Concatenation of pseudorandom binary sequences

Katalin Gyarmati

Eötvös Loránd University, Faculty of Siences,
Department of Algebra and Number Theory,
Hungary, Budapest

gykati@cs.elte.hu

In cryptography it is of basic importance to give good constructions of pseudorandom binary sequences (e.g., they are used in the Vernam cipher).

In the applications it may occur that our initial pseudorandom binary sequence turns out to be not long enough, thus we have to take the concatenation (or merging) of it with other pseudorandom binary sequences. In this way we may obtain several new additional bits.

Here our goal is study when can we form the concatenation of several pseudorandom binary sequences belonging to a given family so that the resulting longer sequence still has strong pseudorandom properties?

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

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

It is an important question when does a binary sequence E_N have good pseudorandom properties?

Mauduit and Sárközy introduced new, quantitative 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 $a, b, t \in \mathbb{N}$ and $1 \leq a \leq a + (t-1)b \leq N$.

The correlation measure of order ℓ of E_N is defined as

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

where the maximum is taken over all $D = (d_1, d_2, \ldots, d_\ell)$ and M such that $1 \le d_1 < d_2 < \cdots < d_\ell < M + d_\ell \le N$.

Then the sequence E_N is considered as a "good" pseudorandom sequence if both these measures $W(E_N)$ and $C_\ell(E_N)$ (at least for small ℓ) are "small" in terms of N (in particular, both are $o(N)$ as $N \to \infty$).

Numerous binary and k-ary sequences have been tested for pseudorandomness by N. Brandstätter, J. Cassaigne, Z. Chen, H. Daboussi, X. Du, É. Fouvry, P. Gaborit, L. Goubin, E. Grant, X. Guozhen, S. Ferenczi, J. Folláth, O. D. King, S. Li, H. Liu, S. R. Louboutin, C. Mauduit, L. Mérai, P. Michel, H. Niederreiter, S. Oon, A. Pethő, W. Philipp, J. Rivat, A. Sárközy, J. Shallit, T. Stoll, R. Tichy, V. Tóth, G. Xiao, C. Yang, W. Zhai and A. Winterhof. In the best constructions we have $W(E_N) \ll N^{1/2}(\log N)^{c_1}$ and $C_\ell(E_N) \ll N^{1/2}(\log N)^{c_\ell}$, where $c_1, c_2, \ldots$ are positive constants.

However, the first constructions produced only a "few" pseudorandom sequences; usually for a fixed integer N, the construction provides only one pseudorandom sequence E_N of length N. First L. Goubin, C. Mauduit and A. Sárközy succeeded in constructing large families of pseudorandom binary sequences. Their construction was the following:

Construction (Goubin, Mauduit, Sárközy)

Suppose that p is a prime number, and $f(x) \in \mathbb{F}_p[x]$ is a polynomial with degree $k > 0$, with leading coefficient 1 and no multiple zero in $\overline{\mathbb{F}}_p$. Define the binary sequence $E_p = \{e_1, \ldots, e_p\}$ by

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

(where $\left(\frac{n}{p}\right)$ denotes the Legendre symbol).

Here we may take about p^k different polynomials $f(x)$: the degree of the polynomial is k, and every coefficients of $f(x)$ may take p different values (here we also count the polynomials with multiple zero, but the number of such polynomials is negligible).

It is a natural question: do all sequences obtained this way have strong pseudorandom properties? Are these sequences independent?

Theorem (Goubin, Mauduit, Sárközy)

If p is a prime and $f(x)$ is a polynomial as it is described in the Construction, then for the sequence E_p defined in the Construction we have

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

Moreover, assume that for $\ell \in \mathbb{N}$ one of the following assumptions holds:
(i) $\ell = 2$;
(ii) $\ell < p$ and 2 is a primitive root modulo p;
(iii) $(4k)^\ell < p$.
Then we also have

$$C_\ell(E_p) < 10k\ell p^{1/2} \log p.$$

In many applications it is not enough to know that the family contains many binary sequences with strong pseudorandom properties; it is also important that the family has a "rich", "complex" structure, there are many "independent" sequences in it. Ahlswede, Khachatrian, Mauduit and Sárközy introduced the *f-complexity* ("*f*" for family):

Definition

The f-complexity $C(\mathcal{F})$ 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 $1 \le i_1 < i_2 < \cdots < i_j \le N$, and for any $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_j \in \{-1,+1\}^j$, we have at least one $E_N = \{e_1, \ldots, e_N\} \in \mathcal{F}$ for which

$$e_{i_1} = \varepsilon_1, \ e_{i_2} = \varepsilon_2, \ldots, e_{i_j} = \varepsilon_j.$$

It is clear from Definition 1 that for $j < C(\mathcal{F})$, there exist at least $2^{C(\mathcal{F})-j}$ sequences $E_N \in \mathcal{F}$ with

$$e_{i_1} = \varepsilon_1, \ e_{i_2} = \varepsilon_2, \ldots, e_{i_j} = \varepsilon_j.$$

However, the high f-complexity ensures only that the family contains many "independent" sequences in this sense, it does not ensure that any pair of sequences in the family are independent. Next we will show an example for a family, where the f-complexity is large, but there are certain connections between almost any pair of sequences.

Example (Gy.)

Let $3 \mid N$ and $E_N = \{e_1, e_2, \ldots, e_N\} \in \{-1, +1\}^N$ be a truly random sequence. Define the family $\mathcal{F}(E_N)$ of binary sequences in the following way:

$$\mathcal{F}(E_N) = \{\{e_1 f_1, e_2 f_2, \ldots, e_N f_N\} : \{f_1, f_2, \ldots, f_N\} \in \{-1, +1\}^N \text{ and } |\{i : f_i = 1\}| = N/3\}.$$

Since $E_N = \{e_1, e_2, \ldots, e_N\}$ is a truly random sequence, for arbitrary sequence $\{f_1, f_2, \ldots, f_N\}$ the sequence $\{e_1 f_1, e_2 f_2, \ldots, e_N f_N\} \in \mathcal{F}(E_N)$ is a random type sequence.

Clearly the f-complexity of $\mathcal{F}(E_N)$ is large: $N/3$. We have seen that every sequence in the family is random-type and the f-complexity is large, but in certain applications, for example in a variant of the Vernam-cipher we may use at most one sequence from $\mathcal{F}(E_N)$ as a keystream. Namely,

the reason of that in certain applications we may use at most one sequence from $\mathcal{F}(E_N)$ is that between any two sequences from $\mathcal{F}(E_N)$ there are certain connections:

The "product" of any two sequences from $\mathcal{F}(E_N)$ is of the form $\{e_1 f_1 e_1 f_1', \ldots, e_N f_N e_N f_N'\} = \{f_1 f_1', \ldots, f_N f_N'\}$. Since in both $\{f_1, \ldots, f_N\}$ and $\{f_1', \ldots, f_N'\}$ the rate of +1's and -1's is $1 : 2$, by a simple computation we see that in the sequence $\{f_1 f_1', \ldots, f_N f_N'\}$ the rate of +1's and -1's is usually around $5 : 4$.

This shows that the f-complexity is not enough to guarantee the secure applicability of the family, one also needs the introduction of further measures. In certain applications we need at least a weak independence of all sequences used in the applications. I expect that the small f-correlation measures (which will be introduced in the next slide) assure this weak independence.

First Anantharam introduced and studied the correlation measure of a family. Here I will extend his definition. I expect that if the correlation measure of a family is small, then the sequences in the family are independent in some sense.

Definition

Let $\mathcal{F} \subseteq \{-1,+1\}^N$ be a large family of pseudorandom binary sequences. The f-correlation measure of order ℓ of $\mathcal{F}$ is defined by

$$C_\ell(\mathcal{F}) \stackrel{def}{=} \max_{1\le t\le \ell,\ E_N^{(1)},E_N^{(2)},\ldots,E_N^{(t)}} C_\ell(\{E_N^{(1)}, E_N^{(2)}, \ldots, E_N^{(t)}\}),$$

where the maximum is taken over all $1 \le t \le \ell$, different sequences $E_N^{(1)}, E_N^{(2)}, \ldots, E_N^{(t)} \in \mathcal{F}$, and where $\{E_N^{(1)}, E_N^{(2)}, \ldots, E_N^{(t)}\} \in \{-1,+1\}^{tN}$ is a binary sequence of length tN obtained by writing the elements of $E_N^{(1)}, E_N^{(2)}, \ldots, E_N^{(t)}$ successively.

First I tested this new measure on the large family of pseudorandom binary sequences introduced by Goubin, Mauduit and Sárközy:

Proposition

Let p be a prime number and $R \in \mathbb{N}$. Consider all the polynomials $f(x) \in \mathbb{F}_p[x]$, which has no multiple roots and

$$0 < \deg f(x) \leq R,$$

where $\deg f(x)$ denotes the degree of $f(x)$. For each of these polynomials $f(x)$, consider the binary sequence $E_p = E_p(f) = \{e_1, e_2, \ldots, e_p\} \in \{-1, +1\}^p$ defined by

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

and let $\mathcal{F}_1$ denote the family of all binary sequences obtained in this way. Then

$$C_2(\mathcal{F}_1) \geq p - 1.$$

Clearly $\mathcal{F}_1$ contains many independent sequences, but a few sequences from $\mathcal{F}_1$ are not independent.

For example

$$E_p(f(x)) = \{e_1, e_2, \ldots, e_p\}$$

and

$$E_p(f(x+1)) = \{e_2, e_3, \ldots, e_p, e_1\}$$

are both members of the family $\mathcal{F}_1$ and we may get $E_p(f(x))$ by shifting the elements of $E_p(f(x+1))$ to left by 1. This property makes the f-correlation of $\mathcal{F}_1$ large.

I remark that altough the f-correlation is large (so there are some sequences in $\mathcal{F}_1$ which are not independent) the f-complexity does not show this, it is also large (so the family also contains many independent sequences):

Theorem (Ahlswede, Khachatrian, Mauduit, Sárközy)

$$C(\mathcal{F}_1) \geq R$$

Later I sharpened the lower bound of Ahlswede, Khachatrian, Mauduit, Sárközy and I determined the correct order of the f-complexity:

Theorem

$$\frac{R}{2\log 2}\log p - O(R\log(R\log p)) \le C(\mathcal{F}_1) \le \frac{R}{\log 2}\log p.$$

I mentioned that the f-correlation measure of order 2 is large, more precisely:

$$\begin{aligned} C_2(\mathcal{F}_1) &\ge C_2(\{E_p(f(x)), E_p(f(x+1))\}) \\ &\ge \left|e_2^2 + e_3^2 + \cdots + e_p^2\right| = p-1. \end{aligned}$$

Similarly it is easy to see that for $a \in \mathbf{F_p}$

$$C_2(\{E_p(f(x)), E_p(f(x+a))\}) \ge \lceil p/2 \rceil .$$

This shows that if we want the f-correlation measure of order 2 to be smaller than $p/2$, then we may use at most one of the polynomials $f(x), f(x+1), \ldots, f(x+p-1)$ in the construction.

$f(x), f(x+1), \ldots, f(x+p-1)$ are polynomials of the same degree r. If this degree $r < p$, then there exists exactly one polynomial $f(x+a)$ with $a \in \mathbb{F}_p$ such that the coefficient of x^{r-1} is 0. Next we restrict our family to such polynomials.

Construction (Gy.)

Let p be a prime number and $R \in \mathbb{N}$, $R < p$. Consider all the polynomials $f(x) \in \mathbb{F}_p[x]$ which have no multiple roots, $0 < \deg f(x) \leq R$ and $f(x)$ is of the form

$$f(x) = a_r x^r + a_{r-2} x^{r-2} + a_{r-3} x^{r-3} + \cdots + a_1 x + a_0$$

with $1 \leq r \leq R$, $a_i \in \mathbb{F}_p$, so that the coefficient of the term $x^{\deg f - 1}$ is $a_{r-1} = 0$. For each of these polynomials $f(x)$, consider the binary sequence $E_p = E_p(f) = \{e_1, e_2, \ldots, e_p\} \in \{-1, +1\}^p$ defined by

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

and let $\mathcal{F}_2$ denote the family of all binary sequences obtained in this way. (Clearly $\mathcal{F}_2 \subseteq \mathcal{F}_1$).

The family $\mathcal{F}_2$ can be used in those applications where only a weak independence of pairs of sequences is required, but we do not need the independence of 3 or more sequences. (Thus small f-correlation measure of order 2 does not give full security.)

Theorem (Gy.)

$$C_2(\mathcal{F}_2) \leq 80Rp^{1/2}\log p.$$

but

$$C_3(\mathcal{F}_2) \geq C_3(\{E_p(x(x+1)), E_p((x+1)(x+2)), E_p(x(x+2))\}) \geq p-3.$$

Fortunately, if we restrict our family to irreducible polynomials then the f-correlation measure of higher order will be also small.

Theorem (Gy.)

Let p be a prime number and $R \in \mathbb{N}$, $R < p$. Consider all the polynomials $f(x) \in \mathbf{F_p}[x]$ which are irreducible, $0 < \deg f(x) \leq R$ and $f(x)$ is of the form

$$f(x) = a_r x^r + a_{r-2} x^{r-2} + a_{r-3} x^{r-3} + \cdots + a_1 x + a_0$$

with $1 \leq r \leq R$, $a_i \in \mathbf{F_p}$, so the coefficient of the term $x^{\deg f - 1} = x^{r-1}$ is $a_{r-1} = 0$. For each of these polynomials $f(x)$, consider the binary sequence $E_p = E_p(f) = \{e_1, e_2, \ldots, e_p\} \in \{-1, +1\}^p$ defined by

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

and let $\mathcal{F}_3$ denote the family of all binary sequences obtained in this way. (Clearly $\mathcal{F}_3 \subseteq \mathcal{F}_1$.) Then for $\ell \geq 2$

$$C_\ell(\mathcal{F}_3) \leq 10 R \ell^2 2^{\ell-1} p^{1/2} \log p.$$

The construction of irreducible polynomials over $\mathbb{F}_p$ (that we needed in the previous theorem) is an important and difficult subject.

Sometimes we won't need that all sequences in the family should be independent, and in this case we can avoid to generate irreducible polynomials. Usually it is enough to consider the correlation measure of those ℓ-tuples, which we indeed use in the application.

Then our question is the following: for a fixed sequence $E_p^{(1)} \in \mathcal{F}$ (where $\mathcal{F}$ is an arbitrary large family of pseudorandom binary sequences) how do we choose further sequences $E_p^{(2)}, E_p^{(3)}, \ldots, E_p^{(t)} \in \mathcal{F}$ such that the longer sequence $\{E_p^{(1)}, E_p^{(2)}, E_p^{(3)}, \ldots, E_p^{(t)}\} \in \{-1, +1\}^{tp}$ will have strong pseudorandom properties?

Recall that our starting point was the following construction:

Construction (Goubin, Mauduit, Sárközy)

Suppose that p is a prime number, and $f(x) \in \mathbb{F}_p[x]$ is a polynomial with degree $k > 0$, with leading coefficient 1 and no multiple zero in $\overline{\mathbb{F}}_p$. Define the binary sequence $E_p = \{e_1, \ldots, e_p\}$ by

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

(where $\left(\frac{n}{p}\right)$ denotes the Legendre symbol).

Goubin, Mauduit and Sárközy proved that under some not too restrictive conditions on the polynomial $f(x)$, the pseudorandom measures of E_p are small.

Now I will give a further condition on the polynomial $f(x)$ which guarantees that the pseudorandom measures are small. It is very easy to generate polynomials which satisfy this condition.

Theorem (Gy.)

Suppose that p is an odd prime, $R \in \mathbb{N}$, $f(x) \in \mathbb{F}_p[x]$ is a polynomial which is of the form

$$f(x) = \left(x^2 - a_1\right)\left(x^2 - a_2\right)\cdots\left(x^2 - a_k\right)(x-1),$$

where $0 \le k \le R$, the a_i's are different quadratic non-residues modulo p so that $\left(\frac{a_i}{p}\right) = -1$ for $1 \le i \le k$. For each of these polynomials f, consider the binary sequence $E_p = E_p(f)$ defined by

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

and let $\mathcal{F}_4$ denote the family of all binary sequences obtained in this way. Then for any $2 \le \ell \in \mathbb{N}$ and for all $f \in \mathcal{F}_4$ we have

$$W(E_p(f)) \ll Rp^{1/2}\log p,$$
$$C_\ell(E_p(f)) \ll R\ell p^{1/2}\log p.$$

Remark

We note that a second degree polynomial $x^2 - a$ is irreducible over $\mathbb{F}_p$ if and only if a is quadratic non-residue modulo p.

It is a natural question why do we not consider polynomials of form $f(x) = \left(x^2 - a_1\right)\left(x^2 - a_2\right)\cdots\left(x^2 - a_k\right)$? In this case the sequence $E_p(f)$ would be symmetric ($e_n = e_{p-n}$ for $1 \leq n \leq p$) which may cause difficulties in the applications.

Next we will adapt the method of this construction for constructing pseudorandom sequences whose concatenation also possesses strong pseudorandom properties.

Next we will give a sufficient condition for the sequence $\{E_p^{(1)}, E_p^{(2)}, E_p^{(3)}, \ldots, E_p^{(t)}\} \in \{-1, +1\}^{tp}$ having strong pseudorandom properties.

Theorem (Gy.)

Suppose that

$$f_1(x) = (x^2 - a_{11})(x^2 - a_{12}) \dots (x^2 - a_{1r_1})(x-1),$$
$$\vdots$$
$$f_t(x) = (x^2 - a_{t1})(x^2 - a_{t2}) \dots (x^2 - a_{tr_t})(x-1),$$

where $a_{i1}, a_{i2}, \dots, a_{ir_i}$ *are different quadratic non-residues modulo* p *for* $1 \le i \le t$. *Moreover suppose that* $a_{i1} = a_{vs}$ *if and only if* $i = v$ *and* $1 = s$. *Let* $R = \max_{1 \le i \le t} \deg f_i(x)$. *Define* $E_p^{(i)} = E_p(f_i)$ *by*

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

Then $E_{tp} = \{E_p^{(1)}, E_p^{(2)}, E_p^{(3)}, \dots, E_p^{(t)}\} \in \{-1, +1\}^{tp}$ *has strong pseudorandom properties: for any* $2 \le \ell \in \mathbb{N}$ *we have*

$$W(E_{tp}) \ll tRp^{1/2} \log p, \qquad C_\ell(E_{tp}) \ll R\ell t 2^{\ell-1} p^{1/2} \log p.$$