*Statistical Science*
2005, Vol. 20, No. 1, 1–31
DOI 10.1214/088342304000000468


# Control Variates for Quasi-Monte Carlo

**Fred J. Hickernell, Christiane Lemieux and Art B. Owen**

*Abstract.* Quasi-Monte Carlo (QMC) methods have begun to displace ordinary Monte Carlo (MC) methods in many practical problems. It is natural and obvious to combine QMC methods with traditional variance reduction techniques used in MC sampling, such as control variates. There can, however, be some surprises. The optimal control variate coefficient for QMC methods is not in general the same as for MC. Using the MC formula for the control variate coefficient can worsen the performance of QMC methods. A good control variate in QMC is not necessarily one that correlates with the target integrand. Instead, certain high frequency parts or derivatives of the control variate should correlate with the corresponding quantities of the target. We present strategies for applying control variate coefficients with QMC and illustrate the method on a 16-dimensional integral from computational finance. We also include a survey of QMC aimed at a statistical readership.

*Key words and phrases:* Digital nets, lattice rules, low discrepancy methods, stratification, variance reduction.

## 1. INTRODUCTION

We consider here the problem of computing the integral $I$ of a function $f$ defined on the $s$-dimensional unit cube $[0,1)^s$:

$$I = \int f(x)\,dx. \tag{1}$$

Here and elsewhere, integrals without explicit ranges are understood to be over $[0,1)^s$. It is very common in applications that the integrals arise in a form other than (1), but are translated into that form.

The basic form of Monte Carlo (MC) sampling simulates independent random vectors $X_1, \ldots, X_n$ that have the $U[0,1)^s$ distribution. Then the MC estimate of $I$ is

$$\hat{I} = \hat{I}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i). \tag{2}$$

*Fred J. Hickernell is Professor and Chair, Department of Applied Mathematics, Illinois Institute of Technology, Room 208, Engineering 1 Building, 10 West 32nd Street Chicago, Illinois 60616, USA. Christiane Lemieux is Assistant Professor, Department of Mathematics and Statistics, and Department of Computer Science, University of Calgary, Calgary, Alberta, Canada T2N 1N4. Art B. Owen is Professor, Department of Statistics, Stanford University, Stanford, California 94305, USA.*

It is elementary that $E(\hat{I}) = I$ and if we suppose that the variance of the integrand $\sigma^2 = \int (f(x) - I)^2\,dx$ satisfies $0 < \sigma^2 < \infty$, then we can write the mean square error as

$$E\big((\hat{I} - I)^2\big) = \mathrm{Var}(\hat{I}) = \sigma^2/n.$$

Many techniques have been developed to improve the accuracy of MC methods. Two such techniques are quasi-Monte Carlo (QMC) sampling, which can be likened to a very intense multiple stratification, and the classical method of control variates. To employ both of these methods at once is an obvious idea and one that is easy to implement. Less obvious is that the control variate strategy for MC applied to QMC points can reduce the accuracy of the QMC method. The optimal control variate coefficient depends on the sampling strategy and even on the sample size. In MC a good control variate is one that correlates with the integrand. In QMC methods it can be better to have some other aspect of the control variate, such as a derivative or a sum of high frequency Fourier components, correlate with the corresponding aspect of the target integrand.

Monte Carlo variance reductions for QMC have been studied earlier. Spanier and Maize (1994) discussed combinations of importance sampling with QMC and mentioned some early work by Chelson (1976).

While our main contribution is on the interplay between QMC and control variates, we also present a brief survey of QMC methods. This survey appears as Section 2. It presents some historical motivations of QMC, and the main techniques in use today, for readers with a statistical background. Section 3 records some basic results on control variates that we use. Section 4 describes how estimating the control variate coefficient becomes a challenge when we combine the two methods. Section 5 describes replication and related ideas that estimate a control variate coefficient for QMC, though possibly tuned to a smaller sample size than the one in use. Section 6 considers the coefficient appropriate in the limit as the sample size tends to infinity. Section 7 describes cases where the MC and QMC coefficients coincide so that the MC coefficient can be estimated from QMC data. Section 8 presents a low dimensional example for which we can compute the variance formulas of this paper. Section 9 illustrates these ideas on a 16-dimensional integral that arises as the value of an Asian call option. Section 10 summarizes our conclusions.

### 1.1 Notation

We complete this introductory section by describing some notation. Some additional notation is introduced at the point where it is used.

The integral $I$ of $f$ is the same over $[0,1]^s$ or $(0,1)^s$ or $[0,1)^s$. We employ $[0,1)^s$ only because it partitions easily into congruent subhypercubes.

A generic point in the unit cube is denoted by $x = (x^1, \ldots, x^s)^T$, while a point used in an integration rule is $X_i = (X_i^1, \ldots, X_i^s)^T$. For a function $g(x)$ on $[0,1)^s$, the term $\mathrm{Var}(g)$ denotes $\int (g(x) - \int g(x)\,dx)^2\,dx$, the variance of $g(X)$ when $X \sim U[0,1)^s$. For a vector $z$, the usual Euclidean norm is denoted $\|z\|_2$, and $\|z\|_1$ denotes the sum of absolute values of components of $z$.

Let $u \subseteq \{1, \ldots, s\}$. We use $|u|$ for the cardinality of $u$ and $-u$ for the complementary set $\{1, \ldots, s\} - u$.

There is an analysis of variance (ANOVA) decomposition for functions on the unit cube that is analogous to the ANOVA decomposition used in factorial experiments. A square integrable function $f$ can be written as a sum $f = \sum_u f_u(x)$ over $2^s$ subsets of $\{1, \ldots, s\}$, where $f_u(x)$ depends on $x$ only through $x^j$ for $j \in u$. Then $\mathrm{Var}(f) = \sum_{|u|>0} \mathrm{Var}(f_u)$. See Hoeffding (1948), Sobol' (1969) and Efron and Stein (1981).

When $s = 1$, the derivative of $g$ is denoted by $g'$. For $s \geq 1$, the gradient of $g$ is $\nabla g$, taken as an $s$-dimensional row vector. For a column vector $h$ of $J$ functions on $[0,1)^s$, the gradient $\nabla h$ is a $J$ by $s$ matrix of partial derivatives.

## 2. QUASI-MONTE CARLO

The Monte Carlo estimate $\hat{I}$ from (2) converges to $I$ with probability 1 by the strong law of large numbers. Quasi-Monte Carlo sampling may be thought of as a way to get a law of large numbers to hold without using randomness. The rate at which $|\hat{I} - I|$ converges to zero may be better for QMC than for MC, at least for functions $f$ with some spatial regularity.

### 2.1 Uniformity and Discrepancy

Quasi-Monte Carlo grew out of the theory of uniformly distributed sequences initiated by Weyl (1914, 1916); see Kuipers and Niederreiter (1974, Chapter 1). Let $a$ and $b$ be two points of $[0,1)^s$ for which $a < b$ holds coordinatewise, let $[a,b)$ be the $s$-dimensional box of points $X \in [0,1)^s$ for which $a \leq X < b$ holds coordinatewise and let $\mathrm{vol}([a,b))$ be the $s$-dimensional volume of that box. For $X_i \in [0,1)^s$ with $1 \leq i < \infty$, the sequence $(X_i)$ is *uniformly distributed* in $[0,1)^s$ if $\lim_{n\to\infty}(1/n)\sum_{i=1}^n \mathbb{1}_{a\leq X_i<b} = \mathrm{vol}([a,b))$ for all $0 \leq a < b \leq 1$.

If the sequence $(X_i)$ is uniformly distributed, then $\lim_{n\to\infty}(1/n)\sum_{i=1}^n f(X_i) = \int f(x)\,dx$ holds for every $f$ that is Riemann integrable on $[0,1)^s$. Thus the uniform distribution provides a deterministic analogue of the law of large numbers. Although Riemann integrability is a more stringent condition than the Lebesgue integrability required for Monte Carlo sampling, Riemann integrability is a very mild condition for applications.

The celebrated Weyl criterion is that $(X_i)$ is uniformly distributed if and only if $\lim_{n\to\infty}(1/n)\cdot\sum_{i=1}^n \exp(2\pi\sqrt{-1}\,k^T X_i) = 0$ for every nonzero vector $k \in \mathbb{Z}^s$. The Weyl criterion provides a way to establish that a given sequence is uniformly distributed.

Given two or more uniformly distributed sequences, it is of interest to decide which is better. Discrepancy measures are used to quantify the uniformity of a sequence of points.

The star discrepancy of a finite sequence $X_1, \ldots, X_n$ is defined as

$$D_n^*(X_1, \ldots, X_n) = \sup_{a\in[0,1)^s} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{0\leq X_i<a} - \mathrm{vol}([0,a)) \right|. \tag{3}$$

The star discrepancy is an $s$-dimensional generalization of the Kolmogorov–Smirnov distance between the discrete uniform distribution taking $X_i$ with probability $1/n$ for $i = 1, \ldots, n$ and the continuous uniform

distribution on $[0,1)^s$. Replacing the supremum over anchored boxes $[0,a)$ in (3) by the supremum over general axis parallel boxes $[a,b)$ yields the extreme, or unanchored, discrepancy $D_n(X_1,\dots,X_n)$. Because $D_n^* \le D_n \le 2^s D_n^*$, asymptotic rates in $n$, for fixed $s$, are identical for these discrepancies. Other discrepancies have been defined by replacing the supremum over boxes by suprema over other collections of subsets of $[0,1)^s$.

A different type of generalization of star discrepancy replaces the supremum with an $L^p$ norm as

$$D_n^{p*}(X_1,\dots,X_n) = \left(\int \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{0\le X_i<x} - \mathrm{vol}([0,x))\right|^p dx\right)^{1/p} \tag{4}$$

for $p \ge 1$, with $p=2$ the most widely studied. Beck and Chen (1987) and Matoušek (1999) provided book length treatments of discrepancy. In yet another generalization, we may interpret the star discrepancy as the worst case integration error over $f$ in the class of indicator functions of anchored boxes. Discrepancies defined with respect to classes of smooth functions appear in Paskov (1993), who considered integrated indicators of anchored boxes, and in Hickernell (1996), who considered functions in reproducing kernel Hilbert spaces.

Measures of discrepancy can be related to the quadrature error $|\hat{I}-I|$. The best known connection is the Koksma–Hlawka inequality

$$|\hat{I}-I| \le D_n^*(X_1,\dots,X_n) V_{\mathrm{HK}}(f), \tag{5}$$

where $V_{\mathrm{HK}}(f)$ denotes total variation of $f$ in the sense of Hardy and Krause. See Niederreiter (1992, Chapter 2) for a discussion of (5), Zaremba (1968) for an analogous inequality based on $D_n^{2*}$, Sobol' (1969, Chapter 8) for an inequality involving $D_n^{p*}$ and Hickernell (1996) for a treatment bounding $|\hat{I}-I|$ by a generalization of discrepancy times a generalization of variation.

Some infinite sequences $(X_i)$ with $D_n^*(X_1,\dots,X_n) = O(n^{-1}(\log n)^s)$ are known. It is suspected that $D_n^*$ cannot be $o(n^{-1}(\log n)^s)$ along an infinite sequence, but it has only been proved for $s=1$ and $s=2$. It is known that $D_n^*(X_1,\dots,X_n) \ge C_s n^{-1}(\log n)^{s/2}$ for infinitely many $n$ for some $C_s > 0$.

The fast convergence of $D_n^*$ combined with (5) shows that QMC is asymptotically superior to MC for functions of bounded variation. When $s$ is large, the quantity $n^{-1}(\log n)^s$ is not small at usual Monte Carlo sample sizes $n$. In empirical investigations (Morokoff and Caflisch, 1995; Sarkar and Prasad, 1987; Schlier, 2002) QMC is sometimes found to be much better than MC; other times the methods are comparable.

There are also triangular array constructions $X_{ni} \in [0,1)^s$ for $1 \le i \le n < \infty$ for which $D_n^*(X_{n1},\dots,X_{nn})$ attains the slightly better rate $O(n^{-1}(\log n)^{s-1})$. A disadvantage of triangular array schemes is that the points of the $n$ point quadrature rule are not necessarily present in the $n+1$ point rule. Rules based on the first $n$ points of an infinite sequence, by constrast, are necessarily extensible. There are many links between extensible rules in $s$ dimensions and nonextensible ones in $s+1$ dimensions. Matoušek (1999, Chapter 1) discussed this point.

Two QMC methods have dominated recent research and practice: digital nets and lattice rules. Digital nets are constructed to integrate the indicator functions of certain axis parallel boxes without error. Lattice rules integrate a class of sinusoidal functions without error. Each method then integrates linear combinations of its ideal integrands without error. Functions that are well approximated by such linear combinations are then integrated with small errors.

In both settings we will write the integrand as $f(x) = f_{\mathrm{G}}(x) + f_{\mathrm{B}}(x)$. Here $f_{\mathrm{G}}$ is a function on which the QMC method does a good job, integrating it without error. The error of QMC is then determined by the function $f_{\mathrm{B}}$ on which it does badly. The definitions of $f_{\mathrm{G}}$ and $f_{\mathrm{B}}$ differ for nets and lattices and depend on the sample size $n$. As $n$ increases, $\int (f(x)-f_{\mathrm{G}}(x))^2\,dx \to 0$. For each method, $\int f_{\mathrm{G}}(x) f_{\mathrm{B}}(x)\,dx = 0$ when $f$ and $g$ are in $L^2$.

### 2.2 Digital Nets

A thorough treatment of digital nets, also known as $(t,m,s)$ nets, was given by Niederreiter (1992). This section presents brief formal definitions of $(t,m,s)$ nets, $(t,s)$ sequences and $(\lambda,t,m,s)$ nets.

The following geometric discussion may be helpful for the reader who is encountering these definitions for the first time. A $(t,m,s)$ net in base $b$ is a form of stratified sample wherein the number of simultaneously balanced strata can be much larger than the sample size. The strata are hyperrectangular cells called elementary intervals or $b$-ary boxes. The sides of these $b$-ary boxes have endpoints that are $b$-adic fractions: integer multiples of $b^{-k}$ for some integer $k \ge 0$ and integer base $b \ge 2$. Given $n$ points $X_1,\dots,X_n$ in an integration rule, we would like every $b$-ary box of volume $b^{-K}$ to contain exactly $nb^{-K}$ of them. Nets manage to do this, at least for small enough $K$.

DEFINITION 1. For integer $b \geq 2$, a $b$-ary box in $[0,1)^s$ is a set of the form

$$\mathcal{B} = \prod_{j=1}^{s} \left[ \frac{\ell_j}{b^{k_j}}, \frac{\ell_j + 1}{b^{k_j}} \right) \tag{6}$$

for nonnegative integers $k_j$ and $\ell_j < b^{k_j}$.

DEFINITION 2. A $(t, m, s)$ net in base $b$ is a finite sequence $X_1, \ldots, X_{b^m}$ for which every $b$-ary box of volume $b^{t-m}$ contains exactly $b^t$ points of the sequence.

It is clear that smaller values of $t$ imply a better stratification. For given values of $b$, $m$ and $s$, there may not exist a net with $t = 0$, and so nets with $t > 0$ are also widely used.

Figure 1 shows the points of a $(0, 3, 5)$ net in base 5 projected onto two coordinates. The unit square can be partitioned into 125 boxes of shape $1/5 \times 1/25$. Each such box has exactly one point of the net. The same is true for partitions of shape $1/25 \times 1/5$. Although the reference lines do not show it, the 5-ary boxes of shape $1 \times 1/125$ and $1/125 \times 1$ also contain one point of the net. Finally, in any three-dimensional projection there are 125 boxes of shape $1/5 \times 1/5 \times 1/5$ with one point each.

The net shown is extensible. One can adjoin another 125 points to it, with the result that each $b$-ary box has two points of the extended sequence. Furthermore, some net constructions are extensible, not just twofold but $r$-fold for any integer $r > 1$. Finally, as some nets are extended, $b$-ary boxes of ever smaller volume contain the proportional number of points. Such extensible digital nets are defined through $(t, s)$ sequences.

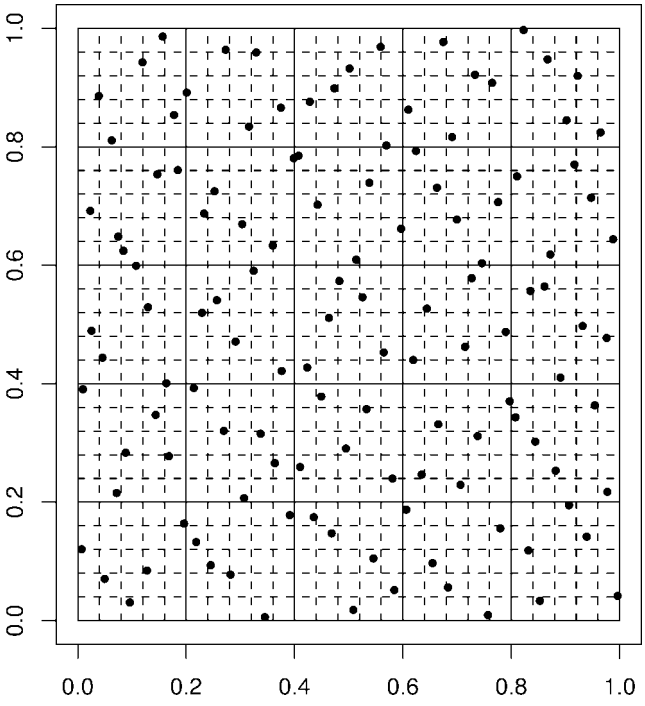


FIG. 1. *The* 125 *points of a digital net in base* 5 *as described in the text.*

DEFINITION 3. A $(t, s)$ sequence in base $b$ is an infinite sequence $X_i$ for $i \geq 1$ such that for all integers $r \geq 0$ and $m \geq t$, the points $X_{rb^m+1}, \ldots, X_{(r+1)b^m}$ form a $(t, m, s)$ net in base $b$.

If one samples a $(t, s)$ sequence with $n$ increasing through values $\lambda b^m$ for $1 \leq \lambda < b$ and $m \geq t$, then every $b$-ary box eventually contains a proportional number of points from the sequence and retains this balance thereafter. The first $\lambda b^m$ points of a $(t, s)$ sequence in base $b$ are a $(\lambda, t, m, s)$ net in base $b$, for any $m \geq t$ and $1 \leq \lambda < b$.

DEFINITION 4. Let $m, t$ and $\lambda$ be integers with $m \geq t \geq 0$ and $1 \leq \lambda < b$. A sequence of $\lambda b^m$ points in $[0,1)^s$ is called a $(\lambda, t, m, s)$ net in base $b$ if every $b$-ary box of volume $b^{t-m}$ contains $\lambda b^t$ points of the sequence and no $b$-ary box of volume $b^{t-m-1}$ contains more than $b^t$ points of the sequence.

The prototypical digital sequences are radical inverse sequences in base $b$, originating in the base 2 sequences of van der Corput (1935a, b). For integer base $b \geq 2$, let the nonnegative integer $n$ have base $b$ expansion $\sum_{k=1}^{\infty} n_k b^{k-1}$, where $n_k \in \{0, 1, \ldots, b-1\}$ and only finitely many $n_k$ are positive. The base $b$ radical inverse function, $\phi_b(n) = \sum_{k=1}^{\infty} n_k b^{-k} \in [0,1)$, reflects the base $b$ digits of $n$ through the base $b$ decimal point. In any $b^m$ consecutive nonnegative integers, all $b^m$ possible trailing digits appear exactly once. Then the corresponding values of $\phi_b$ contain all $b^m$ possible leading digits exactly once. It is customary to start the radical inverse sequence at 0. Thus $X_i = \phi_b(i-1)$ for $i \geq 1$ is a digital sequence with $t = 0$, $s = 1$ and base $b$.

Higher-dimensional digital nets and sequences require number theory to describe and construct, and are beyond the scope of an introductory survey. Faure (1982) presented constructions of $(0, p)$ sequences in prime bases $p$ and Sobol' (1967) constructed $(t, s)$ sequences in base 2, where the quality parameter $t$ depends on $s$. Niederreiter (1987) combined and extended these constructions. Of all presently known $(t, s)$-sequence constructions, those of Niederreiter and Xing (2001, Chapter 8) have the smallest values of $t$ for given values of $b$ and $s$.

To see why nets are effective integration rules, consider the $b$-ary indicator function $\mathbb{1}_{\mathcal{B}}(x)$ that is 1 if $x \in \mathcal{B}$ and 0 otherwise, where $\mathcal{B}$ is the $b$-ary box defined in (6). The volume of $\mathcal{B}$ is $b^{-K}$, where $K = \sum_{j=1}^{s} k_j$. If $X_1, \dots, X_n$ are a $(\lambda, t, m, s)$ net in base $b$ with $m - t \geq K$, then $(1/n)\sum_{i=1}^{n} \mathbb{1}_{\mathcal{B}}(X_i) = \int \mathbb{1}_{\mathcal{B}}(x)\,dx$. The points of a $(\lambda, t, m, s)$ net integrate without error any function that is a linear combination of the $b$-ary indicator functions of volume $b^{t-m}$. A combinatorial argument shows that there are $\binom{m-t+s-1}{s-1} b^{m-t}$ different $b$-ary indicator functions of volume $b^{t-m}$ correctly integrated by the points of a $(\lambda, t, m, s)$ net in base $b$. For example, the 625 points of a $(0, 4, 5)$ net in base 5 correctly integrate the indicators of 43,570 different 5-ary boxes of volume $1/625$.

Let $f_{\mathrm{G}}$ be the linear combination of indicator functions of $b$-ary boxes with volume $b^{t-m}$ that minimizes $\int (f(x) - f_{\mathrm{G}}(x))^2\,dx$. A formula for $f_{\mathrm{G}}$ can be based on tensor products of base $b$ Haar wavelets (Owen, 1997a). The integration error in a $(\lambda, t, m, s)$ net is the corresponding sample average of $f_{\mathrm{B}} = f - f_{\mathrm{G}}$.

### 2.3 Integration Lattices

Lattice methods for integration were introduced by Korobov (1959). Textbooks on the topic include Hua and Wang (1981), Sloan and Joe (1994) and Fang and Wang (1994).

DEFINITION 5. An $s$-dimensional lattice is a set of the form $\{\sum_{j=1}^{s} \alpha_j v_j \mid \alpha_j \in \mathbb{Z}\}$, where $v_1, \dots, v_s$ are linearly independent vectors in $\mathbb{R}^s$.

DEFINITION 6. An $s$-dimensional integration lattice is an $s$-dimensional lattice that contains every member of $\mathbb{Z}^s$.

DEFINITION 7. An $s$-dimensional lattice rule is the intersection of an $s$-dimensional integration lattice with $[0, 1)^s$.

The simplest lattice rule method is that known as "good lattice points." There one selects a sample size $n$ and a vector $\tau = (\tau_1, \dots, \tau_s)$ of nonnegative integers. Then for $i = 1, \dots, n$, let

$$X_i = \frac{(i-1)\tau}{n} \bmod 1, \tag{7}$$

where $z \bmod 1 = z - \lfloor z \rfloor$ and $\lfloor z \rfloor$ is the greatest integer less than or equal to $z$. Integration lattices that can be written in the form (7) are known as rank 1 lattices, because they have one generating vector $\tau$. Lattice rules of ranks 1 through $s$ were described by Sloan and Joe (1994). We emphasize rank 1 rules here. The lattice rules of Korobov (1959) are rank 1 rules for which $\tau = (1, \eta, \eta^2, \dots, \eta^{s-1})$ for some $\eta \in \mathbb{Z}$.

The vectors $(i-1)\tau/n$ are equally spaced on a ray from the origin to $(n-1)\tau/n$. Taking them modulo 1 causes them to "wrap around" the boundary of the unit cube. Careful choices of $\tau$ and $n$, made by combinations of algebra and computer search, lead to points that are very regularly spaced. Figure 2 shows a lattice rule with $\tau = (1, 89)$ and $n = 144$.

Classical lattice rules have a fixed sample size $n$ like a $(t, m, s)$ net. The development of extensible lattice rules, analogous to digital sequences, is fairly recent. The key insight is that one can replace $(i-1)\tau/n \bmod 1$ with $\phi_b(i-1)\tau \bmod 1$, where $\phi_b$ is the radical inverse function. The resulting points lie on a shifted lattice. Extensible shifted lattice rules allow the sample size $n$ to increase through a sequence of values of the form $b^m$ for increasing integers $m$. It has been shown by Hickernell and Niederreiter (2003) that there exist $\infty$-dimensional generating vectors $\tau = (\tau_1, \tau_2, \dots)$ that depend only on some base $b \geq 2$ and that give good lattice rules for all dimensions $s$ and for all $n$ equal to a power of $b$. Computer searches for vectors $\tau$ that give good lattices for a range of $s$ and $n$ have been made by Hickernell, Hong, L'Ecuyer and Lemieux (2000). The viability of component-by-component constructions has been demonstrated by

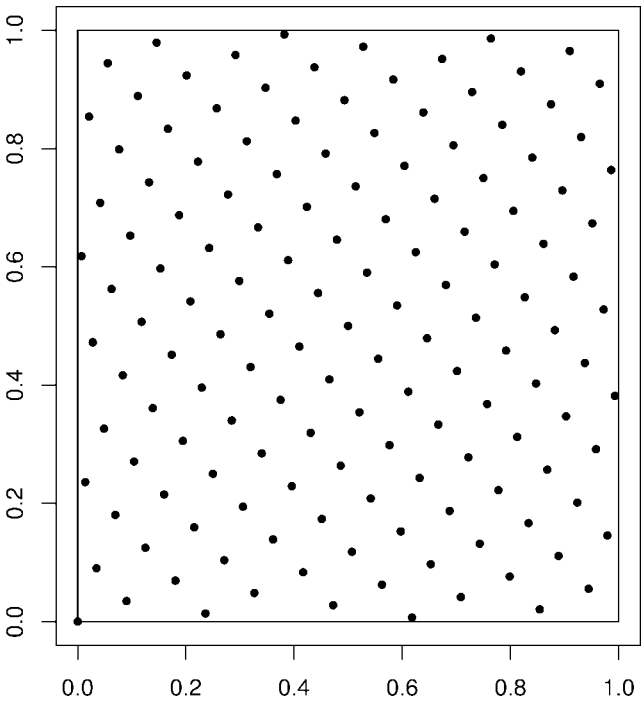


FIG. 2. *The* 144 *points of an integration lattice.*

Sloan and Reztsov (2002) and Sloan, Kuo and Joe (2002a, b).

Whereas nets are designed to integrate indicators of $b$-ary boxes, lattice rules integrate certain sinusoidal functions without error. Consider the multivariate trigonometric polynomials $\exp(2\pi\sqrt{-1}k^Tx)$, where $k \in \mathbb{Z}^s$ is an integer wave number vector. Suppose that $k$ belongs to the dual lattice $L^\perp = \{k : k^T\tau = 0 \bmod n\}$ of a rank 1 lattice rule. Then the function $\exp(2\pi\sqrt{-1}k^Tx)$ is completely aliased with the constant function 1 on the points of the lattice defined by (7). Lattice methods integrate trigonometric functions that correspond to $k \in L^\perp \setminus \{0\}$ with 100% error. However, for $k = 0$ or $k \notin L^\perp$, the function $\exp(2\pi\sqrt{-1}k^Tx)$ is integrated with zero error by lattice methods. For lattices $f_{\mathrm{B}}$ is the sum of the functions $\exp(2\pi\sqrt{-1}k^Tx)$ times the corresponding Fourier coefficients, taken over $k$ in $L^\perp \setminus \{0\}$. Then $f_{\mathrm{G}} = f - f_{\mathrm{B}}$ is the corresponding sum of Fourier contributions for $k \notin L^\perp - \{0\}$.

From the Weyl criterion we might expect that integrating trigonometric polynomials well will lead to a good quadrature rule. On a good sequence of lattice rules, the dual lattice $L^\perp$ becomes sparser as $n$ increases. The star discrepancy can be shown to approach zero at the same rate found for nets. The more rapidly the Fourier coefficients of $f$ decay, the better the asymptotic error rate for $|\hat{I} - I|$. For functions $f$ with $\partial^{rs} f / \prod_j \partial(x^j)^r$ continuous on $[0,1)^s$, the error rate can be made $O(n^{-r+\varepsilon})$ (Niederreiter, 1992), where $n^\varepsilon$ hides powers of $\log n$, although for large $s$ it may take very large $n$ for this rate to be relevant.

### 2.4 Randomized QMC

The law of large numbers is used to justify Monte Carlo methods, but not to compute error estimates. Practical error estimation is based on sample-based variance estimates, sometimes with a calibration via the central limit theorem. Bounds like (5) justify the use of QMC, but they are poorly suited to error estimation. Discrepancy is hard to calculate—total variation is harder still—and the resulting bound on $|\hat{I} - I|$, while tight for some worst-case $f$, can be extremely conservative.

Randomized quasi-Monte Carlo (RQMC) methods have been developed to combine QMC accuracy with the practical error estimation methods of MC. Typical RQMC methods replace a QMC sequence $A_1, \dots, A_n$ by a randomized version $X_1, \dots, X_n$ such that each $X_i \sim U[0,1)^s$ while the ensemble $X_1, \dots, X_n$ still has a QMC property. Because $X_i \sim U[0,1)^s$, it follows that $E(\hat{I}) = I$. The variance of $\hat{I}$ can be estimated through a small number of independent replications of the RQMC method. Studying RQMC also allows us to make sharper comparisons with MC, because variances can be estimated for both. Methods of randomizing nets and lattices were surveyed by Owen (1998a) and by L'Ecuyer and Lemieux (2002). Hong and Hickernell (2003) described software to randomize nets.

A scrambled net is a randomization of the base $b$ digits of the points of a digital net $A_1, \dots, A_n$. Let $A_i^j = \sum_{k=1}^\infty a_{ijk} b^{-k}$, where each $a_{ijk} \in \{0, 1, \dots, b-1\}$. The points of a scrambled net are $X_i^j = \sum_{k=1}^\infty x_{ijk} b^{-k}$, where $x_{ijk}$ are obtained through some random permutations of $a_{ijk}$. In the scrambling method proposed by Owen (1995), $x_{ij1} = \pi_j(a_{ij1})$, then $x_{ij2} = \pi_{j,a_{ij1}}(a_{ij2})$, so that the permutation of the second digit depends on what the first digit was; generally $x_{ijk} = \pi_{j,a_{ij1}\cdots a_{ijk-1}}(a_{ijk})$, where each $\pi$ is a uniform random permutation of 1 through $b-1$. Each $X_i^j$ has the $U[0,1)$ distribution and if $(A_i)$ are a digital net or sequence, then so are $(X_i)$ with probability 1. These scrambling schemes require a lot of permutations, and some derandomizations using fewer permutations have been proposed by Matoušek (1998) and Hong and Hickernell (2003).

For the scrambling method proposed by Owen (1995), as well as random linear scrambling (Matoušek, 1998; Hong and Hickernell, 2003), the variance of scrambled $(0, m, s)$-net quadrature satisfies

$$\mathrm{Var}_{\mathrm{rnet}}(\hat{I}) \le \frac{e}{n} \int f_{\mathrm{B}}(x)^2 \, dx \le e \, \mathrm{Var}_{\mathrm{mc}}(\hat{I}),$$

where $e = \exp(1) \doteq 2.718$. When $t > 0$, the variance bound $e/n$ has to be increased, but we still find $\mathrm{Var}_{\mathrm{rnet}}(\hat{I}) \le C_{b,s,t} \int f_{\mathrm{B}}(x)^2 \, dx / n$ for a constant $C_{b,s,t}$. See Owen (1998b), Niederreiter and Pirsic (2001) or Yue and Hickernell (2002). As $m$ increases, $f_{\mathrm{G}}$ accounts for more of the structure of $f$. In the limit, $\int f_{\mathrm{B}}(x)^2 \, dx \to 0$ and so $\mathrm{Var}_{\mathrm{rnet}}(\hat{I}) = o(1/n)$ for any square integrable $f$. Loh (2003) has proved a central limit theorem for the scramble proposed by Owen (1995).

For smooth functions, the rate at which $\int f_{\mathrm{B}}(x)^2 \, dx \to 0$ can be studied. Owen (1997b) showed that scrambled net integration attains a variance of $O(n^{-3}(\log n)^{s-1})$, so that $|\hat{I} - I| = O_p(n^{-3/2}(\log n)^{(s-1)/2})$, under a mild smoothness condition on $f$, given in Section 6.3. Note that in this setting, scrambling reduces the error of unscrambled nets by approximately a multiple of $n^{1/2}$.

Yue (1999) studied the variance over randomized $(\lambda, t, m, s)$ nets. Hickernell and Yue (2000), Matoušek (1998) and Heinrich, Hickernell and Yue (2004) investigated the discrepancy of scrambled nets and sequences. Owen (2002) studied the variance of scrambled net quadrature, finding that it can depend in a strong way on the details of the scrambling.

The usual randomization of lattice rules is a form of rotation modulo 1, due to Cranley and Patterson (1976). They took

$$X_i = U + \frac{(i-1)\tau}{n} \bmod 1, \tag{8}$$

where $U \sim U[0,1)^s$. Rotated lattice rules are a form of cluster sampling. They do not improve the error rate of lattice rules, but they do allow replication-based error estimates. Rotation affects the aliasing: For $k$ in the dual lattice, $\exp(2\pi\sqrt{-1}\,k^T X_i)$ equals $\exp(2\pi\sqrt{-1}\,k^T U)$ instead of 1.

To study randomized lattice rules, recall that some trigonometric polynomials are integrated exactly by the lattice while the others are constant on $X_1, \ldots, X_n$. For randomized lattice rules, $\mathrm{Var}_{\mathrm{rnet}}(\hat{I}) = \mathrm{Var}(f_{\mathrm{B}}) = \int f_{\mathrm{B}}(x)^2\,dx$. As with nets, the part $f_{\mathrm{G}}$ does not contribute to the error, but unlike nets, there is no $O(1/n)$ factor multiplying the contribution of the aliased part $f_{\mathrm{B}}$. The decay of $\mathrm{Var}_{\mathrm{rlat}}(\hat{I})$ with increasing $n$ is due to increasing sparsity of the dual lattice.

### 2.5 QMC and MCMC

Markov chain Monte Carlo (MCMC) is better known to statisticians than QMC. Both fields have a long history and both have grown tremendously in recent years. We have found only a little overlap between the methods. Liao (1998) reported some results using the Gibbs sampler in a QMC application. Ostland and Yu (1997) applied QMC to estimation of marginal distributions.

One reason why QMC and MCMC are so disjoint is that the integrands used in MCMC are often very spiky. For such problems, not much benefit can be expected from more uniform sampling of the entire space. Even if RQMC errors are like $An^{-3/2}$ while MCMC errors are like $Bn^{-1/2}$, the ratio $A/B$ for a spiky integrand could be much larger than any $n$ we might be able to use.

In some applications a well chosen importance sampling scheme could reduce the spikiness of the integrand to the point where QMC would be beneficial at realistic sample sizes, but effective importance sampling is very problem specific. It is also much more common in MCMC applications for $f(x)$ to be a product $p(x)g(x)$, where $p$ is a density function known only up to a normalizing constant. Then MCMC generates approximate samples from $p$, while QMC would have to fall back on ratio estimation methods.

An important difference between MCMC and QMC algorithms is that for MCMC the number of replications $n$ is small, perhaps one long run, while the dimension $s$ is large, nominally infinite. For QMC, $n$ is usually large and $s$ can be small.

## 3. CONTROL VARIATES

The idea in control variates is to exploit known values of $\int h_j(x)\,dx$ for $j = 1, \ldots, J$ to sharpen the estimate of $I$. The method is particularly compelling when $J = 1$ and $h_1 \doteq f$ with $\theta_1 = \int h_1(x)\,dx$ known. Most books on Monte Carlo methods consider control variates. See, for example, Bratley, Fox and Schrage (1987), Ripley (1987) or Fishman (1996). Essentially the same method goes by the name "regression estimators" in the survey sampling literature. See Cochran (1977) and Lohr (1999). Here we simply summarize some well-known results.

Suppose that we know the values $\int h(x)\,dx = \theta$ for the vector $h = (h_1, \ldots, h_J)^T$ of functions and the vector $\theta = (\theta_1, \ldots, \theta_J)^T$ of scalars. Then for any vector $\beta = (\beta_1, \ldots, \beta_J)^T \in \mathbb{R}^J$, the estimate

$$\hat{I}_\beta = \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - \sum_{j=1}^{J}\beta_j\bigl(h_j(X_i) - \theta_j\bigr)\right) \tag{9}$$

satisfies $E(\hat{I}_\beta) = I$ when $X_i \sim U[0,1)^s$.

To avoid trivialities, we suppose that $\max_{1\le j\le J}\int h_j^2(x)\,dx < \infty$ and that $\mathrm{Var}(\sum_{j=1}^{J}\beta_j \cdot h_j(X)) > 0$ for $X \sim [0,1)^s$ whenever $\beta \neq 0$. If $\mathrm{Var}(\beta^T h(X)) = 0$ for some nonzero $\beta$, then one or more of the functions $h_j$ is redundant and can be dropped.

The MC variance of $\hat{I}_\beta$ is $\mathrm{Var}_{\mathrm{mc}}(\hat{I}_\beta) = \sigma_\beta^2/n$, where

$$\sigma_\beta^2 = E\bigl(\bigl[f(X_i) - I - \beta^T\bigl(h(X_i) - \theta\bigr)\bigr]^2\bigr), \tag{10}$$

a quadratic function of the vector $\beta$. The minimizing value of $\beta$ is given by

$$\beta_{\mathrm{mc}} = \left(\int \bigl(h(x)-\theta\bigr)\bigl(h(x)-\theta\bigr)^T\,dx\right)^{-1} \cdot \int \bigl(h(x)-\theta\bigr)f(x)\,dx. \tag{11}$$

It always holds that $\sigma_{\mathrm{mc}}^2 \equiv \sigma_{\beta_{\mathrm{mc}}}^2 \le \sigma^2$, because $\sigma^2$ corresponds to $\beta = (0, \ldots, 0)^T$. We assume that $\sigma_{\mathrm{mc}}^2 > 0$ to rule out some trivial cases.

The value $\beta_{\mathrm{mc}}$ is typically unknown, and is usually estimated by

$$\hat{\beta}_{\mathrm{mc}} = \left( \sum_{i=1}^{n} (h(X_i) - \hat{H})(h(X_i) - \hat{H})^T \right)^{-1} \cdot \sum_{i=1}^{n} (h(X_i) - \hat{H}) f(X_i), \tag{12}$$

where $\hat{H} = (\hat{H}_1, \ldots, \hat{H}_J)^T$ and

$$\hat{H}_j = \frac{1}{n} \sum_{i=1}^{n} h_j(X_i).$$

The known values $\theta_j$ could possibly be used in place of $\hat{H}_j$, but typically are not. Instead $\hat{\beta}_{\mathrm{mc}}$ is the ordinary least squares estimator of the regression coefficients that relate $f(X_i)$ to $h_j(X_i)$.

The control variate estimator is $\hat{I}_{\hat{\beta}_{\mathrm{mc}}}$, which is obtained by substituting $\hat{\beta}_{\mathrm{mc}}$ for $\beta$ in (9). The resulting error is

$$\begin{aligned} \hat{I}_{\hat{\beta}_{\mathrm{mc}}} - I &= \hat{I}_{\beta_{\mathrm{mc}}} - I + \hat{I}_{\hat{\beta}_{\mathrm{mc}}} - \hat{I}_{\beta_{\mathrm{mc}}} \\ &= \hat{I}_{\beta_{\mathrm{mc}}} - I + (\hat{\beta}_{\mathrm{mc}} - \beta_{\mathrm{mc}})^T (\hat{H} - \theta). \end{aligned} \tag{13}$$

The second term in (13) does not ordinarily have mean zero, so the use of $\hat{\beta}_{\mathrm{mc}}$ typically introduces a small bias. It is ordinarily true that both $\hat{\beta}_{\mathrm{mc}} - \beta_{\mathrm{mc}}$ and $\hat{H} - \theta$ are $O_p(n^{-1/2})$, and then the last term in (13) is $O_p(1/n)$. This small term and the associated bias are customarily ignored. Cross-validatory methods can remove the bias in the estimate of $I$ and also in the variance estimate (Avramidis and Wilson, 1993).

Control variate methods are forgiving of mild errors in the coefficient $\beta$. Because $\sigma^2_\beta$ is a quadratic function of the vector $\beta$ with a minimum at $\beta_{\mathrm{mc}}$, it follows that $\sigma^2_\beta - \sigma^2_{\beta_{\mathrm{mc}}} = O(\|\beta - \beta_{\mathrm{mc}}\|_2^2)$ and, in particular, $\sigma^2_{\hat{\beta}_{\mathrm{mc}}} / \sigma^2_{\beta_{\mathrm{mc}}} = 1 + O_p(n^{-1})$.

## 4. CONTROL VARIATES WITH RQMC

Suppose that $X_1, \ldots, X_n$ are generated by an RQMC rule. Let $f$ be the integrand of interest and let $h = (h_1, \ldots, h_J)^T$ be a vector with $\int h(x)\,dx = \theta = (\theta_1, \ldots, \theta_J)^T$. The estimate $\hat{I}_\beta$ from (9) is still an unbiased estimate of $I$, but now

$$\mathrm{Var}_{\mathrm{rqmc}}(\hat{I}_\beta) = \mathrm{Var}_{\mathrm{rqmc}}\left( \hat{I} - \sum_{j=1}^{J} \beta_j \hat{H}_j \right), \tag{14}$$

where $\hat{H}_j = (1/n) \sum_{i=1}^{n} h_j(X_i)$, as before. Equation (14) does not simplify as in the IID case because the $X_i$ are not independent. This variance is still a quadratic in $\beta$, and the minimizing value is now

$$\beta_{\mathrm{rqmc}} = \mathrm{Cov}_{\mathrm{rqmc}}(\hat{H}, \hat{H})^{-1} \mathrm{Cov}_{\mathrm{rqmc}}(\hat{H}, \hat{I}). \tag{15}$$

There is always a control variate strategy that is at least as good as using no control variates: $\mathrm{Var}_{\mathrm{rqmc}}(\hat{I}_{\beta_{\mathrm{rqmc}}}) \le \mathrm{Var}_{\mathrm{rqmc}}(\hat{I})$ because $\mathrm{Var}_{\mathrm{rqmc}}(\hat{I})$ corresponds to using $\beta = 0$. A suboptimal or poorly estimated coefficient can, however, lead to worse results than obtained from not using the control variate. It is also clear from (14) that a control variate $h_j$ for which $\mathrm{Var}_{\mathrm{rqmc}}(\hat{H}_j) = 0$ is redundant.

As (14) and (15) show, an effective set of control variates must be correlated with $f$ under RQMC sampling. This is not necessarily the same as correlation of $h$ with $f$ under IID sampling. In particular, writing $f = f_{\mathrm{G}} + f_{\mathrm{B}}$ and $h = h_{\mathrm{G}} + h_{\mathrm{B}}$, we find that $f_{\mathrm{G}}$ and $h_{\mathrm{G}}$ do not contribute to (14), and we would rather have $h_{\mathrm{B}}$ correlated with $f_{\mathrm{B}}$ than have $h$ correlated with $f$.

Note that formula (12) for $\hat{\beta}_{\mathrm{mc}}$ applied to an RQMC sample will estimate $\beta_{\mathrm{mc}}$, not $\beta_{\mathrm{rqmc}}$. The use of RQMC sampling does not turn $\hat{\beta}_{\mathrm{mc}}$ into an estimate of $\beta_{\mathrm{rqmc}}$, but instead simply provides a more accurate estimate of $\beta_{\mathrm{mc}}$ than MC sampling would provide.

There is a further complication in that (15) is a moving target. It depends on the sample size $n$. For $n = 1$, we have $\beta_{\mathrm{rqmc}} = \beta_{\mathrm{mc}}$. As the sample size increases, more of the structure from $f$ is integrated exactly, and $\beta_{\mathrm{rqmc}}$ is determined only by the parts of $f$ and $h_j$ not integrated exactly.

### 4.1 Cautionary Example

The following simple example highlights the possible differences between $\beta_{\mathrm{mc}}$ and $\beta_{\mathrm{rqmc}}$. Take $s = 1$ and, for $M > 0$, let $f(x) = (1 + 2\lfloor Mx \rfloor - Mx)/M$ be a sawtooth function with teeth of width $1/M$. Figure 3 shows such a function for $M = 50$. In ordinary Monte Carlo sampling, the linear function $h_1(x) = x$ is an extremely good control variate for $f$. The optimal coefficient can be shown to be $\beta_{\mathrm{mc}} = 1 - 2M^{-2}$ and then $\sigma^2_{\mathrm{mc}} = 4\sigma^2(M^{-2} - M^{-4})$. Thus for $M = 50$, the control variate reduces the variance by a factor of 625.25.

Now consider a randomized $(0, 1, 1)$ net in base $b = n$. This trivially simple net reduces to a stratified sample in which one point is taken uniformly from each of the $n$ intervals $[(i-1)/n, i/n)$ for $i = 1, \ldots, n$. For simplicity suppose that $M = n$. The variance of $\hat{I}$ for this $f(x)$ under this stratified sampling is $1/(12M^3)$. Using the control variate with the coef-

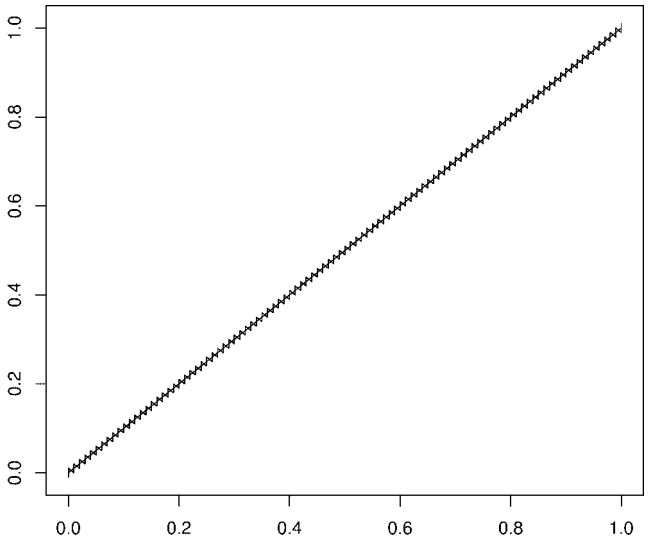


FIG. 3. *A sawtooth function $f$ with tooth width* 0.02 *and a linear function $h_1(x) = x$.*

ficient $\beta_{\mathrm{mc}}$ approximately doubles the variance compared to RQMC without a control variate.

The linear function $h_1(x) = x$ is in fact a good control variate for the sawtooth integrand $f$. Taking $\beta_{\mathrm{rqmc}} = -1$, we find that $\mathrm{Var}_{\mathrm{rqmc}}(\hat{I}_{\beta_{\mathrm{rqmc}}}) = 0$. In this case, using a coefficient $\beta$ optimized for RQMC eliminates the variance, while using the ordinary MC coefficient doubles the RQMC variance.

## 5. REPLICATION AND INTERNAL REPLICATION

In this section we consider the use of $R$ independent replicates of an $\tilde{n}$ point RQMC method. The total sample size is then $n = R\tilde{n}$ and replication allows us to estimate the vector $\beta_{\mathrm{rqmc}}$ appropriate to a sample of $\tilde{n}$ observations. A related idea is to exploit an "internal replication" structure, wherein $n$ consecutive RQMC points can be broken into $R$ consecutive blocks of $\tilde{n}$ points, in which each block constitutes a smaller RQMC rule. As described below, there is a trade-off in choosing $R$.

### 5.1 Replication Estimates of $\beta_{\mathrm{rqmc}}$

For $J$ control variates, let us take $R > J+1$ independent replications of the RQMC method with $\tilde{n}$ points each, producing for $r = 1, \ldots, R$ the estimates $\hat{I}_r$ and $\hat{H}_r = (\hat{H}_{1r}, \ldots, \hat{H}_{Jr})^T$. These estimates depend on $\tilde{n}$, but we suppress that dependence here.

Define $\hat{I}_\bullet = (1/R)\sum_{r=1}^R \hat{I}_r$ and $\hat{H}_\bullet = (1/R) \cdot \sum_{r=1}^R \hat{H}_r$. The combined replication estimate of $I$ is

$$\hat{I}_{\hat{\beta}} = \hat{I}_\bullet - \hat{\beta}^T(\hat{H}_\bullet - \theta), \tag{16}$$

where

$$\hat{\beta} = \left(\sum_{r=1}^R (\hat{H}_r - \hat{H}_\bullet)(\hat{H}_r - \hat{H}_\bullet)^T\right)^{-1} \cdot \left(\sum_{r=1}^R (\hat{H}_r - \hat{H}_\bullet)(\hat{I}_r - \hat{I}_\bullet)\right) \tag{17}$$

is a sample version of (15). The sum of squares

$$\mathrm{SS}(\beta_0, \beta) = \sum_{r=1}^R (\hat{I}_r - \beta_0 - \hat{H}_r^T\beta)^2 \tag{18}$$

is minimized by taking the scalar $\beta_0 = \hat{I}_{\hat{\beta}} - \hat{\beta}^T\theta$ and the vector $\beta = \hat{\beta}$. A natural estimate of $\mathrm{Var}(\hat{\beta})$ is then $\widehat{\mathrm{Var}}(\hat{\beta}) = \mathrm{SS}(\hat{\beta}_0, \hat{\beta})/(R(R-J-1))$.

### 5.2 Choosing $R$

For a given budget of $n = R\tilde{n}$ an important practical problem is to decide whether to use a large $R$ and a small $\tilde{n}$ or vice versa. The QMC error decreases faster in $\tilde{n}$ than in $R$, suggesting that $R$ should ordinarily be taken as small as other considerations allow. If $\beta_{\mathrm{rqmc}}$ is not being estimated from replications, then taking $R$ to be about 5 should give at least a reasonable number of degrees of freedom in a variance estimate. When there are $J$ coefficients in $\beta_{\mathrm{rqmc}}$ to estimate as in Sections 5.1 and 5.3, then taking $R = J + 5$ might suffice, taking note that control variate methods are forgiving of modest errors in $\beta$. The trade-off in picking small $R$ is that $R$ is the sample size for subsidiary tasks of estimating $\beta$ and the replication variance. To attempt an optimal choice of $R$ is a topic for further research.

### 5.3 Internal Replication

QMC schemes can often be considered to be "internally replicated." For example, a $(\lambda, t, m, s)$ net taken from a $(t, s)$ sequence can be decomposed into $R = \lambda b^{m-\tilde{m}}$ consecutive $(t, \tilde{m}, s)$ nets for $0 \le \tilde{m} \le m$. Likewise, an extensible shifted lattice with $n = b^m$ points can be decomposed into $R = b^{m-\tilde{m}}$ consecutive shifted lattices of $\tilde{n} = b^{\tilde{m}}$ points each.

For nets scrambled as described by Owen (1995), the formulas from Section 5.1 can ordinarily be used directly. As Owen (1997a) discussed, variance estimates based on internal replication tend to be conservative. Each internal replicate tends to fill in spaces avoided by the others and this tends to induce negative correlations among quantities such as $\hat{I}_r$ from different replications. Negative correlations among $\hat{I}_r$ reduce the variance of $\hat{I}_\bullet$ while simultaneously increasing the usual variance estimates.

Internal replication is more complicated for shifted lattice rules, owing to the aliasing phenomenon. One consequence of aliasing is that $\mathrm{Var}_{\mathrm{rlat}}(\hat{I}) = \mathrm{Var}(f_{\mathrm{B}})$ and similarly for $\hat{H}$, so that (15) reduces to

$$\beta_{\mathrm{rlat}} = \left(\int h_{\mathrm{B}}(x) h_{\mathrm{B}}^T(x)\,dx\right)^{-1} \cdot \int h_{\mathrm{B}}(x) f_{\mathrm{B}}(x)\,dx. \tag{19}$$

We discuss how to estimate $\int h_{\mathrm{B}}(x) h_{\mathrm{B}}^T(x)\,dx$ from (19); similar comments apply to $\int h_{\mathrm{B}}(x) f_{\mathrm{B}}(x)\,dx$. For lattices, $\hat{H} = \theta + \hat{H}_{\mathrm{B}}$, where $\hat{H}_{\mathrm{B}}$ is the quadrature rule applied to $h_{\mathrm{B}}$. Within replicate $r$ we get $\hat{H}_{\tilde{n},r} = \theta + \hat{H}_{\mathrm{B},\tilde{n},r}$, using notation that recognizes how the function $h_{\mathrm{B},\tilde{n}}$ depends on the within-replicate sample size $\tilde{n}$.

The denominator matrix in (19) may then be estimated by

$$\begin{aligned} &\frac{1}{R}\sum_{r=1}^{R}(\hat{H}_r - \hat{H}_\bullet)(\hat{H}_r - \hat{H}_\bullet)^T \\ &= \frac{1}{R}\sum_{r=1}^{R}\hat{H}_{\mathrm{B},\tilde{n},r}\hat{H}_{\mathrm{B},\tilde{n},r}^T - \hat{H}_{\mathrm{B}}\hat{H}_{\mathrm{B}}^T \\ &= \frac{1}{n}\sum_{i=1}^{n} h_{\mathrm{B},\tilde{n}}(X_i) h_{\mathrm{B},\tilde{n}}^T(X_i) - \hat{H}_{\mathrm{B}}\hat{H}_{\mathrm{B}}^T, \end{aligned} \tag{20}$$

wherein the first equality follows because averages of $h$ reduce to averages of $h_{\mathrm{B},\tilde{n}}$ and the second equality follows from aliasing. Inspecting (20) we see that $\beta_{\mathrm{rlat}}$ from (19) depends on mean squares defined through $f_{\mathrm{B}}$ and $h_{\mathrm{B}}$, while the internal replication estimate reduces to corresponding mean squares of $f_{\mathrm{B},\tilde{n}}$ and $h_{\mathrm{B},\tilde{n}}$. Thus the internal replication estimate $\beta$ is seen to be a direct estimate of $\beta_{\mathrm{rlat},\tilde{n}}$ for $\tilde{n} < n$.

## 6. LIMITING VALUES OF $\beta$

The previous section considered estimates of $\beta_{\mathrm{rqmc}}$ appropriate to sample sizes $\tilde{n} \le n$. In some cases we can compute or approximate

$$\begin{aligned} \beta_{\mathrm{rqmc}}^{\infty} &\equiv \lim_{n\to\infty} \beta_{\mathrm{rqmc}} \\ &= \lim_{n\to\infty} \mathrm{Cov}_{\mathrm{rqmc}}(\hat{H},\hat{H})^{-1}\,\mathrm{Cov}_{\mathrm{rqmc}}(\hat{H},\hat{I}), \end{aligned}$$

and the results provide qualitative insight and suggest some methods for choosing $\beta$.

We present three cases: (1) stratified sampling of $[0,1)$, (2) stratified sampling of $[0,1)^s$ and (3) randomized $(0,m,s)$ nets. For the first two cases the limit is obtained by correlating certain differential operators applied to $f$ and $h$. A similar result by Owen (1992) shows that a good control variate $h$ for Latin hypercube sampling is one whose nonadditive part correlates with that of $f$. The variance expressions for nets do not provide an expression for $\beta_{\mathrm{rqmc}}^{\infty}$, but do suggest a value that can be tested empirically. For extensible shifted lattices, it is not clear when $\beta_{\mathrm{rqmc}}^{\infty}$ exists.

### 6.1 Stratified Sampling of $[0,1)$

Suppose that $h_j$ and $f$ have Lipschitz continuous derivatives $h_j'$ and $f'$ on $[0,1]$. That is, for some $\Delta \in (0,1]$, some $B < \infty$ and all $x, x^* \in [0,1]$, both $|f'(x) - f'(x^*)| \le B|x - x^*|^{\Delta}$ and $\max_j |h_j'(x) - h_j'(x^*)| \le B|x - x^*|^{\Delta}$ hold. In practice this condition may commonly hold with $\Delta = 1$.

We stratify $[0,1)$ into $n$ intervals, and sample independently and uniformly within each of them. Specifically, our sample has independent random variables $X_i$ uniformly distributed on $[(i-1)/n, i/n)$ for $i = 1,\dots,n$.

Let $g$ be a function with Lipschitz continuous derivative $g'$ satisfying $|g'(x) - g'(x^*)| \le B|x - x^*|^{\Delta}$ for all $x, x^* \in [0,1]$. Then from Section 3 of Owen (1997b) we obtain

$$\begin{aligned} &\mathrm{Var}_{\mathrm{strat}}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right) \\ &\quad = \frac{1}{12n^3}\int_0^1 g'(x)^2\,dx + O(n^{-3-\Delta}). \end{aligned} \tag{21}$$

It is natural to substitute $f - \beta^T h$ for $g$ in the lead term of (21) and then minimize over $\beta$. Some care is required with the error term. We show below that this minimization gives the right answer.

LEMMA 1. *Assume that $f$ and $h_j$ have Lipschitz derivatives as described above with common values of $B$ and $\Delta$, and that $\int h'(x)h'(x)^T\,dx$ has full rank $J$. Then the optimal control variate coefficient under stratified sampling satisfies*

$$\lim_{n\to\infty}\beta_{\mathrm{strat}} \equiv \beta_{\mathrm{strat}}^{\infty} = \left(\int_0^1 h'(x)h'(x)^T\,dx\right)^{-1} \cdot \int h'(x) f'(x)\,dx. \tag{22}$$

PROOF. By (21) we get

$$\mathrm{Var}_{\mathrm{strat}}(\hat{I}_\beta) = \frac{1}{12n^3}\int_0^1 \left(f' - \sum_{j=1}^J \beta_j h_j'\right)^2 dx + \left(1 + \sum_{j=1}^J |\beta_j|\right) O(n^{-3-\Delta}), \tag{23}$$

where the constant inside the $O$ symbol is independent of $\beta$. Let that constant be $D/12$ for $0 \le D < \infty$.

Because $\int_0^1 h'(x)h'(x)^T\,dx$ has full rank, the right-hand side of (22) is the unique minimizer of the first term in (23). Let $\delta_J > 0$ be the smallest eigenvalue of $\int_0^1 h'(x)h'(x)^T\,dx$. By a sequence of elementary bounds, for large enough $n$ we have

$$\begin{aligned} &12n^3\big(\mathrm{Var}_{\mathrm{strat}}(\hat{I}_\beta) - \mathrm{Var}_{\mathrm{strat}}(\hat{I}_{\beta^\infty_{\mathrm{strat}}})\big)\\ &\quad\ge \delta_J^2\|\beta - \beta^\infty_{\mathrm{strat}}\|_2^2 - Dn^{-\Delta}(1+\|\beta\|_1)\\ &\quad\ge \delta_J^2\|\beta - \beta^\infty_{\mathrm{strat}}\|_2^2 - Dn^{-\Delta}(1+\|\beta\|_1)\\ &\quad\ge \delta_J^2\|\beta - \beta^\infty_{\mathrm{strat}}\|_1^2/J\\ &\qquad - Dn^{-\Delta}(1+\|\beta\|_1 + \|\beta - \beta^\infty_{\mathrm{strat}}\|_1). \end{aligned}$$

Suppose that $\|\beta - \beta^\infty_{\mathrm{strat}}\|_1 > \varepsilon > 0$. Then $\mathrm{Var}_{\mathrm{strat}}(\hat{I}_\beta) > \mathrm{Var}_{\mathrm{strat}}(\hat{I}_{\beta^\infty_{\mathrm{strat}}})$ holds for large enough $n$. The result follows. □

Lemma 1 shows that the asymptotically optimal control variate coefficient is obtained through the expected cross-products of first derivatives of $f$ and $h_j$. Notice that the averages of $f'$ and $h_j'$ are not first subtracted.

In practice we can estimate $\beta^\infty_{\mathrm{strat}}$ from the stratified sample as

$$\hat{\beta}^\infty_{\mathrm{strat}} = \left(\sum_{i=1}^n h'(X_i)h'(X_i)^T\right)^{-1} \cdot \sum_{i=1}^n h'(X_i)f'(X_i) \tag{24}$$

and replication is not necessary. Here $\beta^\infty_{\mathrm{strat}}$ is obtained by least squares regression, without an intercept term, of $f'$ on $h'$.

A simple special case has $h_1(x) = x$. Then $h_1'(x) = 1$ and $\beta^\infty_{\mathrm{strat}} = \int_0^1 f'(x)\,dx = f(1) - f(0)$, and so

$$\hat{I}_{\beta_{\mathrm{strat}}} = \frac{1}{n}\sum_{i=1}^n \big(f(X_i) - \big(f(1) - f(0)\big)(X_i - 0.5)\big)$$

with variance $(12n^3)^{-1}\,\mathrm{Var}(f'(x)) + O(n^{-3-\Delta})$ instead of $(12n^3)^{-1}\int_0^1 f'(x)^2\,dx + O(n^{-3-\Delta})$. If the variance of $f'(X)$ for $X \sim U[0,1)$ is much smaller than its mean square, then an appreciable variance reduction is obtained.

The stratification scheme above describes a simple special case of randomized nets. A similarly simple special case of lattice rules has $X_i = (i - 1 + U)/n$ for $i = 1, \ldots, n$, where the same random variable $U \sim [0,1)$ is used in all $n$ random values. In this case we also find that (22) is the best regression coefficient, but the factor $1/(12n^3)$ in the variance has to be replaced by $1/(12n^2)$. The stratified sample by using $n$ independent uniform deviations achieves an additional variance reduction factor of $n$ from error cancellation.

### 6.2 Stratified Sampling of $[0,1)^s$

For small $s$ it is feasible to stratify the unit cube into $n = m^s$ congruent subcubes having side dimension $1/m$ and to sample one $X_i$ uniformly within each such cube. For $f$ and $h_j$ smooth enough we find a similar result to the one-dimensional case.

If the real-valued function $g$ has two continuous derivatives, then the variance of $g(X)$ for $X$ sampled uniformly within a hypercube of size $1/m$ with center $c$ is

$$\frac{1}{12m^2}\|\nabla g(c)\|_2^2 + O(m^{-2}),$$

where $\nabla g$ is the 1 by $s$ gradient (row) vector of $g$.

The lead term $\mathrm{Var}_{\mathrm{strat}}(\hat{I}_\beta)$ is then

$$\frac{1}{12n^{1+2/s}}\int_{[0,1)^s}\left\|\nabla\left(f(x) - \sum_{j=1}^J \beta_j h_j(x)\right)\right\|_2^2 dx. \tag{25}$$

The variance rate $n^{-(1+2/s)}$ describes the well-known deterioration of cubic stratification in higher dimensions.

Recalling our definition of $\nabla$ from Section 1.1 we may write the asymptotically optimal coefficient as

$$\beta^\infty_{\mathrm{strat}} = \left(\int_{[0,1)^s}\nabla h\,\nabla h^T\,dx\right)^{-1} \cdot \int_{[0,1)^s}\nabla h\,\nabla f\,dx \tag{26}$$

and estimate it by

$$\hat{\beta}^\infty_{\mathrm{strat}} = \left(\sum_{i=1}^n \nabla h(X_i)\nabla h^T(X_i)\right)^{-1} \cdot \sum_{i=1}^n \nabla h(X_i)\nabla f(X_i). \tag{27}$$

The results for $s$-dimensional stratification generalize those of one-dimensional stratification by replacing the scalar first derivatives $h'$ and $f'$ with the corresponding gradients $\nabla h$ and $\nabla f$. An argument along the lines of Lemma 1 shows that optimizing the dominant term of (25) gives the asymptotically optimal coefficient.

### 6.3 Randomized Nets

Finite sample variance formulas are available for randomized nets, but they appear to be too cumbersome to help us choose $\beta$. The asymptotic variance formulas are not sharp enough to allow us to derive the exact value of $\beta_{\mathrm{rnet}}^{\infty}$, but they do suggest a way to compute a candidate value $\widetilde{\beta}_{\mathrm{rnet}}^{\infty}$. This and other candidates, such as estimates of $\beta_{\mathrm{mc}}$, can then be compared numerically in applications.

Let $\partial^s f/\partial x$ denote the order $s$ mixed partial derivative of $f$ taken once with respect to each component of $x$. Let $\partial^{|u|} f/\partial_u x$ denote the mixed partial derivative of $f$ taken once with respect to each index in $u$. Owen (1997b) defined smooth $s$-dimensional functions as those that satisfy

$$\left|\frac{\partial^s}{\partial x}\bigl(f(x)-f(x^*)\bigr)\right| \le B\|x-x^*\|_2^{\Delta} \tag{28}$$

for finite $B \ge 0$ and $\Delta \in (0,1]$. Then, under a scrambled $(0,m,s)$ net,

$$\begin{aligned}\mathrm{Var}_{\mathrm{rnet}}(\hat{I}) = \biggl[&\frac{(\log n)^{s-1}}{n^3}\frac{\lambda^2}{12^s(s-1)!}\\ &\cdot\left(\frac{b^2-1}{\log b}\right)^{s-1}\int\left(\frac{\partial^s f(x)}{\partial x}\right)^2 dx\biggr]\\ &\cdot\bigl(1+O(1)\bigr)\end{aligned} \tag{29}$$

as $n\to\infty$, for the scrambling in Owen (1995), where the constant in $O(1)$ depends on $B$ and $\Delta$ only.

If we replace $f$ with $f-\beta^T h$ in (29) and minimize the integral there over $\beta$, we obtain

$$\begin{aligned}\widetilde{\beta}_{\mathrm{rnet}}^{\infty} = &\left(\int\frac{\partial^s h(x)}{\partial x}\frac{\partial^s h^T(x)}{\partial x}\,dx\right)^{-1}\\ &\cdot\int\frac{\partial^s f(x)}{\partial x}\frac{\partial^s h^T(x)}{\partial x}\,dx\end{aligned} \tag{30}$$

as the optimizer of an estimate of $\mathrm{Var}_{\mathrm{rnet}}(\hat{I}_\beta)$.

Equation (29) arises in the limit as $n\to\infty$ of a sum

$$\frac{1}{n}\sum_{|u|>0}\bigl(M_u+O(1)\bigr)\int\left(\frac{\partial^{|u|}f_u}{\partial_u x}\right)^2 dx.$$

The sum contains $2^s-1$ terms, one for every nonconstant ANOVA term $f_u$ in $f$. The coefficients $M_u$ can be found in Owen (1997b). As $n\to\infty$ the highest-order ANOVA term dominates, having a coefficient $M_{\{1,\dots,s\}}$ that is larger by powers of $\log(n)$ than any other terms. Equation (30) can be written without an ANOVA component because $\partial^s f_{\{1,\dots,s\}}/\partial x = \partial^s f/\partial x$.

Things simplify considerably if $h_j$ only has one nonzero ANOVA component. If, for example, $J=1$ and $h_1(x)=\prod_{\ell\in u}(x^\ell-0.5)$, then $\partial^{|u|}h_1(x)/\partial_u x\equiv 1$ and then

$$\widetilde{\beta}_{\mathrm{rnet}}^{\infty} = \int\frac{\partial^{|u|}f_u(x)}{\partial_u x}\,dx.$$

In special settings we might know this value or be able to approximate it using sample values of the required partial derivative.

## 7. ORTHOGONAL CONTROL VARIATE COEFFICIENTS

If we can show that $\beta_{\mathrm{rqmc}}=\beta_{\mathrm{mc}}$, then we can expect $\hat{I}_{\hat{\beta}_{\mathrm{mc}}}$ to be effective in RQMC sampling. For a stratified sample, consider a function $h$ such that the average value of $h$ is $\theta$ within every one of the strata. Then $\mathrm{Cov}_{\mathrm{strat}}(\hat{H},\hat{H}) = \mathrm{Cov}_{\mathrm{mc}}(\hat{H},\hat{H})$ and $\mathrm{Cov}_{\mathrm{strat}}(\hat{H},\hat{I}) = \mathrm{Cov}_{\mathrm{mc}}(\hat{H},\hat{I})$, and so $\beta_{\mathrm{strat}}=\beta_{\mathrm{mc}}$.

For a scrambled $(\lambda,0,m,s)$ net in base $b$, there are some integrands known to have exactly the Monte Carlo variance. For a $(0,m,s)$ net in base $b$, it follows from Owen (1997a) that the indicator function of a sufficiently fine $b$-ary box, one with $\sum_{j=1}^s k_j \ge m$, will be integrated with exactly the Monte Carlo variance as will a linear combination of such fine $b$-ary boxes.

The variance of scrambled net integration is known to be a sum of contributions from each nonconstant ANOVA term in the integrand. In examples with smooth integrands (Owen, 1997b; Caflisch, Morokoff and Owen, 1997), one sees that the contribution from a given ANOVA term tends to decay at the MC rate $1/n$ until about $n=b^{|u|+t}$. Then it declines more rapidly. Thus we can expect control variates dominated by their higher-dimensional ANOVA contributions to have $\beta_{\mathrm{rqmc}}$ close to $\beta_{\mathrm{mc}}$.

A good control variate for scrambled nets would be one that matched the high dimensional and fine parts of the function, leaving a difference $f-\beta^T h$ that had primarily low dimensional, and coarse parts. That is, the control variate would leave an integrand of low effective dimension in the superposition sense of Caflisch, Morokoff and Owen (1997).

For shifted extensible lattices a good control variate is one whose aliased part is strongly correlated with the

aliased part of $f$. Aliasing makes it harder to estimate the coefficient for such a control variate. If, however, we know that $\beta_{\text{rlat}} = \beta_{\text{mc}}$, then the strategy from Section 5.3 with a small value of $\tilde{n}$ is reasonable.

## 8. SMALL NUMERICAL EXAMPLE

Here we present a two-dimensional numerical example. Because the dimension is so low and the functions involved are smooth, we can expect the asymptotic variance formulas to be reliable, even for modest sample sizes.

For $x = (x^1, x^2)^T \in [0,1)^2$, let $f(x) = \sin(\pi(x^1 + x^2))$. It is common to select control variates that have a qualitative similarity to the integrand. Here we let $J = 1$ and take $h_1(x) = (x^1 + x^2 - 1)^3 - (x^1 + x^2 - 1)$ as such a similar function. We know that $\int h_1(x)\,dx = \theta_1 = 0$. We also know that $I = 0$, but we will investigate the accuracy of estimates of $I$. The various integrals in the asymptotic variance formulas have been computed by averaging over a 100 by 100 midpoint grid in $[0,1)^2$ and also by averaging over 65,536 points obtained from a scrambled $(0, 15, 2)$ net in base 2 and its antithetic points of the form $(1 - X_i^1, 1 - X_i^2)^T$. These two methods agree for the values reported below.

The simple estimator (2) has variance $1/(2n)$ under MC sampling. The variable $h$ is highly correlated with $f$, and we find $\beta_{\text{mc}} = 2.675$. Equation (26) gives $\beta_{\text{strat}}^{\infty} = 2.809$ and (30) gives $\widetilde{\beta}_{\text{rnet}}^{\infty} = 2.547$. Table 1 records the asymptotic sampling variances of $\hat{I}_\beta$ for all three methods and all four control variate coefficient values. Each method has its own asymptotic rate in $n$. The coefficients are computed through (10), (25) and (29), including the constants $1/12$ and $12^{-2}(2^2 - 1)/\log(2) = 0.0301$ in the latter two.

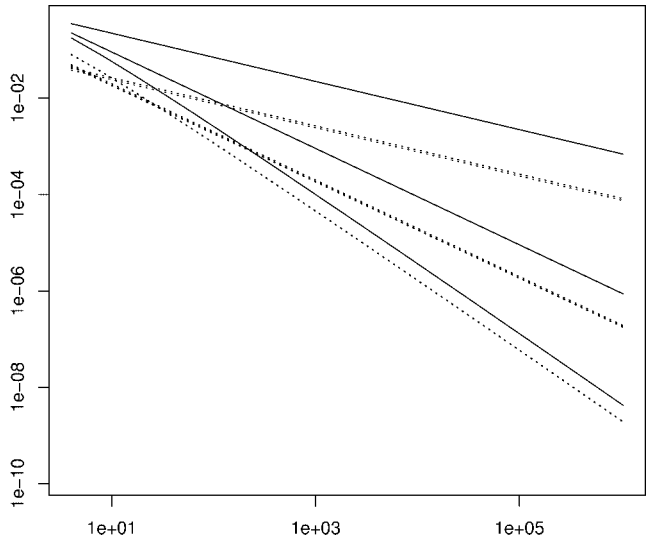


FIG. 4. *The asymptotic standard deviations of $\hat{I}_\beta$ versus $n$ for the methods in Table* 1. *The solid lines are, top to bottom, for MC, stratification and randomized nets. Below these are parallel dotted lines that represent when control variates are employed. Lines for different control variate values largely overlap on this plot.*

Standard deviations found as square roots of the asymptotic variances from Table 1 are plotted in Figure 4. The story for this example is that nets work better than stratification, which works better than IID sampling. For all methods, using the control variate brings an improvement and the amount of improvement does not depend strongly on which coefficient was used. The benefit from using this control variate diminishes as one uses better sampling methods.

These asymptotic variances predict that stratification without control variates will surpass MC with an optimal control variate at roughly $n = 139$, which we ought to round to 144 because stratification requires that $n$ be a perfect square. Scrambled nets without control vari-

TABLE 1
*Asymptotic variances of MC, stratification and QMC for a two-dimensional problem from the text*

| | | CV coefficient | | | | |
|---|---|---|---|---|---|---|
| **Method** | **Rate** | **None** **0** | $\boldsymbol{\beta_{\text{mc}}}$ **2.675** | $\boldsymbol{\beta_{\text{strat}}^{\infty}}$ **2.809** | $\boldsymbol{\widetilde{\beta}_{\text{rnet}}}$ **2.547** | **Gain** |
| MC | $n^{-1}$ | 0.5 | 0.00594 | 0.00718 | 0.00707 | 84.2 |
| Strata | $n^{-2}$ | 0.8245 | 0.0351 | 0.0333 | 0.0402 | 24.7 |
| $(0, m, 2)$ net | $n^{-3}\log(n)$ | 1.464 | 0.297 | 0.307 | 0.294 | 4.98 |

NOTE. The coefficients $\beta_{\text{mc}}$ from (11), $\beta_{\text{strat}}^{\infty}$ from (27) and $\widetilde{\beta}_{\text{rnet}}^{\infty}$ from (30) were computed numerically and are displayed above the table. The asymptotic variance formulas (10), (25) and (29) applied to $f - \beta^T h$ have rates in $n$ given to the left of the table with numerically determined constants given in the body of the table. The rightmost column shows the variance reduction comparing the $\beta = 0$ variance to the smallest variance in the row.

ates overtake MC with the optimal control variate at roughly $n = 29$, which we round to 32 because these nets require $n$ to be a power of 2. For nets without the control variate, to overcome stratification with the control variate takes $n = 241$, which again we round to 256.

For Monte Carlo sampling, the control variate in this example allowed us to reduce the variance by a factor of 84.2. The corresponding factors for stratified sampling and randomized nets are 24.7 and 4.98, respectively. It happened that the better balanced sample points gained less from the control variate and, what is almost the same thing, were more forgiving of inaccurate control variate values.

Matching $h_1$ to $f$ we found that there was a lesser, but still useful, correlation between certain derivatives of $h_1$ and corresponding derivatives of $f$. There was one surprise. Viewing stratification as intermediate between MC and RQMC, we might have expected to find that $\beta^{\infty}_{\text{strat}}$ would lie between $\beta_{\text{mc}}$ and $\beta^{\infty}_{\text{rqmc}}$, but it did not.

Notice that the benefit from a variance reduction is higher for MC sampling than it is for QMC. For example, in MC sampling a variance reduction of 10 is equivalent to a 10-fold increase in the effective sample size. In settings where the variance decreases more quickly, the gain translates into smaller sample size multiples. When the variance decreases proportionally to $n^{-2}$ or $n^{-3}$, then a 10-fold reduction in variance would equate to sample size increases of $10^{1/2} \doteq 3.16$ and $10^{1/3} \doteq 2.15$, respectively. The rate $n^{-3}$ corresponds to scrambled net variance ignoring logarithmic powers, while $n^{-2}$ is appropriate to bivariate stratification and, ignoring logarithmic powers, some other RQMC methods.

## 9. ASIAN OPTION

This section considers an example in $s = 16$ variables. There is no assurance that asymptotic error rates for QMC are relevant for this dimension until $n$ is extremely large. There is, however, empirical evidence that QMC and RQMC methods usually surpass MC methods, well before entering their asymptotic regime.

The integral we study represents the value of an Asian call option. Valuing Asian options is a problem of practical interest in financial applications and is also a widely studied test problem for MC and QMC methods. In this setting there is an underlying asset with price $S(t)$ at time $t$. The option pays an amount $\max(0, (1/s)\sum_{i=1}^{s} S(t_i) - K)$ at time $T$, where $K$ is the strike price and $t_1, \dots, t_s$ are the dates at which the asset's price is recorded. Somebody planning to make regular purchases of the asset between times 0 and $T$ might buy this option as a hedge against high future prices.

Under the Black–Scholes model, the value of this option at time $t = 0$ is the expected value of the payment, assuming that $S(t)$ follows geometric Brownian motion times a discount factor that reflects the time value of money. Geometric Brownian motion at $s$ time points can be expressed through a vector $x \sim U[0,1)^s$ as

$$\begin{aligned} S(t_i) &= S(t_i, x) \\ &= S(0)\exp\Bigg[(r - \sigma^2/2)t_i \\ &\qquad + \sigma\sqrt{T/s}\sum_{j=1}^{i}\Phi^{-1}(x^j)\Bigg], \end{aligned}$$

where the drift parameter $r$ is the risk-free rate, $\sigma$ is the volatility of the asset prices and $\Phi^{-1}$ is the inverse of the standard normal cumulative distribution function. Incorporating the discount we find the value is $\int f(x)\,dx$, where

$$f(x) = e^{-rT}\max\left(0, \frac{1}{s}\sum_{i=1}^{s} S(t_i, x) - K\right).$$

In our experiments, we used an initial price of $S(0) = 100$, an annualized interest rate of $r = 0.05$, an expiration of $T = 1$ year and $s = 16$ equispaced times $t_i = i/16$ for $i = 1, \dots, 16$. The volatility is $\sigma = 0.3$. The strike price is $K = 120$, so that the option is initially out of the money. For this option the probability of a nonzero payout is roughly 0.17. When the payout probability is much smaller than this, then some form of importance sampling becomes helpful.

A widely used control variate for Asian options replaces the arithmetic option by a geometric one:

$$h_1(x) = e^{-rT}\max\left(0, \prod_{i=1}^{s} S(t_i, x)^{1/s} - K\right).$$

The geometric mean inside $h_1(x)$ has a log-normal distribution that allows $\int h_1(x)\,dx$ to be found via a one-dimensional integration that reduces to the Black–Scholes formula

$$\int h_1(x)\,dx = e^{-rT}\big[\exp(a + b^2/2)\Phi(d_1) - K\Phi(d_2)\big],$$

where

$$a = \ln(S(0)) + (r - \sigma^2/2)T(s+1)/(2s),$$
$$b^2 = \sigma^2 T(s+1)(2s+1)/(6s^2),$$
$$d_1 = (-\ln K + a + b^2)/b,$$
$$d_2 = d_1 - b,$$

taking $b \geq 0$, and where $\Phi$ is the standard normal cumulative distribution function. See Ritchken, Sankarasubramanian and Vijh (1993).

The functions

$$A(x) = e^{-rT}\left(\frac{1}{s}\sum_{i=1}^{s} S(t_i, x) - K\right), \tag{31}$$

$$G(x) = e^{-rT}\left(\prod_{i=1}^{s} S(t_i, x)^{1/s} - K\right) \tag{32}$$

are useful in a control variate strategy for QMC. The standard asymptotic results for QMC assume integrands of bounded variation in the sense of Hardy and Krause. The functions $f(x)$ and $h_1(x)$ are unbounded on $[0,1)^s$ and hence are not of bounded variation. The functions $f(x) - A(x)$ and $h_1(x) - G(x)$ are at least bounded, although lacking sufficient smoothness to be of bounded variation. Note also that $f - A$ and $h_1 - G$ represent the discounted payoff from the corresponding put options, which pay $\max(0, K - (1/s)\sum_{i=1}^{s} S(t_i))$ and $\max(0, K - \prod_{i=1}^{s} S(t_i)^{1/s})$, respectively. Both $\int A(x)\,dx$ and $\int G(x)\,dx$ are easily obtainable. For this problem,

$$\int h_1(x)\,dx = 1.916,$$
$$\int A(x)\,dx = -16.454,$$
$$\int G(x)\,dx = -17.191.$$

The Monte Carlo methods we consider are listed in Table 2. They all use IID points $X_i \sim U[0,1)^s$. The $\mathrm{MC}_0$ is plain Monte Carlo with no control variates; $\mathrm{MC}_1$ uses one control variate, $h_1$; $\mathrm{MC}_3$ uses three control variates, $h_1$, $A$ and $G$; $\mathrm{MC}_\mathrm{B}$ uses the bounded function $f - A$; and $\mathrm{MC}_\mathrm{BB}$ uses the bounded function $f - A$ with a bounded control variate $h_1 - G$. The coefficients $\beta_j$ required are estimated by least squares on the Monte Carlo sample.

We also considered (randomized) QMC versions of all of these strategies. For an out of the money option such as this, $f(x) = 0$ for most $x$ and has smaller variance than $f(x) - A(x)$. It is reasonable a priori to expect $\mathrm{MC}_\mathrm{B}$ to be worse than $\mathrm{MC}_0$, but $\mathrm{QMC}_\mathrm{B}$ might be better than $\mathrm{QMC}_0$ due to boundedness in $f - A$.

The RQMC strategies we investigated were based on $(0, m, 16)$ nets in base 17 using the generalized Faure construction described in Tezuka (1995). Our first version used $R = 85$ independent replicates of a randomized $(0, 2, 16)$ net. Our second version used $R = 5$ replicates of a $(0, 3, 16)$ net. Both versions require $n = 5 \times 17^3 = 24{,}565$ function evaluations, and this is also the number of function evaluations used in the MC simulations. The randomization was a random digital shift as described in L'Ecuyer and Lemieux (2002). We denote the methods $\mathrm{QMC}^{(2)}$ and $\mathrm{QMC}^{(3)}$. The superscript shows $m$ and the control variate method is specified through the same list of subscripts used for MC.

For the 85 replicates of the $(0, 2, 16)$ net, the replication strategy in Section 5.1 was used to estimate the control variates and the variance of $\hat{I}_{\hat{\beta}}$. In each of the five replicates of the $(0, 3, 16)$ net, the coefficients $\beta_j$ were estimated using the formula for $\hat{\beta}_{\mathrm{mc}}$ applied to QMC data. These five values were then averaged and the sample standard error was computed.

TABLE 2
*The Monte Carlo methods used in the Asian option example*

| Name | Estimate |
|---|---|
| $\mathrm{MC}_0$ | $\hat{I}(f)$ |
| $\mathrm{MC}_1$ | $\hat{I}(f - \beta_1 h_1) + \beta_1 I(h_1)$ |
| $\mathrm{MC}_3$ | $\hat{I}(f - \beta_2 h_1 - \beta_3 A - \beta_4 G) + \beta_2 I(h_1) + \beta_3 I(A) + \beta_4 I(G)$ |
| $\mathrm{MC}_\mathrm{B}$ | $\hat{I}(f - A) + I(A)$ |
| $\mathrm{MC}_\mathrm{BB}$ | $\hat{I}(f - A - \beta_5(h_1 - G)) + I(A) + \beta_5 I(h_1 - G)$ |

NOTE. In each estimate $\hat{I}(g)$ is the sample average of $g(X_i)$ and $I(g) = \int g(x)\,dx$ is assumed known. The $X_i$ employed are IID from $U[0,1)^s$ and $\beta_j$ are estimated by least squares regression. The mnemonic underlying the first three subscripts is that those methods use 0, 1 and 3 control variates. $\mathrm{MC}_\mathrm{B}$ works directly with a bounded integrand and $\mathrm{MC}_\mathrm{BB}$ uses a bounded integrand and a bounded control variate.

The results of the simulation are shown in Tables 3 and 4. The standard errors in Table 3 were obtained by analyzing the MC and QMC$^{(2)}$ data as 85 replicate samples of size 289 and those for QMC$^{(3)}$ were obtained in an analysis of five replicates of size 4913. As might be expected, $\beta_3$ is close to $-\beta_4$, while the other coefficients are close to 1. The values for $\beta_3$ and $\beta_4$ are quite different for QMC$^{(2)}$ than those for the other methods. The reason is that QMC$^{(3)}$, having only 5 replicates, used estimates of $\beta_{\mathrm{mc}}$, while the 85 replicates in QMC$^{(2)}$ were sufficient to allow estimation of $\beta_{\mathrm{rqmc}}$.

In Table 4 we see that for each set of control variates, QMC$^{(3)}$ is more accurate than QMC$^{(2)}$, which is in turn more accurate than MC. In particular, while QMC$^{(3)}$ could only be used with estimates of suboptimal coefficients, it still outperformed QMC$^{(2)}$.

Without control variates, the root mean square error (RMSE) for MC is about 11.92 times that for QMC$^{(3)}$. For MC to attain that reduced error would require a sample size $11.92^2 \doteq 142$ times as large. The QMC$^{(2)}$ attained a smaller improvement over MC.

The best control variate strategy for MC was to use all three variates. For this function the control variates reduced RMSE by a factor of 21.2 corresponding to a sample size improvement of about 450. In this problem control variates alone bring a better result than QMC alone.

With optimal coefficients, using all three variates would also be the best strategy for QMC, because the other control variate strategies can be obtained as choices of $\beta_2$, $\beta_3$ and $\beta_4$. The QMC$^{(2)}$ used 85 replicates and also had its smallest error with all three control variates. For QMC$^{(3)}$ with QMC estimates of $\beta_{\mathrm{mc}}$, the method $\mathrm{QMC}_{\mathrm{BB}}^{(3)}$ with just one control variate had better accuracy than $\mathrm{QMC}_3^{(3)}$.

The best combined strategy was $\mathrm{QMC}_{\mathrm{BB}}^{(3)}$, with an efficiency gain of $(4.41/0.0735)^2 \doteq 3600$ compared to $\mathrm{MC}_0$. The two best methods for this problem are $\mathrm{QMC}_1^{(3)}$ and $\mathrm{QMC}_{\mathrm{BB}}^{(3)}$. They gave option values of 2.162 and 2.163, respectively, with the standard errors in Table 4.

TABLE 3
*Estimated control variate coefficients for MC and for QMC with m indicated as a superscript* 2 *or* 3
(*standard errors are in parentheses*)

| Coef. | MC | QMC$^{(2)}$ | QMC$^{(3)}$ |
|---|---|---|---|
| $\beta_1$ | 1.10 (4.9e−4) | 1.08 (5.6e−3) | 1.10 (1.1e−3) |
| $\beta_2$ | 1.04 (2.3e−4) | 1.01 (7.7e−3) | 1.04 (4.0e−4) |
| $\beta_3$ | 0.534 (1.5e−3) | 1.33 (1.3e−1) | 0.519 (2.7e−3) |
| $\beta_4$ | −0.525 (1.5e−3) | −1.37 (9.7e−2) | −0.510 (2.7e−3) |
| $\beta_5$ | 0.988 (2.0e−4) | 1.03 (9.0e−3) | 0.987 (1.2e−4) |

TABLE 4
*Estimated root mean squared errors*

| | MC | QMC$^{(2)}$ | QMC$^{(3)}$ |
|---|---|---|---|
| 0 | 4.41e−2 | 2.05e−2 | 3.70e−3 |
| 1 | 2.99e−3 | 2.16e−3 | 1.34e−3 |
| 3 | 2.08e−3 | 1.48e−3 | 1.04e−3 |
| B | 9.05e−2 | 1.69e−2 | 2.94e−3 |
| BB | 2.81e−3 | 1.52e−3 | 7.35e−4 |

NOTE. The row labels describe the control variate strategy as described in Table 2. The column labels describe the sampling strategy: MC or QMC with $m$ indicated as a superscript 2 or 3.

As expected, $\mathrm{MC}_{\mathrm{B}}$ was worse than $\mathrm{MC}_0$. For both QMC methods the bounded function approaches $\mathrm{QMC}_{\mathrm{B}}^{(m)}$ were (slightly) better than the corresponding $\mathrm{QMC}_0^{(m)}$ methods. Similarly there were small advantages for $\mathrm{QMC}_{\mathrm{BB}}^{(m)}$ using the bounded functions $f - A$ and $h_1 - G$ over $\mathrm{QMC}_1^{(m)}$ using corresponding unbounded functions $f$ and $h_1$.

The results discussed above can be brought out in an ANOVA of the logarithms of the numbers in Table 4. An additive model fits with an $R^2$ of 90%. The fitted main effects may be interpreted as follows. Compared to MC, QMC$^{(2)}$ and QMC$^{(3)}$ reduce variance by factors of 4.4 and 33, respectively. Control variates reduce variance by factors of 53 for method 1, 103 for method 3 and 104 for method BB, while method B increases variance by about 1.2. The interaction effects, when exponentiated, result in some synergies, most notably a further 5-fold variance reduction for B with QMC$^{(3)}$ and about a 5.7-fold variance increase for B with MC.

## 10. CONCLUSIONS

In this paper we have investigated the consequences of combining QMC with control variates. Replacing MC with QMC usually improves accuracy. Applying this notion to $f - \beta^T h$, we ordinarily expect the combined method to improve on MC with control variates. Incorporating control variates into MC or QMC also improves accuracy, in general, although for QMC it can be harder to select control variates.

Not surprisingly, in our examples we saw diminishing returns to employing both strategies: the improvement from control variates was smaller for QMC than

for MC. Equivalently, the improvement from QMC was smaller with control variates than without. These results are consistent with Ben Ameur, L'Ecuyer and Lemieux (1999), who reported numerical examples in which control variates improve QMC, but not as much as they improve simpler methods. Furthermore, as remarked in Section 8, a given variance reduction factor corresponds to a larger sample size reduction for MC than for QMC.

In our two numerical examples, using estimates of $\beta_{\mathrm{mc}}$ with QMC gave very good results, and this is reassuring. In the Asian option problem we saw better results using estimated suboptimal coefficients $\beta_{\mathrm{mc}}$ with our best equidistribution strategy $\mathrm{QMC}^{(3)}$ than we saw using a weaker equidistribution $\mathrm{QMC}^{(2)}$ for which we could estimate the corresponding optimal $\beta_{\mathrm{rqmc}}$. The tentative conclusion is that if one is using both QMC and control variates, the quality of the QMC method is more important than that of the control variate coefficient.

In other problems, estimates of $\beta_{\mathrm{mc}}$ could lead to poor performance. In practice this can be tested by comparing standard errors for QMC with and without control variates. Then, if necessary, replicates may be used to estimate $\beta_{\mathrm{rqmc}}$ or, what seems better, internal replicates can be used to estimate the value of $\beta_{\mathrm{rqmc}}$ appropriate to a smaller sample size than the one in use.

We found theoretically that effective control variates for QMC are not necessarily the same as for MC. For MC, a good control variate is one that correlates with the integrand, while for QMC, a good control variate is one wherein certain derivatives or high frequency components correlate with the corresponding aspects of the integrand.

In our derivations we explored control variates for RQMC instead of for QMC per se. An alternative approach is to define the optimal $\beta_{\mathrm{qmc}}$ as one that minimizes an error bound, such as one proportional to the total variation of $f - \beta^T h$. We found that alternative less attractive for several reasons. First, the total variation is a factor in a bound on the error and the value of $\beta$ that minimizes the bound is not necessarily the one that minimizes the error itself. Second, the total variation is not as tractable to optimize as the variance. For smooth enough functions, the total variation may be written as an $L^1$ norm applied to the $s$-dimensional mixed partial derivative $\partial^s/\partial x$, suggesting that we should consider minimizing $\int |\partial^s(f(x) - \beta^T h(x))/\partial x| \, dx$. Thus, qualitatively at least, effective control variates are again those for which a certain derivative is approximately linearly related to the corresponding derivative of the target integrand.

## ACKNOWLEDGMENTS

We thank Rong-Xian Yue and Sut Yue Hung for discussions. Hickernell's work was partially supported by Hong Kong Research Grants Council Project HKBU 2030/99P and Hong Kong Baptist University Grant FRG/00-01/II-62. Lemieux's work was supported by NSERC Grant RGP238959. Owen's work was supported by NSF Grants DMS-00-72445 and DMS-03-06612.

## REFERENCES

AVRAMIDIS, A. and WILSON, J. R. (1993). A splitting scheme for control variates. *Oper. Res. Lett.* **14** 187–198.

BECK, J. and CHEN, W. W. L. (1987). *Irregularities of Distribution*. Cambridge Univ. Press.

BEN AMEUR, H., L'ECUYER, P. and LEMIEUX, C. (1999). Variance reduction of Monte Carlo and randomized quasi-Monte Carlo estimators for stochastic volatility models in finance. In *Proc. 1999 Winter Simulation Conference* **1** 336–343. IEEE Press, New York.

BRATLEY, P., FOX, B. L. and SCHRAGE, L. E. (1987). *A Guide to Simulation*, 2nd ed. Springer, New York.

CAFLISCH, R. E., MOROKOFF, W. and OWEN, A. B. (1997). Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. *J. Comput. Finance* **1** 27–46.

CHELSON, P. (1976). Quasi-random techniques for Monte Carlo methods. Ph.D. dissertation, Claremont Graduate School.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

CRANLEY, R. and PATTERSON, T. (1976). Randomization of number theoretic methods for multiple integration. *SIAM J. Numer. Anal.* **13** 904–914.

EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596.

FANG, K.-T. and WANG, Y. (1994). *Number-Theoretic Methods in Statistics*. Chapman and Hall, London.

FAURE, H. (1982). Discrépance de suites associées à un système de numération (en dimension $s$). *Acta Arith.* **41** 337–351.

FISHMAN, G. (1996). *Monte Carlo*: *Concepts, Algorithms, and Applications*. Springer, New York.

HEINRICH, S., HICKERNELL, F. J. and YUE, R.-X. (2004). Optimal quadrature for Haar wavelet spaces. *Math. Comp.* **73** 259–277.

HICKERNELL, F. J. (1996). Quadrature error bounds with applications to lattice rules. *SIAM J. Numer. Anal.* **33** 1995–2016; corrected printing of Sections 3–6 (1997) **34** 853–866.

HICKERNELL, F. J., HONG, H. S., L'ECUYER, P. and LEMIEUX, C. (2000). Extensible lattice sequences for quasi-Monte Carlo quadrature. *SIAM J. Sci. Comput.* **22** 1117–1138.

HICKERNELL, F. J. and NIEDERREITER, H. (2003). The existence of good extensible rank-1 lattices. *J. Complexity* **19** 286–300.

HICKERNELL, F. J. and YUE, R.-X. (2000). The mean square discrepancy of scrambled $(t, s)$-sequences. *SIAM J. Numer. Anal.* **38** 1089–1112.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

HONG, H. S. and HICKERNELL, F. J. (2003). Algorithm 823: Implementing scrambled digital sequences. *AMS Trans. Math. Software* **29** 95–109.

HUA, L. and WANG, Y. (1981). *Applications of Number Theory to Numerical Analysis*. Springer, Berlin.

KOROBOV, N. M. (1959). The approximate computation of multiple integrals. *Dokl. Akad. Nauk SSSR* **124** 1207–1210.

KUIPERS, L. and NIEDERREITER, H. (1974). *Uniform Distribution of Sequences*. Wiley, New York.

L'ECUYER, P. and LEMIEUX, C. (2002). Recent advances in randomized quasi-Monte Carlo methods. In *Modeling Uncertainty*: *An Examination of Stochastic Theory, Methods, and Applications* (M. Dror, P. L'Ecuyer and F. Szidarovszki, eds.) 419–474. Kluwer, Dordrecht.

LIAO, J. G. (1998). Variance reduction in Gibbs sampler using quasi random numbers. *J. Comput. Graph. Statist.* **7** 253–266.

LOH, W.-L. (2003). On the asymptotic distribution of scrambled net quadrature. *Ann. Statist.* **31** 1282–1324.

LOHR, S. (1999). *Sampling*: *Design and Analysis*. Brooks/Cole, Pacific Grove, CA.

MATOUŠEK, J. (1998). On the $L_2$-discrepancy for anchored boxes. *J. Complexity* **14** 527–556.

MATOUŠEK, J. (1999). *Geometric Discrepancy*: *An Illustrated Guide*. Springer, Heidelberg.

MOROKOFF, W. and CAFLISCH, R. E. (1995). Quasi-Monte Carlo integration. *J. Comput. Phys.* **122** 218–230.

NIEDERREITER, H. (1987). Point sets and sequences with small discrepancy. *Monatsh. Math.* **104** 273–337.

NIEDERREITER, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia.

NIEDERREITER, H. and PIRSIC, G. (2001). The microstructure of $(t, m, s)$-nets. *J. Complexity* **17** 683–696.

NIEDERREITER, H. and XING, C. (2001). *Rational Points on Curves over Finite Fields*: *Theory and Applications. London Math. Soc. Lecture Note Ser.* **285**. Cambridge Univ. Press.

OSTLAND, M. and YU, B. (1997). Exploring quasi Monte Carlo for marginal density approximation. *Statist. Comput.* **7** 217–228.

OWEN, A. B. (1992). A central limit theorem for Latin hypercube sampling. *J. Roy. Statist. Soc. Ser. B* **54** 541–551.

OWEN, A. B. (1995). Randomly permuted $(t, m, s)$-nets and $(t, s)$-sequences. *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing. Lecture Notes in Statist.* **106** 299–317. Springer, New York.

OWEN, A. B. (1997a). Monte Carlo variance of scrambled net quadrature. *SIAM J. Numer. Anal.* **34** 1884–1910.

OWEN, A. B. (1997b). Scrambled net variance for integrals of smooth functions. *Ann. Statist.* **25** 1541–1562.

OWEN, A. B. (1998a). Latin supercube sampling for very high dimensional simulations. *ACM Transactions on Modeling and Computer Simulation* **8** 71–102.

OWEN, A. B. (1998b). Scrambling Sobol' and Niederreiter–Xing points. *J. Complexity* **14** 466–489.

OWEN, A. B. (2002). Scrambled net variance with alternative scramblings. Technical report, Dept. Statistics, Stanford Univ.

PASKOV, S. (1993). Average case complexity of multivariate integration for smooth functions. *J. Complexity* **9** 291–312.

RIPLEY, B. D. (1987). *Stochastic Simulation*. Wiley, New York.

RITCHKEN, P., SANKARASUBRAMANIAN, L. and VIJH, A. M. (1993). The valuation of path dependent contracts on the average. *Management Sci.* **39** 1202–1213.

SARKAR, P. K. and PRASAD, M. A. (1987). A comparative study of pseudo- and quasirandom sequences for the solution of integral equations. *J. Comput. Phys.* **68** 66–88.

SCHLIER, C. (2002). A practitioner's view on QMC integration. Technical report, Fakultät für Physik, Univ. Freiburg.

SLOAN, I. H. and JOE, S. (1994). *Lattice Methods for Multiple Integration*. Oxford Univ. Press, New York.

SLOAN, I. H., KUO, F. Y. and JOE, S. (2002a). Constructing randomly shifted lattice rules in weighted Sobolev spaces. *SIAM J. Numer. Anal.* **40** 1650–1665.

SLOAN, I. H., KUO, F. Y. and JOE, S. (2002b). On the step-by-step construction of quasi-Monte Carlo integation rules that achieve strong tractability error bounds in weighted Sobolev spaces. *Math. Comp.* **71** 1609–1640.

SLOAN, I. H. and REZTSOV, A. V. (2002). Component-by-component construction of good lattice rules. *Math. Comp.* **71** 263–273.

SOBOL', I. M. (1967). The distribution of points in a cube and the accurate evaluation of integrals. *Zh. Vychisl. Mat. i Mat. Fiz.* **7** 784–802. (In Russian.)

SOBOL', I. M. (1969). *Multidimensional Quadrature Formulas and Haar Functions*. Nauka, Moscow. (In Russian.)

SPANIER, J. and MAIZE, E. H. (1994). Quasi-random methods for estimating integrals using relatively small samples. *SIAM Rev.* **36** 18–44.

TEZUKA, S. (1995). *Uniform Random Numbers*: *Theory and Practice*. Kluwer, Boston.

VAN DER CORPUT, J. G. (1935a). Verteilungsfunktionen I. *Nederlandse Akademie van Wetenschappen Proceedings* **38** 813–821.

VAN DER CORPUT, J. G. (1935b). Verteilungsfunktionen II. *Nederlandse Akademie van Wetenschappen Proceedings* **38** 1058–1066.

WEYL, H. (1914). Über ein Problem aus dem Gebeite der diophantischen Approximationen. *Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II* 234–244.

WEYL, H. (1916). Über die Gleichverteilung von Zahlen mod. eins. *Math. Ann.* **77** 313–352.

YUE, R.-X. (1999). Variance of quadrature over scrambled unions of nets. *Statist. Sinica* **9** 451–473.

YUE, R.-X. and HICKERNELL, F. J. (2002). The discrepancy and gain coefficients of scrambled digital nets. *J. Complexity* **18** 135–151.

ZAREMBA, S. K. (1968). Some applications of multidimensional integration by parts. *Ann. Polon. Math.* **21** 85–96.

# Comment

**Pierre L'Ecuyer**

*Randomized quasi-Monte Carlo* (RQMC) is a form of variance reduction technique (VRT) that aims to induce negative dependence between the replicates when a mathematical expectation is estimated by an average. It works in the same spirit as stratification and the method of antithetic variates, for example. Another way to reduce the variance is to exploit the dependence between the estimator and *control variates* (CVs; i.e., other random variables with known expectation, correlated with the original estimator), and make an appropriate correction to the estimator. Of course, these two techniques can be combined. However, combining VRTs often gives rise to complicated synergetic effects that are not always easy to analyze (see, e.g., Avramidis and Wilson, 1996; Glynn and Szechtman, 2002) and this applies to the RQMC–CV setup. Two important observations are that (1) the variance reduction factor for the combined method can be no better than for each method alone, but can also be orders of magnitude better than the product of variance reduction factors of the two methods and (2) the optimal CV coefficients with and without RQMC ($\beta_{\text{rqmc}}$ and $\beta_{\text{mc}}$) may be very different and the former is often harder to estimate.

Hickernell, Lemieux and Owen's interesting paper provides good insight on these issues and, perhaps more importantly, opens the door to attractive and largely unexplored territory. Their paper starts with a nice compact and authoritative overview of QMC methods and their randomizations. Although artificial, their small example in Section 4.1 gives a case where the RQMC–CV estimator reduces the variance to zero with its optimal CV coefficient $\beta_{\text{rqmc}}$, while $\beta_{\text{rqmc}}$ is approximately the *opposite* of $\beta_{\text{mc}}$, so using RQMC–CV with coefficient $\beta_{\text{mc}}$ in this case *increases* the variance. This underlines the importance of estimating the optimal CV coefficient *for the correct setting*.

### ESTIMATING $\beta_{\text{rqmc}}$ WITH $R$ REPLICATES

As the authors rightly point out, estimating $\beta_{\text{rqmc}}$ defined in their equation (15) is harder than estimating $\beta_{\text{mc}}$, because with RQMC the observations $X_i$ are not independent. The easiest way to estimate the variance of an RQMC estimator is to replicate the RQMC scheme $R$ times, independently, and use the sample variance of the $R$ averages as a variance estimator. This same methodology can be used to estimate all the variances and covariances involved in the expression for $\beta_{\text{rqmc}}$. So, instead of an $n$-point RQMC scheme, one uses $R$ independently randomized $\tilde{n}$-point RQMC schemes, where $n = R\tilde{n}$. The difficulty is that taking a large $R$ compromises the effectiveness of RQMC, whereas with a small $R$, the variance and covariance estimators can be very noisy, making the estimator of $\beta_{\text{rqmc}}$ unreliable.

This problem does not occur for the Asian option example in Section 9, where using $\beta_{\text{mc}}$ instead of $\beta_{\text{rqmc}}$ works reasonably well anyway, but it may certainly occur in other applications. Would it be rare or frequent? I guess only experience will tell.

The controlled estimator (16) is also biased in general when $\beta_{\text{rqmc}}$ is estimated from the same data. The bias vanishes when $R \to \infty$. There is no bias for finite $R$ if the distribution of $(\hat{I}_r, \hat{H}_r^T)^T$ is multinormal, but otherwise, for small $R$, there could be significant bias and it becomes more difficult to have reliable variance estimates. We may be interested in finding the value of $R$ that minimizes the mean square error of (16) for a given $n$, for example. The solution is of course highly problem dependent and depends on $n$. To get useful insight on what $R$ should be used in actual applications, it seems that empirical investigations with specific classes of models and RQMC methods are necessary.

To avoid diluting RQMC's effectiveness, one can also use the *internal replications* heuristic discussed by the authors, where an RQMC method is used based on a point set of cardinality $n$ that can be partitioned into $R$ highly uniform point sets of cardinality $\tilde{n}$. The idea is to *pretend* that the estimates obtained with these $R$ different subsets of points are independent, as if these were $R$ independent replicates of a given RQMC scheme. In fact, these $R$ estimates are not independent, so this heuristic provides a *biased* estimator of $\beta_{\text{rqmc}}$.

The authors argue that in certain settings where the $R$ RQMC estimators based on $\tilde{n}$ points are *identically distributed* [this is a key property that underlies the validity of their equation (20), in particular], the method

*Pierre L'Ecuyer is Professor, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Québec, Canada H3C 3J7.*

provides valid estimators of the variances, covariances and optimal $\beta$ that correspond to *one* of these $\tilde{n}$-point RQMC estimators (instead of $n$). If both $n$ and $\tilde{n}$ are large, it should be typically true that these variances and covariances for $n$ and $\tilde{n}$ do not differ much. On the other hand, one can construct examples where they differ by arbitrarily large factors.

Suppose for instance that $f$ has the ANOVA decomposition $f = f_{\mathrm{G}} + f_{\mathrm{B}} + f_{\mathrm{BB}}$, where $f_{\mathrm{G}}$ is integrated with zero error by the $\tilde{n}$-point rule, $f_{\mathrm{B}}$ is integrated with zero error by the $n$-point rule but with 100% error by the $\tilde{n}$-point rule and $f_{\mathrm{BB}}$ is integrated with 100% error by both rules. We may decompose $h$ in a similar way, as $h = h_{\mathrm{G}} + h_{\mathrm{B}} + h_{\mathrm{BB}}$. It may very well happen that $\int h_{\mathrm{B}}(x) f_{\mathrm{B}}(x)\,dx$ differs significantly from $\int h_{\mathrm{BB}}(x) f_{\mathrm{BB}}(x)\,dx$ or that $\int h_{\mathrm{B}}(x) h_{\mathrm{B}}(x)\,dx$ differs significantly from $\int h_{\mathrm{BB}}(x) h_{\mathrm{BB}}(x)\,dx$, so $\beta_{\mathrm{rqmc},n}$ may turn out to be quite different from $\beta_{\mathrm{rqmc},\tilde{n}}$. This did not happen in the Asian option example examined by the authors and perhaps it is unlikely to happens in a majority of practical cases, but the danger still exists.

## STRATIFICATION

In their Section 6 the authors discuss the choice of $\beta$ and provide convergence results for a CV with stratified sampling. Their analysis assumes a single CV coefficient for all strata. However, there are many situations where it is more appropriate to select *different* coefficients $\beta$ for the different strata.

I will illustrate this with an example of a telephone call center, modeled as a queueing system with a nonstationary Poisson arrival process, gamma service times and a single first-in–first-out FIFO queue (see, e.g., Pichitlamken, Deslauriers, L'Ecuyer and Avramidis, 2003). Agents answering calls are the *servers* in the queueing system. Customers have a random *patience time* and *abandon* the queue (are lost) when their waiting time in the queue exceeds this patience time. Two quantities that interest call center managers are $E[L]$ and $E[G(s)]$, where $L$ is the number of abandonments in a day and $G(s)$ is the number of callers who waited less than $s$ seconds in a day. Let $A$ be the total number of arrivals in a day. It is easy to compute $E[A]$ from the model, so in the long run (over an infinite number of days), the fraction of callers who abandon is $E[L]/E[A]$ and the fraction whose waiting time is less than $s$ is $E[G(s)]/E[A]$. These fractions can be estimated by estimating their numerators. Call centers may receive several thousands of calls per day, so these expectations are integrals with a huge (and random) number of dimensions. Nevertheless, RQMC methods, stratification and their combination with CVs can help improve simulation efficiency if we take advantage of the structure of the model.

For example, suppose that $M$ agents do not report to work (and cannot be replaced) on a given day, where $M$ is a random variable that takes value $m$ with probability $q_m$, $m = 0, 1, \ldots, \nu$. Clearly, $L$ and $G(s)$ should be significantly correlated with $M$. So stratifying on $M$ (or using a one-dimensional RQMC scheme with respect to $M$ and independent random numbers elsewhere) immediately comes to mind. Moreover, $A$ is an obvious choice for a CV. Suppose we want to stratify on $M$ with an *optimal allocation*, that is, by doing $n_m$ simulation runs with $M$ fixed at $m$, where $n_0 + \cdots + n_\nu = n$, $n_m$ is approximately proportional to $q_m \sigma_m$ and $\sigma_m^2$ is the variance of the estimator of interest [$L$ or $G(s)$, with the CV $A$] conditional on $M = m$. Here, $\sigma_m^2$ will depend on the CV coefficient $\beta_m$ used in stratum $m$. The optimal $\beta_m$ will also depend on $m$. Thus, one must first estimate the optimal $\beta_m$ and the corresponding value of $\sigma_m$ for each $m$, perhaps by using a fixed fraction of the $n$ simulation runs, and then allocate the remaining runs so that the global allocation approximates the optimal one.

There are cases where we cannot control the allocation to strata. If we *poststratify* only instead of stratifying with a selected allocation, we can still optimize the coefficient $\beta_m$ within each stratum. This can be done if we use RQMC on the random variates that determine the strata, for example. This differs from the authors' setting, in which the same $\beta = \beta_{\mathrm{rqmc}}$ would be used everywhere.

Another place where stratification or RQMC would help in this application is as follows. Empirical evidence shows that a nonstationary Poisson process with *deterministic* rate function does not provide a realistic model for call arrivals to a telephone call center, because the number of calls received in any given time interval is a random variable that typically has a much larger variance than its mean. One model that better fits the data is a doubly stochastic one, where the arrival process on a given day is Poisson with rate function $R(t) = B\lambda(t)$, where $\{\lambda(t),\, t \geq 0\}$ is deterministic and $B$ is a random variable with mean 1 which can be interpreted as the *business* factor for the day. The *gamma* distribution is often a good choice for $B$ (Avramidis, Deslauriers and L'Ecuyer, 2004). Whenever the variance of $B$ is important (which is typical), one would surely want to stratify on $B$, because $L$ and $G(s)$ should be strongly dependent with

it. As a CV to be used jointly with the stratification, one may consider $H = A - E[A|B]$, with a coefficient $\beta(B)$ that depends on the value of $B$. The optimal coefficient is $\beta^*(B) = E[HL|B]/E[H^2|B]$ if the goal is to estimate $E[L]$. To estimate $\beta^*(b)$ as a function of $b$, one could estimate the two functions $q_1(b) = E[HL|B = b]$ and $q_2(b) = E[H^2|B = b]$ from the sample $\{(B_i, H_i, L_i),\ i = 1, \ldots, n\}$ of $n$ values of $(B, H, L)$, for example, using least-squares approximation to fit a curve $\hat{q}_1$ to the points $(B_i, H_i L_i)$ and another curve $\hat{q}_2$ to the points $(B_i, H_i^2)$. The ratio will estimate the function $\beta^*(b)$. In the situations where this function is far from being a constant, this could make a significant difference compared with using the same $\beta$ for all values of $B$.

### CVS FOR FUNCTIONS OF SEVERAL EXPECTATIONS

The authors have considered a setting where linear CVs are used to correct the estimator of a *single* mathematical expectation estimated by a sample average. This could be generalized to the estimation of a function of several expectations, say, $g(\mu) = g(\mu_1, \ldots, \mu_d)$ by

$$g(\hat{X}_1, \ldots, \hat{X}_d) - \beta^T(\hat{H} - \theta),$$

where $g$ is continuously differentiable at $(\mu_1, \ldots, \mu_d)$ and $\sqrt{n}(\hat{X}_1 - \mu_1, \ldots, \hat{X}_d - \mu_d)$ converges to a multinormal with mean zero when $n \to \infty$ (as in Glynn, 1994, e.g.). The asymptotically optimal $\beta$ in this case is $\beta_{\mathrm{mc}} = (\mathrm{Cov}[\hat{H}])^{-1}\,\mathrm{Cov}[\hat{H}, \hat{X}]\nabla g(\mu)$, and similarly for RQMC, where $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_d)$. In other words, in the generalization it suffices to replace $\mathrm{Cov}[\hat{H}, \hat{I}]$ with $\mathrm{Cov}[\hat{H}, \hat{X}]\nabla g(\mu)$ in (15). One simple useful example of this is the estimation of a ratio of expectations, where $g(\mu_1, \mu_2) = \mu_1/\mu_2$.

### ACKNOWLEDGMENTS

The author's work was supported by NSERC-Canada Grant ODGP0110050, NATEQ-Québec Grant 02ER3218, a Killam Research Fellowship and the Canada Research Chair in Stochastic Simulation and Optimization.

# Comment: Computation, Survey and Inference

**Xiao-Li Meng**

### 1. THE SURVEY CONNECTION

#### 1.1 Anticipating the "Surprises"

As someone who has benefited greatly from the sample survey literature, I am particularly pleased to see Hickernell, Lemieux and Owen's (HLO) emphasis on the equivalence between the control variates in Monte Carlo estimation and regression estimators in the sample survey literature. Indeed, the "surprises" described in HLO can be anticipated from similar phenomena in sample survey. For example, suppose that we, as a marketing firm, want to estimate the average household consumption of a certain product for the first six months of this year, based on a simple random sample (SRS) of a well-defined population of households (SRS is too simplistic for most practices, but adequate for the current discussion). Suppose a previous year's population counterpart is available (e.g., from a census source) for covariance adjustment (i.e., as a control variate). Let $Y$ be the variable for the current semiannual consumption and let $X$ represent the same period of the previous year. Given an SRS $\{(x_i, y_i),\ i = 1, \ldots, n\}$, asymptotically our best estimator is the well-known regression estimator

$$\hat{\mu}_y = \bar{y}_n - \hat{\beta}_{y,x}(\bar{x}_n - \mu_x), \tag{1.1}$$

where $\mu_x$ and $\mu_y$ are population averages, and $\hat{\beta}_{y,x}$ is the usual least-squares estimator from regressing $Y$ on $X$.

Suppose, however, that we discover that the population average consumption for the first quarter, denoted by $\mu_{y^{(F)}}$, can be treated as known (e.g., there was a much larger survey for the first quarter by a different marketing firm). Then we can estimate $\mu_y$ by

*Xiao-Li Meng is Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: meng@stat.harvard.edu).*

$\hat{\mu}_y^* = \mu_{y^{(F)}} + \hat{\mu}_{y^{(S)}}$, where $\mu_{y^{(S)}}$ denotes the population average for the second quarter, assuming $\{y_i^{(S)},\ i = 1, \ldots, n\}$ were available (e.g., we collected monthly consumption for the first six months). This setting mimics HLO's setting with $f(x) = f_{\mathrm{G}}(x) + f_{\mathrm{B}}(x)$, where the integration of $f_{\mathrm{G}}$ is done with no error by design, so all the estimation or integration errors come from the second component. [The analogy, of course, is not perfect because in HLO the choice of $f_{\mathrm{G}}$ depends on the design and $f_{\mathrm{G}}$ approaches $f$ (in $L^2$) as the data size increases. In sample surveys, the estimand rarely depends on the choice of designs, including the sample size. Fortunately, these differences are immaterial for our current discussion because the use of control variates is postdesign and with a given finite sample size.]

This hypothetical survey example makes it clearer that as far as the estimation of $\mu_{y^{(S)}}$ goes, neither $X$ nor $\beta_{y,x}$ is necessarily the best choice, even if they are for (1.1). It is likely that a better covariance adjustment for $Y^{(S)}$ is $X^{(S)}$, the second quarter consumption for the same previous year, perhaps due to the seasonality of the product. This is analogous to HLO's discussion in Section 4 with $f = f_{\mathrm{G}} + f_{\mathrm{B}}$ and $h = h_{\mathrm{G}} + h_{\mathrm{B}}$; since $f_{\mathrm{G}}$ and $h_{\mathrm{G}}$ do not contribute to the variance calculation, the goal is not to have $h$ correlated with $f$, but rather $h_{\mathrm{B}}$ correlated with $f_{\mathrm{B}}$. Furthermore, even if the semiannual consumption $X$ is still a better covariance adjustment for $Y^{(S)}$ because $\mathrm{Corr}^2(X, Y^{(S)}) > \mathrm{Corr}^2(X^{(S)}, Y^{(S)})$, the regression slope in (1.1) will need to be changed from $\beta_{y,x}$ to $\beta_{y^{(S)},x}$. Therefore, unless $\mathrm{Corr}^2(X, Y^{(S)}) > \mathrm{Corr}^2(X^{(S)}, Y^{(S)})$ and $\beta_{y,x} = \beta_{y^{(S)},x}$, using $\hat{\beta}_{y,x}(\bar{x}_n - \mu_y)$ to adjust $\bar{y}_n^{(S)}$ will not produce an optimal estimator. This is in agreement with HLO's summary discussion at the beginning of Section 4.

### 1.2 When Does the Wrong Optimality Hurt?

Indeed, it is also well known in the survey literature that using a nonoptimal adjustment may actually do some harm compared to no adjustment, for example, in the context of comparing ratio estimators with SRS estimators (e.g., Cochran, 1977, Chapter 6). The same survey literature inspires the following general result regarding when it becomes harmful to use a wrong optimal regression adjustment compared to making no adjustment.

LEMMA 1. *Let*

$$\hat{\theta}_{\mathrm{opt}}^{(i)} = \hat{\theta}^{(i)} - \beta_{\mathrm{opt}}^{(i)}(\hat{\psi}^{(i)} - \psi^{(i)}), \quad i = 1, 2, \tag{1.2}$$

*be two regression estimators for the same estimand $\theta$, where $\beta_{\mathrm{opt}}^{(i)} = \mathrm{Cov}(\hat{\theta}^{(i)}, \hat{\psi}^{(i)})/\mathrm{Var}(\hat{\psi}^{(i)}) > 0$ is treated as known. Let*

$$\hat{\theta}^{(1,2)} = \hat{\theta}^{(1)} - \beta_{\mathrm{opt}}^{(2)}(\hat{\psi}^{(1)} - \psi^{(1)}) \tag{1.3}$$

*be the "wrong" regression estimator, that is, it uses $\hat{\psi}^{(1)} - \psi^{(1)}$ to adjust $\hat{\theta}^{(1)}$, but with the regression slope from the other estimator. Then $\mathrm{Var}(\hat{\theta}^{(1,2)}) > \mathrm{Var}(\hat{\theta}^{(1)})$ if and only if*

$$\left|\frac{\beta_{\mathrm{opt}}^{(2)}}{\beta_{\mathrm{opt}}^{(1)}} - 1\right| > 1, \quad \textit{that is,} \quad \frac{\beta_{\mathrm{opt}}^{(2)}}{\beta_{\mathrm{opt}}^{(1)}} > 2 \quad \textit{or} \quad \frac{\beta_{\mathrm{opt}}^{(2)}}{\beta_{\mathrm{opt}}^{(1)}} < 0. \tag{1.4}$$

The proof of this lemma follows directly from the fact that

$$\mathrm{Var}(\hat{\theta}^{(1,2)}) = \mathrm{Var}(\hat{\theta}^{(1)}) - [\beta_{\mathrm{opt}}^{(1)}]^2 \mathrm{Var}(\hat{\psi}^{(1)}) + [\beta_{\mathrm{opt}}^{(2)} - \beta_{\mathrm{opt}}^{(1)}]^2 \mathrm{Var}(\hat{\psi}^{(1)}).$$

This result provides theoretical support of HLO's empirical finding that the use of $\beta_{\mathrm{MC}}$ still often leads to useful improvement with QMC, because it assures us that unless the regression slope changes substantially, that is, either it changes the sign or it is at least twice as large in magnitude, the use of the wrong regression slope is still beneficial compared to not making any adjustment, regardless of whether or not we use the same control covariate. For HLO's "cautionary example" (Section 4.1), $\beta_{\mathrm{MC}} = 1 - 2M^{-2} > 0$, but $\beta_{\mathrm{RQMC}} = -1$, so there is a switching of the sign of the regression slope. Consequently, using $\beta_{\mathrm{MC}}$ in place of $\beta_{\mathrm{RQMC}}$ will lead to an estimator with larger variance than the RQMC estimator without adjusting for the control variate. Note that in HLO's example, $\hat{\psi}^{(1)} = \hat{\psi}^{(2)}$; indeed Lemma 1 can be recast with only one regression class estimator, $\hat{\theta}_\beta = \hat{\theta} - \beta(\hat{\psi} - \psi)$, and then using a nonoptimal $\beta$ becomes harmful if and only if $|(\beta/\beta_{\mathrm{opt}}) - 1| > 1$. Also note that in real applications the regression slope is seldom known and will be replaced by its least-squares estimator. This replacement, however, does not affect the conclusion of Lemma 1 asymptotically because of the forgiving nature of the regression estimators to the error in the slope, as discussed toward the end of Section 3 of HLO.

It is also known from the survey literature that the use of regression estimators tends to have diminishing gains for stratified sample designs relative to SRS, be-

cause covariance/regression adjustment is essentially a form of (deep) stratification. Consequently, unless the two stratifying variables are uncorrelated with each other, the stratified design has already "achieved" a part of gain in efficiency intended by the regression adjustment. The degree of the "achievement" depends on how deep the original stratification is in the sampling design. Since QMC designs, especially the more advanced ones as reviewed in HLO, are often very deep stratifications (compared to the types of stratifications in sample surveys), it comes as no surprise that the gains of using control variates tend to be noticeably less pronounced for QMC than for MC, as summarized in Section 10 of HLO.

### 1.3 Why Do We Need to Go beyond the Design-Based Perspective?

The sampling survey, or more generally the design-based perspective, however, does not explain everything. Consider the following question/comparison. In the semiannual consumption example in Section 1.1 we had

$$\hat{\mu}_y = h(\hat{\mu}_{y^{(F)}}, \hat{\mu}_{y^{(S)}}) \equiv \hat{\mu}_{y^{(F)}} + \hat{\mu}_{y^{(S)}}. \tag{1.5}$$

When the true value of $\mu_{y^{(F)}}$ is known, it is almost impossible to resist the temptation to replace $\hat{\mu}_{y^{(F)}}$ with its true value in $h(\hat{\mu}_{y^{(F)}}, \hat{\mu}_{y^{(S)}})$ to form $\hat{\mu}_y^* = h(\mu_{y^{(F)}}, \hat{\mu}_{y^{(S)}}) = \mu_{y^{(F)}} + \hat{\mu}_{y^{(S)}}$ to estimate $\mu_y$. Indeed, why not? How could we get hurt, as far as efficiency/variance goes, by taking advantage of as much truth as we know?

Now consider the regression estimator given in (1.1), which can also be written as

$$\hat{\mu}_y = g(\bar{y}_n, \bar{x}_n, \hat{\beta}_{y,x}) = \bar{y}_n - \hat{\beta}_{y,x}(\bar{x}_n - \mu_x). \tag{1.6}$$

It is legitimate to consider (1.1) as a function of $\bar{y}_n, \bar{x}_n$ and $\hat{\beta}_{y,x}$ only, because only these quantities depend on the sample. Putting it differently, we can give a user a "black-box" software routine that computes the value of $\hat{\mu}_y$, with $\bar{y}_n, \bar{x}_n$ and $\hat{\beta}_{y,x}$ as input, calculated from the user's particular sample. Suppose that the user accidentally discovered that the population true value of $\mu_x$ was actually available from a census source, just as we (hypothetically) discovered that the true value of $\mu_{y^{(F)}}$ was available. Now if the user adopts the same reasoning/intuition as we did with $h$, then she or he would surely input $\mu_x$ in $g$ in place of her or his sample average $\bar{x}_n$. However, this action will completely wipe out the regression adjustment. See Liu, Rubin and Wu (1998) for a similar discussion in the context of viewing the PX–EM algorithm as a covariance adjusted EM algorithm.

One may argue that the problem occurred simply because the user did not understand the actual form of the estimator, but this is exactly the issue: For a general estimation procedure, which can be of arbitrary complexity, how can we tell when it is and when it is not beneficial to substitute a part of our estimation procedure by a more precise estimator (including its true value)? This question is particularly relevant for Monte Carlo estimators, be they quasi or not, because in a simulation setting, nothing is *unknown*, in its original sense. Consequently, the formulation of optimal estimators based on simulated data will depend intricately on how we model what we *ignore*, not what we know—a question that is beyond the realm of any design-based perspective. A different perspective is therefore needed, which is the subject of the next section. In particular, we shall see how the new perspective leads to a new interpretation of control variates and, more importantly, leads to a new control-variate estimator that appears to be difficult to anticipate from the traditional design-based perspective of Monte Carlo integration or of sample survey.

## 2. THE INFERENCE CONNECTION

### 2.1 Why Does Likelihood Inference Appear to Be Useless with Simulated Data?

To define optimality meaningfully, we first need to quantify what data and model assumptions we permit ourselves to use. In a real-data analysis, once the data are collected or provided, the central challenge typically is to postulate a suitable set of reasonable assumptions, parametric or nonparametric, to link our data with our estimand of interest. Once the model is posited and a measure of efficiency is chosen (e.g., variance), the corresponding optimality can then be quantified theoretically, at least asymptotically (e.g., via Fisher information).

The above discussion might lead us to believe that quantifying optimality with simulated data is an easier task, because there is no issue of model uncertainty, for we are the one who generated all the data (or design points). Ironically, the issue turns out to be far more complicated, precisely because we know too much. To illustrate, consider importance sampling, as discussed in HLO. We are interested in the value of $c_1 = \int_\Omega q_1(x)\mu(dx)$, where $q_1(x)$ is our known integrand and $\mu$ is the baseline measure, typically Lebesgue or counting. We have draws from a trial density $p_2 = q_2/c_2$, denoted by $\{X_{i2},\ i = 1, \ldots, n_2\}$. Then

the well-known importance sampling identity

$$r \equiv \frac{c_1}{c_2} = E_2\left[\frac{q_1(X)}{q_2(X)}\right], \tag{2.1}$$

where $E_2$ is the expectation with respect to $p_2$, provides us with an estimation equation from which we arrive at the well-known importance sampling (IS) estimator

$$\hat{r} = \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{q_1(X_{i2})}{q_2(X_{i2})}. \tag{2.2}$$

Note that in common IS settings, as in HLO, $c_2$ is chosen to be 1 and thus $r = c_1$, but in more general settings ratios are of interest; see Meng and Schilling (2002) for a recent discussion of this issue.

So on what basis can we claim (2.2) is optimal? How do we know there is no other estimation equation that can deliver a more efficient estimator than (2.1) can? Since asymptotically the maximum likelihood estimator is most efficient (under standard regularity conditions) and since asymptotic arguments are more relevant for simulated data because the size of data is under our control, we naturally wonder what the well established likelihood theory can tell us for such questions. For simplicity, let us assume that the draws from $p_2 = q_2/c_2$ are i.i.d. Then the density of our "data" $\{X_{i2},\ i = 1, \ldots, n_2\}$ is given by

$$p(X_{12}, \ldots, X_{n_2 2}) = \prod_{i=1}^{n_2}\frac{q_2(X_{i2})}{c_2}. \tag{2.3}$$

The above expression immediately suggests that something is quite amiss. On one hand, our estimand $c_1$ does not even appear in our "likelihood function" (2.3). On the other hand, it is clear that without $\{X_{i2},\ i = 1, \ldots, n_2\}$, we do not even have the IS estimator (2.2). So could this be an obvious counterexample to the likelihood principle?

Take bridge sampling as another example. Bridge sampling is a generalization of importance sampling, as described by Meng and Wong (1996). Here our goal is still to estimate $r = c_1/c_2$, as in the IS setting. The difference is that we now have draws from both $p_1 = q_1/c_1$ and $p_2 = q_2/c_2$, denoted by $\{X_{ij},\ i = 1, \ldots, n_j\},\ j = 1, 2$. Since $q_1$ and $q_2$ are assumed to be known, under the assumption of independent draws, the "likelihood" for $c_1$ and $c_2$ becomes

$$\begin{aligned} &L(c_1, c_2 | \{X_{ij},\ i = 1, \ldots, n_j\},\ j = 1, 2) \\ &\quad = \prod_{j=1}^{2}\prod_{i=1}^{n_j}\frac{q_j(X_{ij})}{c_j} \propto c_1^{-n_1}c_2^{-n_2}, \end{aligned} \tag{2.4}$$

which is free of any data! So once again, the likelihood method seems to fail, whereas estimators based on the estimation equation approach abound (see Meng and Wong, 1996).

One answer to the above paradoxes is simply that likelihood methods are not applicable to simulated data. Whereas logically this is an admissible answer, if it were true, it certainly would be the most disturbing puzzle lying in the foundation of likelihood inference, at least to some of us. How could it be? How could an inferential method so powerful with an uncertain data-generating mechanism becomes completely useless when the mechanism is completely known?

### 2.2 The Answer: Because We Were Looking at the Wrong Parameter!

An astute reader may have already seen a hidden problem with the "likelihood" as given in (2.4). The normalizing constant $c_j$ is deterministically related to $q_j$ via

$$c_j = \int_\Omega q_j(x)\mu(dx), \quad j = 1, 2. \tag{2.5}$$

So when we ignore $q_j(X_{ij})$ from (2.4) because they are known, we actually have also effectively ignored a part of the "parameter" that our likelihood intends to infer. A closer inspection of (2.5) reveals that the problem is far more serious than just appropriately sorting out the connection between $c_j$ and $q_j(X_{ij})$. The problem is that it is impossible to treat $c_j$ as an unknown parameter when we treat $q_j$ as known, unless we can treat the baseline measure $\mu$ as unknown. In other words, when we treat both $q_j$ and $\mu$ as known, there is no statistical inference problem for $c_j$ to speak of, since $c_j$ is completely determined by $q_j$ and $\mu$. Putting it differently, although $c_j$'s or their ratios are what we are after, they cannot be the *only unknown* model parameters for any meaningful statistical modeling.

To resolve this problem, Kong, McCullagh, Meng, Nicolae and Tan (2003) proposed to conduct the likelihood inference by treating the baseline measure $\mu$ as the unknown parameter and then to estimate $c_j$ as a linear functional of $\mu$ via (2.5). With this approach, (2.3) becomes a well-defined and meaningful likelihood in the form of

$$\begin{aligned} &L(\mu | X_{12}, \ldots, X_{n_2 2}) \\ &\quad = \prod_{i=1}^{n_2}\frac{q_2(X_{i2})\mu(X_{i2})}{\int q_2(x)\mu(dx)} \propto \frac{\prod_{i=1}^{n_2}\mu(X_{i2})}{[\int q_2(x)\mu(dx)]^{n_2}}, \end{aligned} \tag{2.6}$$

where $\mu(X) = \mu(\{X\})$ or $\mu(\{dX\})$. The maximum likelihood estimator of $\mu$, among all possible nonnegative measures, is given by $\hat{\mu}(x) \propto P_{n_2}(x)/q_2(x)$, where $P_{n_2}(x)$ is the usual empirical measure, with $n_2^{-1}$ mass at each observed $X_{i2}$. Clearly from (2.6), $\mu$ (and thus $c_j$'s) can only be estimated up to a multiplicative constant. Substituting $\mu$ in (2.5) with $\hat{\mu}$ shows that $\hat{r}$ of (2.2) is indeed the (nonparametric) maximum likelihood estimator (MLE) of $r$ under the likelihood (2.6). This suggests that, without employing any other information, $\hat{r}$ of (2.2) is indeed (asymptotically) the best possible estimator of $r$ given $\{X_{i2},\ i = 1, \ldots, n_2\}$. Similarly, Kong et al. (2003) have shown that the optimal bridge sampling estimator given in Meng and Wong (1996) is the same as the MLE when we have $\{X_{ij},\ i = 1, \ldots, n_j;\ j = 1, 2\}$ as our data.

The reason why this likelihood perspective can easily resolve these paradoxes is that it captures the real inference structure of Monte Carlo integration. Specifically, Monte Carlo simulation means that we use *samples* to represent, and therefore effectively *estimate*, the underlying population $q_j(x)\mu(dx)$, and hence *estimate* $\mu$ since $q_j$ is known. One may find the phrase "estimate" puzzling because we invariably know what $\mu$ is (e.g., Lebesgue or counting). However, our knowledge of $\mu$ is never used in any way, for example, in forming (2.2). This can be best seen by considering that there are two individuals: a simulator and an analyst. The simulator provides the simulated data $\{X_{i2},\ i = 1, \ldots, n_2\}$ to the analyst, who has the task of estimating $r$. The analyst is also given both $q_1$ and $q_2$, but is never told about the actual $\mu$ used in simulation. Nevertheless, the analyst can consistently estimate $r$, which obviously depends on $\mu$, as long as the support of $q_1$ does not exceed that of $q_2$. (This well-known condition on the supports can also be clearly seen from the likelihood perspective, because we can only make inference about $\mu$ on a support that is identifiable from the data $\{X_{i2},\ i = 1, \ldots, n_2\}$.) Consequently, as far as (2.2) goes, $\mu$ is completely unknown; more precisely, no knowledge of $\mu$ is used in (2.2) and thus it is legitimate (and actually necessary) to treat $\mu$ as the unknown model parameter.

The above discussion also suggests that we can use partial knowledge of $\mu$ to improve upon (2.2), as long as the resulting MLE for $r$ is still easy to compute. Clearly we should not use our full knowledge about $\mu$, which will lead us back to the infeasible analytic calculation required by (2.5). For example, since Lebesgue measure is invariant to reflection with respect to the origin, we can restrict our parameter space to all nonnegative measures that satisfy this invariance property, if the true $\mu$ is indeed Lebesgue. The resulting MLE of $r$ is

$$\hat{r}^* = \frac{1}{n_2}\sum_{i=1}^{n_2} \frac{q_1(X_{i2}) + q_1(-X_{i2})}{q_2(X_{i2}) + q_2(-X_{i2})}, \tag{2.7}$$

which is the Rao–Blackwellization treatment of $\hat{r}$ by averaging over the orbit of the reflection group $\{I, -I\}$, and hence its variance never exceeds that of $\hat{r}$ (under the assumption of i.i.d. draws). See Kong et al. (2003) for a general formulation of using group invariance to restrict the parameter space for $\mu$ and hence to improve Monte Carlo efficiency. Also see Casella (1996) for a detailed discussion of the use of Rao–Blackwellization methods in Monte Carlo simulation and, more generally, the interrelationship between statistical inference theory and computational algorithms.

### 2.3 Indeed a Surprise: An Unexpected Control-Variate Estimator and Insight

Another fundamental advantage of this likelihood approach is that it provides a unified framework for investigating variance reduction techniques, including control variates. In the importance sampling context, when we use a $g$ with

$$\int_\Omega g(x)\mu(dx) = 0 \tag{2.8}$$

as a control variate, we effectively put a constraint on the unrestricted parameter space $\Theta_\mu = \{\mu : \text{all nonnegative measures on } \Omega\}$. Consequently, the MLE under this submodel will be more efficient than the MLE under the full model. The resulting MLE for $r$ under this constraint, however, is not the usual regression estimator, albeit asymptotically they are equivalent, as they should be.

Specifically, because any measure with zero mass at any single observation will lead to a zero likelihood in (2.6), the maximization of (2.6) under constraint (2.8) is effectively discrete, as is typical with nonparametric or empirical MLE (e.g., Owen, 2001). The discrete problem we need to solve is

$$\max_{\mu \in \Theta_{n_2}^{(g)}} \left\{ \sum_{i=1}^{n_2} \log(\mu_i) - n_2 \log\left[ \sum_{i=1}^{n_2} q_{2i}\mu_i \right] \right\}, \tag{2.9}$$

where, for simplicity, we have let $\mu_i = \mu(X_{i2})$,

$q_{2i} = q_2(X_{i2})$, $g_i = g(X_{i2})$ and

$$\Theta_{n_2}^{(g)} = \Bigg\{(\mu_1, \ldots, \mu_{n_2}) : \mu_i > 0,\ i = 1, \ldots, n_2; \text{ and } \sum_{i=1}^{n_2} g_i \mu_i = 0\Bigg\}. \tag{2.10}$$

Tan (2003) presented an elegant solution to this maximization problem under the more general setting with multiple control variates. The following is a slightly more elementary recast of Tan's (2003) derivation.

We start by assuming condition (A): $\min_i g_i < 0$ and $\max_i g_i > 0$. This is not a real restriction in view of (2.8) and relatively large $n_2$ in practice, but technically it is a necessary and sufficient condition for (2.9) to have a solution. Clearly it is necessary, because without it, $\Theta_{n_2}^{(g)}$ will be empty. The sufficiency is established by the following argument, which shows that (2.9) has the unique maximizer when condition (A) holds.

First, because $\sum_{i=1}^{n_2} g_i \mu_i = 0$, (2.9) is the same as

$$\max_{\mu \in \Theta_{n_2}^{(g)}} \Bigg\{ \sum_{i=1}^{n_2} \log(\mu_i) - n_2 \log\Bigg[ \frac{1}{n_2} \sum_{i=1}^{n_2} (q_{2i} + \lambda g_i)\mu_i \Bigg] - n_2 \log n_2 \Bigg\} \tag{2.11}$$

for any $\lambda \in \Lambda_{n_2} = \{\lambda : q_{2i} + \lambda g_i > 0,\ i = 1, \ldots, n_2\}$, which is nonempty because it contains at least $\lambda = 0$ since all $q_{2i} > 0$ by our sample design. Consequently, by Jensen's inequality applied to the second log expression in (2.11), we obtain

$$\begin{aligned} &\max_{\mu \in \Theta_{n_2}^{(g)}} \Bigg\{ \sum_{i=1}^{n_2} \log(\mu_i) - n_2 \log\Bigg[ \sum_{i=1}^{n_2} q_{2i} \mu_i \Bigg] \Bigg\} \\ &\quad \le - \sum_{i=1}^{n_2} \log(q_{2i} + \lambda g_i) - n_2 \log n_2, \end{aligned} \tag{2.12}$$

where the equality holds if and only if

$$\mu_i \propto \frac{1}{q_{2i} + \lambda g_i} \quad \text{and} \quad \sum_{i=1}^{n_2} g_i \mu_i = 0. \tag{2.13}$$

Since (2.12) holds for any $\lambda \in \Lambda_{n_2}$, we can minimize the right-hand side over $\lambda$, which leads to

$$\begin{aligned} &\max_{\Theta_{n_2}^{(g)}} \Bigg\{ \sum_{i=1}^{n_2} \log(\mu_i) - n_2 \log\Bigg[ \sum_{i=1}^{n_2} q_{2i} \mu_i \Bigg] \Bigg\} \\ &\quad \le - \max_{\lambda \in \Lambda_{n_2}} \sum_{i=1}^{n_2} \log(q_{2i} + \lambda g_i) - n_2 \log n_2. \end{aligned} \tag{2.14}$$

Second, we can show that the inequality in (2.14) actually is an equality. This is because, under condition (A), $\Lambda_{n_2}$ is a finite open interval containing zero and

$$\ell(\lambda) \equiv \sum_{i=1}^{n_2} \log(q_{2i} + \lambda g_i) \tag{2.15}$$

is a strict concave and differentiable function on $\Lambda_{n_2}$. Consequently, $\ell(\lambda)$ has the unique maximum $\hat{\lambda} \in \Lambda_{n_2}$, which satisfies

$$\frac{d\ell(\hat{\lambda})}{d\lambda} = \sum_{i=1}^{n_2} \frac{g_i}{q_{2i} + \hat{\lambda} g_i} = 0. \tag{2.16}$$

In other words, when we let $\lambda = \hat{\lambda}$ in (2.13), the resulting $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_{n_2})$ indeed satisfies the constraint in (2.13), and therefore this, and only this, choice of $\mu$ equates the two sides of (2.14). Consequently,

$$\hat{\mu}(x) \propto \frac{P_{n_2}(x)}{q_2(x) + \hat{\lambda} g(x)} \tag{2.17}$$

is the unique solution to (2.9), where $P_{n_2}(x)$ is the standard empirical measure based on $\{X_1, \ldots, X_{n_2}\}$. The corresponding MLE of $r$ is given by

$$\hat{r}_{\mathrm{MLE}} = \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{q_1(X_{i2})}{q_2(X_{i2}) + \hat{\lambda} g(X_{i2})}. \tag{2.18}$$

The form of this MLE is rather intriguing. First, unlike the standard regression estimator, which takes a linear form for adjustment, $\hat{r}_{\mathrm{MLE}}$ retains a ratio form. The advantage of the ratio form is that it ensures the nonnegativity of $\hat{r}_{\mathrm{MLE}}$ whenever the integrand $q_1$ is nonnegative. This is, of course, expected because $\hat{r}_{\mathrm{MLE}}$ is an MLE and hence it must be within the original allowable space of $r$ (as determined by our usable knowledge of $q_1$). In contrast, the regression estimator does not have this property. Asymptotically, however, linear adjustment is all one needs, and thus $\hat{r}_{\mathrm{MLE}}$ is equivalent to the regression estimator by a Taylor expansion argument, as given in Tan (2003).

Second, $\hat{r}_{\mathrm{MLE}}$ has the same form as the IS estimator (2.2), but with $q_2(x) + \hat{\lambda} g(x)$ as the "trial" density. This can be seen more clearly when our control variate is introduced by using an unnormalized density $q_3$ such that $\int q_2(x)\mu(dx) = \int q_3(x)\mu(dx)$ (see Kong et al., 2003, for an illustration), that is, $g(x) = q_3(x) - q_2(x)$. Then the function in the denominators in (2.18) becomes a mixture of $q_2$ and $q_3$, $(1 - \hat{\lambda})q_2 + \hat{\lambda} q_3$, where $\hat{\lambda}$ is the MLE of the mixture weight $\lambda$ from fitting the mixture model $(1 - \lambda)q_2 + \lambda q_3$ to the simulated data

$\{X_{i2},\ i = 1, \ldots, n_2\}$. (Note that here $\lambda$ is not restricted to the unit interval, as long as it is inside $\Theta_{n_2}^{(g)}$.)

This fitting aspect is the most intriguing part of the MLE approach because the true value of $\lambda$ is known to be zero, since all the data were drawn from $q_2$. However, with any finite sample, the best fitted $\hat{\lambda}$ under the mixture model will almost surely deviate from the true value $\lambda = 0$, indicating an "imperfection" of the sample to represent the intended population $q_2$. The MLE approach uses this deviation to adjust for the imperfection via the known relationship (2.8), in the same spirit as the regression estimator uses $\bar{x}_n - \mu_x$ to adjust. Specifically, just as the regression estimator (1.1) effectively treats an "imperfect" sample $\{y_1, \ldots, y_n\}$ with mean $\mu_y$ as a "perfect" sample with mean $\mu_y + \beta_{y,x}(\bar{x}_n - \mu_x)$, the MLE treats an imperfect sample from $q_2$ as a perfect sample from $(1 - \hat{\lambda})q_2 + \hat{\lambda}q_3$: It is perfect as far as estimating $\int_\Omega g(x)\mu(dx) = 0$ goes because of (2.16). The MLE then uses this "perfect" model/sample to perform the usual importance sampling, as in (2.18). This construction appears to be difficult to conceive from a purely design-based perspective, which inevitably would only call for inverse-probability weight $1/q_2(X)$, since $X$ was drawn from $q_2$. In particular, this is another example where the use of the fitted value is better than using the truth, as discussed in Section 1.3.

### 2.4 Possible Applications to QMC and Surveys

The discussion so far centers on MC designs, where there is a natural sampling distribution and hence a natural likelihood. The central issue there is to recognize what the correct model parameter is. For deterministic QMC, this approach is not directly applicable since there is no sampling distribution in the design. However, when randomness is reintroduced into QMC, as with the RQMC methods discussed in HLO, the likelihood method appears to be applicable, albeit the implementation could be more complicated in view of the more stratified nature of the design compared to i.i.d. or even the more general MCMC designs, which are typically without stratification. In addition, there appear to be more constraints on $\mu$ such as $\int f_G(x) f_B(x)\mu(dx) = 0$ with the QMC methods (Section 2.1 of HLO). It would be interesting to see the form of the resulting MLE for $\int [f_G(x) + f_B(x)]\mu(dx)$ under the likelihood approach.

For deterministic QMC, although the likelihood approach is not directly applicable (and this time there is no paradox, because there is no random data-generating mechanism to start with), the inference perspective is still very fruitful. This was, for example, discussed by Diaconis (1988), where a Bayesian approach, which does not necessarily require a sampling scheme or a likelihood, was investigated. This approach is to put a prior model—a stochastic process—on the integrand $q$, with $q$'s values at the design points as the observations. The inference is then carried out by computing the conditional distribution of the process, and hence the integration, given the observations. The advantage of this class of methods is that, by choosing appropriate stochastic models, one can take into account known properties of the integrand $q$. In contrast, our likelihood approach takes advantage of usable known properties of the baseline measure, either via group restrictions or other constraints such as control variates. As a result, the Bayesian approach can produce much more efficient results for specific integrands. Indeed, many well-known numerical integration methods can be rederived from this perspective, as shown by Diaconis (1988) and the references therein. On the other hand, the MLEs obtained under the likelihood approach are much more generally applicable, but they can be made more efficient if specific knowledge of the integrand (e.g., differentiability) can be utilized. So the two approaches complement each other and, ideally, we would like to have a combined inference method that will model the usable knowledge of both the baseline measure and the integrand. Research in this direction is very much needed, and HLO's investigation of using control variates with RQMC methods can be viewed as an important step in this direction because it takes into account both the properties of the integrand and the restriction on the baseline measure via the use of the control variates.

Finally, to complete the circle, the new ratio-type control-variate estimator also suggests a possible corresponding counterpart for sample survey applications, where the two standard estimators for covariance adjustments have been the direct ratio estimator [i.e., $\hat{\mu}_y = (\bar{y}_n/\bar{x}_n)\bar{\mu}_x$] and the regression estimator (1.1). Such a counterpart, if it exists, would be of direct practical value, because it retains important advantages of both the ratio estimator and the regression estimator, as we discussed in Section 2.3, especially considering that many survey estimands are positive by nature.

## 3. FURTHER CONNECTIONS BETWEEN MCMC AND QMC

As HLO correctly pointed out in their Section 2.5, both MCMC and QMC have a long history and both

have grown rapidly in recent years, yet there is very little overlap between the two fields. This is certainly a very unfortunate and ironic situation, considering that both fields share exactly the same goal. HLO's paper is certainly a very timely contribution to changing this situation—a change that is much needed, because the two fields can learn a great deal from each other, as HLO's paper clearly demonstrates. Here I want to add two topics from recent work that I was involved in to demonstrate the great benefit of using techniques and ideas from both fields.

The first topic is path sampling, which is a generalization of bridge sampling with infinitely many bridges, as well as a general formulation of thermodynamic integration in statistical physics, as shown by Gelman and Meng (1998). The method is particularly suited for handling some very high-dimensional integrations, as discussed by Ogata (1989). The key identity that underlies path sampling expresses $\log r$, where $r$ is the same as in (2.1), as a low-dimensional integration over a prior parameter of a high-dimensional expectation that is conditional on the parameter. This presents an ideal situation to use both MCMC methods and QMC methods, with the former applied to estimate the high-dimensional expectation and the latter applied to numerically estimate the outside low-dimensional integration. The effectiveness of such a hybrid approach was demonstrated by Gelman and Meng (1998), where very basic numerical approaches (e.g., trapezoidal rule; rectangular lattices) were used for the low-dimensional integrations. It is likely that the effectiveness will be even more impressive if the more advanced QMC methods, such as those reviewed in HLO, are used for these low-dimensional integrations.

The second topic is multiprocess parallel antithetic coupling for backward and forward MCMC (Craiu and Meng, 2005). Using antithetic variates is a very old variance reduction technique in the Monte Carlo literature (e.g., Hammersley and Morton, 1956). However, in the standard MCMC literature, typically only a pair of antithetic variables is used (e.g., Frigessi, Gåsemyr and Rue, 2000). Viewing antithetic variates as a form of stratification, employing more than two strata becomes an obvious next step. However, unlike the case of using a pair, generating a set of $k > 2$ antithetic variates is not a trivial task. This is because there is no unique way to generate $k > 2$ antithetic variates that are *negatively associated* (i.e., preserve negative correlation under monotone transformation) and *extremely antithetical* (i.e., as negatively correlated as possible). Nevertheless, we (Craiu and Meng, 2005) found that Latin hypercube sampling, as mentioned in Section 6 of HLO, as well as an iterative extension of it, serves as an effective general-purpose scheme. The advantages of running multiprocess antithetically coupled MCMC, for both the standard forward implementation and the backward perfect-sampling implementation (see Casella, Lavine and Robert, 2001, for an introduction), include not only further reduction of Monte Carlo variances compared to using $k = 2$, but also reduction of biases due to slow mixing, because antithetically coupled chains can search a state space more thoroughly compared with using $k$ independent chains, which is the current common recommendation (e.g., Gelman and Rubin, 1992).

In conclusion, I thank HLO for writing this timely and inspiring article and the Editor for inviting me to discuss it. Given the clear benefit of cross-fertilization between MCMC and QMC, I hope this set of discussion articles can serve as a successful matchmaker for a long, happy and (re)productive marriage between QMC and MCMC!

### ACKNOWLEDGMENTS

I thank Radu Craiu, Andrew Gelman, Martin Romero and Zhiqiang Tan for helpful comments and exchanges. The research was supported in part by NSF Grant DMS-02-04552.

# Rejoinder

**Fred J. Hickernell, Christiane Lemieux and Art B. Owen**

We thank Professor L'Ecuyer and Professor Meng for their thoughtful remarks. We particularly liked L'Ecuyer's concise summary of combining variance reduction techniques, and Meng's references to combinations of antithetic and Latin hypercube sampling with MCMC. Our reply is organized by topic.

## EFFICIENCY

In the Asian option example only a small inefficiency arose from estimating $\beta_{\text{mc}}$ instead of $\beta_{\text{rqmc}}$. L'Ecuyer asks whether this will be rare or frequent. As he notes, experience will tell. We expect that the inefficiency will very often be small. The cost of having the wrong coefficient is quadratic in the coefficient error, so small coefficient errors are largely forgiven.

From Meng's comment, we learn that even some large coefficient errors have mild effects: for a scalar $\beta \neq \beta_{\text{opt}}$ to give a larger variance than $\beta = 0$ gives, you have to either get the sign wrong or have $|\beta| > 2|\beta_{\text{opt}}|$. Unless $\beta_{\text{opt}}$ is close to zero, there is a wide window to aim for.

An inefficiency that is often small can also be often large. Moreover, one can construct worst-case problems for which the inefficiency is arbitrarily large. Fortunately, in practice one can estimate the error variance both with and without the control variate.

## CONFIDENCE INTERVALS

L'Ecuyer points out that things would be easier for small $R$ if $(\hat{I}_r, \hat{H}_r)$ had a multivariate Gaussian distribution. In the case of scrambled nets, the central limit theorem of Loh (2003) gives reason to suppose that a multivariate Gaussian distribution would be a good approximation. On the other hand, Loh's theorem allows for an extremely slow rate of convergence to the Gaussian distribution. The magnitude of the bias L'Ecuyer mentions is an interesting open issue.

We usually prefer a small $R$, supposing that accuracy in estimating $\hat{I}$ is more important than accuracy in estimating error. That assumption is not valid for all applications. In such cases one can use a larger $R$ to get a more reliable confidence interval around $\hat{I}$ at the cost of less accurate estimation of $\hat{I}$.

## STRATIFICATION

As L'Ecuyer points out, we have used only a single control variate coefficient vector, while in stratified sampling one often prefers to use a different coefficient within each stratum. For scrambled nets, the number of strata is equal to the number of sample values $n$ when $t = 0$ and $s = 1$. For $s > 1$, the number of simultaneously balanced strata can be much larger than $n$. At these extremes, one cannot afford to estimate one coefficient per stratum by least squares.

Quasiregression with coefficient shrinkage is an alternative to least squares that allows for the number of control variates to be larger than $n$. The intercept term in a quasiregression is an estimate of the integral. See Jiang and Owen (2003) and Jiang (2003) for details of quasiregression with plain Monte Carlo methods.

L'Ecuyer's idea of breaking the problem into pieces each with its own control variate seems to be a good compromise between using a single coefficient and using $O(n)$ or more coefficients. The call center example seems well suited to multiple control variates.

It should be possible to incorporate some applications of poststratification and stratum-specific coefficients into the framework of this paper. If there are two strata, taking $h_1$ to be an indicator function for one of those strata captures the benefits of poststratification. Then for a second variable $h_2$, putting $h_3(x) = h_1(x)h_2(x)$ captures the benefits of stratum-specific coefficients.

## NONPARAMETRIC LIKELIHOOD

It is intriguing to see how nonparametric and empirical likelihood can be used in Monte Carlo problems. Meng advocates treating the baseline measure as unknown instead of the parameter. Of course, in the mathematical sense the baseline measure is at least as known as the parameter. The test is whether this point of view helps us to solve problems, and it appears to do so.

Once again, this is a setting where survey sampling researchers have been active. A survey of empirical likelihood methods for complex survey samples appeared in Owen (2001, Chapter 8, Sections 5–8). Key contributions were made by Jing Qin, Jiahua Chen, Randy Sitter, Changbao Wu, Bob Zhong and Jon Rao.

## BAYESIAN CONNECTION

As mentioned by Meng, the Bayesian approach, described by Diaconis (1988), has proven quite useful in studying the problem of integration. Ground-breaking work was done by Sacks and Ylvisaker (1966, 1968, 1970a, b). A comprehensive survey of classical and recent results was given by Ritter (2000). The Bayesian approach has a long history: Diaconis (1988) traced it to Poincaré. That approach is also well suited to approximation.

In the Bayesian approach, the values of the random integrand at two different positions are described by the covariance kernel. The smoothness of this kernel then affects the convergence rate of the numerical integration algorithm. Error analysis in the Bayesian setting

has many parallels to the worst-case analysis for deterministic integrands. In the latter case, the Hilbert space of integrands is often defined by its reproducing kernel (Wahba, 1990; Hickernell, 2000). The error measures for linear numerical integration rules are the same in the Bayesian and worst-case settings when the kernels are the same. However, for the same kernel, the Hilbert space of integrands in the worst-case setting typically corresponds to a subset of measure zero in the space of random integrands in the Bayesian setting. This is due to the worst-case setting's more conservative or pessimistic approach.

## PRIOR KNOWLEDGE

We agree with Meng that it is tricky to know how best to use prior knowledge. Here is a particularly simple example. Suppose that $\theta = \mu_y - \mu_z$, where $(\mu_y, \mu_z) = E((Y_i, Z_i))$. One might naturally use $\hat{\theta} = \bar{Y} - \bar{Z}$. Now suppose $\mu_z$ is known. Then $\bar{Y} - \mu_z$ might be attractive. Of course if $Y_i = Z_i + \theta$, then the original $\hat{\theta}$ has variance zero while the proposed improvement can have infinite variance.

In general we cannot rule out side information that would make (2.2) far from optimal. To take an extreme example, suppose we know that $c_1/c_2$ either equals our phone number or our fax number. Somebody working with this knowledge can do much better than somebody else. It is quite hard to draw a sharp line between side information that we can assume will not be available and side information that might well be available.

We agree with Meng that it is legitimate to use a model in which the baseline measure is treated as an unknown. In examples it leads to sensible answers that can be tested. We do not see how it could be necessary. People using classical Monte Carlo and quasi-Monte Carlo methods can get reliable results and they do so without treating the baseline measure as unknown. As a case in point, the Bayesian methods described above work conditionally on the sample points under a model in which the baseline measure is known, but the target function has a Gaussian distribution. "There's more than one way to skin a cat." No animals were harmed in this rejoinder.

## ADDITIONAL REFERENCES


AVRAMIDIS, A. N., DESLAURIERS, A. and L'ECUYER, P. (2004). Modeling daily arrivals to a telephone call center. *Management Sci.* **50** 896–908.

AVRAMIDIS, A. N. and WILSON, J. R. (1996). Integrated variance reduction strategies for simulation. *Oper. Res.* **44** 327–346.

CASELLA, G. (1996). Statistical inference and Monte Carlo algorithms (with discussion). *Test* **5** 249–344.

CASELLA, G., LAVINE, M. and ROBERT, C. P. (2001). Explaining the perfect sampling. *Amer. Statist.* **55** 299–305.

CRAIU, R. and MENG, X.-L. (2005). Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo. *Ann. Statist.* **33** 661–697.

DIACONIS, P. (1988). Bayesian numerical analysis. In *Statistical Decision Theory and Related Topics IV* (S. Gupta and J. O. Berger, eds.) **1** 163–175. Springer, Berlin.

FRIGESSI, A., GÅSEMYR, J. and RUE, H. (2000). Antithetic coupling of two Gibbs sampler chains. *Ann. Statist.* **28** 1128–1149.

GELMAN, A. and MENG, X.-L. (1998). Computing normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7** 457–511.

GLYNN, P. W. (1994). Efficiency improvement techniques. *Ann. Oper. Res.* **53** 175–197.

GLYNN, P. W. and SZECHTMAN, R. (2002). Some new perspectives on the method of control variates. In *Monte Carlo and Quasi-Monte Carlo Methods 2000* (K.-T. Fang, F. J. Hickernell and H. Niederreiter, eds.) 27–49. Springer, Berlin.

HAMMERSLEY, J. M. and MORTON, K. V. (1956). A new Monte Carlo technique: Antithetic variates. *Proc. Cambridge Philos. Soc.* **52** 449–475.

HICKERNELL, F. J. (2000). What affects the accuracy of quasi-Monte Carlo quadrature? In *Monte Carlo and Quasi-Monte Carlo Methods 1998* (H. Niederreiter and J. Spanier, eds.) 16–55. Springer, Berlin.

JIANG, T. (2003). Data driven shrinkage strategies for quasi-regression. Ph.D. dissertation, Dept. Statistics, Stanford Univ.

JIANG, T. and OWEN, A. B. (2003). Quasi-regression with shrinkage. *Math. Comput. Simulation* **62** 231–241.

KONG, A., MCCULLAGH, P., MENG, X.-L., NICOLAE, D. and TAN, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 585–618.

LIU, C., RUBIN, D. B. and WU, Y. (1998). Parameter expansion to accelerate EM: The PX–EM algorithm. *Biometrika* **85** 755–770.

MENG, X.-L. and SCHILLING, S. (2002). Warp bridge sampling. *J. Comput. Graph. Statist.* **11** 552–586.

MENG, X.-L. and WONG, W. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860.

OGATA, Y. (1989). A Monte Carlo method for high-dimensional integration. *Numer. Math.* **55** 137–157.

OWEN, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC Press, Boca Raton, FL.

PICHITLAMKEN, J., DESLAURIERS, A., L'ECUYER, P. and AVRAMIDIS, A. N. (2003). Modeling and simulation of a telephone call center. In *Proc. 2003 Winter Simulation Conference* **2** 1805–1812. IEEE Press, New York.

RITTER, K. (2000). *Average-Case Analysis of Numerical Problems*. *Lecture Notes in Math.* **1733**. Springer, Berlin.

SACKS, J. and YLVISAKER, D. (1966). Designs for regression problems with correlated errors. *Ann. Math. Statist.* **37** 66–89.

SACKS, J. and YLVISAKER, D. (1968). Designs for regression problems with correlated errors; many parameters. *Ann. Math. Statist.* **39** 49–69.

SACKS, J. and YLVISAKER, D. (1970a). Design for regression problems with correlated errors. III. *Ann. Math. Statist.* **41** 2057–2074.

SACKS, J. and YLVISAKER, D. (1970b). Statistical designs and integral approximation. In *Proc. Twelfth Biennial Seminar of the Canadian Mathematics Congress* (R. Pyke, ed.) 115–136. Canadian Math. Soc., Montreal.

TAN, Z. (2003). A likelihood approach for Monte Carlo integration. Ph.D. dissertation, Dept. Statistics, Univ. Chicago.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.