QMC for MCMC

Radu Craiu
University of Toronto

Conferinta SPSR, Aprilie 2007

Outline

- General MCMC.
- Antithetic coupling and Stratified Sampling.
- Negative Association.
- Latin Hypercube Sampling.
- Implementation for the Gibbs sampler.
- Multiple-Try Metropolis (MTM).
- Randomized QMC for MTM.
- Future directions in MCMC.

General MCMC

- We are interested in computing for $f : R^d \to R$

$$I = \int_{\Omega \subset R^d} f(x)\pi(x)dx.$$

- π is generally known only up to a proportionality constant so direct calculation is impossible.
- MCMC idea: create a Markov chain whose stationary distribution is π. Sample from π using the realizations of this Markov chain.
- Issues:
  - burn-in long enough?
  - chain is mixing well?

The Gibbs Sampler

Suppose π is a d-variate distribution with one-dimensional conditionals $\pi_i(x_i|x_{[-i]})$ from which it is possible to sample for all $i = 1, \ldots, d$. The Gibbs sampler goes through the following steps:

Step 0 Initialize the chain by sampling/selecting $x_0 \in R^d$.

Step t For each $1 \leq i \leq d$ update X_{t-1} to X_t by sampling from

$$X_{t;i} \sim \pi_i(\cdot|x_{t;1}, x_{t;2}, \ldots, x_{t;i-1}, x_{t-1;i+1}, \ldots, x_{t-1;d}).$$

Alternative implementations

- One can (in fact should!) update simultaneously subvector $(x_{i1}, \dots, x_{ik})$ of $(x_1, \dots, x_d)$ if the corresponding conditional distribution can be sampled directly.
- One does not have to go through the components of x in the order $x_1 \to x_2 \dots \to x_d$. Any order would do, in fact the order can be selected at random in each step.

The Metropolis-Hastings Sampler

Given a target π and a proposal distribution T the Metropolis-Hastings sampler is performed in the following manner:

Step 0 Initialize the chain by sampling/selecting $x_0 \in R^d$.

Step t:1 Sample a *proposal* $y \sim T(\cdot|x_{t-1})$; the proposal distribution may depend on the current state of the chain, x_{t-1}.

Step t:2 Compute the *acceptance ratio* $r_t = \min\left\{1, \frac{\pi(y)T(x_{t-1}|y)}{\pi(x_{t-1})T(y|x_{t-1})}\right\}$.

Step t:3 Sample independently $U_t \sim \text{Uniform}(0,1)$. If $U_t \leq r_t$ then $X_t = y$; otherwise $X_t = x_{t-1}$.

Sample processing

- Given the set of realizations $x_0, x_1, \dots, x_m, x_{m+1}, \dots, x_{m+n}$ we discard the first m samples (m is called burn-in time) and we use for inference the last n samples obtained. In many cases m can be HUGE!.
- The desired integral is approximated by

$$\hat{I} = \frac{1}{n} \sum_{j=1}^{n} f(x_{m+j}).$$

- The efficiency of the estimator depends on the size of $\mathrm{Cov}(f(X_{m+t}), f(X_{m+t+s}))$. The auto-covariance can be reduced via: reparametrization of the distribution π, choice of the proposal distribution T, antithetic variates, etc.

Antithetic principle for classical Monte Carlo

- Find $\hat{I}$, $\hat{I}'$, estimators of I such that $\mathrm{Corr}(\hat{I}, \hat{I}') \leq 0$.
- For K=2 processes use the antithetic quantile coupling $X_i^{(1)} = F^{-}(U)$, $X_i^{(2)} = F^{-}(1-U)$.
- Take $\hat{I} = \sum_{i=1}^{n} X_i^{(1)}$ and $\hat{I}' = \sum_{i=1}^{n} X_i^{(2)}$.
- Use $\frac{1}{2}(\hat{I} + \hat{I}')$ as the estimator for I (Hammersley and Morton, 1955).
- Stratification of the input variables state space into two strata. What if we want to use more than two strata?

Antithetic variates for k MCMC processes

We want to estimate $I = \mathrm{E}_\pi f(X)$ using

$$\begin{array}{ccccc} \text{Process 1} & \rightarrow & f(X^{(1)}_{m+1}) & \dots & f(X^{(1)}_{m+n}) \\ \text{Process 2} & \rightarrow & f(X^{(2)}_{m+1}) & \dots & f(X^{(2)}_{m+n}) \\ \dots & \dots & \dots & \dots & \dots \\ \text{Process K} & \rightarrow & f(X^{(K)}_{m+1}) & \dots & f(X^{(K)}_{m+n}) \end{array}$$

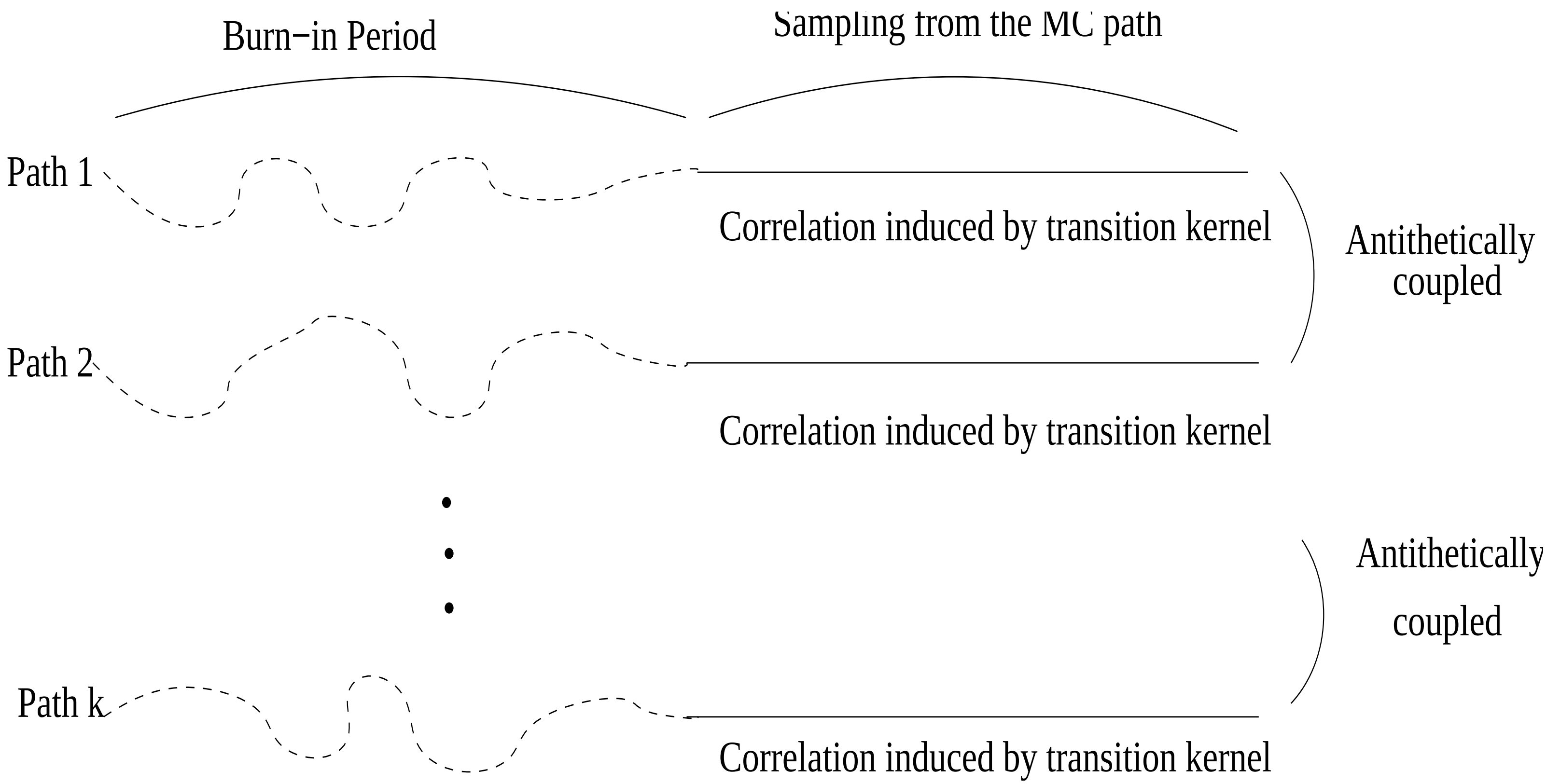
$$\text{If} \quad \gamma_s = \mathrm{Cov}(f(X^{(j)}_{m+r}), f(X^{(j)}_{m+r+s}))$$

$$\beta_s = \mathrm{Cov}(f(X^{(i)}_{m+r}), f(X^{(j)}_{m+r+s})).$$

Let $\hat{I} = \frac{1}{nK} \sum_{r,j} f(X^{(j)}_{m+r})$. We denote $\hat{I}_{ind}$ if the parallel processes are independent.

$$\frac{V(\hat{I})}{V(\hat{I}_{ind})} = 1 + (K-1)\frac{\beta_0 + 2\sum_{r=1}^{n-1}\beta_r(1-\frac{r}{n})}{\gamma_0 + 2\sum_{r=1}^{n-1}\gamma_r(1-\frac{r}{n})}$$

In general, $\gamma_r \geq 0$ so if $\beta_s \leq 0$ then we obtain variance reduction.



Negative Association

- The random variables $X_1, X_2, ..., X_K$ are said to be negatively associated (NA) if for every pair of disjoint subsets A_1, A_2 of $\{1, 2, ..., K\}$

$$\text{Cov}(f_1(X_i, i \in A_1), f_2(X_j, j \in A_2)) \leq 0$$

whenever f_1 and f_2 are increasing in each of the arguments (Joag-Dev and Proschan, 1983).

- The union of two independent sets, each of which is NA, is NA.

NA and MCMC

- A generic MCMC algorithm can be written in the general form

$$X_t = \psi(X_{t-1}, W_t),$$

where ψ is a deterministic map and all randomness is absorbed in the random seed $W^{(t)}$. We can think of W as being a vector of Uniform$(0,1)$ random variables. In the case of Gibbs samplers, ψ is monotone in at least some of the W's.

- Suppose $W_t = (U_t, V_t)$ and ψ is monotone in both U_t and V_t. We can antithetically couple K parallel MCMC processes by generating at the t-th iteration K-dimensional random vectors $(U_t^{(1)}, \ldots, U_t^{(K)})$ and $(V_t^{(1)}, \ldots, V_t^{(K)})$ which are NA and update each chain using $X_t^{(i)} = \psi(X_{t-1}^{(i)}, U_t^{(i)}, V_t^{(i)})$ for any $1 \le i \le K$.

Iterative Latin Hypercube Sampling

- Latin Hypercube Sampling is a traditional method to stratify the input variables used in Monte Carlo experiments. The iterative variant of the classical construction is as follows:

  Step 0 Draw $U^{(0)} = (U_1^{(0)}, ..., U_K^{(0)})$ iid Uniform$(0, 1)$.

  Step t Let $\sigma^{(t)}$ be a random permutation of $\{0, 1, ..., K-1\}$ then take $U^{(t)} = \frac{1}{K}(\sigma^{(t)} + U^{(t-1)}), t = 1, ..., T$.

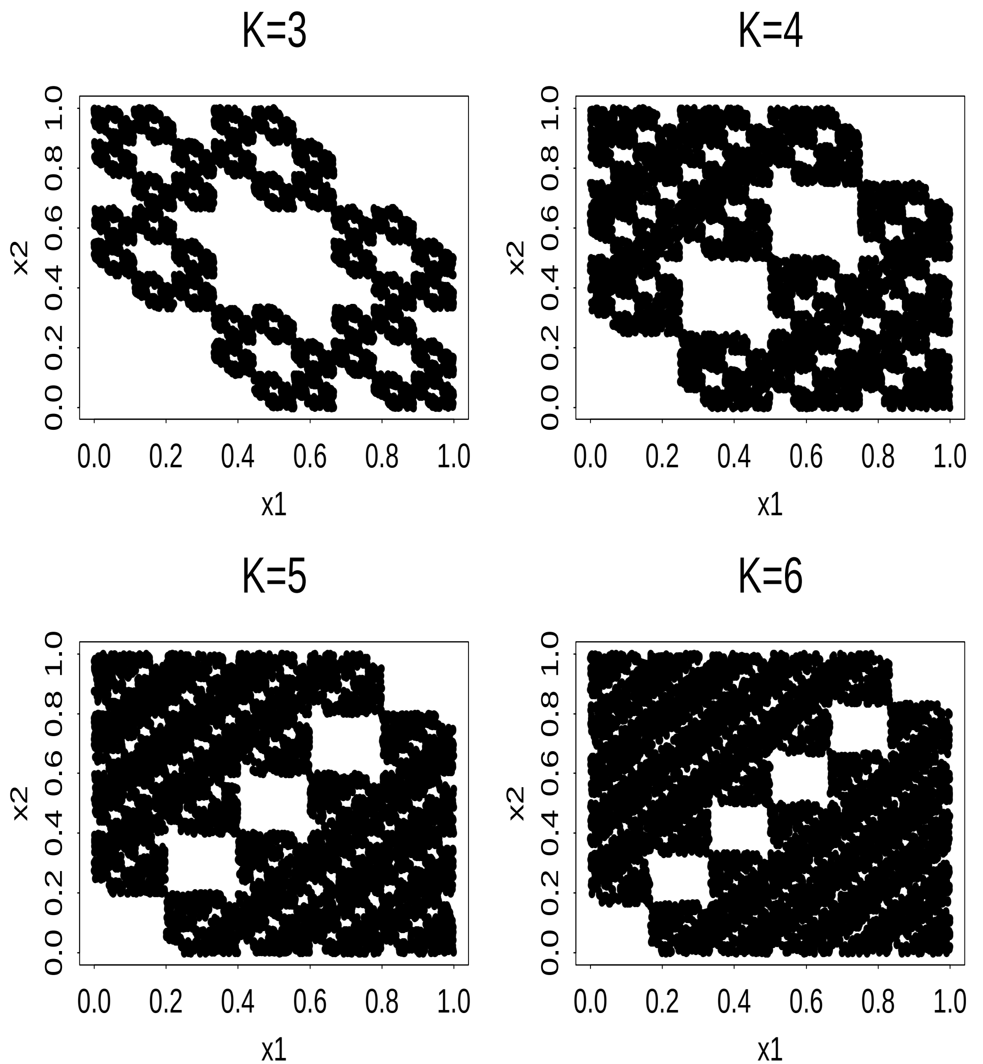
- Marginally, $U_i^{(T)} \sim \text{Uniform}(0, 1)$, $\forall T, i$.

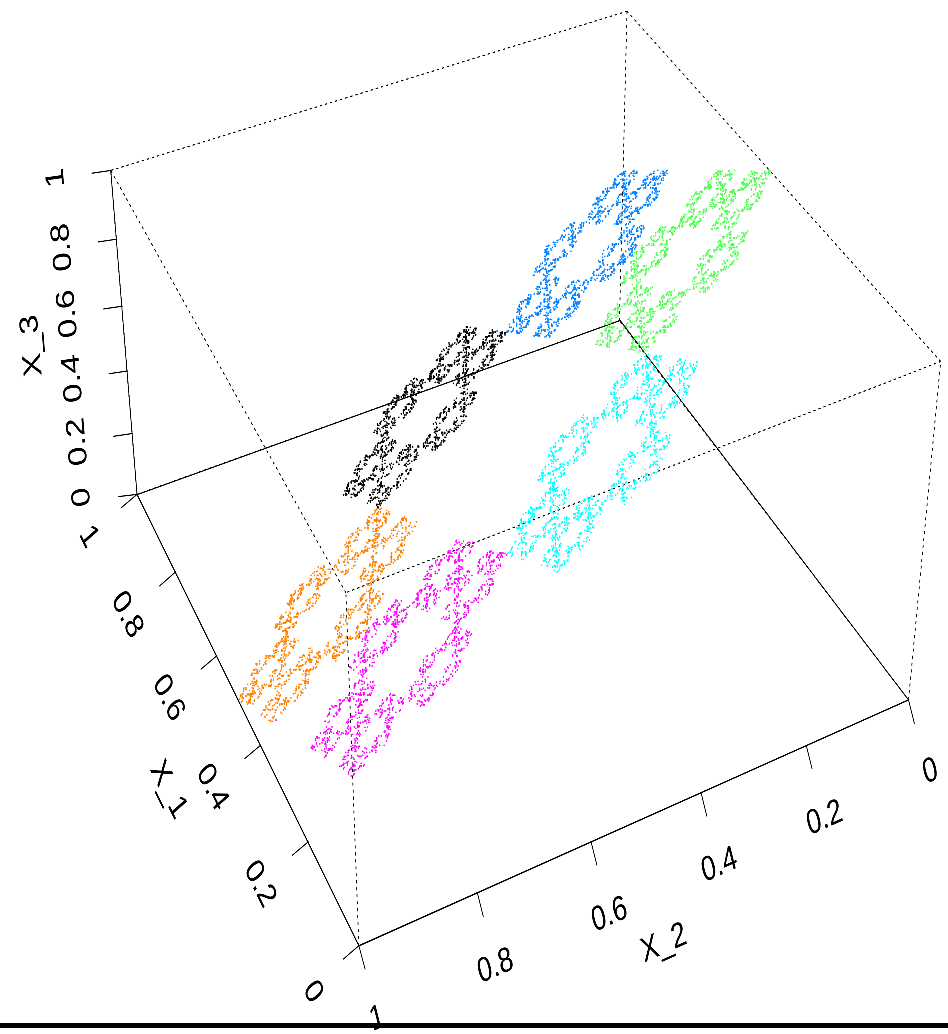
- $\text{Corr}(U_1^{(T)}, U_2^{(T)}) = -\frac{1}{K-1}\left(1 - \frac{1}{K^{2T}}\right) \overset{T\to\infty}{\longrightarrow} -\frac{1}{K-1}$.

- $U_1^{(T)}, U_2^{(T)}, ..., U_K^{(T)}$ are NA, $\forall T > 0$.

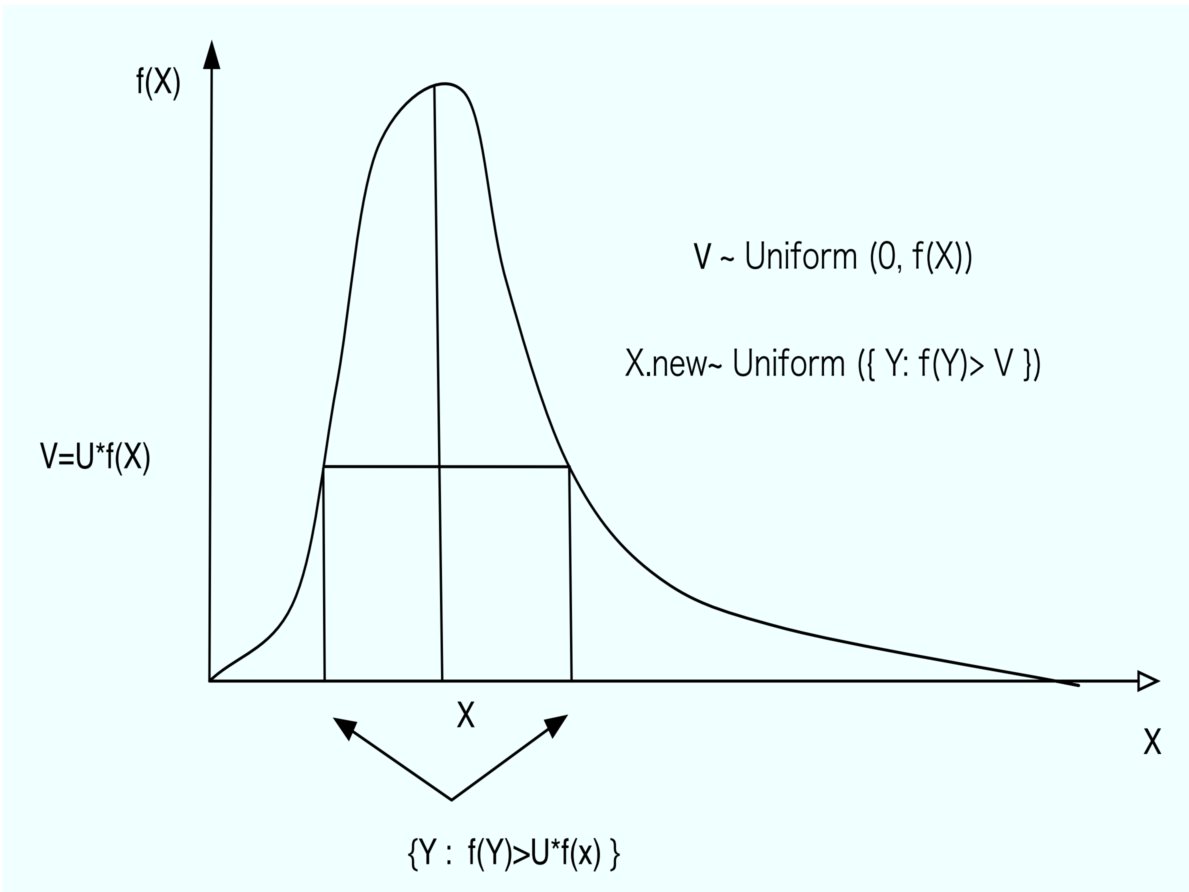
Example K=3

- $U_1^{(0)}, U_2^{(0)}, U_3^{(0)} \sim \mathrm{Uniform}(0,1)$.
- $U_1^{(T)} = \frac{1}{K} + \frac{2}{K^2} + \frac{0}{K^3} + ... + \frac{1}{K^T} + \frac{U_1^{(0)}}{K^T}$.
- $U_2^{(T)} = \frac{0}{K} + \frac{1}{K^2} + \frac{2}{K^3} + ... + \frac{2}{K^T} + \frac{U_2^{(0)}}{K^T}$.
- $U_3^{(T)} = \frac{2}{K} + \frac{0}{K^2} + \frac{1}{K^3} + ... + \frac{0}{K^T} + \frac{U_3^{(0)}}{K^T}$.


K=3
K=4
K=5
K=6
x1
x2
0.0 0.2 0.4 0.6 0.8 1.0


X_1
X_2
X_3
0
0.2
0.4
0.6
0.8
1

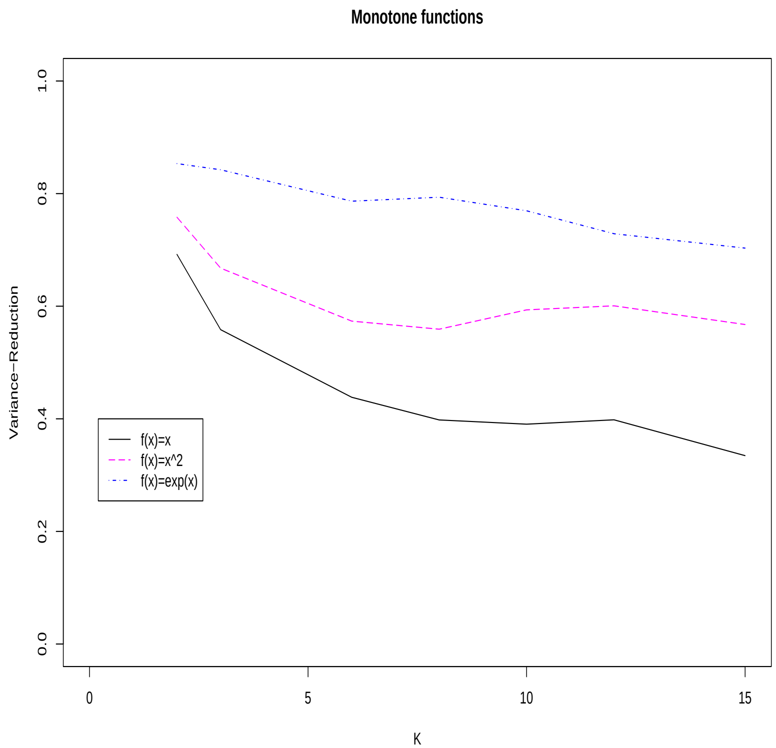
Slice Sampling: $\pi(x) = C \cdot f(x)$

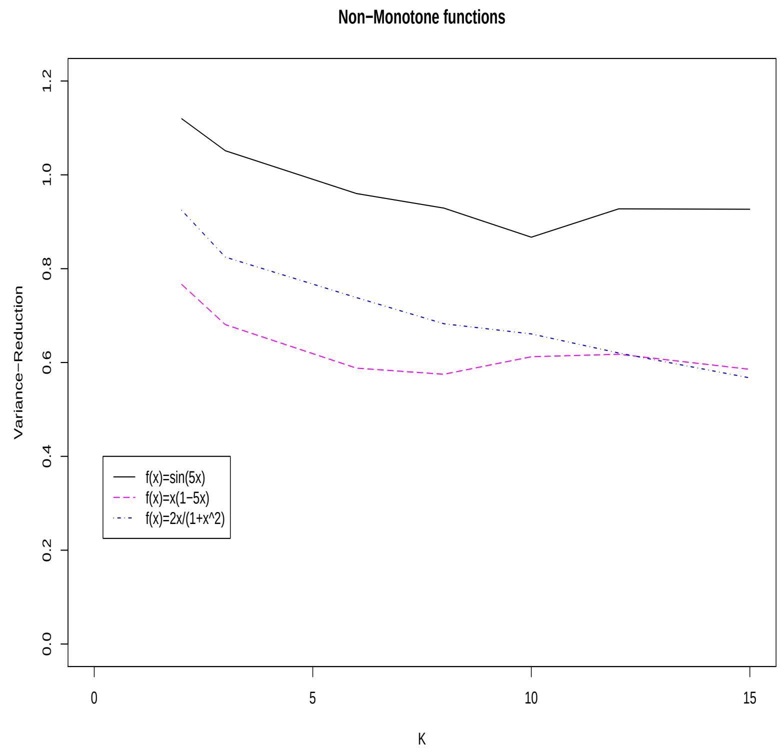


Simple Illustration

- We want $\mathrm{E}_\pi[X]$ where $\pi(x) \propto e^{-e^x}$.
- Slice sampling using the change of variable $v = -\log(u/x^2)$.
- Step 1: $v \sim p(v|x) \propto e^{-v} I_{\{v \geq e^x\}} dv$
- Step 2: $x \sim p(x|v) \propto I_{\{x \leq log(v)\}} dx$.
- Together $X_{t+1} = \psi(X_t, \xi_1, \xi_2) = \xi_1 \log(e^{X_t} - \log(1 - \xi_2))$, where ξ_1 and ξ_2 are i.i.d. $\mathrm{Uniform}(0, 1)$.

Simulation Results





Quasi-Monte Carlo (QMC)

- QMC is a de-randomized MC.
- QMC methods focus on the unit hypercube for uniform and stratified sampling.
- Features of interest: equi-distribution, high-uniformity, technically known as low discrepancy.
- The sequences do not have to be random, in fact they can be completely deterministic.
- More often randomized versions of QMC (denoted RQMC) algorithms are used.
- Adding noise to a deterministic method allows the estimation of the Monte Carlo error. (i.e. $\mathrm{Var}(\hat{I})$).

QMC and RQMC

- If we construct the LHS using at $t = 1$

$$U_i^{(1)} = \frac{\sigma(i) + 0.5}{K}, \ \forall 1 \leq i \leq K$$

then the (iterative) hypercube sampling is deterministic.

- The random variables $U_1^{(0)}, \ldots, U_K^{(0)} \sim \text{Uniform}(0,1)$ used in our construction results in allowing each component of $U^{(1)}$ to be anywhere inside the intervals $(i/K, (i+1)/K)$ for $0 \leq i \leq K-1$.

RQMC for MH

- The antithetic coupling described before fails in the case of M-H algorithms.
- Due to accept-reject behavior, the NA between processes cannot be preserved.
- A different use of RQMC methods is allowed via the Multiple-Try Metropolis

Multiple-Try Metropolis

- Suppose T is such that $T(x|y) > 0 \Leftrightarrow T(y|x) > 0$.
- Draw K trial proposals $Y_1, \ldots, Y_K$ from $T(y|x^{(t)})$. Compute $w(y_j, x^{(t)}) = \pi(y_j)T(x^{(t)}|y_j)\lambda(x^{(t)}, y_j)$ for each j. (we need only $\lambda(x,y) = \lambda(y,x)$)
- Select Y among the K proposals with probability $w(y_j, x^{(t)})/\sum_{i=1}^{K} w(y_i, x^{(t)}), j = 1, \ldots, K$.
- Draw $x_1^*, \ldots, x_{K-1}^* \sim T(\cdot|y)$ and let $x_K^* = x^{(t)}$.
- Accept $x^{(t+1)} = y$ with generalized acceptance probability

$$r_g = \min\left\{1, \frac{w(y_1, x^{(t)}) + \ldots + w(y_K, x^{(t)})}{w(x_1^*, y) + \ldots + w(x_K^*, y)}\right\}.$$

Multiple-Correlated-Try Metropolis

- Suppose we sample K trial proposals $Y_1, \ldots, Y_K$ from $\tilde{T}(y_1, \ldots, y_K | x^{(t)})$ where $\int \tilde{T}(y_1, \ldots, y_K | x^{(t)}) dy_2 \ldots dy_K = T(y_1 | x^{(t)})$.
- The algorithm proceeds as in the independent case with one exception.
- Draw $(X_1^*, \ldots, X_{K-1}^*)$ variates from the conditional transition kernel $\tilde{T}(x_1, \ldots, x_{K-1} | y, x_K = x^{(t)})$ and let $X_K^* = x^{(t)}$
- We have quite a bit of freedom in choosing $\tilde{T}$ as long as we can perform the blue step.

Random Walk Multiple Try Metropolis

- For multivariate targets a common choice is the Random Walk Metropolis.
- $Y_1, \ldots, Y_k \sim N_d(x_t; \Sigma)$.
- Idea: Stratifying the sample of proposals produces a more structured search of the space "around X_t".

Korobov rule

- Choose an integer $a \in \{1, \ldots, K-1\}$ and let

$$P_K = \left\{ \frac{i-1}{K}(1, a, \ldots, a^{r-1}) \bmod 1, i = 1, \ldots, K \right\}.$$

- This type of point set can be randomized by generating a random vector $\mathbf{v}$ uniformly in $[0,1)^r$, and adding it to each point of P_K (modulo 1). That is, let $\tilde{P}_K = \{\tilde{\mathbf{u}}_i, i = 1, \ldots, K\}$, where

$$\tilde{\mathbf{u}}_i = (\mathbf{u}_i + \mathbf{v}) \bmod 1.$$

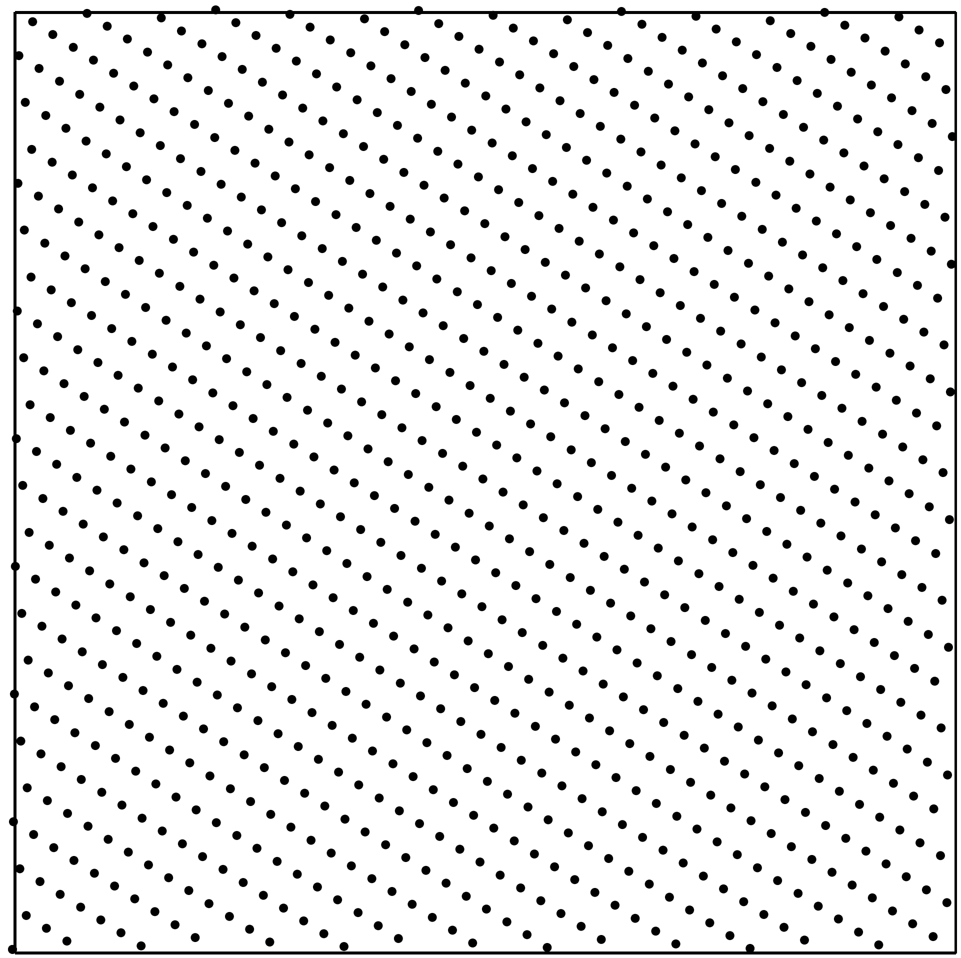


Figure 1: *Two-dimensional Korobov rule with* $K = 1024$ *and* $a = 139$.

Example: Lupus Data

Table 1: *The number of latent membranous lupus nephritis cases, the numerator, and the total number of cases, the denominator, for each combination of the values of the two covariates.*

	IgA				
IgG3-IgG4	0	0.5	1	1.5	2
-3.0	0/ 1	-	-	-	-
-2.5	0/ 3	-	-	-	-
-2.0	0/ 7	-	-	-	0/ 1
-1.5	0/ 6	0/ 1	-	-	-
-1.0	0/ 6	0/ 1	0/ 1	-	0/ 1
-0.5	0/ 4	-	-	1/ 1	-
0	0/ 3	-	0/ 1	1/ 1	-
0.5	3/ 4	-	1/ 1	1/ 1	1/ 1
1.0	1/ 1	-	1/ 1	1/ 1	4/ 4
1.5	1/ 1	-	-	2/ 2	-

The Model

- $logit\ P(Y_i = 1) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$ where $X_i^T = (1, X_{i1}, X_{2i})$ is the vector of covariates for the i-th individual.
- The prior for $\beta = (\beta_0, \beta_1, \beta_2)^T$ is trivariate normal with zero mean and variance $\text{diag}(100^2, 100^2, 100^2)$.
- The posterior density is then proportional to

$$\pi(\beta|x, y) \propto \prod_{j=0}^{2} \frac{e^{-0.5\beta_j/100^2}}{100\sqrt{2\pi}} \prod_{i=1}^{55} \left[\frac{\exp(X_i^T\beta)}{1 + \exp(X_i^T\beta)}\right]^{y_i} \left[\frac{1}{1 + \exp(X_i^T\beta)}\right]^{1-y_i}.$$

Simulation Results

- We report for β_1 and $p_{25} = 1_{\{\beta_1 > 25\}}$, the ratios $R = \frac{\text{MSE}_{\text{anti}}}{\text{MSE}_{\text{ind}}}$ and $R = \frac{\text{MSE}_{\text{qmc}}}{\text{MSE}_{\text{ind}}}$
- If we denote by b_{ij} the j^{th} sample point drawn in the i^{th} replicate from the posterior distribution of β_1 then, using $\bar{b}_{\cdot\cdot} = \frac{\sum_{ij} b_{ij}}{MN}$ and $\bar{b}_{i\cdot} = \frac{\sum_j b_{ij}}{N}$ for all $i = 1, \ldots, M$ the MSE is defined as

$$\text{MSE} = (\bar{b}_{\cdot\cdot} - E[\beta_1|\text{data}])^2 + \frac{\sum_i (\bar{b}_{i\cdot} - \bar{b}_{\cdot\cdot})^2}{(M-1)}.$$

- Similar calculations can be done for p_{25}.

Table 2: *Values of R for β_1/p_{25} in the logit example.*

	Antithetic			QMC		
$K\backslash\sigma$	2	3	4	2	3	4
3	0.92/0.92	0.90/0.86	0.99/0.95	-	-	-
4	0.94/0.87	0.88/0.88	0.91/0.89	-	-	-
5	0.98/0.96	0.81/0.81	0.89/0.86	-	-	-
6	0.91/0.86	0.86/0.78	0.95/0.92	-	-	-
8	0.81/0.70	0.75/0.69	0.83/0.80	0.69/0.72	0.61/0.60	0.59/0.56
16	0.87/0.81	0.97/0.94	0.91/0.88	0.81/0.81	0.82/0.84	0.76/0.75

Monte Carlo: What Else is Hot?

- Adaptive MCMC: how can we change the transition kernel for an MCMC algorithm "on the go". Problems: the adaptation has to take into account a number of the samples already produced so the process loses its Markovian property.

- Sequential Monte Carlo.

$$\text{state equation: } x_t \sim q_t(\cdot|x_{t-1}, \theta)$$

$$\text{observation equation: } y_t \sim f_t(\cdot|x_t, \phi)$$

where y_t are observations arriving sequentially, x_t are the "state variables". Of interest is the "current" posterior distribution of x_t

$$\pi_t(x_t) \propto \int q_t(x_t|x_{t-1}) f_t(y_t|x_t) \pi_{t-1}(x_{t-1}) dx_{t-1}.$$

Places to go and see more

1. Web page of Christiane Lemieux (Waterloo):
   http://www.math.uwaterloo.ca/˜clemieux/
2. Web page of Jun Liu (Harvard):
   http://www.people.fas.harvard.edu/˜junliu/
3. Web page of Art Owen (Stanford):
   http://www-stat.stanford.edu/˜owen/
4. I will post this talk on my website:
   http://fisher.utstat.toronto.edu/craiu/