Combinatorics, Probability and Computing (2006) 15, 1–29. © 2006 Cambridge University Press
doi:10.1017/S0963548305007170 Printed in the United Kingdom

Measures of Pseudorandomness for Finite Sequences: Minimal Values

N. ALON[1],[†] Y. KOHAYAKAWA[2],[‡] C. MAUDUIT[3],
C. G. MOREIRA[4,§] and V. RÖDL[5,¶,‖]

[1] Raymond and Beverly Sackler Faculty of Exact Sciences,
Tel Aviv University, Tel Aviv 69978, Israel
(e-mail: noga@math.tau.ac.il)
[2] Instituto de Matemática e Estatística, Universidade de São Paulo,
Rua do Matão 1010, 05508-090 São Paulo, Brazil
(e-mail: yoshi@ime.usp.br)
[3] Institut de Mathématiques de Luminy, CNRS-UPR9016, 163 av. de Luminy,
case 907, F-13288, Marseille Cedex 9, France
(e-mail: mauduit@iml.univ-mrs.fr)
[4] IMPA, Estrada Dona Castorina 110, 22460-320 Rio de Janeiro, RJ, Brazil
(e-mail: gugu@impa.br)
[5] Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322, USA
(e-mail: rodl@mathcs.emory.edu)

Received 19 March 2004; revised 30 June 2005

For Béla Bollobás on his 60th birthday

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.

[†] Supported by a USA–Israel BSF grant, by a grant from the Israel Science Foundation and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.
[‡] 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).
[§] Partially supported by MCT/CNPq through a ProNEx project (Proc. CNPq 662416/1996-1) and by CNPq (Proc. 300647/95-6).
[¶] Partially supported by NSF grant 0300529.
[‖] Part of this work was done at IMPA, whose hospitality the authors gratefully acknowledge. This research was partially supported by IM-AGIMB/IMPA. 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).

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 $\mathcal{N}(E_N)$.

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 $1 \leqslant d_1 < \cdots < d_k \leqslant N - M + 1$. 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

$$T(E_N, M, X) = \operatorname{card}\{n : 0 \leqslant n < M,\ n + k \leqslant N, \quad \text{and} \quad (e_{n+1}, e_{n+2}, \dots, e_{n+k}) = X\} \tag{1.1}$$

and

$$\begin{aligned} V(E_N, M, D) &= \sum \{e_{n+d_1} e_{n+d_2} \cdots e_{n+d_k} : 0 \leqslant n < M\} \\ &= \sum_{0 \leqslant n < M} \prod_{1 \leqslant i \leqslant k} e_{n+d_i} = \sum_{0 \leqslant n < M} \prod_{d \in D} e_{n+d}. \end{aligned} \tag{1.2}$$

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

$$\mathcal{N}(E_N) = \max_k \max_X \max_M \left| T(E_N, M, X) - \frac{M}{2^k} \right|, \tag{1.3}$$

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

$$C_k(E_N) = \max\{|V(E_N, M, D)| : M \text{ and } D \text{ such that } M - 1 + d_k \leqslant N\}. \tag{1.4}$$

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 [2], 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 $N \geqslant N_0$, then, with probability at least $1-\varepsilon_0$, we have*

$$\frac{2}{5}\sqrt{N\log\binom{N}{k}} < C_k(E_N) < \sqrt{(2+\varepsilon_1)N\log\left(N\binom{N}{k}\right)} < \sqrt{(3+\varepsilon_0)N\log\binom{N}{k}} < \frac{7}{4}\sqrt{N\log\binom{N}{k}} \tag{1.5}$$

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

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 $2 \leqslant k \leqslant \log N - \log\log N$, there is a function $\Gamma(k,N)$ and a constant N_0 for which the following holds. If $N \geqslant N_0$, then the probability that*

$$1-\varepsilon < \frac{C_k(E_N)}{\Gamma(k,N)} < 1+\varepsilon \tag{1.6}$$

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 $2 \leqslant k \leqslant N$, we have*

(i) $\min\{C_k(E_N) : E_N \in \{-1,1\}^N\} = 1$ *if k is odd,*
(ii) $\min\{C_k(E_N) : E_N \in \{-1,1\}^N\} \geqslant \log_2(N/k)$ *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 $k \geqslant 2$ there is a constant $c > 0$ such that for $N \to \infty$ we have

$$\min\{C_k(E_N) : E_N \in \{-1,1\}^N\} \gg N^c, \tag{1.7}$$

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 1.1. *If k and N are natural numbers with k even and $2 \leqslant k \leqslant N$, then*

$$C_k(E_N) > \sqrt{\frac{1}{2}\left\lfloor \frac{N}{k+1} \right\rfloor} \tag{1.8}$$

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

The lower bound given in (1.8) decreases as k increases. One may ask whether, in fact, $C_{2k}(E_N) \geqslant c\sqrt{kN}$ for some absolute constant $c > 0$, or at least $C_{2k}(E_N) \geqslant c\sqrt{N}$ 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 $C_2(E_N), C_4(E_N), \ldots, C_k(E_N)$ (with k again even), then a lower bound of order $\sqrt{kN}$ may indeed be proved.

Theorem 1.2. *There is an absolute constant $c > 0$ for which the following holds. For any positive integers ℓ and N with $\ell \leqslant N/3$, we have*

$$\max\{C_2(E_N), C_4(E_N), \ldots, C_{2\ell}(E_N)\} \geqslant c\sqrt{\ell N} \tag{1.9}$$

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

In view of Theorem A, the lower bound in Theorem 1.2 is best possible apart from a multiplicative factor of $O\big(\sqrt{\log(N/2\ell)}\big)$, for all $\ell \leqslant N/8$.

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

$$\max\{C_{2k-2}(E_N), C_{2k}(E_N)\} \geqslant \sqrt{\frac{1}{2}\left\lfloor \frac{N}{3} \right\rfloor} \tag{1.10}$$

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

Some further results are stated and proved in Section 2.3 (see Theorems 2.7, 2.9, and 2.10).

1.2. Typical and minimal values of normality

We now turn to the normality measure $\mathcal{N}(E_N)$. In [2], the following result is proved.

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

$$\delta\sqrt{N} < \mathcal{N}(E_N) < \frac{1}{\delta}\sqrt{N} \tag{1.11}$$

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 $\mathcal{N}(E_N) = O\big(N^{1/3}(\log N)^{2/3}\big)$.

Theorem 1.4. *For any sufficiently large N, there exists a sequence $E_N \in \{-1,1\}^N$ with*

$$\mathcal{N}(E_N) \leqslant 3N^{1/3}(\log N)^{2/3}. \tag{1.12}$$

A simple argument shows that $\mathcal{N}(E_N) \geqslant (1/2 + o(1)) \log_2 N$ for any $E_N \in \{-1,1\}^N$ (see Proposition 3.1). In view of Theorem 1.4, we have

$$\left(\frac{1}{2} + o(1)\right) \log_2 N \leqslant \min_{E_N \in \{-1,1\}^N} \mathcal{N}(E_N) \leqslant 3N^{1/3}(\log N)^{2/3} \tag{1.13}$$

for all large enough N. It would be interesting to close the rather wide gap in (1.13).

The construction of the sequence $E_N \in \{-1,1\}^N$ in Theorem 1.4 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 (1.12) 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 1.1 that we give in Section 2.2 is based on the following elementary lemma from linear algebra (see, *e.g.*, [1, Lemma 9.1] or [5, Lemma 7]), whose proof we include for completeness.

Lemma 2.1. *For any symmetric matrix $\mathbf{A} = (A_{ij})_{1 \leqslant i,j \leqslant n}$, we have*

$$\mathrm{rk}(\mathbf{A}) \geqslant \frac{(\mathrm{tr}(\mathbf{A}))^2}{\mathrm{tr}(\mathbf{A}^2)} = \frac{\left(\sum_{1 \leqslant i \leqslant n} A_{ii}\right)^2}{\sum_{1 \leqslant i,j \leqslant n} A_{ij}^2}. \tag{2.1}$$

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

$$\mathrm{rk}(\mathbf{A}) \geqslant \frac{n}{1 + \varepsilon^2(n-1)}. \tag{2.2}$$

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

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

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

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

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

as required. Inequality (2.2) follows immediately from (2.1). □

The next lemma, due to the first author [1], improves Lemma 2.1 for larger values of ε.

Lemma 2.2. *Let $\mathbf{A} = (A_{ij})_{1\leqslant i,j\leqslant n}$ be an $n \times n$ real matrix with $A_{ii} = 1$ for all i and $|A_{ij}| \leqslant \varepsilon$ for all $i \neq j$, where $\sqrt{1/n} \leqslant \varepsilon \leqslant 1/2$. Then*

$$\mathrm{rk}(\mathbf{A}) \geqslant \frac{1}{100\varepsilon^2 \log(1/\varepsilon)} \log n. \tag{2.3}$$

If $\mathbf{A}$ is symmetric, then (2.3) holds with the constant $1/100$ replaced by $1/50$.

For completeness, we give the proof of Lemma 2.2. We shall need the following auxiliary lemma [1].

Lemma 2.3. *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 $\mathbf{v}_1 = (v_{1,j})_{j=1}^n$, $\mathbf{v}_2 = (v_{2,j})_{j=1}^n$, …, $\mathbf{v}_d = (v_{d,j})_{j=1}^n$ 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 $k_1, k_2, \ldots, k_d$ range over all nonnegative 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 $k_1, k_2, \ldots, k_d$ whose sum is precisely k. □

Proof of Lemma 2.2. Let us first note that the nonsymmetric 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 (2.3) with the constant $1/100$ replaced by $1/50$.

Let $\delta = 1/16$. Consider first the case in which $\varepsilon \leqslant 1/n^\delta$. 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 $1/\sqrt{m} \geqslant \varepsilon$, and hence Lemma 2.1 applies to $\mathbf{A}'$, and we deduce that $\mathrm{rk}(\mathbf{A}) \geqslant \mathrm{rk}(\mathbf{A}') \geqslant m/2$. It now suffices to check that, because $\varepsilon \leqslant \min\{1/2, 1/n^\delta\}$ and $\delta = 1/16$, we have

$$\frac{1}{2}m \geqslant \frac{3}{8\varepsilon^2} = \frac{3}{2^7\delta\varepsilon^2} > \frac{1}{50\varepsilon^2 \log(1/\varepsilon)} \log n, \tag{2.4}$$

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

$$k = \left\lfloor \frac{\log n}{2\log(1/\varepsilon)} \right\rfloor \geqslant \left\lfloor \frac{1}{2\delta} \right\rfloor = 8, \tag{2.5}$$

and let $m = \lfloor 1/\varepsilon^{2k} \rfloor$. Note that, then, we have $m \leqslant n$. 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

$$\varepsilon^k \leqslant \frac{1}{\sqrt{m}}. \tag{2.6}$$

Let $\mathbf{A}''$ be the matrix obtained from $\mathbf{A}'$ by raising all its entries to the kth power. Because of (2.6) and the hypothesis on the entries of $\mathbf{A}$, Lemma 2.1 applies and tells us that

$$\mathrm{rk}(\mathbf{A}'') \geqslant \frac{1}{2}m = \frac{1}{2}\left\lfloor \frac{1}{\varepsilon^{2k}} \right\rfloor \geqslant \frac{0.49}{\varepsilon^{2k}}, \tag{2.7}$$

where the last inequality follows easily from the fact that $\varepsilon \leqslant 1/2$ and $k \geqslant 8$ (see (2.5)). We now observe that Lemma 2.3 tells us that

$$\mathrm{rk}(\mathbf{A}'') \leqslant \binom{k + \mathrm{rk}(\mathbf{A}')}{k} \leqslant \left(\frac{\mathrm{e}(k + \mathrm{rk}(\mathbf{A}'))}{k}\right)^k. \tag{2.8}$$

Putting together (2.7) and (2.8), we get

$$\mathrm{rk}(\mathbf{A}) \geqslant \mathrm{rk}(\mathbf{A}') \geqslant \frac{k}{\varepsilon^2}\left(\frac{0.49^{1/k}}{\mathrm{e}} - \varepsilon^2\right), \tag{2.9}$$

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

$$\mathrm{rk}(\mathbf{A}) > \frac{1}{50\varepsilon^2 \log(1/\varepsilon)} \log n, \tag{2.10}$$

and we are done. □

2.2. Proof of the lower bounds for correlation

We shall prove Theorems 1.1 and 1.2 in this section. These results will be deduced from suitable applications of Lemmas 2.1 and 2.2; to describe these applications, we first need to introduce some notation.

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

$$v_{L,i} = \prod_{x \in L} e_{i+x} \tag{2.11}$$

for all $0 \leqslant i < M$ (note that $1 \leqslant i + x \leqslant M - 1 + N' = N$ for any x in (2.11)). Let us now define an $\mathscr{L} \times \mathscr{L}$ matrix $\mathbf{A} = (A_{L,L'})_{L,L' \in \mathscr{L}}$, putting

$$A_{L,L'} = \frac{1}{M}\langle \mathbf{v}_L, \mathbf{v}_{L'} \rangle = \frac{1}{M} \sum_{0 \leqslant i < M} v_{L,i} v_{L',i} \tag{2.12}$$

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

$$\begin{aligned} A_{L,L'} = \frac{1}{M}\langle \mathbf{v}_L, \mathbf{v}_{L'} \rangle &= \frac{1}{M} \sum_{0 \leqslant i < M} \left(\prod_{x \in L} e_{i+x}\right)\left(\prod_{y \in L'} e_{i+y}\right) \\ &= \frac{1}{M} \sum_{0 \leqslant i < M} \prod_{z \in L \triangle L'} e_{i+z}, \end{aligned} \tag{2.13}$$

where we write $L \triangle L'$ for the symmetric difference of the sets L and L'. Let $\mathscr{L}^{\triangle} = \{L \triangle L' : L, L' \in \mathscr{L},\ L \neq L'\}$ and let K be the set of the cardinalities of the members

of $\mathscr{L}^{\triangle}$, that is, $K = \{|S| : S \in \mathscr{L}^{\triangle}\}$. It follows from (2.13) and the definition of $C_k(E_N)$ that

$$\max\{C_k(E_N) : k \in K\} \geqslant M \max\{|A_{L,L'}| : L, L' \in \mathscr{L},\ L \neq L'\}. \tag{2.14}$$

Lemma 2.1 and (2.14) imply the following result.

Lemma 2.4. *We have*

$$\max\{C_k(E_N) : k \in K\} > \sqrt{M - \frac{M^2}{|\mathscr{L}|}}. \tag{2.15}$$

Proof. Let $\mathbf{B} = (\mathbf{v}_L^T)_{L \in \mathscr{L}}$ be the $|\mathscr{L}| \times M$ matrix with rows $\mathbf{v}_L^T$ $(L \in \mathscr{L})$. 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 2.1, we get

$$M \geqslant \mathrm{rk}(\mathbf{A}) > \frac{|\mathscr{L}|}{1 + \varepsilon^2 |\mathscr{L}|}, \tag{2.16}$$

where $\varepsilon = \max\{|A_{L,L'}| : L, L' \in \mathscr{L},\ L \neq L'\}$. It follows from (2.16) that

$$\varepsilon > \sqrt{\frac{1}{M} - \frac{1}{|\mathscr{L}|}}. \tag{2.17}$$

Inequality (2.15) follows from (2.14) on multiplying (2.17) by M. □

We are now ready to prove Theorem 1.1.

Proof of Theorem 1.1. Let k, N, and E_N be as in the statement of Theorem 1.1. Set $\ell = k/2$ and $M = \lfloor N/(k+1) \rfloor$ and, as above, let $N' = N - M + 1$. We take for $\mathscr{L} \subset \mathscr{P}([N'])$ a set system of $t = \lfloor N'/\ell \rfloor$ pairwise disjoint ℓ-element subsets $L_1, \ldots, L_t$ of $[N']$. Note that

$$|\mathscr{L}| = t = \left\lfloor \frac{N - \lfloor N/(k+1) \rfloor + 1}{k/2} \right\rfloor \geqslant \left\lfloor \frac{2N}{k+1} \right\rfloor \geqslant 2M. \tag{2.18}$$

Therefore, it follows from (2.15) and (2.18) that

$$C_k(E_N) > \sqrt{M - \frac{M^2}{|\mathscr{L}|}} \geqslant \sqrt{M - \frac{M}{2}} = \sqrt{\frac{1}{2}\left\lfloor \frac{N}{k+1} \right\rfloor}, \tag{2.19}$$

as required. □

Lemma 2.4 was deduced from an application of Lemma 2.1 to the matrix $\mathbf{A} = (A_{L,L'})$; the next lemma will be obtained from an application of Lemma 2.2 to $\mathbf{A}$.

Lemma 2.5. *If* $2M \leqslant |\mathscr{L}| < \mathrm{e}^{50M}$, *then*

$$\max\{C_k(E_N) : k \in K\} \geqslant \min\left\{\frac{1}{2}M, \sqrt{\frac{1}{50}M(\log|\mathscr{L}|) \Big/ \log\frac{50M}{\log|\mathscr{L}|}}\right\}. \tag{2.20}$$

Proof. Let $\varepsilon = \max\{|A_{L,L'}| : L, L' \in \mathscr{L},\ L \neq L'\}$. Inequality (2.2) and the fact that $\mathrm{rk}(\mathbf{A}) \leqslant M$, coupled with $M \leqslant |\mathscr{L}|/2$, give that

$$\varepsilon^2 > \frac{1}{M} - \frac{1}{|\mathscr{L}|} \geqslant \frac{1}{|\mathscr{L}|}, \tag{2.21}$$

and hence $\varepsilon > \sqrt{1/|\mathscr{L}|}$. If $\varepsilon > 1/2$, then (2.20) follows immediately (recall (2.14)). Therefore, we may suppose that $\sqrt{1/|\mathscr{L}|} \leqslant \varepsilon \leqslant 1/2$, and hence we may apply Lemma 2.2 to the symmetric matrix $\mathbf{A}$. Combining the fact that $\mathbf{A}$ has rank at most M with Lemma 2.2, we obtain that

$$M \geqslant \mathrm{rk}(\mathbf{A}) \geqslant \frac{1}{50\varepsilon^2 \log(1/\varepsilon)} \log|\mathscr{L}|, \tag{2.22}$$

whence

$$\varepsilon^2 \log\frac{1}{\varepsilon} \geqslant \frac{1}{50M} \log|\mathscr{L}|. \tag{2.23}$$

Using that $1/\varepsilon \geqslant \log 1/\varepsilon$, we have from (2.23) that

$$\varepsilon \geqslant \varepsilon^2 \log\frac{1}{\varepsilon} \geqslant \frac{1}{50M} \log|\mathscr{L}|. \tag{2.24}$$

Plugging (2.24) into (2.23), we get

$$\varepsilon^2 \log\frac{50M}{\log|\mathscr{L}|} \geqslant \varepsilon^2 \log\frac{1}{\varepsilon} \geqslant \frac{1}{50M} \log|\mathscr{L}|, \tag{2.25}$$

and hence

$$\varepsilon \geqslant \sqrt{\frac{\log|\mathscr{L}|}{50M} \Big/ \log\frac{50M}{\log|\mathscr{L}|}}. \tag{2.26}$$

Inequality (2.20) follows easily from (2.14), (2.26), and the definition of ε. □

We shall now deduce Theorem 1.2 from Lemma 2.5.

Proof of Theorem 1.2. Let ℓ and N with $\ell \leqslant N/3$ be given. Let $M = \lfloor N/3 \rfloor$, and set $N' = N - M + 1 \geqslant 2N/3$. We take for $\mathscr{L}$ the set system of all ℓ-element subsets of $[N']$. Then, clearly, $\mathscr{L}^{\triangle} = \{L \triangle L' : L, L' \in \mathscr{L},\ L \neq L'\}$ is the family of non-empty subsets of $[N']$ of even cardinality not greater than 2ℓ. Hence, $K = \{|S| : S \in \mathscr{L}^{\triangle}\} = \{2, 4, \ldots, 2\ell\}$. Moreover,

$$|\mathscr{L}| = \binom{N'}{\ell} \geqslant N' \geqslant \frac{2N}{3} \geqslant 2M, \tag{2.27}$$

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

$$|\mathscr{L}| \leqslant 2^N = (2^{N/M})^M \leqslant 2^{5M} < \mathrm{e}^{50M}. \tag{2.28}$$

Inequalities (2.27) and (2.28) tell us that Lemma 2.5 may be applied. We deduce from Lemma 2.5 that

$$\max\{C_2(E_N), C_4(E_N), \ldots, C_{2\ell}(E_N)\} \geqslant \min\left\{\frac{1}{2}M, \sqrt{\frac{1}{50}M(\log|\mathscr{L}|) \Big/ \log\frac{50M}{\log|\mathscr{L}|}}\right\}. \tag{2.29}$$

If the minimum on the right-hand side of (2.29) 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

$$\frac{1}{50}M(\log|\mathscr{L}|)\Big/\log\frac{50M}{\log|\mathscr{L}|} \geqslant \frac{1}{50}\left\lfloor\frac{N}{3}\right\rfloor(\log|\mathscr{L}|)\Big/\log\frac{50N/3}{\log|\mathscr{L}|}, \tag{2.30}$$

and, moreover,

$$|\mathscr{L}| = \binom{N'}{\ell} \geqslant \left(\frac{2N}{3\ell}\right)^{\ell}, \tag{2.31}$$

so that

$$\log|\mathscr{L}| \geqslant \ell\log\frac{2N}{3\ell}. \tag{2.32}$$

By (2.30) and (2.32), it suffices to show that

$$\frac{1}{150}N\ell\left(\log\frac{2N}{3\ell}\right)\Big/\log\frac{50N/3}{\ell\log(2N/3\ell)} \geqslant c'N\ell \tag{2.33}$$

for some absolute constant $c' > 0$. Routine calculations show that a suitable constant $c' > 0$ will do in (2.33). We only give a sketch: suppose first that $1 \leqslant \ell = o(N)$. In this case, it is simple to check that the left-hand side of (2.33) is in fact

$$\left(\frac{1}{150} + o(1)\right)N\ell. \tag{2.34}$$

Suppose now that $c''N \leqslant \ell \leqslant N/3$. In this case, the left-hand side of (2.33) is at least

$$\frac{1}{150}N\ell(\log 2)\Big/\log\frac{50/3}{c''\log 2}, \tag{2.35}$$

and (2.33) follows for some small enough $c' > 0$. □

2.3. Some further lower bounds for correlation

In this section, we deduce some further consequences of Lemmas 2.4 and 2.5, using other families $\mathscr{L}$.

2.3.1. Projective plane bounds. We shall prove Theorem 1.3 (see Section 1.1) by making use of systems of sets derived from projective planes. Recall that Theorem 1.3 tells us that, for any $2 \leqslant k \leqslant \sqrt{N/6}$ and any $E_N \in \{-1,1\}^N$, at least one of $C_{2k-2}(E_N)$ and $C_{2k}(E_N)$ is $\geqslant c\sqrt{N}$, 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 2.6. *Let positive integers k and n with $k \leqslant (1/2)\sqrt{n}$ be given. If n is large enough, then there is a family $\mathscr{L}$ of k-element subsets of $[n]$ with $|\mathscr{L}| = n$ and such that $|L \cap L'| \leqslant 1$ for all distinct L and $L' \in \mathscr{L}$.*

One may prove Lemma 2.6 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