Random Sidon Sequences

Anant P. Godbole

Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931, U.S.A. E-mail: anant@mtu.edu

Svante Janson

Department of Mathematics, Uppsala University, PO Box 480, S-751 06 Uppsala, Sweden. E-mail: svante.janson@math.uu.se

Nicholas W. Locantore, Jr.

Department of Statistics, University of North Carolina, Chapel Hill, NC 27599, U.S.A. E-mail: locantor@stat.unc.edu

Rebecca Rapoport

Department of Mathematics, Harvard University, Cambridge, MA 02138, U.S.A. E-mail: rapoport@fas.harvard.edu

Running Head: Random Sidon Sequences

Proofs: Should be sent to Anant Godbole, Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931-1295, U.S.A.

Abstract

A subset A of the set $[n]=\{1,2,\ldots,n\}$, $|A|=k$, is said to form a *Sidon* (or B_h) sequence, $h\geq 2$, if each of the sums $a_1+a_2+\ldots+a_h, a_1\leq a_2\leq\ldots\leq a_h; a_i\in A$, are distinct. We investigate threshold phenomena for the Sidon property, showing that if A_n is a random subset of $[n]$, then the probability that A_n is a B_h sequence tends to unity as $n\to\infty$ if $k_n=|A_n|\ll n^{1/2h}$, and that $\mathbf{P}(A_n \text{ is Sidon})\to 0$ provided that $k_n\gg n^{1/2h}$. The main tool employed is the Janson exponential inequality. The validity of the Sidon property *at* the threshold is studied as well; we prove, using the Stein–Chen method of Poisson approximation, that $\mathbf{P}(A_n \text{ is Sidon})\to\exp\{-\lambda\}$ $(n\to\infty)$ if $k_n\sim\Lambda\cdot n^{1/2h}$ $(\Lambda\in\mathbf{R}^+)$, where λ is a constant that depends in a well-specified way on Λ. Multivariate generalizations are presented.

1. Introduction

A subset A of $[n] = \{1, 2, \ldots, n\}$, $|A| = k$, is said to form a *Sidon* (or B_h) sequence, $h \geq 2$, if each of the $\binom{k+h-1}{h}$ sums $a_1 + a_2 + \ldots + a_h, a_1 \leq a_2 \leq \ldots \leq a_h, a_i \in A$ $(i = 1, 2, \ldots, h)$ are distinct. For example, any two element set $\{a, b\}$ is B_2, since the three sums $a+b, 2a, 2b$ are necessarily distinct, whilst a three element set $\{a, b, c\}$ is B_2 iff a, b, c are not in arithmetic progression. An extensive survey of the properties of Sidon sequences may be found in Halberstam and Roth [5], where it is shown, for example, that B_h sequences are of size at most $O(n^{1/h})$ [for any $h \geq 2$] , and, moreover, that there do exist B_h sequences of order $n^{1/h}$. In particular, Lindström [6] showed that $|A| \leq n^{1/2} + n^{1/4} + 1$ for any B_2 sequence A. Recent papers on finite and infinite Sidon sequences include the ones by Graham [4] and Spencer and Tetali [8].

We consider a set A_n obtained by selecting, without replacement, a random sample of size k_n from the first n integers, and investigate threshold phenomena for the Sidon property, showing, in Theorem 1, that the probability that A_n is B_h tends to unity as $n \to \infty$ if $k_n \ll n^{1/2h}$, and that $\mathbf{P}(A_n \text{ is Sidon}) \to 0$ provided that $k_n \gg n^{1/2h}$, where we write $\varphi(n) \gg \varsigma(n)$ (resp. $\varphi(n) \ll \varsigma(n)$) if $\varphi(n)/\varsigma(n) \to \infty$ (resp. 0) as $n \to \infty$. (The first part has also been shown by Nathanson, see [7], page 37, Exercise 14.) The main tool employed is the Janson exponential inequality (see, e.g., Alon and Spencer [1]). Theorem 1 shows that the Sidon property becomes rare at a level far below that indicated by the above-mentioned extremal results in Halberstam and Roth [5]; it is conceivable, however, that a carefully selected non-uniform measure on the k_n-subsets of $[n]$ will yield a threshold closer to $n^{1/h}$: for example, one may be able to exploit the fact [3,4] that maximal B_2

sequences are uniformly distributed. In Section 3, we investigate the behaviour of the Sidon property *at* the threshold, proving in Theorem 2 that $\mathbf{P}(A_n \text{ is } B_h) \to \exp\{-\lambda\}$ as $n \to \infty$ if $|A_n| \sim \Lambda \cdot n^{1/2h}$, where $\Lambda \in \mathbf{R}^+$ and $\lambda = \kappa_h \Lambda^{2h}$ for a constant depending on h. ($\kappa_2 = 1/12$ and $\kappa_3 = 11/1440$; asymptotically $\kappa_h \sim \sqrt{\frac{3}{4\pi}} h^{-1/2} h!^{-2}$ as $h \to \infty$.) The Stein–Chen method of Poisson approximation [2] is the main technique used in the proof of this result. We also provide multivariate Poisson approximations for the *joint* distribution of the ensemble $\{I_{\mathbf{a},\mathbf{b}} : a_1 + \ldots + a_h = b_1 + \ldots + b_h\}$, where $\mathbf{a} = (a_1, \ldots, a_h)$, $\mathbf{b} = (b_1, \ldots, b_h)$, and where the zero-one variable $I_{\mathbf{a},\mathbf{b}}$ equals one iff $\{a_1, \ldots, a_h\} \subseteq A_n$, $\{b_1, \ldots, b_h\} \subseteq A_n$; this result (Theorem 3) enables one to understand the structure of the set A_n in a global sense, keeping track, as it does, of *all* the episodes when an integer m is obtained by two h-sums of elements of A_n. The Stein–Chen method is used once again as the driving force behind the proof; of special note is the fact that the components of the multivariate Poisson approximant in Theorem 3 are *independent*, whereas the variables $I_{\mathbf{a},\mathbf{b}}$ are clearly not.

We have chosen to employ different methods in Sections 2 and 3, but it should be made clear at the outset that we could have done differently. In fact, Theorem 1 is a simple corollary of Theorem 2, and thus follows by the Stein–Chen method too. (A third possibility is to use Chebyshev's inequality together with estimates derived below.) Conversely, Theorem 2 may be derived using the Janson inequality.

Similar questions can be asked regarding sum-free subsets of the integers, and will be reported on elsewhere, as will be results on B_h sequences where $h \to \infty$ along with n, and on subsets with distinct sums (see [1] for the relevant definitions).

We write $u = O(v)$ or (equivalently) $u \preceq v$ if $u \le Av$ for some constant A that may depend on h but not on n or any other variable.

2. Threshold functions for the Sidon property

The following is the main result of this section:

Theorem 1. *Consider a subset A_n of size k_n chosen at random from the $\binom{n}{k_n}$ such subsets of $[n] = \{1, 2, \ldots, n\}$. Then for any $h \ge 2$,*

$$k_n = o(n^{1/2h}) \Rightarrow \mathbf{P}(A_n \text{ is } B_h) \to 1 \quad (n \to \infty)$$

and

$$n^{1/2h} = o(k_n) \Rightarrow \mathbf{P}(A_n \text{ is } B_h) \to 0 \quad (n \to \infty).$$

Proof. We begin with the easy first half, the proof of which employs nothing more than the Markov inequality. We introduce some notation to be used throughout the paper.

Let $\mathcal{A} = \mathcal{A}_{n,h}$ be the set of all sequences $\mathbf{a} = (a_1, \ldots, a_h)$ with $1 \le a_1 \le a_2 \le \ldots \le a_h \le n$, and let

$$\mathcal{B} = \mathcal{B}_{n,h} = \{(\mathbf{a}, \mathbf{b}) \in \mathcal{A} \times \mathcal{A} : a_1 + \ldots + a_h = b_1 + \ldots + b_h \text{ and } \mathbf{a} < \mathbf{b}\}$$

where $<$ denotes the lexicographic order.

An element $\mathbf{a}$ of $\mathcal{A}$ is thus an (ordered) sequence $(a_1, \ldots, a_h)$, but we will also, when convenient, use $\mathbf{a}$ to denote the corresponding set $\{a_1, \ldots, a_h\}$; for example, $|\mathbf{a}|$ denotes the number of elements of this set, i.e., the number of distinct numbers a_i.

Using this notation, a set $A_n \subset [n]$ is Sidon if and only if A_n does not contain $\mathbf{a} \cup \mathbf{b}$ for any $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}$.

Let, as above, $I_{\mathbf{a},\mathbf{b}}$, $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}$, be the (random) indicator variables defined by $I_{\mathbf{a},\mathbf{b}} = 1$ if $\mathbf{a} \cup \mathbf{b} \subseteq A_n$ (with $I_{\mathbf{a},\mathbf{b}} = 0$ otherwise), and define

$$X = \sum_{(\mathbf{a},\mathbf{b}) \in \mathcal{B}} I_{\mathbf{a},\mathbf{b}}.$$

Thus A_n is Sidon if and only if $I_{\mathbf{a},\mathbf{b}} = 0$ for every pair $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}$, i.e., when $X = 0$.

We define

$$\mathcal{B}(l) = \{(\mathbf{a}, \mathbf{b}) \in \mathcal{B} : |\mathbf{a} \cup \mathbf{b}| = l\}, \qquad l = 1, \dots, 2h,$$

and note that $\mathcal{B}(2h)$ is the set of pairs $(\mathbf{a}, \mathbf{b})$ with $2h$ distinct numbers $a_1, \dots, b_h$. Clearly, for any $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}(l)$,

$$\mathbf{P}(I_{\mathbf{a},\mathbf{b}} = 1) = \binom{n-l}{k-l} \Big/ \binom{n}{k} \le \left(\frac{k}{n}\right)^l,$$

and thus, by Markov's inequality,

$$\begin{aligned} &\mathbf{P}(A_n \text{ is not } B_h) = \mathbf{P}(X \ge 1) \\ &\qquad \le \mathbf{E}(X) = \sum_{l=1}^{2h} |\mathcal{B}(l)| \binom{n-l}{k-l} \Big/ \binom{n}{k} \le \sum_{l=1}^{2h} |\mathcal{B}(l)| \left(\frac{k}{n}\right)^l. \end{aligned} \tag{1}$$

We estimate $|\mathcal{B}(l)|$ as a lemma.

Lemma 1. *$|\mathcal{B}(l)|$, the number of pairs $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}$ containing exactly l different numbers, is $O(n^{l-1})$ for every $l \le 2h$.*

Proof. A pair $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}(l)$ satisfies a pattern of $2h - l$ (non-redundant) coincidences among $\{a_1, \dots, b_h\}$, for example $a_1 = a_2 = b_1$, $a_5 = b_3$, $\dots$. Fix one such pattern. This pattern defines $2h-l$ of the variables $a_1, \ldots, b_h$ in terms of the remaining l 'free' ones. Moreover, the relation $a_1 + \ldots + a_h = b_1 + \ldots + b_h$ yields a linear relation between the free variables, and this relation degenerates only when each free variable occurs equally many times in $\mathbf{a}$ and in $\mathbf{b}$, which means that the pattern implies $\mathbf{a} = \mathbf{b}$ and hence $(\mathbf{a}, \mathbf{b}) \notin \mathcal{B}$. For all other patterns, the pair $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}$ is thus specified by $l-1$ variables $\in [n]$, and the number of pairs $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}$ with a given pattern is thus $\leq n^{l-1}$. This completes the proof, since the number of possible patterns is finite (and bounded independently of n).

Consequently, if $k = o(n^{1/2h})$, then

$$\mathbf{P}(A_n \text{ is not } B_h) \preceq \sum_{l=1}^{2h} n^{l-1} k^l n^{-l} \preceq k^{2h} n^{-1} \to 0,$$

as $n \to \infty$, which proves the first part of the theorem.

Turning to the second half, we note that the main contribution to $\mathbf{E}(X)$ is through h-tuples $\mathbf{a}$ and $\mathbf{b}$ whose $2h$ coordinates are all distinct. Thus we define

$$Y = \sum_{(\mathbf{a}, \mathbf{b}) \in \mathcal{B}(2h)} I_{\mathbf{a}, \mathbf{b}}$$

and see that

$$\mathbf{P}(A \text{ is } B_h) = \mathbf{P}(X = 0) \leq \mathbf{P}(Y = 0).$$

We thus focus on computing $\mathbf{P}(Y = 0)$, and start by changing the underlying model somewhat; we will revert to the original model later in the proof: Let us choose each element of $[n]$ independently with probability $p = k/n$. This yields a set whose *expected* (as opposed to actual) cardinality is k. Such a strategy is necessary due the baseline assumption of independence that is required for the successful application of the Janson inequality, which yields (see e.g. Alon and Spencer [1], Theorem 1.1 in Chapter 8 with $\varepsilon = 1/2$; the version given there has the (not really necessary) assumption $\mathbf{P}_u(I_{\mathbf{a},\mathbf{b}} = 1) = p^{2h} \leq \frac{1}{2}$ for all $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}(2h)$, which we may assume without loss)

$$\mathbf{P}_u(Y = 0) \leq \left(\prod_{(\mathbf{a},\mathbf{b}) \in \mathcal{B}(2h)} \mathbf{P}_u(I_{\mathbf{a},\mathbf{b}} = 0) \right) \exp(\Delta), \tag{2}$$

where $\mathbf{P}_u$ is the probability measure corresponding to the modified model described above and Δ is given by

$$\Delta = \sum_{(\mathbf{a},\mathbf{b}) \sim (\mathbf{c},\mathbf{d})} \mathbf{P}_u(I_{\mathbf{a},\mathbf{b}} I_{\mathbf{c},\mathbf{d}} = 1)$$

with the relation $\sim$ on $\mathcal{B}(2h)$ being defined as follows: We say that $(\mathbf{a}, \mathbf{b}) \sim (\mathbf{c}, \mathbf{d})$ if $(\mathbf{a}, \mathbf{b}), (\mathbf{c}, \mathbf{d}) \in \mathcal{B}(2h)$, $(\mathbf{a}, \mathbf{b}) \neq (\mathbf{c}, \mathbf{d})$ and $(\mathbf{a} \cup \mathbf{b}) \cap (\mathbf{c} \cup \mathbf{d}) \neq \emptyset$. By (2), our result will follow, under the modified model, if we can show that the right hand side of (2) tends to zero for suitable p. Let, for $2h \leq l \leq 4h$,

$$\mathcal{D}(l) = \left\{ \big((\mathbf{a}, \mathbf{b}), (\mathbf{c}, \mathbf{d}) \big) \in \mathcal{B}(2h) \times \mathcal{B}(2h) : (\mathbf{a}, \mathbf{b}) \neq (\mathbf{c}, \mathbf{d}) \text{ and } |\mathbf{a} \cup \mathbf{b} \cup \mathbf{c} \cup \mathbf{d}| = l \right\}.$$

Then $\mathcal{D} := \bigcup_{l=2h}^{4h-1} \mathcal{D}(l)$ is the set of pairs of pairs $\big((\mathbf{a}, \mathbf{b}), (\mathbf{c}, \mathbf{d}) \big)$ with $(\mathbf{a}, \mathbf{b}) \sim (\mathbf{c}, \mathbf{d})$. We have,

$$\begin{aligned} \Delta &= \sum_{(\mathbf{a},\mathbf{b}) \sim (\mathbf{c},\mathbf{d})} \mathbf{P}_u(I_{\mathbf{a},\mathbf{b}} I_{\mathbf{c},\mathbf{d}} = 1) = \sum_{l=2h}^{4h-1} \sum_{((\mathbf{a},\mathbf{b}),(\mathbf{c},\mathbf{d})) \in \mathcal{D}(l)} \mathbf{P}_u(I_{\mathbf{a},\mathbf{b}} I_{\mathbf{c},\mathbf{d}} = 1) \\ &= \sum_{l=2h}^{4h-1} |\mathcal{D}(l)| p^l. \end{aligned} \tag{3}$$

Lemma 2. *For each $l \geq 2h$, $|\mathcal{D}(l)| \preceq n^{l-2}$.*

Proof. We argue as in the proof of Lemma 1. This time each $\big((\mathbf{a}, \mathbf{b}), (\mathbf{c}, \mathbf{d}) \big) \in \mathcal{D}(l)$ satisfies a pattern of $4h - l$ coincidences of the types $a_i = c_j$, $a_i = d_j$, $b_i = c_j$ and $b_i = d_j$, where no variable occurs more than once. (Recall that by assumption, $(\mathbf{a}, \mathbf{b})$ and $(\mathbf{c}, \mathbf{d})$ each contain $2h$ distinct numbers.)

We fix one such pattern. Suppose first that $l > 2h$. Then there are n^{2h-1} choices of $a_1, \ldots, b_{h-1}$, which together determine b_h (possible outside $[n]$ and thus illegal) because $a_1 + \ldots + a_h = b_1 + \ldots + b_h$. The pattern of coincidences then determine $4h - l$ of $c_1, \ldots, d_h$, and of the remaining $2h - (4h - l) = l - 2h > 0$ variables one is determined by the others because of the relation $c_1 + \ldots + c_h = d_1 + \ldots + d_h$; hence there are $\leq n^{l-2h-1}$ choices of $c_1, \ldots, d_h$. Together this gives $\leq n^{2h-1+l-2h-1} = n^{l-2}$ choices for each pattern, and the result for the case $l > 2h$ follows.

In the case $l = 2h$, the pattern determines each c_j and d_j as one of $a_1, \ldots, b_h$. If each c_j coincides with an a_i, then necessarily $\mathbf{c} = \mathbf{a}$ (recall that the sequences are ordered) and $\mathbf{d} = \mathbf{b}$, which violates $(\mathbf{a}, \mathbf{b}) \neq (\mathbf{c}, \mathbf{d})$, and there are no pairs of pairs in $\mathcal{D}(2h)$ satisfying the pattern. Similarly, if each c_j coincides with an b_i, then $\mathbf{c} = \mathbf{b}$ and $\mathbf{d} = \mathbf{a}$, which violates $\mathbf{a} < \mathbf{b}$ and $\mathbf{c} < \mathbf{d}$. Hence we only have to consider patterns where all four types of coincidences $a_i = c_j$, $a_i = d_j$, $b_i = c_j$ and $b_i = d_j$ occur (with different indices, in general), but in this case the relations $a_1 + \ldots + a_h = b_1 + \ldots + b_h$ and $c_1 + \ldots + c_h = d_1 + \ldots + d_h$ give two linearly independent relations between $a_1, \ldots, b_h$, and thus these numbers are determined by $2h - 2 = l - 2$ of them. Consequently, the number of pairs of pairs for each pattern is $\leq n^{l-2}$ in this case too, and the result follows.

We thus have, using (3) and $np = k \geq 1$,

$$\Delta = \sum_{l=2h}^{4h-1} |\mathcal{D}(l)| p^l \preceq \sum_{l=2h}^{4h-1} n^{l-2} p^l \preceq n^{4h-3} p^{4h-1}. \tag{4}$$

Note further that $|\mathcal{B}(2h)| \succeq n^{2h-1}$ (we will prove a more precise estimate in the next section). Returning to (2), we thus obtain, for some positive constants c and C,

$$\begin{aligned}
\mathbf{P}_u(Y=0) &\le \left(\prod_{(\mathbf{a},\mathbf{b})\in\mathcal{B}(2h)} \mathbf{P}_u(I_{\mathbf{a},\mathbf{b}}=0)\right)\exp\{Cn^{4h-3}p^{4h-1}\}\\
&\le \left(1-p^{2h}\right)^{cn^{2h-1}}\exp\{Cn^{4h-3}p^{4h-1}\}\\
&\le \exp\{-cn^{2h-1}p^{2h}+Cn^{4h-3}p^{4h-1}\}\\
&= \exp\{-n^{2h-1}p^{2h}\left(c-Cn^{2h-2}p^{2h-1}\right)\}. \qquad (5)
\end{aligned}$$

Now if

$$\frac{1}{n^{\frac{2h-1}{2h}}} \ll p \ll \frac{1}{n^{\frac{2h-2}{2h-1}}},$$

(5) reveals that $\mathbf{P}_u(Y=0)\to 0$, showing, by monotonocity, that Theorem 1 holds for the altered model if $p \gg 1/n^{(2h-1)/2h}$, i.e., if $\mathbf{E}(|A_n|) \gg n^{1/2h}$. We must now translate this fact into the format of the original problem, and thus need to compute, under the transformed model, $\mathbf{P}_u(A_n \text{ is } B_h \big| |A_n| = np)$, which, again by monotonicity, is smaller than $\mathbf{P}_u(A_n \text{ is } B_h \big| |A_n| \le np)$ and thus than $\mathbf{P}_u(A_n \text{ is } B_h)/\mathbf{P}_u(|A_n| \le np)$. Now the numerator of this last quantity is asymptotically small if $p \gg 1/n^{(2h-1)/2h}$, whilst the denominator is certainly, at least for large n, of magnitude close to $1/2$. The theorem follows.

3. The behavior of the Sidon property at the threshold

As mentioned above, the first result of this section, which finds the asymptotic value of $\mathbf{P}(A_n \text{ is } B_h)$ when $|A_n| \sim \Lambda n^{1/2h}$ *could* have been obtained on using the methods of Section 2. We choose, however, to employ the Stein–Chen method of Poisson approximation [2] (which could, conversely, have been used to establish Theorem 1) to address a wider issue: If X denotes, as before, the number of episodes $(\mathbf{a}, \mathbf{b})$ (under the model P_u) for which A_n contains both the vectors $\mathbf{a}$ and $\mathbf{b}$ whose coordinates sum to the same value, then what can be said about the distribution of X (and not just the value of the point probability $\mathbf{P}_u(X = 0)$?) Let $\mathcal{L}(U)$ denote the probability distribution of the random variable U, and $\mathrm{Po}(\lambda)$ the Poisson distribution with parameter λ. Finally, let $d_{\mathrm{TV}}(\mathcal{L}(U), \mathcal{L}(V))$ be the total variation distance between $\mathcal{L}(U)$ and $\mathcal{L}(V)$, defined by

$$d_{\mathrm{TV}}(\mathcal{L}(U), \mathcal{L}(V)) = \sup_{A \subseteq \mathbf{Z}^+} |\mathbf{P}(U \in A) - \mathbf{P}(V \in A)|.$$

Now for any three random variables U, V and W,

$$d_{\mathrm{TV}}(\mathcal{L}(U), \mathcal{L}(V)) \le d_{\mathrm{TV}}(\mathcal{L}(U), \mathcal{L}(W)) + \mathbf{P}(V \ne W),$$

so that in our context,

$$d_{\mathrm{TV}}(\mathcal{L}(X), \mathrm{Po}(\mathbf{E}_u(Y))) \le d_{\mathrm{TV}}(\mathcal{L}(Y), \mathrm{Po}(\mathbf{E}_u(Y))) + \mathbf{P}_u(X \ne Y),$$

where X and Y are as defined in Section 2. Since, as in the argument leading to (1), and using Lemma 1,

$$\mathbf{P}_u(X \ne Y) \le \mathbf{E}_u(X - Y) = \sum_{l=1}^{2h-1} |\mathcal{B}(l)| p^l \preceq n^{2h-2} p^{2h-1} \to 0 \tag{6}$$

if $p = o(1/n^{(2h-2)/(2h-1)})$, we focus on bounding $d_{\mathrm{TV}}(\mathcal{L}(Y), \mathrm{Po}(\mathbf{E}_u(Y)))$.

Our first task will be to obtain a tight estimate on $\lambda = \mathbf{E}_u(Y)$. Now

$$\lambda = \sum_{(\mathbf{a},\mathbf{b}) \in \mathcal{B}(2h)} \mathbf{P}(I_{\mathbf{a},\mathbf{b}} = 1) = p^{2h} |\mathcal{B}(2h)|. \tag{7}$$

Loosely, we know that $|\mathcal{B}(2h)| \asymp n^{2h-1}$ so that $\lambda \asymp p^{2h}n^{2h-1} = \Lambda^{2h}$ if $p = \Lambda n^{-(2h-1)/2h}$, but we must be more exact.

We define the functions $f_j = \chi_{(0,1]}^{*j}$, $j = 1, 2, \ldots$, to be the convolution powers of the characteristic function of $(0, 1]$, i.e., $f_1(x) = 1$ when $0 < x \leq 1$ and 0 otherwise, and

$$f_{j+1}(x) = \int_{x-1}^{x} f_j(t)\,dt, \qquad j \geq 1.$$

(Note that $f_j(x)$ equals the density function for the distribution of the sum of j independent random variables, each uniformly distributed on $(0, 1]$.)

Lemma 3. *Let $h \geq 1$ and let $N_{m,n}$ be the number of h-subsets of $\{1, \ldots, n\}$ with sum m. Then*

$$N_{m,n} = \frac{1}{h!} f_h(m/n) n^{h-1} + O(n^{h-2}).$$

(Recall our convention that the constant implicit in the O term does not depend on m or n.)

Proof. Let $N^*_{m,n,h}$ be the number of sequences $\mathbf{a} = (a_1, \ldots, a_h)$ with $1 \leq a_i \leq n$ for all i and $a_1 + \ldots + a_h = m$. Since the number of such sequences with distinct elements equals $h!N_{m,n}$, and the number of such sequences with two or more elements coinciding is $O(n^{h-2})$, it suffices to show that

$$N^*_{m,n,h} = f_h(m/n) n^{h-1} + O(n^{h-2}). \tag{8}$$

This is trivially true for $h = 1$. Moreover, collecting sequences according to their last element a_h, it is seen that

$$N^*_{m,n,h} = \sum_{j=1}^{n} N^*_{m-j,n,h-1}$$

and (8) follows easily by induction, and approximating the appropriate integral by its Riemann sum.

Lemma 4. *For every $h \geq 2$,*

$$|\mathcal{B}(2h)| = \kappa_h n^{2h-1} + O(n^{2h-2}),$$

where

$$\kappa_h = \frac{1}{2(h!)^2} \int_0^h f_h^2(x)\, dx > 0.$$

Proof. $2|\mathcal{B}(2h)|$ equals the number of pairs $(\mathbf{a}, \mathbf{b}) \in \mathcal{A} \times \mathcal{A}$ with $a_1 + \ldots + a_h = b_1 + \ldots + b_h$ and $|\mathbf{a} \cup \mathbf{b}| = 2h$. Each such pair thus consists of two h-subsets $\mathbf{a}$ and $\mathbf{b}$ with the same sum m for some $m \leq hn$; conversely, all pairs of two disjoint h-subsets with the same sum arise in this way. Hence

$$2|\mathcal{B}(2h)| \leq \sum_{m=1}^{hn} N_{m,n}^2 \leq 2|\mathcal{B}(2h)| + N', \tag{9}$$

where N' is the number of pairs $(\mathbf{a}, \mathbf{b})$ with $\mathbf{a}, \mathbf{b} \in \mathcal{A}$ and $\mathbf{a} \cap \mathbf{b} \neq \emptyset$, and thus $|\mathbf{a} \cup \mathbf{b}| < 2h$. Considering the three cases $\mathbf{a} < \mathbf{b}$, $\mathbf{a} = \mathbf{b}$ and $\mathbf{a} > \mathbf{b}$, we obtain, using Lemma 1,

$$N' \leq 2 \sum_{l=1}^{2h-1} |\mathcal{B}(l)| + |\mathcal{A}| \preceq n^{2h-2} + n^h \preceq n^{2h-2}. \tag{10}$$

Next we use Lemma 3 and conclude that

$$\begin{aligned} \sum_{m=1}^{hn} N_{m,n}^2 &= \frac{1}{h!^2} \sum_{m=1}^{hn} \left(f_h^2(m/n) n^{2h-2} + O(n^{2h-3}) \right) \\ &= \frac{n^{2h-2}}{h!^2} \sum_{m=1}^{hn} f_h^2(m/n) + O(n^{2h-2}). \end{aligned} \tag{11}$$

Finally we have, using the fact that $f_h'(x) = f_{h-1}(x) - f_{h-1}(x-1)$ is bounded for every $h \geq 2$,

$$\sum_{m=1}^{hn} f_h^2(m/n) = \sum_{m=1}^{hn} n \int_{(m-1)/n}^{m/n} \left(f_h^2(x) + O(n^{-1}) \right) \, dx = n \int_0^h f_h^2(x) \, dx + O(1). \tag{12}$$

The lemma follows by combining (9), (10), (11) and (12).

The function f_h vanishes outside $[0, h]$, and on each interval $[i-1, i]$, $i = 1, \ldots, h$, it equals a polynomial; hence $\int_0^h f_h^2$ can in principle be computed directly for each h. This is easily done for small h, but quickly becomes rather tedious and does not seem to yield a general formula. We thus calculate the integral using Fourier methods.

Lemma 5. *If $h \geq 1$, then*

$$\int_0^h f_h^2(x) \, dx = \frac{1}{(2h-1)!} \sum_{j=0}^{h-1} (-1)^j \binom{2h}{j} (h-j)^{2h-1}.$$

Proof. The Fourier transform of $\chi_{(0,1]}$ is

$$\hat{\chi}_{(0,1]}(t) = \int_0^1 e^{itx} \, dx = \frac{1}{it}(e^{it} - 1).$$

Since $\hat{f}_h = (\hat{\chi}_{(0,1]})^h$, Plancherel's formula yields

$$\int_0^h f_h^2(x) \, dx = \int_{-\infty}^{\infty} f_h^2(x) \, dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\hat{f}_h^2(t)| \, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{|e^{it} - 1|^{2h}}{t^{2h}} \, dt. \tag{13}$$

Denote the numerator $|e^{it} - 1|^{2h} = (e^{it} - 1)^h (e^{-it} - 1)^h$ by $P(t)$. We integrate by parts $2h - 2$ times, obtaining

$$\begin{aligned} \int_0^h f_h^2(x) \, dx &= \frac{1}{\pi} \int_0^{\infty} \frac{P(t)}{t^{2h}} \, dt = \frac{1}{\pi(2h-1)} \int_0^{\infty} \frac{P'(t)}{t^{2h-1}} \, dt = \ldots \\ &= \frac{1}{\pi(2h-1)!} \int_0^{\infty} \frac{P^{(2h-2)}(t)}{t^2} \, dt. \end{aligned} \tag{14}$$

(The integrals converge and the integrated parts vanish because P has a zero of order $2h$ at $t = 0$ and P and all its derivatives are bounded.)

A binomial expansion yields

$$P(t) = (e^{it} - 1)^h (e^{-it} - 1)^h = (-1)^h e^{-ith} (e^{it} - 1)^{2h} = \sum_{j=0}^{2h} \binom{2h}{j} (-1)^{h+j} e^{it(h-j)}$$

and thus (except for an extra constant term in the case $h = 1$)

$$\begin{aligned} P^{(2h-2)}(t) &= \sum_{j=0}^{2h} \binom{2h}{j} (-1)^{j+1} (h-j)^{2h-2} e^{it(h-j)} \\ &= \sum_{j=0}^{h-1} \binom{2h}{j} (-1)^{j+1} (h-j)^{2h-2} 2\cos(h-j)t. \end{aligned}$$

Hence, using also $P^{(2h-2)}(0) = 0$, (14) yields

$$\begin{aligned} \int_0^h f_h^2(x)\, dx &= \frac{1}{(2h-1)!\,\pi} \int_0^\infty \frac{P^{(2h-2)}(t) - P^{(2h-2)}(0)}{t^2}\, dt \\ &= \frac{1}{(2h-1)!\,\pi} \sum_{j=0}^{h-1} \binom{2h}{j} (-1)^{j+1} (h-j)^{2h-2} \int_0^\infty \frac{2\cos(h-j)t - 2}{t^2}\, dt. \end{aligned}$$

Finally, for any $k > 0$,

$$\int_0^\infty \frac{1 - \cos kt}{t^2}\, dt = k \int_0^\infty \frac{1 - \cos u}{u^2}\, du = k \frac{\pi}{2},$$

and the result follows. (The integral $\int_0^\infty \frac{1-\cos u}{u^2}\, du = \frac{\pi}{2}$ is well-known; alternatively, this follows by checking the case $h = 1$ of the lemma.)

We summarize the result.

Lemma 6.

$$\mathbf{E}_u X = \kappa_h n^{2h-1} p^{2h} + O(n^{2h-2} p^{2h-1}) \tag{15}$$

and

$$\mathbf{E}_u Y = \kappa_h n^{2h-1} p^{2h} + O(n^{2h-2} p^{2h}) \tag{16}$$

with

$$\kappa_h = \frac{1}{2(h!)^2(2h-1)!} \sum_{j=0}^{h-1} (-1)^j \binom{2h}{j} (h-j)^{2h-1}. \tag{17}$$

Proof. (16) follows by combining (7) with Lemmas 4 and 5, and (15) by further using the estimate in (6).

In particular, if $p = (\Lambda + o(1))n^{(1/2h)-1}$, then both $\mathbf{E}_u X$ and $\mathbf{E}_u Y$ tend to $\kappa_h \Lambda^{2h}$ as $n \to \infty$.

The sum in (17) involves massive cancellation and does not easily yield asymptotic expressions. We therefore study the asymptotics of κ_h as $h \to \infty$ by other means.

Lemma 7. *As $h \to \infty$, $\int_0^h f_h^2(x)\,dx \sim \sqrt{\frac{3}{\pi h}}$ and thus $\kappa_h \sim \sqrt{\frac{3}{4\pi h}}(h!)^{-2}$.*

Proof. Since $|e^{it} - 1| = 2|\sin(t/2)|$, (13) yields

$$\int_0^h f_h^2(x)\,dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\frac{\sin(t/2)}{t/2}\right)^{2h} dt = \frac{1}{\pi} \int_{-\infty}^{\infty} \left(\frac{\sin t}{t}\right)^{2h} dt.$$

We divide this integral into two parts. First,

$$\int_{|t|\geq 1} \left(\frac{\sin t}{t}\right)^{2h} dt \leq 2 \int_1^{\infty} \frac{dt}{t^{2h}} = \frac{2}{2h-1} = o(h^{-1/2})$$

as $h \to \infty$.

For $|t| \leq 1$ we make the substitution $t = x/\sqrt{h}$. The Taylor series for $\sin t$ shows that $\frac{\sin t}{t} = 1 - \frac{t^2}{6} + O(t^4)$, and thus for each fixed x

$$\left(\frac{\sin(x/\sqrt{h})}{x/\sqrt{h}}\right)^{2h} = \left(1 - \frac{x^2}{6h} + O(h^{-2})\right)^{2h} \to e^{-x^2/3};$$

moreover it follows that, when $|t| \leq 1$, $\left|\frac{\sin t}{t}\right| \leq 1 - t^2/7$ and thus

$$\left(\frac{\sin(x/\sqrt{h})}{x/\sqrt{h}}\right)^{2h} \leq \left(1 - \frac{x^2}{7h}\right)^{2h} \leq e^{-2x^2/7}, \qquad |x| \leq \sqrt{h}.$$

Consequently, by dominated convergence,

$$\sqrt{h}\int_{-1}^{1}\left(\frac{\sin t}{t}\right)^{2h} dt = \int_{-\sqrt{h}}^{\sqrt{h}}\left(\frac{\sin(x/\sqrt{h})}{x/\sqrt{h}}\right)^{2h} dx \to \int_{-\infty}^{\infty} e^{-x^2/3}\, dx = \sqrt{3\pi},$$

and the result follows.

The basic Stein–Chen approximation theorem we employ is as follows:

Poisson approximation theorem for positively related variables *(Corollary 2.E.1 in [2]): Consider a sum $W = \sum_{j\in\mathcal{J}} I_j$ of indicator random variables, and set $\lambda = \mathbf{E}(W)$. Suppose that the variables I_j are increasing functions of some underlying independent random variables. Then*

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq \frac{1-e^{-\lambda}}{\lambda}\left(\mathrm{Var}(W) - \lambda + 2\sum_j \mathbf{P}^2(I_j = 1)\right).$$

Armed with the above result (or alternatively Corollary 2.C.4 in [2] together with a simple explicit coupling), we are ready to prove

Theorem 2. *Consider a subset A_n formed by randomly and independently choosing each element of $[n]$ with probability p_n. Let X and Y be as defined above and set $\lambda = \mathbf{E}_u(Y)$. Then*

$$d_{\mathrm{TV}}(\mathcal{L}(X), \mathrm{Po}(\lambda)) \to 0 \quad (n \to \infty)$$

provided that $p_n = o(1/n^{(2h-2)/(2h-1)})$. In particular, if $\mathbf{E}_u(|A_n|) = (\Lambda + o(1))n^{1/2h}$, then $\mathbf{P}_u(X = 0) \to \exp\{-\kappa_h \Lambda^{2h}\}$ $(n \to \infty)$, where κ_h is given by (17).

Proof. We clearly need to just compute a bound on $d_{\mathrm{TV}}(\mathcal{L}(Y), \mathrm{Po}(\lambda))$. The result quoted above yields immediately (the underlying independent variables are the indicators for the individual numbers in $[n]$)

$$
\begin{aligned}
d_{\mathrm{TV}}(\mathcal{L}(Y), \mathrm{Po}(\lambda)) &\leq \frac{1}{\lambda}\left(\mathrm{Var}_u(Y) - \lambda + 2\sum_{(\mathbf{a},\mathbf{b})\in\mathcal{B}(2h)} \mathbf{P}_u^2(I_{\mathbf{a},\mathbf{b}} = 1)\right) \\
&= \frac{\mathrm{Var}_u(Y)}{\lambda} - 1 + 2p^{2h} \\
&= \frac{1}{\lambda}\sum_{(\mathbf{a},\mathbf{b})\sim(\mathbf{c},\mathbf{d})} \left\{\mathbf{E}_u(I_{\mathbf{a},\mathbf{b}} I_{\mathbf{c},\mathbf{d}}) - p^{4h}\right\} + \frac{1}{\lambda}\sum_{(\mathbf{a},\mathbf{b})\in\mathcal{B}(2h)} \left\{\mathbf{E}_u(I_{\mathbf{a},\mathbf{b}}^2) - p^{4h}\right\} - 1 + 2p^{2h} \\
&\leq \frac{\Delta}{\lambda} + 2p^{2h} \preceq n^{2h-2}p^{2h-1}, \qquad (18)
\end{aligned}
$$

where the last estimate in (18) follows by (4) and (16). This establishes Theorem 2.

Even though Theorem 2 is a result about sets of random size, it can readily be translated into a statement about random subsets of a fixed size:

Corollary. *Consider a subset A_n of size k_n chosen at random from the $\binom{n}{k_n}$ such subsets of $[n] = \{1, 2, \ldots, n\}$. Then for any $h \geq 2$,*

$$k_n = (\Lambda + o(1))n^{1/2h} \Rightarrow \mathbf{P}(A_n \text{ is } B_h) \to e^{-\kappa_h \Lambda^{2h}} \quad (n \to \infty)$$

where κ_h is given by (17).

Proof. Let

$$p_n^+ = \frac{k_n}{n} + \frac{n^{1/4h}\log n}{n}$$

and

$$p_n^- = \frac{k_n}{n} - \frac{n^{1/4h}\log n}{n};$$

these choices are made for convenience only, and are certainly not unique. Then both p_n^+ and p_n^- are of the form $(\Lambda + o(1))n^{-(2h-1)/2h}$; let us use them to generate random sets A_n^+ and A_n^- as in Theorem 2. Note that

$$\mathbf{E}_u(|A_n^+|) = k_n + n^{1/4h} \log n$$

and

$$\mathrm{Var}_u(|A_n^+|) < \mathbf{E}_u(|A_n^+|) = O(n^{1/2h}).$$

Furthermore, by Chebychev's inequality,

$$\mathbf{P}_u(|A_n^+| < k_n) \preceq \frac{1}{\log^2 n} \to 0,$$

and thus for a set A_n^+ of cardinality k_n,

$$\begin{aligned}
\mathbf{P}(A_n^+ \text{ is not a } B_h \text{ set}) &= \mathbf{P}_u(A_n^+ \text{ is not a } B_h \text{ set}\big|\,|A_n^+| = k_n) \\
&\leq \mathbf{P}_u(A_n^+ \text{ is not a } B_h \text{ set}\big|\,|A_n^+| \geq k_n) \\
&\leq \frac{\mathbf{P}_u(A_n^+ \text{ is not a } B_h \text{ set})}{\mathbf{P}_u(|A_n^+| \geq k_n)} \to 1 - e^{-\lambda}
\end{aligned}$$

$(\lambda = \kappa_h \Lambda^{2h})$, so that for a randomly chosen A_n with $|A_n| = k_n$,

$$\limsup_{n\to\infty} \mathbf{P}(A_n \text{ is not a } B_h \text{ set}) \leq 1 - e^{-\lambda}.$$

The opposite inequality, which shows that

$$\liminf_{n\to\infty} \mathbf{P}(A_n \text{ is not a } B_h \text{ set}) \geq 1 - e^{-\lambda}$$

follows on using a similar argument with the set A_n^-. This proves the corollary.

Theorem 3. *Consider, under the model $\mathbf{P}_u$, the ensemble $\{I_{\mathbf{a},\mathbf{b}} : a_1 + \ldots + a_h = b_1 + \ldots + b_h; \mathbf{a} < \mathbf{b}\}$ of dependent indicator random variables. Then*

$$d_{\mathrm{TV}}\left(\mathcal{L}\{I_{\mathbf{a},\mathbf{b}}\}, \prod \mathrm{Po}(\mu_{\mathbf{a},\mathbf{b}})\right) \to 0$$

as $n \to \infty$ provided that $p = o(1/n^{(4h-3)/(4h-1)})$, where $\mu_{\mathbf{a},\mathbf{b}} = \mathbf{E}_u(I_{\mathbf{a},\mathbf{b}}) = p^{2h}$ if $\mathbf{a}, \mathbf{b}$ are two disjoint h-tuples of distinct elements, and $\mu_{\mathbf{a},\mathbf{b}} = 0$ otherwise.

Proof. Let $K_{\mathbf{a},\mathbf{b}} = I_{\mathbf{a},\mathbf{b}}$ if $(\mathbf{a}, \mathbf{b}) \in \mathcal{B}(2h)$, with $K_{\mathbf{a},\mathbf{b}} \equiv 0$ otherwise. Since

$$\begin{aligned}
& d_{\mathrm{TV}}\left(\mathcal{L}\{I_{\mathbf{a},\mathbf{b}}\}, \prod \mathrm{Po}(\mu_{\mathbf{a},\mathbf{b}})\right) \\
& \leq d_{\mathrm{TV}}\left(\mathcal{L}\{K_{\mathbf{a},\mathbf{b}}\}, \prod \mathrm{Po}(\mu_{\mathbf{a},\mathbf{b}})\right) + d_{\mathrm{TV}}\left(\mathcal{L}\{I_{\mathbf{a},\mathbf{b}}\}, \mathcal{L}\{K_{\mathbf{a},\mathbf{b}}\}\right) \\
& \leq d_{\mathrm{TV}}\left(\mathcal{L}\{K_{\mathbf{a},\mathbf{b}}\}, \prod \mathrm{Po}(\mu_{\mathbf{a},\mathbf{b}})\right) + \mathbf{E}_u(X - Y)
\end{aligned}$$

and $p = o(1/n^{(2h-2)/(2h-1)})$ which implies $\mathbf{E}_u(X - Y) \to 0$, we see that the result will follow if we can establish that $d_{\mathrm{TV}}(\mathcal{L}\{K_{\mathbf{a},\mathbf{b}}\}, \prod \mathrm{Po}(\mu_{\mathbf{a},\mathbf{b}})) \to 0$. Now we invoke Corollary 10.J.1 and Theorem 2.E in [2] which yield,

$$d_{\mathrm{TV}}\left(\mathcal{L}\{K_{\mathbf{a},\mathbf{b}}\}, \prod \mathrm{Po}(\mu_{\mathbf{a},\mathbf{b}})\right) \leq \left(\mathrm{Var}_u(Y) - \lambda + 2 \sum_{(\mathbf{a},\mathbf{b}) \in \mathcal{B}(2h)} \mathbf{P}_u^2(I_{\mathbf{a},\mathbf{b}} = 1)\right), \qquad (19)$$

where $\lambda = \mathbf{E}_u(Y)$. Now it is easy to check that the bound in (19) reduces, as in the argument leading to (18), to a term of order $n^{4h-3}p^{4h-1}$; the different rate results due to the absence of the "magic factor" of $(1 - e^{-\lambda})/\lambda$ that is present in the univariate case. This establishes the result; note that

$$\frac{1}{n^{(2h-1)/2h}} \leq \frac{1}{n^{(4h-3)/(4h-1)}} \leq \frac{1}{n^{(2h-2)/(2h-1)}}.$$

Remarks. Theorem 3 can easily be restated in terms of the measure $\mathbf{P}$; we skip the details. In any event, this result provides a nice global view of the presence/absence of various taboo (i.e., B_h-property producing) integer sums in the random set A. Also, since the total variation distance is preserved under any functional, we may use Theorem 3 to estimate probabilities such as $\mathbf{P}(a \leq \Psi \leq b)$, where Ψ equals the number of integers m which can be represented as two or more integer sums.

Acknowledgements. A preliminary version of the work reported here was presented at the IMA Workshop on Discrete Random Structures in Minneapolis in November 1993. The research of AG, NL and RR was supported by NSF grants DMS-9200409 and DMS-9322460; SJ acknowledges the support of the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine. We thank Sid Graham for introducing us to Sidon sequences, and for useful discussions. We thank the referees for finding an error in a previous version of the paper.

References

1. N. ALON and J. SPENCER, "The Probabilistic Method," John Wiley and Sons, Inc., New York, 1992.
2. A.D. BARBOUR, L. HOLST and S. JANSON, "Poisson Approximation," Oxford University Press, Oxford, 1992.
3. P. ERDŐS and S. FREUD, On sums of a Sidon sequence, *J. Number Theory* **38** (1991), 196–205.
4. S.W. GRAHAM, B_h sequences, *in* "Analytic Number Theory", Proceedings of a Conference in Honor of Heini Halberstam, eds. B.C. Berndt, H.G. Diamond and A.J. Hildebrand, Birkhäuser 1996, 431–449.

5. H. HALBERSTAM and K. ROTH, “Sequences,” Springer-Verlag, New York, 1983.

6. B. LINDSTRÖM, An inequality for B_2 sequences, *J. Comb. Theory* **6** (1969), 211–212.

7. M.B. NATHANSON, “Additive Number Theory: Inverse Problems and the Geometry of Sumsets”, Springer-Verlag, New York, 1996.

8. J. SPENCER and P. TETALI, Sidon sets with small gaps, *in* “Discrete Probability and Algorithms,” The IMA Volumes in Mathematics and its Applications, Vol. 72, D. Aldous, P. Diaconis, J. Spencer and J. M. Steele, eds., Springer-Verlag, New York, 1995.