# On the correlation of binary sequences

R. Ahlswede [a] J. Cassaigne [b] A. Sárközy [c,1]

[a] *Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, D-33501 Bielefeld, Germany*

[b] *Institut de Mathématiques de Luminy, 163 avenue de Luminy, Case 907, F-13288 Marseille Cedex 9, France*

[c] *Eötvös Loránd University, Department of Algebra and Number Theory, H-1117 Budapest, Pázmány Péter sétány 1/c, Hungary*

## 1 Introduction

In a series of papers Mauduit and Sárközy (partly with further coauthors) studied finite pseudorandom binary sequences

$$E_N = \{e_1, e_2, \ldots, e_N\} \in \{-1, +1\}^N.$$

In particular, in Part I [6] first they introduced the following measures of pseudorandomness:

The **well-distribution measure** of $E_N$ is defined as

$$W(E_N) = \max_{a,b,t} \left| \sum_{j=0}^{t-1} e_{a+jb} \right|$$

where the maximum is taken over all $a, b, t$ such that $a, b, t \in \mathbb{N}$ (where $\mathbb{N}$ is the set of the positive integers) and $a \le a + (t-1)b \le N$, while the **correlation measure of order** $k$ of $E_N$ is defined as

$$C_k(E_N) = \max_{M,D} \left| \sum_{n=1}^{M} e_{n+d_1} e_{n+d_2} \ldots e_{n+d_k} \right|$$

[1] Research of A. Sárközy partially supported by the Hungarian National Foundation for Scientific Research, Grant No. T043623. This paper was written while the second and third author were visiting the Zentrum für interdisziplinäre Forschung, Universität Bielefeld.

where the maximum is taken over all $D = (d_1, \dots, d_k)$ and $M$ such that $d_1 < \dots < d_k$ are non-negative integers with $M + d_k \le N$.

Then the sequence $E_N$ is considered as a "good" pseudorandom sequence if both these measures $W(E_N)$ and $C_k(E_N)$ (at least for small $k$) are "small" in terms of $N$ (in particular, both are $o(N)$ as $N \to \infty$). This terminology is justified by the fact that, as it is shown in [2] and [5], for a "truly random" $E_N \in \{-1,+1\}^N$ both $W(E_N)$ and, for fixed $\ell$, $C_\ell(E_N)$ are around $N^{1/2}$ with "near to 1" probability.

In [6] is explained why to use the well-distribution measure and correlation measure as measures of pseudorandomness. However, one would expect that there are applications where it suffices to control only some of the pseudorandom measures instead of the full control of all of them. In particular, one of the most important applications of pseudorandomness is cryptography. If, e.g., we want to use a binary sequence $E_N \in \{-1,+1\}^N$ (after transforming it into a bit sequence) as a key stream in the standard Vernam cipher [7], then $E_N$ must possess certain pseudorandom properties. Does $E_N$ need to possess both small well-distribution measure and, for any fixed small $k$, small correlation measure of order $k$? In other words, if $W(E_N)$ is large, resp. $C_k(E_N)$ is large for some fixed small $k$, then can the enemy utilize this fact to break the code? The most natural line of attack is the exhaustive search: the attacker may try all the binary sequences $E_N \in \{-1,+1\}^N$ with large $W(E_N)$, resp. large $C_k(E_N)$, as a potential key stream. Clearly, this attack is really threatening only if the number of sequences $E_N \in \{-1,+1\}^N$ with

(i) large $W(E_N)$, resp.
(ii) large $C_k(E_N)$

is "much less" than the total number $2^N$ of sequences in $\{-1,+1\}^N$, besides one needs a fast algorithm to generate the sequences of type (i), resp. (ii).

The case (i) is easy, thus, for the sake of completeness, here we just present an estimate for the number of sequences $E_N$ with large $W(E_N)$ and we sketch the background, but we leave the details (which are similar but simpler than later in the study of the correlation) to the reader.

For $0 \le x \le 1$ define the function $\xi(x)$ (that is, the binary entropy) by

$$\xi(x) = \begin{cases} -\frac{1}{\log 2}(x \log x + (1-x)\log(1-x)) & \text{for } 0 < x < 1 \\ 0 & \text{for } x = 0 \text{ and } x = 1. \end{cases}$$

For $N \in \mathbb{N}$ and $0 < \alpha < 1$ write

$$\mathcal{V}(N,\alpha) = \{E_N : E_N \in \{-1,+1\}^N, W(E_N) > \alpha N\}.$$

**Proposition 1.** *For* $0 < \alpha < 1$, $\varepsilon > 0$, $N \in \mathbb{N}$, *and* $N > N_o(\alpha, \varepsilon)$ *we have*

$$2^{(\xi((1-\alpha)/2)-\varepsilon)N} < |\mathcal{V}(N,\alpha)| < 2^{(\xi((1-\alpha)/2)+\varepsilon)N}.$$

**Proof of Proposition 1.** Indeed, if $E_N \in \mathcal{V}(N,\alpha)$, then one of the sums in the definition of $W(E_N)$ must be greater than $\alpha N$:

$$\left|\sum_{j=0}^{t-1} e_{a+jb}\right| = \Big|t - 2|\{j : 0 \le j < t, e_{a+jb} = -1\}|\Big| > \alpha N.$$

It follows that either

$$|\{j : 0 \le j < t, e_{a+jb} = -1\}| > \frac{t+\alpha N}{2}$$

or

$$|\{j : 0 \le j < t, e_{a+jb} = -1\}| < \frac{t-\alpha N}{2};$$

we may assume that the latter inequality holds. If $a$, $b$, $t$ are fixed and the number of $j$'s with $e_{a+jb} = -1$ is $s$, then these $j$ values can be chosen in $\binom{t}{s}$ ways, with $s < \frac{t-\alpha N}{2}$. Thus fixing $a$, $b$, $t$, and one of the inequalities above, there are

$$\sum_{0 \le s < \frac{t-\alpha N}{2}} \binom{t}{s}$$

choices of $E_N$. This sum is the greatest when $t$ is the greatest, i.e., $a = b = 1$, $t = N$, and then we have

$$\sum_{0 \le s < \frac{1-\alpha}{2}N} \binom{N}{s} \ge \binom{N}{\left[\frac{1-\alpha}{2}N\right] - 1}$$

so this is a lower bound for $|\mathcal{V}(N,\alpha)|$. To obtain an upper bound observe that there are two inequalities above to consider, $a$, $b$, $t$, $s$ each can be chosen in at most $N$ ways, and the greatest term in the last sum is at most $\binom{N}{\left[\frac{1-\alpha}{2}N\right]}$ so that

$$|\mathcal{V}(N,\alpha)| < 2N^4 \binom{N}{\left[\frac{1-\alpha}{2}N\right]}.$$

It remains to estimate the binomial coefficient in the lower and upper bound for $|\mathcal{V}(N,\alpha)|$, which can be done in the standard way of estimating binomial coefficients (see Lemma 2 later) and then we get the result.

The case (ii), i.e., the case of large correlation is much more interesting: this case will be studied in Section 2.

In Section 3 we will sharpen the results of Section 2 in the special case when the order of the correlation is 2.

Finally, in Section 4 we will study a lemma, which plays a crucial role in the estimation of the correlation in some of the most important constructions of pseudorandom binary sequences.

## 2 The number of binary sequences with large correlation

For $k, N \in \mathbb{N}$, $2 \le k \le N$, and $0 < \alpha < 1$ write

$$\mathcal{F}(k, N, \alpha) = \{E_N : E_N \in \{-1, +1\}^N, C_k(E_N) > \alpha N\}.$$

First we will prove:

**Theorem 1.** *For every $k \in \mathbb{N}$ with $k \ge 2$ and every $\varepsilon > 0$ there are $\delta = \delta(k, \varepsilon) > 0$ and $N_o = N_o(k, \varepsilon)$ so that for $N > N_o$ we have*

$$|\mathcal{F}(k, N, \varepsilon)| < 2^{(1-\delta)N}. \tag{2.1}$$

**Proof of Theorem 1.** The proof will be based on the following estimate:

**Lemma 1.** *For all $k, N \in \mathbb{N}$, $2 \le k \le N$, and $0 < \alpha < 1$ we have*

$$|\mathcal{F}(k, N, \alpha)| \le 2N^{k+1} \max_{M : \alpha N < M < N} 2^{N-M} \sum_{0 \le t < \frac{M - \alpha N}{2}} \binom{M}{t}. \tag{2.2}$$

**Proof of Lemma 1.** If $E_N \in \mathcal{F}(k, N, \alpha)$, then by the definition of $\mathcal{F}(k, N, \alpha)$ and $C_k(E_N)$, there are $M$, $d_1, \dots, d_k$ with

$$1 \le M < N \tag{2.3}$$

$$(0 \le) \quad d_1 < \dots < d_k \quad (\le N - M < N) \tag{2.4}$$

and either

$$\sum_{n=1}^{M} e_{n+d_1} \dots e_{n+d_k} > \alpha N \tag{2.5}$$

or

$$\sum_{n=1}^{M} e_{n+d_1} \dots e_{n+d_k} < -\alpha N. \tag{2.6}$$

Let $\mathcal{F}^+$ and $\mathcal{F}^-$ denote the set of the sequences $E_N \in \mathcal{F}(k,N,\alpha)$ for which (2.5), resp. (2.6), holds for some $D$ and $M$, so that $\mathcal{F}(k,N,\alpha) = \mathcal{F}^+ \cup \mathcal{F}^-$ and whence

$$|\mathcal{F}(k,N,\alpha)| \leq |\mathcal{F}^+| + |\mathcal{F}^-|. \tag{2.7}$$

First we will estimate $|\mathcal{F}^+|$. Assume that $E_N \in \mathcal{F}^+$, and (2.3), (2.4), and (2.5) hold.

It follows from (2.5) that

$$\alpha N < \sum_{n=1}^{M} e_{n+d_1} \dots e_{n+d_k} \leq \sum_{n=1}^{M} 1 = M. \tag{2.8}$$

Let $\mathcal{H} = \{h_1, \dots, h_t\}$ denote the set of the positive integers $h$ with

$$1 \leq h \leq M, \quad e_{h+d_1} \dots e_{h+d_k} = -1. \tag{2.9}$$

By (2.5) we have

$$\alpha N < \sum_{n=1}^{M} e_{n+d_1} \dots e_{n+d_k} = \sum_{\substack{1\leq n \leq M \\ n \notin \mathcal{H}}} 1 - \sum_{\substack{1\leq n \leq M \\ n \in \mathcal{H}}} 1 = (M - |\mathcal{H}|) - |\mathcal{H}| = M - 2|\mathcal{H}|$$

and whence,

$$|\mathcal{H}| < \frac{M}{2} - \frac{\alpha}{2}N. \tag{2.10}$$

Now observe that

$$M, d_1, \dots, d_k, e_1, e_2, \dots, e_{d_k}, e_{M+d_k+1}, e_{M+d_k+2}, \dots, e_N, \mathcal{H} \tag{2.11}$$

determine $E_N = \{e_1, \dots, e_N\}$ uniquely. To prove this, clearly it suffices to show that the numbers in (2.11) determine every $e_{n+d_k}$ with $1 \leq n \leq M$ uniquely. This can be proved by induction on $n$: if $e_1, e_2, \dots, e_{(n-1)+d_k}$ are already given, then

$$e_{n+d_k} = \begin{cases} e_{n+d_1} \dots e_{n+d_{k-1}} & \text{if } n \notin \mathcal{H} \\ -e_{n+d_1} \dots e_{n+d_{k-1}} & \text{if } n \in \mathcal{H}. \end{cases}$$

Thus it remains to count the number of choices of the parameters in (2.11). First we fix an $M$ satisfying (2.3) and (2.8). By (2.4), we may choose $d_1, \dots, d_k$ in at most $N^k$ ways. Each of the $e_i$'s in (2.11) can be chosen in two ways, and their number is

$$d_k + (N - (M + d_k)) = N - M$$

so that they can be chosen in $2^{N-M}$ ways. Finally, by (2.9) and (2.10), the set $\mathcal{H} \subset \{1, 2, \dots, M\}$ can be chosen in at most

$$\sum_{0 \leq t < \frac{M - \alpha N}{2}} \binom{M}{t} \tag{2.12}$$

ways. It follows that for fixed $M$ the remaining parameters in (2.11) can be chosen in at most

$$N^k 2^{N-M} \sum_{0 \le t < \frac{M-\alpha N}{2}} \binom{M}{t}$$

ways. Summation over the $M$ values satisfying (2.3) and (2.8) gives

$$\begin{aligned} |\mathcal{F}^+| &\le \sum_{\alpha N < M < N} N^k 2^{N-M} \sum_{0 \le t < \frac{M-\alpha N}{2}} \binom{M}{t} \\ &\le N^{k+1} \max_{\alpha N < M < N} 2^{N-M} \sum_{0 \le t < \frac{M-\alpha N}{2}} \binom{M}{t}. \end{aligned} \tag{2.13}$$

$|\mathcal{F}^-|$ can be estimated in exactly the same way and we obtain the same upper bound:

$$|\mathcal{F}^-| \le N^{k+1} \max_{\alpha N < M < N} 2^{N-M} \sum_{0 \le t < \frac{M-\alpha N}{2}} \binom{M}{t}. \tag{2.14}$$

(2.2) follows from (2.7), (2.13), and (2.14), and this completes the proof of the lemma.

In order to derive (2.1) from Lemma 1, observe that clearly for all $\varepsilon > 0$ there is a $\gamma = \gamma(\varepsilon) > 0$ such that for all $M < N$ we have

$$\sum_{0 \le t < \frac{M-\varepsilon N}{2}} \binom{M}{t} \le \sum_{0 \le t < \frac{1-\varepsilon}{2} M} \binom{M}{t} < 2^{(1-\gamma)M}.$$

Thus by Lemma 1 we have

$$\begin{aligned} |\mathcal{F}(k, N, \varepsilon)| &\le 2N^{k+1} \max_{\varepsilon N < M < N} 2^{N-M} 2^{(1-\gamma)M} \\ &= 2N^{k+1} \max_{\varepsilon N < M < N} 2^{N-\gamma M} < 2N^{k+1} 2^{(1-\gamma\varepsilon)N} \end{aligned}$$

and whence (2.1) follows with $\delta = \frac{\gamma\varepsilon}{2}$ for $N > N_o(k, \varepsilon)$. This completes the proof of Theorem 1.

Note that the proof above also provides a fast algorithm for generating the family $\mathcal{F}(k, N, \alpha)$ of the sequences of large correlation of order $k$. The steps of this algorithm[2] follow trivially from the proof.

Indeed, the first step is to fix the data in (2.11). First we fix an integer $M$ satisfying (2.3). Next we choose integers $(0 <)\ d_1 < \cdots < d_k\ (\le N - M)$, i.e., we select a subset of cardinality $k$ from the set $\{1, 2, \ldots, N - M\}$; it is a standard combinatorial problem to generate the subsets of given size of a

[2] The algorithm was included to this paper at the 13th September 2006

given set. Next, we choose a binary sequence

$$\{e_1, e_2, \dots, e_{d_k}, e_{M+d_k+1}, e_{M+d_k+2}, \dots, e_N\} \in \{-1,+1\}^{N-M};$$

again, it is a standard and easy problem to generate all binary sequences of a given length. Finally, we select a subset $\mathcal{H}$ of $\{1, 2, \dots, M\}$ with cardinality $|\mathcal{H}|$ satisfying (2.10); it is a basic problem to generate all the subsets $\mathcal{H} \subset \{1, 2, \dots, M\}$ with $|\mathcal{H}| \le H$ and there are standard fast and simple algorithms executing this. To generate all the subsets $\mathcal{H}$ is the most critical and time-consuming step of the algorithm (now we choose many more elements than in the previous step where $k$ elements were chosen with a fixed and usually small $k$). This completes the choice of the data listed in (2.11), and once this data are given then we may complete the algorithm by determining the elements $e_1$, $e_2$, $\dots$, $e_N$ by the recursion described after (2.11).

One might like to know how far the upper bound in (2.1) is from the best possible. Fix a $\delta > 0$ and $k, N \in \mathbb{N}$ with $k < \frac{\delta N}{2}$. Set $H = \lceil \delta N \rceil - 1$, and define the family $\mathcal{G} \subset \{-1,+1\}^N$ by

$$\mathcal{G} = \Big\{\{e_1, \dots, e_N\} : e_1, \dots, e_N \in \{-1,+1\}, e_1 = e_2 = \dots = e_H = +1\Big\}.$$

Then clearly

$$|\mathcal{G}| = 2^{N-H} > 2^{N-\delta N} = 2^{(1-\delta)N}$$

and for $N > N_o(\delta, k)$ and every $E_N = \{e_1, \dots, e_N\} \in \mathcal{G}$ we have

$$C_k(E_N) \ge \sum_{n=1}^{H-k+1} e_n e_{n-1} \dots e_{n+k-1} = \sum_{n=1}^{H-k+1} 1 = H - k + 1 > \frac{\delta}{2} N$$

so that we have

$$|\mathcal{F}(k, N, \delta/2)| > |\mathcal{G}| > 2^{(1-\delta)N}.$$

This example shows that apart from the dependence of $\delta = \delta(k, \varepsilon)$ on $k$ and $\varepsilon$, Theorem 1 is the best possible.

One might like to study the dependence of $\delta = \delta(k, \varepsilon)$ on $k$ and, mostly, $\varepsilon$. One could easily deduce an explicit bound (in terms of $k$ and $\varepsilon$) for the function $\delta = \delta(k, \varepsilon)$ in the theorem, but our proof would certainly not give the optimal $\delta(k, \varepsilon)$. For $k > 2$ it seems to be a very difficult problem to find the exact value of the best $\delta(k, \varepsilon)$ (while the case $k = 2$ will be studied in the next section). However, we can prove a qualitative theorem in this direction:

**Theorem 2.** *For every $k \in \mathbb{N}$ with $k \ge 2$ and every $\varphi > 0$ there are $\psi = \psi(k, \varphi) > 0$ and $N_1 = N_1(k, \varphi)$ so that for $N > N_1$ we have*

$$|\mathcal{F}(k, N, 1-\psi)| = |\{E_N : E_N \in \{-1,+1\}^N, C_k(E_N) > (1-\psi)N\}| < 2^{\varphi N}.$$

**Proof of Theorem 2.** The proof is the same as the proof of Theorem 1 but now $\varepsilon$ is replaced by $1-\psi$. Now (2.8) becomes

$$(1-\psi)N < M, \tag{2.15}$$

and the sum in (2.12) is replaced by

$$\sum_{0\le t<\frac{\psi}{2}M} \binom{M}{t}$$

which is less than $2^{\varphi M/2} < 2^{\varphi N/2}$ if $\psi$ is small enough in terms of $\varphi$. Using also (2.15), the result follows easily for small enough $\psi$.

## 3 The special case $k = 2$

In the special case $k = 2$, i.e., in the case of correlation of order 2 we can improve on the results of Section 2 considerably. Indeed, we will be able to determine the asymptotics of the logarithm of the number of the binary sequences $E_N$ belonging to the family

$$\mathcal{F} = \mathcal{F}(2, N, \alpha) = \{E_N : E_N \in \{-1, +1\}^N, C_2(E_N) > \alpha N\} \tag{3.1}$$

(for all $0 < \alpha < 1$).

We will prove:

**Theorem 3.** *If $0 < \alpha < 1$, $\varepsilon > 0$, $N \in \mathbb{N}$, and $N > N_o(\alpha, \varepsilon)$, then writing $F(N, \alpha) = |\mathcal{F}(2, N, \alpha)|$, we have*

$$2^{(\xi((1-\alpha)/2)-\varepsilon)N} < F(N, \alpha) < 2^{(\xi((1-\alpha)/2)+\varepsilon)N}. \tag{3.2}$$

**Proof of Theorem 3.** First we will prove the lower bound in (3.2). Define $s$ by

$$s = \left[\frac{1-\alpha}{2}N\right] - 1. \tag{3.3}$$

Let $\mathcal{G} = \mathcal{G}(N, \alpha)$ denote the family of the sequences $E_N = \{e_1, \dots, e_N\} \in \{-1, +1\}^N$ with the following properties:

(i) $e_1 = 1$, and
(ii) writing $\mathcal{H} = \mathcal{H}(E_N) = \{n : 1 \le n < N, e_{n+1} = -e_n\}$, we have

$$|\mathcal{H}| = s. \tag{3.4}$$

If $E_N \in \mathcal{G}$ then by (3.3) we have

$$\begin{aligned}
C_2(E_N) &\geq \left| \sum_{n=1}^{N-1} e_n e_{n+1} \right| = \left| \sum_{n=1}^{N-1} 1 - \sum_{\substack{1 \leq n < N \\ e_{n+1} = -e_n}} 2 \right| \\
&= \Big| (N-1) - 2|\mathcal{H}| \Big| = N - 1 - 2s \\
&= N - 1 - 2\left( \left[ \frac{1-\alpha}{2} N \right] - 1 \right) > N - 2\left[ \frac{1-\alpha}{2} N \right] \geq N - (1-\alpha)N \\
&= \alpha N \qquad (3.5)
\end{aligned}$$

(since $s \leq (N-1)/2$ by (3.3)) so that $E_N \in \mathcal{F}(2, N, \alpha)$. It follows that $\mathcal{G} \subset \mathcal{F}(2, N, \alpha)$ and whence

$$|\mathcal{G}| \leq |\mathcal{F}(2, N, \alpha)| = F(N, \alpha). \qquad (3.6)$$

Thus it remains to give a lower bound for $|\mathcal{G}|$.

Clearly, the elements of $\mathcal{H}$ determine $E_N \in \mathcal{G}$ uniquely. These elements can be chosen from the integers $1, 2, \ldots, N-1$, and, by (3.4), their number is $s$. Thus $\mathcal{H}$, i.e., a sequence $E_N \in \mathcal{G}$, can be chosen in $\binom{N-1}{s}$ ways, so that

$$|\mathcal{G}| = \binom{N-1}{s}. \qquad (3.7)$$

We will need the following result:

**Lemma 2.** *Let $0 < a < b$ and $\varepsilon > 0$.*

*There exist a positive number $\delta = \delta(a, b, \varepsilon)$ and a positive integer $m_o(a, b, \varepsilon)$ such that if*

$$m > m_o(a, b, \varepsilon),$$

$$|u - bm| < \delta m,$$

*and*

$$|v - am| < \delta m,$$

*then we have*

$$2^{(b\xi(a/b) - \varepsilon)m} < \binom{u}{v} < 2^{(b\xi(a/b) + \varepsilon)m}.$$

**Proof of Lemma 2.** This is Lemma 2 in [8], and it can be proved easily by using Stirling's formula (it is also well-known in Information Theory and Statistical Physics).

By using Lemma 2 with $N$, $\frac{1-\alpha}{2}$, 1, $N-1$, and $s$ in place of $m$, $a$, $b$, $u$, and $v$, respectively, it follows from (3.3), (3.6), and (3.7) that for $N > N_o(\varepsilon)$ we have

$$F(N,\alpha) \geq |\mathcal{G}| = \binom{N-1}{s} > 2^{(\xi((1-\alpha)/2)-\varepsilon)N},$$

and this proves the lower bound in (3.2).

The upper bound in (3.2) will be proved by using a simple elementary version of the saddle point method (readers familiar with Information Theory know the exponential growth of such quantities).

By Lemma 1 we have

$$F(N,\alpha) = |\mathcal{F}(2,N,\alpha)| \leq 2N^3 \max_{\alpha N < M < N} 2^{N-M} \sum_{0 \leq t < \frac{M-\alpha N}{2}} \binom{M}{t}.$$

Define $\beta$ by $M = \beta N$ so that

$$\alpha < \beta < 1, \tag{3.8}$$

and write

$$x = \frac{\beta+\alpha}{\beta-\alpha} (> 1).$$

Then

$$\begin{aligned} F(N,\alpha) &\leq 2N^3 \sup_{\alpha<\beta<1} 2^{(1-\beta)N} \sum_{0 \leq t < \frac{\beta-\alpha}{2}N} \binom{[\beta N]}{t} x^{\frac{\beta-\alpha}{2}N-t} \\ &\leq 2N^3 \sup_{\alpha<\beta<1} 2^{(1-\beta)N} \sum_{0 \leq t \leq [\beta N]} \binom{[\beta N]}{t} x^{\frac{\beta-\alpha}{2}N-t} \\ &\leq 2N^3 \sup_{\alpha<\beta<1} 2^{(1-\beta)N} \left(1+\frac{1}{x}\right)^{\beta N} x^{\frac{\beta-\alpha}{2}N} \\ &\leq 2N^3 \sup_{\alpha<\beta<1} \exp(g(\alpha,\beta)N), \end{aligned} \tag{3.9}$$

where we have

$$\begin{aligned} g(\alpha,\beta) &= (1-\beta)\log 2 + \beta \log \frac{2\beta}{\beta+\alpha} + \frac{\beta-\alpha}{2} \log \frac{\beta+\alpha}{\beta-\alpha} \\ &= \log 2 + \beta \log \beta - \frac{1}{2}(\beta-\alpha)\log(\beta-\alpha) - \frac{1}{2}(\beta+\alpha)\log(\beta+\alpha). \end{aligned}$$

By (3.8), for any fixed $0 < \alpha < 1$ we have

$$\begin{aligned}\frac{\partial g(\alpha,\beta)}{\partial\beta} &= \log\beta - \frac{1}{2}\log(\beta-\alpha) - \frac{1}{2}\log(\beta+\alpha)\\ &= \log\beta - \frac{1}{2}\log(\beta^2-\alpha^2) > \log\beta - \frac{1}{2}\log\beta^2 = 0,\end{aligned}$$

and $g(\alpha,\beta)$ is continuous in $\alpha < \beta \leq 1$. It follows that we have

$$\begin{aligned}&g(\alpha,\beta) \leq g(\alpha,1) = \log 2 - \frac{1-\alpha}{2}\log(1-\alpha) - \frac{1+\alpha}{2}\log(1+\alpha)\\ &= (\log 2)\xi\big((1-\alpha)/2\big).\end{aligned} \tag{3.10}$$

By (3.9) and (3.10) we have

$$F(N,\alpha) \leq 2N^3 2^{\xi((1-\alpha)/2)N} < 2^{(\xi((1-\alpha)/2)+\varepsilon)N}$$

if $N$ is large enough in terms of $\varepsilon$, which completes the proof of Theorem 3.

## 4 On an addition lemma

Let $\mathbb{Z}_m$ denote the ring of the modulo $m$ residue classes.

There is a lemma whose variants played a crucial role in the estimate of the **correlation** in several important constructions [1, 3, 4]:

**Lemma 3.** *If $p$ is a prime number, $k, \ell \in \mathbb{N}$,*

$$(4\ell)^k < p, \tag{4.1}$$

*$\mathcal{A}$, $\mathcal{B} \subset \mathbb{Z}_p$, $|\mathcal{A}| = k$, and $|\mathcal{B}| = \ell$, then there is a $c \in \mathbb{Z}_p$ which has a unique representation in the form*

$$a + b = c, \quad a \in \mathcal{A},\ b \in \mathcal{B}. \tag{4.2}$$

Any improvement on condition (4.1) would lead to a similar improvement on the upper bounds for the correlation in the papers mentioned above, and it was believed that (4.1) is far from being best possible. Thus Mauduit and Sárközy proposed to try to improve on (4.1). Now we will show that they were wrong and, assuming that there are infinitely many Mersenne primes, i.e., primes of the form $p = 2^q - 1$ where $q$ is also a prime, (4.1) is nearly best possible. Probably there are infinitely many Mersenne primes, but at the present it is hopeless to prove this. Indeed, it would be nearly equally satisfactory to prove that there are infinitely many primes $q$ such that $2^q - 1$ has a large prime factor $p > 2^{\varepsilon q}$ since the following Theorem 4 can be generalized easily to such a $p$.

**Theorem 4.** *If $p = 2^q - 1$ is a Mersenne prime, then there are $\mathcal{A}, \mathcal{B} \subset \mathbb{Z}_p$ such that*

$$|\mathcal{A}| = q, \quad |\mathcal{B}| = q + 1 \tag{4.3}$$

*and there is no $c \in \mathbb{Z}_p$ which has a unique representation in the form (4.2).*

Note that it follows from (4.3) that

$$|\mathcal{B}| > |\mathcal{A}| = q = \frac{\log(p+1)}{\log 2}. \tag{4.4}$$

On the other hand, assume that

$$k, \ell < \frac{\log p}{\log \log p}$$

which is smaller than (4.4) only by a factor $c \log \log p$. Then if $p$ is large enough ($p > e^{e^4}$),

$$\log(4\ell)^k < \frac{\log p}{\log \log p}(\log 4 + \log \log p - \log \log \log p) < \log p$$

so (4.1) is satisfied and Lemma 3 applies. This shows that condition (4.1) is close to optimal.

**Proof of Theorem 4.** Let $\mathcal{A} = \{1, 2, 4, \ldots, 2^{q-1}\}$, $\mathcal{B} = \{0\} \cup \mathcal{A}$. Then (4.3) holds trivially. If $c \in \{2^i + 2^j : 0 \le i < j \le q - 1\}$, then $c$ has exactly two representations in the form (4.2):

$$c = 2^i + 2^j = 2^j + 2^i.$$

If $c \in \{2^i : 0 \le i \le q - 1\}$, then again $c$ has exactly two representations in the form (4.2):

$$c(= 2^i) = 2^{i-1} + 2^{i-1} = 2^i + 0 \quad \text{for} \quad 1 \le i \le q - 1$$

and

$$c(= 1) = 1 + 0 = 2^{q-1} + 2^{q-1} \quad \text{for} \quad i = 0$$

(in $\mathbb{Z}_p$). If, finally,

$$c \in \mathbb{Z}_p \setminus (\{2^i + 2^j : 0 \le i < j \le q - 1\} \cup \{2^i : 0 \le i \le q - 1\}),$$

then $c$ cannot be represented in the form (4.2). Thus for every $c \in \mathbb{Z}_p$, (4.2) has either 0 or 2 solutions, which completes the proof of the theorem.

## References

[1] R. Ahlswede, C. Mauduit, and A. Sárközy, Large families of pseudorandom sequences of $k$ symbols and their complexity, II, General Theory of Information Transfer and Combinatorics, Lecture Notes in Computer Science, Vol. 4123, Springer Verlag, 2006.

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

[3] L. Goubin, C. Mauduit, and A. Sárközy, Construction of large families of pseudorandom binary sequences, J. Number Theory, 106, 56–69, 2004.

[4] K. Gyarmati, On a family of pseudorandom binary sequences, Period. Math. Hungar., 49, 45–63, 2004.

[5] Y. Kohayakawa, C. Mauduit, C. G. Moreira, and V. Rödl, Measures of pseudorandomness for finite sequences: minimum and typical values, Proceedings of WORDS'03, 159–169, TUCS Gen. Publ., 27, Turku Cent. Comput. Sci., Turku, 2003.

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

[7] A. Menezes, P. van Oorschot, and R. Vanstone, Handbook of Applied Cryptography, CRC Press, Inc., 1997.

[8] A. Sárközy, Some metric problems in the additive number theory, II, Annales Univ. Sci. Budapest. Eötvös 20, 111–129, 1977.