

QMC Sampling from Empirical Datasets

Fei Xie, Michael B. Giles, and Zhijian He

Abstract This paper presents a simple idea for the use of quasi-Monte Carlo sampling with empirical datasets, such as those generated by MCMC methods. It also presents and analyses a related idea of taking advantage of the Hilbert space-filling curve. Theoretical and numerical analyses are provided for both. We find that when applying the proposed QMC sampling methods to datasets coming from a known distribution, they give similar performance as the standard QMC method directly sampling from this known distribution.

1 Introduction

If U is a random variable uniformly distributed over the d -dimensional cube $[0, 1]^d$, the standard MC (Monte Carlo) estimate for $\mathbb{E}[f(U)]$ using n i.i.d. sample points is

$$\bar{f}_n = \frac{1}{n} \sum_{j=1}^n f(U_j)$$

If the variance $\mathbb{V}[f]$ is finite then the error is $O(n^{-1/2})$. The QMC (Quasi-Monte Carlo) estimate is similar, except that the n points are chosen carefully [3] so that the error is $O((\log n)^d/n)$ for integrands satisfying appropriate conditions [8]. This extends naturally to more general distributions, with

Fei Xie
Tsinghua University, Beijing, China e-mail: xiefei901108@126.com

Michael B. Giles
Mathematical Institute, University of Oxford, UK e-mail: mike.giles@maths.ox.ac.uk

Zhijian He
South China University of Technology, Guangzhou, China e-mail: hezhijian@scut.edu.cn

$$\mathbb{E}[f(X)] \approx \bar{f}_n = \frac{1}{n} \sum_{j=1}^n f(X_j)$$

where for the MC estimate the X_j are i.i.d. samples from the distribution of the random variable X . For the corresponding QMC estimate $X_j = \psi(U_j)$ with the U_j being the same set of QMC points in the unit cube as before, and $\psi(U)$ is a mapping which gives the desired probability density function for X , i.e., if X has probability density function $\rho(x)$ then it is related to the determinant of the Jacobian of the mapping through

$$\rho(\psi(u)) = \left| \det \left(\frac{\partial \psi}{\partial u} \right) \right|^{-1}.$$

The question we address in this paper is what should we do when we do not know $\rho(x)$, and instead all we have is a large set of samples $X_j, j = 1, \dots, N$ which constitutes an empirical dataset?

The medical decision application [1] which motivated this research has inputs which are multiple independent multi-dimensional random variables. Some of these come from known distributions, but some correspond to Bayesian posterior distributions defined solely by a set of values generated by an MCMC process. In order to obtain a given number of samples from this posterior distribution, currently these MCMC datasets are approximated by multi-variate Gaussian distributions, and then QMC can be used very effectively. However, it might be better to work directly with these empirical datasets without introducing any approximation error, and the question is then how to use QMC. Although the overall application is fairly high-dimensional (15 – 40), the individual MCMC datasets are low-dimensional (2 or 3).

For the purpose of sampling from an empirical dataset using QMC, we propose a novel QMC sampling method based on recursive bisection. The proposed QMC sampling method is described in Section 2, and an error analysis for the case $d = 1$ is given. Section 3 presents an alternative method based on a Hilbert space-winding curve, and a corresponding numerical analysis. Section 4 presents numerical results which compare the two QMC sampling methods with the standard QMC method. Conclusions are given in Section 5.

2 QMC Sampling

Suppose we have an empirical dataset of $N = 2^M$ d -dimensional samples $\{X_j, j = 1, \dots, N\}$, where M is a positive integer. With QMC sampling, the challenge is to construct a mapping from QMC points in the hypercube $(0, 1)^d$ to points X_i in the dataset.

2.1 Recursive Bisection-based Sorting

Borrowing ideas from recursive graph bisection for distributed-memory parallel computing [2, 6] the central idea is to repeatedly bisect the dataset in alternating directions.

For example, for a 2-dimensional dataset $\{(x_1, x_2)_j, j = 1, \dots, N\}$, firstly we find the median value $x_{\text{mid}}^{(1)}$ of their x_1 coordinates by sorting the points based on x_1 , and then re-arrange the set into two halves, the first half with $x_1 < x_{\text{mid}}^{(1)}$ and the second half with $x_1 > x_{\text{mid}}^{(1)}$; if there are two or more points for which $x_1 = x_{\text{mid}}^{(1)}$ they can be assigned to the two halves based on their indices to achieve a bisection. Each of the two halves is then sorted within each half based on x_2 , and bisected and re-arranged using the median value of their x_2 coordinates (with the median being different for each half). In the next step, the points in each quarter are sorted again based on x_1 , and each quarter is further bisected. This is repeated, alternating between the x_1 and x_2 coordinates, until each subset has just one point. The empirical dataset is thus sorted by the recursive bisection procedure. The process is illustrated in Figure 1 for a small dataset with just 16 sample points.

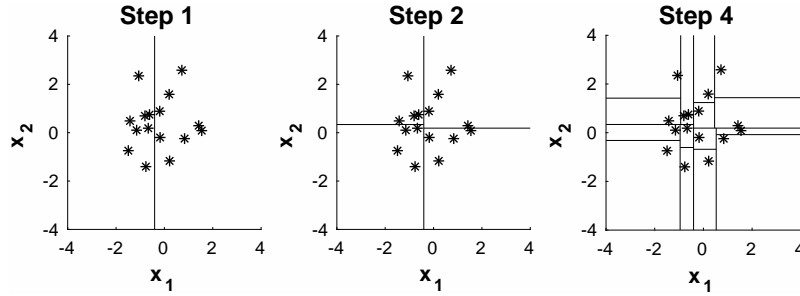


Fig. 1 Steps 1, 2, and 4 in the recursive bisection of a 2-dimensional dataset with 16 points.

2.2 Mapping of QMC Points

For application of QMC on selecting samples from a dataset of size N , firstly the dataset is recursively bisected into N subsets. Then the location of the subsets to be selected is characterized by the relative position of the QMC points in the hypercube $(0, 1)^d$. In this way, we pass the uniformity of QMC points into an integer index between 0 and $N - 1$ representing the particular position in a data sequence. In this subsection, we construct a mapping from points in $(0, 1)^d$ to the dataset sorted by the recursive bisection procedure. It is then straightforward to use QMC (or randomized QMC) points to sample from the sorted dataset.

Let us start with the simple case $M = d \times k$, where k is a positive integer. Suppose the low discrepancy sequence we use is the Sobol' sequence [10], which is a (t, d) -sequence in base 2 for all d . To map a point $Q = (q_1, \dots, q_d) \in (0, 1)^d$ from the hypercube $(0, 1)^d$ to one of the indices $\{0, 1, \dots, 2^{dk} - 1\}$, there are two steps. In the first step, the point Q is mapped to the d -dimensional integer space $\{0, \dots, 2^k - 1\}^d$ by

$$(q_1, \dots, q_d) \mapsto (\lfloor 2^k q_1 \rfloor, \dots, \lfloor 2^k q_d \rfloor),$$

where $\lfloor x \rfloor$ denotes x rounded down to the nearest integer.

In the second step, the d -dimensional integer index $(i_1, \dots, i_d) \in \{0, \dots, 2^k - 1\}^d$ needs to be bijectively mapped onto an integer index $i \in \{0, \dots, N - 1\}$ as the data points are stored in a sequence of length N . To this end, we write each integer coordinate of the index (i_1, \dots, i_d) in its k -bit binary representation $(i_j)_2 = b_{j,k-1} \dots b_{j,0}$, $j = 1, \dots, d$, where

$$i_j = \sum_{\ell=0}^{k-1} b_{j,\ell} 2^\ell.$$

Each bit of i_j 's binary representation can be viewed as the indication of which half the desired point is located in each time the dataset is bisected according to the j -th coordinate. i.e., $b_{j,k-1}$ indicates whether the desired point is in the first half or the second when the dataset is bisected according to the j -th coordinate for the first time, $b_{j,k-2}$ indicates when the dataset is bisected in this direction for the second time, and so on. As we are doing the bisection in alternating directions, interleaving the bits of each integer coordinate gives the complete information of selection in each successive bisection step

$$(i)_2 = b_{1,k-1} \dots b_{d,k-1} b_{1,k-2} \dots b_{d,k-2} \dots b_{1,0} \dots b_{d,0},$$

which can then be easily transformed into the corresponding index

$$i = \sum_{\ell=0}^{k-1} 2^{d\ell} \left(\sum_{j=1}^d b_{j,\ell} 2^{d-j} \right) \in \{0, \dots, N - 1\}.$$

To better understand the above formula, let us consider the 2-dimensional index $(i_1, i_2) = (5, 2)$. Writing it in 3-bit binary representation and interleaving their bits results in the 6-bit binary representation of the index i .

$$\begin{aligned} i_1 &= 1 & 0 & 1 \\ i_2 &= & 0 & 1 & 0 \\ \Rightarrow i &= 1 & 0 & 0 & 1 & 1 & 0 \end{aligned}$$

Thus, the index i corresponding to $(i_1, i_2) = (5, 2)$ is 38 in a dataset with $N = 2^{2 \times 3} = 64$ points.

For computational efficiency, a final step is to re-arrange the dataset once more so that the (i_1, \dots, i_d) point is stored at linear memory location

$$\sum_{\ell=1}^d 2^{(\ell-1)k} i_{\ell}.$$

This simplifies the subsequent implementation of the QMC algorithm.

Now we generalize to the case $M = d \times k - c$ where $0 < c < k$. For this case, the point Q is firstly mapped to the d -dimensional integer space $\Omega(d, k, c) = \{0, \dots, 2^k - 1\}^{d-c} \times \{0, \dots, 2^{k-1} - 1\}^c$, where the last c coordinates are in $\{0, \dots, 2^{k-1} - 1\}$, and the remaining coordinates are in $\{0, \dots, 2^k - 1\}$. This is achieved by

$$(q_1, \dots, q_d) \mapsto (\lfloor 2^k q_1 \rfloor, \dots, \lfloor 2^k q_{d-c} \rfloor, \lfloor 2^{k-1} q_{d-c+1} \rfloor, \dots, \lfloor 2^{k-1} q_d \rfloor).$$

Similar to the case where $c = 0$, each integer coordinate of the index $(i_1, \dots, i_d) \in \Omega(d, k, c)$ is written in its binary representation, but with different bits:

$$(i_j)_2 = \begin{cases} b_{j,k-1} \dots b_{j,0}, & j = 1, \dots, d-c \\ b_{j,k-1} \dots b_{j,1}, & j = d-c+1, \dots, d \end{cases}.$$

Note that here $b_{j,k-l}$ still indicates whether the desired point is in the first half or the second when the dataset is bisected according to the j -th coordinate for the l th time. The difference is that the dataset is bisected for k times in total according to each of the first $d-c$ coordinates, while it is only bisected for $k-1$ times according to each of the last c coordinates. Again, by interleaving the bits of each integer coordinate we get

$$(i)_2 = b_{1,k-1} \dots b_{d,k-1} b_{1,k-2} \dots b_{d,k-2} \dots b_{1,1} \dots b_{d,1} b_{1,0} \dots b_{d-c,0}.$$

The corresponding index can be expressed as

$$i = \sum_{\ell=1}^{k-1} 2^{d\ell-c} \left(\sum_{j=1}^d b_{j,\ell} 2^{d-j} \right) + \left(\sum_{j=1}^{d-c} b_{j,0} 2^{d-c-j} \right) \in \{0, \dots, N-1\}.$$

Again consider the example of the 2-dimensional index $(i_1, i_2) = (5, 2)$. We now calculate its corresponding index i in a dataset with $N = 2^{2 \times 3 - 1} = 32$ points. Writing i_1 and i_2 in 3-bit and 2-bit binary representation respectively, and interleaving their bits results in the 5-bit binary representation of the index i .

$$\begin{array}{rcl} i_1 & = & 1 \quad 0 \quad 1 \\ i_2 & = & \quad 1 \quad 0 \\ \Rightarrow i & = & 1 \quad 1 \quad 0 \quad 0 \quad 1 \end{array}$$

Thus, the index i corresponding to $(i_1, i_2) = (5, 2)$ is 25 in a dataset with $N = 32$ points. We see that for different numbers of points N the same 2-dimensional index (i_1, i_2) is mapped to different indices i .

2.3 Error Analysis

Let $\tau : [0, 1]^d \rightarrow \mathcal{X} = \{X_1, \dots, X_N\}$ be the mapping described above, where $N = 2^M$ for some integer $M \geq 1$. Based on the empirical dataset \mathcal{X} , one may use the following sample average

$$\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^N f(X_i) \quad (1)$$

as an estimate of $\mu = \mathbb{E}[f(X)]$. The QMC sampling for the empirical dataset gives an estimate

$$\hat{\mu}_n^I = \frac{1}{n} \sum_{i=1}^n f(\tau(U_i)), \quad (2)$$

where $U_i \in (0, 1)^d$ are RQMC (randomised QMC) points with the sample size $n \leq N$. We define the difference between these two estimates as the selection error from QMC sampling. We next study the selection error with QMC sampling when $d = 1$.

When $d = 1$, the mapping τ reduces to the inverse function of the empirical CDF of the dataset. More specifically, let the empirical CDF of the dataset be

$$F_N(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{X_i \leq x\}.$$

Then we have

$$\tau(u) = F_N^{-1}(u) := \inf\{x : F_N(x) \geq u\}.$$

Now consider the case of using a scrambled $(0, m, 1)$ -net in base $b = 2$ as inputs in (2), where $n = 2^m$ and $m \leq M$. Let $U_{(i)}$ be the order statistic of U_i , $i = 1, \dots, n$, and similarly for $X_{(i)}$. By the property of a scrambled $(0, m, 1)$ -net, we have $U_{(i)} \sim \mathbb{U}([(i-1)/n, i/n])$ independently with probability one. Note that one-dimensional stratified sampling has the same effect. Our results in this section also apply for stratified sampling.

Let $R = N/n = 2^{M-m}$, and $Y_j := \tau(U_{(i)})$. The estimate becomes

$$\hat{\mu}_n^I = \frac{1}{n} \sum_{i=1}^n f(Y_i), \quad (3)$$

where Y_j is randomly and independently selected from the block $\{X_{(iR-R+1)}, \dots, X_{(iR)}\}$ for $i = 1, \dots, n$.

To fix our idea, let's consider the simple case $X_i \stackrel{\text{iid}}{\sim} \mathbb{U}((0, 1))$ for $i = 1, \dots, N$. For this case, the k th order statistic of the uniform distribution is a Beta random variable, i.e., $X_{(k)} \sim \text{Beta}(k, N+1-k)$. We next bound the selection error $\mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)^2]$. If f is Lipschitz continuous with constant $C > 0$, we find that

$$\mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)^2 | \mathcal{X}] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \left[\left(f(Y_i) - \frac{1}{R} \sum_{j=1}^R f(X_{(iR-R+j)}) \right)^2 \middle| \mathcal{X} \right] \quad (4)$$

$$= \frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \left[\left(\frac{1}{R} \sum_{j=1}^R (f(Y_i) - f(X_{(iR-R+j)})) \right)^2 \middle| \mathcal{X} \right] \quad (5)$$

$$\leq \frac{C^2}{n^2} \sum_{i=1}^n \mathbb{E} \left[(X_{(iR)} - X_{(iR-R+1)})^2 \middle| \mathcal{X} \right]. \quad (6)$$

The first equality (4) is due to the fact that given \mathcal{X} , Y_i is randomly and independently selected from the block $\{X_{(iR-R+1)}, \dots, X_{(iR)}\}$. This gives

$$\begin{aligned} \mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)^2] &\leq \frac{C^2}{n^2} \sum_{i=1}^n \mathbb{E} \left[(X_{(iR)} - X_{(iR-R+1)})^2 \right] \\ &= \frac{C^2}{n^2} \sum_{i=1}^n \frac{(R-1)^2 + R-1}{(N+1)(N+2)} \\ &= \frac{C^2(R^2 - R)}{(N+1)(N+2)n} \\ &\leq \frac{C^2 R^2}{N^2 n} = C^2 n^{-3}, \end{aligned} \quad (7)$$

where we use Lemma 1 from the Appendix in establishing (7). More generally, we have the following theorem.

Theorem 1. Suppose that X_i are iid with a density $q(x)$ on a bounded support $\mathcal{D} = [a, b]$. Assume that $q_{\min} = \inf_{x \in \mathcal{D}} q(x) > 0$. If f is Lipschitz continuous with constant $C > 0$, then

$$\mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)^2] \leq \frac{C^2}{q_{\min}^2 n^3}. \quad (8)$$

If f is piecewise Lipschitz continuous with jumps at position $a_j, j = 1, \dots, k$, then

$$\mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)^2] \leq \frac{C_1^2}{q_{\min}^2 n^3} + \frac{4kC_2^2}{n^2}, \quad (9)$$

where C_1 is the Lipschitz constant and $C_2 = \sup_{x \in \mathcal{D}} |f(x)|$.

Proof. Note that for any $1 \leq i, j \leq N$,

$$|X_i - X_j| = |F(X_i) - F(X_j)| / q(\xi) \leq |F(X_i) - F(X_j)| / q_{\min}, \quad (10)$$

where ξ is a constant between X_i and X_j , and F is the CDF of the X_i . Let $\tilde{Y}_i = F(X_{(i)})$. Then \tilde{Y}_i are the order statistic of the uniform distribution. If f is Lipschitz continuous, by (6) and (10), we have

$$\mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)^2] \leq \frac{C^2}{n^2 q_{\min}^2} \sum_{i=1}^n \mathbb{E}[(\tilde{Y}_{iR} - \tilde{Y}_{iR-R+1})^2] \leq \frac{C^2}{q_{\min}^2} n^{-3}.$$

Now assume that f is piecewise Lipschitz continuous with jumps at position $a_j, j = 1, \dots, k$. There are at most k intervals of the form $[X_{(iR-R+1)}, X_{(iR)}]$ in which f is not Lipschitz continuous. For each of them,

$$\mathbb{E} \left[\left(\frac{1}{R} \sum_{j=1}^R (f(Y_i) - f(X_{(iR-R+j)})) \right)^2 \middle| \mathcal{X} \right] \leq 4C_2^2.$$

We thus establish (9). \square

The mean squared error (MSE) of the QMC estimate (2) can be decomposed into two terms, that is

$$\begin{aligned} \text{MSE}(\hat{\mu}_n^I) &= \mathbb{E}[(\hat{\mu}_n^I - \mu)^2] = \mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N + \hat{\mu}_N - \mu)^2] \\ &= \mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)^2] + \mathbb{E}[(\hat{\mu}_N - \mu)^2] + 2\mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)(\hat{\mu}_N - \mu)] \\ &= \mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)^2] + \mathbb{E}[(\hat{\mu}_N - \mu)^2], \end{aligned} \quad (11)$$

where we use the fact that $\mathbb{E}[\hat{\mu}_n^I - \hat{\mu}_N | \mathcal{X}] = 0$ and thus

$$\begin{aligned} \mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)(\hat{\mu}_N - \mu)] &= \mathbb{E}[\mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)(\hat{\mu}_N - \mu) | \mathcal{X}]] \\ &= \mathbb{E}[(\hat{\mu}_N - \mu) \mathbb{E}[\hat{\mu}_n^I - \hat{\mu}_N | \mathcal{X}]] = 0. \end{aligned}$$

The first term in (11) is the error due to selecting a few samples from the original datasets, and the second term is due to the original sampling error. The mean square selection error $\mathbb{E}[(\hat{\mu}_n^I - \hat{\mu}_N)^2]$ is related to n , the size of the thinned sample set, and the dimension d , and intuitively this quantity should decrease with increasing n . What we are interested in determining is the minimum size of n to ensure the selection error is no bigger than the sampling error, so that at this level of n , the accuracy of the estimate $\hat{\mu}_n^I$ is comparable to $\hat{\mu}_N$. For an iid dataset, the mean square error $\mathbb{E}[(\hat{\mu}_N - \mu)^2] = \sigma^2/N$, where σ^2 is the variance of X_i . To make the two terms in the right-hand side of (11) comparable, the optimal n is

$$n^* = \begin{cases} O(N^{1/3}), & \text{if } f \text{ is Lipschitz,} \\ O(N^{1/2}), & \text{if } f \text{ is piecewise Lipschitz.} \end{cases} \quad (12)$$

for $d = 1$. As a result, the MSE of the QMC estimate is

$$\text{MSE}(\hat{\mu}_n^I) = \begin{cases} O(n^{-3}), & \text{if } f \text{ is Lipschitz,} \\ O(n^{-2}), & \text{if } f \text{ is piecewise Lipschitz,} \end{cases} \quad (13)$$

for $n \leq n^*$ and as $N \rightarrow \infty$.

Things become more complicated for $d > 1$. We leave these cases for future research. Instead, we give numerical analysis for an alternative way of QMC sampling by making use of the Hilbert space-filling curve.

3 Sorting the Dataset via the Hilbert Space-filling Curve

Instead of using the recursive bisection mapping in Section 2, we map the dataset in dimension $d > 2$ to the one-dimensional unit interval via a Hilbert space-filling curve. The data then are simply sorted in natural order. Formally, let $H : [0, 1] \rightarrow [0, 1]^d$ be the Hilbert curve with $H(0) = (0, \dots, 0)$. It is important to noting that the mapping is not a bijection because certain points in $[0, 1]^d$ have more than one preimage through H . However, the set of such points is of Lebesgue measure 0. There exists a one-to-one Borel measurable pseudo-inverse $h : [0, 1]^d \rightarrow [0, 1]$ such that $H(h(x)) = x$ for all $x \in [0, 1]^d$. There are some properties of the Hilbert curve and its inverse h :

- For any measurable set $I \subset [0, 1]$, $\lambda_d(H(I)) = \lambda_1(I)$;
- For any $x, y \in [0, 1]$, then

$$\|H(y) - H(x)\| \leq 2\sqrt{d+3}|y-x|^{1/d}; \quad (14)$$

- If $u \sim \mathbb{U}([0, 1])$, then $H(u) \sim \mathbb{U}([0, 1]^d)$;
- If $U \sim \mathbb{U}([0, 1]^d)$, then $h(U) \sim \mathbb{U}([0, 1])$.

See [5] for the properties of the Hilbert curve. Gerber and Chopin [4] uses the inverse function h to sort the particles in the particle filtering algorithm to combine with QMC. Let \mathcal{D} be the support of X_i , and let

$$x_i = h \circ \psi(X_i) \in [0, 1], \quad (15)$$

where $\psi : \mathcal{D} \rightarrow [0, 1]^d$ is some user-chosen bijection between \mathcal{D} and $\psi(\mathcal{D}) \subset [0, 1]^d$, and X_i s are still the points of an empirical dataset as in Section 2. We may simply take $\psi(X_i) = (\psi_1(x_{i1}), \dots, \psi_d(x_{id}))$, where x_{ij} is the j th component of X_i , and the ψ_i s are continuous and strictly monotone. Particularly, if the X_i are in a bounded cube $\prod_{i=1}^d [a_i, b_i]$, then we can simply take

$$\psi_i(t) = \frac{t - a_i}{b_i - a_i}, \quad i = 1, \dots, d.$$

When $\mathcal{D} = \mathbb{R}^d$, [4] suggested to use the logistic transformation componentwise. Let $\tilde{F}_N(x)$ be the empirical CDF of the transformed dataset $\{x_1, \dots, x_N\}$:

$$\tilde{F}_N(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{x_i \leq x\}. \quad (16)$$

Let $\tilde{\tau}$ be a mapping from $[0, 1]$ to \mathcal{D} , defined by

$$\tilde{\tau}(u) = \psi^{-1}(H(\tilde{F}_N^{-1}(u))).$$

This gives an estimate

$$\hat{\mu}_n^{\text{II}} = \frac{1}{n} \sum_{i=1}^n f(\tilde{\tau}(u_i)), \quad (17)$$

where $u_i \in [0, 1]$ are one-dimensional RQMC points. We focus on the case of using a scrambled $(0, m, 1)$ -net in base $b = 2$ as inputs.

For simplicity, let's consider the case $X_j \stackrel{\text{iid}}{\sim} \mathbb{U}([0, 1]^d)$ for $j = 1, \dots, N$. Now ψ is set to the identity function. By the properties of h listed above, we have $x_j \stackrel{\text{iid}}{\sim} \mathbb{U}([0, 1])$, $j = 1, \dots, N$. Again, by the properties of a scrambled $(0, m, 1)$ -net, we have the order statistic $u_{(i)} \sim \mathbb{U}([(i-1)/n, i/n])$ independently with probability one. Let $Y_i := \tilde{\tau}(u_{(i)})$ $i = 1, \dots, n$, and let $\tilde{Y}_j = \psi^{-1} \circ H(x_{(j)})$, $j = 1, \dots, N$. The Hilbert curve based estimate becomes

$$\hat{\mu}_n^{\text{II}} = \frac{1}{n} \sum_{i=1}^n f(Y_i), \quad (18)$$

where Y_i is randomly and independently selected from the block $\{\tilde{Y}_{iR-R+1}, \dots, \tilde{Y}_{iR}\}$ for $i = 1, \dots, n$. Thus if f is Lipschitz continuous with constant $C > 0$, by (14), we have

$$\begin{aligned} |f(Y_i) - f(\tilde{Y}_{iR-R+j})| &\leq C \|Y_i - \tilde{Y}_{iR-R+j}\| \\ &\leq C \sup_{k, \ell=1, \dots, R} \|\tilde{Y}_{iR-R+k} - \tilde{Y}_{iR-R+\ell}\| \\ &= C \sup_{k, \ell=1, \dots, R} \|\psi^{-1} \circ H(x_{(iR-R+k)}) - \psi^{-1} \circ H(x_{(iR-R+\ell)})\| \\ &= C \sup_{k, \ell=1, \dots, R} \|H(x_{(iR-R+k)}) - H(x_{(iR-R+\ell)})\| \\ &\leq 2C\sqrt{d+3}(x_{(iR)} - x_{(iR-R+1)})^{1/d} \end{aligned} \quad (19)$$

for $j = 1, \dots, R$. Similarly to (5), we have

$$\mathbb{E}[(\hat{\mu}_n^{\text{II}} - \hat{\mu}_N)^2 | \mathcal{X}] \leq \frac{4C^2(d+3)}{n^2} \sum_{i=1}^n \mathbb{E} \left[(x_{(iR)} - x_{(iR-R+1)})^{2/d} \middle| \mathcal{X} \right]. \quad (20)$$

As a result, for any $d \geq 2$,

$$\begin{aligned}\mathbb{E}[(\hat{\mu}_n^\Pi - \hat{\mu}_N)^2] &\leq \frac{4C^2(d+3)}{n^2} \sum_{i=1}^n \mathbb{E}[(x_{(iR)} - x_{(iR-R+1)})^{2/d}] \\ &\leq \frac{4C^2(d+3)}{n^2} \sum_{i=1}^n (\mathbb{E}[x_{(iR)} - x_{(iR-R+1)}])^{2/d}\end{aligned}\quad (21)$$

$$= \frac{4C^2(d+3)}{n^2} \sum_{i=1}^n \left(\frac{R-1}{N+1}\right)^{2/d} \quad (22)$$

$$\begin{aligned}&\leq \frac{4C^2(d+3)}{n} \left(\frac{R-1}{N+1}\right)^{2/d} \\ &\leq 4C^2(d+3)n^{-1-2/d},\end{aligned}\quad (23)$$

where in (21) we use Jensen's inequality (because $2/d \leq 1$), and we use Lemma 1 in establishing (22).

Theorem 2. Suppose that X_i are iid with a density $q(X)$ on a bounded support $\mathcal{D} = \prod_{i=1}^d [a_i, b_i]$. Assume that $q_{\min} = \inf_{X \in \mathcal{D}} q(X) > 0$. Let

$$\psi_i(t) = \frac{t - a_i}{b_i - a_i}, \quad i = 1, \dots, d.$$

If f is Lipschitz continuous with constant $C > 0$, then

$$\mathbb{E}[(\hat{\mu}_n^\Pi - \hat{\mu}_N)^2] \leq \frac{4C^2 L_1^2 (d+3)}{q_{\min}^{2/d} L_2^2 n^{1+2/d}}, \quad (24)$$

where $L_1 = \max_{i=1, \dots, d} \{b_i - a_i\}$ and $L_2 = \min_{i=1, \dots, d} \{b_i - a_i\}$.

Proof. Let $F_1(t)$ be the CDF of $x = h \circ \psi(X)$. For any $t_1 < t_2$, we have

$$\begin{aligned}F_1(t_2) - F_1(t_1) &= P[t_1 < h \circ \psi(X) \leq t_2] = P[\psi(X) \in H((t_1, t_2])] \\ &= \prod_{i=1}^d (b_i - a_i) \int_{Y \in H((t_1, t_2])} q(\psi^{-1}(Y)) dY \\ &\geq q_{\min} \prod_{i=1}^d (b_i - a_i) \lambda_d(H((t_1, t_2])) \\ &= q_{\min}(t_2 - t_1) \prod_{i=1}^d (b_i - a_i).\end{aligned}$$

Let $\tilde{x}_i = F_1(x_i)$, $i = 1, \dots, N$. Then for any $1 \leq i, j \leq N$,

$$|x_i - x_j| \leq \frac{|\tilde{x}_i - \tilde{x}_j|}{q_{\min} \prod_{i=1}^d (b_i - a_i)}. \quad (25)$$

Then $\tilde{x}_{(i)}$ s are the order statistics of the uniform distribution. By (19) and (25), we have

$$\begin{aligned}
|f(Y_i) - f(\tilde{Y}_{iR-R+j})| &\leq C \sup_{k,\ell=1,\dots,R} \|\psi^{-1} \circ H(x_{(iR-R+k)}) - \psi^{-1} \circ H(x_{(iR-R+\ell)})\| \\
&\leq CL_1 \sup_{k,\ell=1,\dots,R} \|H(x_{(iR-R+k)}) - H(x_{(iR-R+\ell)})\| \\
&\leq 2CL_1 \sqrt{d+3} (x_{(iR)} - x_{(iR-R+1)})^{1/d} \\
&\leq \frac{2CL_1 \sqrt{d+3}}{q_{\min}^{1/d} \prod_{i=1}^d (b_i - a_i)^{1/d}} (\tilde{x}_{(iR)} - \tilde{x}_{(iR-R+1)})^{1/d} \\
&\leq \frac{2CL_1 \sqrt{d+3}}{q_{\min}^{1/d} L_2} (\tilde{x}_{(iR)} - \tilde{x}_{(iR-R+1)})^{1/d},
\end{aligned}$$

where $L_1 = \max_{i=1,\dots,d} \{b_i - a_i\}$, $L_2 = \min_{i=1,\dots,d} \{b_i - a_i\}$. If f is Lipschitz continuous, similarly to (23), we have

$$\mathbb{E}[(\hat{\mu}_n^{\text{II}} - \hat{\mu}_N)^2] \leq \frac{4C^2 L_1^2 (d+3)}{n^2 q_{\min}^{2/d} L_2^2} \sum_{i=1}^n \mathbb{E}[(\tilde{x}_{(iR)} - \tilde{x}_{(iR-R+1)})^{2/d}] \leq \frac{4C^2 L_1^2 (d+3)}{q_{\min}^{2/d} L_2^2 n^{1+2/d}}.$$

□

As a result, for Lipschitz functions in dimensions $d \geq 2$, we may take $n = O(N^{d/(d+2)})$ so that $\text{MSE}(\hat{\mu}_n^{\text{II}}) = O(n^{-1-2/d})$. This error rate is in line with the one reported in [5], where for an RQMC quadrature based on the Hilbert curve an MSE of size $O(n^{-1-2/d})$ is found for a class of Lipschitz functions.

4 Numerical Tests

To test the effectiveness of the QMC sampling methods, we apply it to datasets generated from four different distributions to estimate $\mu = \mathbb{E}[f(X_1, \dots, X_d)]$ with two different integrand functions, and compare the accuracy to the standard MC and QMC methods applied to the same problems.

The integrands include a continuous one

$$f(x_1, \dots, x_d) = \prod_{i=1}^d (x_i^2 + 1), \quad (26)$$

and a discontinuous one

$$f(x_1, \dots, x_d) = 1\{\sum_{i=1}^d x_i > d/2\}. \quad (27)$$

The four distributions, based on independent Normals $Z_1, Z_2, Z_3 \stackrel{\text{iid}}{\sim} N(0, 1)$ and uniformly distributed $U \sim U(0, 1)$ are:

- Correlated Normals: $X_1 = Z_1 + 0.5Z_2, X_2 = Z_1 - 0.5Z_2,$

- Distorted Gaussian: $X_1 = Z_1$ and $X_2 = X_1^2 + Z_2$;
- Double well: $X_1 = \text{sign}(U - 0.5) (2 + \Phi^{-1}(2|U - 0.5|))$ and $X_2 = 2Z_2$
- Unit ball surface: $X_i = Z_i / \sqrt{\sum_{j=1}^3 Z_j^2}, i = 1, 2, 3$

The empirical dataset is generated via crude MC sampling. We focus on four schemes of sampling:

- Standard MC (labelled as “std MC”), which estimates it by averaging function values based on n random samples;
- Standard QMC (labelled as “std QMC”), which generates samples from the two or three dimensional Sobol’ points u using the mapping $z = \Phi^{-1}(u)$ to generate the quasi-Normal random vectors z ;
- QMC sampling based on recursive bisection (labelled as “bQMC”), which selects n samples from a dataset of size $N = 2^{20}$ using two or three dimensional Sobol’ points according to the procedure described in Section 2.
- QMC sampling based on the Hilbert curve (labelled as “hQMC”), which also selects n samples from the Hilbert sorted dataset of size $N = 2^{20}$.

Under each scheme, the RMS (root mean squared) error is calculated for different numbers of samples $n = 2^m$, $m = 1, \dots, 19$ and plotted for comparison. To obtain these values, all experiments are replicated $r = 256$ times, resulting in estimates $\hat{\mu}_{N,j}$ (dataset), $\hat{\mu}_{n,j}^I$ (bQMC) and $\hat{\mu}_{n,j}^{II}$ (hQMC), respectively, for $j = 1, \dots, r$. The Sobol’ points are randomized by the digital scrambling [9]. The “true” value μ is taken to be the average over the $r = 256$ estimates based on the datasets when estimating RMS errors for “bQMC” and “hQMC”, while for “std MC” and “std QMC”, we use the standard deviation as an estimate of the RMS error since they are unbiased.

Figure 2 presents the results for estimating the expectation of the continuous integrand (26) with datasets from the four distributions. Generally speaking, the proposed QMC sampling method based on recursive bisection has an accuracy which is comparable to the standard QMC method until it reaches a plateau which corresponds to the inherent error, i.e., the original sampling error. In particular, with the unit ball surface distribution, the QMC sampling method converges faster than the standard QMC method before it reaches the plateau. The mapping via a Hilbert curve gives similar results.

Figure 3 presents the results for the discontinuous integrand (27). Again the two QMC sampling methods result in a comparable accuracy to the standard QMC, with the Hilbert curve mapping scheme showing small advantage over the recursive bisection scheme.

5 Conclusions

The proposed QMC sampling technique provides a novel way in which to improve the accuracy of estimates obtained using empirical datasets, such as those produced by MCMC methods. Error analysis for the $d = 1$ case is provided, and the numerical

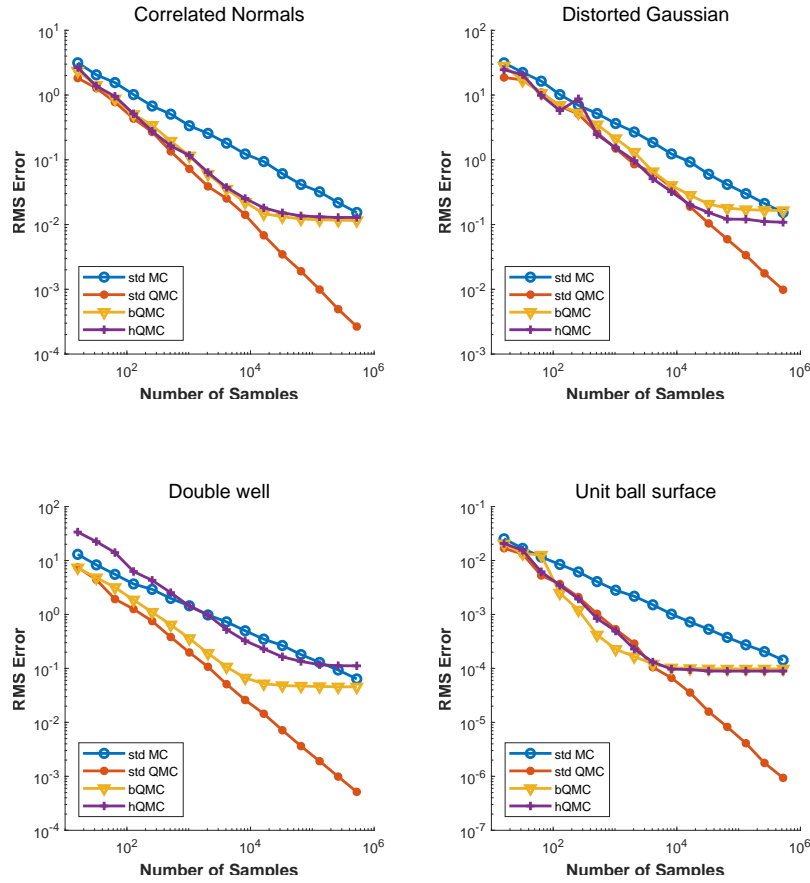


Fig. 2 Results for continuous test function.

results for the 2-dimensional and 3-dimensional cases are presented. These show that when applied to datasets coming from known distributions the effectiveness is very similar to the standard QMC method applied to those distributions. Furthermore, the technique is very easy to apply. In addition, the Hilbert space-filling curve can also be made use of for mapping QMC points to the dataset. Theoretical and empirical error analyses are also provided for this QMC sampling method based on the Hilbert curve. Numerical comparison results show that these two QMC sampling methods demonstrate similar performance, both are comparable to the standard QMC method. Supporting theory for the proposed recursive bisection based QMC sampling method will need to be the subject of future research.

Acknowledgements FX would like to acknowledge support from China Scholarship Council for her visit to Oxford University which led to this research. FX was also supported by the National Science Foundation of China (No. 71471100) and the National Key R&D Program of China (No.

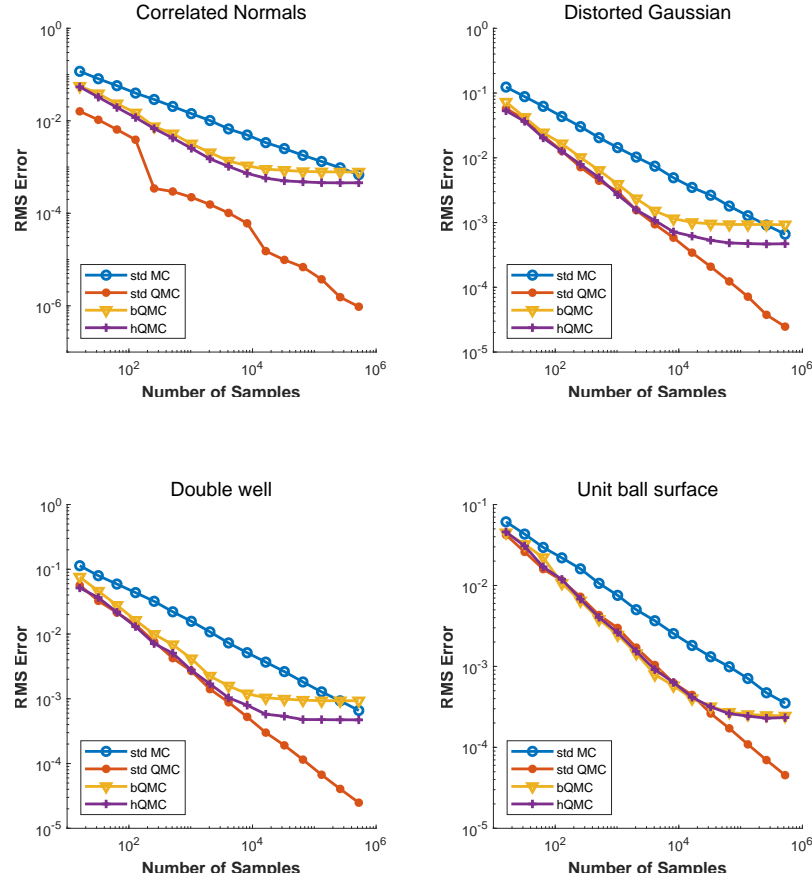


Fig. 3 Results for discontinuous test function.

2016QY02D0301). MBG would like to acknowledge the support of SAMSI (the Statistical and Applied Mathematical Sciences Institute under National Science Foundation Grant DMS-1127914) which hosted the workshop at which he first presented this work, and to thank various participants for their very helpful feedback. ZH was supported by the National Science Foundation of China (No. 71601189) and the Research Funds of South China University of Technology (No. D6191160).

Appendix

Lemma 1. *If Y_k , $k = 1, \dots, N$, are the order statistics of N iid uniformly distributed random variables over $[0, 1]$, then for any $1 \leq i < j \leq N$, we have*

$$\mathbb{E}[Y_j - Y_i] = \frac{j-i}{N+1},$$

and

$$\mathbb{E}[(Y_j - Y_i)^2] = \frac{(j-i)^2 + (j-i)}{(N+1)(N+2)}.$$

Proof. Since $Y_k \sim \text{Beta}(k, N+1-k)$ for any $1 \leq k \leq N$, we have

$$\begin{aligned}\mathbb{E}[Y_k] &= \frac{k}{N+1}, \\ \mathbb{E}[Y_k^2] &= \frac{k(k+1)}{(N+1)(N+2)}.\end{aligned}$$

By [7], we have

$$\mathbb{E}[Y_i Y_j] = \frac{i(j+1)}{(N+1)(N+2)}.$$

We therefore obtain

$$\begin{aligned}\mathbb{E}[(Y_j - Y_i)^2] &= \mathbb{E}[Y_i^2] + \mathbb{E}[Y_j^2] - 2\mathbb{E}[Y_i Y_j] \\ &= \frac{i(i+1) + j(j+1)}{(N+1)(N+2)} - \frac{2i(j+1)}{(N+1)(N+2)} \\ &= \frac{(j-i)^2 + (j-i)}{(N+1)(N+2)}.\end{aligned}$$

□

References

1. Ades, A., Lu, G., Claxton, K.: Expected value of sample information calculations in medical decision modeling. *Medical Decision Making* **24**(2), 207–227 (2004)
2. Barnard, S., Simon, H.: Fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems. *Concurrency and Computation: Practice and Experience* **6**(2), 101–117 (1994)
3. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica* **22**, 133–288 (2013)
4. Gerber, M., Chopin, N.: Sequential quasi Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(3), 509–579 (2015)
5. He, Z., Owen, A.B.: Extensible grids: uniform sampling on a space filling curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**(4), 917–931 (2016)
6. Hendrickson, B., Leland, R.: A multi-level algorithm for partitioning graphs. *Supercomputing* **95**(28) (1995)
7. Moghadam, S., Pazira, H.: The relations among the order statistics of uniform distribution. *Trends in Applied Sciences Research* **6**(7), 719–723 (2011)
8. Niederreiter, H.: *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia (1992)
9. Owen, A.B.: Scrambling Sobol and Niederreiter-Xing points. *Journal of Complexity* **14**(4), 466–489 (1998)
10. Sobol', I.M.: On the distribution of points in a cube and the approximate evaluation of integrals. *Comput. Math. Math. Phys.* **7**(4), 86–112 (1967)