Random Number Generation with the Method of Uniform Sampling: Very High Goodness of Fit and Randomness

S. Gokhun Tanyer, *Member, IAENG*

***Abstract*— System models in general are developed to predict outcomes for given inputs. However, the models used in simulations necessarily involve random variables when knowledge of the system is probabilistic. Various optimization methods require randomly generated populations. Today, various pseudorandom and true random number generators (RNGs) are continually developed to improve performance in various fields of science, including mathematics, physics, and engineering. Here we propose two test metrics to measure the goodness of fit error and the quality of an RNG based on improved empirical cumulative distribution function (IECDF). An RNG based on the method of uniform sampling, MUS-RNG, is proposed and demonstrated to provide high goodness of fit and randomness which is shown to have very small error even for a set of 10. MUS-RNG is compared with various true and pseudo-RNGs and tested on both uniform and standard normal distributions. Two quantitative benchmarking tests are proposed. It is also observed that MUS-RNG is also very successful for discontinuous cumulative distribution functions. The comparative results show that MUS-RNG has very small goodness of fit error and is easy to implement. The algorithm has the potential to provide higher convergence in optimization problems and accuracy in statistical simulations.**

***Index Terms*—Digital signal processing, method of uniform sampling (MUS), optimization, improved empirical cumulative distribution function (IECDF), improved empirical probability density function (IEPDF), probability distribution, random number generation (RNG)**

I. Introduction

THE nature of chance has intrigued mankind since at least 2500 BCE, when the Sumerians played the Game of 20 Squares, which was discovered in the ruins of the ancient Mesopotamian city of Ur. The oldest commonplace random number generator may have been the talus or astragalus bone of animals, which seems to have been used as a four-sided die [1]. Today, probability is a rich branch of pure mathematics with a foundational role in science, engineering, and applied statistics [2]. Statistical mathematics lies at the center of many applications, including signal, image, and video processing [3, 4]; detection and estimation [5]; and optimization. Systems are often analyzed using statistical simulations that rely on reference random data [6, 7]. Random number generators (RNGs) are not only selected for their randomness, but also for their goodness of fit (GOF) to a given distribution. Test statistics provide quantitative performance observations. There are numerous tests available, and generally, batteries of such tests are used [8-10]. The design of numerical simulation techniques is challenging, including practical constraints such as computer processing time and memory [11]. Such limitations can lead developers to use an insufficient number of iterations for Monte Carlo simulations [12-14] or insufficiently large populations for applications such as the genetic [15, 16], particle swarm optimization [17, 18], ant colony [19, 20], invasive weed [21], and gravitational search algorithms [22, 23]. Such critical decisions can easily affect the repeatability of numerical results. Insufficient numbers of trials are often used as in methods of validation without any quantitative confirmation by test statistics.

True RNGs employ a kind of natural stochasticity. Generally, the least significant bits obtained from digital sampling of some physical phenomenon are concatenated to generate truly random numbers. Conversely, pseudo-RNGs are deterministic bit generators that rely on algorithms to generate sequences of numbers with the properties that approximate those of truly random numbers. It has been shown that both types of RNG, of high randomness, can fail GOF tests for observed numbers of samples as large as 100,000 [24]. Random numbers with low GOF error are required in various fields, including mathematics, physics and engineering.

The rest of this paper proceeds as follows:

In Section II, the improved empirical cumulative distribution function (IECDF) and the improved empirical probability density function (IEPDF) are proposed. Test statistics for GOF are reviewed, and a novel statistic for the quality of RNGs is proposed in Section III. The method of uniform sampling (MUS) is improved for better GOF results in Section IV, and a novel random number generator based on MUS is proposed in Section V. Finally, the results of numerical tests are presented in Section VI.

II. The Improved ECDF and The Improved EPDF

The cumulative distribution function (CDF) and the probability density function (PDF) are theoretical asymptotic

Manuscript received April 23, 2017; revised November 30, 2017. This work was supported in part by Baskent University research grant No: BAP-BA14/FM-13.

S. G. Tanyer is on his academic sabbatical leave in the Department of Electrical and Computer Engineering, University of Victoria, Canada. He is with the Department of Electrical-Electronics Engineering, Faculty of Engineering, Baskent University, Eskisehir yolu 18. km, Etimesgut, Ankara, Turkey (phone: +1-250-507-5609; fax: +90-312-246-6707; e-mail: gokhun.tanyer@gmail.com).

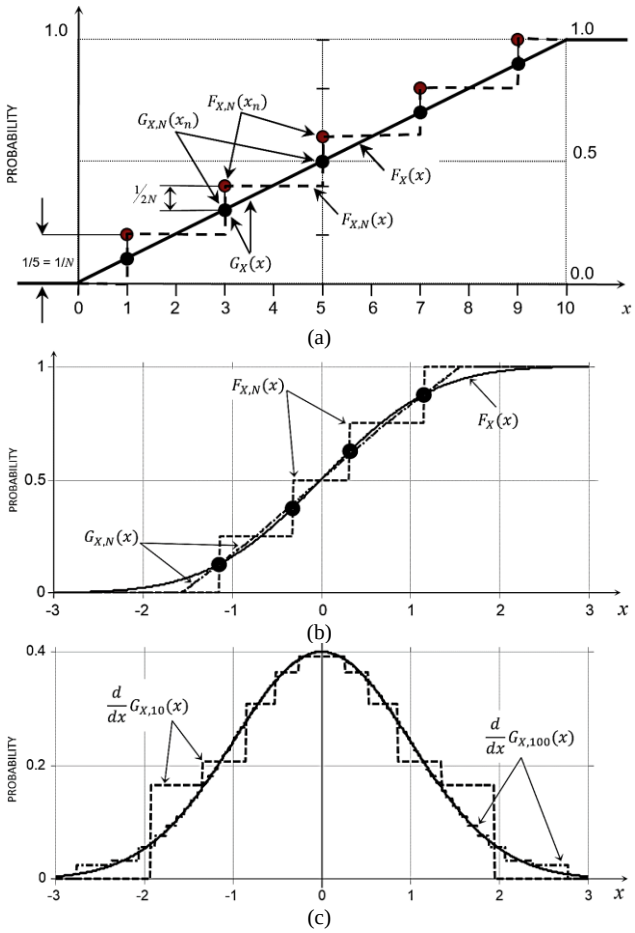

Figure 1. Test metrics defined by (1) and (2). (a) CDF (*solid line*), IECDF $G_{X,N}(x)$ (*dot-dashed line*; this overlaps with the CDF) from (4), and ECDF (*dashed line*). The set {1, 3, 5, 7, 9} was obtained using the method of uniform sampling (MUS) described in Section IV (N = 5). A uniform distribution in (0, 10) is assumed for the CDF. (b) Same as (a), but for normally distributed MUS samples (N = 4). (c) Comparison of the PDF and the IEPDF for N = 10 and N = 100.

functions that are accurate when the number of observed samples in a set approaches infinity. In applied digital signal processing, however, the acquired signals in, acoustics, telecommunications, and optical and geophysical applications are sampled for a limited time and at a given rate. Thus, every acquisition yields a finite number of observed samples, and the CDF and PDF cannot provide exact knowledge. The empirical cumulative distribution function (ECDF) is often used for analysis of finite data sets; the fundamental ideas for such testing are attributed to Kolmogorov, Smirnov, Cramér, and von Mises [25-27].

The ECDF of a random sample is the uniform discrete measure on the observations [26]. For a real-valued independent and identically distributed (i.i.d.) random variable X: $\mathbb{R} \to \mathbb{R}$, the ECDF of the ordered set of observed samples $x_n = \{ x_1, x_2 \ldots x_N\}$ is the function $F_{X,N}: \mathbb{R} \to [0, 1]$ defined as

$$F_{X,N}(x) = \frac{1}{N}\sum_{n=1}^{N} I(x_n \le x) \tag{1}$$

where $I\ (x_n \le x)$ is the indicator function which is equal to 1 if $x_n \le x$ and 0 otherwise, and n = 1, 2 ... N. The ECDF is shown to converge in $\mathbb{R}$: [∞,∞] to the CDF in distribution using Donsker's theorem [26, 28].

The ECDF is a stair-case function and the empirical probability density function (EPDF) is a train of impulses which are not very helpful. Let us consider the piecewise linear (PWL) test function as the improved empirical cumulative distribution function (IECDF)

$$G_{X,N}(x) = \frac{1}{N}\left(n - \frac{1}{2}\right) + \frac{x - x_n}{N\Delta x_n} \tag{2}$$

for $x_n < x < x_{n+1}$ and $1 \le n \le N - 1$, where $\Delta x_n = x_{n+1} - x_n$, and $G_{X,N}(x) = 0$ for $x < x_n$, and 1 for $x_N < x$. The IECDF can be shown to converge $F_X(x)$ using Donsker's theorem since the maximum bias between $G_{X,N}$ and $F_{X,N}$ (ECDF) is $1/(2N)$ at each sample, and converge to 0 for $N \to \infty$ [26] and

$$\begin{aligned} G_X(x) &= \lim_{N\to\infty}\left[G_{X,N}(x)\right] \\ &= \lim_{N\to\infty}\left[\frac{1}{N}\left(n - \frac{1}{2}\right) + \frac{x - x_n}{N\Delta x_n}\right] \end{aligned} \tag{3}$$

where $G_X(x) = F_X(x)$. The CDF, ECDF and the IECDF are illustrated for uniform and standard normal distributions in Fig. 1 (a) and (b). The definition of $G_{X,N}(x)$ requires special care outside the region (x_1, x_N), as illustrated in Fig. 1 (a) for x_1 = 1 and x_N = 9 for N = 5. Equation (2) is nonzero in the leftmost region, where $x_1 - \Delta x_1/2 < x < x_1$ for n = 1, and in the rightmost region where $x_N < x < x_N + \Delta x_{N-1}/2$ for $n = N$. Note that the IECDF has zero bias and perfectly coincides with the CDF for the uniform distribution as shown in Fig. 1 (a) for the samples obtained using the method of uniform sampling (MUS) as described in Section IV.

The IECDF is a continuous PWL function (first-order polynomial function) and its derivative can be analyzed

$$\begin{aligned} g_X(x) &= \frac{d}{dx}G_X(x) \\ &= \lim_{N\to\infty}\left(\frac{x - x_n}{N\Delta x_n}\right) \end{aligned} \tag{4}$$

where $g_X(x) = f_X(x)$ is the probability density function (PDF) of the random variable X. The improved empirical probability density function (IEPDF) can be defined as

$$g_{X,N}(x) = \frac{x - x_n}{N\Delta x_n}. \tag{5}$$

The PDF and the IEPDF are illustrated in Fig. 1 (c) for N = 10 and N = 1000. Note that the IEPDF is very successful in representing of the PDF for finite data.

There are various studies on the estimation of the probability density function such as, design of tunable-kernel models based on orthogonal forward regression procedure [29], over-sampling approach using Parzen-window kernel function [30], a hybrid method of minimum frequency and maximum entropy [31], multi-rate signal processing approach [32]. It is interesting that the novel IEPDF is accurate and practical to be implemented, and yet a priori information and kernel optimization are not necessary for the estimation.

III. Test Metrics for Goodness of Fit

The discrepancy between a continuous one-dimensional CDF and the ECDF can be analyzed using some measure of the lack of fit [25-27]. Statistical tests are preferred to estimate whether a given sample was drawn from a given

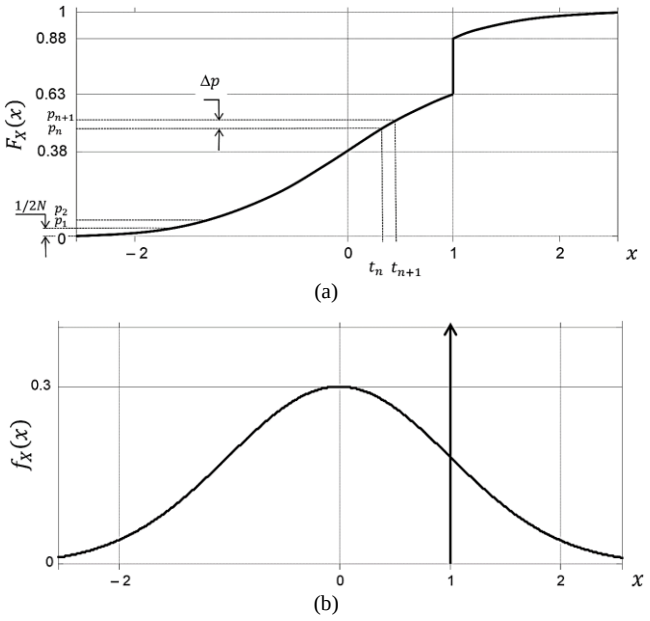

Figure 2. Illustration of the method of uniform sampling. The probability (*y*-axis) of the CDF is sampled for the generation of quantiles t_n (*x*-axis), as shown in (a). The method is illustrated for a discontinuous CDF with the PDF shown in (b), which is the sum of the standard normal distribution (mean and variance of 1) and a delta function at $x = 1$, with observation probabilities of 0.75 and 0.25, respectively.

probability distribution. For example, the area

$$\mathrm{Err} = \int_{x_A}^{x_B} [F_X(x) - F_{X,N}(x)]^2 w(x) dx \tag{6}$$

is expected to be smaller for statistically more representative data, where $F_X(x)$ is the CDF, w is some weighting function and x_A and x_B are selected such that $F_X(x_A)$ and $F_X(x_B)$ are ≈0 and ≈1, respectively, and where the ECDF is given in (1). The potential contribution to the integral from outside (x_A, x_B) can be assumed to be negligible.

The total goodness-of-fit error for a set of observed samples statistic based on the distance between the *n*th observed sample x_n and its theoretical CDF value has previously been proposed [24, 33],

$$\mathrm{Err}_N = \left[\frac{1}{N} \sum_{n=1}^{N} (\mathrm{err}_n)^2 \right]^{1/2} \tag{7}$$

where err_n is the lack of fit of the *n*th sample, given by

$$\mathrm{err}_n = \left| F_X(x_n) - F_{X,N}(x_n) \right| \tag{8}$$

and where F_X and $F_{X,N}$ are the CDF and the ECDF, respectively. The ECDF $F_{X,N}(x) = 0$ for $x < x_1$, n/N for $x_1 < x < x_N$, and 1 for $x < x_N$, respectively. It is shown to converge to the CDF as $N \to \infty$ in [26, 27], and this is illustrated numerically in Section VI.

For a set of N observed samples for a real-valued i.i.d. random variable X, each value x_n is taken to have an equal observation probability of $1/N$, and the empirical probability density function (EPDF) is the sum of N equal delta functions located at the points $x = \{x_1, x_2, \ldots, x_N\}$. Therefore, the ECDF is a stair step function obtained simply from the integral of the EPDF, as shown in Fig. 1 (a). The test metric (8) from [24, 33] is improved upon here. The improved, although similar, metric

$$\mathrm{err}_n = \left| F_X(x_n) - G_{X,N}(x_n) \right| \tag{9}$$

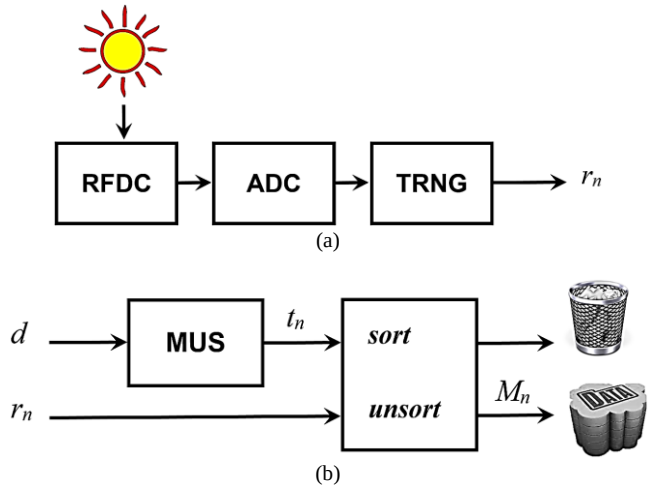

Figure 3. (a) A standard true RNG using solar RF emission; (b) random number generation using MUS-RNG. RFDC, RF down conversion; ADC, analog-to-digital conversion; TRNG, true RNG.

avoids this measurement bias, which can dominate the total GOF error of a set, especially when the number of samples is small.

The ECDF and the IECDF are illustrated Fig. 1 (b) for the standard normal distribution. It can be seen that the ECDF and the CDF have a bias of $1/(2N)$ at each sample point, and these contribute to the total GOF error through (7). This shows that the novel GOF test metric (9) is better than (8), especially for small sets, because the GOF error is not dominated by a superficial bias.

It is possible to calculate the average GOF error for the *n*th sample of a RNG output of length N as

$$\mathrm{Err}_{n,S} = \left[\frac{1}{S} \sum_{s=1}^{S} (\mathrm{err}_{n,S})^2 \right]^{1/2} \tag{10}$$

where $\mathrm{err}_{n,S}$ is calculated for the *n*th sample in the *s*th set using (9). The quality of the GOF of a RNG can be quantified,

$$Q_{N,S} = -10 \log \left\{ \left[\frac{1}{N} \sum_{n=1}^{N} (\mathrm{Err}_{n,S})^2 \right]^{1/2} \right\}. \tag{11}$$

Note that $Q_{n,S} \to \infty$ if every $\mathrm{Err}_{n,S} \to 0$, which is expected.

IV. Method of Uniform Sampling

An ordered set of numbers can be \obtained by simply sampling the CDF according to

$$t_n = F_X^{-1}(p_n) \tag{12}$$

where the t_n, $n = 1, 2, \ldots, N$, are the quantiles for a given CDF obtained by sampling the probability axis, with N the total number of samples (quantiles), and the $p_n = (n - 1/2)/N$ are the probability values, with the sampling interval being $\Delta p = 1/N$ where $p_1 = d = 1/(2N)$. Here the sampling points on the probability (vertical) axis proposed in [24, 33] are improved by shifting them to the middle of each probability region defined by Δp such that $F_X(t_n) = p_n \approx G_{X,N}(t_n)$. The method is corrected by omitting the use of the sampling parameter δ and Δp of [28].

MUS is illustrated in Fig. 2 for the PDF

$$f_X(x) = \frac{3}{4\sqrt{2\pi}} e^{-x^2/2} + \frac{1}{4} \delta(x - 1). \tag{13}$$

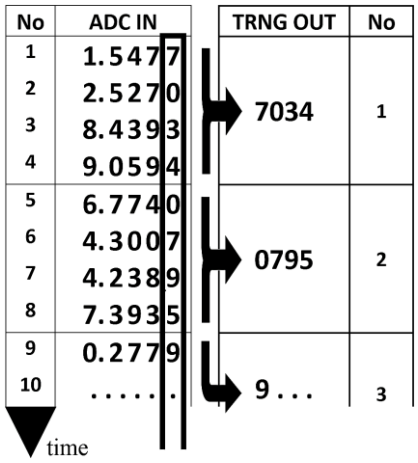

Figure 4. Practical true random number generation using digital samples of solar RF emission observed at the output of the analog-to-digital converter (ADC IN) illustrated in Fig. 3 (a). The least significant digits for every M samples are concatenated to generate each M-digit random value (here M = 4) (TRNG OUT). True random numbers uniformly distributed in (0, 1), r_n, are obtained by normalization to 1.

Method provides solutions for distributions where the PDF tails extend to infinity, $x_A = t_1 \to -\infty$ and $x_B = t_N \to +\infty$, because the potential contribution to the integral outside (x_A, x_B) can be assumed to be negligible in accordance with (6).

V. Random Number Generation Using MUS

True RNGs make use of natural sources of entropy, which are assumed to be unaffected by deterministic human artifacts. Unfortunately, these natural sources provide randomness but can lack GOF to a uniform distribution, especially when the number of observed samples is small. For example, in Monte Carlo or similar simulations, the total number of observed samples for a desired random variable used in the initialization and iteration steps could be insufficient. This is a result of physical limitations¸ namely, computer time and memory, and could threaten the replicability of a study, especially if the GOF error is unexpectedly large.

The proposed method-of-uniform-sampling true RNG (MUS-RNG) is illustrated schematically in Fig. 3. Data from the San Vito Solar Observatory (SVSO) are used as an example of a natural entropy source, but any type of RNG can be used for this initial step. Solar radio emissions and the physical mechanisms behind this RF radiation have been reviewed by Shibasaki et al. [34]. SVSO is one of the observatories of the Radio Solar Telescope Network [35]. In short, the sun can be used as a source of natural entropy to generate truly random numbers. The archival data used in this study consist of 1 s binned radio flux measurements from eight distinct frequency bands obtained over 31 days in January 1990.

RF signals are generally down converted (demodulated and low-pass filtered) to the baseband. They are sampled by an analog-to-digital converter to produce raw digital samples, as shown in Fig. 3 (a). The least significant decimal digits of successive samples are then concatenated to

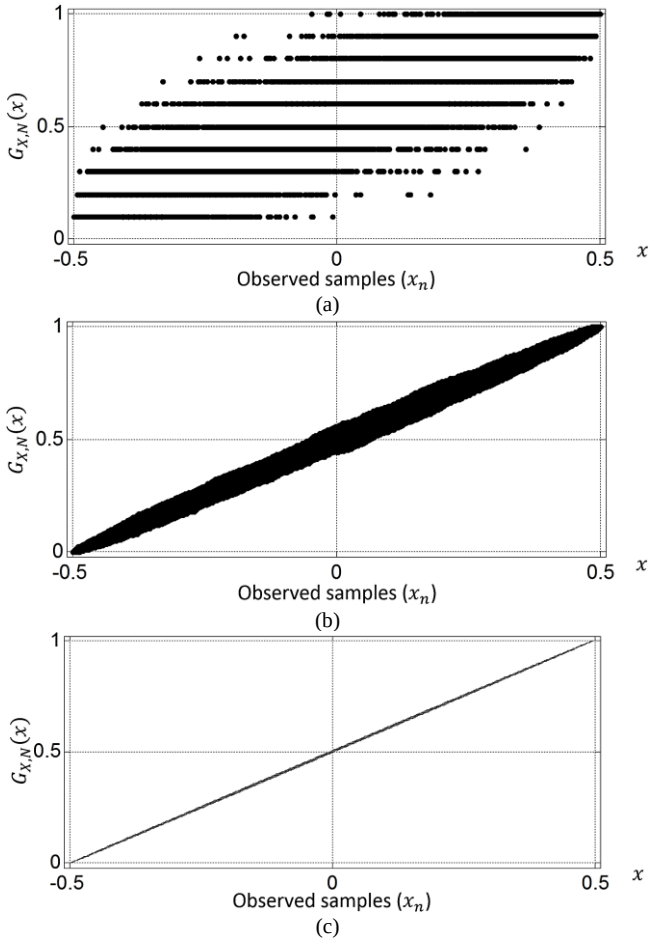

Figure 5. IECDF for the observed samples (uniform in [–0.5, 0.5]). (a) N = 10; (b) $N = 10^3$; (c) $N = 10^5$ (S = 1000). Each dot corresponds to an observed sample in the set. The dots were selected to be very small in (c).

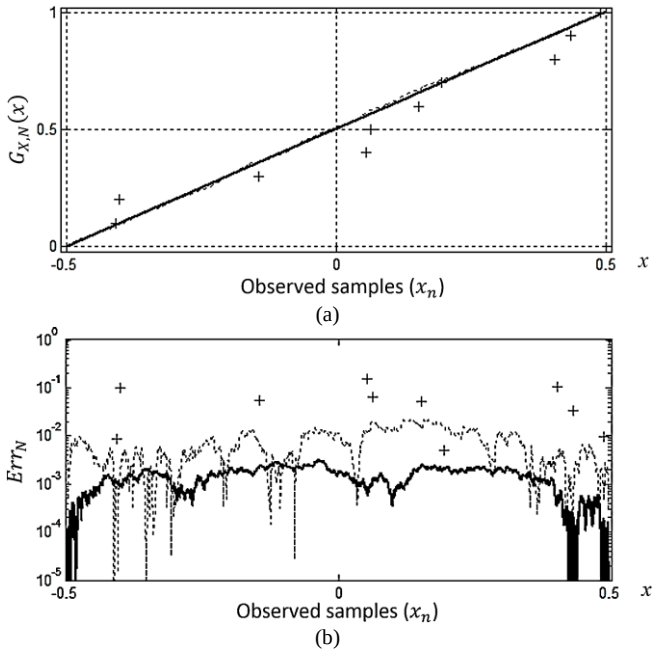

Figure 6. (a) IECDF for the observed (uniform in [−0.5, 0.5]) samples; (b) GOF error Err_N defined by (5). Plus signs, N = 10; dashed lines, $N = 10^3$; solid lines, $N = 10^6$.

produce random numbers r_n as illustrated in Fig. 4. MUS is used to generate the quantiles t_n as shown in Fig. 3 (b). The simple procedure required for sorting the r_n provides the unsorting (shuffling) information for the t_n, which yields the final MUS-RNG samples M_n. The GOF measures for t_n and

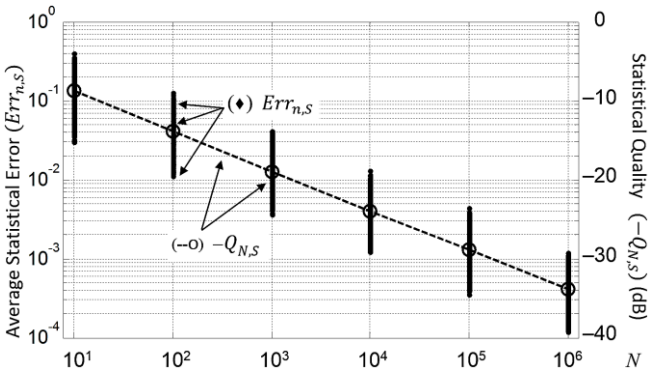

Figure 7. Average statistical error for a uniform distribution in [−0.5, 0.5] and the negative of the statistical quality (S = 10,000) as a function of data length N. The average GOF error, $Err_{n,S}$, and the GOF quality, $Q_{N,S}$, are defined in (10) and (11), respectively. The same set of observed samples as in Figs. 4 and 5 is used.

M_n are equal according to (5), because they both have exactly the same ECDF. In addition to very high GOF qualities, every element of M_n has also very high randomness as a result of the shuffling of the information transferred from the true random numbers r_n, as demonstrated in Section VI.

VI. Results and Discussions

In this section, the GOF and randomness of the proposed MUS-RNG technique [Fig. 3 (b)] are compared with pseudo-RNG data generated using the rand function of the Matlab software package, true-RNG data from the random.org service, and the solar-RF true RNG illustrated in Fig. 3 (a).

A. Goodness of fit tests for RNGs

The IECDFs for uniformly distributed random numbers are shown for total samples of 10 to 100,000 in Fig. 5, and the lack-of-fit errors for the observed samples defined by (9) are illustrated in Fig. 6. The numerical results for the Matlab pseudo-RNG, the true RNG illustrated in Fig. 3 (a), and the random.org true RNG were very similar, and the results for the first two RNGs have been omitted for brevity. The results for (8) and (9) are very similar for $N > 100$. It can be seen that the GOF error decreases ($err_n \to 0$) for increasing data lengths ($N \to \infty$) and that the IECDF ($G_{X,N}$) converges to the CDF (F_X), as expected. It is interesting that there is noticeable error even for N = 1,000,000. The numerical results for the test metric given in (9) show that the GOF error becomes unpredictable for standard RNGs when the total number of samples is decreased, as shown in Fig. 6 (b).

The average statistical error for the uniform distribution in [−0.5, 0.5] and the negative of the statistical quality defined by (10) and (11), respectively, are shown in Fig. 7. The results show that GOF error increases and the quality decreases as the number of samples decreases, as expected. This shows that even true RNGs can generate output with insufficient GOF and quality if the total number of samples is small. This error could be intolerable even for a set of 1,000,000 samples, depending on the specific application.

B. Goodness of fit tests for the MUS output

Method of uniform sampling (MUS) can be used to generate quantiles directly by sampling the CDF [Fig. 2 (a)]. The MUS-generated numbers have zero error as defined by (9) for a uniform distribution, because $F_X(x_n) = G_{X,N}(x_n)$. Next, the MUS-RNG approach is applied to the standard normal distribution. The observed mean and variance for the true-RNG output using 100,000 independent sets obtained from random.org are compared with the MUS-RNG output in Fig. 8. It can be seen that the MUS-RNG numbers are very accurate even for N = 10. It is interesting to observe that this is not the case for the true-RNG output, especially for smaller data lengths. This shows that MUS-RNG is also successful for the standard normal distribution.

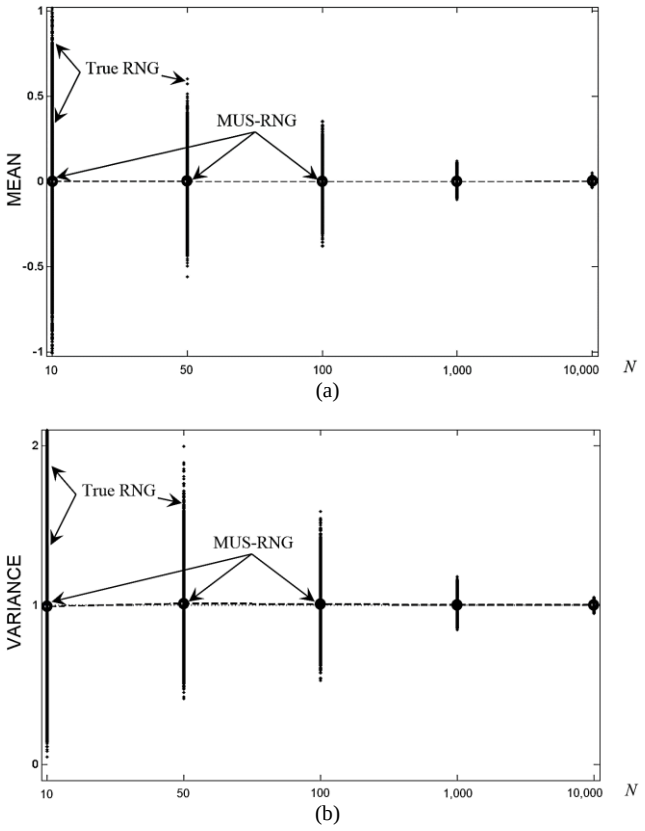

Figure 8. Observed (a) mean μ and (b) variance σ vs. data length N for the standard normal distribution. Dots, true RNG (S = 100,000); dashed lines and circles, the MUS-RNG samples, for N = 10, 50, 100, 10^3, and 10^5.

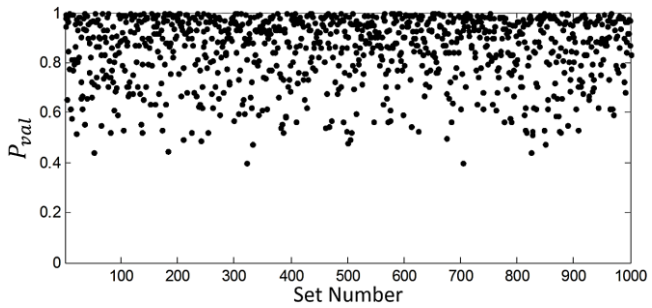

Figure 9. P_{val} for the block frequency test of randomness. There are 1000 sets total, and N = 10. The test results are better when N > 10 and for the monobit test.

C. Tests for randomness

MUS provides quasi-random samples that have very high GOF and quality. Unfortunately, they are not random. MUS-RNG, however, generates a set of numbers with the same values (and EPDF) as the MUS numbers and also the same information as a result of its ordering. MUS-RNG's output

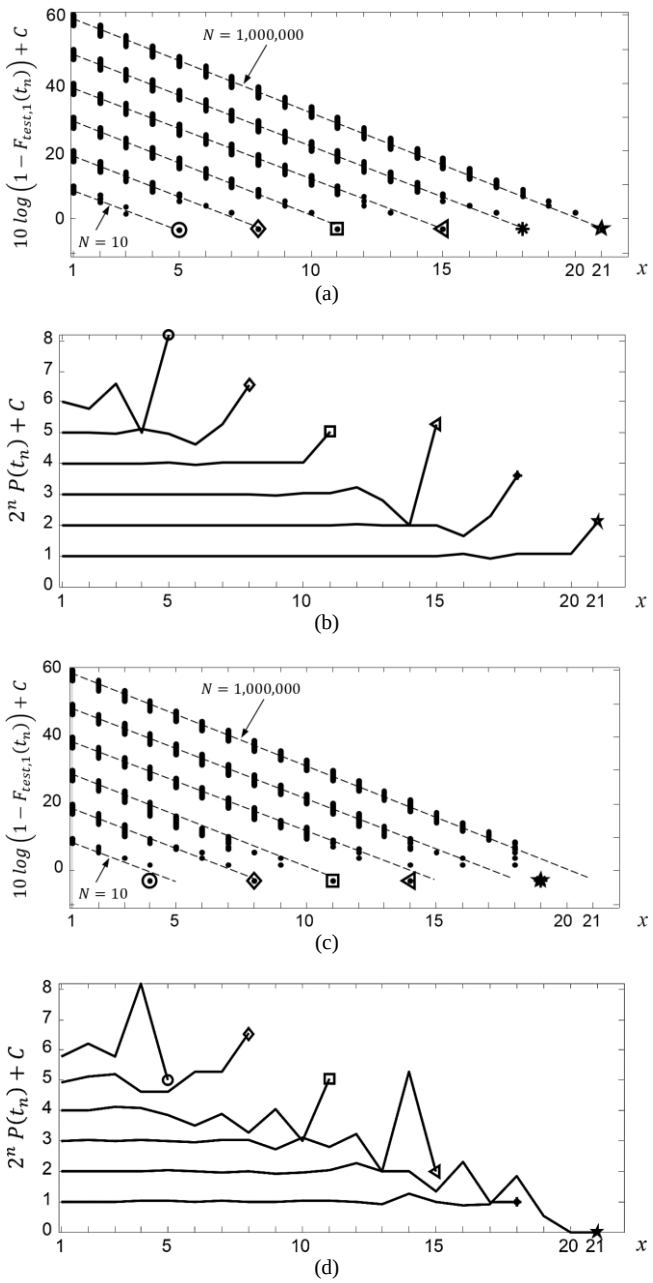

Figure 10. (a) IECDFs and (b) normalized probabilities for the observed samples t_n generated using MUS-RNG. Circles, N = 10; diamonds, N = 10^2; squares, N = 10^3; triangles, N = 10^4; asterisks, N = 10^5; five-pointed stars, N = 10^6. The groups are shifted vertically by a constant C to avoid overlap; in (a) C = 10, 20, …, 60, and in (b) C = 5, 4, …, and 0 for N = 10, 10^2, …, 10^6, respectively. (c, d) Results for the standard true RNG.

has been tested for randomness using NIST's monobit and block frequency tests [10]. It can be seen from Fig. 9 that all test results (P_{val}) are greater than 0.01. This shows that every set is accepted as random. Similar results for different data lengths and the monobit test have been omitted for brevity.

D. Benchmarking tests

It was shown that both uniform and standard normal distributions can be used for GOF testing for the comparison of different RNGs in the Appendix. The GOF test results are generally better for larger sets, and worse for smaller sets. Better test CDFs could be helpful for quantitative comparisons in order to assess the relative performance of two RNGs. This may be possible if a test CDF is selected such that values with low observation probability are present (e.g., the tail regions of continuous probability density functions or discrete random values with low probabilities). It would be almost impossible to observe samples with low probabilities when the total number of observed samples is low. Thus, different performance results can be observed. It would be practical to observe a single parameter as an indicator. Indicators for different RNGs can be used for quantitative comparisons.

A practical benchmark for the GOF can be constructed as follows. First, define a test PDF in terms of increasing positive integers with probabilities converging to zero,

$$f_{\text{test},1}(x) = \sum_{k=1}^{\infty} (1/2)^k \, \delta(x-k). \tag{14}$$

The test CDF can be obtained from the integral of its PDF,

$$F_{\text{test},1}(x) = \sum_{k=1}^{\infty} (1/2)^k \, u(x-k), \tag{15}$$

where $u(x - k) = 1$ for $x > k$ and 0 otherwise. Note that for a finite set of generated random numbers, it is impossible to observe all the positive integers that contribute to the upper limit of the summation. Thus, the larger values of k, with lower probabilities, $(1/2)^k$, are not expected to be observed. It can be assumed that a RNG providing a greater number of observed integers is better in terms of its GOF. The largest observed integer can be selected as the indicator.

A second, similar test CDF can be defined in terms of equal-probability integer observables {1, 2 ... N}

$$F_{\text{test},2}(x) = \frac{1}{N} \sum_{k=1}^{N} u(x-k), \tag{16}$$

where N is the number of observed samples in the set. Note that the probability of each random integer is equal to $1/N$. An integer k that is observed n times in the set has an observation probability of n/N. For this test CDF, the total number of observed integer values can be selected as the GOF indicator. The observed probabilities for each integer can also be analyzed.

Let us use the first benchmarking test to compare the MUS-RNG and the reference RNGs illustrated in Figs. 5 and 6. Equation (15) can be used to generate positive integers for each RNG. As the total number of samples increases, the larger k-values with lower observation probabilities are also expected to be in the set. The largest number in the set can be selected to be the GOF indicator for that RNG (one could also analyze the missing integers).

The ECDF and the observed probability of each integer are illustrated in Fig. 10 for increasing data lengths. The benchmarking indicator of (15) is analyzed. The numerical results for the MUS-RNG and the reference RNGs of Fig. 5 are compared using the GOF indicator in Fig. 10 (a, b) and Fig. 10 (c, d), respectively. The ECDF's slow convergence to 1 becomes clear when 10 log $[1 - F_{\text{test},1}(x)]$ is calculated, as shown in Fig. 10 (a, c). It can be seen that by increasing the total number of samples in a set, integers with lower observation probabilities are also generated. The largest integer in every set is selected to be the GOF indicator. The theoretical probability for the random integer k on the other hand, is $(1/2)k$ as defined by (14) and (15). This probability

can be normalized to 1 by multiplication by $2k$. These normalized probabilities for each observed integer are analyzed for different data lengths in Fig. 10 (b) and (d).

It can be seen that for larger N, the ECDFs show the presence of more observed integers, and similarly, the normalized probabilities for smaller integers are more accurate (closer to 1) for both RNGs as shown in Fig. 10. It is shown that the observed integers for the MUS-RNG output span the probability axis of the EPDF more uniformly, and thus, there are fewer missing integers with MUS-RNG. It is interesting that the largest MUS-RNG integer for each N (indicators) are observed to have normalized observation probabilities that are always greater than 1, as shown in Fig. 10 (b). For example, for N = 10, 10^2, …, 10^6, the MUS-RNG indicators are 5, 8, 11, 15, 18, and 21 with normalized observation probabilities of 3.2, 2.56, 2.048, 3.2768, 2.6214, and 2.0972, respectively. For example, for N = 10, the MUS-RNG outputs are {1, 1, 1, 1, 1, 2, 2, 3, 3, 5}. Integer k = 1 has an exact observation probability of 0.5. The actual probabilities are 5/10, 2/10, 2/10, 0, and 1/10. These become 1, 0.8, 1.6, 0, and 3.2 when normalized with $(1/2)^k$. Note that k = 4 is missing from the set.

The test CDF results for the reference RNG of Fig. 5 are shown in Fig. 10 (c) and Fig. 10 (d). It can be seen that the indicators and corresponding probabilities obtained with MUS are better than those of the standard RNGs. For example, for N = 10, 10^3, and 10^6 (C = 5, 2, and 0), the probabilities of the expected indicators are all 0. This means that the maximum values of the integers are less than the MUS-RNG indicators.

VII. Conclusion

Two novel statistical functions; the improved empirical cumulative distribution function (IECDF) and the improved empirical probability density function (IEPDF) are proposed for the statistical analysis of finite data where the IECDF forms the basis for the Section III. It is shown that these tools are very helpful to represent finite data accurately for uniform and standard normal distributions. It is also shown that apriori information and kernel optimization are not necessary to estimate the PDF accurately.

Novel test metrics for the GOF and quality of RNGs have been proposed, as defined in (10) and (11), respectively. It is shown that the GOF error measures increase for decreasing number of samples, as expected. Analysis was carried out for both uniform and standard normal distributions. It is interesting that this error is still noticeable even when the total number of samples is increased to 1,000,000. It was found that for a set of only 10 samples, the GOF error becomes large, as shown in Figs. 5 (a) and 6 (b).

The method of uniform sampling was proposed using new parameters in Section IV and the sampling parameters given in [24, 33] were improved to use the IECDF. Here the improved quantiles are selected closer to $F_X(x)$, which are the samples of $G_{X,N}$ for the uniform distribution. The method of uniform sampling random number generator (MUS-RNG) was also proposed, as illustrated in Fig. 3 (b). The GOF errors of the MUS-RNG outputs are tested for the standard normal distribution in Fig. 8. It can been seen that MUS is also very accurate for this distribution.

MUS is used to develop a novel random number generator (MUS-RNG) in Section V. MUS-RNG was shown to be successful for the uniform and standard normal distributions, and for discontinuous CDFs. The MUS-RNG output generated for the standard normal distribution is observed to have very accurate (almost exact) mean and variance values. The results show that very high GOF is possible even for a set of 10 samples, as shown in Fig. 8.

The MUS-RNG outputs were tested using NIST's monobit (not shown) and block frequency tests for randomness [10]. The P_{val} values were calculated for 1000 independent sets. The numerical results are shown in Fig. 9. All of the sets yield $P_{val} > 0.01$, and the test results indicate that they are random.

Two test CDFs are proposed for testing GOF, in (15) and (16). The proposed benchmarking indicators are the largest observed integer value and the total number of different observed values. MUS-RNG was tested to generate samples using these two test CDFs. The first test CDF was used to compare MUS-RNG and the reference RNGs in Fig. 10. The numerical results obtained from 1000 independent analyses show that MUS-RNG provides better indicators and more accurate observed probabilities. Both test CDFs and the indicators for the GOF tests are found to be successful.

It can be concluded that MUS-RNG is a valuable source of random numbers, especially for applications where the data length is small and the number of iterations are few, for example, as a result of limitations of CPU cycles or memory. MUS-RNG produces outputs with very high GOF and randomness. Moreover, MUS is simple and thus easy to implement; MUS-RNG can be used for any type of distribution as long as the quantiles can be obtained for a given list of probability values or, similarly, the inverse of the desired CDF is known. It could provide a better approach to generating random numbers for distributions that have mathematically simpler inverse CDF functions.

In the future, MUS-RNG can be demonstrated to use almost any type of entropy source, including computer fan noise, electronic noise, received RF signal, optical/infrared detections, and data received from the web. Providing very high GOF quality, its advantages could be analyzed in different applications, including statistical signal processing and optimization. MUS-RNG could improve initialization and iteration steps, providing faster convergence. Finally, MUS-RNG can be applied to a wide range of probability distributions.

Appendix

Pseudocodes for the GOF error and quality analyses, the generation of the MUS-RNG numbers, and the GOF indicators for the first test function, defined by (15), are given below.

The ECDF and the improved ECDF ($G_{X,N}$):

Given; N, Data → Output; ECDF and improved ECDF

```
sData = Sort(Data)                 /* from small to large
ECDF = [1:N]                       /* probability values
deltap = (1−1/(2N))/(N−1);
```

first = 1/(2N); last = 1−1/(2N)
GXn = [first:deltap:last]
plot(sData, ECDF) /* the ECDF as in Fig. 5
plot(sData, GXn) /* data on x and p_n on the y-axis

Total GOF error for a given set and the reference CDF:

Given; N, sData, ECDF, GXN → Output; Total GOF of the given set error defined by (8)

```
ErrN1 = ErrN2 = 0   /* 1 and 2 denote old and proposed
for in = 1:N
  ErrN1(in) =ErrN1(in) + abs(sData(in) − ECDF(in))^2
  ErrN2(in) =ErrN2(in) + abs(sData(in) − GXn(in))^2
end
ErrN1 =sqrt(ErrN1/N); ErrN2 = sqrt(ErrN2/N)
```

ErrN1 and ErrN2 of (9) are for the novel metrics proposed in this paper.

Quality analysis for a given RNG and the reference CDF:

Given; N, S, sData, ECDF, GXN → Output; GOF quality of the RNG defined by (10) and (11)

```
ErrS1 = ErrS2 = 0        /* 1&2 denote old and proposed
for in = 1:N
  for is = 1:S
    ErrS1(in) =ErrS1(in) +abs(sData(in) −ECDF(in))^2
    ErrS2(in) =ErrS2(in) +abs(sData(in) −GXn(in))^2
  end
ErrS1 = sqrt(ErrS1/S); ErrS2 = sqrt(ErrS2/S)
end
Qtmp = sum(ErrS1^2);                    /* (or ErrS2^2)
QNSlinear = Qtmp/N                      /* as shown in Fig. 7
QNSlog = −5*log10(Qtmp)          /* factor of 1/2 for sqrt
```

Practical calculation of MUS numbers for a desired CDF:

Given; N, CDFinv → Output; MUS numbers using (12)

```
Prob_samples = [first:deltap:last]
MUSdata = CDFinv(Prob_samples)   /* obtain quantiles
```

Random number generation utilizing MUS:

Given; RNGdata → Output; MUSRNGdata as in Fig. 3 (b)

```
Load(RNGdata, N); N = length(RNGdata)
Calculate(MUSdata, N)
Unsort_while_sorting(MUSdata, RNGdata)
                              /* obtain MUS-RNG output
```

Unsorting one data set while sorting the other:

Given; data {MUS, RNG}, N → Output; MUSRNGdata

```
for ii = 1:N
  for jj = N:−1:2
    if (RNGdata(jj) > rRNGdata(jj-1))
      dum1 = RNG(jj); dum2 = MUSdata(jj)
      RNGdata(jj) =RNGdata(ii-1); RNGdata(jj-1)=dum1
      MUSdata(jj) =MUSdata(jj-1); MUSdata(jj-1)=dum2
    end; end; end
MUSRNGdata = MUSdata;
Delete(RNGdata)
```

Benchmarking for GOF test defined by (11):

Let us first calculate the value of the indicator $K = (k)_{\max}$ for a given ECDF value A, where

$$A = F_{\text{test},1}(x_n) = \sum_{k=1}^{K} (1/2)^k u(u-k). \qquad \text{(A.1)}$$

Multiplying each side by 2 yields $2A = 1 + A - (1/2)^K$. This gives $A = 1 - (1/2)^K$, and $K = \log_2[1/(1 - A)]$. Hence, for a given A the corresponding $K = (k)_{\max}$ can be evaluated.

This result can be used to generate the random numbers of (11) for a given set of input random numbers uniform in (0, 1), and also to find the GOF benchmarking indicator, $K = (k)_{\max}$. This process can be iterated for every set of observed samples under the GOF test. Thus, the indicators for different sets generated by the same RNG, different N, and also indicators for different RNGs can be analyzed.

Given; raw random numbers of length N uniform in (0, 1) {RNGdata} → Output; the reference random data with ECDF given in (11) and the indicator $[K = (k)_{\max}]$ for the GOF benchmarking as shown in Figs. 10 (c) and (d))

```
A = Sort(RNGdata)          /* input numbers as the ECDF
for in = 1:N   /*then calculat. the corresponding k-values
  k(in) = floor(log2(1/(1−A(in))))
end
```

K_{max} = max(k) /* all k's and K_{max} can now be analyzed.

ACKNOWLEDGMENT

This work was supported in part by Baskent University research grant No: BAP-BA14/FM-13. Authors wishes to thank Dr. S. C. Inam for his support during the implementation of Baskent University's true random number generator.

REFERENCES

[1] J. Tabak, Probability and Statistics: The Science of Uncertainty, Facts on File, pp. 4–12, 2004

[2] V. F. Hendricks, S. A. Pedersen, K. F. Jørgensen, eds. Probability Theory: Philosophy, Recent History and Relations to Science, Kluwer, 2001

[3] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacci and S. Tubaro, "An overview on video forensics," APSIPA Trans. on Signal and Informat. Process., Vol. 1, August, 2012, pp. 1–18.

[4] M. Rakhshanfar, M. A. Amer, "Estimation of Gaussian, Poissonian-Gaussian, and processed visual noise and its level function," IEEE Trans. On Image Process., vol. 25, no. 9, Sept. 2016, pp. 4172–4185.

[5] A. Whalen, "Statistical theory of signal detection and parameter estimation," IEEE Communications Magazine, vol. 22, no. 2, pp. 37–44, 1984.

[6] P. Cenetobelli, G. Converso, M. Gallo, L. C. Santillo, "From process mining to process design: a simulation model to reduce conformance risk," Engineering Letters, vol. 23, no. 3, pp. 145–155, 2015.

[7] Y. Yu, R. Ru, K. Fang, "Bio-inspired mobility prediction clustering algorithm for ad hoc UAV networks", Engineering Letters, vol. 24, no.3, pp.83 – 92, 2016.

[8] G. Marsaglia, "Random number generators," Journal of Modern & Applied Statistical Methods, vol. 2, no. 1, pp. 2–13, 2003.

[9] Random Number Generators: An Evaluation and Comparison of Random.org and Some Commonly Used Generators, Project Report, Trinity College Dublin, The Distributed Systems Group, 2005.

[10] A. Rukhin et al., "A statistical test suite for random and pseudorandom number generators for cryptographic applications," Special Publ. 800-22, Rev. 1a, National Institute of Standards and Technology, 2010

[11] Y. A. Yudo, N. Shigei, H. Miyajima, "Multiple route construction with path-overlap avoidance for mobile relay on WSN", Engineering Letters, vol. 23, no. 4, pp. 299–306, 2015.

[12] A. Doucet, X. Wang, "Monte Carlo methods for signal processing: A review in the statistical signal processing context," IEEE Signal Processing Magazine, vol. 22, no. 6, pp. 152–170, 2005.

[13] Y. Fayad, C. Wang, Q. Cao, "A developed ESPRIT for moving target 2D-DOAE", Engineering Letters, vol. 24, no. 1, pp. 30–37, 2016.

[14] N. Aunsri, "A Bayesian filtering approach with time-frequency representation for corrupted dual tone multi frequency identification", Engineering Letters, vol. 24, no. 4, pp. 370–377, 2016.

[15] S. G. Tanyer, A. E. Yılmaz, F. Yaman, "Adaptive desirability function for multiobjective design of thinned array antennas," Journal of Electromagnetic Waves and Applications, vol. 26, no. 17–18, pp. 2410–2417, 2012.

[16] G. Manson, E. Papatheou, K. Worden, "Genetic optimisation of a neural network damage diagnostic," The Aeronautical Jour., vol. 112, no. 1131, pp. 267–274, 2010. May 2008.
[17] Y. Fu, M. Ding, C. Zhou, "Phase angle-encoded and quantum-behaved particle swarm optimization applied to three-dimensional route planning for UAV," IEEE Transactions on Systems, Man and Cybernetics A, vol. 42, no. 2, pp. 511–526, 2012.
[18] O. T. Altınoz, S. G. Tanyer, A. E. Yılmaz, "A comparative study of fuzzy-PSO and chaos-PSO," Elektrotehniski Vestnik, vol. 79, no. 1–2, pp. 68–72, 2012.
[19] S. M. Kwang, S. H. Weng, "Ant colony optimization for routing and load-balancing: survey and new directions," IEEE Transactions on Systems and Humans A, vol. 33, no. 5, pp. 560–572, 2003.
[20] M. P. Marques, F. R. Durand, T. Arao, "WDM/OCDM energy-efficient networks based on heuristic ant colony optimization," IEEE Systems Jour., vol. 10, no. 4, pp. 1482–1493, Dec. 2016.
[21] C. Sur, A. Shukla, "Discrete invasive weed optimization algorithm for graph based combinatorial road network management problem," in International Symposium on Computational and Business Intelligence, New Delhi, 2013, pp. 254–257.
[22] O. T. Altınoz, A. E. Yılmaz, G.-W. Weber, "Improvement of the gravitational search algorithm by means of low-discrepancy Sobol quasi random-number sequence based initialization," Advances in Electrical and Computer Engineering, vol. 14, no. 3, pp. 55–62, 2014.
[23] J. Wang, J. Song, "A hybrid algorithm based on gravitational search and particle swarm optimization algorithm to solve function optimization problems", Engineering Letters, vol. 25, no. 1, pp. 22–28, 2017.
[24] S. G. Tanyer, "Generation of quasi-random numbers with exact statistics" (in Turkish), in 22nd Signal Processing and Communications Applications Conference, Trabzon, 2014, pp. 281–284.
[25] The Concise Encyclopedia of Statistics, pp. 283–287, Springer, 2008.
[26] A. W. Van Der Vaart, "Asymptotic statistics,"Cambridge University Press, Cambridge, 2000, Section 19.3, pp. 265–271.
[27] T. W. Anderson, D. A. Darling, "Asymptotic theory of certain goodness of fit criteria based on stochastic processes," Annals of Mathematical Statistics, vol. 23, no. 2, pp. 193–212, 1952.
[28] J. Donsker, "Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems", Annals of Mathematical Statistics, vol. 23, pp. 277–281, 1952.
[29] S. Chen, X. Hong, C. J. Harris, "Propability density estimation with tunable kernels using orthogonal regression", IEEE Trans. On Sys., Man. And Cybernetics, Part B., vol. 40, no. 4, pp. 1101–1114, 2010.
[30] M. Gao, X. Hongi S. Chen, C. J. Harris, "Probability density estimation based over-sampling for imbalanced two-class problems", WCCI 2012 IEEE World Congress on Comput. Intelligence, Brisbane, Australia, 10-15 June, 2012.
[31] D. Li, W. Yan, W. Li, T. Chen, "Estimation of the probability density function of renewable power production using a hybrid method of minimum frequency and maximum entropy", 2016 Int. Conf. on Probabilistic Methods Applied to Power Systems, (PMAPS), 2016.
[32] B-J. Yoon, P. P. Vaidyanathan, "A multirate DSP model for estimation of discrete probability density functions", IEEE Trans. On Signal Process., vol. 53, no. 1, pp. 252–264, 2005.
[33] S. G. Tanyer, "True random number generation of very high goodness-of-fit and randomness qualities," 2014 International Conference on Mathematics and Computers in Science and Industry, Varma, 2014, pp. 213–215.
[34] K. Shibasaki, C. E. Alissandrakis, S. Pohjolainen, "Radio emission of the quiet Sun and active regions," Solar Physics, vol. 273, no. 2, pp. 309–337, 2011.
[35] D. A. Guidice, E. W. Cliver, W. R. Barron, S. Kahler, "The Air Force RSTN system," Bulletin of the American Astronomical Society, vol. 13, no. 2, p. 553, 1981.

S. Gokhun Tanyer received the B. Sc. and the M. Sc. in electrical engineering from Middle East Technical University and Bilkent University, Ankara, Turkey in 1988 and 1990 respectively, and the Ph. D. in electrical engineering from Washington State University, Pullman, WA, United States in 1994. His Ph. D. thesis work received the best research award in 'Sixth Annual Graduate and Professional Student Research Exposition' in 1994. From 1995 to 2000, he was a research fellow, Assistant Professor and Associate Professor in Bilkent University, Ankara University and Başkent University, respectively. During 2000 – 2011, he was with the Scientific and Technological Research Council of Turkey (TUBITAK). He was responsible head for foundation of the Advanced Technology Research Institution (ILTAREN), and Assistant Director for the National Research Institute for Electronics and Cryptology (TUBITAK UEKAE). He was the project manager for more than 20 contracted projects where he worked on various aspects of electromagnetics and radar signal processing. He is working as a tenured professor at Baskent University, and is co-founder of Defense Engineering R & D Technologies (SMART) Inc. both in Ankara, Turkey. His research interests include; antenna design and optimization, radar signal processing, statistical signal processing and music theory.