 doi:10.20944/preprints202011.0436.v1

# Deterministic sampling from uniform distributions with Sierpiński space-filling curves

Hime Aguiar e O. Jr.*

*National Cinema Agency - Rio de Janeiro, Brazil*

**Abstract**

In this paper the problem of sampling from uniform probability distributions is approached by means of space-filling curves (SFCs), a topological concept that has found a number of important applications in recent years. Departing from the theoretical fact that they are surjective but not necessarilly injective, the investigation focused upon the structure of the distributions obtained when their domain is swept in a uniform and discrete manner, and the corresponding values used to build histograms, that are approximations of their true PDFs. This work concentrates on the real interval [0,1], and the Sierpiński space-filling curve was chosen because of its favorable computational properties. In order to validate the results, the Kullback-Leibler distance is used when comparing the obtained distributions in several levels of granularity with other already established sampling methods.

In truth, the generation of uniform random numbers is a deterministic simulation of randomness using numerical operations. In this fashion, sequences resulting from this sort of process are not truly random.

*Keywords:* Space-filling curves; Ergodic Theory. uniform random number generation;

## 1. Introduction

In the last decades the proliferation of low cost and faster digital computers has expanded the development and applicability of techniques using stochastic simulation. The uncertainty in such systems is usually simulated by means of random numbers. Although some time ago the purpose of the simulations was aimed at qualitative understanding, the explosive availability of processing power in the last decades made it possible to reach more quantitative results. In this fashion, large scale simulations made it feasible the estimation of objects with large variances and very precise estimation of variables under study. Accordingly, random numbers with high quality and long periods are needed - portability and efficiency are essential as well. Frequently, simulations of probabilistic models need random variables with specific probability distributions and, in practice, many algorithms for generation of these non-uniform random variables are based on certain transformations of uniform random numbers. Therefore, uniform random numbers

*Corresponding author
*Email address:* hime@engineer.com (Hime Aguiar e O. Jr.)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1

can be considered the basis for probabilistic simulation. In truth, uniform random number generation is a deterministic process which imitates randomness by means of arithmetic operations. Therefore, sequences obtained in this way are radically different from sequences of truly random numbers. The notion of absolute randomness is an idealized concept, and there were initiatives in the sense of "statistically purify" it by means of models of physical processes. While there is some logical basis for that, this kind of approach is not very well accepted in the field of simulation. This is so because numbers generated by using a particular physical process are not truly random numbers, which exist only conceptually. As another example we have the argument that there is no ideal dice without any bias, capable of simulating Bernoulli trials. In addition, there is the infeasibility of reproducing the same sequence of random numbers several times and in different contexts. This type of procedure is fundamental for checking simulation results obtained by others, and code debugging.

However, generating true randomness on digital computers is not a feasible task, taking into account their deterministic nature. In this fashion, sequences obtained from deterministic processes are known as pseudorandom ones. There are at least two different ways of dealing with this situation: the pragmatic one, that is, to use pseudorandom numbers with some optimism or just to give up usinig randomness on computer simulations and converging to the concept of derandomization. The first alternative will demand the execution of statistical tests on every set produced by the chosen generators in different applications. So, they should be "stressed" in several directions, so that randomness be at least approximated, even though never reached. In [13] it is stated that all deterministic methods for producing randomness fail in some application, and only experience and imagination of developers and users can lead to a better understanding of this type of event. Despite all those true statements, considerable progress has been made since the 1940s, when computer-generated random numbers were successfully used in the implementation of the so-called Monte Carlo methods.

For the sake of efficiency, typical algorithms have been based on recurrence relations, which can be viewed as discrete dynamical systems "equipped" with finite memory. These two characteristics of discreteness and finiteness show up whenever considering any type of method for digital computer generation of random objects - actually, present day computers are unable to exactly represent irrational numbers, leaving available only the rational ones, up to a limited numerical precision. Finiteness, by definition, implies in the production of periodic sequences. For example, sometimes it is possible to find people with the belief that certain nonlinear discrete dynamical systems are candidates for deterministic schemes of random number generation. But, on digital computers, it is simply not possible to simulate chaotic behavior at all, considering the finite precision of numbers processed by present day machines, among other things. So, the main problem in random number generation on computers is quick generation, using small amounts of memory, and producing sequences with huge periods.

An alternative approach to handle the difficult situation described above, by means of pseudorandom numbers, is to look for more rational and precise mathematical solutions. According to von Neumann [15], "It is true that a problem that we suspect of being solvable by random methods may be solvable by some rigorously defined sequence." In the same direction, some researchers proposed alternative methods using deterministic versions of Monte Carlo methods, called

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1

quasi-Monte Carlo methods. Also, and in the same line of reasoning, this article presents a deterministic and practical method for asymptotically sampling from a 1-dimensional uniform probability distribution with arbitrary precision levels. The proposal is based on the space-filling curve synthesized by Sierpiński [19], coupled to some simple transformations described below.

Many applications of space-filling curves and their approximations have been recently proposed in the literature [10, 12, 20]. Probably their most important and amazing property is the ability to completely fill compact regions of higher dimensional spaces, in the sense of (typically) being surjective and continuous at the same time.

## 2. Sierpiński space-filling curves

It is possible to define space-filling curves as surjective and continuous functions from [0,1] to compact subsets of finite dimensional vector spaces, usually identified to $\mathbb{R}^n$. These objects were well-studied in the past, with many theoretical results establishing necessary conditions for their existence [19]. In addition, several important mathematicians defined concrete examples and established several interesting properties [19] of SFCs. Decades later, researchers have found significant applications of space-filling curves to several fields, including global optimization of numerical functions and design of certain devices [10]. However, implementation using digital computers causes certain difficulties,mainly due to their finite word length. Space-filling curves are often defined by means of infinite expansions and use, for instance, the property that elements in [0,1] can be represented in the form $0.t_1t_2t_3t_4...$ in a given basis B, where each $t_i$ is an integer between 0 and B-1. Accordingly it is necessary to find precise approximations of SFCs to pursue this kind of approach in real world computers. The most adequate candidate seems to be the Sierpiński curve, taking into account the availability of its precise and simple defining formulas, as shown below [19]:

$$x(t) = f(t), \;\; y(t) = f(t - 1/4), \;\; t \in [0,1] \tag{1}$$

where $f$ is a bounded, even and continuous real function whose expression is given by

$$f(t) = \frac{\Theta(t)}{2} - \frac{\Theta(t)\Theta(\tau_1(t))}{4} + \frac{\Theta(t)\Theta(\tau_1(t))\Theta(\tau_2(t))}{8} - ... \tag{2}$$

The 1-periodic functions $\Theta(t)$ and $\tau_k(t)$ are defined in [19], so that the resulting curve $(x(t), y(t))$ is a 2-dimensional SFC showing good behavior in numerical computations. When constructing higher-dimensional SFCs, certain results were fundamental. Before stating them, however, it is necessary to present some definitions.

**Definition 1.** *A function $\varphi : [0,1] \to \mathbb{R}$ is uniformly distributed with respect to the Lebesgue measure if, for any (Lebesgue) measurable set $A \subset \mathbb{R}$, we have*

$$\Lambda_1(\varphi^{-1}(A)) = \Lambda_1(A) \tag{3}$$

*where $\Lambda_1$ is the Lebesgue measure in the real line.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1

**Definition 2.** *n measurable functions $\varphi_1, \varphi_2, ..., \varphi_n : [0,1] \to \mathbb{R}$ are stochastically independent with respect to Lebesgue measure if, for any n measurable sets $A_1, A_2, ..., A_n \subset \mathbb{R}$,*

$$\Lambda_1(\bigcap_{j=1}^{n} \varphi_j^{-1}(A_j)) = \prod_{j=1}^{n} \Lambda_1(\varphi_j^{-1}(A_j)) \tag{4}$$

**Theorem 1.** *(H. Steinhaus) [19] If $\varphi_1, \varphi_2, ..., \varphi_n : [0,1] \to \mathbb{R}$ are continuous, non-constant and stochastically independent with respect to the Lebesgue measure, then*

$$f = (\varphi_1, \varphi_2, ..., \varphi_n) : [0,1] \to \varphi_1([0,1]) \times \varphi_2([0,1]) \times ... \times \varphi_n([0,1]) \tag{5}$$

*is a SFC.*

**Theorem 2.** *[19] If $f = (\varphi, \psi) : [0,1] \to [0,1] \times [0,1]$ is (Lebesgue) measure-preserving and onto, then its coordinate functions $\varphi, \psi$ are uniformly distributed and stochastically independent.*

Considering that the Sierpiński SFC is measure-preserving ([19], page 111), it is clear that its coordinate functions are uniformly distributed and stochastically independent, and can be used to obtain higher-dimensional SFCs with coordinates

$$\begin{cases} x_1(t) = \varphi(t) \\ x_2(t) = \varphi \circ \psi(t) \\ \quad \cdots\cdots \\ x_n(t) = \varphi \circ \psi \circ \psi \circ \cdots \circ \psi(t) \end{cases} \tag{6}$$
$$\text{where } t \in [0,1]$$

that are not constant, being continuous and stochastically independent. It occurs that, for higher-dimensional sets, approximations of original Sierpiński based SFCs were not able to completely fill up the associated compact domains. This is due to distortions caused by numerical approximations, despite the theoretical curve being a space-filling one.
In order to find a more robust curve, it is possible to compose the Sierpiński function with an invertible measure-preserving transformation that is a natural extension of a particular generalized Lüroth series transformation [7], mapping [0,1)x[0,1) onto itself. Finally, a new SFC mapping [0,1] onto [-1,1]x[-1,1] is generated by means of a linear homeomorphism from [-1,1]x[-1,1] to [0,1]x[0,1] and vice-versa. By denoting such a transformation by $T$, its definition is

$$T(x,y) \stackrel{\Delta}{=} ((k-1)(kx-1), \frac{1}{k}(1 + \frac{y}{k-1})), x \in I_k, y \in [0,1) \tag{7}$$
$$I_k \stackrel{\Delta}{=} [\frac{1}{k}, \frac{1}{k-1}), k \in \{2,3,...\}$$

It is worth highlighting that despite Steinhaus' theorem is stated only for continuous functions, it is also valid for surjective (over [0,1]), piecewise continuous coordinate functions. In this fashion, the following result is true.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1

**Theorem 3.** *(extended Steinhaus) If* $\varphi_1, \varphi_2, \ldots, \varphi_n : [0,1] \to [0,1]$ *are piecewise continuous, surjective, non-constant and stochastically independent with respect to the Lebesgue measure, then*

$$f \stackrel{\Delta}{=} (\varphi_1, \varphi_2, ..., \varphi_n) : [0,1] \to \varphi_1([0,1]) \times \varphi_2([0,1]) \times \cdots \times \varphi_n([0,1]) \quad (8)$$

is a space-filling curve.

## 3. Proposed method

Considering that space-filling curves are, by definition, surjective, it is obvious that every single point in their images is "visited" at least once. The fundamental question is: what kind of distribution corresponds to their full excursion in the respective domain?
In principle, each type of SFC has one specific solution for this issue, and the corresponding theoretical treatment is not easy, for sure. Turning back to the problem at hand, and focusing on the unidimensional case, let us consider a Sierpiński SFC having $[0,1]$ as its image, that is, it passes through every point in this interval when its domain is "swept" from 0 to 1. In this setting, if the resulting continuous frequency distribution (defined as the limit of frequency histograms when the width of classification bins tends to zero and the number of samples tends to infinite) coincides with the uniform PDF in $[0,1]$, it is possible to say that a deterministic sampling process with uniform PDF over $[0,1]$ has been obtained. It is worth mentioning that even irrational numbers (not perfectly representable in finite word length) are generated, differently from existing paradigms. The proposed sampling scheme is very simple and may help to verify theoretical results concerning Sierpiński SFC, giving approximations of the real form of the cited distribution (call it D). To that end, it is proposed an algorithm that, given an integer number N, produces N samples from the distribution D, corresponding to a special SFC S in [0,1] so that, as N tends to infinite, the corresponding Kullback-Leibler distance (or divergence) [6] between $D_N$ and $U(0,1)$ tends to zero, where $D_N$ is the frequency distribution corresponding to the N samples just computed. This notable fact indicates the coincidence of the concept established in Definition 1 with the well-known uniform distribution idea, leading to the experimental verification that this particular SFC is able to produce deterministic samples from $U(0,1)$, mainly due to its measure-theoretical properties ([19], page 111).

The algorithm itself is very compact and consists of a few steps - the inputs are **N** (number of samples) and **B** (number of bins to accumulate the samples)

- **Compute S(i/N), for i = 1 to N**
- **Linearly scale the resulting N – uple so as to keep its components in [0, 1]**
- **Construct the probability mass function, or histogram, of the N – tuple using B bins**
- **Return a vector containing the B obtained values corresponding to the B bins**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1

The frequency histogram relative to the resulting **B**-uple is an approximation for the PDF $U(0,1)$, as demonstrated by the experiments to be presented below.

Another important observation is that whenever the desired number of samples is changed, the values of the generated numbers is drastically altered even when the difference is very small, but the histograms maintain the coherence in terms of convergence to $U(0,1)$.

## 4. Experiment description and numerical results

The proposed method is compared to 4 well-established algorithms and their relative quality is measured by means of the Kullback-Leibler divergence with respect to the unitary uniform distributiom in $[0,1]$. In this fashion, the lower the divergence between two given PDFs, the better the approximation. The methods used for comparison are:

- The Mersenne twister implementation of Randomc library, described in [9]
- The RANDLC implementation in [3]
- The RNGLIB implementation in [2, 11]
- The Ziggurat implementation in [4, 14]

Table 1: Numerical results - histograms with 1000 bins

| Samples | KL(Mersenne) | KL(RANDLC) | KL(RNGLIB) | KL(ZIGGURAT) | KL(SFC) |
|---|---|---|---|---|---|
| $10^4$ | 0.052703657 | 0.052475 | 0.05197726 | 0.049796 | 0.046596252975 |
| $10^5$ | 0.0049975597 | 0.004830718 | 0.00474658 | 0.005328 | 0.003174821773 |
| $10^6$ | 0.000516303655 | 0.000533191 | 0.0004649 | 0.000508 | 0.000141780203 |
| $10^7$ | 4.7645683E-05 | 4.79491E-05 | 5.086E-05 | 4.694E-05 | 1.5440363E-05 |
| $10^8$ | 4.930764E-06 | 5.08564E-06 | 5.1647E-06 | 4.86699E-06 | 2.61962E-07 |

In Table 1 and Figure 1, numerical results of KL divergences are displayed, evidencing the convergence of all PDFs to the uniform distribution. Also, according to simulations, it is possible to observe that the SFC-based distribuition is faster in the convergence process. In order to show more detailed information, histogram sets displayed in Figures 2 through 6 are shown below, resulting from sampling according to the chosen methods. The sequence of figures clearly shows the convergence process taking place, and closes the pictorial presentation. Notice that histograms corresponding to SFCs are ”smoother” than the others. In these simulations, only 50 bins were used, for the sake of better visualization.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1

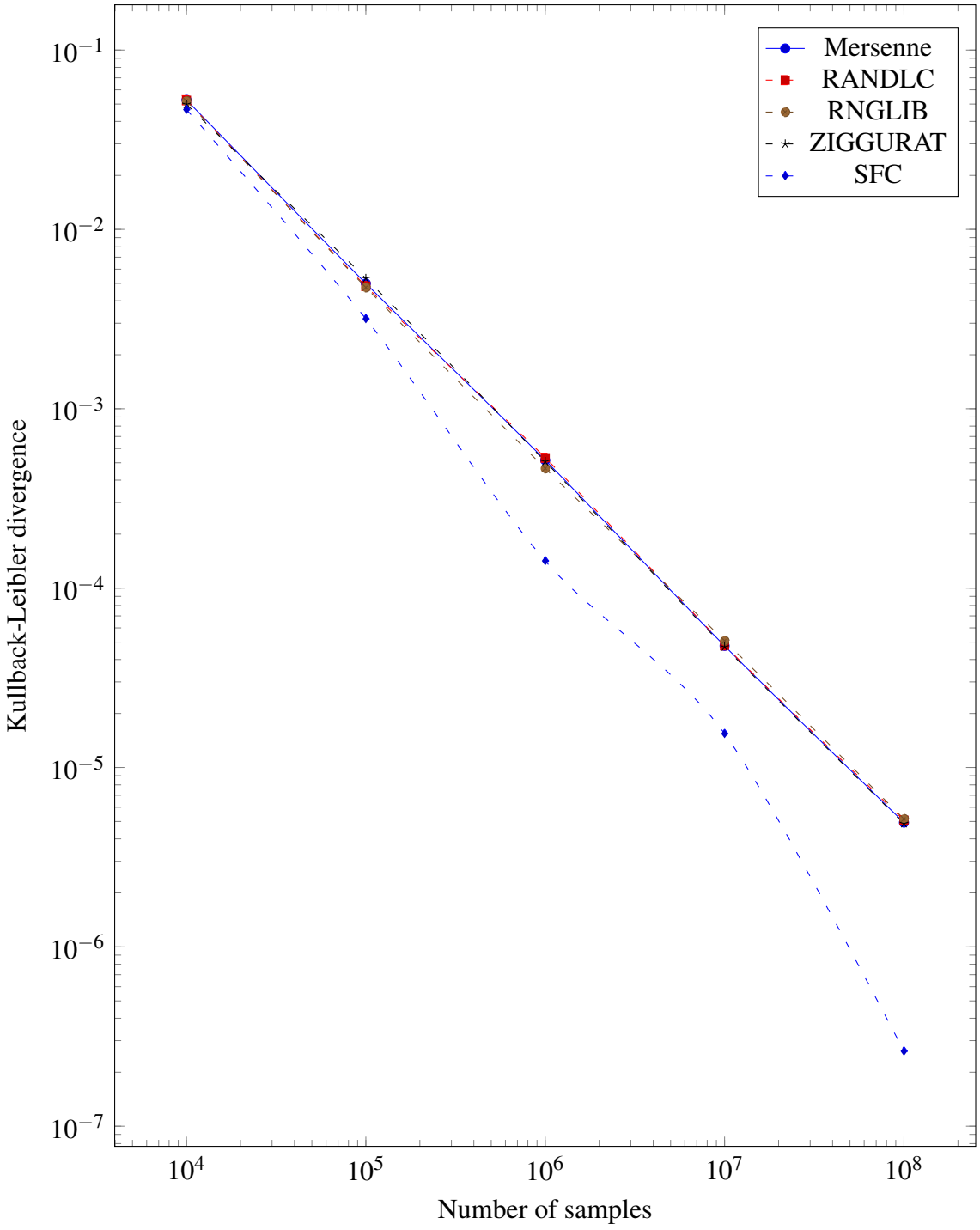
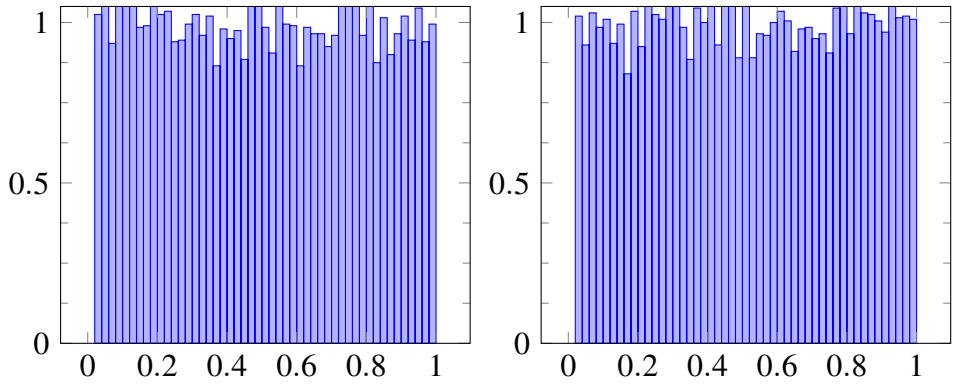
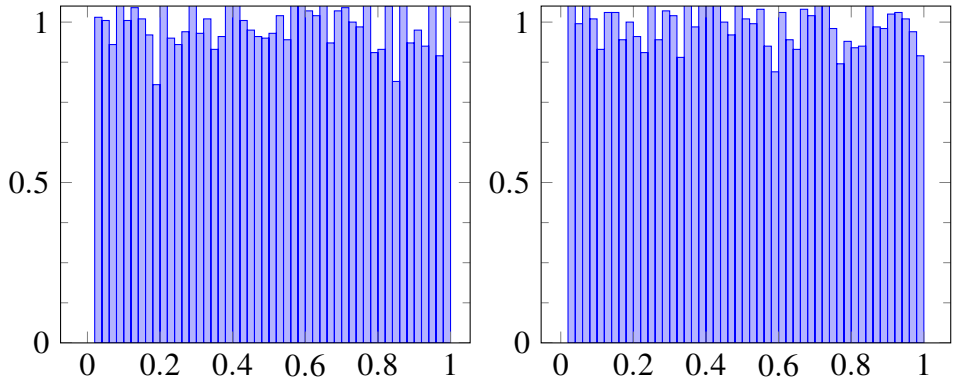



Figure 1: KL divergences to uniform distribution for each method.

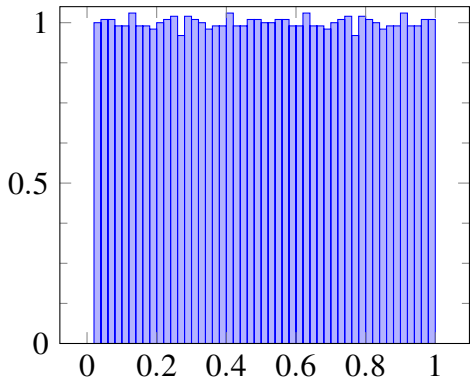
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1




(a) Mersenne and RANDLC



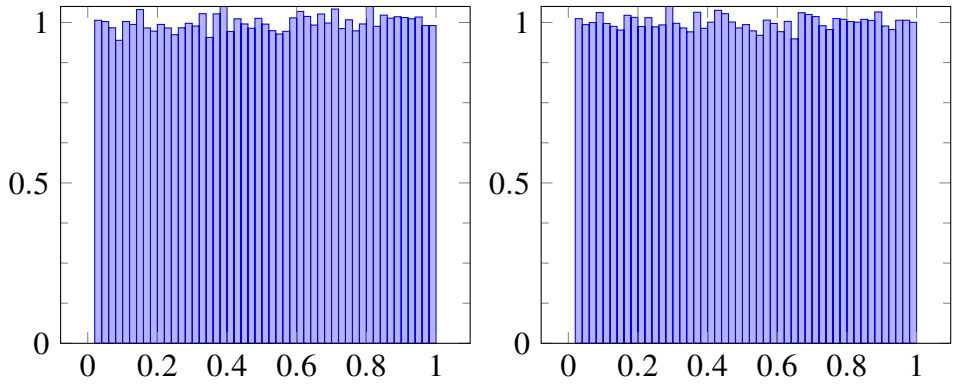

(b) RNGLIB and ZIGGURAT



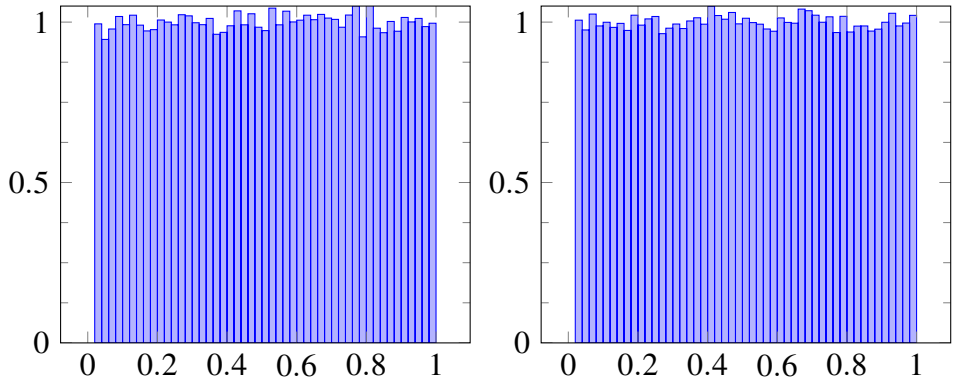

(c) Sierpiński SFC

Figure 2: Histograms for 10000 samples - 50 bins

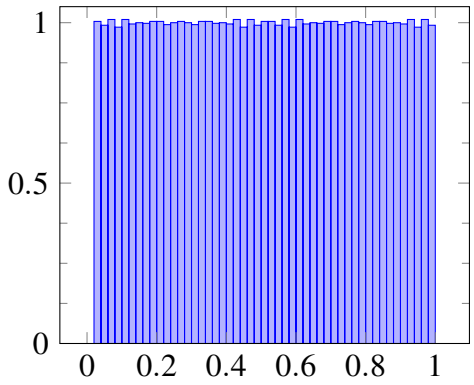
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1




(a) Mersenne and RANDLC



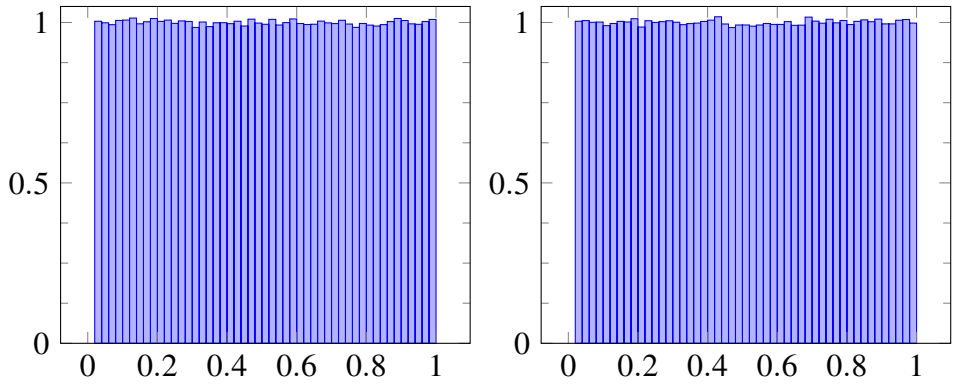

(b) RNGLIB and ZIGGURAT



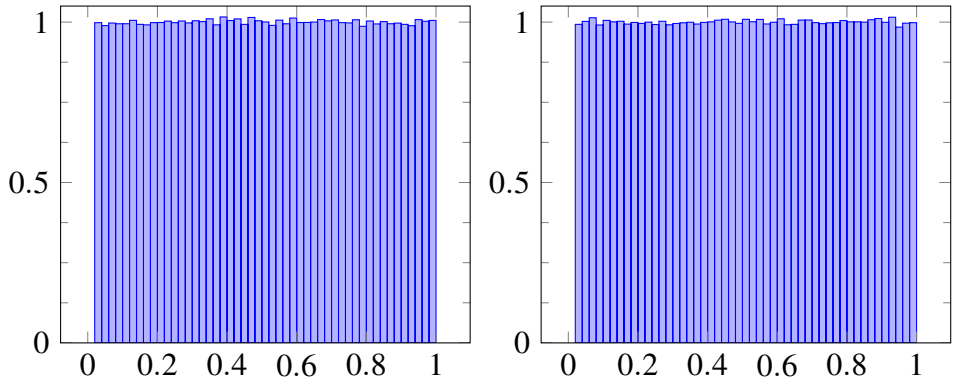

(c) Sierpiński SFC

Figure 3: Histograms for 100000 samples - 50 bins

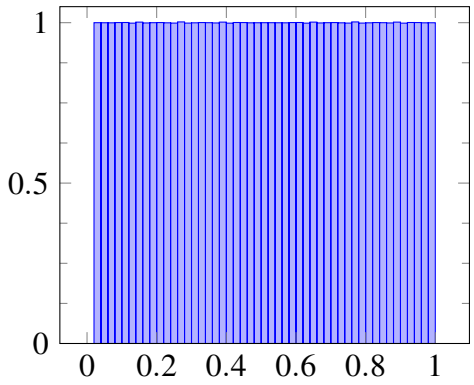
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1




(a) Mersenne and RANDLC



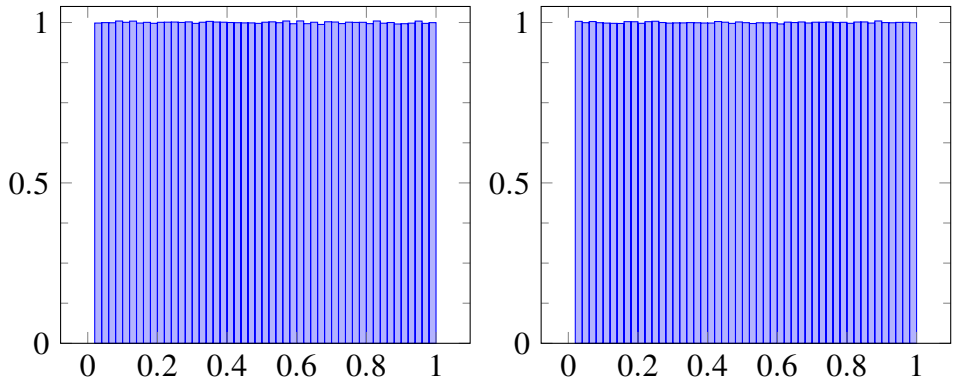

(b) RNGLIB and ZIGGURAT



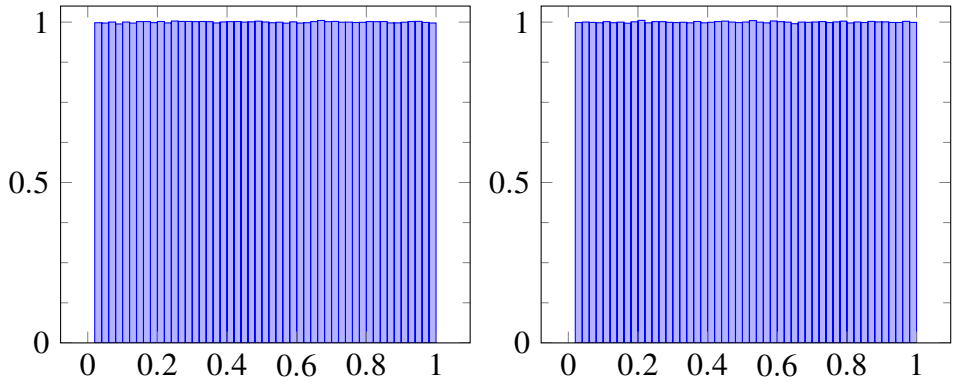

(c) Sierpiński SFC

Figure 4: Histograms for 1000000 samples - 50 bins

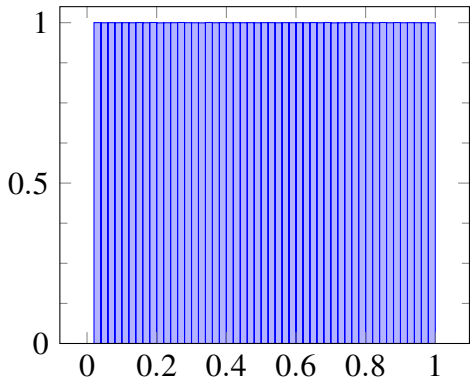
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1




(a) Mersenne and RANDLC



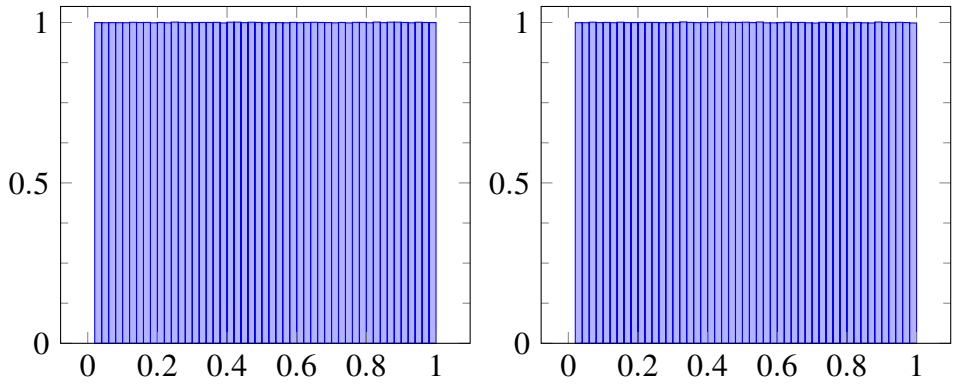

(b) RNGLIB and ZIGGURAT



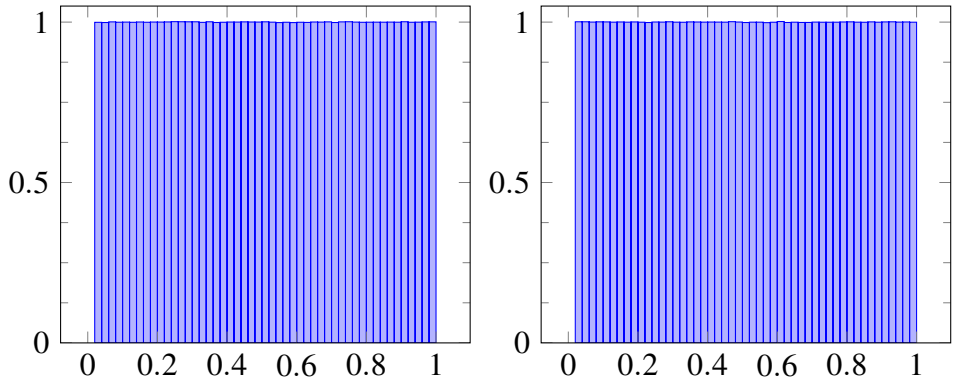

(c) Sierpiński SFC

Figure 5: Histograms for 10000000 samples - 50 bins

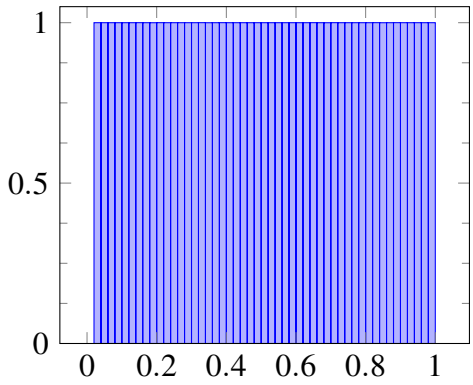
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1




(a) Mersenne and RANDLC




(b) RNGLIB and ZIGGURAT




(c) Sierpiński SFC

Figure 6: Histograms for 100000000 samples - 50 bins

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 November 2020 doi:10.20944/preprints202011.0436.v1

## 5. Conclusions

Through several simulations, this paper has shown that it is possible to (approximately and) deterministically sample from $U(0,1)$ by means of Sierpiński space-filling curves. The proposed method was compared to 4 other well-established algorithms and the "proximity" to the uniform distribution was measured with the Kullback-Leibler divergence, which is widely accepted as a good index of discrepancy between PDFs. By analyzing the numerical results, it is possibe to infer that the presented algorithm converges faster than the other methods. This type of result may be very useful in applied fields, like cryptography, for example.

## References

[1] A. Boyarsky, P. Góra, Laws of Chaos, Invariant Measures and Dynamical Systems in One Dimension. Birkhäuser, Boston, 1997.

[2] J.Burkardt, https://people.sc.fsu.edu/~jburkardt/cpp_src/rnglib/rnglib.html.

[3] J.Burkardt, https://people.sc.fsu.edu/~jburkardt/cpp_src/randlc/randlc.html.

[4] J.Burkardt, https://people.sc.fsu.edu/~jburkardt/cpp_src/ziggurat/ziggurat.html.

[5] G.H. Choe, Computational Ergodic Theory. Springer-Verlag, Berlin, 2005.

[6] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, 1991.

[7] K. Dajani, C. Kraaikamp, Ergodic Theory of Numbers, The Mathematical Association of America, Washington DC, 2002.

[8] G. Edgar, Measure, Topology, and Fractal Geometry, Springer-Verlag, 2008.

[9] A. Fog, www.agner.org/random/

[10] B. Goertzel, Global Optimization with Space-Filling Curves. Applied Mathematics Letters 12 (1999) 133–135.

[11] P. LEcuyer, Serge Cote, Implementing a Random Number Package with Splitting Facilities. ACM Transactions on Mathematical Software, Vol. 17 1 (1991) 98–111.

[12] D. Lera, Y. D. Sergeyev, Lipschitz and Hölder global optimization using space-filling curves. Applied Numerical Mathematics 60 (2010) 115–129.

[13] G. Marsaglia. Remarks on choosing and implementing random number generators. Communications of the ACM 36 (1993) 105–108.

[14] G. Marsaglia, W. W. Tsang, The Ziggurat Method for Generating Random Variables. Journal of Statistical Software. Vol. 5 8 (2000).

[15] J. von Neumann. Various techniques used in connection with random digits. In Collected Works, Vol. 5, 768–770. Pergamon Press, Oxford, 1963.

 doi:10.20944/preprints202011.0436.v1

[16] H. A. Oliveira Jr., L. Ingber, A. Petraglia, M.R. Petraglia, M.A.S. Machado, Stochastic Global Optimization and Its Applications with Fuzzy Adaptive Simulated Annealing, Springer-Verlag, Berlin-Heidelberg, 2012.

[17] H. A. Oliveira Jr., Evolutionary Global Optimization, Manifolds and Applications, Springer-Verlag, Cham Heidelberg New York Dordrecht London, 2016.

[18] H.A. Oliveira Jr., A. Petraglia, Global optimization using space-filling curves and measure-preserving transformations, in: A. Gaspar-Cunha et al. (Eds.), Soft Computing in Industrial Applications, AISC 96, Springer-Verlag, Berlin Heidelberg, 2011, pp. 121-130.

[19] H. Sagan, Space-Filling Curves, Springer-Verlag, New York, 1994.

[20] Y. D. Sergeyev, R.G. Strongin, D. Lera, Introduction to Global Optimization Exploiting Space-Filling Curves, Springer-Verlag, Heidelberg New York Dordrecht London, 2013.