True Random Number Generation of Very High Goodness-of-fit and Randomness Qualities

Süleyman Gökhan Tanyer  
Department of Electrical and Electronics Engineering  
Faculty of Engineering, Başkent University  
Eskisehir yolu 20. km., Etimesgut, 06810 Ankara, Turkey  
Email: gokhun.tanyer@gmail.com, gokhuntanyer@baskent.edu.tr

Abstract—The statistical nature of numerous problems in mathematics, physics and engineering have led to the development of methods for generating random data for a given distribution. Ancient methods include; dice, coin flipping and shuffling of cards. Today, various pseudo, quasi and true random generators (RNGs) are being proposed for their improved properties. In this work, test metrics for goodness-of-fit and randomness are reviewed. The method of uniform sampling (MUS) is modified for improving the randomness without harming the goodness-of-fit qualities. The test results illustrate that very high goodness-of-fit can be obtained even when the number of observed samples is as small as 10.

Keywords—random number generator, statistical signal processing, probability, test statistics.

I. INTRODUCTION

The nature of chance had been a very attractive mystery to mankind even before the Sumerians played the royal game of the 20-squares in the Mesopotamian city of Ur in 4,600 B.C. Probably the oldest random number generator was the talus bone, which was before dice, coins or cards [1]. Today, probability is a rich branch of pure mathematics with the role as a foundation for science; mathematics, engineering and applied statistics [2]. In digital signal processing, statistical mathematics lies in the center of many applications including; detection and estimation, acoustic signal processing and optimization [3]. Systems are often analyzed by statistical simulations which all rely on reference random data. RNGs are mostly selected for their randomness as well as their goodness-of-fit for a given desired distribution. Test statistics provide quantitative observations for selected features of a RNG. There are numerous tests available and generally battery of tests are used instead of relying on a single observation [4]. Designing of numerical simulators are often challenged by practical limitations such as CPU time and computer memory. Those limitations could force the developer to use a few multi runs for a Monte Carlo simulation, a few number of populations for the genetic algorithm etc. Those critical decisions could easily affect the repeatability of numerical results. A few number of trials are often used as a method of validation without being confirmed by test statistics. It is recently shown that even true RNGs of high randomness suffer goodness-of-fit test when the number of observed samples are small. This illustrates the need for a RNG of high randomness and goodness-of-fit qualities for finite observed samples. In this work, the method of uniform sampling is modified and random data is generated (MUS-RNG) with exact statistics (mean, variance). MUS-RNG is later processed with a true RNG of high randomness to yield random numbers of very high goodness-of-fit and randomness qualities.

II. STATISTICAL ANALYSIS OF FINITE DATA

The discrepancy between a theoretical distribution and the observed properties for a finite data can be analyzed using some measures of lack of fit. Test metrics can be defined for measuring the ‘distance’ of a finite data from a given probability distribution. Observed samples tests are proposed using the Kolmogorov–Smirnov, Cramer-von Mises and Anderson–Darling test statistics [5, 6, 7].

For an ordered set of N elements, \( X_n = \{x_n : x_n \in \mathbb{R}, n = 1, 2, \ldots, N \} \) where \( x_p < x_r \) for \( 1 \leq p < r \leq N \), the average statistical error for a single set of observed N elements can be defined as

\[
E(N) = \left[ \frac{1}{N} \sum_{n=1}^{N} e(x_n)^2 \right]^{1/2} \tag{1}
\]

where the lack of fit error for a single sample is given by

\[
e(x_n) = \left| F_X(x_n) - F_{X,N}(x_n) \right| \tag{2}
\]

where \( F_X \) and \( F_{X,N} \) are the cumulative distribution function (CDF) and the N-element cumulative distribution function (N-CDF) respectively [7]. N-CDF representation of a sample set varies for each observation as shown in Fig. 1. The lack of fit error is calculated for uniformly distributed random variable of [8] in Fig. 2. Note that, for \( N \to \infty, e \to 0 \) since \( F_{X,N} \) converges to \( F_X \) as expected.

One can further measure the statistical quality of a RNG by

\[
Q(N, S) = -10 \log \left[ \frac{1}{N} \sum_{n=1}^{N} E_n(N, S)^2 \right] \tag{3}
\]

where \( E_n \) is the average statistical error for the nth order sample for the data generator calculated using S independent data sets of length N, and
and number of observed samples (Fig. 1. 

Fig. 1. The N-CDF $F_{N}(x)$ for the observed samples uniform in [-0.5, 0.5], and number of observed samples (S); (a) $N = 10$, (b) $N = 100$, (c) $N = 1,000$ and (d) $N = 10,000$.

$$E_n(N,S) = \left[ \frac{1}{S} \sum_{n=1}^{S} e_n(x_n)^2 \right]^{1/2}.$$  

(4)

Note that $Q(N,S) \rightarrow \infty$ if every $e_n(x_n) \rightarrow 0$.

III. MODIFIED METHOD OF UNIFORM SAMPLING (MUS)

An ordered set of numbers can be obtained by simply sampling the cumulative distribution function (CDF) itself [7]

$$z_n = F_{X}^{-1}(p_n)$$  

(5)

where $N$ is the data length, $0 < p_n < 1$ ($0 \leq n \leq N$) are the uniform samples obtained from the probability axis, $\Delta \rho/(1 - 2 \delta)$ is the sampling size, $\delta$ is a random variable uniform in {0, 1/2N} ($\delta = 1/(2N)$ in [7]) and $z_n$ are the quantiles for the given CDF (Fig. 3). Note that $z_n$ have approximately zero fitting error and infinite quality factor according to (1) and (3) respectively. This is true because random quantiles satisfy $F_{X}(z_n) = F_{X}^{-1}(z_n) = p_n$ as shown in Fig. 4.

IV. RANDOM NUMBER GENERATOR OF HIGH MERITS OF GOODNESS-OF-FIT AND RANDOMNESS

True random number generators (RNGs) utilize generally chaotic undeterministic natural noise sources which are assumed to have no deterministic human effects. Unfortunately, those natural sources provide randomness but lack goodness-of-fit especially when the observed number of samples is small. The data size and the number of iterations are usually insufficient in statistical analysis utilizing Monte Carlo or similar multi run simulations. It is a good idea to somehow merge very high goodness-of-fit data obtained from MUS-RNG and highly random TRNG data to be able to generate random data of both good merits. Graphical representation of the novel true random number generator is shown in Fig. 5.

Noise source is illustrated by the Sun’s wideband RF emissions as an example. RF signal is later down converted (demodulated and low pass filtered) to the baseband. It is sampled by the analog to digital converter to obtain raw data $r_n$ and $\delta$. Later, MUS-RNG provides the quantiles, $z_n$. Simple procedure required for sorting $r_n$ provides the unsorting (shuffling) procedure for $z_n$ to yield the MUS-TRNG data $m_n$. Note that the Sun’s TRNG data are statistically corrected by MUS-RNG.

V. TEST RESULTS

MUS-TRNG samples uniform in [0, 1] are tested for their randomness and goodness-of-fit properties. Different data lengths of data ($N = 10, 50, 100, 500, 1,000$) are examined using independent set of 30 in Table I. It is observed that both randomness and goodness-of-fit merits can be obtained even for a data length of 10. A complete set of tests is given in [9]. Similar tests on MUS-TRNG [10, 11] show that MUS-TRNG
Table I.
The Sun’s emissions MUS–TRNG Test Results

<table>
<thead>
<tr>
<th>$n$</th>
<th>Run Test*</th>
<th>Goodness-of-Fit</th>
<th>Mean ± SD**</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>30/30</td>
<td>30/30</td>
<td>0.550 ± 0.303</td>
</tr>
<tr>
<td>50</td>
<td>29/30</td>
<td>30/30</td>
<td>0.510 ± 0.291</td>
</tr>
<tr>
<td>100</td>
<td>27/30</td>
<td>30/30</td>
<td>0.505 ± 0.290</td>
</tr>
<tr>
<td>500</td>
<td>27/30</td>
<td>30/30</td>
<td>0.501 ± 0.289</td>
</tr>
<tr>
<td>1000</td>
<td>30/30</td>
<td>30/30</td>
<td>0.501 ± 0.289</td>
</tr>
</tbody>
</table>

(*) a/b; a: Number of positive results, b: Total of 30 sets.
(**) SD; Standard deviation value.

is a valuable source of random data, especially where data length is small and number of multi-run iterations are limited due to CPU time and computer memory.

ACKNOWLEDGMENT

This work is supported by Baskent University Research Grant No: BA14/FM-13 (BAP).

Fig. 3. The method of uniform sampling; sampling the probability axis for the synthesis of the surrogate data.

Fig. 4. (a) N-CDF for the synthesized data for standard normal random variable, (o) $N = 10$, (*) $N = 50$, (–) $N = 100$, (–) $N = 1,000$, (-) $N = 10,000$.

Fig. 5. Novel true random number generation utilizing both the modified method of uniform sampling (MUS) and the standard true random number generation method (MUS-TRNG); RFDC: RF down conversion, ADC: analog to digital conversion, MUS: (modified) method of uniform sampling.

REFERENCES