Control variates for quasi-Monte Carlo

Fred J. Hickernell
Hong Kong Baptist University

Christiane Lemieux
University of Calgary

Art B. Owen
Stanford University

This: April 2003
Orig: August 2001

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.

Keywords: 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 $X_1, \ldots, X_n$ having the $U[0,1)^s$ distribution. Then the MC estimate of I is

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

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

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

Many techniques have been developed for improving 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) discuss combinations of importance sampling with QMC and mention 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 arising 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 sub-hypercubes.

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 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) and 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 ∇g, taken as an s dimensional row vector. For a column vector h of J functions on $[0,1)^s$, the gradient ∇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 one by the strong law of large numbers. Quasi-Monte Carlo sampling may be thought of as a way of getting 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

QMC 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 coordinate-wise, let $[a,b)$ be the s-dimensional box of points $X \in [0,1)^s$ for which $a \le X < b$ holds coordinate-wise, and let $\mathrm{vol}([a,b))$ be the s-dimensional volume of that box. For $X_i \in [0,1)^s$ with $1 \le i < \infty$, the sequence (X_i) is *uniformly distributed* in $[0,1)^s$ if $\lim_{n\to\infty}(1/n)\sum_{i=1}^n 1_{a\le X_i<b} = \mathrm{vol}([a,b))$ for all $0 \le a < b \le 1$.

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 uniform distribution provides a deterministic analogue of the law of large numbers. Though 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 $\lim_{n\to\infty}(1/n)\sum_{i=1}^n e^{2\pi\sqrt{-1}k^TX_i} = 0$ 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 $X_1,\dots,X_n$ is defined as

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

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 $i = 1,\dots,n$ 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 $D_n^* \le D_n \le 2^sD_n^*$, 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 by an L^p norm as follows

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

for $p \geq 1$ with $p = 2$ the most widely studied. Beck and Chen (1987) and Matoušek (1998a) provide 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 considers integrated indicators of anchored boxes, and in Hickernell (1996), who considers 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

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

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 $D_n^*(X_1, \ldots, X_n) = O(n^{-1}(\log n)^s)$ 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 $D_n^*(X_1, \ldots, X_n) \geq C_s n^{-1}(\log n)^{s/2}$ 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 (Caflisch and Morokoff 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 $1 \leq i \leq n < \infty$ for which $D_n^*(X_{n1}, \ldots, X_{nn})$ 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. An infinite sequence, by constrast, is necessarily extensible. There are many links between extensible rules in s dimensions and non-extensible ones in $s+1$ dimensions. Matoušek (1998a, Chapter 1) discusses 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 $f(x) = f_G(x) + f_B(x)$. Here f_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_B on which it does badly. The definitions of f_G and f_B differ for nets and lattices and depend on the sample size n. As n increases, $\int (f(x) - f_G(x))^2 dx \to 0$. For each method $\int f_G(x) f_B(x) dx = 0$ 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, is given in Niederreiter (1992). This section presents brief formal definitions of (t, m, s)–nets, (t, s)–sequences, and (λ, t, m, s)–nets.

The following geometric discussion may be helpful for the reader 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 $k \geq 0$ and integer base $b \geq 2$. Given n points $X_1, \ldots, X_n$ 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*

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

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 $1/25 \times 1/5$. Though 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 2 points of the extended sequence. Further, some net constructions are extensible, not just two-fold 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:

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 λb^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 λb^m points of a (t, s)-sequence in base b are a (λ, t, m, s)–net in base b, for any $m \geq t$ and $1 \leq \lambda < b$.

Definition 4 *Let m, t, λ be integers with $m \geq t \geq 0$, and $1 \leq \lambda < b$. A sequence of λb^m points in $[0, 1)^s$ is called a (λ, t, m, s)–net in base b if every b-ary box of volume b^{t-m} contains λb^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 (van der Corput 1935a; van der Corput 1935b). For integer base $b \geq 2$, let the non-negative 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 non-negative integers, all b^m possible trailing digits appear exactly once. Then the corresponding values of ϕ_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) presents constructions of $(0, p)$-sequences in prime bases p,

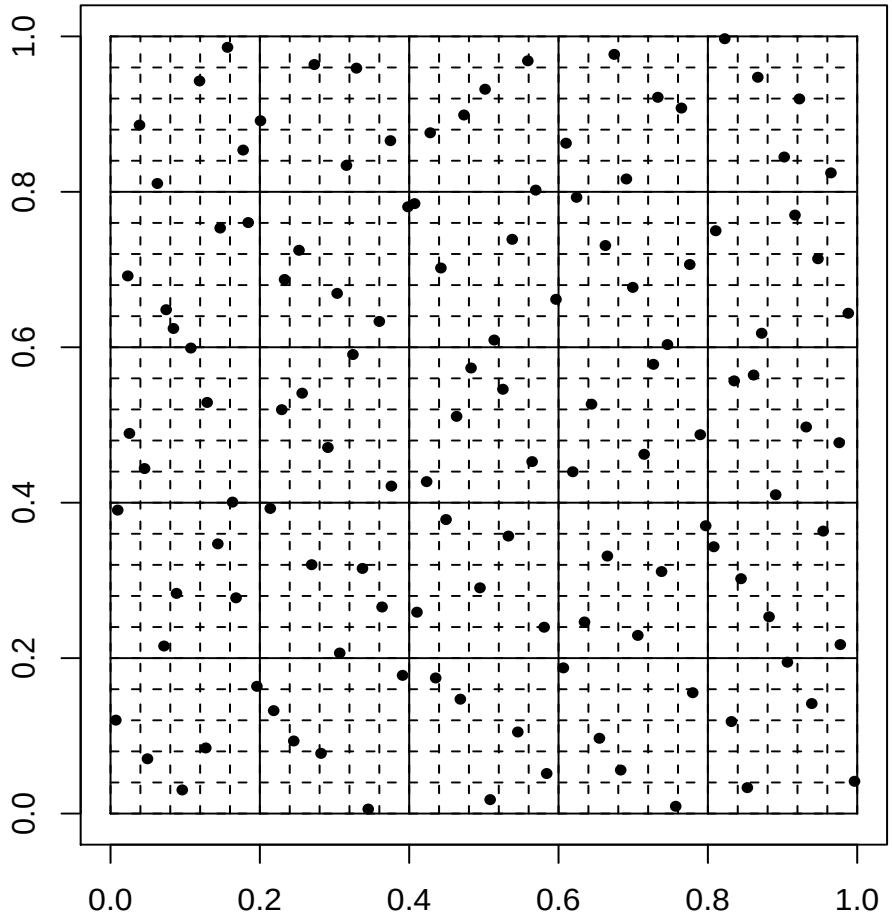

Figure 1: Shown are 125 points of a digital net in base 5 as described in the text.

and Sobol' (1967) constructs (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 $1_{\mathcal{B}}(x)$ 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 $X_1, \ldots, X_n$

are a (λ, t, m, s)–net in base b with $m - t \geq K$, then $(1/n)\sum_{i=1}^{n} 1_{\mathcal{B}}(X_i) = \int 1_{\mathcal{B}}(x)dx$. The points of a (λ, 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 (λ, 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 43570 different 5-ary boxes of volume $1/625$.

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

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 $\left\{\sum_{j=1}^{s} \alpha_j v_j \mid \alpha_j \in \mathbb{Z}\right\}$ where $v_1, \ldots, v_s$ 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 $\tau = (\tau_1, \ldots, \tau_s)$ of nonnegative integers. Then for $i = 1, \ldots, n$ let

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

where $z \mod 1 = z - \lfloor z \rfloor$ 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 τ. Lattice rules of ranks 1 through s are described in Sloan and Joe (1994). We will emphasize rank 1 rules here. The lattice rules of Korobov (1959) are rank 1 rules for which $\tau = (1, \eta, \eta^2, \ldots, \eta^{s-1})$ for some $\eta \in \mathbb{Z}$.

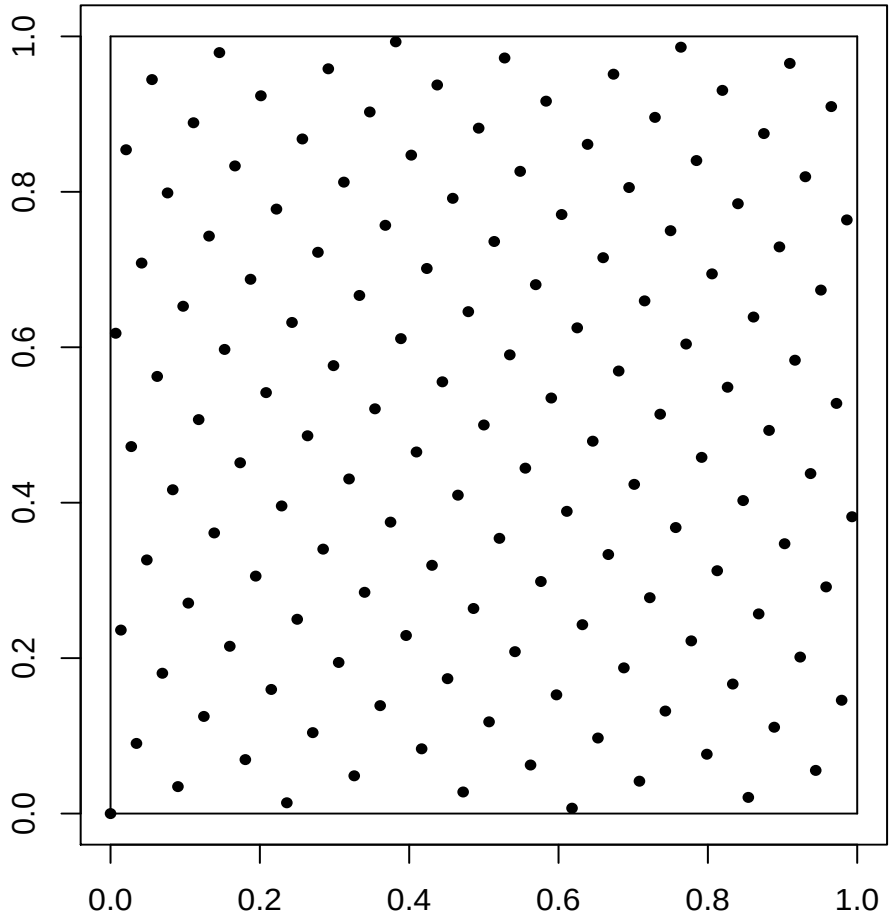

Figure 2: Shown are 144 points of an integration lattice.

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 τ 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 $(i-1)\tau/n \mod 1$ by $\phi_b(i-1)\tau \mod 1$ where ϕ_b is the radical inverse function. The result-

ing 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 ∞-dimensional generating vectors $\tau = (\tau_1, \tau_2, \dots)$ depending only on some base $b \geq 2$, that give good lattice rules for all dimensions s and for all n equal to a power of b. Computer searches for vectors τ 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 have been demonstrated by Sloan and Reztsov (2002), Sloan, Kuo, and Joe (2002a), and Sloan, Kuo, and Joe (2002b).

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

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 $\partial^{rs} f / \prod_j \partial(x^j)^r$ continous on $[0,1)^s$, the error rate can be made $O(n^{-r+\epsilon})$ (Niederreiter 1992) where n^ϵ hides powers of $\log n$, though 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 $A_1, \ldots, A_n$ by a randomized version $X_1, \ldots, X_n$ such that each $X_i \sim U[0,1)^s$ while the ensemble $X_1, \ldots, X_n$ 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 are surveyed by Owen (1998a) and by L'Ecuyer and Lemieux (2002). Hong and Hickernell (2003) describe software to randomize nets.

A scrambled net is a randomization of the base b digits of the points of a digital net $A_1, \ldots, A_n$. Let $A_i^j = \sum_{k=1}^{\infty} a_{ijk} b^{-k}$ where each $a_{ijk} \in \{0, 1, \ldots, b-1\}$. 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 permutions of a_{ijk}. In the scrambling method proposed in Owen (1995), $x_{ij1} = \pi_{1\cdot}(a_{ij1})$, then $x_{ij2} = \pi_{1\cdot a_{ij1}}(a_{ij2})$, so that the permutation of the second digit depends on what the first digit was, and generally $x_{ijk} = \pi_{1\cdot a_{ij1}\cdots a_{ij\,k-1}}(a_{ijk})$, where each π 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 one. These scrambling schemes require a lot of permutations, and some derandomizations using fewer permutations have been proposed by Matoušek (1998b) and Hong and Hickernell (2003).

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

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

where $e = \exp(1) \doteq 2.718$. When $t > 0$ the variance bound e/n has to be increased but we still find $\mathrm{Var}_{\mathrm{rnet}}(\hat{I}) \le C \int f_B(x)^2 dx / n$ for a constant C. See Owen (1998b) or Niederreiter and Pirsic (2001) or Yue and Hickernell (2002). As m increases f_G accounts for more of the structure of f. In the limit $\int f_B(x)^2 dx \to 0$ 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 in Owen (1995).

For smooth functions the rate at which $\int f_B(x)^2 dx \to 0$ can be studied. Owen (1997c) shows 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) studies the variance over randomized (λ, t, m, s)–nets. Hickernell and Yue (2000), Matoušek (1998b), and Heinrich, Hickernell, and Yue (2003) investigate the discrepancy of scrambled nets and sequences. Owen (2002) studies 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 take

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

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, $e^{2\pi\sqrt{-1}\,k^T X_i}$ equals $e^{2\pi\sqrt{-1}\,k^T U}$ 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 $\mathrm{Var}_{\mathrm{rnet}}(\hat{I}) = \mathrm{Var}(f_B) = \int f_B(x)^2 dx$. As with nets, the part f_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_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 is 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) reports some results using the Gibbs sampler in a QMC application. Ostland and Yu (1997) apply 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 spikyness 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 1 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 $j = 1, \dots, J$ 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 (1995). Essentially the same method goes by the name of "regression estimators" in the survey sampling literature. See Cochran (1977) and Lohr (1998). Here we simply summarize some well known results.

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

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

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 $\mathrm{Var}(\sum_{j=1}^{J}\beta_j h_j(X)) > 0$ for $X \sim [0,1)^s$, whenever $\beta \neq 0$. If $\mathrm{Var}(\beta^T h(X)) = 0$ for some nonzero β, 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

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

a quadratic function of the vector β. The minimizing value of β is given by

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

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

The value β_{mc} is typically unknown, and is usually estimated by

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

where $\hat H = (\hat H_1,\dots,\hat H_J)^T$ and

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

The known values θ_j could possibly be used in place of $\hat H_j$, but typically are not used. Instead $\hat\beta_{\mathrm{mc}}$ is the ordinary least squares estimator of the regression coefficients relating $f(X_i)$ to $h_j(X_i)$.

The control variate estimator is $\hat I_{\hat\beta_{\mathrm{mc}}}$ obtained by substituting $\hat\beta_{\mathrm{mc}}$ for β 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, and 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 β. Because σ^2_β is a quadratic function of the vector β with a minimum at β_{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 $X_1,\dots,X_n$ are generated by an RQMC rule. Let f be the integrand of interest, and let $h = (h_1,\dots,h_J)^T$ 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 β, 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 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 equations (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 $f = f_G + f_B$ and $h = h_G + h_B$, we find that f_G and h_G do not contribute to (14), and we would rather have h_B correlated with f_B than have h correlated with f.

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

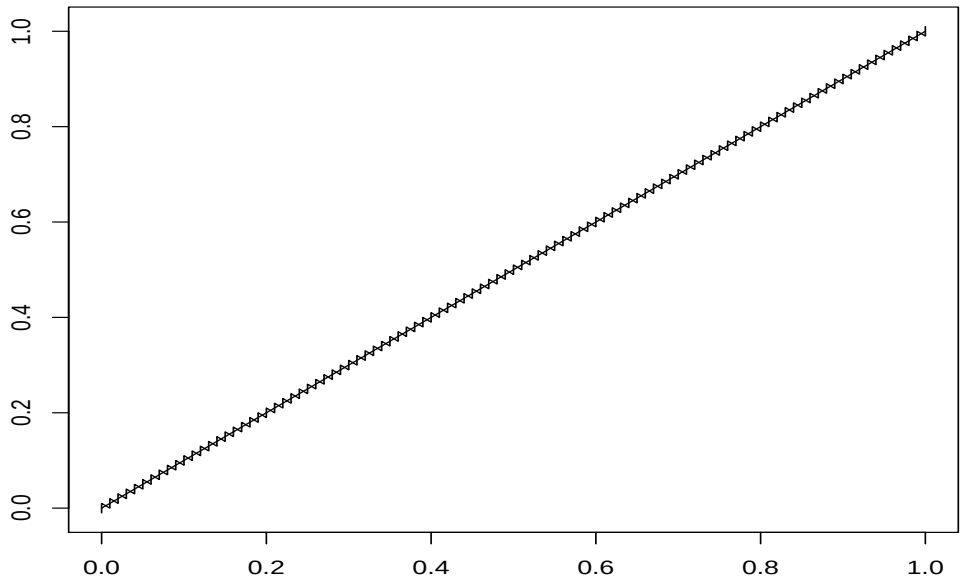

Figure 3: Shown are a sawtooth function f with tooth width 0.02 and a linear function $h_1(x) = x$.

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 1 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 coefficient β_{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 β 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 β_{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 tradeoff in choosing R.

5.1 Replication estimates of $\beta_{\mathbf{rqmc}}$

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

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

where

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

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 $\widehat{\mathrm{Var}}(\hat{\beta}) = \mathrm{SS}(\hat{I}_{\hat{\beta}}, \hat{\beta})/(R(R-J-1))$.

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 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 β_{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 β_{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 β. The tradeoff in picking small R is that R is the sample size for subsidiary tasks of estimating β 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 (λ, 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 in Owen (1995) the formulas from Section 5.1 can ordinarily be used directly. As Owen (1997b) discusses, 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 $\mathrm{Var}_{\mathrm{rlat}}(\hat{I}) = \mathrm{Var}(f_B)$, and similarly for $\hat{H}$, so that equation (15) reduces to

$$\beta_{\mathrm{rlat}} = \left(\int h_B(x) h_B^T(x)\, dx \right)^{-1} \int h_B(x) f_B(x)\, dx. \tag{19}$$

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

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

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

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

6 Limiting values of β

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

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

and the results provide qualitative insight, and suggest some methods for choosing β.

We present three cases: stratified sampling of $[0,1)$, stratified sampling of $[0,1)^s$, and 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 in Owen (1992) shows that a good control variate h for Latin hypercube sampling is one whose non-additive part correlates with that of f. The variance expressions for nets do not provide an expression for $\beta^{\infty}_{\mathrm{rqmc}}$, but do suggest a value that can be tested empirically. For extensible shifted lattices it is not clear when $\beta^{\infty}_{\mathrm{rqmc}}$ 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 $B < \infty$, 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' satisyfing $|g'(x) - g'(x^*)| \le B|x-x^*|^{\Delta}$ for all $x, x^* \in [0,1]$. Then from Section 3 of Owen (1997c) we obtain

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

It is natural to substitute $f - \beta^T h$ for g in the lead term of (21) and then minimize over β. 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 Δ, and that $\int h'(x)h'(x)^T dx$ has full rank J. Then the optimal control variate coefficient under stratified sampling satisfies*

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

Proof: By equation (21) we get

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

where the constant inside the O symbol is independent of β. 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 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

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

Suppose that $\|\beta - \beta^\infty_{\mathrm{strat}}\|_1 > \epsilon > 0$. 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 the 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

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

and replication is not necessary. Here $\beta^{\infty}_{\text{strat}}$ 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 $\beta^{\infty}_{\text{strat}} = \int_0^1 f'(x)dx = f(1) - f(0)$, and so

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

with variance $(12n^3)^{-1}\text{Var}(f'(x))+O(n^{-3-\Delta})$ instead of $(12n^3)^{-1}\int_0^1 f'(x)^2dx+O(n^{-3-\Delta})$. 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 $i = 1, \dots, n$, 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 ∇g is the 1 by s gradient (row) vector of g.

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

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

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

Recalling our definition of ∇ from Section 1.1 we may write the asymptotically optimal coefficient as

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

and estimate it by

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

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 ∇h and ∇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 β. The asymptotic variance formulas are not sharp enough to allow us to derive the exact value of $\beta^{\infty}_{\mathrm{rnet}}$, but they do suggest a way to compute a candidate value $\widetilde{\beta}^{\infty}_{\mathrm{rnet}}$. This and other candidates, such as estimates of β_{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 (1997c) defines smooth s dimensional functions as ones that satisfy

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

for finite $B \ge 0$ and $\Delta \in (0, 1]$. Then, under scrambled $(0, m, s)$–net $\mathrm{Var}_{\mathrm{rnet}}(\hat{I})$ equals

$$\left[\frac{(\log n)^{s-1}}{n^3} \frac{\lambda^2}{12^s (s-1)!} \left(\frac{b^2-1}{\log b}\right)^{s-1} \int \left(\frac{\partial^s f(x)}{\partial x}\right)^2 dx\right]\Big(1 + O(1)\Big) \tag{29}$$

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

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

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

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

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

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

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

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

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 θ 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 $\sum_{j=1}^s k_j \geq m$, 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 1997c; 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 β_{rqmc} close to β_{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, 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 $\beta_{\mathrm{rlat}} = \beta_{\mathrm{mc}}$ 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 having 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'll 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 65536 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 $\beta_{\mathrm{mc}} = 2.675$. Equation (26) gives $\beta^{\infty}_{\mathrm{strat}} = 2.809$ and equation (30) gives $\widetilde{\beta}^{\infty}_{\mathrm{rnet}} = 2.547$. 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 equations (10), (25), and (29), including the constants $1/12$, and $12^{-2}(2^2 - 1)/\log(2) = 0.0301$ in the latter two.

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

Method	CV Coef → Rate ↓	None 0	β_{mc} 2.675	$\beta^{\infty}_{\mathrm{strat}}$ 2.809	$\widetilde{\beta}_{\mathrm{rnet}}$ 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

Table 1: Shown are the asymptotic variances of MC, stratification, and QMC, for a two dimensional problem from the text. The coefficients β_{mc} from (11), $\beta^{\infty}_{\mathrm{strat}}$ from (27), and $\widetilde{\beta}^{\infty}_{\mathrm{rnet}}$ 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.

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 variates 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 $\beta^{\infty}_{\mathrm{strat}}$ would lie between β_{mc} and $\beta^{\infty}_{\mathrm{rqmc}}$, but it did not.

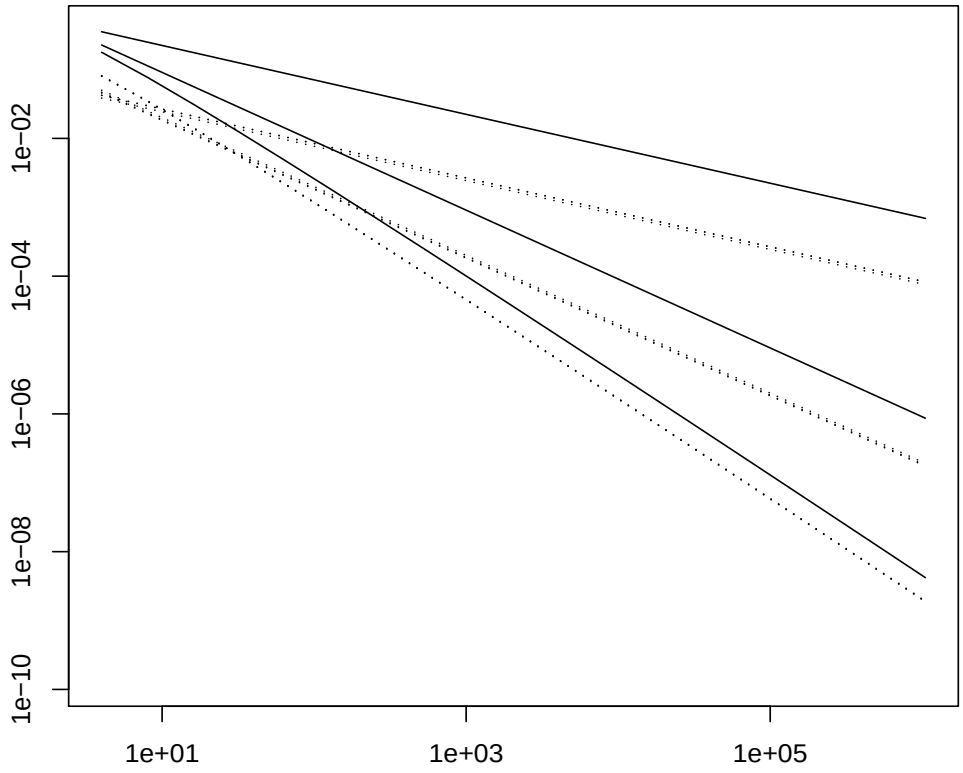

Figure 4: Shown are the asymptotic standard deviations of $\hat{I}_\beta$ versus n for the methods in Table 1. The solid lines are, top to bottom, for MC, stratification, and randomized nets. Below these are parallel dotted lines when control variates are employed. Lines for different control variate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

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 the 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 $t_1, \ldots, t_s$ 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

$$S(t_i) = S(t_i, x) = S(0)\exp\left[(r - \sigma^2/2)t_i + \sigma\sqrt{T/s}\sum_{j=1}^{i}\Phi^{-1}(x^j)\right],$$

where the drift parameter r is the risk-free rate, σ is the volatility of the asset prices, and Φ^{-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 use 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 $i = 1, \ldots, 16$. 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 lognormal distribution allowing $\int h_1(x)dx$ to be found via a one dimensional integration reducing to the Black-Scholes formula,

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

where

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

taking $b \geq 0$, and where Φ is the standard normal CDF. 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. The functions $f(x)$ and $h_1(x)$ are not of bounded variation on $[0,1)^s$, but $f(x) - A(x)$ and $h_1(x) - G(x)$ are 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

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

The Monte Carlo methods we consider are listed in Table 2. They all use IID points $X_i \sim U[0,1)^s$. MC_0 is plain Monte Carlo with no control

Name	*Estimate*
MC_0	$\hat{I}(f)$
MC_1	$\hat{I}(f-\beta_1 h_1)+\beta_1 I(h_1)$
MC_3	$\hat{I}(f-\beta_2 h_1-\beta_3 A-\beta_4 G)+\beta_2 I(h_1)+\beta_3 I(A)+\beta_4 I(G)$
MC_B	$\hat{I}(f-A)+I(A)$
MC_{BB}	$\hat{I}(f-A-\beta_5(h_1-G))+I(A)+\beta_5 I(h_1-G)$

Table 2: Shown are the Monte Carlo methods used in the Asian option example. 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 β_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. MC_B works directly with a bounded integrand and MC_{BB} uses a bounded integrand and a bounded control variate.

variates. MC_1 uses one control variate, h_1. MC_3 uses three control variates, h_1, A, and G. MC_B uses the bounded function $f-A$, and MC_{BB} uses the bounded function $f-A$ with a bounded control variate h_1-G. The coefficients β_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 MC_B to be worse than MC_0 but QMC_B might be better than QMC_0 due to bounded variation 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 $n=5\times 17^3=24565$ 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 5 replicates of the $(0,3,16)$–net the coefficients β_j were

Coef	MC	$\text{QMC}^{(2)}$	$\text{QMC}^{(3)}$
β_1	1.10 (4.9e−4)	1.08 (5.6e−3)	1.10 (1.1e−3)
β_2	1.04 (2.3e−4)	1.01 (7.7e−3)	1.04 (4.0e−4)
β_3	0.534 (1.5e−3)	1.33 (1.3e−1)	0.519 (2.7e−3)
β_4	−0.525 (1.5e−3)	−1.37 (9.7e−2)	−0.510 (2.7e−3)
β_5	0.988 (2.0e−4)	1.03 (9.0e−3)	0.987 (1.2e−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.

	MC	$\text{QMC}^{(2)}$	$\text{QMC}^{(3)}$
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

Table 4: Estimated root mean squared errors. 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.

estimated using the formula for $\hat{\beta}_{\text{mc}}$ applied to QMC data. These 5 values were then averaged and the sample standard error was computed.

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 $\text{QMC}^{(2)}$ data as 85 replicate samples of size 289 and those for $\text{QMC}^{(3)}$ were obtained in an analysis of 5 replicates of size 4913. As might be expected β_3 is close to $-\beta_4$ while the other coefficients are close to 1. The values for β_3 and β_4 are quite different for $\text{QMC}^{(2)}$ than for the other methods. The reason is that $\text{QMC}^{(3)}$, having only 5 replicates, used estimates of β_{mc} while the 85 replicates in $\text{QMC}^{(2)}$ were sufficient to allow estimation of β_{rqmc}.

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

Without control variates the root mean square error (RMSE) for MC is about 11.92 times that for $\text{QMC}^{(3)}$. For MC to attain that reduced error

would require a sample size $11.92^2 \doteq 142$ times as large. $\mathrm{QMC}^{(2)}$ attained a smaller improvement over MC.

The best control variate strategy for MC was to use all 3 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 β_2, β_3 and β_4. $\mathrm{QMC}^{(2)}$ used 85 replicates and also had its smallest error with all three control variates. For $\mathrm{QMC}^{(3)}$ with QMC estimates of β_{mc}, the method $\mathrm{QMC}^{(3)}_{BB}$ with just one control variate had better accuracy than $\mathrm{QMC}^{(3)}_{3}$.

The best combined strategy was $\mathrm{QMC}^{(3)}_{BB}$, with an efficiency gain of $(4.41/.0735)^2 \doteq 3600$ compared to MC_0. The two best methods for this problem are $\mathrm{QMC}^{(3)}_{1}$ and $\mathrm{QMC}^{(3)}_{BB}$. They gave option values of 2.162 and 2.163 respectively with the standard errors in Table 4.

As expected MC_B was worse than MC_0. For both QMC methods the bounded function approaches $\mathrm{QMC}^{(m)}_{B}$ were (slightly) better than the corresponding $\mathrm{QMC}^{(m)}_{0}$ methods. Similarly there were small advantages for $\mathrm{QMC}^{(m)}_{BB}$ using the bounded functions $f - A$ and $h_1 - G$ over $\mathrm{QMC}^{(m)}_{1}$ 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, $\mathrm{QMC}^{(2)}$ and $\mathrm{QMC}^{(3)}$ 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 $\mathrm{QMC}^{(3)}$ 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 by 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, though 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 Ameur, L'Ecuyer, and Lemieux (1999) who report numerical examples in which control variates improve QMC but not as much as they improve simpler methods. Further, 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 β_{mc} with QMC gave very good results, and this is reassuring. In the Asian option problem we saw better results using estimated suboptimal coefficients β_{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 β_{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 β_{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 β_{rqmc}, or what seems better, internal replicates can be used to estimate the value of β_{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 β_{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 β 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 corre-

sponding derivative of the target integrand.

11 Acknowledgements

We thank Rong-Xian Yue and Ms. 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 grant DMS-0072445.

References

Ameur, H. B., P. L'Ecuyer, and C. Lemieux (1999). Variance reduction of Monte Carlo and randomized quasi-Monte Carlo estimators for stochastic volatility models in finance. In *Proceedings of the 1999 Winter Simulation Conference*, pp. 336–343. IEEE Press.

Avramidis, A. and J. R. Wilson (1993). A splitting scheme for control variates. *Operations Research Letters 14*(4), 187–198.

Beck, J. and W. W. L. Chen (1987). *Irregularities of Distribution*. New York: Cambridge University Press.

Bratley, P., B. J. Fox, and L. E. Schrage (1987). *A Guide to Simulation (Second Edition)*. Springer-Verlag.

Caflisch, R. E. and W. Morokoff (1995). Quasi-monte carlo integration. *Journal of Computational Physics 122*, 218–230.

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

Chelson, P. (1976). *Quasi-random techniques for Monte Carlo methods*. Ph. D. thesis, The Claremont Graduate School.

Cochran, W. G. (1977). *Sampling Techniques (3rd Ed)*. John Wiley & Sons.

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

Efron, B. and C. Stein (1981). The jackknife estimate of variance. *Annals of Statistics 9*, 586–596.

Fang, K.-T. and Y. Wang (1994). *Number Theoretic Methods in Statistics*. London: Chapman and Hall.

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

Fishman, G. (1995). *Monte Carlo: Concepts, Algorithms, and Applications*. Springer-Verlag.

Heinrich, S., F. J. Hickernell, and R. X. Yue (2003). Optimal quadrature for Haar wavelet spaces. *Mathematics of Computation*. to appear.

Hickernell, F. J. (1996). Quadrature error bounds and figures of merit for quasi-random points. *SIAM Journal of Numerical Analysis 33*, 1995–2016. corrected printing of Sections 3-6 in ibid., **34** (1997), 853–866.

Hickernell, F. J., H. S. Hong, P. L'Ecuyer, and C. Lemieux (2000). Extensible lattice sequences for quasi-Monte Carlo quadrature. *SIAM Journal on Scientific Computing 22*(3), 1117–1138.

Hickernell, F. J. and H. Niederreiter (2003). The existence of good extensible rank-1 lattices. *J. Complexity 19*. to appear.

Hickernell, F. J. and R. X. Yue (2000). The mean square discrepancy of scrambled (t, s)-sequences. *SIAM Journal of Numerical Analysis 38*, 1089–1112.

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

Hong, H. S. and F. J. Hickernell (2003). Implementing scrambled digital sequences. *AMS Transactions on Mathematical Software*. To appear.

Hua, L. and Y. Wang (1981). *Applications of number theory to numerical analysis*. Berlin: Springer.

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

Kuipers, L. and H. Niederreiter (1974). *Uniform Distribution of Sequences*. New York: Wiley.

L'Ecuyer, P. and C. Lemieux (2002). A survey of randomized quasi-Monte Carlo methods. In M. Dror, P. L'Ecuyer, and F. Szidarovszki (Eds.), *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, pp. 419–474. Kluwer Academic Publishers.

Liao, L. G. (1998). Variance reduction in Gibbs sampler using quasi random numbers. *Journal of Computational and Graphical Statistics 7*, 253–266.

Loh, W.-L. (2003). On the asymptotic distribution of scrambled net quadrature. *Annals of Statistics 31*. To appear.

Lohr, S. (1998). *Sampling: Design and Analysis*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Matoušek, J. (1998a). *Geometric Discrepancy : An Illustrated Guide*. Heidelberg: Springer-Verlag.

Matoušek, J. (1998b). On the L^2–discrepancy for anchored boxes. *Journal of Complexity 14*, 527–556.

Niederreiter, H. (1987). Point sets and sequences with small discrepancy. *Monatshefte fur mathematik 104*, 273–337.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia, PA: S.I.A.M.

Niederreiter, H. and G. Pirsic (2001, Dec.). The microstructure of (t,m,s)-nets. *Journal of Complexity 17*(4), 683–696.

Niederreiter, H. and C. Xing (2001). *Rational Points on Curves over Finite Fields: Theory and Applications*. Number 285 in London Mathematical Society Lecture Note Series. Cambridge: Cambridge University Press.

Ostland and B. Yu (1997). Exploring quasi Monte Carlo for marginal density approximation. *Statistics and Computing 7*, 217–228.

Owen, A. B. (1992). A central limit theorem for Latin hypercube sampling. *Journal of the Royal Statistical Society, Series B 54*, 541–551.

Owen, A. B. (1995). Randomly permuted (t, m, s)-nets and (t, s)-sequences. In H. Niederreiter and P. J.-S. Shiue (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, New York, pp. 299–317. Springer-Verlag.

Owen, A. B. (1997a). Monte Carlo variance of scrambled equidistribution quadrature. *SIAM Journal of Numerical Analysis 34*(5), 1884–1910.

Owen, A. B. (1997b). Monte Carlo variance of scrambled equidistribution quadrature. *SIAM Journal of Numerical Analysis 34*(5), 1884–1910.

Owen, A. B. (1997c). Scrambled net variance for integrals of smooth functions. *Annals of Statistics 25*(4), 1541–1562.

Owen, A. B. (1998a). Latin supercube sampling for very high dimensional simulations. *ACM Transactions on Modeling and Computer Simulation 8*(2), 71–102.

Owen, A. B. (1998b, December). Scrambling Sobol' and Niederreiter-Xing points. *Journal of Complexity 14*(4), 466–489.

Owen, A. B. (2002). Scrambled net variance with alternative scramblings. Technical report, Stanford University, Department of Statistics.

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

Ripley, B. D. (1987). *Stochastic Simulation*. John Wiley & Sons.

Ritchken, P., L. Sankarasubramanian, and A. M. Vijh (1993). The valuation of path dependent contracts on the average. *Management Science 39*(10), 1202–1213.

Sarkar, P. K. and M. A. Prasad (1987). A comparative study of pseudo and quasi random sequences for the solution of integral equations. *Journal of Computational Physics 68*, 66–88.

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

Sloan, I. H. and S. Joe (1994). *Lattice Methods for Multiple Integration*. Oxford: Oxford Science Publications.

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

Sloan, I. H., F. Y. Kuo, and S. Joe (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 A. V. Reztsov (2002). Component-by-component construction of good lattice rules. *Mathematics of Computation 71*, 263–273.

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

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

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

Tezuka, S. (1995). *Uniform random numbers: theory and practice*. Boston: Kluwer Academic Publishers.

van der Corput, J. G. (1935a). Verteilungsfunktionen I. *Nederl. Akad. Wetensch. Proc. 38*, 813–821.

van der Corput, J. G. (1935b). Verteilungsfunktionen II. *Nederl. Akad. Wetensch. Proc. 38*, 1058–1066.

Weyl, H. (1914). Über ein problem aus dem gebeite der diophantischen approximationen. *Nachrichten der Akademie der Wissenschaften in Göttingen. II. Mathematisch-Physikalische Klasse*, 234–244.

Weyl, H. (1916). Über die gleichverteilung von zahlen mod. eins. *Mathematische Annalen 77*, 313–352.

Yue, R.-X. (1999). Variance of quadrature over scrambled unions of nets. *Statistica Sinica 9*(2), 451–473.

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

Zaremba, S. K. (1968). Some applications of multidimensional integration by parts. *Annales Polonici Mathematici 21*, 85–96.

Fred J. Hickernell: Professor, Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong.
Christiane Lemieux: Assistant Professor, Department of Mathematics and Statistics, and Department of Computer Science, University of Calgary, 2500 University Drive N.W., Calgary, Alberta, Canada T2N 1N4.
Art B. Owen: Professor, Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford University, Stanford CA 94305.