Collision and avalanche effect in finite pseudorandom sequences

Theses of PhD Dissertation

Viktória Tóth

Supervisors:
Dr. János Gonda
and
Dr. András Sárközy

PhD School of Informatics
Director: Dr. András Benczúr
PhD Program of Numerical and Symbolical Computations
Program director: Dr. Antal Járai

Department of Computer Algebra
Eötvös Loránd University, Faculty of Informatics
2014

1. Introduction

The notion of pseudorandomness is crucial in computational applications. This notion is used in the numerical analysis and also in pure mathematics. The most important ones are the cryptographycal applications so in this thesis the pseudorandomness is analysed from this point of view.

A standard definition of pseudorandomness is based on the use of the tools of computational complexity. This approach has certain limitations and difficulties, thus recently Mauduit and Sárközy [13] initiated another, more constructive approach. They introduced several measures of pseudorandomness, among them the following two definitions are the most important ones:

1. Definition. *The* ***well-distribution measure*** *of a given* $E_N = (e_1, ..., e_N) \in \{-1, +1\}^N$ *binary sequence is defined as*

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

where the maximum is taken over all a, b, t *with* $a \in \mathbb{Z}, b, t \in \mathbb{N}$ *and* $1 \le a + b \le a + tb \le N$.

2. Definition. *The* ***correlation measure of order*** k *of a given* $E_N = (e_1, ..., e_N) \in \{-1, +1\}^N$ *binary sequence is defined as*

$$C_k(E_N) = \max_{M,D} |V(E_N, M, D)| = \max_{M,D} \left| \sum_{n=0}^{M-1} e_{n+d_1} e_{n+d_2} ... e_{n+d_k} \right|,$$

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

Then E_N is considered as a "good" pseudorandom sequence if both $W(E_N)$ and $C_k(E_N)$ (at least for "small" k) are "small" in terms of N. Indeed, later, Cassaigne, Mauduit and Sárközy 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 $N^{1/2}(\log N)^c$. Later Alon, Kohayokawa, Mauduit, Moreira and Rödl gave the exact value of the constant c. (See also [3].)

2. Further pseudorandom properties: collisions and avalanche effect

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), and to each $s \in \mathcal{S}$ we assign a unique binary sequence

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

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

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

In order to characterize the pseudorandom quality of families of binary sequences Ahlswede, Khachatrian, Mauduit and Sárközy in [1] introduced the notation of family complexity. However, the family complexity measures only one pseudorandom property of families of binary sequences, and there also other pseudorandom properties appearing in the literature. In earlier papers [17], [18] I studied the following pseudorandom properties of families of binary sequences.

3. Definition. *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. If $\mathcal{F}$ is not collision free but the number of collisions is limited, they do not cause many problems. A good measure of the number of collisions is the following:

4. Definition. *The* ***collision maximum*** *$M = M(\mathcal{F}, \mathcal{S})$ is defined by*

$$M = M(\mathcal{F}, \mathcal{S}) = \max_{E_N \in \mathcal{F}} |\{s : s \in \mathcal{S}, E_N(s) = E_N\}|$$

(i.e., M is the maximal number of elements of S representing the same binary sequence E_N).

There is another related notion appearing in the literature, namely, the notion of avalanche effect (see, e.g. [5], [8], and [11]). In [17] I introduced the following related definitions:

5. Definition. *If in (1) we have $S = \{-1, +1\}^l$, and for any $s \in S$, changing any element of 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 $s \in S, s' \in S, s \neq s'$ at least $(\frac{1}{2} - o(1))N$ elements of $E_N(s)$ and $E_N(s')$ are different then $\mathcal{F}$ is said to possess* ***strict avalanche property****.*

To study the avalanche property, I introduced the following measure:

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

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

(a variant of this notion is introduced in [5]; this is 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*

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

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

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

I tested two of the most important constructions, see [17] and [18]. One of them is the extended Legendre symbol construction, proposed by Goubin, Mauduit and Sárközy in [9] and the other one is proposed by Mauduit, Rivat and Sárközy in [12].

My results are presented in the next chapters.

3. A construction using the Legendre symbol

A good candidate for testing these new pseudorandom measures is the Legendre symbol whose random type behaviour has been known for at least a century as the papers of Jacobstahl [10], Davenport [7], Bach [4], Peralta [15] and Damgard [6] and Sárközy's book [16] show.

Mauduit and Sárközy in [13] proved the following results:

7. Theorem (Mauduit és Sárközy, 1997). *There is a number* p_0 *such that if* $p > p_0$ *is a prime number,*$k \in \mathbb{N}$, $k < p$ *and the sequence* $E_{p-1} = (e_1, ..., e_{p-1})$ *is defined by*

$$e_n = \left(\frac{n}{p}\right) \quad (n = 1, 2, ..., p-1) \tag{3}$$

where $\left(\frac{n}{p}\right)$ *denotes the Legendre symbol, then we get*

$$W(E_{p-1}) \leq 9p^{1/2} \log p, \text{ and}$$

$$C_k(E_{p-1}) \leq 9kp^{1/2} \log p.$$

In [9] Goubin, Mauduit and Sárközy extended the (3) Legendre symbol construction in the following way:

for $f(x) \in \mathbb{F}_p[x]$ let

$$e_n = \begin{cases} \left(\frac{f(n)}{p}\right), & \text{ha } (f(n),p)=1 \\ +1, & \text{ha } p|f(n). \end{cases} \tag{4}$$

They proved that under not very strong conditions on $f(x)$ both the well-distribution measure and the correlation measure of small order are small.

Now I will present my results connected to this (4) construction. Namely, I prove that it has a further strong pseudorandom property: a variant of the family described in the Theorem above is collision free, even it possesses a strong form of the avalanche property.

These results were published in [17].

8. Theorem. *Let $\mathcal{S}$ be the set of polynomials $f(x) \in \mathbb{F}_p[X]$ of degree $D \geq 2$ which do not have multiple zeros. Define $E_p = E_p(f) = (e_1, ..., e_p)$ by (4) and $\mathcal{F} = \mathcal{F}(\mathcal{S})$ by (1). Then we have*

$$m(\mathcal{F}) \geq \frac{1}{2}\left(p - (2D-1)p^{1/2} - 2D\right).$$

Note that if $D < \frac{p^{1/2}}{2}$, then it follows from Theorem 8 that

$$m(\mathcal{F}) \geq \frac{1}{2}(p - (2D-1)p^{1/2} - p^{1/2}) = \frac{1}{2}(p - 2Dp^{1/2}) > 0,$$

and thus $\mathcal{F}$ is collision free. This proves the following result:

9. Corollary. *If $\mathcal{S}$, $\mathcal{F}$ are defined as in Theorem 8 and we also have $D < \frac{p^{1/2}}{2}$, then $\mathcal{F}$ is collision free.*

Moreover, if $p \to +\infty$ and $D = o(p^{1/2})$ then Theorem 8 gives

$$m(\mathcal{F}) \geq \left(\frac{1}{2} - o(1)\right) \cdot p,$$

which proves

10. Corollary. *If $\mathcal{S}$, $\mathcal{F}$ are defined as in Theorem 8 and we have $p \to +\infty$, $D = o(p^{1/2})$ then $\mathcal{F}$ possesses the strong avalanche property.*

4. A construction using additive characters

In [12] Mauduit, Rivat and Sárközy presented the following construction:

Let p be an odd prime number, $f(X) \in \mathbb{F}_p[X]$, and define $E_p = (e_1, ..., e_p)$ by

$$e_n = \begin{cases} +1, & \text{ha } 0 \le r_p(f(n)) < p/2 \\ -1, & \text{ha } p/2 \le r_p(f(n)) < p, \end{cases} \tag{5}$$

where $r_p(n)$ denotes the unique $r \in \{0, ..., p-1\}$ such that $n \equiv r (\mathrm{mod}\, p)$.

(The adjective "additive" appears in the title of this section since the prove of the theorem which describes the good properties of this construction uses additive characters.)

They proved:

11. Theorem (Mauduit, Rivat and Sárközy, 2004). *If $f \in \mathbb{F}_p[X]$ is of degree $d \ge 2$ and $E_p = (e_1, ..., e_p)$ is defined as above, we have*

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

and if $2 \le l \le d-1$:

$$C_l(E_p) \ll dp^{1/2}(\log p)^{l+2}.$$

This is the fastest construction but it has the weakness that the correlation of high order can be large:

12. Theorem (Mauduit, Rivat and Sárközy, 2004). *For any $k = 2^t$ there exists a constant $c = c(k) > 0$ such that if p is a prime number large enough, $f \in \mathbb{F}_p[X]$ is of degree k and $E_p = (e_1, ..., e_p)$ is defined as above, then*

$$\max_{\substack{T,M \\ 1 \le T < T+M \le p}} \left| \sum_{n=T}^{T+M} e_n e_{n+1} ... e_{n+k-1} \right| > cp.$$

Now I will show that the family constructed above has a further weak pseudo-random property: there are "many" collisions in it.

First I remark that if $|\mathcal{S}| > 2^N$, then it is trivial that there are collisions in $\mathcal{F}$ and, indeed, we have

13. Theorem. *For any family $\mathcal{F} = \mathcal{F}(\mathcal{S})$ of type (5) the collision maximum $M = M(\mathcal{F}, \mathcal{S})$ satisfies*

$$M \ge \frac{|\mathcal{S}|}{2^N}. \tag{6}$$

Now fix a prime p, and for $k \in \mathbb{N}$ write
$\mathcal{S}_k = \{f(x) : f(x) \in \mathbb{F}_p[x], degf(x) = k\}$ and
$\mathcal{F}_k = \{E_p(f) = (e_1, ...e_p) : f \in \mathcal{S}_k\}$ where
$E_p = E_p(f) = (e_1, ...e_p)$ is defined by (5).
Then it is easy to see that

$$|\mathcal{S}_k| > p^k.$$

Thus it follows from Theorem 13. that if

$$\frac{|\mathcal{S}_k|}{2^p} > \frac{p^k}{2^p} = \exp(k \log p - p \log 2) \to \infty,$$

if

$$\frac{k}{p(\log p)^{-1}} \to \infty. \tag{7}$$

Then we have

$$M(\mathcal{F}_k, \mathcal{S}_k) \to \infty,$$

so that for k satisfying (7) there are many collisions in $\mathcal{F}_k$.

It is much interesting that there are many collisions in $\mathcal{F}_k$ also for small k, even for $k = 2$ (which, besides the property described in Theorem 12, is a further weakness of this construction). $\lfloor x \rfloor$ will be note the integer part of x.

14. Theorem. *If p is a fixed prime and $\mathcal{F}_2$, $\mathcal{S}_2$ are defined as above then we have* $M(\mathcal{F}_2, \mathcal{S}_2) \geq \lfloor \frac{1}{6} \log p \rfloor$.

Summarizing these results:
Mauduit, Rivat and Sárközy proved in [12] if $f \in \mathcal{F}$, than the well-distribution measure and the correlation measure of low order are small (in the term of D). On the other hand, I proved in [17] that there are "many" collisions in $\mathcal{S}_2$ and so in $\mathcal{S}$ as well.

Based on this negative result, one may think that construction (5) is much weaker than (4). Here our goal is to show that this is not completely so and, indeed, in case of construction (5) the situation can be saved by replacing $\mathcal{F} = \mathcal{F}(\mathcal{S})$ by a subfamily of it which is just slightly smaller than $\mathcal{F}$, however, it is collision free, even it possesses the strict avalanche property (and it can be generated as easily as the original family $\mathcal{F}$).

Now I present the collision-free variant of this construction.

Since now let e_p denote the additive character:

$$e_p(n) := e^{2i\pi n/p},$$

and in the binary sequence constructed by (5), generated by the polynomial $f(x)$, let the nth element denoted by $\tilde{e}(f(n))$.
So we have $E_p = (\tilde{e}(f(1)), \tilde{e}(f(2)), \ldots, \tilde{e}(f(p)))$.

So, with this new notion we will focus on the following construction:

$$\tilde{e}(f(n)) = \begin{cases} +1, & \text{ha } 0 \leq r_p(f(n)) < p/2 \\ -1, & \text{ha } p/2 \leq r_p(f(n)) < p. \end{cases} \tag{8}$$

Let $\mathcal{P}_d$ be the set of monic polynomials of degree d whose constant term is 0:

$$\mathcal{P}_d = \{f(x) \in \mathbb{F}_p[x] : f(x) = \sum_{i=0}^{d} a_i x^i \text{ ,ahol } a_0 = 0, a_d = 1\}$$

15. Theorem. *If $f(x) \in \mathcal{P}_d$ and $d < \frac{\sqrt{p}}{16 \log^2 p}$, then the family of binary sequences constructed by (8) is collision free.*

16. Theorem. *If $f(x) \in \mathcal{P}_d$ and $d < \frac{\sqrt{p}}{16 \log^2 p}$, then the family of binary sequences obtained by (8) possesses the strict avalanche property.*

We remark that the family is also collision free and possesses the strict avalanche property if the degrees of the polynomials are not the same but their constant terms are zero. So the results can be extended to the set $\mathcal{U} = \cup_{d=2}^{D} \mathcal{U}_d$ in the following way:

17. Theorem. *If $f(x) \in \mathcal{P}$, where $D < \frac{\sqrt{p}}{16 \log^2 p}$, then the family of binary sequences constructed by (8) is collision free.*

18. Theorem. *If $f(x) \in \mathcal{P}$, where $D < \frac{\sqrt{p}}{16 \log^2 p}$, then the family of binary sequences obtained by (8) possesses the strict avalanche property.*

Conclusion: If a large family of binary sequences with strong pseudorandom properties is given, and it turns out that there are many collisions in it, then this negative fact does not mean that the construction must be discarded immediately. As the construction studied in this paper shows, it may occur that the situation can be saved by replacing the given family by a subfamily of it which is just slightly smaller, it can be generated easily, it is collision free and ideally it possesses even the strict avalanche property.

5. The case of k symbols

Mauduit and Sárközy in [14] extended the study of binary sequences to sequences of k symbols in the following way:

Let $k \in \mathbb{N}, k \geq 2$, and let $\mathcal{A} = a_1, ..., a_k$ be a finite set ("alphabet") of k symbols ("letters"), and consider a sequence $E_N = (e_1, ..., e_N) \in \mathcal{A}^N$ of these symbols.

They introduced the analogue of the binary well-distribution measure and correlation measure to the k-ary case, then they also generalized the Legendre symbol construction to the case of $k \geq 2$ symbols:

Let p be a prime with $p \equiv 1 \pmod{k}$; by Dirichlet's theorem there are infinitely many primes with this property. Write $N = p - 1$, and let $\mathcal{A}$ denote the set of the k-th roots of unity:

$$\mathcal{A} = \left\{ e\left(\frac{j}{k}\right) : j = 0, 1, ..., k-1 \right\}$$

(where $e(\alpha)$ is the standard notation $e(\alpha) = e^{2\pi i \alpha}$). Let g be a primitive root modulo p, and consider the (multiplicative) character χ_1 modulo p with

$$\chi_1(g) = e\left(\frac{1}{k}\right). \tag{9}$$

Clearly, (9) determines this character χ_1 uniquely. Moreover, χ_1 is of order k (so that $\chi_1 \neq \chi_0$ by $k \geq 2$), and for all $1 \leq n \leq N = p - 1$ we have

$$\chi_1(n) \in \left\{ e\left(\frac{j}{k}\right) : j = 0, 1, ..., k-1 \right\} = \mathcal{A}.$$

Now define $E_N = (e_1, ..., e_N)$ by

$$e_n = \chi_1(n) \ ; \ n = 1, 2, ..., N. \tag{10}$$

It turned out that this sequence is a "good" pseudorandom sequence, i.e. both $\delta(E_N)$ and $\gamma_l(E_N)$ are "small":

The definitions of *collision*, *collision maximum* and *avalanche effect* introduced for familes of binary sequences can be adapted to the k-ary case without any change. The notion of *distance* and *distance maximum* can remain also the same as in the binary case (see in [19]). There is a change in the definition of strict avalanche effect:

19. Definition. *If for any $s \in S, s' \in S, s \neq s'$ at least $(\frac{k-1}{k} - o(1)) \cdot N$ elements of $E_N(s)$ and $E_N(s')$ are different then $\mathcal{F}$ is said to possess* ***strict avalanche property****.*

In [14] Mauduit and Sárközy constructed only "a few" k-ary sequences with strong pseudorandom properties. Later Ahlswede, Mauduit and Sárközy in [2] extended their construction to a large family of sequences of k symbols inserting a polynomial $f(x)$ in construction (10):

Define $E_N = (e_1, ..., e_p)$ by

$$e_n = \begin{cases} \chi_1(f(n)), & \text{ha } (f(n), p) = 1 \\ +1 & \text{ha } p|f(n), \end{cases} \tag{11}$$

$n = 1, ..., p$, where χ_1 is a multiplicative character modulo p of order k.

They showed in [2] that this family of sequences possesses strong pseudorandom properties.

Then I studied the case of collisions in this family. In[19] I adapted the extension of the method used in [17], but the fact that now $k > 2$ led certain minor difficulties. However, it turned out that these difficulties can be controlled. Then I proved that this family $\mathcal{F}$ is collision free and it possesses the strong avalanche effect under certain assumptions.

My results are presented below.

Let $\mathcal{H}_D$ be the set of monic polynomials $f(x) \in \mathbb{F}_p[x]$ of degree D, which do not have multiple zeroes.

20. Theorem. *If $f(x) \in \mathcal{H}_D$, then in the family of k-ary sequences constructed above by* (11)*, we have:*

$$m(\mathcal{F}) \geq \frac{k-1}{k} \cdot (p - (2D-1) \cdot p^{1/2}) - 2D.$$

21. Corollary. *If $\mathcal{H}_D$ and $\mathcal{F}$ are defined as above and we also have $k > p$ and $16D^2 < p$, then $\mathcal{F}$ is collision free.*

22. Corollary. *If $\mathcal{H}_D$ and $\mathcal{F}$ are defined as above and we have $p \to \infty$ and $D = o(p^{1/2})$, then $\mathcal{F}$ possesses the strong avalanche property.*

References

[1] 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.

[2] R. Ahlswede, C. Mauduit, A. Sárközy, *Large families of pseudorandom sequences of k symbols and their complexity - Part I.*, Book Title: General Theory of Information Transfer and Combinatorics, Series Title: Lecture Notes in Computer Science, Series Volume 4123 (2006), 308–325.

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

[4] E. Bach, *Realistic analysis of some randomized algorithms*, 19th ACM Sympos. on Theory of Computing (1987)

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

[6] I. Damgard, *On the randomness Legendre and Jacobi sequences*, Lect. Notes in Comp. Sci. 403, Springer-Verlag, Berlin (1990), 163–172.

[7] H. Davenport, *On the distribution of quadratic residues (mod p)*, J. London Math. Soc. 6 (1931), 49–54.

[8] H. Feistel, W. A. Notz, J. L. Smith, *Some cryptographic techniques for machine-to-machine data communications*, Proceedings of the IEEE, 63 (1975), 1545–1554.

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

[10] E. Jacobstahl, *Anwendungen einer Formel aus der Theorie der quadratischen Reste*, Dissertation, Berlin (1906), 26–32.

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

[12] C. Mauduit, J. Rivat, A. Sárközy, *Construction of pseudorandom binary sequences using additive characters*, Monatshefte Math. 141 (2004), 197–208.

[13] C. Mauduit, A. Sárközy, *On finite pseudorandom binary sequences I: The measures of pseudorandomness, the Legendre symbol*, Acta Arith. 82 (1997) 365–377.

[14] C. Mauduit, A. Sárközy, *On finite pseudorandom sequences of k symbols*, Indag. Math. 13 (2002) 89–101.

[15] R. Peralta, *On the distribution of quadratic residues and nonresidues modulo a prime number*, Math. Comp. 58 (1992), 433–440.

[16] Sárközy A., *Számelmélet és alkalmazásai*, Műszaki Könyvkiadó, Budapest, 1978.

[17] V. Tóth, *Collision and avalanche effect in families of pseudorandom binary sequences*, Periodica Math. Hungar. 55. (2007) 2, 185–196.

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

[19] V. Tóth, *Extension of the notion of collision and avalanche effect to sequences of k symbols*, Periodica Math. Hungar. 65. (2012) 2, 229–238.