On pseudo-random subsets of the set of the integers not exceeding N

Cécile Dartyge, Andras Sarkozy

(13/2/2007, 9h45)

On pseudo-random subsets of the set of the integers not exceeding N

Cécile Dartyge (Nancy) and András Sárközy (Budapest) *

Abstract. The notion of pseudo-randomness of subsets of $\{1, 2, \dots, N\}$ is defined, and the measures of pseudo-randomness are introduced. Then three special examples are studied. In two cases it turns out that the subset in question possesses strong pseudo-random properties, while the third example is a negative one.

Key words and phrases: pseudo-random, subset.
2000 Mathematics Subject Classification: 05A05, 11Z05

1. Introduction

In many applications (cryptography, simulation, etc.) we need a **random** subset $\mathcal{R}$ of the positive integers not exceeding a certain fixed integer N. (E. g. in [18] and [22] large families of finite binary sequences with strong pseudo-random properties are presented with potential use in cryptography, and in both constructions we start out from polynomials $f(n)$ of form $f(n) = \prod_{a\in\mathcal{A}}(n-a)$ where $\mathcal{A}$ is a random subset of a given size of $\{1, 2, \dots, p\}$ for some prime p.) In most cases we replace the random subset by a **pseudo-random (=PR) subset**, *i. e.*, by a subset which is of "random type", which "looks random", and which has been constructed by a suitable algorithm. But when is a subset a "good" PR subset, when it is "of random type"? Clearly, the subset $\mathcal{R}$ which consists of the even integers not exceeding N, or, in case of subsets containing half of the elements of the given set, subset $\mathcal{R}$ which never contains n, $n+1$ and $n+2$ simultaneously cannot be considered as a "good" PR subset; if we end up with such a subset then it must be discarded. Note that in both examples the special "non-random type" structure is related to the ordering of the integers and, indeed, the starting point of the study of pseudo-randomness of subsets of finite ordered sets must be their ordering. Clearly, it suffices to study subsets of $\{1, 2, \dots, N\}$, the study of subsets of other finite ordered sets can be reduced to this case. So we will study properties of subsets of $\{1, 2, \dots, N\}$ which are related to the ordering of the integers. Clearly, to do this, we will have to use number theoretic tools intensively.

In the last 20 years numerous papers have been written on random structures. In particular, random (Bollobás [2]), pseudo-random (Thomason [25], [26]) and quasi-random graphs (Chung, Graham and Wilson [8], [9], Simonovits and T. Sós [23]), pseudo-random (Haviland and Thomason [12], [13], [25], [26]) and quasi-random hypergraphs (Chung and Graham [4], [5], Kohayakawa, Rödl and Skokan, [15]), quasi-random set systems (Chung and Graham [6]), quasi-random subsets of $\mathbb{Z}_n$ (Chung and Graham [7]) are studied. In these papers typically structures without ordering are considered, correspondingly, combinatorial tools dominate. The study of quasi-random subsets of $\mathbb{Z}_n$ [7] is closest to our subject, we will return to this question in a subsequent paper.

Pseudo-randomness of finite binary sequences has also been studied intensively, mostly in connection with cryptography (see e. g., [19]); since the elements of sequences are

* Research partially supported by the Hungarian National Foundation for Scientific Research, Grant T043623 and T049693, and by French-Hungarian EGIDE-OMKFHÁ exchange program Balaton F-2/03.

ordered, thus this field is closer to our subject. Usually the pseudo-randomness of **algorithms** generating sequences ("pseudo-random generators") is studied and not that of individual sequences, and the tools of computational complexity are used (see, e. g. [11]). This approach has certain limitations and weak points which were analyzed in [21]. Thus in [16] Mauduit and Sárközy proposed another, more constructive approach. Later this approach was used and extended in numerous papers (see the survey paper [21]). In this field the number theoretic tools dominate (but combinatorial, probabilistic, algebraic and analytic tools are also used). In this paper our goal will be to show that the tools introduced in [16] and extended by Hubert and Sárközy in [14] can be adapted easily to study the pseudo-randomness of subsets of $\{1, 2, \dots, N\}$. First in section 2 we will introduce the measures of pseudo-randomness of subsets of $\{1, 2, \dots, N\}$. In the rest of the paper we will study special examples; both positive and negative examples will be presented.

2. The measure of pseudo-randomness of subsets of $\{1, 2, \dots, N\}$.

In [16] Mauduit and Sárközy introduced the following measures of pseudo-randomness of finite binary sequences.

Consider a finite binary sequence

$$E_N = (e_1, \dots, e_n) \in \{-1, 1\}^N. \tag{2·1}$$

Then the **well-distribution measure** of E_N is defined as

$$W(E_N) = \max_{a,b,t} \Big| \sum_{j=0}^{t-1} e_{a+jb} \Big|, \tag{2·2}$$

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

$$C_k(E_N) = \max_{M,D} \Big| \sum_{n=1}^{M} e_{n+d_1} e_{n+d_2} \cdots e_{n+d_k} \Big|, \tag{2·3}$$

where the maximum is taken over all $D = (d_1, \dots, d_k)$ and M such that $O \leqslant d_1 < \dots < d_k \leqslant N - M$. Then the sequence is considered as a **"good" pseudo-random 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 $N \to \infty$.) Indeed Cassaigne, Mauduit and Sárközy [3] showed that for a "truly random" $E_N \in \{-1, +1\}^N$, both $W(E_N)$ and, for fixed k, $C_k(E_N)$ are around $N^{1/2}$ with "near 1" probability (see also [1]). Thus for a "really good" PR sequence we expect the measures (2·2), (2·3) to be not much greater than $N^{1/2}$.

The pseudo-randomness of a sequence of form (2·1) can be interpreted in the following way: suppose ξ is a random variable distributed according to the law

$$P(\xi = 1) = P(\xi = -1) = \frac{1}{2}; \tag{2·4}$$

e. g., we obtain such a ξ if we toss a coin and put $+1$ if it shows head, -1 if it is tail. Now suppose that ξ_1, ξ_2, $\dots$, ξ_N are independent random variables each of distribution (2·4), e. g., we toss the coin N times. We record the outcome of each coin toss, and let e_i denote the value of ξ, *i. e.*, $e_i = +1$ if the i-th coin toss is head and $e_i = -1$ if it is tail. In this way we get a binary sequence of form (2·1). The question is: what can we say about a "typical"

sequence $(e_1, \dots, e_N)$ obtained in this way, what are its most important properties? We say "important" in two senses: important in the applications, and also important in the sense that our sequence possesses it with probability "near 1". Definitions (2·2) and (2·3), and the theorems of Cassaigne, Mauduit and Sárközy described above can be considered as the answer to these questions.

Hubert and Sárközy [14] generalised this model and this notion of pseudo-randomness in the following way.

Replace ξ in (2·4) by a random variable which again may assume two values only, but now they are not equally probable. Suppose they occur with probability p, resp. $1-p$, and by technical reasons, let ξ be defined so that its mean value is 0. In other words, replace (2·4) by, say,

$$P(\xi = 1-p) = p, \quad P(\xi = -p) = 1-p, \tag{2·5}$$

so that, once more, the expected value $M(\xi)$ verifies

$$M(\xi) = 0. \tag{2·6}$$

Then by (2·6), one may define the notion of **p-pseudo-randomness** (pseudo-randomness with respect to the distribution in (2·5)) again by formulas (2·2) and (2·3). In other words, Hubert and Sárközy [14] define the notion of p-pseudo-randomness in the following way.

Consider a finite binary sequence

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

Then the **p-well-distribution measure** of E_N is defined as

$$W(E_N, p) = \max_{a,b,t} \Big| \sum_{j=0}^{t-1} e_{a+jb} \Big|, \tag{2·7}$$

while the **p-correlation measure of order** k of E_N is defined as

$$C_k(E_N, p) = \max_{M,D} \Big| \sum_{n=1}^{M} e_{n+d_1} e_{n+d_2} \cdots e_{n+d_k} \Big|. \tag{2·8}$$

(The maximum in (2·7) and (2·8) is taken in the same way as in (2·2) and (2·3), respectively.)

Then again the sequence is considered as a "good" **p-pseudo-random sequence** if both these measures $W(E_N, p)$ and $C_k(E_N, p)$ (at least for "small " k) are small in terms of N. Again, this terminology is justified by the fact that, as it is proved in [14], for a p-random E_N both $W(E_N, p)$ and $C_k(E_N, p)$ are around $N^{1/2}$. Clearly, this notion of p-pseudo-randomness generalises the notion of pseudo-randomness.

The definition of p-pseudo-randomness can be adapted easily to define the pseudo-randomness of subsets of $\{1, \dots, N\}$. Suppose we want to check a subset $\mathcal{R} \subset \{1, \dots, N\}$ for pseudo-randomness. Let $|\mathcal{R}| = h$. Then a random integer $n \in \{1, \dots, N\}$ belongs to $\mathcal{R}$ with probability $\frac{h}{N}$, which corresponds to the $p = h/N = |\mathcal{R}|/N$ case of pseudo-randomness (but it is not identical with it!) Thus defining the sequence

$$E_N = E_N(\mathcal{R}) = (e_1, e_2, \dots, e_N) \in \Big\{1 - \frac{|\mathcal{R}|}{N}, -\frac{|\mathcal{R}|}{N}\Big\}^N \tag{2·9}$$

by

$$e_n = \begin{cases} 1 - \frac{|\mathcal{R}|}{N} & \text{for } n \in \mathcal{R} \\ -\frac{|\mathcal{R}|}{N} & \text{for } n \notin \mathcal{R} \end{cases} \quad (n = 1, 2, \dots, N), \tag{2·10}$$

we may define the **well-distribution measure** and the **correlation measure of order** k of the subset $\mathcal{R}$ by formulas (2·7), and (2·8) respectively:

$$W(\mathcal{R}, N) = W\big(E_N(\mathcal{R}), \frac{|\mathcal{R}|}{N}\big) = \max_{a,b,t} \Big| \sum_{j=0}^{t-1} e_{a+jb} \Big| \tag{2·11}$$

and

$$C_k(\mathcal{R}, N) = C_k\big(E_N(\mathcal{R}), \frac{|\mathcal{R}|}{N}\big) = \max_{M,D} \Big| \sum_{n=1}^{M} e_{n+d_1} \cdots e_{n+d_k} \Big| \tag{2·12}$$

where $E_N(\mathcal{R})$ is defined by (2·9) and (2·10).

One would expect and might like to show that these measures are "small" (are around $N^{1/2}$) for a "truly random subset" $\mathcal{R}$ of $\{1, \dots, N\}$; this fact does not follow from the analogous results on p-pseudo-randomness, and there are difficulties in adapting their proofs in [14]. We remark that here the natural definition of "truly random subset" would be to take every $\mathcal{R} \subset \{1, \dots, N\}$ with uniform probability 2^{-N}. However, in the applications (e. g., in [18] and [22]) $h = |\mathcal{R}|$ is fixed, thus it is better to show the smallness of the PR measures in the sharper form that we fix h (with $h \to +\infty$, $N - h \to +\infty$) and then we consider the h-element subsets of $\{1, \dots, N\}$ with uniform probability $\binom{N}{h}^{-1}$. We will return to this question in a subsequent paper.

We emphasize that the notions of "p-pseudo-randomness" and "pseudo-randomness of subset" of $\{1, \dots, N\}$ are very close but not identical. Indeed, we use the same tools and formulas but, on the other hand, in the first case we study binary sequences while in the second subsets. Besides, in the first case p is fixed, and then for a "good" p-pseudo-random binary sequence the frequencies of the two elements need not be exactly p, resp. $1 - p$, it is enough if they are near these values; on the other hand, in the second case first we consider a subset $\mathcal{R} \subset \{1, \dots, N\}$, and then the associated p value is exactly $p = |\mathcal{R}|/N$, so that the proportion of the elements selected is exactly p. Finally, in the first case we typically consider a fixed p with $\varepsilon < p < 1 - \varepsilon$ (for some $\varepsilon >$), while in the second case typically we are interested also in subsets $\mathcal{R}$ with $|\mathcal{R}| = o(N)$ so that now $|\mathcal{R}|/N$ (which corresponds to p) is $o(1)$.

3. Subsets formed by mod p residues of polynomials

Let p be a prime number, $\mathbb{F}_p$ the field of modulo p residue classes, and $\overline{\mathbb{F}}_p$ its algebraic closure.

The first "good " pseudo-random sequence studied by Mauduit and Sárközy in [16] was the sequence defined by the Legendre symbol:

$$e_n = \left(\frac{n}{p}\right) \text{ for } 1 \leqslant n \leqslant p - 1.$$

They showed that this sequence has good PR properties, the well-distribution measure (2·2) and the correlation measure of order k (2·3) are $O_k(\sqrt{p} \log p)$ (with a good and

explicite control of the dependance in k). In [17] they extended these results to sequences of the form

$$e_n = \begin{cases} \left(\frac{f(n)}{p}\right) & \text{if } f(n) \not\equiv 0 \,(\mathrm{mod}\, p) \\ 1 & \text{if } f(n) \equiv 0 \,(\mathrm{mod}\, p), \end{cases}$$

for $1 \leqslant n \leqslant p$ where f is a permutation polynomial whose unique zero in $\mathbb{F}_p$ has odd multiplicity. A permutation polynomial $f \in \mathbb{F}_p[X]$ is a polynomial whose associated polynomial function $x \mapsto f(x)$ is a permutation of $\mathbb{F}_p$. For example if $(k, p-1) = 1$, the monomial x^k is a permutation polynomial (see [17] for other examples).

In this section we generalize this construction to power residues.

Let $d|p-1$ and f be a permutation polynomial. The equation $f(x) = 0$ has a unique solution x_0 in $\mathbb{F}_p$, and in $\overline{\mathbb{F}}_p$ we have the factorization:

$$f(x) = (x - x_0)^{r_0}(x - \alpha_1)^{r_1} \cdots (x - \alpha_{s-1})^{r_{s-1}},$$

where $\alpha_1, \ldots, \alpha_s \notin \mathbb{F}_p$. We suppose:

$$(d, r_0) = 1. \tag{3·1}$$

We will study the pseudo-random properties of the following set V

$$V := \{x \in \mathbb{F}_p, \exists y \in \mathbb{F}_p \smallsetminus \{0\} : f(x) \equiv y^d \,(\mathrm{mod}\, p)\}. \tag{3·2}$$

The cardinality of V is $(p-1)/d$. The associated sequence $E(V) = \{e_n\}_{1 \leqslant n \leqslant p}$ defined by (2·9) and (2·10) satisfies:

$$e_n = \begin{cases} 1 - \alpha & \text{if } n \in V \\ -\alpha & \text{if } n \notin V, \end{cases} \tag{3·3}$$

with

$$\alpha = \frac{\operatorname{card} V}{p} = \frac{p-1}{dp}. \tag{3·4}$$

We will show that V has strong PR properties:

Theorem 3.1. *Under (3·1), we have*

$$W(V, p) \leqslant 1 + 9\big(\frac{d-1}{d}\big) s\sqrt{p} \log p, \tag{3·5}$$

and for $k \geqslant 2$,

$$C_k(V, p) \leqslant k + 9\big(\frac{d-1}{d}\big)^k \big(1 + \frac{1}{p}\big)^k ks\sqrt{p} \log p. \tag{3·6}$$

First we prove (3·5). Recall that

$$W(V, p) = \max_{\substack{a,b,t \\ b+a(t-1) \leqslant p}} \Big| \sum_{j=0}^{t-1} e_{aj+b} \Big|.$$

By definition of the sequence $E(V)$ we have:

$$\sum_{j=0}^{t-1} e_{aj+b} = (1-\alpha) \sum_{\substack{j=0 \\ aj+b \in V}}^{t-1} 1 - \alpha \sum_{\substack{j=0 \\ aj+b \notin V}}^{t-1} 1. \tag{3·7}$$

Next in (3·7) we use the formula

$$\sum_{\substack{j=0\\ aj+b\notin V}}^{t-1} 1 = t - \sum_{\substack{j=0\\ aj+b\in V}}^{t-1} 1,$$

and we obtain:

$$\sum_{j=0}^{t-1} e_{aj+b} = \sum_{\substack{j=0\\ aj+b\in V}}^{t-1} 1 - t\alpha. \tag{3·8}$$

We introduce character sums to detect d-power residues. Let χ_0 denote the principal character modulo p. When $(x,p)=1$, we have

$$\sum_{\chi^d=\chi_0} \chi(x) = \begin{cases} d & \text{if } \exists y\in\mathbb{F}_p\setminus\{0\} : x=y^d,\\ 0 & \text{otherwise.}\end{cases} \tag{3·9}$$

Thus we have:

$$\sum_{j=0}^{t-1} e_{aj+b} = \frac{1}{d}\sum_{\chi^d=\chi_0}\sum_{0\leqslant j\leqslant t-1} \chi(f(aj+b)) - \alpha t.$$

The contribution of χ_0 is t/d if $aj+b\not\equiv x_0 \pmod p$ for all $0\leqslant j\leqslant t-1$ and $t/d-1$ otherwise. We easily check that

$$\Big|\frac{1}{d}\sum_{j=0}^{t-1}\chi_0(f(aj+b)) - \alpha t\Big| \leqslant 1.$$

Thus we have:

$$\Big|\sum_{j=0}^{t-1} e_{aj+b}\Big| \leqslant 1 + \frac{1}{d}\Big|\sum_{\substack{\chi^d=\chi_0\\ \chi\neq\chi_0}}\sum_{0\leqslant j\leqslant t-1}\chi(f(aj+b))\Big|. \tag{3·10}$$

To evaluate the character sum we will use the following lemma:

Lemma 3.2. *Suppose that p is a prime number, χ is a non-principal character modulo p of order d (so that $d|p-1$), $f(x)\in\mathbb{F}_p[X]$ has the factorization $f(X)=b(X-x_1)^{d_1}\cdots(X-x_s)^{d_s}$ (where $x_i\neq x_j$ for $i\neq j$) in $\overline{\mathbb{F}}_p$ with*

$$(d,d_1,\dots,d_s)=1.$$

Let X,Y be real numbers with $0<Y\leqslant p$. Then

$$\Big|\sum_{X<n\leqslant X+Y}\chi(f(n))\Big| < 9s\sqrt{p}\log p.$$

This is Lemma 2 in [20], it is a slightly modified form of Theorem 2 in [16], and it was derived from A. Weil's theorem [28]. By Lemma 3.2 we obtain:

$$\begin{aligned}\Big|\sum_{j=0}^{t-1} e_{aj+b}\Big| &\leqslant 1 + \frac{1}{d}\sum_{\substack{\chi^d=\chi_0\\ \chi\neq\chi_0}} 9s\sqrt{p}\log p\\ &\leqslant 1 + 9\big(\frac{d-1}{d}\big)s\sqrt{p}\log p,\end{aligned}$$

this ends the proof of (3·5).

Now we study the correlation measures. Let $k \geqslant 2$. We have to compute:

$$C_k(V,p) = \max_{M,D} \Big| \sum_{n=1}^{M} e_{n+d_1} \cdots e_{n+d_k} \Big|,$$

with M and $D = (d_1, \ldots, d_k)$, $0 \leqslant d_1 \leqslant \ldots \leqslant d_k$ such that $M + d_k \leqslant p$.

If $n \not\equiv x_0 - d_j \pmod p$ for any $1 \leqslant j \leqslant k$, we have by (3·9)

$$\begin{aligned} e_{n+d_1} \cdots e_{n+d_k} &= \prod_{j=1}^{k} \Big[(1-\alpha)\frac{1}{d} \sum_{\chi^d = \chi_0} \chi(f(n+d_j)) - \alpha\big(1 - \frac{1}{d} \sum_{\chi^d=\chi_0} \chi(f(n+d_j))\big)\Big]. \\ &= \prod_{j=1}^{k} \Big[\frac{1}{d} \sum_{\chi^d=\chi_0} \chi(f(n+d_j)) - \alpha \Big] \\ &= \prod_{j=1}^{k} \Big[\frac{1}{d} \sum_{\substack{\chi^d=\chi_0 \\ \chi \neq \chi_0}} \chi(f(n+d_j)) + \frac{1}{dp} \Big] \\ &= \frac{1}{d^k} \prod_{j=1}^{k} \Big[\sum_{\substack{\chi^d=\chi_0 \\ \chi \neq \chi_0}} \chi(f(n+d_j)) + \frac{1}{p} \Big]. \end{aligned}$$

If there exists some j, $1 \leqslant j \leqslant k$ such that $n + d_j \equiv x_0 \pmod p$ then this j is unique and

$$e_{n+d_\ell} = \begin{cases} -\alpha & \text{if } \ell = j \\ \frac{1}{d}\Big[\sum_{\substack{\chi^d=\chi_0 \\ \chi \neq \chi_0}} \chi(f(n+d_\ell)) + \frac{1}{p} \Big] & \text{if } \ell \neq j, \end{cases}$$

and

$$\Big| e_{n+d_1} \cdots e_{n+d_k} - \frac{1}{d^k} \sum_{n=1}^{M} \prod_{j=1}^{k} \Big[\sum_{\substack{\chi^d=\chi_0 \\ \chi \neq \chi_0}} \chi(f(n+d_j)) + \frac{1}{p} \Big] \Big| \leqslant |-\alpha + \frac{1}{dp}| \leqslant 1.$$

There exists at most k integers $n \leqslant M$ such that $n + d_j \equiv x_0 \pmod p$ for some $1 \leqslant j \leqslant k$. Thus we have (see also [17] section 8):

$$\Big| \sum_{n=1}^{M} e_{n+d_1} \cdots e_{n+d_k} - \frac{1}{d^k} \sum_{n=1}^{M} \prod_{j=1}^{k} \Big[\sum_{\substack{\chi^d=\chi_0 \\ \chi \neq \chi_0}} \chi(f(n+d_j)) + \frac{1}{p} \Big] \Big| \leqslant k. \tag{3·11}$$

We define

$$Z = \frac{1}{d^k} \sum_{n=1}^{M} \prod_{j=1}^{k} \Big[\sum_{\substack{\chi^d=\chi_0 \\ \chi \neq \chi_0}} \chi(f(n+d_j)) + \frac{1}{p} \Big].$$

We develop the above product:

$$Z = \frac{1}{d^k} \sum_{r=0}^{k} \frac{1}{p^{k-r}} \sum_{1 \leqslant j_1 < \ldots < j_r \leqslant k} \sum_{\substack{\chi_{j_1} \neq \chi_0 \\ \chi_{j_1}^d = \chi_0}} \cdots \sum_{\substack{\chi_{j_r} \neq \chi_0 \\ \chi_{j_r}^d = \chi_0}} \sum_{n=1}^{M} \chi_{j_1}(f(n+d_{j_1})) \cdots \chi_{j_r}(f(n+d_{j_r})).$$

To evaluate this sum we do the same operations as in [20] p. 382-384. We will not give all the details. The only differences arise from the permutation polynomial f. Since $\mathbb{F}_p^*$ is cyclical we may write each χ_{j_ℓ} like $\chi_{j_\ell} = \chi^{\delta_\ell}$ where χ is a character of order $p-1$.

Let $\delta = (\delta_1, \dots, \delta_r)$, and $\delta_i = \delta D_i$ for $1 \leqslant i \leqslant r$. Since $\chi_{j_i}^d = \chi_0$, we have $d\delta_i \equiv 0 \,(\mathrm{mod}\, p-1)$.

Thus

$$\delta \equiv 0 \,(\mathrm{mod}\, (p-1)/d). \tag{3·12}$$

We write $\chi^* = \chi^\delta$. It is proved in [20] (16), that $\chi^* \neq \chi_0$, more precisely, the order D of χ^* is $D = (p-1)/(p-1, \delta)$ and in our case, $D|d$. The computations p. 383 of [20] yield to:

$$\sum_{n=1}^{M} \chi_{j_1}(f(n+d_{j_1})) \cdots \chi_{j_r}(f(n+d_{j_r})) = \sum_{n=0}^{M-1} \chi^*\big(f(n+d_{j_1})^{D_1} \cdots f(n+d_{j_r})^{D_r}\big).$$

We apply Lemma 3.2 with χ^* instead of χ and with the polynomial $F(n) = f(n+d_{j_1})^{D_1} \cdots f(n+d_{j_r})^{D_r}$. In $\bar{\mathbb{F}}_p$ we have

$$F(x) = \prod_{i=1}^{r} (x + d_{j_i} - x_0)^{r_0 D_i} \prod_{\ell=1}^{s-1} \prod_{i=1}^{r} (x + d_{j_i} - \alpha_\ell)^{r_\ell D_{j_i}}.$$

Since $(r_0, d) = 1$ and $(D_1, \dots, D_r) = 1$, we have $(d, r_0 D_1, \dots, r_0 D_r) = 1$, and since by the assumptions of the theorem and the definition of x_0 the α's do not belong to $\mathbb{F}_p$, the condition of Lemma 3.2 is satisfied.

By Lemma 3.2 we obtain:

$$\sum_{n=0}^{M-1} \chi_{j_1}(f(n+d_{j_1})) \cdots \chi_{j_r}(f(n+d_{j_r})) \leqslant 9ks\sqrt{p}\log p \tag{3·13}.$$

We apply this upper bound in Z, and by (3·11) we end the proof of the theorem:

$$Z \leqslant 9ks\sqrt{p}\log p \sum_{r=0}^{k} \frac{1}{p^{k-r}} \binom{k}{r} = 9ks\sqrt{p}\log p \big(1 + \frac{1}{p}\big)^k.$$

4. A construction using the index

In this section we will give a construction for subsets with strong PR properties which will be based on the notion of index (discrete logarithm) and which is a variant of the construction given in [20]. Thus we will refer to [20] repeatedly, and we will leave some details to the reader.

Let p be an odd prime, g a fixed primitive root, and let $\mathrm{ind}\, a$ denote the modulo p index (discrete logarithm) of a to the base g so that

$$g^{\mathrm{ind}\, a} \equiv a \,(\mathrm{mod}\, p)$$

and, to make the index unique,

$$1 \leqslant \mathrm{ind}\, a \leqslant p-1.$$

Theorem 4.1. *Let h, ℓ be integers with $0 \leqslant h < h + \ell \leqslant p - 1$, and define the subset $\mathcal{R}$ of $\{1, \dots, p-1\}$ by*

$$\mathcal{R} = \{n : 1 \leqslant n \leqslant p-1,\ h < \operatorname{ind} n \leqslant h + \ell\}.$$

Then we have

$$|\mathcal{R}| = \ell, \tag{4·1}$$

$$W(\mathcal{R}, p-1) < 2\sqrt{p}(\log p)^2 \tag{4·2}$$

and, for all $k \in \mathbb{N}$, $k < p$,

$$C_k(\mathcal{R}, p-1) < 9k2^k\sqrt{p}(\log p)^{k+1}. \tag{4·3}$$

Proof. The equality (4·1) is trivial. The proof of (4·2) is based on the Pólya-Vinogradov inequality:

Lemma 4.2. *If p is a prime number, χ a non-principal character modulo p and X, Y are real numbers with $X < Y$, then we have*

$$\Big|\sum_{X<n\leqslant Y} \chi(n)\Big| < \sqrt{p}\log p.$$

(See, e. g., [10], p. 135 for a proof.)
Assume that

$$1 \leqslant a \leqslant a + (t-1)b \leqslant p-1, \tag{4·4}$$

and define $E_{p-1} = (e_1, \dots, e_{p-1})$ by (2·9) and (2·10) where now $|\mathcal{R}| = \ell$ by (4·1), and $N = p - 1$. Then the sum in (2·11) is

$$\begin{aligned}\Big|\sum_{j=0}^{t-1} e_{aj+b}\Big| &= \Big|\sum_{0\leqslant j<t}\Big(\sum_{\substack{h<i\leqslant h+\ell\\ g^i\equiv a+jb\,(\mathrm{mod}\,p)}} 1 - \frac{\ell}{p-1}\Big)\Big| \\ &= \Big|\sum_{0\leqslant j<t}\sum_{\substack{h<i\leqslant h+\ell\\ g^i\equiv a+jb\,(\mathrm{mod}\,p)}} 1 - \frac{\ell t}{p-1}\Big|.\end{aligned} \tag{4·5}$$

By the formula

$$\frac{1}{p-1}\sum_{\chi}\bar{\chi}(a)\chi(b) = \begin{cases} 1 & \text{if } a \equiv b\,(\mathrm{mod}\,p) \text{ and } (a,p) = 1 \\ 0 & \text{otherwise,}\end{cases}$$

here we have

$$\sum_{0\leqslant j<t}\sum_{\substack{h<i\leqslant h+\ell\\ g^i\equiv a+jb\,(\mathrm{mod}\,p)}} 1 = \frac{1}{p-1}\sum_{\chi}\sum_{j=0}^{t-1}\sum_{i=h+1}^{h+\ell}\bar{\chi}(a+jb)\chi(g^i).$$

The contribution of the principal character is

$$\frac{1}{p-1}\sum_{j=0}^{t-1}\sum_{i=h+1}^{h+\ell}1=\frac{\ell t}{p-1}.$$

Thus it follows from (4·5) that

$$\begin{aligned}\Big|\sum_{j=0}^{t-1}e_{aj+b}\Big|&=\frac{1}{p-1}\Big|\sum_{\chi\neq\chi_0}\sum_{j=0}^{t-1}\sum_{i=h+1}^{h+\ell}\bar{\chi}(a+jb)\chi(g^i)\Big|\\&=\frac{1}{p-1}\Big|\sum_{\chi\neq\chi_0}\Big(\sum_{j=0}^{t-1}\bar{\chi}(a+jb)\Big)\Big(\sum_{i=h+1}^{h+\ell}\chi^i(g)\Big)\Big|\\&\leqslant\frac{1}{p-1}\sum_{\chi\neq\chi_0}\Big|\sum_{j=0}^{t-1}\chi(aj+b)\Big|\Big|\sum_{i=h+1}^{h+\ell}\chi^i(g)\Big|.\end{aligned}\tag{4·6}$$

The first inner sum can be estimated by Lemma 4.2 as in (7) in [20], and we obtain

$$\Big|\sum_{j=0}^{t-1}\chi(aj+b)\Big|<\sqrt{p}\log p,\tag{4·7}$$

and as in (8) in [20], the second inner sum is

$$\Big|\sum_{i=h+1}^{h+\ell}\chi^i(g)\Big|<\frac{2}{|1-\chi(g)|}.\tag{4·8}$$

By (4·7) and (4·8) we get from (4·6) that

$$\Big|\sum_{j=0}^{t-1}e_{aj+b}\Big|\leqslant 2\frac{\sqrt{p}\log p}{p-1}\sum_{\chi\neq\chi_0}\frac{1}{|1-\chi(g)|}.\tag{4·9}$$

By (10) in [20] the last sum is

$$\sum_{\chi\neq\chi_0}\frac{1}{|1-\chi(g)|}<(p-1)\log p.\tag{4·10}$$

It follows from (4·9) and (4·10) that

$$\Big|\sum_{j=0}^{t-1}e_{aj+b}\Big|<2\sqrt{p}(\log p)^2.$$

This holds for every a,b,t satisfying (4·4) which, by (2·11), completes the proof of (4·2).

The proof of (4·3) is based on Lemma 3.2 stated in the previous section.

In order to prove (4·3) consider any $D=(d_1,\dots,d_k)$ with non-negative integers $d_1<\ldots<d_k$ and positive integer M with $M+d_k\leqslant p-1$. Then, as in (4·5) and

(4·6), the sum in (2·12) is

$$\begin{aligned}
\Big|\sum_{n=1}^{M} e_{n+d_1}e_{n+d_2}\cdots e_{n+d_k}\Big| &= \Big|\sum_{n=1}^{M}\prod_{j=1}^{k}\Big(\sum_{\substack{h<i\leqslant h+\ell\\ g^i\equiv n+d_j \pmod p}} 1-\frac{\ell}{p-1}\Big)\Big| \\
&= \Big|\sum_{n=1}^{M}\prod_{j=1}^{k}\Big(\frac{1}{p-1}\sum_{\chi_j}\sum_{i_j=h+1}^{h+\ell}\bar\chi_j(n+d_j)\chi_j(g^{i_j})-\frac{\ell}{p-1}\Big)\Big| \\
&= \Big|\sum_{n=1}^{M}\prod_{j=1}^{k}\Big(\frac{1}{p-1}\sum_{\chi_j\neq\chi_0}\sum_{i_j=h+1}^{h+\ell}\bar\chi_j(n+d_j)\chi_j(g^{i_j})\Big)\Big| \\
= \frac{1}{(p-1)^k}\Big|\sum_{\chi_1\neq\chi_0}\cdots\sum_{\chi_k\neq\chi_0}\Big(\sum_{n=1}^{M}\prod_{j=1}^{k}\bar\chi_j(n+d_j)\Big)\prod_{j=1}^{k}\Big(\sum_{i_j=h+1}^{h+\ell}\chi_j(g^{i_j})\Big)\Big| \\
\leqslant \frac{1}{(p-1)^k}\sum_{\chi_1\neq\chi_0}\cdots\sum_{\chi_k\neq\chi_0}\Big|\sum_{n=1}^{M}\prod_{j=1}^{k}\chi_j(n+d_j)\Big|\prod_{j=1}^{k}\Big|\sum_{i_j=h+1}^{h+\ell}\chi_j(g^{i_j})\Big|.
\end{aligned}$$

This expression, apart from a missing factor 2^k, is of nearly the same form as the upper bound in (12) in [20] and, by using Lemma 3.2, it can be estimated in the same way. Thus we end up with an upper bound less by a factor 2^k than the one in [20]:

$$\Big|\sum_{n=1}^{M} e_{n+d_1}\cdots e_{n+d_k}\Big| < 9k2^k p^{1/2}(\log p)^{k+1}$$

which, by (2·12) completes the proof of (4·3).

5. Subsets obtained by sifting

In this last section we will study a subset of $\{1,\dots,N\}$ obtained by sifting: the subset of the square free numbers. Let $Q(N)$ be the set of the square free numbers less than N. The cardinality of $Q(N)$ is (see for example [24] section I.3.7)

$$Q(N) = \frac{6}{\pi^2}N + O(\sqrt{N}). \tag{5·1}$$

We define the rate

$$q_N = \frac{\operatorname{card} Q(N)}{N}.$$

We have

$$q_N = \frac{6}{\pi^2} + O\Big(\frac{1}{\sqrt{N}}\Big). \tag{5·2}$$

By (2·9) and (2·10), the corresponding sequence is $E_N(Q(N)) = (e_1,\dots,e_N)$ with

$$\begin{aligned}
e_n &= \begin{cases} 1-q_N & \text{if } \mu^2(n)=1,\\ -q_N & \text{if } \mu(n)=0\end{cases} \\
&= (1-q_N)\mu^2(n) - q_N(1-\mu^2(n)) \\
&= \mu^2(n) - q_N.
\end{aligned} \tag{5·3}$$

It is easy to see that this sequence does not have good PR properties. First the well-distribution measure is large, for example for each $n \leqslant N/4$, we have $e_{4n} = -q_N$. Thus we have

$$\begin{aligned} W\big(Q(N),N\big) &\geqslant \Big|\sum_{n=1}^{[N/4]} e_{4n}\Big| \\ &\geqslant \frac{Nq_N}{4} - 1. \end{aligned} \tag{5·4}$$

More generally we can see by elementary means that this sequence is not well-distributed in every arithmetic progression of modulus > 1:

Lemma 5.1. *Let $a \geqslant 2$, b, t such that $b + a(t-1) \leqslant N$. We have:*

$$\sum_{j=0}^{t-1} e_{aj+b} = \frac{6t}{\pi^2}(F(a,b)-1) + O\big(\sqrt{N}\big),$$

with

$$F(a,b) = \begin{cases} 0 & \text{if there exists } p \text{ such that } p^2|(a,b), \\ \prod_{p|a}\left(1-\frac{1}{p^2}\right)^{-1} \prod\limits_{\substack{p|(a,b)\\ a\not\equiv 0 \,(\mathrm{mod}\, p^2)}} \left(1-\frac{1}{p}\right) & \text{otherwise.} \end{cases}$$

By this lemma we see that $W(Q(N),N)$ is given by the arithmetic progression modulo 4.

Corollary 5.2. *For $N \geqslant 2$, we have:*

$$W(Q(N),N) = \frac{3N}{2\pi^2} + O(\sqrt{N}).$$

Proof of Corollary 5.2. By Lemma 5.1, we have

$$\begin{aligned} W(Q(N),N) &= \max_{b+a(t-1)\leqslant N} \Big(\frac{6t}{\pi^2}|F(a,b)-1|\Big) + O\big(\sqrt{N}\big) \\ &= \max_{a,b} \frac{(N-b)}{a}\frac{6}{\pi^2}|F(a,b)-1| + O(\sqrt{N}). \end{aligned}$$

By (5·2) and (5·4), to prove Corollary 5.2, it is sufficient to show that

$$|F(a,b)-1|\frac{1}{a} \leqslant \frac{1}{4}. \tag{5·5}$$

By the definition of F, it is clearly the case when $F(a,b) = 0$. We see that the possible other large values of $|F(a,b)-1|$ are obtained for $b = 1$ and we have

$$0 \leqslant \frac{F(a,1)-1}{a} \leqslant (\frac{\pi^2}{6} - 1)\frac{1}{a}$$

which is less than $1/4$ for all $a \geqslant 3$ and we have $(F(2,1)-1)/2 = 1/6$. This proves that Corollary 5.2 is a consequence of that Lemma 5.1.

Proof of Lemma 5.1. We have by (5·3)

$$\sum_{j=0}^{t-1} e_{aj+b} = \sum_{j=0}^{t-1} \mu^2(aj+b) - tq_N.$$

Next we use the formula

$$\mu^2(n) = \sum_{d^2|n} \mu(d),$$

and exchange the order of summations:

$$\begin{aligned}\sum_{j=0}^{t-1} e_{aj+b} &= \sum_{j=0}^{t-1} \sum_{d^2|aj+b} \mu(d) - tq_N \\ &= \sum_{d^2 \leqslant a(t-1)+b} \mu(d) \sum_{\substack{0\leqslant j\leqslant t-1 \\ aj+b\equiv 0 \,(\mathrm{mod}\, d^2)}} 1 - tq_N.\end{aligned}$$

The congruence $aj + b \equiv 0 \,(\mathrm{mod}\, d^2)$ has a solution if and only if $(a, d^2)|b$ and we have

$$\sum_{j=0}^{t-1} e_{aj+b} = \sum_{\substack{d^2\leqslant a(t-1)+b \\ (d^2,a)|b}} \frac{\mu(d)t(a,d^2)}{d^2} - \frac{6t}{\pi^2} + O(\sqrt{N}).$$

It remains to compute the sum over d:

$$\sum_{\substack{d^2\leqslant a(t-1)+b \\ (d^2,a)|b}} \frac{\mu(d)(a,d^2)}{d^2} = \sum_{(d^2,a)|b} \frac{\mu(d)(a,d^2)}{d^2} + O\Big(\sum_{d>\sqrt{at+b}} \frac{\mu(d)^2(a,d^2)}{d^2}\Big).$$

The error term above is clearly less than $O(\sqrt{a/t})$. Let

$$f(d) = \begin{cases} \frac{\mu(d)(a,d^2)}{d^2} & \text{if } (d^2,a)|b \\ 0 & \text{otherwise.}\end{cases}$$

This function is multiplicative and we have

$$f(p) = \begin{cases} -1 & \text{if } p^2|(a,b) \\ -\frac{1}{p^2} & \text{if } (p,a)=1 \\ -\frac{1}{p} & \text{if } p\|a \text{ and } p|b \\ 0 & \text{otherwise.}\end{cases}$$

We have

$$\sum_{(d^2,a)|b} \frac{\mu(d)(a,d^2)}{d^2} = \sum_d f(d) = \prod_p (1+f(p)) = \frac{6}{\pi^2} F(a,b).$$

This ends the proof of (Lemma 5.1).

The correlation measures are also large for every k.

Lemma 5.3. *For all $k \geqslant 2$, we have*

$$C_k\big(Q(N), N\big) \gg_k N.$$

Proof.

Let M and $d_1, \dots d_k$ be some positive integers such that $0 \leqslant d_1 < d_2 < \ldots < d_k$ and $M + d_k \leqslant N$. We have to evaluate

$$C(M, d_1, \dots, d_k) = \sum_{n=1}^{M} e_{n+d_1} \cdots e_{n+d_k}.$$

By (5·3) we have:

$$C(M,d_1,\dots,d_k)=\sum_{n=1}^{M}\prod_{j=1}^{k}\big(\mu^2(n+d_j)-q_N\big).$$

We take the term-by-term product:

$$C(M,d_1,\dots,d_k)=(-q_N)^kM+\sum_{1\leqslant r\leqslant k}(-q_N)^{k-r}\sum_{1\leqslant j_1<\dots<j_r\leqslant k}\sum_{n=1}^{M}\prod_{i=1}^{r}\mu^2(n+d_{j_i}).$$

This sum will be estimated with this following result of Tsang [27]:

Lemma 5.4. *Let $d_1,\dots,d_r$ be distinct integers such that $|d_i|\leqslant cx$ for $1\leqslant i\leqslant r$, $r\leqslant \log x/(25\log\log x)$, and c an absolute constant. For any $x\geqslant 3$, we have*

$$\sum_{n\leqslant x}\mu^2(n+d_1)\cdots\mu^2(n+d_k)=A(d_1,\dots,d_r)x+O_c(x^{7/11}(\log x)^8), \tag{5·6}$$

where

$$A(d_1,\dots,d_r)=\prod_{p}\big(1-\frac{u(p,d_1,\dots,d_r)}{p^2}\big),$$

and $u(p,d_1,\dots,d_r)$ is the number of distinct residue classes modulo p^2 represented by the numbers $d_1,\dots,d_r$. The implied constant in the error term in (5·6) depends only on c.

This is a slightly weaker form of Theorem 2 of Tsang [27].

By Lemma 5.4 we have for $d_k\leqslant cM$:

$$\begin{aligned}C(M,d_1,\dots,d_k)=\frac{(-6)^k}{\pi^{2k}}M+M\sum_{1\leqslant r\leqslant k}\frac{(-6)^{k-r}}{\pi^{2(k-r)}}\sum_{1\leqslant j_1<\dots<j_r\leqslant k}A(d_{j_1},\dots,d_{j_r})\\+O(k^rM^{7/11}(\log M)^8).\end{aligned}$$

For $z\geqslant 2$ we define $D_z=\prod_{p<z}p^2$. We take $d_j=jD_z$ for $j=1,\dots,k$ so that for $p<z$, $u(p,d_1,\dots,d_r)=1$. We have

$$A(d_1,\dots,d_r)=\prod_{p<z}\big(1-\frac{1}{p^2}\big)\big(1+O(\frac{k}{z})\big)=\frac{6}{\pi^2}\big(1+O(\frac{k}{z})\big).$$

Finally we obtain with this choice of $d_1,\dots,d_r$:

$$\begin{aligned}C(M,d_1,\dots,d_k)&=\frac{(-6)^k}{\pi^{2k}}M+\frac{6M}{\pi^2}\big(1+O(\frac{k}{z})\big)\sum_{1\leqslant r\leqslant k}\frac{(-6)^{k-r}}{\pi^{2(k-r)}}\binom{k}{r}\\&\quad+O(k^2M^{7/11}(\log M)^8)\\&=M\Big[\big(\frac{-6}{\pi^2}\big)^k\big(1-\frac{6}{\pi^2}\big)+\frac{6}{\pi^2}\big(1-\frac{6}{\pi^2}\big)^k\Big](1+o(1)).\end{aligned}$$

This ends the proof of Lemma 5.3.

We may also ask if in general, the subset obtained by sifting condition has poor PR properties. This is clearly the case for the distribution if we sieve with small numbers. For example the set of the integers which are sums of two squares is poorly distributed modulo p for $p\equiv 3\,(\mathrm{mod}\,4)$ and the correlation are not small either.

References

[1] N. Alon, Y. Kohayakawa, C. Mauduit, C. G. Moreira and V. Rödl, Measure of pseudorandomnes for finite sequences: typical values, preprint.

[2] B. Bollobás, *Random Graphs*, Academic Press, London, 1985.

[3] J. Cassaigne, C. Mauduit and A. Sárközy, On finite pseudorandom binary sequences VII: the measure of pseudorandomnes, *Acta Arith.* 103 (2002), 97-118.

[4] F. R. K. Chung, Quasi-random classes for hypergraphs, *Random Structures Algorithms 1* (1990), 363-382.

[5] F. R. K. Chung and R. L. Graham, Quasi-random hypergraphs, *Random Structures Algorithms 1* (1990), 105-124.

[6] F. R. K. Chung and R. L. Graham, Quasi-random set systems, *J. American Math. Soc.* 4 (1991), 151-196.

[7] F. R. K. Chung and R. L. Graham, Quasi-random subsets of $\mathbb{Z}_n$, *J. Combin. Theory* Ser. A 61 (1992), 64-86.

[8] F. R. K. Chung, R. L. Graham and R. M. Wilson, Quasi-random graphs,*Proc. Nat. Acad. Sci. U.S.A* 85 (1988), 969-970.

[9] F. R. K. Chung, R. L. Graham and R. M. Wilson, Quasi-random graphs, *Combinatorica 9* (1989), 345-362.

[10] H. Davenport, *Multiplicative Number Theory*, 2nd ed., revised by H. L. Montgomery, Graduate Texts in Math. 74, Springer, New-York, 1980.

[11] S. Goldwasser, Mathematical Foundations of Modern Cryptography: Computational Complexity Perspective, ICM 2002, vol. I, 245-272.

[12] J. Haviland and A. Thomason, Pseudo-random hypergraphs, *Discrete Math.* 75 (1989), 255-278.

[13] J. Haviland and A. Thomason, On testing the "pseudo-randomnes" of a hypergraph, *Discrete Math.* 103 (1992), 321-327.

[14] P. Hubert and A. Sárközy, On p-pseudorandom binary sequences, *Periodica Math. Hungar.* 49 (2004), 73-91.

[15] Y. Kohayakawa, V. Rödl and J. Skokan, Hypergraphs, quasi-randomness and conditions for regularity, *J. Combin. Theory* Ser. A 97 (2002), 307-352.

[16] C. Mauduit and A. Sárközy, On finite pseudorandom binary sequences, I. Measure of pseudo-randomness, the Legendre symbol, *Acta Arith.* 82 (1997), 365-377.

[17] C. Mauduit and A. Sárközy, On finite pseudorandom binary sequences II. The Champernowne, Rudin-Shapiro, and Thue-Morse sequences: a further construction. *J. number theory* 72 (1998), 1-21.

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

[19] A. Menezes, P. van Oorshot and S. A. Vanstone, *Handbook of Applied Cryptography*, CRS Press, Boca Roton, (1997).

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

[21] A. Sárközy, On finite pseudorandom binary sequences and their applications in cryptography, *Tatra Mountains J.*, to appear.

[22] A. Sárközy and C. L. Stewart, On pseudorandomness in families of sequences derived from the Legendre symbol, preprint.

[23] M. Simonovits and V. T. Sós, Szemerédi partitions and quasi-randomness, *Random Structures Algorithms 2* (1991), 1-10.

[24] G. Tenenbaum, *Introduction to analytic and probabilistic number theory*, Cambridge studies in advanced mathematics, 46, Cambridge University Press (1995).

[25] A. Thomason, Pseudo-random graphs, in *Random Graphs '85 (Poznań, 1985)* (M. Karónski ed.) Ann. Discrete Math. 33 (1987), North Holland, Amsterdam; 307-331.

[26] A. Thomason, Random graphs, strongly regular graphs and pseudo-random graphs, in: *Surveys in Combinatorics 1987* (C. Whitehead, ed.), London Math. Soc. Lecture Notes Ser., Vol. 123, Cambridge Univ. Press, (1987), 173-196.

[27] K. M. Tsang, The distribution of r-tuples of squarefree numbers, *Mathematika* 32 (1985)2, 265-275.

[28] A. Weil, *Sur les courbes algébriques et les variétés qui s'en déduisent*, Publ. Inst. Math. Univ. Strasbourg 7 (1945), Hermann, Paris, (1948).

Cécile Dartyge
Institut Élie Cartan
Université Henri Poincaré–Nancy 1
BP 239
54506 Vandœuvre Cedex
France
dartyge@iecn.u-nancy.fr

András Sárközy
Department of Algebra and Number Theory
Eötvös Loránd University
H-1518 Budapest, Pf. 120
1117 Budapest, Pázmány Péter Sétány 1/C
Hungary
sarkozy@cs.elte.hu