MEASURES OF PSEUDORANDOMNESS FOR FINITE SEQUENCES: MINIMAL VALUES

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

Dedicated to Professor Béla Bollobás 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'. Two of these parameters are the *normality measure* $\mathcal{N}(E_N)$ and the *correlation measure* $C_k(E_N)$ *of order* k, which focus on different combinatorial aspects of E_N. In their work, amongst others, Mauduit and Sárközy investigated the minimal possible value of these parameters.

In this paper, we continue the work in this direction and prove a lower bound for the correlation measure $C_k(E_N)$ (k even) for arbitrary sequences E_N, establishing one of their conjectures. We also give an algebraic construction for a sequence E_N with small normality measure $\mathcal{N}(E_N)$.

Contents

Date: Copy produced on June 29, 2005.

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 supported by a USA Israeli BSF grant, by a grant from the Israel Science Foundation and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University. The second author was partially supported by FAPESP and CNPq through ProNEx projects (Proc. CNPq 664107/1997–4 and Proc. FAPESP 2003/09925–5) and by CNPq (Proc. 306334/2004–6 and 479882/2004–5). 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 0300529. The authors gratefully acknowledge the support of a CNPq/NSF cooperative grant (910064/99–7, 0072064) and the Brazil/France Agreement in Mathematics (Proc. CNPq 69–0014/01–5 and 69–0140/03–7).

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] certain 'measures of pseudorandomness', to be defined shortly. We restrict ourselves to the Mauduit–Sárközy parameters directly relevant to the present note, and refer the reader to [10] and [11] for detailed discussions concerning the definitions below, related measures, and further related literature.

Let $k \in \mathbb{N}$, $M \in \mathbb{N}$, and $X \in \{-1, 1\}^k$ be given. Also, let $D = \{d_1, \dots, d_k\}$, where the d_i are integers with $1 \le d_1 < \cdots < d_k \le N - M + 1$. 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

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

and

$$V(E_N, M, D) = \sum \{e_{n+d_1} e_{n+d_2} \dots e_{n+d_k} : 0 \le n < M\} \\ = \sum_{0 \le n < M} \prod_{1 \le i \le k} e_{n+d_i} = \sum_{0 \le n < M} \prod_{d \in D} e_{n+d}. \quad (2)$$

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. On the other hand, one may think of the quantity $V(E_N, M, D)$ as 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

$$\mathcal{N}(E_N) = \max_k \max_X \max_M \left| T(E_N, M, X) - \frac{M}{2^k} \right|, \quad (3)$$

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

$$C_k(E_N) = \max\{|V(E_N, M, D)|\colon M \text{ and } D \text{ such that } M - 1 + d_k \le N\}. \quad (4)$$

In what follows, we shall sometimes make use of terms commonly used in the area of combinatorics on words. In particular, sequences will sometimes be referred to as *words*. Moreover, a word u *occurs* in a word w if w contains u as a 'contiguous segment' (that is, $w = tuv$, where t is a 'prefix' of w and v is a 'suffix' of w).

In Section 1.1 we shall state and discuss our results concerning the correlation measure C_k, while in Section 1.2 we shall state and discuss our results on the normality measure $\mathcal{N}$.

1.1. **Typical and minimal values of correlation.** In [4], Cassaigne, Mauduit, and Sárközy studied, amongst others, the typical value of $C_k(E_N)$ for random binary sequences E_N, with all the 2^N sequences in $\{-1,1\}^N$ equiprobable, and the minimal possible value for $C_k(E_N)$. The investigation of the typical value of $C_k(E_N)$ is continued in [1], where Theorems A and B below are proved. (In what follows, we write log for the natural logarithm.)

Theorem A. *Let $0 < \varepsilon_0 < 1/16$ 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 $N \geq N_0$, then, with probability at least $1 - \varepsilon_0$, we have*

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

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

Note that Theorem A establishes the typical order of magnitude of $C_k(E_N)$ for a wide range of k, including values of k proportional to N. The next result tells us that $C_k(E_N)$ is concentrated in the case in which k is small.

Theorem B. *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*

$$1 - \varepsilon < \frac{C_k(E_N)}{\Gamma(k,N)} < 1 + \varepsilon \tag{6}$$

holds is at least $1 - \varepsilon$.

Clearly, Theorem A tells us that $\Gamma(k,N)$ is of order $\sqrt{N\log\binom{N}{k}}$. Let us now turn to the minimal possible value of the parameter $C_k(E_N)$. In [4], the following result is proved.

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

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

Theorem C(i) follows simply from the observation that the alternating sequence $E_N = (1, -1, 1, -1, \dots)$ is such that $C_k(E_N) = 1$ for odd k. Owing to Theorem C(i), when concerned with minimal values of $C_k(E_N)$, we are only interested in even k. In [4], it is conjectured that for any even $k \geq 2$ there is a constant $c > 0$ such that for $N \to \infty$ we have

$$\min\left\{C_k(E_N)\colon E_N \in \{-1,1\}^N\right\} \gg N^c, \tag{7}$$

which would be a considerable strengthening of Theorem C(ii). In this paper, we prove the conjecture above in a more general form. We shall prove the following result.

Theorem 1. *If k and N are natural numbers with k even and $2 \leq k \leq N$, then*

$$C_k(E_N) > \sqrt{\frac{1}{2}\left\lfloor\frac{N}{k+1}\right\rfloor} \tag{8}$$

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

The lower bound given in (8) decreases as k increases. One may ask whether, in fact, $C_{2k}(E_N) \geq c\sqrt{kN}$ for some absolute constant $c > 0$, or at least $C_{2k}(E_N) \geq c\sqrt{N}$ for some absolute constant $c > 0$. The results below (and the results in Section 2.3) are partial answers in this direction.

It turns out that if we look at the maximum of $C_2(E_N), C_4(E_N), \dots, C_k(E_N)$ (with k again even), then a lower bound of order $\sqrt{kN}$ may indeed be proved.

Theorem 2. *There is an absolute constant $c > 0$ for which the following holds. For any positive integers ℓ and N with $\ell \leq N/3$, we have*

$$\max\{C_2(E_N), C_4(E_N), \dots, C_{2\ell}(E_N)\} \geq c\sqrt{\ell N} \tag{9}$$

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

In view of Theorem A, the lower bound in Theorem 2 is best possible apart from a multiplicative factor of $O\left(\sqrt{\log(N/2\ell)}\right)$, for all $\ell \leq N/8$.

One may also prove lower bounds of the form $c\sqrt{N}$ for some absolute constant $c > 0$ if one considers correlations of two consecutive even orders $2k-2$ and $2k$ (with k not too large).

Theorem 3. *Let positive integers k and N with $2 \leq k \leq \sqrt{N/6}$ be given. If N is large enough, then*

$$\max\{C_{2k-2}(E_N), C_{2k}(E_N)\} \geq \sqrt{\frac{1}{2}\left\lfloor\frac{N}{3}\right\rfloor} \tag{10}$$

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

Some further results are stated and proved in Section 2.3 (see Theorems 11, 13, and 14).

1.2. **Typical and minimal values of normality.** We now turn to the normality measure $\mathcal{N}(E_N)$. In [1], the following result is proved.

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

$$\delta\sqrt{N} < \mathcal{N}(E_N) < \frac{1}{\delta}\sqrt{N} \tag{11}$$

with probability at least $1-\varepsilon$.

Here, we shall give an explicit construction for sequences $E_N \in \{-1,1\}^N$ with $\mathcal{N}(E_N)$ small. Theorem D tells us that, typically, $\mathcal{N}(E_N)$ is of order $\sqrt{N}$. We shall exhibit a sequence E_N with $\mathcal{N}(E_N) = O\left(N^{1/3}(\log N)^{2/3}\right)$.

Theorem 4. *For any sufficiently large N, there exists a sequence $E_N \in \{-1,1\}^N$ with*

$$\mathcal{N}(E_N) \leq 3N^{1/3}(\log N)^{2/3}. \tag{12}$$

A simple argument shows that $\mathcal{N}(E_N) \geq (1/2+o(1))\log_2 N$ for any $E_N \in \{-1,1\}^N$ (see Proposition 16 in Section 3.1). In view of Theorem 4, we have

$$\left(\frac{1}{2}+o(1)\right)\log_2 N \leq \min_{E_N\in\{-1,1\}^N} \mathcal{N}(E_N) \leq 3N^{1/3}(\log N)^{2/3} \tag{13}$$

for all large enough N. It would be interesting to close the rather wide gap in (13).

The construction of the sequence $E_N \in \{-1,1\}^N$ in Theorem 4 may be generalized to larger alphabets Σ, as long as the cardinality of Σ is a power of a prime (see Section 3.3). Finally, we remark that one of the ingredients in the proof of (12) for our sequence E_N allows one to give a short proof of the celebrated Pólya–Vinogradov inequality on incomplete character sums (see Section 3.4), which is somewhat simpler than the known proofs.

2. The minimum of the correlation measure

2.1. **Auxiliary lemmas from linear algebra.** The proof of Theorem 1 that we give in Section 2.2 is based on the following elementary lemma from linear algebra (see, e.g., [2, Lemma 9.1] or [5, Lemma 7]), whose proof we include for completeness.

Lemma 5. *For any symmetric matrix $\mathbf{A} = (A_{ij})_{1\leq i,j\leq n}$, we have*

$$\operatorname{rk}(\mathbf{A}) \geq \frac{(\operatorname{tr}(\mathbf{A}))^2}{\operatorname{tr}(\mathbf{A}^2)} = \frac{\left(\sum_{1\leq i\leq n} A_{ii}\right)^2}{\sum_{1\leq i,j\leq n} A_{ij}^2}. \tag{14}$$

Consequently, if $A_{ii}=1$ for all i and $|A_{ij}|\leq\varepsilon$ for all $i\neq j$, then

$$\operatorname{rk}(\mathbf{A}) \geq \frac{n}{1+\varepsilon^2(n-1)}. \tag{15}$$

In particular, if $\varepsilon = \sqrt{1/n}$, then $\operatorname{rk}(\mathbf{A}) \geq n/2$.

Proof. Let $r = \mathrm{rk}(\mathbf{A})$. Then $\mathbf{A}$ has exactly r non-zero eigenvalues, say, $\lambda_1, \dots, \lambda_r$. By the Cauchy–Schwarz inequality, we have

$$(\mathrm{tr}(\mathbf{A}))^2 = (\lambda_1 + \cdots + \lambda_r)^2 \le r(\lambda_1^2 + \cdots + \lambda_r^2) = r\,\mathrm{tr}(\mathbf{A}^2),$$

and it now suffices to notice that, because $\mathbf{A}$ is symmetric, we have

$$\mathrm{tr}(\mathbf{A}^2) = \sum_{1\le i\le n}\left(\sum_{1\le j\le n} A_{ij}A_{ji}\right) = \sum_{1\le i,j\le n} A_{ij}^2,$$

as required. Inequality (15) follows immediately from (14). □

The next lemma, due to the first author [2], improves Lemma 5 for larger values of ε.

Lemma 6. *Let $\mathbf{A} = (A_{ij})_{1\le i,j\le n}$ be an $n \times n$ real matrix with $A_{ii} = 1$ for all i and $|A_{ij}| \le \varepsilon$ for all $i \ne j$, where $\sqrt{1/n} \le \varepsilon \le 1/2$. Then*

$$\mathrm{rk}(\mathbf{A}) \ge \frac{1}{100\varepsilon^2 \log(1/\varepsilon)} \log n. \tag{16}$$

If $\mathbf{A}$ is symmetric, then (16) holds with the constant $1/100$ replaced by $1/50$.

For completeness, we give the proof of Lemma 6. We shall need the following auxiliary lemma [2].

Lemma 7. *Let $\mathbf{A} = (A_{i,j})$ be an $n \times n$ matrix of rank d, and let $P(x)$ be an arbitrary polynomial of degree k. Then the rank of the $n \times n$ matrix $(P(A_{i,j}))$ is at most $\binom{k+d}{k}$. Moreover, if $P(x) = x^k$, then the rank of $(P(A_{i,j})) = (A_{i,j}^k)$ is at most $\binom{k+d-1}{k}$.*

Proof. Let $\mathbf{v}_1 = (v_{1,j})_{j=1}^n$, $\mathbf{v}_2 = (v_{2,j})_{j=1}^n, \dots, \mathbf{v}_d = (v_{d,j})_{j=1}^n$ be a basis of the row space of $\mathbf{A}$. Then the vectors $(v_{1,j}^{k_1} v_{2,j}^{k_2} \cdots v_{d,j}^{k_d})_{j=1}^n$, where $k_1, k_2, \dots, k_d$ range over all non-negative integers whose sum is at most k, span the row space of the matrix $(P(A_{i,j}))$. If $P(x) = x^k$, then it suffices to take all these vectors corresponding to $k_1, k_2, \dots, k_d$ whose sum is precisely k. □

Proof of Lemma 6. Let us first note that the non-symmetric case follows from the symmetric case: if $\mathbf{A}$ is not symmetric, it suffices to consider the symmetric matrix $(\mathbf{A}^T + \mathbf{A})/2$, whose rank is at most twice the rank of $\mathbf{A}$. We therefore suppose that $\mathbf{A}$ is symmetric, and proceed to prove (16) with the constant $1/100$ replaced by $1/50$.

Let $\delta = 1/16$. Consider first the case in which $\varepsilon \le 1/n^\delta$. In this case, let $m = \lfloor 1/\varepsilon^2 \rfloor$, and let $\mathbf{A}'$ be the submatrix of $\mathbf{A}$ consisting of the, say, first m rows and first m columns of $\mathbf{A}$. By the choice of m, we have that $1/\sqrt{m} \ge \varepsilon$, and hence Lemma 5 applies to $\mathbf{A}'$, and we deduce that $\mathrm{rk}(\mathbf{A}) \ge \mathrm{rk}(\mathbf{A}') \ge m/2$. It now suffices to check that, because $\varepsilon \le \min\{1/2, 1/n^\delta\}$ and $\delta = 1/16$, we have

$$\frac{1}{2}m \ge \frac{3}{8\varepsilon^2} = \frac{3}{2^7\delta\varepsilon^2} > \frac{1}{50\varepsilon^2\log(1/\varepsilon)} \log n, \tag{17}$$

and we are done in this case. We now suppose that $1/n^\delta \le \varepsilon \le 1/2$. In this case, we let

$$k = \left\lfloor \frac{\log n}{2\log(1/\varepsilon)} \right\rfloor \ge \left\lfloor \frac{1}{2\delta} \right\rfloor = 8, \tag{18}$$

and let $m = \lfloor 1/\varepsilon^{2k} \rfloor$. Note that, then, we have $m \le n$. We again let $\mathbf{A}'$ be the submatrix of $\mathbf{A}$ consisting of the first m rows and first m columns of $\mathbf{A}$. We now have

$$\varepsilon^k \le \frac{1}{\sqrt{m}}. \tag{19}$$

Let $\mathbf{A}''$ be the matrix obtained from $\mathbf{A}'$ by raising all its entries to the kth power. Because of (19) and the hypothesis on the entries of $\mathbf{A}$, Lemma 5 applies and tells us that

$$\mathrm{rk}(\mathbf{A}'') \ge \frac{1}{2}m = \frac{1}{2}\left\lfloor \frac{1}{\varepsilon^{2k}} \right\rfloor \ge \frac{0.49}{\varepsilon^{2k}}, \tag{20}$$

where the last inequality follows easily from the fact that $\varepsilon \le 1/2$ and $k \ge 8$ (see (18)). We now observe that Lemma 7 tells us that

$$\mathrm{rk}(\mathbf{A}'') \le \binom{k + \mathrm{rk}(\mathbf{A}')}{k} \le \left(\frac{\mathrm{e}(k + \mathrm{rk}(\mathbf{A}'))}{k} \right)^k. \tag{21}$$

Putting together (20) and (21), we get

$$\mathrm{rk}(\mathbf{A}) \ge \mathrm{rk}(\mathbf{A}') \ge \frac{k}{\varepsilon^2}\left(\frac{0.49^{1/k}}{\mathrm{e}} - \varepsilon^2 \right), \tag{22}$$

which, because $0.49^{1/8}/\mathrm{e} \ge 1/3$ and $\varepsilon^2 \le 1/4$, implies that $\mathrm{rk}(\mathbf{A}) \ge k/12\varepsilon^2$. Therefore, we have

$$\mathrm{rk}(\mathbf{A}) > \frac{1}{50\varepsilon^2 \log(1/\varepsilon)} \log n, \tag{23}$$

and we are done. □

2.2. **Proof of the lower bounds for correlation.** We shall prove Theorem 1 and 2 in this section. These results will be deduced from suitable applications of Lemmas 5 and 6; to describe these applications, we first need to introduce some notation.

Let $E_N = (e_i)_{1 \le i \le N} \in \{-1,1\}^N$ be given. Let a positive integer $M \le N$ be fixed and set $N' = N - M + 1$. Moreover, fix a family $\mathcal{L}$ of subsets of $[N']$. We now define a vector $\mathbf{v}_L = (v_{L,i})_{0 \le i < M} \in \{-1,1\}^M$ for all $L \in \mathcal{L}$, letting

$$v_{L,i} = \prod_{x \in L} e_{i+x} \tag{24}$$

for all $0 \le i < M$ (note that $1 \le i + x \le M - 1 + N' = N$ for any x in (24)). Let us now define an $\mathcal{L} \times \mathcal{L}$ matrix $\mathbf{A} = (A_{L,L'})_{L,L' \in \mathcal{L}}$, putting

$$A_{L,L'} = \frac{1}{M}\langle \mathbf{v}_L, \mathbf{v}_{L'} \rangle = \frac{1}{M}\sum_{0 \le i < M} v_{L,i} v_{L',i} \tag{25}$$

for all L, $L' \in \mathcal{L}$. Clearly, the diagonal entries of $\mathbf{A}$ are all 1. Suppose now that $L \neq L'$. Then

$$A_{L,L'} = \frac{1}{M}\langle \mathbf{v}_L, \mathbf{v}_{L'}\rangle = \frac{1}{M}\sum_{0\le i<M}\Big(\prod_{x\in L} e_{i+x}\Big)\Big(\prod_{y\in L'} e_{i+y}\Big)$$
$$= \frac{1}{M}\sum_{0\le i<M}\prod_{z\in L\triangle L'} e_{i+z}, \tag{26}$$

where we write $L \triangle L'$ for the symmetric difference of the sets L and L'. Let $\mathcal{L}^{\triangle} = \{L \triangle L'\colon L,\ L' \in \mathcal{L},\ \ L \neq L'\}$ and let K be the set of the cardinalities of the members of $\mathcal{L}^{\triangle}$, that is, $K = \{|S|\colon S \in \mathcal{L}^{\triangle}\}$. It follows from (26) and the definition of $C_k(E_N)$ that

$$\max\{C_k(E_N)\colon k \in K\} \geq M \max\{|A_{L,L'}|\colon L,\ L' \in \mathcal{L},\ L \neq L'\}. \tag{27}$$

Lemma 5 and (27) imply the following result.

Lemma 8. *We have*

$$\max\{C_k(E_N)\colon k \in K\} > \sqrt{M - \frac{M^2}{|\mathcal{L}|}}. \tag{28}$$

Proof. Let $\mathbf{B} = (\mathbf{v}_L^T)_{L\in\mathcal{L}}$ be the $|\mathcal{L}| \times M$ matrix with rows $\mathbf{v}_L^T$ $(L \in \mathcal{L})$. Observing that $\mathbf{A} = M^{-1}\mathbf{B}\mathbf{B}^T$, we see that $\mathbf{A}$ has rank at most M. Combining this with the lower bound for the rank of $\mathbf{A}$ given by Lemma 5, we get

$$M \geq \mathrm{rk}(\mathbf{A}) > \frac{|\mathcal{L}|}{1+\varepsilon^2|\mathcal{L}|}, \tag{29}$$

where $\varepsilon = \max\{|A_{L,L'}|\colon L,\ L' \in \mathcal{L},\ \ L \neq L'\}$. It follows from (29) that

$$\varepsilon > \sqrt{\frac{1}{M} - \frac{1}{|\mathcal{L}|}}. \tag{30}$$

Inequality (28) follows from (27) on multiplying (30) by M. $\square$

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let k, N, and E_N be as in the statement of Theorem 1. Set $\ell = k/2$ and $M = \lfloor N/(k+1)\rfloor$ and, as above, let $N' = N - M + 1$. We take for $\mathcal{L} \subset \mathcal{P}([N'])$ a set system of $t = \lfloor N'/\ell\rfloor$ pairwise disjoint ℓ-element subsets $L_1, \dots, L_t$ of $[N']$. Note that

$$|\mathcal{L}| = t = \left\lfloor \frac{N - \lfloor N/(k+1)\rfloor + 1}{k/2}\right\rfloor \geq \left\lfloor\frac{2N}{k+1}\right\rfloor \geq 2M. \tag{31}$$

Therefore, it follows from (28) and (31) that

$$C_k(E_N) > \sqrt{M - \frac{M^2}{|\mathcal{L}|}} \geq \sqrt{M - \frac{M}{2}} = \sqrt{\frac{1}{2}\left\lfloor\frac{N}{k+1}\right\rfloor}, \tag{32}$$

as required. $\square$

Lemma 8 was deduced from an application of Lemma 5 to the matrix $\mathbf{A} = (A_{L,L'})$; the next lemma will be obtained from an application of Lemma 6 to $\mathbf{A}$.

Lemma 9. *If $2M \le |\mathcal{L}| < \mathrm{e}^{50M}$, then*

$$\max\{C_k(E_N)\colon k \in K\} \ge \min\left\{\frac{1}{2}M, \sqrt{\frac{1}{50}M(\log|\mathcal{L}|)\Big/\log\frac{50M}{\log|\mathcal{L}|}}\right\}. \tag{33}$$

Proof. Let $\varepsilon = \max\{|A_{L,L'}|\colon L,\, L' \in \mathcal{L},\ L \ne L'\}$. Inequality (15) and the fact that $\mathrm{rk}(\mathbf{A}) \le M$, coupled with $M \le |\mathcal{L}|/2$, give that

$$\varepsilon^2 > \frac{1}{M} - \frac{1}{|\mathcal{L}|} \ge \frac{1}{|\mathcal{L}|}, \tag{34}$$

and hence $\varepsilon > \sqrt{1/|\mathcal{L}|}$. If $\varepsilon > 1/2$, then (33) follows immediately (recall (27)). Therefore, we may suppose that $\sqrt{1/|\mathcal{L}|} \le \varepsilon \le 1/2$, and hence we may apply Lemma 6 to the symmetric matrix $\mathbf{A}$. Combining the fact that $\mathbf{A}$ has rank at most M with Lemma 6, we obtain that

$$M \ge \mathrm{rk}(\mathbf{A}) \ge \frac{1}{50\varepsilon^2\log(1/\varepsilon)}\log|\mathcal{L}|, \tag{35}$$

whence

$$\varepsilon^2\log\frac{1}{\varepsilon} \ge \frac{1}{50M}\log|\mathcal{L}|. \tag{36}$$

Using that $1/\varepsilon \ge \log 1/\varepsilon$, we have from (36) that

$$\varepsilon \ge \varepsilon^2\log\frac{1}{\varepsilon} \ge \frac{1}{50M}\log|\mathcal{L}|. \tag{37}$$

Plugging (37) into (36), we get

$$\varepsilon^2\log\frac{50M}{\log|\mathcal{L}|} \ge \varepsilon^2\log\frac{1}{\varepsilon} \ge \frac{1}{50M}\log|\mathcal{L}|, \tag{38}$$

and hence

$$\varepsilon \ge \sqrt{\frac{\log|\mathcal{L}|}{50M}\Big/\log\frac{50M}{\log|\mathcal{L}|}}. \tag{39}$$

Inequality (33) follows easily from (27), (39), and the definition of ε. □

We shall now deduce Theorem 2 from Lemma 9.

Proof of Theorem 2. Let ℓ and N with $\ell \le N/3$ be given. Let $M = \lfloor N/3 \rfloor$, and set $N' = N - M + 1 \ge 2N/3$. We take for $\mathcal{L}$ the set system of all ℓ-element subsets of $[N']$. Then, clearly, $\mathcal{L}^{\triangle} = \{L \triangle L'\colon L,\, L' \in \mathcal{L},\ L \ne L'\}$ is the family of non-empty subsets of $[N']$ of even cardinality not greater than 2ℓ. Hence, $K = \{|S|\colon S \in \mathcal{L}^{\triangle}\} = \{2, 4, \ldots, 2\ell\}$. Moreover,

$$|\mathcal{L}| = \binom{N'}{\ell} \ge N' \ge \frac{2N}{3} \ge 2M, \tag{40}$$

and, as $M = \lfloor N/3 \rfloor \geq N/5$ because $N \geq 3$, we have

$$|\mathcal{L}| \leq 2^N = (2^{N/M})^M \leq 2^{5M} < \mathrm{e}^{50M}. \tag{41}$$

Inequalities (40) and (41) tell us that Lemma 9 may be applied. We deduce from that lemma that

$$\max\{C_2(E_N), C_4(E_N), \dots, C_{2\ell}(E_N)\} \geq \min\left\{\frac{1}{2}M, \sqrt{\frac{1}{50}M(\log|\mathcal{L}|) \Big/ \log\frac{50M}{\log|\mathcal{L}|}}\right\}. \tag{42}$$

If the minimum on the right-hand side of (42) is achieved by $M/2 = \lfloor N/3 \rfloor/2$, then we are already done; suppose therefore that the minimum is given by the other term. Observe that

$$\frac{1}{50}M(\log|\mathcal{L}|) \Big/ \log\frac{50M}{\log|\mathcal{L}|} \geq \frac{1}{50}\left\lfloor\frac{N}{3}\right\rfloor(\log|\mathcal{L}|) \Big/ \log\frac{50N/3}{\log|\mathcal{L}|}, \tag{43}$$

and, moreover,

$$|\mathcal{L}| = \binom{N'}{\ell} \geq \left(\frac{2N}{3\ell}\right)^{\ell}, \tag{44}$$

so that

$$\log|\mathcal{L}| \geq \ell\log\frac{2N}{3\ell}. \tag{45}$$

By (43) and (45), it suffices to show that

$$\frac{1}{150}N\ell\left(\log\frac{2N}{3\ell}\right) \Big/ \log\frac{50N/3}{\ell\log(2N/3\ell)} \geq c'N\ell \tag{46}$$

for some absolute constant $c' > 0$. Routine calculations show that a suitable constant $c' > 0$ will do in (46). We only give a sketch: suppose first that $1 \leq \ell = o(N)$. In this case, it is simple to check that the left-hand side of (46) is in fact

$$\left(\frac{1}{150} + o(1)\right)N\ell. \tag{47}$$

Suppose now that $c''N \leq \ell \leq N/3$. In this case, the left-hand side of (46) is at least

$$\frac{1}{150}N\ell(\log 2) \Big/ \log\frac{50/3}{c''\log 2}, \tag{48}$$

and (46) follows for some small enough $c' > 0$. $\square$

2.3. **Some further lower bounds for correlation.** In this section, we deduce some further consequences of Lemmas 8 and 9, using other families $\mathcal{L}$.

2.3.1. *Projective plane bounds.* We shall prove Theorem 3 (see Section 1.1) by making use of systems of sets derived from projective planes. Recall that Theorem 3 tells us that, for any $2 \leq k \leq \sqrt{N/6}$ and any $E_N \in \{-1,1\}^N$, at least one of $C_{2k-2}(E_N)$ and $C_{2k}(E_N)$ is $\geq c\sqrt{N}$, for some absolute constant $c > 0$. (We shall not try to obtain the best value of c in what follows.) We shall use the following fact.

Lemma 10. *Let positive integers k and n with $k \leq (1/2)\sqrt{n}$ be given. If n is large enough, then there is a family $\mathcal{L}$ of k-element subsets of $[n]$ with $|\mathcal{L}| = n$ and such that $|L \cap L'| \leq 1$ for all distinct L and $L' \in \mathcal{L}$.*

One may prove Lemma 10 by considering suitable projective planes on m points, with m only slightly larger than n: one may first delete $m - n$ points from the plane at random, to obtain a system with n points and $\geq n$ 'lines' of cardinality only slightly smaller than $\sqrt{n}$, and then one may remove some points from these 'lines' to turn them into k-element sets. (The constant $1/2$ in the upper bound for k in Lemma 10 may in fact be replaced by any constant < 1.)

Proof of Theorem 3. Let k and N as in the statement of the theorem be given. Let $M = \lfloor N/3 \rfloor$ and

$$N' = N - M + 1 \geq \frac{2}{3}N \geq 2M. \tag{49}$$

Observe that $k \leq \sqrt{N/6} = (1/2)\sqrt{2N/3} \leq (1/2)\sqrt{N'}$. We now use that N is supposed to be large and invoke Lemma 10, to obtain a family $\mathcal{L}$ of k-element subsets of $[N']$ with $|\mathcal{L}| = N'$ and $|L \cap L'| \leq 1$ for any two distinct L and $L' \in \mathcal{L}$.

By (49), we have

$$M - \frac{M^2}{|\mathcal{L}|} \geq \frac{1}{2}M = \frac{1}{2}\left\lfloor \frac{N}{3} \right\rfloor. \tag{50}$$

Moreover, $|L \triangle L'| \in \{2k-2, 2k\}$ for all distinct L and $L' \in \mathcal{L}$. Inequality (10) follows from (28). $\square$

If a projective plane of order k exists, then one may give a lower bound of order $\sqrt{N}$ for $C_{2k}(E_N)$.

Theorem 11. *For any constant $1/\sqrt{2} < \alpha < 1$, there is a constant $c = c(\alpha) > 0$ for which the following holds. Given any $\varepsilon > 0$, there is N_0 such that if $N \geq N_0$ and k is a power of a prime and $|k - \alpha\sqrt{N}| \leq \varepsilon\sqrt{N}$, then*

$$C_{2k}(E_N) \geq c\sqrt{N} \tag{51}$$

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

Proof. We only give a sketch of the proof. Let k be a large prime power as in the statement of our result, and set

$$N' = k^2 + k + 1 \quad \text{and} \quad M = N - N' + 1. \tag{52}$$

Using that $k = (\alpha + o(1))\sqrt{N}$, we have

$$N' = (\alpha^2 + o(1))N \quad \text{and} \quad M = (1 - \alpha^2 + o(1))N. \tag{53}$$

We now use that k is a prime power, and let $\mathcal{L}$ be the family of lines of a projective plane with point set $[N']$. Clearly, every member of $\mathcal{L}$ has $k + 1$ elements and

$$|\mathcal{L}| = N' = (\alpha^2 + o(1))N. \tag{54}$$

We shall now apply Lemma 8. By (53), we have

$$\begin{aligned} M - \frac{M^2}{|\mathcal{L}|} &= (1 - \alpha^2 + o(1))N - \frac{(1 - \alpha^2 + o(1))^2 N^2}{(\alpha^2 + o(1))N} \\ &= \left(1 - \frac{1 - \alpha^2 + o(1)}{\alpha^2 + o(1)}\right)(1 - \alpha^2 + o(1))N \\ &= (1 + o(1))\left(2 - \frac{1}{\alpha^2}\right)(1 - \alpha^2)N. \end{aligned} \tag{55}$$

Clearly, $|L \triangle L'| = 2k$ for all distinct L and $L' \in \mathcal{L}$. Therefore, inequality (28) in Lemma 8, together with the hypothesis that $1/\sqrt{2} < \alpha < 1$, imply the desired result. $\square$

The proof of Theorem 11 above is based on Lemma 8; one may use Lemma 9 instead, which would give a somewhat different value for the constant c in (51). A bound of the form (51) for k of order N may also be proved in the case in which there exists a $4k \times 4k$ Hadamard matrix. Indeed, it suffices to consider such a matrix as the incidence matrix of a system $\mathcal{L}$ of $2k$-element subsets of a $4k$-element set; the system $\mathcal{L}$ would then have the property that all pairwise symmetric differences of its members are of cardinality $2k$.

The condition that k should be a power of a prime in Theorem 11 may be removed by making use of Vinogradov's three primes theorem (to be more precise, we use a strengthening of that result). The key observation is the following.

Lemma 12. *For any $\varepsilon > 0$, there is an integer k_0 for which the following holds. If $k \geq k_0$ is an odd integer, then there there is a family $\mathcal{L}$ of $(k+3)$-element subsets of $[n]$, where $|n - k^2/3| \leq \varepsilon k^2$, such that $\big||\mathcal{L}| - n/3\big| \leq \varepsilon n$ and*

$$|L \triangle L'| = 2k \tag{56}$$

for all distinct L and $L' \in \mathcal{L}$. If $k \geq k_0$ is even, then there is a family $\mathcal{L}$ of $(k+4)$-element subsets of $[n]$, where $|n - k^2/4| \leq \varepsilon k^2$, such that $\big||\mathcal{L}| - n/4\big| \leq \varepsilon n$ and (56) holds for all distinct L and $L' \in \mathcal{L}$.

Proof. We give a sketch of the proof. Let $\varepsilon > 0$ be fixed and suppose first that k is a large odd integer.

We use a strengthening of Vinogradov's theorem, according to which any large enough odd integer k may be written as a sum of three primes p_1,

p_2, and p_3 that satisfy $p_i = (1/3 + o(1))k$, where $o(1) \to 0$ as $k \to \infty$ (an old theorem of Haselgrove [8] implies this result). Let $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ be projective planes of order p_1, p_2, and p_3, respectively, and suppose that $p_1 \le p_2$ and p_3. We take the $\mathcal{L}_i$ on pairwise disjoint point sets X_i and let $X = X_1 \cup X_2 \cup X_3$. Clearly, $n = |X| = 3(1/3 + o(1))^2 k^2 = (1/3 + o(1))k^2$. Let the lines of $\mathcal{L}_i$ be $L_1^{(i)}, \cdots, L_{n_i}^{(i)}$, where $n_i = p_i^2 + p_i + 1 = (1/3 + o(1))^2 k^2 = (1/3 + o(1))n$. We let $\mathcal{L}$ be the set system on X given by

$$\mathcal{L} = \left\{ L_j^{(1)} \cup L_j^{(2)} \cup L_j^{(3)} : 1 \le j \le n_1 \right\}. \tag{57}$$

The members of $\mathcal{L}$ are therefore $(k+3)$-element subsets of X, with $|L \triangle L'| = 2(p_1 + p_2 + p_3) = 2k$ for all distinct L and $L' \in \mathcal{L}$, and the case in which k is a large odd integer follows.

For even k, it suffices to let $p_4 = (1/4 + o(1))k$ be an odd prime (whose existence follows from the prime number theorem) and apply Haselgrove's result to $k - p_4$, and then construct $\mathcal{L}$ as the union of 4 suitable projective planes. We omit the details. □

Lemmas 8 and 12 imply the following result.

Theorem 13. *For all $\varepsilon > 0$, there are constants $c > 0$, k_0, and N_0 for which the following hold.*

(*i*) *If $k \ge k_0$ is an odd integer with*

$$\left(\frac{3}{2} + \varepsilon\right)\sqrt{N} \le k \le \left(\sqrt{3} - \varepsilon\right)\sqrt{N}, \tag{58}$$

then $C_{2k}(E_N) \ge c\sqrt{N}$ for all $E_N \in \{-1, 1\}^N$ as long as $N \ge N_0$.

(*ii*) *If $k \ge k_0$ is an even integer with*

$$\left(\frac{4}{5}\sqrt{5} + \varepsilon\right)\sqrt{N} \le k \le (2 - \varepsilon)\sqrt{N}, \tag{59}$$

then $C_{2k}(E_N) \ge c\sqrt{N}$ for all $E_N \in \{-1, 1\}^N$ as long as $N \ge N_0$.

We omit the proof of Theorem 13. We only remark that it suffices to take for $\mathcal{L}$ in Lemma 8 the systems given by Lemma 12. One may prove results similar to Theorem 13 for other ranges of k of order $\sqrt{N}$ using the method above: one simply proves variants of Lemma 12 by writing k as the sum of h nearly equal primes, for other values of h.

We close by making the following remark. In the discussion above, we have used the family of lines in projective planes; it is easy to check that one may also use hyperplanes in projective d-spaces for other values of d, to obtain lower bounds for $C_{2k}(E_N)$ for any k of order $N^{1-1/d}$, in certain ranges (as in Theorem 13). Furthermore, for any k of order N, again in certain ranges, we may use Hadamard matrices arising from quadratic residues modulo primes to prove lower bounds of order $\sqrt{N}$ for $C_{2k}(E_N)$. We omit the details.

2.3.2. *A variant of Theorem 2.* In this section, we shall prove a result similar in nature to Theorem 2.

Theorem 14. *There is an absolute constant $c > 0$ for which the following holds. For any positive integers ℓ and N with $\ell \leq N/25$ and N large enough, we have*

$$\max\{C_{2\ell+2}(E_N), C_{2\ell+4}(E_N), \ldots, C_{4\ell}(E_N)\} \geq c\sqrt{\ell N} \tag{60}$$

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

The proof of Theorem 14 is based on the following lemma.

Lemma 15. *Let $1 \leq \ell \leq n/9\mathrm{e}$. Then there is a system $\mathcal{L}$ of 2ℓ-element subsets of $[n]$ with*

$$|\mathcal{L}| \geq \frac{1}{2}\binom{n/9\mathrm{e}}{\ell} \tag{61}$$

and

$$|L \triangle L'| \geq 2\ell + 2 \tag{62}$$

for all distinct L and $L' \in \mathcal{L}$.

Proof. We give a sketch of the proof. Comparing with a geometric series, one may check that, say,

$$\sum_{\ell \leq j \leq 2\ell} \binom{2\ell}{j}\binom{n-2\ell}{2\ell-j} \leq 2\binom{2\ell}{\ell}\binom{n-2\ell}{\ell}. \tag{63}$$

Let $\mathcal{L}$ be a maximal family of 2ℓ-element subsets of $[n]$, with any two of its members satisfying (62) for all distinct L and $L' \in \mathcal{L}$. Then, clearly,

$$2\binom{2\ell}{\ell}\binom{n-2\ell}{\ell}|\mathcal{L}| \geq |\mathcal{L}| \sum_{\ell \leq j \leq 2\ell} \binom{2\ell}{j}\binom{n-2\ell}{2\ell-j} \geq \binom{n}{2\ell}. \tag{64}$$

Therefore,

$$\begin{aligned} |\mathcal{L}| &\geq \frac{1}{2}\binom{n}{2\ell} \Big/ \binom{2\ell}{\ell}\binom{n-2\ell}{\ell} = \frac{(n)_\ell (n-\ell)_\ell (\ell!)^2}{2(2\ell)_\ell (n-2\ell)_\ell (2\ell)!} \\ &\geq \frac{(n-\ell)_\ell}{2(2\ell)_\ell 4^\ell} \geq \frac{1}{2}\left(\frac{n-\ell}{8\ell}\right)^\ell \geq \frac{1}{2}\left(\frac{n}{9\ell}\right)^\ell \geq \frac{1}{2}\binom{n/9\mathrm{e}}{\ell}, \end{aligned} \tag{65}$$

as required. ∎

Proof of Theorem 14. This follows from Lemmas 9 and 15; we shall only give a sketch of the proof, because the argument is simple and very similar to the argument given in the proof of Theorem 2. Let ℓ and N be as given in the statement of Theorem 14. The case in which $\ell = 1$ is covered by Theorem 1 (in fact, the case in which ℓ is bounded follows from that result). Therefore, we suppose $\ell \geq 2$. Let $M = \lfloor N/50 \rfloor$ and $N' = N - M + 1$. Let $\mathcal{L}$ be a family of 2ℓ-element subsets of $[N']$ of maximal cardinality satisfying (62) for all distinct L and $L' \in \mathcal{L}$. By Lemma 15, we have

$$2M \leq \frac{1}{2}\left(\frac{N'/9\mathrm{e}}{2}\right)^2 \leq \frac{1}{2}\binom{N'/9\mathrm{e}}{\ell} \leq |\mathcal{L}| \leq 2^{N'} < \mathrm{e}^{50M} \tag{66}$$

for all large enough N. Therefore, Lemma 9 applies and we deduce that, for all $E_N \in \{-1,1\}^N$, we have

$$\begin{aligned}\max\{C_{2\ell+2}(E_N), C_{2\ell+4}(E_N), \dots, C_{4\ell}(E_N)\}\\ \geq \min\left\{\frac{1}{2}M, \sqrt{\frac{1}{50}M(\log|\mathcal{L}|) \Big/ \log\frac{50M}{\log|\mathcal{L}|}}\right\}.\end{aligned} \tag{67}$$

If the minimum on the right-hand side of (67) is achieved by $M/2$, we are done. In the other case, we may check that (60) follows for a suitable absolute constant $c > 0$ by, say, analysing the cases $1 \leq \ell = o(N)$ and $\ell \geq c'N$ separately (see the proof of Theorem 2). $\square$

2.4. **Bounds from coding theory.** We observe that one may prove lower bounds for the parameter $C_k(E_N)$ by invoking upper bounds for the size of codes with a given minimum distance (bounds in the range that we are interested in are given in [9, p. 565] (see also [12])). For simplicity, let us take the case in which $k = 2$. A sequence with $C_2(E_N)$ small gives rise to a large number of nearly orthogonal $\{-1,1\}$-vectors of a given length: it suffices to consider all the $N - M + 1$ segments of E_N of length M, where we take $M = (\alpha + o(1))N$ for a suitable positive constant α. From the fact that $C_2(E_N)$ is small, we may deduce that these $N - M + 1$ vectors are pairwise nearly orthogonal. Therefore, these binary vectors have pairwise Hamming distance at least $M/2 - \Delta$, for some small $\Delta > 0$. On the other hand, bounds from the theory of error correcting codes give us lower bounds for Δ, because we have a family of $N - M + 1$ such vectors. The bounds one deduces with this approach are somewhat weaker than the bounds obtained above.

However, we mention that the argument above applies in a more general setting. For $E_N \in \{-1,1\}^N$, let

$$\widetilde{C}_k(E_N) = \max\{V(E_N, M, D)\colon M \text{ and } D \text{ with } M - 1 + d_k \leq N\}, \tag{68}$$

where $D = \{d_1, \dots, d_k\}$ is as in Section 1; the only difference between $C_k(E_N)$ and $\widetilde{C}_k(E_N)$ is that, in the definition of $\widetilde{C}_k(E_N)$, we do not take $V(E_N, M, D)$ in absolute value (cf. (4) and (68)). Clearly, $\widetilde{C}_k(E_N) \leq C_k(E_N)$. The argument from coding theory briefly sketched in the previous paragraph applies to $\widetilde{C}_k(E_N)$ as well.

3. The minimum of the normality measure

3.1. **Remarks on** $\min \mathcal{N}(E_N)$**.** We start with two observations on $\mathcal{N}(E_N)$. Put

$$\mathcal{N}_k(E_N) = \max_X \max_M \left| T(E_N, M, X) - \frac{M}{2^k} \right|, \tag{69}$$

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

Proposition 16. *(i) We have* $\min_{E_N} \mathcal{N}_k(E_N) = 1 - 2^{-k}$ *for any* $k \geq 1$ *and any* $N \geq 2^k$*. (ii) We have*

$$\min_{E_N} \mathcal{N}(E_N) \geq \left(\frac{1}{2} + o(1)\right) \log_2 N. \tag{70}$$

Proof. To prove (i), we simply consider powers of appropriate de Bruijn sequences [6]. 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 an appropriate number of times. The fact that $\mathcal{N}_k(E_N) \geq 1 - 2^{-k}$ for this sequence E_N may be seen by taking $M = 1$ in (69) with X the prefix of E_N of length k. We leave the other inequality for the reader.

Let us now prove (ii). If a sequence $E_N \in \{-1, 1\}^N$ contains no segment of length $k = \lfloor \log_2 N - \log_2 \log_2 N \rfloor$ of repeated 1s, then

$$\mathcal{N}_k(E_N) \geq \frac{N - k + 1}{2^k} = (1 + o(1)) \frac{N}{2^k} \geq (1 + o(1)) \log_2 N, \tag{71}$$

as required. Suppose now that $E_N = (e_i)_{1 \leq i \leq N}$ does contain such a segment, say, $(e_{M_0}, \ldots, e_{M_0+k-1}) = (1, \ldots, 1)$. Fix $\ell = \ell(N) \to \infty$ as $N \to \infty$ with $\ell = o(k)$, and let X_ℓ be the sequence of ℓ consecutive 1s. Let $M_1 = M_0 + k - \ell$, and note that then

$$\begin{aligned} &T(E_N, M_1, X_\ell) - T(E_N, M_0, X_\ell) \\ &\qquad = M_1 - M_0 + 1 = k - \ell + 1 = (1 + o(1))k. \end{aligned} \tag{72}$$

Therefore

$$\begin{aligned} &\left(T(E_N, M_1, X_\ell) - \frac{M_1}{2^\ell}\right) - \left(T(E_N, M_0, X_\ell) - \frac{M_0}{2^\ell}\right) \\ &\qquad = (1 + o(1))k - (M_1 - M_0)2^{-\ell} = (1 + o(1))k. \end{aligned} \tag{73}$$

It follows from (73) that for some $M_0 \leq M^* \leq M_1$ we have

$$\left| T(E_N, M^*, X_\ell) - \frac{M^*}{2^\ell} \right| \geq \left(\frac{1}{2} + o(1)\right) k = \left(\frac{1}{2} + o(1)\right) \log_2 N. \tag{74}$$

Therefore, $\mathcal{N}(E_N) \geq \mathcal{N}_\ell(E_N) \geq (1/2 + o(1)) \log_2 N$, as required. □

We suspect that the logarithmic lower bound in Proposition 16(ii) is far from the truth.

Problem 17. *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?

3.2. **A sequence E_N with small $\mathcal{N}(E_N)$.** Our aim in this section is to prove Theorem 4. We start by describing the construction of E_N.

Let s be a positive integer and let $\mathbb{F}_{2^s} = \mathrm{GF}(2^s)$ be the finite field with 2^s elements. Fix a primitive element $x \in \mathbb{F}_{2^s}^*$, and let $m = |\mathbb{F}_{2^s}^*| = 2^s - 1$. We consider $\mathbb{F}_{2^s}$ as a vector space over $\mathbb{F}_2$, and fix a non-zero linear functional

$$b\colon \mathbb{F}_{2^s} \to \mathbb{F}_2. \tag{75}$$

We now let

$$\widetilde{E}_m = (b(x), b(x^2), \dots, b(x^m)) \in \mathbb{F}_2^m = \{0,1\}^m \tag{76}$$

and let

$$E_m = ((-1)^{b(x)}, (-1)^{b(x^2)}, \dots, (-1)^{b(x^m)}) \in \{-1,1\}^m. \tag{77}$$

Finally, set

$$E_N = E_m^q = E_m \dots E_m \qquad (q \text{ factors}), \tag{78}$$

where E_m^q denotes the concatenation of q copies of E_m; clearly, E_N has length $N = qm$.

Theorem 18. *Let $s \geq 2$. With E_N as defined in (78), we have*

$$\mathcal{N}(E_N) \leq q + 2(\log_2(m-1))\sqrt{m}. \tag{79}$$

Theorem 4 will be deduced from Theorem 18 in Section 3.2.2 below. Let us now give a rough outline of the proof of Theorem 18. Essentially all the work will concern the sequence E_m defined above.

In what follows, we shall first prove that any reasonably long segment of E_m has small 'discrepancy'; we shall show that the entries of segments of E_m of length k add up to $O\big((\log k)\sqrt{m}\big)$ (see Corollary 22). We shall then show two results concerning the number of occurrences of (short) words in E_m. We shall first show that all the words of length $k \leq s$ (except for the word $(0,\dots,0)$) occur exactly the same number of times in E_m (see Lemma 23). We shall then prove a similar fact for *segments* of E_m, although for segments the conclusion will be weaker (see Lemma 25). Theorem 18 will then be deduced from these facts in Section 3.2.2.

3.2.1. *Auxiliary lemmas.* We start with a well known lemma concerning the 'discrepancy' of matrices whose rows have uniformly bounded norm and, pairwise, have non-positive inner product (see, e.g., [7, Theorem 15.2] for a similar statement).

Lemma 19. *Let $H = (h_{ij})_{1\leq i,j\leq M}$ be an M by M real matrix and let $\mathbf{v}_i$ be the ith row of H $(1 \leq i \leq M)$. Let A, $B \subset [M]$ be given, and suppose that*

$$\|\mathbf{v}_a\| \leq \sqrt{m} \tag{80}$$

for all $a \in A$ and

$$\langle \mathbf{v}_a, \mathbf{v}_{a'} \rangle = \sum_{1 \le b \le M} h_{ab} h_{a'b} \le 0 \tag{81}$$

for all $a \ne a'$ with a, $a' \in A$. Then

$$\left| \sum_{a \in A,\, b \in B} h_{ab} \right| \le \sqrt{m|A||B|}. \tag{82}$$

Proof. Let $\mathbf{1}_B \in \{0,1\}^M$ be the characteristic vector of B. By the Cauchy–Schwarz inequality, we have

$$\left| \sum_{A,\,B} h_{ab} \right| = \left| \left\langle \sum_{a \in A} \mathbf{v}_a, \mathbf{1}_B \right\rangle \right| \le \left\| \sum_{a \in A} \mathbf{v}_a \right\| \sqrt{|B|}. \tag{83}$$

From (80) and (81), we have

$$\left\| \sum_{a \in A} \mathbf{v}_a \right\|^2 = \sum_{a \in A} \|\mathbf{v}_a\|^2 + \sum_{a \in A} \sum_{a \ne a' \in A} \langle \mathbf{v}_a, \mathbf{v}_{a'} \rangle \le m|A|. \tag{84}$$

Plugging (84) into (83), we have

$$\left| \sum_{A,\,B} h_{ab} \right| \le \sqrt{m|A||B|},$$

as required. □

We now define a matrix $\mathbf{E}$ from E_m; we shall apply Lemma 19 to $\mathbf{E}$ to deduce the discrepancy property we seek for E_m. Let

$$\mathbf{E} = (E_{ij})_{1 \le i,j \le m} = \begin{bmatrix} (-1)^{b(x)} & (-1)^{b(x^2)} & \dots & (-1)^{b(x^m)} \\ (-1)^{b(x^2)} & (-1)^{b(x^3)} & \dots & (-1)^{b(x)} \\ \vdots & \vdots & \ddots & \vdots \\ (-1)^{b(x^m)} & (-1)^{b(x)} & \dots & (-1)^{b(x^{m-1})} \end{bmatrix}. \tag{85}$$

Note that $\mathbf{E}$ is an $m \times m$ circulant, symmetric $\{-1,1\}$-matrix whose first row is E_m. For convenience, let $\mathbf{e}_i = (E_{ij})_{1 \le j \le m}$ $(1 \le i \le m)$ denote the ith row of $\mathbf{E}$. Moreover, if $\mathbf{v} = (v_j)_{1 \le j \le m}$ and $\mathbf{w} = (w_j)_{1 \le j \le m}$ are two real m-vectors, let $\mathbf{v} \circ \mathbf{w}$ denote the m-vector $(v_j w_j)_{1 \le j \le m}$.

Lemma 20. *The following hold for $\mathbf{E}$:*

- (*i*) *Every row of $\mathbf{E}$ adds up to -1, that is, $\sum_{1 \le j \le m} E_{ij} = -1$ for all $1 \le i \le m$.*
- (*ii*) *For all $i \ne i'$ $(1 \le i, i' \le m)$, we have $\mathbf{e}_i \circ \mathbf{e}_{i'} = \mathbf{e}_{i''}$ for some $1 \le i'' \le m$.*
- (*iii*) *The matrix $\mathbf{E}$ satisfies*

$$\mathbf{E}\mathbf{E}^T = -\mathbf{J} + (m+1)\mathbf{I}, \tag{86}$$

where $\mathbf{J}$ is the $m \times m$ matrix with all entries 1 and $\mathbf{I}$ is the $m \times m$ identity matrix.

(*iv*) *For all A and $B \subset [m]$, we have*

$$\left|\sum_{a\in A,\, b\in B} E_{ab}\right| \le \sqrt{m|A||B|}. \tag{87}$$

Proof. Since $b\colon \mathbb{F}_{2^s} \to \mathbb{F}_2$ is a non-zero linear functional, $b^{-1}(0)$ is a hyperplane in $\mathbb{F}_{2^s}$ and hence has cardinality 2^{s-1}. Given that $\mathbb{F}_{2^s}^* = \{x^j\colon 1 \le j \le m = 2^s - 1\}$, we conclude that $b(x^j) = 1$ for 2^{s-1} values of j with $1 \le j \le m$ and $b(x^j) = 0$ for all the other $2^{s-1}-1$ values of j with $1 \le j \le m$. Therefore, statement (i) follows. Let now $1 \le i < i' \le m$ be fixed. Then

$$\begin{aligned}
&\mathbf{e}_i \circ \mathbf{e}_{i'}\\
&\quad = ((-1)^{b(x^i)+b(x^{i'})}, (-1)^{b(x^{i+1})+b(x^{i'+1})}, \dots, (-1)^{b(x^{i-1})+b(x^{i'-1})})\\
&\quad = ((-1)^{b(x^i+x^{i'})}, (-1)^{b(x^{i+1}+x^{i'+1})}, \dots, (-1)^{b(x^{i-1}+x^{i'-1})})\\
&\quad = ((-1)^{b(x^i(1+x^{i'-i}))}, (-1)^{b(x^{i+1}(1+x^{i'-i}))}, \dots, (-1)^{b(x^{i-1}(1+x^{i'-i}))}).
\end{aligned} \tag{88}$$

However, as $0 < i' - i < m$, we have $1 + x^{i'-i} \ne 0$, and hence $1 + x^{i-i'} = x^k$ for some $1 \le k \le m$. Therefore, we have from (88) that

$$\mathbf{e}_i \circ \mathbf{e}_{i'} = ((-1)^{b(x^{i+k})}, (-1)^{b(x^{i+1+k})}, \dots, (-1)^{b(x^{i-1+k})}) = \mathbf{e}_{i+k} \tag{89}$$

(naturally, the index of $\mathbf{e}_{i+k}$ is modulo m). Equation (89) proves (ii).

Equation (86) is an immediate consequence of (i) and (ii), and hence (iii) is clear. Finally, for (iv), it suffices to notice that $\|\mathbf{e}_i\| = \sqrt{m}$ for all i and that, from the above discussion, $\langle \mathbf{e}_i, \mathbf{e}_{i'}\rangle = -1 < 0$ for all $i \ne i'$. Therefore, Lemma 19 applies and (87) follows. □

Lemma 20(iv) tells us that 'rectangles' in the matrix $\mathbf{E}$ have small discrepancy (in the sense of (87)). We shall now deduce a similar result for 'triangles' in $\mathbf{E}$, which will later be used to show that *segments* of E_m have small discrepancy.

Lemma 21. *Let A and $B \subset [m]$ be given and suppose $A = \{a_1, \dots, a_t\}$, $B = \{b_1, \dots, b_t\}$, where $a_1 < \dots < a_t$ and $b_1 < \dots < b_t$. The following assertions hold for the matrix $\mathbf{E} = (E_{ij})_{1\le i,j\le m}$.*

(*i*) *We have*

$$\left|\sum_{i+j\le t+1} E_{a_i b_j}\right| \le (t\log_2 t + 1)\sqrt{m}. \tag{90}$$

(*ii*) *Similarly,*

$$\left|\sum_{i+j\ge t+1} E_{a_i b_j}\right| \le (t\log_2 t + 1)\sqrt{m}. \tag{91}$$

Proof. Inequality (90) follows from Lemma 20(iv), by induction on t. Note first that (90) holds for $t=1$. Now suppose that $t>1$ and that (90) holds for smaller values of t. By the triangle inequality, we have

$$\left|\sum_{i+j\le t+1} E_{a_ib_j}\right| \le \left|\sum_{(i,j)\in S} E_{a_ib_j}\right| + \left|\sum_{(i,j)\in T_1} E_{a_ib_j}\right| + \left|\sum_{(i,j)\in T_2} E_{a_ib_j}\right|, \tag{92}$$

where $S=\{(i,j)\colon i,j\le\lceil t/2\rceil\}$, $T_1=\{(i,j)\colon i\le\lceil t/2\rceil,\ j>\lceil t/2\rceil\}$, and $T_2=\{(i,j)\colon j\le\lceil t/2\rceil,\ i>\lceil t/2\rceil\}$. We now estimate the three terms on the right-hand side of (92) by using (87) and the induction hypothesis twice. We have

$$\begin{aligned}\left|\sum_{i+j\le t+1} E_{a_ib_j}\right| &\le \left\lceil\frac{t}{2}\right\rceil\sqrt{m} + 2\left(\left\lfloor\frac{t}{2}\right\rfloor\log_2\left\lfloor\frac{t}{2}\right\rfloor+1\right)\sqrt{m}\\ &\le \left\lceil\frac{t}{2}\right\rceil\sqrt{m} + (t(\log_2 t-1)+2)\sqrt{m}\\ &\le (t\log_2 t+1)\sqrt{m} + \left\lceil\frac{t}{2}\right\rceil\sqrt{m} - (t-1)\sqrt{m}\\ &\le (t\log_2 t+1)\sqrt{m}, \end{aligned}\tag{93}$$

which completes the induction step, and (i) is proved. The proof of assertion (ii) is similar, and hence it is omitted. □

We shall now show that segments of E_m have small discrepancy, in the sense that they have the same number of 1s as -1s, up to a small error. We observe that Corollary 22 also considers segments of E_m that “wrap around” the end of E_m; equivalently, that result considers E_m as a circular sequence.

Corollary 22. *For any $1\le r\le m$ and $2\le k\le m$, we have*

$$\left|\sum_{0\le i<k}(-1)^{b(x^{r+i})}\right| \le \left(\log_2 k + \left(1-\frac{1}{k}\right)\log_2(k-1)+\frac{2}{k}\right)\sqrt{m}. \tag{94}$$

In particular, for all $1\le r\le m$ and $2\le k\le m$, we have

$$\left|\sum_{0\le i<k}(-1)^{b(x^{r+i})}\right| \le 2(\log_2 k)\sqrt{m}. \tag{95}$$

Proof. Note that

$$\left(1-\frac{1}{k}\right)\log_2(k-1)+\frac{2}{k}\le \log_2 k \tag{96}$$

if and only if

$$(k-1)^{1-1/k}2^{2/k}\le k \tag{97}$$

if and only if

$$\left(1-\frac{1}{k}\right)^k \le \frac{1}{4}(k-1), \tag{98}$$

$$
\begin{array}{ccccccc}
 & & & E_{1,r} & E_{1,r+1} & \dots & E_{1,r+k-1} \\
 & & E_{2,r-1} & E_{2,r} & & \iddots & \\
 & \iddots & \vdots & \vdots & \iddots & & \\
E_{k,r-k+1} & \dots & E_{k,r-1} & E_{k,r} & & &
\end{array}
$$

FIGURE 1. Portion of the matrix $\mathbf{E}$ to which Lemma 21(i) and (ii) are applied. Note that $E_{1,r} = E_{2,r-1} = \cdots = E_{k,r-k+1} = (-1)^{b(x^r)}$, $E_{1,r+1} = E_{2,r} = \cdots = E_{k,r-k+2} = (-1)^{b(x^{r+1})}$, etc.

which holds if $k \geq 3$. Therefore, (95) follows directly from (94) for $3 \leq k \leq m$. If $k = 2$, then (95) holds by inspection. To prove (94), we apply Lemma 21(i) and (ii). For the application of (i), we consider the sets $A = \{1, 2, \dots, k\}$, and $B = \{r, r+1, \dots, r+k-1\}$, whereas for the application of (ii) we consider $A' = \{2, 3, \dots, k\}$ and $B' = \{r-k+1, r-k+2, \dots, r-1\}$. Taking into account that $\mathbf{E} = (E_{ij})$ is circulant, we deduce that

$$
\begin{aligned}
k \sum_{0 \leq i < k} (-1)^{b(x^{r+i})} = \sum \{E_{ab} \colon a \in A,\ b \in B,\ a + b \leq k + r\} \\
+ \sum \{E_{a'b'} \colon a' \in A',\ b' \in B',\ a' + b' \geq r + 1\} \qquad (99)
\end{aligned}
$$

(see Figure 1). Therefore, by the triangle inequality, we have

$$
\begin{aligned}
k \left| \sum_{0 \leq i < k} (-1)^{b(x^{r+i})} \right| \leq \left| \sum \{E_{ab} \colon a \in A,\ b \in B,\ a + b \leq k + r\} \right| \\
+ \left| \sum \{E_{a'b'} \colon a' \in A',\ b' \in B',\ a' + b' \geq r + 1\} \right| \\
\leq (k \log_2 k + 1)\sqrt{m} + ((k-1)\log_2(k-1) + 1)\sqrt{m}, \qquad (100)
\end{aligned}
$$

and (94) follows on dividing (100) by k. □

The next lemma states that the number of occurrences of shorter words in $\widetilde{E}_m$ is basically equal to the expectation of this number in the case of the random sequence of length m. To state this precisely, we introduce some notation. Let $1 \leq k \leq s$ be fixed. For all $1 \leq r \leq m$, let $\widetilde{E}_m^{(r)}$ denote the segment of $\widetilde{E}_m$ of length k starting at its rth letter, that is,

$$
\widetilde{E}_m^{(r)} = (b(x^r), b(x^{r+1}), \dots, b(x^{r+k-1})) \qquad (101)
$$

($\widetilde{E}_m$ is considered as a cyclic sequence). Now, for all $X \in \{0,1\}^k$, let $f_X = f_X(\widetilde{E}_m)$ denote the number of occurrences of X as a segment in $\widetilde{E}_m$, where we consider $\widetilde{E}_m$ as a cyclic sequence; that is,

$$f_X = \mathrm{card}\{r\colon 1 \le r \le m \text{ and } \widetilde{E}_m^{(r)} = X\}. \tag{102}$$

Lemma 23. *For all $1 \le k \le s$, we have*

$$f_X = f_X(\widetilde{E}_m) = \begin{cases} (m+1)2^{-k} - 1 = 2^{s-k} - 1 & \textit{if } X = (0,\dots,0) \in \{0,1\}^k \\ (m+1)2^{-k} = 2^{s-k} & \textit{otherwise.} \end{cases} \tag{103}$$

Proof. Let $1 \le r \le m$ and $\delta = (\delta_i)_{1\le i\le k} \in \{0,1\}^k$ be given. Note that

$$\langle \delta, \widetilde{E}_m^{(r)} \rangle = \sum_{1\le i\le k} \delta_i b(x^{r+i-1}) = b\left(x^r \sum_{1\le i\le k} \delta_i x^{i-1} \right). \tag{104}$$

We shall now use the fact that x does not satisfy a polynomial over $\mathbb{F}_2$ of degree less than s (indeed, if $p(x) = 0$ for a polynomial p over $\mathbb{F}_2$ of degree t, then a standard argument shows that $1, x, \dots, x^{t-1}$ spans $\mathbb{F}_{2^s}$ as a vector space over $\mathbb{F}_2$ and hence $\deg(p) = t \ge s$). We use this fact in (104): as $k \le s$, we see that $\sum_{1\le i\le k} \delta_i x^{i-1} \ne 0$ as long as $\delta \ne (0,\dots,0)$, and hence this sum is x^t for some $1 \le t \le m$ independent of r. Therefore, $\langle \delta, \widetilde{E}_m^{(r)} \rangle = b(x^{r+t})$, and we have

$$\sum_{1\le r\le m} (-1)^{\langle \delta, \widetilde{E}_m^{(r)} \rangle} = \sum_{1\le r\le m} (-1)^{b(x^{r+t})} = -1 \tag{105}$$

by Lemma 20(i), since we have in (105) above the sum of the entries of the $(t+1)$st row of $\mathbf{E}$. If $\delta = (0,\dots,0)$, then clearly the sum in (105) is m.

Let us now observe that the left-hand side of (105) may also be written as

$$\sum_X (-1)^{\langle \delta, X \rangle} f_X, \tag{106}$$

where the sum is over all $X \in \{0,1\}^k$. Therefore, we have established a system of 2^k linear equation for the f_X ($X \in \{0,1\}^k$):

$$\sum_{X\in\{0,1\}^k} (-1)^{\langle \delta, X \rangle} f_X = \begin{cases} m & \text{if } \delta = (0,\dots,0) \in \{0,1\}^k \\ -1 & \text{otherwise.} \end{cases} \tag{107}$$

The matrix associated to the system of equations (107) is the $2^k \times 2^k$ Hadamard matrix $\mathbf{H}_k = [(-1)^{\langle \delta, X \rangle}]_{\delta, X\in\{0,1\}^k}$. For convenience, let $\mathbf{f} = (f_X)_{X\in\{0,1\}^k}$ and let $\mathbf{g} = (g_\delta)_{\delta\in\{0,1\}^k}$, where $g_\delta = m$ if $\delta = (0,\dots,0)$ and $g_\delta = -1$ otherwise. Then (107) may be written as

$$\mathbf{H}_k \mathbf{f} = \mathbf{g}. \tag{108}$$

Now, since

$$\sum_{\delta\in\{0,1\}^k} (-1)^{\langle \delta, X \rangle} (-1)^{\langle \delta, Y \rangle} = \sum_{\delta\in\{0,1\}^k} (-1)^{\langle \delta, X \triangle Y \rangle} = 0 \tag{109}$$

if $X \neq Y$, we have

$$\mathbf{H}_k^T \mathbf{H}_k = 2^k \mathbf{I}, \tag{110}$$

where, naturally, $\mathbf{I}$ is the $2^k \times 2^k$ identity matrix. Therefore, from (108) and (110) we have

$$2^k \mathbf{f} = \mathbf{H}_k^T \mathbf{H}_k \mathbf{f} = \mathbf{H}_k^T \mathbf{g}. \tag{111}$$

The last product in (111) may be computed explicitly, and one obtains that

$$\mathbf{H}_k^T \mathbf{g} = \begin{bmatrix} m - 2^k + 1 \\ m+1 \\ \vdots \\ m+1 \end{bmatrix}, \tag{112}$$

where the entry $m - 2^k + 1$ corresponds to $X = (0, \dots, 0)$. Equation (103) now follows from (111) and (112). □

Setting $k = s$ in Lemma 23, we see that words of length s occur in $\widetilde{E}_m$ at most once. Since every occurrence of a word of length at least s gives us an occurrence of its prefix of length s, we conclude that words longer than s occur no more than once in $\widetilde{E}_m$. We thus have the following corollary to Lemma 23, to be used later in the proof of Theorem 18.

Corollary 24. *Suppose $\ell \geq s = \log_2(m+1)$. Any $Y \in \{0,1\}^\ell$ occurs at most once in $\widetilde{E}_m$, even considering $\widetilde{E}_m$ as cyclic sequence; that is,*

$$\mathrm{card}\{r\colon 1 \leq r \leq m \textit{ and } (b(x^r), \dots, b(x^{r+\ell-1})) = Y\} \leq 1. \tag{113}$$

As it turns out, not only has $\widetilde{E}_m$ the property that shorter words occur evenly in it (as shows Lemma 23), but $\widetilde{E}_m$ has this property on its longer *segments* (in a weaker sense): for $k \leq s = \log_2(m+1)$, every k-letter word $X \in \{0,1\}^k$ occurs roughly $n2^{-k}$ times in any segment of $\widetilde{E}_m$ of length n, as long as n is reasonably large.

To make the above statement precise, we introduce some notation. Let $1 \leq r \leq m$ and $1 \leq n \leq m$ be given. Let $\widetilde{E}_m^{(r,n)}$ be the segment of $\widetilde{E}_m$ of length n starting at the rth letter of $\widetilde{E}_m$, that is, set

$$\widetilde{E}_m^{(r,n)} = (b(x^r), b(x^{r+1}), \dots, b(x^{r+n-1})). \tag{114}$$

Now let $1 \leq k \leq s$. We shall be interested in the segments $\widetilde{E}_m^{(t,k)}$ of length k of $\widetilde{E}_m$, for $r \leq t < r+n$. For $X \in \{0,1\}^k$, set

$$f_X = f_X(\widetilde{E}_m^{(r,n)}) = \mathrm{card}\{t\colon r \leq t < r+n \text{ and } \widetilde{E}_m^{(t,k)} = X\}. \tag{115}$$

In what follows, we write $O_1(a)$ for any term b such that $|b| \leq a$. We are now ready to state our lemma on the frequency of words in segments of $\widetilde{E}_m$.

Lemma 25. *For any $1 \leq r \leq m$, $2 \leq n \leq m$, and $1 \leq k \leq s$, we have*

$$f_X = f_X(\widetilde{E}_m^{(r,n)}) = n2^{-k} + O_1\left(2(\log_2 n)\sqrt{m}\right) \tag{116}$$

for all $X \in \{0,1\}^k$.

The proof of Lemma 25 will be similar to the proof of Lemma 23, except that we shall now make use of Corollary 22, instead of using the fact that the sum of the entries of the whole sequence E_m is -1.

Proof of Lemma 25. Let $\delta = (\delta_i)_{1\le i\le k} \in \{0,1\}^k$ be fixed. As before, we have

$$\langle \delta, \widetilde{E}_m^{(t,k)} \rangle = \sum_{1\le i\le k} \delta_i b(x^{t+i-1}) = b\left(x^t \sum_{1\le i\le k} \delta_i x^{i-1} \right) = b(x^{t+u}), \tag{117}$$

for some $1 \le u \le m$ independent of t. Therefore, by Corollary 22,

$$\begin{aligned} \left| \sum_{X\in\{0,1\}^k} (-1)^{\langle \delta, X\rangle} f_X \right| &= \left| \sum_{r\le t<r+n} (-1)^{\langle \delta, \widetilde{E}_m^{(t,k)}\rangle} \right| \\ &= \left| \sum_{r\le t<r+n} (-1)^{b(x^{t+u})} \right| \le 2(\log_2 n)\sqrt{m}. \end{aligned} \tag{118}$$

As before, let $\mathbf{H}_k$ be the $2^k \times 2^k$ Hadamard matrix $[(-1)^{\langle \delta, X\rangle}]_{\delta, X\in\{0,1\}^k}$, and let $\mathbf{f} = (f_X)_{X\in\{0,1\}^k}$. If $\mathbf{g} = \mathbf{H}_k\mathbf{f}$ and $\mathbf{g} = (g_\delta)_{\delta\in\{0,1\}^k}$, then (118) implies that

$$g_\delta = \begin{cases} n & \text{if } \delta = (0,\dots,0) \\ O_1\left(2(\log_2 n)\sqrt{m}\right) & \text{otherwise.} \end{cases} \tag{119}$$

Using that $\mathbf{H}_k^T\mathbf{H}_k = 2^k\mathbf{I}$, we have

$$\mathbf{f} = 2^{-k}\mathbf{H}_k^T\mathbf{H}_k\mathbf{f} = 2^{-k}\mathbf{H}_k^T\mathbf{g}. \tag{120}$$

One may easily observe that the entries of $\mathbf{H}_k^T\mathbf{g}$ are all equal to

$$n + O_1\left(2^{k+1}(\log_2 n)\sqrt{m}\right). \tag{121}$$

The asserted conclusion (116) follows from (120) and (121). $\square$

3.2.2. *Proof of Theorems 4 and 18.* We shall prove Theorem 18 using Lemmas 23 and 25 and Corollary 24, whereas we shall deduce Theorem 4 from Theorem 18 by making a suitable choice for q and m in the construction of E_N. Let us start with the proof of Theorem 18.

Proof of Theorem 18. Let E_N be as defined in (78), and let $X \in \{-1,1\}^k$ with $1 \le k \le \log_2 N$ be given. Let $1 \le M \le N - k + 1$ and let us compute $T(E_N, M, X)$; our aim is to compare $T(E_N, M, X)$ and $M2^{-k}$.

We first suppose $k \le s$, so that we may apply Lemmas 23 and 25.

Let $M = \alpha m + \beta$, where α and β are integers with $0 \le \beta < m$. Clearly, $0 \le \alpha \le q$. We use the following notation below, for conciseness: if P is some property, then $[P] = 0$ if P is false and $[P] = 1$ if P is true.

By definition (1), we have $T(E_m, \beta, X) \le \beta$. Suppose for a moment that $\beta \ge 2$. Then, by Lemma 25 applied with $r = 1$ and $n = \beta \ge 2$, we have

$$\begin{aligned} T(E_m, \beta, X) &\le \beta 2^{-k} + 2(\log_2 \beta)\sqrt{m} \\ &\le \beta 2^{-k} + 2(\log_2 (m-1))\sqrt{m}. \end{aligned} \tag{122}$$

As $m = 2^s - 1 \ge 3$, the upper bound (122) for $T(E_m, \beta, X)$ does hold for $\beta = 0$ and $\beta = 1$ as well. Lemma 23 tells us that $T(E_m, m, X) \le (m+1)2^{-k} - [X = \mathbf{1}]$ (note that the 'exceptional' sequence in (103), which concerns $\widetilde{E}_m \in \{0,1\}^m$, is the zero sequence $\mathbf{0} \in \{0,1\}^k$, which translates to the all 1 sequence $\mathbf{1} \in \{-1,1\}^k$ when considering $E_m \in \{-1,1\}^m$). We conclude from this and (122) that

$$\begin{aligned} T(E_N, M, X) &= \alpha T(E_m, m, X) + T(E_m, \beta, X) \\ &\le \alpha(m2^{-k} + 2^{-k} - [X = \mathbf{1}]) + \beta 2^{-k} + 2(\log_2(m-1))\sqrt{m} \\ &= \alpha m 2^{-k} + \beta 2^{-k} + \alpha(2^{-k} - [X = \mathbf{1}]) + 2(\log_2(m-1))\sqrt{m} \\ &\le M2^{-k} + q + 2(\log_2(m-1))\sqrt{m}. \end{aligned} \tag{123}$$

Similarly, by Lemmas 23 and 25, we have

$$\begin{aligned} T(E_N, M, X) &= \alpha T(E_m, m, X) + T(E_m, \beta, X) \\ &\ge \alpha(m2^{-k} + 2^{-k} - [X = \mathbf{1}]) + \beta 2^{-k} - 2(\log_2(m-1))\sqrt{m} \\ &= \alpha m 2^{-k} + \beta 2^{-k} + \alpha(2^{-k} - [X = \mathbf{1}]) - 2(\log_2(m-1))\sqrt{m} \\ &\ge M2^{-k} - q - 2(\log_2(m-1))\sqrt{m}. \end{aligned} \tag{124}$$

From (123) and (124), we have

$$\left| T(E_N, M, X) - \frac{M}{2^k} \right| \le q + 2(\log_2(m-1))\sqrt{m}. \tag{125}$$

We have thus completed the analysis for the case in which $k \le s$. Suppose now that $k > s$. Recall that Corollary 24 tells us that, in this case, X occurs in E_m at most once, that is, $T(E_m, m, X) \le 1$ and hence $0 \le T(E_N, M, X) \le q$. Note also that

$$0 \le \frac{M}{2^k} \le \frac{N}{2^{s+1}} = \frac{N}{2(m+1)} < \frac{1}{2}q. \tag{126}$$

Therefore,

$$\left| T(E_N, M, X) - \frac{M}{2^k} \right| \le q. \tag{127}$$

Inequality (79) follows from (125) and (127). □

We shall now prove Theorem 4.

Proof of Theorem 4. Let an integer N be given. In what follows, we may suppose that N is suitably large for our inequalities to hold. We start by choosing an integer s so that $m = 2^s - 1$ satisfies

$$\frac{14}{17}\left(\frac{N}{\log_2 N}\right)^{2/3} \le m \le \frac{5}{3}\left(\frac{N}{\log_2 N}\right)^{2/3}. \tag{128}$$

We now let

$$q = \left\lfloor \frac{11}{9} N^{1/3}(\log_2 N)^{2/3} \right\rfloor, \tag{129}$$

set $N' = qm$, and consider $E_{N'} = E_m \dots E_m = E_m^q$. We have

$$N' = qm \ge \left\lfloor \frac{11}{9} N^{1/3}(\log_2 N)^{2/3} \right\rfloor \times \frac{14}{17}\left(\frac{N}{\log_2 N}\right)^{2/3} \ge N \tag{130}$$

for all large enough N. We let E_N be the prefix of $E_{N'}$ of length N.

We claim that E_N satisfies (12). Clearly, it suffices to show that $E_{N'}$ is such that $\mathcal{N}(E_{N'}) \le 3N^{1/3}(\log_2 N)^{2/3}$. To prove this last inequality, we simply show that the right-hand side of (79) is at most $3N^{1/3}(\log_2 N)^{2/3}$.

We have

$$\log_2(m-1) < \log_2\left(\frac{5}{3}\left(\frac{N}{\log_2 N}\right)^{2/3}\right) < \frac{2}{3}\log_2 N. \tag{131}$$

Moreover,

$$\sqrt{m} \le \left(\frac{5}{3}\left(\frac{N}{\log_2 N}\right)^{2/3}\right)^{1/2} < \frac{31}{24}\left(\frac{N}{\log_2 N}\right)^{1/3} \tag{132}$$

for all large enough N. Therefore,

$$\begin{aligned} q + 2(\log_2(m-1))\sqrt{m} &< \frac{11}{9}N^{1/3}(\log_2 N)^{2/3} + \frac{31}{18}(\log_2 N)\left(\frac{N}{\log_2 N}\right)^{1/3} \\ &< 3N^{1/3}(\log_2 N)^{2/3}, \end{aligned} \tag{133}$$

implying that the right-hand side of (79) is at most $3N^{1/3}(\log_2 N)^{2/3}$, as required. □

We close with a remark concerning some recent work of Carpi and de Luca [3], generalizing de Bruijn sequences [6]. Those authors have proved a number of interesting results on *uniform words*: words w such that for any two words u and v of the same length, the number of occurrences of u and v in w differ by at most 1. It would be interesting to see whether their constructions could be used to obtain words with small normality measure.

3.3. **Larger alphabets.** We now sketch a generalization of the construction in Section 3.2 to alphabets of cardinality larger than 2. As it turns out, the construction generalizes easily to alphabets of cardinality that are powers of primes.

Let s be a positive integer and q a power of a prime, and let $\mathbb{F}_{q^s} = \mathrm{GF}(q^s)$ be the finite field with q^s elements. Fix a primitive element $x \in \mathbb{F}_{q^s}^*$, and let $m = |\mathbb{F}_{q^s}^*| = q^s - 1$. We consider $\mathbb{F}_{q^s}$ as a vector space over $\mathbb{F}_q$, and fix a non-zero linear functional

$$b\colon \mathbb{F}_{q^s} \to \mathbb{F}_q. \tag{134}$$

Let $\psi\colon \mathbb{F}_q \to S^1 \subset \mathbb{C}$ be an additive character with $\mathrm{card}\{\psi(y)\colon y \in \mathbb{F}_q\} = q$ (that is, we take ψ injective), and put

$$\widetilde{E}_m = (b(x), b(x^2), \ldots, b(x^m)) \in \mathbb{F}_q^m \tag{135}$$

and

$$E_m = (\psi(b(x)), \psi(b(x^2)), \ldots, \psi(b(x^m))) \in (S^1)^m. \tag{136}$$

Finally, set

$$E_N = E_m^\ell = E_m \ldots E_m \qquad (\ell \text{ factors}), \tag{137}$$

where E_m^ℓ denotes the concatenation of ℓ copies of E_m; clearly, E_N has length $N = \ell m$. The sequence E_N, considered as a word over the q-letter alphabet

$$\Sigma_q = \{\psi(y)\colon y \in \mathbb{F}_q\}, \tag{138}$$

is such that

$$\mathcal{N}^{(q)}(E_N) = O\left(N^{1/3}(\log N)^{2/3}\right), \tag{139}$$

where

$$\mathcal{N}^{(q)}(E_N) = \max_k \max_X \max_M \left| T(E_N, M, X) - \frac{M}{q^k} \right|, \tag{140}$$

and the maxima are taken over all $1 \le k \le \log_q N$, $X \in \Sigma_q^k$, and $0 < M \le N + 1 - k$.

Let us sketch the proof of (139). This time, we let

$$\mathbf{E} = (E_{ij})_{1\le i,j\le m} = \begin{bmatrix} \psi(b(x)) & \psi(b(x^2)) & \ldots & \psi(b(x^m)) \\ \psi(b(x^2)) & \psi(b(x^3)) & \ldots & \psi(b(x)) \\ \vdots & \vdots & \ddots & \vdots \\ \psi(b(x^m)) & \psi(b(x)) & \ldots & \psi(b(x^{m-1})) \end{bmatrix}. \tag{141}$$

Then $\mathbf{E}$ is an $m \times m$ circulant, complex matrix whose first row is E_m. Again, let $\mathbf{e}_i = (E_{ij})_{1\le j\le m}$ $(1 \le i \le m)$ denote the ith row of $\mathbf{E}$. Moreover, if $\mathbf{v} = (v_j)_{1\le j\le m}$ and $\mathbf{w} = (w_j)_{1\le j\le m}$ are two complex m-vectors, let $\mathbf{v} \circ \mathbf{w}$ denote the m-vector $(v_j\overline{w_j})_{1\le j\le m}$, where $\overline{z}$ denotes the complex conjugate of $z \in \mathbb{C}$.

It turns out that Lemma 20 generalizes to the the matrix $\mathbf{E}$ defined in (141), in the following way.

Lemma 26. *The following hold for* $\mathbf{E}$:

(*i*) *Every row of* $\mathbf{E}$ *adds up to* -1, *that is,* $\sum_{1\le j\le m} E_{ij} = -1$ *for all* $1 \le i \le m$.

(*ii*) *For all* $i \ne i'$ $(1 \le i, i' \le m)$, *we have* $\mathbf{e}_i \circ \mathbf{e}_{i'} = \mathbf{e}_{i''}$ *for some* $1 \le i'' \le m$.

(*iii*) *The matrix* $\mathbf{E}$ *satisfies*

$$\mathbf{E}\mathbf{E}^* = -\mathbf{J} + (m+1)\mathbf{I}, \tag{142}$$

where $\mathbf{E}^*$ *is the adjoint of* $\mathbf{E}$.

(*iv*) *For all* A *and* $B \subset [m]$, *we have*

$$\left|\sum_{a\in A,\, b\in B} E_{ab}\right| \le \sqrt{m|A||B|}. \tag{143}$$

Lemma 26(*i*)–(*iii*) may be checked easily. For Lemma 26(*iv*), one observes that Lemma 19 may be generalized in a natural way to complex matrices, with exactly the same proof.

Lemma 27. *Let* $H = (h_{ij})_{1\le i,j\le M}$ *be an* M *by* M *complex matrix and let* $\mathbf{v}_i$ *be the* i*th row of* H $(1 \le i \le M)$. *Let* A, $B \subset [M]$ *be given, and suppose that*

$$\|\mathbf{v}_a\| = \sqrt{\sum_{1\le j\le m} |h_{aj}|^2} \le \sqrt{m} \tag{144}$$

for all $a \in A$ *and*

$$\langle \mathbf{v}_a, \mathbf{v}_{a'}\rangle = \sum_{1\le b\le m} h_{ab}\overline{h_{a'b}} \le 0 \tag{145}$$

for all $a \ne a'$ *with* a, $a' \in A$. *Then*

$$\left|\sum_{a\in A,\, b\in B} h_{ab}\right| \le \sqrt{m|A||B|}. \tag{146}$$

To prove Lemma 26(*iv*), one applies Lemma 27 to the matrix $\mathbf{E}$ given in (141). The remainder of the argument is as before, with some small changes. The $2^k \times 2^k$ Hadamard matrix $\mathbf{H}_k = [(-1)^{\langle\delta,X\rangle}]_{\delta,X\in\{0,1\}^k}$ that occurs later in the proof should be replaced by the $q^k \times q^k$ matrix $\mathbf{H}_k = [\psi(\langle\delta,X\rangle)]_{\delta,X}$, where δ and X vary over $\mathbb{F}_q^k$, which is a unitary matrix, up to a multiplicative constant: $\mathbf{H}_k\mathbf{H}_k^* = mI$. We omit the details.

3.4. **The Pólya–Vinogradov inequality.** Let p be a prime and let $\chi\colon \mathbb{F}_p = \mathbb{Z}/p\mathbb{Z} \to S^1 \subset \mathbb{C}$ be a multiplicative character, where, as usual, $\chi(0) = 0$. With the methods in Section 3.2.1 (and Lemma 27 above) one may easily prove the celebrated Pólya–Vinogradov inequality, in the following form.

Theorem 28. *For all integers* r *and* $2 \le k \le p$, *we have*

$$\left|\sum_{0\le h<k} \chi(r+h)\right| \le 2(\log_2 k)\sqrt{p-1}. \tag{147}$$

We give an outline of the proof of Theorem 28. This time, we let $\mathbf{E} = (e_{ij})_{i,j} = (\chi(i-j))_{0\le i,j<p}$. Note that $\mathbf{E}$ is circulant: $e_{00} = e_{11} = e_{22} = \cdots$, $e_{01} = e_{12} = e_{23} = \cdots$, $e_{10} = e_{21} = e_{32}\cdots$, etc. The rows $\mathbf{v}_i$ $(0 \le i < p)$ of $\mathbf{E}$ have Euclidean norm $\sqrt{p-1}$. Moreover, one may check that

$$\langle \mathbf{v}_i, \mathbf{v}_{i'}\rangle = -1 \tag{148}$$

for all $i \neq i'$. Indeed,

$$\langle \mathbf{v}_i, \mathbf{v}_{i'} \rangle = \sum_{0 \leq j < p} \chi(i-j)\overline{\chi(i'-j)} = \sum_{0 \leq j < p,\, j \neq i,\, i'} \chi\left(\frac{i-j}{i'-j}\right)$$
$$= \sum_{0 \leq j < p,\, j \neq i,\, i'} \chi\left(1 - \frac{i'-i}{i'-j}\right). \quad (149)$$

As j varies over $\mathbb{F}_p \setminus \{i, i'\}$, the argument $1-(i'-i)/(i'-j)$ of χ in the last term in (149) varies over $\mathbb{F}_p \setminus \{0, 1\}$. Since $\chi(1) = 1$ and $\sum_{0 \leq j < p} \chi(j) = 0$, we conclude from (149) that (148) does indeed hold.

Therefore, by Lemma 27, we have

$$\left| \sum_{a \in A,\, b \in B} \chi(a-b) \right| \leq \sqrt{(p-1)|A||B|} \quad (150)$$

for all A and $B \subset \{0, \ldots, p-1\}$. Theorem 28 now follows from (150) in the same way that (95) follows from (87) (and the fact that $\mathbf{E}$ is circulant).

Acknowledgements

The authors are grateful to Eduardo Tengan and Norihide Tokushige for their careful reading of this paper and for their many comments. The authors are also most pleased to thank the referee for his or her very meticulous work.

References

1. N. Alon, Y. Kohayakawa, C. Mauduit, C. G. Moreira, and Rödl, *Measures of pseudorandomness for finite sequences: typical values*, in preparation. 1.1, 1.2
2. Noga Alon, *Problems and results in extremal combinatorics. I*, Discrete Math. **273** (2003), no. 1-3, 31–53, EuroComb'01 (Barcelona). MR **2005a:**05208 2.1, 2.1, 2.1
3. Arturo Carpi and Aldo de Luca, *Uniform words*, Adv. in Appl. Math. **32** (2004), no. 3, 485–522. MR **2005a:**68164 3.2.2
4. Julien Cassaigne, Christian Mauduit, and András Sárközy, *On finite pseudorandom binary sequences. VII. The measures of pseudorandomness*, Acta Arith. **103** (2002), no. 2, 97–118. MR **2004c:**11139 1.1, 1.1
5. Bruno Codenotti, Pavel Pudlák, and Giovanni Resta, *Some structural properties of low-rank matrices related to computational complexity*, Theoret. Comput. Sci. **235** (2000), no. 1, 89–107, Selected papers in honor of Manuel Blum (Hong Kong, 1998). MR **2001e:**05078 2.1
6. N. G. de Bruijn, *A combinatorial problem*, Nederl. Akad. Wetensch., Proc. **49** (1946), 758–764, (Indagationes Math. **8** (1946), 461–467). MR 8,247d 3.1, 3.2.2
7. Paul Erdős and Joel 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 3.2.1
8. C. B. Haselgrove, *Some theorems in the analytic theory of numbers*, J. London Math. Soc. **26** (1951), 273–277. MR 13,438e 2.3.1

9. F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes. II*, North-Holland Publishing Co., Amsterdam, 1977, North-Holland Mathematical Library, Vol. 16. MR 57 #5408b 2.4
10. Christian Mauduit, *Finite and infinite pseudorandom binary words*, Theoret. Comput. Sci. **273** (2002), no. 1-2, 249–261, WORDS (Rouen, 1999). MR **2002m:**11072 1
11. Christian Mauduit and András 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
12. Aimo Tietäväinen, *Bounds for binary codes just outside the Plotkin range*, Inform. and Control **47** (1980), no. 2, 85–93. MR **83f:**94042 2.4

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