Statistica Sinica **15**(2005), 3-18

IMPROVING ON THE INDEPENDENT METROPOLIS-HASTINGS ALGORITHM

Yves F. Atchadé and François Perron

Harvard University and University of Montreal

Abstract: This paper proposes methods to improve Monte Carlo estimates when the Independent Metropolis-Hastings Algorithm (IMHA) is used. Our first approach uses a control variate based on the sample generated by the proposal distribution. We derive the variance of our estimator for a fixed sample size n and show that, as n tends to infinity, this variance is asymptotically smaller than the one obtained with the IMHA. Our second approach is based on Jensen's inequality. We use a Rao-Blackwellization and exploit the lack of symmetry in the IMHA. An upper bound on the improvements that we can obtain by these methods is derived.

Key words and phrases: Control variates, Metropolis-hastings algorithm, Rao-Blackwellization, symmetry.

1. Introduction

Let Π be a target distribution on some subset $\mathcal{X}$ of $\mathbb{R}^d$ equiped with its Borel subsets. In Markov Chain Monte Carlo (MCMC) we are interested in having a Markov chain (X_n) with stationary distribution Π. There are several algorithms to achieve this goal. For an introduction to Markov Chain Monte Carlo algorithms, we refer the reader to Robert and Casella (1999) or Liu (2001). The Independent Metropolis-Hastings (IMH) algorithm is among the most popular algorithms. This algorithm works best when there is another probability measure Q, easy to sample from, that is close to Π. Assume that Π is absolutely continuous with respect to Q and let $\omega/c_0 = d\Pi/dP$ be its Radon Nikodým derivative for some (maybe unknown) normalizing constant c_0. The IMH algorithm works as follows. Assume that $X_n = x$, then we sample independently $U_{n+1} \sim \mathcal{U}(0,1)$ and $Y_{n+1} \sim Q$, and X_{n+1} is set as:

$$X_{n+1} = x + (Y_{n+1} - x)\mathbf{1}_{[0,\alpha(x,Y_{n+1})]}(U_{n+1}),$$

where $\mathbf{1}_A$ is the indicator function of the set A and $\alpha(x,y) = \min(1, \omega(y)/\omega(x))$ ($\alpha(x,y) = 1$ if $\omega(x) = 0$).

This paper is about variance reduction. We are interested in the problem of estimating $\Pi(f) := \int f(x)\Pi(dx)$ for some integrable real-valued function f defined on $\mathcal{X}$. The basic estimator from the IMH sample (X_n) is given by:

$\hat{\mu}_0 = \sum_{i=1}^n f(X_i)/n$. The performance of $\hat{\mu}_0$ as an estimator of $\Pi(f)$ will depend on Π, ω, f. In general, Π and f are fixed by the problem at hand. Finding the best possible choice for ω in practice is still a largely unsolved problem. However, for ω fixed, it is possible to improve on the basic estimator $\hat{\mu}_0$. Typically, Q is a well-known distribution. This means that it is easy to generate random variables from Q but it is also easy to select a control variate, that is a function $h: \mathcal{X} \to \mathbb{R}$ such that $\mathrm{E}_Q(h(X)) = 0$ and $\mathrm{Var}_Q(h(X)) = 1$. Therefore, in our first approach, we propose and study the family of estimators $\hat{\mu}_\beta = \sum_{i=1}^n (f(X_i) - \beta h(Y_i))/n$. We study the estimator $\hat{\mu}_\beta$ in the slightly more general setting of a Markov chain defined recursively as $X_{n+1} = F(X_n, U_{n+1}, Y_{n+1})$ where (U_n) is a sequence of i.i.d. random variables with distribution $\mathcal{U}(0,1)$, (Y_n) a sequence of i.i.d. random variables with distrbution Q and $F: \mathcal{X} \times [0,1] \times \mathbb{R}^d \longrightarrow \mathcal{X}$ is a measurable function. In the case of the IMH algorithm, our main result says that $n\{\mathbb{V}\mathrm{ar}(\hat{\mu}_0) - \mathbb{V}\mathrm{ar}(\hat{\mu}_\beta)\} \to \beta_0^2 - (\beta - \beta_0)^2$ with $\beta_0 = \mathrm{Cov}_\Pi(f(X), h(X))$ as $n \to \infty$. Suprisingly, the expression for β_0 is rather simple in comparison to $\lim_{n\to\infty} n\mathbb{V}\mathrm{ar}(\hat{\mu}_0)$ and this is a key element in this paper. We also discuss the best possible choice for h. This leads us to propose the estimator:

$$\hat{\mu}_1 = \hat{\mu}_0 - \frac{\sum_{i=1}^n \omega(Y_i)\left(f_1(Y_i) - \Pi(f_1)\right)}{\sum_{i=1}^n \omega(Y_i)},$$

where f_1 is some function which is close to f and for which $\Pi(f_1)$ is known. If such function f_1 is not available, one can use $f_1 = f$ and run a pilot simulation to obtain a crude estimate of $\Pi(f)$. We use this approach in our examples. The results are very good, around 50% improvement over $\hat{\mu}_0$.

As a by-product of the study of $\hat{\mu}_\beta$ we derive a simple expression for $\mathbb{V}\mathrm{ar}(\hat{\mu}_0)$ for fixed n, not only for our problem but for any stationary and reversible Markov chain.

Our second approach uses Rao-Blackwellizations and symmetry. The idea of Rao-Blackwellization has been introduced in MCMC simulations by Gelfand and Smith (1990) in the context of the Gibbs sampler. Suppose that (X_n, Y_n) is a Markov chain from a Gibbs sampler with stationary distribution $\pi(x,y)$. Suppose that we want to estimate some marginal quantity $E_\pi(f(X))$. The usual estimator is $(1/n)\sum_{k=1}^n f(X_k)$. But, in the context of the Gibbs sampler, $E_\pi(f(X)|Y = y)$ is typically available and Gelfand and Smith (1990) introduced the Rao-Blackwellized version $(1/n)\sum_{k=1}^n E\left(f(X)|Y = Y_k\right)$, later proved by Liu (1994) to be always better than $(1/n)\sum_{k=1}^n f(X_k)$. Later on, Rao-Blackwellized versions of general Metropolis-Hastings samplers have been proposed by Casella and Robert (1996) in the form $\mathbb{E}((1/n)\sum_{k=1}^n f(X_k)|Y_1, \ldots, Y_n)$, where (X_n) is a sample from a Metropolis-Hastings algorithm and (Y_n) is the sequence of proposals. In Section 3, and following Perron (1999), we consider the IMH algorithm with the more efficient Rao-Blackwellized version $\mathbb{E}((1/n)\sum_{k=1}^n f(X_k)|Y_{(1)}, \ldots, Y_{(n)})$, where $Y_{(1)}, \ldots, Y_{(n)}$ are the order statistics. We also work on the symmetry of the problem. We can rearrange the estimator by introducing more symmetry and this will reduce its variance. The second major result of this paper shows that if we combine the approach based on a covariate with a Rao-Blackwellization, the variance reduction obtained by the two approaches will add when combined.

The rest of the paper is organized as followed. In Section 2, we study the control variate estimator $\hat{\mu}_\beta$. We apply our results to two examples. In Section 3, we consider the use of symmetry and Rao-Blackwellization. A simulation example is provided to illustrate our methods. Proofs are given in Section 4.

2. Variance Reduction with a Control Covariate

2.1. Control variates for Markov Chains

Throughout this paper, $\mathcal{X}$ denotes a nonempty subspace of $\mathbb{R}^d$ equiped with its Borel subsets. Let (X_n) be a $\mathcal{X}-$valued Markov chain defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$: $X_0 \sim \nu_0$ for some intial distribution ν_0 and, for $n \geq 1$,

$$X_n = F(X_{n-1}, U_n, Y_n), \tag{2.1}$$

where (U_n) is an i.i.d. sequence of uniformly distributed random variables, (Y_n) an i.i.d. sequence of $\mathbb{R}^d-$valued random variables with distribution Q, and $F : \mathcal{X} \times [0,1] \times \mathbb{R}^d \longrightarrow \mathcal{X}$ is a measurable function. Assume that (X_n) is ergodic with invariant distribution Π and that we are interested in the estimation of $\Pi(f) := \int f(x)\Pi(dx)$. It is well known that

$$\hat{\mu}_0 = \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \tag{2.2}$$

is an asymptotically unbiaised estimate of $\Pi(f)$. Given the recursive representation of (X_n), we wish to explore the possibility of using the i.i.d. sequence (Y_n) as a control variable to improve on $\hat{\mu}_0$. Commonly, much is known about the proposal distribution Q. For example we may assume that there is a real-valued function h such that $Q(h) = 0$ and $Q(h^2) = 1$. We propose the linear control variate estimate $\hat{\mu}_\beta = \hat{\mu}_0 - \beta(1/n)\sum_{k=0}^{n-1} h(Y_k)$.

For definiteness, let $\mathbb{E}$, $\mathbb{V}$ar and $\mathbb{C}$ov denote the expectation, variance and covariance, respectively, with respect to the probability measure $\mathbb{P}$. Whenever a probability distribution is known, we write $E_\nu(t(X))$ for the expectation of the random $t(X)$ when the law of X is ν. Let $P(x, A) := \mathbb{E}\left(X_n \in A | X_{n-1} = x\right)$ be the transition kernel of the chain (X_n). Clearly, $P(x, A) = \int Q(dy) \int_0^1 \mathbf{1}_A \left(F(x, u, y)\right) du$. P induces a linear operator K on real-valued function space by $Kf(x) := \int P(x, dy)f(y)$. Also, define $K_0 := K - \pi(\cdot)$. We assume that $\Pi(f^2) < \infty$ and

$$\sum_{k \geq 0} K_0^n f \quad \text{converges in} \quad L^2(\pi). \tag{2.3}$$

Define $(I-K_0)^{-1}f := \sum_{k\geq 0} K_0^n f$. The transition kernel $R(x,A) := \int \Pi(dz) \int_0^1 \mathbf{1}_A\left(F(z,u,x)\right) du$ will also play an important role in what follows. Actually, $R(x,A) = \mathbb{E}\left(X_n \in A | Y_n = x, X_0 \sim \Pi\right)$, and it induces a bounded operator T on real-valued functions $Tf(x) := \int R(x,dy)f(y)$. Note that for any $f \in L^2(\Pi)$, $Tf \in L^2(Q)$. Here is our main result.

Theorem 2.1. *Assume that (X_n) given by (2.1) is Harris ergodic with invariant distribution Π and that $f \in L^2(\Pi)$ satisfies (2.3).*

(i) *If $X_0 \sim \Pi$, then*

$$n\mathbb{V}ar\left(\hat{\mu}_0\right) = \mathrm{Cov}_{\Pi}\left((I-K_0)^{-1}(I+K_0)f(X), f(X)\right) - \frac{2}{n}\mathrm{Cov}_{\Pi}\left(f_{0,n}(X), f(X)\right), \tag{2.4}$$

$$n\mathbb{V}ar\left(\hat{\mu}_\beta\right) = n\mathbb{V}ar\left(\hat{\mu}_0\right) - \beta_0^2 + (\beta-\beta_0)^2 + \frac{2\beta}{n}\mathrm{Cov}_{Q}\left(Tf_{0,n}(X), h(X)\right), \tag{2.5}$$

where $f_{0,n} = (I-K_0)^{-2}(I-K_0^n)K_0 f$ and $\beta_0 = \mathrm{Cov}_Q(T(I-K_0)^{-1}f(X), h(X))$.

(ii) *Irrespective of the initial distribution of X_0, $n\left(\mathbb{V}ar\left(\hat{\mu}_\beta\right) - \mathbb{V}ar\left(\hat{\mu}_0\right)\right) \longrightarrow \beta_0^2 - (\beta-\beta_0)^2$.*

Proof. See Section 4.

Theorem 2.1 (ii) implies that the best (asymptotically) β for $\hat{\mu}_\beta$ to improve on $\hat{\mu}_0$ is β_0. This result is not very useful in practice because β_0 is not known and its estimation seems even more challenging than $\Pi(f)$. But it turns out that in the case of the IMH algorithm, some simplifications are possible.

Remark 2.1.

(i) The representation (2.1) can suit many MCMC Markov chains. For example, for the Independent Metropolis-Hastings algorithm (IMH), $F(x,u,y) = y\mathbf{1}_{[0,\alpha(x,y)]}(u) + x\mathbf{1}_{(\alpha(x,y),1]}(u)$; for the Random Walk Metropolis algorithm, $F(x,u,y) = (x+y)\mathbf{1}_{[0,\alpha(x,y)]}(u) + x\mathbf{1}_{(\alpha(x,y),1]}(u)$, where $\alpha(x,y) = \min(1, (\pi(x+y)/\pi(x)))$ with the obvious assumption that Π is absolutely continuous with respect to the Lebesgue measure with density π.

(ii) In the assumption (2.3), $\sum_{k\geq 0} K_0^n f$ is a solution of the so-called Poisson equation $g - K_0 g = f$. The purpose of the assumption is to guarantee that a central limit theorem holds for $(1/n)\sum_{k=0}^{n-1}\left(f(X_k) - \Pi(f)\right)$. This is satisfied in many practical situations. For example, if (X_n) is geometrically ergodic and $\Pi-$reversible then (2.3) holds for any $f \in L^2(\Pi)$ (see Roberts and Rosenthal (1997)). Note that $(I-K_0)\sum_{k\geq 0} K_0^n f = \left(\sum_{k\geq 0} K_0^n f\right)(I-K_0) = f$, where I is the identity operator of $L^2(\Pi)$. Therefore it makes sense to define $(I-K_0)^{-1}f := \sum_{k\geq 0} K_0^n f$. Also note that (2.3) implies that $\sum_{k\geq 0} K_0^n g$ converges in $L^2(\Pi)$ when g is any finite linear combination of the functions $K^i f$.

(iii) Formula (2.4) gives an alternative formula for the variance of $\hat{\mu}_0$ for fixed n. As $n \to \infty$, it follows from (2.3) that $f_{0,n}$ converges to some function in $L^2(\Pi)$. Therefore $n\mathbb{V}\mathrm{ar}\,(\hat{\mu}_0) - \mathrm{Cov}_\Pi\,((I-K_0)^{-1}(I+K_0)f(X), f(X))$ dies out at the rate $1/n$. From the definition of $(I-K_0)^{-1}f$, it easily follows that

$$\begin{aligned}&\mathrm{Cov}_\Pi\left((I-K_0)^{-1}(I+K_0)f(X), f(X)\right)\\ &= \mathbb{C}\mathrm{ov}(f(X_0), f(X_0)) + 2\sum_{k=1}^{\infty} \mathbb{C}\mathrm{ov}\left(K_0^k f(X_0), f(X_0)\right),\end{aligned}$$

which is the form the asymptotic variance for additive functional of Markov chains usually takes (Chan and Geyer (1994)).

2.2. Application to the independent metropolis-hastings sampler

Here we assume that Π is absolutely continuous with respect to Q and write $\omega(x)/c_0 = \Pi(dx)/Q(dx)$ for some normalizing constant c_0 that is not necessarily known. In the IMH algorithm, if $X_n = x$, a proposal move Y_{n+1} is sampled from Q and uniformly distributed random variable U_{n+1} is sampled. Then $X_{n+1} = F(X_n, U_{n+1}, Y_{n+1})$ where $F(x,u,y) = y\mathbf{1}_{[0,\alpha(x,y)]}(u) + x\mathbf{1}_{(\alpha(x,y),1]}(u)$. Clearly, if the acceptance rate of the algorithm is high, (X_n) and (Y_n) will be highly correlated.

Corollary 2.1. *Assume that IMH chain (X_n) is ergodic and $f \in L^2(\Pi)$ satisfies* (2.3).

(i) *One has $n\,(\mathbb{V}ar\,(\hat{\mu}_\beta) - \mathbb{V}ar\,(\hat{\mu}_0)) \longrightarrow \beta_0^2 - (\beta-\beta_0)^2$, where $\beta_0 = Cov_\Pi(f(X), h(X))$. The choice $\beta = \beta_0$ is asymptotically optimal.*

(ii) *With the choice $h(x) = (f(x) - E_Q(f(X)))/(Var_Q(f(X)))$, $\beta_0 > 0$, and for any $\beta \in (0, 2\beta_0)$, $\lim_{n\to\infty} n\,(\mathbb{V}ar\,(\hat{\mu}_\beta) - \mathbb{V}ar\,(\hat{\mu}_0)) < 0$.*

Proof. See Section 4.

Corollary 2.2. *The best possible choice for h for reducing the asymptotic variance of $\hat{\mu}_\beta$ is given by $h^*(y) = \omega(y)(f(y) - \Pi(f))/\sqrt{Var_Q\,(\omega(Y)(f(Y) - \Pi(f)))}$. For this choice of h we obtain: $\hat{\mu}_{\beta_0} = \hat{\mu}_0 - (1/c_0)((1/n)\sum_{i=1}^n \omega(Y_i)(f(Y_i) - \Pi(f)))$, where c_0 is the normalizing constant of ω.*

Proof. See Section 4.

Remark.

(i) First of all, Corollary 2.1 (i) shows that the applicability of control variates for the IMH algorithm as proposed in this paper is of the same order of difficulty as in the i.i.d. Monte Carlo setting, namely the computation of β_0. Indeed, if we choose $Q = \Pi$ so that $X_i = Y_i$, then our estimate $\hat{\mu}_\beta$ is exactly the classical control variate estimate in the i.i.d. Monte Carlo setting and the optimal β is β_0 as given above.

(ii) Corollary 2.1 (ii) suggests using the estimator

$$\hat{\mu}_{\beta_0} = \hat{\mu}_0 - \frac{\beta_0}{n}\sum_{k=1}^{n}\frac{(f(Y_k)-\eta)}{\sigma}, \tag{2.6}$$

where $\eta = \mathrm{E}_Q\left(f(X)\right)$ and $\sigma^2 = \mathrm{Var}_Q\left(f(X)\right)$. Typically, η and σ are known. Different strategies can be used to estimate $\beta_0 = \mathrm{Var}_\Pi(f(X))/\mathrm{Var}_Q(f(X))$: pilot simulation can be used; one can estimate β_0 directly from the sample (X_n) generated by the IMH algorithm (call this approach the in-line strategy). In our simulations we find this second approach interesting because it is computationally cheaper and still compares very well with the pilot simulation approach. But when the acceptance rate of the algorithm in low, these two estimates derived from (2.6) perform poorly.

(iii) Corollary 2.2 is not very useful either; because c_0 is not known and of course $\Pi(f)$ is not known. But it suggests that if we can find a function f_1 that is close in shape to f and for which $\Pi(f_1)$ is known then

$$\hat{\mu}_1 = \hat{\mu}_0 - \frac{\sum_{i=1}^n \omega(Y_i)\left(f_1(Y_i) - \Pi(f_1)\right)}{\sum_{i=1}^n \omega(Y_i)}$$

could perform well compared to $\hat{\mu}_0$. Finding such function in practice can be difficult. Another possible strategy is to run a pilot simulation to obtain an estimate $\hat{\Pi}(f)$ of $\Pi(f)$ and to use

$$\hat{\mu}_2 = \hat{\mu}_0 - \frac{\sum_{i=1}^n \omega(Y_i)(f(Y_i) - \hat{\Pi}(f))}{\sum_{i=1}^n \omega(Y_i)}. \tag{2.7}$$

In our simulation results below, we find that this strategy works quite well.

(iv) If (X_n) is geometrically ergodic then (2.3) is automatically satisfied for any $f \in L^2(\Pi)$. The geometric ergodicity of the IMH algorithm is well understood. It happens if and only if ω is essentially bounded (Mengersen and Tweedie (1996)).

2.3. Simulation examples

2.3.1. Bayesian estimation of a correlation ρ

This is Example 2.3 in Chen, Shao and Ibrahim (2000). Let $D = \{Y_1, \ldots, Y_n\}$, where $Y_i = (Y_{i,1}, Y_{i,2})'$ is an i.i.d. sample from $N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

We want to estimate ρ. We assume a flat prior $\mathcal{U}(-1,1)$ for ρ. Let $S_1 = \sum_{i=1}^n Y_{i,1}^2$, $S_2 = \sum_{i=1}^n Y_{i,1}Y_{i,2}$ and $S_3 = \sum_{i=1}^n Y_{i,2}^2$. The posterior density of ρ can be written

$$\pi(\rho|D) \propto (1-\rho^2)^{-n/2} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(S_1 - 2\rho S_2 + S_3\right)\right\}.$$

We make the change of variable $\rho = (-1+e^\tau)/(1+e^\tau)$ to obtain the following distribution on $\mathbb{R}$:

$$\pi(\tau|D) = \pi(\rho|D)\frac{2e^\tau}{(1+e^\tau)^2}.$$

For the simulations, we used data simulated from a $N(0,\Sigma)$ with $\rho = 0.5$.

Table 1. Simulated data set from $N(0,\Sigma)$ with $\rho = 0.5$.

Y_1	-1.066	0.274	1.257	-0.203	-0.420	1.328	0.255	-0.561	1.336	-0.536
Y_2	-1.468	-0.013	0.152	-0.597	0.137	2.130	-1.820	0.604	0.271	-0.900

To sample from $\pi(\tau|D)$, we used an IMH algorithm with a normal proposal having mean given by the maximum likelihood estimate of τ and variance equal to the inverse of the Fisher information. These quantities were estimated numerically. We ran a pilot simulation for $5,000$ iterations to obtain estimates $\hat{\tau}_0$ of $E_\pi(\tau)$ and $\hat{\theta}_0$ of $\text{Var}_\pi(\tau)$. These were $\hat{\tau}_0 = 0.883$ and $\hat{\theta}_0 = 0.305$.

We compare different estimators of τ and the results are presented in table 2. These results are based on 100 independent replications of the IMH algorithm each run for 5,000 iterations. The acceptance rate of the algorithm is around 60%. We normalize the proposal and use it as a control variate. The estimate $\hat{\tau}_1$ represents the basic estimate (2.2); $\hat{\tau}_2$ is the estimator $\hat{\mu}_{\beta_0}$ given in (2.6) where the optimal β_0 is estimated from the results of the pilot simulation; $\hat{\tau}_3$ is the same as $\hat{\tau}_2$ but the optimal β_0 is estimated from the same sample that estimates $\hat{\tau}_1$ (this is what we called the in-line strategy in the remarks and discussion of Section 2.2); $\hat{\tau}_4$ is the estimator in (2.7) where $\Pi(f)$ is estimated from the pilot simulation. In all our examples, we measure the improvement of an estimator $\hat{\mu}$ over the IMH algorithm estimator $\hat{\mu}_0$ by $1 - \text{Var}(\hat{\mu})/\text{Var}(\hat{\mu}_0)$.

Table 2. Performance of the different estimates of τ. The improvement are computed with respect to $\hat{\tau}_1$.

	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$
mean	0.868	0.868	0.868	0.883
Variance($10^{-3}\times$)	0.102	0.093	0.093	0.050
Improv. (in %)		8.56	8.58	50.6

Two strategies stand out from these results: $\hat{\tau}_3$ and $\hat{\tau}_4$. The estimate $\hat{\tau}_3$ is interesting because it does not use results from the pilot simulation, but it likely performs poorly if the acceptance rate of the algorithm is low. When information from a pilot simulation is available, $\hat{\tau}_4$ is expected to perform very well compared to $\hat{\tau}_1$.

2.3.2. A nonlinear regression example

We consider a data set given in Bates and Watts (1988, p.307). The data has been modeled by Newton and Raftery (1994) as follows: response Y_i to input X_i is

$$Y_i = \beta_1 + \frac{\beta_2}{1 + \exp\{-\beta_4(X_i - \beta_3)\}} + \varepsilon_i,$$

where ε_i are i.i.d. $N(0, \sigma^2)$ $i = 1, \ldots, 21$. Following Gilks, Roberts and Sahu (1998), we use the prior $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4)$. We integrate analytically over σ to obtain the posterior

$$\pi(\boldsymbol{\beta}|D) \propto \Big\{ \sum_{i=1}^{n} \Big(Y_i - \beta_1 - \frac{\beta_2}{1 + \exp\{-\beta_4(X_i - \beta_3)\}} \Big)^2 \Big\}^{-2-\frac{n}{2}}.$$

We are interested in the posterior mean of β_4. Although it is not the best sampler for this problem, we used the IMH algorithm with a normal proposal with mean given by the maximum likelihood of $\boldsymbol{\beta}$, computed numerically. The variance-covariance matrix of the proposal was zero off-diagonal and with diagonal equal to the inverse of the diagonal of the Fisher information matrix, also computed numerically. We ran 100 independent chains each with $8,000$ iterations. The acceptance rate of the algorithm was low, around 15%. As a control variate, we used the normalized version of the fouth component of the proposal. We ran a pilot simulation to estimate the optimal β_0 and $E(\beta_4)$. We obtained $\hat{\beta}_0 = 0.5$ and $\hat{\beta}_{4,0} = -1.285$. As in the last example, we compare different strategies. Now $\hat{\beta}_{4,1}$ is the basic estimate $\hat{\mu}_0$; $\hat{\beta}_{4,2}$ is the estimate (2.6) with the pilot estimate of β_0; $\hat{\beta}_{4,3}$ is the inline version of (2.6); $\hat{\beta}_{4,4}$ is (2.7) where $\hat{\Pi}(f) = \hat{\beta}_{4,0} = -1.285$, the estimate of $\Pi(f)$ from the pilot simulation.

Table 3. Performance of the different estimates of β_4. The improvements are computed with respect to $\hat{\beta}_{4,1}$.

	$\hat{\beta}_{4,1}$	$\hat{\beta}_{4,2}$	$\hat{\beta}_{4,3}$	$\hat{\beta}_{4,4}$
mean	-1.267	-1.267	-1.266	-1.260
Variance	0.0128	0.0129	0.0129	0.0068
Improv. (in %)		-0.42	-0.32	46.78

As expected, $\hat{\beta}_{4,2}$ and $\hat{\beta}_{4,3}$ performed poorly because in this case the acceptance rate of the algorithm was low, while $\hat{\beta}_{4,4}$ performed very well.

3. Variance Reduction via Rao-Blackwellizations and Symmetry

3.1. Variance reduction via symmetry

Consider the setting of the IMH algorithm given in the introduction. Let $\mathcal{S}$ be the group of permutations on $\{1,\ldots,n\}$. For $s \in \mathcal{S}$, $(s(1),\ldots,s(n))$ is a permutation of $(1,\ldots,n)$. The Metropolis-Hasting estimator is based on $((U_1,Y_1),\ldots,(U_n,Y_n))$. For any $s \in \mathcal{S}$ we say that $\tilde{\mu}_0(s)$ is equal to $\hat{\mu}_0$ evaluated at $((U_{s(1)},Y_{s(1)}),\ldots,(U_{s(n)},Y_{s(n)}))$. The estimators $\{\tilde{\mu}_0(s)\colon s \in \mathcal{S}\}$ are different from one another but they share the same distribution. For example, Y_1 may appears up to n times in $\tilde{\mu}_0((1,\ldots,n))$ but it can appear at most once in $\tilde{\mu}_0((2,3,\ldots,n,1))$.

Let π be a distribution on $\mathcal{S}$. Let $\pi\tilde{\mu}_0 = \sum_{s\in\mathcal{S}} \pi(s)\tilde{\mu}_0(s)$. Since all of the $\tilde{\mu}_0(s)$ have the same distribution, we obtain from Jensen's inequality that

$$\mathbb{E}[\pi\tilde{\mu}_0] = \mu \text{ and } \mathbb{V}\mathrm{ar}[\pi\tilde{\mu}_0] < \mathbb{V}\mathrm{ar}[\hat{\mu}_0],$$

if π is not a Dirac distribution. A natural thing to do is to consider the uniform distribution on $\mathcal{S}$. However, if we cannot find any simplifications, the algorithm will involve $n!$ evaluations and it will become untractable for large values of n. The simplifications that we have found so far do not help that much. A second approch would be to replace $\pi\tilde{\mu}_0$ by an approximation. For example, it could be a Monte Carlo simulation of fixed size or a numerical approach such as quasi Monte Carlo. A third approach that we shall develop consists of taking π to be the uniform distribution on some small subset of $\mathcal{S}$. Define $[i] = 1 + (i-1)\mathrm{mod}_n$ and let $\mathcal{S}_0 = \{(k,[k+1],\ldots,[k+n-1])\colon k = 1,\ldots,n\}$. Let $\tilde{\mu}_k = \tilde{\mu}_0((k,[k+1],\ldots,[k+n-1]))$. Now consider

$$\begin{aligned}
z(i) &= \sum_{j=0}^{n-1}\prod_{k=0}^{j} \mathrm{I}(U_{[i+k]}w(X_0) > w(Y_{[i+k]})),\\
m(i) &= 1 + \sum_{j=1}^{n-1}\prod_{k=1}^{j} \mathrm{I}(U_{[i+k]}w(Y_i) > w(Y_{[i+k]})),\\
s_j(i) &= \begin{cases} z([i]) & \text{if } j = 1,\\ s_{j-1}(i) + m([i+s_{j-1}(i)]) & \text{if } j > 1,\end{cases}\\
\ell(i) &= \max\{j\colon s_j(i) \le n, j \ge 1\}.
\end{aligned}$$

We obtain

$$\tilde{\mu}_k = \frac{1}{n}\{z(k)h(X_0) + \sum_{j=1}^{\ell(k)} m([k+s_j(k)])h(Y_{[k+s_j(k)]}) - (s_{\ell(k)+1}(k)-n)h(Y_{[k+s_{\ell(k)}(k)]})\}$$

and, if π is the uniform distribution on $\mathcal{S}_0$, then $\pi\tilde{\mu}_0 = 1/n \sum_{k=1}^{n} \tilde{\mu}_k$.

3.2. Variance reduction via Rao-Blackwellization

Cesella and Roberts (1996) suggest taking the conditional expectation of the Metropolis-Hastings estimator given that $Y_1, \ldots, Y_n$ are fixed. Let us call this estimator $\hat{\mu}_0^*$. It is easy to improve $\hat{\mu}_0^*$ by simply considering the conditional expectation given that $Y_{(1)}, \ldots, Y_{(n)}$ are fixed, as Perron (1999) did for a different problem. However, in our context, conditioning on the order statistics will involve too many calculations, infeasable for large values of n. Here again, the evaluation of $\hat{\mu}_0^*$ on $(Y_1, \ldots, Y_n)$ will be different than the evaluation of $\hat{\mu}_0^*$ on $(Y_2, Y_3, \ldots, Y_1)$. Let us say that $\tilde{\mu}_k^*$ is equal to $\hat{\mu}_0^*$ evaluated at $(Y_k, Y_{[k+1]}, \ldots, Y_{[k+n-1]})$, for example, $\tilde{\mu}_1^* = \hat{\mu}_0^*$. In general, if we set

$$\begin{aligned}
p^*(i,j) &= 1 \wedge w(Y_j)/w(Y_i) \text{ for } i,j = 1, \ldots, n,\\
p^*(0,j) &= 1 \wedge w(Y_j)/w(X_0)) \text{ for } j = 1, \ldots, n,\\
q^*(i,j) &= 1 - p^*(i,j),\\
f^*(i,\ell) &= \prod_{j=1}^{\ell} q^*(i,[i+j]) \text{ with } f^*(i,0) = 1,\\
\varphi^*(k,\ell) &= \prod_{j=1}^{\ell} q^*(0,[k+j-1]),\\
\delta^*(k,1) &= p^*(0,[k]) \text{ and for } \ell > 1,\\
\delta^*(k,\ell) &= \varphi^*(k,\ell-1)p^*(0,k+\ell-1)\\
&\quad + \sum_{j=1}^{\ell-1} \delta^*(k,j) f^*([k+j],\ell-j-1) p^*([k+j],[k+\ell]),
\end{aligned}$$

we can write

$$\tilde{\mu}_k^* = \frac{1}{n} \sum_{\ell=1}^{n} \{\varphi^*(k,\ell)h(X_0) + [\sum_{j=0}^{n-\ell} f^*([k+\ell],j)]\delta^*(k,\ell)h(Y_{[k+\ell]})\}$$

and, if π is the uniform distribution on $\mathcal{S}_0$, then $\pi\tilde{\mu}_0^* = \frac{1}{n} \sum_{k=1}^{n} \tilde{\mu}_k^*$.

Remark 3.1. As is mentioned in Casella and Robert (1996) the evaluations of f^* are very time consuming. For $\hat{\mu}_0^*$, $f^*(i,\ell)$ has to be evaluated at $i = 1, \ldots, n$, $\ell = 1, \ldots, n-i$. For $\pi\tilde{\mu}_0^*$, $f^*(i,\ell)$ has to be evaluated at i, $\ell = 1, \ldots, n$. Thus, even if $\pi\tilde{\mu}_0^*$ is an average of n estimators similar to $\hat{\mu}_0^*$ it requires only twice the number of evaluations of f^* than $\hat{\mu}_0^*$.

Now we shall see that the reduction in the variance produced by a Rao-Blackwellization is rather limited. In fact, $\mathrm{Var}[\hat{\mu}_0^*]$ is of the order of $O(1/n)$.

Remark 3.2. If we take into account how time consuming is the Rao-Blackwellization for large values of n, it is better to increase the sample size than to perform a Rao-Blackwellization when we want to reduce the variance in case n is large. A better strategy might be to run several parallel chains based on small sample sizes with Rao-Blackwellizations instead of running one chain with a large sample size.

Lemma 3.1. *Let Z be a satistic based on $(U_1, Y_1), \dots, (U_n, Y_n)$ such that the vector of the order statistics $(Y_{(1)}, \dots, Y_{(n)})$ is a function of Z, and assume that h is a covariate. We obtain that*

$$\mathbb{C}\mathrm{ov}\Big[\mathbb{E}\Big(\sum_{i=1}^{n} f(X_i)|Z\Big), \sum_{j=1}^{n} h(Y_j)\Big] = \mathbb{C}\mathrm{ov}\Big[\sum_{i=1}^{n} f(X_i), \sum_{j=1}^{n} h(Y_j)\Big].$$

Proof. See Section 4.

Theorem 3.1. *Under the conditions of Lemma* 3.1*,* $\liminf_{n\to\infty} n\mathbb{V}ar\,[\mathbb{E}[\hat{\mu}_0|Z]] \geq \mathrm{Var}_{\,Q}\,[\omega(Y)(f(Y)-\mu)]$.

Proof. See Section 4.

Remark 3.3. Finally, Lemma 3.1 tells us that it is possible to combine the result of this section with that of the previous section. It suggests the estimator $\pi\hat{\mu}^*_{\beta_0} = \pi\tilde{\mu}^*_0 - \beta_0\bar{g}$. Moreover, from Lemma 3.1 we see that the improvement of $\pi\hat{\mu}^*_{\beta_0}$ over $\hat{\mu}_0$ is the improvement due to the use of a covariate plus the improvement due to the use of symmetry combined with the Rao-Blackwellization.

3.3. Simulation example: Computation of the mean μ of a Gamma distribution

This example comes from Casella and Roberts (1996) and will be used to illustrate the methods developed in Section 3. The target is a Gamma distribution with pdf $\pi(x) \propto x^\alpha e^{-x}$. We use a proposal with the same mean but easier to sample from: $Q(x) \propto xe^{-(2/\alpha)x}$. Clearly, if α is close to 2, we shall have very few rejections in the IMH algorithm. In these situations, the method using a covariate will be very efficient for reducing the variance. However, if α is much larger than 2 then we shall have many repetitions in the IMHA and the methods developed in Section 3 will take advantage of these repetitions. We want to compute μ, the mean of π. A possible control variate is $h(x) = \sqrt{2}(x-\alpha)/\alpha$ so $\mathrm{E}_Q\,(h(X)) = 0$, $\mathrm{Var}_{\,Q}\,(h(X)) = 1$. Here the (asymptotically) optimal value of β can be computed exactly, $\beta_0 = \sqrt{2}$. Our analysis is based on an estimation of these variances based on 10,000 replications. The parameter α will vary from 2.2 to 22.0.

Figures 1 and 2 present the improvements (computed as $1-\mathbb{V}\mathrm{ar}(\hat{\mu})/\mathbb{V}\mathrm{ar}(\hat{\mu}_0)$) of each method for different values of the parameter α. In Figure 1 we consider the case $n = 100$. This graph shows that the covariate approach is very good when the proposal is close to the target and the method provides a small improvement even when there is a big difference between the proposal and the target. Finally, the estimation of the best possible choice for β_0 works well. Initially, we had considered the case $n = 1,000$, but in this situation it was too difficult to see a difference between the case where β_0 is fixed and the one where it is estimated.

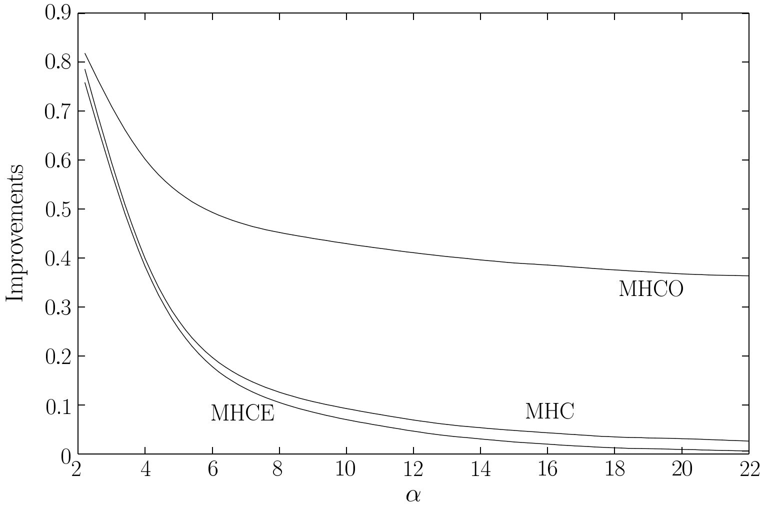

Figure 1. Improvements of the control variate estimators over the classical estimator.
MHC: optimal β.
MHCE: estimated β.
MHCO: optimal β optimal covariate.

In Figure 2 we study the different estimators proposed in this paper for the case $n = 25$. The case $n = 25$ has been chosen to keep Rao-Blackwellization manageable. We see that the Rao-Blackwellizations helps a lot when there are many rejections in the IMHA and, to a lesser extent, this is also true for the symmetric versions. We see also, graphically, that the improvement given by the approach developed in Section 2 is added to the one given by an approach developed in Section 3 when the two approaches are combined. It is surprising that the use of a covariate is still good, considering that the choice of the covariate has been made on the basis of asymptotic considerations. If we take into account evaluation time and algorithm complexity, perhaps the use of a symmetric version with a covariate would be a nice approach.

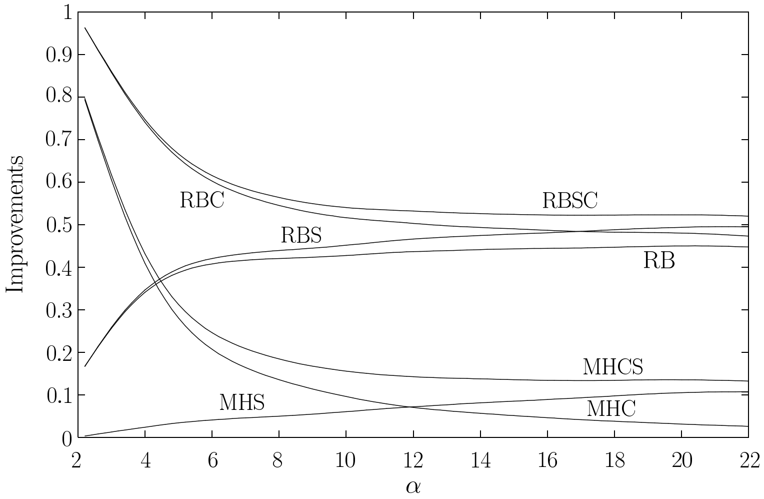

Figure 2. Improvements of the new estimators over the classical estimator. MHC: control variate. MHS: symmetry. MHCS: control variate and symmetry.
RB: Rao-Blackwellization. RBC: Rao-Blackwellization and control variate.
RBSC: Rao-Blackwellization and symmetry and control variate.
RBS: Rao-Blackwellization and symmetry.

4. Proofs of the Results

Proof of Theorem 2.1. First, it is easy to verify that

$$\sum_{\ell=0}^{n-1}(n-\ell)K_0^{\ell} = n(I-K_0)^{-1} - (I-K_0)^{-2}(I-K_0^n)K_0.$$

(i) Since $X_0 \sim \Pi$, (X_n) is stationary and we have

$$\begin{aligned}
n\mathbb{V}\mathrm{ar}[\hat{\mu}_0] &= \frac{2}{n}\sum_{1\le i\le j\le n}\mathbb{C}\mathrm{ov}[f(X_i),f(X_j)] - \frac{1}{n}\sum_{1\le i\le n}\mathbb{V}\mathrm{ar}[f(X_i)] \\
&= \frac{2}{n}\sum_{\ell=0}^{n-1}(n-\ell)\mathbb{C}\mathrm{ov}[f(X_{\ell+1}),f(X_1)] - \mathbb{V}\mathrm{ar}[f(X_0)] \\
&= \frac{2}{n}\mathbb{C}\mathrm{ov}\left[\sum_{\ell=0}^{n-1}(n-\ell)K_0^{\ell}f(X_0),f(X_0)\right] - \mathbb{V}\mathrm{ar}[f(X_0)] \\
&= \mathbb{C}\mathrm{ov}\left[(I-K_0)^{-1}(I+K_0)f(X_0),f(X_0)\right] \\
&\quad -\frac{2}{n}\mathbb{C}\mathrm{ov}\left[(I-K_0)^{-2}(I-K_0^n)K_0f(X_0),f(X_0)\right].
\end{aligned}$$

For $\hat{\mu}_\beta$, we have

$$n\mathbb{V}\text{ar}[\hat{\mu}_\beta] = n\mathbb{V}\text{ar}[\hat{\mu}_0] - 2\frac{\beta}{n}\mathbb{C}\text{ov}\left[\sum_{i=1}^{n} f(X_i), \sum_{j=1}^{n} h(Y_j)\right] + \beta^2.$$

For $\ell \geq 0$, taking the conditional expectation gives

$$\begin{aligned}\mathbb{E}[f(X_{\ell+1})|Y_1 = y] &= \mathbb{E}\left[\mathbb{E}\left(f(X_{\ell+1})|Y_1 = y, X_1\right)|Y_1 = y\right]\\ &= \mathbb{E}\left[E\left(f(X_{\ell+1})|X_1\right)|Y_1 = y\right]\\ &= \mathbb{E}\left(K^\ell f(X_1)|Y_1 = y\right)\\ &= TK^\ell f(y).\end{aligned}$$

Using this expression we obtain

$$\begin{aligned}\mathbb{C}\text{ov}\Big[\sum_{i=1}^{n} f(X_i), \sum_{j=1}^{n} h(Y_j)\Big] &= \sum_{1\leq j\leq i\leq n} \mathbb{C}\text{ov}[f(X_i), h(Y_j)]\\ &= \sum_{\ell=0}^{n-1}(n-\ell)\mathbb{C}\text{ov}[f(X_{\ell+1}), h(Y_1)]\\ &= \sum_{\ell=0}^{n-1}(n-\ell)\mathbb{C}\text{ov}\left[\mathbb{E}[f(X_{\ell+1})|Y_1], h(Y_1)\right]\\ &= \sum_{\ell=0}^{n-1}(n-\ell)\text{Cov}_Q\left[TK_0^\ell f(X), h(X)\right]\\ &= \text{Cov }_Q\left[\sum_{\ell=0}^{n-1}(n-\ell)TK_0^\ell f(X), h(X)\right]\\ &= n\beta_0 - \text{Cov}_Q[T(I-K_0)^{-2}(I-K_0^n)K_0 f(X), h(X)].\end{aligned}$$

(ii) The Harris recurrence of (X_n) implies the Harris recurrence of (X_n, Y_n) which implies that the asymptotic distribution of (X_n, Y_n) doesn't depend on the initial distribution of this chain. So, without any loss of generality, we can assume that $X_0 \sim \Pi$. Then from (i) it is sufficient to show that $f_{0,n} = (I-K_0)^{-2}(I-K_0^n)K_0 f$ converges in $L^2(\Pi)$. But this obviously follows from the fact that the series $\sum K_0^n f$ converges in $L^2(\Pi)$.

Proof of Corollary 2.1.

(i) We only have to show that in the case of the IMH algorithm, $\beta_0 = \text{Cov}_\Pi$ $(f(X), h(X))$. From Theorem 2.1, we have $\beta_0 = \text{Cov}_Q(T(I-K_0)^{-1}f(X),$ $h(X))$. Therefore, since $\Pi\left((I-K_0)^{-1}f\right) = 0$, it is sufficient to show that $Tf(x) = \omega(x)(I-K_0)f(x)$ for any $f \in L^2(\Pi)$ with $\Pi(f) = 0$.

For $x \in \mathcal{X}$ and A measurable,

$$\begin{aligned} R(x,A) &= \int \Pi(dz) \int_0^1 \mathbf{1}_A(F(z,u,x)du \\ &= \int \Pi(dz) \left\{\alpha(z,x)\mathbf{1}_A(x) + (1-\alpha(z,x))\mathbf{1}_A(z)\right\} \\ &= \mathbf{1}_A(x) \int \alpha(z,x)\Pi(dz) + \int_A (1-\alpha(z,x))\Pi(dz). \end{aligned}$$

Now, noting that $\omega(z)\alpha(z,x) = \omega(x)\alpha(x,z)$ and with some straightforward computations we get for $f \in L^2(\Pi)$ such that $\Pi(f) = 0$, $Tf(x) = \omega(x)(I - K_0)f(x)$.

(ii) With the choice $h(x) = (f(x) - \mathrm{E}_Q(f(X)))/(\mathrm{Var}_Q(f(X)))$, $\beta_0 = \mathrm{Var}_\Pi(f(X)) / \mathrm{Var}_Q(f(X)) > 0$ and $\lim_{n\to\infty} n\left(\mathbb{V}\mathrm{ar}\left(\hat{\mu}_\beta\right) - \mathbb{V}\mathrm{ar}\left(\hat{\mu}_0\right)\right) = (\beta-\beta_0)^2 - \beta_0^2 < 0$ for $\beta \in (0, 2\beta_0)$.

Proof of Corollary 2.2. From Corollary 2.1 the asymptotic variance will be minimized if we can maximize β_0^2. We have

$$\begin{aligned} \beta_0^2 &= \mathrm{Cov}_\Pi^2\left(f(X), h(X)\right) \\ &= \mathrm{Cov}_\Pi^2\left((f(X) - \Pi(f)), h(X)\right) \\ &= \mathrm{Cov}_Q^2\left(\omega(X)(f(X) - \Pi(f)), h(X)\right)/c_0^2 \\ &\leq \mathrm{Var}_Q\left(\omega(X)(f(X) - \Pi(f))\right)/c_0^2 \end{aligned}$$

with equality if $h(y) = (\omega(y)(f(y) - \Pi(f)))/(c_0\sqrt{\mathrm{Var}_Q(\omega(X)(f(X) - \Pi(f)))})$.

Proof of Lemma 3.1.

$$\begin{aligned} \mathbb{C}\mathrm{ov}\Big(\sum_{i=1}^n f(X_i), \sum_{j=1}^n h(Y_j)\Big) = \mathbb{C}\mathrm{ov}\Big[&\mathbb{E}\Big(\sum_{i=1}^n f(X_i)|Z\Big), \sum_{j=1}^n h(Y_j)\Big] \\ &+\mathbb{E}\Big[\mathbb{C}\mathrm{ov}\Big(\Big\{\sum_{i=1}^n f(X_i), \sum_{j=1}^n h(Y_j)\Big\}|Z\Big)\Big], \end{aligned}$$

but $\sum_{j=1}^n h(Y_j)$ is fixed when Z is fixed.

Proof of Theorem 3.1. Assume that $g(y) = w(y)(f(y) - \mu)$, with $\mu = \Pi(f)$ and let $\bar{g} = \sum_{j=1}^n g(Y_j)/n$. We obtain

$$\begin{aligned} n\mathbb{V}\mathrm{ar}[\mathbb{E}[\hat{\mu}_0|Z]] &= n\mathbb{V}\mathrm{ar}[\mathbb{E}[\hat{\mu}_0|Z] - \frac{\mathbb{C}\mathrm{ov}[\mathbb{E}[\hat{\mu}_0|Z], \bar{g}]}{\mathbb{V}\mathrm{ar}[\bar{g}]}\bar{g}] + n\frac{(\mathbb{C}\mathrm{ov}[\mathbb{E}[\hat{\mu}_0|Z], \bar{g}])^2}{\mathbb{V}\mathrm{ar}[\bar{g}]} \\ &\geq n\frac{(\mathbb{C}\mathrm{ov}[\mathbb{E}[\hat{\mu}_0|Z], \bar{g}])^2}{\mathbb{V}\mathrm{ar}[\bar{g}]} \\ &= n\frac{(\mathbb{C}\mathrm{ov}[\hat{\mu}_0, \bar{g}])^2}{\mathbb{V}\mathrm{ar}[\bar{g}]} \\ &\to \mathrm{Var}_Q[\omega(Y)(f(Y) - \mu)] \text{ as } n \to \infty, \end{aligned}$$

where the last equality comes from Lemma 3.1 and the limiting result is explained in the proof of Theorem 2.1.

Acknowledgement

We are grateful to the referee and an associate editor for their very helpful comments.

References

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications.* Wiley, New York.

Casella, G. and Robert, C. P. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika* **83**, 81-94.

Chan, K. and Geyer, G. (1994). Discussion paper. *Ann. Statist.* **22**, 1747-1758.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000). *Monote Carlo Methods in Bayesian Computation.* Springer-Verlag, New York.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Gilks, W. R. and Roberts, G. O. and Sahu, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration/ *J. Amer. Statist. Assoc.* **93**, 1045-1054.

Liu, J. S. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27-40.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer Verlag, New-York.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101-121.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with Discussion). *J. Roy. Statist. Soc. Ser. B* **56**, 3-48.

Perron, F. (1999). Beyond accept-reject sampling. *Biometrika* **86**, 803-813.

Robert, C. P. and Casella, G. (1999). *Monte Carol Statistical Methods.* Spring Verlag, New York.

Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab.* **2**, 13-25 (electronic).

Department of Statistics, Harvard University, Science Center, One Oxford Street, Cambridge, MA 02138-2901, U.S.A.

E-mail: atchade@stat.harvard.edu

Department of Mathematics and Statistics, University of Montreal, P.O. Box 6128, Station Centre-Ville, Montreal QC, Canada, H3C 3J7.

E-mail: perronf@dms.umontreal.ca

(Received June 2003; accepted May 2004)