Monte Carlo and quasi-Monte Carlo for image synthesis

Art B. Owen,

Stanford University

These are the slides I presented in a keynote at the 25th Eurographics symposium on rendering in June 2014 in Lyon.

Many/most of the figures are from my work-in-progress book, ten chapters of which are posted at `statweb.stanford.edu/~owen/mc`

I have made a few changes to the slides, adding notes about things said aloud, adding a few references, and making a few corrections.

It was a delightful meeting, socially and scientifically, and I am very grateful to Wojciech Jarosz and Pieter Peers for inviting me and to Victor Ostromoukhov, the local organizer.

-Art Owen, June 2014

Sampling for graphics

- light travels from sources to retina/lens
- bouncing off of objects
- each pixel is an average over light paths
- sampling those paths fits naturally (at least since Kajiya (1988))
- and converges numerically
- Helmholtz: we can even sample the reversed paths

Sampling challenges

- High or infinite dimensional integrands.
- Singular integrands.
- Lack of smoothness.
- Visual artifacts, despite good numerical accuracy

Rendering has all of these challenges.

Bidirectional research

- Start at a rendering problem, adapt sampling ideas, or,
- Start at a sampling idea, adapt it to rendering, or,
- Start at both ends, and meet in the middle

Research path occlusion

- computational cost
- unforeseen nastiness of the integrand
- unforeseen visual artifacts
- patents

A lot can go wrong between idea and implementation. It is hard to see around corners. This talk shows some sampling ideas, old and new, selected for their potential to be useful in sampling.

A tour of some sampling ideas

1) MC Monte Carlo
2) QMC Quasi-Monte Carlo
3) RQMC Randomized Quasi-Monte Carlo
4) MCMC Markov chain Monte Carlo
5) MLMC Multilevel Monte Carlo

Additionally

1) New multiple importance sampling method
2) New low discrepancy sampling in the triangle
3) New results for Hilbert curve sampling

Important but omitted

Sequential Monte Carlo Particle methods

Blue noise (Fiume & McCool, Ostromoukhov, Mitchell, Keller, $\cdots$)

(Crude) Monte Carlo

We want $$\mu \equiv \int_{\mathbb{R}^d} f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

We use $$\hat{\mu} \equiv \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i), \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} p$$

Computationally

Get $\mathbf{U}[0,1]$ random variables via Mersenne Twister (or other RNG)

Matsumoto & Nishimura (1988)

Turn them into samples from p

Devroye (1986)

Monte Carlo Properties

Law of large numbers

$$\mathbb{P}\Bigl(\lim_{n\to\infty}\hat{\mu}=\mu\Bigr)=1 \quad \text{if } \mu \text{ exists}$$

Central Limit Theorem

$$\sqrt{n}(\hat{\mu}-\mu)\xrightarrow{\mathrm{d}}\mathcal{N}(0,\sigma^2), \quad \text{if } \sigma^2=\int(f(\boldsymbol{x})-\mu)^2p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}<\infty$$

Root mean square error $\mathbb{E}((\hat{\mu}-\mu)^2)^{1/2}=\sigma/\sqrt{n}$

vs classic quadrature $O(n^{-r/d})$ using r derivatives in d dimensions.

Good news

It is easy to estimate error

Competitive when dimension is high or smoothness low

Monte Carlo problems/fixes

Accuracy may be too low

1) Quasi-Monte Carlo
2) Importance sampling
3) Other variance reductions

For InDevroyable* distributions

1) Markov chain Monte Carlo
2) Multilevel Monte Carlo
3) Sequential Monte Carlo

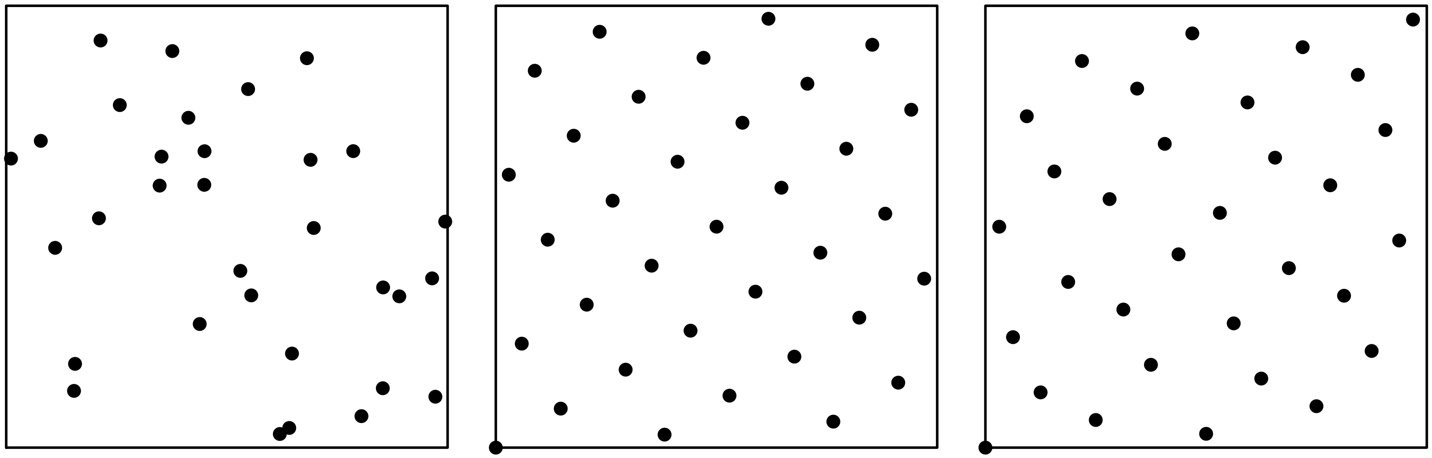
*I.E. not available by methods of Devroye (1986)

InDevroyable could have been UnDevroyable, but the former works well in both English and French.

Quasi-Monte Carlo

Monte Carlo simulates randomness. We don't need that, we just need an accurate answer. QMC chooses points more uniformly than MC does.

MC and two QMC methods in the unit square



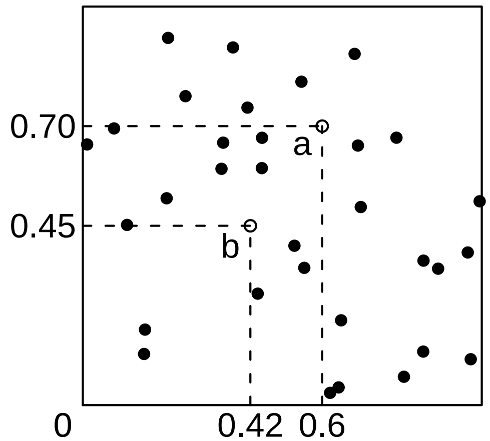
Monte Carlo Fibonacci lattice Hammersley sequence

MC yields holes and clumps in random places.

QMC used in graphics by Keller, Niederreiter, Heinrich, Kollig, Shirley, Grunschloss and others.

Discrepancies

Local discrepancy at a, b



$$\delta(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n} 1_{\boldsymbol{x}_i \in [0,\boldsymbol{a})} - \mathbf{vol}([0,\boldsymbol{a}))$$
$$= \frac{13}{32} - 0.6 \times 0.7 = -0.01375$$

Star discrepancy

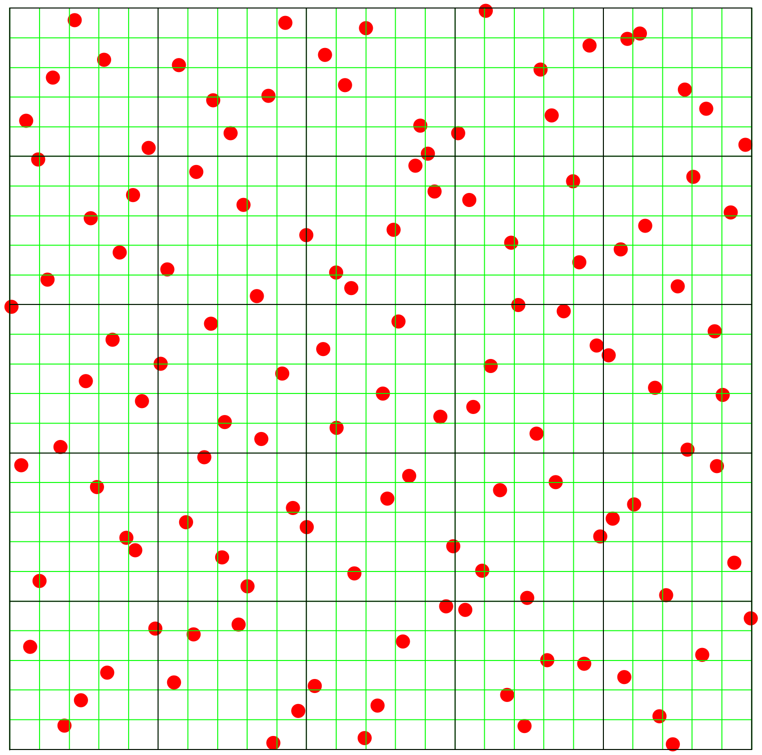
$$D_n^*(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sup_{\boldsymbol{a}\in[0,1)^d} |\delta(\boldsymbol{a})|.$$

I.E., worst of the local discrepancies

Uniformly distributed points

$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ are uniformly distributed (u.d.) iff $D_n^*(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \to 0$

Digital nets



Stratified $\frac{1}{125} \times 1$ and $\frac{1}{25} \times \frac{1}{5}$ and $\frac{1}{5} \times \frac{1}{25}$ and $1 \times \frac{1}{125}$ and $\frac{1}{5} \times \frac{1}{5} \times \frac{1}{5}$

Generalizes to more dimensions. Extensible in n.

Constructions by Sobol', Faure, Niederreiter, Niederreiter-Xing.

QMC properties

Counterpart to LLN

If $f(\boldsymbol{x})$ is Riemann integrable and $D_n^* \to 0$, then $\hat{\mu} \to \mu$

If f is **not** Riemann integrable then $\hat{\mu} \not\to \mu$ for **some** u.d. points

Counterpart to CLT

$$|\hat{\mu} - \mu| \leqslant D_n^*(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) V_{\mathrm{HK}}(f)$$

This is the Koksma-Hlawka inequality. It is a 100% bound on error.

V_{HK} is total variation, in the sense of Hardy and Krause.

QMC vs MC

In favor of QMC:

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

$D_n^* = O(n^{-1+\epsilon})$ is possible

Then $|\hat{\mu} - \mu| = O(n^{-1+\epsilon})$ vs $n^{-1/2}$ for MC

Against QMC:

V_{HK} far harder to estimate than μ, so no error estimate

$V_{\mathrm{HK}} = \infty$ for singular integrands

$V_{\mathrm{HK}} = \infty$ for most discontinuities (e.g., occlusion)

$$C \times n^{-1+\epsilon} \times \mathrm{Unknown} = \mathrm{Unknown}$$

QMC in high dimensions

QMC can lose effectiveness in high dimensions

or it can remain effective

It all depends on f

Low effective dimension

If f is nearly a sum of functions of a few variables then QMC remains effective[1]

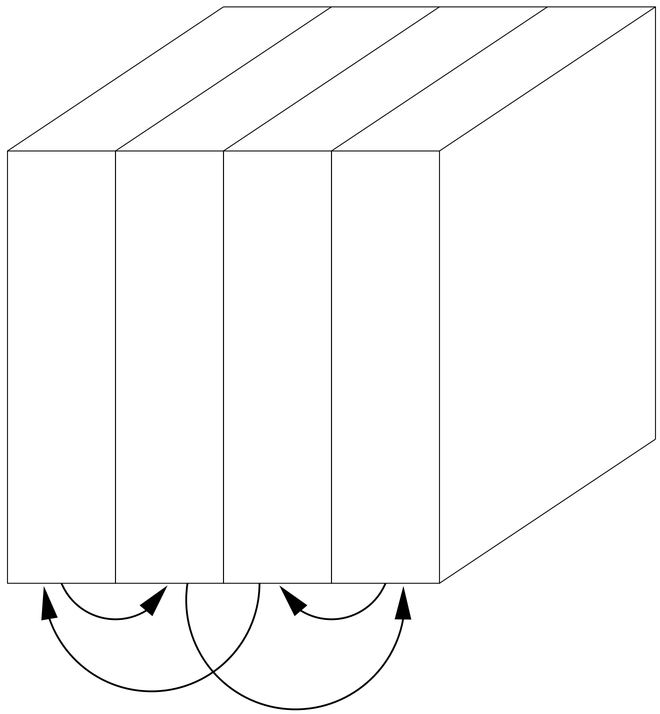
E.g. $f(x_1, x_2, \ldots, x_d) \doteq f_a(x_1, x_2) + f_b(x_{17}) + f_c(x_2, x_8, x_{1000}) + \cdots$
all functions with only a few inputs

Caflisch, Morokoff & O (1997), Sloan & Woźniakowski (1998)

Low effective dimension is surprisingly common.

[1] about as effective as the low dimensional component integrations.

Randomized QMC



This method yields “scrambled nets”

Chop $[0, 1]^d$ into b pieces

Randomly shuffle

Chop the pieces and recurse

Apply to all d axes

Scrambled net properties

Each $\boldsymbol{x}_i \sim \mathbf{U}[0,1]^d$, so

$$\mathbb{E}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\int f(\boldsymbol{x}_i)\,\mathrm{d}\boldsymbol{x}_i = \mu.$$

If f is smooth

$$\sqrt{\mathrm{Var}(\hat{\mu})} = O(n^{-3/2+\epsilon}) \quad \text{vs } O(n^{-1+\epsilon}) \quad \text{for QMC}$$

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

$$\sqrt{\mathrm{Var}(\hat{\mu})} = o(n^{-1/2})$$

even if $V_{\mathrm{HK}}(f) = \infty$.

Error estimation

Use independent replications.

Summary

	$\lvert\mu\rvert < \infty$	$\sigma^2 < \infty$	$V_{\mathrm{HK}} < \infty$	Smooth[1]
MC	$o(1)$	$O(n^{-1/2})$	$O(n^{-1/2})$	$O(n^{-1/2})$
QMC	×	×	$O(n^{-1+\epsilon})$	$O(n^{-1+\epsilon})$
RQMC	?	$o(n^{-1/2})$	$O(n^{-1+\epsilon})$	$O(n^{-3/2+\epsilon})$

Table shows error and RMSE rates.

[1]Finite mean square for ∂f once with respect to each x_j.

Higher order nets of Dick exploit greater smoothness and get better rates.

Markov chain Monte Carlo

Google scholar: About 136,000 results (0.05 sec) (June 2014)

We may not be able to generate $\boldsymbol{x}_i \sim p$.

Instead we take

$$\boldsymbol{x}_i = \phi(\boldsymbol{x}_{i-1}, \boldsymbol{u}_i), \quad \boldsymbol{u}_i \overset{\text{iid}}{\sim} \mathbf{U}[0,1]^s$$

with ϕ chosen so that $\boldsymbol{x}_i \xrightarrow{\text{d}} p$

Choosing ϕ

Incredible variety of methods

Estimation

$$\mu = \mathbb{E}(f(\boldsymbol{x}) \mid \boldsymbol{x} \sim p) \qquad \hat{\mu} = \frac{1}{n}\sum_{i=b+1}^{b+n} f(\boldsymbol{x}_i)$$

There are Markov chain laws of large numbers, central limit theorems and variance estimates. b is 'burn-in'.

Metropolis-Hastings

$\boldsymbol{x}_i$ represents a light path.

At $\boldsymbol{x}_i$, make a random proposal $\boldsymbol{y}$ from distribution $Q(\boldsymbol{x}_i \to \boldsymbol{y})$

Accept with probability

$$A(\boldsymbol{x}_i \to \boldsymbol{y}) = \min\left(1, \frac{p(\boldsymbol{y})}{p(\boldsymbol{x}_i)} \times \frac{Q(\boldsymbol{y} \to \boldsymbol{x}_i)}{Q(\boldsymbol{x}_i \to \boldsymbol{y})}\right)$$

If accepted $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{y}$ else $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i$

Intuition

Large	$\dfrac{p(\boldsymbol{y})}{p(\boldsymbol{x}_i)}$	to favor uphill moves
Large	$\dfrac{Q(\boldsymbol{y} \to \boldsymbol{x}_i)}{Q(\boldsymbol{x}_i \to \boldsymbol{y})}$	to avoid getting trapped

Symmetric proposals

If $Q(\boldsymbol{x} \to \boldsymbol{y}) = Q(\boldsymbol{y} \to \boldsymbol{x})$ we get $A(\boldsymbol{x} \to \boldsymbol{y}) = \min(1, p(\boldsymbol{y})/p(\boldsymbol{x}))$, of Metropolis et. al (1953).

MCMC issues

Used to good effect by Veach & Guibas (1997)

New work by Hachisuka, Kaplanyan & Dachsbarcher (2014)

combining with multiple IS

Can be very effective. Can also fail to 'mix'.

E.G.: a random walk on N steps takes $O(N^2)$ time to go back and forth.

So exploring a big poorly connected space takes lots of time.

Remedies

Proposals to embed QMC into MCMC.

Chentsov (1967), Liao (1998), Tribble & O (2005), Chen, Dick & O (2011), Chopin & Gerber (2014), Bornn, de Freitas, Eskelin, Fang & Welling (2013)

Momentum (hybrid or Hamiltonian MC) counters random walk-ness

Originated in physics: Duane, Kennedy, Pendleton & Roweth (1987).

In the STAN statistics software Hoffman & Gelman (2011)

Multilevel MC

This is a relatively new Monte Carlo technique.

There were dozens of presentations on it at MCQMC 2014 in Belgium.

Original use for sampling stochastic differential equations

Many more uses now

Key references

Giles (2008)

A 2-level precursor Heinrich (2001)

Stochastic process context

- We want to simulate a random $S(t)$ function on $t \in [0,1]$
- We simulate it at only T positions $f(1/T)$, $f(2/T), \dots, f(1)$.
- Get truncated realization $S_T(t)$, and $Y^{(T)} = f(S_T(\cdot))$
- Do N Monte Carlo simulations

Estimator

$$\hat{\mu}_T = \frac{1}{N}\sum_{i=1}^{N} Y_i^{(T)}$$

Typically $\mathbb{E}\big((\hat{\mu}_T) - \mu)^2\big) = \frac{c_1}{N} + \frac{c_2}{T^r} \quad \equiv$ variance $+$ bias^2

The cost is $C = O(NT)$

Root mean squared error is worse than $C^{-1/2}$ due to bias-variance tradeoff.

E.g., Euler method has $r = 2$ and optimized RMSE is $O(C^{-1/3})$ not $O(C^{-1/2})$.

Multilevel idea

Do simulations at $T = T_\ell$ e.g., $T_\ell = 2^\ell$, for $\ell = 0, 1, 2, \ldots, L$.

Let $\mu_\ell = \mathbb{E}(f(S_{T_\ell}(\cdot))$.

Telescoping sums

$$\mu_L = \mu_0 + (\mu_1 - \mu_0) + (\mu_2 - \mu_1) + \ldots + (\mu_L - \mu_{L-1}) \qquad \equiv \sum_{\ell=0}^{L} \delta_\ell$$

$$\hat{\mu}_L = \hat{\mu}_0 + \widehat{\mu_1 - \mu_0} + \widehat{\mu_2 - \mu_1} + \ldots + \widehat{\mu_L - \mu_{L-1}} \qquad \equiv \sum_{\ell=0}^{L} \widehat{\delta}_\ell$$

Estimates

$$\hat{\delta}_\ell = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \hat{\delta}_{\ell,i}$$

Optimal allocation

$$N_\ell \propto \sqrt{\frac{\mathrm{Var}(\hat{\delta}_\ell)}{\text{Cost}(\hat{\delta}_{\ell,i})}}.$$

Coupling

To make multilevel work we need

1) Unbiased estimate of $\mu_\ell - \mu_{\ell-1}$

2) Very close paths $S_{T_\ell}(t) \doteq S_{T_{\ell-1}}(t)$

The second step is coupling.

E.G., the path $S_{256}(\cdot)$ should not be sampled independently of $S_{128}(\cdot)$.

Instead it should be a refinement with very small $S_{256}(\cdot) - S_{128}(\cdot)$.

For stochastic differential equations

Do a small number of expensive simulations at very large T

Increase that number as T decreases,

doing a large number of low cost simulations at very small T.

For favorable smoothness and coupling accuracy, RMSE can be $O(C^{-1/2})$

Other multilevel uses

Continuous time Markov chains

Biochemical kinetics Anderson & Higham (2012)

FPGA's with T bits

Use 2-bit, 4-bit, 8-bit $\cdots$ 64-bit computation

Liu (2012), Brugger et. al (2014)

De-biasing

Rhee & Glynn (2012), McLeish (2011)

$$Y = \sum_{\ell=0}^{\infty} X_\ell \qquad \mu = \sum_{\ell=0}^{\infty} \delta_\ell \qquad \delta_\ell = \mathbb{E}(X_\ell)$$

Choose random $L > 1$ independent of X_ℓ

$$\mathbb{E}(Y) = \mathbb{E}\left(\sum_{\ell=0}^{\infty} \frac{X_\ell 1_{L \geqslant \ell}}{\mathbb{P}(L \geqslant \ell)}\right) = \mathbb{E}\left(\sum_{\ell=0}^{L} \frac{X_\ell}{\mathbb{P}(L \geqslant \ell)}\right)$$

Three new directions

Hera He, Stanford

Optimal mixing in multiple importance sampling

Zhijian He, Tsinghua

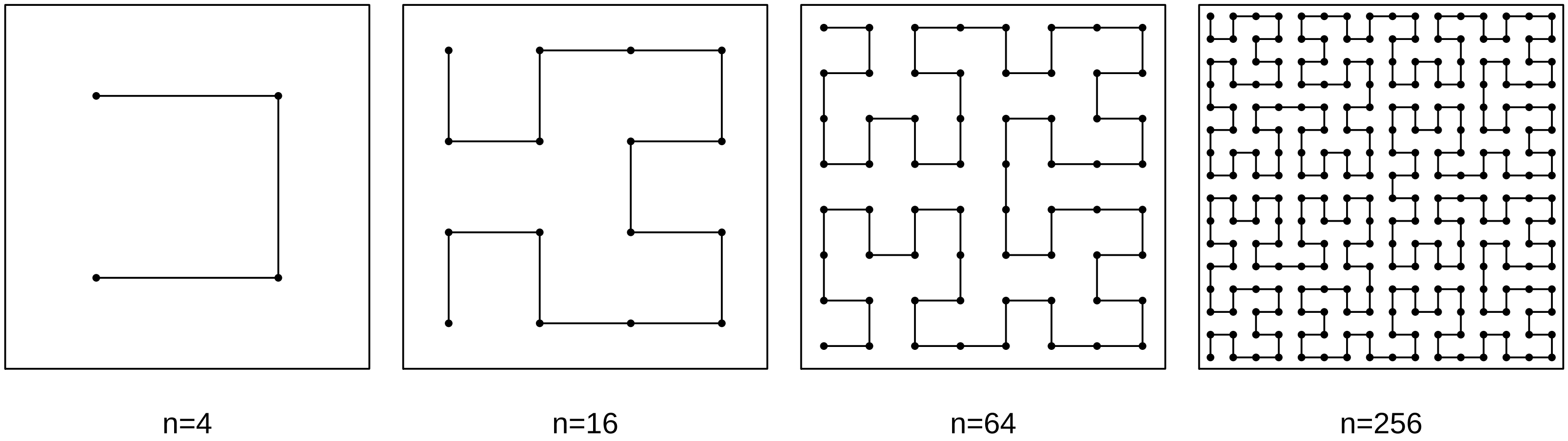
Sampling along a Hilbert curve

Kinjal Basu, Stanford

QMC sampling in the triangle

Hilbert sampling

Hilbert curve samples



$H(x)$ maps $[0,1]$ onto $[0,1]^d$. Commonly used in graphics.

If $x \sim \mathbf{U}[0,1]$ then $H(x) \sim \mathbf{U}[0,1]^d$

Recently used by Chopin & Gerber (2014) for QMC particle sampling

We will sample $x_i \in [0,1]$ and use $H(x_i)$ as QMC points.

Discrepancy

Take $x_i \in [(i-1)/n, i/n] \quad i = 1, \dots, n$ (random or not)

$$D_n^* = O(n^{-1/d})$$

- QMC gets $O(n^{-1+\epsilon})$
- Hilbert gets same rate as sampling on an $n = m^d$ grid
- Available at **any** n
- extensible via van der Corput sampling of $[0, 1]$

Z. He & O (2014)

Integration

Take $x_i \sim \mathbf{U}[(i-1)/n, i/n] \quad i = 1, \dots, n$

f is Lipshitz continuous

$$\mathrm{Var}(\hat{\mu}) = O(n^{-1-2/d})$$

- Better than MC. Optimal rate for Lipshitz.
- Same rate as stratified sampling in an $n = m^d$ grid
- Available at **any** n
- van der Corput sampling of $[0, 1]$ gives extensible sequence with this rate.

He & O (2014)

Hilbert Integration

Take $x_i \sim \mathbf{U}[(i-1)/n, i/n] \quad i = 1, \dots, n$

$f(\boldsymbol{x}) = g(\boldsymbol{x}) + 1_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x})$, g, h Lipshitz continuous Ω well behaved set

This models occlusion.

$$\mathrm{Var}(\hat{\mu}) = O(n^{-1-1/d})$$

- Rate seems new.
- May be useful in low dimensions.
- Available at **any** n, extensible

He & O (2014)

QMC in the triangle

We want to integrate over a triangular region

$$\int_{\triangle} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

or maybe $\triangle^k$

$$\int_{\triangle}\int_{\triangle}\cdots\int_{\triangle} f(\boldsymbol{x}_1,\dots,\boldsymbol{x}_k)\,\mathrm{d}\boldsymbol{x}_1\dots\,\mathrm{d}\boldsymbol{x}_k$$

for a path connecting k triangular regions

Mapping

Arvo (1995) maps $[0,1]^2$ onto $\triangle$

More mappings in Devroye (1986) Also Pillards & Cools have 5 mappings.

Area preserving mappings; Jacobian may give infinite variation.

Brandolini, Colzani, GIgante, Travaglini (2013) have discrepancy and Koksma-Hlawka inequality for $\triangle$ but $\cdots$ no constructions.

Triangular van der Corput

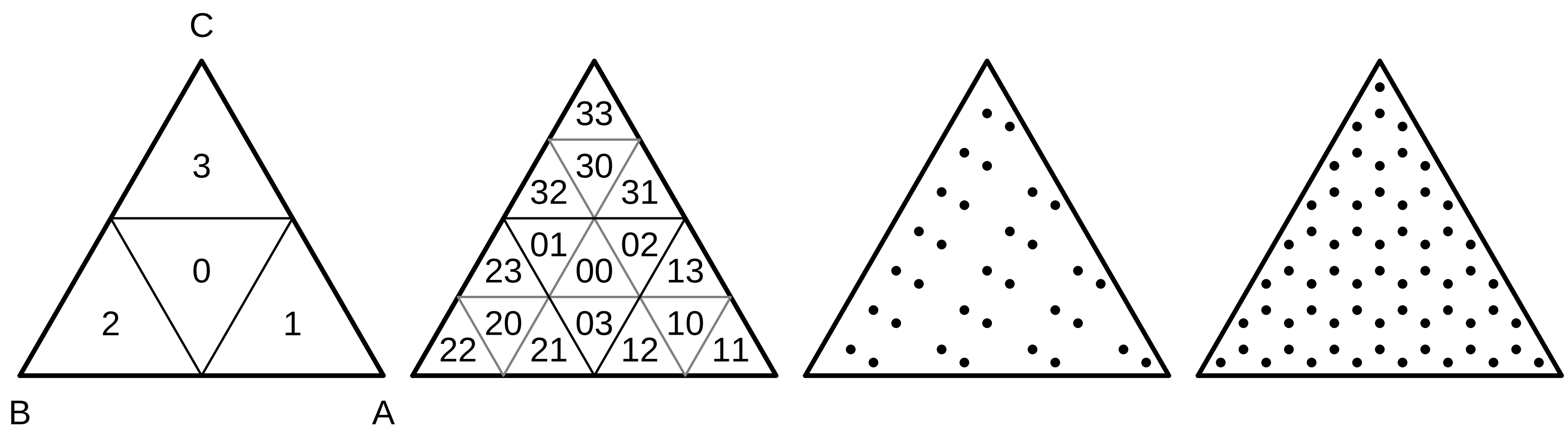
$n = \sum_{k \geqslant 1} d_k 4^{k-1}, d_k \in \{0, 1, 2, 3\}$

Place into subtriangle corresponding to d_1

Then sub-subtriangle corresponding to d_2, etc.

$n \to \boldsymbol{x}_n \in \triangle$

First 32 and 64 points



Triangular van der Corput

For $n = 4^k$, $D_{\triangle}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \leqslant \frac{2}{\sqrt{n}} - \frac{1}{n}$

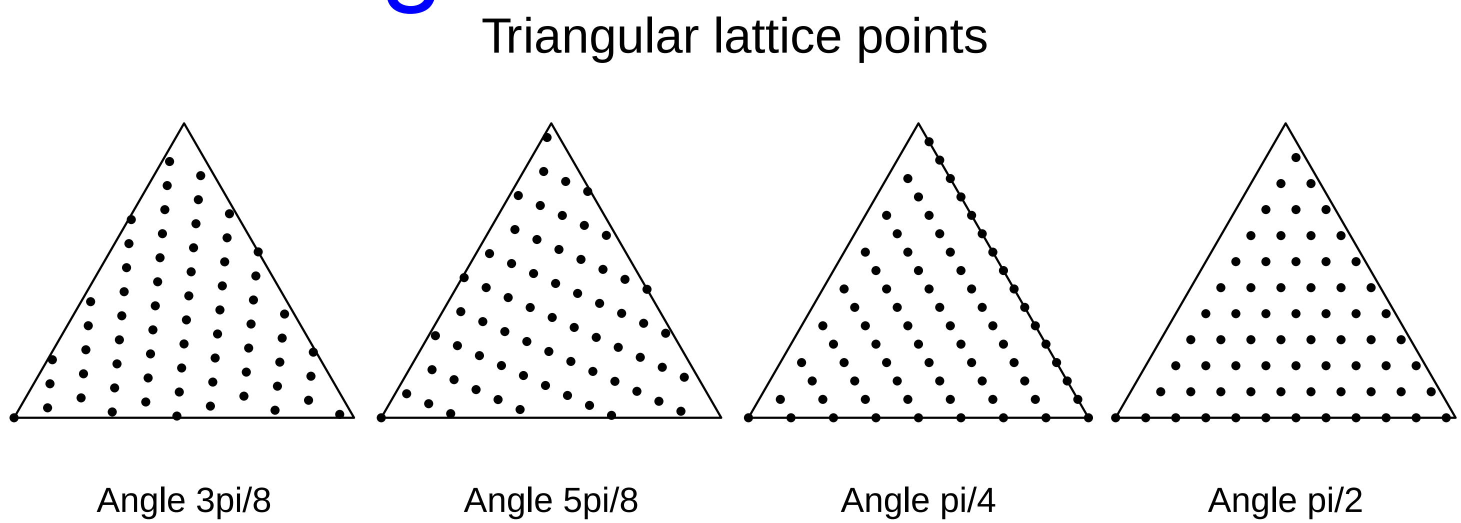
Generally $D_{\triangle}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \leqslant \frac{12}{\sqrt{n}}$

Result: consistent estimation for any Riemann integrable function on $\triangle$

Deterministic $O(n^{-1/2})$ estimation for bounded variation

Basu & O (2013)

Triangular Kronecker lattice



1) Take integer grid $\mathbb{Z}^2$
2) Rotate clockwise by angle $\alpha = 3\pi/8$
3) Shrink by $\sqrt{2n}$
4) Remove points not in $\triangle$ bounded by $(0,0), (0,1), (1,0)$
5) Add/remove $O(\log(n))$ points to get exactly n in $\triangle$
6) Map linearly to desired triangle

$D_\triangle(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) < C\log(n)/n$ Basu & O (2014)

In the previous figure angles $3\pi/8$ and $5\pi/8$ work well. Angles $\pi/4$ and $\pi/2$ are examples of what goes wrong when the angle is poorly chosen. They have big empty trapezoids.

The good angles are those for which $\tan(\alpha)$ is an irrational number 'badly approximable' by rational numbers. The best examples of these are the quadratic irrationals, any number of the form $(a + b\sqrt{c})/d$ where a is an integer, b and d are nonzero integers, and c is a positive integer that is not a perfect square.

Prior to the paper with Basu, one could deduce that good points did exist, but there was no explicit recipe for them.

$\triangle$ van der Corput

RQMC version gets RMSE $O(n^{-1})$

Also base 4 digital nets in $[0,1]^k$ lead to quadrature over $\triangle^k$

That is, sampling for paths

$$\triangle \to \triangle \to \triangle \to \cdots \to \triangle$$

or

$$\triangle \to \square \to \triangle \to \square \to \cdots \to \square \to \triangle$$

Potential graphics use

Computing form factors or throughput Schröder & Hanrahan (1993)

Hanrahan (1993) Rendering Concepts in Cohen & Wallace (1993)

Importance sampling

Often f is singular or only nonzero in a set A with $p\{A\} = \int_A p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \epsilon$.

We need lots of $\boldsymbol{x}_i$ in the 'important' region.

Choose $q(\boldsymbol{x})$ with $q(\boldsymbol{x}) > 0$ whenver $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$.

$$\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int \Bigl(f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\Bigr)q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

Importance sampler

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \qquad \boldsymbol{x}_i \stackrel{\text{iid}}{\sim} q$$

Sample from q; correct by multiplying by p/q

$\therefore$ we must be able to compute the ratio p/q

NB

Chapter 9 of `statweb.stanford.edu/~owen/mc` is on importance sampling. Chapter 10 includes material on adaptive IS.

Importance sampling properties

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \qquad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q$$

Variance

$$\text{Var}(\hat{\mu}_q) = \frac{1}{n}\left[\int \frac{f^2p^2}{q^2}q - \mu^2\right] = \frac{1}{n}\left[\int \frac{f^2p^2}{q} - \mu^2\right] = \frac{1}{n}\int \frac{(fp-\mu q)^2}{q}$$

Consequences

1) Perfect q is $\propto fp$ (when $f \geqslant 0$)

2) Good $q \mathrel{\dot{\propto}} fp$

3) Watch out for q that gets small

Defensive importance sampling

Hesterberg (1988, 1995)

Potential trouble if $q \ll p$. So use $q_\alpha = \alpha p + (1-\alpha) q$

$$\hat{\mu}_\alpha = \hat{\mu}_{q_\alpha} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{\alpha p(\boldsymbol{x}_i) + (1-\alpha)q(\boldsymbol{x}_i)} \qquad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q_\alpha$$

Bounded importance ratio

$$\frac{p(\boldsymbol{x}_i)}{\alpha p(\boldsymbol{x}_i) + (1-\alpha)q(\boldsymbol{x}_i)} \leqslant \frac{1}{\alpha} \quad \forall \boldsymbol{x}_i$$

Bounded variance

$$\mathrm{Var}(\hat{\mu}_\alpha) \leqslant \frac{1}{\alpha}\mathrm{Var}(\hat{\mu}_p) + \frac{1}{n}\frac{1-\alpha}{\alpha}\mu^2$$

Not much worse than using p.

But could be much worse than q.

Multiple importance sampling

Combine J different densities q_j, e.g. bidirectional path sampling

Veach & Guibas (1994), Lafortune & Willems (1993)

$$\hat{\mu}_{\boldsymbol{\alpha}} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{f(\boldsymbol{x}_{ij})p(\boldsymbol{x}_{ij})}{\sum_{j=1}^{J}\alpha_j q_j(\boldsymbol{x}_{ij})} \qquad \boldsymbol{x}_{ij} \sim q_j$$

This is the 'balance heuristic' of Veach & Guibas (1995)

Also a Horvitz-Thompson estimator

$$\mathrm{Var}(\hat{\mu}_{\boldsymbol{\alpha}}) \leqslant \mathrm{Var}(\hat{\mu}_{\mathrm{other}}) + \Bigl(\frac{1}{\min_j n_j} - \frac{1}{n}\Bigr)\mu^2$$

$\hat{\mu}_{\mathrm{other}}$ another weighting.

E.G., all weight on q_j, get $\mathrm{Var}(\hat{\mu}_{q_j})/\alpha_j$

Adding control variates

We know that

$$\int q_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int \frac{q_j(\boldsymbol{x})}{q_{\boldsymbol{\alpha}}(\boldsymbol{x})} q_{\boldsymbol{\alpha}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1.$$

Using this

$$\hat{\mu}_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} \frac{f(\boldsymbol{x}_{ij})p(\boldsymbol{x}_{ij}) - \sum_{j=1}^{J}\beta_j q_j(\boldsymbol{x}_{ij})}{q_{\boldsymbol{\alpha}}(\boldsymbol{x}_{ij})} + \sum_{j=1}^{J}\beta_j$$

Method

$\boldsymbol{\beta} = \mathbf{0}$ recovers balance-heuristic.

Optimize over $\boldsymbol{\beta}$ by least squares to reduce variance further.

$$\mathrm{Var}(\hat{\mu}_{\boldsymbol{\alpha},\boldsymbol{\beta}}) \leqslant \min_{1\leqslant j\leqslant J} \mathrm{Var}(\hat{\mu}_{q_j})/\alpha_j$$

O & Zhou (2000)

Shaves the multiple of μ^2 off of the variance bound

Optimizing over $\boldsymbol{\alpha}$

What if there are 1000's of densities q_j?

Sampling equally can be a waste

We can optimize over α

Convex optimization

The variance is **jointly convex** in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ H. He & O (2014) (in preparation)

Adaptive importance sampling, alternates between learning $\boldsymbol{\alpha}$ and using it

Constraints

We can constrain each $\alpha_j \geqslant \epsilon_j > 0$

Singularity example

This example was from work in progress with Hera He. It considers a 5 dimensional integrand that just barely has finite mean square.

Preliminary results were shown. At time of writing they are not final enough (still some checking to do).

What we saw was a large variance reduction from sequential multiple importance sampling. About 3×10^6-fold. Optimizing the weights gave a further variance reduction of about 5-fold.

The mixture components included some centered near the singularity and some distractors centered far away. This is to model the setting where we have imperfect knowledge of where the singularities might be. The gain from adaptive importance sampling should be roughly equal to the fraction of non-distractor variance components. When one mixes thousands of sampling distributions of which a small number are extremely helpful, then adapting α has the most potential to pay off.

Rare event example

This example was from work in progress with Hera He. It featured a rare event with probability on the order of 10^{-8}.

The model included some importance samplers based on approximate knowledge of where the rare event takes place as well as some additional samplers based on incorrect guesses about where the rare event takes place.

This time, mixture importance sampling reduced variance by about 2×10^5-fold and optimizing the mixture component gained a further 7-fold.

Thanks

- Wojciech Jarosz & Pieter Peers

  Organizers

- Victor Ostromoukhov

  Local organization

- Alex Keller, Eugene Fiume, Pat Hanrahan

  Additional advice on material to include

- Hera He, Kinjal Basu, Zhijian He

  Co-authors

- NSF DMS-0906056

  Funding