

# Monte Carlo Integration With Acceptance-Rejection

Zhiqiang TAN

This article considers Monte Carlo integration under rejection sampling or Metropolis-Hastings sampling. Each algorithm involves accepting or rejecting observations from proposal distributions other than a target distribution. While taking a likelihood approach, we basically treat the sampling scheme as a random design, and define a stratified estimator of the baseline measure. We establish that the likelihood estimator has no greater asymptotic variance than the crude Monte Carlo estimator under rejection sampling or independence Metropolis-Hastings sampling. We employ a subsampling technique to reduce the computational cost, and illustrate with three examples the computational effectiveness of the likelihood method under general Metropolis-Hastings sampling.

**Key Words:** Importance sampling; Metropolis-Hastings sampling; Rao-Blackwellization; Rejection sampling; Stratification; Variance reduction.

## 1. INTRODUCTION

In many problems of statistics, it is of interest to compute expectations with respect to a probability distribution. For certain situations, it is also necessary to estimate its normalizing constant. Specifically, let  $q(x)$  be a nonnegative function on a state space  $\mathcal{X}$  and consider the probability distribution whose density is

$$p(x) = \frac{q(x)}{Z}$$

with respect to a baseline measure  $\mu_0$ , where  $Z$  is the normalizing constant  $\int q(x) d\mu_0$ . Monte Carlo is a useful method for solving the aforementioned problems, and typically has two parts, simulation and estimation, in its implementation.

First, a sequence of observations  $x_1, \dots, x_n$  are simulated from the distribution  $p(\cdot)$ . Then the expectation  $E_p(\varphi)$  of a function  $\varphi(x)$  with respect to  $p(\cdot)$  can be estimated by the

---

Zhiqiang Tan is Assistant Professor, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205 (E-mail: [ztan@jhsph.edu](mailto:ztan@jhsph.edu)).

© 2006 American Statistical Association, Institute of Mathematical Statistics,  
and Interface Foundation of North America

*Journal of Computational and Graphical Statistics*, Volume 15, Number 3, Pages 735–752  
DOI: 10.1198/106186006X142681

sample average or the crude Monte Carlo (CMC) estimator

$$\frac{1}{n} \sum_{i=1}^n \varphi(x_i). \quad (1.1)$$

By letting  $\varphi(x) = q_1(x)/q(x)$ , the normalizing constant  $Z$  can be estimated by

$$\left( \frac{1}{n} \sum_{i=1}^n \frac{q_1(x_i)}{q(x_i)} \right)^{-1}, \quad (1.2)$$

where  $q_1(x)$  is a probability density on  $\mathcal{X}$ . This estimator is called reciprocal importance sampling (RIS); see DiCiccio, Kass, Raftery, and Wasserman (1997) and Gelfand and Dey (1994).

This article considers rejection sampling or Metropolis-Hastings sampling for the simulation part. Rejection sampling requires a probability density  $\rho(x)$  and a constant  $C$  such that  $q(x) \leq C\rho(x)$  on  $\mathcal{X}$ , which implies  $Z \leq C$  (von Neumann 1951). At each time  $t \geq 1$ ,

- *Sample  $y_t$  from  $\rho(\cdot)$ ;*
- *accept  $y_t$  with probability  $q(y_t)/[C\rho(y_t)]$  and move to the next trial otherwise.*

The second step can be implemented by generating  $u_t$  from uniform  $(0, 1)$  and accepting  $y_t$  if  $u_t \leq q(y_t)/[C\rho(y_t)]$ . Then the accepted  $y_t$  are independent and identically distributed (iid) as  $p(\cdot)$ . To compare, Metropolis-Hastings sampling requires a family of probability densities  $\{\rho(\cdot; x) : x \in \mathcal{X}\}$  (Metropolis et al. 1953; Hastings 1970). At each time  $t \geq 1$ ,

- *Sample  $y_t$  from  $\rho(\cdot; x_{t-1})$ ;*
- *accept  $x_t = y_t$  with probability  $1 \wedge \beta(y_t; x_{t-1})$  and let  $x_t = x_{t-1}$  otherwise, where*

$$\beta(y; x) = \frac{q(y)\rho(x; y)}{q(x)\rho(y; x)}.$$

The second step can also be implemented by generating  $u_t$  from uniform  $(0, 1)$  and accepting  $x_t = y_t$  if  $u_t \leq 1 \wedge \beta(y_t; x_{t-1})$ . Under suitable regularity conditions, the Markov chain  $(x_1, x_2, \dots)$  converges to the target distribution  $p(\cdot)$ . In the so-called independence case (IMH), the proposal density  $\rho(\cdot; x) \equiv \rho(\cdot)$  is independent of  $x$ . Then the chain is uniformly ergodic if  $q(x)/\rho(x)$  is bounded from above on  $\mathcal{X}$  and is not even geometrically ergodic otherwise (Mengersen and Tweedie 1996). The condition that  $q(x) \leq C\rho(x)$  on  $\mathcal{X}$  is assumed henceforth.

Neither rejection sampling nor Metropolis-Hastings sampling requires the value of the normalizing constant  $Z$ . However, each algorithm involves accepting or rejecting observations from proposal distributions. Acceptance or rejection depends on uniform random variables. By integrating out these uniform random variables, Casella and Robert (1996) proposed a Rao-Blackwellized estimator that has no greater variance than the crude Monte Carlo estimator, but they mostly disregarded the issue of computational time increased by Rao-Blackwellization.

Recently, Kong et al. (2003) formulated Monte Carlo integration as a statistical model using simulated observations as data. The baseline measure is treated as the parameter and

estimated as a discrete measure by maximum likelihood. Consequently, integrals of interest are estimated as finite sums by substituting the estimated measure. We take the likelihood approach, and develop a method for estimating simultaneously the normalizing constant  $Z$  and the expectation  $E_p(\varphi)$  under rejection sampling (Section 2) or Metropolis-Hastings sampling (Section 3). This work can be considered as a concrete case of Tan's (2003) methodology, but its significance is worth singling out.

Under rejection sampling or independence Metropolis-Hastings sampling, computation of the estimated measure requires negligible effort. We establish theoretical results that the likelihood estimator of  $E_p(\varphi)$  has no greater asymptotic variance than the crude Monte Carlo estimator under each scheme. These two facts together imply that the likelihood method is always computationally more effective than crude Monte Carlo under rejection sampling or independence Metropolis-Hastings sampling.

Under general Metropolis-Hastings sampling, computation of the estimated measure involves intensive evaluations of proposal densities. Therefore we employ a subsampling technique to reduce this computational cost. We provide empirical evidence for the computational effectiveness of the likelihood method with three examples. We also introduce approximate variance estimators for the point estimators. These estimators agree with the empirical variances from repeated simulations in all the examples.

The proofs of all lemmas and theorems are collected in the Appendix. Computations in Sections 2.3 and 3.3 are programmed in MATLAB on Pentium 4 machines with CPU speed of 2.00 GHz.

## 2. REJECTION SAMPLING

For clarity, let us consider two cases separately. In one case, the number of trials is fixed while that of acceptances is random. In the other case, the number of acceptances is fixed while that of trials is random. For illustration, an example is provided.

### 2.1 FIXED NUMBER OF TRIALS

Suppose that rejection sampling is run for a fixed number, say  $n$ , of trials. In the process,  $y_1, \dots, y_n$  are generated from  $\rho(\cdot)$ ,  $u_1, \dots, u_n$  are from uniform  $(0, 1)$ , and acceptance occurs at  $t_1, \dots, t_L$ . Then the number of acceptances  $L$  has the binomial distribution  $(n, Z/C)$ . The expectation  $E_p(\varphi)$  can be estimated by

$$\frac{1}{L} \sum_{i=1}^L \varphi(y_{t_i}). \quad (2.1)$$

This estimator is defined to be 0 if  $L = 0$ , which has positive probability for any finite  $n$ .

**Lemma 1.** *Assume that  $\text{var}_\rho(\varphi) < \infty$ , where  $\text{var}_\rho$  denotes the variance under the distribution  $\rho(\cdot)$ . The estimator (2.1) has asymptotic variance  $n^{-1}(Z/C)\text{var}_p(\varphi)$  as  $n$  tends to infinity, where  $\text{var}_p$  is the variance under the distribution  $p(\cdot)$ .*

In the likelihood approach, we ignore the baseline measure ( $\mu_0$ ) being Lebesgue or counting and treat the ignored measure ( $\mu$ ) as the parameter in a model. The parameter

space consists of nonnegative measures on  $\mathcal{X}$  such that  $\int \rho(x) d\mu$  is finite and positive. The model states that the data are generated as follows. For each  $t \geq 1$ ,

- $y_t$  has the distribution  $\rho(\cdot) d\mu / \int \rho(y) d\mu$ ;
- $u_t$  has the distribution uniform  $(0, 1)$ ;
- accept  $y_t$  if  $u_t \leq q(y_t) / [C\rho(y_t)]$  and reject otherwise.

It is easy to show that the accepted  $y_t$  are independent and identically distributed as  $q(\cdot) d\mu / \int q(y) d\mu$ . The likelihood of the process at  $\mu$  is proportional to

$$\prod_{i=1}^n \left[ \rho(y_i) \mu(\{y_i\}) / \int \rho(y) d\mu \right].$$

Any measure that does not place mass at each of the points  $y_1, \dots, y_n$  has zero likelihood. By Jensen's inequality, the maximizing measure is given by

$$\hat{\mu}(\{y\}) \propto \frac{\hat{\Gamma}(\{y\})}{\rho(y)},$$

where  $\hat{\Gamma}$  is the empirical distribution placing mass  $n^{-1}$  at each of the points  $y_1, \dots, y_n$ . Consequently, the expectation  $E_p(\varphi)$  is estimated by  $\int \varphi(y) q(y) d\hat{\mu} / \int q(y) d\hat{\mu}$ , or

$$\sum_{i=1}^n \frac{\varphi(y_i) q(y_i)}{\rho(y_i)} / \sum_{i=1}^n \frac{q(y_i)}{\rho(y_i)}. \quad (2.2)$$

Computation of this estimator is easy because the ratios  $q(y_i) / \rho(y_i)$  are already evaluated in the simulation. By the delta method (Ferguson 1996, sec. 7), the estimator (2.2) has asymptotic variance  $n^{-1} \text{var}_\rho[(\varphi - E_p(\varphi))p/\rho]$  as  $n$  tends to infinity.

**Theorem 1.** *The likelihood estimator (2.2) has no greater asymptotic variance than the crude Monte Carlo estimator (2.1).*

In practice, the estimator (2.1) is loosely referred to as rejection sampling and the estimator (2.2) as importance sampling. Liu (1996) argued that importance sampling can be asymptotically more efficient than rejection sampling in many cases. However, Theorem 1 indicates that importance sampling is asymptotically at least as efficient as rejection sampling in all cases.

The estimator (2.2) effectively achieves Rao-Blackwellization and averages over random variables  $u_1, \dots, u_n$  in the following sense:

$$E \left[ \frac{L}{n} \mid y_1, \dots, y_n \right] = \frac{1}{nC} \sum_{i=1}^n \frac{q(y_i)}{\rho(y_i)},$$

$$E \left[ \frac{1}{n} \sum_{i=1}^L \varphi(y_{t_i}) \mid y_1, \dots, y_n \right] = \frac{1}{nC} \sum_{i=1}^n \frac{\varphi(y_i) q(y_i)}{\rho(y_i)}.$$

Note that the conditional expectation of (2.1) given  $y_1, \dots, y_n$  is not equal to (2.2).

## 2.2 FIXED NUMBER OF ACCEPTANCES

Suppose that rejection sampling is run for  $N$  trials until a fixed number, say  $l$ , of proposals  $y_t$  are accepted. Then the number of trials  $N$  has the negative binomial distribution  $(l, Z/C)$ .

The estimator

$$\frac{1}{l} \sum_{i=1}^l \varphi(y_{t_i}) \tag{2.3}$$

of  $E_p(\varphi)$  is unbiased, and has variance  $n^{-1} \text{var}_p(\varphi)$ . Despite the sequential stopping of the process, consider the same Monte Carlo model as in Section 2.1. The resulting estimator of  $E_p(\varphi)$  is

$$\sum_{i=1}^N \frac{\varphi(y_i)q(y_i)}{\rho(y_i)} / \sum_{i=1}^N \frac{q(y_i)}{\rho(y_i)}. \tag{2.4}$$

As before, computation of (2.4) is easy because the ratios  $q(y_i)/\rho(y_i)$  are already evaluated in the simulation. Theorem 2 gives a similar comparative result as Theorem 1. It is interesting that the efficiency factor of the likelihood estimator over the crude Monte Carlo estimator is the same whether the number of trials or that of acceptances is fixed.

**Lemma 2.** *Assume that  $\text{var}_\rho(\varphi) < \infty$ . The estimator (2.4) has asymptotic variance  $l^{-1}(Z/C)\text{var}_\rho[(\varphi - E_p(\varphi))p/\rho]$  as  $l$  tends to infinity.*

**Theorem 2.** *The likelihood estimator (2.4) has no greater asymptotic variance than the crude Monte Carlo estimator (2.3).*

Casella and Robert (1996) considered the Rao-Blackwellized estimator

$$\frac{1}{l} E \left[ \sum_{i=1}^l \varphi(y_{t_i}) \mid N, y_1, \dots, y_N \right],$$

and gave a recursive formula with  $O(N^2)$  operations. To compare, the estimator (2.4) involves  $O(N)$  operations and is a function only of  $(N, y_1, \dots, y_N)$ . In fact,  $l$  and  $\sum_{i=1}^l \varphi(y_{t_i})$  are Rao-Blackwellized without conditioning on the event that there are  $l$  acceptances. We believe that the likelihood estimator is asymptotically as efficient as the Rao-Blackwellized estimator, and provide some empirical evidence in Section 2.3.

## 2.3 ILLUSTRATION

Consider Example 2.1 of Casella and Robert (1996). The target distributions are gamma (2.434, 4.868) and gamma (20.62, 41.24) with mean 0.5, and the proposal distribution is gamma (2, 4) with the same mean. The acceptance rate is 0.9 for the first case and 0.3 for the second one. The mean 0.5 and tail probability 0.05 are estimated; see Tables 1–2. The likelihood estimator has smaller mean squared error than the crude Monte Carlo estimator, considerably at the lower acceptance rate. The magnitude of decrease in mean squared error appears similar to that achieved by Rao-Blackwellization in Casella and Robert (1996).

Table 1. Comparison of Estimators of Mean 0.5

Size	Acceptance rate 0.9					Acceptance rate 0.3				
	CMC	LIK	CMC	LIK	%	CMC	LIK	CMC	LIK	%
	Mean	Mean	MSE	MSE	Decrease	Mean	Mean	MSE	MSE	Decrease
10	0.5004	0.5011	10.13	8.045	20.59	0.5002	0.5000	1.229	0.4814	60.83
25	0.4995	0.4997	4.053	3.181	21.51	0.5001	0.4997	0.4811	0.1864	61.26
50	0.4994	0.4998	2.077	1.611	22.46	0.4999	0.4998	0.2426	0.08967	63.04
100	0.4990	0.4992	1.062	0.8251	22.31	0.4999	0.4999	0.1221	0.04620	62.16

Table 2. Comparison of Estimators of Tail Probability 0.05

Size	Acceptance rate 0.9					Acceptance rate 0.3				
	CMC	LIK	CMC	LIK	%	CMC	LIK	CMC	LIK	%
	Mean	Mean	MSE	MSE	Decrease	Mean	Mean	MSE	MSE	Decrease
10	0.0492	0.0498	4.753	3.579	24.71	0.0509	0.0504	4.847	1.088	77.55
25	0.0494	0.0496	1.879	1.394	25.80	0.0502	0.0497	1.905	0.4113	78.41
50	0.0499	0.0501	0.9808	0.7260	25.98	0.0496	0.0498	0.9288	0.2085	77.55
100	0.0496	0.0497	0.4926	0.3654	25.81	0.0498	0.0500	0.4663	0.1061	77.24

Note: Size is the number of acceptances by rejection sampling. Mean or MSE is the mean or mean squared error of the point estimates, and %Decrease is the percentage decrease in MSE achieved by the likelihood estimator, based on 7,500 simulations.

### 3. METROPOLIS-HASTINGS SAMPLING

We now turn to Metropolis-Hastings sampling. Given  $q(\cdot)$  and  $\rho(\cdot; x)$  for  $x \in \mathcal{X}$ , consider the following model:

- $y_t$  has the distribution  $\rho(\cdot; x_{t-1}) d\mu / \int \rho(y; x_{t-1}) d\mu$ ;
- $u_t$  has the distribution uniform  $(0, 1)$ ;
- accept  $x_t = y_t$  if  $u_t \leq 1 \wedge \beta(y_t; x_{t-1})$  and let  $x_t = x_{t-1}$  otherwise.

The likelihood of the process at  $\mu$  is proportional to

$$\prod_{i=1}^n \left[ \rho(y_i; x_{i-1}) \mu(\{y_i\}) / \int \rho(y; x_{i-1}) d\mu \right].$$

Under support and connectivity conditions (Vardi 1985), the maximum likelihood estimator has finite support  $\{y_1, \dots, y_n\}$  and satisfies

$$\hat{\mu}(\{y\}) = \frac{\hat{\Gamma}(\{y\})}{n^{-1} \sum_{j=1}^n \hat{Z}^{-1}(x_{j-1}) \rho(y; x_{j-1})},$$

where  $\hat{Z}(x_{j-1}) = \int \rho(y; x_{j-1}) d\hat{\mu}(y)$ , and  $\hat{\Gamma}$  is the empirical distribution placing mass  $n^{-1}$  at each of the points  $y_1, \dots, y_n$ . Next, we substitute the true value  $\int \rho(y; x_{j-1}) d\mu_0(y) = 1$  for  $\hat{Z}(x_{j-1})$  and obtain the closed-form estimator

$$\tilde{\mu}(\{y\}) = \frac{\hat{\Gamma}(\{y\})}{n^{-1} \sum_{j=1}^n \rho(y; x_{j-1})}.$$

Consequently, the integral  $Z = \int q(y) d\mu_0$  is estimated by

$$\tilde{Z} = \int q(y) d\tilde{\mu} = \sum_{i=1}^n \frac{q(y_i)}{\sum_{j=1}^n \rho(y_i; x_{j-1})}.$$

Note that the same estimator also holds for a real-valued integrand  $q(y)$ . The expectation  $E_p(\varphi) = \int \varphi(y)q(y) d\mu_0 / \int q(y) d\mu_0$  is estimated by

$$\begin{aligned} \tilde{E}(\varphi) &= \int \varphi(y)q(y) d\tilde{\mu} / \int q(y) d\tilde{\mu} \\ &= \sum_{i=1}^n \frac{\varphi(y_i)q(y_i)}{\sum_{j=1}^n \rho(y_i; x_{j-1})} / \sum_{i=1}^n \frac{q(y_i)}{\sum_{j=1}^n \rho(y_i; x_{j-1})}. \end{aligned}$$

Computations of  $\tilde{Z}$  and  $\tilde{E}(\varphi)$  are straightforward once  $\tilde{\mu}$  is evaluated.

Metropolis-Hastings sampling can be subsumed within the framework of Tan (2003), by identifying  $x_{i-1}$  as an index and  $y_i$  as a draw given the index. The sampling scheme basically provides a random design: an index  $x_{i-1}$  is stochastically selected and then a draw  $y_i$  is made from  $\rho(\cdot; x_{i-1})$  for  $1 \leq i \leq n$ . The estimator  $\tilde{Z}$  is a stratified importance sampling estimator using one observation  $y_i$  per distribution  $\rho(\cdot; x_{i-1})$ . The estimator  $\tilde{E}(\varphi)$  is a weighted Monte Carlo estimator: the observations  $y_i$  are given weights proportional to

$$w(y_i) = \frac{q(y_i)}{n^{-1} \sum_{j=1}^n \rho(y_i; x_{j-1})}.$$

Further, Tan (2003) proposed that the asymptotic variance of  $\tilde{Z}$  be estimated by

$$n^{-1} \int (w(y) - \tilde{Z})^2 d\hat{\Gamma},$$

and that of  $\tilde{E}(\varphi)$  be estimated by

$$n^{-1} \int (\varphi(y) - \tilde{E}(\varphi))^2 w^2(y) d\hat{\Gamma} / \tilde{Z}^2.$$

Despite their paradoxical appearance, the formulas can be justified with strong approximation of a Markov chain by a corresponding regression process.

Now suppose that a function  $q_1(y)$  is given on  $\mathcal{X}$  and its integral  $Z_1$  is 1 with respect to  $\mu_0$  as in reciprocal importance sampling. Following Cochran (1977), two alternative estimators of  $Z$  are the ratio estimator

$$\tilde{Z}_{\text{ratio}} = \sum_{i=1}^n w(y_i) / \sum_{i=1}^n w_1(y_i),$$

where  $w_1(y_i) = q_1(y_i)/[n^{-1} \sum_{j=1}^n \rho(y_i; x_{j-1})]$ , and the regression estimator

$$\tilde{Z}_{\text{reg}} = \tilde{Z} - \tilde{\beta}(\tilde{Z}_1 - 1),$$

where  $\tilde{\beta}$  is the regression coefficient of  $w(y_i)$  on  $w_1(y_i)$ ,  $1 \leq i \leq n$ . The asymptotic variance of  $\tilde{Z}_{\text{ratio}}$  can be estimated by

$$n^{-1} \int (w(y) - \tilde{Z}w_1(y))^2 d\hat{\Gamma},$$

and that of  $\tilde{Z}_{\text{reg}}$  can be estimated by

$$n^{-1} \int (w(y) - \tilde{Z} - \tilde{\beta}(w_1(y) - 1))^2 d\hat{\Gamma}.$$

The effect of variance reduction is such that these estimators have zero variance if  $q(y)$  is proportional to  $q_1(y)$  on  $\mathcal{X}$ . In the case of iid sampling, the regression estimator has no greater asymptotic variance than both the basic and ratio estimators. Kong et al. (2003) and Tan (2004) considered submodels that incorporate linear constraints on the baseline measure, and showed that regression estimators are first-order approximations to likelihood estimators in various situations.

If the detailed-balance equation is satisfied so that no proposals are rejected [i.e.,  $q(x)\rho(y; x) = q(y)\rho(x; y)$  for  $x, y \in \mathcal{X}$ ], then the average of successive proposal densities

$$\frac{1}{n} \sum_{j=1}^n \rho(y; x_{j-1})$$

converges pointwise to the stationary density  $p(y)$ , which is proportional to the integrand  $q(y)$ . The asymptotic proportionality suggests that the estimator  $\tilde{Z}$  can converge faster than at the standard rate  $n^{-1/2}$ . This super-efficiency for estimating the normalizing constant was observed by Kong et al. (2003) and Tan (2003). It remains an open question whether a super-efficient estimator of  $Z$  exists under general Metropolis-Hastings sampling.

### 3.1 BLOCKING AND SUBSAMPLING

In general, computation of  $\tilde{\mu}$  involves  $n^2$  evaluations of  $\rho(y_i; x_{j-1})$  for  $1 \leq i, j \leq n$ . [There are only  $n$  evaluations under independence Metropolis-Hastings sampling (Section 4).] This computation becomes intensive for large  $n$ , even though each individual  $\rho(y_i; x_{j-1})$  is easy to evaluate. Tan (2003) proposed a subsampling technique to reduce the computational cost. Here we compare it with a blocking technique.

We divide the Markov chain into  $m$  blocks each of length  $b$  ( $n = bm$ ). Applying  $\tilde{\mu}$  to the  $j$ th block  $(y_{(j-1)b+1}, y_{(j-1)b+2}, \dots, y_{jb})$ , we obtain

$$\tilde{\mu}_{\text{col}j}(\{y\}) = \frac{\hat{\Gamma}_{\text{col}j}(\{y\})}{b^{-1} \sum_{k=1}^b \rho(y; x_{(j-1)b+k-1})},$$

where  $\hat{\Gamma}_{\text{col}j}$  is the empirical distribution on the  $j$ th block. To use all the  $m$  blocks, we take the average

$$\tilde{\mu}_{\text{col}}(\{y\}) = \frac{1}{m} \sum_{j=1}^m \tilde{\mu}_{\text{col}j}(\{y\}).$$

Computation of  $\tilde{\mu}_{\text{col}}$  involves  $b^2m$  evaluations of proposal densities, only a fraction  $1/m$  of those for computation of  $\tilde{\mu}$ . In one extreme  $(b, m) = (n, 1)$ ,  $\tilde{\mu}_{\text{col}}$  becomes  $\tilde{\mu}$ . In the other extreme  $(b, m) = (1, n)$ ,  $\tilde{\mu}_{\text{col}}$  becomes  $n^{-1} \sum_{i=1}^n \rho^{-1}(y_i; x_{i-1}) \delta_{y_i}$ , where  $\delta_{y_i}$  denotes point mass at  $y_i$ , and the resulting estimator of  $Z$  is

$$\frac{1}{n} \sum_{i=1}^n \frac{q(y_i)}{\rho(y_i; x_{i-1})}.$$

This estimator is unbiased if the support of  $q(\cdot)$  is contained in that of  $\rho(\cdot; x)$ . Casella and Robert (1996) considered a similar estimator and its Rao-Blackwellization.

Alternatively, applying  $\tilde{\mu}$  to the  $i$ th subsampled sequence  $[(\xi_i, x_i), (\xi_{b+i}, x_{b+i}), \dots, (\xi_{(m-1)b+i}, x_{(m-1)b+i})]$ , we obtain

$$\tilde{\mu}_{\text{row}i}(\{y\}) = \frac{\hat{\Gamma}_{\text{row}i}(\{y\})}{m^{-1} \sum_{k=1}^m \rho(y; x_{(k-1)b+i-1})},$$

where  $\hat{\Gamma}_{\text{row}i}$  is the empirical distribution on the  $i$ th subsampled sequence. To use all the  $b$  subsequences, we take the average

$$\tilde{\mu}_{\text{row}}(\{y\}) = \frac{1}{b} \sum_{i=1}^b \tilde{\mu}_{\text{row}i}(\{y\}).$$

Computation of  $\tilde{\mu}_{\text{row}}$  involves  $bm^2$  evaluations of proposal densities, only a fraction  $1/b$  of those for computation of  $\tilde{\mu}$ . The resulting estimator of  $Z$  is unbiased if  $(b, m) = (n, 1)$  or if observations are independent every  $b$  iterations (subject to suitable support conditions).

In our simulation studies, a subsampled estimator generally has smaller mean squared error than a blocked estimator at equal computational cost. For subsampling, it is necessary that  $m$  be large enough, say  $\geq 50$ , and the mixture

$$\frac{1}{m} \sum_{k=1}^m \rho(\cdot; x_{(k-1)b+i-1})$$

cover sufficiently the target distribution  $p(\cdot)$ . As  $n$  increases to infinity,  $m$  can remain constant, which not only makes each subsampled sequence approximately independent, but also allows the computational cost to grow linearly.

### 3.2 INDEPENDENCE METROPOLIS-HASTINGS SAMPLING

Computation of  $\tilde{\mu}$  involves only  $n$  evaluations of  $\rho(y_i)$  under independence Metropolis-Hastings sampling. The likelihood estimator of  $E_p(\varphi)$  is in fact identical to the importance sampling estimator (2.2). Theorem 3 says that this estimator is asymptotically at least as efficient as the ergodic average (1.1). For Example 3.1 of Casella and Robert (1996), the likelihood estimator yields a 40–50% decrease in mean squared error over the ergodic average, which is similar to that achieved by Rao-Blackwellization.

**Theorem 3.** *The likelihood estimator (2.2) has no greater asymptotic variance than the ergodic average estimator (1.1) under independence Metropolis-Hastings sampling.*

As a corollary, the ratio estimator

$$\sum_{i=1}^n \frac{q(y_i)}{\rho(y_i)} / \sum_{i=1}^n \frac{q_1(y_i)}{\rho(y_i)}$$

has no greater asymptotic variance than the reciprocal importance sampling estimator (1.2). We can further reduce the variance by using the regression estimator

$$\frac{1}{n} \sum_{i=1}^n \frac{q(x_i)}{\rho(y_i)} - \tilde{\beta} \left( \frac{1}{n} \sum_{i=1}^n \frac{q_1(y_i)}{\rho(y_i)} - 1 \right),$$

where  $\tilde{\beta}$  is the regression coefficient of  $q(y_i)/\rho(y_i)$  on  $q_1(y_i)/\rho(y_i)$ ,  $1 \leq i \leq n$ . The likelihood method allows more efficient use of the fact that the integral of  $q_1(y)$  is 1 with respect to  $\mu_0$ .

### 3.3 EXAMPLES

First we present an example where analytical answers are available. Then we apply our method to Bayesian computation and provide two examples for logit regression and nonlinear regression.

#### 3.3.1 Illustration

Consider the bivariate normal distribution with zero mean and variance

$$V = \begin{pmatrix} 1 & 4 \\ 4 & 5^2 \end{pmatrix}.$$

Let  $q(x)$  be  $\exp(-x^\top V^{-1}x/2)$ . The normalizing constant  $Z$  is  $2\pi\sqrt{\det(V)}$ . In our simulations, the random walk Metropolis sampler is started at  $(0, 0)$  and run for 500 iterations. The proposal  $\rho(\cdot; x)$  is bivariate normal with mean  $x$  and variance  $(1.5)^2V$ .

We estimate  $\log Z$  by two reciprocal importance sampling estimators and five likelihood estimators (including the basic estimator, two ratio and two regression estimators); see Table 3. Two choices of  $q_1(x)$  are bivariate normal densities with zero mean and variance

Table 3. Comparison of Estimators of  $\log Z$

	RIS		LIK				
	Case 1	Case 2	Basic	Ratio 1	Ratio 2	Reg 1	Reg 2
Time Ratio	— < 0.01 —				— 1.9 —		
Bias	0.023	-0.0023	-0.0041	-0.00052	0.00096	0.0037	-0.00050
Std Dev	0.192	0.0471	0.0433	0.0316	0.0218	0.0219	0.0108
Sqrt MSE	0.194	0.0472	0.0435	0.0316	0.0218	0.0222	0.0108
Approx Err	NA	NA	0.0430	0.0311	0.0214	0.0219	0.0107

Note: Time Ratio is the ratio of CPU seconds for integral evaluation against Metropolis sampling. Bias, Std Dev, and Sqrt MSE are the bias, standard deviation, and  $\sqrt{\text{mean squared error}}$  of the point estimates, and Approx Err is  $\sqrt{\text{mean of the variance estimates}}$ , based on 5,000 simulations.

Table 4. Comparison of Estimators of Expectations

	$E(x^1)$		$pr(x^1 > 0)$		$pr(x^1 > 1.645)$	
	CMC	LIK	CMC	LIK	CMC	LIK
Bias	-0.00093	0.00012	-0.0023	0.00021	-0.00050	0.00011
Std Dev	0.122	0.0457	0.0562	0.0305	0.0228	0.00792
Sqrt MSE	0.122	0.0457	0.0562	0.0305	0.0228	0.00792
Approx Err	0.0442	0.0478	0.0222	0.0309	0.00965	0.00815

$(1.5)^2V$  and  $(0.8)^2V$ . The basic likelihood estimator assumes no knowledge of  $q_1(x)$  and is still more accurate than the better reciprocal importance sampling estimator. The regression estimator has mean squared error reduced by a factor of  $(.0472/.0108)^2 \approx 19$  compared with the better reciprocal importance sampling estimator. This factor is computationally worthwhile, because the total computational time (for simulation and evaluation) of the regression estimator is 2.9 times as large as that of the reciprocal importance sampling estimator.

We also estimate moments and probabilities by the crude Monte Carlo estimator and the likelihood estimator; see Table 4. The likelihood estimator has mean squared error reduced by an average factor of 7.1 for two marginal means, 7.8 for three second-order moments, and 4.5 for 38 marginal probabilities (ranging from 0.05 to 0.95 by 0.05), while it requires total computational time only 2.9 times as large as the crude Monte Carlo estimator.

The square root of the mean of the variance estimates is close to the empirical standard deviation for all the likelihood estimators. For comparison, the sample variance divided by  $n$  is computed as a variance estimator for the crude Monte Carlo estimator. As expected, the square root of the mean of these variance estimates is seriously below the empirical standard deviation of the crude Monte Carlo estimator.

Finally, we study the effects of blocking and subsampling with different  $(b, m)$ ; see Figure 1. The subsampled estimator of  $\log Z$  has nearly constant variance and negligible bias as  $m$  decreases from 500 to as small as 5. By contrast, the blocked estimator of  $\log Z$  has increasingly serious bias as  $b$  decreases from 500 to 5, and the bias is then reduced to zero as  $b$  decreases from 5 to 1. This pattern does not appear in the blocked estimators of moments and probabilities. Overall, a subsampled estimator has smaller mean squared error than a blocked estimator at equal computational cost.

### 3.3.2 Logit Regression

For the data in van Dyk and Meng (2001, Table 1), consider the logit regression

$$\text{logit pr}(y_i = 1) = x_i^\top \beta,$$

where  $y_i$  is the disease indicator and  $x_i$  is the vector of constant 1 and two covariates. Let the prior on  $\beta = (\beta^0, \beta^1, \beta^2)$  be trivariate normal with zero mean and variance

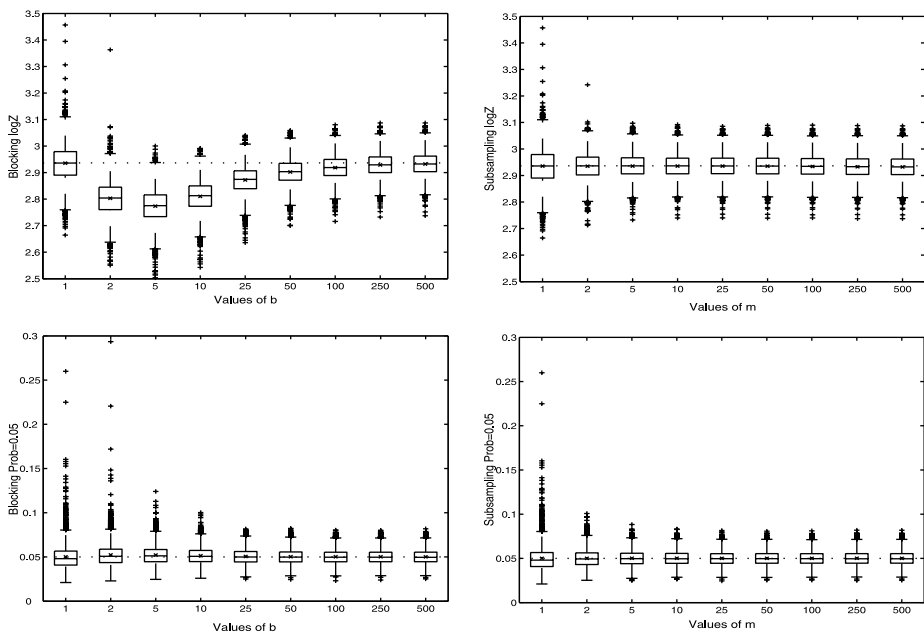


Figure 1. Boxplots of blocked and subsampled estimators.

diag(100<sup>2</sup>, 100<sup>2</sup>, 100<sup>2</sup>). The posterior density is proportional to

$$q(\beta) = \prod_{j=0}^2 \frac{1}{\sqrt{2\pi}100} e^{-(\beta^j/100)^2/2} \prod_{i=1}^{55} \left[ \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \right]^{y_i} \left[ \frac{1}{1 + \exp(x_i^T \beta)} \right]^{1-y_i}$$

We are interested in computing the normalizing constant and the posterior expectations.

The random walk Metropolis sampler is used in our 1,000 simulations of size 5,000. The proposal  $\rho(\cdot; \beta)$  is trivariate normal with mean  $\beta$  and variance  $2^2V$ , where  $V$  is the variance of the normal approximation. In Figure 2, we give the autocorrelation plot of each marginal chain and the scatterplot of  $(\beta^1, \beta^2)$  superimposed with the normal approximation contour (with relative levels 3, 1, 0.1, 0.01) from one of the simulations. We estimate  $\log Z$  by the

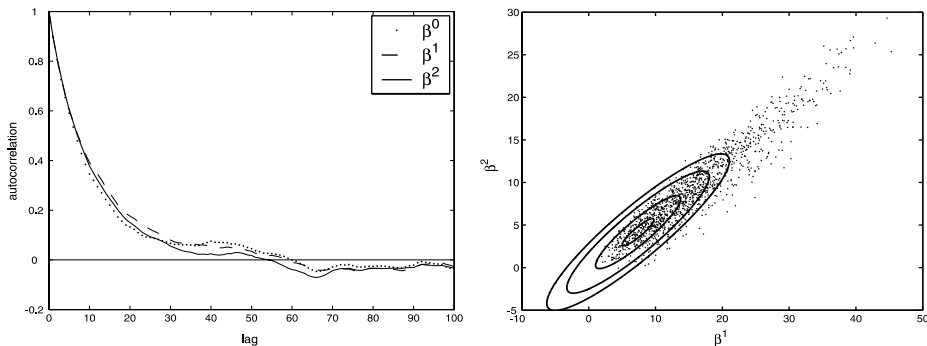


Figure 2. Autocorrelation plot and scatterplot of Metropolis draws.

Table 5. Comparison of Estimators of log Z

	$(b, m) = (1, 5000)$				$(b, m) = (100, 50)$		
	<i>RIS</i>	<i>Basic</i>	<i>Ratio</i>	<i>Reg</i>	<i>Basic</i>	<i>Ratio</i>	<i>Reg</i>
Time Ratio	<0.01	— 12.0 —			— .79 —		
Mean+17	-0.09184	-0.1677	-0.1658	-0.1668	-0.1664	-0.1655	-0.1661
Std Dev	0.124	0.0217	0.0193	0.0127	0.0217	0.0194	0.0128
Approx Err	NA	0.0216	0.0199	0.0129	0.0217	0.0201	0.0131

Table 6. Comparison of Estimators of Expectations

	$E(\beta^1 y) - 13$			$pr(\beta^1 > 25 y)$		
	<i>CMC</i>	$b = 1$	$b = 100$	<i>CMC</i>	$b = 1$	$b = 100$
Mean	0.5528	0.5674	0.5721	0.07232	0.07268	0.07287
Std Dev	0.491	0.145	0.162	0.0153	0.00372	0.00411
Approx Err	0.101	0.152	0.167	0.00366	0.00394	0.00430

Note: The true values are  $\log Z = -17.16604 \pm 0.00006$ ,  $E(\beta^1|y) = 13.5714 \pm 0.0005$ ,  $pr(\beta^1 > 25|y) = 0.07305 \pm 0.00002$ .

reciprocal importance sampling estimator, where  $q_1(x)$  is the normal approximation, and three likelihood estimators; see Table 5. The variances of the basic, ratio, and regression estimators are smaller than that of the reciprocal importance sampling estimator by a factor of 32.7, 40.9, and 93.5 under subsampling  $(b, m) = (100, 50)$ . The reduction factors are similar to those without subsampling. Moreover, the reciprocal importance sampling estimator is pseudo biased, because the lower left area in the scatterplot has negligible probability under the posterior but has nonnegligible probability under the normal approximation.

We also estimate the posterior expectations of  $(\beta^1, \beta^2)$  by the crude Monte Carlo estimator and the subsampled estimator  $[(b, m) = (100, 50)]$ ; see Table 6. The variance is reduced by an average factor of 9.1 for two means and 7.5 for 22 probabilities (ranging from 0.05 to 0.96). For this problem, the crude Monte Carlo estimator slightly underestimates the means and the upper tail probabilities of  $(\beta^1, \beta^2)$ .

### 3.3.3 NONLINEAR REGRESSION

Following Bates and Watts (1988), we consider the nonlinear regression model in which the response is normal with mean

$$E(y_i) = \beta^1(1 - e^{-\beta^2 x_i})$$

and variance  $\sigma^2$ , for the biochemical oxygen demand (BOD) data. We take the prior  $1/(360\sigma^2)$  on  $(\beta^1, \beta^2, \sigma^2)$  over the region  $(0, 60) \times (0, 6) \times (0, \infty)$ . After integrating

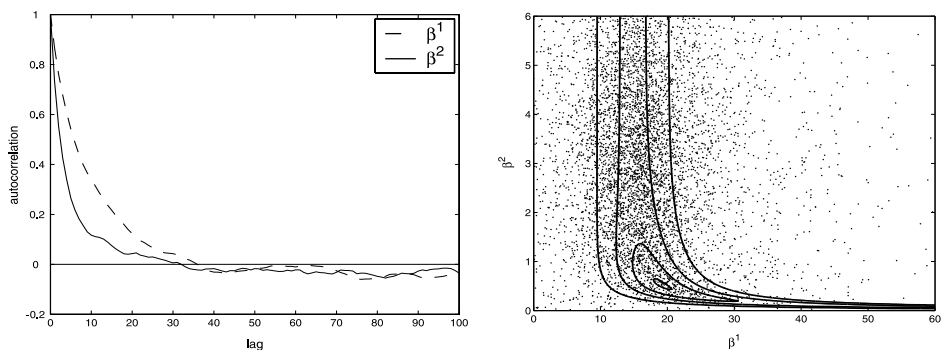


Figure 3. Autocorrelation plot and scatterplot of Metropolis draws.

out  $\sigma^2$ , the posterior density of  $\beta = (\beta^1, \beta^2)$  is proportional to

$$q(\beta) = \left[ \frac{1}{6} \sum_{i=1}^6 (y_i - \beta^1 (1 - e^{-\beta^2 x_i}))^2 \right]^{-3},$$

and has a nonelliptical contour. Bayesian computation for this problem was considered previously by several authors including DiCiccio et al. (1997) and Genz and Kass (1997).

The random walk Metropolis sampler is used in our 1,000 simulations of size 10,000. The proposal  $\rho(\cdot; \beta)$  is uniform over the intersection of  $(\beta^1 - 10, \beta^1 + 10) \times (\beta^2 - 3, \beta^2 + 3)$  and  $(0, 60) \times (0, 6)$ . In Figure 3, we give the autocorrelation plot of each marginal chain and the scatterplot of  $(\beta^1, \beta^2)$  superimposed with the posterior contour (with relative levels 1, 0.1, 0.01, 0.001). There appears to be no obvious sign of nonconvergence.

The ergodic average estimator is seriously biased even after 10,000 iterations for this problem. By contrast, the likelihood estimator has negligible bias and small variance; see Table 7. The accuracy is very competitive to those achieved by using subregion-adaptive quadrature (Genz and Kass 1997) and other Monte Carlo methods (DiCiccio et al. 1997). In Figure 4, we demonstrate the corrective effect of the likelihood method by overlaying the histogram of the 10,000 Metropolis draws, the weighted histogram of the 10,000 proposals, and the true marginal posterior density. In the weighted histogram, a bin's height is the ratio of the sum of the weights against the bin's width.

Table 7. Estimators of log Z and Posterior Means

$(b, m) = (100, 100)$ Time ratio = 1.6	$\log Z$	$E(\beta^1 y)$		$E(\beta^2 y)$	
	LIK	CMC	LIK	CMC	LIK
Mean	-3.5945	18.2927	18.7430	2.5647	1.1665
Std Dev	0.0446	0.5336	0.1816	0.0542	0.0299
Approx Err	0.0468	0.0869	0.1759	0.0167	0.0319

Note: The true values are  $\log Z = -3.5920 \pm 0.0002$ ,  $E(\beta^1|y) = 18.7789 \pm 0.0005$ ,  $E(\beta^2|y) = 1.1637 \pm 0.0001$ .

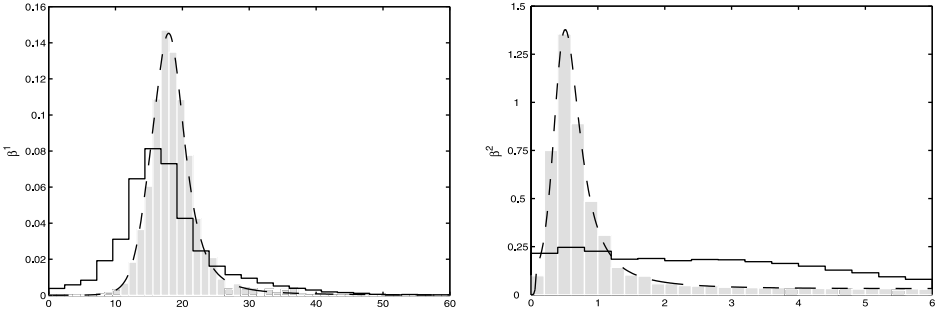


Figure 4. Histograms of Metropolis draws and weighted proposals. Dashed line: true density, solid line: Metropolis draws, bar: weighted proposals.

### 4. SUMMARY

First, consider rejection sampling or independence Metropolis-Hastings sampling, where only one proposal distribution  $\rho(\cdot)$  is involved. Given a sample  $y_1, \dots, y_n$  from this distribution, there are three possible estimators for an expectation  $E_p(\varphi)$  with respect to the target distribution  $p(\cdot)$ —the rejection sampling estimator (2.1), the independence Metropolis-Hastings sampling estimator (1.1), and the importance sampling estimator (2.2). The first two estimators require additional uniform random variables and acceptance-rejection operations. We establish that the importance sampling estimator is asymptotically most efficient among the three estimators. In fact, we derive the importance sampling estimator as a likelihood estimator in the approach of Kong et al. (2003).

Second, observations are generated and accepted/rejected from a family of proposal distributions under general Metropolis-Hastings sampling. While taking the likelihood approach of Kong et al. (2003), we basically treat the sampling scheme as a random design, and define a stratified estimator of the baseline measure. We employ a subsampling technique to reduce the computational cost in evaluating this estimator, and illustrate the computational effectiveness of the likelihood method with three examples. Further work is desirable on extension and application of this methodology to a broader range of problems such as hierarchical Bayesian computation.

### APPENDIX

**Proof of Lemma 1 and Theorem 1:** Using indicator functions, we write

$$L = \sum_{i=1}^n 1_{u_i \leq q(y_i)/[C\rho(y_i)]},$$

$$\sum_{i=1}^L \varphi(y_{t_i}) = \sum_{i=1}^n \varphi(y_i) 1_{u_i \leq q(y_i)/[C\rho(y_i)]}.$$

By the delta method (Ferguson 1996, sec. 7), the estimator (2.1) has asymptotic variance

$$n^{-1} \text{var} \left[ (\varphi(Y) - E_p(\varphi)) 1_{U \leq q(Y)/[C\rho(Y)]} \right] / \left( \frac{Z}{C} \right)^2,$$

where  $Y \sim \rho(\cdot)$  and  $U \sim \text{uniform}(0, 1)$  independently. By the relationship between conditional and unconditional variances,

$$\begin{aligned} \text{var} \left[ (\varphi(Y) - E_p(\varphi)) 1_{U \leq q(Y)/[C\rho(Y)]} \right] / \left( \frac{Z}{C} \right)^2 \\ &= E_p \left[ (\varphi - E_p(\varphi)) \frac{p}{\rho} \right]^2 + E_p \left[ (\varphi - E_p(\varphi))^2 \frac{p}{\rho} \left( \frac{C}{Z} - \frac{p}{\rho} \right) \right] \\ &= \frac{C}{Z} E_p \left[ (\varphi - E_p(\varphi))^2 \frac{p}{\rho} \right] = \frac{C}{Z} \text{var}_p(\varphi). \end{aligned} \tag{A.1}$$

The theorem follows from the decomposition (A.1). □

**Proof of Lemma 2 and Theorem 2:** The theorem follows from the lemma and the decomposition (A.1). To prove the lemma, we write

$$\sqrt{N} \left[ \frac{\sum_{i=1}^N \frac{\varphi(y_i)q(y_i)}{\rho(y_i)}}{\sum_{i=1}^N \frac{q(y_i)}{\rho(y_i)}} - E_p(\varphi) \right] = \frac{1}{\sqrt{N}} \sum_{i=1}^N \left[ (\varphi(y_i) - E_p(\varphi)) \frac{p(y_i)}{\rho(y_i)} \right] \cdot \frac{1}{\frac{1}{N} \sum_{i=1}^N \frac{p(y_i)}{\rho(y_i)}}.$$

By Doeblin-Anscombe theorem (Chow and Teicher 1988, sec. 9.4), the first factor converges to the normal distribution with zero mean and variance

$$\text{var}_\rho \left[ (\varphi - E_p(\varphi)) \frac{p}{\rho} \right].$$

By the law of large numbers, the numerator of the second factor converges to 1 with probability one. Moreover,  $N/l$  converges to  $C/Z$  with probability one because  $N$  is a sum of  $l$  independent geometric random variables with mean  $C/Z$ . Combining these three pieces and applying Slutsky's theorem, we conclude that

$$\sqrt{(C/Z)l} \left[ \frac{\sum_{i=1}^N \frac{\varphi(y_i)q(y_i)}{\rho(y_i)}}{\sum_{i=1}^N \frac{q(y_i)}{\rho(y_i)}} - E_p(\varphi) \right]$$

converges to the normal distribution with zero mean and the above variance. □

**Proof of Theorem 3:** We introduce the following notation. Denote by  $P$  the probability distribution with density  $p(\cdot)$ , and by  $\Gamma$  the probability distribution with density  $\rho(\cdot)$ , with respect to  $\mu_0$ . Denote  $w(x) = p(x)/\rho(x)$ . If the current state is  $x$ , the probability of rejection is  $\lambda(w(x))$ , where

$$\lambda(u) = \int_{w(y) \leq u} \left( 1 - \frac{w(y)}{u} \right) d\Gamma(y).$$

We show that  $\lambda(u)$  is bounded from below by  $1 - 1/u$ :

$$\lambda(u) = 1 - \frac{1}{u} + \int_{w(y) > u} \left( \frac{w(y)}{u} - 1 \right) d\Gamma(y) \geq 1 - \frac{1}{u}.$$

For  $k \geq 1$  and  $u \geq 0$ , define

$$T_k(u) = \int_u^\infty \frac{k\lambda^{k-1}(v)}{v^2} dv.$$

Then the  $k$ -step transition kernel of  $y$  given  $x$  is (Smith and Tierney 1996)

$$T_k(w(x) \vee w(y))P(dy) + \lambda^k(w(x))\delta_x(dy),$$

where  $\delta_x$  denotes point mass at  $x$  and  $w_1 \vee w_2 = \max\{w_1, w_2\}$ .

The boundedness of  $q(x)/\rho(x)$  implies that the IMH Markov chain is uniformly ergodic (Mengersen and Tweedie 1996, theorem 2.1). Then the asymptotic variance of the ergodic average (1.1) is

$$n^{-1} \left( \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \right),$$

where  $\gamma_0$  is the variance and  $\gamma_k$  is the lag- $k$  autocovariance under stationarity. By the formula of transition kernels,

$$\begin{aligned} \gamma_k &= \int \varphi^\dagger(x) \left[ \int \varphi^\dagger(y) T_k(w(x) \vee w(y)) dP(y) + \varphi^\dagger(x) \lambda^k(w(x)) \right] dP(x) \\ &= \iint \varphi^\dagger(x) \varphi^\dagger(y) T_k(w(x) \vee w(y)) dP(y) dP(x) + \int \varphi^{\dagger 2}(x) \lambda^k(w(x)) dP(x), \end{aligned}$$

where  $\varphi^\dagger(x) = \varphi(x) - E_p(\varphi)$ . First, the first term above is nonnegative:

$$\begin{aligned} &\iint \varphi^\dagger(x) \varphi^\dagger(y) T_k(w(x) \vee w(y)) dP(y) dP(x) \\ &= \iint \varphi^\dagger(x) \varphi^\dagger(y) \left[ \int_{w(x) \vee w(y)}^\infty \frac{k\lambda^{k-1}(u)}{u^2} du \right] dP(y) dP(x) \\ &= \iint \varphi^\dagger(x) \varphi^\dagger(y) \left[ \int_0^\infty 1_{u \geq w(x)} 1_{u \geq w(y)} \frac{k\lambda^{k-1}(u)}{u^2} du \right] dP(y) dP(x) \\ &= \int_0^\infty \frac{k\lambda^{k-1}(u)}{u^2} \left[ \iint 1_{u \geq w(x)} 1_{u \geq w(y)} \varphi^\dagger(x) \varphi^\dagger(y) dP(y) dP(x) \right] du \\ &= \int_0^\infty \frac{k\lambda^{k-1}(u)}{u^2} \left[ \int 1_{u \geq w(x)} \varphi^\dagger(x) dP(x) \right]^2 du \geq 0. \end{aligned}$$

Second, the sum of  $\gamma_0$  and twice the sum of the second term for  $k = 1, 2, \dots$  is no greater than the  $n^{-1}$ -normalized asymptotic variance of the estimator (2.2):

$$\begin{aligned} &\int \varphi^{\dagger 2}(x) dP(x) + 2 \sum_{k=1}^{\infty} \int \varphi^{\dagger 2}(x) \lambda^k(w(x)) dP(x) \\ &= \int \varphi^{\dagger 2}(x) \frac{1 + \lambda(w(x))}{1 - \lambda(w(x))} dP(x) \\ &\geq \int \varphi^{\dagger 2}(x) \frac{1}{1 - \lambda(w(x))} dP(x) \\ &\geq \int \varphi^{\dagger 2}(x) w(x) dP(x), \end{aligned}$$

because  $1/(1 - \lambda(w)) \geq w$ . We conclude the proof by combining these two results.  $\square$

## ACKNOWLEDGMENTS

This work was part of the author's doctoral thesis at the University of Chicago. The author is grateful to Peter McCullagh and Xiao-Li Meng for their advice and support.

[Received April 2005. Revised November 2005.]

## REFERENCES

- Bates, D. M., and Watts, D. G. (1988), *Nonlinear Regression Analysis and its Applications*, New York: Wiley.
- Casella, G., and Robert, C. P. (1996), "Rao-Blackwellization of Sampling Schemes," *Biometrika*, 83, 81–94.
- Chow, Y. S., and Teicher, H. (1988), *Probability Theory: Independence, Interchangeability, Martingales* (2nd ed.), New York: Springer.
- Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.
- DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of American Statistical Association*, 92, 902–915.
- Ferguson, T. S. (1996), *A Course in Large Sample Theory*, London: Chapman & Hall.
- Gelfand, A. E., and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal Royal Statistical Society, Series B*, 56, 501–514.
- Genz, A., and Kass, R. E. (1997), "Subregion-Adaptive Integration of Functions Having a Dominant Peak," *Journal of Computational and Graphical Statistics*, 6, 92–111.
- Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.
- Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., and Tan, Z. (2003), "A Theory of Statistical Models for Monte Carlo Integration" (with discussion), *Journal Royal Statistical Society, Series B*, 65, 585–618.
- Liu, J. S. (1996), "Metropolized Independent Sampling with Comparisons to Rejection Sampling and Importance Sampling," *Statistics and Computing*, 6, 113–119.
- Mengersen, K. L., and Tweedie, R. L. (1996), "Exact Convergence Rates for the Hastings and Metropolis Sampling Algorithms," *The Annals of Statistics*, 24, 101–121.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1091.
- Smith, R. L., and Tierney, L. (1996), "Exact Transition Probabilities for the Independence Metropolis Sampler," Technical Report, Department of Statistics, University of North Carolina.
- Tan, Z. (2003), "Monte Carlo Integration With Markov Chain," Working Paper, Department of Biostatistics, Johns Hopkins University.
- (2004), "On a Likelihood Approach for Monte Carlo Integration," *Journal of the American Statistical Association*, 99, 1027–1036.
- van Dyk, D., and Meng, X.-L. (2001), "The Art of Data Augmentation" (with discussion), *Journal of Computational and Graphical Statistics*, 10, 1–111.
- Vardi, Y. (1985), "Empirical Distributions in Selection Bias Models," *The Annals of Statistics*, 25, 178–203.
- von Neumann, J. (1951), "Various Techniques used in Connection With Random Digits," *National Bureau of Standards Applied Mathematics Series*, 12, 36–38.