MEASURES OF PSEUDORANDOMNESS FOR FINITE SEQUENCES: TYPICAL VALUES

N. ALON, Y. KOHAYAKAWA, C. MAUDUIT, C. G. MOREIRA, AND V. RÖDL

Abstract. Mauduit and Sárközy introduced and studied certain numerical parameters associated to finite binary sequences $E_N \in \{-1,1\}^N$ in order to measure their 'level of randomness'. Those parameters, the *normality measure* $\mathcal{N}(E_N)$, the *well-distribution measure* $W(E_N)$, and the *correlation measure* $C_k(E_N)$ *of order* k, focus on different combinatorial aspects of E_N. In their work, amongst others, Mauduit and Sárközy (i) investigated the relationship among those parameters and their minimal possible value, (ii) estimated $\mathcal{N}(E_N)$, $W(E_N)$, and $C_k(E_N)$ for certain explicitly constructed sequences E_N suggested to have a 'pseudorandom nature', and (iii) investigated the value of those parameters for genuinely random sequences E_N.

In this paper, we continue the work in the direction of (iii) above and determine the order of magnitude of $\mathcal{N}(E_N)$, $W(E_N)$, and $C_k(E_N)$ for typical E_N. We prove that, for most $E_N \in \{-1,1\}^N$, both $W(E_N)$ and $\mathcal{N}(E_N)$ are of order $\sqrt{N}$, while $C_k(E_N)$ is of order $\sqrt{N \log \binom{N}{k}}$ for any given $2 \le k \le N/4$.

Date: Copy produced on April 12, 2006.

1991 *Mathematics Subject Classification.* 68R15.

Key words and phrases. Random sequences, pseudorandom sequences, finite words, normality, correlation, well-distribution, discrepancy.

Part of this work was done at IMPA, whose hospitality the authors gratefully acknowledge. This research was partially supported by IM-AGIMB/IMPA. The first author was partially supported by the Israel Science Foundation, by a USA-Israeli BSF grant, by NSF grant CCR-0324906, by the James Wolfensohn fund and by the State of New Jersey. The second author was partially supported by FAPESP and CNPq through a Temático–ProNEx project (Proc. FAPESP 2003/09925–5) and by CNPq (Proc. 306334/2004–6 and 479882/2004–5). The third author was partially supported by the Brazil/France Agreement in Mathematics (Proc. CNPq 60-0014/01-5 and 69-0140/03-7). The fourth author was partially supported by MCT/CNPq through a ProNEx project (Proc. CNPq 662416/1996–1) and by CNPq (Proc. 300647/95–6). The fifth author was partially supported by NSF Grant DMS 0300529. The authors gratefully acknowledge the support of a CNPq/NSF cooperative grant (910064/99–7, 0072064).

Contents

1. Introduction and statement of results

1.1. Measures of pseudorandomness for finite binary sequences. In a series of papers, Mauduit and Sárközy studied finite pseudorandom binary sequences $E_N = (e_1, \dots, e_N) \in \{-1,1\}^N$. In particular, they investigated in [11] the 'measures of pseudorandomness' to be defined shortly. The readers interested in detailed discussions concerning the definitions below and further related literature are referred to [10] and [11].

Let $k \in \mathbb{N}$, $M \in \mathbb{N}$, $X \in \{-1,1\}^k$, $a \in \mathbb{Z}$, $b \in \mathbb{N}$, $b > 0$, and $D = (d_1, \dots, d_k) \in \mathbb{N}^k$ with $0 \le d_1 < \dots < d_k < N$ be given. Below, we write $\operatorname{card} S$ for the cardinality of a set S, and if S is a set of numbers, then we write $\sum S$ for the sum $\sum_{s \in S} s$. We let

$$T(E_N, M, X) = \operatorname{card}\{n\colon 0 \le n < M,\ n+k \le N,\ \text{and} \\ (e_{n+1}, e_{n+2}, \dots, e_{n+k}) = X\}, \quad (1)$$

$$U(E_N, M, a, b) = \sum\{e_{a+jb}\colon 1 \le j \le M,\ 1 \le a+jb \le N \text{ for all } j\}, \quad (2)$$

and

$$V(E_N, M, D) = \sum\{e_{n+d_1} e_{n+d_2} \dots e_{n+d_k}\colon 1 \le n \le M,\ n+d_k \le N\}. \quad (3)$$

In words, $T(E_N, M, X)$ is the number of occurrences of the pattern X in E_N, counting only those occurrences whose first symbol is among the first M elements of E_N. The quantity $U(E_N, M, a, b)$ is the 'discrepancy' of E_N on an M-element arithmetic progression contained in $\{1, \dots, N\}$. Finally, $V(E_N, M, D)$ is the 'correlation' among k length M segments of E_N 'relatively positioned' according to $D = (d_1, \dots, d_k)$.

The *normality measure* of E_N is defined as

$$\mathcal{N}(E_N) = \max_k \max_X \max_M \left| T(E_N, M, X) - \frac{M}{2^k} \right|, \tag{4}$$

where the maxima are taken over all $k \le \log_2 N$, $X \in \{-1,1\}^k$, and $0 < M \le N + 1 - k$. The *well-distribution measure* of E_N is defined as

$$W(E_N) = \max\{|U(E_N, M, a, b)|\colon\ a,\ b, \text{ and } M \text{ such that } 1 \le a + b < a + Mb \le N\}. \tag{5}$$

Finally, the *correlation measure of order* k of E_N is defined as

$$C_k(E_N) = \max\{|V(E_N, M, D)|\colon M \text{ and } D \text{ such that } M + d_k \le N\}. \tag{6}$$

1.2. **Typical values of $W(E_N)$, $C_k(E_N)$, and $\mathcal{N}(E_N)$.** In [4], Cassaigne, Mauduit, and Sárközy studied, amongst others, the values of $W(E_N)$ and $C_k(E_N)$ for random binary sequences E_N, with all the 2^N sequences in $\{-1,1\}^N$ equiprobable, and the minimum possible values for $W(E_N)$ and $C_k(E_N)$. They proved the following theorems. (Below and elsewhere in this paper, we write log for the natural logarithm.)

Theorem A. *For all $\varepsilon > 0$, there are numbers $N_0 = N_0(\varepsilon)$ and $\delta = \delta(\varepsilon) > 0$ such that for $N \ge N_0$ we have*

$$\delta\sqrt{N} < W(E_N) < 6\sqrt{N \log N} \tag{7}$$

with probability at least $1 - \varepsilon$.

Theorem B. *For every integer $k \ge 2$ and real $\varepsilon > 0$, there are numbers $N_0 = N_0(\varepsilon, k)$ and $\delta = \delta(\varepsilon, k) > 0$ such that for all $N \ge N_0$ we have*

$$\delta\sqrt{N} < C_k(E_N) < 5\sqrt{kN \log N} \tag{8}$$

with probability at least $1 - \varepsilon$.

As it turns out, an improvement of the upper bound in Theorem A may be deduced from a proof of a closely related result of Erdős and Spencer. Indeed, an argument in [7, Chapter 8] tells us that one may drop the logarithmic factor in (7), at the expense of increasing the multiplicative constant.

In this paper, we give stronger versions of Theorems A and B.

Theorem 1. *For any given $\varepsilon > 0$ there exist N_0 and $\delta > 0$ such that if $N \ge N_0$, then*

$$\delta\sqrt{N} < W(E_N) < \frac{1}{\delta}\sqrt{N} \tag{9}$$

with probability at least $1 - \varepsilon$.

Theorem 1 above is essentially proved in Erdős and Spencer [7, Chapter 8]. However, we give our alternative proof for this result because an idea in this proof is also used in the proofs of Theorems 4, 5, and 6 below.

We next state a result that establishes the typical order of magnitude of $C_k(E_N)$ for a wide range of k, including values of k proportional to N.

Theorem 2. *Let $0 < \varepsilon_0 \le 1/16$ be fixed and let $\varepsilon_1 = \varepsilon_1(N) = (\log\log N)/\log N$. There is a constant $N_0 = N_0(\varepsilon_0)$ such that if $N \ge N_0$, then, with probability at least $1-\varepsilon_0$, we have*

$$\frac{2}{5}\sqrt{N\log\binom{N}{k}} < C_k(E_N) < \sqrt{(2+\varepsilon_1)N\log\left(N\binom{N}{k}\right)} < \sqrt{(3+\varepsilon_0)N\log\binom{N}{k}} < \frac{7}{4}\sqrt{N\log\binom{N}{k}} \tag{10}$$

for every integer k with $2 \le k \le N/4$.

Our next result tells us that $C_k(E_N)$ is concentrated around its mean $\mathbb{E}(C_k)$ in the case in which k is small.

Theorem 3. *For any fixed constant $\varepsilon > 0$ and any integer function $k = k(N)$ with $2 \le k \le \log N - \log\log N$, there is a constant N_0 for which the following holds. If $N \ge N_0$, then the probability that*

$$1-\varepsilon < \frac{C_k(E_N)}{\mathbb{E}(C_k)} < 1+\varepsilon \tag{11}$$

holds is at least $1-\varepsilon$.

We suspect that the upper bound on k in Theorem 3 may be weakened. We now turn to the normality measure $\mathcal{N}(E_N)$.

Theorem 4. *For any given $\varepsilon > 0$ there exist N_0 and $\delta > 0$ such that if $N \ge N_0$, then*

$$\delta\sqrt{N} < \mathcal{N}(E_N) < \frac{1}{\delta}\sqrt{N} \tag{12}$$

with probability at least $1-\varepsilon$.

We shall show that the lower bounds in Theorems 1 and 4 (i.e., the lower bounds in (9) and (12)) are in a sense best possible. In Section 3 we prove the following two results.

Theorem 5. *For any $\delta > 0$, there is $c(\delta) > 0$ and $N_0 = N_0(\delta)$ such that, for any $N \ge N_0$, we have*

$$\mathbb{P}(W(E_N) < \delta\sqrt{N}) > c(\delta). \tag{13}$$

Theorem 6. *For any $\delta > 0$, there is $c(\delta) > 0$ and $N_0 = N_0(\delta)$ such that, for any $N \ge N_0$, we have*

$$\mathbb{P}(\mathcal{N}(E_N) < \delta\sqrt{N}) > c(\delta). \tag{14}$$

In Section 4, we make some simple remarks to show that the upper bounds in (9) and (12) are in a sense best possible. Those remarks and Theorems 5 and 6 tells us that $W(E_N)/\sqrt{N}$ and $\mathcal{N}(E_N)/\sqrt{N}$ do not converge in distribution as $N \to \infty$ to a distribution that is concentrated at a point.

1.3. **Minimal values of $W(E_N)$, $C_k(E_N)$, and $\mathcal{N}(E_N)$.** The minimal possible values of $W(E_N)$, $C_k(E_N)$, and $\mathcal{N}(E_N)$ have also been investigated in the literature, see, e.g., [1, 4, 9, 15]. We include this brief section for the convenience of the reader. It follows from the classical results of Roth [15] and Matoušek and Spencer [9] that the order of magnitude of

$$\min\left\{W(E_N)\colon E_N \in \{-1,1\}^N\right\} \tag{15}$$

is $N^{1/4}$. For any fixed even $k \geq 2$, the order of magnitude of $\min_{E_N} C_k(E_N)$, where the minimum is again taken over all $E_N \in \{-1,1\}^N$, has been established up to a polylogarithmic factor recently:

$$\sqrt{\frac{1}{2}\left\lfloor\frac{N}{k+1}\right\rfloor} < \min_{E_N} C_k(E_N) \leq \frac{7}{4}\sqrt{kN\log N}, \tag{16}$$

where we suppose $N \geq N_0(k)$ for the upper bound. The lower bound in (16) is proved in [1], whereas the upper bound in (16) follows from Theorem 2 above. It is easy to see that, for any odd k, we have $\min_{E_N} C_k(E_N) = 1$. Some further results giving lower bounds for $C_k(E_N)$ are given in [1].

Turning to the normality measure $\mathcal{N}(E_N)$, we mention that the best bounds that we know of for the minimal value of $\mathcal{N}(E_N)$ ($E_N \in \{-1,1\}^N$) are as follows:

$$\left(\frac{1}{2}+o(1)\right)\log_2 N \leq \min_{E_N}\mathcal{N}(E_N) \leq 3N^{1/3}(\log N)^{2/3} \tag{17}$$

for all large enough N. The upper bound in (17) is proved in [1] constructively, where a suitable, explicit algebraic construction based on finite fields is given. It is interesting to compare the upper bounds in (17) and (12) in Theorem 4. The lower bound in (17) may be proved simply (see the remarks at the end of Section 2.4). It would be interesting to close the rather wide gap in (17). The following problem was already raised in [1].

Problem 7. *Is there an absolute constant $\alpha > 0$ for which we have*

$$\min_{E_N}\mathcal{N}(E_N) > N^{\alpha} \tag{18}$$

for all large enough N?

The authors believe that the answer to Problem 7 is positive.

2. Estimates for $W(E_N)$, $C_k(E_N)$, and $\mathcal{N}(E_N)$ for random sequences E_N

We shall prove Theorems 1–4 in this section. Recall that these results concern random elements E_N from the uniform space $\{-1,1\}^N$. In this section, unless stated otherwise, E_N will always stand for such a random sequence.

2.1. **Estimates for the binomial distribution.** We give in this section some standard facts about the binomial distribution.

We start with the following form of the de Moivre–Laplace theorem (see, e.g., [3, Chapter I, Theorem 6]). Let $\mathrm{Bi}(n,p)$ denote the binomial distribution with parameters n and p. Moreover, write $S(n,p)$ for the sum of n independent Bernoulli random variables with mean p. We first consider the symmetric case $p=1/2$.

Fact 8. *(i) For any $\ell=\ell(n)\in\mathbb{Z}$ with $\ell=o(n^{2/3})$, we have*

$$\mathbb{P}\left(S(n,1/2)=\left\lfloor\frac{n}{2}\right\rfloor+\ell\right)=\frac{1}{2^n}\binom{n}{\lfloor n/2\rfloor+\ell}$$
$$=\sqrt{\frac{2}{\pi n}}\,e^{-2\ell^2/n}(1+o(1)), \tag{19}$$

(ii) For any $c=c(n)>0$ with $c=o(n^{1/6})$, we have

$$\mathbb{P}\left(S(n,1/2)\geq\left\lfloor\frac{n}{2}\right\rfloor+c\sqrt{n}\right)=\sum_{\ell\geq c\sqrt{n}}\frac{1}{2^n}\binom{n}{\lfloor n/2\rfloor+\ell}$$
$$=\left(\sqrt{\frac{2}{\pi}}+o(1)\right)\left(\int_c^\infty e^{-2x^2}\mathrm{d}x\right). \tag{20}$$

In particular, if we further have that $c\to\infty$, then

$$\mathbb{P}\left(S(n,1/2)\geq\left\lfloor\frac{n}{2}\right\rfloor+c\sqrt{n}\right)=\frac{e^{-2c^2}}{2c\sqrt{2\pi}}(1+o(1)). \tag{21}$$

(iii) The estimates (20) and (21) hold for the lower tail

$$\mathbb{P}\left(S(n,1/2)\leq\left\lfloor\frac{n}{2}\right\rfloor-c\sqrt{n}\right)$$

as well.

In what follows, we shall often be concerned with sums of ±1 independent random variables. We let

$$S^{\pm}(n)=\sum_{1\leq i\leq n}X_i \tag{22}$$

where X_i $(1\leq i\leq n)$ are independent random variables with mean 0, that is,

$$\mathbb{P}(X_i=-1)=\mathbb{P}(X_i=+1)=1/2.$$

Clearly, $(S^{\pm}(n)+n)/2$ is binomially distributed with parameters n and $1/2$. Let us now state a well known estimate for large deviations of $S^{\pm}(n)$ (see, e.g., [2, Appendix A]).

Fact 9. *Let X_i $(1\leq i\leq n)$ be independent ±1 random variables with mean 0. Let $S^{\pm}(n)=\sum_{1\leq i\leq n}X_i$. For any real number $a>0$, we have*

$$\mathbb{P}(S^{\pm}(n)>a)<e^{-a^2/2n}. \tag{23}$$

We now prove a lower estimate for the symmetric binomial distribution. As usual, we let $\{x\} = x - \lfloor x \rfloor$.

Fact 10. *Let n and ℓ be integers with*

$$-\left\lfloor \frac{n}{2} \right\rfloor \le \ell \le \left\lceil \frac{n}{2} \right\rceil. \tag{24}$$

If n is sufficiently large, then

$$\mathbb{P}\left(S(n,1/2) = \left\lfloor \frac{n}{2} \right\rfloor + \ell\right) = \frac{1}{2^n}\binom{n}{\lfloor n/2\rfloor + \ell}$$
$$\ge 2^{-4(\ell+\{n/2\})^2/n-n}\binom{n}{\lfloor n/2\rfloor} = (1+o(1))2^{-4(\ell+\{n/2\})^2/n}\sqrt{\frac{2}{\pi n}}. \tag{25}$$

Proof. We shall in fact prove the following statement, which implies Fact 10:

(†) *if $0 \le \ell \le n/2$ and n is large enough, then*

$$\binom{n}{\lfloor n/2\rfloor - \ell} \ge 2^{-4(\ell+\{n/2\})^2/n}\binom{n}{\lfloor n/2\rfloor}. \tag{26}$$

We remark that if n is even, or else if n is odd and $\ell \le 0$, then (†) is equivalent to what is claimed in Fact 10. If n is odd and $\ell > 0$, then (†) is slightly stronger, because in this case we have

$$\binom{n}{\lfloor n/2\rfloor + \ell} = \binom{n}{\lceil n/2\rceil - \ell} = \binom{n}{\lfloor n/2\rfloor - \ell + 1}$$
$$\ge 2^{-4(\ell-1+\{n/2\})^2/n}\binom{n}{\lfloor n/2\rfloor} \ge 2^{-4(\ell+\{n/2\})^2/n}\binom{n}{\lfloor n/2\rfloor}, \tag{27}$$

where we used (26) in the first inequality in (27).

We shall now prove (†). Let

$$a(\ell) = 2^{4(\ell+\{n/2\})^2/n}\binom{n}{\lfloor n/2\rfloor - \ell} \tag{28}$$

for all $0 \le \ell \le n/2$. It is easy to check that

$$a(0),\ a(\lfloor n/2\rfloor) \ge \binom{n}{\lfloor n/2\rfloor} \tag{29}$$

for all large enough n. Therefore, it suffices to show that

$$\min\{a(\ell) \colon 0 \le \ell \le n/2\}$$

is achieved either at $\ell = 0$ or at $\ell = \lfloor n/2\rfloor$. In particular, if we show that $a(\ell)$ is unimodal, we are done. Let

$$b(\ell) = \frac{a(\ell+1)}{a(\ell)} = 2^{4(2\ell+1+2\{n/2\})/n}\frac{\lfloor n/2\rfloor - \ell}{\lceil n/2\rceil + \ell + 1} \tag{30}$$

for all $0 \le \ell \le n/2 - 1$. Observe that, for all large enough n, we have

$$b(0) = 2^{4/n}\frac{\lfloor n/2\rfloor}{\lceil n/2\rceil + 1} > 1 \tag{31}$$

and

$$b(\lfloor n/2 \rfloor - 1) = \frac{1}{n} 2^{4(1-1/n)} < 1. \tag{32}$$

We now show that $b(\ell)$ is itself unimodal. Let

$$\begin{aligned} c(\ell) = \frac{b(\ell+1)}{b(\ell)} &= 2^{8/n} \left(\frac{\lfloor n/2 \rfloor - \ell - 1}{\lfloor n/2 \rfloor - \ell} \right) \left(\frac{\lceil n/2 \rceil + \ell + 1}{\lceil n/2 \rceil + \ell + 2} \right) \\ &= 2^{8/n} \left(1 - \frac{n+1}{(\lfloor n/2 \rfloor - \ell)(\lceil n/2 \rceil + \ell + 2)} \right) \end{aligned} \tag{33}$$

for all $0 \leq \ell \leq n/2 - 2$. Note that $c(0) > 1$ for sufficiently large n and $c(\ell)$ is decreasing. Therefore, $b(\ell)$ is unimodal, 'starts' with $b(0) > 1$, and 'ends' with $b(\lfloor n/2 \rfloor - 1) < 1$. Therefore we conclude that $a(\ell)$ is indeed unimodal, as required. □

In Section 2.4, we shall need results similar to Fact 9 for the tail of $\mathrm{Bi}(n, p)$ for arbitrary $0 < p < 1$. We shall use the following estimates (see [8, Theorem 2.1]).

Fact 11. *Let X have binomial distribution $\mathrm{Bi}(n, p)$ and set $\lambda = \mathbb{E}(X) = pn$. Then, for any $t \geq 0$, we have*

$$\mathbb{P}\big(X \geq \mathbb{E}(X) + t\big) \leq \exp\left(-\frac{t^2}{2(\lambda + t/3)} \right) \tag{34}$$

and

$$\mathbb{P}\big(X \leq \mathbb{E}(X) - t\big) \leq \exp\left(-\frac{t^2}{2\lambda} \right). \tag{35}$$

2.2. **The well-distribution measure W.** Our aim in this section is to prove Theorem 1. Let $E_N = (e_1, \ldots, e_N) \in \{-1, 1\}^N$ be drawn uniformly at random; we wish to estimate

$$W(E_N) = \max_{a,b,M} \sum_{1 \leq j \leq M} e_{a+jb}, \tag{36}$$

where the maximum is taken over all integers a, b, and M with $1 \leq a + b < a + bM \leq N$.

We start with the lower bound in (9). Observe that Fact 8 tells us that, for any fixed $C > 0$, we have

$$\lim_{N \to \infty} \mathbb{P} \left(\left| \sum_{1 \leq j \leq N} e_j \right| \geq C\sqrt{N} \right) = 2\sqrt{\frac{2}{\pi}} \int_{C/2}^{\infty} e^{-2x^2} \mathrm{d}x > 0. \tag{37}$$

The lower bound for $W(E_N)$ for a typical E_N follows from (37) and the fact that

$$2\sqrt{\frac{2}{\pi}} \int_{C/2}^{\infty} e^{-2x^2} \mathrm{d}x \to 1$$

as $C \to 0$. Indeed, the facts above show that, for any $\varepsilon > 0$, there is $\delta > 0$ so that the lower bound in (9) holds with probability at least $1-\varepsilon$. It remains to prove the upper bound for $W(E_N)$ for typical sequences E_N.

Let K be an integer. The arithmetic progressions of the form $1 \le a+b < a+2b < \cdots < a+Tb \le K$, where $-b+1 \le a \le 0$ and T is the largest integer satisfying $a+Tb \le K$, will be called *complete arithmetic progressions in* $[1,K]\cap\mathbb{Z}$. Our first auxiliary lemma gives an estimate, as a function of K, for the maximal value of the 'discrepancy'

$$S_{a,b}^{(K)} = \left| \sum_{1\le a+jb\le K} e_{a+jb} \right| \tag{38}$$

over complete arithmetic progressions in $[1,K]$, for typical sequences

$$(e_1,\dots,e_K) \in \{-1,1\}^K.$$

Lemma 12. *Let K be a positive integer and C a positive real number. Let $(e_1,\dots,e_K)$ be drawn uniformly at random from $\{-1,1\}^K$. The probability that there are integers a and b with $-b+1 \le a \le 0$ such that*

$$S_{a,b}^{(K)} = \left| \sum_{1\le a+jb\le K} e_{a+jb} \right| > C\sqrt{K} \tag{39}$$

is $O\big(e^{-C^2/4}\big)$, as long as, say, $C \ge 2$.

Proof. For any given integers a and b with $-b+1 \le a \le 0$, the complete arithmetic progression $1 \le a+b < a+2b < \cdots \le K$ in $[1,K]$ has either $\lfloor K/b \rfloor$ or $\lceil K/b \rceil$ terms. Fix a large constant C, and observe that (23) in Fact 9 tells us that, if $b \le K$, then the probability that we have

$$S_{a,b}^{(K)} = \left| \sum_{1\le a+jb\le K} e_{a+jb} \right| > C\sqrt{K}$$

is less than $2e^{-bC^2/4}$.

Now, given b, there are b arithmetic progressions as above, and the possible values for b are $1,\dots,K-1$. Therefore, the probability that there are a and b for which $S_{a,b}^{(K)}$ is larger than $C\sqrt{K}$ is

$$O\Bigg(\sum_{1\le b<K} be^{-bC^2/4} \Bigg) = O\big(e^{-C^2/4}\big), \tag{40}$$

where we used that $C \ge 2$ and hence the terms in the sum in (40) drop geometrically. $\square$

Let us now consider intervals of integers of the form

$$B_{m,r} = (m2^r, (m+1)2^r] \cap \mathbb{Z}, \tag{41}$$

where m and r are non-negative integers. Clearly, $|B_{m,r}| = 2^r$. We refer to the $B_{m,r}$ as *blocks*. Put

$$k = 1 + \lfloor \log_2 N \rfloor, \tag{42}$$

and observe that all blocks $B_{m,r}$ contained in $[1,N]$ have $r<k$.

Lemma 12 above tells us that, given a block $B_{m,r}$, the probability that the discrepancy on some complete arithmetic progression contained in $B_{m,r}$ is greater than $C(k-r)\sqrt{2^r}$ is $O(e^{-C^2(k-r)^2/4})$; that is, the probability that we have

$$\max_{a,b}\left|\sum\{e_{a+jb}\colon a+jb\in B_{m,r}\}\right|>C(k-r)\sqrt{2^r} \tag{43}$$

is $O(e^{-C^2(k-r)^2/4})$. On the other hand, for any r, we have at most $N/2^r\le 2^{k-r}$ blocks $B_{m,r}$ contained in $[1,N]$. Therefore, if C is large enough, the probability that (43) holds for *some* block $B_{m,r}\subset[1,N]$ is

$$O\bigg(\sum_{0\le r<k}2^{k-r}e^{-C^2(k-r)^2/4}\bigg)=O\big(e^{-C^2/4}\big), \tag{44}$$

which tends to 0 as $C\to\infty$. Therefore, to complete our proof, we may suppose that (43) does not hold for any $B_{m,r}$, that is, we may suppose that

(*) *for all m and r with $B_{m,r}\subset[1,N]$, we have*

$$\max_{a,b}\left|\sum\{e_{a+jb}\colon a+jb\in B_{m,r}\}\right|\le C(k-r)\sqrt{2^r}.$$

If we are able to deduce from (*) that $W(E_N)=O(\sqrt{N})$, then we shall be done. Let us now state and prove the following combinatorial observation.

Fact 13. *Given any integers $1\le p<q\le N$, it is possible to write the integer interval $I_{p,q}=[p,q]\cap\mathbb{Z}$ as the disjoint union of blocks $B_{m,r}$ so that, for any r, we use at most two blocks of the form $B_{m,r}$.*

Proof. To see this, first consider the smallest f for which $[p,q]\subset[1,2^f)$ (that is, $f=\lceil\log_2(q+1)\rceil$), and then consider the 'dyadic' ruler in the interval $[1,2^f)$: the level 0 mark in our ruler is 1; the *level* 1 *mark* is 2^{f-1}; the set of *level* 2 *marks* is $\{2^{f-2},2^{f-1}+2^{f-2}\}$; the set of *level* 3 *marks* is

$$\{2^{f-3},2^{f-2}+2^{f-3},2^{f-1}+2^{f-3},2^{f-1}+2^{f-2}+2^{f-3}\};$$

and so on. Let L be the mark of smallest level contained in $[p,q]$. One may then use the binary expansion of $q-L$ to obtain a decomposition of $(L,q]\cap\mathbb{Z}$ into blocks, using, for each r, at most one block of the form $B_{m,r}$. Similarly, one may use the binary expansion of $L-p+1$ to obtain such a decomposition of $[p,L]\cap\mathbb{Z}$. □

We now use (*) and Fact 13 above to prove that $W(E_N)=O(\sqrt{N})$. Given a, b, and M with $1\le a+b<a+Mb\le N$, we write $[a+b,a+Mb]\cap\mathbb{Z}$ as a disjoint union of blocks such that, for each r, we make use of at most two blocks of the form $B_{m,r}$. Let $\mathcal{B}$ be the collection of blocks $B_{m,r}$ that we

have used in our partition of $[a+b, a+Mb]\cap\mathbb{Z}$. We have

$$\begin{aligned}\Bigg|\sum_{1\leq j\leq M} e_{a+jb}\Bigg| &\leq \sum_{B_{m,r}\in\mathcal{B}}\Big|\sum\{e_{a+jb}\colon a+jb\in B_{m,r}\}\Big| \\ &\leq \sum_{0\leq r<k} 2\times C(k-r)\sqrt{2^r} \\ &= 2C2^{k/2}\sum_{1\leq \ell\leq k}\ell 2^{-\ell/2} = O(2^{k/2}) = O(\sqrt{N}).\end{aligned}$$

This completes the proof of Theorem 1.

2.3. **The correlation measure C_k.** In this section we prove Theorems 2 and 3.

2.3.1. *Proof of Theorem 2.* We shall first prove the upper estimate for $C_k(E_N)$ for typical sequences E_N. Indeed, we prove the following lemma.

Lemma 14. *Let $\varepsilon=\varepsilon(N)=(\log\log N)/\log N$. With probability tending to 1 as $N\to\infty$, we have*

$$C_k(E_N) < \sqrt{(2+\varepsilon)N\log\left(N\binom{N}{k}\right)} \tag{45}$$

for every integer k with $2\leq k\leq N/4$.

Proof. We first consider a fixed integer k with $2\leq k\leq N/4$. Fix a sequence $D=(d_1,\dots,d_k)$ with $0\leq d_1<\dots<d_k<N$. Since $E_N=(e_i)_{1\leq i\leq N}$ is drawn uniformly at random from $\{-1,1\}^N$, the sequence

$$(e_{1+d_1}e_{1+d_2}\dots e_{1+d_k}, e_{2+d_1}e_{2+d_2}\dots e_{2+d_k},\dots, e_{N-d_k+d_1}e_{N-d_k+d_2}\dots e_N)$$

is a random element of the uniform space $\{-1,1\}^{N-d_k}$. Recall (6), and let an integer M with $M+d_k\leq N$ be fixed. Clearly, $V(E_N,M,D)$ has the same distribution as $S^{\pm}(M)$. We now use (23) with

$$a=\sqrt{(2+\varepsilon)N\log\left(N\binom{N}{k}\right)},$$

where $\varepsilon=(\log\log N)/\log N$ is as in the statement of our lemma, to deduce that

$$|V(E_N,M,D)| > \sqrt{(2+\varepsilon)N\log\left(N\binom{N}{k}\right)} \tag{46}$$

holds with probability less than

$$\exp\left\{-\frac{1}{2M}(2+\varepsilon)N\log\left(N\binom{N}{k}\right)\right\}\leq\left\{N\binom{N}{k}\right\}^{-(1+\varepsilon/2)}. \tag{47}$$

Summing over all possible choices of M and D, we get an extra factor of $N\binom{N}{k}$, which gives an upper bound of $\left\{N\binom{N}{k}\right\}^{-\varepsilon/2}$ for the probability of the event that (46) should hold for *some* M and D. We shall now sum these bounds over $2 \le k \le N/4$. For estimating this sum, we observe that

$$\frac{1}{1-1/3^\delta} \le \frac{1}{\delta} \tag{48}$$

for all $0 < \delta \le \delta_0$, where $\delta_0 > 0$ is some suitably small absolute constant, and

$$\binom{N}{k-1} \Big/ \binom{N}{k} \le \frac{1}{3} \tag{49}$$

for $2 < k \le N/4$. Using (48) and (49), we see that the probability that (46) holds for some M, D, and k with k in the range $2 \le k \le N/4$ is at most

$$\begin{aligned} \sum_{2\le k\le N/4} \left\{ N\binom{N}{k} \right\}^{-\varepsilon/2} &\le \left\{ N\binom{N}{2} \right\}^{-\varepsilon/2} \sum_{\ell\ge 0} \left(\frac{1}{3}\right)^{\ell\varepsilon/2} \\ &\le \frac{2}{\varepsilon} \left\{ N\binom{N}{2} \right\}^{-\varepsilon/2} = O\left(\frac{1}{(\log\log N)\sqrt{\log N}} \right) = o(1), \end{aligned} \tag{50}$$

and our result follows. □

The proof of Theorem 2 will be complete if we prove the following result.

Lemma 15. *With probability tending to* 1 *as* $N \to \infty$, *we have*

$$C_k(E_N) > \frac{2}{5}\sqrt{N \log \binom{N}{k}} \tag{51}$$

for every integer k *with* $2 \le k \le N/4$.

We start with an auxiliary result.

Fact 16. *Let* $m = \lfloor N/3 \rfloor$. *For every sufficiently large* N, *the following hold.*

(*i*) *If* $2 \le k \le \log m$, *then*

$$\sqrt{N \log \binom{N/3}{k}} \ge 0.99\sqrt{N \log \binom{N}{k}}. \tag{52}$$

(*ii*) *If* $\log m < k \le N/4$, *then*

$$\sqrt{N \log \binom{N/3}{k}} \ge \sqrt{\frac{1-10^{-10}}{3} N \log \binom{N}{k}}. \tag{53}$$

We include the proof of Fact 16 for completeness. The reader may prefer to go directly to the proof of Lemma 15, given below.

Proof of Fact 16. Throughout this proof, we suppose that N is larger than a suitably large absolute constant.

Let us start with the proof of (i). Suppose $2 \le k \le \log m$. Using that $(a/b)^b \le \binom{a}{b} \le (ea/b)^b$, we have

$$\binom{N/3}{k} \ge \left(\frac{N}{3k}\right)^k \ge \left(\frac{1}{3e}\right)^k \left(\frac{eN}{k}\right)^k \ge \left(\frac{1}{3e}\right)^k \binom{N}{k}. \tag{54}$$

Therefore,

$$\begin{aligned}\log \binom{N/3}{k} &\ge \left(1 - \frac{k\log(3e)}{\log\binom{N}{k}}\right) \log \binom{N}{k} \\ &\ge \left(1 - \frac{1+\log 3}{\log(N/k)}\right) \log \binom{N}{k} \ge \left(1 - \frac{3}{\log N}\right) \log \binom{N}{k}. \end{aligned}\tag{55}$$

Inequality (52) follows from (55), and (i) is proved.

Let us now turn to (ii). Suppose $\log m < k \le N/4$. Since $\binom{N/3}{k} \ge \binom{\lfloor N/3 \rfloor}{k} = \binom{m}{k}$, it clearly suffices to prove that

$$\binom{m}{k} \ge \binom{N}{k}^{(1-10^{-10})/3} \tag{56}$$

for all large enough N. If $k < m/2$, then $m-k > m/2 > k$. Moreover,

$$\binom{m}{k} = \binom{m}{m-k} \tag{57}$$

and

$$\binom{N}{m-k}^{(1-10^{-10})/3} \ge \binom{N}{k}^{(1-10^{-10})/3}. \tag{58}$$

Therefore, it suffices to consider k with

$$\frac{1}{2}\left\lfloor \frac{N}{3} \right\rfloor = \frac{1}{2}m \le k \le \frac{N}{4}. \tag{59}$$

Now, the left-hand side of (56) is decreasing in the range (59), while the right-hand side is increasing in that range. Therefore, it suffices to verify (56) at $k = \lfloor N/4 \rfloor$.

To check (56) for $k = \lfloor N/4 \rfloor$, we may use the fact that if $r = r(n)$ is an integer function of n with $\lim_{n\to\infty} r(n)/n = x$, then

$$\binom{n}{r} = 2^{(H(x)+o(1))n}, \tag{60}$$

where $o(1) \to 0$ as $n \to \infty$, and

$$H(x) = -x\log_2 x - (1-x)\log_2(1-x) \tag{61}$$

for all $0 \le x \le 1$ ($H(0) = H(1) = 0$). Of course, this is an immediate consequence of Stirling's formula

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + O\left(\frac{1}{n}\right)\right). \tag{62}$$

We now note that

$$\frac{1}{3}H\left(\frac{3}{4}\right)=\frac{1}{3}H\left(\frac{1}{4}\right)>\frac{1-10^{-10}}{3}H\left(\frac{1}{4}\right), \tag{63}$$

and hence (56) does indeed follow for $k=\lfloor N/4\rfloor$ as long as N is large enough. This completes the proof of Fact 16. □

We now give the proof of the main result in this section.

Proof of Lemma 15. Set $m=\lfloor N/3\rfloor$, and recall that we write $S(m,1/2)$ for a random variable with binomial distribution $\mathrm{Bi}(m,1/2)$. Fix $2\le k\le N/4$. We are interested in the largest integer r for which

$$\mathbb{P}\left(S(m,1/2)\ge\frac{1}{2}(m+r)\right)\ge k^2(\log N)\binom{m+1}{k-1}^{-1} \tag{64}$$

holds. Indeed, we let

$$r(m)=r_k(m)=\max\left\{r\in\mathbb{N}\colon \text{inequality (64) holds}\right\}. \tag{65}$$

We need the following lower bounds for $r(m)$.

Fact 17. *For every sufficiently large N, the following hold.*

(*i*) *If $2\le k\le\log m$, then*

$$r(m)\ge 0.99\sqrt{2m\log\binom{m+1}{k-1}}. \tag{66}$$

(*ii*) *If $\log m<k\le N/4$, then*

$$r(m)\ge(1-10^{-10})\sqrt{\frac{m}{\log 2}\log\binom{m+1}{k-1}}. \tag{67}$$

(*iii*) *If $2\le k\le N/4$, then*

$$r(m)\ge\frac{2}{5}\sqrt{N\log\binom{N}{k}}. \tag{68}$$

Proof. In this proof, we may and shall suppose that N is larger than a suitably large absolute constant for the inequalities below to hold.

We start observing that we have

$$k^2(\log N)\binom{m+1}{k-1}^{-1}=\binom{m+1}{k-1}^{-1+o(1)} \tag{69}$$

uniformly in k in the range $2\le k\le N/4$; that is, for all $\eta>0$ there is N_0 such that if $N\ge N_0$, then, for all $2\le k\le N/4$, we have

$$k^2\log N\le\binom{m+1}{k-1}^{\eta}. \tag{70}$$

To check this assertion, note that if $2\le k\le(\log N)^2$, then the left-hand side of (70) is polylogarithmic in N, whereas $\binom{m+1}{k-1}\ge N/3$. On the other hand, if $(\log N)^2 < k \le N/4$, then the left-hand side of (70) is $O(N^2 \log N)$, whereas $\binom{m+1}{k-1} \ge \big((m+1)/(k-1)\big)^{k-1} \ge (4/3+o(1))^{k-1} \ge (5/4)^{(\log N)^2-1}$, which is superpolynomial in N. Therefore, (69) does hold.

We now turn to the proof of (i). Suppose $2 \le k \le \log m$. Set

$$r = \left\lceil 0.99\sqrt{2m\log\binom{m+1}{k-1}}\right\rceil. \tag{71}$$

We shall show that (64) holds for the value of r given in (71). Let

$$c = \frac{r+1}{2\sqrt{m}} = (1+o(1))0.99\sqrt{\frac{1}{2}\log\binom{m+1}{k-1}}. \tag{72}$$

We now use (21) in Fact 8(ii) to deduce that

$$\begin{aligned}\mathbb{P}\left(S(m,1/2) \ge \frac{1}{2}(m+r)\right) &\ge \mathbb{P}\left(S(m,1/2) \ge \left\lfloor\frac{m}{2}\right\rfloor + c\sqrt{m}\right)\\ = \frac{e^{-2c^2}}{2c\sqrt{2\pi}}(1+o(1)) &\ge \frac{1}{100}\binom{m+1}{k-1}^{-0.99}\left(\log\binom{m+1}{k-1}\right)^{-1/2}, \end{aligned} \tag{73}$$

which is clearly larger than

$$k^2(\log N)\binom{m+1}{k-1}^{-1} \tag{74}$$

for large enough N (see (69)). This completes the proof of (64), and hence (i) follows. We now turn to (ii). Suppose $\log m < k \le N/4$. Set

$$r = \left\lceil (1-10^{-10})\sqrt{\frac{m}{\log 2}\log\binom{m+1}{k-1}}\right\rceil. \tag{75}$$

We shall show that (64) holds for the value of r given in (75). Let

$$\ell = \left\lceil\frac{r+1}{2}\right\rceil = (1+o(1))\frac{1-10^{-10}}{2\sqrt{\log 2}}\sqrt{m\log\binom{m+1}{k-1}}. \tag{76}$$

We now use (25) in Fact 10 to deduce that

$$\begin{aligned}\mathbb{P}\Big(S(m,1/2) \ge \frac{1}{2}(m+r)\Big) &\ge \mathbb{P}\Big(S(m,1/2) \ge \left\lfloor\frac{m}{2}\right\rfloor + \ell\Big)\\ &\ge (1+o(1))2^{-4(\lceil (r+1)/2\rceil+1/2)^2/m}\sqrt{\frac{2}{\pi m}}\\ &\ge (1+o(1))2^{-r^2/m}2^{-(6r+9)/m}\sqrt{\frac{2}{\pi m}}\\ &\ge (1+o(1))\binom{m+1}{k-1}^{-1+10^{-10}} \ge k^2(\log N)\binom{m+1}{k-1}^{-1}\end{aligned}$$

for all large enough N, where, again, we used (69) for the last inequality. This completes the proof of (64), and (ii) follows.

We now turn to (iii). Suppose first that $2 \le k \le \log m$, so that (66) holds. Using that $\binom{m+1}{k-1} \ge \binom{N/3}{k-1} \ge \binom{N/3}{k}^{1/2}$ in this range of k and Fact 16(i), we deduce that

$$\begin{aligned} r(m) &\ge 0.99\sqrt{2\left\lfloor\frac{N}{3}\right\rfloor \log\binom{m+1}{k-1}} \\ &\ge (1+o(1))\frac{0.99}{\sqrt{3}}\sqrt{N\log\binom{N/3}{k}} \ge \frac{2}{5}\sqrt{N\log\binom{N}{k}}. \end{aligned} \tag{77}$$

Suppose now that $\log m < k \le N/4$, so that (67) holds. Using that $\binom{m+1}{k-1} \ge \binom{N/3}{k}^{1-o(1)}$ in this range of k and Fact 16(ii), we deduce that

$$\begin{aligned} r(m) &\ge \frac{1-10^{-10}}{\sqrt{\log 2}}\sqrt{\left\lfloor\frac{N}{3}\right\rfloor \log\binom{m+1}{k-1}} \\ &\ge (1+o(1))\frac{1-10^{-10}}{\sqrt{3\log 2}}\sqrt{N\log\binom{N/3}{k}} \ge \frac{2}{5}\sqrt{N\log\binom{N}{k}}. \end{aligned} \tag{78}$$

Inequality (68) follows from (77) and (78), and (iii) is proved. □

To prove Lemma 15, we shall show that, with probability $\le 2/k^2 \log N$, we have

$$C_k(E_N) \le r_k(m), \tag{79}$$

and then we shall sum over all $2 \le k \le N/4$. This gives that (79) holds for *some* k with $2 \le k \le N/4$ with probability $O(1/\log N) = o(1)$. Therefore, (79) *fails* for *all* $2 \le k \le N/4$ with probability $1 - o(1)$, and Lemma 15 will be proved.

Let $2 \le k \le N/4$ be fixed. Our strategy to estimate the probability that (79) should hold will be as follows. Recall $E_N = (e_1, \dots, e_N)$ and let $u = (e_1, \dots, e_m)$. Let $\mathcal{D}_k$ be the set of $(k-1)$-tuples $D = (d_1, \dots, d_{k-1})$ with $m \le d_1 < \dots < d_{k-1} \le 2m$. For each $D \in \mathcal{D}_k$, let

$$v_D = (e_{1+d_1}e_{1+d_2}\dots e_{1+d_{k-1}}, \dots, e_{m+d_1}e_{m+d_2}\dots e_{m+d_{k-1}}), \tag{80}$$

and let $A(D)$ be the event that

$$|\langle u, v_D\rangle| > r(m) = r_k(m) \tag{81}$$

holds. It suffices to show that some $A(D)$ $(D \in \mathcal{D}_k)$ holds with probability at least $1 - 2/k^2 \log N$. For convenience, let $X = X(E_N)$ be the number of events $A(D)$ $(D \in \mathcal{D}_k)$ that hold for E_N. Let

$$p = p(m) = \mathbb{P}\left(S(m, 1/2) \ge \frac{1}{2}(m + r(m))\right). \tag{82}$$

Because of (65), we have

$$\mathbb{E}(X) = p|\mathcal{D}_k| = p\binom{m+1}{k-1} \ge k^2 \log N. \tag{83}$$

We have now arrived at the key claim: the events $A(D)$ $(D \in \mathcal{D}_k)$ are pairwise independent.

Claim 18. *For all distinct D and $D' \in \mathcal{D}_k$, we have*

$$\mathbb{P}(A(D) \cap A(D')) = p^2. \tag{84}$$

We postpone the proof of Claim 18. To complete the proof of Lemma 15, we make use of the following result, which gives a lower bound for the probability of a union of pairwise independent events. We shall in fact state a stronger lemma, which has as hypothesis that the events should be asymptotically negatively correlated. Versions of this lemma may be found in [5] and [6]. More recently, Petrov used this result to generalize the Borel–Cantelli lemma [13, 14]

Lemma 19. *Let $A_1, \dots, A_M$ be events in a probability space, each with probability at least p. Let $\varepsilon \geq 0$ be given, and suppose that*

$$\mathbb{P}(A_i \cap A_j) \leq p^2(1+\varepsilon) \tag{85}$$

for all $i \neq j$. Then

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{1\leq j\leq M} A_j\Big) &\geq \frac{Mp}{1+(M-1)p(1+\varepsilon)} \\ &= 1 - \frac{1-p+(M-1)p\varepsilon}{1+(M-1)p(1+\varepsilon)} > 1 - \varepsilon - \frac{2}{Mp}.\end{aligned} \tag{86}$$

The proof of Lemma 19 is given below. We conclude the proof of Lemma 15 combining Claim 18 and Lemma 19. It suffices to notice that we have $M = \binom{m+1}{k-1}$ pairwise independent events $A(D)$ (that is, $\varepsilon = 0$), and $pM \geq k^2 \log N$ (see (83)). Inequality (86) then tells us that, with probability greater than $1 - 2/k^2 \log N$, the event $A(D)$ happens for some $D \in \mathcal{D}_k$. We conclude that (79) occurs with probability at most $2/k^2 \log N$, and hence, as observed above, summing over all $2 \leq k \leq N/4$, Lemma 15 follows. □

We shall now prove Claim 18 and Lemma 19.

Proof of Claim 18. Let us consider the events $A(D)$ and $A(D')$ for two distinct D, $D' \in \mathcal{D}_k$. Let $D = (d_1, \dots, d_{k-1})$ and $D' = (d'_1, \dots, d'_{k-1})$, and recall that

$$m \leq d_1 < \dots < d_{k-1} \leq 2m, \qquad m \leq d'_1 < \dots < d'_{k-1} \leq 2m,$$

and that

$$\begin{aligned} v_D &= (e_{1+d_1} e_{1+d_2} \dots e_{1+d_{k-1}}, \dots, e_{m+d_1} e_{m+d_2} \dots e_{m+d_{k-1}}), \\ v_{D'} &= (e_{1+d'_1} e_{1+d'_2} \dots e_{1+d'_{k-1}}, \dots, e_{m+d'_1} e_{m+d'_2} \dots e_{m+d'_{k-1}}). \end{aligned}$$

Let $v_i^D = e_{i+d_1} \dots e_{i+d_{k-1}}$ and $v_i^{D'} = e_{i+d'_1} \dots e_{i+d'_{k-1}}$ $(1 \leq i \leq m)$ be the components of v_D and $v_{D'}$. For convenience, let us write $\omega \in_U \Omega$ if ω is a uniformly distributed random element of Ω. We shall prove more than (84); we shall prove that

$$(e_1v_1^D,\dots,e_mv_m^D,e_1v_1^{D'},\dots,e_mv_m^{D'})\in_U\{-1,1\}^{2m}. \tag{87}$$

Clearly, this proves Claim 18, because $\langle u,v_D\rangle=\sum_{1\le i\le m}e_iv_i^D$ and $\langle u,v_D\rangle=\sum_{1\le i\le m}e_iv_i^{D'}$.

To check (87), we make the following observation: we have

$$(x_1,\dots,x_m,y_1,\dots,y_m)\in_U\{-1,1\}^{2m} \tag{88}$$

if and only if

$$(x_1,\dots,x_m,x_1y_1,\dots,x_my_m)\in_U\{-1,1\}^{2m}. \tag{89}$$

This observation follows immediately from the fact that the map that takes the vector in (88) to the vector in (89) is a bijection (it is in fact an involution).

We apply the observation above to

$$x_i=e_iv_i^D\quad\text{and}\quad y_i=e_iv_i^{D'}\quad(1\le i\le m) \tag{90}$$

to obtain that (87) holds if and only if

$$\begin{aligned}(x_1,\dots,x_m,&x_1y_1,\dots,x_my_m)\\&=(e_1v_1^D,\dots,e_mv_m^D,v_1^Dv_1^{D'},\dots,v_m^Dv_m^{D'})\in_U\{-1,1\}^{2m}.\end{aligned} \tag{91}$$

However, (91) does hold. Indeed, since $D\ne D'$, a moment's thought shows that

$$(v_1^Dv_1^{D'},\dots,v_m^Dv_m^{D'})\in_U\{-1,1\}^m. \tag{92}$$

(To see (92), note that the value of $v_i^Dv_i^{D'}$ $(1\le i\le m)$ is determined by the e_j with

$$j\in i+D\triangle D'=i+(D\setminus D')\cup(D'\setminus D) \tag{93}$$

alone, and the family of sets $i+D\triangle D'$ $(1\le i\le m)$ form a "hypertree". We shall not elaborate on this simple argument.) Fact (92) and the fact that the e_i $(1\le i\le m)$ are not involved in (92) at all imply (91), and hence (87) does hold. This completes the proof of Claim 18. □

Proof of Lemma 19. Let $f=\sum_{1\le j\le M}\chi_j$, where χ_j denotes the characteristic function of the event A_j. We have

$$\int f\,\mathrm{d}\mathbb{P}=\sum_{1\le j\le M}\mathbb{P}(A_j)\ge Mp, \tag{94}$$

and

$$\begin{aligned}\int f^2\,\mathrm{d}\mathbb{P}&=\int\Big(\sum_{1\le j\le M}\chi_j+2\sum_{1\le i<j\le M}\chi_i\chi_j\Big)\,\mathrm{d}\mathbb{P}\\&\le\int f\,\mathrm{d}\mathbb{P}+M(M-1)p^2(1+\varepsilon).\end{aligned} \tag{95}$$

Let $X = \bigcup_{1\le j\le M} A_j$. By the Cauchy–Schwarz inequality, we have

$$\Big(\int_X 1 \,\mathrm{d}\mathbb{P}\Big)\Big(\int_X f^2 \,\mathrm{d}\mathbb{P}\Big) \ge \Big(\int_X f \,\mathrm{d}\mathbb{P}\Big)^2,$$

which, together with (95), gives

$$\mathbb{P}(X)\Big(\int f \,\mathrm{d}\mathbb{P} + M(M-1)p^2(1+\varepsilon)\Big) \ge \Big(\int_X f \,\mathrm{d}\mathbb{P}\Big)^2.$$

Therefore

$$\begin{aligned}\mathbb{P}(X) &\ge \Big(\int_X f \,\mathrm{d}\mathbb{P}\Big)^2 \Big/ \Big(\int f \,\mathrm{d}\mathbb{P} + M(M-1)p^2(1+\varepsilon)\Big)\\ &\ge \frac{M^2p^2}{Mp + M(M-1)p^2(1+\varepsilon)} = \frac{Mp}{1+(M-1)p(1+\varepsilon)}, \qquad (96)\end{aligned}$$

and the lemma follows. □

2.3.2. *Proof of Theorem 3.* Theorem 3 follows simply from some well known results concerning concentration of measure. Let ε and $k = k(N)$ as in the statement of that theorem be given. We now observe that if $E_N = (e_j)_{1\le j\le N}$ and $E'_N = (e'_j)_{1\le j\le N} \in \{-1,1\}^N$ differ in exactly one coordinate, then

$$|V(E_N, M, D) - V(E'_N, M, D)| \le 2k \qquad (97)$$

for any M and D. Indeed, suppose $e'_{j_0} = -e_{j_0}$ and $e_j = e'_j$ for all $j \ne j_0$. Note that e_{j_0} is involved in at most k summands in the definition of $V(E_N, M, D)$ (see (3)), and hence changing its value to $-e_{j_0}$ will change $V(E_N, M, D)$ by at most $2k$. Therefore,

$$|C_k(E_N) - C_k(E'_N)| \le 2k \qquad (98)$$

as well. We may now use, for instance, Lemma 1.2 from [12] to deduce that, for any $t > 0$, we have

$$\mathbb{P}\big(|C_k(E_N) - \mathbb{E}(C_k)| \ge t\big) \le 2\exp\left(-t^2/2k^2N\right). \qquad (99)$$

It now suffices to check that if

$$k \le \log N - \log\log N, \qquad (100)$$

then, taking $t = \varepsilon\mathbb{E}(C_k)$, we have that the right-hand side of (99) is $o(1)$.

Theorem 2 tells us that, if N is large enough, then, say,

$$\mathbb{E}(C_k) \ge \frac{1}{5}\sqrt{N\log\binom{N}{k}}. \qquad (101)$$

Condition (100) on k and (101) easily imply that

$$\frac{t^2}{2k^2N} = \frac{\varepsilon^2\mathbb{E}(C_k)^2}{2k^2N} \to \infty \qquad (102)$$

as $N \to \infty$, and this completes the proof of Theorem 3.

2.4. **The normality measure $\mathcal{N}$.** Recall that the normality measure of $E_N = (e_1, \dots, e_N) \in \{-1,1\}^N$ is defined as

$$\mathcal{N}(E_N) = \max_k \max_X \max_M \left| T(E_N, M, X) - \frac{M}{2^k} \right|, \tag{103}$$

where the maxima are taken over all $k \le \log_2 N$, $X \in \{-1,1\}^k$, and $0 < M \le N + 1 - k$, and $T(E_N, M, X)$ is the number of occurrences of the pattern X in E_N, counting only those occurrences starting with some e_j with $j \le M$ (see (1)). Our aim in this section is to prove Theorem 4.

Proof of Theorem 4. We start with the lower bound in (12). We take $k = 1$ in (103); in fact, we consider $T(E_N, N, (1))$, the number of occurrences of 1 in E_N. Fact 8 tells us that

$$\lim_{N\to\infty} \mathbb{P}\left(\left| T(E_N, N, (1)) - \frac{N}{2} \right| \ge C\sqrt{N} \right) = 2\sqrt{\frac{2}{\pi}} \int_C^\infty e^{-2x^2} \mathrm{d}x > 0. \tag{104}$$

Since

$$2\sqrt{\frac{2}{\pi}} \int_C^\infty e^{-2x^2} \mathrm{d}x \to 1$$

as $C \to 0$, the lower bound in Theorem 4 follows. Indeed, the facts above show that, for any $\varepsilon > 0$, there is $\delta > 0$ so that the lower bound in (12) holds with probability at least $1 - \varepsilon$. It remains to prove the upper bound for $\mathcal{N}(E_N)$ for typical sequences E_N.

The basic lemma that we shall use is Lemma 20 below. Recall that we refer to the sets $B_{m,r}$ as *blocks* (recall (41)). For an integer $k \ge 1$, $X \in \{-1,1\}^k$, and $B_{m,r}$ a block with

$$\max B_{m,r} = (m+1)2^r \le N - k + 1, \tag{105}$$

we shall write $T(E_N, B_{m,r}, X)$ for the number of occurrences of the pattern X in E_N, counting only those occurrences starting in $B_{m,r}$, that is,

$$T(E_N, B_{m,r}, X) = \mathrm{card}\{n \in B_{m,r} \colon E_N^{(n)} = X\}, \tag{106}$$

where

$$E_N^{(n)} = (e_n, e_{n+1}, \dots, e_{n+k-1}), \tag{107}$$

and, as usual, $E_N = (e_1, \dots, e_N)$.

Lemma 20. *Let m and r be fixed non-negative integers with $B_{m,r} \subset [1, N]$. For all $D > 0$, the probability that there is $X \in \{-1,1\}^k$ with $k \le \log_2 N$ satisfying (105) such that*

$$\left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| > D\sqrt{2^r} \tag{108}$$

is at most

$$O\left(e^{-2D^2/9}\right) + 2(\log_2 N)^2 N \exp\left(-\frac{3D}{4\log_2 N} 2^{r/2} \right). \tag{109}$$

We postpone the proof of Lemma 20 and continue with the proof of Theorem 4. Let us apply Lemma 20 with

$$D = C(K-r) \tag{110}$$

for all blocks $B_{m,r} \subset [1,N]$, where

$$K = 1 + \lfloor \log_2 N \rfloor, \tag{111}$$

and C is a large constant. There are at most $N/2^r < 2^{K-r}$ blocks $B_{m,r}$ contained in $[1,N]$ and any such block is such that $r < K$. Let us call a block $B_{m,r}$ *large* if

$$r \geq 4 \log_2 \log_2 N \tag{112}$$

and *small* otherwise. We claim that, with the value of D given in (110) with a large enough constant C, inequality (108) holds for some large block $B_{m,r}$ and some $X \in \{-1,1\}^k$ ($k \leq \log_2 N$) with probability $O\left(e^{-2C^2/9}\right)$.

To prove our claim, for convenience, let $\sum_r'$ indicate sum over all $r < K$ satisfying (112). Adding up (109) over all large blocks $B_{m,r} \subset [1,N]$, we get

$$\begin{aligned}
&\sum\nolimits_r' 2^{K-r} O\left(e^{-2C^2(K-r)^2/9}\right) \\
&\qquad\qquad + \sum\nolimits_r' 2^{K-r+1} (\log_2 N)^2 N \exp\left(-\frac{3C(K-r)}{4\log_2 N}\, 2^{r/2}\right) \\
&\quad \leq \sum_{1\leq \ell \leq K} 2^{\ell} O\left(e^{-2C^2\ell^2/9}\right) \\
&\qquad\qquad\quad + (\log_2 N)^2 N^2 \sum\nolimits_r' \exp\left(-\frac{3C}{4\log_2 N}\, 2^{r/2}\right) \\
&\quad = O\left(e^{-2C^2/9}\right) + O\left(\frac{1}{N}\right) = O\left(e^{-2C^2/9}\right),
\end{aligned} \tag{113}$$

as long as C is a large enough constant and N is large enough with respect to C, proving our claim.

Since the bound in (113) tends to 0 as $C \to \infty$, we may and shall suppose henceforth that

(**) *for all integers m, r, and $k \leq \log_2 N$ with $B_{m,r} \subset [1,N]$ satisfying (105) and (112), and every $X \in \{-1,1\}^k$, we have*

$$\left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| \leq C(K-r)\sqrt{2^r}.$$

Now fix $k \leq \log_2 N$ and fix M with $1 \leq M \leq N-k+1$. Observe that we may write $[1,M]$ as a disjoint union of blocks $B_{m,r}$ ($r \leq \log_2 M < K$) with at most one block of the form $B_{m,r}$ for each r. Indeed, such a decomposition of $[1,M]$ may be read out from the binary expansion of M. Let us write I for the set of the pairs (m,r) for which $B_{m,r}$ occurs in this decomposition of $[1,M]$. Furthermore, let $I = I_+ \cup I_-$ be the partition of I with

$$I_+ = \{(m,r) \in I \colon r \text{ satisfies } (112)\}. \tag{114}$$

For later reference, observe that

$$|I_-| < 4\log_2\log_2 N. \tag{115}$$

Observe also that if $(m,r)\in I_-$, then

$$\left|T(E_N, B_{m,r}, X) - 2^{r-k}\right| \le 2^r < (\log_2 N)^4 \tag{116}$$

for any $X\in\{-1,1\}^k$. Using (**), (115), and (116), we see that, for any $X\in\{-1,1\}^k$, we have

$$\begin{aligned}
\left|T(E_N, M, X) - M2^{-k}\right| &\le \sum_{(m,r)\in I}\left|T(E_N, B_{m,r}, X) - 2^{r-k}\right| \\
&= \sum_{(m,r)\in I_+}\left|T(E_N, B_{m,r}, X) - 2^{r-k}\right| + \sum_{(m,r)\in I_-}\left|T(E_N, B_{m,r}, X) - 2^{r-k}\right| \\
&\le \sum_{0\le r<K} C(K-r)2^{r/2} + |I_-|(\log_2 N)^4 \qquad (117)\\
&\le C2^{K/2}\sum_{1\le \ell\le K}\ell 2^{-\ell/2} + 4(\log_2\log_2 N)(\log_2 N)^4 \\
&= O(2^{K/2}) = O(\sqrt{N}),
\end{aligned}$$

as required. It now remains to prove Lemma 20. □

Proof of Lemma 20. Let non-negative integers m and r with $B_{m,r}\subset[1,N]$ be fixed, and let D be a positive real. Because of the form of the upper bound in (109), we may suppose that $D\ge D_0$ for some conveniently large absolute constant D_0. Indeed, if we prove an upper bound of the form (109) for $D\ge D_0$, then we may simply adjust the absolute constant in the big O term to take into account the case in which $D<D_0$ (making (109) greater than 1, say).

Let us now fix $k\le\log_2 N$ for which (105) holds, and let $X\in\{-1,1\}^k$ be given. We shall first show that the probability that (108) happens for this fixed X is suitably small.

Let us decompose $B_{m,r}$ in arithmetic progressions with difference k: for $0\le s<k$, let

$$I_s = \{j\in B_{m,r} \colon j\equiv s \pmod k\}, \tag{118}$$

so that $B_{m,r} = \bigcup_{0\le s<k} I_s$ is a partition of $B_{m,r}$. Let

$$T(E_N, I_s, X) = \operatorname{card}\{n \colon n\in I_s \text{ and } E_N^{(n)} = X\}, \tag{119}$$

and observe that, clearly, $T(E_N, B_{m,r}, X) = \sum_{0\le s<k} T(E_N, I_s, X)$. In particular, if

$$\left|T(E_N, I_s, X) - |I_s|2^{-k}\right| \le \frac{1}{k}D\sqrt{2^r} \tag{120}$$

for all $0\le s<k$, then (108) fails. Let us fix $0\le s<k$, and let us estimate the probability that (120) should fail.

Clearly, $T(E_N, I_s, X)$ has binomial distribution $\mathrm{Bi}(|I_s|, 2^{-k})$. Therefore, using Fact 11, we see that the probability that (120) should fail is at most

$$2\exp\left(-\frac{t^2}{2(\lambda+t/3)}\right), \tag{121}$$

where $t = (D/k)2^{r/2}$ and $\lambda = |I_s|2^{-k} \le 2^{r-k}$. If $\lambda \ge t/3$, then (121) is at most

$$2\exp\left(-\frac{t^2}{4\lambda}\right) \le 2\exp\left(-\frac{D^2 2^r/k^2}{2^{r-k+2}}\right) = 2e^{-D^2 2^{k-2}/k^2}. \tag{122}$$

If, on the other hand, $\lambda < t/3$, then (121) is at most

$$2\exp\left(-\frac{3}{4}t\right) = 2\exp\left(-\frac{3D}{4k}2^{r/2}\right) \le 2\exp\left(-\frac{3D}{4\log_2 N}2^{r/2}\right). \tag{123}$$

Below, rather crudely, we bound (121) by the sum of (122) and (123). Indeed, we add up (122) and (123) over all choices of s ($0 \le s < k$) and $X \in \{-1,1\}^k$, and get that the probability that (108) should hold for some $X \in \{-1,1\}^k$ ($k \le \log_2 N$) is

$$2k2^k\left(e^{-D^2 2^{k-2}/k^2} + \exp\left(-\frac{3D}{4\log_2 N}2^{r/2}\right)\right). \tag{124}$$

We now sum (124) over all $1 \le k \le \lfloor \log_2 N \rfloor = K-1$. We add up the terms in (124) separately.

We have

$$2\sum_{1\le k<K} k2^k e^{-D^2 2^{k-2}/k^2} = 4e^{-D^2/2} + 2^4 e^{-D^2/4} + 3\times 2^4 e^{-2D^2/9} S, \tag{125}$$

where S is dominated by the sum of a convergent geometric series as long as D is large enough. Therefore, the sum in (125) is

$$O\left(e^{-2D^2/9}\right). \tag{126}$$

Moreover,

$$2\sum_{1\le k<K} k2^k \exp\left(-\frac{3D}{4\log_2 N}2^{r/2}\right) \le 2(\log_2 N)^2 N \exp\left(-\frac{3D}{4\log_2 N}2^{r/2}\right). \tag{127}$$

Lemma 20 follows from (125)–(127). □

3. Small $W(E_N)$ and $\mathcal{N}(E_N)$

3.1. **The probability of having $W(E_N)$ small.** Our aim in this section is to prove Theorem 5, which tells us that, given any $\delta > 0$, the probability of the event $W(E_N) \le \delta\sqrt{N}$ is bounded away from 0.

The proof of Theorem 5 is given in Section 3.1.2, but a key 'positive correlation' result is first proved in Section 3.1.1. To be a little more precise, Claim 26, needed in the proof of Theorem 5, is proved by invoking a general correlation result given in Corollary 22 in Section 3.1.1.

3.1.1. *Preliminary lemmas.* We start with some preliminary results concerning distributions on the integers, that is, vectors $(p_n)_{n\in\mathbb{Z}}$ with $\sum_{n\in\mathbb{Z}} p_n = 1$ and $p_n \geq 0$ for all n. Let us say that the distribution $(p_n)_{n\in\mathbb{Z}}$ is *symmetric* if $p_n = p_{-n}$ for all n and *monotone* if $p_n \geq p_{n+1}$ for all $n \geq 0$.

Let $(p_n) = (p_n)_{n\in\mathbb{Z}}$ and $(q_n) = (q_n)_{n\in\mathbb{Z}}$ be two distributions. We shall say that (p_n) is *more concentrated than* (q_n) if, for all $k \geq 0$, we have

$$\sum_{|n|\leq k} p_n \geq \sum_{|n|\leq k} q_n. \tag{128}$$

We shall write $(p_n) \geq (q_n)$ to indicate that (p_n) is more concentrated than (q_n).

If $P = (p_n)_{n\in\mathbb{Z}}$ and $Q = (q_n)_{n\in\mathbb{Z}}$ are two distributions, we write $P + Q$ for the distribution $(s_n)_{n\in\mathbb{Z}}$ given by the convolution of (p_n) and (q_n), that is, $s_n = \sum_{k\in\mathbb{Z}} p_k q_{n-k}$. Clearly,

(‡) if X is an integer-valued random variable with distribution P and Y is an integer-valued random variable with distribution Q, then $X+Y$ has distribution $P + Q$.

In what follows, we shall have integer-valued random variables that *have parity*, that is, they will take even values or odd values only. For such random variables, we need to change the notion of monotonicity introduced above. We shall say that the distribution (p_n) of a random variable that has parity is monotone if $p_n \geq p_{n+2}$ for all $n \geq 0$. The notions of symmetry and concentration do not need to be changed.

Lemma 21. *(i) Let $P = (p_n)_{n\in\mathbb{Z}}$, $Q = (q_n)_{n\in\mathbb{Z}}$, and $R = (r_n)_{n\in\mathbb{Z}}$ be monotone and symmetric distributions with $Q \geq R$. Then $P + Q$ and $P + R$ are monotone and symmetric and, moreover, $P + Q \geq P + R$.*

(ii) Suppose now that P, Q, and R have parity, are monotone and symmetric, and $Q \geq R$. Then $P + Q$ and $P + R$ have parity, are monotone and symmetric, and $P + Q \geq P + R$.

Proof. We shall consider (i) first. Let $P + Q = (s_n)_{n\in\mathbb{Z}}$. Then, using that P and Q are symmetric, for any n we have $s_{-n} = \sum_{k\in\mathbb{Z}} p_k q_{-n-k} = \sum_{k\in\mathbb{Z}} p_{-k} q_{n+k} = \sum_{k\in\mathbb{Z}} p_k q_{n-k} = s_n$, and hence (s_n) is symmetric. Now let $n \in \mathbb{Z}$ be fixed. We have

$$\begin{aligned} s_n &= p_0 q_n + \sum_{k=1}^{\infty} (p_k q_{n-k} + p_{-k} q_{n+k}) \\ &= p_0 q_n + \sum_{k=1}^{\infty} p_k (q_{n-k} + q_{n+k}) = \sum_{k=0}^{\infty} (p_k - p_{k+1}) \sum_{j=-k}^{k} q_{n+j}. \end{aligned} \tag{129}$$

Since P is monotone, we have $p_k - p_{k+1} \geq 0$ for all $k \geq 0$; since Q is monotone and symmetric, we have $q_{n-k} \geq q_{n+1+k}$ for all $n \geq 0$ and all $k \geq 0$, which is equivalent to $\sum_{j=-k}^{k} q_{n+j} \geq \sum_{j=-k}^{k} q_{n+1+j}$. It follows that, for all $n \geq 0$, we have

$$s_n = \sum_{k=0}^{\infty}(p_k - p_{k+1}) \sum_{j=-k}^{k} q_{n+j} \geq \sum_{k=0}^{\infty}(p_k - p_{k+1}) \sum_{j=-k}^{k} q_{n+1+j} = s_{n+1}, \quad (130)$$

and hence (s_n) is also monotone.

Now recall that $R = (r_n)$ and let $P + R = (\widetilde{s}_n)_{n\in\mathbb{Z}}$. Observe that, for all ℓ and k, the sum $\sum_{n=-\ell}^{\ell}\sum_{j=-k}^{k} q_{n+j}$ may be written as a linear combination of terms of the form $\sum_{i=-m}^{m} q_i$, with non-negative coefficients (this may be checked by induction on $\min\{\ell, k\}$). This observation and the fact that $Q \geq R$ let us deduce from (129) that

$$\sum_{n=-\ell}^{\ell} s_n = \sum_{k=0}^{\infty}(p_k - p_{k+1}) \sum_{n=-\ell}^{\ell}\sum_{j=-k}^{k} q_{n+j}$$

$$\geq \sum_{k=0}^{\infty}(p_k - p_{k+1}) \sum_{n=-\ell}^{\ell}\sum_{j=-k}^{k} r_{n+j} = \sum_{n=-\ell}^{\ell} \widetilde{s}_n, \quad (131)$$

and hence $P + Q \geq P + R$, as required.

We now turn to (ii) (we shall be somewhat sketchy now). Suppose that P, Q, and R have parity, are monotone and symmetric, and $Q \geq R$. The fact that $P + Q$ and $P + R$ have parity follows easily from (‡). The symmetry of the distributions $P + Q$ and $P + R$ follows easily from the symmetry of P, Q, and R, as in (i). It remains to consider monotonicity and concentration.

We need the following variants of (129). Suppose first that P is a distribution supported on the *even* integers. If n has the parity of Q (that is, Q is supported on the integers with the same parity as n), then

$$s_n = \sum_{k=0}^{\infty}(p_{2k} - p_{2k+2}) \sum_{j=-k}^{k} q_{n+2j}. \quad (132)$$

If n does not have the parity of Q_n, then $s_n = 0$. Now suppose that P is a distribution supported on the *odd* integers. If n has the parity of Q, then $s_n = 0$. On the other hand, if n does not have the parity of Q, then

$$s_n = \sum_{k=1}^{\infty}(p_{2k-1} - p_{2k+1}) \sum_{j=1}^{k}(q_{n-(2j-1)} + q_{n+(2j-1)}). \quad (133)$$

The monotonicity of $P + Q$ follows from (132) and (133), as in (i), and, of course, the monotonicity of $P + R$ follows similarly. The fact that $P + Q \geq P + R$ follows from (132) and (133), as in (i) above. □

We shall now discuss a consequence of Lemma 21 that will be needed later. Let

$$A_1, B_1, \ldots, A_k, B_k, C \quad (134)$$

be pairwise disjoint subsets of $[n] = \{1, \dots, n\}$. Let $(e_i)_{i\in[n]} \in \{-1, 1\}^n$ be given, and write $\widehat{A}_j$, $\widehat{B}_j$, and $\widehat{C}$ for the sum of the e_i over the respective sets; e.g.,

$$\widehat{A}_j = \sum_{i\in A_j} e_i \tag{135}$$

for all $1 \le j \le k$. In what follows, we suppose that $(e_i)_{i\in[n]} \in \{-1, 1\}^n$ is chosen uniformly at random. Note that, then, the random variables $\widehat{A}_j$ have parity $|A_j| \bmod 2$, and have distributions that are symmetric and monotone, and similarly for $\widehat{B}_j$ and $\widehat{C}$. We are interested in the following consequence of Lemma 21.

Corollary 22. *For all $r, s_1, \dots, s_k \ge 0$, we have*

$$\mathbb{P}(|\widehat{A}_1 + \dots + \widehat{A}_k + \widehat{C}| \le r \;\big|\; |\widehat{A}_j + \widehat{B}_j| \le s_j \text{ all } 1 \le j \le k) \\ \ge \mathbb{P}(|\widehat{A}_1 + \dots + \widehat{A}_k + \widehat{C}| \le r). \tag{136}$$

Proof. Let $P^{(j)} = (p_n^{(j)})$ be the distribution of $\widehat{A}_j$ conditioned on $|\widehat{A}_j + \widehat{B}_j| \le s_j$. Let $\widetilde{P}^{(j)}$ be the (unconditional) distribution of $\widehat{A}_j$.

Claim 23. *The distributions $P^{(j)}$ and $\widetilde{P}^{(j)}$ have parity $|A_j| \bmod 2$ and are symmetric and monotone. Moreover, $P^{(j)} \ge \widetilde{P}^{(j)}$.*

Let us postpone the proof of Claim 23. To prove (136), it suffices to apply Lemma 21 to conclude that

$$P^{(1)} + \dots + P^{(k)} + Q \ge \widetilde{P}^{(1)} + \dots + \widetilde{P}^{(k)} + Q, \tag{137}$$

where Q is the distribution of $\widehat{C}$. $\square$

We now prove Claim 23.

Proof of Claim 23. Let $A = A_j$ and $B = B_j$. We shall first prove that $P^{(j)} \ge \widetilde{P}^{(j)}$. We start by observing that

$$\mathbb{P}(|\widehat{A} + \widehat{B}| \le k \;\big|\; \widehat{A} = \ell) \qquad (\ell \equiv |A| \bmod 2,\ |\ell| \le |A|) \tag{138}$$

is monotone decreasing in $|\ell|$. Here and in what follows we shall tacitly suppose that ℓ 'has the right parity' (that is, $\ell \equiv |A| \bmod 2$) and $|\ell| \le |A|$. To see that (138) is decreasing in $\ell \equiv |A| \bmod 2$, it suffices to note that the quantity in (138) is equal to $\sum_{n=-\ell-k}^{-\ell+k} \mathbb{P}(\widehat{B} = n)$ and to recall that the distribution of $\widehat{B}$ is monotone.

We now show that

$$\mathbb{P}(|\widehat{A} + \widehat{B}| \le k \mid |\widehat{A}| \le r) \ge \mathbb{P}(|\widehat{A} + \widehat{B}| \le k). \tag{139}$$

To prove (139), we first notice that the left-hand side of (139) is

$$\begin{aligned}\mathbb{P}(|\widehat{A}+\widehat{B}| \le k \mid |\widehat{A}| \le r) &= \mathbb{P}(|\widehat{A}+\widehat{B}| \le k \text{ and } |\widehat{A}| \le r)/\mathbb{P}(|\widehat{A}| \le r)\\ &= \mathbb{P}(|\widehat{A}| \le r)^{-1} \sum_{\ell=-r}^{r} \mathbb{P}(|\widehat{A}+\widehat{B}| \le k \text{ and } \widehat{A} = \ell)\\ &= \sum_{\ell=-r}^{r} \mathbb{P}(|\widehat{A}+\widehat{B}| \le k \mid \widehat{A} = \ell)\frac{\mathbb{P}(\widehat{A}=\ell)}{\mathbb{P}(|\widehat{A}| \le r)}, \end{aligned} \tag{140}$$

and the right-hand side of (139) is

$$\mathbb{P}(|\widehat{A}+\widehat{B}| \le k) = \sum_{\ell \in \mathbb{Z}} \mathbb{P}(|\widehat{A}+\widehat{B}| \le k \mid \widehat{A} = \ell)\mathbb{P}(\widehat{A} = \ell). \tag{141}$$

From (140) and (141) we see that the left-hand side of (139) is a weighted average of the $\mathbb{P}(|\widehat{A}+\widehat{B}| \le k \mid \widehat{A} = \ell)$ with $-r \le \ell \le r$ with weights $\mathbb{P}(\widehat{A} = \ell)/\mathbb{P}(|\widehat{A}| \le r)$, whereas the right-hand side of (139) is a weighted average of the $\mathbb{P}(|\widehat{A}+\widehat{B}| \le k \mid \widehat{A} = \ell)$ *over all* ℓ with weights $\mathbb{P}(|\widehat{A}| = \ell)$.

Now notice that the fact that (138) is monotone decreasing in $|\ell|$ implies that, for all $\ell_2 \in \mathbb{Z} \setminus [-r, r]$ and all $\ell_1 \in [-r, r] \cap \mathbb{Z}$ with $\ell_1 \equiv \ell_2 \equiv |A| \bmod 2$, we have

$$\mathbb{P}(|\widehat{A}+\widehat{B}| \le k \mid |\widehat{A}| = \ell_2) \le \mathbb{P}(|\widehat{A}+\widehat{B}| \le k \mid |\widehat{A}| = \ell_1). \tag{142}$$

Inequality (139) follows from (140), (141), and (142).

Inequality (139) is equivalent to

$$\mathbb{P}(|\widehat{A}+\widehat{B}| \le k \text{ and } |\widehat{A}| \le r)/\mathbb{P}(|\widehat{A}| \le r) \ge \mathbb{P}(|\widehat{A}+\widehat{B}| \le k). \tag{143}$$

It follows from (143) that

$$\begin{aligned}&\mathbb{P}(|\widehat{A}| \le r \mid |\widehat{A}+\widehat{B}| \le k)\\ &\quad = \mathbb{P}(|\widehat{A}+\widehat{B}| \le k \text{ and } |\widehat{A}| \le r)/\mathbb{P}(|\widehat{A}+\widehat{B}| \le k) \ge \mathbb{P}(|\widehat{A}| \le r), \end{aligned} \tag{144}$$

and hence $P^{(j)} \ge \widetilde{P}^{(j)}$ does indeed hold.

It remains to show that the distribution of $\widehat{A}$ conditioned on $|\widehat{A}+\widehat{B}| \le r$ is symmetric and monotone. The fact that this distribution is symmetric is immediate. To see that this distribution is monotone, note that $\mathbb{P}(|\widehat{A}| = \ell+2) \le \mathbb{P}(|\widehat{A}| = \ell)$ for all $\ell \ge 0$ and recall that $\mathbb{P}(|\widehat{A}+\widehat{B}| \le r \mid |\widehat{A}| = \ell+2) \le \mathbb{P}(|\widehat{A}+\widehat{B}| \le r \mid |\widehat{A}| = \ell)$ for all $\ell \ge 0$ (recall that (138) is decreasing in $|\ell| \equiv |A| \bmod 2)$. It follows that

$$\begin{aligned}
&\mathbb{P}(|\widehat{A}| = \ell + 2 \mid |\widehat{A} + \widehat{B}| \le r) \\
&\qquad = \mathbb{P}(|\widehat{A} + \widehat{B}| \le r \mid |\widehat{A}| = \ell + 2) \frac{\mathbb{P}(|\widehat{A}| = \ell + 2)}{\mathbb{P}(|\widehat{A} + \widehat{B}| \le r)} \\
&\qquad \le \mathbb{P}(|\widehat{A} + \widehat{B}| \le r \mid |\widehat{A}| = \ell) \frac{\mathbb{P}(|\widehat{A}| = \ell)}{\mathbb{P}(|\widehat{A} + \widehat{B}| \le r)} \\
&\qquad = \mathbb{P}(|\widehat{A}| = \ell \mid |\widehat{A} + \widehat{B}| \le r), \qquad (145)
\end{aligned}$$

as required. □

3.1.2. *Proof of Theorem 5.* We are now ready to prove Theorem 5. The reader may find it useful to recall the proof of Theorem 1, given in Section 2.2. For convenience, let us fix the following notation: given $E_N = (e_1, \dots, e_N) \in \{-1, 1\}^N$ and $Q \subset [1, N] \cap \mathbb{Z}$, we let

$$U(E_N; Q) = \sum_{q \in Q} e_q. \qquad (146)$$

In what follows, Q will usually be an arithmetic progression.

Let us now start the proof of Theorem 5 proper. Let $\delta > 0$ be given. To define the constant $c(\delta) > 0$ as promised in Theorem 5, we define some auxiliary constants. Let us first fix a constant $C(\delta) \ge 2$ for which we have

$$40 e^{-C(\delta)^2/4} \le \frac{1}{3}. \qquad (147)$$

Let $\ell(\delta)$ be the smallest integer for which we have

$$2C(\delta) \sum_{\ell > \ell(\delta)} \ell 2^{-\ell/2} \le \frac{1}{2}\delta. \qquad (148)$$

We now let $b(\delta)$ be the smallest integer such that

$$2^{\ell(\delta)+2} \sum_{b > b(\delta)} b e^{-\sqrt{b}/4} \le \frac{1}{3} \qquad (149)$$

and

$$\frac{8}{b(\delta)^{1/4}} \le \frac{1}{2}\delta. \qquad (150)$$

Unfortunately, it will still take a little while for us to deliver $c(\delta)$ (see (163)).

We consider the case $N = 2^k$ (the case $2^{k-1} < N \le 2^k$ is similar). As in Section 2.2, for all $0 \le r \le k$ and $0 \le m < 2^{k-r}$, we consider the blocks $B_{m,r} = (m2^r, (m+1)2^r] \cap \mathbb{Z} \subset [1, N]$ defined in (41). In what follows, we shall always have the relation

$$\ell = k - r. \qquad (151)$$

(Later on in this proof we shall focus on "large blocks", for which k and r differ by a constant that depends only on δ, and hence it will be convenient to consider the "gap" $\ell = k - r$.) We also set

$$r_0 = k - \ell(\delta) \quad \text{and} \quad T = T(\delta) = b(\delta)!. \tag{152}$$

Let us now consider the partition

$$[1, N] \cap \mathbb{Z} = \bigcup_{0 \le m < 2^{\ell(\delta)}} B_{m,r_0} \tag{153}$$

of $[1, N] \cap \mathbb{Z}$. For all $0 \le m < 2^{\ell(\delta)}$ and all $0 \le a < T$, let

$$P(m, a) = \{x \in B_{m,r_0} \colon x \equiv a \pmod{T}\}. \tag{154}$$

The progressions $P(m, a)$ then partition B_{m,r_0}:

$$B_{m,r_0} = \bigcup_{0 \le a < T} P(m, a). \tag{155}$$

Also, putting together (153) and (155), we have a partition of $[1, N] \cap \mathbb{Z}$ into the progressions $P(m, a)$:

$$[1, N] \cap \mathbb{Z} = \bigcup_m \bigcup_a P(m, a), \tag{156}$$

where

$$0 \le m < 2^{\ell(\delta)} \quad \text{and} \quad 0 \le a < T = T(\delta). \tag{157}$$

Clearly, the number of $P(m, a)$ is $2^{\ell(\delta)}T(\delta)$. Given m and a as in (157), we define $\mathcal{G}(m, a)$ to be the event

$$\mathcal{G}(m, a) = \left\{ |U(E_N; P(m, a))| \le \frac{\delta}{2^{\ell(\delta)+1}T(\delta)} \sqrt{N} \right\} \tag{158}$$

and let

$$\mathcal{G} = \bigcap_{m,a} \mathcal{G}(m, a), \tag{159}$$

where m and a range over all values in (157). (As the reader may have already guessed, the event in (159) is a "good" event, and hence the notation.) Let

$$\widetilde{\eta}_m(\delta) = \min_{0 \le a < T} \mathbb{P}(\mathcal{G}(m, a)). \tag{160}$$

Note first that the right-hand side of (160) is independent of m and hence we may simply write $\widetilde{\eta}(\delta)$ for $\widetilde{\eta}_m(\delta)$. A simple but crucial observation is that $\widetilde{\eta}(\delta)$ is bounded away from 0 (as $N \to \infty$), as is

$$\eta(\delta) = \widetilde{\eta}(\delta)^{2^{\ell(\delta)}T(\delta)}. \tag{161}$$

Note that

$$\mathbb{P}(\mathcal{G}) \ge \eta(\delta). \tag{162}$$

We finally set

$$c = c(\delta) = \frac{1}{4} \liminf_{N \to \infty} \eta(\delta) > 0. \tag{163}$$

We claim that the choice of $c = c(\delta)$ in (163) will do in Theorem 5. More precisely, we make the following claim.

Claim 24. *We have*

$$\mathbb{P}(W(E_N) \le \delta\sqrt{N}) \ge \frac{1}{3}\eta(\delta). \tag{164}$$

We verify Claim 24 in the remainder of this proof. Let us say that a block $B_{m,r}$ is *large* if $|B_{m,r}| = 2^r \ge 2^{r_0}$. Following the convention in (151) and recalling the definition of r_0 (see (152)), we see that $B_{m,r}$ is large if and only if $\ell = k - r \le \ell(\delta)$. We shall also say that $B_{m,r}$ is *small* if it is not large, that is, if $|B_{m,r}| = 2^r < 2^{r_0}$, or, equivalently, if $\ell = k - r > \ell(\delta)$.

Let us now define two "bad" events $\mathcal{F}_1$ and $\mathcal{F}_2$. To define $\mathcal{F}_1$, we adapt a definition used in Section 2.2; an arithmetic progression contained in $B_{m,r}$ will be called *complete* if it cannot be extended within $B_{m,r}$ keeping the same common difference. We shall say that $\mathcal{F}_1$ occurs if for some small block $B_{m,r}$ and some complete arithmetic progression Q contained in $B_{m,r}$ we have

$$|U(E_N; Q)| > C(\delta)(k - r)\sqrt{2^r}. \tag{165}$$

We shall say that $\mathcal{F}_2$ occurs if for some large block $B_{m,r}$ and some complete arithmetic progression Q contained in $B_{m,r}$ with difference $b > b(\delta)$ we have

$$|U(E_N; Q)| > 2^{-\ell/2} b^{-1/4}\sqrt{N}. \tag{166}$$

With the definitions of $\mathcal{F}_1$ and $\mathcal{F}_2$ at hand, we may state two auxiliary results that will complete the proof of Theorem 5.

Claim 25. *If $E_N \in \mathcal{G} \setminus (\mathcal{F}_1 \cup \mathcal{F}_2)$, then*

$$W(E_N) \le \delta\sqrt{N}. \tag{167}$$

Claim 26. *We have*

$$\mathbb{P}(\mathcal{G} \setminus (\mathcal{F}_1 \cup \mathcal{F}_2)) \ge \frac{1}{3}\eta(\delta). \tag{168}$$

Clearly, Claims 25 and 26 imply Claim 24 and hence our proof is reduced to verifying those two claims.

Proof of Claim 25. Fix $E_N \in \mathcal{G} \setminus (\mathcal{F}_1 \cup \mathcal{F}_2)$ and fix an arithmetic progression $Q \subset [1, N] \cap \mathbb{Z}$; say,

$$Q = \{a + jb \colon 1 \le j \le M\}. \tag{169}$$

We wish to show that $|U(E_N; Q)| \le \delta\sqrt{N}$. Let us decompose $[a + b, a + Mb] \cap \mathbb{Z}$ into blocks $B_{m,r}$ as in Fact 13, that is, using at most two blocks $B_{m,r}$ with the same r for any r:

$$[a + b, a + Mb] \cap \mathbb{Z} = \bigcup_{\mathcal{B}} B_{m,r}, \tag{170}$$

where $\mathcal{B}$ in (170) simply denotes the family of blocks used in our special decomposition. Put

$$\mathcal{B}_{\mathrm{S}} = \{B_{m,r} \in \mathcal{B} \colon B_{m,r} \text{ is small}\} \tag{171}$$

and

$$\mathcal{B}_{\mathrm{L}} = \{B_{m,r} \in \mathcal{B}\colon B_{m,r} \text{ is large}\}. \tag{172}$$

Intersecting the progression Q with the $B_{m,r} \in \mathcal{B}$, we obtain complete arithmetic progressions

$$Q_{m,r} = Q \cap B_{m,r} \tag{173}$$

partitioning Q, that is, such that $Q = \bigcup Q_{m,r}$. We have

$$|U(E_N; Q)| \le \Sigma_1 + \Sigma_2, \tag{174}$$

where

$$\Sigma_1 = \Big| \sum_{B_{m,r} \in \mathcal{B}_{\mathrm{S}}} U(E_N; Q_{m,r}) \Big| \tag{175}$$

and

$$\Sigma_2 = \Big| \sum_{B_{m,r} \in \mathcal{B}_{\mathrm{L}}} U(E_N; Q_{m,r}) \Big|. \tag{176}$$

We claim that

$$\Sigma_i \le \frac{1}{2}\delta\sqrt{N} \tag{177}$$

for both $i = 1$ and 2. Let us verify (177) for $i = 1$ first. Since $E_N \notin \mathcal{F}_1$, we know that (165) fails for all small blocks $B_{m,r}$; in particular, for all $B_{m,r} \in \mathcal{B}_{\mathrm{S}}$, we have

$$|U(E_N; Q_{m,r})| \le C(\delta)(k-r)\sqrt{2^r}. \tag{178}$$

Therefore, we have

$$\begin{aligned} \Sigma_1 &\le \sum_{B_{m,r} \in \mathcal{B}_{\mathrm{S}}} C(\delta)(k-r)\sqrt{2^r} \le 2 \sum_{0 \le r < r_0} C(\delta)(k-r)\sqrt{2^{r-k}}\sqrt{2^k} \\ &\le 2 \sum_{\ell > \ell(\delta)} C(\delta)\ell 2^{-\ell/2}\sqrt{N} \le \frac{1}{2}\delta\sqrt{N}, \end{aligned} \tag{179}$$

where in (179) above we used the inequality in (148) and the convention in (151).

We now prove (177) for $i = 2$. We shall consider two cases.

Case 1. $b > b(\delta)$

Since $E_N \notin \mathcal{F}_2$, for any $B_{m,r} \in \mathcal{B}_{\mathrm{L}}$, we have, in this case,

$$|U(E_N; Q_{m,r})| \le 2^{-\ell/2} b^{-1/4}\sqrt{N} \tag{180}$$

(see (166)). Therefore, recalling (150) and (151), we have

$$\begin{aligned} \Sigma_2 &\le \sum_{B_{m,r} \in \mathcal{B}_{\mathrm{L}}} 2^{-\ell/2} b^{-1/4}\sqrt{N} \le 2 \sum_{r_0 \le r \le k} 2^{-\ell/2} b^{-1/4}\sqrt{N} \\ &= 2 \sum_{0 \le \ell \le \ell(\delta)} 2^{-\ell/2} b^{-1/4}\sqrt{N} \le \frac{8}{b^{1/4}}\sqrt{N} \le \frac{1}{2}\delta\sqrt{N}, \end{aligned} \tag{181}$$

and the proof of (177) for $i = 2$ is complete in this case.

Case 2. $b \le b(\delta)$

Set $Q_{\mathrm{L}} = \bigcup Q_{m,r}$, with the union ranging over all pairs (m,r) with $B_{m,r} \in \mathcal{B}_{\mathrm{L}}$. Since $b \le b(\delta)$, we have that b divides $T = T(\delta) = b(\delta)!$. It therefore follows that we have a partition

$$Q_{\mathrm{L}} = \bigcup_{(m,a)\in I} P(m,a) \tag{182}$$

of Q_{L} into some $P(m,a)$, where I is some index set. Then

$$\Sigma_2 \le \sum_{(m,a)\in I} |U(E_N; P(m,a))|, \tag{183}$$

and, using that $E_N \in \mathcal{G}$ and recalling that there are $2^{\ell(\delta)}T(\delta)$ progressions $P(m,a)$ in total (see (156) and (157)), we have

$$\Sigma_2 \le \sum_{(m,a)\in I} \frac{\delta}{2^{\ell(\delta)+1}T(\delta)}\sqrt{N} \le \frac{1}{2}\delta\sqrt{N}, \tag{184}$$

which verifies (177) for $i = 2$ in this case.

Claim 25 follows from (174) and (177) for $i = 1$ and 2. □

Proof of Claim 26. We aim at bounding $\mathbb{P}(\mathcal{G} \setminus (\mathcal{F}_1 \cup \mathcal{F}_2))$ from below. Let us first bound $\mathbb{P}(\mathcal{F}_1)$ and $\mathbb{P}(\mathcal{F}_2)$ from above. We first consider $\mathcal{F}_1$. Calculations similar to the ones leading to (*) in the proof of Lemma 12 show that

$$\mathbb{P}(\mathcal{F}_1) \le 40e^{-C(\delta)^2/4} \le \frac{1}{3}. \tag{185}$$

Now consider $\mathcal{F}_2$. Fix $B_{m,r}$ and Q as in the definition of $\mathcal{F}_2$. Using Fact 9, we obtain that

$$\begin{aligned}\mathbb{P}(|U(E_N;Q)| > 2^{-\ell/2}b^{-1/4}\sqrt{N}) &< 2\exp\left(-2^{-\ell}b^{-1/2}N/2|Q|\right)\\ &\le 2\exp\left(-2^{-\ell}b^{-1/2}N/4(2^r/b)\right) = 2e^{-\sqrt{b}/4}.\end{aligned} \tag{186}$$

Summing the bound in (186) over all choices of $B_{m,r}$ and Q and using (149), we have

$$\mathbb{P}(\mathcal{F}_2) \le 2\sum_{0\le\ell\le\ell(\delta)} 2^{\ell} \sum_{b(\delta)<b\le 2^r} be^{-\sqrt{b}/4} \le 2^{\ell(\delta)+2}\sum_{b>b(\delta)} be^{-\sqrt{b}/4} \le \frac{1}{3}. \tag{187}$$

Unfortunately, the bounds in (185) and (187) are not quite enough to complete the proof, as $\mathcal{G}$ is an event of rather small probability—indeed, those bounds do not even guarantee that $\mathcal{G} \setminus (\mathcal{F}_1 \cup \mathcal{F}_2) \neq \emptyset$. We thus refine our argument, and give upper estimates for the conditional probabilities $\mathbb{P}(\mathcal{F}_1 \mid \mathcal{G})$ and $\mathbb{P}(\mathcal{F}_2 \mid \mathcal{G})$. To obtain those estimates, we make use of Corollary 22.

Note that, to obtain the estimates (185) and (187), we used the so called union bound: $\mathbb{P}(A) \le \sum_{\lambda\in\Lambda}\mathbb{P}(A_\lambda)$ if $A \subset \bigcup_{\lambda\in\Lambda} A_\lambda$. Of course, we may also consider this bound conditioning on $\mathcal{G}$: $\mathbb{P}(A \mid \mathcal{G}) \le \sum_{\lambda\in\Lambda}\mathbb{P}(A_\lambda \mid \mathcal{G})$. If we repeat the arguments given above for (185) and (187), but conditioning on $\mathcal{G}$, we have to bound probabilities of the form $\mathbb{P}(|U(E_N;Q)| > t \mid \mathcal{G})$

(for instance, see (186) and (187)). However, a little meditation reveals that Corollary 22 implies that, for any t, we have

$$\mathbb{P}(|U(E_N;Q)|>t\mid\mathcal{G})\le\mathbb{P}(|U(E_N;Q)|>t). \tag{188}$$

In particular, the inequalities used to obtain the bounds for $\mathbb{P}(\mathcal{F}_1)$ and $\mathbb{P}(\mathcal{F}_2)$ above would equally hold for $\mathbb{P}(\mathcal{F}_1\mid\mathcal{G})$ and $\mathbb{P}(\mathcal{F}_2\mid\mathcal{G})$. Therefore, we have

$$\mathbb{P}(\mathcal{F}_1\mid\mathcal{G})\le\frac{1}{3}\quad\text{and}\quad\mathbb{P}(\mathcal{F}_2\mid\mathcal{G})\le\frac{1}{3}, \tag{189}$$

and hence

$$\mathbb{P}(\mathcal{F}_1\cup\mathcal{F}_2\mid\mathcal{G})\le\frac{2}{3}. \tag{190}$$

Thus

$$\frac{\mathbb{P}(\mathcal{G}\setminus(\mathcal{F}_1\cup\mathcal{F}_2))}{\mathbb{P}(\mathcal{G})}=\mathbb{P}(\mathcal{G}\setminus(\mathcal{F}_1\cup\mathcal{F}_2)\mid\mathcal{G})\ge\frac{1}{3}, \tag{191}$$

whence, by (162),

$$\mathbb{P}(\mathcal{G}\setminus(\mathcal{F}_1\cup\mathcal{F}_2))\ge\frac{1}{3}\mathbb{P}(\mathcal{G})\ge\frac{1}{3}\eta(\delta), \tag{192}$$

as required. □

As observed before, Claims 25 and 26 imply Claim 24, and the proof of Theorem 5 is complete.

3.2. **The probability of having $\mathcal{N}(E_N)$ small.** Our aim in this section is to prove Theorem 6, which gives a lower estimate for the probability that $\mathcal{N}(E_N)$ should be small. Our argument here will be based on the arguments used in the proof of Theorem 4, but an additional idea will be crucial. In what follows, we shall be sketchy in parts.

Let us start with some variants of some results given in Section 2.4. The following is a simple variant of Lemma 20.

Lemma 27. *Let m and r be fixed non-negative integers with $B_{m,r}\subset[1,N]$. For all $D>0$, the probability that there is $X\in\{-1,1\}^k$ with $k\le\log_2 N$ and $k\le r$ satisfying (105) such that*

$$\left|T(E_N,B_{m,r},X)-2^{r-k}\right|>D2^{-k/4}\sqrt{2^r} \tag{193}$$

is at most

$$O\left(e^{-D^2/18}\right)+2(\log_2 N)^2N\exp\left(-\frac{3D}{4\log_2 N}\,2^{r/4}\right). \tag{194}$$

Remark 28. *If $k>r$, then $2^{r-k}<1$, and hence, using the above estimate for $k=r$, we get that the probability that*

$$\left|T(E_N,B_{m,r},X)-2^{r-k}\right|>D2^{r/4}+1 \tag{195}$$

for some $X\in\{-1,1\}^k$ is at most as given in (194).

Again, for simplicity, we shall assume that $N = 2^K$. In this proof, a block $B_{m,r}$ is *large* if

$$r \geq 8 \log_2 \log_2 N \tag{196}$$

and is *small* otherwise. As in the proof of Theorem 4, we may and shall assume that

(***) *for all integers m, r, and $k \leq \log_2 N$ with $B_{m,r} \subset [1,N]$ and $r \geq 8 \log_2 \log_2 N$ satisfying (105) and every $X \in \{-1,1\}^k$, we have the following: if $k \leq r$, then*

$$\left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| \leq C 2^{-k/4} (K-r) \sqrt{2^r},$$

and if $k > r$, then

$$|T(E_N, B_{m,r}, X) - 2^{r-k}| \leq C(K-r) 2^{r/4} + 1.$$

To see that (***) does hold with probability approaching 1 as $C \to \infty$, we follow the argument given in the proof of Theorem 4 (see (113)), but using Lemma 27 and Remark 28 instead of Lemma 20. In the proof of Theorem 4, the relevant probability was shown to be $O\left(e^{-2C^2/9}\right)$; repeating the same calculations with the bounds in Lemma 27 and Remark 28, one obtains $O\left(e^{-C^2/18}\right)$.

We now point out another fact that may be read out from calculations we have already seen. Making use of (***) instead of (**), the calculations in (117) show that, for any $\ell \geq 1$, the contribution of blocks of cardinality $\leq 2^{K-\ell}$ to the discrepancy $|T(E_N, M, X) - M2^{-k}|$ is

$$O\left(C 2^{K/2} \ell 2^{-\ell/2}\right) + O\left((\log_2 \log_2 N)(\log_2 N)^8\right). \tag{197}$$

This observation shows that

(‡) *the contribution of the blocks $B_{m,r}$ with $r \leq K - \ell_0$ to the estimate in (117) is, say, $\leq (\delta/2)\sqrt{N}$ if $\ell_0 = \ell_0(\delta)$ is some large enough constant that depends only on δ.*

Having cleared some of the pre-requisites, we start the proof of Theorem 6.

Proof of Theorem 6. Let $\delta > 0$ be given. We let $k_0 = k_0(\delta)$ be a suitably large integer constant that depends only on δ, for the inequalities below to hold. We wish to show that (14) holds for some suitably small positive constant $c(\delta)$.

Fix $k \leq \log_2 N$ and M with $1 \leq M \leq N - k + 1$. As before, write $[1, M]$ as a disjoint union of blocks $B_{m,r}$ ($r \leq \log_2 M \leq K$) with at most one block of the form $B_{m,r}$ for each r. Let us write I for the set of the pairs (m, r) for which $B_{m,r}$ occurs in this decomposition of $[1, M]$. Furthermore, let $I = I_+ \cup I_-$ be the partition of I with

$$I_+ = \{(m,r) \in I \colon r \text{ satisfies } (196)\}. \tag{198}$$

For later reference, observe that

$$|I_-| < 8 \log_2 \log_2 N. \tag{199}$$

Observe also that if $(m, r) \in I_-$, then

$$\left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| \le 2^r < (\log_2 N)^8 \tag{200}$$

for any $X \in \{-1, 1\}^k$.

Let us now fix $X \in \{-1, 1\}^k$. We wish to estimate $|T(E_N, M, X) - M2^{-k}|$. As in (117), we start with

$$\begin{aligned} \left| T(E_N, M, X) - M2^{-k} \right| &\le \sum_{(m,r) \in I} \left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| \\ = \sum_{(m,r) \in I_+} &\left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| + \sum_{(m,r) \in I_-} \left| T(E_N, B_{m,r}, X) - 2^{r-k} \right|. \end{aligned} \tag{201}$$

Because of (199) and (200), the last sum in (201) is $o(\sqrt{N})$, and hence we may ignore it. We thus focus on the sum over the pairs $(m, r) \in I_+$, that is, we consider the contribution of the large blocks $B_{m,r}$ (i.e., blocks with $r \ge 8 \log_2 \log_2 N$).

Case 1. $k \ge k_0 = k_0(\delta)$

We shall be able to dispose of this case by invoking some observations that we have already discussed. Let us start by observing that

$$\begin{aligned} \sum_{(m,r) \in I_+} &\left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| \\ &= \sum\nolimits_1 \left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| \\ &\qquad + \sum\nolimits_2 \left| T(E_N, B_{m,r}, X) - 2^{r-k} \right|, \end{aligned} \tag{202}$$

where $\sum_1$ indicates sum over all pairs $(m, r) \in I_+$ with $r < k$ and $\sum_2$ indicates sum over all pairs $(m, r) \in I_+$ with $r \ge k$. Making use of (***), one may check that calculations very similar to the ones in (117) show that the last but one sum in (202) is $O(C2^{K/4}) = o(\sqrt{N})$ and the last sum is $\le (\delta/2)\sqrt{N}$, as long as $k_0 = k_0(\delta)$ is large enough. The proof is therefore complete in this case.

Case 2. $k < k_0 = k_0(\delta)$

This case will require considerable more work. Let

$$r_0 = K - k_0 \tag{203}$$

and observe that

$$\sum_{(m,r)\in I_+} \left|T(E_N, B_{m,r}, X) - 2^{r-k}\right| = \sum\nolimits_1 \left|T(E_N, B_{m,r}, X) - 2^{r-k}\right| + \sum\nolimits_2 \left|T(E_N, B_{m,r}, X) - 2^{r-k}\right|, \quad (204)$$

where $\sum_1$ indicates sum over all pairs $(m,r) \in I_+$ with $r < r_0$ and $\sum_2$ indicates sum over all pairs $(m,r) \in I_+$ with $r \geq r_0$.

The estimate of the first sum on the right-hand side of (204) is based on (‡): since in that sum we are considering blocks $B_{m,r}$ with $r < r_0$, according to (‡), that sum is $\leq (\delta/2)\sqrt{N}$, as long as k_0 is large enough.

In view of the above discussion, we are left with estimating the last sum in (204). The key fact that we shall use is the following, which deals with blocks B_{m,r_0} and sequences X of length k_0.

Claim 29. *There is a constant $c'(\delta) > 0$ for which the following holds. Let $\mathcal{E}$ be the event that, for all m with $B_{m,r_0} \subset [1,N]$ and $\max B_{m,r_0} = (m+1)2^{r_0} \leq N - k_0 + 1$ and every $X \in \{-1,1\}^{k_0}$, we have*

$$\left|T(E_N, B_{m,r_0}, X) - 2^{r_0-k_0}\right| = O\left(\delta 2^{-5k_0/2}\sqrt{N}\right). \quad (205)$$

Then $\mathbb{P}(\mathcal{E}) \geq c'(\delta)$.

We leave the proof of Claim 29 for later and complete the proof of Theorem 6. Suppose E_N is such that (205) does hold for all B_{m,r_0} and all X of length k_0 as specified in the definition of the event $\mathcal{E}$. Observe that, then, for all $Y \in \{-1,1\}^k$ with $k \leq k_0$, we have

$$\left|T(E_N, B_{m,r_0}, Y) - 2^{r_0-k}\right| = O\left(\delta 2^{-3k_0/2}\sqrt{N}\right). \quad (206)$$

Indeed, given a sequence Y of length $k \leq k_0$, it suffices to consider all $2^{k_0-k} \leq 2^{k_0}$ extensions of Y to sequences X of length k_0, with Y a prefix of X, and then apply (205). We therefore assume that (206) does hold for all Y of length $k \leq k_0$.

Recall that we have to estimate the last sum in (204), where X is a fixed sequence of length $k < k_0$. Note that that sum is at most

$$\sum_{m\in J} \left|T(E_N, B_{m,r_0}, X) - 2^{r_0-k}\right|, \quad (207)$$

where the sum ranges over some set $J \subset [0, 2^{k_0})$. Clearly, from (206) applied to $Y = X$, we see that the sum in (207) is $O(\delta 2^{-k_0/2}\sqrt{N}) \leq (\delta/3)\sqrt{N}$, as long as $k_0 = k_0(\delta)$ is large enough, and this completes the proof of this case.

To complete the proof of Theorem 6, we need to prove Claim 29. □

Proof of Claim 29. We shall be somewhat brief in parts of this proof. Let $\delta > 0$ be given. We may suppose that δ is sufficiently small. In this proof, for simplicity, we write k for $k_0 = k_0(\delta)$, which is supposed to be a sufficiently large constant for the inequalities below to hold. Recall we assume that $N = 2^K$. We now let $M = 2^{K-7k}$ and consider the distribution of

$$\mathbf{s}(E_M) = \left(\frac{T(E_M, M, X) - M2^{-k}}{\sqrt{M}}\right)_{X\in\{-1,1\}^k} \in \mathbb{R}^{2^k}, \tag{208}$$

with E_M chosen from $\{-1,1\}^M$ uniformly at random. (We remark that it would be more natural to have

$$\frac{T(E_M, M-k+1, X) - (M-k+1)2^{-k}}{\sqrt{M-k+1}} \tag{209}$$

as the entries of $\mathbf{s}(E_M)$, but, for simplicity, we shall use the definition given in (208), and we shall ignore the additive error term, of order $O(k/\sqrt{M})$, introduced by using M instead of $M-k+1$ in (208).)

Let us say that $E_M \in \{-1,1\}^M$ is *good* if

$$|T(E_M, B_{m,r}, X) - 2^{r-\ell}| \le k(K-7k-r)2^{-\ell/4}\sqrt{2^r} \tag{210}$$

for all m and r with $B_{m,r} \subset [M-\ell+1]$ and $r \ge 8\log\log M$ and all $X \in \{-1,1\}^\ell$ with $\ell \le r$ and $\ell \le \log_2 M$. Let $\mathcal{G} \subset \{-1,1\}^M$ be the set of good sequences, and let us write $\mathcal{G}^{\mathrm{c}}$ for the complement of $\mathcal{G}$. We now state the following claim.

Claim 30. *We have*

$$\mathbb{P}(\mathcal{G}^{\mathrm{c}}) = O(e^{-k^2/18}). \tag{211}$$

Claim 30 above follows from calculations similar to the ones that let us assume (***) (see also the proof of Theorem 4).

In what follows, we wish to ignore the sequences E_M in $\mathcal{G}^{\mathrm{c}}$. The next claim tells us that the sequences in $\mathcal{G}^{\mathrm{c}}$ do not influence the expectation of $\mathbf{s}(E_M)$ too much.

Claim 31. *For any $X \in \{-1,1\}^k$,*

$$\mathbb{E}(|T(E_M, M, X) - M2^{-k}|\chi_{\mathcal{G}^{\mathrm{c}}}) = O(\sqrt{M}e^{-k^2/18}), \tag{212}$$

where $\chi_{\mathcal{G}^{\mathrm{c}}}$ is the characteristic function of $\mathcal{G}^{\mathrm{c}}$.

Proof. Fix $X \in \{-1,1\}^k$ and let $Y = Y(E_M) = |T(E_M, M, X) - M2^{-k}|$. Note first that Fact 11 tells us that

$$\mathbb{P}(Y \ge M2^{-k}) \le 2e^{-M/2^{k+2}}. \tag{213}$$

Since we always have $Y \le M$, we conclude from (213) that, setting $A = \{E_M : Y(E_M) \ge M/2^k\}$. we have

$$\mathbb{E}(Y\chi_A) = O(Me^{-M/2^{k+2}}) = O(\sqrt{M}e^{-k^2/18}). \tag{214}$$

We shall make use of (214) in a short while. Let $V(E_M)$ be given by

$$V(E_M) = \begin{cases} Y(E_M) & \text{if } Y > k\sqrt{M/2^k} \\ 0 & \text{otherwise.} \end{cases} \tag{215}$$

Then we have

$$Y\chi_{\mathcal{G}^c} \le k\sqrt{\frac{M}{2^k}}\chi_{\mathcal{G}^c} + V, \tag{216}$$

whence, recalling (211), we deduce that

$$\mathbb{E}(Y\chi_{\mathcal{G}^c}) \le k\sqrt{\frac{M}{2^k}}\mathbb{P}(\chi_{\mathcal{G}^c}) + \mathbb{E}(V) = k\sqrt{\frac{M}{2^k}}O(e^{-k^2/18}) + \mathbb{E}(V). \tag{217}$$

Now let $R = k^{-1}\sqrt{M/2^k}$. Using (214), we see that

$$\mathbb{E}(V) \le O(\sqrt{M}e^{-k^2/18}) + \sum_{1\le r\le R} \mathbb{P}\left(Y \ge rk\sqrt{\frac{M}{2^k}}\right)(r+1)k\sqrt{\frac{M}{2^k}}. \tag{218}$$

From Fact 11, we obtain

$$\mathbb{P}\left(Y \ge rk\sqrt{\frac{M}{2^k}}\right) \le 2e^{-r^2k^2/4}. \tag{219}$$

Putting together (218) and (219), we get

$$\begin{aligned} \mathbb{E}(V) &\le O(\sqrt{M}e^{-k^2/18}) + 2k\sqrt{\frac{M}{2^k}}\sum_{1\le r\le R}(r+1)e^{-r^2k^2/4} \\ &= O(\sqrt{M}e^{-k^2/18}) + O\left(k\sqrt{\frac{M}{2^k}}e^{-k^2/4}\right) \\ &= O(\sqrt{M}e^{-k^2/18}), \end{aligned} \tag{220}$$

as required. ☐

Note that the vector $\mathbf{s}(E_M)$ has entries with expectation 0. Combined with Claim 31, this implies that, for any $X \in \{-1,1\}^k$, the expected value of

$$\frac{T(E_M, M, X) - M2^{-k}}{\sqrt{M}} \tag{221}$$

conditioned on $E_M \in \mathcal{G}$ is $O(e^{-k^2/18}) = o_{\delta\to 0}(\delta/64^k)$. In this proof, we write $o_{\delta\to 0}(x)$ for any term y such that $y/x \to 0$ as $\delta \to 0$.

For $E_M \in \mathcal{G}$, we have

$$\begin{aligned} \mathbf{s}(E_M) = \left(\frac{T(E_M, M, X) - M2^{-k}}{\sqrt{M}}\right)_{X\in\{-1,1\}^k} \\ \in [-k2^{-k/4}, k2^{-k/4})^{2^k} \subset [-1,1)^{2^k}. \end{aligned} \tag{222}$$

Let us partition $[-1,1)^{2^k}$ in $b = (2\times 64^k/\delta)^{2^k}$ blocks of side length $\delta/64^k$ (that is, blocks of the form $\prod_{1\le j\le 2^k}[x_j, x_j + \delta/64^k)$; we may assume without loss of generality that $64^k/\delta$ is an integer). Let these blocks be C_i ($1 \le i \le b$). For each i, let $P_i \in C_i$ be the average of the points $\mathbf{s}(E_M)$ that belong to C_i and are such that $E_M \in \mathcal{G}$; if $\mathbf{s}^{-1}(C_i) \cap \mathcal{G} = \emptyset$, let P_i be the centre of C_i. If

$$q_i = \mathbb{P}(P \in C_i \bigm| P \in \mathcal{G}), \tag{223}$$

then $\mathbf{E} = \sum_i q_i P_i$ is the expectation of $\mathbf{s}(E_M)$ conditional on $E_M \in \mathcal{G}$. Recall that all entries of $\mathbf{E}$ are $o_{\delta\to 0}(\delta/64^k)$.

Let

$$J = \left\{ 1 \le j \le b\colon q_j \ge \frac{\delta}{k64^k}\left(\frac{\delta}{64^k}\right)^{2^k} \right\}. \tag{224}$$

Then

$$\sum_{j\notin J} q_j P_j \tag{225}$$

has all entries that are $O(\delta/k64^k) = o_{\delta\to 0}(\delta/64^k)$. Let

$$\widetilde{\mathbf{E}} = \sum_{j\in J} \widetilde{q}_j P_j, \tag{226}$$

where $\widetilde{q}_j = q_j/\sum_{i\in J} q_i$. Then all the entries of $\widetilde{\mathbf{E}}$ are $o_{\delta\to 0}(\delta/64^k)$, because all the entries of $\mathbf{E}$ are $o_{\delta\to 0}(\delta/64^k)$ and

$$\mathbf{E} = \Big(\sum_{i\in J} q_i\Big)\widetilde{\mathbf{E}} + \sum_{j\notin J} q_j P_j. \tag{227}$$

Clearly, $\widetilde{\mathbf{E}}$ may be written as a convex combination of the $P_j \in \mathbb{R}^{2^k}$ ($j \in J$). Therefore, there are points $Q_j = P_{i_j}$ ($0 \le j \le 2^k$ and $i_j \in J$ for all j) and real numbers $t_j \ge 0$ with $\sum_{j=0}^{2^k} t_j = 1$ such that $\widetilde{\mathbf{E}} = \sum_{j=0}^{2^k} t_j P_{i_j}$.

We next approximate the t_j by rationals $m_j/64^k$ in such a way that

$$\left|\frac{m_j}{64^k} - t_j\right| \le \frac{1}{64^k} \qquad \text{and} \qquad \sum_{j=0}^{2^k} m_j = 64^k. \tag{228}$$

(We can take $m_j = \lfloor 64^k t_j \rfloor$ if $\sum_{i<j} m_i \ge 64^k \sum_{i<j} t_i$) and $m_j = \lceil 64^k t_j \rceil$ otherwise.)

We now partition $[N]$ into the 2^k blocks $B_{m,r_0} = B_{m,K-k}$ ($0 \le m < 2^k$) and partition each of these blocks into 64^k segments of length $M = 2^{K-7k}$. We consider $E_N \in \{-1,1\}^N$ satisfying certain conditions. We first require that E_N should be such that, in each of these segments of length M, we have a good sequence (a sequence in $\mathcal{G}$). Moreover, we require that, in each of the 2^k blocks B_{m,r_0} of length $2^{r_0} = 2^{K-k}$, the sequence E_N should be composed by 64^k consecutive segments of length M in such a way that in the first m_0 segments we have $\mathbf{s}(E_M) \in C_{i_0}$, in the next m_1 segments we have $\mathbf{s}(E_M) \in C_{i_1}$, etc. The probability that all these requirements are met is at least

$$c(\delta) = \left(\frac{1}{2k} \left(\frac{\delta}{64^k} \right)^{2^k+1} \right)^{64^k \times 2^k}. \tag{229}$$

In each of the 2^k blocks $B_{m,r_0} = B_{m,K-k}$, we have

$$\begin{aligned} T(E_N, B_{m,r_0}, X) - |B_{m,r_0}|2^{-k} &= \sum_{i=0}^{2^k} m_i \left(T(E_M^{(i)}, M, X^{(i)}) - M2^{-k} \right) \\ &= \sum_{j=0}^{2^k} m_j \left(Q_j + O\left(\frac{\delta}{64^k} \right) \right) \sqrt{M} \\ &= O\left(\frac{\delta}{64^k} \sum_{j=0}^{2^k} m_j \sqrt{M} \right) + \sum_{j=0}^{2^k} m_j Q_j \sqrt{M} \\ &= O\left(\delta\sqrt{M} \right) + \sum_{j=0}^{2^k} m_j Q_j \sqrt{M} \\ &= O\left(\delta\sqrt{M} \right) + \sum_{j=0}^{2^k} \left(64^k t_j + O(1) \right) Q_j \sqrt{M} \qquad (230) \\ &= O\left(\delta\sqrt{M} \right) + O(2^k\sqrt{M}) + \sum_{j=0}^{2^k} 64^k t_j Q_j \sqrt{M} \\ &= O\left(2^k\sqrt{M} \right) + O\left(64^k \frac{\delta}{64^k} \sqrt{M} \right) \\ &= O\left(2^k\sqrt{M} \right) = O\left(2^k 2^{-7k/2}\sqrt{N} \right) = O\left(2^{-5k/2}\sqrt{N} \right), \end{aligned}$$

as required. □

Remark 32. *We may prove upper bounds for $c(\delta)$ in Theorems 5 and 6 as follows. Partition $[N]$ into k intervals, each of length N/k. In each such interval, the probability that the sum exceeds $(1/2)\sqrt{N/k}$ in absolute value is $\geq 1/2$, whence it follows that the probability that this event occurs in none of these intervals is $\leq 1/2^k$. Taking $\delta = 1/2\sqrt{k}$, we obtain $c(\delta) \leq (1/2)^{1/4\delta^2}$.*

4. Concluding remarks

The upper bounds in Theorems 1 and 4 are best possible in the following sense. Let us consider $W(E_N)$. We claim that, for any $C > 0$, there is $\varepsilon_0 > 0$ such that

$$\mathbb{P}\Big(W(E_N) < C\sqrt{N} \Big) \leq 1 - \varepsilon_0 \tag{231}$$

for all large enough N. Therefore, the fact that the constant $1/\delta$ in the upper bound in Theorem 1 depends on ε cannot be avoided.

Inequality (231) follows simply from Fact 8. Indeed, that result tells us that, for any fixed $C > 0$, we have

$$\lim_{N\to\infty} \mathbb{P}\left(\left|\sum_{1\le j\le N} e_j\right| \ge C\sqrt{N}\right) = 2\sqrt{\frac{2}{\pi}}\int_{C/2}^{\infty} e^{-2x^2}\mathrm{d}x > 0, \tag{232}$$

and this clearly gives (231) for any C for a suitably small but positive ε_0.

One may prove similar facts concerning the upper bound in Theorem 4 by considering $T(E_N, N, (1))$, the number of occurrences of 1 in E_N. Indeed, it suffices to observe that, for any fixed $C > 0$, we have

$$\lim_{N\to\infty} \mathbb{P}\left(\left|T(E_N, N, (1)) - \frac{N}{2}\right| \ge C\sqrt{N}\right) = 2\sqrt{\frac{2}{\pi}}\int_{C}^{\infty} e^{-2x^2}\mathrm{d}x > 0 \tag{233}$$

(we omit the details).

Problem 33. *Investigate the existence of the limiting distributions of*

$$\left\{\frac{W(E_N)}{\sqrt{N}}\right\}_{N\ge1} \quad \text{and} \quad \left\{\frac{\mathcal{N}(E_N)}{\sqrt{N}}\right\}_{N\ge1}$$

and

$$\left\{\frac{C_k(E_N)}{\sqrt{N\log\binom{N}{k}}}\right\}_{N\ge1}.$$

Investigate these distributions.

It is most likely that all three sequences in Problem 33 have limiting distributions. Note that Theorem 3 tell us that $\{C_k(E_N)/\mathbb{E}(C_k)\}_{N\ge1}$ has a limiting distribution that is concentrated at a point, as long as

$$k = k(N) \le \log N - \log\log N \tag{234}$$

(the condition in (234) may probably be weakened).

References

[1] N. Alon, Y. Kohayakawa, C. Mauduit, C. G. Moreira, and V. Rödl, *Measures of pseudorandomness for finite sequences: minimal values*, Combin. Probab. Comput. **15** (2006), no. 1–2, 1–29. 1.3, 1.3, 1.3

[2] N. Alon and J. H. Spencer, *The probabilistic method*, second ed., Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience [John Wiley & Sons], New York, 2000, With an appendix on the life and work of Paul Erdős. MR 2003f:60003 2.1

[3] B. Bollobás, *Random graphs*, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1985. MR 87f:05152 2.1

[4] J. Cassaigne, C. Mauduit, and A. Sárközy, *On finite pseudorandom binary sequences. VII. The measures of pseudorandomness*, Acta Arith. **103** (2002), no. 2, 97–118. MR 2004c:11139 1.2, 1.3

[5] K. L. Chung and P. Erdős, *On the application of the Borel-Cantelli lemma*, Trans. Amer. Math. Soc. **72** (1952), 179–186. MR 13:567b 2.3.1

[6] P. Erdős and A. Rényi, *On Cantor's series with convergent* $\sum 1/q_n$, Ann. Univ. Sci. Budapest. Eötvös. Sect. Math. **2** (1959), 93–109. MR 23:A3710 2.3.1

[7] P. Erdős and J. Spencer, *Probabilistic methods in combinatorics*, Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1974, Probability and Mathematical Statistics, Vol. 17. MR 52 #2895 1.2, 1.2

[8] S. Janson, T. Łuczak, and A. Ruciński, *Random graphs*, Wiley-Interscience, New York, 2000. MR 2001k:05180 2.1

[9] J. Matoušek and J. Spencer, *Discrepancy in arithmetic progressions*, J. Amer. Math. Soc. **9** (1996), no. 1, 195–204. MR 96c:11089 1.3

[10] C. Mauduit, *Finite and infinite pseudorandom binary words*, Theoret. Comput. Sci. **273** (2002), no. 1-2, 249–261, WORDS (Rouen, 1999). MR 2002m:11072 1.1

[11] C. Mauduit and A. Sárközy, *On finite pseudorandom binary sequences. I. Measure of pseudorandomness, the Legendre symbol*, Acta Arith. **82** (1997), no. 4, 365–377. MR 99g:11095 1.1

[12] C. McDiarmid, *On the method of bounded differences*, Surveys in combinatorics, 1989 (Norwich, 1989), London Math. Soc. Lecture Note Ser., vol. 141, Cambridge Univ. Press, Cambridge, 1989, pp. 148–188. MR 91e:05077 2.3.2

[13] V. V. Petrov, *A note on the Borel-Cantelli lemma*, Statist. Probab. Lett. **58** (2002), no. 3, 283–286. MR MR1921874 (2003e:60005) 2.3.1

[14] ______, *A generalization of the Borel-Cantelli lemma*, Statist. Probab. Lett. **67** (2004), no. 3, 233–239. MR MR2053525 (2005a:60011) 2.3.1

[15] K. F. Roth, *Remark concerning integer sequences*, Acta Arith. **9** (1964), 257–260. MR 29 #5806 1.3

Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel
E-mail address: noga@math.tau.ac.il

Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão 1010, 05508–090 São Paulo, Brazil
E-mail address: yoshi@ime.usp.br

Institut de Mathématiques de Luminy, CNRS-UPR9016, 163 av. de Luminy, case 907, F-13288, Marseille Cedex 9, France
E-mail address: mauduit@iml.univ-mrs.fr

IMPA, Estrada Dona Castorina 110, 22460–320 Rio de Janeiro, RJ, Brazil
E-mail address: gugu@impa.br

Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322, USA
E-mail address: rodl@mathcs.emory.edu