ON THE SECURITY OF RANDOM SOURCES

Jean-Sébastien Coron

Ecole Normale Supérieure
45 rue d'Ulm
Paris, F-75230, France
coron@clipper.ens.fr

Gemplus Card International
34 rue Guynemer
Issy-les-Moulineaux, F-92447, France
coron@gemplus.com

Abstract. Many applications rely on the security of their random number generator. It is therefore essential that such devices be extensively tested for malfunction. The purpose of a statistical test is to detect *specific* weaknesses in random sources.
Maurer's universal test is a very common randomness test, capable of detecting a wide range of statistical defects. The test is based on the computation of a function which is asymptotically related to the source's entropy, which measures the effective key-size of block ciphers keyed by the source's output.
In this work we develop a variant of Maurer's test where the test function is in theory *exactly* equal to the source's entropy, thereby enabling a better detection of defects in the tested source.

1 Introduction

Random number generators are probably the most basic cryptographic primitives. They are widely used for block cipher, public-key (*e.g.* RSA-moduli), keystream generation and as passwords sources. In some algorithms (*e.g.* DSA) or protocols (*e.g.* zero-knowledge), random numbers are intrinsic to the computation. In all these applications, security tightly depends on the randomness of the source.

A *pseudo-random* generator is a deterministic polynomial time algorithm that expands short seeds into longer bit sequences, which distribution is polynomially-indistinguishable from the uniform probability distribution. In other words, the output bits must appear to be statistically independent and uniformly distributed. The first pseudo-random generator was constructed and proved by Blum and Micali, under the assumption that the discrete logarithm problem is intractable on a non-negligible fraction of instances [2]. In the light of their practical and theoretical value, constructing pseudo-random generators is a major concern. Procedures for ensuring the security of random number generators are becoming of great importance with the increased usage of electronic communication [4].

It is nevertheless difficult to give a general and reliable measure of the cryptographic quality of a pseudo-random sequence. In practice, many different tests

are carried on sequences generated by the random source to evaluate its performance. These practical tests are divided into two groups : complexity tests and statistical tests. Complexity tests evaluate how much of a generated string is required to reconstruct the whole string [8] while statistical tests evaluate whether the generator's behaviour matches a specific probabilistic model. We refer the reader to [5] for a general treatment of randomness tests.

Maurer's universal test is based on the stationary ergodic source with finite memory statistical model [6]. This model allows the computation of the source's entropy, which, in turn, measures the number of bits of "unpredictability". Failure to provide such unpredictability can weaken severely the security of a cryptosystem, as an attacker could use the reduction in entropy to speed-up exhaustive search on an otherwise secure encryption algorithm.

However, Maurer's universal test only provides an asymptotic measure of the source's entropy. In this paper, we show that with a simple transformation, Maurer's test function can yield *the source's entropy*. Therefore the new test enables a more accurate detection of defects in the tested source.

The paper is organized as follows: we first recall the basic definitions of the stationary ergodic source model and the asymptotic relation between Maurer's test function and the source's entropy. Then we propose a simple transformation of Maurer's test so that the test function yields the source's entropy. Then we study the distribution of the modified test and give a sample program. Finally, we compare the performance of the two tests with respect to different random sources.

2 Statistical model for a random source

2.1 Definition

Consider an information source S emitting a sequence $U_1, U_2, U_3, \ldots$ of binary random variables. S is a *finite memory source* if there exists a positive integer M such that the conditional probability distribution of U_n, given $U_1, \ldots, U_{n-1}$, only depends on the last M bits emitted [6]:

$$P_{U_n|U_1 \ldots U_{n-1}}(u_n|u_1 \ldots u_{n-1}) = P_{U_n|U_{n-M} \ldots U_{n-1}}(u_n|u_{n-M} \ldots u_{n-1})$$

for $n > M$ and for every binary sequence $[u_1, \ldots, u_n] \in \{0,1\}^n$. The smallest M is called the *memory* of the source. The probability distribution of U_n is thus determined by the source's *state* $\Sigma_n = [U_{n-M}, \ldots, U_{n-1}]$ at step n.

The source is *stationary* if it satisfies :

$$P_{U_n|\Sigma_n}(u|\sigma) = P_{U_1|\Sigma_1}(u|\sigma)$$

for all $n > M$, for $u \in \{0,1\}$ and $\sigma \in \{0,1\}^M$.

The state-sequence of a stationary source with memory M forms a finite Markov chain : the source can be in a finite number (actually 2^M) of states σ_i,

$0 \leq i \leq 2^M - 1$, and there is a set of transition probabilities $\Pr(\sigma_j|\sigma_i)$, expressing the odds that if the system is in state σ_i it will next go to state σ_j. For a general treatment of Markov chains, the reader is referred to [1].

For a general Markov chain with r states, let $P_i^{(n)}$ be the probability of being in state σ_i at time $t = n$ and let $P^{(n)}$ be the "state distribution vector" at time n, *i.e.*, $P^{(n)} = [P_1^{(n)}, \ldots, P_r^{(n)}]$.

Let Π be the transition matrix of the chain, *i.e.*, $\Pi_{i,j} = \Pr(\sigma_j|\sigma_i)$ where $\Pi_{i,j}$ is the element in row i and column j of Π.

For state σ_j at time n the source may originate from any state σ_i at time $n-1$ and thus :

$$P_j^{(n)} = \Pr(\sigma_j|\sigma_1)P_1^{(n-1)} + \ldots + \Pr(\sigma_j|\sigma_r)P_r^{(n-1)}$$

which becomes in matrix notations :

$$P^{(n)} = P^{(n-1)}\Pi$$

For the class of *ergodic* Markov processes the probabilities $P_j^{(n)}$ of being in state σ_j after n emitted bits, approach (as $n \to \infty$) an equilibrium P_j which must satisfy the system of r linear equations :

$$\begin{cases} \sum_{j=1}^{r} P_j = 1 \\ \\ P_j = \sum_{i=1}^{r} P_i \ \Pr(\sigma_j|\sigma_i) \qquad \text{for } 1 \leq j \leq r-1 \end{cases}$$

In the case of a source with memory M, each of the 2^M states has at most two successor states with non-zero probability, depending on whether a zero or a one is emitted. The transition probabilities are thus determined by the set of conditional probabilities $p_i = \Pr(1|\sigma_i)$, $0 \leq i \leq 2^M - 1$ of emitting a one from each state σ_i. The transition matrix Π is thus defined by :

$$\Pi_{i,j} = \begin{cases} p_i & \text{if } j = 2i+1 \mod 2^M \\ 1-p_i & \text{if } j = 2i \mod 2^M \\ 0 & \text{otherwise} \end{cases}$$

The entropy of state σ_i is then $H_i = H(p_i)$, where H is the binary entropy function :

$$H(x) = -x\log_2 x - (1-x)\log_2(1-x)$$

The source's entropy is then the average of the entropies H_i (of states σ_i) weighted with the state-probabilities P_i :

$$H_S = \sum_i P_i H_i$$

Let us now assume that the random source is used to generate the N-bit key of a block cipher and let $n(q)$ be the number of N-bit keys that must be tested (in decreasing probability order) in order to reach an overall success probability of q. Shannon proved (see [7], theorem 4) that for $q \neq 0$ and $q \neq 1$:

$$\lim_{N \to \infty} \frac{\log_2 n(q)}{N} = H_S$$

This shows that when an ergodic stationary source is used to key a block cipher, the entropy H_S is closely related to the number of keys an attacker has to try in order to find the right key. In other words, the entropy measures the effective key-size of a cryptosystem keyed by the source's output.

2.2 Probability of a bit sequence

In this section we compute the probability of emitting a bit sequence, which will be used in section 7.2. Starting from a state distribution vector $W = [W_1, \ldots, W_r]$, the probability of emitting a bit $b \in \{0, 1\}$ is :

$$\Pr[b|W] = \sum W_i \Pi_{i,j} \tag{1}$$

where the sum is taken over the couples $\{i, j\}$ for which b is emitted during the transition from σ_i to σ_j.

Let $\Pi(b)$ be the transition matrix corresponding to an emitted bit b :

$$\Pi(b)_{i,j} = \begin{cases} \Pi_{i,j} & \text{if bit } b \text{ is emitted from } \sigma_i \text{ to } \sigma_j \\ 0 & \text{otherwise} \end{cases}$$

It follows that $\Pi = \Pi(0) + \Pi(1)$ and equation (1) becomes :

$$\Pr[b|W] = W \Pi(b) U \quad \text{where} \quad U = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

By iteration, the probability of emitting the sequence $b = [b_0, \ldots, b_n]$ from the state distribution vector W is :

$$\Pr[b|W] = W \Pi(b_0) \Pi(b_1) \ldots \Pi(b_n) U$$

and with $\Pi(b) = \Pi(b_0) \Pi(b_1) \ldots \Pi(b_n)$ the probability of appearance of sequence b is :

$$\Pr[b] = P \Pi(b) U$$

3 Maurer's universal test and the source's entropy

3.1 Maurer's test

Maurer's universal test [6] takes as input three integers $\{L, Q, K\}$ and a $(Q+K) \times L = N$-bit sample $s^N = [s_1, \ldots, s_N]$ generated by the tested source. The parameter L is chosen from the interval $[6, 16]$. The sequence s^N is partitioned into non-overlapping L-bit blocks. For $1 \leq n \leq Q + K$, let $b_n(s^N) = [s_{L(n-1)+1}, \ldots, s_{Ln}]$ denote the n-th L-bit block of s^N.

The first Q blocks of the sequence are used to initialize the test; Q should be chosen to be at least 10×2^L in order to have a high likelihood that each of the 2^L blocks of L bits occurs at least once in the first Q blocks. The remaining K blocks are used to compute the test function $f_{T_U} : B^N \to \mathbb{R}$:

$$f_{T_U}(s^N) = \frac{1}{K} \sum_{n=Q+1}^{Q+K} \log_2 A_n(s^N) \tag{2}$$

where B denotes the set $\{0, 1\}$ and $A_n(s^N)$ the minimum distance between the n-th block and any similar preceding block :

$$A_n(s^N) = \begin{cases} n & \text{if } \forall i < n, b_{n-i}(s^N) \neq b_n(s^N) \\ \min\{i : i \geq 1, b_n(s^N) = b_{n-i}(s^N)\} & \text{otherwise.} \end{cases} \tag{3}$$

3.2 Asymptotic entropy relation

As will be justified later, Maurer's test function is closely related to the source's entropy. It follows that Maurer's universal test is able to detect any of the statistical defects that can be modeled by an ergodic stationary source with finite memory.

Let K_L be the entropy of L-bit blocks, G_L the per-bit entropy of blocks of L bits and F_L the entropy of the L-th order approximation of the source (see Shannon [7]) :

$$K_L = -\sum_{b \in B^L} \Pr[b] \log_2 \Pr[b] \tag{4}$$

$$F_L = -\sum_{b \in B^{L-1},\ j \in B} \Pr[b, j] \log_2 \Pr[j|b] \tag{5}$$

$$G_L = \frac{K_L}{L} = \frac{1}{L} \sum_{i=1}^{L} F_i \tag{6}$$

In [3] we proved the following asymptotic relation between the expectation of Maurer's test function for a stationary ergodic source S outputting a sequence U_S^N of random binary variables and the entropy of L-bit blocks of S :

$$\lim_{L\to\infty}\Big[E[f_{T_U}(U_S^N)] - K_L\Big] = C \triangleq \int_0^\infty e^{-\xi}\log_2\xi\; d\xi \cong -0.8327462 \qquad (7)$$

In the next section we improve the performance of Maurer's test by modifying the test function so that its expectation yields the source's entropy, instead of having an asymptotical relation.

4 Improving Maurer's universal test

Maurer's test function is defined as the average of the logarithm to the base two of the minimum distances between two similar blocks. Here we generalize the definition of the test parameter to any function $g : \mathbb{N} \to \mathbb{R}$ of the minimum distance between two similar blocks :

$$f_{T_U}^g(s^N) = \frac{1}{K}\sum_{n=Q+1}^{Q+K} g(A_n(s^N))$$

The mean of $f_{T_U}^g(U_S^N)$ for S is given by :

$$E[f_{T_U}^g(U_S^N)] = \sum_{i\geq 1}\Pr[A_n(U_S^N) = i]g(i)$$

with

$$\Pr[A_n(U_S^N) = i] = \sum_{b\in B^L}\Pr[b_n = b, b_{n-1}\neq b, \ldots, b_{n-i+1}\neq b, b_{n-i} = b] \qquad (8)$$

If we assume that the L-bit blocks are statistically independent, the above probability factors into :

$$\Pr[A_n(U_S^N) = i] = \sum_{b\in B^L}\Pr[b]^2\times(1-\Pr[b])^{i-1}$$

and we get :

$$E[f_{T_U}(U_S^N)] = \sum_{b\in B^L}\Pr[b]\times\gamma_g(\Pr[b]) \qquad (9)$$

where :

$$\gamma_g(x) = x\sum_{i=1}^{\infty}(1-x)^{i-1}g(i)$$

Equation (9) shows that the mean value of the generalized test may be interpreted as the expectation of a random variable $W = W(X)$ which hits the value

$\gamma_g(\Pr[b])$ with probability $\Pr[b]$. However, the entropy of L-bit blocks K_L (equation (4)) can be viewed as the expectation of a random variable $W' = W'(X)$ which takes the value $-\log_2(\Pr[b])$ with probability $\Pr[b]$.

In order to determine the expectation of the test with the entropy of L-bit blocks, we have to solve the following equation :

$$\gamma_g(x) = -\log_2(x) \tag{10}$$

Letting $t = 1 - x$, equation (10) yields :

$$(1-t)\sum_{i=1}^{\infty} t^{i-1} g(i) = -\log_2(1-t) = \frac{1}{\log(2)} \sum_{i=1}^{\infty} \frac{t^i}{i}$$

and we get :

$$\begin{cases} g(1) = 0 \\ g(i+1) - g(i) = \frac{1}{i\log(2)} \quad \text{for } i \geq 1, \end{cases}$$

Hence we can define a modified version of Maurer's test which test parameter $f_{T_U}^H(s^N)$ is computed using :

$$f_{T_U}^H(s^N) = \frac{1}{K} \sum_{n=Q+1}^{Q+K} g(A_n(s^N)) \tag{11}$$

$$g(i) = \frac{1}{\log(2)} \sum_{k=1}^{i-1} \frac{1}{k} \tag{12}$$

and equation (3) for the definition of $A_n(s^N)$.

The mean value of this new test function taking as input a sequence U_S^N generated by an ergodic stationary source S is equal to the entropy of L-bit blocks of S :

$$E[f_{T_U}^H(U_S^N)] = K_L \tag{13}$$

5 Distribution of the modified test parameter

To tune the test's rejection rate, one must first know the distribution of $f_{T_U}^H(R^N)$, where R^N denotes a sequence of N bits emitted by a binary symmetric source (BSS, *i.e.* a truly random source). A sample s^N would then be rejected if the number of standard deviations separating its $f_{T_U}^H(s^N)$ from $E[f_{T_U}^H(R^N)]$ exceeds a reasonable constant.

In this section we compute the mean and standard deviation of the modified test parameter for a BSS under the reasonable assumption that $Q \to \infty$ (in practice, Q should be larger than 10×2^L).

From equations (11 and 12) the expected value $E[f_{T_U}^H(R^N)]$ of the test parameter $f_{T_U}^H$ for a BSS is given by :

$$E[f_{T_U}^H(R^N)] = \frac{1}{\log(2)} \sum_{i=2}^{\infty} \Pr[A_n(R^N) = i] \sum_{k=1}^{i-1} \frac{1}{k} \tag{14}$$

Using equation (8) we have for a BSS :

$$\Pr[A_n(R^N) = i] = 2^{-L}(1 - 2^{-L})^{i-1} \quad \text{for } i \geq 1 \tag{15}$$

and with equation (14) :

$$E[f_{T_U}^H(R^N)] = \frac{2^{-L}}{\log(2)} \sum_{i=2}^{\infty} (1 - 2^{-L})^{i-1} \sum_{k=1}^{i-1} \frac{1}{k} = L$$

Thus the mean of the test parameter for a truly random source is simply equal to L, the length of the blocks in the tested sequence. Note that this result is straightforward considering equation (13) since the entropy K_L of L-bit blocks is equal to L for a truly random source.

For statistically independent random variables the variance of a sum is the sum of variances but the A_n-terms in (11) are heavily inter-dependent; of course, the same holds for Maurer's original test function (2). Consequently, Maurer introduced in [6] a corrective factor $c(L, K)$ by which the standard deviation of f_{T_U} is reduced compared to what it would have been if the A_n-terms were independent :

$$\mathrm{Var}[f_{T_U}(R^N)] = c(L, K)^2 \times \frac{\mathrm{Var}[\log_2 A_n(R^N)]}{K}$$

Similarly, we can define $c^H(L, K)$ to be the corrective factor by which the standard deviation of the modified test parameter $f_{T_U}^H$ is reduced compared to what it would have been if the A_n-terms were independent :

$$\mathrm{Var}[f_{T_U}^H(R^N)] = c^H(L, K)^2 \times \frac{\mathrm{Var}[g(A_n(R^N))]}{K}$$

The variance of the A_n-terms can be easily computed using equation (15) :

$$\begin{aligned} \mathrm{Var}[g(A_n(R^N))] &= E[(g(A_n(R^N))^2] - \Big(E[g(A_n(R^N))]\Big)^2 \\ &= 2^{-L} \sum_{i=2}^{\infty} (1 - 2^{-L})^{i-1} \Big(\sum_{k=1}^{i-1} \frac{1}{k \log(2)} \Big)^2 - L^2 \end{aligned}$$

In [3] we have computed the exact value of the factor $c(L, K)$, while only a heuristic estimate of $c(L, K)$ was given in [6].

The expression of $c^H(L, K)$ is very similar to the one of $c(L, K)$ given in [3] as one should simply replace the terms in the formulae containing $\log_2 i$ by :

$$g(i) = \frac{1}{\log(2)} \sum_{k=1}^{i-1} \frac{1}{k}.$$

As in [3], the factor $c^H(L,K)$ can be approximated for $K \geq 33 \times 2^L$ by :

$$c^H(L,K)^2 = d(L) + \frac{e(L) \times 2^L}{K}$$

and $\mathrm{Var}[g(A_n(R^N))]$, $d(L)$ and $e(L)$ are listed in table 1 for $3 \leq L \leq 16$ and $L \to \infty$.

This approximation is sufficient because the test must be performed with $K \geq 1000 \times 2^L$.

To summarize, the distribution of $f^H_{T_U}(R^N)$ can be approximated by the normal distribution of mean $E[f^H_{T_U}(R^N)] = L$ and standard deviation :

$$\sigma = c(L,K)\sqrt{\mathrm{Var}[g(A_n(R^N))]/K}$$

L	$\mathrm{Var}[g(A_n(R^N))]$	$d(L)$	$e(L)$
3	2.5769918	0.3313257	0.4381809
4	2.9191004	0.3516506	0.4050170
5	3.1291382	0.3660832	0.3856668
6	3.2547450	0.3758725	0.3743782
7	3.3282150	0.3822459	0.3678269
8	3.3704039	0.3862500	0.3640569
9	3.3942629	0.3886906	0.3619091
10	3.4075860	0.3901408	0.3606982
11	3.4149476	0.3909846	0.3600222
12	3.4189794	0.3914671	0.3596484
13	3.4211711	0.3917390	0.3594433
14	3.4223549	0.3918905	0.3593316
15	3.4229908	0.3919740	0.3592712
16	3.4233308	0.3920198	0.3592384
∞	3.4237147	0.3920729	0.3592016

Table 1. $\mathrm{Var}[g(A_n(R^N))]$, $d(L)$ and $e(L)$ for $3 \leq L \leq 16$ and $L \to \infty$

A source is then rejected if and only if either $f^H_{T_U}(s^N) < t_1$ or $f^H_{T_U}(s^N) > t_2$ where the thresholds t_1 and t_2 are defined by :

$$t_1 = L - y\sigma \quad \text{and} \quad t_2 = L + y\sigma,$$

where y is the number of standard deviations σ from the mean allowed for $f^H_{T_U}(s^N)$. The parameter y must be chosen such that $\mathcal{N}(-y) = \rho/2$, where ρ is the rejection rate expressing the probability that a sequence emitted by a truly random source will be rejected. $\mathcal{N}(x)$ is the integral of the normal density function :

$$\mathcal{N}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\xi^2/2} d\xi$$

[6] recommends to choose the parameters L between 6 and 16, $Q \simeq 10 \times 2^L$ and $K \simeq 1000 \times 2^L$, and to take a rejection rate $\rho \simeq 0.01, \ldots, 0.001$, obtained by setting $y = 2.58$ or $y = 3.30$ respectively. We suggest to keep these bounds for the new test.

6 A sample program

As pointed out in [6], the test can be implemented efficiently by using a table `tab` of size $V = 2^L$ that stores for each L-bit block the time index of its most recent occurrence. At step n the program gets the L-bit block $b_n(s^N)$ from the random source, computes the minimum distance $A_n(s^N) \leftarrow n - \texttt{tab}[b_n(s^N)]$, adds $g(A_n(s^N))$ to an accumulator and updates the most recent occurrence table with $\texttt{tab}[b_n(s^N)] \leftarrow n$.

To improve efficiency, the coefficients computed by the function $g(i)$ are approximated for large i using (16). For $i \geq 23$ the error is smaller than 10^{-8}.

$$\sum_{i=1}^{n} \frac{1}{i} = \log n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \mathcal{O}(\frac{1}{n^4}) \tag{16}$$

where γ is Euler's constant :

$$\gamma = -\int_0^{\infty} e^{-x} \log x \; dx \simeq 0.577216$$

The sample program calls the function `fsource(L)` which returns an L-bit integer produced by the random source.

```
double fcoef(int i)
{
       double l=log(2),s=0,C=-0.8327462;
       int k,j=i-1,limit=23;
       if(i<limit) {
              for(k=1;k<i;k++) { s=s+1./k; }
              return s/l;
       }
       return log(j)/l-C+(1./(2*j)-1./(12.*j*j))/l;
}

double NewUniversalTest(int L,int Q, int K)
{
       int V=(1 << L),i,n,k;
       int *tab=new int[V];
```

```
        double sum=0;

        for(i=0;i<V;i++) {
              tab[i]=0;
        }
        for(n=1;n<=Q;n++) {
              tab[fsource(L)]=n;
        }
        for(n=Q+1;n<=(Q+K);n++) {
              k=fsource(L);
              sum=sum+fcoef(n-tab[k]);
              tab[k]=n;
        }
        delete tab;
        return sum/K;
}
```

7 A comparative analysis of the two tests

In section 4 we assumed the block sequences of length L to be statistically independent, *i.e.* that the probability of appearance of a block does not depend on the preceding ones. But this assumption is valid only if the tested source is a binary memoryless source BMS_p (random binary source which emits ones with probability p and zeroes with probability $1-p$). In section 7.1 we compare the performance of Maurer's test and the modified test for a BMS_p.

In the general case of a source with finite (non-zero) memory, the blocks are not statistically independent and the expectation of the modified test function is not equal to the source's entropy of L-bit blocks. However, if the statistics of the tested random source differ from the statistics of a truly random source, the tested source will be rejected with high probability. Only random sources with small statistical bias will pass the test. As shown in section 7.2, this small bias will still make the difference between the expectation of the modified test function and the source's entropy negligible.

7.1 Comparison with respect to a BMS_p.

In this section we compute the expectation of Maurer's test function for a BMS_p and compare it with the expectation of the modified test function and with the actual source's entropy. The expectation of Maurer's test function for a BMS_p with output sequence $U^N_{\mathrm{BMS}_p}$ is given by :

$$E[f_{T_U}(U^N_{\mathrm{BMS}_p})] = \sum_{i=1}^{\infty} \Pr[A_n(U^N_{\mathrm{BMS}_p}) = i] \log_2(i)$$

while equation (8) and :

$$\Pr[b_n(U^N_{\mathrm{BMS}_p}) = b] = p^{w(b)}(1-p)^{L-w(b)}$$

(where $w(b)$ denotes the Hamming weight of $b \in \{0,1\}^L$) yield :

$$E[f_{T_U}(U^N_{\mathrm{BMS}_p})] = \sum_{k=0}^{L} \binom{L}{k} p^k (1-p)^{L-k} \alpha\big(p^k(1-p)^{L-k}\big) \tag{17}$$

where

$$\alpha(x) = x \sum_{i=1}^{\infty} (1-x)^{i-1} \log_2 i$$

One can show that :

$$\lim_{x \to 0}[\alpha(x) + \log_2 x] = -\frac{\gamma}{\log 2} \triangleq C \tag{18}$$

where γ is Euler's constant.

From equations (17 and 18) we recover the result given in [6] :

$$\lim_{L \to \infty} E[f_{T_U}(U^N_{\mathrm{BMS}_p}) - L \times H(p)] = C$$

Note that this result is straightforward using equation (7) as $K_L = L \times H(p)$ for a BMS_p.

In the case of a BMS_p the assumption of statistical independence between the blocks in section 4 is valid and thus equation (13) leads to :

$$E[f^H_{T_U}(U^N_{\mathrm{BMS}_p})] = L \times H(p) \tag{19}$$

Equation (19) shows that the modified test is more accurate than the original one, as it measures the entropy of a BMS_p whereas the relation is only asymptotical in the original one. This is illustrated in table 2, which summarizes the expectation of Maurer's test function, the expectation of the modified test function, and the entropy of a BMS_p, for $L = 4$, $L = 8$, $L = 16$ and several values of p.

7.2 Comparison in the general case.

The mean of the modified test for an ergodic stationary source S is given by :

$$E[f^H_{T_U}(U^N_S)] = \sum_{b \in B^L} \sum_{i \geq 2} \Pr[b(\neg b)^{i-1} b] \sum_{k=1}^{i-1} \frac{1}{k \log(2)}$$

where $\Pr[b(\neg b)^{i-1} b]$ denotes $\Pr[b_n = b, b_{n-1} \neq b, \ldots, b_{n-i+1} \neq b, b_{n-i} = b]$.

Using the fact that $\Pr[b(\neg b)^i] = \Pr[b(\neg b)^i b] + \Pr[b(\neg b)^{i+1}]$, we get :

L	p	$E[f_{T_U}(U^N_{\mathrm{BMS}_p})] - C$	$E[f^H_{T_U}(U^N_{\mathrm{BMS}_p})]$	$L \times H(p)$
4	0.5	4.14397	4.00000	4.00000
4	0.4	4.04187	3.88380	3.88380
4	0.3	3.73034	3.52516	3.52516
8	0.5	8.01641	8.00000	8.00000
8	0.4	7.78833	7.76760	7.76760
8	0.3	7.08957	7.05033	7.05033
16	0.5	16.00012	16.00000	16.00000
16	0.4	15.53542	15.53521	15.53521
16	0.3	14.10161	14.10065	14.10065

Table 2. Comparison between the expectation of Maurer's test $E[f_{T_U}(U^N_{\mathrm{BMS}_p})]$, the expectation of the modified test $E[f^H_{T_U}(U^N_{\mathrm{BMS}_p})]$ and the L-bit block entropy of a BMS_p.

$$E[f^H_{T_U}(U^N_S)] = \sum_{b \in B^L} \sum_{i \geq 1} \Pr[b(\neg b)^i] \frac{1}{i \log(2)}$$

From section 2.2 we obtain the expectation of the modified function in the general case of an ergodic stationary source S with finite memory :

$$E[f^H_{T_U}(U^N_S)] = \sum_{b \in \{0,1\}^L} \sum_{i \geq 1} P \Pi(b) \left(\Pi^L - \Pi(b)\right)^i U \frac{1}{i \log(2)}$$

where Π is the transition matrix of S and $\Pi(b)$ the transition matrix associated to sequence b as defined in section 2.2.

Table 3 gives $E[f^H_{T_U}(U^N_S)]$ for an STP_p, a random binary source for which a bit is followed by its complement with probability p. An STP_p is thus a source with one bit of memory and two equally-probable states. It follows (5 and 6) that $F_1 = H(1/2) = 1$, $H_S = H(p)$, and $K_L = 1 + (L-1)H(p)$. Table 3 compares the mean of Maurer's function, the mean of the modified function and the entropy of L-bit block of an STP_p for $L = 4$ and $L = 8$ and various values of p. As expected, the new test is closer to the source's entropy than the original one.

Moreover, the difference between the expectation of the modified test function and the source's entropy becomes negligible when p is close to 0.5. This is due to the fact that the L-bit blocks become statistically independent as the source's bias disappears. Extensive experiments performed with random sources with memory bigger than one all led the same result.

8 Conclusion and further research

We have introduced a modification in Maurer's universal test that improves its performance. The modification is very simple to implement (a few lines of code)

L	p	$E[f_{T_U}(U^N_{\mathrm{STP}_p}] - C$	$E[f^H_{T_U}(U^N_{\mathrm{STP}_p})]$	$(L-1)H(p)+1$
4	0.5	4.14397	4.00000	4.00000
4	0.49	4.14321	3.99914	3.99913
4	0.45	4.12488	3.97831	3.97832
4	0.4	4.06677	3.91196	3.91285
4	0.3	3.82175	3.62743	3.64387
8	0.5	8.01641	8.00000	8.00000
8	0.49	8.01443	7.99798	7.99798
8	0.45	7.96671	7.94942	7.94942
8	0.4	7.81679	7.79665	7.79665
8	0.3	7.20403	7.16848	7.16904

Table 3. Numerical comparison between the expected value of Maurer's original test $E[f_{T_U}(U^N_{\mathrm{STP}_p}]$, the expected value of the modified test $E[f^H_{T_U}(U^N_{\mathrm{STP}_p}]$ and the L-bit block entropy of an STP_p.

and does not increase the computation time. The new test is more closely related to the source's entropy and therefore enables a more accurate detection of the possible defects in the tested source.

We have not found an analytic expression of the modified test's variance, although the expectation for a truly random source is simply equal to the block length. In addition, an interesting generalization would consist of extending the exact correspondence between the modified test function and the source's entropy to the general class of stationary ergodic random sources with finite (non necessarily zero) memory.

References

1. R. Ash, *Information theory*, Dover publications, New-York, 1965.
2. M. Blum, S. Micali, *How to generate cryptographically strong sequences of pseudo-random bits*. SIAM J. Comput., vol. 13, no. 4, pp. 850-864, 1984
3. J.-S. Coron, D. Naccache, *An accurate evalutation of Maurer's universal test*. Proceedings of SAC'98, Lecture notes in computer science, springer-verlag, 1998. To appear. Available at `http://www.eleves.ens.fr:8080/home/coron/index.html`
4. FIPS 140-1, *Security requirements for cryptographic modules*, Federal Information Processing Standards Publication 140-1, U.S. Department of Commerce / N.I.S.T., National Technical Information Service, Springfield, Virginia, 1994.
5. D. Knuth, *The art of computer programming, Seminumerical algorithms*, vol. 2, Addison-Wesley publishing company, Reading, pp. 2–160, 1969.
6. U. Maurer, *A universal statistical test for random bit generators*, Journal of cryptology, vol. 5, no. 2, pp. 89–105, 1992.
7. C. Shannon, *A mathematical theory of communication*, The Bell system technical journal, vol. 27, pp. 379–423, 623–656, July-October, 1948.

8. J. Ziv, *Compression tests for randomness and estimating the statistical model of an individual sequence*, Sequences, pp. 366-373, 1990.

This article was processed using the LaTeX macro package with LLNCS style