AN ACCURATE EVALUATION OF MAURER'S UNIVERSAL TEST

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

David Naccache
Gemplus Card International
34 rue Guynemer
Issy-les-Moulineaux, F-92447, France
naccache@compuserve.com

Abstract. Maurer's universal test is a very common randomness test, capable of detecting a wide gamut of statistical defects. The algorithm is simple (a few Java code lines), flexible (a variety of parameter combinations can be chosen by the tester) and fast.

Although the test is based on sound probabilistic grounds, one of its crucial parts uses the heuristic approximation :

$$c(L,K) \cong 0.7 - \frac{0.8}{L} + \left(1.6 + \frac{12.8}{L}\right) K^{-4/L}$$

In this work we compute the precise value of $c(L,K)$ and show that the inaccuracy due to the heuristic estimate can make the test 2.67 times more permissive than what is theoretically admitted.

Moreover, we establish a new asymptotic relation between the test parameter and the source's entropy.

1 Introduction

In statistics, *randomness* refers to these situations where care is taken to see that *each individual has the same chance of being included in the sample group.* In practice, random sampling is not easy : being after a random sample of people, it's not good enough to stand on a street corner and select every fifth person who passes as this would exclude habitual motorists from the sample; call on 50 homes in different areas, and you may end up with only housewives' opinions, their husbands being at work; pin a set of names from a telephone directory, and you exclude *in limine* those who do not have a telephone.

Whilst the use of random samples proves helpful in literally thousands of fields, non-random sampling is fatally disastrous in cryptography. Assessing the randomness of noisy sources is therefore crucial and a variety of tests for doing so exists. Interestingly, most if not all such tests are designed around a common skeleton, called *the monkey paradigm.* Informally, the idea consists in measuring the expectation at which a monkey playing with a typewriter would create a meaningful text. Although one can easily conclude that a complex text (*e.g.* the IACR's bylaws) has a negligible monkey probability, a simple word such as `cat`

is expected to appear more frequently (each $\cong 17,576$ keystrokes) and could be used as a basic (yet very insufficient) randomness test.

However, analyzing *textual features* is much more efficient than pattern-scanning where inter-pattern information is wasted without being re-cycled for deriving additional monkeyness evidence.

Usually, parameters such as the average inter-symbol distance or the length of sequences containing the complete alphabet are measured in a sample and a parameter is calculated from the difference between the measure and its corresponding expectation when a monkey, theorized as a binary symmetric source (BSS), is given control over the keyboard. A BSS is a random source which outputs statistically independent and symmetrically distributed binary random variables. Based on the expected distribution of the BSS' parameter, the test succeeds or fails.

We refer the reader to [2, 4] for a systematic treatment of randomness tests and focus the following sections on a particular test, suggested by Maurer in [5].

2 Maurer's universal test

Maurer's universal test [5] 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.

Let B denote the set {0,1}. Denoting by $b_n(s^N) = [s_{L(n-1)+1}, \ldots, s_{Ln}]$ the n-th L-bit block of s^N, the test function $f_{T_U} : B^N \to \mathbb{R}$ is defined by :

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

where,

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

To tune the test's rejection rate, one must first know the distribution of $f_{T_U}(R^N)$, where R^N denotes a sequence of N bits emitted by a BSS. A sample would then be rejected if the number of standard deviations separating its f_{T_U} from $E[f_{T_U}(R^N)]$ exceeds a reasonable constant[1].

For statistically independent random variables the variance of a sum is the sum of variances but the A_n-terms in (1) are heavily inter-dependent; consequently, [5] introduces 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 :

[1] the precise value of $E[f_{T_U}(R^N)]$ is computed in [5] and recalled in section 3.3.

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

A heuristic estimate of $c(L,K)$ is given for practical purposes in [5] :

$$c(L,K) \cong c'(L,K) = 0.7 - \frac{0.8}{L} + \left(1.6 + \frac{12.8}{L}\right) K^{-4/L}$$

In the next section we compute the precise value of $c(L,K)$, under the admissible assumption that $Q \to \infty$ (in practice, Q should be larger than 10×2^L); this enables a much better tuning of the test's rejection rate (according to [5] the precise computation of $c(L,K)$ should have required a considerable if not prohibitive computing effort).

3 An accurate expression of $c(L,K)$

3.1 Preliminary computations

For any set of random variables, we have :

$$\mathrm{Var}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}[X_i, X_j] \tag{3}$$

where $\mathrm{Cov}[X_i, X_j]$ is the covariance of X_i and X_j :

$$\mathrm{Cov}[X_1, X_2] = E[X_1 X_2] - E[X_1] \times E[X_2] \tag{4}$$

Throughout this paper the notation $a_i = \log_2 A_i$ will be extensively used and, unless specified otherwise, A_i will stand for $A_i(R^N)$.

Formulae (1), (2) and (3) yield :

$$c(L,K)^2 = 1 + \frac{2}{K \times \mathrm{Var}[a_n]} \sum_{1 \le i < j \le K} \mathrm{Cov}[a_{Q+i}, a_{Q+j}]$$

Assuming that $Q \to \infty$ (in practice, $Q > 10 \times 2^L$), the covariance of a_i and a_j is only a function of $k = j - i$ and by the change of variables $k = j - i$ we get :

$$c(L,K)^2 = 1 + \frac{2}{\mathrm{Var}[a_n]} \times \sum_{k=1}^{K-1} (1 - \frac{k}{K}) \times \mathrm{Cov}[a_n, a_{n+k}] \tag{5}$$

whereas (4) yields :

$$\mathrm{Cov}[a_n, a_{n+k}] = \sum_{i,\, j \ge 1} \log_2 i \log_2 j \Pr[A_{n+k} = j, A_n = i] \; - E[a_n]^2 \tag{6}$$

Considering a source emitting the random variables $U^N = U_1, U_2, \ldots, U_N$, and letting $b_n = b_n(U^N)$, we get :

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

and, when the $b_n(U^N)$-blocks are statistically independent and uniformly distributed,

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

For a BSS we thus have :

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

3.2 Expression of $\Pr[A_{n+k} = j, A_n = i]$

Deriving the BSS' $\Pr[A_{n+k} = j, A_n = i]$ for a fixed $i \geq 1$ and variable $j \geq 1$ is somewhat more technical and requires the separate analysis of five distinct cases :

- **Disjoint blocks** $1 \leq j \leq k - 1$

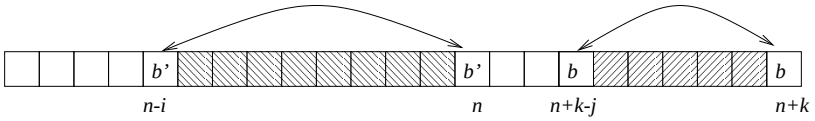

Fig. 1. DISJOINT SEQUENCES.

When $1 \leq j \leq k - 1$, the events $\langle A_{n+k} = j \rangle$ and $\langle A_n = i \rangle$ are independent, as there is no overlap between $[b_{n+k-j} \ldots b_{n+k}]$ and $[b_{n-i} \ldots b_n]$ (figure 1); consequently,

$$\Pr[A_{n+k} = j, A_n = i] = \Pr[A_{n+k} = j] \times \Pr[A_n = i]$$

$$\Pr[A_{n+k} = j, A_n = i] = 2^{-2L}(1 - 2^{-L})^{i+j-2}$$

- **Adjacent blocks** $j = k$

Letting $b = b_{n+k} = b_n = b_{n-i}$ and letting $\mathcal{E}_{j=k}[b]$ be the event (figure 2) :

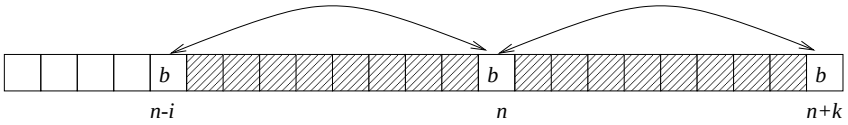

Fig. 2. ADJACENT SEQUENCES.

$$\begin{array}{ll} \mathcal{E}_{j=k}[b] = & \Pr[\mathcal{E}_{j=k}[b]] = \\ \quad \{b_{n+k} = b, & \quad \Pr[b_{n+k} = b] \times \\ \quad b_{n+k-1} \neq b, \ldots, b_{n+1} \neq b, & \quad \Pr[b_{n+k-1} \neq b, \ldots, b_{n+1} \neq b] \times \\ \quad b_n = b, \quad \Rightarrow & \quad \Pr[b_n = b] \times \\ \quad b_{n-1} \neq b, \ldots, b_{n-i+1} \neq b, & \quad \Pr[b_{n-1} \neq b, \ldots, b_{n-i+1} \neq b] \times \\ \quad b_{n-i} = b\} & \quad \Pr[b_{n-i} = b] \end{array}$$

we get,

$$\Pr[\mathcal{E}_{j=k}[b]] = \Pr[b_n = b]^3 \times \Pr[b_n \neq b]^{k+i-2} = 2^{-3L}(1 - 2^{-L})^{k+i-2}$$

$$\Pr[A_{n+k} = k, A_n = i] = \sum_{b \in B^L} \Pr[\mathcal{E}_{j=k}[b]]$$

$$\Pr[A_{n+k} = k, A_n = i] = 2^{-2L}(1 - 2^{-L})^{i+k-2}$$

• **Intersecting blocks** $k + 1 \leq j \leq k + i - 1$

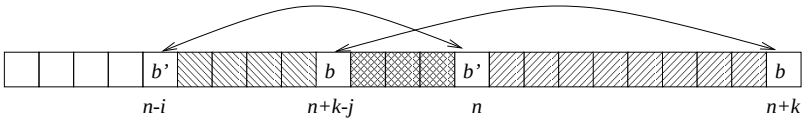

Fig. 3. INTERSECTING SEQUENCES.

For $k+1 \leq j \leq k+i-1$, the sequence $[b_{n+k-j} \ldots b_{n+k}]$ intersects $[b_{n-i} \ldots b_n]$ as illustrated in figure 3. Letting $b = b_{n+k} = b_{n+k-j}$ and $b' = b_n = b_{n-i}$, we get the following configuration, denoted $\mathcal{E}_{k+1 \leq j \leq k+i-1}[b, b']$:

$$\begin{array}{ll} \mathcal{E}_{k+1 \leq j \leq k+i-1}[b, b'] = \{b_{n+k} = b, \\ \qquad b_{n+k-1} \neq b, \ldots, b_{n+1} \neq b, \\ \qquad b_n = b', \\ \qquad b_{n-1} \notin \{b, b'\}, \ldots, b_{n+k-j+1} \notin \{b, b'\}, \\ \qquad b_{n+k-j} = b, \\ \qquad b_{n+k-j-1} \neq b', \ldots, b_{n-i+1} \neq b', \\ \qquad b_{n-i} = b'\} \end{array}$$

whereby :

$$\Pr[A_{n+k} = j, A_n = i] = \sum_{\substack{b,b' \in B^L \\ b \neq b'}} \Pr[\mathcal{E}_{k+1 \leq j \leq k+i-1}[b, b']]$$

$$\begin{aligned} \text{for} \qquad \Pr[b_n = b] &= \Pr[b_n = b'] = 2^{-L} \\ \Pr[b_n \neq b] &= 1 - 2^{-L} \\ \Pr[b_n \notin \{b, b'\}] &= 1 - 2 \times 2^{-L} \end{aligned}$$

and finally :

$$\Pr[A_{n+k} = j, A_n = i] = 2^{-2L}(1 - 2^{-L})^{i+k-2} \left(1 - \frac{1}{2^L - 1}\right)^{j-k-1}$$

• **The forbidden case** $j = k + i$

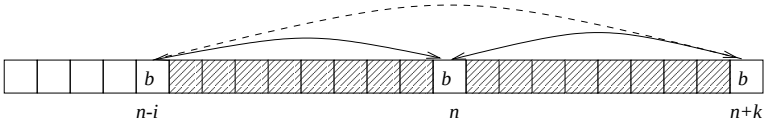

Fig. 4. THE FORBIDDEN CASE.

If $A_n = i$, A_{n+k} can not be equal to $k + i$, as shown in figure 4.

$$\Pr[A_{n+k} = k + i, A_n = i] = 0$$

• **Inclusive blocks** $j \geq k + i + 1$

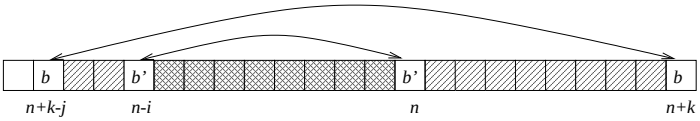

Fig. 5. INCLUSIVE SEQUENCES.

For $j \geq k + i + 1$, the sequence $[b_{n-i} \ldots b_n]$ is included in $[b_{n+k-j} \ldots b_{n+k}]$. As depicted in figure 5, the blocks of $[b_{n+1} \ldots b_{n+k-1}]$ differ from b, those of

$[b_{n-i+1} \dots b_{n-1}]$ differ from both b and b' and those of $[b_{n+k-j+1} \dots b_{n-i-1}]$ differ from b. Letting $\mathcal{E}_{j \geq k+i+1}[b, b']$ be the event :

$$\begin{aligned}\mathcal{E}_{j \geq k+i+1}[b, b'] = \{ & b_{n+k} = b, \\ & b_{n+k-1} \neq b, \dots, b_{n+1} \neq b, \\ & b_n = b', \\ & b_{n-1} \notin \{b, b'\}, \dots, b_{n-i+1} \notin \{b, b'\}, \\ & b_{n-i} = b', \\ & b_{n-i-1} \neq b, \dots, b_{n+k-j+1} \neq b, \\ & b_{n+k-j} = b\}\end{aligned}$$

$$\Pr[A_{n+k} = j, A_n = i] = \sum_{\substack{b, b' \in B^L \\ b \neq b'}} \Pr[\mathcal{E}_{j \geq k+i+1}[b, b']]$$

we obtain :

$$\Pr[A_{n+k} = j, A_n = i] = 2^{-2L}(1 - 2^{-L})^{j-2} \left(1 - \frac{1}{2^L - 1}\right)^{i-1}$$

3.3 Expression of $c(L, K)$

Let us now define the function :

$$h(z, k) = (1 - z) \sum_{i=1}^{\infty} \log_2(i + k) z^{i-1}$$

For a fixed z, the sequence $\Big\{h(z, k)\Big\}_{k \in \mathbb{N}}$ has the inductive property :

$$h(z, k) = (1 - z) \log_2(k + 1) + z \times h(z, k + 1) \tag{7}$$

Let

$$u = 1 - 2^{-L} \qquad \text{and} \qquad v = 1 - \frac{1}{2^L - 1}$$

The expected value $E[f_{T_U}(R^N)]$ of the test parameter $f_{T_U}(R^N)$ for a BSS is given by :

$$E[f_{T_U}(R^N)] = E[a_n] = \sum_{i=1}^{\infty} \log_2 i \times \Pr[A_n = i] = h(u, 0)$$

and the variance of a_n is :

$$\begin{aligned}\mathrm{Var}[a_n] &= E[(a_n)^2] - (E[a_n])^2 \\ &= 2^{-L} \sum_{i=1}^{\infty} (\log_2 i)^2 (1 - 2^{-L})^{i-1} - h(u, 0)^2\end{aligned}$$

From equation (6) and the expressions of $\Pr[A_{n+k} = j, A_n = i]$, one can derive the following expression :

$$\mathrm{Cov}[a_n, a_{n+k}] = u^k \Bigg(h(u,0)\big(h(v,k) - h(u,k)\big) + 2^{-L} \sum_{i=1}^{\infty} \log_2 i \; u^{i-1} v^{i-1} \Big(h(u,k+i) - h(v,k+i-1) \Big) \Bigg)$$

and, using equation (5), finally obtain :

$$c(L,K)^2 = 1 - \frac{2}{\mathrm{Var}[a_n]} \Big(p(L,1) - p(L,K) - \frac{q(L,1) - q(L,K)}{K} \Big)$$

where :

$$p(L,K) = u^{K-1} \sum_{l=1}^{\infty} F(l,L,K) u^{l-1} \quad , \quad q(L,K) = u^{K-1} \sum_{l=1}^{\infty} G(l,L,K) u^{l-1} \quad ,$$

$$F(l,L,K) = u^2 \Big(h(v,l+K-1) - h(u,l+K) \Big) \Big(h(v,0) - v^l h(v,l) \Big) + u \times h(u,0) \Big(h(u,l+K-1) - h(v,l+K-1) \Big)$$

and

$$G(l,L,K) = u \Big(h(v,l+K-1) - h(u,l+K) \Big) \Bigg(u\,(l+K)\,\big(h(v,0) - v^l h(v,l)\big) - 2^{-L} \sum_{i=1}^{l} i \log_2 i \; v^{i-1} \Bigg) + u \Big(l+K-1 \Big) h(u,0) \Big(h(u,l+K-1) - h(v,l+K-1) \Big)$$

3.4 Computing $c(L,K)$ in practice

The functions $h(u,k)$, $h(v,k)$, $p(L,K)$ and $q(L,K)$ are all power series in u or v and converge rapidly ($t = 33 \times 2^L$ terms are experimentally sufficient).

To speed things further,

$$\Big\{ h(u,k) \Big\}_{1 \le k \le 2t} \quad \text{and} \quad \Big\{ h(v,k) \Big\}_{1 \le k \le 2t}$$

could be tabulated to compute $c(L,K)$ in $\mathcal{O}(2^L)$.

For $K \ge t$, we get with an excellent approximation :

$$c(L,K)^2 \cong d(L) + \frac{e(L) \times 2^L}{K} \tag{8}$$

$$\text{where} \quad d(L) = 1 - 2\frac{p(L,1)}{\mathrm{Var}[a_n]} \quad \text{and} \quad e(L) = \frac{q(L,1)}{\mathrm{Var}[a_n]} \times 2^{-L+1}$$

In most cases approximation (8) is sufficient, as [5] recommends to choose $K \geq 1000 \times 2^L > 33 \times 2^L$.

Although rather complicated to prove (ten pages omitted for lack of space), it is interesting to note that asymptotically :

$$\lim_{L\to\infty} (E[f_{T_U}(R^N)] - L) = C \stackrel{\triangle}{=} \int_0^\infty e^{-\xi} \log_2 \xi \, d\xi \cong -0.8327462$$

$$\lim_{L\to\infty} \mathrm{Var}[a_n] = \frac{\pi^2}{6 \ln^2 2} \cong 3.4237147$$

$$\lim_{L\to\infty} d(L) = 1 - \frac{6}{\pi^2} \cong 0.3920729$$

$$\lim_{L\to\infty} e(L) = \frac{2}{\pi^2}(4 \ln 2 - 1) \cong 0.3592016$$

The distribution of $f_{T_U}(R^N)$ can be approximated by the normal distribution of mean $E[f_{T_U}(R^N)]$ and standard deviation :

$$\sigma = c(L,K)\sqrt{\mathrm{Var}[a_n]/K} \tag{9}$$

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

4 How accurate is Maurer's test ?

Let $c'(L,K)$ be Maurer's approximation for $c(L,K)$, and let σ' be the standard deviation calculated under this approximation.

$$c'(L,K) = 0.7 - \frac{0.8}{L} + \left(1.6 + \frac{12.8}{L}\right) K^{-\frac{4}{L}} \tag{10}$$

$$\sigma' = c'(L,K)\sqrt{\mathrm{Var}[a_n]/K}$$

Letting y' be the approximated number of standard deviations away from the mean allowed for $f_{T_U}(s^N)$, a device is rejected if and only if $f_{T_U}(s^N) < t_1$ or $f_{T_U}(s^N) > t_2$, where t_1 and t_2 are defined by :

$$t_1 = E[f_{T_U}(R^N)] - y'\sigma' \quad \text{and} \quad t_2 = E[f_{T_U}(R^N)] + y'\sigma'$$

L	$E[f_{T_U}(R^N)]$	$\mathrm{Var}[a_n]$	$d(L)$	$e(L)$
3	2.4016068	1.9013347	0.2732725	0.4890883
4	3.3112247	2.3577369	0.3045101	0.4435381
5	4.2534266	2.7045528	0.3296587	0.4137196
6	5.2177052	2.9540324	0.3489769	0.3941338
7	6.1962507	3.1253919	0.3631815	0.3813210
8	7.1836656	3.2386622	0.3732189	0.3730195
9	8.1764248	3.3112009	0.3800637	0.3677118
10	9.1723243	3.3564569	0.3845867	0.3643695
11	10.1700323	3.3840870	0.3874942	0.3622979
12	11.1687649	3.4006541	0.3893189	0.3610336
13	12.1680703	3.4104380	0.3904405	0.3602731
14	13.1676926	3.4161418	0.3911178	0.3598216
15	14.1674884	3.4194304	0.3915202	0.3595571
16	15.1673788	3.4213083	0.3917561	0.3594040
∞	$L - 0.8327462$	3.4237147	0.3920729	0.3592016

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

y' is chosen such that $\mathcal{N}(-y') = \rho'/2$, where ρ' is the approximated rejection rate. $\mathcal{N}(x)$ is the integral of the normal density function [3] defined as :

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

The actual number of allowed standard deviations is consequently given by $y = y'\ \sigma'/\sigma$, yielding a rejection rate of $\rho = 2\mathcal{N}(-y) = 2\mathcal{N}(-y'\ \sigma'/\sigma)$.

The worst and average *rationes* ρ'/ρ are listed in table 2 for $3 \leq L \leq 16$ and $1000 \times 2^L \leq K \leq 4000 \times 2^L$ and $\rho' = 0.001$ (*i.e.* $y' = 3.30$), as suggested in [5]. Figures show that the inaccuracy due to (10) can make the test 2.67 times more permissive than what is theoretically admitted.

The correct thresholds t_1 and t_2 can now be precisely computed using formulae (8), (9) and :

$$t_1 = E[f_{T_U}(R^N)] - y\sigma \quad \text{and} \quad t_2 = E[f_{T_U}(R^N)] + y\sigma$$

where y is chosen such that $\mathcal{N}(-y) = \rho/2$ and ρ is the rejection rate.

5 The entropy conjecture

Maurer's test parameter is closely related to the source's per-bit entropy, which measures the effective key-size of a cryptosystem keyed by the source's output. [5] gives the following result, which applies to every binary ergodic stationary source S with finite memory :

L	$\lim_{K\to\infty} c'(L,K)$	$\lim_{K\to\infty} c(L,K)$	worst ρ'/ρ	average ρ'/ρ
3	0.4333333	0.5227547	0.1541921	0.1547350
4	0.5000000	0.5518244	0.3462276	0.3464583
5	0.5400000	0.5741591	0.5058411	0.5097624
6	0.5666667	0.5907426	0.6245271	0.6394724
7	0.5857143	0.6026454	0.7215661	0.7565605
8	0.6000000	0.6109165	0.8118111	0.8775954
9	0.6111111	0.6164930	1.0607613	1.0117992
10	0.6200000	0.6201505	1.2317137	1.1634270
11	0.6272727	0.6224903	1.4245388	1.3337681
12	0.6333333	0.6239543	1.6386583	1.5223726
13	0.6384615	0.6248524	1.8723810	1.7278139
14	0.6428571	0.6253941	2.1234364	1.9481901
15	0.6466667	0.6257157	2.3893840	2.1814850
16	0.6500000	0.6259042	2.6678142	2.4257316

Table 2. A comparison of Maurer's $\{c', \rho'\}$ and the actual $\{c, \rho\}$ values.

$$\lim_{L\to\infty} \frac{E[f_{T_U}(U_S^N)]}{L} = H_S \tag{11}$$

where H_S is the source's per-bit entropy. Moreover, [5] conjectures that (11) can be further refined as :

$$\lim_{L\to\infty} \Big[E[f_{T_U}(U_S^N)] - LH_S\Big] \stackrel{\mathrm{c}}{=} C \stackrel{\triangle}{=} \int_0^\infty e^{-\xi} \log_2 \xi \; d\xi \cong -0.8327462$$

In this section we show that the conjecture is false and that the correct asymptotic relation between $E[f_{T_U}(U_S^N)]$ and the source's entropy is :

$$\lim_{L\to\infty} \Big[E[f_{T_U}(U_S^N)] - \sum_{i=1}^{L} F_i\Big] = C$$

where F_i is the entropy of the i-th order approximation of the source, and :

$$\lim_{L\to\infty} F_L = H_S$$

5.1 Statistical model for a random source

Consider a 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 emitted bits :

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

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].

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

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

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 2^M linear equations :

$$\begin{cases} \sum_{j=0}^{2^M-1} P_j = 1 \\ \\ P_j = \sum_{i=0}^{2^M-1} P_i \ \Pr[\sigma_j|\sigma_i) \qquad \text{for } 0 \leq j \leq 2^M - 2 \end{cases}$$

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

$$H_S = \sum_i P_i H_i \tag{12}$$

5.2 Asymptotic relation between $E[f_{T_U}(U_S^N)]$ and H_S

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

$$E[f_{T_U}(U_S^N)] = \sum_{i \geq 1} \Pr[A_n(U_S^N) = i] \log_2 i \tag{13}$$

with

$$\Pr[A_n(U_S^N) = i] = \sum_{b \in B^L} \Pr[b_n = b, b_{n-1} \neq b, \ldots, b_{n-i+1} \neq b, b_{n-i} = b] \tag{14}$$

Following [6] (theorem 3), the sequences of length L can be looked upon as independent for a sufficiently large L :

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

and

$$E[f_{T_U}(U_S^N)] = \sum_{b \in B^L} \Pr[b]^2 \sum_{i \geq 1} \log_2 i \ (1 - \Pr[b])^{i-1}$$

Re-using the function $v(r)$ defined in [5],

$$v(r) = r \sum_{i=1}^{\infty} (1-r)^{i-1} \log_2 i \tag{15}$$

we have

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

wherefrom one can show that,

$$\lim_{r \to 0} [v(r) + \log_2 r] = \int_0^{\infty} e^{-\xi} \log_2 \xi \, d\xi \stackrel{\triangle}{=} C \cong -0.8327462 \tag{16}$$

which yields :

$$\lim_{L \to \infty} \left[E[f_{T_U}(U_S^N)] + \sum_{b \in B^L} \Pr[b] \log_2 \Pr[b] \right] = C \tag{17}$$

Let G_L be the per-bit entropy of L-bit blocks :

$$G_L = -\frac{1}{L} \sum_{b \in B^L} \Pr[b] \log_2 \Pr[b]$$

then,

$$\lim_{L \to \infty} \left[E[f_{T_U}(U_S^N)] - L \times G_L \right] = C$$

Shannon proved ([6], theorem 5) that

$$\lim_{L\to\infty} G_L = H_S$$

which implies that :

$$\lim_{L\to\infty} \frac{E[f_{T_U}(U_S^N)]}{L} = H_S$$

Let $\Pr[b, j]$ be the probability of a binary sequence b followed by the bit $j \in \{0, 1\}$ and $\Pr[j|b] = \Pr[b, j]/\Pr[b]$ be the conditional probability of bit j after b. Let,

$$F_L = -\sum_{b,j} \Pr[b, j] \log_2 \Pr[j|b] \tag{18}$$

where the sum is taken over all sequences b of length $L-1$ and $j \in \{0, 1\}$. We have :

$$F_L = \sum_{b\in B^{L-1}} \Pr[b] H(\Pr[1|b])$$

and, by virtue of Shannon's sixth theorem (*op. cit.*) :

$$F_L = L \times G_L - (L-1)G_{L-1} \ , \quad G_L = \frac{1}{L}\sum_{i=1}^{L} F_i$$

and

$$\lim_{L\to\infty} F_L = H_S$$

wherefrom

$$\lim_{L\to\infty} \left[E[f_{T_U}(U_S^N)] - \sum_{i=1}^{L} F_i\right] = C$$

5.3 Refuting the entropy conjecture

F_L is in fact the entropy of the L-th order approximation of S [1, 6]. Under such an approximation, only the statistics of binary sequences of length L are considered. After a sequence b of length $L-1$ has been emitted, the probability of emitting the bit $j \in \{0, 1\}$ is $\Pr[j|b]$. The L-th order approximation of a source is thus a binary stationary source with less than $L-1$ bits of memory, as defined in section 5.1. A source with M bits of memory is then equivalent to its L-th order approximation for $L > M$, and thus $\forall i > M, F_i = H_S$, and :

$$\lim_{L\to\infty} \left[E[f_{T_U}(U_S^N)] - \sum_{i=1}^{M} F_i - (L-M)H_S\right] = C$$

For example, considering a BMS_p (random binary source which emits ones with probability p and zeroes with probability $1-p$ and for which $M=0$ and $H_S = H(p)$), we get the following result given in [5] :

$$\lim_{L\to\infty}\left[E[f_{T_U}(U_S^N)] - LH(p)\right] = C$$

The conjecture is nevertheless refuted by considering an STP_p which is a random binary source where 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 0 and 1. It follows (12 and 18) that $F_1 = H(1/2) = 1$, $H_S = H(p)$, and :

$$\lim_{L\to\infty}\left[E[f_{T_U}(U_S^N)] - (L-1)H_S - 1\right] = C$$

which contradicts Maurer's (7-years old) entropy conjecture :

$$\lim_{L\to\infty}\left[E[f_{T_U}(U_S^N)] - LH_S\right] \stackrel{\mathrm{c}}{=} C$$

6 Further research

Although the universal test is now precisely tuned, a deeper exploration of Maurer's paradigm still seems in order : for instance, it is possible to design a $c(L,K)$-less test by using a newly-sampled random sequence for each $A_n(s^N)$ (since in this setting the $A_n(s^N)$ are truly independent, $c(L,K)$ could be replaced by one). Note however that this approach increases considerably the total length of the random sequence; other theoretically interesting generalizations consist in extending the test to non-binary sources or designing tests for comparing generators to biased references (non-BSS ones).

References

1. R. Ash, *Information theory*, Dover publications, New-York, 1965.
2. D. Knuth, *The art of computer programming, Seminumerical algorithms*, vol. 2, Addison-Wesley publishing company, Reading, pp. 2–160, 1969.
3. R. Langley, *Practical statistics*, Dover publications, New-York, 1968.
4. G. Marsaglia, *Monkey tests for random number generators*, Computers & mathematics with applications, vol. 9, pp. 1–10, 1993.
5. U. Maurer, *A universal statistical test for random bit generators*, Journal of cryptology, vol. 5, no. 2, pp. 89–105, 1992.
6. C. Shannon, *A mathematical theory of communication*, The Bell system technical journal, vol. 27, pp. 379–423, 623–656, July-October, 1948.

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