MEASURES OF PSEUDORANDOMNESS FOR FINITE SEQUENCES: MINIMUM AND TYPICAL VALUES (EXTENDED ABSTRACT)

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

Dedicated to Professor Imre Simon on the occasion of his 60th birthday

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'. These 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 these 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 these 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 $2 \le k \le N/4$. We also prove a lower bound for the correlation measure $C_k(E_N)$ (k even) for arbitrary sequences E_N, which, in particular, gives that $\min_{E_N} C_k(E_N) \ge c_k\sqrt{N}$ for some $c_k > 0$ for any constant even k.

1. Introduction and statement of results

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

Date: 15 August 2003.

1991 *Mathematics Subject Classification.* 68R15.

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

This research was partially supported by IM-AGIMB/IMPA. The first author was partially supported by MCT/CNPq a ProNEx project (Proc. CNPq 664107/1997–4) and by CNPq (Proc. 300334/93–1 and 468516/2000–0). The fourth author was partially supported by NSF Grant 0300529. The collaboration of the Brazilian authors and the US author is supported by a CNPq/NSF cooperative grant (910064/99–7, 0072064).

and further related literature are referred to [10] (WORDS, Rouen 1999) 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 $0 \leq d_1 < \cdots < d_k < N$ 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

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

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

and

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

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

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

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

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

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

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

In [3], Cassaigne, Mauduit, and Sárközy studied, amongst others, the value 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 minimal possible values for $W(E_N)$ and $C_k(E_N)$. They proved the following theorems. (Below and elsewhere in this abstract, we write log for the natural logarithm.)

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

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

with probability at least $1 - \varepsilon$.

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

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

with probability at least $1 - \varepsilon$.

Theorem C. *For all k and $N \in \mathbb{N}$ with $2 \leq k \leq N$, we have*

(*i*) $\min\{C_k(E_N)\colon E_N \in \{-1,1\}^N\} = 1$ *if k is odd,*
(*ii*) $\min\{C_k(E_N)\colon E_N \in \{-1,1\}^N\} \geq \log_2(N/k)$ *if k is even.*

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.

Let us also observe that it follows from the results of Roth [14] and Matoušek and Spencer [9] that the order of magnitude of

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

is $N^{1/4}$. In [3], it is conjectured that for any even $k \geq 2$ there is a constant $c > 0$ such that for $N \to \infty$ we have

$$\min\left\{C_k(E_N)\colon E_N \in \{-1,1\}^N\right\} \gg N^c,$$

which would be a considerable strengthening of Theorem C(*ii*).

In this extended abstract, we state stronger versions of Theorems A and B, we prove the conjecture above in a more general form, and we give a result concerning the typical value of $\mathcal{N}(E_N)$.

Theorem 1. *For any given $\varepsilon > 0$ there exist N_0 and $\delta > 0$ such that if $N \geq N_0$, 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]. The reader is referred to [8] for an alternative proof, the main idea of which is also used in the proof of Theorem 4 below (see Section 2.2).

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 $0 < \varepsilon_0 \leq 1$ 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 $N \geq N_0$, 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 $2 \leq k \leq N/4$.

The proof of Theorem 2 is sketched in Section 2.1. Our next result tells us that $C_k(E_N)$ is concentrated in the case in which k is small.

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

$$1 - \varepsilon < \frac{C_k(E_N)}{\Gamma(k,N)} < 1 + \varepsilon \tag{11}$$

holds is at least $1 - \varepsilon$.

Theorem 3 follows from standard arguments involving martingales (see [8] for details). Clearly, Theorem 2 tells us that $\Gamma(k, N)$ is of order $\sqrt{N \log \binom{N}{k}}$. 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 $N \geq N_0$, then*

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

with probability at least $1 - \varepsilon$.

The proof of Theorem 4 is sketched in Section 2.2. Finally, we state our result concerning the minimal possible value for the parameter $C_k(E_N)$.

Theorem 5. *If k and N are natural numbers with k even and $2 \leq k \leq N$, then*

$$C_k(E_N) > \sqrt{\frac{N}{3(k+1)}} \tag{13}$$

for any $E_N \in \{-1, 1\}^N$.

The proof of Theorem 5 is given in Section 3. In Section 4, we remark that the upper bounds in Theorems 1 and 4 are in a certain sense best possible.

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

We shall sketch the proofs of Theorems 2 and 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. **Proof of Theorem 2.** The upper estimate for $C_k(E_N)$ follows from standard estimates for the binomial distribution, and we refer the reader to [8] for the details. Here, we shall concentrate on the following result.

Lemma 6. *With probability tending to* 1 *as* $N \to \infty$, *we have*

$$C_k(E_N) > \frac{2}{5}\sqrt{N \log \binom{N}{k}} \tag{14}$$

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

We start by stating a technical result without proof.

Fact 7. *Let* $m = \lfloor N/3 \rfloor$. *For every sufficiently large* N, *the following hold.*

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

$$\sqrt{N \log \binom{N/3}{k}} \geq 0.99\sqrt{N \log \binom{N}{k}}. \tag{15}$$

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

$$\sqrt{N \log \binom{N/3}{k}} \geq \sqrt{\frac{1 - 10^{-10}}{3} N \log \binom{N}{k}}. \tag{16}$$

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

Proof of Lemma 6. 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 $2 \leq k \leq N/4$. We are interested in the largest integer r for which

$$\mathbb{P}\left(S(m, 1/2) \geq \frac{1}{2}(m + r)\right) \geq k^2(\log N)\binom{m+1}{k-1}^{-1} \tag{17}$$

holds. Indeed, we let

$$r(m) = r_k(m) = \max\left\{r \in \mathbb{N} \colon \text{inequality (17) holds}\right\}. \tag{18}$$

We need the following fact concerning $r(m)$.

Fact 8. *For every sufficiently large* N, *the following hold.*

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

$$r(m) \geq 0.99\sqrt{2m \log \binom{m+1}{k-1}}. \tag{19}$$

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

$$r(m) \geq (1 - 10^{-10})\sqrt{\frac{m}{\log 2} \log \binom{m+1}{k-1}}. \tag{20}$$

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

$$r(m) \geq \frac{2}{5}\sqrt{N \log \binom{N}{k}}. \tag{21}$$

We shall not prove Fact 8 here. Inequality (21) in Fact 8, which follows from (19) and (20) and the inequalities in Fact 7, will be used shortly.

To prove Lemma 6, we shall show that, with probability $\leq 2/k^2 \log N$, we have

$$C_k(E_N) \leq r_k(m), \tag{22}$$

and then we shall sum over all $2 \leq k \leq N/4$. This gives that (22) holds for *some* k with $2 \leq k \leq N/4$ with probability $O(1/\log N) = o(1)$. Therefore, (22) *fails* for *all* $2 \leq k \leq N/4$ with probability $1 - o(1)$, and hence, taking into account (21), Lemma 6 will be proved.

Let $2 \leq k \leq N/4$ be fixed. Our strategy to estimate the probability that (22) 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 $m \leq d_1 < \cdots < d_{k-1} \leq 2m$. For each $D \in \mathcal{D}_k$, let

$$v_D = (e_{1+d_1} e_{1+d_2} \dots e_{1+d_{k-1}}, \dots, e_{m+d_1} e_{m+d_2} \dots e_{m+d_{k-1}}), \tag{23}$$

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

$$|\langle u, v_D \rangle| > r(m) = r_k(m) \tag{24}$$

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

$$p = p(m) = \mathbb{P}\left(S(m, 1/2) \geq \frac{1}{2}(m + r(m)) \right). \tag{25}$$

Because of (18), we have

$$\mathbb{E}(X) = p|\mathcal{D}_k| = p\binom{m+1}{k-1} \geq k^2 \log N. \tag{26}$$

We have now arrived at the key claim: the events $A(D)$ ($D \in \mathcal{D}_k$) are pairwise independent. Although this somewhat surprising fact is crucial, we shall omit its proof because of space constraints. (The reader is invited to amuse himself or herself proving this; for a one-page proof, see [8]).

Claim 9. *For all distinct D and $D' \in \mathcal{D}_k$, we have $\mathbb{P}(A(D) \cap A(D')) = p^2$.*

To complete the proof of Lemma 6, 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 [4] and [6]. More recently, Petrov used this result to generalize the Borel–Cantelli lemma [12, 13].

Lemma 10. *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 $\mathbb{P}(A_i \cap A_j) \leq p^2(1+\varepsilon)$ for all $i \neq j$. Then*

$$\mathbb{P}\Big(\bigcup_{1 \leq j \leq M} A_j\Big) \geq \frac{Mp}{1 + (M-1)p(1+\varepsilon)} > 1 - \varepsilon - \frac{2}{Mp}.$$

We conclude the proof of Lemma 6 combining Claim 9 and Lemma 10. 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 (26)). Lemma 10 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 (22) 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 6 follows. □

2.2. **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

$$\mathcal{N}(E_N) = \max_k \max_X \max_M \left| T(E_N, M, X) - \frac{M}{2^k} \right|, \tag{27}$$

where the maxima are taken over all $k \leq \log_2 N$, $X \in \{-1,1\}^k$, and $0 < M \leq N + 1 - k$, 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 $j \leq M$ (see (1)). In this section, we outline the proof of Theorem 4.

Proof of Theorem 4 (Outline). We start with a sketch of the proof of the lower bound in (12). We take $k = 1$ in (27); in fact, we consider $T(E_N, N, (1))$, the number of occurrences of 1 in E_N. Then a simple application of the de Moivre–Laplace theorem on the binomial distribution (see, *e.g.*, [2]) tells us 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 11 below. We shall consider intervals of integers of the form $B_{m,r} = (m2^r, (m+1)2^r] \cap \mathbb{Z}$, where m and r are non-negative integers. Clearly, $|B_{m,r}| = 2^r$. We refer to the $B_{m,r}$ as *blocks*. For an integer $k \geq 1$, $X \in \{-1,1\}^k$, and $B_{m,r}$ a block with

$$\max B_{m,r} = (m+1)2^r \leq N - k + 1, \tag{28}$$

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, $T(E_N, B_{m,r}, X) = \operatorname{card}\{n \in B_{m,r} \colon E_N^{(n)} = X\}$, where $E_N^{(n)} = (e_j)_{n \leq j < n+k}$, and, as usual, $E_N = (e_1, \dots, e_N)$.

Lemma 11. *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 $k \leq \log_2 N$ satisfying* (28) *such that*

$$\left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| > D\sqrt{2^r} \tag{29}$$

is at most

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

We shall not prove Lemma 11 here. Let us continue with the proof of Theorem 4. Let us apply Lemma 11 with

$$D = C(K - r) \tag{30}$$

for all blocks $B_{m,r} \subset [1, N]$, where $K = 1 + \lfloor \log_2 N \rfloor$, 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

$$r \geq 4 \log_2 \log_2 N \tag{31}$$

and *small* otherwise. If C is a large enough constant, then, with the value of D given in (30), it follows from Lemma 11 that inequality (29) holds for some large block $B_{m,r}$ and some $X \in \{-1, 1\}^k$ ($k \leq \log_2 N$) with probability

$$O\left(e^{-2C^2/9}\right). \tag{32}$$

Since the bound in (32) tends to 0 as $C \to \infty$, we may and shall suppose henceforth that (**) *for all integers m, r, and k with $B_{m,r} \subset [1, N]$ satisfying* (28) *and* (31), *and every $X \in \{-1, 1\}^k$ ($k \leq \log_2 N$), we have*

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

Now observe that, for any $1 \leq M \leq N - k + 1$, 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

$$I_+ = \{(m, r) \in I \colon r \text{ satisfies (31)}\}. \tag{33}$$

For later reference, observe that

$$|I_-| < 4 \log_2 \log_2 N. \tag{34}$$

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

$$\left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| \leq 2^r < (\log_2 N)^4 \tag{35}$$

for any $X \in \{-1, 1\}^k$. Using (**), (34), and (35), one may check that, for any $X \in \{-1, 1\}^k$ ($k \leq \log_2 N$), we have

$$\left| T(E_N, M, X) - M2^{-k} \right| \leq \sum_{(m,r) \in I} \left| T(E_N, B_{m,r}, X) - 2^{r-k} \right| = O(\sqrt{N}),$$

as required (we omit the details). □

2.2.1. *Two remarks and a problem on $\mathcal{N}$.* We close this section with two observations about the normality measure $\mathcal{N}$ and an open problem that we find quite attractive. Put

$$\mathcal{N}_k(E_N) = \max_X \max_M \left| T(E_N, M, X) - \frac{M}{2^k} \right|, \tag{36}$$

where the maxima are taken over all $X \in \{-1,1\}^k$ and $0 < M \leq N+1-k$. Note that, then, we have $\mathcal{N}(E_N) = \max\{\mathcal{N}_k(E_N) \colon k \leq \log_2 N\}$.

Our first remark is that $\min_{E_N} \mathcal{N}_k(E_N) = 1$ for any fixed k: we consider powers of appropriate de Bruijn sequences. More precisely, we take a circular sequence in which every member of $\{-1,1\}^k$ occurs exactly once, open it up (turning it into a linear sequence), and repeat it many times.

Our second remark is that $\min_{E_N} \mathcal{N}(E_N) \geq (1/2 - o(1)) \log_2 N$: if a sequence $E_N \in \{-1,1\}^N$ contains no segment of length $k = \log_2 N - \log_2 \log_2 N$ of repeated 1s (say), then $\mathcal{N}_k(E_N) \geq \log_2 N = N/2^k$. If E_N contains such a segment, then $\mathcal{N}_1(E_N) \geq k/2 = (1/2)(\log_2 N - \log_2 \log_2 N)$.

Problem 12. *Is there an absolute constant $\alpha > 0$ for which we have*

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

for all large enough N?

3. The minimum of the correlation measure

The proof of Theorem 5 that we give in this section is based on the following elementary lemma from linear algebra [5, Lemma 7], whose proof we include for completeness.

Lemma 13. *For any symmetric matrix $\mathcal{M} = (\mathcal{M}_{ij})_{1\leq i,j\leq n}$, we have*

$$\operatorname{rank}(\mathcal{M}) \geq \frac{(\operatorname{trace}(\mathcal{M}))^2}{\operatorname{trace}(\mathcal{M}^2)} = \frac{\left(\sum_{1\leq i\leq n} \mathcal{M}_{ii}\right)^2}{\sum_{1\leq i,j\leq n} \mathcal{M}_{ij}^2}. \tag{37}$$

Proof. Let $r = \operatorname{rank}(\mathcal{M})$. Then $\mathcal{M}$ has exactly r non-zero eigenvalues, say, $\lambda_1, \ldots, \lambda_r$. By the Cauchy–Schwarz inequality, we have

$$(\operatorname{trace}(\mathcal{M}))^2 = (\lambda_1 + \cdots + \lambda_r)^2 \leq r(\lambda_1^2 + \cdots + \lambda_r^2) = r \operatorname{trace}(\mathcal{M}^2).$$

As $\mathcal{M}$ is symmetric, we have $\operatorname{trace}(\mathcal{M}^2) = \sum_{1\leq i\leq n} \left(\sum_{1\leq j\leq n} \mathcal{M}_{ij}\mathcal{M}_{ji}\right) = \sum_{1\leq i,j\leq n} \mathcal{M}_{ij}^2$, as required. □

Proof of Theorem 5. First we remark that $C_N(E_N) = 1$, so that (13) is true for $k = N$, and we can suppose for a contradiction that there exist k with $2 \leq k \leq N-1$ and $E_N \in \{-1,1\}^N$ such that

$$C_k(E_N) \leq \sqrt{\frac{N}{3(k+1)}}. \tag{38}$$

Set $k = 2\ell$, put

$$M = \left\lfloor \frac{N}{k+1} \right\rfloor, \tag{39}$$

and consider the $2M$ vectors $v_0, \dots, v_{2M-1} \in \mathbb{R}^M$ given by

$$v_i = \left(\prod_{1\le j\le\ell} e_{i\ell+j}, \prod_{1\le j\le\ell} e_{i\ell+j+1}, \dots, \prod_{1\le j\le\ell} e_{i\ell+j+M-1} \right) \tag{40}$$

for $i \in \{0, \dots, 2M-1\}$.

Define the $2M \times 2M$ matrix $\mathcal{M} = (\mathcal{M}_{ij})_{1\le i,j\le 2M}$ putting

$$\mathcal{M}_{ij} = \langle v_i, v_j \rangle \tag{41}$$

for all $(i,j) \in \{1, \dots, 2M\}^2$. Then $\mathcal{M}$ has the following properties:

(i) $\operatorname{rank}(\mathcal{M}) \le M$,
(ii) $\mathcal{M}_{ii} = M$ for any $i \in \{1, \dots, 2M\}$,
(iii) $|\mathcal{M}_{ij}| \le N/3(k+1)$ for any $(i,j) \in \{1, \dots, 2M\}^2$ with $i \ne j$.

Indeed, (i) and (ii) are clear and (iii) follows from (38). It follows from Lemma 13 that

$$\operatorname{rank}(\mathcal{M}) \ge \frac{\left(2M^2\right)^2}{2M \times M^2 + 2M(2M-1)N/3(k+1)}$$
$$= \frac{2M^3}{M^2 + (2M-1)N/3(k+1)},$$

so that, by (39), we have

$$\operatorname{rank}(\mathcal{M}) > \frac{2M^3}{M^2 + (2M-1)(M+1)/3} = M\frac{6M^2}{5M^2 + M - 1} > M,$$

which contradicts property (i) above. This contradiction shows that for no k with $2 \le k \le N-1$ there is $E_N \in \{-1, 1\}^N$ such that (38) holds, and our result follows. □

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

$$\mathbb{P}\Big(W(E_N) < C\sqrt{N} \Big) \le 1 - \varepsilon_0 \tag{42}$$

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 (42) follows simply from the de Moivre–Laplace theorem on the binomial distribution (we omit the details). 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.

Problem 14. *Investigate the existence of the limiting distributions of*

$$\left\{\frac{W(E_N)}{\sqrt{N}}\right\}_{N\geq 1} \quad \textit{and} \quad \left\{\frac{\mathcal{N}(E_N)}{\sqrt{N}}\right\}_{N\geq 1}$$

and

$$\left\{\frac{C_k(E_N)}{\sqrt{N\log\binom{N}{k}}}\right\}_{N\geq 1}.$$

Investigate these distributions.

It is most likely that all three sequences in Problem 14 have limiting distributions. Note that Theorems 2 and 3 tell us that there is a function $\Gamma(k,N)$, whose order of magnitude is $\sqrt{N\log\binom{N}{k}}$, for which $\{C_k(E_N)/\Gamma(k,N)\}_{N\geq 1}$ has a limiting distribution that is concentrated on a point, as long as $k = k(N) \leq \log N - \log\log N$. Finally, we believe that Problem 12 merits investigation.

Note added in the revision. Very recently, Noga Alon has obtained an algebraic construction for a sequence E_N with $\mathcal{N}(E_N)$ of order at most $N^{1/3}(\log N)^{O(1)}$. He has also observed that Theorem 9.3 from [1] may be used to prove further lower estimates for $\min_{E_N} C_k(E_N)$. We shall come back to these results in the near future.

References

[1] N. Alon. Problems and results in extremal combinatorics I. *Discrete Math.* To appear.

[2] B. Bollobás. *Random graphs.* Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1985.

[3] J. Cassaigne, C. Mauduit, and A. Sárközy. On finite pseudorandom binary sequences. VII. The measures of pseudorandomness. *Acta Arith.*, 103(2):97–118, 2002.

[4] K. L. Chung and P. Erdös. On the application of the Borel-Cantelli lemma. *Trans. Amer. Math. Soc.*, 72:179–186, 1952.

[5] B. Codenotti, P. Pudlák, and G. Resta. Some structural properties of low-rank matrices related to computational complexity. *Theoret. Comput. Sci.*, 235(1):89–107, 2000. Selected papers in honor of Manuel Blum (Hong Kong, 1998).

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

[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.

[8] Y. Kohayakawa, C. Mauduit, C. G. Moreira, and Rödl. Measures of pseudorandomness for finite sequences: minimum and typical values. In preparation, 2003.

[9] J. Matoušek and J. Spencer. Discrepancy in arithmetic progressions. *J. Amer. Math. Soc.*, 9(1):195–204, 1996.

[10] C. Mauduit. Finite and infinite pseudorandom binary words. *Theoret. Comput. Sci.*, 273(1-2):249–261, 2002. WORDS (Rouen, 1999).

[11] C. Mauduit and A. Sárközy. On finite pseudorandom binary sequences. I. Measure of pseudorandomness, the Legendre symbol. *Acta Arith.*, 82(4):365–377, 1997.

[12] V. V. Petrov. A generalisation of the Borel–Cantelli lemma. manuscript, 2002, 8pp.

[13] V. V. Petrov. A note on the Borel-Cantelli lemma. *Statist. Probab. Lett.*, 58(3):283–286, 2002.

[14] K. F. Roth. Remark concerning integer sequences. *Acta Arith.*, 9:257–260, 1964.

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