property (i.e. satisfying 1.1 and 1.2). We have*

$$\lambda(\text{sum}(S)) = \lim_{n\to\infty} 2^{n+1}\left(1 - \sum_{m\leq n} s_m\right). \tag{1.5}$$

What makes this case special, and allows this direct computation, is that there are no 'overlaps' between sums over different subsets.

Lemma 1.3. *Let $(s_n) \subset [0,1]$ be any countable set with the small tail sum property. If $x \in$ sum(S), then there is a unique subset $I \subset \mathbb{N}$ such that*

$$x = \sum_{i\in I} s_i. \tag{1.6}$$

Moreover, I is obtained by applying the greedy algorithm to x: for $i \in \mathbb{N}$,

$$i \in I \iff \left(\sum_{j \in I, j < i} s_j \right) + s_i < x. \tag{1.7}$$

Proof. Suppose $I, J \subset \mathbb{N}$ are two distinct subsets, and by reversing I and J if necessary, let

$$k = \min\{l \in \mathbb{N} : l \in I, l \notin J, \text{ and } J \cap [l] \subset I \cap [l]\}. \tag{1.8}$$

By the tail bound 1.2 and the definition of k,

$$\sum_{i \in I} s_i - \sum_{j \in J} s_j \geq s_k - \sum_{j > k} s_j > 0. \tag{1.9}$$

So $\sum_{i \in I} s_i \neq \sum_{j \in J} s_j$.

If $x \in \mathrm{sum}(S)$, then the greedy algorithm succeeds. Indeed, choose I such that 1.6 holds, and suppose I does not agree with the greedy algorithm, i.e. for some n,

$$\left(\sum_{i \in I \cap [n-1]} s_i \right) + s_n < x \text{ but } n \notin I. \tag{1.10}$$

Then by 1.2,

$$\begin{aligned} \sum_{i \in I} s_i &< \left(\sum_{i \in I \cap [n-1]} s_i \right) + \sum_{j > i} s_j && (1.11) \\ &< \left(\sum_{i \in I \cap [n-1]} s_i \right) + s_n && (1.12) \\ &< x, && (1.13) \end{aligned}$$

contradicting 1.6.

□

An immediate corollary is:

Corollary 1.4. *For S satisfying 1.2, the greedy algorithm is a bijection between the power set $2^{\mathbb{N}}$ and $\mathrm{sum}(S)$.*

The next lemma describes the complement of $\mathrm{sum}(S)$ as a countable union of intervals. For $I \subset \mathbb{N}$, use s_I to denote the sum over I:

$$s_I = \sum_{i \in I} s_i. \tag{1.14}$$

Lemma 1.5. *Let $S \subset [0,1]$ be a countable probability distribution satisfying 1.2. The set of reals not in the sumset of S can be written as a union of (open) intervals:*

$$[0,1] \setminus \mathrm{sum}(S) = \bigcup_{n \geq 0} \bigcup_{I \subset [n-1]} \left(s_I + \sum_{i > n} s_i, s_I + s_n \right) := \bigcup_{n} \bigcup_{I \subset [n-1]} A_I(n), \tag{1.15}$$

where $[n] = \{0, 1, \dots, n\}$*, and by convention* $[-1] = \emptyset$*. Moreover, the intervals appearing in the union are all pairwise disjoint.*

Proof. Suppose $y \notin \text{sum}(S)$. By 1.3, the greedy algorithm must fail at some finite stage, i.e. for some n and $I \subset [n-1]$,

$$s_I + \sum_{i>n} s_i < y < s_I + s_n. \tag{1.16}$$

So it suffices to show that the $A_I(n)$ are disjoint. Let $I, J \subset \mathbb{N}$ be distinct finite subsets, and consider the intervals $A_I(n)$ and $A_J(m)$. By reversing I and J if necessary, let

$$k = \min\{l \in \mathbb{N} : l \in I, l \notin J, \text{ and } J \cap [l] \subset I \cap [l]\}, \tag{1.17}$$

as in the proof of 1.3. Then

$$s_I \geq s_{I\cap[n]} \geq s_{J\cap[n]} + \sum_{j>n} s_j \geq s_J + s_{m+1}, \tag{1.18}$$

which implies $\inf A_I(n) > \sup A_J(m)$.

□

Lemma 1.5 immediately leads to a computation for the measure of $\text{sum}(S)$.

Proof of 1.2. By lemma 1.5 and 1.1,

$$1 - \lambda(\text{sum}(S)) = \sum_{n\geq 0} \sum_{I\subset[n-1]} \left(s_I + s_n - s_I - \sum_{i>n} s_i \right) \tag{1.19}$$

$$= \sum_{n\geq 0} 2^n \left(s_n - \sum_{i>n} s_i \right) \tag{1.20}$$

$$= \sum_{n\geq 0} 2^n \left(s_n - \left(1 - \sum_{i=0}^{n} s_i \right) \right) \tag{1.21}$$

$$= \lim_{N\to\infty} \left(\sum_{n=0}^{N} 2^n \left(s_0 + s_1 + \cdots + 2s_n - 1 \right) \right) \tag{1.22}$$

$$= \lim_{N\to\infty} \left(-2^{N+1} + 1 + \sum_{k=0}^{N} s_k (2 \cdot 2^k + 2^{k+1} + \cdots + 2^N) \right) \tag{1.23}$$

$$= \lim_{N\to\infty} 1 - 2^{N+1} \left(1 - \sum_{k=0}^{N} s_k \right) \tag{1.24}$$

Thus

$$\lambda(\text{sum}(S)) = \lim_{N\to\infty} 2^{N+1} \left(1 - \sum_{n=0}^{N} s_n \right). \tag{1.25}$$

□

For example, when $s_n = (1-p)p^n$ for $p < 1/2$, we get

$$\lambda(\mathrm{sum}(S)) = \lim_{N\to\infty} 2^{N+1}\left(1-(1-p^{N+1})\right) = \lim_{N\to\infty}(2p)^{N+1} = 0. \tag{1.26}$$

This is somewhat surprising: the measure of the sumset is 1 for $p = 1/2$, but it jumps down to 0 for $p < 1/2$ – there is a sharp phase transition. To obtain a 'smoother' transition, we can look at the Hausdorff dimension of sum(S) instead. By the usual ideas to compute the Hausdorff dimension of a Cantor set, and ignoring issues of convergence, we get the general formula

$$\mathcal{H}(\mathrm{sum}(S)) = \frac{\log 2}{-\lim_{n\to\infty}\frac{1}{n}\log\sum_{m>n}s_m}. \tag{1.27}$$

It seems that this limit should always exist, perhaps as a consequence of:

Lemma 1.6. *Any countable discrete probability distribution S with the small tail sum property 1.2 satisfies $s_n \le 2^{-n}$ for n sufficiently large.*

Proof. ??? □

(Example) For the geometric distributions $s_n = (1-p)p^n$, the Hausdorff dimension is exactly

$$\mathcal{H}(\mathrm{sum}(S)) = \frac{\log 2}{-\log p}. \tag{1.28}$$

This gives a smooth phase transition for $p \in (0, \frac{1}{2})$.

2 Multiple independent samples

In this section we address the following:

Question 2.1. *Given a discrete probability distribution F, how many independent samples of F are needed to generate any Bernoulli(p) for $p \in (0,1)$?*

The small tail sum property will be relevant here, too. Given k iid samples $X_1, X_2, \ldots, X_k$ sampled from the distribution F, we can look at events of the form $(X_i = x_i, i \in [k])$, and sum over countably many such events to get an event with probability p. Thus

$$\{p : \exists A \subset \mathbb{N}^k, \mathbb{P}((X_1,\ldots,X_k)\in A) = p\} = \mathrm{sum}(S), \tag{2.1}$$

where S is the multiset of all possible atoms $\mathbb{P}(X_i = x_i, i \in [k])$.

Writing $F = \sum_n u_n\delta_n$, the atoms of the product measure are of the form $\prod_{i=1}^k u_{x_i}$ for any $x = (x_i : i \in [k]) \subset \mathbb{N}^k$, and each such atom occurs perm(x) many times in S, where perm(x) is the number of distinct permutations of the sequence x (exactly $k!$ occurrances if the entries of x are distinct, but fewer otherwise). Because of these multiplicities in S, S will not have the small tail sum property at most levels, and correspondingly, sum(S) will be larger:

Proposition 2.2. *Suppose S fails the small tail property for all $n \ge n_0$. Then $\lambda(sum(S)) \ge \sum_{m\ge n_0} s_m > 0$. In particular, if S fails the small tail property for all levels, then $sum(S) = [0,1]$.*

Proof. In short: when the small tail property fails, the greedy algorithm succeeds. Note that if the small tail property fails at level m, then $s_m > 0$.

Claim 2.3. *For any $x \in [0, \sum_{m \geq n_0} s_m]$, applying the greedy algorithm to x yields a subset $I \subset \{n_0, n_0+1, \ldots\}$ with $\sum_{i \in I} s_i = x$.*

If not, then the greedy algorithm fails at some stage, i.e. for some set $I \subset \{n_0, n_0+1, \ldots, \ell\}$ with $\ell \in I$,

$$\sum_{i \in I} s_i + s_{\ell+1} > x > \sum_{i \in I} s_i + \sum_{i > \ell+1} s_i, \tag{2.2}$$

contradicting the fact that the small tail property fails at $\ell + 1$. $\square$

2.1 Geometric

Computations for general distributions F are difficult, so we focus on the class of Geometric(p) random variables for $p \in (0, 1/2)$. This case is particularly nice becasuse the atom sizes are constant on hyperplanes normal to $(1, \ldots, 1) \in \mathbb{N}^k$, i.e. with $u_y = qp^y$, we have

$$\prod_{i=1}^{k} u_{x_i} = q^k p^{\sum_i x_i}$$

And, the total number of such atoms is just the size of the hyperplane, or the number of compositions of $r = \sum_i x_i$ with k parts, namely $\binom{r+k-1}{r}$. Mathematica gives that

Lemma 2.4. $p^r \leq \sum_{w > r} \binom{w+k-1}{w} p^w$ *for all $r \geq 0$ if and only if $p \geq 1 - 2^{-k^{-1}}$.*

Thus the sequence S of atoms obtained this way fails the small tail sum property at every level if p is large enough. We obtain:

Proposition 2.5. *k iid samples of Geometric(p) generate all possible Bernoulli(p') for $p' \in [0, 1]$ if and only if $p \geq 1 - 2^{-k^{-1}}$.*

2.2 Geometric, $k = 2$ independent samples

In the special case $k = 2$ we can do some exact computations for the Lebesgue measure of sum(S). Some algebra shows that the set of atoms S satisfies the small tail sum property once for each level $q^2 p^r$ with $r = 0, 1, \ldots, R-1$, where

$$R = \left\lceil \frac{1 - 4p + 2p^2}{p(1-p)} \right\rceil. \tag{2.3}$$

It follows that for each $r = 0, 1, \ldots, R-1$, sum(S) is missing has exactly $(r+1)!$ intervals of size

$$q^2 \left(p^r - \sum_{w > r} (w+1) q^2 p^w \right) \tag{2.4}$$

(There are $w + 1$ copies of the atom $q^2 p^w$ for each w, and thus $(w + 1)$ many choices for each atom size's contribution to the sum). So the Lebesgue measure of the set of Bernoulli's that can be simulated with $k = 2$ independent Geo(p)'s is

$$1 - \sum_{r < R} q^2 (r+1)! \left(p^r - \sum_{w > r} (w+1) p^w \right). \tag{2.5}$$

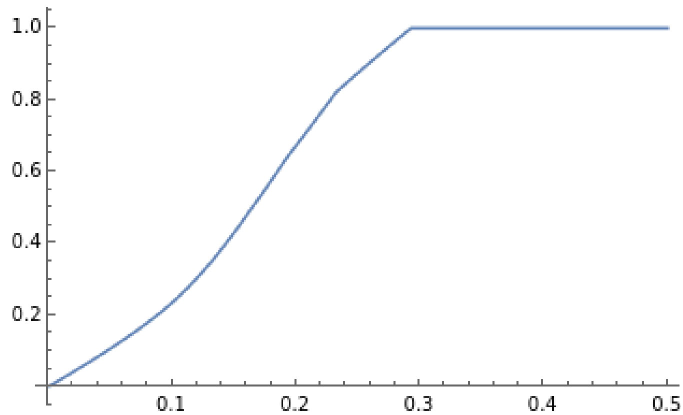

Figure 1: A plot of the Lebesgue measure 2.5 of sum(S) as a function of p, where S is the atoms of the product measure with two independent Geometric(p) random variables. Note that the measure is 1 for $p > 1 - \frac{1}{\sqrt{2}} \approx .29289$ by Proposition 2.5.

3 Further questions

- Is there a countable set $S \subset [0,1]$ such that property 1.2 fails for infinitely many n, and such that $\lambda(\text{sum}(S)) = 0$?
- Is there a discrete distribution F such that k iid random variables distributed according to F are not enough to generate all possible Bernoulli(p)'s? (One possible idea: given p that is not constructible, form a dynamical system by thinking about what happens when $k \to k+1$. Given atoms S at stage k, for p not to be constructible in the next step, it is sufficient that $(s_{\ell+1})^{-1}(p - \sum_I s_i)$ is not in sum(S), where I is the greedy construction in S which fails at level ℓ. If one can show that this map $p \mapsto g_S(p) = (s_{\ell+1})^{-1}(p - \sum_I s_i)$ – which trivially has $g_S(p) \leq p$ with strict inequality if $p \neq 0$ – always hits 0 in a globally bounded number of iterations, we'd be in business.)
- Prove that if S is uncountable, then $\lambda(\text{sum}(S)) > 0$. Can the measure be arbitrarily close to 0 in this case?
- Is there an uncountable set $S \subset [0,1]$ such that $\lambda(2S) = 0$? (The cantor set C satisfies $2S = [0,2]$.)
- Given a set S and $p \in \text{sum}(S)$, where $p = \sum_{i \in I} s_i$, give a notion of 'efficiency' of the representation – e.g. Komolgorov complexity of the set I. Is there some tradeoff between the worst case or average efficiency and the size of sum(S)? Is the number of ways to write p as a sum of elements of S related to such a tradeoff? What about the entropy of the sequence S?
- Suppose we construct S in a random way: for example, fix a distribution function F on $[0,1]$, sample $X_0, X_1, \ldots$ i.i.d. $\sim F$, set $S_0 = X_0$ and recursively define $S_n = X_n S_{n-1}$; or let (X_n) have the Poisson-Dirichlet distribution. What is the probability that the random sequence $(S_n)_n$ satisfies 1.1? 1.2? What is the distribution of $\lambda(\text{sum}(S))$? (Expectation?)