ACTA STEREOL 1996; 15/1: 71-76
ORIGINAL SCIENTIFIC PAPER

A STEREOLOGICAL ALGORITHM TO GENERATE UNIFORM POINTS OVER A DOMAIN IN A HIGH DIMENSIONAL SPACE

Christian LANTUEJOUL
Centre de Géostatistique, Ecole des Mines, 35 rue Saint-Honoré
77305 Fontainebleau, France

ABSTRACT

A sequential algorithm is proposed to generate uniform points over a d-dimensional domain. This algorithm is applicable whatever the shape of the domain and remains efficient even if d is large. This paper presents the algorithm, proves its validity and gives bounds for the rate of convergence.

Keywords: multivariate uniform generation, Markov chain.

INTRODUCTION

The objective of the present work is to devise conditional simulation algorithms for random sets or random functions, specifically simulations that take some prespecified values at certain data points. To a certain extent, this amounts to generating a uniform point over a domain which has a dimension equal to the number of conditioning data points. As this number can be rather large (several hundred data points are not uncommon), the classical acceptance-rejection method is usually inefficient, and consequently another algorithm is required. The large number of dimensions suggests a stereological approach.

The stereological approach adopted here consists in simulating uniform points over a d-dimensional domain only by generating uniform points over unidimensional bounded sets. The idea of a stereological simulation dates back many years. It can be found in a paper by Turčin (1971). There are however different ways to implement this idea and the one considered here seems to be original.

The paper starts with a description of the algorithm and the proof of its validity. Since the algorithm is sequential, the problem of its rate of convergence will be then addresssed. A discussion where various implementations are compared and several generalizations are proposed, concludes this paper.

THE ALGORITHM

Let D_0 be an open subset of $\mathbb{R}^d$ with finite volume. D_0 has an arbitrary shape. It is not necessarily bounded, nor convex, nor even connected. The following sequential algorithm is proposed for generating a uniform point over D_0:

i) let x be an arbitrary point in D_0.
ii) generate a uniformly oriented line L passing through x.
iii) generate a uniform point y over $L \cap D_0$.
iv) take $x = y$ and goto ii).

To illustrate how this algorithm works, we have taken D_0 to be a two-dimensional domain, made up of two connected components. The initial point x has been chosen in the non simply connected component (cf. Fig. 1.1). 1000 uniformly oriented lines have then been generated through x, resulting in 1000 points at the first iteration (cf. Fig. 1.2). The distribution of these points is far from uniform. Notice in particular that a high density of points is observed along the lines through x having a short intersection with D_0. The 1000 points are used as input for a second iteration, and Fig. 2.3 to 2.6 show the evolution of the simulation at iterations 2, 3, 5 and 10. A gradual convergence toward uniformity is observed. The rate of convergence seems to be quite fast since uniformity has been practically reached as early as at the 5^{th} iteration.

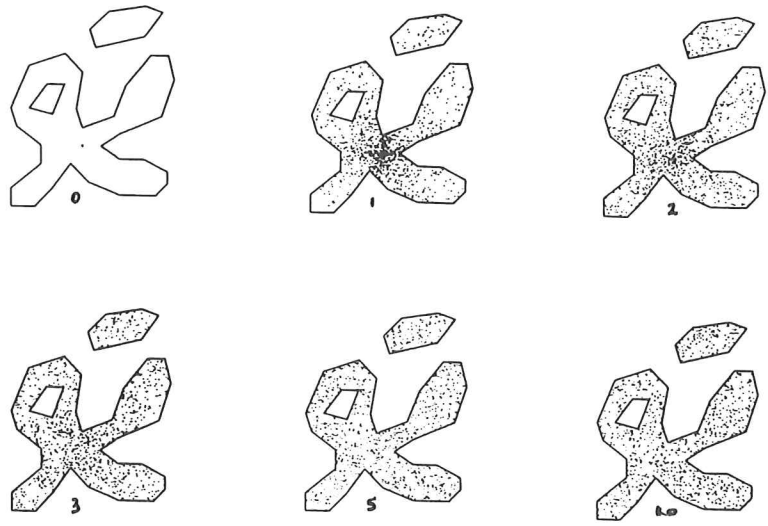

Fig. 1. Simulation of 1000 points at iterations 0, 1, 2, 3, 5 and 10.

PROOF OF THE ALGORITHM

Let X_n be the random point at the n^{th} iteration. $X_n \in D_0$ with $X_0 = x$. The sequence $\{X_n\}_{n \geq 0}$ is a homogeneous Markov chain. It is ergodic since any point is accessible from

any other at each iteration. Consequently, the Markov chain admits a unique stationary distribution, denoted by p. Furthermore, p is the only distribution over D_0 which is invariant under the transition distribution of the Markov chain. In other words, we have

$$\int_{D_0} p(x)\, P(x, D)\, dx \;=\; p(D), \qquad D \in \mathcal{B}(D_0) \tag{1}$$

($\mathcal{B}(D_0)$ is the family of Borel subsets of D_0), with

$$P(x, D) \;=\; P\{X_1 \in D \,|\, X_0 = x\}. \tag{2}$$

To prove the algorithm, it therefore suffices to verify that the uniform distribution over D_0 is invariant under the transition distribution. A direct calculation shows that $P(x, D)$ admits the density function

$$f(x, y) \;=\; \frac{2}{d\omega_d} \frac{1}{l(x, y, D_0)\, |x - y|^{d-1}}, \qquad y \neq x, \tag{3}$$

where ω_d denotes the volume of the unit ball in $\mathbb{R}^d$, and where $l(x, y, D_0)$ stands for the length of the intersection between D_0 and the line through x and y. Since f is symmetric, we have

$$\begin{aligned} \int_{D_0} P(x, D)\, \frac{dx}{|D_0|} \;&=\; \frac{1}{|D_0|} \int_{D_0} \int_D f(x, y)\, dy\, dx \\ &=\; \frac{1}{|D_0|} \int_D \int_{D_0} f(y, x)\, dx\, dy \\ &=\; \frac{1}{|D_0|} \int_D dy = \frac{|D|}{|D_0|} \end{aligned}$$

and the proof has been completed. Thus, we have

$$p(D) \;=\; \frac{|D|}{|D_0|}, \qquad D \in \mathcal{B}(D_0). \tag{4}$$

RATE OF CONVERGENCE

Let $P^{(n)}(x, D)$ be the transition distribution of order n of the Markov chain

$$P^n(x, D) \;=\; P\{X_n \in D \,|\, X_0 = x\}. \tag{5}$$

What has been established in the previous section can be also written as

$$\lim_{n \longrightarrow +\infty} P^n(x, D) \;=\; p(D), \qquad D \in calB(D_0). \tag{6}$$

It remains to see how the quantity

$$h^{(n)}(x, D) \;=\; P^{(n)}(x, D) - p(D), \qquad D \in calB(D_0), \tag{7}$$

converges to 0 as n tends to $+\infty$. This will be done under the assumption that

$$\inf_{\substack{x, y \in D_0 \\ x \neq y}} f(x, y) \geq \frac{\delta}{|D_0|} > 0, \tag{8}$$

which holds if D_0 is bounded. Integrating $f(x,.)$ over D_0 gives $\delta \leq 1$. Actually, some continuity argument on f yields the stronger inequality $\delta < 1$.

In what follows, we are going to show that

$$|h^{(n)}(x,D)| \leq (1-\delta)^n \tag{9}$$

holds for any $n \geq 1$. The proof inspired by Doob (1953) is done by induction.

Consider first the case $n = 1$. Define

$$h(x,D) = P(x,D) - p(D), \tag{10}$$

$h(x,D)$ satisfies two inequalities

$$\begin{aligned} h(x,D) &\geq p(D)(\delta-1) \geq \delta - 1, \\ h(x,D) &= -h(x,D_0 \setminus D) \leq -p(x,D_0 \setminus D)(\delta-1) \leq 1-\delta, \end{aligned}$$

which can be summarized into

$$|h(x,D)| \leq 1-\delta. \tag{11}$$

Suppose now that the inequality (9) is true at order n. Since p is invariant under the transition distribution of order n, we can write

$$h^{(n+1)}(x,D) = \int_{D_0} h^{(n)}(x,dy)\, h(y,D). \tag{12}$$

Notice that $h^{(n)}$ is a signed measure with a zero integral. According to the Jordan-Hahn theorem (Neveu,1964), there exists a Borel subset D_n of D_0 such that

$$D \subset D_n \Longrightarrow h^{(n)}(x,D) \geq 0 \qquad D \subset D_0 \setminus D_n \Longrightarrow h^{(n)}(x,D) \leq 0. \tag{13}$$

Now the positive and the negative part are separated,

$$h^{(n+1)}(x,D) = \int_{D_n} h^{(n)}(x,dy)\, h(y,D) + \int_{D_0 \setminus D_n} h^{(n)}(x,dy)\, h(y,D), \tag{14}$$

which allows $h^{(n+1)}(x,D)$ to be upper bounded

$$\begin{aligned} h^{(n+1)}(x,D) &\leq p(D_0 \setminus D)(1-\delta)h^{(n)}(x,D_n) - p(D)(1-\delta)h^{(n)}(x,D_0 \setminus D_n) \\ &= p(D_0 \setminus D)(1-\delta)h^{(n)}(x,D_n) + p(D)(1-\delta)h^{(n)}(x,D_n) \\ &= (1-\delta)h^{(n)}(x,D_n) \\ &\leq (1-\delta)^{n+1}, \end{aligned}$$

as well as lower bounded

$$\begin{aligned} h^{(n+1)}(x,D) &\geq -p(D)(1-\delta)h^{(n)}(x,D_n) + p(D_0 \setminus D)(1-\delta)h^{(n)}(x,D_0 \setminus D_n) \\ &= -p(D)(1-\delta)h^{(n)}(x,D_n) - p(D_0 \setminus D)(1-\delta)h^{(n)}(x,D_n) \\ &= -(1-\delta)h^{(n)}(x,D_n) \\ &\geq -(1-\delta)^{n+1}, \end{aligned}$$

and finally

$$|P^{(n+1)}(x, D) - p(D)| \leq (1 - \delta)^{n+1}, \tag{15}$$

which is the desired result.

This formula suggests a fast rate of convergence, but the coefficient δ can be quite small. However, it should be pointed out that the proof does not really account for the geometry of D_0, and the obtained bounds are certainly quite loose. Practical experience shows that the rate of convergence is faster than the formula suggests. Consider for instance the pyramid defined in $\mathbb{R}^d$ by the set of inequalities $x_i \geq 0$ for $i = 1, ..., d$ and $x_1 + \cdots + x_d \leq 1$. Experiments have been carried out for several workspace dimensions, namely $d = 3, 10, 20$ and 100. For each d value, 1000 points have been generated using 1000 iterations starting from the point of coordinates $x_i = 1/2\sqrt{d}$ $(i = 1, ..., d)$. A criterion must be introduced in order to judge the quality of the simulations. This is the distribution function of the distance of a uniform point in the pyramid to the hyperplane of equation $\sum_{i=1}^{d} x_i = 1$. The theoretical curves (plain line) and the experimental ones (dotted line) are compared on Fig. 2. The results look quite satisfactory in spite of the limited number of measurements. It is also worthwhile mentioning that the simulation of a uniform point using the acceptance rejection method by enclosing the pyramid within the d-dimensional unit cube would require $d!$ attempts on average.

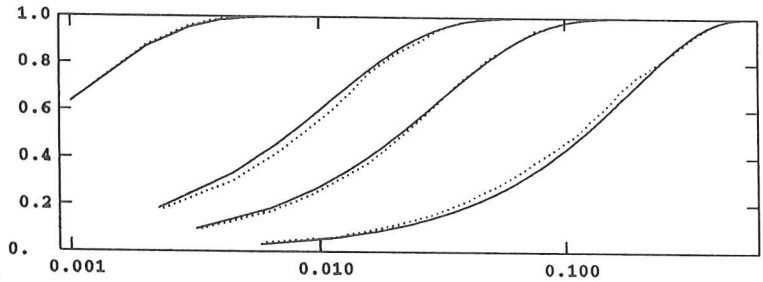

Fig. 2. Distribution function of the distance of a uniform point in a d-dimensional pyramid to its base. From right to left $d = 3, 10, 20$ and 100.

DISCUSSION

Turčin's original idea was to generate lines with a finite set of possible directions. This simplifies the determination of the linear sections, and makes the algorithm easier to implement. Note however that an ergodic problem can occur if the set of directions has not been suitably chosen, especially in the case where the domain D_0 is not connected (cf. Fig. 3).

Resorting to uniformly oriented lines has three major advantages:

i) ergodic problems are avoided,

ii) the Markov chain is more mixing, which implies a faster rate of convergence,

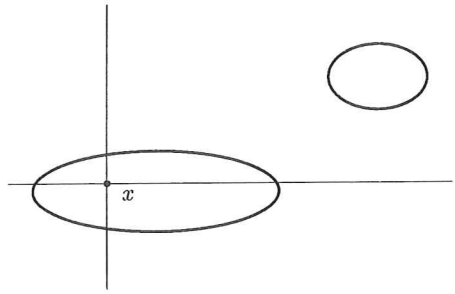

Fig. 3. In the case where the lines are only allowed to be either horizontal or vertical, transitions between the connected components of D_0 are not possible. The Markov chain is not ergodic.

iii) with uniformly oriented lines, the rate of convergence depends only on the geometry of the domain D_0.

Regarding ii), replacing lines with fixed directions by uniformly oriented lines is not the only way to improve the mixing properties of the Markov chain. Another possibility is to consider i-planes instead of lines. Steps ii) and iii) of the algorithm then consist in generating a uniform point over the section of D_0 with an i-plane through x, the orientation of which is either uniform or chosen at random within some prespecified set of orientations. The larger the i value, the more mixing the Markov chain. More generally, there is no inconvenience in working with a transition distribution that involves planes of different dimensions. This is of special interest in the case where the domain under study presents some degree of symmetry. Consider, for example, the cone C_d defined in $\mathbb{R}^d$ as the subgraph of the function

$$f(x) = (1 - |x|)1_{|x|\leq 1} \qquad x \in \mathbb{R}^{d-1}$$

A fast way to simulate uniform points within C_d is to run the algorithm with "horizontal" hyperplanes (simulating uniform points within a ball of $\mathbb{R}^{d-1}$ does not cause any difficulty) and "vertical" lines. Since two successive simulations in the same horizontal hyperplane or the same vertical line are useless and time consuming, the successive sections should be taken in a sequential order and not at random.

REFERENCES

Doob JL. Stochastic Processes. New York: Wiley, 1953: 190-218.
Neveu J. Bases mathématiques du Calcul des Probabilités. Paris: Masson, 1964: 99-101.
Turčin VF. On the computation of multidimensional integrals by the Monte Carlo me thod. Theory Prob Appl 1971; 16:720-4.

Presented at the 9th International Congress for Stereology, Copenhagen, August 20th to 25th, 1995.