Randomized Quasi-Monte Carlo

Art B. Owen
Stanford University
These slides are from a series of four lectures given at the Johann Radon Institute for Computational and Applied Mathematics (RICAM) held on March 24 and March 25 2021.

It was an honor to be asked to present on quasi-Monte Carlo (QMC) sampling in Austria, from where so much of QMC comes and has come. The talks were virtual; I would have otherwise made sure to get some Linzertorte. That will have to wait.

1. Quasi-Monte Carlo
2. Randomized Quasi-Monte Carlo
3. QMC Beyond the Cube
4. QMC and Variable Importance

A small number of corrections have been made since then.
Quasi-Monte Carlo

Estimate $\mu = \int_{[0,1]^d} f(x) \, dx$ by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

getting

$$|\hat{\mu} - \mu| \leq D_n^*(x_1, \ldots, x_n) \times V_{HK}(f)$$

with

$$D_n^*(x_1, \ldots, x_n) = O\left(\frac{\log(n)^{d-1}}{n}\right) = o(n^{-1+\epsilon}), \text{ any } \epsilon > 0$$

Three problems

1) How to estimate $|\hat{\mu} - \mu|$ for given $f$ and $n$?

2) What about all those logs?

3) What if $V_{HK}(f) = \infty$?
I will leave out most of randomized lattice methods focussing on scrambling of digital constructions

For lattices see works by

Hickernell, Joe, Kritzer, Kuo, L’Ecuyer, Lemieux, Nuyens, Sloan, Tuffin, Ulrich

and many others
Problem 1

Estimating $|\hat{\mu} - \mu|$
QMC is deterministic

Repeat $f(x_1), \ldots, f(x_n)$ get same $\hat{\mu}$

$|\hat{\mu}_n - \hat{\mu}/2|$ might describe accuracy of $\hat{\mu}_n/2$

Or it could be too small

**Randomized QMC (RQMC)**

Inject some randomness into $x_i$

Keeping them at low discrepancy

E.G. $\mathbb{P}\left(D_n^* < B \frac{(\log(n))^d}{n}\right) = 1$

Then do independent repeats
RQMC

1) Make \( x_i \sim U[0, 1]^d \) individually,

2) keeping \( D_n^*(x_1, \ldots, x_n) = O(n^{-1+\epsilon}) \) collectively.

Then

\[
\mathbb{E}(f(x_i)) = \int_{[0,1]^d} f(x) \, dx = \mu
\]

\[
\mathbb{E}(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(f(x_i)) = \mu
\]

\[
|\hat{\mu} - \mu| \leq D_n^*(x_1, \ldots, x_n)V_{HK}(f)
\]

Ensure that

\[
D_n^* \leq Cn^{-1+\epsilon} \quad \text{with probability 1}
\]

Therefore

\[
\mathbb{E}((\hat{\mu} - \mu)^2) = O(n^{-2+2\epsilon})
\]
\( R \) independent replicates

\[
\hat{\mu} = \frac{1}{R} \sum_{r=1}^{R} \hat{\mu}_r
\]

\[
\overline{\text{Var}}(\hat{\mu}) = \frac{1}{R(R-1)} \sum_{r=1}^{R} (\hat{\mu}_r - \hat{\mu})^2
\]

If \( V_{\text{HK}}(f) < \infty \) then

\[
\mathbb{E}((\hat{\mu} - \mu)^2) = O\left(\frac{n^{-2+\epsilon}}{R}\right)
\]

Random shift Cranley & Patterson (1976)
Linear scramble Matousek (1998)
Survey in L'Ecuyer & Lemieux (2005)
Rotation modulo 1

Cranley–Patterson rotation

Shift the points by $\mathbf{u} \sim \mathbf{U}[0, 1)^s$ with wraparound:

$$\mathbf{x}_i \rightarrow \mathbf{x}_i + \mathbf{u} \pmod{1}.$$

Shown for a lattice; can work with nets.

At least it removes $x_1 = 0$:

$$P(x_i = 0) = 0.$$
Digit scrambling

1) Chop space into $b$ slabs. Shuffle.
2) Repeat within each of $b$ slabs.
3) Then within $b^2$ sub-slabs.
4) Ad infinitum $b^3, b^4, \ldots$
5) And the same for all $s$ coordinates.

Each $x_i \sim \mathbb{U}[0, 1)^s$ and
\[
P((x_1, \ldots, x_n) \text{ are a net}) = 1 \quad O(1995)
\]
So Niederreiter (1987,1992) still apply

Cheaper scrambles: digital shift and random linear.
Example scrambles

Two components of the first 530 points of a Faure \((0, 53)\)-net in base 53.

Randomized Faure points

The digital shift is much like a Cranley-Patterson rotation.
It uses just one random \(u\) for all points: \(\tilde{x}_i = x_i \oplus u\). (bitwise)

Nested linear Matousek (1998) and nested uniform O (1995)
have the same \(\text{Var}(\hat{\mu})\).
nested linear takes less memory
Unscrambled Faure

First $n = 11^2 = 121$ points of Faure $(0, 11)$-net in $[0, 1]^{11}$.

Two projections of 121 Faure points

Unscrambled points are very structured.
Scrambling breaks it up.
Confidence intervals

Loh (2003):

central limit theorem for $\hat{\mu}$ when $t = 0$

\[ \hat{\mu} \approx \mathcal{N}(\mu, \text{Var}(\hat{\mu})) \]

with nested uniform sampling

RMSE $O(n^{-1+\epsilon}R^{-1/2}) \implies$ small $R$ popular

The bootstrap $t$ of Efron (1982) gives good confidence intervals for small sample sizes $O(1992)$
Problem 2

Those logs

1) present the ANOVA

2) describe wavelet expansion
   for QMC from O (1997)

3) control the variance
ANOVA Example

\[ f(x_1, x_2) = 6x_1 x_2^2 \]

Write it as \( \sum_{u \subseteq \{1,2\}} f_u(x) \)

\[ f_u(x) = f_u(x_u) \quad x_u = (x_j)_{j \in u} \]

Recurse

\[ f_{\emptyset}(x_1, x_2) = \int_0^1 \int_0^1 6x_1 x_2^2 \, dx_1 \, dx_2 = 1 \]

\[ f_{\{1\}}(x_1, x_2) = \int_0^1 (6x_1 x_2^2 - 1) \, dx_2 = 2x_1 - 1 \]

\[ f_{\{2\}}(x_1, x_2) = \int_0^1 (6x_1 x_2^2 - 1) \, dx_1 = 3x_2^2 - 1 \]

\[ f_{\{1,2\}}(x_1, x_2) = 6x_1 x_2^2 - 1 - (2x_1 - 1) - (3x_2^2 - 1) \]
\[ = 6x_1 x_2^2 - 2x_1 - 3x_2^2 + 1 \]

For \( f_{\{1,2\}} \) integrate over \( x_\emptyset \), i.e., no variables
ANOVA decomposition

\[ f(x) = \sum_{u \subseteq 1:d} f_u(x) \]

where \( f_u \) depends on \( x \) only through \( x_u = (x_j)_{j \in u} \).

Don’t attribute to \( x_u \) what can be explained by \( x_v \) for \( v \subsetneq u \)

Typographical conveniences

\[ -u \equiv 1:d \setminus u \quad -j \equiv -\{j\} \quad dx_u = \prod_{j \in u} dx_j \]

Recursive definition

\[ f_{\emptyset}(x) = \int_{[0,1]^d} f(x) \, dx = \mu \quad \text{(constant)} \]

\[ f_{\{j\}}(x) = \int_{[0,1]^{d-1}} (f(x) - \mu) \, dx_{-j} \quad \text{(function of } x_j) \]

\[ f_u(x) = \int_{[0,1]^{d-|u|}} \left( f(x) - \sum_{v \subsetneq u} f_v(x) \right) \, dx_{-u} \quad \text{(function of } x_u) \]
ANOVA properties

\[ j \in u \implies \int_0^1 f_u(x) \, dx_j = 0 \]
\[ u \neq v \implies \int_{[0,1]^d} f_u(x)f_v(x) \, dx = 0 \]

Variance components

\[ \sigma^2_u = \text{Var}(f_u(x)) = \begin{cases} 
0, & u = \emptyset \\
\int f_u(x)^2 \, dx, & \text{else.}
\end{cases} \]

Decomposition

\[ f(x) = \sum_{u \subseteq 1:d} f_u(x) \]
\[ \sigma^2 = \text{Var}(f(x)) = \sum_{u \subseteq 1:d} \sigma^2_u \]
Randomized QMC

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(x_i) = \frac{1}{n} \sum_{i=1}^{n} \sum_{u \subseteq 1:d} f_u(x_i) = \sum_{u \subseteq 1:d} \hat{\mu}_u
\]

\[
\hat{\mu}_u = \frac{1}{n} \sum_{i=1}^{n} f_u(x_i)
\]

For any RQMC method

\[
\text{Var}(\hat{\mu}) = \sum_{u \subseteq 1:d} \text{Var}(\hat{\mu}_u)
\]

O (2019) QMC notes online

Requires

\(x_{1j}, \ldots, x_{nj}\) uniform and

independent of \(x_{1j'}, \ldots, x_{nj'}\)
Elementary intervals

If $f$ is constant with elementary intervals of volume $b^{t-m}$ then $\hat{\mu} = \mu$.

Same if $f$ is a sum of such functions.
Wavelet expansion

From O (1997) SINUM

\[ \nu_{u,k}(x) \text{ constant within cells of volume } \prod_{j \in u} b^{-k_j} \quad k \in \mathbb{N}^{|u|} \]

\[ f_u(x) = \sum_{k \in \mathbb{N}^{|u|}} \nu_{u,k}(x) \quad \text{(orthogonal)} \]

Finding

\[ \text{Var}(\hat{\mu}) = \sum_{u \subseteq 1:d} \sum_{k \in \mathbb{N}^{|u|}} \text{Var}(\hat{\mu}_{u,k}) \]

\[ \hat{\mu}_{u,k} = \frac{1}{n} \sum_{i=1}^{n} \nu_{u,k}(x_i) \]

\[ \sum_{j \in u} k_j \leq m - t \implies \text{Var}(\hat{\mu}_{u,k}) = 0 \]

One can use Walsh functions for digital nets

Gain coefficients

\[
\text{Var}_{\text{RQMC}}(\hat{\mu}) = \frac{1}{n} \sum_u \sum_k \Gamma_{u,k} \sigma_{u,k}^2 \\
\text{Var}_{\text{MC}}(\hat{\mu}) = \frac{1}{n} \sum_u \sum_k \sigma_{u,k}^2
\]

For \( t = 0 \)

\[
\Gamma_{u,k} \leq \Gamma = \left( \frac{b}{b-1} \right)^{d-1} \leq \exp(1) \approx 2.718
\]

NB: When \( t = 0 \) we must have \( b \geq d \)

For \( t > 0 \)

\[
\Gamma_{u,k} \leq \Gamma = b^t \left( \frac{b+1}{b-1} \right)^s
\]

O (1998)

Extends to first \( \lambda b^m \) points of a \((t, s)\)-sequence in base \( b \) \( 1 \leq \lambda < b \)

Improved bounds in Niederreiter & Pirsic (2001)
Those logs

Let $f(x)$ have variance $\sigma^2$

$$\text{Var}(\hat{\mu}) \leq \frac{\Gamma}{n} \sigma^2$$

For finite $n$

$$\text{RMSE} \leq \sigma \Gamma^{1/2} n^{-1/2}$$

Without powers of $\log(n)$

Later

‘sMOOTHNESS’ yields $\text{RMSE} = O(n^{-3/2} \log(n)^{(d-1)/2})$

Powers of $\log(n)$ can reappear only when they are negligible
Problem 3

What if $V_{HK}(f) = \infty$?
Converse

Riemann integrable $f$ and $D^*_n(x_1, \ldots, x_n) \to 0$

$\implies \hat{\mu}_n \to \mu$

Less well known

$|\hat{\mu}_n - \mu| \to 0$ whenever $D^*_n \to 0$

$\implies f$ Riemann integrable \hspace{1cm} \text{Niederreiter (1977)}$

$f$ not BVHK

1) integrable singularities


2) ‘most’ $f(x) = 1\{x \in A\}$ \hspace{0.5cm} $A \subset [0, 1]^d$

He & Wang (2015)

3) $f$ with ‘kinks and jumps’

Sloan, Kuo, Griebel, Griewank
Gains $\Gamma_{u,k} \to 0$

$$\text{Var}(\hat{\mu}) = \frac{1}{n} \sum_{u} \sum_{k} \Gamma_{u,k} \sigma_{u,k}^2$$

$$\lim_{n=b^m \to 0} \Gamma_{u,k} = 0$$

If $f \in L^2[0, 1]^d$

$$\text{RMSE}(\hat{\mu}) = \text{Var}(\hat{\mu}) = o\left(\frac{1}{n}\right)$$

This covers $f(x) = 1_{x \in A}$ for measurable $A$

Square integrable singularities / kinks / jumps
Laws of large numbers

From O & Rudolf (2020)

\[ f \in L^{1+\delta}[0, 1]^d \quad \text{some } \delta > 0 \]

Weak law

\[
\lim_{n=b^m \to \infty} \mathbb{P}(|\hat{\mu}_n - \mu| > \epsilon) = 0
\]

Strong law

\[
\mathbb{P}\left(\limsup_{n \to \infty} |\hat{\mu}_n - \mu| > \epsilon\right) = 0, \quad \text{i.e.}
\]

\[
\mathbb{P}\left(\lim_{n \to \infty} \hat{\mu}_n = \mu\right) = 1
\]

Needed by Balandat et al. (2020) for Bayesian optimization in NeurIPS

First prove for \( f \in L^2 \) adapting Etemadi’s (1981) SLLN
then use Riesz-Thorin interpolation
Three problems

We can estimate the error · · · more work needed

For $f \in L^2$, RQMC cannot be powers of $\log n$ worse than MC

Don’t need $f$ to be BVHK
Smoothness

\[ \partial^u f(x) = \prod_{j \in u} \frac{\partial}{\partial x_j} f(x) \] is continuous for all \( u \subseteq 1:d \)

Consequence

\[ \text{Var}(\nu_{u,k}) \text{ decays quickly with } \sum_{j \in u} k_j \]

\[ \text{Var}(\hat{\mu}) = O\left(n^{-3} \log(n)^{d-1}\right) \]

\[ \text{Var}(\hat{\mu}) \leq \overline{\Gamma} \sigma^2 n^{-1} \] still holds

O (1997, 2008)
Scrambled net properties

Using $\sigma^2 = \int (f(x) - \mu)^2 \, dx$

<table>
<thead>
<tr>
<th>If</th>
<th>Then</th>
<th>N.B.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f \in L^{1+\delta}$</td>
<td>$\mathbb{P}(\lim_{n \to \infty} \hat{\mu}_n = \mu) = 1$</td>
<td>O &amp; Rudolf (2020)</td>
</tr>
<tr>
<td>$f \in L^2$</td>
<td>$\text{Var}(\hat{\mu}) = o(1/n)$</td>
<td>even if $V_{HK}(f) = \infty$</td>
</tr>
<tr>
<td>$f \in L^2$</td>
<td>$\text{Var}(\hat{\mu}) \leq \bar{\Gamma} \sigma^2/n$</td>
<td>if $t = 0$, $\bar{\Gamma} \leq e \approx 2.718$</td>
</tr>
<tr>
<td>all $\partial^u f \in L^2$</td>
<td>$\text{Var}(\hat{\mu}) = O(\log(n)^{d-1}/n^3)$</td>
<td>O (1997, 2008)</td>
</tr>
</tbody>
</table>

Geometrically

Scrambling Faure breaks up the diagonal striping of the nets.

Scrambling Sobol’ points moves the full / empty blocks around.
Scrambling vs shifting

Consider $n = 2^m$ points in $[0, 1)$.

**QMC**

van der Corput points

shift all points by $U \sim U(0, 1)$ with wraparound.

Get a stratified sample, independent $x_i \sim U[(i-1)/n, i/n)$

Random errors cancel yielding an $O(n^{-1/2})$ improvement.
New software

Three python language projects were mentioned at MCQMC 2020

- QMCPy Fred Hickernell++
- PyTorch Max Balandat++
- Scipy Pamphile Roy++

Especially scrambled Sobol'
Sobol’

First 8 Sobol’ points in $[0, 1]^4$

\[
\begin{array}{cccc}
0.000 & 0.000 & 0.000 & 0.000 \\
0.500 & 0.500 & 0.500 & 0.500 \\
0.250 & 0.750 & 0.750 & 0.750 \\
0.750 & 0.250 & 0.250 & 0.250 \\
0.125 & 0.625 & 0.375 & 0.125 \\
0.625 & 0.125 & 0.875 & 0.625 \\
0.375 & 0.375 & 0.625 & 0.875 \\
0.875 & 0.875 & 0.125 & 0.375 \\
\end{array}
\]

Many implementations did not like the leading $0 = (0, 0, 0, 0, 0)$

They dropped it and used $x_2, \ldots, x_{2^m+1}$

starting at $(1/2, 1/2, 1/2, 1/2)$

This is not a net
Dropping initial point

17 Sobol' points

17 scrambled Sobol' points

Can worsen the *rate* of convergence
Just one point

Can make a big difference:

- includes $x_1$
- skips $x_1$

Easy function: $f(x) = \sum_{j=1}^{5} (\exp(x_j) - e + 1)$

Reference lines $O(n^{-1})$ and $O(n^{-3/2})$
Reason

\[|\hat{\mu}_n - \mu| \approx n^{-3/2}\]

Changing one point is \(O(1/n)\) change larger than the error

Sobol’ points out that

\[|\hat{\mu}_n - \mu| \leq \frac{C}{n} \quad \text{and} \quad |\hat{\mu}_{n+1} - \mu| \leq \frac{C}{n + 1}\]

\[\implies \quad |f(\mathbf{x}_{n+1}) - \mu| = O\left(\frac{1}{n}\right)\]

Consequences

1) Don’t use \(n = 10^k\) for a net in base 2

2) Don’t burn in / warm up RQMC

3) Don’t take every \(k\)’th RQMC

See O (2020) about “Dropping the first Sobol’ point”
Wing weight function

Surjanovic & Bingham (2013)

\[ 0.036S_w^{0.758}W_{fw}^{0.0035} \left( \frac{A}{\cos^2(\Lambda)} \right)^{0.6} q^{0.006} \lambda^{0.04} \left( \frac{100t_c}{\cos(\Lambda)} \right)^{-0.3} (N_x W_{dg})^{0.49} + S_w W_p \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Range</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_w )</td>
<td>[150, 200]</td>
<td>wing area (ft(^2))</td>
</tr>
<tr>
<td>( W_{fw} )</td>
<td>[220, 300]</td>
<td>weight of fuel in the wing (lb)</td>
</tr>
<tr>
<td>( A )</td>
<td>[6, 10]</td>
<td>aspect ratio</td>
</tr>
<tr>
<td>( \Lambda )</td>
<td>[−10, 10]</td>
<td>quarter-chord sweep (degrees)</td>
</tr>
<tr>
<td>( q )</td>
<td>[16, 45]</td>
<td>dynamic pressure at cruise (lb/ft(^2))</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>[0.5, 1]</td>
<td>taper ratio</td>
</tr>
<tr>
<td>( t_c )</td>
<td>[0.08, 0.18]</td>
<td>aerofoil thickness to chord ratio</td>
</tr>
<tr>
<td>( N_z )</td>
<td>[2.5, 6]</td>
<td>ultimate load factor</td>
</tr>
<tr>
<td>( W_{dg} )</td>
<td>[1700, 2500]</td>
<td>flight design gross weight (lb)</td>
</tr>
<tr>
<td>( W_p )</td>
<td>[0.025, 0.08]</td>
<td>paint weight (lb/ft(^2))</td>
</tr>
</tbody>
</table>
Wing weight function

This function on \([a, b] \subset \mathbb{R}^{10}\) has low effective dimension.

We don’t normally want the average weight of a randomly made wing.

This function has physical origin.
Accuracy from RQMC

Coordinate projection of a \((t, m, d)\)-net

\[ x_{1,u}, x_{2,u}, \ldots, x_{n,u} \in [0, 1]|u| \quad u \subseteq 1:d \]

It is a \((t_u, m, |u|)\)-net with \(t_u \leq t\)

For Sobol’ and \(|u| = 1\) we get \(t_u = 0\)


Projection-based bound

\[
\text{Var}(\hat{\mu}) = \sum_{u \subseteq 1:d} \text{Var}(\hat{\mu}_u) \leq \frac{1}{n} \sum_{u \subseteq 1:d} \bar{\Gamma}_{|u|,m} \sigma_u^2 \\
\leq \frac{1}{n} \sum_{u \subseteq 1:d} b^{t_u} \left( \frac{b + 1}{b - 1} \right) |u| \sigma_u^2 \\
\leq \frac{1}{n} \sum_{u \subseteq 1:d} 2^{t_u} 3|u| \sigma_u^2 \quad \text{for } b = 2
\]

Favorable cases dominated by \(\sigma_u^2\) for small \(|u|\)
Rotated lattices

Cannot get $\text{Var}(\hat{\mu}) \leq \Gamma \sigma^2 / n$ for all $f \in L^2$

$\text{Var}(\hat{\mu}) = \sigma^2$ possible (for adversarial $f$)

Random shifts do not reduce error by $\approx n^{1/2}$

See however Kritzer, Kuo, Nuyens, Ullrich (2019) who use random $n$

No apparent central limit theorem

L’Ecuyer

No strong law of large numbers

(yet)
Challenges

1)

$$\mu = \int_{\Omega} f(\mathbf{x}) \, d\mathbf{x} \quad \Omega \neq [0, 1]^d$$

E.g. triangles, balls, spheres, simplices, Markov chains

2)

$$\sup_{\mathbf{x}} |f(\mathbf{x})| = \infty$$

Cannot easily get $O(n^{-3/2+\epsilon})$

Work in progress with S. Liu
Thanks

- Johannes Kepler Universität Linz
- RICAM: Johann·Radon·Institute for Computational and Applied Mathematics
- MCQMC series & Harald Niederreiter
- U.S. NSF: grants up to and including IIS-1837931
- Invitation: Peter Kritzer, Gerhard Larcher, Lucia Del Chicca
- Introductions: Peter Kritzer, Gerhard Larcher, Gunther Leobacher
- Organization: Melanie Traxler

Especially Daniel Rudolf