Deterministic sampling from univariate normal distributions with Sierpiński space-filling curves

Hime Aguiar e O. Jr.*

National Cinema Agency - Rio de Janeiro, Brazil

Abstract

This work addresses the problem of sampling from Gaussian probability distributions by means of uniform samples obtained deterministically and directly from space-filling curves (SFCs), a purely topological concept. To that end, the well-known inverse cumulative distribution function method is used, with the help of the probit function,which is the inverse of the cumulative distribution function of the standard normal distribution. Mainly due to the central limit theorem, the Gaussian distribution plays a fundamental role in probability theory and related areas, and that is why it has been chosen to be studied in the present paper. Numerical distributions (histograms) obtained with the proposed method, and in several levels of granularity, are compared to the theoretical normal PDF, along with other already established sampling methods, all using the cited probit function. Final results are validated with the Kullback-Leibler and two other divergence measures, and it will be possible to draw conclusions about the adequacy of the presented paradigm. As is amply known, the generation of uniform random numbers is a deterministic simulation of randomness using numerical operations. That said, sequences resulting from this kind of procedure are not truly random. Even so, and to be coherent with the literature, the expression ”random number” will be used along the text to mean ”pseudo-random number”.

Keywords: Space-filling curves; Ergodic Theory; random number generation; Gaussian distribution.

1. Introduction

Recently the field of stochastic simulation has experienced a high level of visibility, mainly due to decreasing cost and high speed of present day digital computers and the resulting application of so many available computational techniques in this important field. Such tecniques typically depend on good (pseudo)random number generators in order to work properly, that is, generators based on solid paradigms and theoretical results. Considering that simulations of probabilistic models use random variables obeying several types of probability distributions, and many methods for sampling from them (often non-uniform ones) are based on special manipulations of uniform random numbers, it seems sensible to investigate more precise paradigms for their generation. Taking into account that

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

the Gaussian distribution appears very frequently in diverse scenarios, nothing more natural than to concentrate efforts for improving existing algorithms aimed at sampling from this important model.

By following this line of reasoning, it is possible to draw the conclusion that uniform random number generation can be considered fundamental for probabilistic modelling and simulation. Actually, uniform random number generation is a fully deterministic process which mimics randomness by means of numerical operations. Therefore, numerical streams obtained in this way do not contain true random numbers.

In summary, generating true randomness on digital computers (using existing techniques) is not a feasible task. In this fashion, sequences obtained from deterministic processes are known as pseudorandom ones. In [14] it is cited that all deterministic methods for producing randomness will fail in some application, and only experience and imagination of researchers may lead to an improvement of this state of affairs. An approach to deal with this issue could be to look for other types of precise mathematical solutions. According to von Neumann [16], "It is true that a problem that we suspect of being solvable by random methods may be solvable by some rigorously defined sequence."

Based on the previous considerations, this article presents a deterministic and simple method for asymptotically sampling from a Gaussian probability distribution with arbitrary precision. The general idea is to use the classical inverse function algorithm along with the probit function and the uniform generator proposed in [20], which is based on the space-filling curve synthesized by the great mathematician W. Sierpiński [21]. It is worth to highlight that this composition may be changed, provided the uniform generator be the indicated one.

Several algorithms for generating pseudorandom samples from normal distributions have been created over the decades [11] and any uniform random number generating method can be used with the inverse cumulative distribution function (or a reasonable approximation) to generate the desired stream of observations. On the other hand, although there are some simple, efficient and fast methods of simulation for the task at hand, many of them present some undesirable features, like cyclic behavior or imperfect asymptotic sampling, for example. Among the well-known and mature approaches, it is possible to cite Box-Muller, the corresponding Marsaglia-Bray's improvement, Acceptance-Rejection and Ahrens-Dieter methods [11].

The exposition below aims to offer evidences that the empirical distributions obtained with the proposed method converge to the theoretical PDF characteristic of the Gaussian distribution. This will be made by means of geometrical concepts, with pictures of histograms and two true metrics in the sense of general topology, namely, the Euclidean and Manhattan distances on $\mathbb{R}^{1000}$, considering histograms as vectors in $\mathbb{R}^{1000}$ with sum of components equal to 1. The dimension 1000 represents the number of bins used in the construction of histograms.

2. Sierpiński space-filling curves

Although [20] gives more detailed information about space-filling curves and their use in the conception of uniform distribution samplers, it seems opportune to offer some basic facts about their main characteristics, for the sake of better understanding of the overall "landscape".

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

Space-filling curves may be defined as surjective and continuous functions from the interval [0,1] to compact subsets of finite dimensional vector spaces, usually identified with $\mathbb{R}^n$. They were well-studied in the past and many theoretical results establishing necessary conditions for their existence [21] are known. Recently, some researchers found several applications of space-filling curves, including global optimization of numerical functions [10]. In another dimension, their computation using digital computers creates some difficulties, due to the finite word length of existing digital computing machines. In many cases they are defined by means of infinite expansions and pass through every point of their images, including those with irrational coordinates, not exactly representable with a finite number of digits. Due to its availability, with precise defining formulas, the most adequate candidate for certain important tasks is the Sierpiński space-filling curve [21, 19].

In the case of the proposed approach, sampling from specific probability distributions, the ”generation” process is totally atypical and its dynamics is not recursive or iterative. In addition it is fully parallelizable, in the sense it is feasible to assign one processor to each calculation step, resulting in one pseudorandom number per CPU. In this manner, if 1000 samples are needed and there are 1000 integrated CPUs, it is logically possible to obtain the whole set within just 1 time step. As said above, a SFC by definition is surjective, meaning that each point of the codomain is”visited” at least once, even those not having exact representation in present day digital computers, like irrational numbers or n-tuples containing them. Accordingly, the lower bound for the number of ”visits” is known, but the upper bound can vary and will have direct influence in the respective PDFs. In truth, SFCs are fractals and may behave in a somewhat ”strange” way, at least when compared to the usual, smooth curves. The SFC used in the uniform sampling is nothing more than the first component of the original Sierpinski SFC mapped from [-1,1] into [0,1] by simple linear and affine operations. But any other SFC in the same conditions might be a candidate to arrive at the same final result.

3. Proposed method

The presented method consists of a deterministic procedure for asymptotically sampling from a Gaussian probability distribution with arbitrary precision, that can be chosen by users, depending on their needs. The overall idea is absolutely similar to most implementations in the literature and uses the classical inverse function algorithm, along with the probit function, and as its uniform generator the one proposed in [20], based on the space-filling curve due to W. Sierpiński [21]. Naturally, it is possible to synthesize other equivalent algorithms by combining different components, provided the uniform generator be the indicated one. Accordingly, more efficient paradigms can be constructed by replacing the inverse function algorithm, or the application of the probit function, or both of them, with alternatives that may seem more interesting when applied to different scenarios. The only immutable condition is the use of the indicated SFC uniform generator. Here the focus is on the standard normal distribution, with mean 0 and standard deviation 1, taking into account that all other (nonstandard) Gaussian distributions may be easily derived from it.

Given this context, once the resulting continuous frequency distribution (defined

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

as the limit of frequency histograms when the width of classification bins tends to zero and the number of samples tends to infinite) coincides, in the limit, with a standard Gaussian PDF, it seems reasonable to say that a deterministic sampling process for that PDF has been found. The quality of approximation is directly related to the total number of desired samples, that is, finer overall discretizations lead to more proximity to the theoretical PDF. Another important feature is the extreme parallelism degree made possible by the proposed algorithm, considering its independence from previous samples - in the limit, it is possible to compute, for instance, millions of samples at once, provided the availability of enough processors to get the job done.

In general terms, the method looks like any other, except that its input parameter is not an initial seed, or parameter used between iterations, but the granularity of the discretization (or quantization) of samples - as said before, more subdivisions mean greater precision, in terms of similarity with the Gaussian PDF. This calibration may be very important in most tasks, and should be done in advance in order to obtain optimal results - in any case, a high level of discretization is always a good choice, for obvious reasons, not to mention that in simulations needing a substantial number of samples (over 1 billion, for instance), the approximation quality is automatically high. Another subtlety that deserves mentioning, is that the input parameter needs to be an upper bound for the global number of samples needed in a given task.. By doing so, it is possible to guarantee a good probabilistic approximation amd the overall quality of the produced samples. Let us now describe how a typical implementation could look like, using the Gaussian sampler:

Assumptions :

1. An upper bound for the global number of samples to be used in a given task has been determined (**N**)
2. **N** is sufficiently large to get the desired probabilistic precision for the task at hand

Algorithm structure :

(Input = **N** , Output = approximate sample of standard Gaussian distribution)

1. *Compute a sample from the SFC uniform generator, using* **N**
2. *Compute the value of Probit function at the result of previous step*
3. *Return the previous result*

The frequency histograms resulting from sampling with the suggested algorithm are approximations for the PDF of the standard normal distribution, as will be demonstrated by the experiments to be presented below.

It is worth noting that whenever the parameter **N** is changed, the values of generated numbers are radically changed even when the difference is very small. Histograms maintain the coherence in terms of convergence to the target probability distribution function.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

4. Numerical simulations

The proposed method is compared to 3 other possible implementations, using well-established algorithms for uniform number generation, and their relative quality is measured by means of the Kullback-Leibler divergence, Euclidean and Manhattan (or taxicab) metrics [1] for the sake of estimating the discrepancy with respect to the discretized standard normal distribution. In this fashion, the lower the divergence or distance between two given PDFs, the better the approximation. The methods used for sampling from the uniform distribution were:

1. A PRNG implemented through the routine **r8_uniform_01** included in the PROB library [4]
2. A RNGLIB implementation [3, 12]
3. A Ziggurat implementation [5, 15]

It is very important to highlight that, when using the Euclidean and Manhattan true metrics, histograms under comparison have been regarded as vectors in the metric space $\mathbb{R}^{1000}$, endowed with each one of the cited metrics - experiments were done by accumulation in 1000 bins. Also, in order to get histograms corresponding to the theoretical normal PDF, one specific location in each bin was arbitrated for the sake of uniformizing the comparisons - in the present case, these points correspond to the end of each bin. In this fashion, once the true PDF is sampled, the obtained value is considered constant along the corresponding compartment. The $[-5,5]$ interval was used as the domain for the final histograms, considering that outside this region the occurrence of samples is very rare.

In Tables 1-3 and Figures 1-3, numerical results corresponding to all divergences and metrics are displayed, evidencing the gradual approximation of PDFs relatively to the standard normal distribution - 1000 bins were used for displaying histograms as well. It is possible to observe that the SFC-based approximation is faster in the convergence process. In addition, for the sake of showing more detailed information, histogram sets displayed in Figures 4 through 13 are shown below, resulting from sampling according to the chosen methods. The sequence of figures clearly shows the approximation process taking place by also including magnified highly nonlinear subregions, and completes the pictorial presentation. Notice that histograms corresponding to SFCs become "smoother" faster than the others.

Table 1: Numerical results - histograms with 1000 bins (Kullback-Leibler)

Samples	Ziggurat	SFC	RNGLIB	PROB
10^5	2.9188952E-3	1.856744E-3	2.95566E-3	2.6505E-3
10^6	3.86377E-4	1.8702E-4	3.7718E-4	3.92914E-4
10^7	5.32132E-5	3.73552E-5	5.77279E-5	5.41478E-5
10^8	1.76505E-5	2.17663E-5	1.79864E-5	1.70388E-5
$4x10^9$	1.2424E-5	1.308E-5	1.2707E-5	1.25E-5

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

Table 2: Numerical results - histograms with 1000 bins (Euclidean metric)

Samples	Ziggurat	SFC	RNGLIB	PROB
10^5	3.175E-3	2.38899E-3	3.23947E-3	2.99377E-3
10^6	1.0342E-3	4.131E-4	1.01419E-3	1.03278E-3
10^7	3.5838E-4	2.2566E-4	3.723321E-4	3.634463E-4
10^8	2.109041E-4	1.886261E-4	2.097718E-4	2.173273E-4
$4x10^9$	1.87266E-4	1.877E-4	1.8907E-4	1.877E-4

Table 3: Numerical results - histograms with 1000 bins (Manhattan metric)

Samples	Ziggurat	SFC	RNGLIB	PROB
10^5	5.5874E-2	4.7139E-2	5.73232E-2	5.36011E-2
10^6	1.8208E-2	8.7893E-3	1.8514E-2	1.8801E-2
10^7	6.4539E-3	4.83845E-3	6.75553E-3	6.736433E-3
10^8	4.2127843E-3	4.067017E-3	4.196966E-3	4.342839E-3
$4x10^9$	3.98162E-3	3.99E-3	4.0E-3	3.99E-3

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

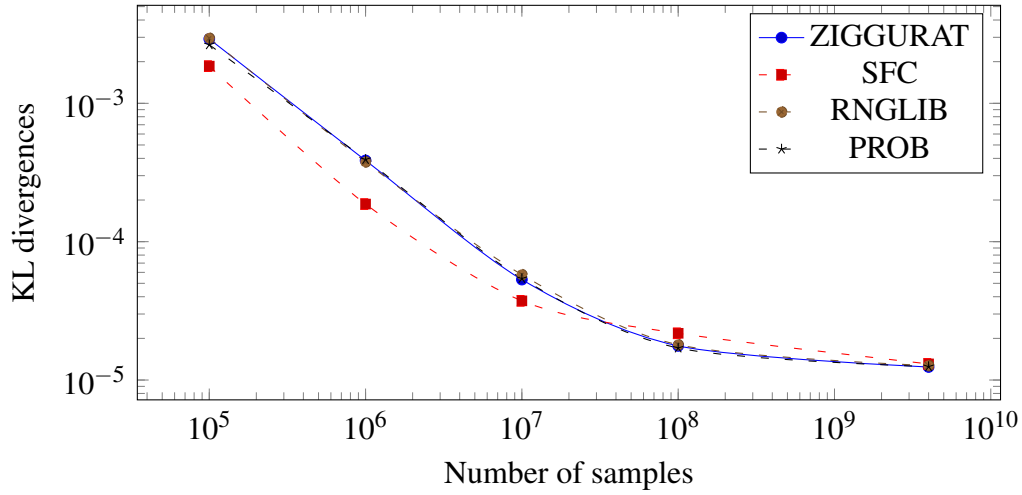

Figure 1: KL divergences to standard Gaussian PDF.

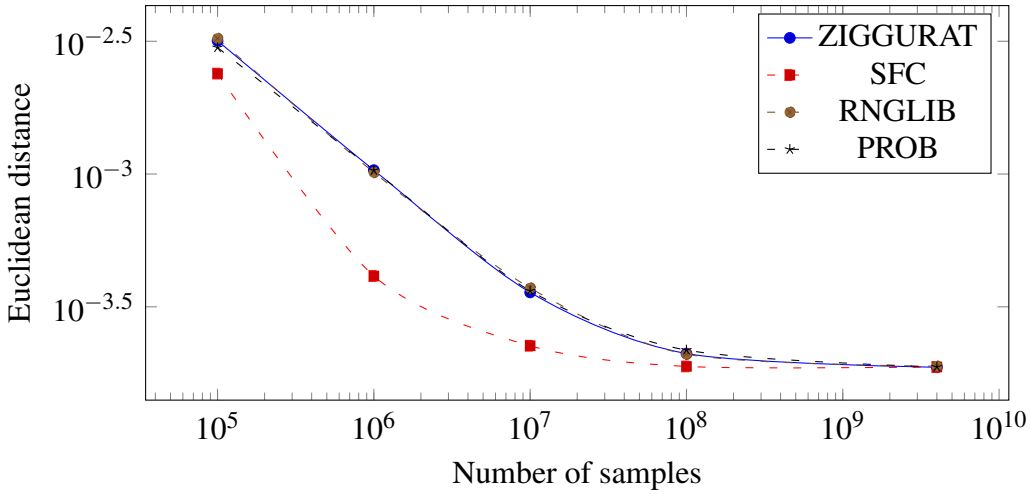

Figure 2: Euclidean distances to standard Gaussian PDF.

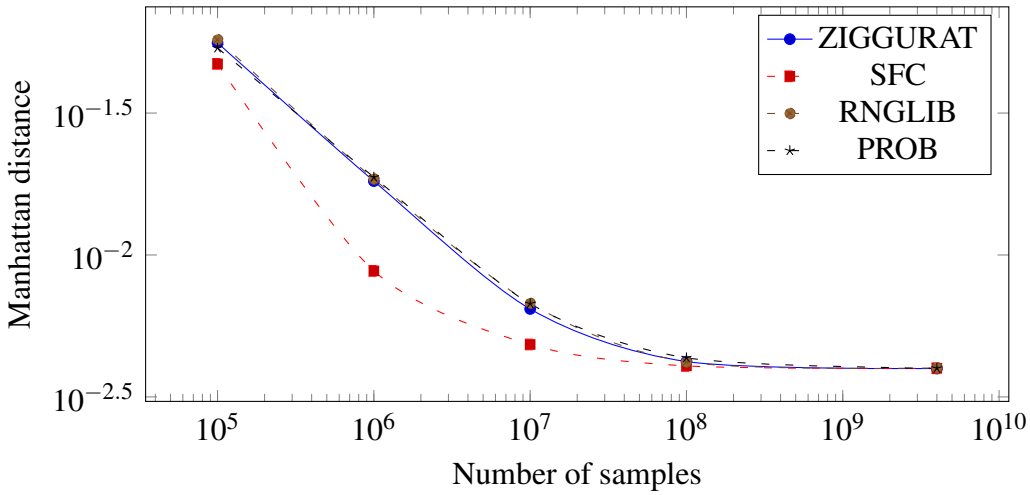

Figure 3: Manhattan distances to standard Gaussian PDF.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

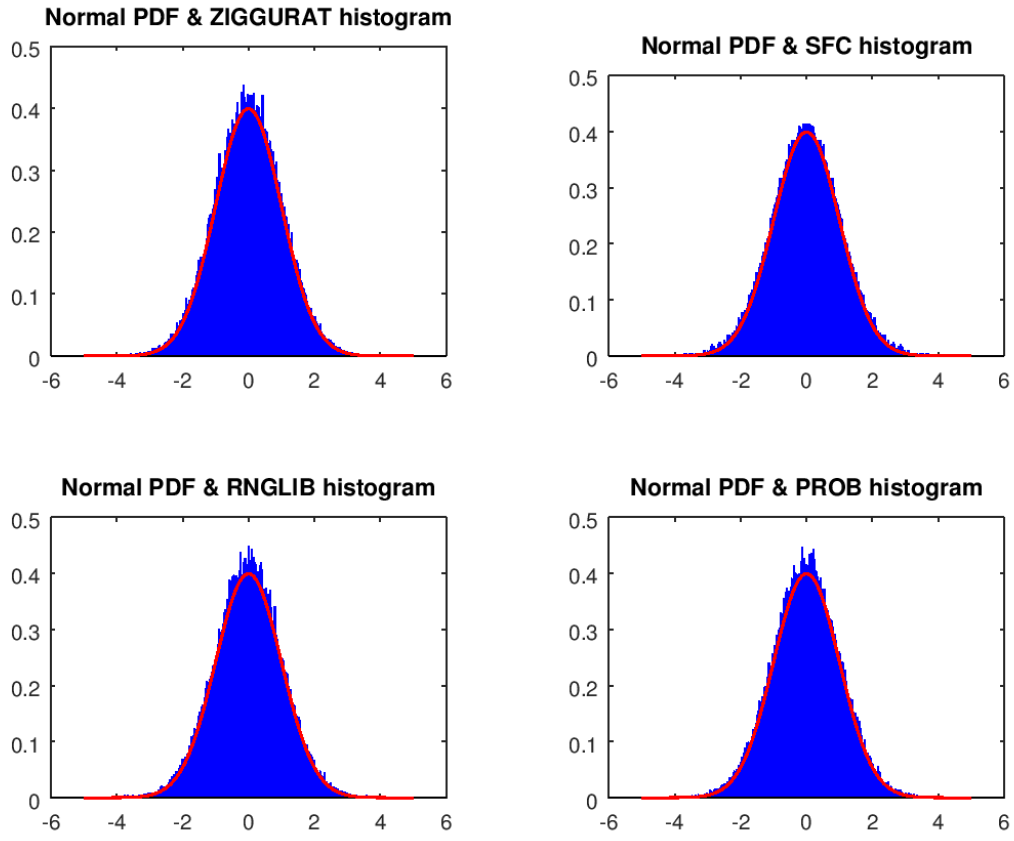

Figure 4: General view - 100000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

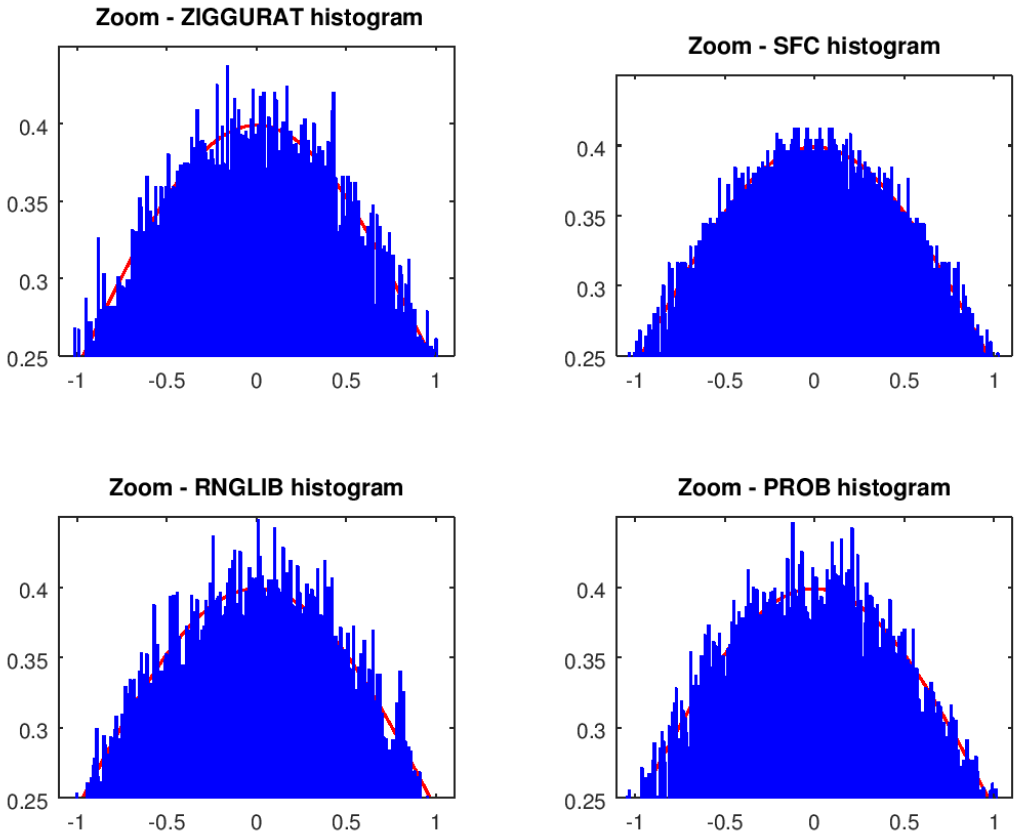

Figure 5: Zoom - 100000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

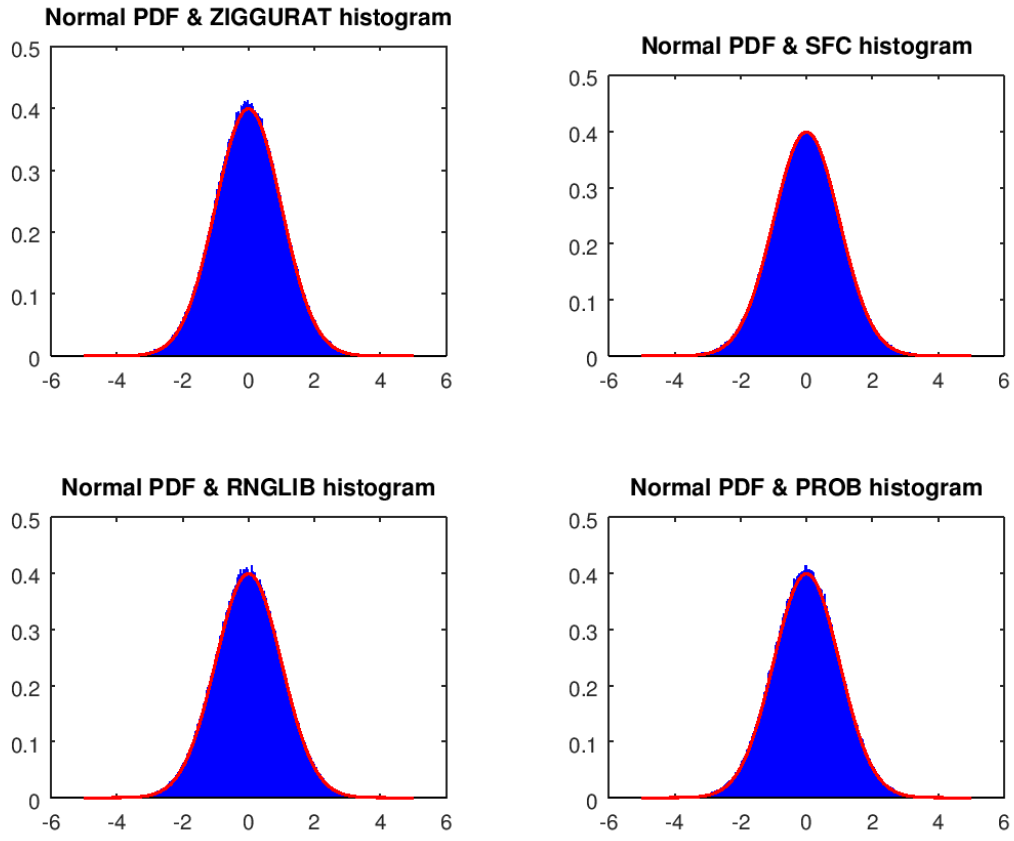

Figure 6: General view - 1000000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

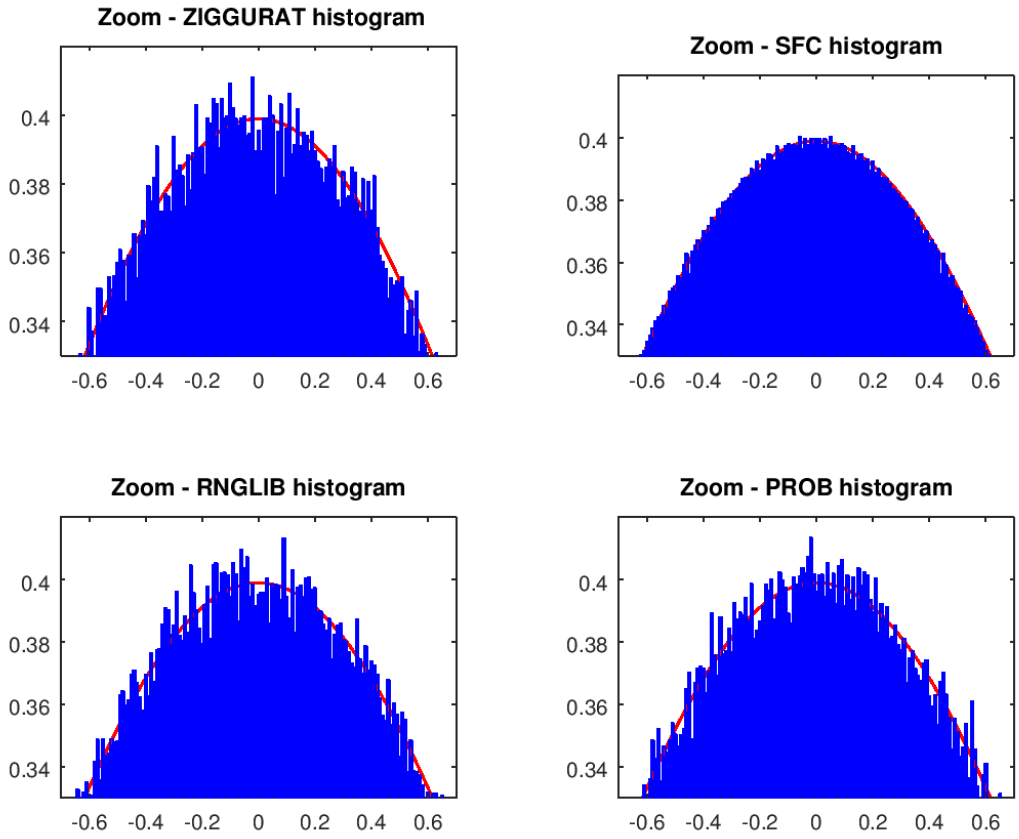

Figure 7: Zoom - 1000000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

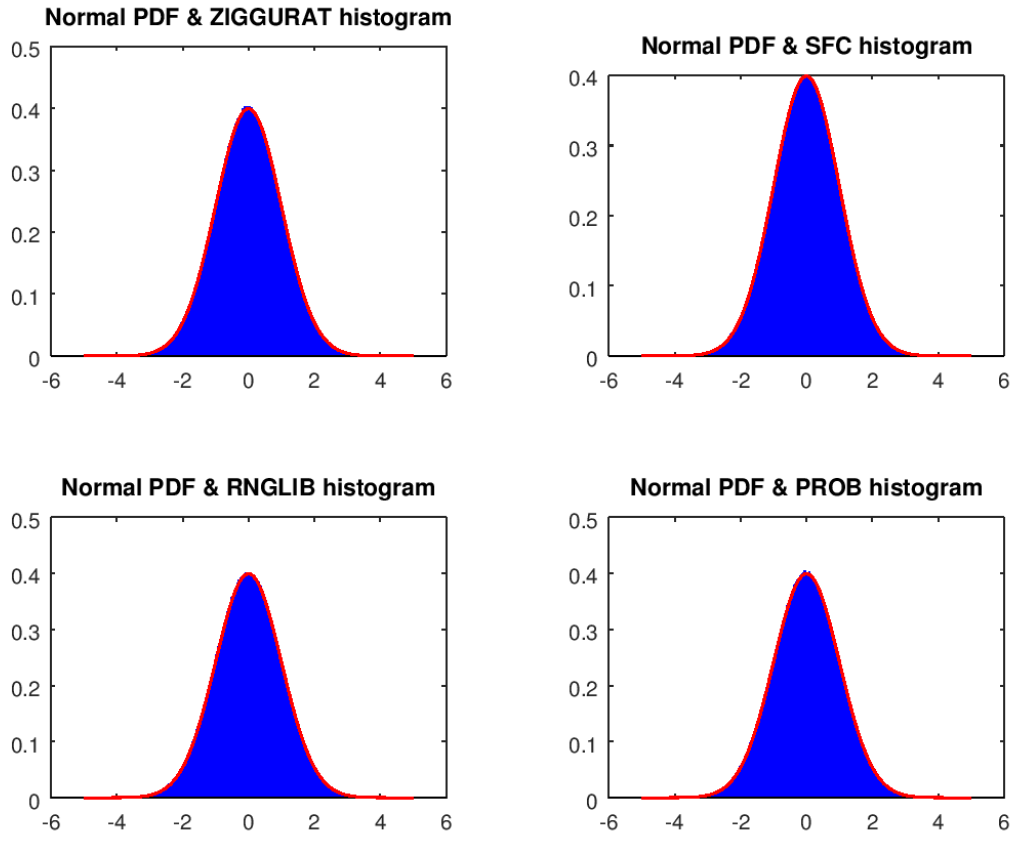

Figure 8: General view - 10000000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

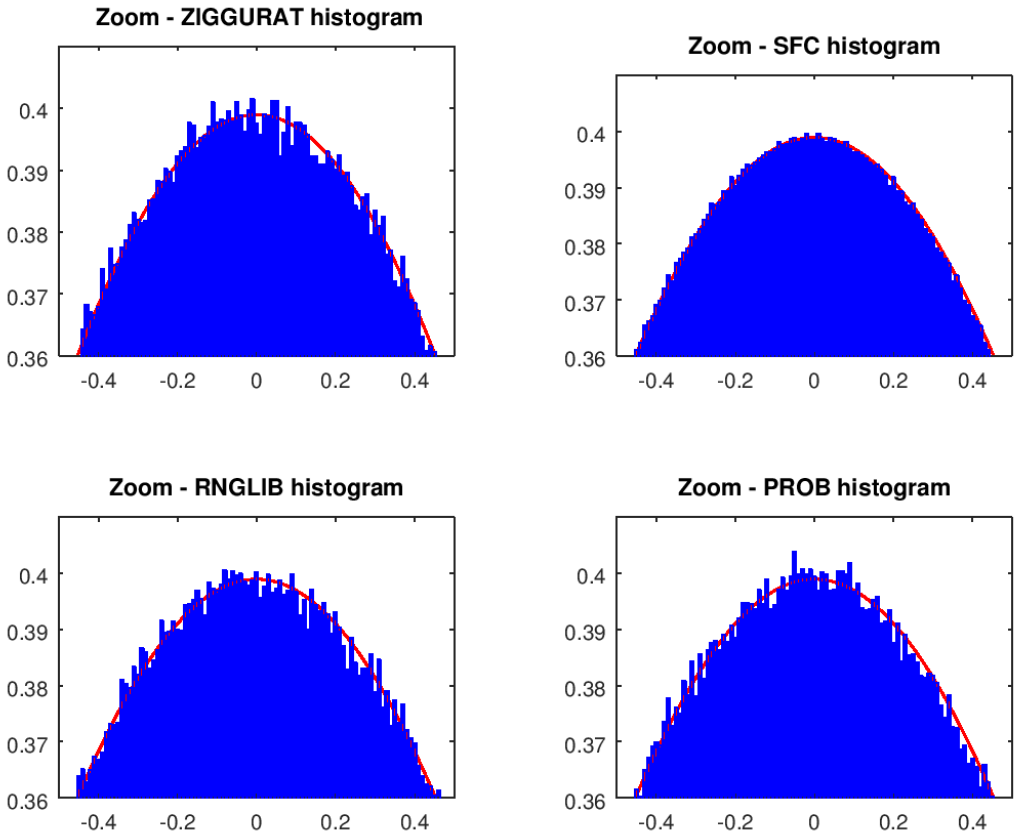

Figure 9: Zoom - 10000000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

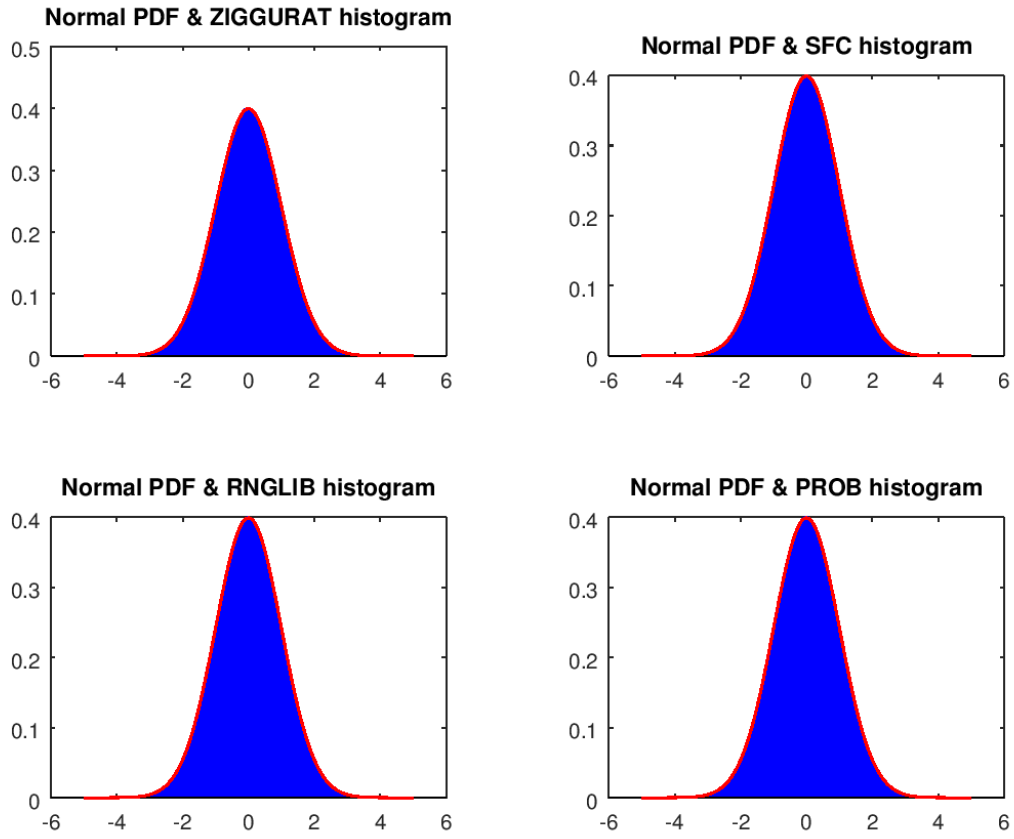

Figure 10: General view - 100000000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

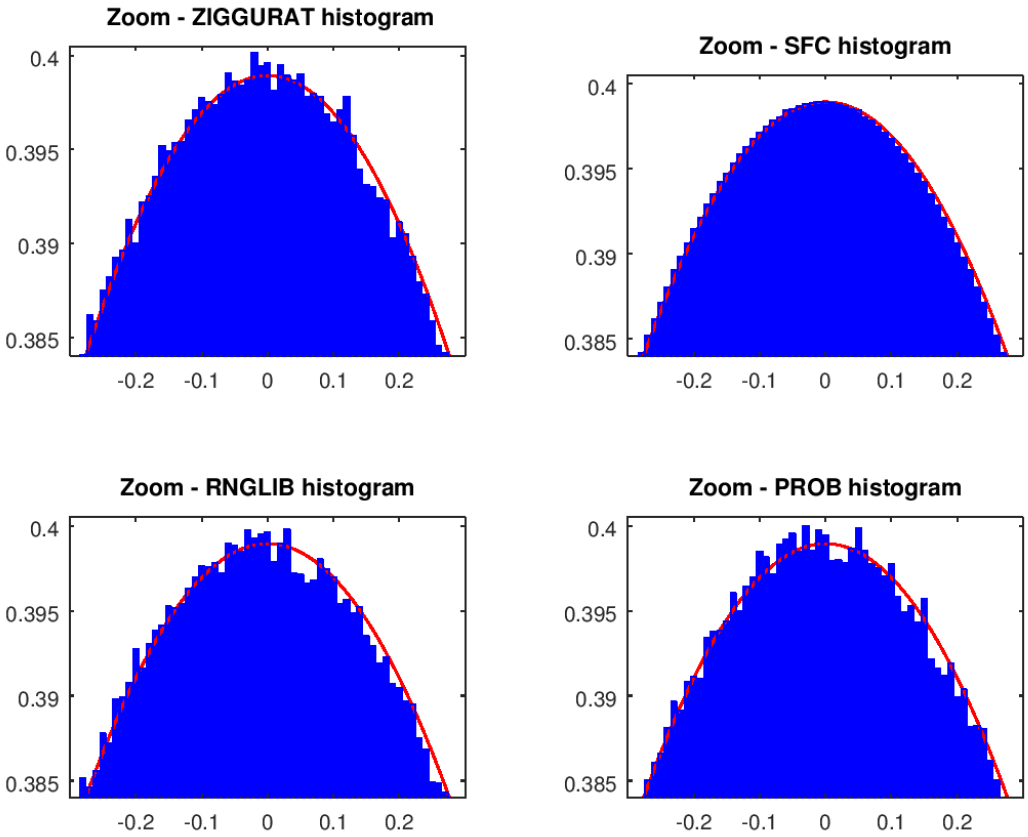

Figure 11: Zoom - 100000000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

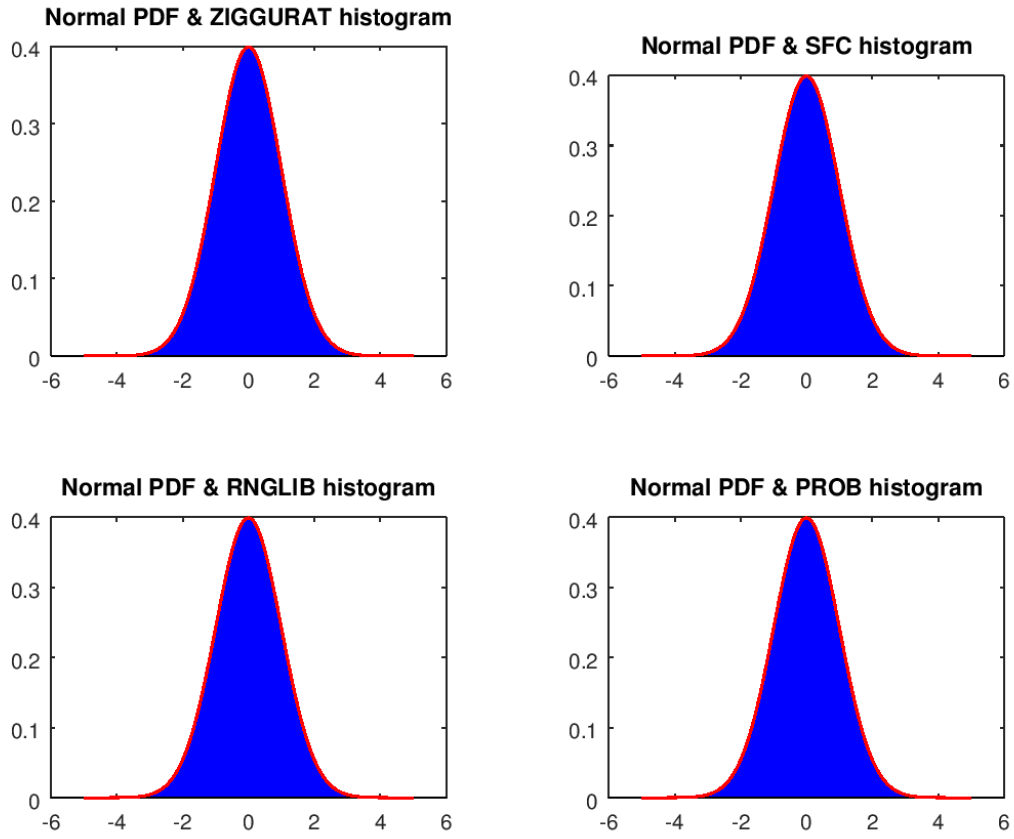

Figure 12: General view - 4000000000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

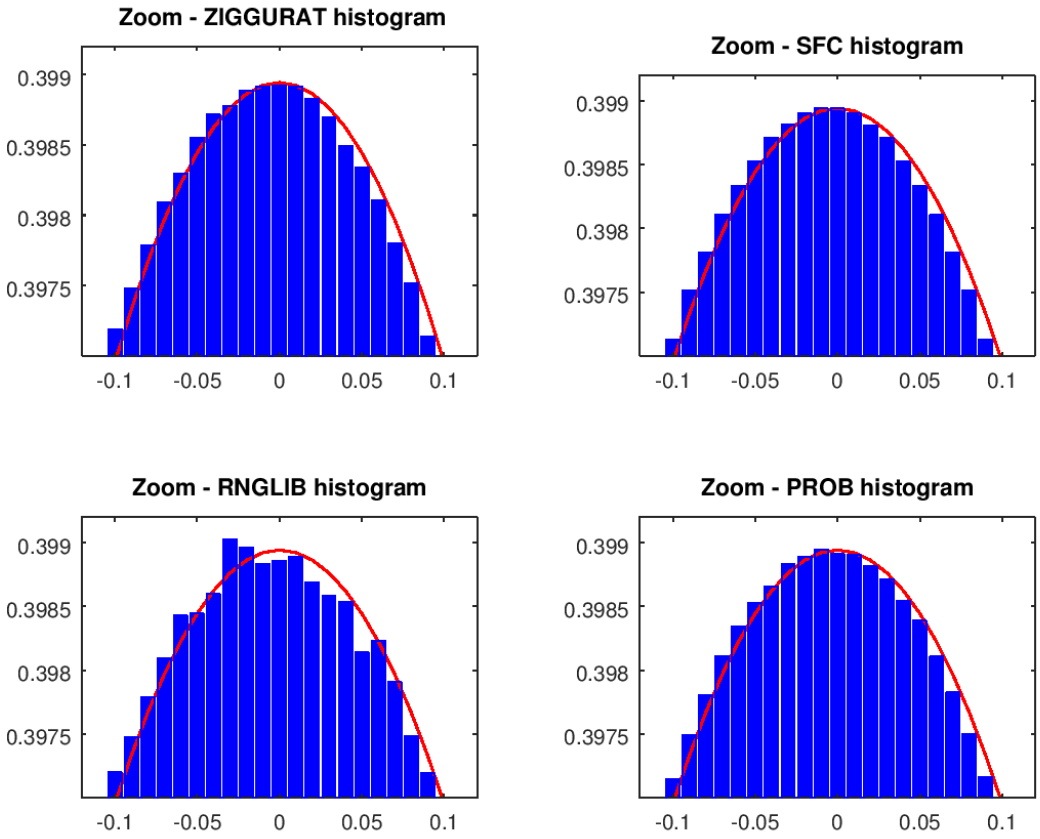

Figure 13: Zoom - 4000000000 samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 September 2021 doi:10.20944/preprints202109.0460.v1

5. Conclusions

Through several simulations, this paper has shown that it is possible to (approximately and) deterministically sample from Gaussian distributions by means of Sierpiński space-filling curves. The proposed method was compared to 3 other well-established algorithms and the discrepancy between the empirical histograms and the true normal PDF was evaluated with the Kullback-Leibler divergence and other two true metrics which represent good indices of disparity between probability distribution functions. By analyzing the numerical results and corresponding graphs, it is possible to infer that the presented algorithm converges faster than the other methods. This type of result may be very useful in applied fields.

References

[1] A. Basu, H. Shioya, C. Park, Statistical Inference - The Minimum Distance Approach, CRC Press, Boca Raton, 2011.

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

[3] J.Burkardt, RNGLIB method implementation, https://people.sc.fsu.edu/~jburkardt/cpp_src/rnglib/rnglib.html, accessed in Apr 21 2021.

[4] J.Burkardt, PROB library, https://people.sc.fsu.edu/~jburkardt/c_src/prob/prob.html, accessed in Sep 21 2021.

[5] J.Burkardt, Ziggurat method implementation, https://people.sc.fsu.edu/~jburkardt/cpp_src/ziggurat/ziggurat.html, accessed in Apr 21 2021.

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

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

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

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

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

[11] N. L. JOHNSON , S. KOTZ , N. BALAKRISHNAN ,Continuous Univariate Distributions (Vol. 1), Wiley, New York, 1994.

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

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

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

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

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

[17] 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.

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

[19] 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.

[20] H.A.Oliveira, Deterministic sampling from uniform distributions with Sierpiński space-filling curves, Computational Statistics, https://doi.org/10.1007/s00180-021-01128-w

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

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