QMC for MCMC:
Background and recent results

Art B. Owen            Su Chen
Stanford University    Stanford University

With contributions from:
Josef Dick, Makoto Matsumoto, Takuji Nishimura
Note:

I have corrected a few typos on the slides that I presented in Warsaw. Also, some important parts of the presentation were spoken, and not read off of the slides. Some of these are inserted on interstitial slides like this one. A few more things have been added in hindsight.

Art Owen, August 2010
Simple Monte Carlo

Used in virtually all sciences

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

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

Recall

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

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, \ldots)

But there’s hope:

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

MCMC broadens **applicability**
Note:

Compared to MC, MCMC offers a much bigger world. QMC by contrast, offers a much better world. Naturally we want to combine these. It is unrealistic to believe that QMC levels of accuracy can be brought to every problem where MCMC is applied. After all, each of QMC and MCMC fails from time to time on its own, the former not always beating MC by much if any, and the latter sometimes failing to mix.

On the other hand, it is unreasonably pessimistic to suppose that there can be no successes. Therefore the objective is to find where the successes might be, combining empirical and theoretical approaches.
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 $\mathbf{x}_i = \phi(\mathbf{x}_{i-1}, \mathbf{v}_i)$, $\mathbf{v}_i \sim U(0, 1)^d$ (Markov property)

Design $\phi(\cdot, \cdot)$ so that $\text{distsn}(\mathbf{x}_i) \to p$

LLN for reasonable conditions

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

- Metropolis-Hastings
  1) While at $x_i$ propose move to $y_{i+1} \sim Q(x_i \rightarrow y_{i+1})$
  2) Accept with probability $A(x_i \rightarrow y_{i+1})$
  3) If accepted, then $x_{i+1} = y_{i+1}$ else $x_{i+1} = x_i$

- Gibbs sampler
  1) Sample component $j$ from $p(x_j \mid x_{ik}, k \neq j)$
  2) Cycle through $j$'s (sequentially, or randomly)
Recipe for QMC in MCMC

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

2) $n$ steps require $u_1, \ldots, u_{nd} \in (0, 1)$

3) MCMC uses $u_i \sim \mathcal{U}(0, 1)$

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
Note:

Non-Markovian simulations are also used in adaptive MCMC.
MCMC looks like $\mathbf{QMC}^\top$

<table>
<thead>
<tr>
<th>Method</th>
<th>Rows</th>
<th>Columns</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>QMC</td>
<td>$n$ points</td>
<td>$d$ variables</td>
<td>$1 \leq d \ll n \rightarrow \infty$</td>
</tr>
<tr>
<td>MCMC</td>
<td>$r$ replicates</td>
<td>$n$ steps</td>
<td>$1 \leq r \ll n \rightarrow \infty$</td>
</tr>
</tbody>
</table>

QMC based on equidistribution  
MCMC based on ergodicity

\[ \text{MCQMC 2010, August 2010, Warsaw} \]
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
Note:

The heat particle might exit stage right. When the proposed $\Delta x$ is large then $U$ is small so acceptance ($U \leq A$) is more likely. When $\Delta x$ is small then $U$ is high. The van der Corput points would be really bad for random walk Metropolis.
Completely uniformly distributed

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

\[ D^*_n(z_1, \ldots, z_n) \to 0, \quad \text{where} \]

\[ z_i = (u_i, \ldots, u_{i+d-1}) \]

For all \( d \geq 1 \)

**Overlapping blocks**

\[ z_1 = (u_1, \ldots, u_d) \]
\[ z_2 = (u_2, \ldots, u_{d+1}) \]
\[ \vdots \]
\[ z_n = (u_n, \ldots, u_{n+d-1}) \]

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

\[ v_i = (u_{d(i-1)+1}, \ldots, u_{di}) \forall d \]
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 $O(1/n)$ · · · if transition probabilities are $a/2^b$ for integers $a, b$
Note:

Chentsov’s paper is remarkable and well worth reading after 40+ 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 $3\epsilon$ argument. The details are in his paper, also in Chen, Dick and Owen (2010) where it is embedded in the ‘Rosenblatt-Chentsov’ transformation.
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
Note:

The discussion followed the thread from theory for finite state spaces, to apparent rate improvements, to theory for continuous spaces.

The exact references for the articles cited are given in the papers. See for example MCMC with QMC papers at stat.stanford.edu/~owen/reports
## Related ideas

<table>
<thead>
<tr>
<th>Related idea</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reordering heat particles</td>
<td>Lécot (1989)</td>
</tr>
<tr>
<td>Manually adaptive QMC</td>
<td>Ostland, Yu (1997)</td>
</tr>
<tr>
<td>QMC for particle filters</td>
<td>Lemieux, Ormoneit, Fleet (2001), UAI</td>
</tr>
<tr>
<td>MCMC ∩ antithetics</td>
<td>Frigessi, Gäsemyr, Rue (2000)</td>
</tr>
<tr>
<td>MCMC ∩ Latin hypercubes</td>
<td>Craiu, Meng (2004)</td>
</tr>
<tr>
<td>Quasi-random walks on balls</td>
<td>Karaivanova, Chi, Gurov (2007)</td>
</tr>
</tbody>
</table>
Results from Tribble

<table>
<thead>
<tr>
<th>Data sets</th>
<th>(n = 2^{10})</th>
<th>(n = 2^{12})</th>
<th>(n = 2^{14})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min</td>
<td>max</td>
<td>min</td>
</tr>
<tr>
<td>Pumps ((d = 11))</td>
<td>286</td>
<td>1543</td>
<td>304</td>
</tr>
<tr>
<td>Vasorestriction ((d = 42))</td>
<td>14</td>
<td>15</td>
<td>56</td>
</tr>
</tbody>
</table>

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.

But the only consistency results $\hat{\mu}_n \to \mu$ were for discrete state spaces, where only small improvements are seen empirically.
Chen, Dick & O extend consistency to continuous state spaces.

\[ \text{MCMC remains consistent when driven by } u_1, u_2, \ldots, \text{ if} \]

1) \( u_i \) are CUD (or CUD in probability)

2) \( m \)-step transitions are Riemann integrable \( \forall m \geq 1 \), 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)

Josef Dick’s talk will show more
Mini-twisters

Matsumoto & Nishimura sent us some small RNGs based on the same principles as the Mersenne twister.

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

$u_1, u_2, \ldots, u_M$

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
**Note:**

I think of them as mini-twisters. More precisely, Makoto Matsumoto tells me that they are small versions of random number generators which obey hypercubical equidistribution properties comparable to those that the twister does. So they’re in the same family as the Mersenne twister, but don’t necessarily share all the same construction ideas.

The $M$ points can be used to define $M$ $k$-tuples of consecutive points, using wraparound. The $k$ dimensional unit cube can be partitioned into $2^{vk}$ subcubes. If $vk \leq m$, then all of those subcubes except the one at the origin have $2^{-vk}(M + 1)$ of the $k$-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.
Same as for finite dimensional QMC.
Gaussian Gibbs sampler

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

Alternate

\[ X_1 \sim \text{DIST}(X_1 \mid X_2 = x_2) = \mathcal{N}(\rho x_2, 1 - \rho^2) \]
\[ X_2 \sim \text{DIST}(X_2 \mid X_1 = x_1) = \mathcal{N}(\rho x_1, 1 - \rho^2) \]
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 $\rho = 0.999$

Simulations of Gaussian Gibbs Sampling, $\rho = 0.999$
Solid = CUD, Dotted = IID

Length of Simulation
Mean Squared Error

Estimate $E(X)$ 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 $\theta_i$ 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 $\text{vol}(\text{triangle})$
M/M/1 queue initial transient

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

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

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

Average wait of first $n$ customers is

$$
\bar{W}_n = \frac{1}{n} \sum_{i=1}^{n} W_i \quad \text{we simulate for} \quad \mathbb{E}(\bar{W}_n)
$$
Variance of average wait

500 simulations of Lindley's formula
Solid=CUD   Dotted=IID

Number of customers
Variance of average wait

MCQMC 2010, August 2010, Warsaw
Heston’s stochastic volatility

\[ dS = rS dt + \sqrt{V} S \, dW_1, \quad 0 < t < T \]
\[ dV = \kappa (\theta - V) dt + \sigma \sqrt{V} \, dW_2 \]

Parameters from J. Zhu (2008):

For \( S \):
\[ S(0) = 100 \quad r = 0.04 \quad T = 6 \]

For \( V \):
\[ V(0) = 0.025 \quad \theta = 0.04 \quad \kappa = 2 \quad \sigma = 0.3 \]

\[ \text{Corr}(dW_1, dW_2) \equiv \rho = -0.5 \]

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

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

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

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

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

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

Compare to exact answer
\[ K = 100 = S(0), \ dt = T/2^8 \]
$K = 70 \quad dt = T/2^8$

Simulations of Heston’s Stochastic Volatility Model for Option Pricing

Spot Price = 100, Strike = 70
Solid = CUD, Dotted = IID
GARCH(1, 1) model

\[ \log\left(\frac{X_t}{X_{t-1}}\right) = r + \lambda \sqrt{h_t} - \frac{1}{2} h_t + \varepsilon_t, \quad 1 \leq t \leq T \]

\[ \varepsilon_t \sim \mathcal{N}(0, h_t) \]

\[ h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} \]

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

For \( X_t \):

\( r = 0 \quad \lambda = 7.452 \times 10^{-3} \quad T = 30 \)

For \( h_t \):

\( \alpha_0 = 1.524 \times 10^{-5} \quad \alpha_1 = 0.1883 \quad \beta_1 = 0.7162 \)

\( h \) starts at \( 0.64 \times 0.2413 \)

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 $\epsilon_t$)

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

Use 100 rotations (adding $U(0, 1)^{30}$)
Garch $X_0 = 0.9$

Simulations of GARCH Option Pricing Model, Spot Price = 0.9
Solid = CUD, Dotted = IID

Length of Simulation
Mean Squared Error
Garch $X_0 = 1.0$

Simulations of GARCH Model for Option Pricing, Spot Price = 1.0
Solid = CUD, Dotted = IID

Mean Squared Error vs. Length of Simulation
GARCH $X_0 = 1.2$

Simulations of GARCH Model for Option Pricing, Spot Price = 1.2
Solid = CUD, Dotted = IID
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: Henryk Wozniakowski, Leszek Plaskota