Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

CONFIDENCE INTERVALS FOR RANDOMIZED QUASI-MONTE CARLO ESTIMATORS

Pierre L'Ecuyer

DIRO, Université de Montréal
C.P.6128, Succ. Centre-Ville
Montréal, QC H3C 3J7, CANADA

Marvin K. Nakayama

Computer Science Department
New Jersey Institute of Technology
Newark, NJ 07102, USA

Art B. Owen

Department of Statistics
Stanford University
Stanford, CA, USA

Bruno Tuffin

Inria, Univ Rennes, CNRS, IRISA
Campus de Beaulieu, 263 Avenue Général Leclerc
35042 Rennes, FRANCE

ABSTRACT

Randomized Quasi-Monte Carlo (RQMC) methods provide unbiased estimators whose variance often converges at a faster rate than standard Monte Carlo as a function of the sample size. However, computing valid confidence intervals is challenging because the observations from a single randomization are dependent and the central limit theorem does not ordinarily apply. A natural solution is to replicate the RQMC process independently a small number of times to estimate the variance and use a standard confidence interval based on a normal or Student t distribution. We investigate the standard Student t approach and two bootstrap methods for getting nonparametric confidence intervals for the mean using a modest number of replicates. Our main conclusion is that intervals based on the Student t distribution are more reliable than even the bootstrap t method on the integration problems arising from RQMC.

1 INTRODUCTION

Integration is a fundamental task with a ubiquitous role in scientific computation. In very special settings a definite integral can be found in closed form, but in many others we must use numerical methods. When the integration problem is of very high dimension, then stochastic simulation (Monte Carlo or MC) methods dominate practice. There has been considerable progress in replacing those MC computations by deterministic *quasi-Monte Carlo* (QMC) methods, which can significantly improve accuracy. But those QMC methods do not support practical error estimation strategies. *Randomized QMC* (RQMC) methods have been developed that provide independent and unbiased estimates with accuracy equal to or even better than QMC provides. So far those replicates have primarily been used to estimate the variance of an RQMC estimate. Much less has been written about how to use such replicates to construct approximate confidence intervals. In this paper we explore various approaches to construct confidence intervals, including bootstrap methods, in RQMC. The motivation is as follows: any practical problem requiring that an answer must be estimated with accuracy further necessitates providing some evidence of the accuracy attained; e.g., the U.S. Nuclear Regulatory Commission (2007) mandates that risk studies include an uncertainty quantification.

Stochastic simulations get their randomness from a random number generator that "simulates" realizations of independent random variables uniformly distributed over the interval $(0,1)$ (denoted $\mathscr{U}(0,1)$). Typically, each simulation run consumes a vector of $\boldsymbol{U}$ of d successive random numbers from this sequence and outputs a simulated random variable $f(\boldsymbol{U})$, where f represents all the computations made in the simulation.

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

The aim is to estimate the expectation

$$\mu = \mathbb{E}[f(\boldsymbol{U})] = \int_{[0,1]^d} f(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}. \tag{1}$$

QMC replaces the independent replicates of the d-dimensional random vector $\boldsymbol{U}$ by a (deterministic) *low-discrepancy sequence* of points $\boldsymbol{u}_0, \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$ in $[0,1)^d$. The first n of these points cover the unit hypercube $[0,1)^d$ far more evenly than random points would. When n is fixed a priori, we talk of a *low-discrepancy point set*. QMC points are typically constructed for special values of n, such as powers of 2. The value of μ is then approximated by the average $\hat{\mu}_n = \frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{u}_i)$. Here $\boldsymbol{u}_i = (u_{i,1}, \ldots, u_{i,d}) \in [0,1)^d$.

Under certain assumptions on the integrand f, there are point sets and sequences for which the QMC approximation error $|\hat{\mu} - \mu|$ converges as $\mathcal{O}(n^{-\alpha+\varepsilon})$ as $n \to \infty$ for any $\varepsilon > 0$, where $\alpha > 0$ depends on the smoothness assumptions made on f, which are expressed in terms of the integrals over the unit hypercube of its squared mixed partial derivatives with respect to the coordinates of $\boldsymbol{u}$ (Dick and Pillichshammer 2010; Niederreiter 1992; Lemieux 2009; L'Ecuyer 2009). As a special case, if f is assumed to have bounded variation in the sense of Hardy and Krause, the rate is $\mathcal{O}(n^{-1}(\ln n)^d)$. Explicit constructions of these point sets and sequences are also available, mainly in the form of lattice point sets and digital nets and sequences (Dick and Pillichshammer 2010; L'Ecuyer et al. 2022; Dick et al. 2022). One limitation, however, is that QMC fails to provide practical (easily computable) error bounds, so it is difficult to assess its accuracy (Tuffin 2004; Nakayama and Tuffin 2021).

RQMC methods randomize the points in a way that each randomized point has the uniform distribution over $[0,1)^d$ while the whole point set retains its high uniformity over this unit hypercube (L'Ecuyer and Lemieux 2002; L'Ecuyer 2018; Owen 1995). The RQMC estimator

$$\widehat{\mu}_{n,\mathrm{rqmc}} = \frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{U}_i) \tag{2}$$

is unbiased for μ, and under similar smoothness assumptions on f as for QMC, its variance converges as $\mathcal{O}(n^{-2\alpha+\varepsilon})$ for any $\varepsilon > 0$, where α depends on the smoothness of f and on the randomization method (Dick and Pillichshammer 2010; Owen 2003). But equation (2) averages *dependent* outputs, which complicates estimation of the variance of $\widehat{\mu}_{n,\mathrm{rqmc}}$ from a single randomization. A simple solution instead constructs $R > 1$ independent and identically distributed (iid) replicates of $\widehat{\mu}_{n,\mathrm{rqmc}}$ from independent randomizations, and computes the empirical variance of the R iid estimators, which is an unbiased variance estimator.

A natural next step is to compute the standard confidence interval for μ based on a central limit theorem (CLT) for the average of the R replicates. Unfortunately, the justification of this step is asymptotic in R and, as explained below, we would prefer not to use large R. More specifically, let y_r, $r = 1, 2, \ldots, R$, denote the R iid realizations of $\widehat{\mu}_{n,\mathrm{rqmc}}$ in (2) and let $\bar{y}_{n,R}$ be their average. Typically, $\mathrm{Var}[y_r] \approx Kn^{-2\alpha}$ for some $\alpha > 1/2$, so the variance of the global average is $\mathrm{Var}[\bar{y}_{n,R}] \approx KR^{-1}n^{-2\alpha}$. For a given upper bound on the total number Rn of simulation runs (total computing budget), this variance is minimized by taking R as small as possible ($R = 1$). On the other hand, a small $R > 1$ gives a poor estimator of the variance, and a questionable CLT approximation. A compromise must be made if we want a reasonable estimate of the variance to assess the accuracy of our estimator of the mean. In practice, R is often chosen in the range from 5 to 20. The global average $\bar{y}_{n,R}$ obeys a classical CLT when $R \to \infty$ for fixed n, but only in a few special cases when $n \to \infty$ for fixed R, so the standard intervals may fail to attain their nominal coverage. This is discussed in more detail in Sections 2 and 3.

In this paper, we explore empirically the distribution of $\widehat{\mu}_{n,\mathrm{rqmc}}$ for various integrands f, various choices of point sets and randomizations, and different values of n. We then compare the standard confidence interval with two others that have a nonparametric derivation. One of those is the bootstrap t of Efron (1981). It has remarkably good asymptotic coverage (Hall 1988) and exceptional empirical coverage in small samples (Owen 1988). The other is the percentile bootstrap method, possibly the best known bootstrap. For

background on the bootstrap, see Efron and Tibshirani (1994). In this work we see some interesting cases where the distribution of $\widehat{\mu}_{n,\text{rqmc}}$ is very far from normal and has a small but non-negligible probability mass which is quite far from the mean. Such cases are very far from those where nonparametric confidence intervals have been previously studied.

The remainder is organized as follows. In Section 2, we recall the definitions of the point sets and randomizations considered in this paper, and the variance bounds that are known for the relevant combinations. In Section 3, we discuss several ways of computing confidence intervals RQMC estimators, including via a Student t approximation and by applying the bootstrap t. In Section 4, we illustrate numerically the behavior of these methods for some functions f. Our conclusions are in Section 5.

2 POINT SETS AND RANDOMIZATION METHODS

Here we define the QMC point sets and the randomizations considered in this paper. For each type of RQMC points, we define a short tag (acronym) to use later when reporting empirical results.

Our first type of QMC point set corresponds to a *lattice rule of rank 1*. It is defined by selecting the number n of points and a generating vector $\boldsymbol{a} = (a_1, \ldots, a_d)$ of integers $a_j \in \{1, \ldots, n-1\}$. Then the ith point $\boldsymbol{u}_i$ is defined by $\boldsymbol{u}_i = (i\boldsymbol{a} \bmod n)/n$, for $i = 0, \ldots, n-1$. In our experiments, the vector $\boldsymbol{a}$ was selected using LatNet Builder (L'Ecuyer and Munger 2016; L'Ecuyer et al. 2022) with the $\mathscr{P}_\alpha$ criterion with $\alpha = 2$, product weights, and a weight of $\gamma_j = 2/(2+j)$ for coordinate j. See L'Ecuyer et al. (2022) for the details.

The standard randomization for a lattice rule is a random shift modulo 1 (Cranley and Patterson 1976; Tuffin 1998; L'Ecuyer and Lemieux 2000): we generate one random point $\boldsymbol{U}$ uniformly in $(0,1)^d$ and add it to each point $\boldsymbol{u}_i$ modulo 1, coordinate-wise. Adding the same uniform $\boldsymbol{U}$ to each of the n points preserves the lattice structure of the point set, with the n resulting points dependent. Moreover, each randomly shifted point is uniformly distributed on $(0,1)^d$, making the average over all n randomly shifted points an unbiased estimator of μ by the linearity of expectation. The tag "Lat-RS" will refer to this type of RQMC point set. It is often recommended to add a baker (or tent) transformation after the random shift. We replace each $u_{i,j}$ by $\text{tent}(u_{ij}) = 1 - 2|u_{i,j} - 1/2|$. As u ranges from 0 to 1, $\text{tent}(u)$ goes from 0 to 1 and then back to 0. With this transformation, a continuous function over $[0,1]^d$ becomes continuous and periodic over the d-dimensional unit torus, and this greatly improves the convergence rate of the RQMC estimator (Hickernell 2002; L'Ecuyer 2018). We refer to this type of point set by "Lat-RSB." For a non-periodic integrand f for which all mixed partial derivatives of order 1 with respect to any subset of coordinates are square integrable over $(0,1)^d$, for instance, the RQMC variance converges as $\mathscr{O}(n^{-2+\varepsilon})$ with Lat-RS, and as $\mathscr{O}(n^{-4+\varepsilon})$ with Lat-RSB. This is a huge gain. It also changes substantially the distribution of the RQMC estimator, which can have a significant impact when computing a confidence interval.

The next QMC point set we describe is a *digital net in base* $b = 2$. The coordinate j of point i is

$$u_{i,j} = \sum_{\ell=1}^{w} u_{i,j,\ell} 2^{-\ell}, \quad i = 0, \ldots, 2^k - 1 \text{ and } j = 1, \ldots, d,$$

where

$$\begin{pmatrix} u_{i,j,1} \\ \vdots \\ u_{i,j,w} \end{pmatrix} = \boldsymbol{C}_j \begin{pmatrix} a_{i,0} \\ \vdots \\ a_{i,k-1} \end{pmatrix} \bmod 2 \quad \text{with } i = \sum_{\ell=0}^{k-1} a_{i,\ell} 2^\ell,$$

and the $\boldsymbol{C}_j$ are $w \times k$ *generating matrices* of rank k, with elements in $\{0,1\}$. See Dick and Pillichshammer (2010), Niederreiter (1992), L'Ecuyer (2018) for more details. This gives $n = 2^k$ points. Our implementation used $w = 31$. For the $\boldsymbol{C}_j$, we use the popular Sobol' constructions, with the direction numbers proposed by Joe and Kuo (2008).

There are many ways of randomizing a digital net while keeping the digital net structure and respecting the RQMC conditions (Owen 2003; L'Ecuyer 2018). The simplest is a *random digital shift* in base $b = 2$:

generate a single random point $\boldsymbol{U} = (U_1, \ldots, U_j)$ uniformly over $(0,1)^d$, and perform a bitwise xor of each coordinate of this point with the corresponding coordinates of the points $\boldsymbol{u}_i$ of the digital net. We refer to this one by "Sob-DS." It already provides an unbiased RQMC estimator. The variance of this estimator can often be improved by also randomizing the generating matrices $\boldsymbol{C}_j$ before generating the points (Owen 2003). The most popular way to do that is to pre-multiply each $\boldsymbol{C}_j$ (mod 2) by a random lower triangular $w \times w$ matrix $\boldsymbol{L}_j$ with 1's on the diagonal and independent random binary entries below the diagonal (Matoušek 1998). We refer to it by "Sob-LMS," where LMS stands for "left matrix scramble." The rationale is that the Sobol' (or other given) generating matrices often have too much structure, and randomizing them improves their quality on average. Note that to obtain an unbiased RQMC estimator with LMS, it is essential to perform a random digital shift after randomizing the generating matrices. A third approach is the *nested uniform scramble* (NUS) from Owen (1995) also described in Dick and Pillichshammer (2010) and L'Ecuyer (2018), for example. We refer to it by "Sob-NUS." For both NUS and LMS, under sufficient smoothness conditions on f, the RQMC variance converges as $\mathcal{O}(n^{-3}(\log n)^{d-1})$ (Owen 1997; Yue and Hickernell 2002). There is also a strong law of large numbers for some randomizations of digital nets (Owen and Rudolf 2021).

For all these RQMC methods, a standard CLT with a normal limiting distribution holds when n is fixed, $\widehat{\mu}_{n,\mathrm{rqmc}}$ has finite positive variance, and $R \to \infty$, because the RQMC estimator is an average of R iid unbiased estimators. But for $n \to \infty$ for fixed R, this type of CLT has been proved only for NUS (Loh 2003), and only for special constructions known as $(0,m,d)$-nets and smooth integrands; see, e.g., Section 15.7 of Owen (2023) for details. For randomly-shifted lattice rules, when $n \to \infty$ for $R = 1$, L'Ecuyer et al. (2010) have shown that the RQMC estimator usually has a limiting distribution whose properly-scaled cumulative distribution function (CDF) is a spline function, not a normal distribution. L'Ecuyer (2018), Figure 10, gives an example of an Lat-RS estimator whose distribution is far from normal. Nakayama and Tuffin (2021) provide *sufficient* conditions for a CLT to hold in an asymptotic regime where both n and R increase together (R must increase fast enough).

3 CONFIDENCE INTERVALS FOR RQMC

As outlined above, the challenge in forming confidence intervals for RQMC comes from a harsh tradeoff. When using R replicates of an RQMC method with n points, at cost nR that is constrained to not exceed a given computation budget, we expect the greatest accuracy in our estimate for large n and small R, while confidence intervals will become more reliable with larger R. The sample size for our confidence intervals is R. We therefore would like reliable confidence intervals for the mean based on a small sample of R iid variables. It is known since Bahadur and Savage (1956) that there do not exist nonparametric confidence intervals with exact coverage for the mean. The usable methods have asymptotic justifications, but our setting considers R not large. Our confidence intervals take the form $[L,U]$ for statistics $L = L(y_1, \ldots, y_R)$ and $U = U(y_1, \ldots, y_R)$, where the nominal confidence level is $1-\alpha$ for a given $\alpha \in (0,1)$, e.g., $\alpha = 0.05$. Because these intervals cannot be perfect, we study their coverage error

$$\varepsilon := \Pr(L \le \mu \le U) - (1-\alpha).$$

Here we describe confidence intervals for the mean of IID values y_r. In our setting, $y_r = \hat{\mu}_r$ for $r = 1, \ldots, R$. Because RQMC estimates are unbiased, the common mean $\mathbb{E}(y_r)$ is the integral value μ that we seek. Given $y_1, \ldots, y_R$, let

$$\bar{y} = \frac{1}{R}\sum_{r=1}^{R} y_r \quad \text{and} \quad S^2 = \frac{1}{R-1}\sum_{r=1}^{R}(y_r - \bar{y})^2.$$

The standard approximate $100(1-\alpha)\%$ confidence interval for μ based on the t-test has

$$L = \bar{y} - t_{(R-1)}^{1-\alpha/2} S R^{-1/2} \quad \text{and} \quad U = \bar{y} + t_{(R-1)}^{1-\alpha/2} S R^{-1/2}$$

where $t_{(R-1)}^{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Student t distribution on $R-1$ degrees of freedom. This interval is constructed to be exact for iid Gaussian data $y_r \overset{\text{iid}}{\sim} N(\mu,\sigma^2)$, i.e., normal with mean μ and variance $\sigma^2>0$. It has an asymptotic justification based on the CLT when y_r has finite variance.

The second method we consider is a kind of *percentile bootstrap*. A bootstrap sample of $y_1,\dots,y_R$ is a list of values $y_1^*,\dots,y_R^*$ sampled independently (with replacement) from $y_1,\dots,y_R$. If there are repeated values among the y_r, the sampling respects that multiplicity. Operationally, we sample indices $\ell_1^*,\dots,\ell_R^* \overset{\text{iid}}{\sim} \mathscr{U}\{1,2,\dots,R\}$ (i.e., uniformly from the discrete set $\{1,2,\dots,r\}$), then take $y_r^* = y_{\ell_r^*}$, and compute $\bar{y}^* = (1/R)\sum_{r=1}^R y_r^*$. The idea is that in this process the sample values y_r^* are to the empirical distribution of y_r what those sample values are to the true distribution of y_r. In the percentile method we repeat this resampling B times independently, getting $\bar{y}^{*b}$ for $b=1,\dots,B$. Sorting them yields $\bar{y}^{*(1)} \le \bar{y}^{*(2)} \le \cdots \le \bar{y}^{*(B)}$. The confidence interval endpoints are then the quantiles $\bar{y}^{*(\lfloor B\alpha/2\rfloor)}$ and $\bar{y}^{*(\lceil B(1-\alpha)/2\rceil)}$, where $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ are the floor and ceiling functions. Confidence intervals based on small sample sizes R commonly use $B=1000$. So, for $B=1000$ and $\alpha=0.05$, we would take $L=\bar{y}^{*(25)}$ and $U=\bar{y}^{*(975)}$. Some authors prefer to take $B=999$ in this case and then use $L=\bar{y}^{*(25)}$ and $U=\bar{y}^{*(975)}$. The rationale is that the B points $\bar{y}^{*b}$ partition $\mathbb{R}$ into $B+1$ intervals and the confidence interval is then the union of the central 95% of those intervals.

The third method we consider is the *bootstrap t* method of Efron (1981). It is recommended for RQMC without much analysis in Owen (2023). Choquet et al. (1999) showed that it outperforms the standard method based on the Student t distribution when estimating a ratio of expectations in a regenerative simulation setting. It is known to be highly accurate for bootstrapping the mean, both asymptotically (Hall 1988) and for small sample sizes, where for all considered distributions but one, Owen (1988) finds empirically that it works reasonably well for sample sizes as low as 5. It also has a stronger rationale than the percentile method. The reasoning behind the bootstrap t is that the sample distribution of the t statistic $\sqrt{R}(\bar{y}-\mu)/S$ is well approximated by the sample distribution of a bootstrapped t statistic $\sqrt{R}(\bar{y}^*-\bar{y})/S^*$, where S^* is the standard deviation of $y_1^*,\dots,y_R^*$. We take B independent bootstrap t values t^{*b} for $b=1,\dots,B$, sort them, and then let t_L^* and t_U^* be the $\alpha/2$ and $1-\alpha/2$ quantiles of the t^{*b} values. By taking B large enough, we can make

$$\Pr\Bigl(t_L^* \le \sqrt{R}\frac{\bar{y}^*-\bar{y}}{S^*} \le t_U^*\Bigr)$$

as close as we like to $1-\alpha$. Then if we reason that $t_L^* \le \sqrt{R}(\bar{y}-\mu)/S \le t_U^*$ has approximately $1-\alpha$ probability of holding, we take

$$L = \bar{y} - St_U^*R^{-1/2} \quad \text{and} \quad U = \bar{y} - St_L^*R^{-1/2}$$

to form our approximate confidence interval for μ. Note that the upper limit for t^* appears in the lower limit for μ and vice versa.

Hall (1988) gives conditions under which the coverage error ε for these methods is $\mathscr{O}(1/R)$. It is remarkable that the confidence interval is more accurate than $\bar{y}$ itself, which has error $\mathscr{O}_p(1/\sqrt{R})$ as $R\to\infty$. It is difficult to connect the tabular values in Hall (1988) to the names that others use for those methods. The bootstrap t method is called 'STUD' for studentized; it is also called 'percentile t' because it used percentiles of a bootstrapped t statistic. The percentile method is called "the other percentile method" by Hall (1992), who then notes that it is called 'BACK' in Hall (1988), who interprets the non-reversal of quantiles to mean that it is looking at the quantile approximations backwards. Hall (1988) does not include the usual t-test confidence intervals but does include their normal theory counterparts which use quantiles of the $N(0,1)$ distribution in place of the Student t distribution. With these assumptions, the coverage errors from Table 1 of Hall (1988) are

$$\begin{aligned}
\text{Normal theory:} \quad & (1/R)\varphi(z^{1-\alpha/2})\bigl[\ \ 0.14\kappa - 2.12\gamma^2 - 3.35\bigr] + \mathscr{O}(1/R^2),\\
\text{Percentile:} \quad & (1/R)\varphi(z^{1-\alpha/2})\bigl[-0.72\kappa - 0.37\gamma^2 - 3.35\bigr] + \mathscr{O}(1/R^2), \quad \text{and}\\
\text{Bootstrap } t\text{:} \quad & (1/R)\varphi(z^{1-\alpha/2})\bigl[-2.84\kappa + 4.25\gamma^2 \qquad\ \ \bigr] + \mathscr{O}(1/R^2),
\end{aligned}$$

where $\gamma = \mathbb{E}((\hat{\mu}_r - \mu)^3)/\sigma^3$ and $\kappa = \mathbb{E}((\hat{\mu}_r - \mu)^4)/\sigma^4 - 3$ are the skewness and kurtosis, respectively, of $\hat{\mu}_r$, φ is the $N(0,1)$ density function, and $z^{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$-quantile of the $N(0,1)$ CDF Φ. These rates hold for central confidence intervals that are our focus here. For one-sided intervals, the percentile method has errors that are $\mathcal{O}(R^{-1/2})$, while the bootstrap t method has one-sided errors that are $\mathcal{O}(R^{-1})$. Hall's conditions involve 8 finite moments for $\hat{\mu}_r$ and an assumption that the support of $\hat{\mu}_r$ is not an arithmetic sequence. Hall's α is twice ours.

The normal theory method above has error $-3.35\varphi(z^{1-\alpha/2})/R + \mathcal{O}(1/R^2)$ for Gaussian data (because it should have used instead the Student t threshold). The bootstrap t has an advantage in missing the -3.35 component that the others have. It has error $\mathcal{O}(R^{-2})$ on Gaussian data since $\gamma = \kappa = 0$. It has a large positive coefficient for γ^2 where the others have negative coefficients. This gives it extra coverage on highly skewed data. The bootstrap t is well known to produce very long intervals; in addition to the asymptotics, there is an intuitive explanation in that the denominator S^* can become small. It can even be zero.

The asymptotics predict that the bootstrap t will undercover when κ is large and $\gamma = 0$. In our context, γ and κ are the skewness and kurtosis of RQMC estimates based on $n = 2^k$ function evaluations. In plain Monte Carlo sampling, the average of n random iid variables with skewness γ has skewness $\gamma/\sqrt{n}$ and the average of n iid random variables with kurtosis κ has kurtosis κ/n. Much less is known about the skewness and kurtosis of RQMC estimates. Indeed, the work presented here is the most thorough exploration of that issue yet.

For continuously distributed y_r, one can show that $\Pr(S^* = 0) = R^{1-R}$. This is not negligible for $R = 5$ which we consider. Such S^* ordinarily provide $|t^*| = \infty$. As long as there are fewer than $B\alpha/2$ bootstrap samples with $t^* = \infty$, the lower confidence interval will be finite, and similarly, fewer than $B\alpha/2$ bootstrap samples with $t^* = -\infty$ means that the upper confidence interval will be finite. We revisit this issue in our numerical work in Section 4 and again in our conclusions in Section 5.

4 NUMERICAL EXPERIMENTS

For our numerical experiments, we made a strategic selection of test functions f described in Section 4.1. They can all be parameterized by the dimension d, and the integral μ in equation (1) can be computed exactly for any d. Subtracting μ from f gives a function with integral 0 in all dimensions, which we can use to estimate the coverage of our confidence intervals. For each selected (centered) function f, we tried the five types of RQMC point sets described in Section 2, with $n = 2^k$ points for $k = 6, 8, 10, 12, 14$, and in $d = 4, 8, 16, 32$ dimensions.

Our computations had two steps. In the first step, for each combination of (f, method, k, d), we generated a pool of $N = 10^4$ independent realizations $y_1, \ldots, y_N$ of the RQMC estimator $y = \widehat{\mu}_{n,\text{rqmc}}$ in (2) and stored the sorted values in files. To do this, we used the packages `hups` and `mcqmctools` from the SSJ Java library (L'Ecuyer 2023). These large samples were used in the second step to visualize the distribution of y and estimate its variance, skewness, and kurtosis, in each case.

In the second step, we generated small samples of R replicates (without replacement) from the pool of N values and used them to form all three kinds of confidence intervals from Section 3. The sample sizes we used were $R \in \{5, 10, 20, 30\}$. We did that 1000 times. This strategy lets us generate more than N/R confidence intervals using our pool of RQMC estimates. For each such sample, we computed a 95% confidence interval with each of the three procedures. From these 1000 replicated confidence intervals we could record the proportion of the intervals that covered 0. We also computed the lengths of all the confidence intervals. Each of our 1000 bootstrap confidence intervals was based on $B = 1000$ bootstrap resamples. The whole process of forming 1000 confidence intervals of each type was done separately for all the integrands, methods, dimensions d, and RQMC sample sizes $n = 2^k$ described above.

4.1 Selected Functions

Here are the functions we consider. In each case, the integrand depends on a point $\boldsymbol{u} \in (0,1)^d$ and the integral we seek is $\int_{(0,1)^d} f(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$. For each function, we subtract an appropriate constant to make the integral equal to zero in all cases. This constant is already included in the definitions below. Some of the test functions are constructed using Φ and Φ^{-1}, the CDF and inverse CDF of the standard normal distribution. Some of the functions are easy for RQMC, while others are more difficult. A function that is easy for RQMC is not necessarily one that is easy to get a confidence interval for.

1. SumUeU: $f(\boldsymbol{u}) = -d + \sum_{j=1}^d u_j \exp(u_j)$. This is smooth and additive, making it very easy for RQMC methods to integrate; e.g., see Section 17.2 of Owen (2023).
2. MC2: $f(\boldsymbol{u}) = -1 + (d-1/2)^d \prod_{j=1}^d (x_j - 1/2)$. This function is from Morokoff and Caflisch (1995). It is smooth and very nearly, but not exactly, additive. This makes it typical of the integrands in weighted function spaces, such as the spaces of Sloan and Woźniakowski (1998).
3. PieceLinGauss: $f(\boldsymbol{u}) = \max\left(d^{-1/2}\sum_{j=1}^d \Phi^{-1}(u_j) - \tau, 0\right) - \varphi(\tau) + \tau\Phi(-\tau)$. This function is piecewise linear and continuous in d Gaussian inputs. It has two sources of infinite variation in the sense of Hardy and Krause (Niederreiter 1992, Section 2.2). It is unbounded and it has a cusp that brings infinite variation for $d \geq 3$. It shares these properties with some integrands encountered in computational finance, which are analytically intractable. Our construction gives this integral a known mean for all d. We use $\tau = 1$. Larger values for τ would give the problem a "rare events" character that is outside the scope of this article.
4. IndSumNormal: $f(\boldsymbol{u}) = -\Phi(-\tau) + \mathbb{I}\{d^{-1/2}\sum_{j=1}^d \Phi^{-1}(u_j) \geq \tau\}$, where $\mathbb{I}$ denotes the indicator function. This function is discontinuous with infinite variation in the sense of Hardy and Krause for $d \geq 2$. The discontinuities are not axis parallel so that they are not QMC-friendly. We use $\tau = 1$. The discreteness in the integrand produces a discrete distribution for the integral estimates. That lets us test a violation of one of the conditions of Hall (1988).
5. SmoothGauss: $f(\boldsymbol{u}) = -\Phi(1/\sqrt{2}) + \Phi(1 + d^{-1/2}\sum_{j=1}^d \Phi^{-1}(u_j))$. This function is smooth and bounded and monotone in each u_j. It has bounded variation.
6. RidgeJohnsonSU: $f(\boldsymbol{u}) = -\eta + F^{-1}(d^{-1/2}\sum_{j=1}^s u_j)$ where F is the CDF of the Johnson's SU distribution (Johnson 1949) with parameters $\gamma = \delta = \lambda = 1$ and $\xi = 0$, and η is the mean of that distribution. It has skewness -5.66 and kurtosis 96.8 (for any d) making it heavy tailed.

The last four examples use "ridge functions" of the form $g(\theta^{\mathsf{T}}\Phi^{-1}(\boldsymbol{u}))$ for a unit vector $\theta \in \mathbb{R}^d$ and a function $g : \mathbb{R} \to \mathbb{R}$, where a superscript T denotes transpose. In a plain MC setting they, before centering, have the distribution $g(N(0,1))$. If g is Lipschitz, then the ridge function has bounded mean dimension (a favorable property for RQMC) as the nominal dimension $d \to \infty$ (Hoyt and Owen 2020), while if g is discontinuous the mean dimension can grow proportionally to $\sqrt{d}$. In the ridge functions above, every variable is equally important, which tends to make integration harder. To introduce some sparsity, let $\eta_j = 2^{-j}$ and $\theta_j = \eta_j / \sqrt{\sum_{j'=1}^d \eta_{j'}^2}$. Then we can have sparse versions of the last three test functions. Sparsity does not affect the value of the integral. We do not include sparse versions in this work.

4.2 Comparison of Student and Bootstrap Confidence Intervals

Our main interest is in coverage accuracy of the confidence intervals. No nonparametric method can get exactly 95% coverage in finite samples. But if we are reasonably confident that the true coverage is below 94%, then we deem the method to have failed. In our experiments, we judged any method that attained less than 92.7% coverage to have failed. From 1000 independent draws, the binomial distribution would give less than a 4% chance of seeing this failure if the true coverage were 94% (or higher). Our 1000 draws from the pool involve some overlap so this computation is not exact, but we think it is accurate enough.

Each confidence interval method was tested in 2400 tasks: 6 integrands, 5 RQMC methods, 4 dimensions, 5 RQMC sample sizes n, and 4 values of the replication size R. The percentile method failed 1698 (70.75%) of those tasks. This is not very surprising. Its undercoverage was seen in Owen (1988), and it is not well suited to very small sample sizes like we would want for RQMC. Despite it being a pretty standard method for some bootstrap computations, it is not well regarded for setting confidence intervals for the mean. We do not consider it further.

The bootstrap t method failed 81 times, and the plain Student t confidence interval method failed only 3 times. The lowest 15 coverage levels for the plain Student t interval were all for $R = 5$ replicates. The bootstrap t had coverage over 95% in all those cases, but it also uses very long confidence intervals in those sets.

Of the 81 failures for bootstrap t, 74 were for Sob-LMS on either SumUeU (44 times) or MC2 (30 times). These are for extremely spiky histograms, which are included in the illustrations in Section 4.3. There were 6 failures for Sob-NUS and 1 for Sob-DS.

As noted in the last paragraph of Section 3, the bootstrap t has the possibility of giving an interval of infinite length. When given a sample of R distinct values, it has probability R^{1-R} of yielding $S^* = 0$. For $R = 5$, this becomes 0.0016. To get an infinite confidence interval we would need that to happen at least 25 times out of $B = 1000$. A binomial calculation shows that this event has probability approximately 1.35×10^{-21}. The probability of $S^* = 0$ is much higher if two or more of the replicate values it gets are equal to each other. There were 21 bootstrap t confidence intervals with infinite length. They were all for the integrand IndSumNormal with $R = 5$. The sample value for this integrand counts the number of events that happen. It has a discrete distribution that may have many fewer than 2^k different values. Getting two equal values among the R samples makes it much more likely that we see $S^* = 0$ and an infinite length interval. The bootstrap t had many very long interval lengths for the IndSumNormal integrand. This problem could then arise in bootstrap t confidence intervals for the value of a digital option in finance.

A second kind of failure is having coverage too high. This means that the confidence intervals could have been shorter. It is usually considered a lesser failure than having coverage too low. The standard intervals had coverage higher than 97% 81 times. This is the threshold at which we are confident that the actual coverage is above 96%. That only happened for the smooth additive integrand SumUeU and the smooth nearly additive integrand MC2. Of those 81 cases, 80 of them were for Sob-LMS and one was for Sob-NUS. For Sob-LMS we know that the underlying kurtosis of the RQMC points diverges to infinity as n increases from results in Pan and Owen (2023).

4.3 Some Histograms

We looked at many histograms made up of $N = 10^4$ integral estimates from the pools of values that we created. The majority of them looked nearly Gaussian. Some of that is to be expected when one averages $n = 2^k$ different random values. If those had been iid values, then the CLT would apply and the averages would have skewness $\gamma/\sqrt{n}$ and kurtosis κ/n, where γ and κ are the skewness and kurtosis of a single random integrand value. The n values in RQMC sampling are identically distributed but far from independent.

Figure 1 shows a selection of histograms, primarily the unusual ones. The upper left entry is a somewhat negatively skewed example from the RidgeJohnsonSU integrand. It is shown for the Sob-DS method, but that skewed pattern appeared for other methods too and not just for small d and k. The upper right entry shows the "spike plus outliers" distribution mentioned earlier. It also arises for nearly additive MC2 at somewhat higher values of k for each d. The phenomenon was pointed out by Pan and Owen (2023) in one dimension and in Pan and Owen (2022) for multiple dimensions which includes some numerical illustrations. We expect the interval length to be quite variable in settings like this because, for example, $R = 10$ replicates might not include any of the outliers. The middle left entry shows a bimodal distribution that frequently appears for lattice samples with a baker transformation. It was also illustrated in L'Ecuyer (2018) and in Chapter 17.4 of Owen (2023). This bimodality did not always happen in our simulations. In

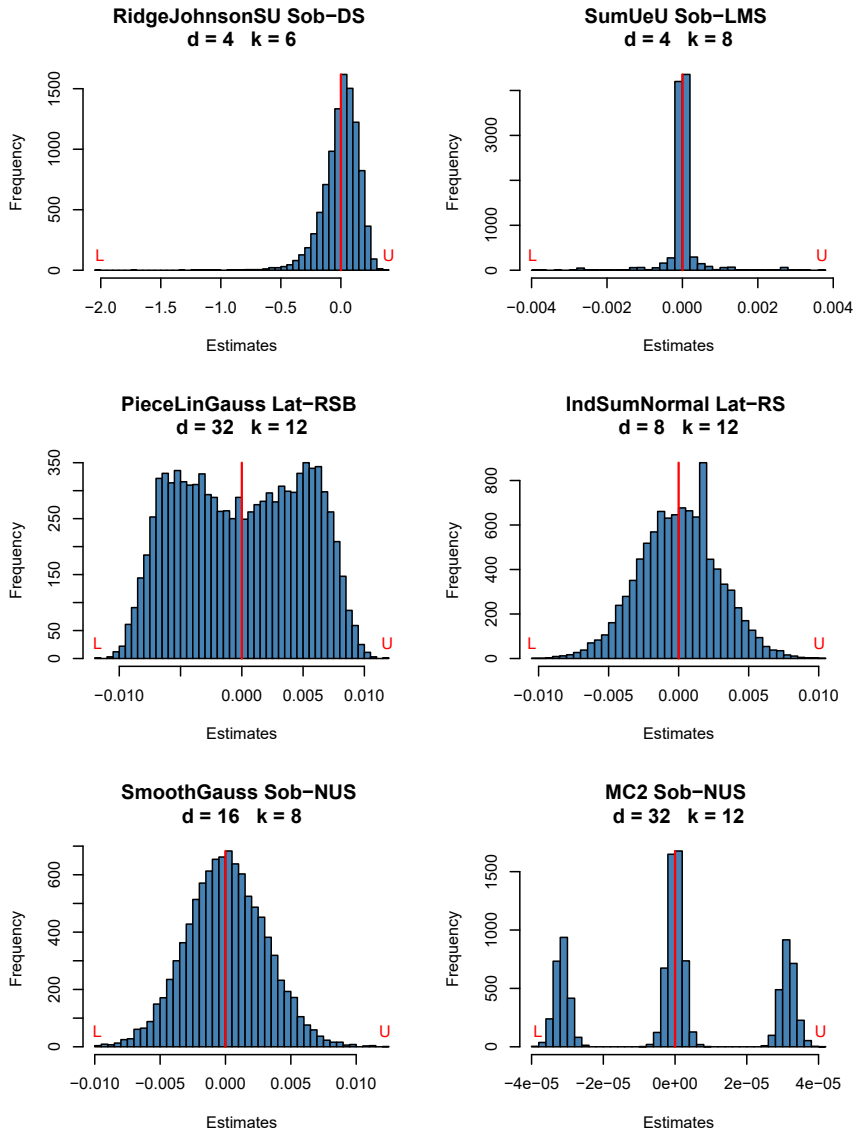

Figure 1: Some histograms of $N = 10{,}000$ integral estimates, with $k = \log_2(n)$. In each case L and U mark the lower and upper exteme sample values and a red vertical line marks the integrand value 0.

more extreme settings, the distribution is nearly U shaped. The middle right entry shows a phenomenon for the IndSumNormal integrand. It is roughly Gaussian plus a spike near one value. For smaller values of k, this integrand gives a discrete distribution with a roughly binomial appearance. The lower left entry is a roughly Gaussian histogram, similar to most of those in the data set. The lower right entry is a histogram for Sob-NUS that looks a lot like many that arise for Sob-LMS but which we have never before seen for Sob-NUS.

Some of our examples showed very extreme kurtosis. None of them showed extreme skewness. The standard confidence interval is known to have robust coverage in response to kurtosis at the expense of having very long confidence intervals, while being vulnerable to skewness. Figure 2 shows a summary of interval length and coverage probability versus kurtosis and skewness for the case $R = 10$. There we see that kurtosis brings above nominal coverage for the standard t intervals. We see interval length decreasing

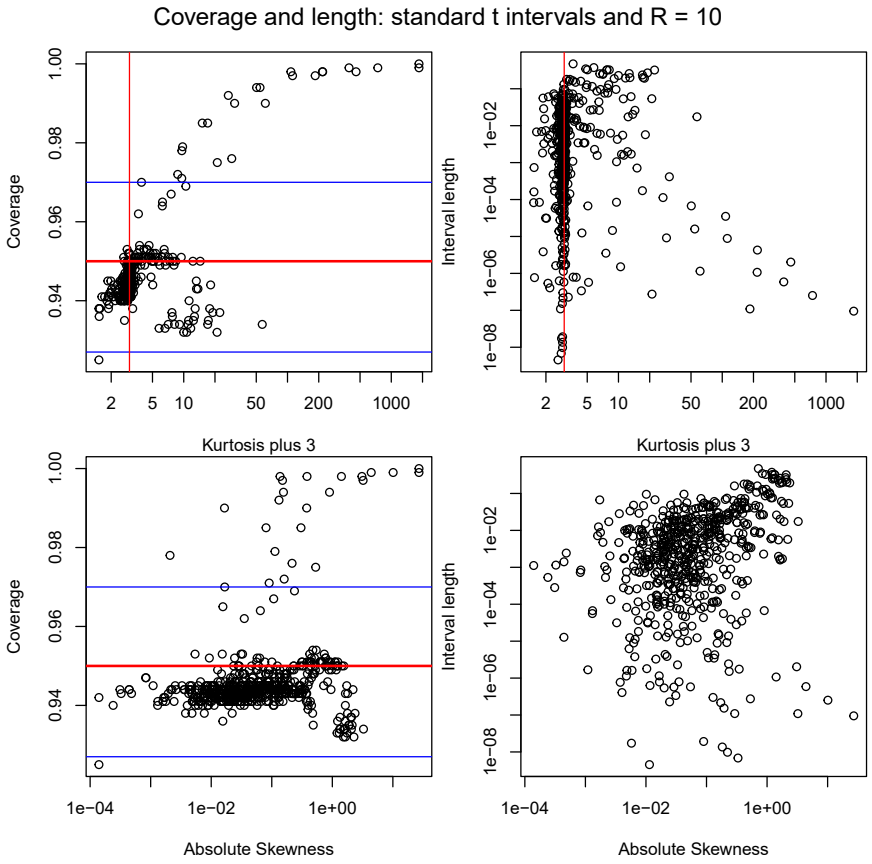

Figure 2: These figures describe coverage and length of normal theory confidence intervals for $R = 10$ replicates. Vertical kurtosis lines are at 0 (for the normal distribution). Horizontal reference lines are at the nominal level as well as at levels where we conclude that coverage is outside the interval $[0.94, 0.96]$.

with extreme kurtosis. That is what happens for Sob-LMS with SumUeU and MC2 where the RQMC pool is made up of spikes plus outliers. See the upper right histogram in Figure 1. For small R, the method only rarely gets an outlier, so most of those confidence intervals are extremely short. It is then a pleasant outcome that it covers the true mean often enough. We are cautious about expecting that to generalize beyond our examples.

5 CONCLUSION

In our examples, we saw that plain normal theory confidence intervals for RQMC performed best overall. Our criterion was simply whether the two-sided interval covered well. This was somewhat surprising as the bootstrap t method had much better coverage in the simulations reported in Choquet et al. (1999) and Owen (1988), even for numbers of replications as low as 5. We can understand this by noticing that the RQMC estimates we considered never had extreme skewness. That would happen if a CLT applied to them. However, the CLT does not always apply and several of the histograms are strongly non-Gaussian while still nearly symmetric. A similar phenomenon of symmetric but non-Gaussian limiting distributions arises in a strategy that recycles physically generated random numbers (Owen 2009). Pan and Owen (2022) show that the RQMC error in Sob-LMS is a randomly weighted sum of the average of some Walsh functions. Those Walsh function averages have a symmetric distribution, but because the random weights are dependent, it does not automatically mean that the RQMC error has a symmetric distribution.

The case of SumNormalInd highlights an issue with the bootstrap t not working on discrete random variables. Discrete random variables supported on an arithmetic sequence are not included in the theory of Hall (1988). A distribution that is a mixture of a discrete and a continuous component will also face difficulties due to the discrete component yielding more frequent $S^* = 0$ when R is small.

The standard normal theory intervals are known to underperform the bootstrap t asymptotically in the case where one-sided intervals are desired. There the coverage error is $O(1/\sqrt{n})$ for the standard method and $O(1/n)$ for the bootstrap t. If however, symmetry is a ubiquitious property of RQMC estimates, then that advantage of the bootstrap t disappears.

We have avoided running any examples that include rare events. For those it is necessary to employ a method such as importance sampling in order to get reasonable accuracy from RQMC. It is possible that rare-event settings without importance sampling will produce very skewed histograms of RQMC points.

ACKNOWLEDGMENTS

The work of P. L'Ecuyer was supported by a Discovery Grant RGPIN-2018-05795 from NSERC-Canada. The work of A. B. Owen was supported by NSF grants IIS-1837931 and DMS-2152780.

REFERENCES

Bahadur, R. R., and L. J. Savage. 1956. "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems". *Annals of Mathematical Statistics* 27(4):1115–1122.

Choquet, D., P. L'Ecuyer, and C. Léger. 1999. "Bootstrap Confidence Intervals for Ratios of Expectations". *ACM Transactions on Modeling and Computer Simulation* 9(4):326–348.

Cranley, R., and T. N. L. Patterson. 1976. "Randomization of Number Theoretic Methods for Multiple Integration". *SIAM Journal on Numerical Analysis* 13(6):904–914.

Dick, J., P. Kritzer, and F. Pillichshammer. 2022. *Lattice Rules: Numerical Integration, Approximation, and Discrepancy*. Springer Nature.

Dick, J., and F. Pillichshammer. 2010. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge, U.K.: Cambridge University Press.

Efron, B. 1981. "Nonparametric Standard Errors and Confidence Intervals". *Canadian Journal of Statistics* 9(2):139–158.

Efron, B., and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.

Hall, P. 1988. "Theoretical Comparison of Bootstrap Confidence Intervals". *Annals of Statistics* 16(3):927–953.

Hall, P. 1992. *The Bootstrap and Edgeworth Expansions*. New York: Springer.

Hickernell, F. J. 2002. "Obtaining $O(N^{-2+\varepsilon})$ Convergence for Lattice Quadrature Rules". In *Monte Carlo and Quasi-Monte Carlo Methods 2000*, edited by K.-T. Fang, F. J. Hickernell, and H. Niederreiter, 274–289. Berlin: Springer-Verlag.

Hoyt, C. R., and A. B. Owen. 2020. "Mean Dimension of Ridge Functions". *SIAM Journal on Numerical Analysis* 58(2):1195–1216.

Joe, S., and F. Y. Kuo. 2008. "Constructing Sobol Sequences with Better Two-Dimensional Projections". *SIAM Journal on Scientific Computing* 30(5):2635–2654.

Johnson, N. L. 1949. "Systems of Frequency Curves Generated by Methods of Translation". *Biometrika* 36(12):149–176.

L'Ecuyer, P. 2009. "Quasi-Monte Carlo Methods with Applications in Finance". *Finance and Stochastics* 13(3):307–349.

L'Ecuyer, P. 2018. "Randomized Quasi-Monte Carlo: An Introduction for Practitioners". In *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC 2016*, edited by P. W. Glynn and A. B. Owen, 29–52. Berlin: Springer.

L'Ecuyer, P. 2023. "SSJ: Stochastic Simulation in Java". https://github.com/umontreal-simul/ssj, accessed 27th April 2023.

L'Ecuyer, P., and C. Lemieux. 2000. "Variance Reduction via Lattice Rules". *Management Science* 46(9):1214–1235.

L'Ecuyer, P., and C. Lemieux. 2002. "Recent Advances in Randomized Quasi-Monte Carlo Methods". In *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, edited by M. Dror, P. L'Ecuyer, and F. Szidarovszky, 419–474. Boston: Kluwer Academic.

L'Ecuyer, P., P. Marion, M. Godin, and F. Puchhammer. 2022. "A Tool for Custom Construction of QMC and RQMC Point Sets". In *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC 2020*, edited by A. Keller, 51–70. Berlin: Springer.

L'Ecuyer, P., and D. Munger. 2016. "Algorithm 958: Lattice Builder: A General Software Tool for Constructing Rank-1 Lattice Rules". *ACM Transactions on Mathematical Software* 42(2):Article 15.

L'Ecuyer, P., D. Munger, and B. Tuffin. 2010. "On the Distribution of Integration Error by Randomly-Shifted Lattice Rules". *Electronic Journal of Statistics* 4:950–993.

Lemieux, C. 2009. *Monte Carlo and Quasi-Monte Carlo Sampling*. Series in Statistics. New York: Springer.

Loh, W.-L. 2003. "On the Asymptotic Distribution of Scrambled Net Quadrature". *Annals of Statistics* 31(4):1282–1324.

Matoušek, J. 1998. "On the L_2-discrepancy for Anchored Boxes". *J. of Complexity* 14:527–556.
Morokoff, W. J., and R. E. Caflisch. 1995. "Quasi-Monte Carlo Integration". *Journal of Computational Physics* 122:218–230.
Nakayama, M. K., and B. Tuffin. 2021. "Sufficient Conditions for Central Limit Theorems and Confidence Intervals for Randomized Quasi-Monte Carlo Methods". Technical report hal-03196085, INRIA. https://hal.inria.fr/hal-03196085, Accessed 14th August 2023.
Niederreiter, H. 1992. *Random Number Generation and Quasi-Monte Carlo Methods*, Volume 63. Philadelphia: SIAM.
Owen, A. B. 1988. "Small Sample Central Confidence Intervals for the Mean". Technical Report 302, Stanford University, Department of Statistics. https://purl.stanford.edu/mz765np4744, Accessed 14th August 2023.
Owen, A. B. 1995. "Randomly Permuted (t,m,s)-Nets and (t,s)-Sequences". In *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing: Lecture Notes in Statistics*, Volume 106, 299–317. Springer.
Owen, A. B. 1997. "Monte Carlo Variance of Scrambled Net Quadrature". *SIAM Journal of Numerical Analysis* 34:1884–1910.
Owen, A. B. 2003. "Variance with Alternative Scramblings of Digital Nets". *ACM Transactions on Modeling and Computer Simulation* 13(4):363–378.
Owen, A. B. 2009. "Recycling Physical Random Numbers". *Electronic Journal of Statistics* 3:1531–1541.
Owen, A. B. 2023. *Practical Quasi-Monte Carlo*. Draft available at https://artowen.su.domains/mc/practicalqmc.pdf. Accessed 14th August 2023.
Owen, A. B., and D. Rudolf. 2021. "A Strong Law of Large Numbers for Scrambled Net Integration". *SIAM Review* 63(2):360–372.
Pan, Z., and A. B. Owen. 2022. "Super-Polynomial Accuracy of Multidimensional Randomized Nets Using the Median-of-Means". Technical report, arXiv:2208.05078.
Pan, Z., and A. B. Owen. 2023. "Super-Polynomial Accuracy of One Dimensional Randomized Nets Using the Median of Means". *Mathematics of Computation* 92(340):805–837.
Sloan, I. H., and H. Woźniakowski. 1998. "When are Quasi-Monte Carlo Algorithms Efficient for High Dimensional Integration?". *Journal of Complexity* 14:1–33.
Tuffin, B. 1998. "Variance Reduction Order Using Good Lattice Points in Monte Carlo Methods". *Computing* 61(4):371–378.
Tuffin, B. 2004. "Randomization of Quasi-Monte Carlo Methods for Error Estimation: Survey and Normal Approximation". *Monte Carlo Methods and Applications* 10(3-4):617–628.
U.S. Nuclear Regulatory Commission 2007. "Guidelines for Evaluating Fatigue Analyses Incorporating the Life Reduction of Metal Components Due to the Effects of the Light-Water Reactor Environment for New Reactors". U.S. Nuclear Regulatory Commission Regulatory Guide 1.207, U.S. Nuclear Regulatory Commission, Washington, DC.
Yue, R.-X., and F. J. Hickernell. 2002. "The Discrepancy and Gain Coefficients of Scrambled Digital Nets". *Journal of Complexity* 18(1):135–151.

AUTHOR BIOGRAPHIES

PIERRE L'ECUYER is a Professor in the Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Canada. He is a member of the CIRRELT and GERAD. His main research interests are random number generation, quasi-Monte Carlo methods, efficiency improvement via variance reduction, sensitivity analysis and optimization of discrete-event stochastic systems. He has published over 300 scientific articles and has developed software libraries and systems for random number generation and stochastic simulation. Web page: http://www.iro.umontreal.ca/~lecuyer. Email: lecuyer@iro.umontreal.ca.

MARVIN K. NAKAYAMA is a professor in the Department of Computer Science at the New Jersey Institute of Technology. He received an M.S. and Ph.D. in operations research from Stanford University and a B.A. in mathematics-computer science from U.C. San Diego. He is an associate editor for *ACM Transactions on Modeling and Computer Simulation*, and served as the simulation area editor for the *INFORMS Journal on Computing* from 2007–2016. His research interests include simulation, modeling, statistics, risk analysis, and energy. His email: marvin@njit.edu.

ART B. OWEN is the Max H. Stein Professor of Statistics at Stanford University. His research interests include quasi-Monte Carlo sampling, nonparametric confidence intervals, and variable importance measures. Web page: https://artowen.su.domains/. Email: owen@stanford.edu.

BRUNO TUFFIN received his PhD degree in applied mathematics from the University of Rennes 1, France, in 1997. Since then, he has been with Inria in Rennes. His research interests include developing Monte Carlo and quasi-Monte Carlo simulation techniques for the performance evaluation of telecommunication systems and telecommunication-related economical models. He is currently Area Editor for *INFORMS Journal on Computing* and Associate Editor for *ACM Transactions on Modeling and Computer Simulation* and *Queuing Systems*. His email: bruno.tuffin@inria.fr.