Maximal Arithmetic Progressions in Random Subsets

Itai Benjamini* Ariel Yadin † Ofer Zeitouni‡

July 24, 2007

Abstract

Let $U^{(N)}$ denote the maximal length of arithmetic progressions in a random uniform subset of $\{0,1\}^N$. By an application of the Chen-Stein method, we show that $U^{(N)} - 2\log N/\log 2$ converges in law to an extreme type (asymmetric) distribution. The same result holds for the maximal length $W^{(N)}$ of arithmetic progressions $\pmod N$. When considered in the natural way on a common probability space, we observe that $U^{(N)}/\log N$ converges almost surely to $2/\log 2$, while $W^{(N)}/\log N$ does not converge almost surely (and in particular, $\limsup W^{(N)}/\log N \geq 3/\log 2$).

1 Introduction and Statement of Results

In this note we study the length of maximal arithmetic progressions in a random uniform subset of $\{0,1\}^N$. That is, let $\xi_1, \xi_2, \ldots, \xi_N$ be a random word in $\{0,1\}^N$, chosen uniformly. Consider the (random) set Ξ_N of elements i such that $\xi_i = 1$. Let $U^{(N)}$ denote the maximal length arithmetic progression in Ξ_N, and let $W^{(N)}$ denote the maximal length aperiodic arithmetic progression $\pmod N$ in Ξ_N. A consequence of our main result (Theorem 1) is that the expectation of both $U^{(N)}$ and $W^{(N)}$ is roughly $2\log N/\log 2$, twice the expectation of the longest run in Ξ_N, see [3],[4]. We also show that the limit law of the centered version of both $W^{(N)}$ and $U^{(N)}$ is of the same extreme type as that of the longest run in Ξ_N.

We observe two interesting phenomena:

*Faculty of Mathematics and Computer Science, The Weizmann Institute of Science, POB 26, Rehovot 76100, ISRAEL; itai.benjamini@weizmann.ac.il

†Faculty of Mathematics and Computer Science, The Weizmann Institute of Science, POB 26, Rehovot 76100, ISRAEL; ariel.yadin@weizmann.ac.il

‡Department of Mathematics, University of Minnesota, MN 55455, USA; zeitouni@math.umn.edu

- Theorem 1 states that the tails of the distribution of $W^{(N)}$ behave differently for positive and negative deviations from the mean. In particular, the probability that $W^{(N)}$ deviates by x from its mean, behaves roughly like $1-\exp\left(-2^{-(x+2)}\right)$ for positive x, and like $\exp\left(-2^{-(x+2)}\right)$ for negative x. Thus, on the positive side of the mean the tail decays exponentially, and on the negative side of the mean the tail decays doubly-exponential.

- One may construct the sets Ξ_N on the same probability space by considering an infinite sequence of i.i.d., Bernoulli random variables $\{\xi_i\}_{i=1}^{\infty}$. Proposition 2 states that with such a construction, the sequence $W^{(N)}/\log N$ converges in *probability* to the constant $2/\log 2$, but a.s. convergence does not hold. This contrasts with the behavior of $U^{(N)}$, where a.s. convergence of $U^{(N)}/\log N$ to $2/\log 2$ holds. The seemingly small change of taking arithmetic progressions that "wrap around" the torus, changes the behavior of the $\limsup$ of the sequence.

The notoriously hard extremal problem, showing that a set of integers of upper positive density contains unbounded arithmetic progressions, and its finite quantitative versions, is a well studied topic reviewed recently in [7].

1.1 The Model

Let $\xi_1, \ldots, \xi_N, \ldots,$ be i.i.d. Bernoulli random variables of mean $\mathbb{E}[\xi_i] = \frac{1}{2}$. For a non-negative integer N and $s, p \in \{1, 2, \ldots, N\}$ define

$$W_{s,p} = W_{s,p}^{(N)} \stackrel{\text{def}}{=} \max\left\{1 \le k \le N \mid \xi_s = 0\ ,\ \prod_{i=1}^{k} \xi_{s+ip \pmod N} = 1\right\}.$$

That is, we consider all arithmetic progressions $\pmod N$ in $\{1, 2, \ldots, N\}$ starting at s, with difference p, and check for the longest one of the form $0, 1, 1, \ldots$ (the role of the 0 is to avoid considering periodic progressions). $W_{s,p}$ is the length of such progression. We set

$$W^{(N)} \stackrel{\text{def}}{=} \max_{s,p} W_{s,p},$$

which is the size of the maximal arithmetic progression $\pmod N$ in $\{1, 2, \ldots, N\}$ of the form $0, 1, 1, \ldots$.

Similarly, define

$$U_{s,p} = U_{s,p}^{(N)} \stackrel{\text{def}}{=} \max\left\{1 \le k \le \lfloor\frac{N-s}{p}\rfloor \mid \xi_s = 0\ ,\ \prod_{i=1}^{k} \xi_{s+ip} = 1\right\},$$

and

$$U = U^{(N)} \stackrel{\text{def}}{=} \max_{s,p \le N} U_{s,p},$$

that is we only consider s, p, k such that $\{s + ip \mid i = 0, 1, \ldots, k\} \subseteq \{1, \ldots, N\}$.

1.2 Results

Throughout, we set $C = 2/\log 2$. Our first main result is the following extreme type limit theorem.

Theorem 1. *Let $\{x_N\}$ be a sequence such that $C \log N + x_N \in \mathbb{Z}$ for all N, and $\inf_N x_N \ge b$, for some $b \in \mathbb{R}$. Then, with $\lambda(x) = 2^{-(x+2)}$, we have*

$$\lim_{N\to\infty} \exp\left(\lambda(x_N)\right) \mathbb{P}\left[W^{(N)} \le C \log N + x_N\right] = 1\,. \tag{1}$$

Similarly, let $\{y_N\}$ be a sequence such that $C \log N - \log(2C \log N) + y_N \in \mathbb{Z}$ for all N, and $\inf_N y_N \ge b$, for some $b \in \mathbb{R}$. Then,

$$\lim_{N\to\infty} \exp\left(\lambda(y_N)\right) \mathbb{P}\left[U^{(N)} \le C \log N - \log(2C \log N) + y_N\right] = 1\,. \tag{2}$$

In particular, both $W^{(N)}/\log N$ and $U^{(N)}/\log N$ converge in probability to C.

The dichotomy in the sequential behavior of $W^{(N)}$ and $U^{(N)}$ is captured in the following proposition.

Proposition 2. *$U^{(N)}/C \log N$ converges a.s. to 1, while*

$$\limsup_{N\to\infty} \frac{W^{(N)}}{C \log N} \ge \frac{3}{2}\,.$$

In particular, $W^{(N)}/C \log N$ does not *converge a.s. to 1.*

The structure of the note is as follows. In the next section, we introduce dependency graphs and the Arratia-Goldstein-Gordon version of the Chen-Stein method, and perform preliminary computations. After these preliminary computations are in place, the short Section 3 is devoted to the proof of Theorem 1. Section 4 is devoted to the proof of Proposition 2.

2 Preliminaries and auxilliary computations

We introduce the notion of dependency graphs, and the method of Chen and Stein to prove Poisson convergence, that will play an important role in our proof.

2.1 Dependency Graphs

Let $X_1, X_2, \ldots, X_N$ be N random variables. Let G be a graph with vertices $1, 2, \ldots, N$. We use the notation $i \sim j$ to denote two vertices connected by an edge. As X_i is not independent of itself, we define $i \sim i$ for all i (this can be thought of as requiring G to have a self loop at each vertex). G is called a *dependency graph* of $\{X_i\}_{i=1}^N$ if for any vertex i,

$$X_i \text{ is independent of the set } \{X_j \ : \ j \not\sim i\}. \tag{3}$$

The notion of dependency graphs has been introduced in connection with the Lovásc Local Lemma, see [1], Chapter 5. Some other results concerning dependency graphs are [5], [6]. We emphasize that there can be many dependency graphs associated to a collection of random variables $\{X_i\}_{i=1}^N$.

We define two quantities associated with a dependency graph G of $\{X_i\}_{i=1}^N$.

$$B_1 = B_1(G) = \sum_{i=1}^N \sum_{j:X_j \sim X_i} \mathbb{E}[X_i]\,\mathbb{E}[X_j], \tag{4}$$

$$B_2 = B_2(G) = \sum_{i=1}^N \sum_{j\neq i:X_j \sim X_i} \mathbb{E}[X_i X_j]. \tag{5}$$

The following is a simplified version of Theorem 1 in [2], which in turn is an effective way to apply the Chen-Stein method:

Theorem 3 (Arratia, Goldstein, Gordon)**.** *Let $\{X_i\}_{i=1}^N$ be N Bernoulli random variables with $p_i = \mathbb{E}[X_i] > 0$. Set*

$$S_N = \sum_{i=1}^N X_i, \qquad \textit{and} \qquad \lambda = \mathbb{E}[S_N] = \sum_{i=1}^N p_i.$$

Let G be a dependency graph of $\{X_i\}_{i=1}^N$, and define B_1 and B_2 as in (4) and (5).

Let Z be a Poisson random variable with mean $\mathbb{E}[Z] = \lambda$. Then, for any $A \subset \mathbb{N}$,

$$|\mathbb{P}[S_N \in A] - \mathbb{P}[Z \in A]| \leq B_1 + B_2.$$

Theorem 3 is useful in proving convergence of sums of "almost" independent variables to the Poisson distribution.

2.2 Auxilliary Calculations

Recall that $C = 2/\log 2$. Fix $\varepsilon > 0$ and set $M = \lfloor (C+\varepsilon)\log N \rfloor$. Define

$$W'_{s,p} = W'^{(N)}_{s,p} \stackrel{\text{def}}{=} \max\left\{1 \le k \le M \;\middle|\; \xi_s = 0\ ,\ \prod_{i=1}^{k} \xi_{s+ip \pmod N} = 1\right\}, \tag{6}$$

and $W' = \max_{s,p} W'_{s,p}$. That is, we take truncated versions of $W_{s,p}$ and W.

For $1 \le s,p \le N$ and $x \in \mathbb{R}$ define

$$I_{s,p}(x) \stackrel{\text{def}}{=} \mathbf{1}_{\left\{W'_{s,p} > C\log N + x\right\}},$$

and set

$$S(x) \stackrel{\text{def}}{=} \sum_{s,p} I_{s,p}(x).$$

Note that $W' > C\log N + x$ iff $S(x) > 0$. For s,p, let $A(s,p) = \{s+ip\}_{i=0}^{M}$ be the arithmetic progression corresponding to $I_{s,p}$.

Let G be the graph with vertex set $\{(s,p)\}_{s,p=1}^{N}$, and edges defined by the relations

$$(s,p) \sim (t,q) \quad \Longleftrightarrow \quad A(s,p) \cap A(t,q) \neq \emptyset.$$

Fix $x \in \mathbb{R}$ such that $x < \varepsilon \log N$ (for large enough N this is always possible). Note that $I_{s,p}(x)$ is independent of $\left\{\xi_{j \pmod N}\ :\ j \notin A(s,p)\right\}$. Thus, G is a dependency graph of $\{I_{s,p}(x)\}_{s,p=1}^{N}$.

Define $\mathcal{D}_{s,p}(k)$ to be the number of pairs t,q with $q \neq p$ such that $|A(s,p) \cap A(t,q)| = k$.

The following combinatorial proposition proves to be useful.

Proposition 4. *For all s,p the following holds:*

$$\mathcal{D}_{s,p}(k) \le \begin{cases} (M+1)^2 N & k = 1 \\ (M+1)^2 M^2 & 2 \le k \le \frac{M}{2}+1 \\ 0 & k > \frac{M}{2}+1 \end{cases}$$

Proof. Fix $1 \le s,p \le N$.

Let $k \ge 2$. Assume that $A(s,p) \cap A(t,q) = \{x_1 < x_2 < \cdots < x_k\}$. Let $L = \operatorname{lcm}(p,q) \stackrel{\text{def}}{=} \min\left\{L \;\middle|\; \exists\, a,b \in \mathbb{Z}\ :\ L = ap = bq\right\}$.

Claim. $\{x_i\}_{i=1}^{k}$ is an arithmetic progression with $x_{i+1} - x_i = L$.

Proof. Assume that $L = ap = bq$, for $a \neq b \geq 1$. Fix $1 \leq i \leq k-1$. Since $x_{i+1} > x_i$ are both in $A(s,p) \cap A(t,q)$, we get that $x_{i+1} - x_i = a'p = b'q$ for nonnegative integers $a' \neq b' \geq 1$. So $x_{i+1} - x_i \geq L$.

Consider $x_i + L$. Since $x_i \in A(s,p) \cap A(t,q)$, and $L = ap = bq$, and since $x_i < x_i + L \leq x_{i+1}$, it follows that $x_i + L \in A(s,p) \cap A(t,q)$. So $x_i + L \geq x_{i+1}$, concluding the proof of the claim. $\square$

Let $a \neq b \geq 1$ be such that $L = ap = bq = \operatorname{lcm}(p,q)$. We have the following constraints:

$$s + (k-1)ap \leq x_1 + (k-1)L \leq s + Mp \quad \text{and} \quad t + (k-1)bq \leq x_1 + (k-1)L \leq t + Mq.$$

Thus, $2 \leq \max\{a,b\} \leq \frac{M}{k-1}$, or: $k \leq \frac{M}{2} + 1$.

So for $k > \frac{M}{2} + 1$ we get that $\mathcal{D}_{s,p}(k) = 0$.

Consider $k \leq \frac{M}{2} + 1$. Since there are at most $\frac{M}{k-1}$ choices for a and for b, and since a choice of a, b determines q, we have at most $\frac{M^2}{(k-1)^2}$ choices for q.

Remark. This can be improved to $\frac{2}{k+1} \cdot \frac{M^2}{(k-1)^2}$, with a slightly more careful analysis. We will not need this improvement.

Since $t = x_1 - iq = s + jp - iq$ for some $0 \leq i, j \leq M$, there are at most $(M+1)^2$ choices for t, once we have fixed q.

Thus, altogether, for $2 \leq k \leq \frac{M}{2} + 1$,

$$\mathcal{D}_{s,p}(k) \leq \frac{(M+1)^2 M^2}{(k-1)^2} \leq (M+1)^2 M^2.$$

If $|A(s,p) \cap A(t,q)| = 1$ then there are at most N choices for q and $(M+1)^2$ choices for t, so $\mathcal{D}_{s,p}(1) \leq (M+1)^2 N$. $\square$

Recall G defined above, a dependency graph of $\{I_{s,p}(x)\}_{s,p=1}^{N}$. Set

$$B_1 = B_1(x,G) = \sum_{s,p} \sum_{\substack{t,q \\ I_{t,q} \sim I_{s,p}}} \mathbb{E}[I_{s,p}]\, \mathbb{E}[I_{t,q}],$$

as in (4). Also, set

$$B_2 = B_2(x,G) = \sum_{s,p} \sum_{\substack{(s,p) \neq (t,q) \\ I_{t,q} \sim I_{s,p}}} \mathbb{E}[I_{s,p} I_{t,q}],$$

as in (5).

Proposition 5. *For any $\delta > 0$,*

$$\sup_{x\in(-\infty,\varepsilon\log N)} B_1(x,G) + B_2(x,G) = O(N^{\delta-1}).$$

Proof. We have that $\mathbb{E}\left[I_{t,q}\right] \le 2^{-(C\log N + x+1)}$, for all t, q.

Fix s, p. There is at most one value of t such that $|A(s,p)\cap A(t,p)| = k$. Hence, the number of pairs t, q such that $|A(s,p)\cap A(t,q)| = k$ is at most $\mathcal{D}_{s,p}(k)+1$

Thus,

$$\begin{aligned}
B_1 &\le \sum_{s,p}\sum_{k=1}^{M+1}(\mathcal{D}_{s,p}(k)+1)2^{-2(C\log N+x+1)} \\
&\le 2^{-2(x+1)}\cdot\frac{1}{N^4}\sum_{s,p}\left((M+1)^2N + 1 + \sum_{2\le k\le \frac{M}{2}+1}((M+1)^2M^2+1)\right) \\
&= O\left(\frac{M^2N+M^5}{N^2}\right) = O\left(N^{\delta-1}\right),
\end{aligned}$$

for all $\delta > 0$.

For s, p and t, q such that $|A(s,p)\cap A(t,q)| = k$ we have $\mathbb{E}\left[I_{s,p}I_{t,q}\right] \le 2^{-2(C\log N+x+1)+k}$. Also, if $q = p$ and $A(s,p)\cap A(t,q)\neq\emptyset$, then either $t\in A(s,p)$ or $s\in A(t,q)$. Thus, if $t\neq s$,

$$\mathbb{E}\left[I_{s,p}I_{t,p}\right] \le \mathbb{P}\left[\xi_s\xi_t = 0\ ,\ \xi_s\xi_t = 1\right] = 0.$$

Hence,

$$\begin{aligned}
B_2 &\le \sum_{s,p}\sum_{k=1}^{M}\mathcal{D}_{s,p}(k)2^{-2(C\log N+x+1)+k} \\
&\le 2^{-2(x+1)}\cdot\frac{1}{N^4}\sum_{s,p}\left(2(M+1)^2N + (M+1)^2M^2\cdot\sum_{2\le k\le\frac{M}{2}+1}2^k\right) \\
&= O\left(\frac{M^2N+M^4 2^{M/2}}{N^2}\right) = O\left(N^{\delta-1}\right),
\end{aligned}$$

for all $\delta > 0$. □

3 Arithmetic Progressions: Proof of Theorem 1

Since the proofs are very similar, we only consider the slightly harder $W^{(N)}$. We write W for $W^{(N)}$ whenever no confusion can occur.

We begin with the following lemma:

Lemma 6. *The sequence $W^{(N)}/C\log N$ converges to 1 in probability; i.e. for any $\delta > 0$,*

$$\lim_{N\to\infty} \mathbb{P}\left[\left|\frac{W^{(N)}}{C\log N} - 1\right| > \delta\right] = 0.$$

Further, the convergence is almost sure on the subsequence $N_k = 2^k$. Finally, the statements hold with $U^{(N)}$ replacing $W^{(N)}$.

Proof of Lemma 6. Again, we consider only $W^{(N)}$. Fix $\varepsilon > 0$. Note that

$$\mathbb{P}\left[W_{s,p} > (C+\varepsilon)\log N\right] \leq 2^{-(C+\varepsilon)\log N - 1}.$$

Thus,

$$\mathbb{P}\left[W > (C+\varepsilon)\log N\right] \leq N^2 \cdot 2^{-(C+\varepsilon)\log N - 1} = \frac{1}{2N^{\varepsilon}} \longrightarrow 0. \tag{7}$$

Now let $x = -\varepsilon \log N$, and let $Z(x)$ be a Poisson random variable with mean

$$\mathbb{E}\left[Z(x)\right] = \lambda(x) = \mathbb{E}\left[S(x)\right] = N^2 \cdot 2^{-\lfloor (C\log N + x + 2)\rfloor} \geq 2^{\varepsilon \log N - 2}.$$

Note that $\{W \leq (C-\varepsilon)\log N\}$ implies that $\{W' \leq (C-\varepsilon)\log N\}$, so using Theorem 3 and Proposition 5,

$$\begin{aligned}
& \mathbb{P}\left[W \leq (C-\varepsilon)\log N\right] \leq \mathbb{P}\left[S(x) = 0\right] \\
\leq\; & B_1(x,G) + B_2(x,G) + \mathbb{P}\left[Z(x) = 0\right] \\
\leq\; & 2^{-2(x+1)} \cdot \frac{\log^5 N}{N} + \exp\left(-2^{\varepsilon \log N - 2}\right) \longrightarrow 0,
\end{aligned} \tag{8}$$

for $\varepsilon < \frac{1}{2\log 2}$.

So for any positive $\delta < \frac{1}{4}$, we get from (7) and (8) that

$$\lim_{N\to\infty} \mathbb{P}\left[\left|\frac{W^{(N)}}{C\log N} - 1\right| > \delta\right] = 0.$$

Further, from the same estimates one has that with $Y_k = W^{(2^k)}/C\log(2^k)$, for any positive $\delta < \frac{1}{4}$,

$$\sum_{k=1}^{\infty} \mathbb{P}\left[|Y_k - 1| > \delta\right] < \infty.$$

One then deduces from the Borel-Cantelli lemma the claimed almost sure convergence. $\square$

Proof of Theorem 1. As in the proof of Lemma 6, for $x \in \mathbb{R}$, let $Z(x)$ be a Poisson random variable with mean

$$\mathbb{E}\left[Z(x)\right] = \lambda(x) = \mathbb{E}\left[S(x)\right] = N^2 \cdot 2^{-\lfloor (C\log N + x + 2)\rfloor}.$$

If $C\log N + x \in \mathbb{Z}$, then $\lambda(x) = 2^{-(x+2)}$.

Note that $W' > C\log N + x$ iff $S(x) > 0$. By Theorem 3 and Proposition 5,

$$\begin{aligned} & \left|\mathbb{P}\left[W' > C\log N + x\right] - \mathbb{P}\left[Z(x) \neq 0\right]\right| = \left|\mathbb{P}\left[S(x) > 0\right] - \mathbb{P}\left[Z(x) > 0\right]\right| \\ \leq \; & B_1(x, G) + B_2(x, G) = O\left(N^{\delta - 1}\right). \end{aligned}$$

We also have the equality

$$\{W > C\log N + x\} = \{W > (C+\varepsilon)\log N\} \bigcup \{W' > C\log N + x\}.$$

Thus, for $0 < \delta < 1$,

$$\begin{aligned} & \left|\mathbb{P}\left[W \leq C\log N + x\right] - e^{-\lambda(x)}\right| \\ \leq \; & \mathbb{P}\left[W > (C+\varepsilon)\log N\right] + \left|\mathbb{P}\left[W' > C\log N + x\right] - \mathbb{P}\left[Z(x) \neq 0\right]\right| \\ \leq \; & O\left(N^{-\varepsilon}\right) + O\left(N^{\delta - 1}\right). \end{aligned}$$

Let $\{x_N\}$ be a sequence such that $C\log N + x_N \in \mathbb{Z}$ for all N. If $\inf_N x_N \geq b \in \mathbb{R}$, then $\exp\left(\lambda(x_N)\right)$ is a bounded sequence. Thus,

$$\left|\exp\left(\lambda(x_N)\right)\mathbb{P}\left[W^{(N)} \leq C\log N + x\right] - 1\right| = O\left(N^{-\varepsilon}\right) \longrightarrow 0.$$

$\square$

4 Convergence in Probability vs. a.s. Convergence

We begin with the following easy consequence of Lemma 6 applied to $U^{(N)}$.

Proposition 7. *$U^{(N)}/C\log N$ converges a.s. to 1.*

Proof. The main observation is that $U^{(N)}$ is a monotone increasing sequence. That is, a.s. for all N, $U^{(N)} \leq U^{(N+1)}$. Thus, a.s. for all N, setting $k = \lfloor \log_2 N \rfloor$, we have

$$\frac{U^{(2^k)}}{\log(2^k)} \cdot \frac{\log(2^k)}{\log(2^{k+1})} \quad \leq \quad \frac{U^{(N)}}{\log N} \leq \frac{U^{(2^{k+1})}}{\log(2^{k+1})} \cdot \frac{\log(2^{k+1})}{\log(2^k)}$$

Since $U^{(2^k)}/\log(2^k)$ converges a.s. to C and $\log(2^k)/\log(2^{k+1})$ converges a.s. to 1, we get a.s. convergence of $\frac{U^{(N)}}{\log N}$ to C. $\square$

We turn to the

Proof of Proposition 2. In view of Proposition 7, it remains only to consider the statement concerning $W^{(N)}$. Toward this end, fix $0 < \beta < 1$. Let $M_N = \lceil (2+\beta)\log_2 N \rceil$. So $N^{2+\beta} \le 2^{M_N} \le 2N^{2+\beta}$. Define

$$I(s,p,N) = \mathbf{1}_{\left\{W_{s,p}^{(N)} \ge M_N\right\}}.$$

That is, $I(s,p,N)$ is the indicator function of the event that $\xi_s = 0$ and $\prod_{i=1}^{M_N} \xi_{s+ip \pmod N} = 1$. Set

$$\operatorname{Cov}((s,p,N),(s',p',N')) = \mathbb{E}\left[I(s,p,N)I(s',p',N')\right] - \mathbb{E}\left[I(s,p,N)\right]\mathbb{E}\left[I(s',p',N')\right].$$

For simplicity of notation, for $a,b \in \mathbb{Z}$, we denote $[a,b] = \left\{ z \in \mathbb{Z} \;\middle|\; a \le z \le b \right\}$. Let $\mathcal{L}_n = \left\{(s,p,N) : N \in [n,2n], s \in [1,N], p \in [N/6, N/5]\right\}$.

Lemma 8. *The following holds:*

$$\sum_{(s,p,N)\in\mathcal{L}_n} \sum_{(s',p',N')\in\mathcal{L}_n} \operatorname{Cov}((s,p,N),(s',p',N')) \le O\left(n^{1-\beta}\log^7(n)\right).$$

Proof. For any $N, N' \in [n,2n]$, $s,p \in [1,N]$ and $s',p' \in [1,N']$, we have that

$$\operatorname{Cov}((s,p,N),(s',p',N')) \le 2^{-(M_N+M_{N'})}\left(2^k - 1\right), \tag{9}$$

with

$$k = \left| \{s+ip \pmod N\}_{i\in[1,M_N]} \cap \{s'+jp' \pmod{N'}\}_{j\in[1,M_{N'}]} \right|.$$

Thus, the proof of the lemma is based on controlling the cardinality of the collection of triples $(s',p',N') \in \mathcal{L}_n$, whose associated arithmetic progression intersects in a prescribed number of points the arithmetic progression associated with a given triple $(s,p,N) \in \mathcal{L}_n$. We divide our estimates into three: intersection at one point, intersection at two points or more, and intersection at $2C\log(2n)/5$ points or more.

For $N, N' \in [n,2n]$ and $p \in [1,N]$ define $\mathcal{T}(N,N',p)$ to be the set of all triples (s,s',p') such that $s \in [1,N]$, $s',p' \in [1,N']$ and

$$\left| \{s+ip \pmod N\}_{i\in[1,M_N]} \cap \{s'+jp' \pmod{N'}\}_{j\in[1,M_{N'}]} \right| = 1. \tag{10}$$

Similarly, define $\mathcal{S}(N,N',p)$ to be the set of all such triples (s,s',p') such that

$$\left| \{s+ip \pmod N\}_{i\in[1,M_N]} \cap \{s'+jp' \pmod{N'}\}_{j\in[1,M_{N'}]} \right| \ge 2. \tag{11}$$

Finally, define $\mathcal{U}(s,p,N)$ the set of all triples $(s',p',N')\in\mathcal{L}_n$ such that

$$\left|\{s+ip \pmod N\}_{i\in[1,M_N]}\cap\{s'+jp' \pmod{N'}\}_{j\in[1,M_{N'}]}\right|\geq 2M_{N'}/5. \tag{12}$$

We have the following estimates.

Proposition 9. *For large enough n, the following holds: for all $N,N'\in[n,2n]$ and $p\in[1,N]$,*

$$|\mathcal{T}(N,N',p)|\leq n^2\log^5(n).$$

Proposition 10. *For large enough n, the following holds: for all $N,N'\in[n,2n]$ and $p\in[1,N]$,*

$$|\mathcal{S}(N,N',p)|\leq n\log^9(n).$$

Proposition 11. *For large enough n, the following holds: for all $(s,p,N)\in\mathcal{L}_n$,*

$$|\mathcal{U}(s,p,N)|\leq\log^7(n).$$

Assuming Propositions 9, 10, 11, we have

$$\begin{aligned}
& \sum_{(s,p,N)\in\mathcal{L}_n}\sum_{(s',p',N')\in\mathcal{L}_n}\mathrm{Cov}((s,p,N),(s',p',N')) \qquad (13)\\
\leq\ & \sum_{N,N'=n}^{2N}\sum_{p=1}^{N}2^{-(M_N+M_{N'})}\,|\mathcal{T}(N,N',p)|+\sum_{N,N'=n}^{2n}\sum_{p=1}^{N}2^{-(M_N+M_{N'})}\,|\mathcal{S}(N,N',p)|\,2^{2M_{N'}/5}\\
& +\sum_{(s,p,N)\in\mathcal{L}_n}|\mathcal{U}(s,p,N)|\,2^{-M_n}\\
\leq\ & O(n^{1-2\beta}\log^5(n))+O(n^{4/5-8\beta/5}\log^9(n))+O(n^{1-\beta}\log^7(n))\leq O(n^{1-\beta}\log^7(n))\,,
\end{aligned}$$

which completes the proof of the lemma. □

Returning to the proof of Proposition 2, let

$$\Lambda(n)=\sum_{(s,p,N)\in\mathcal{L}_n}I(s,p,N)$$

and note that for all large enough n,

$$\mathbb{E}[\Lambda(n)]=\sum_{(s,p,N)\in\mathcal{L}_n}\mathbb{E}\left[I(s,p,N)\right]\quad=\quad\Omega\left(n^{1-\beta}\right) \tag{14}$$

while from Lemma 8,

$$\mathrm{Var}\left(\Lambda(n)\right)=\mathrm{Var}\left(\sum_{(s,p,N)\in\mathcal{L}_n}[I(s,p,N)]\right)\quad=\quad O\left(n^{1-\beta}\log^7(n)\right)\,. \tag{15}$$

Thus,

$$\mathbb{P}\left[\Lambda(n)=0\right] \le \frac{\mathrm{Var}\left(\Lambda(n)\right)}{\left(\mathbb{E}\,\Lambda(n)\right)^2} = O\left(n^{\beta-1}\log^7(n)\right). \tag{16}$$

Since

$$\mathbb{P}\left[\exists\ N\in[n,2n]\ :\ W^{(N)} > \frac{2+\beta}{\log 2}\log N\right] \ge \mathbb{P}\left[\Lambda(n)>0\right],$$

it follows from (16) that

$$\sum_{k=0}^{\infty}\mathbb{P}\left[\max_{N\in[2^k,2^{k+1}]}\frac{W^{(N)}}{\log N}\le\frac{2+\beta}{\log 2}\right]<\infty.$$

By the Borel-Cantelli lemma, we get that a.s.

$$\limsup_{N\to\infty}\frac{W^{(N)}}{\log N}\ge\limsup_{k\to\infty}\max_{N\in[2^k,2^{k+1}]}\frac{W^{(N)}}{\log N}>\frac{2+\beta}{\log 2}.$$

Since $\beta\in(0,1)$ is arbitrary, this completes the proof of Proposition 2. □

Proof of Proposition 9. If $(s,s',p')\in\mathcal{T}(N,N',p)$ then (10) implies that there exist $i\in[1,M_N]$, $j\in[1,M_{N'}]$, $k_i\in[0,M_N]$ and $k'_j\in[0,M_{N'}]$ such that

$$s+ip-k_iN=s'+jp'-k'_jN' \tag{17}$$

There are at most $(2n)^2$ choices for s' and p'. There exists some universal constant K such that there are at most $K\log(n)$ choices for each of i,j,k_i,k'_j. Choosing s',p',i,j,k_i,k'_j determines s. Thus, we have shown that $|\mathcal{T}(N,N',p)|\le 4Kn^2\log^4(n)\le n^2\log^5(n)$ for large enough n. □

Proof of Proposition 10. If $(s,s',p')\in\mathcal{S}(N,N',p)$ then (11) implies that there exist $i,r\in[1,M_N]$ and $j,\ell\in[1,M_{N'}]$ such that $(i,j)\neq(r,\ell)$ and

$$\begin{aligned} s+ip \pmod N &= s'+jp' \pmod{N'} \\ \text{and}\quad s+rp \pmod N &= s'+\ell p' \pmod{N'}. \end{aligned} \tag{18}$$

Note that for any $i\in[1,M_N]$ there exists $k_i\in[0,M_N]$ such that $s+ip \pmod N = s+ip-k_iN$. Similarly, for any $j\in[1,M_{N'}]$ there exists $k'_j\in[0,M_{N'}]$ such that $s'+jp' \pmod{N'} = s'+jp'-k'_jN'$. Plugging this into (18), and subtracting equations, we get that there exist $i,r\in[1,M_N]$, $j,\ell\in[1,M_{N'}]$, $k_i,k_r\in[0,M_N]$ and $k'_j,k'_\ell\in[0,M_{N'}]$ such that

$$(r-i)p+(k_i-k_r)N=(\ell-j)p'+(k'_j-k'_\ell)N'. \tag{19}$$

There exists some universal constant K such that there are at most $K\log(n)$ choices for each of $i, r, j, \ell, k_i, k_r, k'_j, k'_\ell$, and $2n$ choices for s. After choosing $i, r, j, \ell, k_i, k_r, k'_j, k'_\ell, s$, (18) and (19) determine s' and p'. Thus, we have shown that for large enough n,

$$|\mathcal{S}(N, N', p)| \le 2n\left(K\log(n)\right)^8 \le n\log^9(n).$$

$\square$

Proof of Proposition 11. Let $A = \left\{s + ip \pmod N \;\middle|\; i \in [1, M_N]\right\}$, and let $(s', p', N') \in \mathcal{U}(s, p, N)$. For $i = 1, 6, 11, \ldots$, let

$$Z_i = \left\{s' + (i+r)p' \pmod{N'} \;\middle|\; r = 0, 1, \ldots, 4\right\}.$$

This is a partition of the arithmetic progression into packets of five elements. We then have, by the definition of $\mathcal{U}(s, p, N)$,

$$\frac{2M_{N'}}{5} \le \sum_i |Z_i| \le \frac{M_{N'}}{5}\max_i |Z_i|.$$

So there exists some set Z_i such that $|Z_i| \ge 2$. This implies that there exist $x < y \in A \cap [1, N']$, $i \in [1, M_{N'}]$, and $r \in [1, 4]$ such that

$$\begin{aligned} s + ip' \pmod{N'} &= x\ , \\ s + (i+r)p' \pmod{N'} &= y. \end{aligned} \tag{20}$$

Subtracting equations, and using the fact that $rp' < N'$, we get that

$$rp' = y - x. \tag{21}$$

Moreover, (12) also implies that there must exist an integer j (perhaps negative) with $\frac{1}{5}M_{N'} \le |j| \le M_{N'}$, and $z \in A \cap [1, N']$, such that

$$s + (i+j)p' \pmod{N'} = z. \tag{22}$$

For large enough n, we have that $|j| \ge 7$. Since $7p' > N'$, we get by subtracting (20) from (22),

$$jp' + kN' = z - x, \tag{23}$$

for some $k \ne 0$, such that $|k| \le M_{N'}$.

Since $kr \ne 0$, equations (21) and (23) have at most one solution for p', N', in terms of x, y, z, r, j and k. Since there are at most $|A|^3 \le M_N^3$ choices for x, y and z, at most 4 choices for r, and at most $4M_{N'}^2$ choices for j and k, we get that there are at most $16|A|^3M_{N'}^2$ choices for p', N'. Also, there are at most $M_{N'}$ choices for i in (20), and fixing p', N', x and i determines s'. Thus,

$$|\mathcal{U}((s,p,N))| \le 16M_N^3M_{N'}^3.$$

□

Open Problem. We conjecture that in fact,

$$\limsup_{N\to\infty}\frac{W^{(N)}}{\log N} = \frac{3}{2}.$$

Acknowledgements. We thank Tim Austin, Ori Gurel-Gurevich and Gady Kozma for useful discussions.

References

[1] N. Alon, J. H. Spencer. *The Probabilistic Method* (2000), John Wiley & Sons. New York.

[2] R. Arratia, L. Goldstein, L. Gordon. Two moments suffice for Poisson approximations: The Chen-Stein method. *Ann. Probab.* **17** (1989), 9–25.

[3] P. Erdös and A. Rényi, On a new law of large numbers, *J. Analyse Math.* **23** (1970) 103–111.

[4] P. Erdös and P. Revesz, On the length of the longest head-run. *Colloq. Math. Soc. Janos Bolyai* **16** (1977) 219–228.

[5] R. Gradwohl, A. Yehudayoff, t-Wise Independence with Local Dependencies. `arXiv:math.PR/0706.1637`

[6] S. Janson, Large deviations for sums of partly dependent random variables. *Random Structures Algorithms* **24** (2004), no. 3, 234–248.

[7] T. Tao, What is good mathematics. (2007). `arXiv:math.HO/0702.5396`