The cross-correlation measure for families of binary sequences

Katalin Gyarmati
Eötvös Loránd University
Department of Algebra and Number Theory
and MTA-ELTE Geometric and Algebraic Combinatorics Research Group
H-1117 Budapest, Pázmány Péter sétány 1/C, Hungary
E-mail: gykati@cs.elte.hu

and

Christian Mauduit
Université Aix-Marseille
Institut de Mathématiques de Luminy
CNRS, FRE3529,
163 avenue de Luminy, 13288 Marseille cedex 9, France
E-mail: mauduit@iml.univ-mrs.fr

and

András Sárközy
Eötvös Loránd University
Department of Algebra and Number Theory
H-1117 Budapest, Pázmány Péter sétány 1/C, Hungary
E-mail: gykati@cs.elte.hu and sarkozy@cs.elte.hu

Dedicated to Professor Harald Niederreiter on the occasion of his 70th birthday

2010 Mathematics Subject Classification: Primary 11K45.

Keywords and phrases: pseudorandom, cross-correlation, families of binary sequences, family complexity, Legendre symbol.

Research partially supported by ERC-AdG.228005, Hungarian National Foundation for Scientific Research, Grants No. and K100291 and NK104183, the János Bolyai Research Fellowship, the Agence Nationale de la Recherche grant ANR-10-BLAN 0103 called MUNUM and the MTA-ELTE Geometric and Algebraic Combinatorics Research Group.

Abstract

Large families of binary sequences of the same length are considered and a new measure, the cross-correlation measure of order k is introduced to study the connection between the sequences belonging to the family. It is shown that this new measure is related to certain other important properties of families of binary sequences. Then the size of the cross-correlation measure is studied. Finally, the cross-correlation measures of two important families of pseudorandom binary sequences are estimated.

1 Introduction

Pseudorandom binary sequences have many applications, in particular, they play a crucial role in modern cryptography. The pseudorandomness of the individual binary sequences is usually characterized by using the notion of linear complexity, and tests based on mathematical statistics ("poker test", "runs test", etc.) are also used. However, these requirements usually study just a single property of the sequence, and they also have other weak points. Thus recently a more comprehensive theory of pseudorandomness of binary sequences has been initiated by Mauduit and Sárközy [34]. They introduced the following notations and definitions:

Consider a binary sequence

$$E_N = (e_1, \dots, e_N) \in \{-1, +1\}^N.$$

Then 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 with $a, b, t \in \mathbb{N}$, $1 \leq a \leq a + (t-1)b \leq N$, while the *correlation measure of order* k of E_N is defined as

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

where the maximum is taken over all $D = (d_1, \dots, d_k)$ with non-negative integers $d_1 < \cdots < d_k$ and $M \in \mathbb{N}$ with $M + d_k \leq N$.

Then E_N is considered a "good" pseudorandom sequence if both of 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 $N \to \infty$). Indeed, later Cassaigne, Mauduit and Sárközy [9] showed that this terminology is justified since for almost all $E_N \in \{-1,+1\}^N$ both $W(E_N)$ and $C_k(E_N)$ are less than $N^{1/2}(\log N)^c$ (and they are also greater than $\varepsilon N^{1/2}$; see also [5], [6] and [25]). It was also shown in [34] that the Legendre symbol forms a "good" pseudorandom sequence. Since that many constructions have been given for binary sequences with strong pseudorandom properties (see, e.g., [10], [11], [16], [17], [29], [30], [31], [36], [37], [39]).

However, these "good" constructions produce only a "few" good sequences while in many applications, e.g., in cryptography one needs *"large" families* of "good" pseudorandom binary sequences. If these sequences are constructed by an algorithm, then we usually speak of pseudorandom generator, and the algorithm is considered a "good" one if it satisfies the "next bit" test; this approach has certain weak points. Large families consisting of binary sequences which are "good" in terms of the pseudorandom measures defined above have been also constructed; see, e.g., [19], [20], [27], [28], [31], [32], [38]. In these constructions it is guaranteed that the individual sequences belonging to the family possess strong pseudorandom properties. However, in many applications it is not enough to know this; it can be much more important to know that the given family has a "rich", "complex" structure, there are many "independent" sequences in it. In order to handle this requirement Ahlswede, Khachatrian, Mauduit and Sárközy [2] (see also [3], [4], [22], [33]) introduced the notion of *family complexity* or briefly *f-complexity* (which can be especially useful in cryptography):

Definition 1 *The f-complexity $\Gamma(\mathcal{F})$ of a family $\mathcal{F}$ of binary sequences $E_N \in \{-1,+1\}^N$ is defined as the greatest integer j so that for any* specification

$$e_{i_1} = \varepsilon_1, \ \ldots, \ e_{i_j} = \varepsilon_j \ (1 \le i_1 < \cdots < i_j \le N)$$

(with $\varepsilon_1, \ldots, \varepsilon_j \in \{-1,+1\}$) there is at least one $E_N = (e_1, \ldots, e_N) \in \mathcal{F}$ which satisfies it. The f-complexity of $\mathcal{F}$ is denoted by $\Gamma(\mathcal{F})$. (If there is no $j \in \mathbb{N}$ with the property above then we set $\Gamma(\mathcal{F}) = 0$.)

There are also other properties of families which play an important role in the applications. Such a property is the existence of *collisions* in the given family. This notion appears, e.g., in [8], [35], [40], [41]; we will follow here Tóth's [40] presentation. Assume that $N \in \mathbb{N}$, $\mathcal{S}$ is a given set (e.g., a set of certain polynomials or the set of all the binary sequences of a given length much less than N), to each $s \in S$ we assign a unique binary sequence

$$E_N = E_N(s) = (e_1, \ldots, e_N) \in \{-1,+1\}^N,$$

and let $\mathcal{F} = \mathcal{F}(\mathcal{S})$ denote the family of the binary sequences obtained in this way:

$$\mathcal{F} = \mathcal{F}(\mathcal{S}) = \{E_N(s) : \ s \in \mathcal{S}\}. \tag{1}$$

Definition 2 *If $s \in \mathcal{S}$, $s' \in \mathcal{S}$, $s \neq s'$ and*

$$E_N(s) = E_N(s'), \tag{2}$$

then (2) is said to be a *collision* in $\mathcal{F} = \mathcal{F}(\mathcal{S})$. If there is no collision in $\mathcal{F} = \mathcal{F}(\mathcal{S})$, then $\mathcal{F}$ is said to be *collision free.*

In other words, $\mathcal{F} = \mathcal{F}(\mathcal{S})$ is collision free if we have $|\mathcal{F}| = |\mathcal{S}|$. An ideally good family of pseudorandom binary sequences is collision free.

There is another related notion appearing in the literature, namely, the notion of *avalanche effect* (see, e.g., [8], [15], [24], [40], [41]); here we will present Tóth's definition):

Definition 3 *If $\mathcal{F}$ is a family of form* (1)*, and for any $s \in \mathcal{S}$ changing s for any $s' \in \mathcal{S}$ with $s' \neq s$ changes "many" elements of $E_N(s)$ (i.e., for $s \neq s'$ many elements of the sequences $E_N(s)$ and $E_N(s')$ are different), then we speak about* avalanche effect*, and we say that $\mathcal{F} = \mathcal{F}(\mathcal{S})$ possesses the* avalanche property*. If for any $s, s' \in \mathcal{S}$, $s \neq s'$ at least $\left(\frac{1}{2} - o(1)\right) N$ elements of $E_N(s)$ and $E_N(s')$ are different then $\mathcal{F}$ is said to possess the* strict avalanche property*.*

We will also need

Definition 4 *If $N \in \mathbb{N}$, $E_N = (e_1, \dots, e_N) \in \{-1, +1\}^N$ and $E'_N = (e'_1, \dots, e'_N) \in \{-1, +1\}^N$, then the* distance $d(E_N, E'_N)$ *between E_N and E'_N is defined by*

$$d(E_N, E'_N) = |\{n : \ 1 \leq n \leq N, \ e_n \neq e'_n\}|$$

(so that $d(E_N, E'_N)$ is a variant of the Hamming distance). Moreover if $\mathcal{F}$ is a family of form (1)*, then the* distance minimum $m(\mathcal{F})$ *of $\mathcal{F}$ is defined by*

$$m(\mathcal{F}) = \min_{\substack{s, s' \in \mathcal{S} \\ s \neq s'}} d\left(E_N(s), E_N(s')\right).$$

Applying this notion we may say that the family $\mathcal{F}$ in (1) *is collision free if and only if $m(\mathcal{F}) > 0$, and $\mathcal{F}$ possesses the strict avalanche property if*

$$m(\mathcal{F}) \geq \left(\frac{1}{2} - o(1)\right) N. \tag{3}$$

The notions introduced in Definitions 3 and 4 can be also used when the family $\mathcal{F}$ is not of form (1), i.e., no parameter set $\mathcal{S}$ is given; e.g., we may say that $\mathcal{F}$ possesses the avalanche property if for any $E_N \in \mathcal{F}$, $E'_N \in \mathcal{F}$, $E_N \neq E'_N$ the sequences E_N and E'_N have many different elements, and the distance minimum can be defined as

$$m(\mathcal{F}) = \min_{\substack{E_N, E'_N \in \mathcal{F} \\ E_N \neq E'_N}} d(E_N, E'_N).$$

We will use these notions in this extended sense.

In this paper our goal is to study a further important property of families of binary sequences. First in Section 2 we will introduce a measure called *cross-correlation measure*, and we will study the connection between this new measure and the other related notions listed above. Then in Section 3 we will study the connection between the size of the family and its cross-correlation measure. Finally, in Sections 4 and 5 we will estimate the cross-correlation of two important families of pseudorandom binary sequences.

2 The definition of the cross-correlation measure

In Section 1 we mentioned $C_k(E_N)$, the correlation measure of order k of the binary sequence E_N which is, perhaps, the most important measure of pseudorandomness of a *single* binary sequence; in the definition of this measure we consider a fixed sequence and we compare different elements of it (so that this is an *autocorrelation* type quantity). If, instead of a *single* sequence we want to characterize a *family* of sequences, then it is quite natural to compare elements of *different* sequences taken from the family, i.e., to consider a *correlation* type quantity involving different sequences. Thus we suggest to use the following definition:

Definition 5 *Let $N \in \mathbb{N}$, $k \in \mathbb{N}$, and for any k binary sequences $E_N^{(1)}, \dots, E_N^{(k)}$ with*

$$E_N^{(i)} = \left(e_1^{(i)}, \dots, e_N^{(i)}\right) \in \{-1, +1\}^N \text{ (for } i = 1, 2, \dots, k)$$

and any $M \in \mathbb{N}$ and k-tuple $D = (d_1, \dots, d_k)$ of non-negative integers with

$$0 \leq d_1 \leq \dots \leq d_k < M + d_k \leq N, \tag{4}$$

write

$$V_k\left(E_N^{(1)},\dots,E_N^{(k)},M,D\right)=\sum_{n=1}^{M} e_{n+d_1}^{(1)}\cdots e_{n+d_k}^{(k)} \tag{5}$$

Let

$$\tilde{C}_k\left(E_N^{(1)},\dots,E_N^{(k)}\right)=\max_{M,D}\left|V_k\left(E_N^{(1)},\dots,E_N^{(k)},M,D\right)\right| \tag{6}$$

where the maximum is taken over all $D=(d_1,\dots,d_k)$ and $M\in\mathbb{N}$ satisfying (4) *with the additional restriction that if $E_N^{(i)}=E_N^{(j)}$ for some $i\neq j$, then we must not have $d_i=d_j$. Then the* cross-correlation measure of order k *of the family $\mathcal{F}$ of binary sequences $E_N\in\{-1,+1\}^N$ is defined as*

$$\Phi_k(\mathcal{F})=\max \tilde{C}_k\left(E_N^{(1)},\dots,E_N^{(k)}\right) \tag{7}$$

where the maximum is taken over all k-tuples of binary sequences $\left(E_N^{(1)},\dots,E_N^{(k)}\right)$ with

$$E_N^{(i)}\in\mathcal{F} \text{ for } i=1,\dots,k.$$

(Note that other cross-correlation type quantities also occur in [7], [18], [21].)

Then observe first that by the definition of $\tilde{C}_k$, for every $E_N\in\{-1,+1\}^N$ we have

$$\tilde{C}_k\left(E_N,\dots,E_N\right)=C_k(E_N),$$

thus it follows from (7) that

Proposition 1 *We have*

$$\Phi_k(\mathcal{F})\geq \max_{E_N\in\mathcal{F}} C_k(E_N). \tag{8}$$

This means that an upper bound for the cross-correlation of order k of the *family* $\mathcal{F}$ is also an upper bound for correlation of order k of *every sequence* $E_N\in\mathcal{F}$. Thus it suffices to estimate $\Phi_k(\mathcal{F})$: if we have a "good" upper bound for $\Phi_k(\mathcal{F})$, then this guarantees that $\mathcal{F}$ consists of *sequences possessing strong pseudorandom properties.*

In Section 1 we said that in the applications it is "important to know that given family has a rich, complex structure, there are many independent sequences in it". Can one use the cross-correlation measure of a family to show that, indeed, this is the case? We will show that already the small cross-correlation measure of order 2 is enough to guarantee that the sequences in the family are far apart (literally):

Proposition 2 *If $N \in \mathbb{N}$ and $E_N = (e_1, \dots, e_N) \in \mathcal{F}$, $E'_N = (e'_1, \dots, e'_N) \in \mathcal{F}$, $\mathcal{F} \subset \{-1, +1\}^N$, then we have*

$$\left| d(E_N, E'_N) - \frac{N}{2} \right| \leq \frac{1}{2}\tilde{C}_2(E_N, E'_N) \leq \frac{1}{2}\Phi_2(\mathcal{F}). \tag{9}$$

Proof. Clearly we have

$$\frac{(e_n - e'_n)^2}{4} = \begin{cases} 0 & \text{if } e_n = e'_n \\ 1 & \text{if } e_n \neq e'_n \end{cases} \quad \text{for } n = 1, 2, \dots, N$$

thus

$$d(E_N, E'_N) = \sum_{n=1}^{N} \frac{(e_n - e'_n)^2}{4} = \frac{N}{2} - \frac{1}{2}\sum_{n=1}^{N} e_n e'_n$$

whence, by (5), (6) and (7),

$$\left| d(E_N, E'_N) - \frac{N}{2} \right| = \frac{1}{2}\left| \sum_{n=1}^{N} e_n e'_n \right| \leq \frac{1}{2}\tilde{C}_2(E_N, E'_N) \leq \Phi_2(\mathcal{F})$$

which proves (9).

If the cross-correlation of order 2 of the family $\mathcal{F} \subseteq \{-1, +1\}^N$ is $o(N)$:

$$\Phi_2(\mathcal{F}) = o(N), \tag{10}$$

then it follows from Definition 4, (9) and (10) that

$$m(\mathcal{F}) = \min_{\substack{E_N, E'_N \in \mathcal{F} \\ E_N \neq E'_N}} d(E_N, E'_N) \geq \frac{N}{2} - \frac{1}{2}\Phi_2(\mathcal{F}) = \frac{N}{2} - o(N)$$

so that (3) holds. This proves

Proposition 3 *If $N \in \mathbb{N}$, $\mathcal{F} \subset \{-1, +1\}^N$ and* (10) *holds then the family $\mathcal{F}$ possesses the strict avalanche property.*

So far we have seen that there is a close connection between collision, distance minimum and avalanche property in a family of binary sequences on one hand and its cross-correlation on the other hand. It remains to see whether there is any connection between the family complexity of a family and its cross-correlation. We will show by two examples that these two measures are independent in the sense that it may occur that a family $\mathcal{F}$ is "good" concerning its family complexity, i.e., $\Gamma(\mathcal{F})$ is large but it is "bad" considering its cross-correlation, i.e., $\Phi_k(\mathcal{F})$ is also large for every small k;

on the other hand it is also possible that $\mathcal{F}$ is considered "good" since $\Phi_k(\mathcal{F})$ is small, however $\mathcal{F}$ is "bad" concerning its small family complexity. (This means that it is not enough to study only one of $\Gamma(\mathcal{F})$ and $\Phi_k(\mathcal{F})$, we have to estimate both of them.)

Example 1 *Let $N \in \mathbb{N}$ and let $\mathcal{F}$ be the set of all the binary sequences of length N: $\mathcal{F} = \{-1, +1\}^N$. Then clearly $\Gamma(\mathcal{F})$ is maximal: $\Gamma(\mathcal{F}) = N$. On the other hand, $E_N = (e_1, \ldots, e_N) = (1, \ldots, 1) \in \mathcal{F}$ thus by (8), for $k \in \mathbb{N}$, $k \leq N$ we have*

$$\Phi_k(\mathcal{F}) \geq C_k(E_N) = \sum_{n=1}^{N-k+1} e_n e_{n+1} \cdots e_{n+k-1} = \sum_{n=1}^{N-k+1} 1 = N - k + 1$$

which is also large.

Example 2 *Consider any family $\mathcal{F}$ of binary sequences of length N with small cross-correlation of order k for any small k; e.g., we may take $N = p =$ prime and the family $\mathcal{F}_1$ which will be constructed later in Theorem 1 and which will satisfy the inequality*

$$\Phi_k(\mathcal{F}_1) < 10kdp^{1/2} \log p$$

(for any $1 < k < p$). Then for at least half of the sequences $E_p = (e_1, \ldots, e_p) \in \mathcal{F}_1$ either $e_1 = +1$ or $e_1 = -1$ holds; we may assume that the first equality. Then let $\mathcal{F}_1' = \{E_p = (e_1, \ldots, e_p) : \ e_1 = +1\}$ so that $|\mathcal{F}_1'| \geq \frac{|\mathcal{F}_1|}{2}$, we have

$$\Phi_k(\mathcal{F}_1') \leq \Phi_k(\mathcal{F}_1) < 10kdp^{1/2} \log p$$

(which is small) and

$$\Gamma\left(\mathcal{F}_1'\right) = 0$$

(which is also small) since there is no $E_p = (e_1, \ldots, e_p) \in \mathcal{F}_1'$ satisfying the specification

$$e_1 = -1.$$

3 The size of the cross-correlation measure

When we introduce a new pseudorandom measure of sequences or a new family measure, then it is a question of basic importance that what is the expected size of the new measure, and what is the size that we hope to achieve? In case of the measures of pseudorandomness of binary *sequences* our starting point was the study of the behaviour of a truly random binary sequence of a given length N. In case of families the situation is more complex: usually not only the length N of the sequences is given but also the size of the family $\mathcal{F}$ plays an important role. Our optimal goal is usually to construct a possible large family $\mathcal{F}$ of "good" pseudorandom binary sequences with the property that it possesses the strong avalanche property, i.e., (3) holds: $m(\mathcal{F}) \geq \left(\frac{1}{2} - o(1)\right) N$, and by Proposition 3 this is the case if (9) holds: $\Phi_2(\mathcal{F}) = o(N)$. It follows from the results of coding theory [42] that requirement (3) can hold for an $\mathcal{F}$ with $|\mathcal{F}| > 2^{c_1 N}$ with some $0 < c_1 < 1/2$ (e.g., $c_1 = 0.11$ can be taken) but it is known that there is a c_2 such that $c_1 < c_2 < 1/2$ and a family with $|\mathcal{F}| > 2^{c_2 N}$ cannot satisfying (3) (e.g., $c_2 = 0.18$ can be taken). If we relax (3), then the size of $\mathcal{F}$ may grow. However, one should not forget that the sequences in $\mathcal{F}$ must also possess strong pseudorandom properties; it is not at all easy to combine this requirement with (3) and a good lower bound for $|\mathcal{F}|$. In the practice it is quite satisfactory to construct a family $\mathcal{F}$ with $|\mathcal{F}| > \exp\left(N^{c_1}\right)$, $\Phi_k(\mathcal{F}) < N^{c_2}$ (for all small k) with some $0 < c_1 < 1$, $c_2 < 1$ (note that by (8) it also follows from the upper bound for $\Phi_k(\mathcal{F})$ that every $E_N \in \mathcal{F}$ possesses small correlations of small order). It remains to present constructions for families with these properties. This will be done in Sections 4 and 5, but first we will study the extremal values of $\Phi_k(\mathcal{F})$. (One also might like to study the behaviour of the cross-correlation measures for a truly random family of given size. This seems to be a rather difficult task; perhaps we will return to this problem in a subsequent paper.) It was shown in [5], [25] that for $N \in \mathbb{N}$, $k \in \mathbb{N}$ we have

$$\min_{E_N \in \{-1,+1\}^N} C_{2k}(E_N) > \left(\frac{1}{2}\left[\frac{N}{k+1}\right]\right)^{1/2}.$$

By (8) the same lower bound can be given for $\Phi_{2k}(\mathcal{F})$. On the other hand, it was shown in [9] that for all $N \in \mathbb{N}$, $k \in \mathbb{N}$, $2k+1 < N$ we have

$$\min_{E_N \in \{-1,+1\}^N} C_{2k+1}(E_N) = 1. \tag{11}$$

It is a natural question to ask: what about the extremal values of $\Phi_{2k+1}(\mathcal{F})$? If E'_N denotes the binary sequence of N whose every element is $+1$, $\mathcal{F}$ contains the sequence E'_N, and $2k+1 < N$, then by (8) we have

$$\Phi_{2k+1}(\mathcal{F}) = C_{2k+1}\left(E'_N\right) = N - 2k.$$

On the other hand, if $2k+1 < N$ then by (11) there is a binary sequence E''_N of length N with $C_{2k+1}(E''_N) = 1$. If $\mathcal{F}$ consists of the single sequence E''_N

then we have

$$\Phi_{2k+1}(\mathcal{F}) = C_{2k+1}\left(E_N''\right) = 1.$$

But can $\Phi_{2k+1}(\mathcal{F})$ be also small for greater families? We will show that the answer is affirmative if $|\mathcal{F}|$ is "much smaller" than N, the length of the sequences in $\mathcal{F}$ (but we do not know what happens in larger families):

Proposition 4 *Let $N \in \mathbb{N}$, $k \in \mathbb{N}$, $2k+1 < N$, $\ell \in \mathbb{N}$ and $\ell < N$. For $i = 1, \ldots, \ell$, define the binary sequence $E_N^{(i)} = \left(e_1^{(i)}, \ldots, e_N^{(i)}\right)$ of length N by*

$$e_n^{(i)} = (-1)^{\left[\frac{n+i}{\ell}\right]} \text{ for } n = 1, \ldots, N,$$

and let $\mathcal{F}$ be the family $\mathcal{F} = \left\{E_N^{(1)}, \ldots, E_N^{(\ell)}\right\}$. Then we have

$$\Phi_{2k+1}(\mathcal{F}) \le 2\ell. \tag{12}$$

Proof. Using notation (5), for any M, $1 \le i_1, \ldots, i_{2k+1} \le \ell$ and $2k+1$-tuple $D = (d_1, \ldots, d_{2k+1})$ satisfying (4) (with $2k+1$ in place of k) we have

$$\begin{aligned}
&\left|V_{2k+1}\left(E_N^{(i_1)}, \ldots, E_N^{(i_{2k+1})}, M, D\right)\right| = \left|\sum_{n=1}^{M} e_{n+d_1}^{(i_1)} e_{n+d_2}^{(i_2)} \cdots e_{n+d_{2k+1}}^{(i_{2k+1})}\right| \\
&\quad = \left|\sum_{n=1}^{M} (-1)^{\left[\frac{n+d_1+i_1}{\ell}\right]} \cdots (-1)^{\left[\frac{n+d_{2k+1}+i_{2k+1}}{\ell}\right]}\right| \\
&\quad = \left|\sum_{r=1}^{\ell} \sum_{m=0}^{\left[\frac{M}{\ell}\right]-1} (-1)^{\left[\frac{\ell m+r+d_1+i_1}{\ell}\right]} \cdots (-1)^{\left[\frac{\ell m+r+d_{2k+1}+i_{2k+1}}{\ell}\right]}\right. \\
&\qquad \left. + \sum_{\ell\left[\frac{M}{\ell}\right] < n \le M} (-1)^{\left[\frac{n+d_1+i_1}{\ell}\right]} \cdots (-1)^{\left[\frac{n+d_{2k+1}+i_{2k+1}}{\ell}\right]}\right| \\
&\quad \le \left|\sum_{r=1}^{\ell} \sum_{m=0}^{\left[\frac{M}{\ell}\right]-1} (-1)^{m+\left[\frac{r+d_1+i_1}{\ell}\right]} \cdots (-1)^{m+\left[\frac{r+d_{2k+1}+i_{2k+1}}{\ell}\right]}\right| + \sum_{\ell\left[\frac{M}{\ell}\right] < n \le M} 1 \\
&\quad \le \left|\sum_{r=1}^{\ell} (-1)^{\left[\frac{r+d_1+i_1}{\ell}\right]+\cdots+\left[\frac{r+d_{2k+1}+i_{2k+1}}{\ell}\right]} \sum_{m=0}^{\left[\frac{M}{\ell}\right]-1} (-1)^{(2k+1)m}\right| + \ell
\end{aligned}$$

$$\leq \left|\sum_{r=1}^{\ell} 1\right| \left|\sum_{m=0}^{[\frac{M}{\ell}]-1} (-1)^m\right| + \ell \leq \ell \cdot 1 + \ell = 2\ell. \tag{13}$$

(12) follows from (6), (7) and (13).

We do not know what happens in larger families:

Problem 1 *Estimate* $\min \Phi_{2k+1}(\mathcal{F})$ *for any fixed* N, k *and* $|\mathcal{F}|$.

4 A family with small cross-correlation constructed by using the Legendre symbol

The first construction for large families of binary sequences with strong pseudorandom properties (in terms of the measures described in Section 1) was given by Goubin, Mauduit and Sárközy [19] and it used the Legendre symbol (this is, perhaps, still the best construction of this type). They proved

Theorem A. *If* p *is a prime number,* $f(x) \in \mathbb{F}_p[x]$ *has degree* d (>0)*,* $f(x)$ *has no multiple zero in* $\overline{\mathbb{F}}_p$*, and the binary sequence* $E_p = E_p(f) = (e_1, \dots, e_p)$ *is defined by*

$$e_n = \begin{cases} \left(\frac{f(n)}{p}\right) & \text{for } (f(n), p) = 1, \\ +1 & \text{for } p \mid f(n), \end{cases} \qquad (for\ n = 1, 2, \dots, p) \tag{14}$$

then we have

$$W(E_p) < 10dp^{1/2} \log p,$$

and if either
(i) $k = 2$*,*
(ii) 2 *is primitive root modulo* p *and* $k < p$*, or*
(iii) we have

$$k < \frac{p^{1/d}}{4}, \tag{15}$$

then

$$C_k(E_p) < 10kdp^{1/2} \log p$$

also holds.

Indeed, this is a combination of Theorems 1 and 2 in [19]. (Note that (15) is a corrected form of the inequality appearing in Corollary 2, (ii) in [19]; namely, there the exponent of p is $1/4$, while the right exponent provided by the proof is $1/d$ as in (15).)

Let $\mathcal{F}$ denote the family of the binary sequences $E_p(f)$ assigned to the polynomials satisfying the conditions in Theorem A. In Sections 4 and 5 we will show two different ways to modify the definition of the family slightly so that one should also have reasonable control over the cross-correlations of the family.

Theorem 1 *Let $d \in \mathbb{N}$, p a prime number, $d < p$, consider all the irreducible polynomials $f(x) \in \mathbb{F}_p[x]$ of the form*

$$f(x) = x^d + a_2 x^{d-2} + a_3 x^{d-3} + \cdots + a_d \tag{16}$$

(so that there is no x^{d-1} term) and let $\mathcal{F}_1$ denote the family of the binary sequences $E_p = E_p(f)$ assigned to these polynomials f by the formula (14). *Then we have*
(i)

$$\Phi_k(\mathcal{F}_1) < 10kdp^{1/2}\log p \tag{17}$$

for all $k \in \mathbb{N}$, $1 < k < p$, and
(ii) if $d < p^{1/2}/20\log p$, then

$$|\mathcal{F}_1| \geq p^{[d/3]-1}. \tag{18}$$

Proof. (i) By using the notations in Definition 5, we have to estimate

$$\left|V_k\left(E_p^{(1)}, \ldots, E_p^{(k)}, M, D\right)\right| = \left|\sum_{n=1}^{M} e_{n+d_1}^{(1)} \ldots e_{n+d_k}^{(k)}\right|$$

for

$$E_p^{(i)} = E_p^{(i)}(f_i) \in \mathcal{F}_1 \quad (i = 1, 2, \ldots, k)$$

and M, D satisfying the conditions in Definition 5. Clearly,

$$f_i(n + d_i) \equiv 0 \pmod{p},\ 1 \leq n \leq M,\ 1 \leq i \leq k$$

has at most dk solutions (in pairs (n, i)). Thus defining $\left(\frac{a}{p}\right)$ as 0 for $p \mid a$, we have

$$\begin{aligned}\left|V_k\left(E_p^{(1)}, \ldots, E_p^{(k)}, M, D\right)\right| &= \left|\sum_{n=1}^{M} e_{n+d_1}^{(1)} \ldots e_{n+d_k}^{(k)}\right| \\ &\leq \left|\sum_{n=1}^{M} \left(\frac{f_1(n+d_1)}{p}\right) \cdots \left(\frac{f_k(n+d_k)}{p}\right)\right| + dk\end{aligned}$$

$$= \left| \sum_{n=1}^{M} \left(\frac{f_1(n+d_1) \cdots f_k(n+d_k)}{p} \right) \right| + dk. \quad (19)$$

If for some $1 \le i < j \le k$ we have $f_i(x) \ne f_j(x)$, then

$$f_i(x+d_i) \ne f_j(x+d_j) \quad (20)$$

since both f_i and f_j are of form (16). If $1 \le i < j \le k$ and $f_i(x) = f_j(x)$, then by the conditions on D in Definition 5 we cannot have $d_i = d_j$, thus again (20) holds. Then writing

$$h(x) = f_1(x+d_1) \cdots f_k(x+d_k), \quad (21)$$

this polynomial is the product of k distinct monic irreducible polynomials, thus it is squarefree. Now we will need the following lemma:

Lemma 1 *If p is a prime number, χ is a non-principal character modulo p of order t, $h(x) \in \mathbb{F}_p[x]$ has degree r and it is not of the form $h(x) = cg(x)^t$ with $c \in \mathbb{F}_p$, $g(x) \in \mathbb{F}_p[x]$, and X, Y are real numbers with $0 < Y \le p$, then*

$$\left| \sum_{X < n \le X+Y} \chi(h(n)) \right| < 9rp^{1/2} \log p.$$

Proof. This lemma can be derived from Weil's theorem [44] by using a method of Vinogradov [43]; see Theorem 2 and Corollary 1 in [34] and Lemma 2 in [3]. (Note that combining Weil's theorem and Vinogradov's inequality with Cochrane's and Peral's result [12], we obtain that the absolute constant 9 in this upper bound can be replaced by $\frac{4}{\pi^2} + o(1)$ for $p \to \infty$, and then the absolute constant 10 in (17) in Theorem 1 can also be replaced by $\frac{4}{\pi^2} + o(1)$.)

Since the polynomial in (21) is squarefree, thus we may use this lemma with the quadratic character

$$\chi(n) = \begin{cases} \left(\frac{n}{p}\right) & \text{if } (n,p) = 1 \\ 0 & \text{if } p \mid n, \end{cases}$$

the polynomial $h(x)$ in (21) and $t = 2$. Then we get from (19) that

$$\left| V_k\left(E_p^{(1)}, \ldots, E_p^{(k)}, M, D\right) \right| < 9kdp^{1/2} \log p. \quad (22)$$

(17) follows from (6), (7) and (22).

In order to prove (18) we need a result of S. D. Cohen [13]:

Lemma 2 *Given a prime power $q > 3$ and arbitrary positive integers n and $m \leq n/3$, there exists a primitive polynomial $x^n + a_1x^{n-1} + \cdots + a_n \in \mathbb{F}_q[x]$ with the first m coefficients $a_1, \ldots, a_m$ prescribed in advance, with the exception that there is no primitive cubic over $\mathbb{F}_4$ with zero first coefficient.*

Proof. This is Theorem 3 in [13].

Now assume that d satisfies the given condition, and consider two distinct irreducible polynomials f_1, f_2 of form (16). Write

$$E_p(f_1) = \left(e_1^{(1)}, \ldots, e_p^{(1)}\right), \ E_p(f_2) = \left(e_1^{(2)}, \ldots, e_p^{(2)}\right).$$

Then the proof of (19) gives that

$$\begin{aligned}\left|\sum_{n=1}^{p} e_n^{(1)} e_n^{(2)}\right| &\leq \left|\sum_{n=1}^{p} \left(\frac{f_1(n)f_2(n)}{p}\right)\right| + 2d \\ &< 18dp^{1/2}\log p + 2d < 20dp^{1/2}\log p < p\end{aligned}$$

thus $E_p(f_1) \neq E_p(f_2)$. It follows that $|\mathcal{F}_1|$ is at least the number of irreducible polynomials of form (16). For any fixed

$$a_2 \in \mathbb{F}_p, a_3 \in \mathbb{F}_p, \ldots, a_{[d/3]} \in \mathbb{F}_p, \tag{23}$$

there is at least one primitive polynomial $f(x)$ of form (16) with these prescribed coefficients (note that $p > 3$ follows from the conditions in the theorem), and these polynomials are also irreducible (since primitive polynomials are irreducible). Since different a_i's determine different irreducible polynomials of form (16) and distinct polynomials determine sequences E_p, thus the number of these sequences is at least the number of choices of the a_i's in (23):

$$|\mathcal{F}_1| \geq p^{[d/3]-1}$$

which proves (18).

5 Another construction

Theorem 1 gives a good upper bound for the cross-correlation, and the size of the family $\mathcal{F}_1$ is also large. However, this theorem has a weakness: since no good algorithm is known for constructing "many" irreducible polynomials over $\mathbb{F}_p$ (see [1], [14], [23], [26]) thus Theorem 1 proves only existence but it does not provide an explicit construction. Thus now we will present another construction which will be more explicit, but the price paid for this is that we will be able to control the cross-correlation of order k only if $k = 2$ or k is odd.

Theorem 2 *Let $d \in \mathbb{N}$, d odd, $d < p$, and consider all the polynomials $f(x) \in \mathbb{F}_p[x]$ of the form*

$$f(x) = (x - x_1)(x - x_2) \cdots (x - x_d) \tag{24}$$

where

$$x_1, x_2, \dots, x_d \text{ are distinct elements of } \mathbb{F}_p \tag{25}$$

and

$$x_1 + x_2 + \cdots + x_d = 0. \tag{26}$$

Let $\mathcal{F}_2$ denote the family of the binary sequences $E_p = E_p(f)$ assigned to these polynomials by formula (14)*. Then we have*
(i)

$$\Phi_k(\mathcal{F}_2) < 10kdp^{1/2} \log p \tag{27}$$

if $k = 2$ or k is odd, and
(ii)

$$|\mathcal{F}_2| = \frac{1}{d}\binom{p-1}{d-1}. \tag{28}$$

Proof. (i) Since a considerable part of the proof is similar to the proof of (17) in Theorem 1 thus we will leave some details to the reader.

As in the proof of Theorem 1, we have

$$\left|V_k\left(E_p^{(1)}, \dots, E_p^{(k)}, M, D\right)\right| \le \left|\sum_{n=1}^{M} \left(\frac{f_1(n + d_1) \cdots f_k(n + d_k)}{p}\right)\right| + dk \tag{29}$$

where $f_1, f_2, \dots, f_k$ are of form (24) (with $x_1, x_2, \dots, x_d$ satisfying (25) and (26)). It follows from (26) that if $f(x)$ is of this form, $c \in \mathbb{F}_p$ and $c \neq 0$, then $f(x) \neq f(x + c)$, thus by the restriction on the d_i's in Definition 5, for $k = 2$ we cannot have $f_1(x + d_1) = f_2(x + d_2)$ in the sum in (29) thus the monic polynomial $f_1(x + d_1) f_2(x + d_2)$ is not a square. If k is odd then the degree of the (monic) polynomial $f_1(x + d_1) \cdots f_k(x + d_k)$ is kd which is odd (since both k and d are odd), thus again this polynomial cannot be a square. In both cases we may use Lemma 1 to estimate the sum in (29), and we get the same upper bound as in the proof of Theorem 1 which proves (27).

(ii) As in the proof of Theorem 1, it follows from (24), (26) and the proof of (27) (with $k = 2$) that for two distinct polynomials f_1, f_2 of form (24) (with $x_1, x_2, \dots, x_d$ satisfying (25) and (26)) we have $E_p(f_1) \neq E_p(f_2)$. Thus

$|\mathcal{F}_2|$ is equal to the number of the polynomials $f(x)$ which satisfy (24), (25) and (26). The number of d-tuples $x_1, x_2, \dots, x_d$ satisfying (25) and

$$x_1 + x_2 + \cdots + x_d = c$$

is independent of c since there is a bijection between the solutions for different c values (note that $0 < d < p$). Thus the number of solutions of (25) and (26) is the total number of d-tuples satisfying (25) divided by p: $\frac{1}{p}\binom{p}{d} = \frac{1}{d}\binom{p-1}{d-1}$, and this proves (28).

Acknowledgement. We would like to thank Arne Winterhof and the anonymous referee for their valuable suggestions and comments.

References

[1] S. Abrahamyan, M. Alizadeh, M. K. Kyureghyan, *Recursive constructions of irreducible polynomials over finite fields*, Finite Fields Appl. 18 (2012), 738-745.

[2] R. Ahlswede, L. H. Khachatrian, C. Mauduit and A. Sárközy, *A complexity measure for families of binary sequences*, Period. Math. Hungar. 46 (2003), 107-118.

[3] R. Ahlswede, C. Mauduit and A. Sárközy, *Large families of pseudorandom sequences of k symbols and their complexity. I*, Lecture Notes in Comput. Sci. 4123, General Theory of Information Transfer and Combinatorics, Springer, Berlin, 2006; pp. 293-307.

[4] R. Ahlswede, C. Mauduit and A. Sárközy, *Large families of pseudorandom sequences of k symbols and their complexity. II*, Lecture Notes in Comput. Sci. 4123, General Theory of Information Transfer and Combinatorics, Springer, Berlin, 2006; pp. 308-325.

[5] N. Alon, Y. Kohayakawa, C. Mauduit, C. G. Moreira and V. Rödl, *Measures of pseudorandomness for finite sequences: minimal values*, Combin., Probab. Comput. 15 (2005), 1-29.

[6] N. Alon, Y. Kohayakawa, C. Mauduit, C. G. Moreira and V. Rödl, *Measures of pseudorandomness for finite sequences: typical values*, Proc. London Math. Soc. 95 (2007), 778-812.

[7] V. Anantharam, *A technique to study the correlation measures of binary sequences*, Discrete Math. 308 (2008), 6203-6209.

[8] A. Bérczes, J. Ködmön and A. Pethő, *A one-way function based on norm form equations*, Period. Math. Hungar. 49 (2004), 1-13.

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

[10] Z.-X. Chen, *Elliptic curve analogue of Legendre sequences*, Monatsh. Math. 154 (2008), 1-10.

[11] Z. Chen, S. Li and G. Xiao, *Construction of pseudorandom binary sequences from elliptic curves by using the discrete logarithms*, in: Sequences and their applications - SETA 2006, LNCS 4086, Springer, 2006; pp. 285-294.

[12] T. Cochrane and J. C. Peral, *An asymptotic formula for a trigonometric sum of Vinogradov*, J. Number Theory 91 (2001), 1-19.

[13] S. D. Cohen, *Primitive polynomials over small fields*, in: Finite Fields and Applications, Seventh International Conference, Toulouse, 2003, eds. G. Mullen et al., LNCS 2948, Springer, Berlin, 2006; pp. 293-307.

[14] S. D. Cohen, *The explicit construction of irreducible polynomials over finite fields*, Designs Codes Cryptogr. 2 (1992), 169-174.

[15] H. Feistel, W. A. Notz and J. L. Smith, *Some cryptographic techniques for machine -to- machine data communications*, Proc. IEEE 63 (1975), 1545-1554.

[16] J. Folláth, *Construction of pseudorandom binary sequences using additive characters over $GF(2^k)$*, Period. Math. Hungar. 57 (2008), 73-81.

[17] J. Folláth, *Construction of pseudorandom binary sequences using additive characters over $GF(2^k)$. II*, Period. Math. Hungar. 60 (2010), 127-135.

[18] G. Gong, *Character Sums and Polyphase Sequence Families with Low Correlation, Discrete Fourier Transform (DFT), and Ambiguty*, in: Finite Fields and Their Applications, Radon Series on Computational and Applied Mathematics 11, eds. C. Pascale et al., 2013, de Gruyter, Berlin; 1-42.

[19] L. Goubin, C. Mauduit and A. Sárközy, *Construction of large families of pseudorandom binary sequences*, J. Number Theory 106 (2004), 56-69.

[20] K. Gyarmati, *On a family of pseudorandom binary sequences*, Period. Math. Hungar. 49 (2004), 45-63.

[21] K. Gyarmati, *Concatenation of pseudorandom binary sequences*, Period. Math. Hungar. 58 (2009), 99-120.

[22] K. Gyarmati, *On the complexity of a family related to the Legendre symbol*, Period. Math. Hungar. 58 (2009), 209-215.

[23] S. Huczynska, *Existence results for finite field polynomials with special properties*, in: Finite Fields and Their Applications, Character Sums and Polynomials, Radon Series on Computational and Applied Mathematics 11, eds.: P. Charpin et al., 2013, de Gruyter, Berlin, 65-87.

[24] J. Kam and G. Davida, *Structured design of substitution-permutation encryption networks*, IEEE Transactions on Computers 28 (1979), 747-753.

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

[26] M. K. Kyureghyan, *Recurrent methods for constructing irreducible polynomials over $\mathbb{F}_q$ of odd characterics, I, II*, Finite Fields Appl. 9 (2003), 39-58; 12 (2006), 357-378.

[27] H. N. Liu, *A family of pseudorandom binary sequences constructed by the multiplicative inverse*, Acta Arith. 130 (2007), 167-180.

[28] H. Liu, *A large family of pseudorandom binary lattices*, Proc. Amer. Math. Soc. 137 (2009), 793-803.

[29] H. Liu, *New pseudorandom sequences constructed by quadratic residues and Lehmer numbers*, Proc. Amer. Math. Soc. 135 (2007), 1309-1318.

[30] H. Liu, *New pseudorandom sequences constructed using multiplicative inverses*, Acta Arith. 125 (2006), 11-19.

[31] C. Mauduit, J. Rivat and A. Sárközy, *Construction of pseudorandom binary sequences using additive characters*, Monatsh. Math. 141 (2004), 197-208.

[32] C. Mauduit and A. Sárközy, *Construction of pseudorandom binary sequences by using the multiplicative inverse*, Acta Math. Hungar. 108 (2005), 239-252.

[33] C. Mauduit and A. Sárközy, *Family Complexity and VC-dimension*, in: Ahlswede Festschrift, eds. H. Aydinian et al., LNCS 7777, Springer, Berlin, 2013; pp. 346-363.

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

[35] A. Menezes, P. C. van Oorschot and S. Vanstone, *Handbook of Applied Cryptography*, CRS Press, Boca Raton, 1997.

[36] L. Mérai, *A construction of pseudorandom binary sequences using both additive and multiplicative characters*, Acta Arith. 139 (2009), 241-252.

[37] L. Mérai, *A construction of pseudorandom binary sequences using rational functions*, Unif. Distrib. Theory 4 (2009), 35-49.

[38] L. Mérai, *Construction of large families of pseudorandom binary sequences*, Ramanujan J. 18 (2009), 341-349.

[39] A. Sárközy, *A finite pseudorandom binary sequence*, Studia Sci. Math. Hungar. 38 (2001), 377-384.

[40] V. Tóth, *Collision and avalanche effect in families of pseudorandom binary sequences*, Period. Math. Hungar. 55 (2007), 185-196.

[41] V. Tóth, *The study of collision and avalanche effect in a family of pseudorandom binary sequences*, Period. Math. Hungar. 59 (2009), 1-8.

[42] M. Tsfasman, S. Vlăduţ and D. Nogin, *Algebraic Geometric Codes:* Basic Notions, in: Mathematical Surveys and Monographs, Vol. 139, AMS, 2007.

[43] I. M. Vinogradov, *Elements of Number Theory*, Dover 1954.

[44] A. Weil, *Sur les courbes algébriques et les variétés qui s'en déduisent*, Act. Sci. Ind. 1041, Hermann, Paris, 1948.