Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

USING LATTICE RULES FOR VARIANCE REDUCTION IN SIMULATION

Christiane Lemieux

Department of Mathematics and Statistics
University of Calgary, 2500 University Drive N.W.
Calgary, AB, T2N 1N4, CANADA

Pierre L'Ecuyer

Département d'Informatique et de
Recherche Opérationnelle
Université de Montréal, C.P. 6128, Succ. Centre-Ville
Montréal, QC, H3C 3J7, CANADA

ABSTRACT

Quasi-Monte Carlo methods are designed to improve upon the Monte Carlo method for multidimensional numerical integration by using a more regularly distributed point set than the i.i.d. sample associated with Monte Carlo. Lattice rules are one family of quasi-Monte Carlo methods, originally proposed by Korobov in 1959. In this paper, we explain how randomized lattice rules can be used to construct efficient estimators for typical simulation problems, and we give several numerical examples. We are interested in two main aspects: Studying the variance of these estimators and finding which properties of the lattice rules should be considered when defining a selection criterion to rate and choose them. Our numerical results for three different problems illustrate how this methodology typically improves upon the usual Monte Carlo simulation method.

1 INTRODUCTION

Usually, in simulation problems, the goal is to estimate the expectation of a measure of performance defined over a stochastic system. The random input in the simulation program generally comes from a pseudorandom generator that output numbers between 0 and 1. These numbers are assumed to imitate a sequence of i.i.d. uniform random variables between 0 and 1, that are then transformed to follow the probability distributions specified by the model.

Thus, we can say that the goal of the simulation is to estimate an integral of the form

$$\mu = \int_{[0,1)^s} f(\mathbf{u}) d\mathbf{u}, \tag{1}$$

where f is a real-valued function that can be seen as a "black box" taking as an input a sequence of s numbers between 0 and 1, represented by the point $\mathbf{u} \in [0,1)^s$, and outputting an observation $f(\mathbf{u})$ of the measure of performance we are interested in. From this point of view, each simulation can be represented by a point $\mathbf{u}$ in $[0,1)^s$. Moreover, the method that consists in doing n independent simulations (where the ith simulation uses a random and uniform point $\mathbf{u}_i$), and then taking

$$\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{u}_i) \tag{2}$$

as an estimator of μ corresponds to using the Monte Carlo (MC) method for estimating the integral μ.

The MC method has some features that are important in the simulation context. The estimator (2) for μ is unbiased since each $\mathbf{u}_i$ has the uniform distribution over $[0,1)^s$ and thus $\mathrm{E}(f(\mathbf{u}_i)) = \mu$. The variance of (2) is equal to σ^2/n, where $\sigma^2 = \int_{[0,1)^s} f^2(\mathbf{u}) d\mathbf{u} - \mu^2$ is the variance of $f(\mathbf{U})$ when $\mathbf{U}$ is uniform over $[0,1)^s$, and can be estimated easily. In addition, confidence intervals for μ can be computed via the central limit theorem.

On the other hand, it would make sense to take advantage of more accurate methods to estimate (1) than what is achieved by MC. A natural way to improve upon MC is to try to sample the unit hypercube $[0,1)^s$ in a more regular way than by taking a set of n i.i.d. points $\mathbf{u}_i$ that have the uniform distribution over $[0,1)^s$. This improved (and deterministic) sampling scheme roughly defines *low-discrepancy point sets*, which are at the basis of *quasi-Monte Carlo* (QMC) methods. Among these methods we have the lattice rules, which are based on point sets in $[0,1)^s$ that have a lattice structure, i.e., which are closed under addition and subtraction modulo 1. We refer the reader to Niederreiter (1992) for a detailed overview of QMC methods.

Once we have a low-discrepancy point set, we can feed the simulation program with it instead of using a pseudorandom generator. A disadvantage of this approach is that the use of a deterministic source means that the unbiasedness and the usual error estimation methods are lost. However, this problem can be easily fixed by randomizing in an appropriate way the low-discrepancy point sets on which the QMC methods are based. We then have a method that can compete with MC to run our simulations.

In this paper, we recall how this methodology can be easily implemented with randomized lattice rules and summarize the theoretical results underlying this method. We give simulation examples where this method is tested against the MC method. Our goal is to show that randomized lattice rules can be used as a variance reduction technique that can be added to the toolbox of the simulation practitioner. More precisely, we explain in Section 2 how to construct a special case of randomized lattice rules and give results on the variance of the estimators based on this method. In Section 3, we give the definition of a selection criterion for lattice rules and argue why it should provide estimators with reduced variance for many problems. We provide three simulation examples, in Section 4, where this methodology is applied successfully.

2 RANDOMIZED LATTICE RULES

To construct the low-discrepancy point sets needed for the QMC method, we use *Korobov lattice rules* (Korobov 1959), which are point sets of the form

$$P_n = \left\{ \frac{(i-1)}{n}(1, a, \ldots, a^{s-1}) \bmod 1,\ i = 1, \ldots, n \right\}, \tag{3}$$

where n is the number of points in P_n, a is an integer between 1 and $n-1$ and the modulo 1 operation is done component-wise. We discuss in the next section how the integer a can be chosen. It can be shown that this P_n is the intersection of an *integration lattice* L with the unit hypercube $[0,1)^s$, an integration lattice being a discrete subset of $\mathbb{R}^s$ closed under addition and subtraction and containing the set of integer vectors $\mathbb{Z}^s$ as a subset. This property of P_n is the one defining lattice rules in general (see, e.g., Sloan and Joe 1994).

As we discussed in the introduction, it is convenient to randomize low-discrepancy point sets for the purpose of error estimation. For lattice rules, a simple way of doing this has been proposed by Cranley and Patterson (1976). Their method requires to generate randomly and uniformly a vector Δ in $[0,1)^s$, which is then used to translate all the points in P_n modulo 1. This provides the estimator

$$\hat{\mu}_{\mathrm{LR}} = \frac{1}{n}\sum_{i=1}^{n} f((\mathbf{u}_i + \Delta) \bmod 1)$$

for μ. We assume here that the set $P_n = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ defining $\hat{\mu}_{\mathrm{LR}}$ is a Korobov lattice rule. To estimate the variance of $\hat{\mu}_{\mathrm{LR}}$, one just needs to repeat this procedure with m independent random shifts, thus obtaining m i.i.d. copies of $\hat{\mu}_{\mathrm{LR}}$. This is what we do in Section 4 to compare empirically the variance of the randomized lattice rule estimators with the MC estimators.

We now recall some results on the theoretical variance of $\hat{\mu}_{\mathrm{LR}}$ and compare it with the variance of the MC estimator $\hat{\mu}_{\mathrm{MC}}$ based on the same number of points n. An important concept required to analyze the variance of $\hat{\mu}_{\mathrm{LR}}$ is that of the *dual lattice* to the lattice L such that $P_n = L \cap [0,1)^s$. It is defined as the set $L^\perp = \{\mathbf{h} \in \mathbb{Z}^s : \mathbf{h}\cdot\mathbf{u}_i \in \mathbb{Z},\ i = 1, \ldots, n\}$. It is a subset of L and it has the property of containing n times less points than $\mathbb{Z}^s$. We can show (L'Ecuyer and Lemieux 2000):

Proposition 1 *If f is square-integrable (i.e., $\sigma^2 = \mathrm{Var}(f) < \infty$), then $\mathrm{E}(\hat{\mu}_{\mathrm{LR}}) = \mu$ and*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = \sum_{\mathbf{0}\neq\mathbf{h}\in L^\perp} |\hat{f}(\mathbf{h})|^2,$$

where $\hat{f}(\mathbf{h}) = \int_{[0,1)^s} f(\mathbf{u}) e^{2\pi\sqrt{-1}\mathbf{h}\cdot\mathbf{u}} d\mathbf{u}$ is the Fourier coefficient of f evaluated in $\mathbf{h}$.

Proposition 1 shows that $\hat{\mu}_{\mathrm{LR}}$ is an unbiased estimator whose variance is smaller than the variance of the MC estimator $\hat{\mu}_{\mathrm{MC}}$ based on n points if and only if "on average" the squared Fourier coefficients $|\hat{f}(\mathbf{h})|^2$ are smaller on $L^\perp$ than on $\mathbb{Z}^s$. To see this, remember that $\mathbb{Z}^s$ is n times as "dense" as $L^\perp$ and observe that

$$\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}) = \frac{1}{n}\sum_{\mathbf{0}\neq\mathbf{h}\in\mathbb{Z}^s} |\hat{f}(\mathbf{h})|^2$$

(L'Ecuyer and Lemieux 2000). This result warns us that there exist worst-case functions for which $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = \sigma^2 = n\,\mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$, which means that stronger assumptions on f must be made to guarantee that $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \le \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$. This is what we do in the following proposition (Lemieux 2000) where the special case of linear functions (i.e., multivariate polynomials of degree 1) is considered:

Proposition 2 *If f is a square-integrable linear function over $[0,1)^s$ and if $\gcd(n,a) = 1$, then*

$$\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) = \frac{1}{n}\mathrm{Var}(\hat{\mu}_{\mathrm{MC}}).$$

Of course, analytical formulas for μ can be derived when f is a linear function. Nevertheless, this result is interesting because it tells us that if f is a complicated function that can be well approximated by a linear function, then an important part of f will benefit greatly from the use of a lattice rule instead of MC, and we should expect $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}}) \ll \mathrm{Var}(\hat{\mu}_{\mathrm{MC}})$ in this case. An example of this behavior can be found in L'Ecuyer and Lemieux (2000), Section 10.1, where the convergence rate of $1/n^2$ for $\mathrm{Var}(\hat{\mu}_{\mathrm{LR}})$ implied by the preceding proposition is observed empirically for a nonlinear function.

Another way to support the idea that lattice rules can do better than MC is to look at the average variance for a certain class of lattice rules and compare it to the variance

of $\hat{\mu}_{\mathrm{MC}}$. We do this in the following proposition (Lemieux 2000), where $\mathcal{A}$ represents the set of all *rank-1 lattice rules* that have n points. A rank-1 lattice rule is obtained by replacing the vector $(1, a, \ldots, a^{s-1})$ in (3) by any vector in $[1, \ldots, n-1]^s$.

Proposition 3 *If f is square-integrable and n is a prime number, then*

$$\frac{1}{|\mathcal{A}|}\sum_{i=1}^{|\mathcal{A}|}\mathrm{Var}\,(\hat{\mu}_{\mathrm{LR,i}}) \leq \frac{n}{n-1}\mathrm{Var}\,(\hat{\mu}_{\mathrm{MC}}) + \frac{n-2}{n-1}\mathrm{Var}\,(\hat{\mu}_{\mathrm{grid}}),$$

where $\hat{\mu}_{\mathrm{LR,i}}$ is based on the ith lattice rule in $\mathcal{A}$,

$$\hat{\mu}_{\mathrm{grid}} = \frac{1}{n^s}\sum_{m_1,\ldots,m_s=0}^{n-1} f\left(\left(\left(\frac{m_1,\ldots,m_s}{n}\right)+\Delta\right) \bmod 1\right),$$

and Δ is uniformly distributed over $[0,1)^s$.

Notice that the estimator $\hat{\mu}_{\mathrm{grid}}$ is based on n^s evaluations of f in comparison with $\hat{\mu}_{\mathrm{MC}}$ (and $\hat{\mu}_{\mathrm{LR,i}}$) where only n evaluations are done. This suggests that $\mathrm{Var}\,(\hat{\mu}_{\mathrm{grid}})$ should be much smaller than $\mathrm{Var}\,(\hat{\mu}_{\mathrm{MC}})$ for most functions f. If this is true, Proposition 3 means that on the average, lattice rules cannot do much worse than the MC estimator. Moreover, since "bad" lattice rules can be found among the set $\mathcal{A}$, i.e., lattice rules that yield high variance estimators for most functions (e.g., the Korobov lattice rule based on $a = 1$), one can infer that there must be some "good" lattice rules in $\mathcal{A}$, i.e., lattice rules that should do well on most functions. In the next section, we discuss how to find these "good" lattice rules.

3 SELECTION CRITERION

The goal here is to find, for a given number of points n, a lattice rule that should provide estimators reducing the variance compared to MC for most functions encountered in practice. Note that in the case of Korobov rules, choosing a lattice means choosing a generator a in (3). Thus, one possible approach to find a "good" Korobov rule is to define a figure of merit for P_n that can be computed reasonably fast, and then to search over a prespecified set of generators a to find the one whose associated set P_n optimizes the figure of merit. In order to reach our goal, the figure of merit should also imitate (in some sense) the variance of $\hat{\mu}_{\mathrm{LR}}$ for a large class of functions. Hence, we must make some assumptions about the functions we are likely to encounter in practice. As a first approximation, we assume that in most cases, f is a function that can be well approximated by a sum of low-dimensional functions. Let us explain what we mean by that, and we will see later how this is translated into the definition of our selection criterion.

We use the ANOVA functional decomposition of f (Hoeffding 1948; Efron and Stein 1981; Owen 1998), which writes any square-integrable function f as a sum

$$f(\mathbf{u}) = \sum_{I\subseteq\{1,\ldots,s\}} f_I(\mathbf{u}),$$

where $f_I(\mathbf{u})$ is the part of f that depends on the variables u_j whose index j is in I. These components $f_I(\mathbf{u})$ are defined so that $\int_{[0,1)^s} f_I(\mathbf{u})d\mathbf{u} = 0$ if I is non-empty and $\int_{[0,1)^s} f_\emptyset(\mathbf{u})d\mathbf{u} = \mu$. This decomposition is orthogonal, i.e., $\int_{[0,1)^s} f_I(\mathbf{u})f_J(\mathbf{u})d\mathbf{u} = 0$ if $I \neq J$, and thus

$$\sigma^2 = \mathrm{Var}\,(f) = \sum_{I\subseteq\{1,\ldots,s\}} \sigma_I^2,$$

where $\sigma_I^2 := \mathrm{Var}\,(f_I)$. The function f is said to have an *effective dimension of d in the superposition sense* (Caflish, Morokoff, and Owen 1997) if $\sum_{I:|I|\leq d}\sigma_I^2 \approx \sigma^2$. When this holds, it suffices to make sure that the d-dimensional structure of P_n is good in order to define an estimator $\hat{\mu}_{\mathrm{LR}}$ having a small variance, i.e., one must verify that for each subset I such that $|I| \leq d$, the projection $P_n(I)$ of P_n over the subspace of $[0,1)^s$ indexed by the coordinates in I is well distributed. This means that even if a function is defined over a very large dimensional space, point sets P_n providing better estimators than with MC can still be constructed if f has a low effective dimension.

The second hypothesis that we make about f is to assume that its most important squared Fourier coefficients $|\hat{f}(\mathbf{h})|^2$ are those associated with small vectors $\mathbf{h}$. Using Proposition 1, this means (roughly) that the fewer short vectors can be found in the dual lattice $L^\perp$ of a rule P_n, the smaller is the variance of the estimator $\hat{\mu}_{\mathrm{LR}}$ provided by P_n. One way to verify that P_n has this property is to apply the *spectral test* (Coveyou and MacPherson 1967, L'Ecuyer and Couture 1997), i.e., to compute the quantity $\ell_s = \min_{\mathbf{0}\neq\mathbf{h}\in L^\perp}\|\mathbf{h}\|_2$, and to make sure that ℓ_s is as large as possible. In addition, the spectral test can be applied to any projection $P_n(I)$, i.e, one can define $\ell_I = \min_{\mathbf{0}\neq\mathbf{h}\in L_I^\perp}\|\mathbf{h}\|_2$, where $L_I^\perp$ is the dual lattice to the lattice L_I such that $P_n(I) = L_I \cap [0,1)^s$. For more on the link between the spectral test and lattice rules, we refer the reader to Entacher, Hellekalek, and L'Ecuyer (2000).

Combining this test to the preceding arguments about the choice of subsets I for which the projections $P_n(I)$ should be inspected leads to the following selection criterion (L'Ecuyer and Lemieux 2000):

$$M_{t_1,\ldots,t_d} = \min\left[\min_{2\leq j\leq t_1}\frac{\ell_j}{\ell_j^*(n)}, \min_{2\leq j\leq d,\, I\in S(j,t_j)}\frac{\ell_I}{\ell_{|I|}^*(n)}\right],$$

where $S(j,t_j) = \{I = \{i_1,\ldots,i_j\} : 1 = i_1 < \cdots < i_j \leq t_j\}$, and $\ell_j^*(n)$ is an upper bound on ℓ_I for $|I| = j$ that can be found in L'Ecuyer (1999). We thus want $M_{t_1,\ldots,t_d}$ to be as close to 1 as possible. For more on this criterion and its links with criteria previously used to choose *linear congruential generators*, see L'Ecuyer and Lemieux (2000).

We give in Table 1 the best Korobov rules we found using the criterion $M_{32,24,12,8}$. The search has been done for two different values of n (which are prime numbers close to some power of 2), and over all integers a that are primitive element modulo n. Such generators a lead to especially simple implementations for the point sets P_n (Entacher, Hellekalek, and L'Ecuyer 2000; L'Ecuyer 1999). The rules in this table are the ones we use in the next section for our numerical examples. The best rules for other values of n and with respect to other criteria can be found in Lemieux (2000) and L'Ecuyer and Lemieux (2000).

Table 1: Best a's with respect to $M_{32,24,12,8}$

n	251	1021
a	184	76
$M_{32,24,12,8}$	0.31632	0.29344

4 EXAMPLES

In this section, we present three simulation problems on which we successfully use shifted lattice rules (LR) to construct estimators having a smaller empirical variance than the MC estimators based on the same total number of replications. Since the LR estimator is faster to compute, it means that the *efficiency* is improved by the LR method. For all problems, we always use inversion to generate the non-uniform random variables.

4.1 The Bank Example

This example is taken from Bratley, Fox, and Schrage (1987). Consider a bank that opens at 10:00 and closes at 15:00 (or 3 p.m.). Customers arrive from 9:45 to 15:00 according to a non-stationary Poisson process having a rate of 1 customer per minute from 11:00 to 14:00, and 0.5 otherwise. The number of tellers varies between 1 and 3, with a probability $q_3 = 0.85$ of being 3, $q_2 = 0.15$ of being 2, and $q_1 = 0.05$ to be 1. If a customer arrives and finds a waiting queue of ten customers, he *balks* and goes away. If only $q < 10$ customers are waiting, he balks with probability $(q-5)/10$ if $q > 5$ and stays otherwise. All customers that have arrived before 15:00 and did not balk are served. The service times are independent random variables having a gamma distribution with parameters $\alpha = 2$ and $\lambda = 1$. We are interested in estimating the expected number of customers served in one day of operation of this bank.

In the context of (1), the dimension s of this problem is random and related to the number C of customers arriving in a day, which has no upper bound, i.e., formally, $s = \infty$. At least one random variable is required per customer for the interarrival time, with possibly a second one to generate the service time and a third one when a decision of balking or not must be taken. An additional random variable is required to generate the number of tellers. Thus, the dimension s is between $C+1$ and $3C+1$. The function f in (1) is the transformation that maps a point of $[0,1)^s$ to an observation of the number of customers served during a day.

Many variance reduction techniques can be used for this problem (Bratley, Fox, and Schrage 1987). In our experiments, we use some of them and also look at different combinations. Let X denote the number of customers served in a day. The naive estimator for $\mathrm{E}(X)$ is X. With *indirect estimation* (ind), one uses instead $\mathrm{E}(C) - B$, where $\mathrm{E}(C)$ can be computed exactly and B is the number of customers that have balked. The random variable C can also be used as a *control variable* (cv) for this problem, i.e., one replaces X by $X + \beta(\mathrm{E}(C) - C)$, where β is a coefficient to be estimated. When *stratification* is used, the total number of simulations N is allocated to the three different possible values for z, the number of tellers. Denote by N_z the number of simulations done with z tellers. Then one replaces X by $NX(0.85\mathbf{1}_{z=1}/N_1 + 0.15\mathbf{1}_{z=2}/N_2 + 0.05\mathbf{1}_{z=3}/N_3)$. The value of z can either be generated randomly at the beginning of the simulation or can be fixed in a deterministic way (and then the dimension s decreases by 1). Taking $N_z = Nq_z$ is called *proportional allocation*, and choosing the N_z's in a way that minimizes the variance is called *optimal allocation* (str-op). In the latter case, the optimal values of the N_z must be estimated through pilot runs and depend on the variance of $X\mathbf{1}_{z=j}$ for $j = 1,2,3$. For a general review of these variance reduction techniques, see, e.g., L'Ecuyer (1994).

In Table 2, we give the variance reduction factors obtained when using MC or LR in combination with the above techniques. Two different numbers of points n for LR are considered, and in each case 100 independent random shifts are used to estimate the variance of the LR estimators. The MC estimators are based on $100n$ independent simulations. The coefficient β required to apply the control variable technique is estimated with 600 pilot runs, and so are the numbers N_z defining the estimator using stratification with optimal allocation. With LR, N equals n in the definition of the stratification estimator (and we get 100 copies of this estimator), whereas N equals $100n$ with MC. It seems that the variance reduction factors brought by LR do not increase with the number of points n for this example. This can happen when the dimension is large and it is the case here, as we computed the average dimension to be around 500 for this problem. When no stratification is done, LR reduces the variance by significant factors. It also brings

Table 2: Variance Reduction Factors for the Bank

	$n = 251$		$n = 1021$	
	MC	LR	MC	LR
naive	1.0	6.8	1.0	4.7
ind	1.4	20	1.4	19
cv	1.4	23	1.4	21
str-op	3.1	5.7	3.0	5.8
str-op + ind	46	55	44	56
str-op + ind + cv	71	109	70	114

some additional improvement after the stratification, but by factors not larger than 2. We think that this is still interesting, especially if we take into account the fact that LR is faster than MC and thus, that efficiency is clearly improved by using LR for this problem.

4.2 A Time-Sharing Computer System

This example is taken from Law and Kelton (2000), Section 2.5. Consider an old-fashioned time-sharing computer system with T independent and identical terminals using the same CPU. The user of each terminal submits a task to the CPU, waits for the answer, and then thinks for a random length of time before submitting a new task, and so on. The thinking time follows an exponential distribution of mean ν, and the CPU time required by a task has a Weibull distribution of parameters α and λ. The waiting tasks form a queue for the CPU, and are served in a "round-robin" fashion with quantum of size q. When a task gets the CPU, if less than q seconds are required to finish it, it occupies the CPU until its completion. Otherwise, the task occupies the CPU for q seconds and goes back to the end of the waiting line. In both cases, h seconds of overhead are needed to remove the task from the CPU and, if required, put a new one.

The *response time* of a task is defined as the elapsed time between the instant when it leaves the terminal and the instant when it leaves the CPU upon completion. We want to estimate the expected response time in steady state. This is achieved by simulating this system until N tasks are completed, assuming that initially, all T terminals are in the thinking state. The initial bias due to this hypothesis is decreased by simulating N_0 tasks until completion before starting to accumulate the statistics. See L'Ecuyer (2000) for more details on the simulation of this system.

When translating this problem in terms of (1), we get a large dimensional problem because the measure of performance of interest is a long-term average and thus each simulation must run for a long time. More precisely, the dimension s is at least equal to $2N$ since each of the N tasks is defined by two random variables (thinking and processing times). It is smaller or equal to $2(N+T-1)$ because when the Nth task ends, at most two random variables have been required to simulate the $T-1$ other terminals that are either thinking or waiting for the end of a task. Thus, the dimension is random for this problem too, but bounded, and one can take, e.g., $s = 2(N+T-1)$. The function f transforms a point of $[0,1)^s$ into an observation of the average response time for the last $N - N_0$ tasks simulated. This estimator is biased, which means that the integral $\int_{[0,1)^s} f(\mathbf{u})d\mathbf{u}$ is not equal to the expectation we are looking for, but it converges to it as $N \to \infty$.

In Table 3, we give the variance reduction factors obtained by the LR estimators on this problem, for two different number of points n, and two different values for the quantum size q. The other parameters are set to $h = 0.001$, $\nu = 5.0, \alpha = 0.5$, and $\lambda = 1$. We chose $N = 1100$ and $N_0 = 100$, which means that the problem has a dimension of at least 2200. We used 100 random shifts to estimate the variance of the LR estimators, and the MC estimators are based on $100n$ independent simulations. For this problem, LR reduces the variance by factors of at least 13 in comparison with MC.

Table 3: Variance Reduction Factors for the Time-Sharing System

	$q = 0.1$	$q = 0.2$
$n = 251$	13	13
$n = 1021$	36	29

We considered the possibility of using common random numbers (CRN) in combination with LR in order to estimate the difference between the expected response time for the two values of q given in Table 3. Using CRN with LR simply means that we use the same random shifts to estimate both quantities. When we tried it, we obtained better results for the CRN estimator than for the LR+CRN estimator. However, for other sets of parameters leading to more important differences in the two configurations compared, the combination LR+CRN did better than CRN. Note that configurations with say, different mean processing times, present a bigger challenge for CRN because the synchronization between the two systems is unlikely to be achieved for a given simulation, whereas configurations that only differ by their quantum size as those studied in Table 3 are more likely to be synchronized. In the next subsection, we will see another example where the two systems are perfectly synchronized for any simulation and where LR+CRN does significantly better than CRN alone.

4.3 Estimating Sensitivities (Derivatives) of Asian Option Prices

The third problem is taken from Broadie and Glasserman (1996), and consists in doing sensitivity analysis in the context of option pricing. An *option* is a financial contract whose value depends on the price of some *underlying asset*.

We are interested in estimating the sensitivity of an option's fair price (or value) with respect to different parameters. We thus need to estimate the derivative of the value of this option with respect to each of the chosen parameters. In this paper, the word "derivative" is used in the ordinary mathematical sense (i.e., rate of variation), not to denote a class of financial products as commonly done in the world of finance. In the case we describe, there is no analytical expression for these derivatives and simulation is an appropriate tool to estimate them.

The contract considered is an *Asian option* on a dividend-paying asset that follows a geometric Brownian motion. Denote by $S(t)$ the price of this asset at time t. Under the *risk-neutral measure*, we have

$$dS(t) = (r - \delta)S(t)dt + \sigma S(t)dB(t),$$

where $B(\cdot)$ is a standard Brownian motion, r is the risk-free interest rate, δ is the dividend rate, and $\sigma > 0$ is the volatility parameter. The option contract is assumed to have its expiration date at time T, and the current time is set to 0. Its payoff is defined as

$$C(T) = \max(0, \bar{S} - K),$$

where $\bar{S} = (1/s)\sum_{i=1}^{s} S(t_i)$, the t_i's are some prespecified observation dates for the price of the asset, and K is the *strike price*, a parameter specified in the contract. The *value* of the option at time 0 is

$$C(0) = \mathrm{E}(e^{-rT}C(T)),$$

where the expectation is taken under the risk-neutral measure. For more on option pricing, see Duffie (1996).

In L'Ecuyer and Lemieux (2000), we used lattice rules to reduce the variance of MC estimators of $C(0)$. Here, we are interested in estimating the partial derivatives of $C(0)$ with respect to the initial value $S(0)$ of the underlying asset and with respect to the volatility σ, i.e.,

$$\mu_d = \frac{\partial C(0)}{\partial S(0)} \text{ and } \mu_v = \frac{\partial C(0)}{\partial \sigma}.$$

These quantities are called *delta* and *vega*, respectively, in the world of computational finance.

We experiment here with two of the three methods presented in Broadie and Glasserman (1996) to estimate these derivatives; the *pathwise method* (or *infinitesimal perturbation analysis (IPA)*; see Glasserman (1991) for more on this method), and *resimulation* (or *finite differences with common random numbers (FDC)*). With the IPA method, the derivative estimators are obtained by interchanging the expectation and derivative operators. This yields the estimators (Broadie and Glasserman 1996):

$$\begin{aligned}
\hat{\mu}_d &= e^{-rT}\frac{\partial C(T)}{\partial S(0)} = e^{-rT}\mathbf{1}_{\bar{S}\ge K}\frac{\bar{S}}{S(0)} \\
\hat{\mu}_v &= e^{-rT}\frac{\partial C(T)}{\partial \sigma} \\
&= e^{-rT}\mathbf{1}_{\bar{S}\ge K}\frac{1}{s\sigma}\sum_{i=1}^{s} S(t_i)\left[\ln(S(t_i)/S(0))\right. \\
&\qquad \left. -(r-\delta+0.5\sigma^2)t_i\right].
\end{aligned}$$

It is shown in Broadie and Glasserman (1996) that $\hat{\mu}_d$ and $\hat{\mu}_v$ are unbiased estimators of μ_d and μ_v, respectively. The quantities that need to be simulated in order to compute $\hat{\mu}_d$ and $\hat{\mu}_v$ are the prices $S(t_i)$, which follow a lognormal distribution with parameters $\ln S(0) + (r - \delta - 0.5\sigma^2)\,t_i$ and $\sigma^2 t_i$. In the framework of (1), it means that for both derivatives, the dimension s is equal to the number of prices entering the mean value $\bar{S}$ (since we need to generate one normal random variable per observation $S(t_i)$), with the function f being the transformation that maps a point of $[0, 1)^s$ into an observation of $\hat{\mu}_d$ or $\hat{\mu}_v$.

With the FDC method, the partial derivative $\partial C(0)/\partial\theta$ (for some parameter θ) evaluated at $\theta = \theta_0$ is estimated as follows: In addition to a simulation done at $\theta = \theta_0$, a second one is done at $\theta = \theta_0 + h$, where h is a small quantity. If we denote by $e^{-rT}C(T, \theta)$ the discounted payoff obtained with the simulation done at θ, the estimator is defined as $(e^{-rT}C(T, \theta_0 + h) - e^{-rT}C(T, \theta_0))/h$. These two simulations at θ_0 and $\theta_0 + h$ are performed with *common random numbers* (L'Ecuyer and Perron 1994, Law and Kelton 2000, Broadie and Glasserman 1996). We thus have an s-dimensional integral: The integrand f in this case is the function that maps a point of $[0, 1)^s$ into an observation of $e^{-rT}(C(T, \theta_0 + h) - C(T, \theta_0))/h$. The FDC method yields *biased* estimators, i.e., $\int_{[0,1)^s} f(\mathbf{u})d\mathbf{u} \neq \partial C(0)/\partial\theta$, but the bias converges to 0 as $h \to 0$. Using the results of L'Ecuyer and Perron (1994), it can be shown that these FDC estimators have the same asymptotic convergence rates as the IPA estimators if $h \to 0$ fast enough when $n \to \infty$. We should therefore expect the two methods to give almost identical results when h is very small.

As pointed out in Broadie and Glasserman (1996), variance reduction techniques that apply to the estimation of the option's price can also be used for the purpose of derivative estimation. Following this, these authors have used the terminal underlying asset price $S(T)$ as a control variable for the different problems they considered. In the special case of Asian options, another possibility is to use the price of an Asian option based on the *geometric* average as a control variable (Kemna and Vorst 1990).

In Tables 4 and 5, we give the variance reduction factors obtained by the LR estimators in comparison to the MC approach for the quantities *delta* and *vega*, respectively, with IPA. We consider four different estimators: The *naive* one does not use any control variable, CV1 uses the final underlying asset price, CV2 uses the price of the option on the geometric average, and CV12 uses these two control variables together. As expected, the results obtained with the FDC method (not shown in the tables), with $h = 0.0001$, were practically identical to those obtained with IPA. This has also been observed by Broadie and Glasserman (1996). The variance of the LR estimators is estimated with 100 random shifts, and the MC estimators are based on $100n$ independent simulations, where $n = 1021$ is the number of points in the lattice rule.

Table 4: Variance Reduction Factors for Delta, with IPA

	MC			LR		
$S(0)$	90	100	110	90	100	110
naive	(1.7e-6)	(2.8e-6)	(1.3e-6)	4.8	7.8	9.2
CV1	(9.5e-7)	(1.1e-6)	(6.5e-7)	2.4	2.5	3.5
CV2	(7.2e-7)	(1.3e-6)	(7.1e-7)	3.5	2.1	3.3
CV12	(6.3e-7)	(1.1e-6)	(6.5e-7)	2.2	2.0	3.2

Table 5: Variance Reduction Factors for Vega, with IPA

	MC			LR		
$S(0)$	90	100	110	90	100	110
naive	(5.0e-3)	(6.2e-3)	(1.2e-2)	10	27	19
CV1	(2.7e-3)	(2.2e-3)	(4.5e-3)	2.4	3.3	3.9
CV2	(7.2e-4)	(8.9e-5)	(2.0e-3)	7.2	16	7.0
CV12	(6.5e-4)	(4.6e-5)	(8.7e-4)	2.2	0.99	1.3

Each column in Tables 4 and 5 corresponds to a different initial price $S(0)$. The numbers in parentheses in the MC columns represent the estimated variance of these estimators (the reduction factors are equal to 1 for all of them). We give these numbers to indicate how well each control variable performs in comparison with the naive approach. The numbers in the columns labeled by LR correspond to the variance reduction factors obtained with the LR estimator in comparison with the corresponding MC estimator (i.e., whose variance is on the same line, three columns to the left).

The parameters of the option are as in Broadie and Glasserman (1996): $r = 0.1$, $K = 100$, $\delta = 0.03$, $\sigma = 0.25$, $T = 0.2$ year and the average is taken over the last 30 days before the expiration. This means that the dimension of the problem is $s = 30$.

Reduction factors as large as 27 are observed by using LR instead of MC for *vega*. As it was noted for the estimation of Asian option prices in L'Ecuyer and Lemieux (2000), the LR estimator brings more improvement with the naive method than when a control variable is used. Here, we can also say that the improvement by LR is more important when only one control variable is used rather than the combination CV12. CV2 works better than CV1 for the estimation of *vega* whereas for *delta*, the two control variables give similar results. CV2 works especially well for *vega* when $S(0) \approx K$, and the added benefit of LR is also more important in this case. The combination CV12 improves upon each of CV1 and CV2 for *vega*, and has about the same variance as the best between CV1 and CV2 for *delta*.

Given the reduction factors obtained by LR for this problem and knowing that the FDC method gave essentially the same results as those shown in Tables 4 and 5, we can say that the combination of CRN (which is used in FDC) with LR is successful for this application.

In the context where sensitivities need to be estimated for optimization purposes, e.g., to feed a stochastic approximation algorithm, a method like LR that leads to estimators with smaller variance without increasing the computation time would lead to a faster convergence towards the optimum, as explained in L'Ecuyer and Yin (1998).

ACKNOWLEDGMENTS

This work has been supported by NSERC-Canada grant No. ODGP0110050 and FCAR Grant No. 00ER3218 to the second author, and via an FCAR-Québec postdoctoral scholarship to the first author.

REFERENCES

Bratley, P., B. L. Fox, and L. E. Schrage. 1987. *A guide to aimulation*. Second ed. New York: Springer-Verlag.

Broadie, M. and P. Glasserman. 1996. Estimating security price derivatives using simulation. *Management Science*, 42:269–285.

Caflish, R. E., W. Morokoff, and A. Owen. 1997. Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. *The Journal of Computational Finance*, 1(1):27–46.

Coveyou, R. R. and R. D. MacPherson. 1967. Fourier analysis of uniform random number generators. *Journal of the ACM*, 14:100–119.

Cranley, R. and T. N. L. Patterson. 1976. Randomization of number theoretic methods for multiple integration. *SIAM Journal on Numerical Analysis*, 13(6):904–914.

Duffie, D. 1996. *Dynamic asset pricing theory*. Second ed. Princeton University Press.

Efron, B and C. Stein. 1981. The jackknife estimator of variance. *Annals of Statistics*, 9:586–596.

Entacher, K., P. Hellekalek, and P. L'Ecuyer. 2000. Quasi-Monte Carlo node sets from linear congruential generators. In *Monte Carlo and Quasi-Monte Carlo Methods*

1998, ed. H. Niederreiter and J. Spanier, 188–198, Berlin. Springer.

Glasserman, P. 1991. *Gradient estimation via perturbation analysis*. Norwell: Kluwer Academic.

Hoeffding, W. 1948. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19:293–325.

Kemna, A. G. Z. and A. C. F. Vorst. 1990. A pricing method for options based on average asset values. *Journal of Banking and Finance*, 14:113–129.

Korobov, N. M. 1959. The approximate computation of multiple integrals. *Dokl. Akad. Nauk SSSR*, 124:1207–1210. In Russian.

Law, A. M. and W. D. Kelton. 2000. *Simulation modeling and analysis*. Third ed. New York: McGraw-Hill.

L'Ecuyer, P. 1994. Efficiency improvement via variance reduction. In *Proceedings of the 1994 Winter Simulation Conference*, 122–132. IEEE Press.

L'Ecuyer, P. 1999. Tables of linear congruential generators of different sizes and good lattice structure. *Mathematics of Computation*, 68(225):249–260.

L'Ecuyer, P. 2000. *SIMOD: Définition fonctionnelle et guide d'utilisation*. Université de Montréal.

L'Ecuyer, P. and R. Couture. 1997. An Implementation of the Lattice and Spectral Tests for Multiple Recursive Linear Random Number Generators. *INFORMS Journal on Computing*, 9(2):206–217.

L'Ecuyer, P. and C. Lemieux. 2000. Variance reduction via lattice rules. *Management Science*. To appear.

L'Ecuyer, P. and G. Perron. 1994. On the convergence cates of IPA and FDC derivative estimators. *Operations Research*, 42(4):643–656.

L'Ecuyer, P. and G. Yin. 1998. Budget-Dependent Convergence Rate for Stochastic Approximation. *SIAM Journal on Optimization*, 8(1):217–247.

Lemieux, C. 2000. L'utilisation de règles de réseau en simulation comme technique de réduction de la variance. PhD thesis, Université de Montréal.

Niederreiter, H. 1992. *Random Number Generation and Quasi-Monte Carlo Methods*. Volume 63 of *SIAM CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia: SIAM.

Owen, A. B. 1998. Latin supercube sampling for very high-dimensional simulations. *ACM Transactions of Modeling and Computer Simulation*, 8(1):71–102.

Sloan, I. H. and S. Joe. 1994. *Lattice Methods for Multiple Integration*. Oxford: Clarendon Press.

AUTHOR BIOGRAPHIES

CHRISTIANE LEMIEUX is an assistant professor at the Department of Mathematics and Statistics at the University of Calgary. She works on efficiency improvement by quasi-Monte Carlo methods. She completed her Ph.D. in computer science at the Université de Montréal in May 2000. Her web address is `<www.math.ucalgary.ca/~lemieux>`.

PIERRE L'ECUYER is a professor in the "Département d'Informatique et de Recherche Opérationnelle", at the University of Montreal. He received a Ph.D. in operations research in 1983, from the University of Montréal. He obtained the *E. W. R. Steacie* grant from the Natural Sciences and Engineering Research Council of Canada for the period 1995–97. His main research interests are random number generation, efficiency improvement via variance reduction, sensitivity analysis and optimization of discrete-event stochastic systems, and discrete-event simulation in general. He is an Area Editor for the *ACM Transactions on Modeling and Computer Simulation*. More details at: `<www.iro.umontreal.ca/~lecuyer>`, where several of his recent research articles are available on-line.