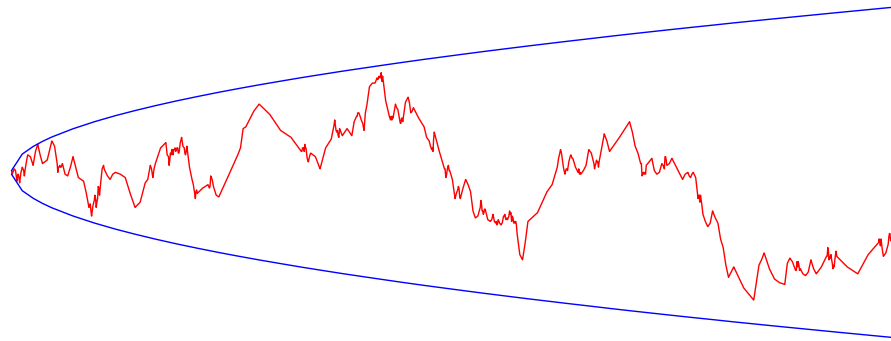


The First Iraqi-French Mathematics Conference  
Salahaddin University - Erbil, 14-18 November 2009

## Measures of pseudorandomness for finite binary sequences

Christian MAUDUIT

Institut de Mathématiques de Luminy, Marseille, France.

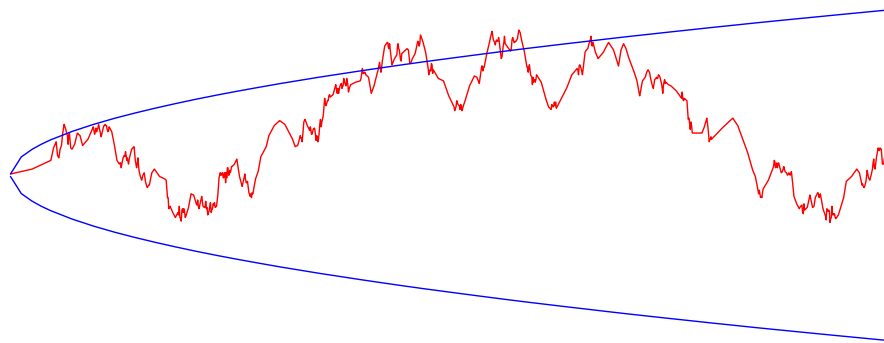


## Authors

Our work on pseudorandom sequences was initiated by Mauduit and Sárközy and developed in a collaboration between Ahlswede, Alon, Cassaigne, Daboussi, Ferenczi, Fouvry, Goubin, Hubert, Khachatryan, Kohayakawa, Mauduit, Michel, Moreira, Rivat, Rödl, Sárközy.

# Part one

## Pseudorandom sequences for cryptography



## Introduction

In the last century numerous papers have been written on pseudorandom (briefly, PR) sequences. In these papers a wide range of goals, approaches, tools is presented, even the concept of “pseudorandomness” is interpreted in different ways (depending mostly on the applications in mind), cf. Knuth (1981), vol. 2.

Two types of sequences are of special interest:

1. sequences of elements of  $[0, 1[$ ,
2. binary sequences (for example 0's or 1's).

In the most papers PR  $[0, 1)$  sequences are considered, much less is known on PR binary sequences, although PR sequences of this type are also needed in applications (simulations, cryptography).

## Cryptosystem – One time pad

Suppose that the message to crypt is a binary sequence of length  $\leq N$ .

We can suppose that the two symbols are  $-1$  and  $+1$ , so we have a sequence  $E_N = \{e_1, \dots, e_N\} \in \{-1, +1\}^N$ . (In practice one uses  $0$  and  $1$ , and the addition  $\oplus$  modulo  $2$ , equivalent but this is less convenient for theory.)

Now we have a *secret* key  $F_N = \{f_1, \dots, f_N\} \in \{-1, +1\}^N$ , so a classical method consist to encrypt by computing

$$\varphi(E_N) = \{e_1 f_1, \dots, e_N f_N\} \stackrel{\text{def}}{=} \{g_1, \dots, g_N\},$$

and decrypt using

$$\varphi(\varphi(E_N)) = \{f_1 g_1, \dots, f_N g_N\} = \{e_1, \dots, e_N\}.$$

This method of *strong* cryptography needs very good pseudorandom sequences.

## Elementary tests

Usually, a sequence is tested for pseudorandomness using some statistical tests. Knuth calls this method *a posteriori* or *empirical* testing. These tests are based on simple properties of “true” random sequences:

1. (*Frequency*) is the number of 0's and 1's approximately the same?
2. (*Serial test*) is the number of 00, 01, 10, 11 approximately the same?
3. (*Poker test*) generalization of the preceding with words of length at most  $c \log n$ , with  $c$  small.
4. (*Runs*) do successive 1's conform with the behavior of a “true” random sequence.
5. (*Autocorrelation*) are there correlations between the sequence and shifted versions of it ?

## Sophisticated tests

More delicate tools can also be used to measure the “quality” of a sequence:

1. (*DFT*) the discrete Fourier transform can be used to detect periodicity properties in the sequence.
2. (*Lempel-Ziv*) “compression” capacity of the sequence: if one can reduce the size of the sequence significantly without loss of information that means that it is not rich enough, and therefore not pseudorandom.
3. (*Maurer’s “universal” test*) based on another form of “compression”.

Elementary tests are implied by “compression” tests.

## *a priori or a posteriori testing*

*a posteriori* testing suffer some caveats. One can deplore in particular

1. more or less arbitrary criterias: why should one use this special property instead of another one ?
2. each new sequence is supposed to be tested, which implies long computations...

One would like to have at disposal what Knuth calls *a priori* or *theoretical* tests, i.e. to propose constructions of sequences for which we **know** that they will satisfy certain pseudorandom properties.

There exists very few results on this subject.

## Elementary generators

The Linear Congruential Generator defines sequences  $(z_n)$  and  $(b_n)$  by

$$z_{n+1} = a \cdot z_n \bmod (2^{31} - 1), \quad b_n = \begin{cases} 0 & \text{if } z_n < 2^{30}, \\ 1 & \text{otherwise.} \end{cases}$$

where  $a$  and  $z_0$  are given initial values.

Of course the LCG is very predictable!

A Quadratic Congruential Generator can be defined by a 512 bits seed  $z_0$  and

$$z_{n+1} = 2z_n^2 + 3z_n + 1 \bmod (2^{512}), \quad b_n = \begin{cases} 0 & \text{if } z_n < 2^{511}, \\ 1 & \text{otherwise.} \end{cases}$$

## The Blum-Micali generator

The Blum-Micali (BM) algorithm defines a binary sequence

$$e_1, \dots, e_N \in \{-1, +1\}$$

by the following recursion

$$x_n = g^{x_{n-1}} \bmod p \quad \text{and} \quad e_n = \begin{cases} +1 & \text{if } 1 \leq x_{n-1} \leq (p-1)/2, \\ -1 & \text{if } (p+1)/2 \leq x_{n-1} \leq p-1. \end{cases}$$

where  $p$  an odd prime number,  $g$  a primitive root modulo  $p$  and  $x_0$  an integer (the *seed*) such that  $1 \leq x_0 < p$ .

The BM generator is cryptographically secure under the assumption that the discrete logarithm problem is intractable.

## The Blum-Blum-Shub generator

The Blum-Blum-Shub (BBS) algorithm defines a sequence

$$b_1, \dots, b_N \in \{0, 1\}$$

by the following recursion:

$$x_i = x_{i-1}^2 \bmod n, \quad b_i = \text{least significant bit of } x_i.$$

where  $n = pq$ ,  $p$  and  $q$  being two distinct (random) prime numbers congruent to 3 modulo 4, and  $x_0 = s^2 \bmod n$ , with the seed  $s$ ,  $1 \leq s < n$  satisfying  $(s, n) = 1$ .

A usual  $\pm 1$  sequence can be deduced immediately by writing  $e_i = 2b_i - 1$ .

The BBS generator is cryptographically secure under the assumption that factorization is intractable.

Friedlander, Pomerance and Shparlinski have interesting results on the period and the distribution of the BBS generator.

## Cryptographic security

By a result of Shannon, a necessary condition for a pseudorandom sequence to be completely secure is that the size of the “seed” is the same as the size of the output sequence.

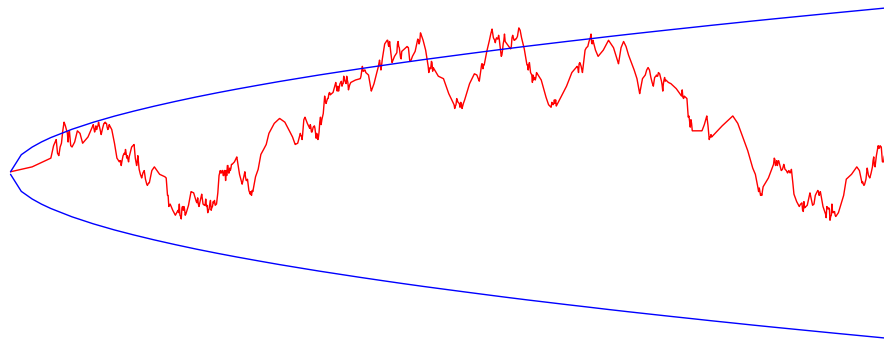
From this point of view, it is clear that all pseudorandom generators based on a simple recursion formula are as far as possible from secure.

However, some notions of cryptographic security have been defined to study the “security” of such generators, the most important being the celebrated “next-bit test”. The widely used Blum-Blum-Shub generator is an example of pseudorandom generator proved to pass the next-bit test, if factorization is untractable, and its recursive formula is fast.

We would like to emphasize that the result of Shannon should not be underestimated, and that the size of the “seed” is an important issue when security is to be taken seriously, and that assumptions should be possibly eliminated or at least their failure should not lead to complete disaster.

# Part two

## New measures of pseudorandomness



## Well distribution and correlation

Mauduit–Sárközy (1997) have defined new measures of pseudorandomness for binary sequences. For

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

the *well-distribution measure* of  $E_N$ :

$$W(E_N) = \max_{a,b,M} \left| \sum_{n=1}^M e_{an+b} \right|$$

with  $1 \leq a + b \leq aM + b \leq N$ , and the *correlation measure* of order  $k$  of  $E_N$ :

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

where  $1 \leq M \leq N$  and  $0 \leq d_1 < d_2 < \cdots < d_k \leq N - M$ .

$E_N$  is considered to be a “good” pseudorandom sequence if both  $W(E_N)$  et  $C_k(E_N)$  are “small”.

## The “random case”

Improving results of Cassaigne-Mauduit-Sárközy, Alon-Kohayakawa-Mauduit-Moreira-Rödl have proved that for any  $\varepsilon > 0$ , there are  $N_0 > 0$  and  $\delta > 0$  such that for  $N > N_0$ , we have with probability at least  $1 - \varepsilon$

$$\delta\sqrt{N} < W(E_N) < \frac{1}{\delta}\sqrt{N}$$

and

$$\frac{2}{5}\sqrt{N \log \binom{N}{k}} < C_k(E_N) < \frac{7}{4}\sqrt{N \log \binom{N}{k}}$$

for all  $2 \leq k \leq N/4$ .

## Research program

Using the new measures  $W$  and  $C_k$ , one would like to:

1. study the pseudorandom behavior of some classical sequences both theoretically and numerically.
2. find out the potential connexions between these measures and the traditional tests and tools.
3. give some applications, hopefully without any unproved assumption.

## The Legendre Symbol

Let  $p$  be an odd prime number. We define \* the Legendre symbol by

$$\left(\frac{n}{p}\right) = \begin{cases} +1 & \text{if } \exists x \in \mathbb{F}_p \text{ such that } n \equiv x^2 \pmod{p}, \\ -1 & \text{otherwise.} \end{cases}$$

We consider the sequence

$$E_p = \left\{ \left(\frac{1}{p}\right), \left(\frac{2}{p}\right), \dots, \left(\frac{p}{p}\right) \right\}$$

Mauduit–Sárközy (1997) have proved that this construction produces a very good pseudorandom sequence:

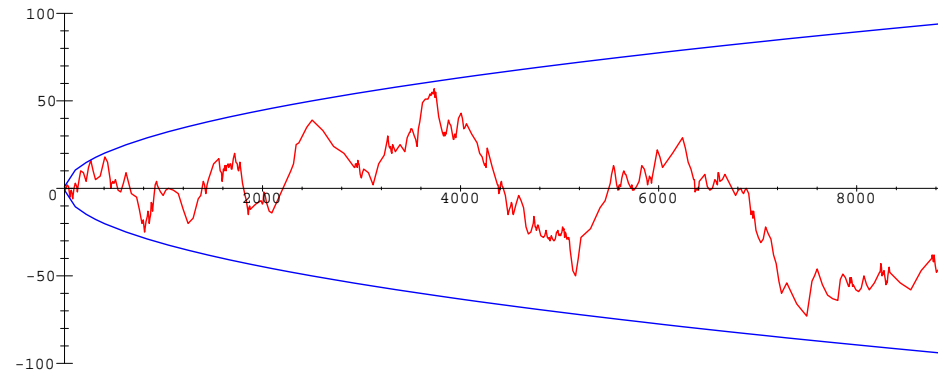
$$\begin{aligned} W(E_N) &\ll N^{1/2} \log N \\ C_k(E_N) &\ll k N^{1/2} \log N \end{aligned}$$

These estimates would be obtained by a “true” random sequence with very high probability, so in some sense the random case is “beaten”.

\*usually  $\left(\frac{0}{p}\right) = 0$ , but in the sequel we will take  $\left(\frac{0}{p}\right) = +1$ .

## “Random walk” of the Legendre Symbol

$$p = 100003, \quad p - 1 = 2 \cdot 3 \cdot 7 \cdot 2381.$$



## $W$ and $C_k$ control on elementary tests

Rivat-Sárközy (2004) have proved that the measures  $W$  and  $C_k$  give a satisfactory control on the elementary statistical tests (frequency, serial, poker, runs, autocorrelation).

In fact  $C_k$  even permits to control long range correlations, which are not usually tackled by the tests, though they can happen in practice.

The Legendre Symbol, for which  $W$  and  $C_k$  are estimated quasi optimally constitute a powerful tool to construct good pseudorandom sequences.

However, for cryptographic applications, we need not only one sequence, but a family of sequences as large as possible!

## Frequency test (monobit test)

Let  $n_-, n_+$  denote the number of  $-1$ 's and  $+1$ 's in  $E_N$ , respectively. The statistic used is

$$X_1 = \frac{(n_- - n_+)^2}{N}$$

which approximately follows a  $\chi^2$  distribution with 1 degree of freedom if  $N \geq 10$ .

Rivat-Sárközy (2001) have proved that

$$X_1 \leq \frac{1}{N}(W(E_N))^2;$$

## Serial test (two bit test)

Let  $n_-$ ,  $n_+$  denote the number of  $-1$ 's and  $+1$ 's in  $E_N$ , respectively, and let  $n_{--}$ ,  $n_{-+}$ ,  $n_{+-}$ ,  $n_{++}$  denote the number of occurrences of  $(-1, -1)$ ,  $(-1, +1)$ ,  $(+1, -1)$ ,  $(+1, +1)$  in  $E_N$ , respectively. (...) The statistic used is

$$X_2 = \frac{4}{N-1}(n_{--} + n_{-+} + n_{+-} + n_{++}) - \frac{2}{N}(n_- + n_+) + 1$$

which approximately follows a  $\chi^2$  distribution with 2 degree of freedom if  $N \geq 21$ .

Rivat-Sárközy (2001) have proved that

$$X_2 \leq \frac{2}{N} \left( (C_2(E_N))^2 + (W(E_N))^2 \right) + 21;$$

## Poker test

Let  $m$  be a positive integer such that  $\lfloor N/m \rfloor \geq 5 \cdot 2^m$  and let  $k = \lfloor N/m \rfloor$ . Divide the sequence  $E_N$  into  $k$  non-overlapping parts each of length  $m$ , and let  $n_i$  be the number of occurrences of the  $i$ -th type of sequence of length  $m$ ,  $1 \leq i \leq 2^m$ . The statistic used is

$$X_3 = \frac{2^m}{N} \left( \sum_{i=1}^{2^m} n_i^2 \right) - k$$

which approximately follows a  $\chi^2$  distribution with  $2^m - 1$  degrees of freedom. Note that the poker test is a generalization of the frequency test: setting  $m = 1$  in the poker test yields the frequency test.

Trying to estimate the statistic  $X_3$  in the poker test, the difficulty is that we have to divide  $E_N$  into *non-overlapping* parts; it is for this reason that it is not enough to use correlations of different orders.

However,  $X_3$  can be handled by the combined PR-measure introduced by Mauduit-Sárközy (1997).

## Runs test

The expected number of runs of  $-1$ 's (or  $+1$ 's) of length  $i$  in a random sequence of length  $N$  is  $m_i = (N - i + 3)/2^{i+2}$ . Let  $k$  be equal to the largest integer  $i$  for which  $m_i \geq 5$ . Let  $B_i, G_i$  be the number of runs of  $-1$ 's (or  $+1$ 's) of length  $i$  in  $E_N$  for each  $i, 1 \leq i \leq k$ . The statistic used is

$$X_4 = \sum_{i=1}^k Y_i, \quad Y_i = \frac{(B_i - m_i)^2 + (G_i - m_i)^2}{m_i}$$

which approximately follows a  $\chi^2$  distribution with  $2k-2$  degrees of freedom.

Rivat-Sárközy (2001) have proved that for  $i \leq W(E_N)$ ,

$$Y_i \leq \frac{2}{m_i} \left( 3 + \frac{i+2}{2^{i+2}} W(E_N) + \frac{1}{2^{i+2}} \sum_{\ell=2}^{i+2} \binom{i+2}{\ell} C_\ell(E_N) \right)^2$$

and for  $i > W(E_N)$ ,

$$Y_i = 2m_i.$$

## Autocorrelation test

Let  $d$  be a fixed integer,  $1 \leq d \leq \lfloor n/2 \rfloor$ . The number of bits in  $E_N$  not equal to their  $d$ -shifts is

$$A(d) = - \sum_{i=1}^{N-d} \frac{e_i e_{i+d} - 1}{2} = \frac{N-d}{2} - \frac{1}{2} \sum_{i=1}^{N-d} e_i e_{i+d}$$

The statistic used is

$$X_5 = 2 \left( A(d) - \frac{N-d}{2} \right) / (N-d)^{1/2}$$

which approximately follows a  $\mathcal{N}(0, 1)$  distribution if  $N-d \geq 10$ . Since small values of  $A(d)$  are as unexpected as large values of  $A(d)$ , a two-sided test should be used.

Rivat-Sárközy (2001) have proved that

$$X_5 \leq \frac{C_2(E_N)}{(N-d)^{1/2}}$$

# Part three

Construction of large families of  
pseudorandom sequences

## The Q construction

Goubin-Mauduit-Sárközy (2003), in conjunction with a paper of Ahlswede-Khachatrian-Mauduit-Sárközy (2003), have generalised the Legendre symbol construction to a large family of sequences.

Let  $p$  be an odd prime and  $f(x) \in \mathbb{F}_p[x]$ . We suppose that  $\deg f = k > 0$  and that  $f$  has no multiple zero in the algebraic closure of  $\mathbb{F}_p$ .

We define  $E_p = \{e_1, \dots, e_p\}$  by  $e_n = \left(\frac{f(n)}{p}\right)$ .

Then

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

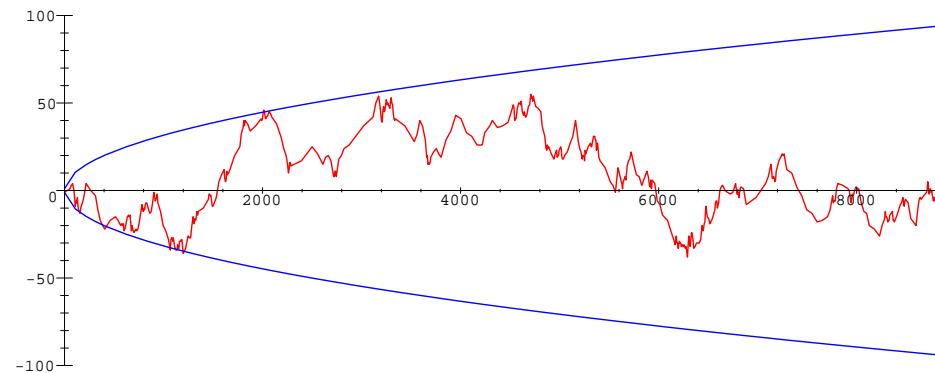
and if  $4k < p^{1/\ell}$ , or if 2 is a primitive root modulo  $p$  then

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

## Example

We take for the sake of illustration only  $p = 1000003$  and  $e_n = \left( \frac{f(n)}{p} \right)$  with

$$f(x) = x^{32} + 637854x^9 + 514861x^8 + 755545x^7 + 883229x^6 + 237063x^5 + 741922x^4 + 631773x^3 + 687734x^2 + 928348x + 283971.$$



## Statistics

We used the software `sts-1.4` “Statistical Test Suite for random and pseudorandom number generators for cryptographic applications” from the National Institute of Standards and Technology (NIST) to analyze the “quality” of our sequences.

we produce 20 sequences of 50000 bits each, et we obtained the following output:

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	P-VALUE	PROPORTION	STATISTICAL TEST
2	2	1	1	4	2	0	3	3	2	0.739918	1.0000	Frequency
4	1	2	2	1	0	0	4	3	3	0.350485	1.0000	Block-Frequency
2	1	2	3	3	3	0	2	1	3	0.834308	1.0000	Cusum
1	2	3	3	3	3	1	1	1	2	0.911413	1.0000	Cusum
6	2	2	1	0	2	2	2	0	3	0.162606	1.0000	Runs
2	3	2	3	1	4	3	2	0	0	0.534146	1.0000	Long-Run
1	0	2	0	2	4	0	3	4	4	0.162606	1.0000	Rank
0	1	5	3	0	1	2	4	2	2	0.213309	1.0000	FFT
0	3	4	3	1	1	3	3	0	2	0.437274	1.0000	Apen
2	6	1	1	3	0	1	1	1	4	0.090936	1.0000	Serial
6	2	2	1	0	3	1	2	0	3	0.122325	1.0000	Serial