Quasi-Monte Carlo

Art B. Owen

Stanford University

A tutorial reflecting personal views on what is important in QMC.

MCQMC 2016 will be at Stanford, August 14-19

mcqmc2016.stanford.edu

These are the slides I presented at MCMSki 5 in Lenzerheide Switzerland on Thursday January 8. I have added a few interstitial slides like this one after the fact. I correct some typos too.

It was my good fortune that the tutorial I gave was opposite the Imposteriors rehearsal in the other room. That gave the tutorial a near-plenary audience. While I regret not reaching the Imposteriors as well, I count myself lucky for having such good attendance.

What I present is a statistician's view of QMC, and so my emphases are different in some ways from the customary presentation of QMC.

Software Matlab has an implementation of scrambled Sobol' and Halton points. Start with the documentation for `sobolset` and `haltonset`. They support skipping initial points and leaping (taking every k'th point for $k > 1$). I know of no reason to ever use leaping. Skipping over b^M values might be helpful to get non-overlapping QMC sets of size b^m for $m \leqslant M$. Skipping can also improve Halton points, but then again, randomizing without skipping should be as good or better.

Outline

1) What QMC is (stratification on steroids)
2) Why it works (discrepancy, variation and Koksma-Hlawka)
3) Digital constructions: van der Corput, Halton and nets
4) Lattices
5) Randomized QMC
6) Curse of dimension, tractability, weighted spaces
   (room for Bayesian thinking here)
7) QMC $\cap$ Particles
8) QMC $\cap$ MCMC

But first: orientation

There are landmark papers where Monte Carlo was introduced/adapted for some specific problems. Here is a subset of examples.

- Physics Metropolis et al. (1953)
- Chemistry (reaction equations) Gillespie (1977)
- Financial valuation Boyle (1977)
- Resampling Efron (1979)
- OR (discrete event simulation) Tocher & Owen (1960)
- Bayes (maybe 5 landmarks in early days)
- Nonsmooth optimization Kirkpatrick et al. (1983)
- Computer graphics (path tracing) Kajiya (1988)

Quasi-Monte Carlo

QMC is used in plain quadrature.

Particle transport methods in physics Jerome Spanier++

Financial valuation, some early examples Paskov & Traub 1990s

Graphical rendering Alex Keller++

(They got an Oscar!)

Solving PDEs Frances Kuo, Christoph Schwab++, 2015

Particle methods Chopin & Gerber (2015)

What next

I expect that there are undiscovered landmark applications of QMC, where somebody applies/adapts/extends it for new problems, especially

- machine learning
- Bayes
- uncertainty quantification

The next slide presents a spectrum of methods to suit problems ranging from those so easy that they are considered solved to those so hard that they may never be solved. Things at the top can be embedded in your toaster. Things at the bottom may be nearly impossible or may require a team of PhDs to operate. There are no PhDs inside your toaster.

This is a riff on a comment Stephen Boyd made.

Plain QMC handles problems a bit better behaved than plain MC and delivers better results. To get at the really cutting edge problems tackled by methods at the bottom of the list may require switching 'plain QMC' to something else.

A spectrum

The harder the problem the further down this list we go.

- Everybody just knows it, e.g., 42
- We have closed form expression, e.g., $\sin(2\pi x_1) \times x_2$
- $\exists$ exact deterministic black box, $b(\boldsymbol{x})$
- classic quadratures, e.g., Simpson's or Gauss rule
- plain quasi-Monte Carlo
- plain Monte Carlo
- MCMC, SMC
- ABC, aMCMC, variational MC

  $\vdots$

- Noone will ever know

MC and QMC

We estimate

$$\mu = \int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \text{by} \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i), \quad \boldsymbol{x}_i \in [0,1]^d$$

In plain MC, the $\boldsymbol{x}_i$ are IID $\mathbf{U}[0,1]^d$. In QMC they're 'spread evenly'.

Non uniform

$$\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \text{and} \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\psi(\boldsymbol{x}_i)), \quad \boldsymbol{x}_i \in [0,1]^d$$

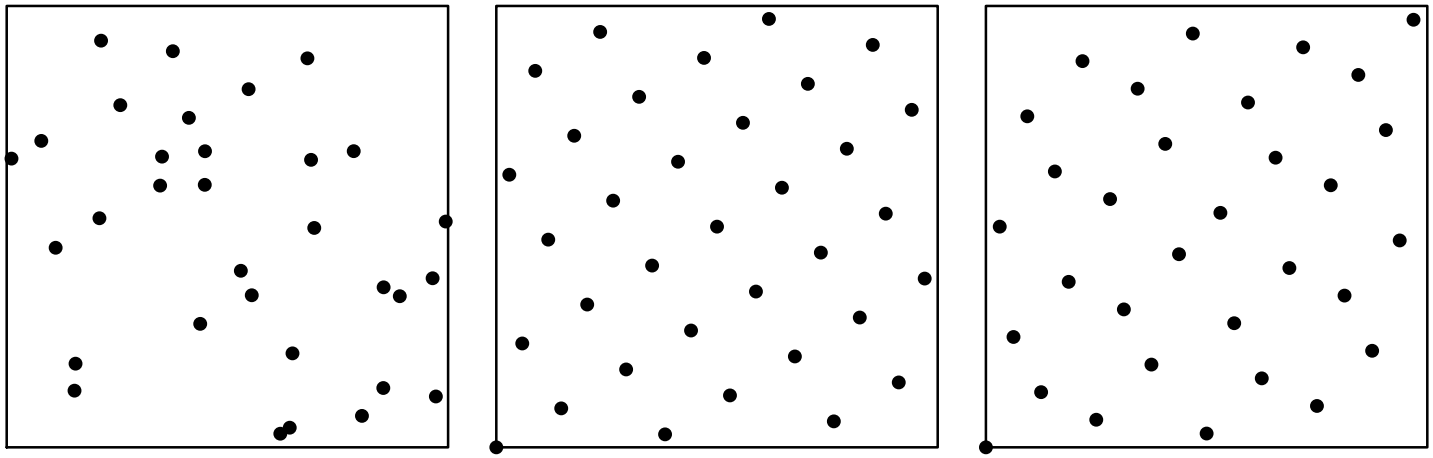
Given suitable ψ

$$\psi(\boldsymbol{x}) \sim p \text{ whenever } \boldsymbol{x} \sim \mathbf{U}[0,1]^d$$

Many methods fit this framework. Acceptance-rejection is awkward.

Illustration

MC and two QMC methods in the unit square



Monte Carlo | Fibonacci lattice | Hammersley sequence

MC points always have clusters and gaps. What is random is where they appear.

QMC points avoid clusters and gaps to the extent that mathematics permits.

Philosophies I don't share

1) Zaremba (1968): The proper justification of the normal practice of Monte Carlo integration must be based not on the randomness of the procedure, which is spurious, but on equidistribution properties of the sets of points at which the integrand values are computed.

I very much like how randomness lets you estimate your errors.

2) Also: it is possible to object to frequentist tools (LLN and CLT) being introduced into Bayesian problems.

If it works, why not use it? This complaint is now very rare.

3) Pseudo-random numbers aren't really random.

Yes, but they're very well tested. We get more trouble from floating point numbers representing reals than we do from pseudo-randomness representing randomness.

Measuring uniformity

We need a way to verify that the points $\boldsymbol{x}_i$ are 'spread out' in $[0,1]^d$.

The most fruitful way is to show that

$$\mathbf{U}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \doteq \mathbf{U}[0,1]^d$$

Discrepancy

A discrepancy is a distance $\|F - \hat{F}_n\|$ between measures $F = \mathbf{U}[0,1]^d$ and $\hat{F}_n = \mathbf{U}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$.

There are many discrepancies.

Technicality

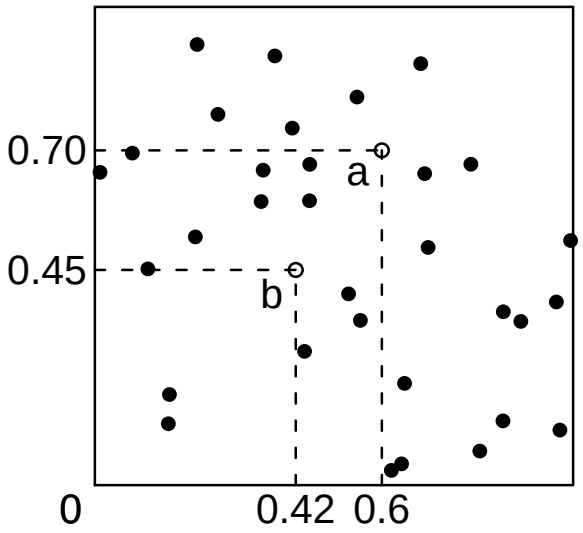
More properly: Distn$(\boldsymbol{x}_i) \doteq \mathbf{U}[0,1]^d$, for $i \sim \mathbf{U}\{1, 2, \ldots, n\}$

There could be ties.

Local discrepancy

Did the box $[0, \boldsymbol{a})$ get it's fair share of points?

Local discrepancy at a, b



$$\delta(\boldsymbol{a}) = \hat{F}_n([0, \boldsymbol{a})) - F([0, \boldsymbol{a})) = \frac{13}{32} - 0.6 \times 0.7 = -0.01375$$

Star discrepancy

$$D_n^* = D_n^*(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sup_{\boldsymbol{a} \in [0,1)^d} |\delta(\boldsymbol{a})|$$

For $d = 1$ this is Kolmogorov-Smirnov.

More discrepancies

$$D_n^* = \sup_{\boldsymbol{a}\in[0,1)^d} \left|\hat{F}_n([0,\boldsymbol{a})) - F([0,\boldsymbol{a}))\right|$$

$$D_n = \sup_{\boldsymbol{a},\boldsymbol{b}\in[0,1)^d} \left|\hat{F}_n([\boldsymbol{a},\boldsymbol{b})) - F([\boldsymbol{a},\boldsymbol{b}))\right|$$

$$D_n^* \leqslant D_n \leqslant 2^d D_n^*$$

L^p discrepancies

$$D_n^{*p} = \left(\int_{[0,1)^d} |\delta(\boldsymbol{a})|^p \,\mathrm{d}\boldsymbol{a}\right)^{1/p}$$

Also

Wrap-around discrepancies Hickernell

Discrepancies over (triangles, rotated rectangles, balls $\cdots$ convex sets $\cdots$).

Beck, Chen, Schmidt

Best results are **only** for axis-aligned hyper-rectangles.

Koksma's inequality

For $d = 1$ $\quad |\hat{\mu} - \mu| \leqslant D_n^*(x_1, \ldots, x_n) \times \int_0^1 |f'(x)| \, \mathrm{d}x$

NB: $\int_0^1 |f'(x)| \, \mathrm{d}x$ is the total variation of f.

Koksma-Hlawka theorem

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) - \int_{[0,1)^d} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right| \leqslant D_n^* \times V_{\mathrm{HK}}(f)$$

V_{HK} is the **total variation** in the sense of Hardy (1905) and Krause (1903)

Puzzler

Is this a 100% confidence interval?

For $d = 1$, Koksma

WLOG $0 \equiv x_0 \leqslant x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n \leqslant x_{n+1} \equiv 1$

Integration and summation by parts

$$\int_0^1 f(x)\,\mathrm{d}x = f(1) - \int_0^1 x f'(x)\,\mathrm{d}x$$

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) = f(1) - \frac{1}{n}\sum_{i=0}^{n} i(f(x_{i+1}) - f(x_i))$$

$$= f(1) - \frac{1}{n}\sum_{i=0}^{n} i \int_{x_i}^{x_{i+1}} f'(x)\,\mathrm{d}x$$

After simplification, for continuous f'

$$\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int_0^1 f(x)\,\mathrm{d}x = -\int_0^1 \delta(x) f'(x)\,\mathrm{d}x$$

Upshot

$$|\hat{\mu} - \mu| \leqslant \|\delta\|_p \times \|f'\|_q \quad 1/p + 1/q = 1$$

Rates of convergence

It is possible to get $D_n^* = \dfrac{\log(n)^{d-1}}{n}$.

Then

$$|\hat{\mu} - \mu| = o(n^{-1+\epsilon}) \quad \text{vs} \quad O_p(n^{-1/2}) \text{ for MC}$$

What about those logs?

$\log(n)^{d-1}/n$ can be very large if d is large.

The s dimensional coordinate projections (i.e., margins) of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ typically have discrepancy $O(\log(n)^{s-1}/n)$

Log factors are not material for functions of low effective dimension (later).

Roth (1954)

$$D_n^* \quad \text{of} \quad o\left(\frac{\log(n)^{(d-1)/2}}{n}\right) \quad \text{is unattainable}$$

Randomization

Can also control the log factors. (later)

Tight and loose bounds

They are not mutually exclusive.

Koksma-Hlawka is tight

$$|\hat{\mu} - \mu| \leqslant (1-\epsilon) D_n^*(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \times V_{\mathrm{HK}}(f) \quad \textbf{fails} \text{ for some } f$$

KH holds as an equality for a worst case function like $f' \doteq \pm\delta$.

Koksma-Hlawka is also very loose

It can greatly over-estimate actual error. Usually δ and f' are dissimilar.

$$\hat{\mu} - \mu = -\langle \delta, f' \rangle$$

Just like Chebychev's inequality

It is also tight and very loose. E.g., $\Pr\bigl(|\mathcal{N}(0,1)| \geqslant 10\bigr) \leqslant 0.01$ is loose.

Yes: $1.5 \times 10^{-23} \leqslant 10^{-2}$

Variation

Multidimensional variation can be a bit more subtle than one dimensional.

$$f(x_1, x_2) = \begin{cases} 1, & x_1 + x_2 \leqslant 1/2 \\ 0, & \text{else} \end{cases}$$

$$V_{\mathrm{HK}}(f) = \infty \quad \text{on } [0,1]^2$$

$$V_{\mathrm{HK}}(f_\epsilon) < \infty, \quad \text{for } \|f - f_\epsilon\|_1 < \epsilon$$

Vitali variation

It is almost $\int_{[0,1]^d} |\frac{\partial^d}{\partial x_1, \ldots, \partial x_d} f(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x}$

Vanishes if f does not depend on x_j for some $j \in 1, \ldots, d$.

Hardy-Krause variation

Sum of Vitali variations on 'upper faces' of $[0,1]^d$ with 0 to $d-1$ components fixed at 1.

QMC-friendly discontinuities

Axis parallel. Xiaoqun Wang++ $V_{\mathrm{HK}} < \infty$.

Effective dimension

Functional ANOVA decomposition Hoeffding (1948), Sobol' (1969)

$$f(\boldsymbol{x}) = \mu + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d) + f_{1,2}(x_1, x_2) + \cdots$$

Each of these sub-f's integrates to zero.

Then

$$\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^{n} f_1(x_{i1}) + \cdots + \frac{1}{n}\sum_{i=1}^{n} f_d(x_{id}) + \frac{1}{n}\sum_{i=1}^{n} f_{1,2}(x_{i1}, x_{i2}) + \cdots$$

If f is dominated by its main effects and low order interactions (Caflisch, Morokoff & O (1997)) then the error is like a lower dimensional QMC error.

With some notational license

$$\begin{aligned} |\hat{\mu} - \mu| &\leqslant \sum_{u \subseteq \{1,\ldots,d\}} D_n^*(\boldsymbol{x}_{1u}, \ldots, \boldsymbol{x}_{nu}) V_{\mathrm{HK}}(f_u) \\ &= \sum_{u \subseteq \{1,\ldots,d\}} O\Big(\frac{\log(n)^{|u|-1}}{n}\Big) V_{\mathrm{HK}}(f_u) \end{aligned}$$

For $d = 1$

Points $x_i = (i - 1/2)/n$ minimize both D_n and D_n^*.

But **where** do we put the $\boldsymbol{n+1}$'st point?

Extensible sequences

Take first n points of $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \ldots$.
Then we can get $D_n^* = O((\log n)^d/n)$.

No known extensible constructions get $O((\log n)^{d-1}/n)$.

van der Corput

i				$\phi_2(i)$
1	1	0.1	1/2	0.5
2	10	0.01	1/4	0.25
3	11	0.11	3/4	0.75
4	100	0.001	1/8	0.125
5	101	0.101	5/8	0.625
6	110	0.011	3/8	0.375
7	111	0.111	7/8	0.875
8	1000	0.0001	1/16	0.0625
9	1001	0.1001	9/16	0.5625

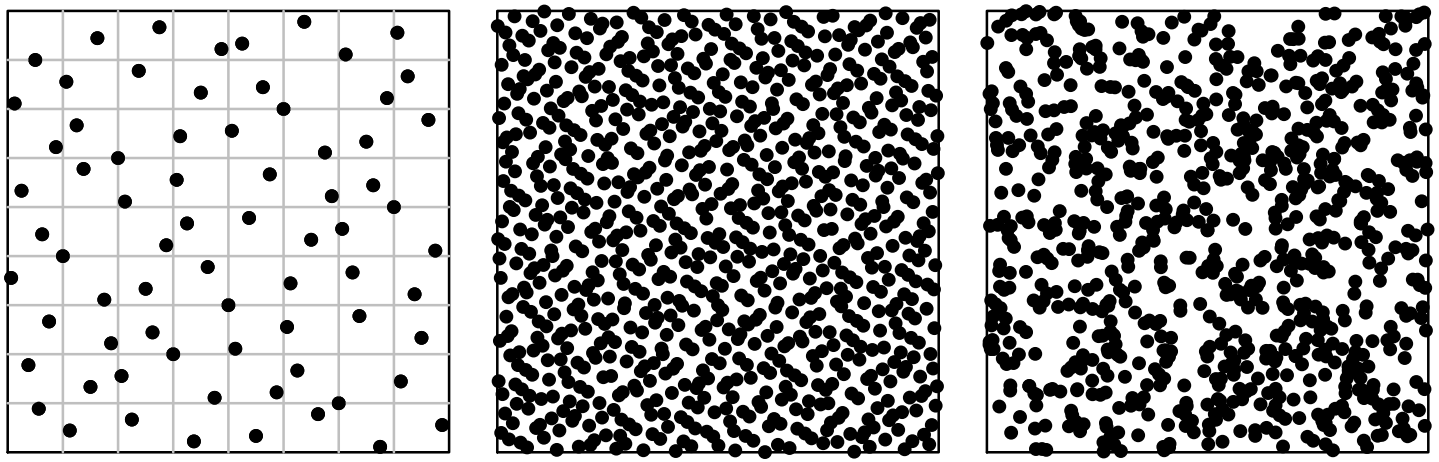
Take $x_i = \phi_2(i)$. Extensible with $D_n^* = O(\log(n)/n)$.

Commonly $x_i = \phi_2(i-1)$ starts at $x_1 = 0$.

Halton sequences

The van der Corput trick works for any base. Use bases $2, 3, 5, 7, \ldots$

Halton sequence in the unit square



72 Halton points | 864 Halton points | 864 random points

Via base b digital expansions

$$i = \sum_{k=0}^{K} b^k a_{ik} \qquad \rightarrow \qquad \phi_b(i) \equiv \sum_{k=0}^{K} b^{-1-k} a_{ik}$$

$$\boldsymbol{x}_i = (\phi_2(i), \phi_3(i), \ldots, \phi_p(i))$$

The Halton sequence is easy to implement. It is a good tool for testing whether QMC will help on your problem. It is also well suited to homework problems.

If you need large prime numbers, then scrambling the digits in the Halton sequence is recommended. Simply using a random permutation of them, as Giray Okten advocates, will bring an improvement.

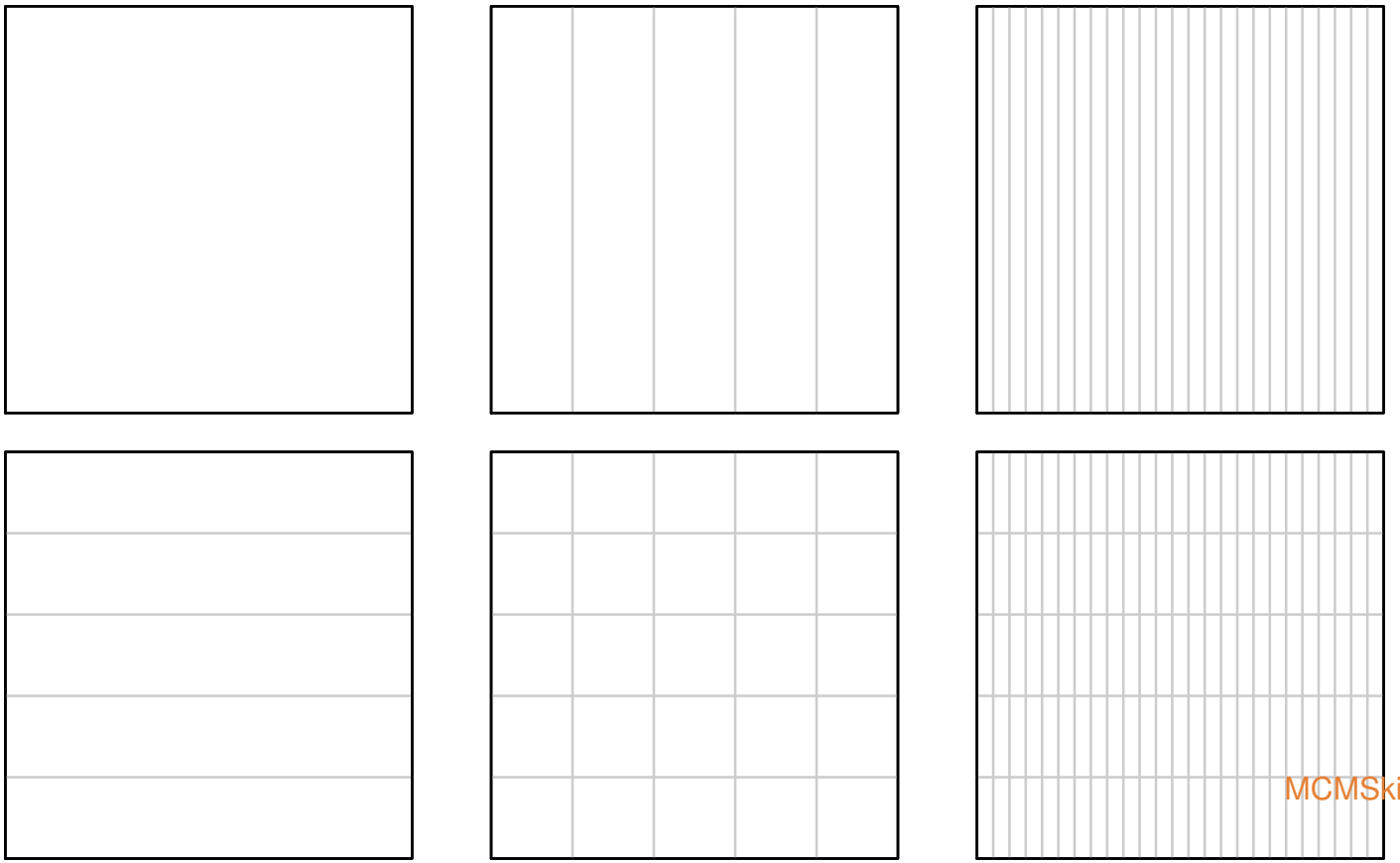
Digital nets

Halton sequences are balanced if n is a multiple of 2^a and 3^b and 5^c . . .

Digital nets use just one base $b \implies$ balance all margins equally.

Elementary intervals

Some elementary intervals in base 5



Digital nets

$$E = \prod_{j=1}^{s} \Big[\frac{a_j}{b^{k_j}}, \frac{a_j + 1}{b^{k_j}} \Big), \quad 0 \leqslant a_j < b^{k_j}$$

$(0, m, s)$-net

$n = b^m$ points in $[0, 1)^s$. If $\mathbf{vol}(E) = 1/n$ then E has one of the n points.

e.g. Faure (1982) points, prime base $b \geqslant s$

(t, m, s)-net

If E deserves b^t points it gets b^t points. Integer $t \geqslant 0$.

e.g. Sobol' (1967) points base 2

Smaller t is better (but a construction might not exist).

minT project

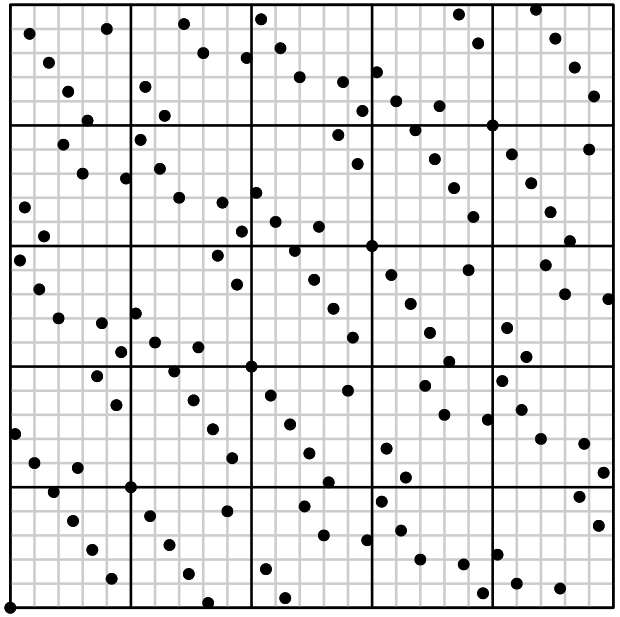
Schürer & Schmid give bounds on t given b, m and s

Monographs

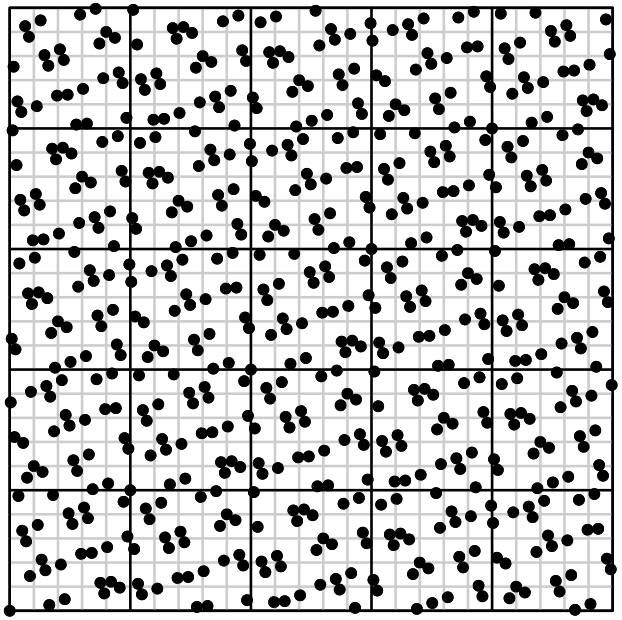
Niederreiter (1992) Dick & Pillichshammer (2010)

Example nets

Two digital nets in base 5



A (0,3,2) net



A (0,4,2) net

The $(0, 4, 2)$-net is a bivariate margin of a $(0, 4, 5)$-net.

The parent net has $5^4 = 625$ points in $[0, 1)^5$.

It balances $43{,}750$ elementary intervals.

Think of $43{,}750$ control variates for 625 obs.

We should remove that diagonal striping artifact (later).

Extensible nets

Nets can be extended to larger sample sizes.

It raises D_n^* from $O((\log n)^{s-1}/n)$ to $O((\log n)^s/n)$

(t, s)-sequence in base b

Infinite sequence of (t, m, s)-nets.

$$\underbrace{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{b^m}}\ \underbrace{\boldsymbol{x}_{b^m+1}, \ldots, \boldsymbol{x}_{2b^m}}\ \cdots\ \underbrace{\boldsymbol{x}_{kb^m+1}, \ldots, \boldsymbol{x}_{(k+1)b^m}}\ \cdots$$

Simultaneously for all $t \geqslant m$.

$$\underbrace{\underbrace{(t,m,s)-\text{net}}_{\text{1st}}\ \underbrace{(t,m,s)-\text{net}}_{\text{2nd}}\ \cdots\ \underbrace{(t,m,s)-\text{net}}_{\text{b'th}}\ \cdots}_{(t,m+1,s)-\text{net}}$$

Examples

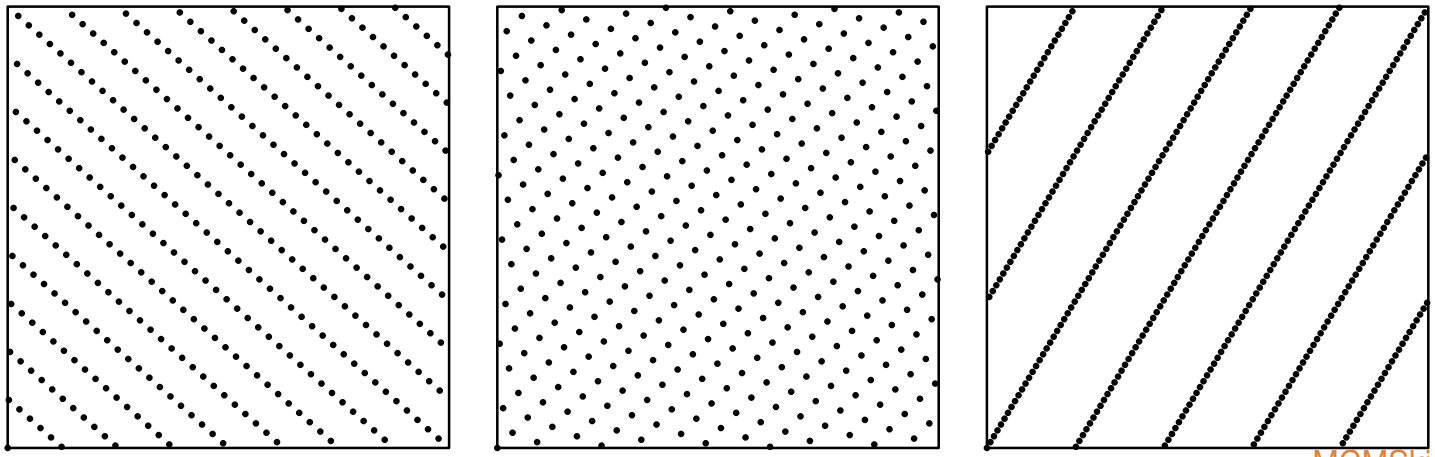
Sobol' $b = 2$ Faure $t = 0$ Niederreiter & Xing $b = 2$ (mostly)

Lattices

$$\boldsymbol{x}_i = \left(\frac{i}{n}, \frac{Z_2 i}{n}, \frac{Z_3 i}{n}, \ldots, \frac{Z_d i}{n}\right) \pmod 1 \quad Z_j \in \mathbb{N}$$
$$\boldsymbol{Z} = (1, Z_2, Z_3, \ldots, Z_d)$$

- the other main family of QMC points
- an extensive literature, e.g., Sloan & Joe also Kuo, Nuyens, Dick, Cools, Hickernell, $\cdots$
- less benefit from randomization

Some lattice rules for n=377



z = (1,41) z = (1,233) z = (1,253)

Lattices ctd

The choice of $\boldsymbol{Z} = (1, Z_2, Z_3, \ldots, Z_d) \in \mathbb{N}^d$ is critical.

We **get to** choose $\boldsymbol{Z}$. $\Longrightarrow$ We can tune/optimize.

We **have to** choose $\boldsymbol{Z}$. $\Longrightarrow$ We must search.

Korobov (1959) lattices

$$\boldsymbol{Z} = (1, Z, Z^2, \ldots, Z^{d-1}), \quad Z \in \mathbb{N}$$

Reduced search space. Minor performance penalty.

Extensibility

Lattices are not extensible. But 'shifted lattices' can be extended.

Hickernell, Hong, L'Ecuyer & Lemieux (2000)

Lattice strategy

Sloan & Joe (1994) approach to lattices.

1) Suppose that f is periodic.

2) Upper bound $|\hat{\mu} - \mu|$ for smooth periodic functions.

3) Find lattice with small upper bound.

4) "Make f periodic."

The last step involves replacing f by periodic $\widetilde{f}$ with $\int \widetilde{f}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$.

I skipped over some of these slides, to just show and explain the dual lattices a few slides hence.

This presentation of lattices is based on the book by Sloan and Joe. There has been considerable progress since then and I'm hopeful that somebody will put it together into a new monograph.

Smooth periodic functions

$$f(\boldsymbol{x}+\boldsymbol{z}) = f(\boldsymbol{x}), \quad \forall \boldsymbol{z} \in \mathbb{Z}^d$$

Fourier

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{h}\in\mathbb{Z}^d} \hat{f}(\boldsymbol{h}) e^{2\pi\sqrt{-1}\boldsymbol{h}^{\mathsf{T}}\boldsymbol{x}} \quad \text{(in mean square)}$$

$$\hat{f}(\boldsymbol{h}) = \int_{[0,1]^d} f(\boldsymbol{x}) e^{-2\pi\sqrt{-1}\boldsymbol{h}^{\mathsf{T}}\boldsymbol{x}}$$

Substitute

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) = \sum_{\boldsymbol{h}\in\mathbb{Z}^d} \hat{f}(\boldsymbol{h}) \left(\frac{1}{n}\sum_{i=1}^{n} e^{2\pi\boldsymbol{h}^{\mathsf{T}}\boldsymbol{x}_i} \right)$$

$$\mu = \int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \sum_{\boldsymbol{h}\in\mathbb{Z}^d} \hat{f}(\boldsymbol{h}) \int_{[0,1]^d} e^{2\pi\boldsymbol{h}^{\mathsf{T}}\boldsymbol{x}}\,\mathrm{d}\boldsymbol{x} = \hat{f}(0)$$

Averaging sinusoids

$$\boldsymbol{x}_i = \frac{i\boldsymbol{Z}}{n} \bmod 1$$

$$\frac{1}{n}\sum_{i=1}^{n} e^{2\pi \boldsymbol{h}^{\mathsf{T}}\boldsymbol{x}_i} = \begin{cases} 1 & \boldsymbol{h}^{\mathsf{T}}\boldsymbol{Z} = 0 \pmod n \\ 0 & \text{else.} \end{cases}$$

$$\hat{\mu} - \mu = \sum_{\substack{\boldsymbol{h}\in\mathbb{Z}^d \\ \boldsymbol{h}^{\mathsf{T}}\boldsymbol{Z}=0 \pmod n}} \hat{f}(\boldsymbol{h}) - \hat{f}(\boldsymbol{0})$$

Dual lattices

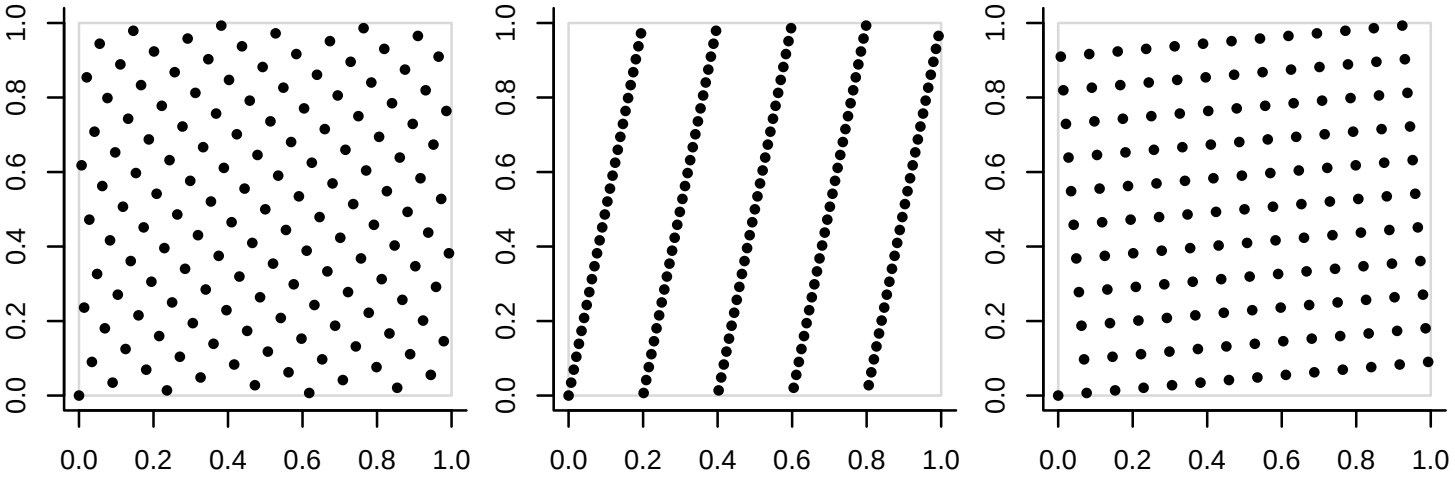
$$L^{\perp} = \{\boldsymbol{h} \in \mathbb{Z}^d \mid \boldsymbol{h}^{\mathsf{T}}\boldsymbol{Z} = 0 \pmod n\}$$

Smooth $f \implies \hat{f}$ decays with $|\hat{f}(\boldsymbol{h})|$.

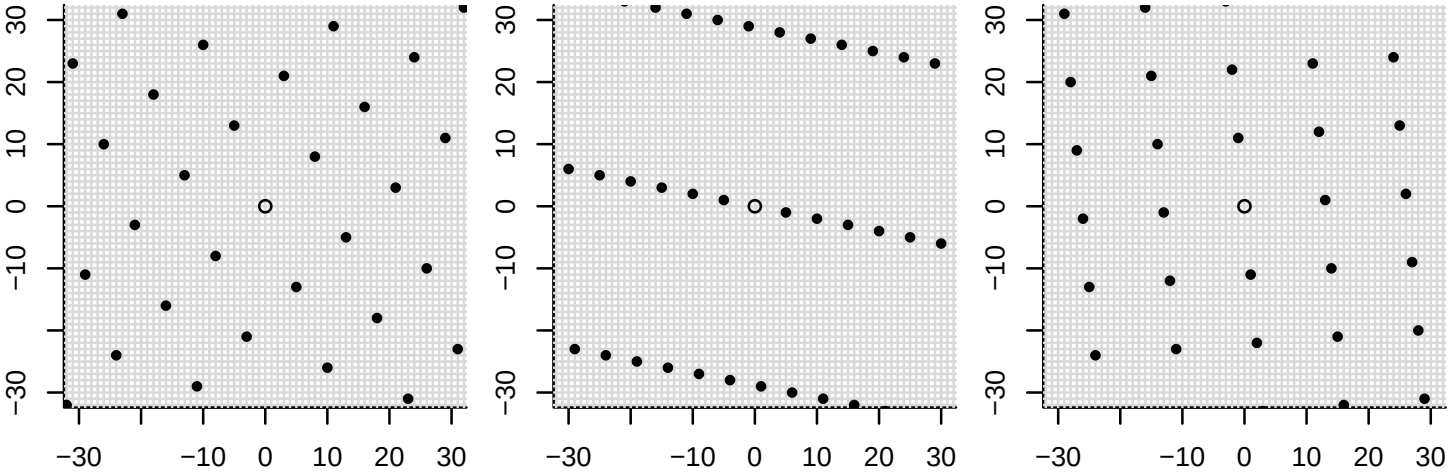
So search for $\boldsymbol{Z}$ with $L^{\perp} \setminus \{\boldsymbol{0}\}$ 'far from origin'.

Dual lattice

Some integration lattices



n=144 z = (1,89) n=144 z = (1,5) n=144 z = (1,131)



and their dual lattices

Periodizing transformation

$$\int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{[0,1]^d} f(\tau(\boldsymbol{x}))J(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

$$\tau = \text{transformation}$$

$$J = \text{Jacobian}$$

Pick τ so that $J(\boldsymbol{x})$ vanishes on the boundary of $[0,1]^d$.

Better: make r derivatives of $f \circ \tau \times J$ vanish on $\partial[0,1]^d$.

Upshot

Good news: get very good convergence rate

Bad news: get strong curse of dimension in the constant.

QMC error estimation

$$|\hat{\mu} - \mu| \leqslant D_n^* \times V_{\mathrm{HK}}(f)$$

Not a 100% confidence interval

- D_n^* is hard to compute
- V_{HK} harder to get than μ
- $V_{\mathrm{HK}} = \infty$ is common, e.g., $f(x_1, x_2) = 1_{x_1+x_2 \leqslant 1}$
- We either get $|\hat{\mu} - \mu| < \infty$ or $|\hat{\mu} - \mu| \leqslant \infty$
  (and maybe we already knew)

Also

Koksma-Hlawka is worst case. It can be very conservative.

Randomized QMC

1) Make $\boldsymbol{x}_i \sim \mathbf{U}[0,1)^d$ individually,

2) keeping $D_n^* = O(n^{-1+\epsilon})$ collectively.

R independent replicates

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

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{R(R-1)}\sum_{r=1}^{R}(\hat{\mu}_r - \hat{\mu})^2$$

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

$$\mathbb{E}((\hat{\mu}-\mu)^2) = O(n^{-2+\epsilon})$$

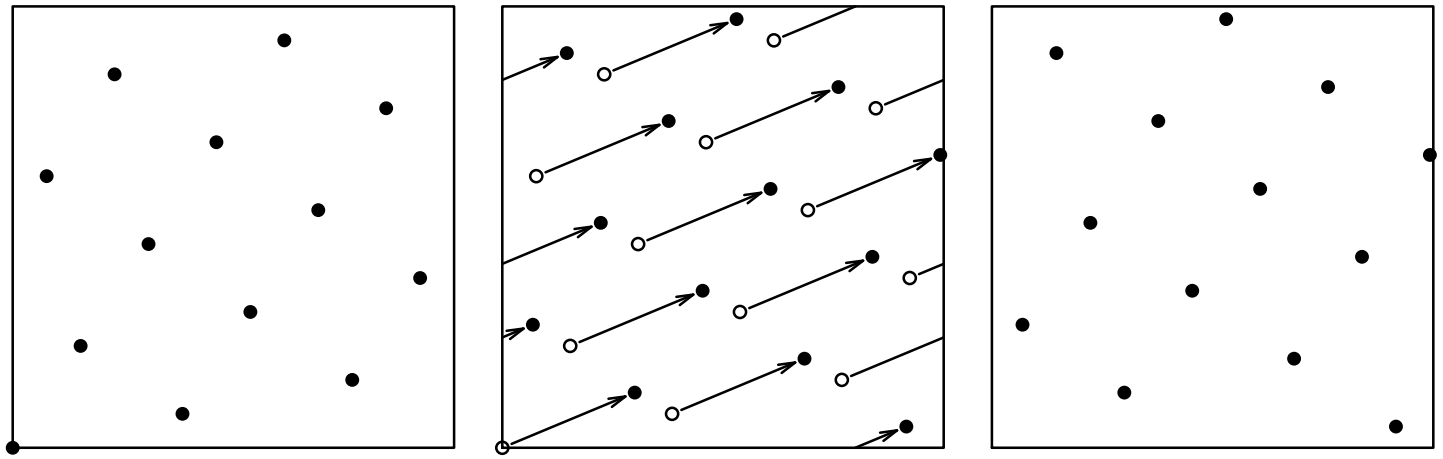
Random shift Cranley & Patterson (1976)

Scrambled nets O (1995,1997,1998)

Survey in L'Ecuyer & Lemieux (2005)

Rotation modulo 1

Cranley–Patterson rotation



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

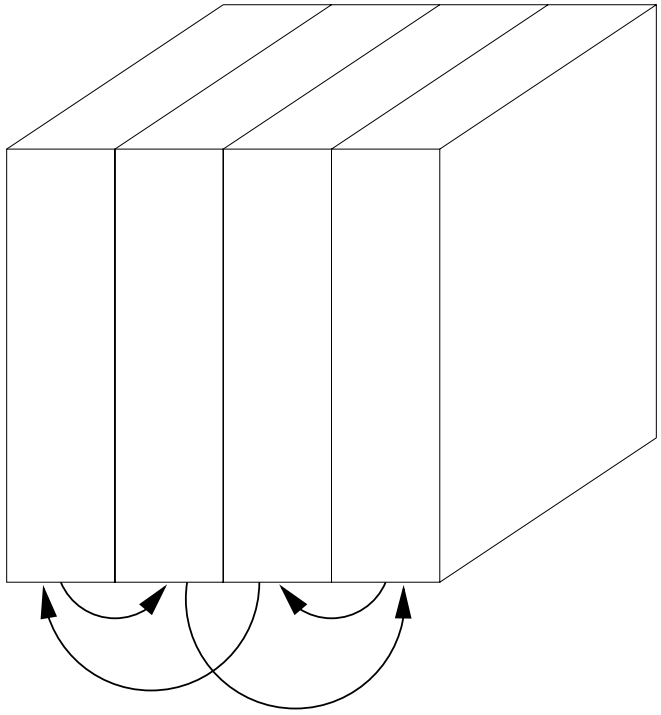
$$\boldsymbol{x}_i \to \boldsymbol{x}_i + \boldsymbol{u} \pmod 1.$$

Commonly used on lattice rules.

Can also be used with nets.

At least it removes $\boldsymbol{x}_1 = 0$.

Digit scrambling



1) Chop the space into b slabs. Shuffle them.
2) Do the same within each of those b slabs.
3) And so on within b^2, b^3, $\cdots$ sub-slabs.
4) And the same for all s coordinates.

This operation yields $\boldsymbol{x}_i \sim \mathbf{U}[0,1)^s$ and preserves the net property. O (1995)

Digit scrambling

Mathematically it is a permutation operation on the base b digits of

$$x_{ij} = \sum_{k=1}^{K} a_{ijk}\, b^{-k} \quad \to \quad \tilde{x}_{ij} = \sum_{k=1}^{K} \tilde{a}_{ijk}\, b^{-k}$$

The mapping $a_{ijk} \to \tilde{a}_{ijk} = \pi_{jk}(a_{ijk})$ preserves the net property.

In the previous 'nested' scramble π_{jk} also depends on $a_{ij1}, \ldots, a_{ij,k-1}$

Simpler scrambles

Random linear permutations $a \to g + h \times a \bmod p$ (prime $p = b$)

$g \sim \mathbf{U}\{0, 1, \ldots, b-1\}$, $h \sim \mathbf{U}\{1, 2, \ldots, b-1\}$

Matousek (1998) has nested linear permutations

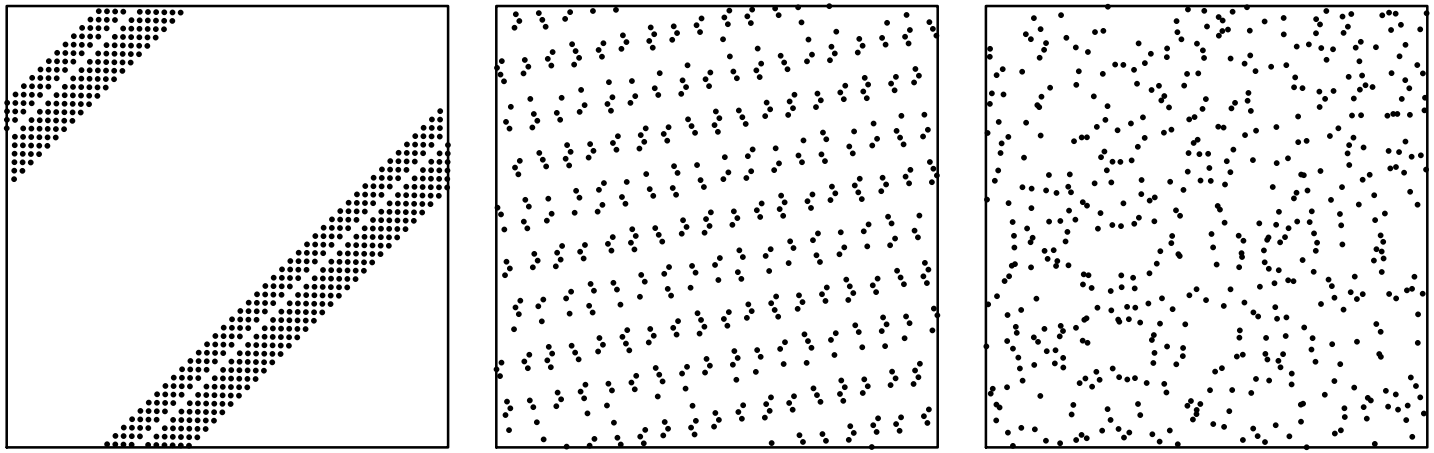
Digital shift $a_{ijk} \to a_{ijk} + g_{ij} \pmod p$

For $b = 2$ $\boldsymbol{x}_i \to \widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i \oplus \boldsymbol{u}$ bitwise XOR

Example scrambles

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

Randomized Faure points



Digital shift | Random linear | Nested uniform

The digital shift is much like a Cranley-Patterson rotation.

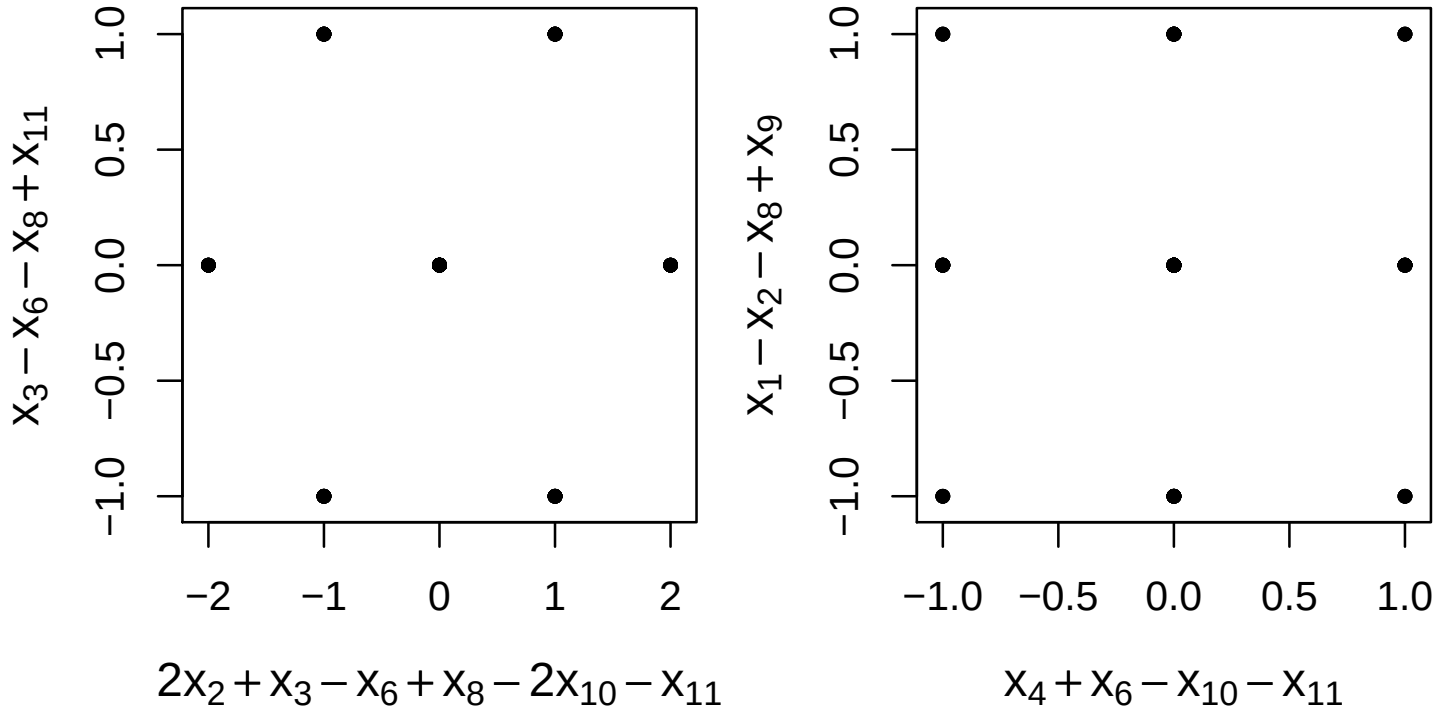
It uses just one random $\boldsymbol{u}$ for all points: $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i \oplus \boldsymbol{u}$.

Random linear Matousek (1998) and nested uniform O (1995) yield the same variance.

Unscrambled Faure

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

Two projections of 121 Faure points



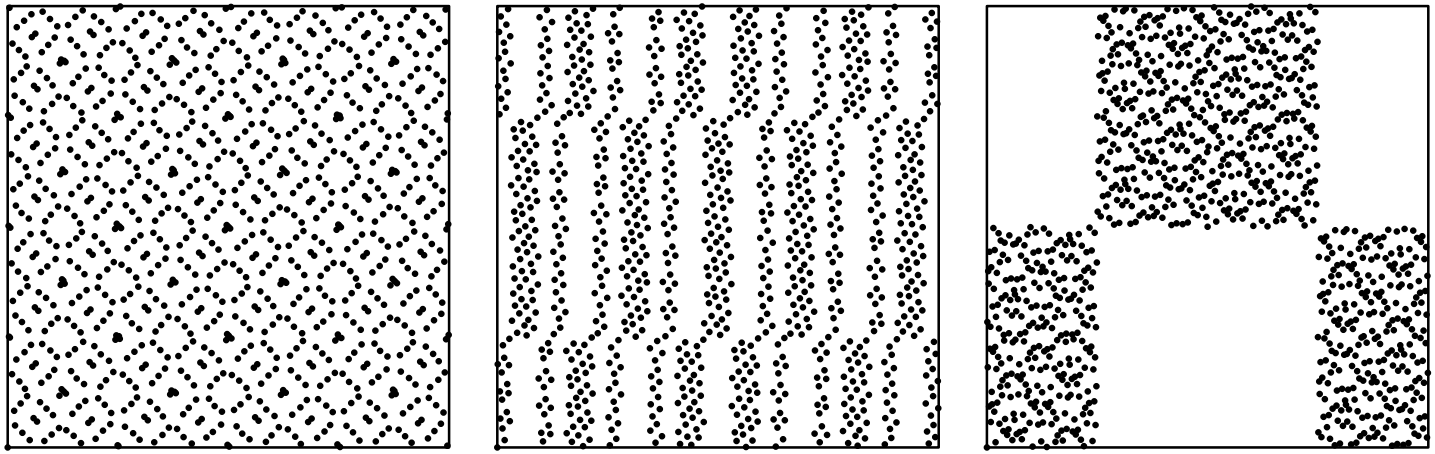
Unscrambled points are very structured.

Sobol' projections

Using code from Bratley & Fox (1988).

Implementations differ based on choices of 'direction numbers'.

Three projections of 1024 Sobol points



At some higher sample size the empty blocks in x_{26} vs x_{31} fill in.

Expect empty blocks in higher order margins to last longer than those in low dimensional margins.

Scrambled net properties

Using $\sigma^2 = \int (f(\boldsymbol{x}) - \mu)^2 \, \mathrm{d}\boldsymbol{x}$

If	Then	N.B.
$f \in L^2$	$\mathrm{Var}(\hat{\mu}) = o(1/n)$	even if $V_{\mathrm{HK}}(f) = \infty$
$f \in L^2$	$\mathrm{Var}(\hat{\mu}) \leqslant \Gamma_{t,b,s} \, \sigma^2/n$	for $t = 0$, $\Gamma \leqslant \exp(1) \doteq 2.718$
$\partial^{1,2,\ldots,s} f \in L^2$	$\mathrm{Var}(\hat{\mu}) = O(\log(n)^{s-1}/n^3)$	O (1997,2008)

$\Gamma < \infty$ rules out $(\log n)^{s-1}$ catastrophe at finite n.

Loh (2003) has a CLT for $t = 0$ (and **fully** scrambled points).

Geometrically

Scrambling breaks up the horizontal striping of the Faure nets.

Scrambling Sobol' points moves the full/empty blocks around.

Improved rate

RMSE is $O(n^{-1/2})$ better than QMC rate (cancellation).

Holds for nested uniform and nested linear scrambles.

I suspect that the CLT will **not** hold for random linear scrambles.

Hickernell and Yue found fourth moment differences in discrepancies for the two different kinds of scrambles.

I have a paper on recycling physical random numbers EJS (2009) where the result always has an asymptotically symmetric distribution which is never Gaussian. That is what lead me to guess (still unverified) that the CLT might not hold for the nested linear scramble of Matousek.

Scrambling vs shifting

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

QMC

van der Corput points $(i-1)/n$ for $i = 1, \ldots, n$.

Shift

Shift all points by $U \sim \mathbf{U}(0, 1)$ with wraparound.

Get one point in each $[(i-1)/n, i/n)$

Scramble

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

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

Higher order nets

Results from Dick, Baldeaux

Start with a net in $2s$ dimensions.

'Interleave' digits of two variables to make a new one

$$0.g_1g_2g_3\cdots \text{ and } 0.h_1h_2h_3\cdots \;\to\; 0.g_1h_2g_2h_2\cdots .$$

Get a 'higher order' net in s dimensions.

Error is $O(n^{-2+\epsilon})$ under increased smoothness.

(partial twice wrt each input)

Scrambling gets RMSE $O(n^{-2-1/2+\epsilon})$

Even higher

Start with ks dimensions interleave down to s.

Get $O(n^{-k+\epsilon})$ and $O(n^{-k-1/2+\epsilon})$ (under still higher smoothness)

Upshot

Not yet widely used. Takes a lot of smoothness and a lot of inputs.

The curse of dimension

Curse of dimension: larger d makes integration harder.

$$C_M^r = \left\{ f : [0,1]^d \to \mathbb{R} \;\middle|\; \left| \prod_j \frac{\partial^{\alpha_j}}{\partial x_j^{\alpha_j}} f \right| \leqslant M,\ \sum_j \alpha_j = r,\ \alpha_j \geqslant 0 \right\}$$

Bahkvalov I:

For any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in [0,1]^d$ there is $f \in C_M^r$ with $|\hat{\mu}_n - \mu| \geqslant k n^{-r/d}$

Ordinary QMC like $r = d$

Bahkvalov II:

Random points can't beat RMSE $O(n^{-r/d-1/2})$

Ordinary MC like $r = 0$

What if we beat those rates?

Sometimes we get high accuracy for large d.

It does not mean we beat the curse of dimensionality.

Bahkvalov never promised universal failure.

Only the existence of hard cases.

We may have just had an easy, non-worst case function.

Two kinds of easy

- Truncation: only the first $s \ll d$ components of $\boldsymbol{x}$ matter
- Superposition: the components only matter "s at a time"

Either way

f might not be "fully d-dimensional".

Lifted curse

Two main tools to describe it

- Weighted spaces and tractability
- ANOVA and effective dimension

Implications

Neither causes the curse to be lifted. They describe the happy circumstance where the curse did not apply.

Both leave important gaps described below.

Weighted spaces

Hickernell (1996), Sloan & Wozniakowski (1998)

$$\partial^u \equiv \prod_{j \in u} \frac{\partial}{\partial x_j} \qquad \text{assume } \partial^{1:d} f \text{ exists}$$

Inner product, weights $\gamma_u > 0$

$$\langle f, g \rangle = \sum_{u \subseteq 1:d} \frac{1}{\gamma_u} \int_{[0,1]^u} \left(\int_{[0,1]^{-u}} \partial^u f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}_{-u} \right) \left(\int_{[0,1]^{-u}} \partial^u g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}_{-u} \right) \mathrm{d}\boldsymbol{x}_u$$

Function ball $B_{\gamma,C} = \{ f \mid \langle f, f \rangle_\gamma \leqslant C \}$

Small $\gamma_u \implies$ small $\| \partial^u f \|$ in ball.

Product weights

$\gamma_u = \prod_{j \in u} \gamma_j$ where γ_j decrease rapidly with j.
Now $f \in B_{\gamma,C}$ implies $\partial^u f$ small when $|u|$ large.

Tractability

$\mathrm{Err}(d, n) \equiv$ Worst case error in function ball B_γ

$n_*(d, \epsilon) \equiv$ First n with $\mathrm{Err}(d, n) \leqslant \epsilon \times \mathrm{Err}(d, 0)$

Weak tractability

$$n_*(d, \epsilon) = \mathrm{poly}(d, 1/\epsilon)$$

$$\limsup_{d \to \infty} \frac{\sum_{j=1}^{d} \gamma_j}{\log d} < \infty \quad \text{suffices}$$

Strong tractability

$$n_*(d, \epsilon) = \mathrm{poly}(1/\epsilon)$$

$$\sum_{j=1}^{\infty} \gamma_j < \infty \quad \text{suffices}$$

Many versions and many contributors

Conditions above are for spaces described by Sloan & Wozniakowski (1998)

Novak, Wasilkowski, Hickernell, Dick, Kuo . . .

ANOVA and scrambled nets

Write the ANOVA of f

$$f(\boldsymbol{x}) = \sum_{u \subseteq \{1,2,\ldots,s\}} f_u(\boldsymbol{x})$$

Term by term

$$\hat{\mu} = \sum_{u \subseteq \{1,2,\ldots,s\}} \bar{f}_u \quad \text{where} \quad \bar{f}_u = \frac{1}{n} \sum_{i=1}^{n} f_u(\boldsymbol{x}_i)$$

$$\mathrm{Var}(\hat{\mu}) = \sum_{u \subseteq \{1,2,\ldots,s\}} \mathrm{Var}(\bar{f}_u)$$ O (1997)

Expect $O(n^{-3+\epsilon})$ from small $|u|$ and $O(n^{-1})$ from large $|u|$

Lower order f_u are formed by integration. Can be smoother than f.

Griebel, Kuo, Sloan (2010, 2013, 2014)

Effective dimension $s \leqslant d$

Caflisch, Morokoff & O (1997)

Often f is dominated by its low order interactions.

Then RQMC may make a huge improvement.

Let $\sigma_u^2 = \mathrm{Var}(f_u)$ variance component

Truncation sense

$$\sum_{u \subseteq 1:s} \sigma_u^2 \geqslant 0.99 \sum_{u \subseteq 1:d} \sigma_u^2$$

Superposition sense

$$\sum_{|u| \leqslant s} \sigma_u^2 \geqslant 0.99 \sum_{u \subseteq 1:d} \sigma_u^2$$

Mean dimension

$$\frac{\sum_u |u| \sigma_u^2}{\sum_u \sigma_u^2}$$

Liu & O (2006) (Easier to estimate $\cdots$ another story)

Example

Kuo, Schwab, Sloan (2012) consider quadrature for

$$f(\boldsymbol{x}) = \frac{1}{1 + \sum_{j=1}^{d} x_j^{\alpha}/j!}, \quad 0 < \alpha \leqslant 1.$$

For $\alpha = 1$ and $d = 500$

$R = 50$ replicated estimates of $\sum_v |v| \sigma_v^2/\sigma^2$ using $n = 10{,}000$

had mean 1.0052 and standard deviation 0.0058.

Upshot

$f(\boldsymbol{x})$ is nearly additive

mean dimension between 1.00356 and 1.00684

(± 2 standard errors)

Weighted spaces

- ☺ γ_u ensure good performance for a class of f.
- ☹ Tractability conditions greatly restrict f.
- ☺ QMC can be tuned to given γ e.g., Dick & Pillichshammer (2010)
- ☹ We don't measure/observe a true γ.

Effective dimension

- ☺ Can be measured for a given f.
- ☹ Does not imply smooth interactions f_u.
- ☺ Anova components can be smoother than f Sloan, Kuo, Griebel.

Description vs engineering

Effective dimension and weighted spaces both **describe** settings where f can be integrated well.

Some authors try to **engineer** changes in f to make it more suited to QMC.
E.g., cram importance into first few components of $\boldsymbol{x}$.

MC vs QMC

MC places lots of effort on variance reduction.

For QMC we gain by reducing effective dimension.

Or Hardy-Krause variation (but there asymptotics are slow).

E.g., turning $\boldsymbol{u} \sim \mathbf{U}[0,1]^d$ into $\boldsymbol{z} \sim \mathcal{N}(0, \Sigma)$

The choice of $\Sigma^{1/2}$ affects QMC performance.

Caflisch, Morokoff & O (1997)

Acworth, Broadie & Glasserman (1998)

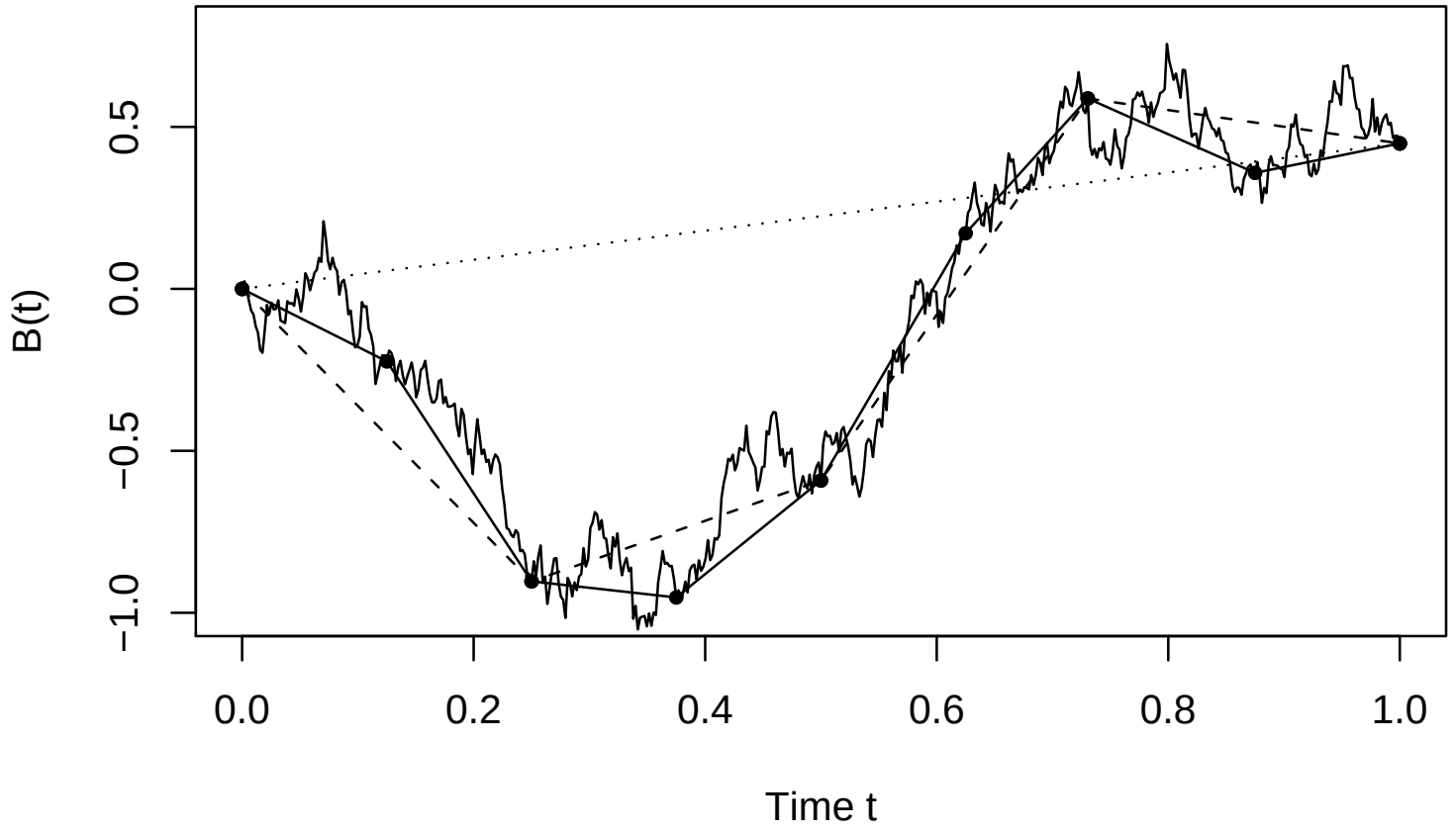
Imai & Tan (2014)

Sampling Brownian motion

Feynman-Kac/Brownian bridge

First few variables define a ‘skeleton’. The rest fill in.

Brownian bridge construction of Brownian motion



See also Mike Giles++ on multi-level MC.

Acworth, Broadie and Glasserman simulate BM in terms of its principal components. If the first few (or any few) PCs dominate the evaluation function, then it reduces effective dimension.

There can be no universal BM generator that is best for all integrands. Papageorgiou (2002) exhibits a financial option for which the ordinary step by step construction of Brownian motion works better than the Brownian bridge construction.

Imai and Tan have a series of papers tuning the generation of Brownian motion to specific integrands.

The book edited by Schreider (1964) “Method of statistical testing, Monte Carlo method” includes a Chapter II by I. M. Sobol’ in which the Brownian bridge construction is described. Sobol’ cites a 1958 article by Gel’fand, Frolov and Chentsov “The evaluation of Wiener integrals by the Monte Carlo method”. Izv. vys. uch. zav. ser. matem., No. 5, 32–45

Choosing γ

Each γ corresponds to a reproducing kernel Hilbert space (RKHS)

The question

Which RKHS should we use in a given problem?

$\mathcal{H}_1$ or $\mathcal{H}_2$ or $\cdots$ or $\mathcal{H}_J$ $\cdots$

1) sometimes $f \in \mathcal{H}_j$ all $j = 1, \ldots, J$
   and $f \in \mathcal{H}_1$ vs $\mathcal{H}_2$ have very different implications

2) sometimes f belongs to none of them.
   while $|f - \widetilde{f}| \leqslant \epsilon$ where $\widetilde{f} \in \mathcal{H}$

Bayes and empirical Bayes ideas might help choose $\mathcal{H}$

Maybe we want an $\mathcal{H}$ where f is 'typical'.

A natural γ has

$$\gamma_u \propto \int_{[0,1]^d} \bigl(\partial^u f(\boldsymbol{x})\bigr)^2 \,\mathrm{d}\boldsymbol{x}$$

NB: the constant of proportionality is also important.

Asymptotics

Two ways to think about asymptotics.

View 1

We really want to know what happens in the limit as $n \to \infty$.

View 2

We want an approximation for finite n, like $\frac{1}{n} \approx 0$.

For QMC

If the dimension is large, then the asymptotics solve view 1 not view 2.

Practical n are likely to be on an 'initial transient'

Sometimes the asymptote 'sets in' at $n \approx 4^d$.

For MC

$\mathrm{Var}(\hat{\mu}) = \sigma^2/n$ 'sets in' for $n \geqslant 1$.

Sequential QMC

JRSS-B discussion paper by Chopin & Gerber (2015)

N particles in $\mathbb{R}^d$ for T time steps

Advance $\boldsymbol{x}_{nt}$ to $\boldsymbol{x}_{n,t+1} = \phi(\boldsymbol{x}_{nt}, \boldsymbol{u}_{n,t+1})$ each $\boldsymbol{u} \in [0,1)^S$.

Do them all with $U_{t+1} \in [0,1)^{N \times S}$. Usually IID.

For QMC

We want rows of U_{t+1} to be 'balanced' wrt position of $\boldsymbol{x}_t$

Easy if $d = S = 1$ Lécot & Tuffin (2004)

Take $U_{t+1} \in [0,1)^{N \times 2}$

Sort first column along $\boldsymbol{x}_{n,t}$. Update points with remaining column.

What to do if $d > 1$?

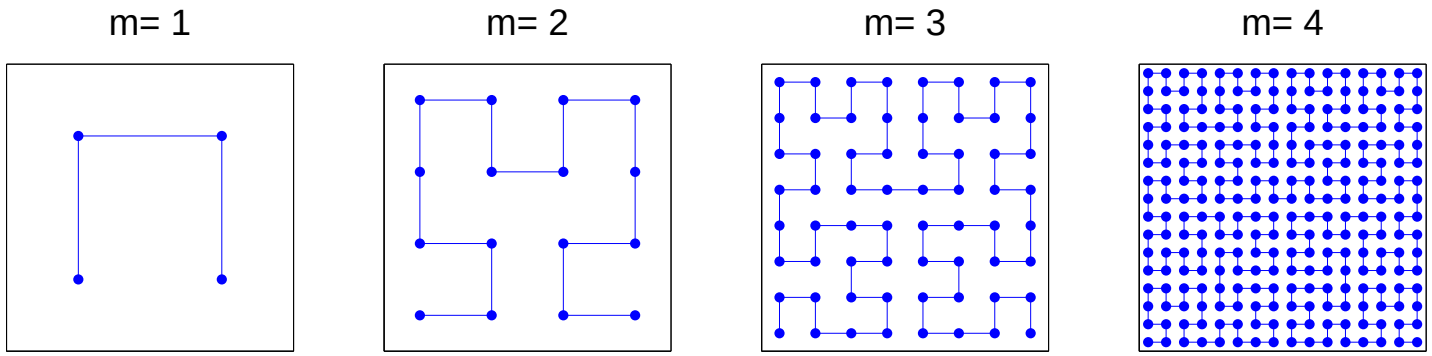
Alignment is tricky because $\mathbb{R}^d$ is not ordered.

Related work by L'Ecuyer, Lécot, L'Archevêque-Gaudet on array-RQMC.

d-dimensional alignments. Huge variance reductions. Limited theory.

Using Hilbert curves

These are space-filling curves as $m \to \infty$:



Chopin & Gerber thread a curve through $\boldsymbol{x}_{n,t} \in \mathbb{R}^d$

Align particles via first column of RQMC points $U_{t+1} \in [0,1)^{N \times (1+S)}$.

Use remaining S columns to advance them.

Results

Squared error is $o(N^{-1})$.

Good empirical results for modest d

NB:

I left out lots of details about particle weighting.

Recipe for QMC in MCMC

For MCMC we follow one particle.

1) Step $\boldsymbol{x}_i \leftarrow \psi(\boldsymbol{x}_{i-1}, \boldsymbol{v}_i)$ for $\boldsymbol{v}_i \in (0,1)^d$.
2) For $\boldsymbol{v}_1 = (u_1, \ldots, u_d)$, $\boldsymbol{v}_2 = (u_{d+1}, \ldots, u_{2d})$ etc.
   n steps require $u_1, \ldots, u_{nd} \in (0,1)$
3) MCMC uses $u_i \sim \mathbf{U}(0,1)$
4) Replace IID by $N = nd$ 'balanced' points

Reasons for caution

1) We're using 1 point in $[0,1]^{nd}$ with $n \to \infty$
2) Our $\boldsymbol{x}_i$ won't be Markovian

QMC ∩ MCMC

Early references

Chentsov (1967)

Plugs in 'completely uniformly distributed' points. (defined later)

Samples in finite state space by inversion.

Shows consistency.

Beautiful coupling argument. (Resembles coupling from the past.)

Sobol' (1974)

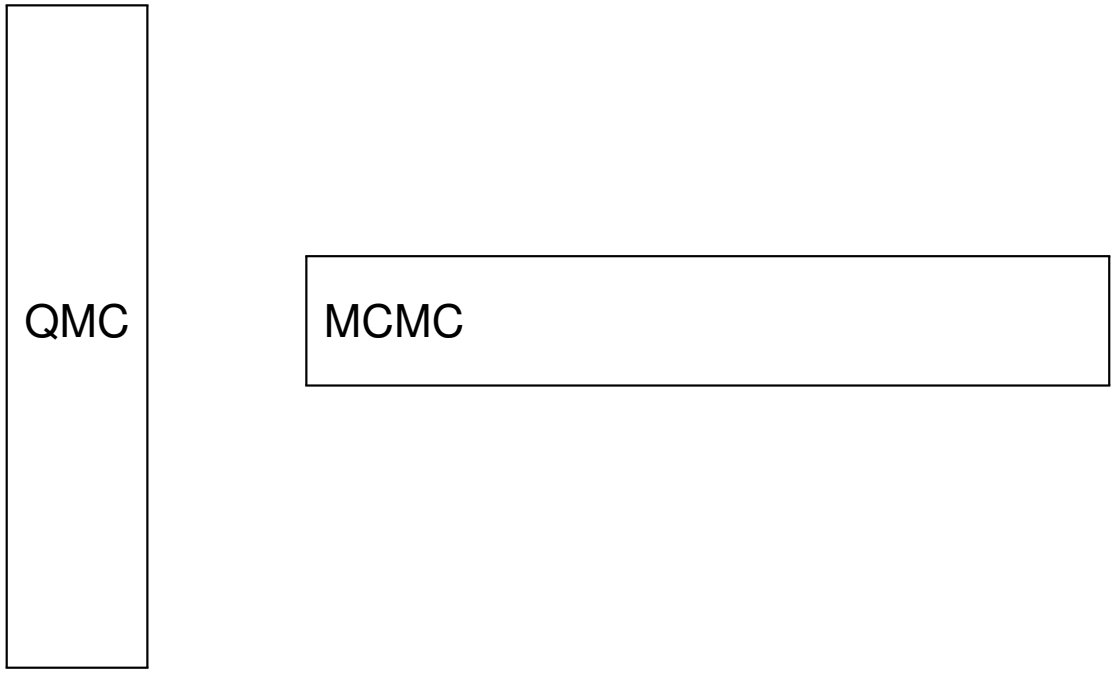
Has $n \times \infty$ points $x_{ij} \in [0, 1]$

Samples from a row until a return to start state, then goes to next row

Gets rate $O(1/n) \cdots$ if transition probabilities are $a/2^b$ for integers a, b

MCMC $\approx$ QMC$^{\mathsf{T}}$

Method	Rows	Columns	
QMC	n points	d variables	$1 \leqslant d \ll n \to \infty$
MCMC	r replicates	n steps	$1 \leqslant r \ll n \to \infty$

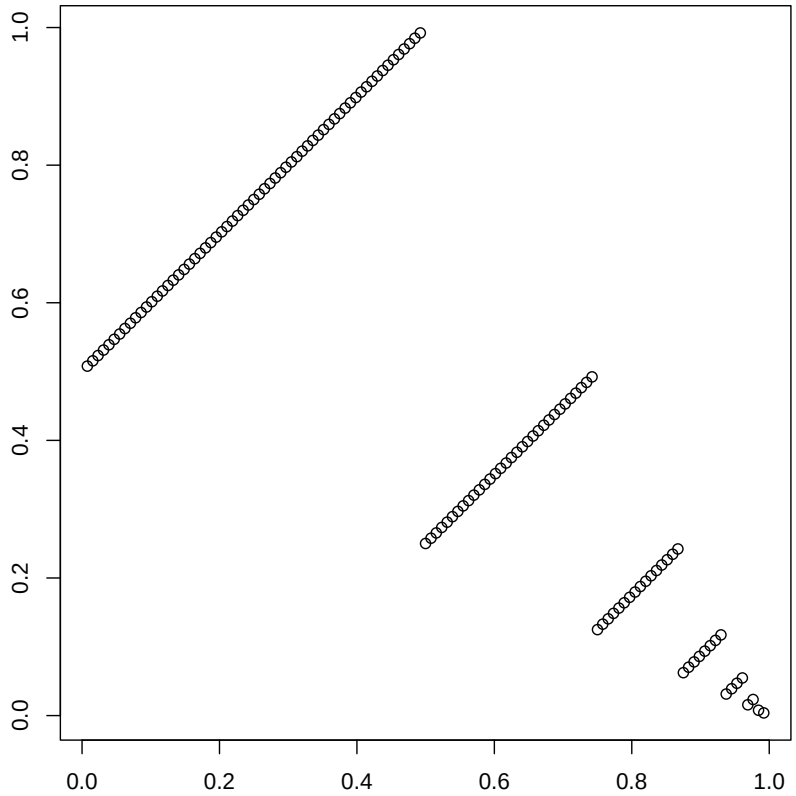


QMC based on equidistribution MCMC based on ergodicity

Severe failure is possible

van der Corput $u_i \in [0, 1/2) \iff u_{i+1} \in [1/2, 1)$

u_{i+1} vs u_i



High proposal $\iff$ low acceptance and vice versa

Morokoff and Caflisch (1993) describe heat particle leaving region

Completely uniformly distributed

$u_1, u_2, \cdots \in [0,1]$ are CUD if

$D_n^*(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \to 0,$ where

$\boldsymbol{z}_i = (u_i, \ldots, u_{i+d-1})$

For all $d \geqslant 1$

Overlapping blocks

$$\begin{aligned}
\boldsymbol{z}_1 &= (u_1, \ldots, u_d) \\
\boldsymbol{z}_2 &= (u_2, \ldots, u_{d+1}) \\
&\vdots \qquad \vdots \\
\boldsymbol{z}_n &= (u_n, \ldots, u_{n+d-1})
\end{aligned}$$

Chentsov (1967) shows we can use non-overlapping blocks

$$\boldsymbol{v}_i = \left(u_{d(i-1)_1}, \ldots, u_{di}\right) \forall d$$

CUD ctd

CUD $\equiv$ one of Knuth's **definitions** of randomness

Recommendations

1) Use all the d-tuples from your RNG

2) Be sure to pick a small RNG

As considered in

Niederreiter (1986)

Entacher, Hellekalek, and L'Ecuyer (1999)

L'Ecuyer and Lemieux (1999)

Some results

1) Metropolis on CUD points consistent in finite state space O & Tribble (2005)

2) Weakly CUD random points ok Tribble & O (2008)

3) Consistency in continuous problems (Gibbs and Metropolis)
   Chen, Dick & O (2010)

4) Chen (2011) rate $O(n^{-1+\epsilon})$ for some smooth ARMAs

Related refs

Liao (1998)	reorders QMC points for MCMC
Craiu & Lemieux (2007)	QMC in multiple-try Metropolis
Lemieux & Sidorsky (2006)	QMC in exact sampling
Lemieux, Ormoneit, Fleet (2001)	QMC for particle filters
Frigessi, Gäsemyr, Rue (2000)	MCMC $\cap$ antithetics
Craiu, Meng (2004)	MCMC $\cap$ Latin hypercubes
Propp (2004)	Rotor-Router

Thesis of Tribble (2007)

Results on CUD and weak CUD

Constructions of small RNGs

Examples with Gibbs and Metropolis

Variance reductions from Tribble

	$n = 2^{10}$		$n = 2^{12}$		$n = 2^{14}$	
Data sets	min	max	min	max	min	max
Pumps $(d = 11)$	286	1543	304	5003	1186	16089
Vasorestriction $(d = 42)$	14	15	56	76	108	124

Pumps: hierarchical Poisson-Gamma model.

Vasorestriction: probit model 3 coefficients, 39 latent variables.

Min & max variance reductions for all pump and all non-latent vaso. parameters.

Details

Targets are posterior means of parameters.

CUD points were LFSR, with Cranley-Patterson rotations.

Chen, Dick & O Annals Stat.

Chen, Dick & O extend consistency to continuous state spaces.

MCMC remains consistent when driven by $u_1, u_2, \ldots,$ if

1) u_i are CUD (or CUD in probability)

2) m-step transitions are Riemann integrable $\forall m \geqslant 1$, and

3)
- for Metropolis-Hastings: there is a coupling region
  (Independence sampler can have one)
- for Gibbs: there is a contraction property
  (Gibbs for probit model proven to contract)

Either way, the chain has to forget it's past.

That could be by regeneration or exponential decay.

Aside: Henri Lebesgue made our lives so much easier!

Small RNGs

Matsumoto & Nishimura sent us some small linear feedback shift register RNGs

Similar equidistribution properties to "small Mersenne twisters" but not necessarily the same constructions.

They come in sizes $M = 2^m - 1$ for $10 \leqslant m \leqslant 32$.

$$u_1, u_2, \ldots, u_M$$

We explore them for some simulations.

Prepend one or more 0s:

$$0, \ldots, 0, u_1, \ldots, u_M$$

put into a matrix and apply Cranley-Patterson rotations

Summary

Bivariate Gaussian	apparent better convergence rate for mean
Hit and run, volume estimator	no improvement
M/M/1 queue, average wait	mixed results
Garch	some big improvements
Heston stochastic volatility	big improvements for in the money case

Synopsis

The smoother the problem, the more CUD points can improve.

Same as for finite dimensional QMC.

What next?

QMC helps MCMC the most when transitions are smooth.

i.e., no acceptance rejection.

Some thoughts

- Gibbs is promising. Maybe tweak JAGS.
- Maybe Hamiltonian dynamics using high acceptance.
- When there is acceptance-rejection, maybe the optimal acceptance rate for QMC is higher.

Thanks

- QMC ∩ MCMC co-authors:

  Seth Tribble, Su Chen, Josef Dick, Makoto Matsumoto, Takuji Nishimura

- MCMSki organizers

- Nicolas Chopin

- NSF