QMC for MCMC:
Background and recent results

Art B. Owen
Stanford University

Su Chen
Stanford University

With contributions from:

Josef Dick, Makoto Matsumoto, Takuji Nishimura

Simple Monte Carlo

Used in virtually all sciences

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

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

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) Sometimes we can, using QMC points that are "completely uniformly distributed" (CUD)

3) Previously:

   (a) convergence required a finite state space, but

   (b) the largest improvements were for continuous examples

4) Now:

   (a) proven consistency on continuous state spaces

   (b) more examples (some good, some disappointing)

Markov chain Monte Carlo

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

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

LLN for reasonable conditions

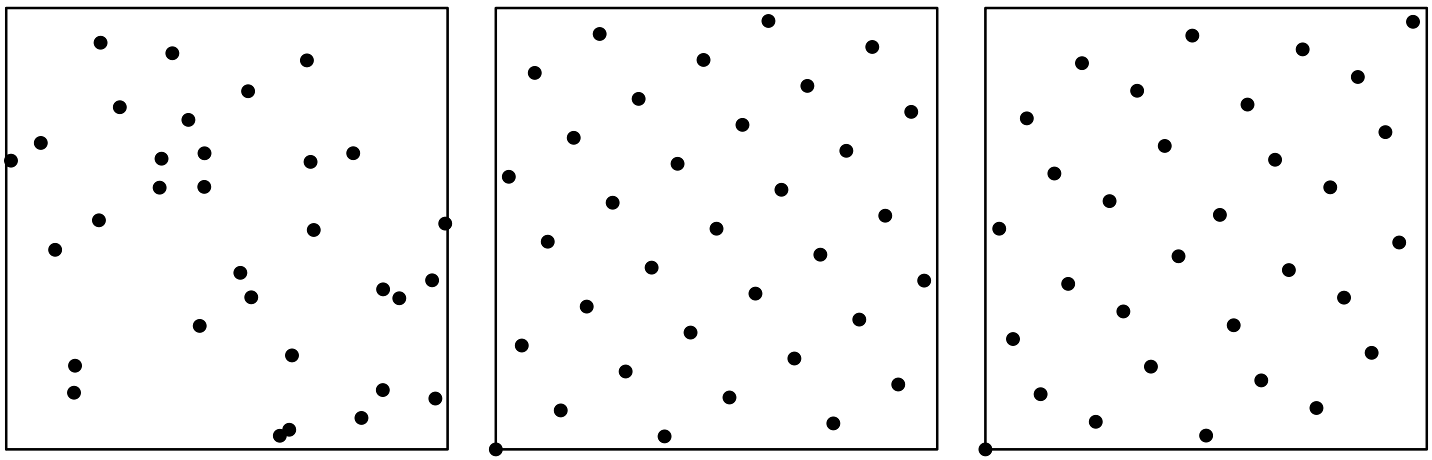
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) \to \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \equiv \mu$$

Main MCMC algorithms

- Metropolis-Hastings
  1) While at $\boldsymbol{x}_i$ propose move to $\boldsymbol{y}_{i+1} \sim Q(\boldsymbol{x}_i \to \boldsymbol{y}_{i+1})$
  2) Accept with probability $A(\boldsymbol{x}_i \to \boldsymbol{y}_{i+1})$
  3) If accepted, then $\boldsymbol{x}_{i+1} = \boldsymbol{y}_{i+1}$ else $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i$
- Gibbs sampler
  1) Sample component j from $p(x_j \mid x_{ik},\ k \neq j)$
  2) Cycle through j's (sequentially, or randomly)

Quasi-Monte Carlo

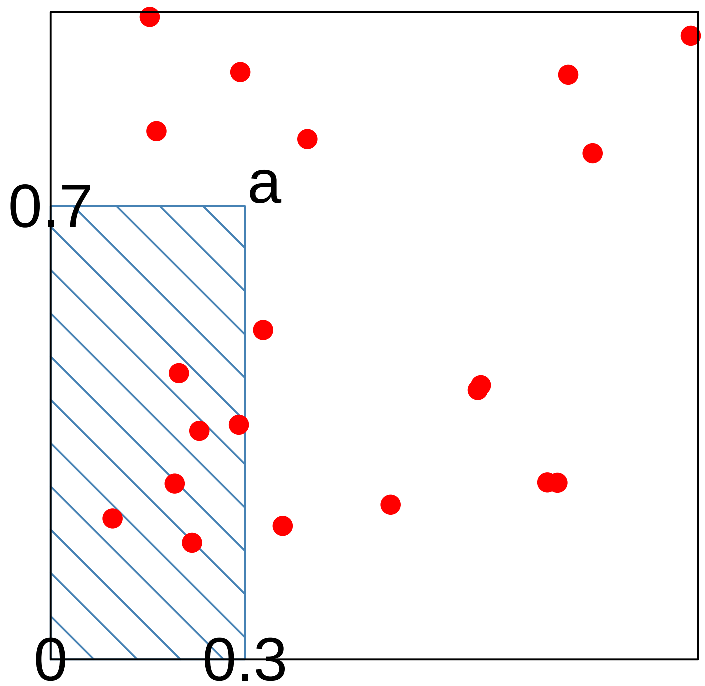
MC and two QMC methods in the unit square



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.

The local discrepancy is

$$\delta(a) = \widehat{F}[0, 1) - F[0, 1) = \frac{6}{20} - .3 \times .7 = .09$$

Here F is $\mathbf{U}[0, 1]^2$ and $\widehat{F}$ is $\mathbf{U}\{x_1, \ldots, x_n\}$

Star discrepancy

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

For $d = 1$ D_n^* is Kolmogorov-Smirnov distance between $\mathbf{U}\{x_1, \dots, x_n\}$ and $\mathbf{U}[0,1)$

Koksma-Hlawka inequality

$$|\hat{\mu}_n - \mu| \leqslant D_n^*(x_1, \dots, x_n) \times \|f\|_{\mathrm{HK}}$$

$\|f\|_{\mathrm{HK}}$ is total variation (Hardy & Krause)

$\exists$ points with $D_n^* = O(n^{-1+\epsilon})$, $\forall \epsilon > 0$

Hence QMC can attain $O(n^{-1+\epsilon})$ error vs $O_p(n^{-1/2})$ for MC.

van der Corput

i	1	2	3	4	5	6	7	$\cdots$
i base 2	1.0	10.0	11.0	100.0	101.0	110.0	111.0	$\cdots$
$\phi_2(i)$ base 2	0.1	0.01	0.11	0.001	0.101	0.011	0.111	$\cdots$
x_i	1/2	1/4	3/4	1/8	5/8	3/8	7/8	$\cdots$

$$
\begin{aligned}
i &= \cdots a_3 a_2 a_2 a_1 \quad \text{base } b \\
\phi_b(i) &= 0.a_1 a_2 a_3 \cdots \quad \text{base } b \\
D_n^*(x_1, \dots, x_n) &= O(\log(n)/n)
\end{aligned}
$$

Halton sequence

$$
x_i = (\phi_{p_1}(i), \phi_{p_2}(i), \dots, \phi_{p_d}(i)) \in [0,1]^d
$$

p_j relatively prime, usually first d primes

Recipe for QMC in MCMC

1) Each step takes d numbers in $(0, 1)$.

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

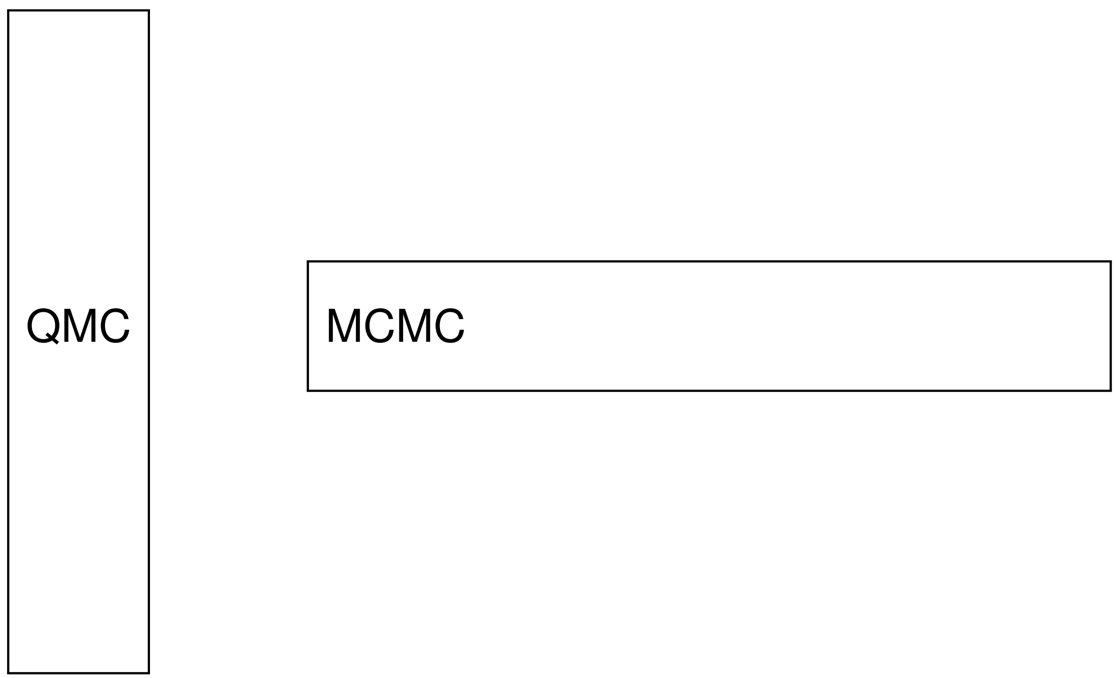
Reasons for caution

1) We're using 1 point in $[0, 1]^{nd}$ with $n \to \infty$

2) Our sequence won't be Markovian

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

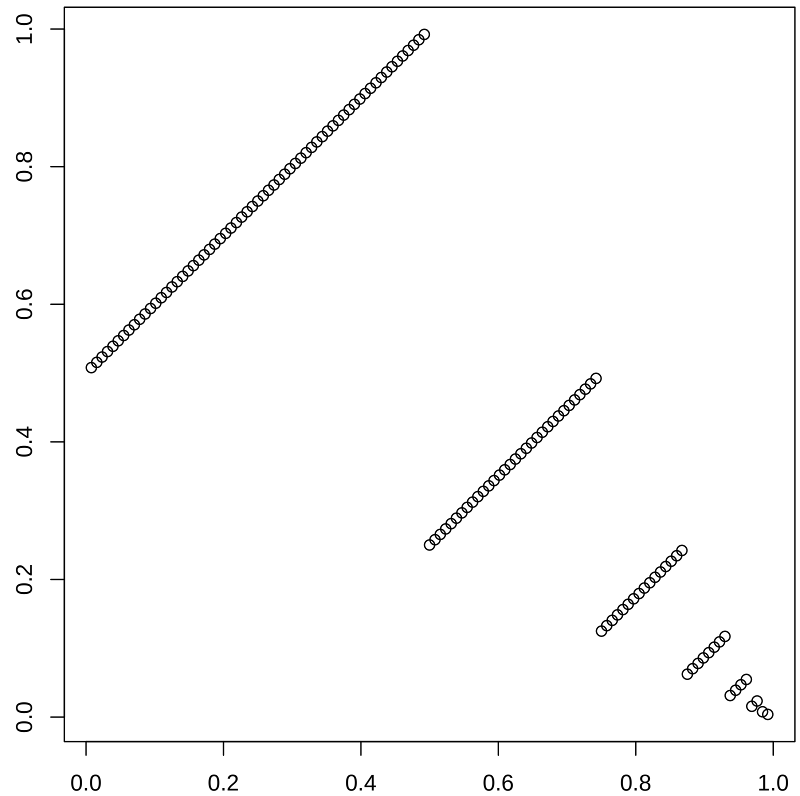


QMC based on equidistribution   MCMC based on ergodicity

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

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

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

For all $d \geqslant 1$

Overlapping blocks

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

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

$$\boldsymbol{v}_i = (u_{d(i-1)_1}, \dots, u_{di}) \;\forall d$$

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)

QMC ∩ 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

Recent QMC ∩ MCMC

	Liao (1998)	reorders QMC points
	Chaudary (2004)	QMC wts on rejected proposals
√	O & Tribble (2005)	CUD pts in Metropolis, finite state space
	Tribble & O (2008)	constructions for weakly CUD pts
√	Tribble (2007)	theory and examples
	Craiu & Lemieux (2007)	QMC in multiple-try Metropolis
	Lemieux & Sidorsky (2006)	QMC in exact sampling
√	Chen, Dick & O (2010)	CUD pts in cts state spaces

Related ideas

Reordering heat particles	Lécot (1989)
Manually adaptive QMC	Ostland, Yu (1997)
QMC for particle filters	Lemieux, Ormoneit, Fleet (2001), UAI
MCMC ∩ antithetics	Frigessi, Gäsemyr, Rue (2000)
MCMC ∩ Latin hypercubes	Craiu, Meng (2004)
array-RQMC	L'Ecuyer, Lécot, Tuffin (2004)
Rotor-Router	Propp (2004)
Quasi-random walks on balls	Karaivanova, Chi, Gurov (2007)
array-RQMC	L'Ecuyer, Lécot, L'Archevêque-Gaudet (2008)
Rao-Blackwellized MH	Douc, Robert (2009)

Results 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

Variance reduction factors from Tribble (2007) for two Gibbs sampling problems. For the pumps data, the greatest and least variance reduction for a randomized CUD sequence versus IID sampling is shown. For the vasorestriction data, greatest and least variance reductions for the three regression parameters are shown. See Tribble (2007) for simulation details.

Targets are posterior means of parameters.
CUD points were LFSR,
with Cranley-Patterson rotations.

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 for $\hat{\mu}_n \to \mu$ were in discrete state spaces, where only small improvements are seen empirically.

Chen, Dick & O Annals Stat.

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

MCMC remains consistent when driven by $u_1, u_2, \dots,$ 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)

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 $10 \leqslant m \leqslant 32$.

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

We explore them for some simulations.

Prepend one or more 0s:

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

put into a matrix and apply Cranley-Patterson rotations

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.

Same as for finite dimensional QMC.

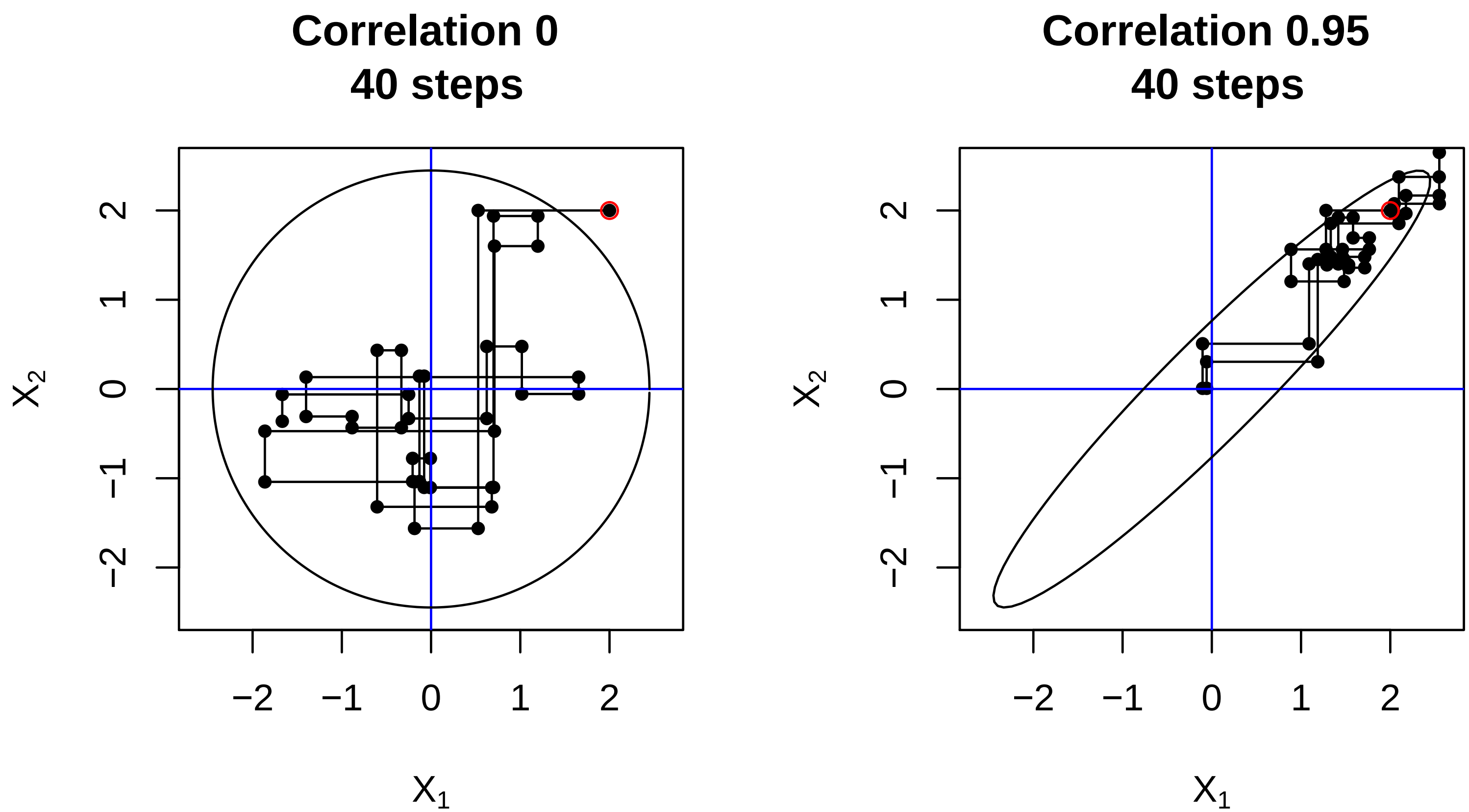
Gaussian Gibbs sampler

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

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

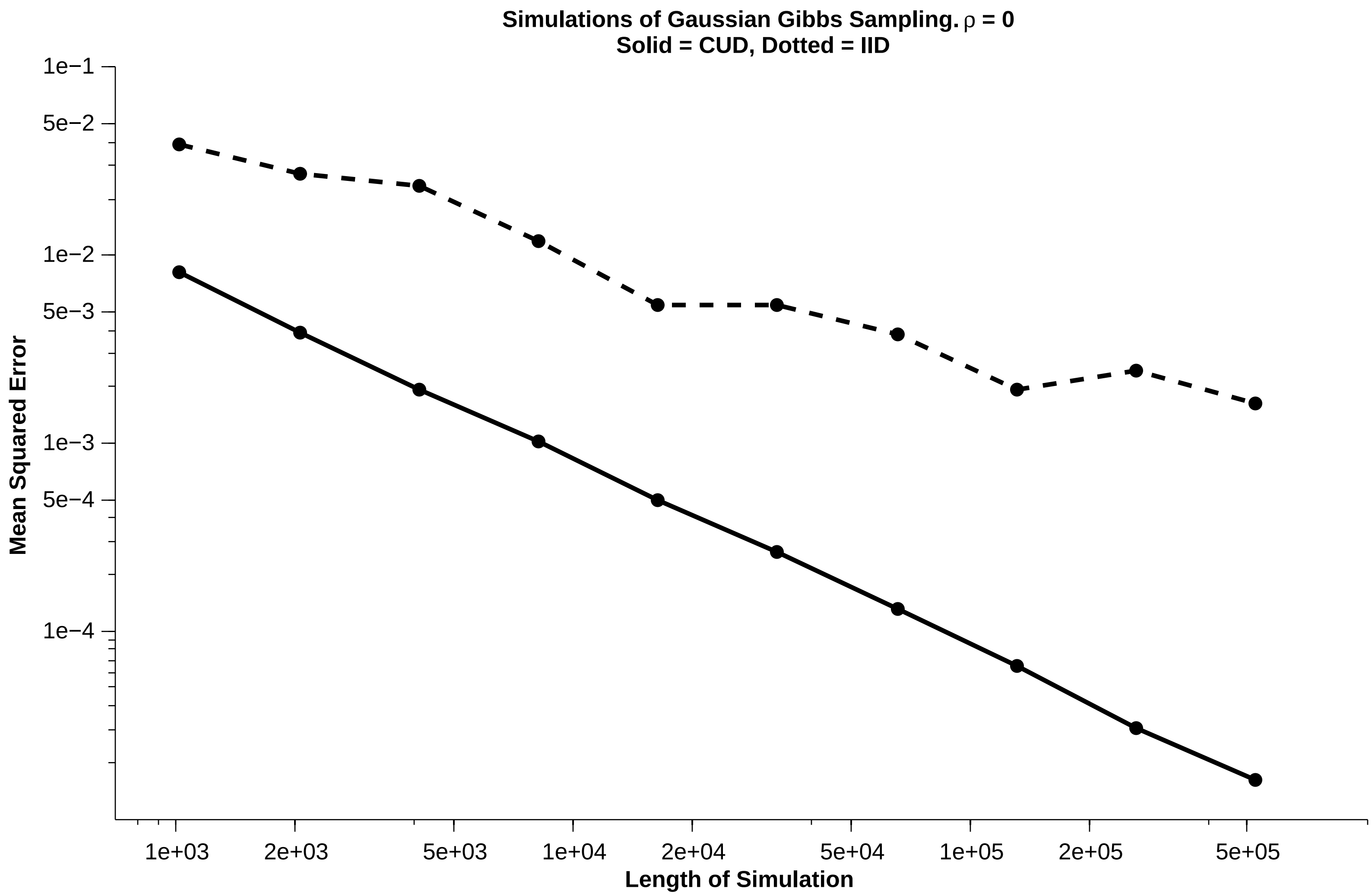


Sampling, $i = 1, \ldots, n$

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

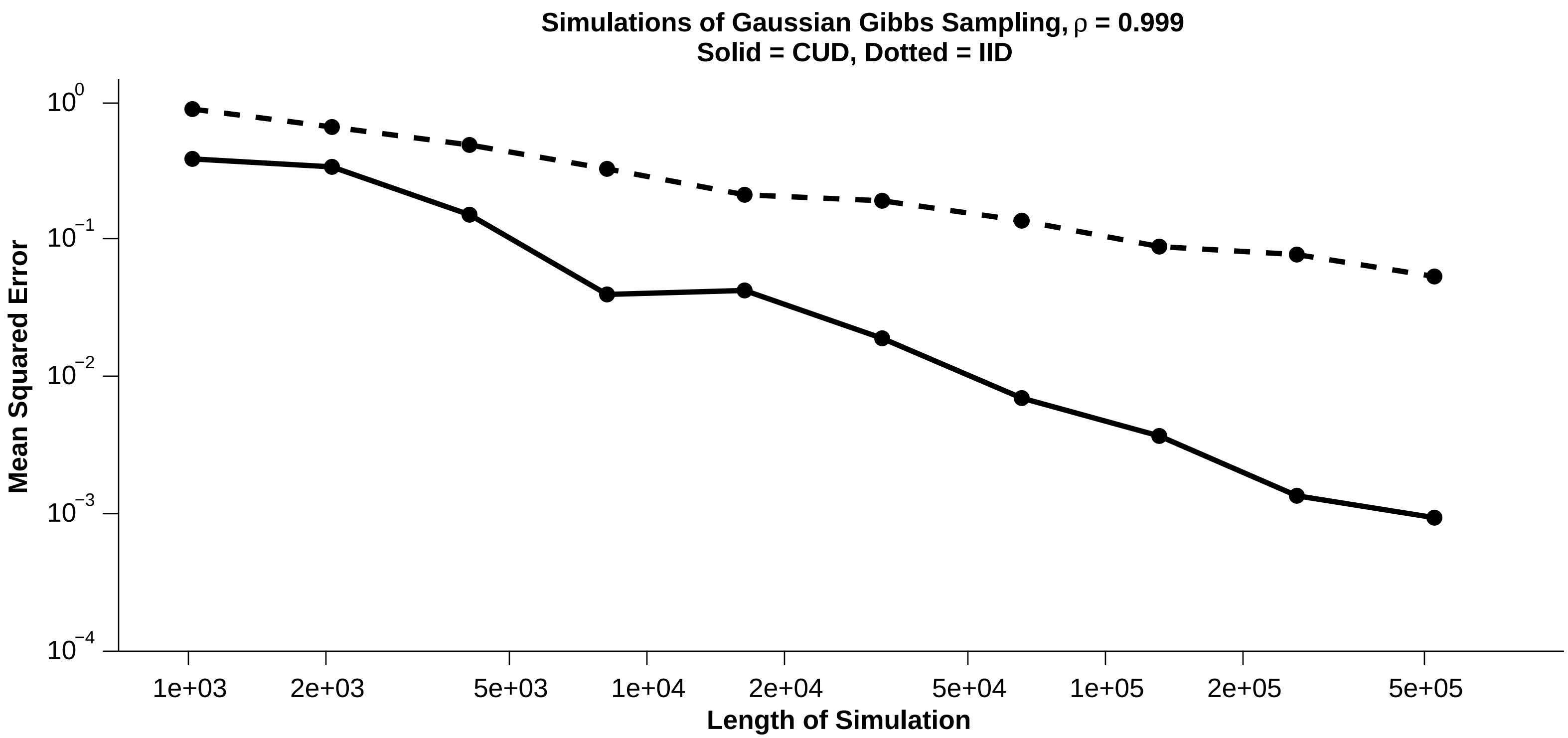
$$X_{i2} \leftarrow \rho X_{i1} + \sqrt{1-\rho^2}\,\Phi^{-1}(u_{2i})$$

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

Hit and run MCMC

The hit and run sampler generates points uniformly inside a convex region R.

Given x_i it picks a random direction θ_i and chooses x_{i+1} at random on

$$R \cap \{x_i + r\theta \mid -\infty < r < \infty\}.$$

You can use it to estimate the ratio of two nested convex regions.

The best known way to estimate volume of high dimensional convex regions uses a cascade of nested regions.

Unfortunately

CUD brought no significant improvement for $\mathbf{vol}(\mathrm{triangle})$

M/M/1 queue initial transient

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

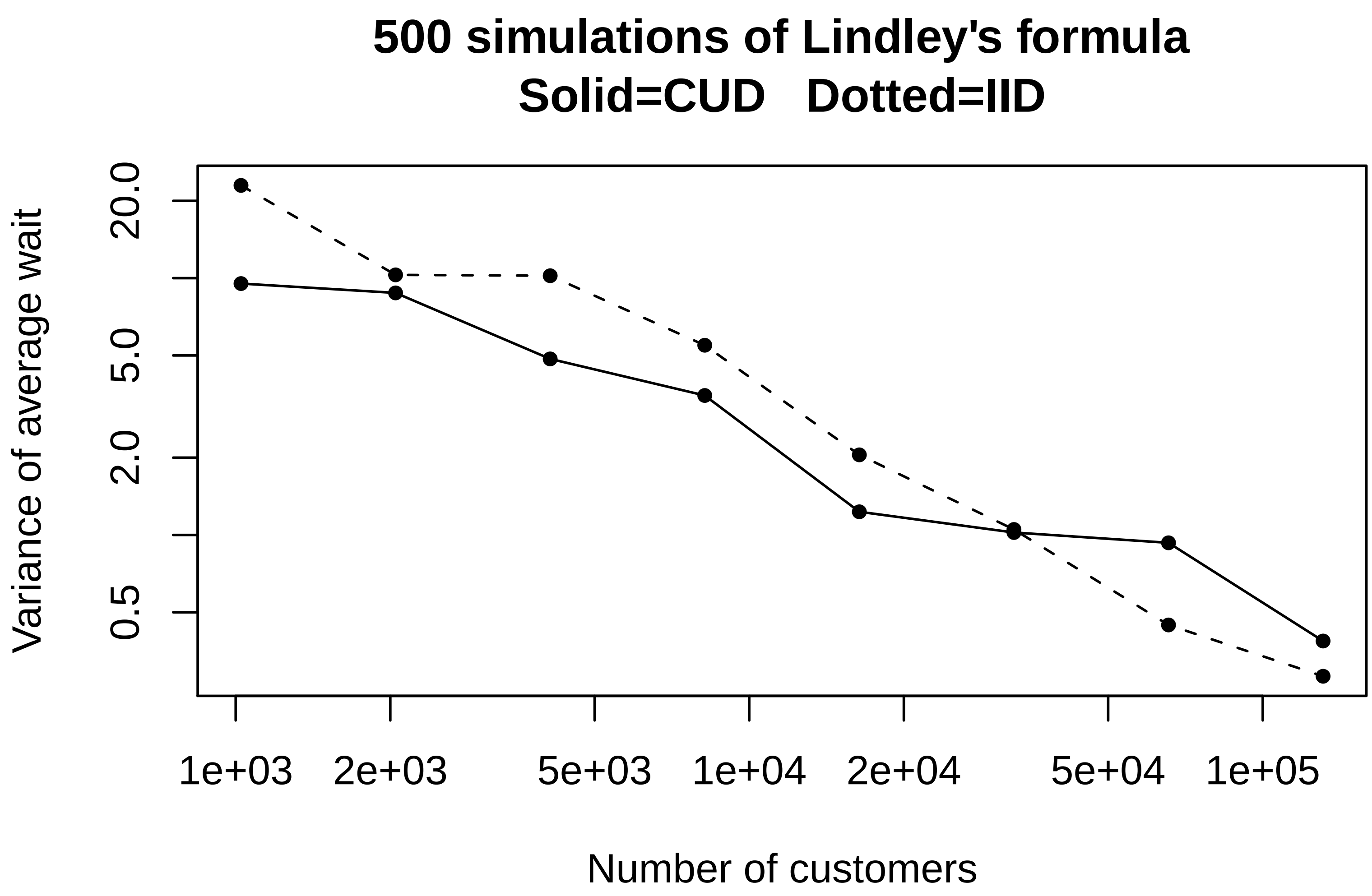
Customer $i \geqslant 1$ has arrival time A_i, the service time S_i, and waiting time W_i, where

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

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



Heston's stochastic volatility

$$\mathrm{d}S = rS\mathrm{d}t + \sqrt{V}S\,\mathrm{d}W_1, \quad 0 < t < T$$

$$\mathrm{d}V = \kappa(\theta - V)\mathrm{d}t + \sigma\sqrt{V}\,\mathrm{d}W_2$$

Parameters from J. Zhu (2008):

For S:	$S(0) = 100$	$r = 0.04$	$T = 6$	
For V:	$V(0) = 0.025$	$\theta = 0.04$	$\kappa = 2$	$\sigma = 0.3$

$$\mathrm{Corr}(\mathrm{d}W_1, \mathrm{d}W_2) \equiv \rho = -0.5$$

Price a European call: expected discounted value of $(S(T) - K)_+$

Heston simulations

Split $[0, T]$ into 2^k intervals $\quad k \in \{8, 10\}$

Use 2^{k+1} CUD numbers per simulation (update price and volatility)

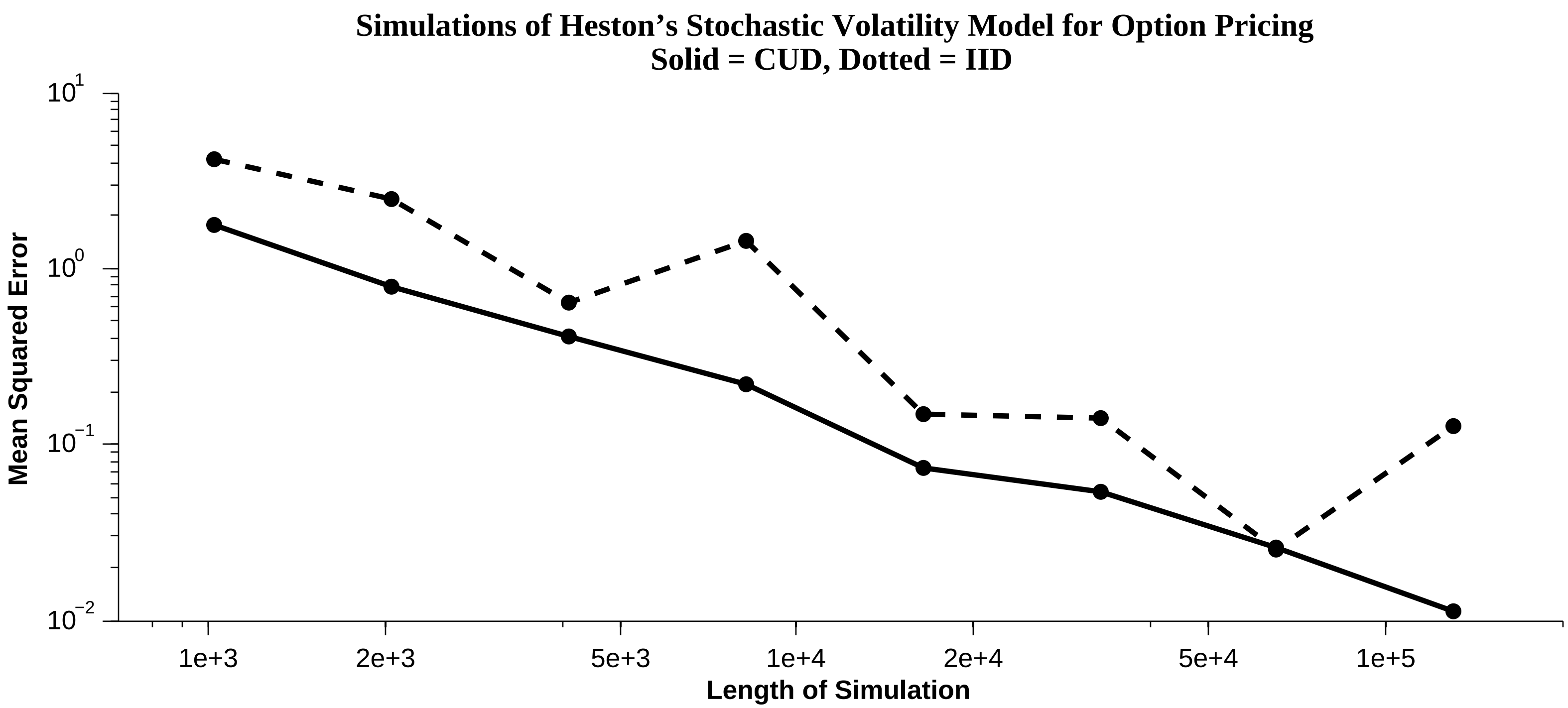
Use 2^{r+k+1} for 2^r simulations $10 \leqslant r \leqslant 17$

Update via 'variance form', not 'volatility form'

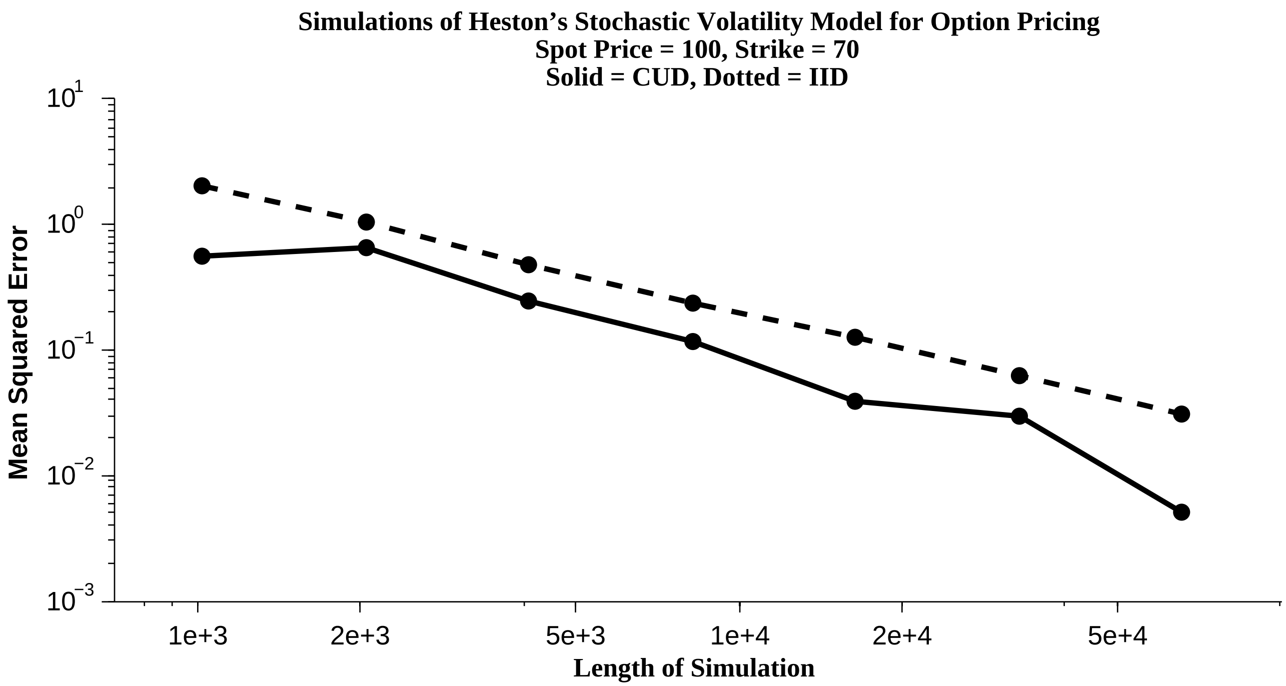
Use 100 rotations (adding $\mathbf{U}(0, 1)^2$)

Compare to exact answer

$K = 100 = S(0), \mathrm{d}t = T/2^8$



$$K = 70 \quad \mathrm{d}t = T/2^8$$



GARCH$(1,1)$ model

$$\log\Bigl(\frac{X_t}{X_{t-1}}\Bigr) = r + \lambda\sqrt{h_t} - \frac{1}{2}h_t + \varepsilon_t, \quad 1 \leqslant t \leqslant T$$
$$\varepsilon_t \sim \mathcal{N}(0, h_t)$$
$$h_t = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \beta_1 h_{t-1}$$

Parameters from J.-C. Duan (1995):

For X_t:	$r = 0$	$\lambda = 7.452 \times 10^{-3}$	$T = 30$
For h_t:	$\alpha_0 = 1.524 \times 10^{-5}$	$\alpha_1 = 0.1883$	$\beta_1 = 0.7162$

h starts at 0.64×0.2413

0.2413 is the stationary variance

European call, strike $K = 1$

Garch simulations

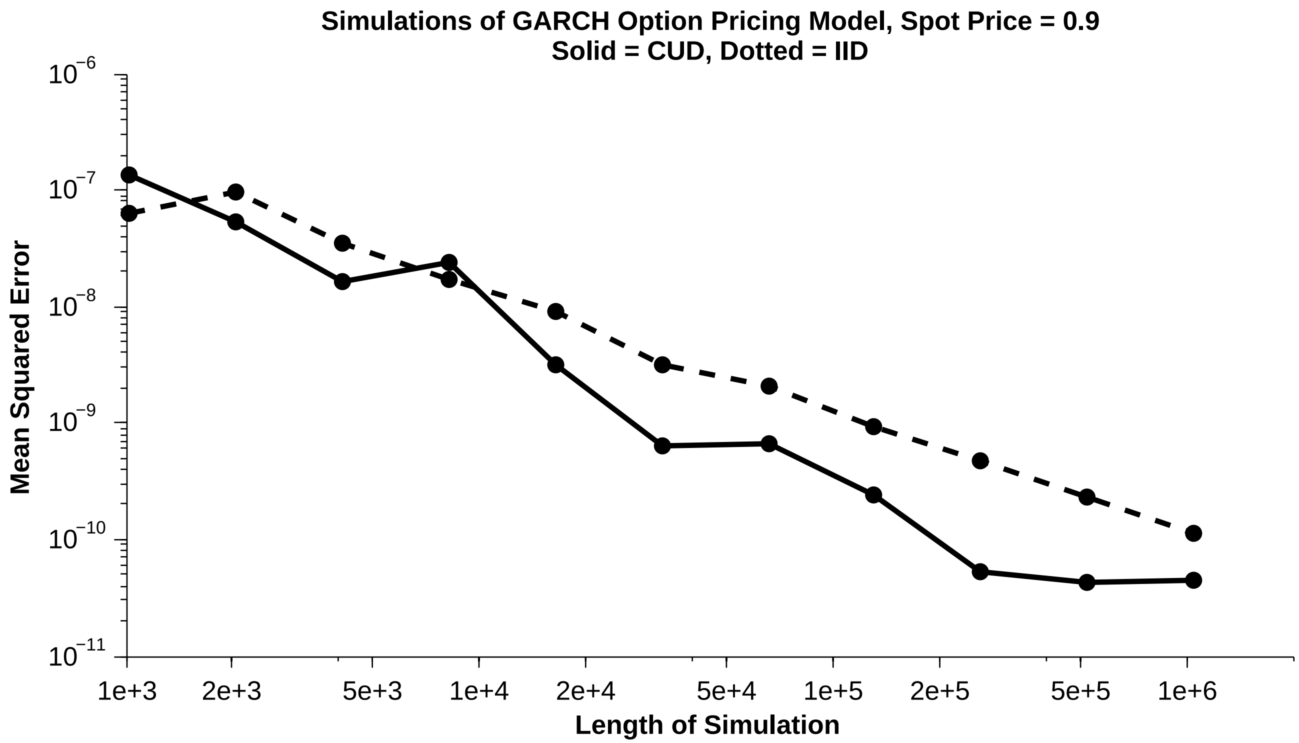
Problem has 30 intervals

Use 30 CUD numbers per simulation (update one price change ε_t)

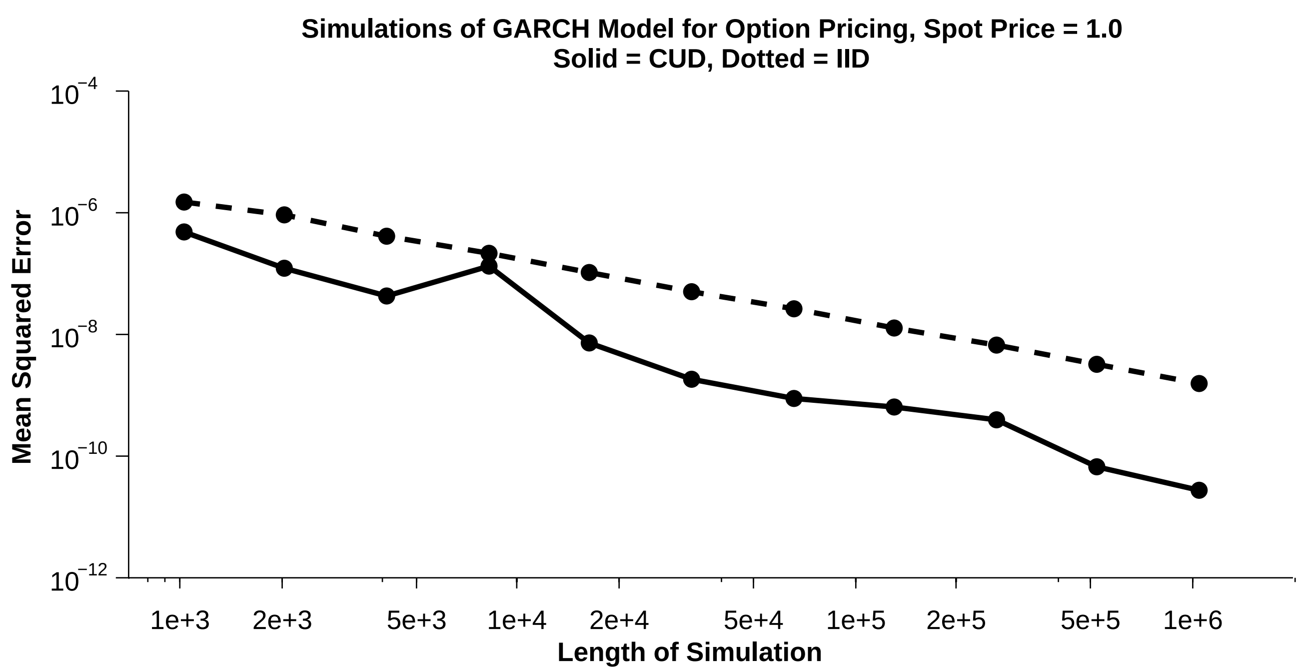
Get $\lfloor 2^m/30 \rfloor$ prices per CUD sequence

Use 100 rotations (adding $\mathbf{U}(0,1)^{30}$)

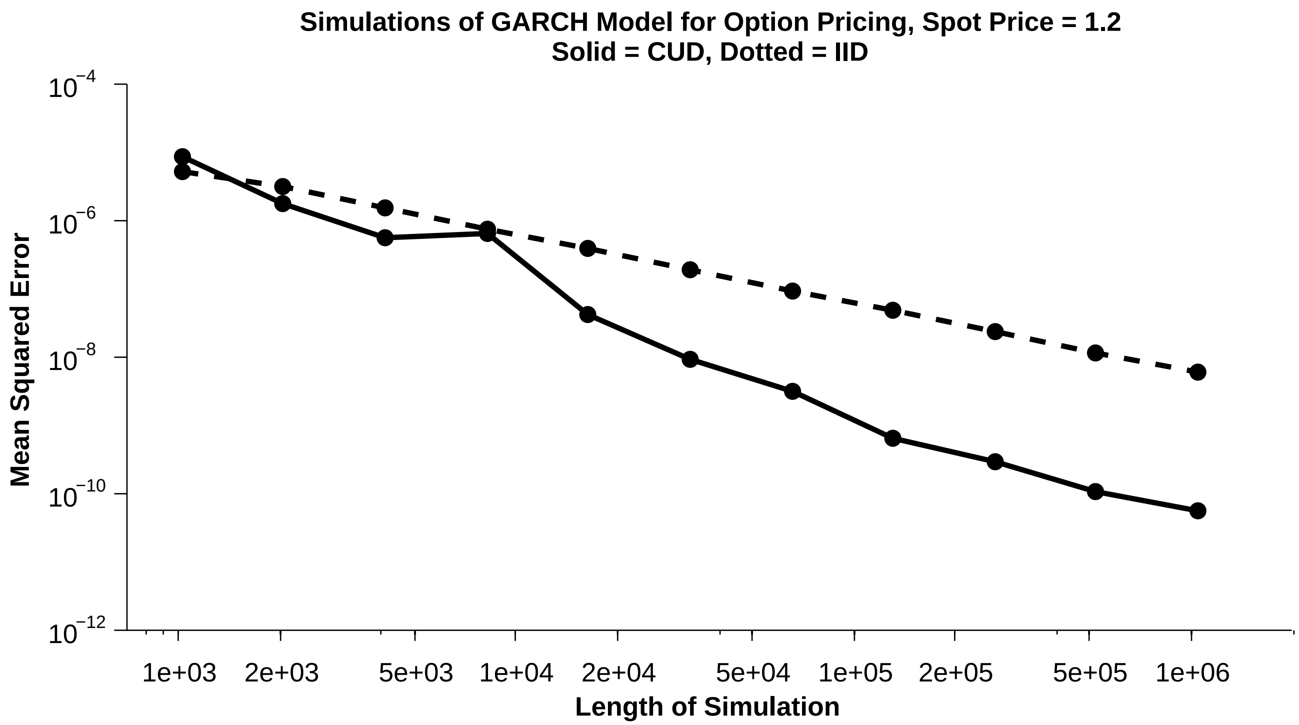
Garch $X_0 = 0.9$


Simulations of GARCH Option Pricing Model, Spot Price = 0.9
Solid = CUD, Dotted = IID
Mean Squared Error
10^−6
10^−7
10^−8
10^−9
10^−10
10^−11
1e+3
2e+3
5e+3
1e+4
2e+4
5e+4
1e+5
2e+5
5e+5
1e+6
Length of Simulation

Garch $X_0 = 1.0$


Simulations of GARCH Model for Option Pricing, Spot Price = 1.0
Solid = CUD, Dotted = IID
Mean Squared Error
10^-4
10^-6
10^-8
10^-10
10^-12
1e+3
2e+3
5e+3
1e+4
2e+4
5e+4
1e+5
2e+5
5e+5
1e+6
Length of Simulation

Garch $X_0 = 1.2$


Simulations of GARCH Model for Option Pricing, Spot Price = 1.2
Solid = CUD, Dotted = IID
Mean Squared Error
10^-4
10^-6
10^-8
10^-10
10^-12
1e+03
2e+03
5e+03
1e+04
2e+04
5e+04
1e+05
2e+05
5e+05
1e+06
Length of Simulation

Conclusions

Some QMC works in MCMC

Improvements range from modest to powerful

Just like QMC in MC

We thank

- NSF for funding
- Coworkers: Josef Dick, Makoto Matsumoto, Takuji Nishimura
- Organizers, especially: Akimichi Takemura, Byeung U. Park, Runze Li

Cassella and George example

Also used by Liao

Gibbs sampling $(v, w) \in [0, 5]^2$ with

$$f(v \mid w) \propto we^{-wv}$$
$$f(w \mid v) \propto ve^{-vw}$$

Same methods, also 3000 replications:

	Mean $\overline{v}$	Var $(\overline{v})$	Mean $\overline{w}$	Var $(\overline{w})$
MC	1.263	0.0066	1.264	0.0067
QMC	1.266	0.0035	1.262	0.0035

Variance ratios are about 1.89 and 1.91

Efficiency gains

1) Small so far
2) Same often happens with finite dimensional QMC
3) Big improvements come from "low effective dimension"

Next steps

1) Find rates of convergence
2) Find/create the niche (if any) where QMC-MCMC brings huge gain
3) Find sufficient conditions for weak CUD

Chentsov's Theorem 1

1) If there are $K < \infty$ states, and,

2) all K^2 transition probabilities are positive, and,

3) u_i are CUD, and,

4) x_0 is arbitrary

5) x_i given x_{i-1} sampled by inversion via u_i, 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 $\hat{\mu}_n \to \mu$

QMC and MCMC

Both subjects

1) have long histories   Metropolis et al. 1953   Weyl 1914

2) are undergoing tremendous growth

A key difference

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

MCMC extends applicability of MC   Eg Bayes methods, stat. phys.

But . . . QMC $\cap$ MCMC $\doteq \varnothing$

Findings

1) We use QMC to drive Metropolis Hastings
2) It reduces variance in our examples
3) We **prove** consistency for some QMC points

Key idea

QMC points that work resemble the whole period of a small RNG

Regular proposals

Recall A is Jordan measureable if 1_A is Riemann integrable

Proposals are regular if

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

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

Suppose that $x_i \to x_{i+1} = \omega$ for any x_i, whenever

$$(u_1, \dots, u_d) \in [0,1]^d \in \mathcal{B}_\omega = \prod_{j=1}^d [a_j, b_j]$$

Then ω is a "home state" with return box $\mathcal{B}_\omega$

Need a "home state" ω with gauranteed return box $\mathcal{B}_\omega$ of positive volume

Example

$z_{i+1} \mid x_i$ Gaussian "pre-proposal" by inversion

y_{i+1} discretized z_{i+1}

Small box around $(1, \dots, 1) \in [0,1]^{d-1}$ always gives "highest proposal" ω

If $\min_{\widetilde{\omega}} A(\widetilde{\omega} \to \omega) > 0$ then ω is a home state

Theorem

For Metropolis-Hastings sampling, if

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

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

Idea of proof

Compare x_{i+m} to $\widetilde{x}_{i,m,m}$ where $\widetilde{x}_{i,m,0}$ is sampled by inversion using u_{id} and the transitions $\widetilde{x}_{i,m,t} \to \widetilde{x}_{i,m,t+1}$ 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 $\widetilde{x}_{i,m,m} \sim p$.