# On Monte-Carlo Methods & Applications in Geoscience

Zhan Zhang,
Dr. J. A. R. Blais
University of Calgary, AB

# *OBJECTIVES*

- To give the geophysicists and geo-scientists a support in terms of the mathematical backgrounds hidden behind the random number generations and relating Monte Carlo algorithm, with several methods are given.
- To gather several problems in assorted branches of geoscience, where the Monte Carlo method can be utilized, general descriptions of each problem will be provided.

## PART I

# Random number & Monte Carlo method

# *PSEUDO RANDOM NUMBER GENERATORS*

- All types of the PRNG will give the uniformly distributed random numbers in ( 0, 1 ), whose idea rooted in the number theory instead of pure theory of probability .

  1. ***Linear Congruential Generator***:

  $$X_n \equiv aX_{n-1} + c \pmod{M}$$

  Certain restriction apply on the choice of a, c & M in order that the sequence has a full period.
  However, LCG has a high serial correlation.

2. ***Lagged Fibonacci Generator***:

$$X_n \equiv X_{n-j} + X_{n-k} \pmod{M}$$

j, k are chosen specially so that the sequence will have a long cycle.
In other cases, the sum in the generator may be replaced by a product or even a linear combination of all the numbers of previous states.

With the generated sequence {X}, the uniformly distributed number sequence {U} is given as:

$$U_n = \frac{X_n}{M}$$

# *CHAOTIC RANDOM NUMBER GENERATOR*

- The generator, known as ***logistic map***, has the form:

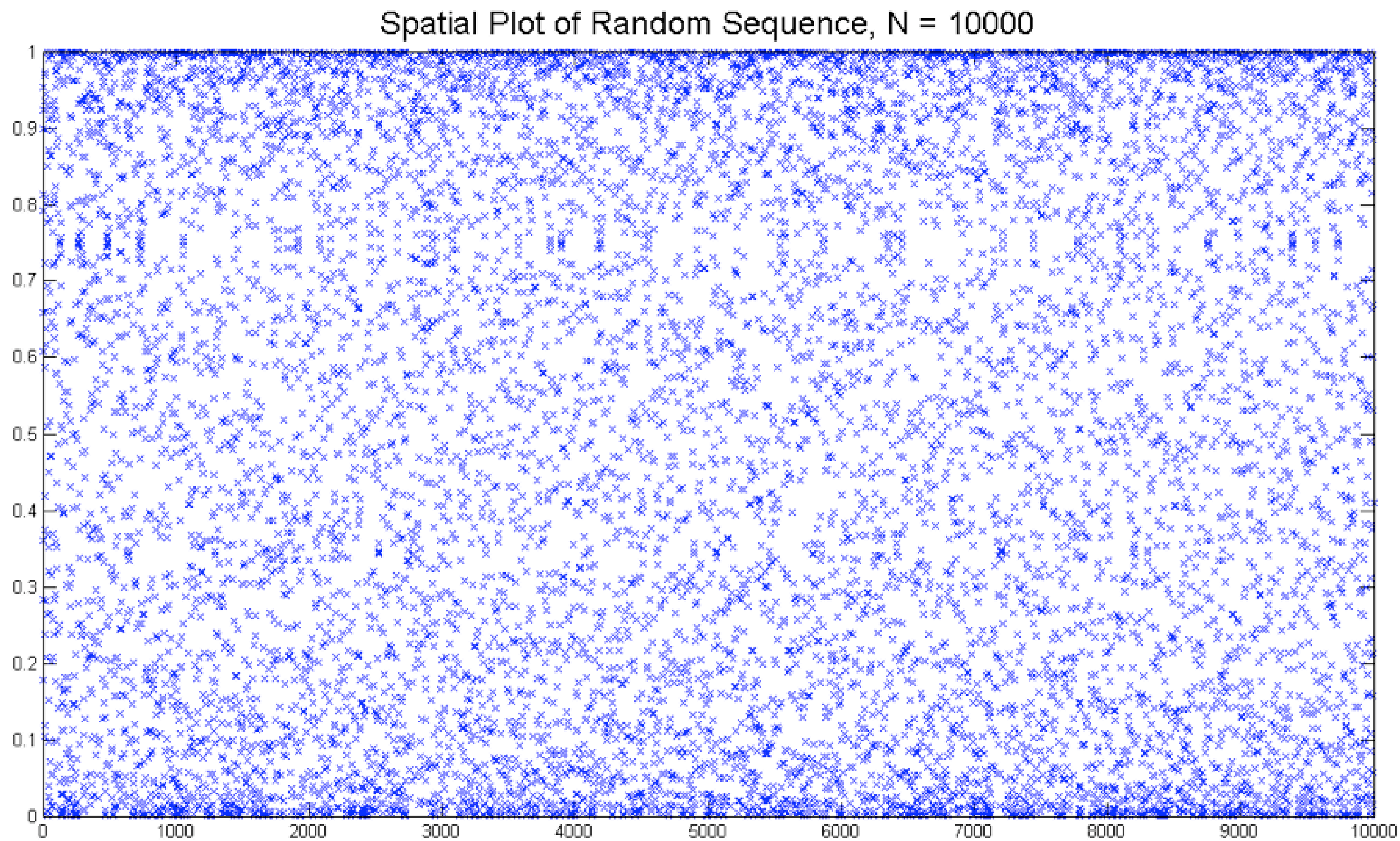
  $$X_n = 4X_{n-1}(1 - X_{n-1})$$

  Given a random seed, this map will produce a sequence of random numbers, which are all absorbed into (0, 1), whose PDF is recognized as:

  $$\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}$$

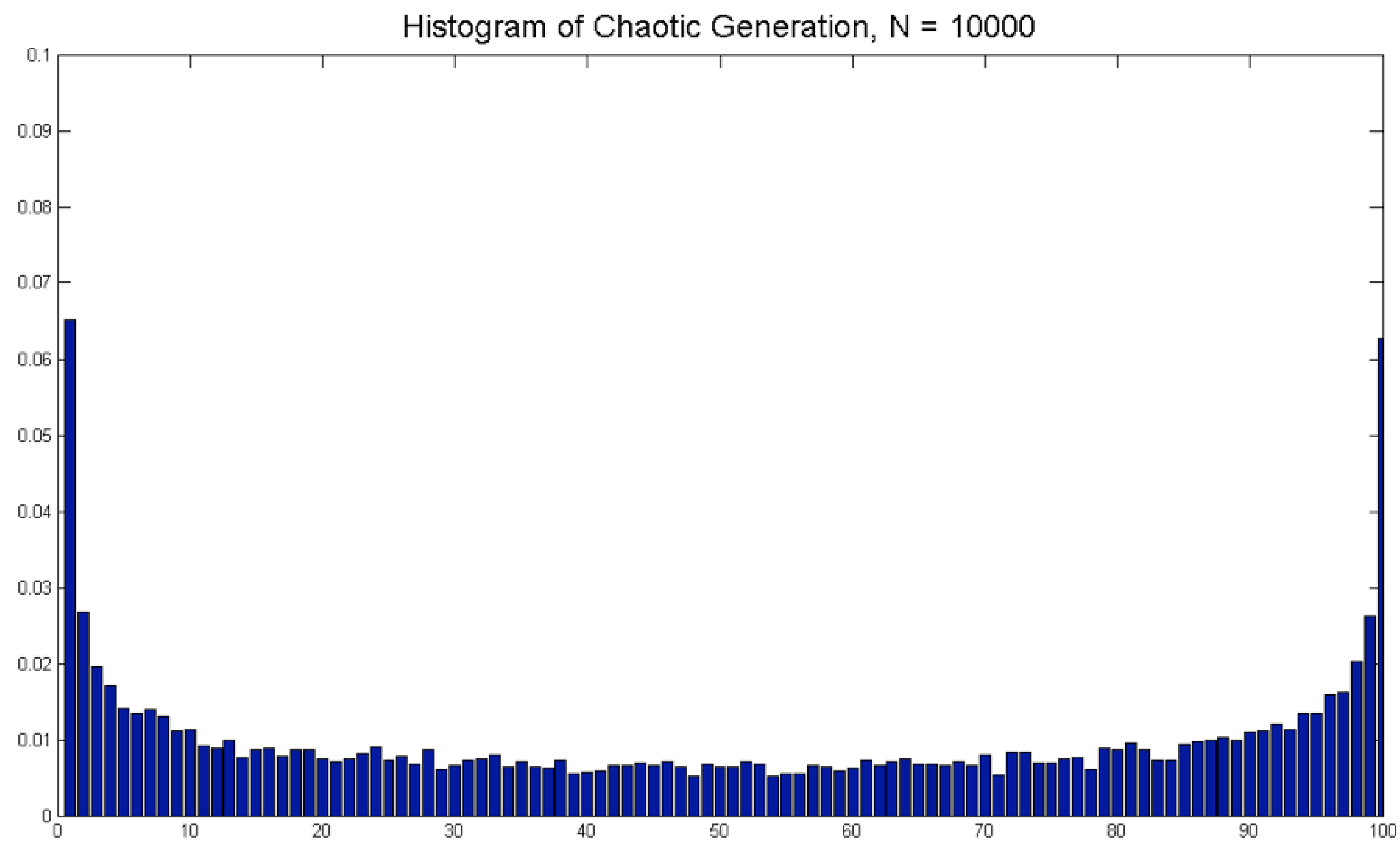
  ***exactly solvable chaos***

  this distribution provides a population which is dense at two extremes 0 and 1.
  This generator is ergodic, and sensitive to the seed.




This gives the spatial plot of chaotic sequence

seed: 0.792207329559554 N = 10000




This gives the histogram of chaotic generation
seed: 0.157613081677548 N = 10000

- In some studies, this logistic map is solved using the well-known trigonometric relation, as:

$$X_n = \sin^2(\theta \pi z^n) \qquad \text{where z is a real number}$$

for z > 1, the fractional z, in principle, provides more unpredicability, leading to more randomness, due to a multiplicity of the solution.

- ***Lyapounov exponent***:
A measure of the predictability of the dynamical system, as:

$$|\delta X(t)| \approx e^{\lambda t} |\delta X(0)|$$

This measures the deviation of the separation of a system in t from that of its initial state.

# *QUASI MONTE CARLO METHOD*

- ***BASIC IDEA***: The random sampling in MC method introduces a certain error in the evaluation, to diminish this error, a set of deterministic points with uniform property is utilized. Therefore, in addition to MC method, '***quasi***' is coined.

- ***Uniform Distribution***:
$\mathcal{E} : 0 \leq u < 1$ and $E : \alpha \leq u < \beta$; where $E \subset \mathcal{E}$
$u_1, u_2, ......, u_N$ be N numbers falling in $\mathcal{E}$, and in which N' numbers in E, then the uniformity is defined:

$$\frac{N'}{N} \rightarrow \beta - \alpha \text{ as } N \rightarrow \infty$$

- ***Discrepancy***: This quantity is a measure for the deviation from the uniform distribution:

$$D(N) = \sup_{(\mathcal{E})} \left| \frac{N'(\mathcal{E})}{N} - (\beta - \alpha) \right|$$

Generally, a sequence of points with the property that this defined discrepancy being low is known as ***low discrepancy sequence***.

- For any sequence, the order of the discrepancy is:

$$O(N^{-1} \log N)$$

- Several low discrepancy sequences are known to us: ***van der Corput's***, ***Hammersley's***, ***Halton's*** and ***Sobol's***.

- *Quasi Monte Carlo vs. Monte Carlo*:
Quasi Monte Carlo is especially used in the numerical integration, since the disadvantage of Monte Carlo method is that it brings the probabilistic uncertainty into the evaluation. Being random is not quite essential in the evaluation, the more important thing is the choice of uniform distributed sequence in the domain, so that the error bound is determined as small as possible.

  Thus, to have a more accurate approximation of the integral of any dimension, quasi-MC method is always suggested than an ordinary MC.

# *MARKOV CHAIN MONTE CARLO*

- ***BASIC IDEA***: In order to sample from some unknown distribution, a Markov Chain (transition matrix) is constructed which preserves the distribution for any number of iteration steps.

- ***Properties of Markov Chain in MC***:
  1. Homogeneity (same P for every transition state) ;
  2. Irreducibility (any transition is possible);
  3. Regularity (existence of limiting absolute probability);
  4. Reversibility ($\pi_i p_{ij} = \pi_j p_{ji}$)

***Metropolis-Hastings Algorithm***:

Let $p_{ij} = q_{ij}\alpha_{ij}$ , where $Q = \{q_{ij}\}$ is another transition matrix, and $\alpha_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$

***Procedure***:

1. Assume an initial value $X(0)$;
2. Let $X(t) = i$ , select the state j based on the distribution given by i-th row of Q ;
3. Take $X(t+1) = j$ with probability $\alpha_{ij}$; or $X(t+1) = i$ with probability $1 - \alpha_{ij}$ .

The whole idea is to construct a reversible Markov Chain ( i.e. a posteriori = a priori in any two states)

# *PSEUDO RANDOM NUMBERS WITH PRESCRIBED DISTRIBUTION*

- Once the pseudo random numbers in (0, 1) are generated, it is sometimes desired to have the random numbers satisfying certain PDF, the general theorem concerning this transformation is:
  ***Theorem***:
  Let $\eta$ be the desired random numbers, with its CDF of the form F(x), $\xi$ satisfies U(0, 1), then the numbers $\eta$ ( with PDF dF(x) ) are obtained by:

$$\eta = F^{-1}(\xi)$$

- ***Complex random numbers with Normal Distribution***:
  To generate a circular normal distributed complex random number (or a normal random point in a plane), one can use ***Rayleigh distribution***:
  Let $\xi$ : U(0, 1), the pair $\eta$ is given by:

$$\eta_1 = (-2\ln\xi_1)^{1/2}\cos(2\pi\xi_2)$$
$$\eta_2 = (-2\ln\xi_1)^{1/2}\sin(2\pi\xi_2)$$

then $(\eta_1, \eta_2)$ is a pair of i.i.d. N(0, 1) random numbers.

This case can be generalized to 3-D space, on generating a i.i.d. N(0,1) random point using ***Maxwell-Boltzmann distribution***.

# PART II
# Applications

# *INTEGRATION*

- ***Definite integral***:

$$\hat{I} = \sum_{i=1}^{N} f(x_i)\Delta x_i$$

x(i): generated sequence in (a, b)

- ***Variance reduction methods for PRNG***:

1. Symmetrization: $$f_1(x_i) = \frac{f(x_i) + f(b - a - x_i)}{2}$$

2. Sub-interval:
Divide [a, b] into M sub-intervals $[a_k, b_k]$ ( k = 1.. M) with fixed probabilistic measure c,

$$\hat{I} = \sum_{k=1}^{N} I_k \qquad \text{where} \quad \int_{a_k}^{b_k} \rho(x)dx = c_k$$

- 3. Approximation of integrand:
If h(x) is close to f(x), with $\int_a^b h(x)dx = J$
then one can use the formula:
$$\hat{I} = J + \sum_{i=1}^{N} |f(x_i) - h(x_i)|\Delta x_i$$
- The same idea can be generalized into multiple integral.
- ***QRNG: Koksma' inequality***:
Using Quasi-MC method, this inequality shows that the error of evaluation is bounded by two terms:
1. the bounded variation of the integrand itself;
2. the discrepancy of the sequence.

- ***Choice of interval***:
The integration value also depends on how to choose the infinitesimal interval $dx_i$ .
(e.g. chaotic generation gives a histogram dense at the extremes, then the equidistance is not a suitable choice)
By a simple deduction, the relation is given by:

$$\Delta x_i = \frac{1}{N \cdot \rho(x_i)}$$

where N is the total number of generated points, $\rho(x)$ is the PDF of the random points.
For uniform distribution, this reduces to $\Delta x_i = \dfrac{1}{N}$

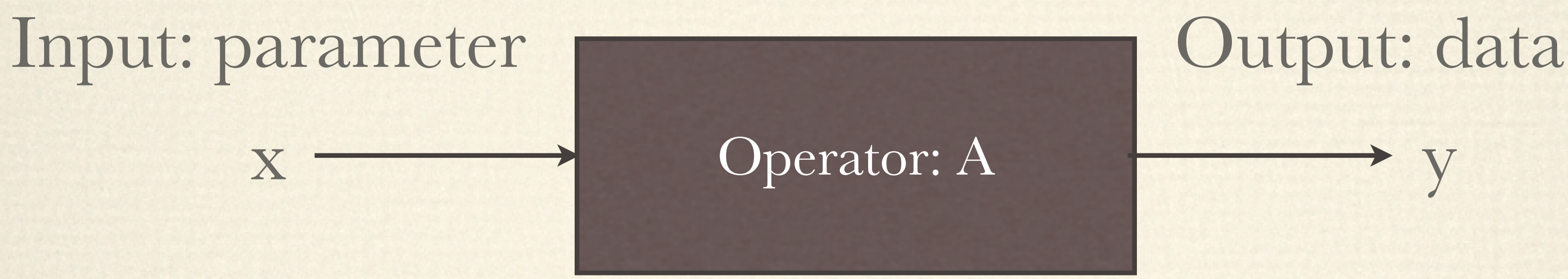
For chaotic distribution, $\Delta x_i = \dfrac{\pi \sqrt{x_i(1 - x_i)}}{N}$

- ***Some Remarks on Integration***:

---1. To compare the efficiency of the calculation by each method, we recommend using hypothesis testing, namely, grouping each method together as experiments, and compare that if they have the same mean, and also their contribution to the variances, in a confidence level.

---2. ***Chaotic generation*** does not give universally more satisfactory results than the normal methods, however, as is shown by Umeno (1999), if the integrand can be approximated by a constrained combination of Tchebyshev polynomials, a superefficiency with the error of the order $O(\frac{1}{N})$ can be reached.

# *INVERSE PROBLEM -- MCMC*




- The inverse problem is to find the input, given an output and an operator, or finding: $x = A^{-1}(y)$
- Various geophysical problems involve this problem:
-- ***Model Optimization***: e.g. randomly search the Earth's model consistent with seismological model;
-- ***Resolution Analysis***.

- ***Simulated Annealing***:
  This ***Bayesian*** technique involves a global optimization of the solution of inverse problem using a Markov random walk (with ***Gibbs distribution***), adapted based on ***Metropolis-Hastings Algorithm***, the final theoretical solution reaches the MAP estimate:

  $$\hat{x} = \arg\max_x P(x|y)$$

  In each iteration, a local change (local characteristics) is made based on the current value and their neighbours (sampling), with the temperature T involving in the calculation, and gradually cooled down so that the local optimization is avoided.

- ***Reliability of Prediction in GIS***:
  The ill-posed queries in GIS can be solved in the linear case (i.e. Gaussian 1st and 2nd moments are used in the solution). However, with the reliability is smaller than one (PR < 1) in the non-linear case, it is needed to give the level for this PR (an error bound). The statistics in this case changes to non-Gaussian, and the a posteriori distribution is required.
  By this assumption, the complexity of the prediction reliability has greatly increased, and only a Monte-Carlo method can approximate this non-linear case.

- ***SAR image processing***:
Given an intensity or complex data image, it is essential to provide a segmentation so that each homogeneous region could be recognized.
*The basic approach*: the image can be understood as a combination of different speckle process, and the grouping process is an Markovian random field, the SA method involves an iterative estimation of parameter, and the conditional posteriori distribution of each pixel is updated at every step, with only dependent on its neighbouring pixels. After reasonable steps, the MAP estimate is approximated.

# ***Bibliography -- Mathematics***


- Gantmacher, F. R.: *The Theory of Matrices*, vol.2, Chelsea Publishing Co., 1959
- Gauss, C.F. (1801): *Untersuchungen über die höhere Arithmetik* (trsl.*Disquisitiones Arithmeticae*) Chelsea Publishing Co., 1965
- Hastings, W.K. (1970): *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57, 1, p.97-109
- Knuth, D.: *The Art of Computer Programming*, vol.2, 3rd Edi., The Art of Computer Programming Series.
- Koksma, J.F. (1964): *The theory of asymptotic distribution modulo one*, Compositio Mathematica, tome 16, p.1-22

- Kolmogoroff, A. (1936): *Zur Theorie der Markoffschen Ketten*, Mathematische Annalen, Vol.112, Nr.1, Springer-Verlag.
- Niederreiter, H.: *Random number generation and quasi-Monte Carlo methods*, CBMS-NSF, Regional Conference Series in Applied Mathematics.
- Ulam, S.M. & von Neumann, J. (1947), Bulletin of AMS. Soc. 53, 1120.
- Umeno, K. (1999): *Chaotic Monte Carlo computation: a dynamical effect of random-number generations*, Japanese Journal of Applied Physics, part I. Vol.39
- Weyl, H. (1916): *Über die Gleichverteilung von Zahlen mod.Eins.*, Mathematische Annalen, Vol.77, Nr.3, Springer-Verlag.

# *Bibliography* -- Geoscience


- Blais, J.A.R. (2002): *Reliability considerations in geospatial information systems*, Geomatica, Vol.56, No.4, p.314-350
- Derin, H., Kelly, P.A., Veniza, G. & Labbit, S.G. (1990): *Modeling and Segmentation of Speckled Images using Complex Data*, IEEE Transactions on Geoscience and Remote Sensing, Vol.28 No.1, 76-87
- Geman, S.& Geman, D. (1984): *Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.Pami-6, No.6, 721-741
- Glimm, J. & Sharp, D.H.(1998): *Prediction and the Quantification of Uncertainty*, Los Alamos, tome 16, p.1-22

- Mosegaard, K. & Tarantola, A. (1995): *Monte Carlo sampling of solutions to inverse problems*, Journal of Geophysical Research, Vol. 100, No.B7, p12431-12447


*End of Presentation,*
*Thank you very much!*
*La fin de la présentation,*
*Merci Beaucoup!*