1981 Winter Simulation Conference Proceedings
T.I. Ören, C.M. Delfosse, C.M. Shub (Eds.)

RECENT RESULTS IN NON-UNIFORM RANDOM VARIATE GENERATION

Luc Devroye
School of Computer Science
McGill University
805 Sherbrooke Street West
Montreal, Canada H4A 2K6

ABSTRACT

A selective survey is given of new methods in non-uniform random variate generation.

We consider the problem of the computer generation of random variables with a given continuous distribution function F on the real line. We are not concerned here with the uniform distribution on $[0,1]$, but rather, we assume that an infinite sequence of independent uniform $[0,1]$ variates is available to us. We also make the unrealistic assumption that real numbers can be stored in a computer.

The following issues are discussed:

(i) the average time required per random variate (here we need the assumption that the basic operations $\times$, $/$, $+$, $-$, exp, log, mod, compare, move take a constant time each);
(ii) the storage requirements for the algorithm;
(iii) the portability and flexibility of the algorithms (if F belongs to a parameterized family of distributions, how easy is it to change the parameter at each call?)

The method one uses depends upon the application and the manner in which F is given to us. For example, in complex simulations, the average time per random variate is less important. But it is essential that the generators are simple, short and easy to understand. The distribution can be given to us in many ways:

A. F is explicitly known and easy to invert: the inversion method seems a prime candidate: generate a uniform $[0,1]$ random variate U and exit with $X \leftarrow F^{-1}(U)$. The time taken is proportional to the "complexity" of F^{-1}. For the exponential or Cauchy distributions, F^{-1} requires a logarithm or a tangent, respectively. Even when F^{-1} is known, its complexity can make the inversion method undesirable. On the other hand, the antithetic variate principle requires anti-correlated random variates, and these are ideally obtainable by the inversion method (Fox, 1980).

B. F is explicitly known and hard to invert: if no other method is applicable, one could try to solve the equation $F(X) = U$ for X (where U is a uniform $[0,1]$ random variate), and stop when a given accuracy is obtained. Using the bisection method, it is possible to obtain an algorithm that is convergent under any circumstances. More sophisticated methods (secant method, Newton-Raphson method) require some additional information about F needed to insure the convergence. For example, the Newton-Raphson method requires the knowledge of F'. A theoretical and comparative study of these techniques for random variate generation is now carried out by Colleen Yuen

81CH1709-5/81/0000-0517$00.75

at McGill University (Yuen, 1981). We cite a couple of results: (i) If bisection is applied as follows: "start at 0 and find an interval containing the solution of $F(X) = U$ by successive doubling, i.e. $[0,1]$, $[1,2]$, $[2,4]$, $[4,8]$, ... or $[-1,0]$, $[-2,-1]$, ... are the sequences of intervals under consideration, and use the bisection method until an interval of length at most δ is obtained", then the average time is close to a constant times $E(\log_+|X|)+\log(1/\delta)$; thus, there are distributions for which the average time is infinite (note however that $E(\log_+|X|) = \infty$ implies that the average number of digits in the integer part of $|X|$ is infinite !), and a constant number of additional iterations adds a constant number of digits to the accuracy of X . (ii) The Newton-Raphson method converges for unimodal distributions when the search is started at the mode. The average time taken in that case is equal to a constant (possibly ∞) depending upon F only plus $\log \log 1/\delta$, i.e. the choice of δ has little influence on the average time. (iii) The Newton-Rapson method is extremely well suited for certain families of distributions such as $F(x) = 1 - \exp(-a_1x - a_2x^2 - \ldots a_dx^d)$ where all $a_i \geq 0$ and $x \geq 0$. Random variates from this distribution can be obtained as the solution of $a_1X + \ldots + a_dX^d = E$ or as $X \leftarrow \min_{1\leq i\leq d} ((E_i/a_i)^{1/i})$ where $E,E_1,\ldots,E_d$ are independent exponential random variates. Both methods are competitive.

Less flexible programs were developed by numerous authors (e.g., Butler (1970), Akima (1970), Guerra, Tapia and Thompson (1978) and Ahrens and Kohrt (1981)). In essence, they store $(x_1,F(x_1)),\ldots,(x_n,F(x_n))$ in a large table; on each interval (x_i,x_{i+1}) , F is approximated by a polynomial. Random variate generation is extremely fast at the cost of longer programs, less flexibility with respect to parameter changes, large storage requirements and, in some cases, reduced accuracy. Yuen (1981) has developed an attractive general purpose adaptive program in which a table is constructed as random variates are being generated. Frequent parameter changes have a tolerable impact on the average time per random variate, while no parameter changes cause faster average times per random variate as longer sequences are produced.

C. The density f is explicitly given. This is the best developed case. Often f can be decomposed into $a_1f_1 + \ldots + a_df_d$ where $f_1,\ldots,f_d$ are component densities and $a_1,\ldots,a_d$ are probability weights. Or another density g is known such that $f(x) \leq c\, g(x)$ for some $c \geq 1$. Or one recognizes that $f(x) = \max(a_1f_1(x),\ldots, a_df_d(x))$ where $a_1,\ldots,a_d$ are constants and $f_1,\ldots,f_d$ are densities. We know how to obtain random variates from f if we we know how to obtain random variates from g , $f_1,\ldots,f_d$, by the composition and rejection methods (for a survey, see Schmeiser (1980) or Vaduva (1977)). We should point out the benefit (in time) of using a squeeze step (as demonstrated, for example, in the fast and short gamma generators of Marsaglia (1977) and Best (1978)). Ahrens and Dieter (1981) report some success with the following new method: let g be a density close to f and let $T = \{x : g(x) > f(x)\}$; generate a random variate X with density g ; if $X \notin T$, exit; if $X \in T$, generate a uniform $[0,1]$ random variate U ; if $U \leq f(x)/g(x)$, exit; otherwise, generate a new X from the density $c(f(x) - g(x))$, $x \in T$, where c is a normalization constant. Finally, we would like to point out that most well-known densities can be dominated by a function of the form $\min(c_1,c_2/x^2)$ (which is proportional to the density of $\sqrt{\frac{c_2}{c_1}}\frac{V_1}{V_2}$ where V_1,V_2 are independent uniform $[-1,1]$ random variables). For these densities, the rejection method is essentially equivalent to the ratio-of-uniforms method (Kinderman and Monahan, 1977).

D. The density is given as a series. When $f(x)$ is given as a convergent series, such as

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \frac{1}{\sqrt{2\pi}} \sum_{i=0}^{\infty} \frac{1}{i!}\left(-\frac{x^2}{2}\right)^i ,$$

one can use the method of consecutive acceptance and rejection. For example, in the example of the normal density, we generate X from a dominating density g ($f \leq c\, g$ for some constant c) and have to compute $f(X)$ at some point in the standard rejection algorithm. Of course, since $f(x) \geq \frac{1}{\sqrt{2\pi}}(1 - \frac{x^2}{2})$, we can accept after considering only one term of the series with fairly high probability. Similarly, we may reject X by using $f(x) \leq \frac{1}{\sqrt{2\pi}}(1 - \frac{x^2}{2} + \frac{x^4}{8})$, etcetera. In other words, we use only a partial sum to make our decision. For the details, see Devroye (1981a). For the normal distribution, this method is not competitive with other methods. Several examples are given in Devroye (1981a) where this method is either the fastest known method of reasonable size or where this method seems to be the only feasible one, e.g. when f is the derivative of

$$F(x) = \sum_{i=-\infty}^{+\infty} (-1)^i \; e^{-2i^2x^2} \quad ,$$

the Kolmogorov-Smirnov distribution.

E. The characteristic function ϕ is given. When ϕ, ϕ'' are absolutely integrable and absolutely continuous, then $f(x) \le \frac{1}{2\pi} \int |\phi(t)| dt$ and $f(x) \le \frac{1}{2\pi x^2} \int |\phi''(t)| dt$. Thus, f is dominated by a function of the type $\min(c_1, c_2/x^2)$. The latter function is proportional to the density of $\sqrt{\frac{c_2}{c_1}} \frac{V_1}{V_2}$ where V_1, V_2 are independent uniform $[-1,1]$ random variables. Thus, in essence, we can use the rejection method here if it is feasible to compute the integral $\frac{1}{2\pi} \int \phi(t) \; e^{ixt} \, dt$. In the case of a perfect rejection algorithm, this integral needs to be computed once for every random variate X . For a discussion with further details, see Devroye (1981b). It is an open problem whether the integral computation can be avoided altogether. An acceptable solution of this problem could lead to fast algorithms for the symmetric stable distributions $\phi(t) = e^{-|t|^\alpha}$, $\alpha \in (0,2]$ that are competitive with the method of Chambers, Mallows and Stuck (1976) [generate a uniform $[-\frac{\pi}{2}, +\frac{\pi}{2}]$ random vairate U and an independent exponential random variate E , and exit with

$$X \leftarrow \frac{\sin(\alpha U)}{(\cos U)^{1/\alpha}} \cdot \left[\frac{\cos((1-\alpha)U)}{E}\right]^{\frac{1-\alpha}{\alpha}} \quad , \quad \alpha \neq 1 \; ,$$

and $X \leftarrow \tan U$ when $\alpha = 1$.] . Or it could be used for the family of unimodal distributions with characteristic function $1/(1 + |t|^\alpha)$, $0 < \alpha \le 2$ (given in Lukacs, 1970, pp. 96).

We note here that the sin-cos method for stable distributions is time consuming. Perhaps an intelligent application of the series method using the Bergström series for the stable densities (Bergström, 1952; see Bartels (1981) for truncation bounds for the series) will lead to a faster method (but certainly not to a shorter method!).

F. The hazard rate is given. The hazard rate $h(x)$ is given by $f(x)/(1 - F(x))$, and the cumulative hazard rate is $H(x) = \int_{-\infty}^{x} h(y)dy = \log (1 - F(x))$. Thus when E is exponential, $H^{-1}(E)$ has cumulative hazard rate H ; sometimes, we need to solve $H(X) = E$ for X . If $X_1, \ldots, X_n$ are independent with hazard rates $h_1, \ldots, h_n$, then $\min(X_1, \ldots, X_n)$ has hazard rate Σh_i . Finally, if $g \ge h$ is a hazard rate on $(0,\infty)$, and we use the thinning algorithm (Lewis and Shedler, 1979):

1. $X \leftarrow 0$.
2. Generate Y with hazard rate $g(x - X)$. Generate an independent uniform $[0,1]$ random variate U . Set $X \leftarrow X + Y$.
3. If $U \le \frac{h(X)}{g(X)}$, exit with X . Go to 2 ,

then X has hazard rate h on $(0,\infty)$. In other words, the inversion, composition and rejection principles have analogues in the case that h is given.

G. An approximation of f on F is known. One can be given a good approximate density g or a close distribution function G . The problems that we are faced with here are:

(1) Is the replacement of f by g (or of F by G) for the purposes of simulation allowable?
(2) If not, how can we use this information in the generator itself?

In Devroye (1981c), several arguments are given for the use of $\int |f(x) - g(x)| dx$ as a measure of the goodness of the approximation, rather than $\sup_x |F(x) - G(x)|$, or $\sup_{0<y<1} |F^{-1}(y) - G^{-1}(y)|$. The main argument was based upon the relation

$$\sup_A \left| \int_A f(x)dx - \int_A g(x)dx \right| = \frac{1}{2} \int |f(x)-g(x)| dx$$

where A ranges over all Borel sets on the real line. For example, if the said integral is 0.001 , then, no matter how one chooses A (e.g., an interval, a union of intervals, etcetera), the probability of $[X \in A]$ will differ from the desired probability of $[X \in A]$ by at most 0.0005 (0.05%) . In experiments with sample size not exceeding 4000 , the replacement of f log g would hardly be noticeable.

If F - G is small, we can try to solve $F(X) = U$ starting with the estimate $X \leftarrow G^{-1}(U)$. Thus, in the inversion method, the information contained in G can be of some use (see Yuen (1981)).

In the rejection method, we may use the information contained in g as follows: since $f \le g + |f - g|$, we generate two independent uniform $[0,1]$ random variates U and V . Then, if $V < \frac{1}{1+\alpha}$ ($\alpha = \int |f-g|$ is known, but small compared to 1), generate X with density g . Otherwise, generate X with density proportional to $|f - g|$. If $f(X) \ge g(X)$, exit with X . If $f(X) < g(X)$, and $U \cdot [2\frac{g(X)}{f(X)} - 1] \le 1$, exit with X . Otherwise, start all over

again. The average number of restarts is $1 + \alpha$ (including the original start). Note that it is not necessary that $f \le c\,g$ for some constant c.

H. The density is given as an integral. Khintchine's theorem states that X has a unimodal density f if and only if $X = UY$ where U is uniform on (a,b) and Y is a random variable, independent of U. If $(a,b) = (0,1)$ and Y has density f, then X has density

$$g(x) = \int_x^\infty \frac{f(z)}{z}\,dz .$$

From this and related results one can deduce an enormous number of facts that are useful in random variate generation (see Lux (1978), Mikhailov (1965) and Bryson and Johnson (1981) for details and extensions), for example, we have the relations:

density$_f$ of Y	density$_g$ of X
exponential (e^{-x})	exponential-integral $(\int_x^\infty \frac{e^{-z}}{z}\,dz)$
Maxwell $(\sqrt{\frac{1}{2\pi}}\,x^2 e^{-x^2/2})$	normal $(\frac{1}{\sqrt{2\pi}}\,e^{-x^2/2})$
density of max (U,V) or $\sqrt{U}$ (U,V are independent uniform (0,1) random variables)	density of min(U,V) (U,V are independent (0,1) random variables)
gamma $(2,1)$ (xe^{-x})	exponential (e^{-x})
beta $(2,b)$	beta $(1, b+1)$
uniform $(0,1)$	logarithmic $(-\log x,\ 0 \le x \le 1)$
density of $U^{\frac{a}{a+1}}$ $(a>0)$	$(a+1)(1-x^{1/a})$, $0 \le x \le 1$
density of Z^a where $a>0$ and Z is a gamma $(b,1)$ random variable	distribution of Johnson, Tietjen and Beckman (1980)
(same with $a = 1/2$)	distribution of Johnson and Johnson (1978)

A similar theorem leads to the well-known method for generating random variates from the stable distribution (Chambers, Mallows and Stuck, 1976): if U,E are independent uniform $[0,1]$ and exponential random variables, and $g : [0,1] \to [0,\infty)$ is given, then $E/g(U)$ has density

$$f(x) = \int_0^1 g(u)\, e^{-xg(u)}\,du .$$

I. The moments of the distribution are given. Assume that we are given the moments $\mu_1,\mu_2,\mu_3,\ldots$ of the distribution. By the celebrated Carleman criterion (Shohat and Tamarkin, 1943, pp. 19), these moments determine the distribution in a unique way if

$$\sum_{n=1}^{\infty} \frac{1}{(\mu_{2n})^{1/2n}} = \infty .$$

How does one generate X with this distribution? If only a finite colection of moments is specified, the distribution is not uniquely determined of course. Still, one might ask for a simple procedure to generate a random variate X with any distribution having the said moments. No satisfactory answer is known at this moment.

REFERENCES

Ahrens, J.H. and Kohrt, K.D. (1981), Computer methods for efficient sampling from largely arbitrary statistical distributions, Computing, Vol. 26, pp. 19-31.

Ahrens, J.H. and Dieter, U. (1981), Generating gamma variates by comparison of probability densities. Manuscript, Institute of Statistics, Technical University, Graz, Austria.

Akima, H. (1970), A new method of interpolation and smooth curve fitting based on local procedures, Journal of the ACM, Vol. 17, pp. 589-602.

Bartels, R. (1981), Truncation bounds for infinite expansions for the stable distributions, Journal of Statistical Computation and Simulation, Vol. 12, pp. 293-302.

Bergström, H. (1952), On some expansions of stable distributions, Arkiv für Matematik, II, Vol. 18, pp. 375-378.

Best, D.J. (1978), Letter to the Editor, Applied Statistics, Vol. 27, pp. 181.

Bryson, M.C. and Johnson, M.E. (1981), Constructing and simulating multivariate distributions using Khintchine's theorem, Los Alamos Scientific Laboratory, Technical Report LA-UR 81-442.

Butler, E.L. (1970), Algorithm 370. General random number generator, Communications of the ACM, Vol. 13, pp. 49-52.

Chambers, J.M., Mallows, C.L. and Stuck, B.W. (1976), A method for simulating stable random variables, Journal of the American Statistical Association, Vol. 71, pp. 340-344.

Devroye, L. (1981a), The series method for random variate generation and its applications, to appear.

Devroye, L. (1981b), On the computer generation of random variables with a given characteristic function, Computers and Mathematics with Applications, Vol. 7, pp. .

Devroye, L. (1981c), A note on approximations in random variate generation, Journal of Statistical Computation and Simulation, Vol. 12, pp. .

Fox, B.L. (1980), Monotonicity, extremal correlations, and synchronization: implications for nonuniform random numbers, Département d'Informatique, Université de Montréal, Montreal, Canada.

Guerra, V.O., Tapia, R.A. and Thompson, J.R., A random number generator for continuous random variables based on an interpolation procedure of Akima, Proceedings of the 1978 Winter Simulation Conference, pp. 228-230.

Johnson, M.E. and Johnson, M.M. (1978), A new probability distribution with applications in Monte Carlo studies, Los Alamos Scientific Laboratory, Technical Report LA-7095-MS.

Johnson, M.E., Tietjen, G.L. and Beckman, R.J. (1980), A new family of probability distributions with applications to Monte Carlo studies, Journal of the American Statistical Association, Vol. 75, pp. 276-279.

Kinderman, A.J. and Monahan, J.T. (1977), Computer generation of random variables using the ratio of uniform deviates, ACM Transactions on Mathematical Software, Vol. 3, pp. 257-260.

Lewis, P.A.W. and Shedler, G.S. (1979), Simulation of nonhomogeneous Poisson processes by thinning, Naval Research Logistics Quarterly, Vol. 26 , pp. 403-413.

Lukacs, E. (1970), Characteristic Functions, Griffin, London.

Lux, I. (1978), A special method to sample some probability density functions, Computing, Vol. 20, pp. 183-188.

Marsaglia, G. (1977), The squeeze method for generating gamma variates, Computers and Mathematics with Applications, Vol. 3, pp. 321-325.

Mikhailov, G.A. (1965), On modelling random variables for one class of distribution laws, Theory of Probability and its Applications, Vol. 10, pp. 681-682.

Schmeiser, B.W. (1980), Random variate generation: a survey, Proceedings of the 1980 Winter Simulation Conference, Orlando, Florida.

Shohat, J.A. and Tamarkin, J.D. (1943), The problem of moments, Math. Survey No. 1, American Mathematical Society, New York.

Vaduva, I. (1977), On computer generation of gamma random variables by rejection and composition procedures, Mathematische Operationsforschung und Statistik, Series Statistics, Vol. 8, pp. 545-576.

Yuen, C. (1981), The inversion method in random variate generation, M.Sc. Thesis, School of Computer Science, McGill University, Montreal, Canada (in preparation).