Sum Constructions

Jacob Richey

September 23, 2021

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

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

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

$$F = \sum_{i \in \mathbb{N}} s_i \delta_i. \tag{0.3}$$

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

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

Let's begin by considering a special class of discrete probability sequences: let $S = \{s_i\}_{i \in \mathbb{N}}$ be countable, and assume

1. S is a probability distribution:
$$\sum_{i\in\mathbb{N}} s_n = 1 \tag{0.5}$$

2. Every tail of (s_n) has smaller sum than the previous term:
$$s_i > \sum_{j>i} s_j \text{ for all } m \in \mathbb{N} \tag{0.6}$$

(Example 1) The geometric series $s_i = (1-\alpha)\alpha^i$ satisfies these conditions if and only if $\alpha < 1/2$. Indeed,

$$(1-\alpha)\alpha^i > \sum_{j>i}(1-\alpha)\alpha^j = \alpha^{i+1} \iff 1-\alpha > \alpha \iff 1/2 > \alpha. \tag{0.7}$$

Equality holds exactly when $\alpha = 1/2$ for every i; 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 0.4. *Let $S \subset [0,1]$ be a countable set satisfying 0.5 and 0.6. We have*

$$\lambda(\text{sum}(S)) = \lim_{n\to\infty} 2^{n+1}\left(1 - \sum_{i=0}^{n} s_n\right). \tag{0.8}$$

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

Lemma 0.5. *Let $S \subset [0,1]$ be a countable set satisfying 0.6. If $x \in sum(S)$, then there is a unique subset $I \subset \mathbb{N}$ such that*

$$x = \sum_{i\in I} s_i. \tag{0.9}$$

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

$$i \in I \iff \left(\sum_{j\in I\cap[i-1]} s_j\right) + s_i < x. \tag{0.10}$$

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

$$k = \min\{l \in \mathbb{N} : l \in I, l \notin J, \text{ and } J \cap [l] \subset I \cap [l]\}. \tag{0.11}$$

By the tail bound 0.6 and the definition of k,

$$\sum_{i\in I} s_i - \sum_{j\in J} s_j \geq s_k - \sum_{j>k} s_j > 0. \tag{0.12}$$

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

If $x \in \text{sum}(S)$, then the greedy algorithm succeeds. Indeed, choose I such that 0.9 holds, and suppose I does not agree with the greedy algorithm, i.e. for some n,

$$\left(\sum_{i \in I \cap [n-1]} s_i \right) + s_n < x \text{ but } n \notin I. \tag{0.13}$$

Then by 0.6,

$$\sum_{i \in I} s_i < \left(\sum_{i \in I \cap [n-1]} s_i \right) + \sum_{j > i} s_j \tag{0.14}$$

$$< \left(\sum_{i \in I \cap [n-1]} s_i \right) + s_n \tag{0.15}$$

$$< x, \tag{0.16}$$

contradicting 0.9.

□

An immediate corollary is:

Corollary 0.6. *For S satisfying 0.6, the greedy algorithm is a bijection between the power set $2^{\mathbb{N}}$ and* $\text{sum}(S)$*.*

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

$$s_I = \sum_{i \in I} s_i. \tag{0.17}$$

Lemma 0.7. *Let $S \subset [0, 1]$ be a countable set satisfying 0.6. The set of reals not in the sumset of S can be written as a union of (open) intervals:*

$$[0, 1] \setminus \text{sum}(S) = \bigcup_{n \geq 0} \bigcup_{I \subset [n-1]} \left(s_I + \sum_{i > n} s_i, s_I + s_n \right) := \bigcup_{n} \bigcup_{I \subset [n-1]} A_I(n), \tag{0.18}$$

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

Proof. Suppose $y \notin \text{sum}(S)$. By 0.5, the greedy algorithm must fail at some finite stage, i.e. for some n and $I \subset [n-1]$,

$$s_I + \sum_{i > n} s_i < y < s_I + s_n. \tag{0.19}$$

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

$$k = \min\{l \in \mathbb{N} : l \in I, l \notin J, \text{ and } J \cap [l] \subset I \cap [l]\}, \tag{0.20}$$

as in the proof of 0.5. Then

$$s_I \geq s_{I\cap[n]} \geq s_{J\cap[n]} + \sum_{j>n} s_j \geq s_J + s_{m+1}, \tag{0.21}$$

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

□

Lemma 0.7 immediately leads to a computation for the measure of sum(S).

Proof of 0.4. By lemma 0.7 and 0.5,

$$1 - \lambda(\mathrm{sum}(S)) = \sum_{n\geq 0} \sum_{I\subset[n-1]} \left(s_I + s_n - s_I - \sum_{i>n} s_i \right) \tag{0.22}$$

$$= \sum_{n\geq 0} 2^n \left(s_n - \sum_{i>n} s_i \right) \tag{0.23}$$

$$= \sum_{n\geq 0} 2^n \left(s_n - \left(1 - \sum_{i=0}^{n} s_i \right) \right) \tag{0.24}$$

$$= \lim_{N\to\infty} \left(\sum_{n=0}^{N} 2^n \left(s_0 + s_1 + \cdots + 2s_n - 1 \right) \right) \tag{0.25}$$

$$= \lim_{N\to\infty} \left(-2^{N+1} + 1 + \sum_{k=0}^{N} s_k(2\cdot 2^k + 2^{k+1} + \cdots + 2^N) \right) \tag{0.26}$$

$$= \lim_{N\to\infty} 1 - 2^{N+1} \left(1 - \sum_{k=0}^{N} s_k \right) \tag{0.27}$$

Thus

$$\lambda(\mathrm{sum}(S)) = \lim_{N\to\infty} 2^{N+1} \left(1 - \sum_{n=0}^{N} s_n \right). \tag{0.28}$$

□

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

$$\lambda(\mathrm{sum}(S)) = \lim_{N\to\infty} 2^{N+1} \left(1 - (1 - \alpha^{N+1}) \right) = \lim_{N\to\infty} (2\alpha)^{N+1} = 0. \tag{0.29}$$

This is somewhat surprising: the measure of the sumset is 1 for $\alpha = 1/2$, but it jumps down to 0 for $\alpha < 1/2$ – there is a sharp phase transition.

Before moving on, we present a probabilistic version of the proof of 0.4, which gives some intuition for the quantities appearing in the first proof.

Probabilistic proof of 0.4. Let U be a uniform random variable on $(0,1)$, and note that by 0.5,

$$\lambda(\mathrm{sum}(S)) = \mathbb{P}(U \in \mathrm{sum}(S)). \tag{0.30}$$

To compute the latter probability, we construct $\text{sum}(S)$ by removing intervals from $[0,1]$ in stages, a la the Cantor set construction, and at each stage compute the probability that U was in one of the removed intervals. In the first stage, the interval $s_0 - \sum_{i>0} s_i$ is removed; at the n^{th} stage, we remove the 2^n intervals $A_I(n)$ for $I \subset [n-1]$. By 0.7, these intervals are all disjoint. Note that $\lambda(A_I(n)) = s_n - \sum_{i>n} s_i$ for every $I \subset [n-1]$. It follows that, letting F_n denote the event that U is removed at stage n,

$$\mathbb{P}(F_n) = 2^n \left(s_n - \sum_{i>n} s_i \right). \tag{0.31}$$

Thus

$$\mathbb{P}(U \notin \text{sum}(S)) = \sum_{n \geq 0} \mathbb{P}(F_n). \tag{0.32}$$

The same computation as in the previous proof finishes the proof.

□

Further questions:

- This analysis deals with the case where we have access to a single sample of a discrete distribution function F, and want to simulate a Bernoulli r.v. What if we have access to two independent samples of F? Then, if $F = \sum_i s_i \delta_i$, we can simulate events with probabilities in the set

$$\text{sum}(S \times S) = \left\{ \sum_{(x,y) \in K} xy : K \subset S \times S \right\}. \tag{0.33}$$

  Can we describe this set in a simple way? For which S does it have positive measure? Zero measure?

- Is there a countable set $S \subset [0,1]$ such that property 0.6 fails for infinitely many n, and such that $\lambda(\text{sum}(S)) = 0$?

- 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 $2S = [0,2]$.)

- 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 0.5? 0.6? What is the distribution of $\lambda(\text{sum}(S))$? (Expectation?)