Statistics and Computing (1996) **6**, 113–119

Metropolized independent sampling with comparisons to rejection sampling and importance sampling

JUN S. LIU

Department of Statistics, Stanford University, Stanford, CA 94305, USA

Received August 1994 and accepted March 1995

Although Markov chain Monte Carlo methods have been widely used in many disciplines, exact eigen analysis for such generated chains has been rare. In this paper, a special Metropolis–Hastings algorithm, *Metropolized independent sampling*, proposed first in Hastings (1970), is studied in full detail. The eigenvalues and eigenvectors of the corresponding Markov chain, as well as a sharp bound for the total variation distance between the nth updated distribution and the target distribution, are provided. Furthermore, the relationship between this scheme, rejection sampling, and importance sampling are studied with emphasis on their relative efficiencies. It is shown that Metropolized independent sampling is superior to rejection sampling in two respects: asymptotic efficiency and ease of computation.

Keywords: Coupling, delta method, eigen analysis, importance ratio

1. Introduction

Monte Carlo methods for evaluating integrals and simulating stochastic systems have been well understood and widely accepted for many years. More recently, people began to realize the potentials of Markov chain Monte Carlo, such as the Metropolis algorithm and the Gibbs sampler, in facilitating Bayesian inferences (Geman and Geman, 1984; Tanner and Wong, 1987; Gelfand and Smith, 1990). Many results on theoretical properties and applications of these methods have been obtained. See Smith and Roberts (1993) for an overview. However, sharp quantitative bounds on the convergence rates of these methods and their efficiency analyses are rare. An early efficiency analysis of Metropolis–Hastings algorithms is due to Peskun (1973). More recently, Liu *et al.* (1994) and Liu (1994) have provided efficiency comparisons for different Gibbs sampling schemes. In special cases, Lovasz and Simonovits (1990) obtain good bounds on the rate of convergence for their Metropolis algorithm; and Rosenthal (1991) did so for some Gibbs samplers.

One of the main tasks of sampling-based methods is to sample from a density or probability distribution function $\pi(x)$ that is usually complicated. Directly generating independent samples from such an arbitrary distribution is in general not possible. Several methods are available for indirect sampling. It is often the case that either the generated samples have to be dependent, or the distribution used to generate the sample has to be different from π. Rejection sampling (von Neumann, 1951) and importance sampling (Marshall, 1956) are schemes that use independent samples generated from a trial distribution similar to π. The Metropolis–Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) is, however, the one that generates dependent samples from a Markov chain with π as its equilibrium distribution.

In this paper we study a special Metropolis–Hastings algorithm, i.e. *Metropolized independent sampling*. This scheme was first discussed by Hastings (1970, Section 2.5) as one way to do importance sampling. Tierney (1991) generalizes the discussion under the heading 'independence chains', establishing irreducibility and aperiodicity of the Metropolized independent sampling kernel under a necessary and sufficient condition that the trial distribution p is positive almost everywhere. Tierney (1991) and Gelman and Rubin (1993) propose inserting a step of Metropolized independent sampling into Gibbs sampling iterations when correctly sampling from conditional distributions is

impossible. The idea is potentially useful in many practical problems with complicated hierarchical structures. An alternative to their proposal is to use rejection sampling directly. Besag and Green (1993) mention that a comparison of the two methods is of interest, but do not provide further analysis. Here we provide refined results for the Metropolis sampling chain: its eigenvalues, the corresponding eigenvectors, and an upper bound for the L^1 distance between the target and the updated distributions. As a byproduct, a study of relative efficiency of the three sampling plans—rejection sampling, importance resampling, and Metropolized independent sampling, is conducted.

Suppose we wish to generate samples from $\pi(x)$. What we have at hand, however, is a trial distribution $p(x)$, believed to be 'similar' to the target π, from which we can generate independent samples easily. Assume that $\mathcal{X}$ is the state space on which both distributions π and p are defined. We call π the target distribution, and p the trial distribution. Three schemes for generating samples from the target distribution π through the use of a trial distribution, $p(x)$, are described as follows.

Metropolis–Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970). The general scheme is described in Hastings (1970); here we concentrate on a special variation of the general algorithm, called *Metropolized independent sampling*. Suppose that $\pi(x)$ is known up to a norming constant, and we are able to draw independent samples from $p(x)$. A Markov chain $\{X_1, X_2, \ldots\}$ can be constructed with the transition function

$$K(x,y) = \begin{cases} p(y)\min\left\{1, \dfrac{w(y)}{w(x)}\right\} & \text{if } y \neq x, \\ 1 - \int_{z\neq x} p(z)\min\left\{1, \dfrac{w(z)}{w(x)}\right\}\mathrm{d}z & \text{if } y = x, \end{cases}$$

where $w(x) = \pi(x)/p(x)$ is called the *importance ratio* (or *importance weight*). Intuitively, the transition from $X_n = x$ to $X_{n+1} = y$ is accomplished by generating an independent sample from $p(\cdot)$, and then 'thinning' it down based on a comparison of the corresponding importance ratios $w(y)$ and $w(x)$. It can be shown that π is an invariant distribution of the constructed Markov chain. Note that the above scheme is only a special example, that more serious Metropolis–Hastings algorithms most commonly make dependent local moves instead of independent global jumps.

Rejection sampling (von Neumann, 1951). Suppose $\pi(x)$ is computable, and we can find a constant c such that the 'envelope property', i.e. $cp(x) \geq \pi(x)$, is satisfied. Then

(a) draw a sample X from $p(x)$ and compute the ratio $r = \pi(X)/\{cp(X)\}$;

(b) flip a coin with success probability r. If the head turns up, we accept the X; and if the tail turns up we reject the X.

It can be shown that the accepted sample follows the target distribution π.

Importance sampling

(a) Draw N independent values $X_1, \ldots, X_N$ from the trial distribution $p(\cdot)$.

(b) Calculate the importance ratios $w_i \propto \pi(X_i)/p(X_i)$ for each sampled value X_i in step (a).

(c) If $\mu = \mathrm{E}_\pi\{h(X)\}$, for any function $h(\cdot)$, is the quantity of interest, approximate it by

$$\hat{\mu} = \frac{w_1 h(X_1) + \cdots + w_N h(X_N)}{w_1 + \cdots + w_N}.$$

A detailed analysis of the Metropolized independent sampling chain, including its eigenvalues, eigenvectors, and bounds on variation distance, is provided in Section 2. In Section 3, a brief comparison between the Metropolis method and rejection sampling is made. Section 4 contains an efficiency analysis for importance sampling. Section 5 concludes the paper with a brief discussion.

2. Eigen analysis for Metropolized independent sampling

In this section, we assume that our sample space $\mathcal{X}$ is a finite set. Without loss of generality, we let $\mathcal{X} = \{1, \ldots, m\}$. Two probability measures $\pi(\cdot)$ and $p(\cdot)$ are then abbreviated as $\pi_i = \pi(i)$, and $p_i = p(i)$, $i = 1, \ldots, m$. We introduce the notation $F_\pi(k) = \pi_1 + \cdots + \pi_k$, $S_\pi(k) = 1 - F_\pi(k-1) = \pi_k + \cdots + \pi_m$; $F_p(k) = p_1 + \cdots + p_k$, and $S_p(k) = 1 - F_p(k-1)$.

Independent samples are generated from the trial distribution $p(\cdot)$, and thinned down by the Metropolis algorithm. More precisely, for any $i, j \in \mathcal{X}$, the transition probability from i to j for the Metropolis sampling can be written as

$$K(i,j) = \begin{cases} p_j \min\left\{1, \dfrac{w_j}{w_i}\right\}, & \text{if } j \neq i, \\ p_i + \Sigma_k p_k \max\left\{0, 1 - \dfrac{w_k}{w_i}\right\}, & \text{if } j = i, \end{cases}$$

where $w_i = \pi_i/p_i$ is defined as the *importance ratio*.

2.1. *Solution for eigenvalues and eigenvectors*

Without loss of generality, we assume that the states are sorted according to the magnitudes of their importance ratios, i.e. the elements in $\mathcal{X}$ are labelled so that

$$w_1 \geq w_2 \geq \cdots \geq w_m.$$

The transition matrix can then be written as

$$K = \begin{bmatrix} p_1 + \lambda_1 & \pi_2/w_1 & \pi_3/w_1 & \cdots & \pi_{m-1}/w_1 & \pi_m/w_1 \\ p_1 & p_2 + \lambda_2 & \pi_3/w_2 & \cdots & \pi_{m-1}/w_2 & \pi_m/w_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1 & p_2 & p_3 & \cdots & p_{m-1} + \lambda_{m-1} & \pi_m/w_{m-1} \\ p_1 & p_2 & p_3 & \cdots & p_{m-1} & p_m \end{bmatrix}$$

where

$$\lambda_k = \sum_{i=k}^{m} \left(p_i - \frac{\pi_i}{w_k} \right) = S_p(k) - \frac{S_\pi(k)}{w_k}, \tag{1}$$

which is just the probability of being rejected in the next step if the chain is currently at state k. For any function $f(x)$, we denote

$$f^+(x) = \begin{cases} f(x), & \text{if } f(x) > 0; \\ 0, & \text{if } f(x) \leqslant 0. \end{cases}$$

Then it is noted that λ_k has another expression, i.e.

$$\lambda_k = \sum_{i \geqslant k} \left(\frac{\pi_i}{w_i} - \frac{\pi_i}{w_k} \right) = \mathrm{E}_\pi \left\{ \frac{1}{w(X)} - \frac{1}{w_k} \right\}^+,$$

where the expectation is taken with respect to X. Apparently, if two states i and $i+1$ have equal importance ratios, then $\lambda_i = \lambda_{i+1}$. Let $\boldsymbol{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$ denote the column vector of trial probabilities, and $\boldsymbol{e} = (1, \ldots, 1)^{\mathrm{T}}$. Then K can be expressed as

$$K = G + \boldsymbol{e}\boldsymbol{p}^{\mathrm{T}},$$

where G is an upper triangular matrix of the form

$$G = \begin{bmatrix} \lambda_1 & \dfrac{p_2(w_2 - w_1)}{w_1} & \cdots & \cdots & \dfrac{p_{m-1}(w_{m-1} - w_1)}{w_1} & \dfrac{p_m(w_m - w_1)}{w_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{m-1} & \dfrac{p_m(w_m - w_{m-1})}{w_{m-1}} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

It is noted that $\boldsymbol{e}$ is a common right eigenvector for both K and $K - G$, corresponding to the largest eigenvalue 1. Since $K - G$ is of rank 1, the rest of the eigenvalues of K and G have to be the same. Hence the eigenvalues for K are $1 > \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{m-1}$. When m is fixed and the number of iterations goes to infinity, the mixing rate of this Metropolis Markov chain is asymptotically dominated by the second largest eigenvalue λ_1, which equals $1 - 1/w_1$. All the eigenvectors of G can be found explicitly. We first note that the vector $\tilde{\boldsymbol{v}}_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$ is a right eigenvector corresponding to λ_1. Checking one more step, we find that $\tilde{\boldsymbol{v}}_2 = (\pi_2, 1 - \pi_1, 0, \ldots, 0)^{\mathrm{T}}$ is a right eigenvector of λ_2. Generalizing, we have the following result.

Lemma 2.1 *The eigenvectors and eigenvalues of G are λ_k, and $\tilde{\boldsymbol{v}}_k = (\pi_k, \ldots, \pi_k, S_\pi(k), 0, \ldots, 0)^T$, for $k = 1, \ldots, m-1$, where there are k non-zero entries in $\tilde{\boldsymbol{v}}_k$.*

Proof When $j > k$, it is obvious that jth element of the vector $G \cdot \tilde{\boldsymbol{v}}_k$ is zero. The kth element is clearly $(G \cdot \tilde{\boldsymbol{v}}_k)_k = \lambda_k S_\pi(k)$. For $j < k$,

$$\begin{aligned} (G \cdot \tilde{\boldsymbol{v}}_k)_j &= \pi_k \lambda_j + \pi_k \frac{p_{j+1}(w_{j+1} - w_j)}{w_j} + \cdots \\ &\quad + \pi_k \frac{p_{k-1}(w_{k-1} - w_j)}{w_j} + S_\pi(k) \frac{p_k(w_k - w_j)}{w_j} \\ &= \pi_k \left\{ \frac{\pi_{j+1} + \cdots + \pi_{k-1}}{w_j} - (p_{j+1} + \cdots + p_{k-1}) \right. \\ &\quad \left. + \lambda_j + \frac{S_\pi(k)}{w_j} \right\} - p_k S_\pi(k) \end{aligned}$$

Here we use the fact that $w_k p_k = \pi_k$. Moreover, by the fact that $\lambda_j = S_p(j) - S_\pi(j)/w_j$, we find

$$(G \cdot \tilde{\boldsymbol{v}}_k)_j = \pi_k S_p(k) - p_k S_\pi(k) = \lambda_k \pi_k.$$

Thus $\tilde{\boldsymbol{v}}_k$ is a right eigenvalue corresponding to λ_k. □

Theorem 2.1 *For Metropolized independent sampling, all the eigenvalues for the transition matrix are $1 > \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{m-1} \geqslant 0$, where $\lambda_k = \Sigma_{i=k}^m (p_i - \pi_i/w_k) = E_\pi\{1/w(X) - 1/w_k\}^+$. The right eigenvector $\boldsymbol{v}_k$ corresponding to λ_k is*

$$\boldsymbol{v}_k \propto (0, \ldots, 0, S_\pi(k+1), -\pi_k, \ldots, -\pi_k)^{\mathrm{T}},$$

where there are $k-1$ zero entries.

Proof Since $K = G + \boldsymbol{e}\boldsymbol{p}^{\mathrm{T}}$, then $K\tilde{\boldsymbol{v}}_k = G\tilde{\boldsymbol{v}}_k + \boldsymbol{e}(\boldsymbol{p}^{\mathrm{T}}\tilde{\boldsymbol{v}}_k)$. It is further noted that

$$\boldsymbol{p}^{\mathrm{T}}\tilde{\boldsymbol{v}}_k = S_\pi(k)\pi_k + p_k S_p(k) = \pi_k(1 - \lambda_k).$$

Hence $K\tilde{\boldsymbol{v}}_k = \lambda_k \tilde{\boldsymbol{v}}_k + \pi_k(1 - \lambda_k)\boldsymbol{e}$. Since $\boldsymbol{e}$ is a right eigenvector of K with eigenvalue 1, we have, for any t,

$$K(\tilde{\boldsymbol{v}}_k - t\boldsymbol{e}) = \lambda_k \left\{ \tilde{\boldsymbol{v}}_k - \frac{t - \pi_k(1 - \lambda_k)}{\lambda_k} \boldsymbol{e} \right\}.$$

Solving $t = \{t - \pi_k(1 - \lambda_k)\}/\lambda_k$, we find that $\boldsymbol{v}_k = \tilde{\boldsymbol{v}}_k - \pi_k \boldsymbol{e}$ is a right eigenvector of K corresponding to λ_k. □

2.2. *Total variation bound for convergence*

The total variation distance between two distributions, say, p and π, is defined as

$$\|p - \pi\| = \frac{1}{2} \sum_{i=1}^{m} |p_i - \pi_i|.$$

A useful Cauchy–Schwarz type inequality to bound this distance between the target and the nth updated distributions can be obtained by making use of all the eigenfunctions and eigenvalues. The following lemma from Diaconis and Hanlon (1992) is useful.

Lemma 2.2 *Let $K(x,y)$ be the Markov transition function with $\pi(\cdot)$ as its invariant measure. Then for any starting state x, the total variation $4||K^{*n}(x,\cdot)-\pi(\cdot)||^2$ is bounded above by*

$$\sum_{y=1}^{m-1}\lambda_y^{2n}f_y^2(x)\leqslant\frac{\lambda_1^{2n}}{\pi(x)} \tag{2}$$

where $f_y(x)$ is the xth coordinate of the right eigenfunction f_y for eigenvector λ_y, and all $f_y(\cdot)$ are assumed to form an orthonormal basis in $L^2(\pi)$.

When the chain is started from any state $k>1$, an upper bound of $4||K^n(k,\cdot)-\pi(\cdot)||^2$ for the Metropolis chain can be obtained by renormalizing those eigenvectors obtained in Theorem 2.1:

$$b_k=\left\{\frac{1}{S_\pi(2)}-\frac{1}{S_\pi(1)}\right\}\lambda_1^{2n}+\cdots+\left\{\frac{1}{S_\pi(k)}-\frac{1}{S_\pi(k-1)}\right\}\lambda_{k-1}^{2n}+\left\{\frac{1}{\pi_k}-\frac{1}{S_\pi(k)}\right\}\lambda_k^{2n}.$$

For $k=1$, the above formula changes slightly:

$$b_1=\left\{\frac{1}{S_\pi(2)}-\frac{1}{S_\pi(1)}\right\}\lambda_1^{2n}+\cdots+\left\{\frac{1}{\pi_1}-\frac{1}{S_\pi(2)}\right\}\lambda_{m-1}^{2n}.$$

Hence, if the chain is started from a distribution $p(\cdot)$ instead of a point, an upper bound for the distance between the nth updated distribution $p^{(n)}$ and the target distribution π, is a weighted mixture of all b_k's. This is equivalent to the following theorem.

Theorem 2.2 *If, in running a Metropolized independent sampling chain, the starting distribution is $p(\cdot)$, then the total variation distance between updated and target distributions has an upper bound*

$$4||p^{(n)}-\pi||^2\leqslant\sum_{k=1}^{m}p_kb_k=\sum_{k=1}^{m-1}\left\{\frac{S_p(k)}{S_\pi(k)}-\frac{S_p(k+1)}{S_\pi(k+1)}+\frac{1}{w_k}\right\}\lambda_k^{2n}.$$

Since the eigenvalues have expressions as in Theorem 2.1, $\lambda_k=S_p(k)-S_\pi(k)/w_k$, it is seen that large eigenvalues, λ_k, $k\geqslant k_0$, can practically dominate the mixing rate of the Markov chain only when the associated masses π_k, $k\geqslant k_0$, and $p_k, k\geqslant k_0$ are substantially large.

Example 2.1 Consider the case when $p(\cdot)$ is uniform, i.e., $p_i=1/m$; and $\pi_k=(2m+1-2k)/m^2$. Then, $S_p(k)=(m-k+1)/m$, $k=1,\ldots,m$, $S_\pi(k)=(m-k+1)^2/m^2$, $k=1,\ldots,m$, and $w(k)=(2m+1-2k)/m$. Hence the eigenvalues are $\lambda_k=(m-k+1)(m-k)/m(2m+1-2k)$. Thus the upper bound for total variation distance is

$$\begin{aligned}4||p^{(n)}-\pi||^2&\leqslant\sum_{k=1}^{m-1}\left(\frac{m}{m-k+1}-\frac{m}{m-k}+\frac{m}{2m+1-2k}\right)\times\left\{\frac{(m-k+1)(m-k)}{m(2m+1-2k)}\right\}^{2n}\\&=\sum_{k=1}^{m-1}\left(\frac{m-k+1}{m}\right)^{2n-1}\left(\frac{m-k}{2m+1-2k}\right)^{2n}\times\left(\frac{m-k+1}{2m+1-2k}-\frac{1}{m-k}\right)\\&\leqslant\frac{1}{2^{2n+1}}\int_0^1 mx^{2n-1}\mathrm{d}x=\frac{m}{n4^{n+1}}.\end{aligned}$$

If we only use the second largest eigenvalue, however, the upper bound for $4||p^{(n)}-\pi||^2$ computed according to Lemma 2.2 is inflated to $m/4^n$ with the starting state being $[m/2]$.

Example 2. Suppose the target distribution $\pi(x)$ is binomial, i.e. $\mathrm{Bin}(m,\theta)$, and $p(x)=1/(m+1)$ is uniform. Then

$$w(x)=(m+1)\frac{m!}{x!(m-x)!}\theta^x(1-\theta)^{m-x}.$$

By standard normal approximation, we find

$$\max_{0\leqslant x\leqslant m}w(x)\approx\sqrt{\frac{m}{2\pi\theta(1-\theta)}}.$$

Hence the second largest eigenvalue $\lambda_1\approx 1-c_0/\sqrt{m}$. By Lemma 2.2, in order to make $4||K^{*n}(x,\cdot)-\pi(\cdot)||^2$ small when $x=[m/2]$, $n=O(\sqrt{m}\log(m))$ iterations are needed.

The coupling method, invented by Doeblin and treated in many monographs and books, for example Diaconis (1988) and Lindvall (1992), can also be used to resolve the convergence rate problem for Metropolized independent sampling.

Suppose that two such Metropolis chains $\{X_0,X_1,\ldots\}$ and $\{Y_0,Y_1,\ldots\}$ are started, of which the X-chain starts from a fixed point $X_0=x_0$ (or a distribution), and the Y-chain starts from the equilibrium distribution π. The transition for the two chains can be 'coupled' in the following way. Suppose $X_k=x$ and $Y_k=y$, then at stage $k+1$, a sample i is generated from the trial distribution $p(\cdot)$ and its associated importance ratio w_i is compared to w_x and w_y, respectively, to decide the values of X_{k+1} and Y_{k+1}. Precisely, an independent sample u is generated from Uniform[0, 1]. If $u\leqslant\min\{w_i/w_x,w_i/w_y\}$, then both chains accept i as their next states, i.e., $X_{k+1}=Y_{k+1}=i$; if, on the other hand, $u\geqslant\max\{w_i/w_x,w_i/w_y\}$, then both chains reject so that $X_{t+1}=x$ and $Y_{k+1}=y$; otherwise the chain with larger ratio accepts, and the chain with smaller ratio rejects. It is clear that the first time both chains accept a

'move' simultaneously is their *coupling time*, the time from which the realizations of the chains are identical.

For $X_k = x, Y_k = y$, the probability of simultaneous acceptance in the next step is

$$\Pr(\text{accept}|X_k = x, Y_k = y) = \sum_{i=1}^{m} p_i \min\left\{1, \frac{w_i}{w_x}, \frac{w_i}{w_y}\right\}$$
$$= \sum_{i=1}^{m} \pi_i \min\left\{\frac{1}{w_i}, \frac{1}{w_x}, \frac{1}{w_y}\right\} \geqslant \frac{1}{w_1},$$

where w_1 is the largest importance ratio. Hence from the Markov property the number of steps for the chains to be coupled is bounded by a geometric distribution

$$\Pr(N \geqslant n) \leqslant \left(1 - \frac{1}{w_1}\right)^n.$$

From the coupling inequality, we have

$$|\Pr(X_n = j) - \Pr(Y_n = j)| = |p_j^{(n)} - \pi_j| \leqslant \Pr(N \geqslant n)$$
$$\leqslant \left(1 - \frac{1}{w_1}\right)^n,$$

and $\|p^{(n)} - \pi\| \leqslant 2\Pr(N \geqslant n) \leqslant 2(1 - 1/w_1)^n$. This result makes the eigenvalue solutions concrete. Note that for Example 1, the coupling method gives a bound $\|p^{(n)} - \pi\| \leqslant 2[1 - (m-1)/(2m-1)]^n \approx 2^{-(n-1)}$, which, when $n > m/64$, is worse than $\sqrt{mn^{-1}}2^{-(n+2)}$, the bound we obtained previously using all the eigenvalues and eigenvectors. In Example 2 where we only used the largest eigenvalue, however, the forgoing coupling argument gives a better bound $(1 - c_0/\sqrt{m})^n$.

2.3 *Infinite state space and other extensions*

It is worth noting that extensions of the above results to an infinite state space are useful and relatively straightforward.

2.3.1. *Countable state space*

Without loss of generality, we assume that $\mathcal{X} = \{1, 2, \ldots, \}$. We also assume that p has a longer tail than π, that is, there exists a i_0 such that the importance ratio w_i, $i \geqslant i_0$, is monotone decreasing. Under this assumption, we can relabel the sample space so that the importance ratio is monotone decreasing: $w_1 \geqslant w_2 \geqslant \cdots$. Then the same argument as in the finite state space case works and leads to the same answer. The coupling argument works as well, so the result $\|p^{(n)} - \pi\| \leqslant 2(1 - 1/w_1)^n$ still holds.

3.2.2. *Continuous state space*

Suppose $p(\cdot)$ and $\pi(\cdot)$ are densities defined on a compact region B in R^n, $w(x) = \pi(x)/p(x)$ is the importance ratio at x, and let $w = \sup_{x \in B} \pi(x)/p(x)$. It is conjectured that the spectral gap between the largest and the second largest eigenvalues of the Markovian transition operator K is $1/w$, and the operator K has continuous spectrum

$$\lambda_x = \Pr\{w(Y) \leqslant w(x)\} - \pi\{w(Y) \leqslant w(x)\}/w(x)$$
$$= \mathrm{E}_\pi\left\{\frac{1}{w(Y)} - \frac{1}{w(x)}\right\}^+,$$

where the expectation is taken over Y.

Since the coupling argument in Section 2.1 and the upper bound $2(1 - 1/w_1)^n$ for $\|p^{(n)} - \pi\|$ does not depend on the number of states, this bound still holds for a continuous state space case. Specifically, one can approximate the two densities on the compact region B by discrete distributions and use a limiting argument. Smith (1994) obtains an exact expression for n-step transition probability of the Metropolized independent sampling by making use of the eigenvalues and eigenvectors provided in Section 2.1. He also extends his result to a continuous state space case, making some of the arguments here more rigorous.

3. Comparing rejection sampling and Metropolized independent sampling

Suppose that of interest is to estimate $\mu = \mathrm{E}_\pi\{h(X)\} \leqslant \infty$ by a Monte Carlo method, where it is assumed without loss of generality that $\mathrm{var}_\pi(h) = 1$. We define the *efficiency* of such a method as the reciprocal of the variance of the sample mean estimator of μ normalized by the size of the generated sample. For example, if N i.i.d. draws are generated from π to estimate μ, its efficiency is 1. In this section, we show that the *efficiency* of rejection sampling is at best $1/w_1$, where w_1 is the largest importance ratio defined previously, while that of Metropolized independent sampling is at least $1/w_1$.

Since $\pi(x) \leqslant w_1 p(x)$, the 'envelope condition' is satisfied by taking $c = w_1$, and w_1 is the best such constant. As was described in Section 1, to obtain samples from the target distribution π we successively draw independent random samples, x from $p(\cdot)$, and u from Uniform(0, 1), until we first encounter a pair such that $w_x/w_1 \leqslant u$. It is easily shown that such obtained samples follow the distribution π. Therefore if N independent samples have been generated from the trial distribution p, the expected sample size remaining after the rejection treatment is

$$N\left(p_1 \frac{w_1}{w_1} + \cdots + p_m \frac{w_m}{w_1}\right) = \frac{N}{w_1}.$$

Hence, the variance of the Monte Carlo estimate of the quantity μ by using rejection sampling is approximately

$$\mathrm{var}_\pi(\hat{\mu}_1) \approx \frac{w_1}{N}.$$

On the other hand, the function h has a spectral expansion, $h(x) = \mu + a_1 f_1(x) + \cdots + a_{m-1} f_{m-1}(x)$ where $f_i(x) \propto v_i$, $i = 2, \ldots, m$, are orthonormal eigenvectors corresponding to the set of egenvalues $\{\lambda_i, i = 1, \ldots, m-1\}$. Hence

$1 = \text{var}_\pi\{h(X)\} = a_1^2 + \cdots + a_{m-1}^2$. Let $X_1, \ldots, X_N$ be samples obtained from a stationary Metropolized independent sampling chain, then

$$\begin{aligned}
&\text{var}_\pi\left\{\frac{h(X_1) + \cdots + h(X_N)}{N}\right\} \\
&= \frac{1}{N^2}\left\{N\sum_{i=1}^{m-1} a_i^2 + 2(N-1)\sum_{i=1}^{m-1}\lambda_i a_i^2 + \cdots \right. \\
&\quad \left. + 2\sum_{i=1}^{m-1}\lambda_i^{N-1} a_i^2\right\} \\
&\leqslant \frac{1}{N}\sum_{i=1}^{m-1}(1 + 2\lambda_i + \cdots + 2\lambda_i^{N-1})a_i^2 \\
&\leqslant \frac{1}{N}\sum_{i=1}^{m-1}\frac{1+\lambda_i}{1-\lambda_i}a_i^2 \leqslant \frac{2\text{var}_\pi\{h(X)\}}{N(1-\lambda_1)} = \frac{2w_1}{N},
\end{aligned}$$

where the last inequality is strict if and only if one of $a_i \neq 0$ for some $\lambda_i < \lambda_1$. Hence we have shown that the asymptotic efficiency of rejection sampling is comparable with that of Metropolis sampling. Since $f_1 \propto \boldsymbol{v}_1$, it is interesting to see that

$$a_1 = \frac{\text{E}_\pi(\boldsymbol{v}_1 h)}{\sqrt{\text{var}_\pi(\boldsymbol{v}_1)}} = \sqrt{\frac{\pi_1}{1-\pi_1}}\{h(1) - \mu\},$$

where $h(1)$ is $h(x)$ evaluated at $x = 1$. In this manner, we obtain a lower as well as an upper bound for the asymptotic variance of the Monte Carlo estimate using Metropolis sampling:

$$\begin{aligned}
\frac{w_1\pi_1}{1-\pi_1}\{h(1) - \mu\}^2 &\leqslant \lim_{N\to\infty} N\text{var}_\pi\left\{\frac{h(X_1) + \cdots + h(X_N)}{N}\right\} \\
&\leqslant 2w_1\text{var}_\pi(h).
\end{aligned}$$

4. The 'rule of thumb' for importance sampling

Importance sampling suggests estimating

$$\mu = \text{E}_\pi\{h(X)\} = \sum_{i=1}^{m}\pi_i h(i),$$

by first generating independent samples $X_1, X_2, \ldots, X_N$ from an easy-to-sample distribution, $\boldsymbol{p}^{\text{T}} = (p_1, \ldots, p_m)$, and then calculating

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} h(X_k)\frac{\pi(X_k)}{p(X_k)}$$

to serve as an approximation. By properly choosing $p(\cdot)$, one substantially decreases the variance of the estimate. A good candidate for $p(\cdot)$ is one that is close to the shape of $h(x)\pi(x)$. Therefore, the importance sampling method can be super-efficient, that is, the resulting variance of $\hat{\mu}$ can be smaller than that obtained using independent samples from π. The method is generalized to the case of, say, evaluating $\text{E}_\pi\{h(X)\}$, where sampling from $\pi(\cdot)$ directly is difficult, but generating from $p(\cdot)$ and computing the importance ratio $w(X) = \pi(X)/p(X)$ are easy. The efficiency of such a method can be measured by a 'rule of thumb': the *effective sample size* resulting from an importance sampling procedure is approximately

$$\frac{N}{1 + \text{var}_p\{w(X)\}}.$$

This approximation can be justified by the delta method, based on a note of Kong (1992), as follows. Note that $\text{E}_p\{w(X)\} = 1$, hence

$$\tilde{\mu} = \frac{\frac{1}{N}\Sigma_{k=1}^{N} h(X_k)w(X_k)}{\frac{1}{N}\Sigma_{k=1}^{N} w(X_k)} = \frac{\bar{Z}}{\bar{W}}, \tag{3}$$

where $Z = h(X)w(X)$, and $W = w(X)$, is also a possible estimate of μ. There can be advantages for choosing (3): the importance sampling ratios only need to be evaluated up to an unknown constant; and (3) may have smaller MSE than $\hat{\mu}$. The variance of $\tilde{\mu}$ can be explored by using standard delta method for ratio statistics:

$$\text{var}_p(\tilde{\mu}) \approx \frac{1}{N}\{\mu^2\text{var}_p(W) + \text{var}_p(Z) - 2\mu\text{cov}_p(W, Z)\}. \tag{4}$$

Since

$\text{cov}_p(W, Z) = \text{E}_\pi(HW) - \mu = \text{cov}_\pi(W, H) + \mu E_\pi(W) - \mu$, where $H = h(X)$, and similarly

$$\begin{aligned}
\text{var}_p(Z) &= \text{E}_\pi(WH^2) - \mu^2 \\
&\approx \text{E}_\pi(W)E_\pi^2(H) + \text{var}_\pi(H)E_\pi(W) \\
&\quad + 2\mu\text{cov}_\pi(W, H) - \mu^2,
\end{aligned}$$

where the approximation is made based on the delta method involving the first two moments of W and H. It is easy to show that the remainder term in the above approximation is

$$\text{E}_\pi[\{W - \text{E}_\pi(W)\}(H - \mu)^2]. \tag{5}$$

By reformulating (4), we find that

$$\text{var}_p(\tilde{\mu}) \approx \text{var}_\pi(H)\{1 + \text{var}_p(W)\}/N.$$

Roughly speaking, if μ were estimated by $\hat{\mu}_0 = \Sigma_{i=1}^{N} h(Y_i)/N$ where $Y_i \sim \pi$, then the efficiency of $\tilde{\mu}$ relative to $\hat{\mu}_0$ is,

$$\frac{\text{var}_\pi\{h(Y)\}}{\text{var}_p\{h(X)w(X)\}} \approx \frac{1}{1 + \text{var}_p\{w(X)\}}.$$

Thus if $h(x)$ is relatively flat, the above relative efficiency can be approximated by the rule of thumb. Obviously, the rule of thumb approximation can be substantially off if the remainder term (5) is large. The advantage of the 'rule' is that it does not involve $h(X)$, which makes it particularly useful as a measure of the relative efficiency of

the method when many different h's are of potential interest.

It is seen that the variance of importance ratios plays an important role in measuring efficiency. Since $1 + \mathrm{var}_p(w) = w_1\pi_1 + w_2\pi_2 + \cdots + w_m\pi_m \leqslant w_1$, it is clear that

$$\frac{1}{1 + \mathrm{var}_p(w)} \geqslant \frac{1}{w_1},$$

where the equality holds only if π degenerates or $p \equiv \pi$. Hence, importance sampling can be asymptotically more efficient than rejection sampling in many cases.

5. Discussion

The above analysis suggests that importance sampling performs differently according to the function $h(x)$ of interest, whereas Metropolized independent sampling is asymptotically comparable to rejection sampling, in terms of the efficiency of estimation. The distribution of samples obtained from importance sampling is, however, always biased from the target, and the distribution of that obtained from Metropolized independent sampling theoretically converges to the target when sample sizes grow to infinity, but practically still differs from the target distribution. In contrast, the distribution of samples generated by rejection sampling is always the desired one. Nevertheless, a computational merit of both importance sampling and Metropolis sampling is that the envelope constant c need not be known, while rejection sampling requires knowing c in advance, which might be a substantial effort in some problems.

Acknowledgements

The author is very grateful to Professors Persi Diaconis and Neal Madras for inspiring discussions, and to Professor Augustine Kong for providing his unpublished manuscript. Part of the manuscript was prepared when the author was Assistant Professor, Department of Statistics, Harvard University. This research was also partially supported by NSF Grant DMS-9404344.

References

Besag, J. and Green, P.J. (1993). Spatial statistics and Baysian computation. *Journal of the Royal Statistical Society* B, **55**, 24–35.

Diaconis, P. (1988) *Group Representations in Probability and Statistics*, Lecture Notes-Monograph Series **11**, IMS, Hayward, California.

Diaconis, P. and Hanlon, P. (1992) Eigen analysis for some examples of the Metropolis algorithm. *Contemporary Mathematics*, **138**, 99–117.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A. and Rubin, D. B. (1993) Discussion on Gibbs sampler and other MCMC methods. *Journal of the Royal Statistical Society* B, **55**, 73–73.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–41.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Kong, A. (1992) A note on importance sampling using renormalized weights. *Technical report*, Department of Statistics, University of Chicago.

Lindvall, T. (1992) *Lectures on the Coupling Method*. Wiley, New York.

Liu, J. S. (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958–66.

Liu, J. S., Kong, A. and Wong, W. H. (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.

Lovasz, L. and Simonovits, M. (1990) The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. Preprint 27, Hungarian Academy of Sciences.

Marshall, A. W. (1956) The use of multi-stage sampling schemes in Monte Carlo computations. In *Symposium on Monte Carlo Methods*, ed. M. A. Meyer, pp. 123–40, Wiley, New York.

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**, 1087–91.

von Neumann, J. (1951) Various techniques used in connection with random digits. *National Bureau of Standards Applied Mathematics Series*, **12**, 36–8.

Peskun, P. H. (1973) Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, **60**, 607–12.

Rosenthal, J. S. (1995) Rates of convergence for Gibbs sampler for variance components models. *Ann. Statist.*, **23**, 740–61.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian Computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*, B, **55**, 3–23.

Smith, R. L. (1994) Exact transition probabilities for Metropolized independent sampling. Technical Report, Dept. Statistics, Univ. of North Carolina.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of American Statistical Association*, **82**, 528–50.

Tierney, L. (1991) Markov chains for exploring posterior distributions. In *Computer Science and Statistics: Proc. 23rd Symp. Interface*.

Yoida, K. (1978) *Functional Analysis*. Springer-Verlag, New York.