QMC for MCMC

Art B. Owen

Stanford University

Based on joint work with:

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

Simple Monte Carlo

Used in virtually all sciences

$$\mu = \mathbb{E}(f(x)), \qquad x \sim p$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i), \quad x_i \quad \text{IID} \quad p$$

Recall

$\mathbb{P}(\hat{\mu}_n \to \mu) = 1$ by Strong Law Large Numbers

If $\mathbb{E}(f(x)^2) < \infty$ then RMSE = $O(n^{-1/2})$

If $\mathbb{E}(f(x)^2) < \infty$ then Central Limit Theorem

Unfortunately:

MC is SLOW: one more digit accuracy $\equiv$ 100 fold more work

MC is HARD: getting $x_i \sim p$ is challenging (for Boltzmann, Bayes, $\cdots$)

But there's hope:

QMC improves accuracy from $O(n^{-1/2})$ to $O(n^{-1+\epsilon})$

MCMC broadens applicability

Talk in one slide

1) We want to combine the benefits of QMC and MCMC.
2) We can, via QMC points that are “completely uniformly distributed” (CUD)
3) Like using up all of your RNG
4) Involves a beautiful coupling argument from Chentsov (1967)
5) Greatest improvements for continuous example (e.g., Gibbs)
6) Sometimes a better rate
7) Interesting software engineering challenge

Markov chain Monte Carlo

Let $\boldsymbol{x}_i = \psi(\boldsymbol{x}_{i-1}, \boldsymbol{v}_i)$ $\quad \boldsymbol{v}_i \sim \mathbf{U}(0,1)^d$ (Markov property)

Design $\psi(\cdot,\cdot)$ so that distn($\boldsymbol{x}_i$) $\to p$

LLN for reasonable conditions

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) \to \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \equiv \mu$$

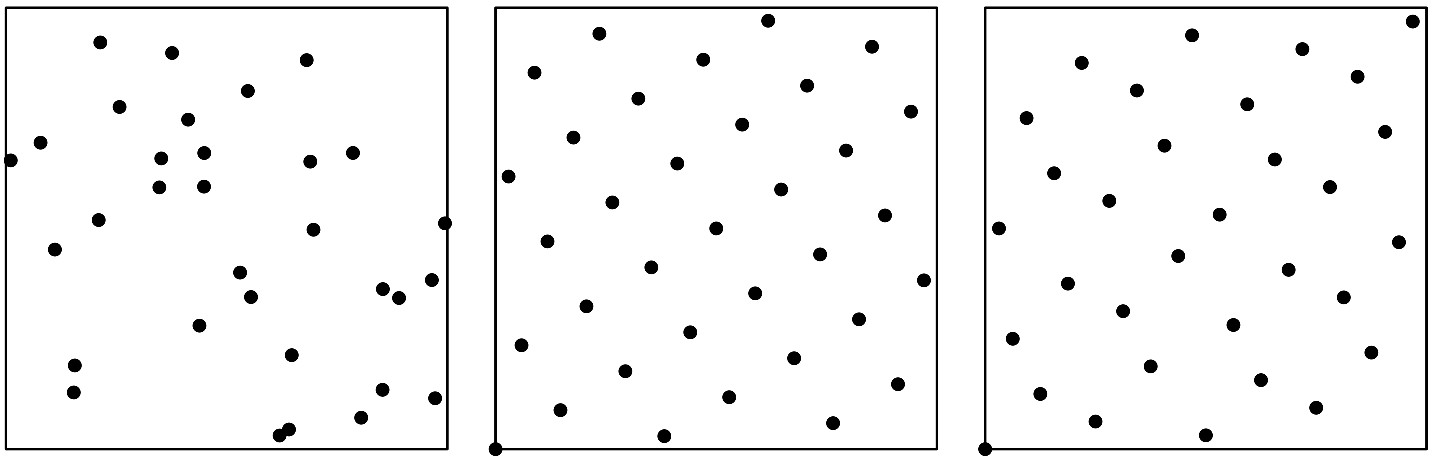
What we will do

$\boldsymbol{v}_i$ come from $u_1, u_2, u_3, \cdots \in (0,1)$.

We will replace IID u_i by balanced ones.

Quasi-Monte Carlo

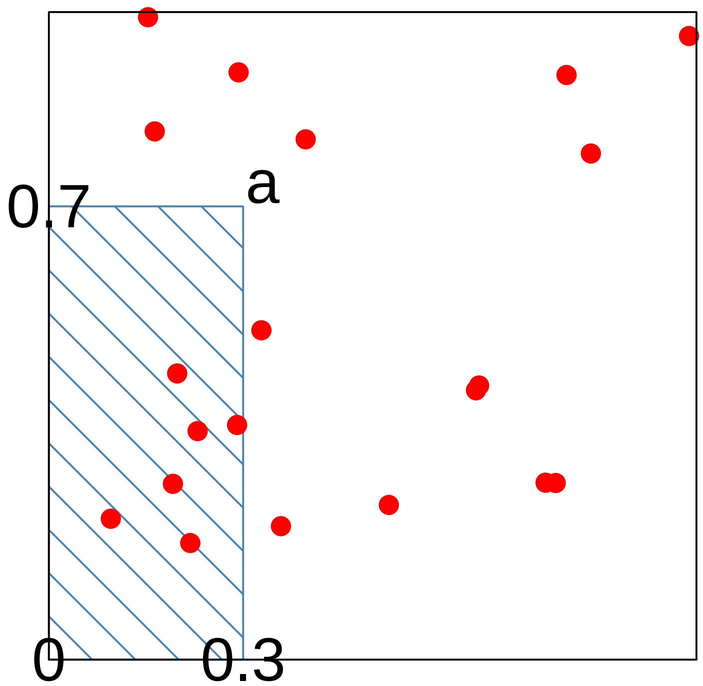
MC and two QMC methods in the unit square



Monte Carlo Fibonacci lattice Hammersley sequence

QMC places the points $x_i \in [0,1]^d$ **more uniformly** than Monte Carlo does.

Local discrepancy



The box $[0, a)$ contains $6/20$ points and has $.3 \times .7 = .21$ of the area.

$$\delta(a) = \frac{6}{20} - .3 \times .7 = .09$$

Star discrepancy

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

Recipe for QMC in MCMC

1) Each step $\boldsymbol{x}_i \leftarrow \psi(\boldsymbol{x}_{i-1}, \boldsymbol{v}_i)$ takes d numbers: in $\boldsymbol{v}_i \in (0,1)^d$.
2) n steps require $u_1, \dots, u_{nd} \in (0,1)$
3) MCMC uses $u_i \sim \mathbf{U}(0,1)$
4) Replace IID by balanced points

Reasons for caution

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

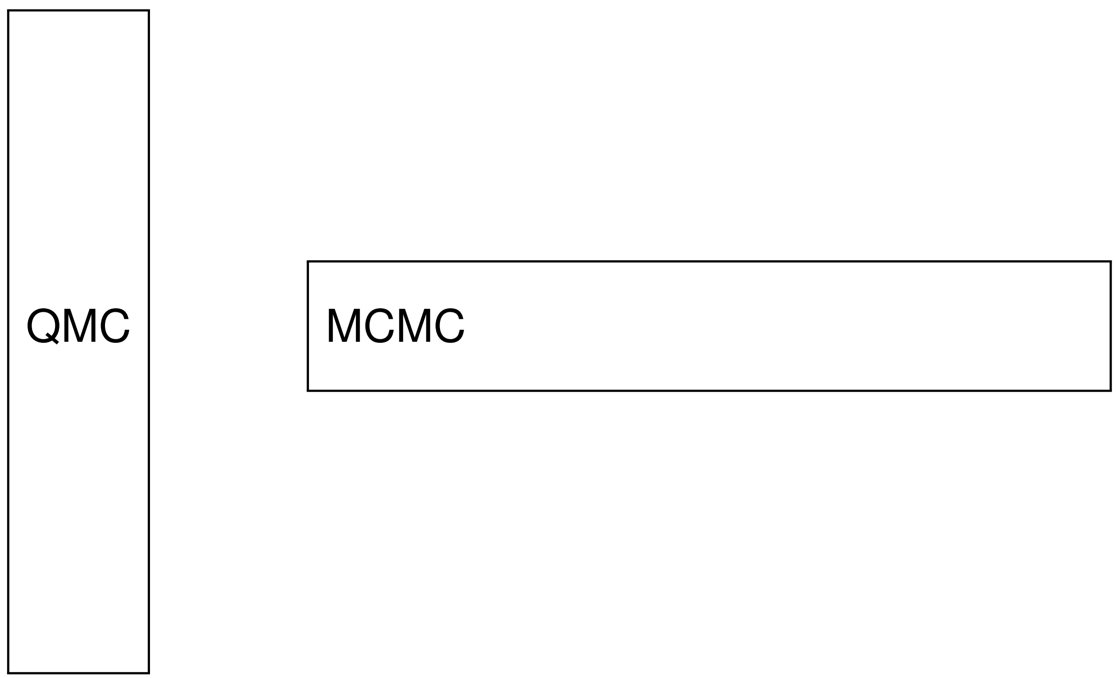
Recipe ctd

$$\underbrace{u_1, u_2, \ldots, u_d}_{\boldsymbol{v}_1} \; \underbrace{u_{d+1}, u_{d+2}, \ldots, u_{2d}}_{\boldsymbol{v}_2} \cdots \underbrace{u_{(n-1)d+1}, u_{(n-1)d+2}, \ldots, u_{nd}}_{\boldsymbol{v}_n}$$

We will replace IID u_i by ‘balanced’ inputs.

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	$1 \leqslant r \ll n \to \infty$

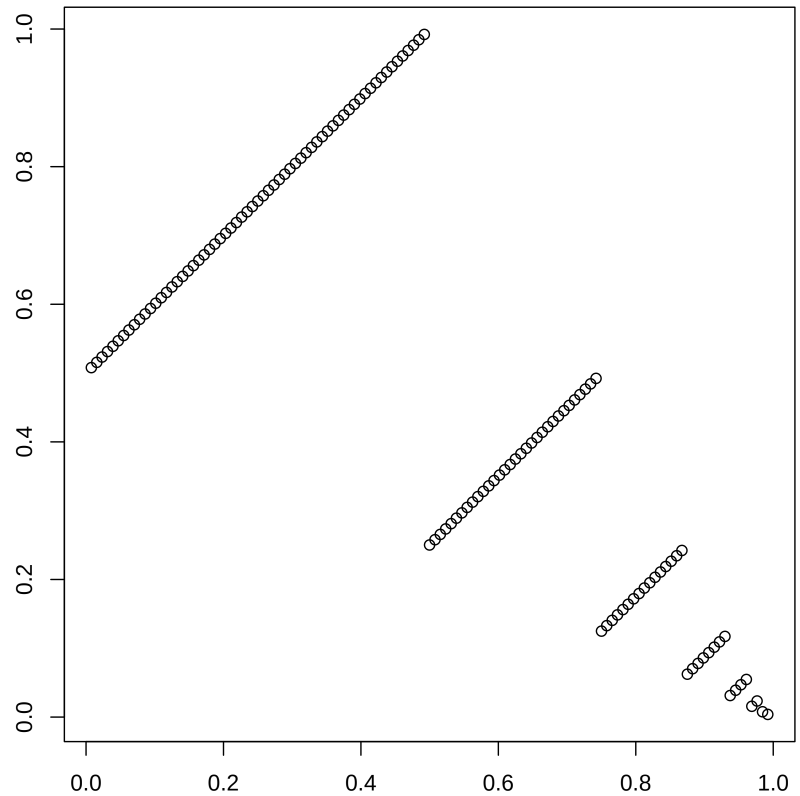


QMC based on equidistribution MCMC based on ergodicity

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

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

$\boldsymbol{z}_i = (u_i, \dots, u_{i+k-1})$ (k-tuples)

For all $d \geqslant 1$

Overlapping blocks

$$
\begin{aligned}
\boldsymbol{z}_1 &= (u_1, \dots, u_k)\\
\boldsymbol{z}_2 &= (u_2, \dots, u_{k+1})\\
\vdots\ \ & \quad \vdots\\
\boldsymbol{z}_n &= (u_n, \dots, u_{n+k-1})
\end{aligned}
$$

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

$$\boldsymbol{v}_i = \big(u_{d(i-1)_1}, \dots, u_{ki}\big)\ \forall k$$

CUD ctd

Originates with Korobov (1950)

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)

Weakly CUD

For random $u_1, u_2, u_3, \dots \in (0,1)$, let

$$\boldsymbol{z}_i = (u_i, u_{i+1}, \dots, u_{i+k-1}) \in (0,1)^k$$

They are weakly CUD if

$$D_n^{k*}(\boldsymbol{z}_1, \dots, \boldsymbol{z}_n) \xrightarrow{\mathrm{d}} 0 \quad \text{for all } k$$

Construction of Liao (1989)

1) Take QMC points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in (0,1)^d$
2) Put in random order $\boldsymbol{x}_{\pi(1)}, \dots, \boldsymbol{x}_{\pi(n)}$
3) Concatenate to get $u_1, \dots, u_{n\times d}$
4) Let $n \to \infty$

More constructions

Tribble (2007) wrote some tiny RNGs and used rotation modulo 1

Chen, Matsumoto, Nishimura & O (2012)

made small linear feedback shift register RNGs

Equidistribution like “small Mersenne twisters”

but not necessarily the same constructions.

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

$$u_1, u_2, \dots, u_M$$

Prepend one or more 0s:

$$0, \dots, 0, u_1, \dots, u_M$$

put into a matrix and apply random rotations mod 1

Software

It would be nice to embed CUD into Stan or JAGS or BUGS etc.

Then users can try lots of examples switching between IID and CUD.

For best results, remove acceptance-rejection where possible.

It seems like a big engineering task.

We have done ‘hand-tuned’ examples.

QMC $\cap$ MCMC

Early references

Chentsov (1967)

Plugs in CUD points.

Samples in finite state space by inversion.

Shows consistency.

Uses very nice coupling argument.

Sobol' (1974)

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

Samples from a row until a return to start state

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

Chentsov's Theorem 1

Law of large numbers via CUD

1) $K < \infty$ states, and,

2) For all $x, y \in \Omega, \quad P(x \to y) > 0$

3) u_i are CUD, and,

4) x_0 is arbitrary

5) $x_i \leftarrow \phi(x_{i-1}, u_i)$ by inversion, $u_i \in (0,1)$, then

$$\hat{p}_n(\omega_k) \equiv \frac{1}{n} \sum_{i=1}^{n} 1_{x_i = \omega_k} \longrightarrow p(\omega_k)$$

and so $\hat{\mu}_n \to \mu$

Remember it was 1967

Chentsov

Chentsov's paper is remarkable and well worth reading after $60+$ years. He wrote before Hastings generalized the Metropolis algorithm and before exact sampling methods were developed for MCMC. The impact of his paper was perhaps limited by studying finite state chains whose transitions can be sampled by inversion.

Chentsov's coupling argument has an intriguing feature. He couples the evolving chain to itself in a particularly elegant way that sets up a 3ϵ argument. The details are in his paper, also in Chen, Dick and O (2011) where it is embedded in the 'Rosenblatt-Chentsov' transformation.

Metropolis

O & Tribble (2004) use $K < \infty$ states and Metropolis-Hastings sampling

$d-1$ variables to propose and 1 to accept or reject:

$$x_{i+1} \leftarrow \phi(x_i, \boldsymbol{v}_{i+1}), \quad \boldsymbol{v}_{i+1} \in (0,1)^d$$

Proposal Ψ, acceptance A

$$y_{i+1} \leftarrow \Psi(x_i, \boldsymbol{v}_{i,1:d-1})$$

$$x_{i+1} \leftarrow \begin{cases} y_{i+1}, & \boldsymbol{v}_{i,d} \leqslant A(x_i \to y_{i+1}) \\ x_i, & \text{else.} \end{cases}$$

Regular proposals

Recall

Set $A \subset \mathbb{R}^d$ is **Jordan** measureable if indicator 1_A is **Riemann** integrable

Proposals are **regular** if

$$S_{\omega_k \to \omega_\ell} = \left\{(u_1, \dots, u_{d-1}) \in [0,1]^{d-1} \mid \Psi(\omega_k, u_1, \dots, u_{d-1}) = \omega_\ell\right\}$$

is Jordan measurable all k, ℓ

Regular proposals in $[0,1]^{d-1}$ give

1) Regular (one step) transition $x_i \to x_{i+1}$ sets in $[0,1]^d$
2) Regular path $x_i \to x_{i+1} \to \cdots \to x_{i+k}$ sets in $[0,1]^{dk}$
3) Regular multi-step transitions $x_i \to x_{i+k}$ sets in $[0,1]^{dk}$

Home state

A set $\mathcal{B}_\omega = \prod_{j=1}^d (a_j, b_j) \subset (0,1)^d$ such that

$$\boldsymbol{v}_i \in \mathcal{B}_\omega \implies x_i = \phi(x_{i-1}, \boldsymbol{v}_i) = \omega$$

A (very) small set $\{\omega\}$. $\quad \omega \in \{1, 2, \dots, K\}$.

Coupling

Wherever you are, the chance of going to ω next would be positive,

for random $\boldsymbol{v} \sim \mathbf{U}(0,1)^d$.

Our $\boldsymbol{v}_i$ can be deterministic.

Theorem

For Metropolis-Hastings sampling, if

1) There are $K < \infty$ states,
2) u_i are CUD,
3) x_0 is arbitrary,
4) y_{i+1} is a regular proposal, and
5) there is a home state ω with $\mathbf{vol}(\mathcal{B}_\omega) > 0$, then

$$\hat{p}_n(\omega_k) \equiv \frac{1}{n}\sum_{i=1}^{n} 1_{x_i=\omega_k} \longrightarrow p(\omega_k)$$

Theorem 2, O & Tribble (2004)

Idea of proof

Compare x_{i+m} to $x_{i,m,m}$ where $x_{i,m,0}$ is sampled by inversion of π using u_{id} and the transitions $x_{i,m,t} \to x_{i,m,t+1}$ use Metropolis-Hastings with the same rule that x_i uses.

For large m, $\widetilde{x}_{i,m,m}$ is usually x_{i+m}. Also $\widetilde{x}_{i,m,m} \sim p$.

Coupling

$$x_1 \to x_2 \to \cdots \to x_i \to x_{i+1} \to x_{i+2} \to \cdots \to x_{i+m}$$

$\downarrow$ inversion "$\pi^{-1}(u_{id})$"

$$x_{i,m,0} \to x_{i,m,1} \to x_{i,m,2} \cdots \to x_{i,m,m}$$

3 epsilon

$$\left|\frac{1}{n}\sum_{i=1}^{n} \pi(\omega) - 1\{x_{i,m,m} = \omega\}\right| + \left|\frac{1}{n}\sum_{i=1}^{n} 1\{x_{i,m,m} = \omega\} - 1\{x_{i+m} = \omega\}\right|$$

$$+ \left|\frac{1}{n}\sum_{i=1}^{n} 1\{x_{i+m} = \omega\} - 1\{x_i = \omega\}\right|$$

Weakly CUD

For Metropolis-Hastings sampling, if

1) There are $K < \infty$ states,
2) u_i are Weakly CUD,
3) x_0 is arbitrary,
4) y_{i+1} is a regular proposal, and
5) IID sampling would have worked

$$\hat{p}_n(\omega_k) \equiv \frac{1}{n}\sum_{i=1}^{n} 1_{x_i=\omega_k} \xrightarrow{\mathrm{d}} p(\omega_k)$$

Theorem 3, O & Tribble (2004)

Some additional ref.s

QMC in multiple-try Metropolis	Craiu & Lemieux (2007)
QMC in exact sampling	Lemieux & Sidorsky (2006)

Related

Reordering heat particles	Lécot (1989)
MCMC $\cap$ antithetics	Frigessi, Gäsemyr, Rue (2000)
MCMC $\cap$ Latin hypercubes	Craiu, Meng (2004)
array-RQMC	L'Ecuyer, Lécot, Tuffin (2004)
array-RQMC	L'Ecuyer, Lécot, L'Archevêque-Gaudet (2008)
Rotor-Router	Propp (2004)
Quasi-random walks on balls	Karaivanova, Chi, Gurov (2007)
Rao-Blackwellized MH	Douc, Robert (2009)

Results from Tribble

Variance reduction factors from Tribble (2007) for two Gibbs sampling problems.

Pumps: hierarchical Poisson-Gamma model.

Vasorestriction: probit model 3 coefficients, 39 latent variables.

	$n = 2^{10}$		$n = 2^{12}$		$n = 2^{14}$	
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

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

Randomized CUD sequence versus IID sampling.

See Tribble (2007) for simulation details.

Targets are posterior means of parameters.

Mackey & Gorham

Continuous state spaces

Tribble's best results were for a smooth setting: continuous state space and the Gibbs sampler, which has no accept-reject component.

This makes sense: QMC wins its biggest improvements on smooth functions

The consistency results in O & Tribble (2005) for $\hat{\mu}_n \to \mu$ were in discrete state spaces, where only small improvements are seen empirically.

Continuous cases

Chen, Dick & O (2011) 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)

Convergence rates

Thesis of Su Chen (2011)

Sometimes we see a better convergence rate.

Conditions for error $O(n^{1-\delta})$, all $\delta > 0$

1) Strong contracting mapping

$$\|\psi(\boldsymbol{x},\boldsymbol{v}) - \psi(\boldsymbol{x}',\boldsymbol{v})\| \leqslant \alpha\|\boldsymbol{x}-\boldsymbol{x}'\|, \quad \text{some } 0 \leqslant \alpha < 1$$

2) Bounded set Ω for $\boldsymbol{x}$
3) $f(\boldsymbol{x})$ Lipschitz continuous
4) ψ infinitely differentiable
5) Irreducible Harris convergent chain
6) Decay conditions

Conditions satisfied for some ARMA models and Sobol' sequence inputs.

Warwick thinker

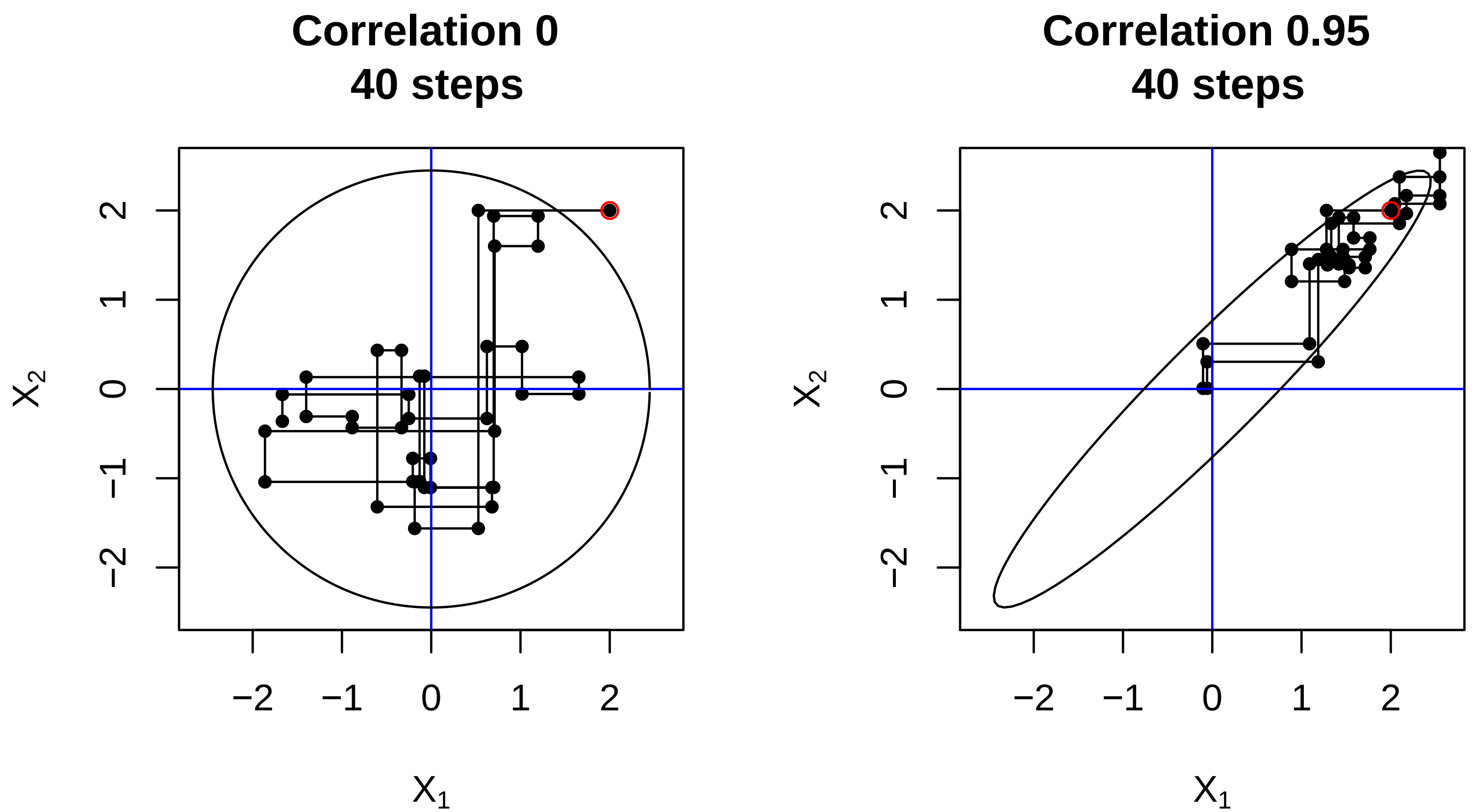
Gaussian Gibbs sampler

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \in \mathbb{R}^2$$

Alternate

$$X_1 \sim \mathsf{DIST}(X_1 \mid X_2 = x_2) = \mathcal{N}(\rho x_2, 1 - \rho^2)$$
$$X_2 \sim \mathsf{DIST}(X_2 \mid X_1 = x_1) = \mathcal{N}(\rho x_1, 1 - \rho^2)$$

Gaussian Gibbs sampler

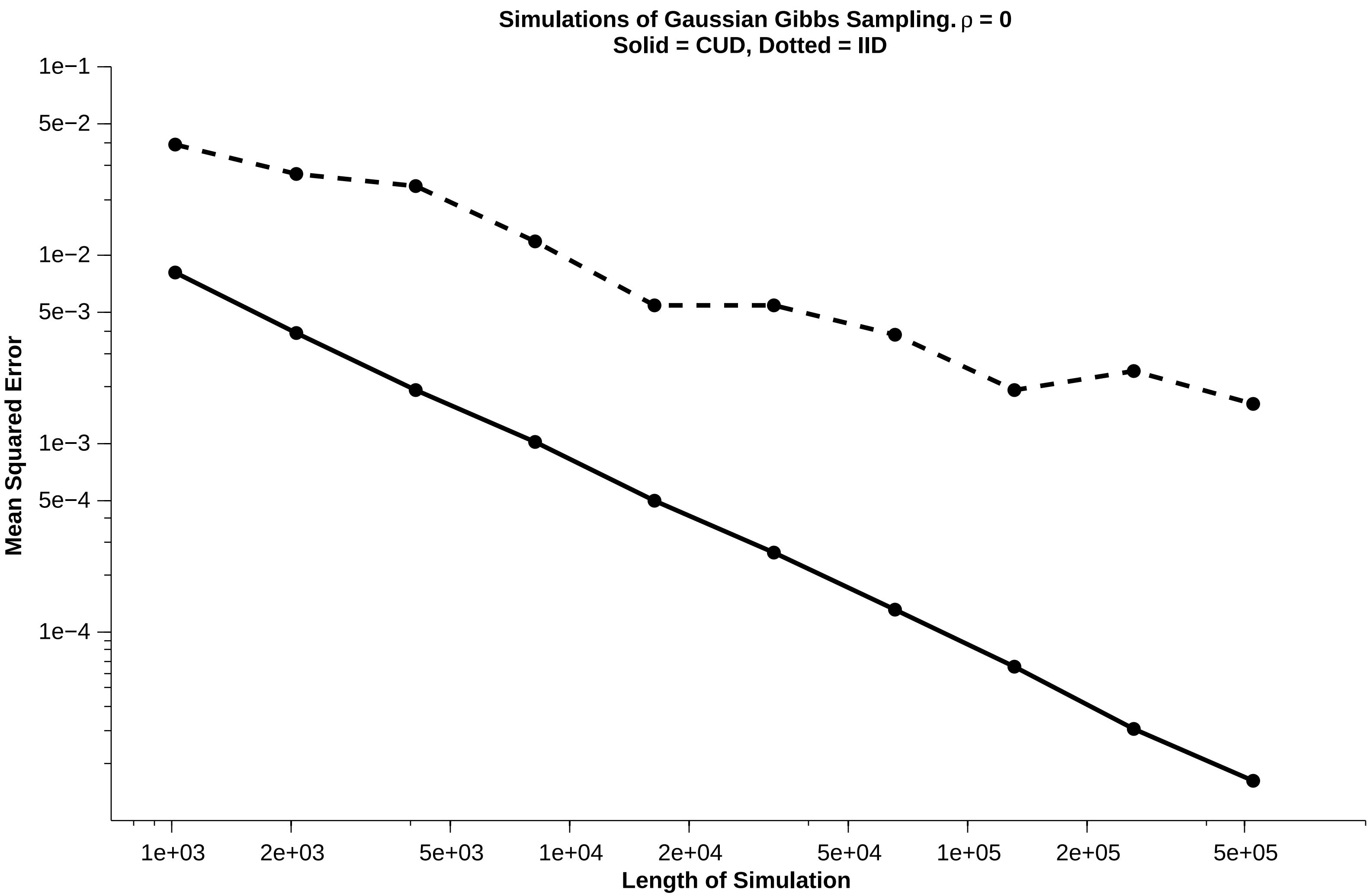


Sampling, $i = 1, \dots, n$

$$X_{i1} \leftarrow \rho X_{i-1,2} + \sqrt{1-\rho^2}\,\Phi^{-1}(u_{2i-1})$$

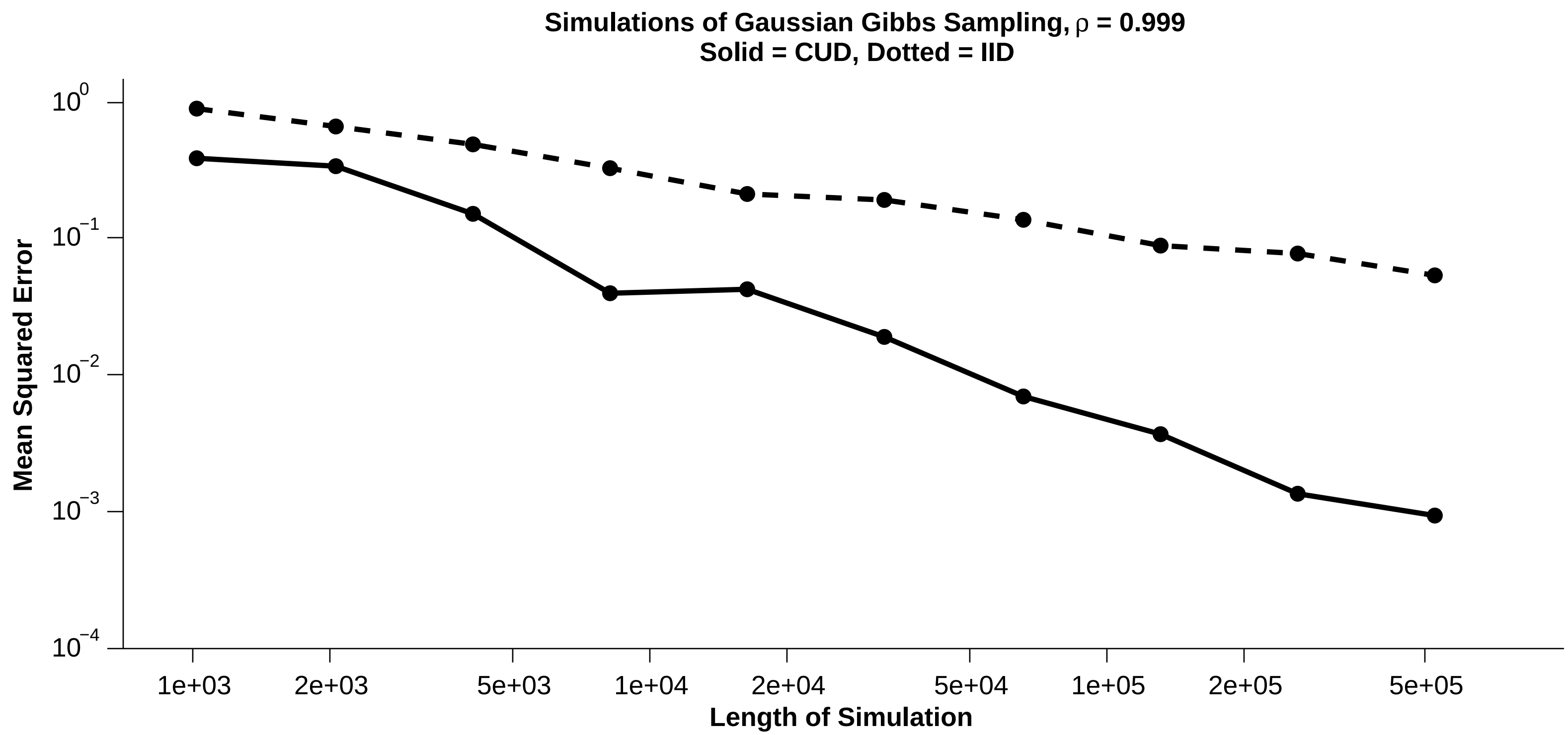
$$X_{i2} \leftarrow \rho X_{i1} + \sqrt{1-\rho^2}\,\Phi^{-1}(u_{2i})$$

Gaussian Gibbs $\rho = 0$



Estimate $\mathbb{E}(X)$ start at $(1, 1)$

Gaussian Gibbs $\rho = 0.999$



Estimate $\mathbb{E}(X)$ start at $(1, 1)$

$\therefore$ models like AR(1) are promising

M/M/1 queue initial transient

Exponential arrivals at rate $\rho = 0.9$ and service times at rate 1

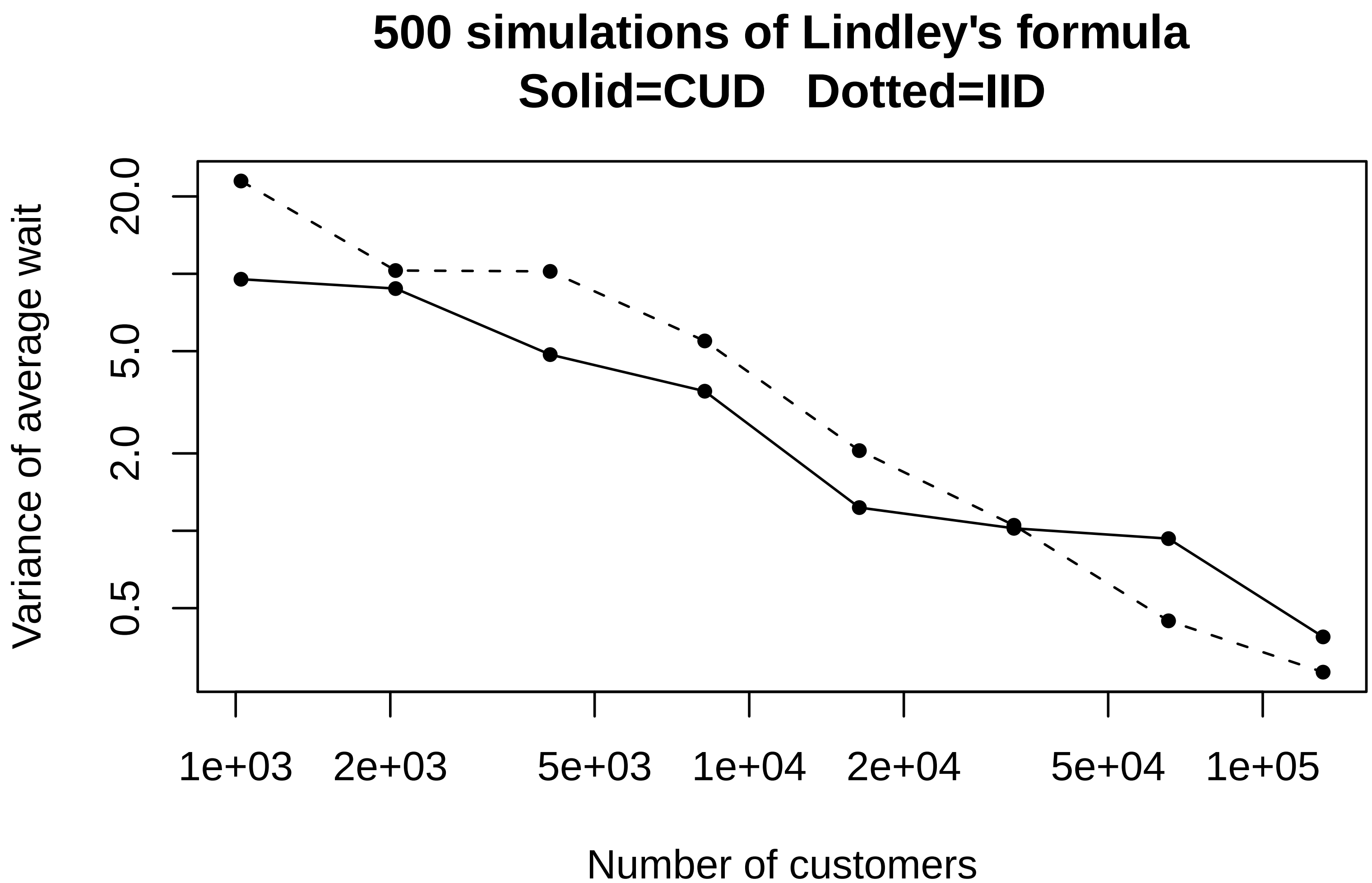
Customer $i \geqslant 1$ has arrival time A_i, the service time S_i, and waiting time W_i, where

$$\begin{aligned}
A_0 &= 0 \\
A_i &= A_{i-1} - \log(1 - u_{2i-1})/\rho \\
S_i &= -\log(1 - u_{2i}) \\
W_1 &= 0 \\
W_i &= \bigl(W_{i-1} + S_{i-1} - A_i\bigr)_+ \qquad \text{(Lindley)}
\end{aligned}$$

Average wait of first n customers is

$$\overline{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i \quad \text{we simulate for} \quad \mathbb{E}(\overline{W}_n)$$

Variance of average wait



Variance reductions

Chen, Matsumoto, Nishimura, O (2012)

Antithetic

$$u_1, u_2, \cdots, u_n, 1-u_1, 1-u_2, \cdots, 1-u_n$$

Round trip

$$u_1, u_2, \cdots, u_n, 1-u_n, 1-u_{n-1}, \cdots, 1-u_1$$

1) Preserves CUD structure ($\approx$ no harm)

2) Sometimes big gains vs plain CUD, sometimes none

3) Can also reverse d-tuples

Summary

Bivariate Gaussian	apparent better convergence rate for mean
Bivariate Gaussian	not much improvement for discrepancy
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.

Improvements range from modest to powerful.

Same as for finite dimensional QMC.

The latest

Tobias Schwedes & Ben Calderhead (2018) on arXiv

and at MCQMC 2018 in Rennes.

Multi-proposal MCMC. Like Craiu & Lemieux (2007)

Extend MP-MCMC of Calderhead (2014).

1) Burn in
2) 511 iterations
3) $N \to \infty$ proposals per iteration
4) Reweight them, and then pick one
5) Using QMC $\cap$ CUD gets empirical error $O(N^{-1})$ (Bayesian logistic regression)

It has MCMC, particles, importance sampling, adaptation, QMC, MALA . . .

Thanks

- Lecturers: Nicolas Chopin, Mark Huber, Jeffrey Rosenthal
- Guest speakers: Michael Giles, Gareth Roberts
- The London Mathematical Society: Elizabeth Fisher, Iain Stewart
- CRISM & The University of Warwick, Statistics
- Sponsors: Amazon, Google
- Partners: ISBA, MCQMC, BAYSM
- Poster: Talissa Gasser, Hidamari Design
- NSF: DMS-1407397 & DMS-1521145
- Planners: Murray Pollock, Christian Robert, Gareth Roberts
- Support: Paula Matthews, Murray Pollock, Shahin Tavakoli