Variance reduction of estimators arising from Metropolis–Hastings algorithms

George Iliopoulos* Sonia Malefaki†

Abstract

The Metropolis–Hastings algorithm is one of the most basic and well-studied Markov chain Monte Carlo methods. It generates a Markov chain which has as limit distribution the target distribution by simulating observations from a different proposal distribution. A proposed value is accepted with some particular probability otherwise the previous value is repeated. As a consequence, the accepted values are repeated a positive number of times and thus any resulting ergodic mean is, in fact, a weighted average. It turns out that this weighted average is an importance sampling-type estimator with random weights. By the standard theory of importance sampling, replacement of these random weights by their (conditional) expectations leads to more efficient estimators. In this paper we study the estimator arising by replacing the random weights with certain estimators of their conditional expectations. We illustrate by simulations that it is often more efficient than the original estimator while in the case of the independence Metropolis–Hastings and for distributions with finite support we formally prove that it is even better than the "optimal" importance sampling estimator.

Key words and phrases: Metropolis–Hastings algorithm, importance sampling, weighted estimators, variance reduction.

1 Introduction

The last decades Markov chain Monte Carlo (MCMC) algorithms have become very popular and widely used tools in Computational Statistics as a way of sampling from complex multidimensional probability distributions. The most basic MCMC method is the Metropolis–Hastings (MH) algorithm which generates a Markov chain with limit distribution the target distribution by drawing observations from a proposal distribution (cf. Metropolis et al., 1953, Hastings, 1970). A proposed value is accepted with a certain probability otherwise the previous accepted value is repeated. As a consequence, the accepted values are repeated a positive number of times. Thus, a positive weight corresponds to any accepted value.

To be formal, let π be the density of the target distribution with respect to some underlying measure μ and q be another distribution with at least the same support as π. The MH algorithm

*Corresponding author; Department of Statistics and Insurance Science, University of Piraeus, 80 Karaoli & Dimitriou str., 18534 Piraeus, Greece, e–mail: geh@unipi.gr

†Department of Engineering Sciences, University of Patras, 26500 Rio, Patras, Greece, smalefaki@upatras.gr

with proposal distribution q generates a Markov chain $\boldsymbol{Y} = (Y_t)_{t\in\mathbb{Z}_+}$ ($\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$) through the following transition. It starts from some (either deterministic or randomly chosen) state y_0 in the support of π and at time $t+1$, given $Y_t = y$, proposes $z \sim q(z|y)$ and sets

$$Y_{t+1} = \begin{cases} z, & \text{with probability } a(y,z) \\ y, & \text{with probability } 1 - a(y,z), \end{cases}$$

where

$$a(y,z) = \min\left\{1, \frac{\pi(z)q(y|z)}{\pi(y)q(z|y)}\right\}.$$

It is well–known that the generated Markov chain $\boldsymbol{Y}$ is reversible with limit distribution π. Malefaki and Iliopoulos (2008, Section 3) studied $\boldsymbol{Y}$ from a different perspective. More specifically, they considered $\boldsymbol{Y}$ as a discrete time Markov jump process with embedded Markov chain the sequence of accepted states and sojourn times the corresponding number of repetitions.

Let $\boldsymbol{X} = (X_n)_{n\in\mathbb{Z}_+}$ be the sequence of accepted states and $\boldsymbol{\xi} = (\xi_n)_{n\in\mathbb{Z}_+}$ be their corresponding numbers of appearances in $\boldsymbol{Y}$. This means that X_i is repeated ξ_i times in $\boldsymbol{Y}$ until the acceptance of the next stated X_{i+1}. The results of Malefaki and Iliopoulos (2008) can be collected in the following proposition (see also Douc and Robert, 2011).

Proposition 1. *(a) The conditional distribution of ξ_n given $X_n = x_n$ is geometric with probability of success $\int \alpha(x_n, z)q(z|x_n)\mu(\mathrm{d}z)$, i.e.,*

$$p(\xi|x_n) = \left\{\int a(x_n,z)q(z|x_n)\mu(\mathrm{d}z)\right\}\left\{1 - \int a(x_n,z)q(z|x_n)\mu(\mathrm{d}z)\right\}^{\xi-1},\ \xi = 1, 2, \ldots.$$

(b) The sequence $\boldsymbol{X} = (X_n)_{n\in\mathbb{Z}_+}$ is a Markov chain with transition density

$$g(x_n|x_{n-1}) = \frac{a(x_{n-1},x_n)q(x_n|x_{n-1})}{\int a(x_{n-1},z)q(z|x_{n-1})\mu(\mathrm{d}z)} = \frac{\min\{\pi(x_{n-1})q(x_n|x_{n-1}), \pi(x_n)q(x_{n-1}|x_n)\}}{\int \min\{\pi(x_{n-1})q(z|x_{n-1}), \pi(z)q(x_{n-1}|z)\}\mu(\mathrm{d}z)}.$$

(c) The Markov chain $\boldsymbol{X} = (X_n)_{n\in\mathbb{Z}_+}$ is reversible with stationary distribution

$$g(x) \propto \int \min\{\pi(x)q(z|x), \pi(z)q(x|z)\}\mu(\mathrm{d}z). \tag{1}$$

(d) The sequence $(X_n, \xi_n)_{n\in\mathbb{Z}_+}$ is properly weighted with respect to π, i.e., $E(\xi_i|X_i = x) = \kappa w(x)$, where $w(x) = \pi(x)/g(x)$ and $\kappa = \left(\iint \min\{\pi(x)q(z|x), \pi(z)q(x|z)\}\mu(\mathrm{d}z)\mu(\mathrm{d}x)\right)^{-1}$ is the normalizing constant of g.

Let $Y_0, Y_1, \ldots, Y_T$ be the sequence generated from the MH algorithm. Then, the standard estimator of $E_\pi(h) = \int h(x)\pi(x)\mu(\mathrm{d}x)$ is the ergodic mean

$$\hat{h}^*_{MH} = \frac{1}{T}\sum_{t=1}^{T} h(Y_t).$$

In practice, the estimate is evaluated after discarding a certain part of the first Y-values that corresponds to the burn-in period. If this is the case, denote by Y_1 the first value after that period. Note that an alternative expression for the above estimator is

$$\hat{h}^*_{MH} = \frac{1}{T}\Big\{ \sum_{i=1}^{n} \xi_i h(X_i) + \Big(T - \sum_{i=1}^{n} \xi_i\Big) h(X_{n+1}) \Big\}.$$

Here $n = n(T)$ is the largest integer for which $T \geqslant \sum_{i=1}^{n} \xi_i$. Note that an alternative estimator of $E_\pi(h)$ that takes into account only the first n states is

$$\hat{h}_{MH} = \frac{\sum_{i=1}^{n} \xi_i h(X_i)}{\sum_{i=1}^{n} \xi_i}. \tag{2}$$

Part (d) of Proposition 1 suggests that if the evaluation of $w(x)$ is possible, another estimator of $E_\pi(h)$ can be

$$\hat{h}_{IS} = \frac{\sum_{i=1}^{n} w(X_i) h(X_i)}{\sum_{i=1}^{n} w(X_i)}.$$

In fact, $\hat{h}_{IS}$ is more efficient than $\hat{h}_{MH}$. This has been realized by Malefaki and Iliopoulos (2008) who have shown that if

$$n^{1/2}\{\hat{h}_{IS} - E_\pi(h)\} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2_{IS}(h)) \quad \text{with} \quad \sigma^2_{IS} < \infty,$$

then

$$n^{1/2}\{\hat{h}_{MH} - E_\pi(h)\} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2_{MH}(h)),$$

where

$$\sigma^2_{MH}(h) = \sigma^2_{IS}(h) + \frac{1}{\kappa^2} E_g\big[Var(\xi|X)\{h(X) - E_\pi(h)\}^2\big]. \tag{3}$$

In the proof of Theorem 2 of Douc and Robert (2011) it is shown that $n/T \xrightarrow{\mathrm{p}} E_\pi(w) \in (0, \infty)$ and $T^{-1/2}(T - \sum_{i=1}^{n} \xi_i) \xrightarrow{\mathrm{p}} 0$ as $T \to \infty$. This implies that the asymptotic distributions of $n^{1/2}\{\hat{h}_{MH} - E_\pi(h)\}$ and $n^{1/2}\{\hat{h}^*_{MH} - E_\pi(h)\}$ coincide. Due to this asymptotic equivalence, in the sequel we will consider $\hat{h}_{MH}$ as the MH estimator instead of $\hat{h}^*_{MH}$. Moreover, since overall it does not matter whether n is random or fixed, whenever it is convenient (as in the presentation of our result) we will assume it to be fixed so that $T = \sum_{i=1}^{n} \xi_i$.

Note that $\hat{h}_{IS}$ is actually a standard importance sampling (IS) estimator with the only difference that it is produced by a Markov chain with limit distribution g rather than direct iid sampling from g. Hence, it is justified to refer to $w(x)$'s as "importance weights". Unfortunately, $w(x)$ cannot be evaluated except for some toy examples as those presented in Section 3. Therefore, Malefaki and Iliopoulos (2008) suggested to estimate the importance weights and plug their estimates into $\hat{h}_{IS}$. They further illustrated by simulations that the resulting estimator, $\hat{h}_{\hat{w}}$, say, behaves well and in some cases almost the same as $\hat{h}_{IS}$. In this paper we move

one step further and show that $\hat{h}_{\hat{w}}$ is often more efficient not only than $\hat{h}_{MH}$ but than $\hat{h}_{IS}$ as well. This is formally proven in the case of independence MH (IMH), i.e., when $q(z|y) = q(z)$, and finite state space. However, simulations indicate that for the IMH this holds for continuous state space as well.

In the literature there are several approaches to improve on $\hat{h}_{MH}$. By considering the above point of view, Douc and Robert (2011) replaced the geometric distributed weights $\xi_1, \ldots, \xi_n$ by some other (also random) weights $\hat{\xi}_1, \ldots, \hat{\xi}_n$ in such a way that $E(\hat{\xi}_i|X_i) = E(\xi_i|X_i)$ and $Var(\hat{\xi}_i|X_i) < Var(\xi_i|X_i)$. The resulting estimator is more efficient than $\hat{h}_{MH}$ due to (3). Alternative methods for improving MH estimators based on Rao–Blackwellization have been presented by Casella and Roberts (1996) and Atchadé and Perron (2005). However, their evaluation requires high computational cost and therefore makes them practically useless. As Atchadé and Perron (2005) say,

> "... 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 Rao–Blackwellization when we want to reduce the variance in case n is large ..."

Recently, a different approach is used by Jacob et al. (2011) in the special case of IMH algorithm. These authors run blocks of several IMH algorithms which share permutations of the same proposed values. The resulting estimator is a weighted average of all of them with corresponding weights their total number of appearances and is more efficient than $\hat{h}_{MH}$.

The rest of the paper is organized as follows. In Section 2, we review the approach of Malefaki and Iliopoulos (2008) and state our theoretical result. We also describe an algorithm for the efficient calculation of our importance weights' estimates in the special case of IMH. In Section 3 we give some illustrative toy examples as well as a real data example. Moreover, we compare by simulation the performance of our estimator with those of Douc and Robert (2011) and Jacob et al. (2011). Section 4 contains a brief discussion. The paper concludes with an appendix containing the proof of our result.

2 Estimation of the importance weights and main result

Using the definition of the distribution g in (1) and its normalizing constant κ we get that

$$\begin{aligned}\{\kappa w(x)\}^{-1} &= \int a(x,z) q(z|x) \mu(\mathrm{d}z) \\ &= \int \min\left\{\frac{q(z|x)}{\pi(z)}, \frac{q(x|z)}{\pi(x)}\right\} \pi(z)\mu(\mathrm{d}z) = E_\pi\left[\min\left\{\frac{q(Z|x)}{\pi(Z)}, \frac{q(x|Z)}{\pi(x)}\right\}\right].\end{aligned}$$

Thus, the inverse of importance weights is, in fact, an expectation with respect to the target distribution π. Hence, they can be estimated using the original MH sequence $\boldsymbol{Y}$. Since $1/w(x) \neq 0$ for all x, for each fixed x the estimator

$$\kappa\hat{w}(x) = \frac{\sum_{j=1}^{n} \xi_j}{\sum_{j=1}^{n} \xi_j \min\{q(X_j|x)/\pi(X_j), q(x|X_j)/\pi(x)\}} \tag{4}$$

converges almost surely to $\kappa w(x)$. So, the idea of Malefaki and Iliopoulos (2008) is as follows: Use $\hat{w}(X_i)$ in the place of the importance weights $w(X_i)$ and estimate $E_\pi(h)$ with

$$\hat{h}_{\hat{w}} = \frac{\sum_{i=1}^{n} \hat{w}(X_i)h(X_i)}{\sum_{i=1}^{n} \hat{w}(X_i)}.$$

Note here that the weight $\hat{w}(X_i)$ is a function of all $(X_i, \xi_i)_{1 \leqslant i \leqslant n}$ and not only of X_i. This causes a problem, since the asymptotic properties of $\hat{h}_{\hat{w}}$ are not straightforward. In particular, it is not even clear whether $\hat{h}_{\hat{w}}$ converges, in general, to $E_\pi(h)$. Moreover, if this is the case and if further $n^{1/2}\{\hat{h}_{\hat{w}} - E_\pi(h)\} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2_{\hat{w}}(h))$ holds, then the variance $\sigma^2_{\hat{w}}(h)$ does not have an expression similar to (3).

We have run extensive simulations and all of them suggest that $\hat{h}_{\hat{w}}$ converges indeed to $E_\pi(h)$. Moreover, there are cases where $\hat{h}_{\hat{w}}$ is more efficient not only than $\hat{h}_{MH}$ but also than the "optimal estimator" $\hat{h}_{IS}$. These facts are stated in the following theorem for the special case of finite state space.

Theorem 1. *If the state space is finite:*
(a) The estimator $\hat{h}_{\hat{w}}$ converges to $E_\pi(h)$ almost surely.
(b) $n^{1/2}\{\hat{h}_{\hat{w}} - E_\pi(h)\} \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\hat{w}}(h))$, where

$$\sigma^2_{\hat{w}}(h) = Var_g\{w(X_0)h(X_0)\} - E_g\left\{ \frac{E\big[w(X_1)\{h(X_1) - E_\pi(h)\}|X_0\big]^2}{\kappa w(X_0)} \right\}.$$

(c) In the case of IMH it holds $\sigma^2_{\hat{w}}(h) < \sigma^2_{IS}(h)$ and thus, $\sigma^2_{\hat{w}} < \sigma^2_{MH}(h)$.

The (rather lengthy) proof of Theorem 1 can be found in the Appendix.

One may argue against the practical usefulness of the estimator $\hat{h}_{\hat{w}}$ due to the extra computational cost needed for the estimation of the weights. There are some tricks one could use in order to reduce the computation time. Malefaki and Iliopoulos (2008) defined the sets $A_x := \{z : \pi(x)q(z|x) \leqslant \pi(z)q(x|z)\}$, and expressed the estimate of $\kappa w(x_i)$ as

$$\kappa\hat{w}(x_i) = \sum_{j=1}^{n} \xi_j \Big/ \left\{ \sum_{j=1}^{n} \xi_j \frac{q(x_j|x_i)}{\pi(x_j)} I(x_j \in A_{x_i}) + \frac{1}{\pi(x_i)} \sum_{j=1}^{n} \xi_j q(x_i|x_j) I(x_j \in A^c_{x_i}) \right\}.$$

In several examples, the sets A_x have a simple form, so obtaining the estimates is not too time consuming. In particular, the calculation of the weights' estimates can be considerebly accelerated in the special case of IMH by using the following procedure.

Algorithm 1. (*Calculation of $\hat{w}$ for IMH*)

Given the IMH output $(x_i, \xi_i)_{1 \leqslant i \leqslant n}$:

Step 0: Sort $r_i = q(x_i)/\pi(x_i)$, $i = 1, \ldots, n$, in ascending order. Let $r_{(1)} \leqslant \cdots \leqslant r_{(n)}$ be the corresponding ordered values and denote also $x_{(i)}$, $\xi_{(i)}$ the values of x and ξ that correspond to $r_{(i)}$.

Step 1: Set $C_1 = 0$, $C_2 = r_{(1)} \sum_{j=1}^{n} \xi_j$ and $\kappa \hat{w}(x_{(1)}) = \sum_{j=1}^{n} \xi_j / (C_1 + C_2)$.

Step i, for $i = 2, \ldots, n$: Update $C_1 = C_1 + \xi_{(i-1)} r_{(i-1)}$, $C_2 = (C_2 - \xi_{(i-1)} r_{(i-1)}) r_{(i)} / r_{(i-1)}$, and set $\kappa \hat{w}(x_{(i)}) = \sum_{j=1}^{n} \xi_j / (C_1 + C_2)$.

Algorithm 1 delivers the estimated weights very quickly. Given that the ratios r_i have been already calculated in the original MH run since they are needed for the acceptance probabilities, the most heavy part of the algorithm becomes the sorting procedure. Nervertheless, in all of the examples we have run, the additional time for the weights' estimation was very short compared to the time needed for the IMH run. Unfortunately, for other MH algorithms this is not the case because most of the ratios $q(x_j|x_i)/\pi(x_j)$ are not available from the original run. This makes the weights' estimation procedure slow and so the possible gain in variance can be counterbalanced by increasing the length of the MH chain.

3 Examples

The examples in this section illustrate the performance of our estimator compared with $\hat{h}_{MH}$ and $\hat{h}_{IS}$ in terms of asymptotic variance. In our simulations we have included the estimator $\hat{h}_{DR}$ proposed by Douc and Robert (2011), the asymptotic variance of which is known to lie between those of the above estimators. In all of the examples we ran independently $m = 200$ chains of length $T = 10000$ and we recorded the corresponding estimators of $E_\pi(X)$ and $E_\pi(X^2)$. (For the real data example in Subsection 3.4 we used $m = 500$ chains.) All simulations have been programmed in Fortran 95 and run on a 2.4 GHz Intel Core 2 Processor with 2 GB RAM. The variances comparison of any two estimators is made as follows. Denote by $(\hat{h}_1^{(1)}, \hat{h}_2^{(1)}), \ldots, (\hat{h}_1^{(m)}, \hat{h}_2^{(m)})$ the m realizations of the estimators $\hat{h}_1$, $\hat{h}_2$, where each pair corresponds to the same chain. It is easy to see that $Var(\hat{h}_1) > Var(\hat{h}_2)$ if and only if $\hat{h}_1 + \hat{h}_2$ and $\hat{h}_1 - \hat{h}_2$ are positively correlated. So, in order to compare the variances we consider the transformed pairs $(\hat{h}_1^{(j)} + \hat{h}_2^{(j)}, \hat{h}_1^{(j)} - \hat{h}_2^{(j)})$, $j = 1, \ldots, m$, and test the one-sided significance of their sample correlation coefficient.

θ	$h(x)$	$\hat{\sigma}_{MH}$	$\hat{\sigma}_{DR}$	$\hat{\sigma}_{IS}$	$\hat{\sigma}_{\hat{w}}$	$\bar{n}$	$r_{IS,\hat{w}}$	$z(r)$
0.1	x	.0349	.0325	.0304	.0218	1813.8	.7777	14.6
	x^2	.1242	.1147	.1096	.0728		.8119	15.9
0.5	x	.0149	.0144	.0141	.0119	6670.3	.9014	20.8
	x^2	.0569	.0561	.0557	.0478		.8747	19.0
0.9	x	.0108	.0106	.0106	.0103	9471.8	.9614	27.6
	x^2	.0455	.0450	.0450	.0441		.8121	15.9

Table 1: (Example 3.1) Estimated standard errors of the four estimators of the first and second moment of the target distribution $\pi(x) = e^{-x}$, $x > 0$, with proposal distribution $q(x) = \theta e^{-\theta x}$, $x > 0$, based on $m = 200$ independent runs. The columns labelled $r_{IS,\hat{w}}$ and $z(r)$ show the sample correlation coefficient for testing $\sigma^2_{IS} > \sigma^2_{\hat{w}}$ and the corresponding Fisher's z-value, respectively.

3.1 Independence Metropolis–Hastings: Exponential target distribution

Consider the IMH algorithm with target distribution $\pi(x) = e^{-x}$, $x > 0$, and proposal $q(z|x) \equiv q(z) = \theta e^{-\theta z}$, $z > 0$, with $\theta < 1$. In this case it can be shown that $g(x) \propto e^{-x(\theta+1)}(\theta - 1 + e^{\theta x})$, $x > 0$, and thus, $w(x) = \pi(x)/g(x) = \{e^{-\theta x}(\theta - 1 + e^{\theta x})\}^{-1}$, $x > 0$. It can be easily seen that in this example it holds $A_x = (0, x]$ and that the ordering of r_i's is the same as that of x_i's. So, the estimates of the weights become

$$\kappa\hat{w}(x_{(i)}) = \sum_{j=1}^{n} \xi_j \Big/ \Big\{ \sum_{j=1}^{i-1} \xi_{(j)} e^{(1-\theta)x_{(j)}} + e^{(1-\theta)x_{(i)}} \sum_{j=i}^{n} \xi_{(j)} \Big\}, \quad i = 1, \ldots, n,$$

where $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$ are the ordered x values and $\xi_{(1)}, \ldots, \xi_{(n)}$ their corresponding weights.

We considered three values of θ, namely, 0.1, 0.5 and 0.9. Note that as θ increases, the proposal distribution gets closer to the target distribution. The estimated standard errors of the estimators are presented in Table 1 together with the average length, $\bar{n}$, of the accepted states sequence. As expected, the "estimator" $\hat{h}_{IS}$ is more efficient than $\hat{h}_{MH}$ while $\hat{h}_{DR}$ lies in between. On the other hand, the standard error of our estimator, $\hat{h}_{\tilde{w}}$, is smaller not only than those of $\hat{h}_{MH}$ and $\hat{h}_{DR}$ but of $\hat{h}_{IS}$ as well. The high significance of the sample correlation coefficients described in the beginning of the section confirms that in all cases of Table 1 it holds $\sigma^2_{IS} > \sigma^2_{\hat{w}}$. Indeed, the corresponding Fisher's z transforms (which are supposed to come from a standard normal distribution when the population correlation coefficient equals zero) imply that all p-values are practically zero. Note that this agrees with Theorem 1, although here the state space is continuous.

As mentioned in the Introduction, Jacob et al. (2011) proposed a new method in order to improve estimation in the case of independence MH. It is based on a different idea in that many

		$\hat{h}_{JRS}(20,500)$		$\hat{h}_{JRS}(100,100)$		$\hat{h}_{JRS}(1000,10)$		$\hat{h}_{\hat{w}}$	
θ	$h(x)$	$\hat{\sigma}_{JRS}$	CPU	$\hat{\sigma}_{JRS}$	CPU	$\hat{\sigma}_{JRS}$	CPU	$\hat{\sigma}_{\hat{w}}$	CPU
0.1	x	.0156	1.144	.0181	.2186	.0232	.0225	.0218	.0031
	x^2	.0414	(.0469)	.0487	(.0405)	.0775	(.0078)	.0728	(.0062)
0.5	x	.0093	1.075	.0087	.2116	.0098	.0208	.0119	.0045
	x^2	.0290	(.0102)	.0268	(.0088)	.0310	(.0073)	.0478	(.0070)
0.9	x	.0098	.9474	.0094	.1769	.0085	.0184	.0103	.0051
	x^2	.0375	(.0262)	.0378	(.0078)	.0346	(.0064)	.0441	(.0073)

Table 2: (Example 3.1) Estimated standard errors of $\hat{h}_{JRS}$ for selected (b,p)'s and mean CPU times along with their standard deviations based on $m = 200$ independent runs. The last two columns contain the corresponding results for $\hat{h}_{\hat{w}}$.

parallel sequences sharing the same proposals are simulated rather than a single Markov chain. More specifically, in the beginning the T independent proposals are split into b blocks of length p (so that $T = bp$). For each block p random permutations of the particular proposals are considered and the corresponding MH sequences of length p are simulated. As soon as the block is finished, one among the last states of the p sequences is randomly selected and it is used as starting point for the p sequences of the next block. Finally, the estimator of $E_\pi(h)$, $\hat{h}_{JRS}(b,p)$ say, averages over all bp^2 simulated values.

In order to compare their approach with ours we consider both the standard errors of estimators and the CPU time needed to obtain them. As we can see in Table 2, $\hat{\sigma}_{JRS}$ is in almost all cases less than $\hat{\sigma}_{\hat{w}}$ with the largest improvement being about 44% (for $\theta = 0.1$ and $(b,p) = (20,500)$). However, when p is small and the proposal is far from the target distribution we may get $\hat{\sigma}_{JRS} > \hat{\sigma}_{\hat{w}}$. In general, the larger the value of p the higher the improvement. Note though that in all of the examples we ran, the improvement in standard error was never more than 50%. On the other hand, the CPU times needed for the method of Jacob et al. (2011) are considerably greater than ours even for small p. For instance, in the case of the above mentioned largest standard error improvement, the CPU time was over 350 times higher. In fact, the simulations showed that the CPU time is roughly linear in p. This is not surprising since the method repeats the MH procedure p times. Given the small standard error improvement of $\hat{h}_{JRS}$ and the huge difference in CPU times, it is clear that one can simulate a single MH sequence with larger T in order to make $\hat{h}_{\hat{w}}$ more efficient and still keep the required time lower.

θ	$h(x)$	$\hat{\sigma}_{MH}$	$\hat{\sigma}_{DR}$	$\hat{\sigma}_{IS}$	$\hat{\sigma}_{\hat{w}}$	$\bar{n}$	$r_{IS,\hat{w}}$	$z(r)$
1.5	x	.01231	.01136	.01123	.01123	7488.2	.0182	0.26
	x^2	.01831	.01810	.01730	.01520		.9306	23.3
5.0	x	.02375	.02161	.01827	.01826	2508.8	.0294	0.41
	x^2	.03690	.03374	.03012	.02283		.7418	13.4
10.0	x	.03537	.03394	.02481	.02469	1267.8	.1679	2.38
	x^2	.05604	.05352	.04516	.03218		.7865	14.9

Table 3: (Example 3.2) Estimated standard errors of the four estimators of the first and second moment of the target distribution $\pi(x) \propto e^{-x^2/2}$, $x \in \mathbb{R}$,, with proposal distribution $q(z) \propto e^{-z^2/2\theta^2}$, $z \in \mathbb{R}$, based on $m = 200$ independent runs. The columns labelled $r_{IS,\hat{w}}$ and $z(r)$ show the sample correlation coefficient for testing $\sigma^2_{IS} > \sigma^2_{\hat{w}}$ and the corresponding Fisher's z-value, respectively.

3.2 Independence Metropolis–Hastings: Normal target distribution

Consider the IMH algorithm with target distribution $\pi(x) \propto e^{-x^2/2}$, $x \in \mathbb{R}$, and proposal distribution $q(z) \propto e^{-z^2/2\theta^2}$, $z \in \mathbb{R}$, with $\theta > 1$. It can be verified that in this example the limit distribution of the accepted states has density

$$\begin{aligned} g(x) &\propto \int \min\{\pi(x)q(z), \pi(z)q(x)\}\,\mu(\mathrm{d}z) \\ &\propto \int \min\{e^{-x^2/2-z^2/2\theta^2}, e^{-z^2/2-x^2/2\theta^2}\}\mathrm{d}z \\ &\propto \theta[2\Phi(|x|/\theta) - 1]e^{-x^2/2} + 2\Phi(-|x|)e^{-x^2/2\theta^2}, \end{aligned}$$

where Φ denotes the cdf of the standard normal distribution, and that $A_x = [-|x|, |x|]$.

We considered several values of θ and repeated the procedure of Example 3.1. The estimated standard errors of the four estimators are presented in Table 3. We can see that their ordering is the same as before. Note that the insignificance of the sample correlation coefficients used to test for $\sigma^2_{IS} > \sigma^2_{\hat{w}}$ in the case of estimating $E_\pi(X)$ is due to the small number m of independent chains. We increased m and concluded that it must be at least 10000 so that $r_{IS,\hat{w}}$ to become significant. This means that the standard error of $\hat{h}_{\hat{w}}$ is indeed smaller but the difference is marginal. We also mention that the comparison of our method with the one of Jacob et al. (2011) gave analogous results with Example 3.1 which are not presented here for brevity.

θ	$h(x)$	$\hat{\sigma}_{MH}$	$\hat{\sigma}_{DR}$	$\hat{\sigma}_{IS}$	$\hat{\sigma}_{\hat{w}}$	$\bar{n}$	$r_{IS,\hat{w}}$	$z(r)$
1.5	x	.02284	.02206	.02119	.01150	5900.9	.8631	18.3
	x^2	.03137	.02998	.02898	.01782		.8364	17.0
5.0	x	.02397	.02330	.01913	.01802	2426.9	.6292	10.4
	x^2	.04022	.03893	.03506	.02614		.8057	15.6
10.0	x	.03570	.03506	.02745	.02702	1257.6	.4163	6.2
	x^2	.05869	.05210	.04652	.03391		.7777	14.6

Table 4: (Example 3.3) Estimated standard errors of the four estimators of the first and second moment of the target distribution $\pi(x) \propto e^{-x^2/2}$, $x \in \mathbb{R}$, with proposal distribution $q(z|x) \propto e^{-(z-x)^2/2\theta^2}$, $z \in \mathbb{R}$, based on $m = 200$ independent runs. The columns labelled $r_{IS,\hat{w}}$ and $z(r)$ show the sample correlation coefficient for testing $\sigma^2_{IS} > \sigma^2_{\hat{w}}$ and the corresponding Fisher's z-value, respectively.

3.3 Random walk Metropolis–Hastings: Normal target distribution

Let $\pi(x) \propto e^{-x^2/2}$, $x \in \mathbb{R}$, and $q(z|x) \propto e^{-(z-x)^2/2\theta^2}$, $z \in \mathbb{R}$. Then, the sequence of the accepted states of the corresponding MH algorithm has limit distribution

$$\begin{aligned} g(x) \propto & \int \min\{\pi(x)q(z|x), \pi(z)q(x|z)\}\, \mu(\mathrm{d}z) \\ \propto & \int \min\{e^{-x^2/2-(z-x)^2/2\theta^2}, e^{-z^2/2-(x-z)^2/2\theta^2}\}\mathrm{d}z \\ \propto & \{2\Phi(2|x|/\theta) - 1/2\}\theta e^{-x^2/2} + \left\{\Phi\left(-\frac{(\theta^2+2)|x|}{\theta\sqrt{\theta^2+1}}\right) + \Phi\left(-\frac{\theta|x|}{\sqrt{\theta^2+1}}\right)\right\}\frac{\theta e^{-x^2/2(\theta^2+1)}}{\sqrt{1+\theta^2}}. \end{aligned}$$

In this case, it holds $A_x = [-|x|, |x|]$ as well. The standard errors of all estimators of the first and second moments of the target distribution for several values of θ appear in Table 4. The conclusions are similar to those in the previous examples.

3.4 A real data example

We applied our approach to the dataset `Pima.te` which is available in libary `MASS` of `R` and has been used as a benchmark dataset by many authors such as Marin and Robert (2010) (for the evaluation of model choice techniques), Douc and Robert (2011) and Jacob et al. (2011) (for the comparison of several proposed estimators based on MH algorithms). The dataset consists of $n = 332$ observations on a population of females at least 21 years old of Pima Indian heritage living near Phoenix, AZ, which have been tested for diabetes. Let $s_i = 1$ or 0 be the response for the ith subject depending on whether she is diabetic or not according to the World Health Organization criteria (variable `type`). For illustration purposes we will use the following explanatory variables:

		$\hat{h}_{MH}$		$\hat{h}_{\hat{w}}$				
	MLE	$\hat{E}(\theta_i \mid \boldsymbol{s}, \boldsymbol{Z})$	Std. error	$\hat{E}(\theta_i \mid \boldsymbol{s}, \boldsymbol{Z})$	Std. error	$\bar{n}$	$r_{MH,\hat{w}}$	$z(r)$
θ_0	-5.0137	-5.0169	2.25×10^{-2}	-5.0173	1.56×10^{-2}	1685.4	.4832	11.7
θ_1	0.0218	0.0218	8.52×10^{-5}	0.0218	6.26×10^{-5}		.4078	9.7
θ_2	0.0024	0.0024	2.01×10^{-4}	0.0024	1.48×10^{-4}		.4052	9.6
θ_3	0.5878	0.5860	6.72×10^{-3}	0.5859	4.88×10^{-3}		.4352	10.4
θ_4	0.0412	0.0412	3.64×10^{-4}	0.0412	2.66×10^{-4}		.4338	10.4

Table 5: (Example 3.4) Estimates of the posterior means of the five regression parameters and their standard errors based on $m = 500$ independent runs. The columns labelled $r_{MH,\hat{w}}$ and $z(r)$ show the sample correlation coefficient for testing $\sigma^2_{MH} > \sigma^2_{\hat{w}}$ and the corresponding Fisher's z-value, respectively.

- z_1: plasma glucose concentration in an oral glucose tolerance test (variable `glu`)
- z_2: diastolic blood pressure in mmHg (variable `bp`)
- z_3: diabetes pedigree function (variable `ped`)
- z_4: body mass index (variable `bmi`)

Denote by $\boldsymbol{Z}$ the 332×5 matrix consisting of a column of ones corresponding to the "intercept" and four columns containing the explanatory variables. Let $\boldsymbol{z}_i$ be its ith row. We use a standard probit model, i.e. it is assumed that given the unknown parameter $\boldsymbol{\theta}$ the s_i's are independent Bernoulli random variables with $P(s_i = 1) = \Phi(\boldsymbol{z}_i'\boldsymbol{\theta})$, where Φ is the standard normal cdf. We use the five-dimensional normal distribution with mean vector zero and covariance matrix $n(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ as a prior distribution for $\boldsymbol{\theta}$. It follows that the posterior distribution of $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{s}, \boldsymbol{Z}) \propto \exp\{-\boldsymbol{\theta}'(\boldsymbol{Z}'\boldsymbol{Z})\boldsymbol{\theta}/2n\} \prod_{i=1}^{n} \{1 - \Phi(\boldsymbol{z}_i'\boldsymbol{\theta})\}^{1-s_i} \Phi(\boldsymbol{z}_i'\boldsymbol{\theta})^{s_i}.$$

In order to estimate the posterior mean $E(\boldsymbol{\theta} \mid \boldsymbol{s}, \boldsymbol{Z})$, we apply the IMH algorithm with proposal distribution $\mathcal{N}_5(\hat{\boldsymbol{\theta}}, c\hat{\boldsymbol{\Sigma}})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\Sigma}}$ is its asymptotic covariance matrix and c is a scale parameter. In our runs we set $c = 3$.

We compare the estimators $\hat{h}_{MH}$ and $\hat{h}_{\hat{w}}$ by running the IMH algorithm $m = 500$ times. The results are presented in Table 5. We can see that $\hat{h}_{\hat{w}}$ is more efficient with the standard error reduction estimated from 25% to 30%. The high significance of the sample correlation coefficients $r_{MH,\hat{w}}$ indicates that $\sigma^2_{MH} > \sigma^2_{\hat{w}}$ for all estimators of the posterior means.

$h(x)$	$\hat{h}_{MH}$	$\hat{h}_{DR}$	$\hat{h}_{IS}$	$\hat{h}_{\hat{w}}$	$\bar{n}$
x	.003321	.003204	.003113	.003427	8457.6
x^2	.003574	.003385	.003289	.003572	

Table 6: (Example 3.5) Estimated standard errors of the four estimators of the first and second moment of the standard uniform target distribution with proposal distribution $q(z|x) \sim \mathcal{U}(0,1)$ if $x \leqslant 1/2$ and $\mathcal{B}eta(1/2,1)$ if $x > 1/2$ based on $m = 200$ independent runs.

3.5 Nonstandard Metropolis–Hastings: Uniform target distribution

Although in all of the previous examples the estimator $\hat{h}_{\hat{w}}$ was the most efficient one, this is not always the case. In this example we will see that it can actually be the worst.

Let us consider the MH algorithm with target distribution $\pi \sim \mathcal{U}(0,1)$, i.e., the standard uniform distribution and proposal distribution $q(z|x) \sim \mathcal{U}(0,1)$ if $x \leqslant 1/2$ and $\mathcal{B}eta(1/2,1)$ if $x > 1/2$. After some calculations it can be shown that the limit distribution of the accepted states is

$$g(x) \propto \begin{cases} 1, & 0 < x \leqslant 1/4, \\ \frac{1}{4}(2 + x^{-1/2}), & 1/4 < x \leqslant 1/2, \\ \frac{1}{4}\{3 + 2\sqrt{2} - x^{-1/2}(1 + 2x)\}, & 1/2 < x < 1. \end{cases}$$

Our simulation results are presented in Table 6. We can see that the standard error of $\hat{h}_{\hat{w}}$ is larger than those of the other estimators when $h(x) = x$. In particular, $\hat{h}_{\hat{w}}$ is less efficient even than $\hat{h}_{MH}$. On the other hand, when $h(x) = x^2$, the estimator $\hat{h}_{\hat{w}}$ is better than $\hat{h}_{MH}$ but worse than the other two estimators.

4 Discussion

The reduction of variance of MCMC estimators and in particular of estimators arising from Metropolis-Hastings algorithms has attracted in the last twenty years the attention of many researchers. In this paper, we considered the modified estimator proposed by Malefaki and Iliopoulos (2008) and showed via illustrative examples that it often performs better not only than the standard MH estimator but also than the "optimal" (i.e., importance sampling) estimator with respect to the particular proposal distribution. We also gave an explicit proof about the properties of this estimator, namely, strong consistency and asymptotic normality in the special case where the state space of the target distribution if finite. Moreover, we proved that in the case of IMH (and finite state space), our estimator is indeed better than the optimal one.

Having run many simulations beyond those presented in Section 3, we strongly believe that

the estimator converges for general state space as well while the efficiency result holds always in the case of IMH. Unfortunately, we were not able to give a formal proof mainly due to the complicated form of the estimated weights. However, we would suggest to the practitioners who run IMH algorithms to apply our method to the generated samples because apparently improves the original MH estimators and is not time consuming.

Our approach can be extended to the case of Metropolis-within-Gibbs sampling schemes. Assume for instance that the target distribution is $\pi_{U,Y}(\cdot,\cdot)$ with full conditionals $\pi_{U|Y}(\cdot|y)$ and $\pi_{Y|U}(\cdot|u)$. Suppose that it is hard to sample directly from $\pi_{Y|U}(\cdot|u)$ and so, a Metropolis step with target this full conditional is applied using some proposal distribution $q(\cdot|u,y)$. If (U_t,Y_t), $t=1,2,\ldots$, is the generated sequence, then in the marginal sequence Y_t, $t=1,2,\ldots$, consists of repetitions of corresponding accepted proposals X_n, $n=1,2,\ldots$, say, that appear ξ_n, $n=1,2,\ldots$, times, respectively. Then, a result similar to Proposition 1 for the sequence (X_n,ξ_n), $n=1,2,\ldots$, can be shown. In particular, it holds $E(\xi_n|X_n=x)=\kappa w(x)$ where

$$\begin{aligned}\{\kappa w(x)\}^{-1} &= \iint \frac{\min\{\pi_{Y|U}(x|u)q(z|u,x),\pi_{Y|U}(z|u)q(x|u,z)\}}{\pi_Y(x)\pi_{Y|U}(z|u)}\,\pi_{U,Y}(u,z)\mathrm{d}u\mathrm{d}z\\ &= \frac{1}{\pi_Y(x)}E_\pi\left[\pi_{Y|U}(x|U)\min\left\{\frac{q(Z|U,x)}{\pi_{Y|U}(Z|U)},\frac{q(x|U,Z)}{\pi_{Y|U}(x|U)}\right\}\right].\end{aligned}$$

Clearly, this quantity can be estimated via the original sequence (U_t,Y_t), $t=1,2,\ldots$. Suppose now that one wishes to estimate the expectation of a function h depending solely on Y. Since its standard estimator has the form $\hat{h}_{MH}$ in (2), everything works like in the previous sections. In fact, in many toy examples we ran, we got results similar to those in Section 3.

Appendix: Proof of Theorem 1

(a) Assume without loss of generality that the state space is $\mathcal{X}=\{1,\ldots,m\}$ for some $m\geq 2$ and set for convenience $h_k=h(k)$, $\pi_k=\pi(k)$, $g_k=g(k)$, $g_{kl}=g(l|k)$ and $w_k=w(k)$. Then,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \hat{w}(X_i)h(X_i) &= \frac{1}{n}\sum_{i=1}^n \frac{h(X_i)}{\sum_{j=1}^n \frac{\xi_j}{\sum_{l=1}^n\xi_l}\min\left\{\frac{q(X_j|X_i)}{\pi(X_j)},\frac{q(X_i|X_j)}{\pi(X_i)}\right\}}\\ &= \sum_{k=1}^m \frac{1}{n}\sum_{i=1}^n \frac{h_k}{\sum_{j=1}^n \frac{\xi_j}{\sum_{l=1}^n\xi_l}\min\left\{\frac{q(X_j|k)}{\pi(X_j)},\frac{q(k|X_j)}{\pi_k}\right\}}I(X_i=k)\\ &= \sum_{k=1}^m \frac{h_k\frac{1}{n}\sum_{i=1}^n I(X_i=k)}{\sum_{\ell=1}^m \min\left\{\frac{q_{k\ell}}{\pi_\ell},\frac{q_{\ell k}}{\pi_k}\right\}\sum_{j=1}^n\frac{\xi_j}{\sum_{l=1}^n\xi_l}I(X_j=\ell)}\end{aligned} \tag{5}$$

By the ergodic theorem it holds $n^{-1}\sum_{i=1}^n I(X_i=k) \xrightarrow{\text{a.s.}} g_k$ and $\sum_{j=1}^n \frac{\xi_j}{\sum_{l=1}^n \xi_l} I(X_j=\ell) \xrightarrow{\text{a.s.}} \pi_\ell$. Since (5) contains only finite sums we conclude that it converges almost surely to

$$\sum_{k=1}^m \frac{h_k g_k}{\sum_{\ell=1}^m \min\left\{\frac{q_{k\ell}}{\pi_\ell}, \frac{q_{\ell k}}{\pi_k}\right\}\pi_\ell} = \sum_{k=1}^m \frac{h_k g_k}{\frac{1}{\pi_k}\sum_{\ell=1}^m \min\{\pi_k q_{k\ell}, \pi_\ell q_{\ell k}\}} = \kappa \sum_{k=1}^m h_k \pi_k = \kappa E_\pi(h)$$

because $g_k = \kappa \sum_{i=1}^m \min\{\pi_k g_{k\ell}, \pi_\ell g_{\ell k}\}$. By taking $h \equiv 1$ we get that $n^{-1}\sum_{i=1}^n \hat{w}(X_i) \xrightarrow{\text{a.s.}} \kappa$ and thus

$$\hat{h}_{\hat{w}} = \frac{\sum_{i=1}^n \hat{w}(X_i)h(X_i)/n}{\sum_{i=1}^n \hat{w}(X_i)/n} \xrightarrow{\text{a.s.}} E_\pi(h).$$

(b) Let us define

$$\bar{U}_k = \frac{1}{n}\sum_{i=1}^n I(X_i=k), \quad k=1,\ldots,m,$$

$$\bar{V}_k = \frac{1}{n}\sum_{i=1}^n \xi_i I(X_i=k), \quad k=1,\ldots,m,$$

and

$$\rho_{kl} = \min\left\{\frac{q_{lk}}{\pi_k}, \frac{q_{lk}}{\pi_l}\right\}, \quad k,l=1,\ldots,m.$$

Billingsley (1961) proved that

$$n^{1/2}(\bar{\boldsymbol{U}} - \boldsymbol{g}) \to \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{11})$$

where $\bar{\boldsymbol{U}} = (\bar{U}_1,\ldots,\bar{U}_m)^T$, $\boldsymbol{g} = (g_1,\ldots,g_m)^T$ and the ij entry of $\boldsymbol{\Sigma}_{11}$ is

$$\sigma_{11}(ij) = \delta_{ij}g_i - g_i g_j + g_i \sum_{n=1}^\infty (g_{ij}^{(n)} - g_j) + g_j \sum_{n=1}^\infty (g_{ji}^{(n)} - g_i).$$

Here $g_{ij}^{(n)}$ denotes the n-step transition probability from state i to state j and δ_{ij} is Kronecker's delta. Using similar arguments it can be proven that

$$n^{1/2}\Big(\begin{pmatrix}\bar{\boldsymbol{U}}\\ \bar{\boldsymbol{V}}\end{pmatrix} - \begin{pmatrix}\boldsymbol{g}\\ \kappa\boldsymbol{\pi}\end{pmatrix}\Big) \to \mathcal{N}_{2m}\Big(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\Big)$$

where the ij entry of the submatrix $\boldsymbol{\Sigma}_{12}$ is $\sigma_{12}(ij) = \kappa w_j \sigma_{11}(ij)$ while the ij entry of the submatrix $\boldsymbol{\Sigma}_{22}$ is $\sigma_{22}(ij) = \delta_{ij}g_i\kappa w_i(\kappa w_i - 1) + \kappa^2 w_i w_j \sigma_{11}(ij)$.

It is clear that the asymptotic distribution of $n^{1/2}\{\hat{h}_{\hat{w}} - E_\pi(h)\}$ can be found via the standard delta method. Observe that $\hat{h}_{\hat{w}}$ can be expressed as

$$\hat{h}_{\hat{w}} = \sum_{k=1}^m \frac{h_k \bar{U}_k}{\sum_{l=1}^m \rho_{kl}\bar{V}_l} \Big/ \sum_{k=1}^m \frac{\bar{U}_k}{\sum_{l=1}^m \rho_{kl}\bar{V}_l} = f_1(\bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}), \tag{6}$$

say. By differentiation we get

$$\frac{\partial}{\partial u_i} f_1(\boldsymbol{u},\boldsymbol{v})\Big|_{(\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{g},\kappa\boldsymbol{\pi})} = w_i\{h_i - E_\pi(h)\}$$

and

$$\frac{\partial}{\partial v_i} f_1(\boldsymbol{u},\boldsymbol{v})\Big|_{(\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{g},\kappa\boldsymbol{\pi})} = \frac{1}{\kappa w_i} E\left\{w(X_1)(h_i - E_\pi(h))|X_0 = i\right\}.$$

The variance of the asymptotic normal distribution of $n^{1/2}\{\hat{h}_{\hat{w}} - E_\pi(h)\}$ is

$$\sigma^2_{\hat{w}}(h) = \nabla_{\boldsymbol{u}} f_1^T \boldsymbol{\Sigma}_{11} \nabla_{\boldsymbol{u}} f_1 + 2\nabla_{\boldsymbol{u}} f_1^T \boldsymbol{\Sigma}_{12} \nabla_{\boldsymbol{v}} f_1 + \nabla_{\boldsymbol{u}} f_1^T \boldsymbol{\Sigma}_{22} \nabla_{\boldsymbol{u}} f_1,$$

where $\nabla_{\boldsymbol{u}} f_1$, $\nabla_{\boldsymbol{v}} f_1$ are the vectors containing the derivatives with respect to $\boldsymbol{u}$, $\boldsymbol{v}$, respectively, evaluated at $(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{g},\kappa\boldsymbol{\pi})$. We will see below that the above expression is in fact as stated in the theorem. To this end, let us also consider the IS "estimator" $\hat{h}_{IS}$ and evaluate its asymptotic variance too. After some algebra we get that

$$\hat{h}_{IS} = \sum_{k=1}^m \frac{h_k \bar{U}_k}{\sum_{l=1}^m \rho_{kl}\pi_l} \Big/ \sum_{k=1}^m \frac{\bar{U}_k}{\sum_{l=1}^m \rho_{kl}\pi_l} = f_2(\bar{\boldsymbol{U}}), \tag{7}$$

say. By (6) and (7) we see that the only difference between the two "estimators" is the replacement of π_l by its unbiased estimate $\kappa^{-1}\bar{V}_l$ in $\hat{h}_{\hat{w}}$. By differentiation we get

$$\frac{\partial}{\partial u_i} f_2(\boldsymbol{u})\Big|_{\boldsymbol{u}=\boldsymbol{g}} = w_i\{h_i - E_\pi(h)\}.$$

Since the corresponding variance of the importance sampling estimator is

$$\sigma^2_{IS}(h) = \nabla_{\boldsymbol{u}} f_2^T \boldsymbol{\Sigma}_{11} \nabla_{\boldsymbol{u}} f_2 \equiv \nabla_{\boldsymbol{u}} f_1^T \boldsymbol{\Sigma}_{11} \nabla_{\boldsymbol{u}} f_1,$$

it is clear that

$$\sigma^2_{\hat{w}}(h) - \sigma^2_{IS}(h) = -2\nabla_{\boldsymbol{u}} f_1^T \boldsymbol{\Sigma}_{12} \nabla_{\boldsymbol{v}} f_1 - \nabla_{\boldsymbol{u}} f_1^T \boldsymbol{\Sigma}_{22} \nabla_{\boldsymbol{u}} f_1.$$

Set now $A_1 = -2\nabla_{\boldsymbol{u}} f_1^T \boldsymbol{\Sigma}_{12} \nabla_{\boldsymbol{v}} f_1$, $A_2 = -\nabla_{\boldsymbol{u}} f_1^T \boldsymbol{\Sigma}_{22} \nabla_{\boldsymbol{u}} f_1$, $B_n = w(X_n)h(X_n)$ for $n = 0, 1, \ldots$, and $b_i = w_i h_i$, $\mu_i = E(B_1|X_0 = i)$ for $i = 1, \ldots, m$. Consider further without loss of generality that $E_\pi(h) = 0$ and note that under stationarity, it holds that $E_\pi(h) = E_g(B_n)$ for all n. Then,

$$\begin{aligned}
A_1 = & -2\sum_{i=1}^m \sum_{j=1}^m b_i \Big(- \frac{\mu_j}{\kappa w_j}\Big)\kappa w_j \sigma_{11}(ij) \\
= & -2\sum_{i=1}^m \sum_{j=1}^m b_i \mu_j \Big\{ \delta_{ij} g_i - g_i g_j + g_i \sum_{n=1}^\infty (g_{ij}^{(n)} - g_j) + g_j \sum_{n=1}^\infty (g_{ji}^{(n)} - g_i) \Big\} \\
= & \ 2\Big\{ \sum_{i=1}^m b_i \mu_i g_i - \sum_{i=1}^m b_i g_i \sum_{j=1}^m \mu_j g_j + \sum_{i=1}^m b_i g_i \sum_{j=1}^m \mu_j \sum_{n=1}^\infty (g_{ij}^{(n)} - g_j) + \\
& \sum_{j=1}^m \mu_j g_j \sum_{i=1}^m b_i \sum_{n=1}^\infty (g_{ji}^{(n)} - g_i) \Big\} \\
= & \ 2\Big\{ \sum_{i=1}^m b_i \mu_i g_i - \sum_{i=1}^m b_i g_i \sum_{j=1}^m \mu_j g_j + \sum_{i=1}^m b_i g_i \sum_{n=1}^\infty \Big(\sum_{j=1}^m \mu_j g_{ij}^{(n)} - \sum_{j=1}^m \mu_j g_j \Big) +
\end{aligned}$$

$$\sum_{j=1}^{m}\mu_j g_j \sum_{n=1}^{\infty}\Big(\sum_{i=1}^{m} b_i g_{ji}^{(n)} - \sum_{i=1}^{m} b_i g_i\Big)\Big\}.$$

But

$$\sum_{i=1}^{m}\mu_i g_i = \sum_{i=1}^{m} E(B_i|X_0=i)g_i = E_g\{E(B_1|X_0)\} = E_g(B_1) = 0$$

and since X is reversible, i.e., it holds $g_i g_{ij} = g_j g_{ji}$ and more generally $g_i g_{ij}^{(n)} = g_j g_{ji}^{(n)}$ for all n, we conclude that

$$A_1 = 2\Big\{\sum_{i=1}^{m} b_i\mu_i g_i + 2\sum_{n=1}^{\infty}\sum_{i=1}^{m}\sum_{j=1}^{m} b_i g_i \mu_j g_{ij}^{(n)}\Big\}.$$

Moreover,

$$\sum_{i=1}^{m} b_i\mu_i g_i = \sum_{i=1}^{m} b_i\Big(\sum_{k=1}^{m} b_k g_{ik}\Big)g_i = E_g(B_0B_1)$$

and for all n, we have that

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{m} b_i g_i\mu_j g_{ij}^{(n)} &= \sum_{i=1}^{m}\sum_{j=1}^{m} b_i g_i\Big(\sum_{k=1}^{m} b_k g_{jk}\Big)g_{ij}^{(n)}\\ &= \sum_{i=1}^{m}\sum_{k=1}^{m} b_i b_k g_i\sum_{j=1}^{m} g_{jk}g_{ij}^{(n)} - \sum_{i=1}^{m}\sum_{k=1}^{m} b_i b_k g_i g_{ik}^{(n+1)} = E_g(B_0B_{n+1}).\end{aligned}$$

Thus,

$$A_1 = 2E_g(B_0B_1) + 4\sum_{n=1}^{\infty} E_g(B_0B_{n+1}) = 2\sum_{n=1}^{\infty} E_g(B_0B_n) + 2\sum_{n=1}^{\infty} E_g(B_0B_{n+1}). \tag{8}$$

On the other hand,

$$\begin{aligned}A_2 = & -\sum_{i=1}^{m}\sum_{j=1}^{m}\Big(-\frac{\mu_i}{\kappa w_i}\Big)\Big(-\frac{\mu_j}{\kappa w_j}\Big)\{\delta_{ij}g_i\kappa w_i(\kappa w_i - 1) + \kappa^2 w_i w_j\sigma_{ij}\}\\ = & -\sum_{i=1}^{m}\frac{\mu_i^2}{\kappa^2 w_i^2} g_i\kappa w_i(\kappa w_i - 1)\\ & -\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_i\mu_j\Big\{\delta_{ij}g_i - g_i g_j + g_i\sum_{n=1}^{\infty}(g_{ij}^{(n)} - g_j) + g_j\sum_{n=1}^{\infty}(g_{ji}^{(n)} - g_i)\Big\}\\ = & \sum_{i=1}^{m}\frac{\mu_i^2 g_i}{\kappa w_i} - 2\sum_{i=1}^{m}(\mu_i)^2 g_i - \sum_{i=1}^{m}\sum_{j=1}^{m}\mu_i\mu_j g_i g_j\\ & -\sum_{i=1}^{m}\mu_i g_i\sum_{j=1}^{m}\mu_j\sum_{n=1}^{\infty}(g_{ij}^{(n)} - g_j) - \sum_{j=1}^{m}\mu_j g_j\sum_{i=1}^{m}\mu_i\sum_{n=1}^{\infty}(g_{ji}^{(n)} - g_i)\\ = & \sum_{i=1}^{m}\frac{\mu_i^2 g_i}{\kappa w_i} - 2\sum_{i=1}^{m}(\mu_i)^2 g_i - \sum_{i=1}^{m}\sum_{j=1}^{m}\mu_i\mu_j g_i g_j\\ & -\sum_{n=1}^{\infty}\sum_{i=1}^{m}\mu_i g_i\Big(\sum_{j=1}^{m}\mu_j g_{ij}^{(n)} - \sum_{j=1}^{m}\mu_j g_j\Big) - \sum_{n=1}^{\infty}\sum_{j=1}^{m}\mu_j g_j\Big(\sum_{i=1}^{m}\mu_i g_{ji}^{(n)} - \sum_{i=1}^{m}\mu_i g_i\Big)\end{aligned}$$

$$= \sum_{i=1}^{m} \frac{\mu_i^2 g_i}{\kappa w_i} - 2\sum_{i=1}^{m} \mu_i^2 g_i - 2\sum_{n=1}^{\infty}\sum_{i=1}^{m}\sum_{j=1}^{m} \mu_i g_i \mu_j g_{ij}^{(n)}.$$

But

$$\sum_{i=1}^{m} \mu_i^2 g_i = \sum_{i=1}^{m} \Big(\sum_{k=1}^{m} b_k g_{ik}\Big)\Big(\sum_{\ell=1}^{m} b_\ell g_{i\ell}\Big) g_i = \sum_{k=1}^{m}\sum_{\ell=1}^{m} b_k b_\ell \sum_{i=1}^{m} g_i g_{ik} g_{i\ell} =$$
$$\sum_{k=1}^{m}\sum_{\ell=1}^{m} b_k b_\ell g_k \sum_{i=1}^{m} g_{ki} g_{i\ell} = \sum_{k=1}^{m}\sum_{\ell=1}^{m} b_k b_\ell g_k g_{k\ell}^{(2)} = E_g(B_0 B_2),$$

and for all n,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \mu_i g_i \mu_j g_{ij}^{(n)} = \sum_{i=1}^{m}\sum_{j=1}^{m} \Big(\sum_{k=1}^{m} b_k g_{ik}\Big)\Big(\sum_{\ell=1}^{m} b_\ell g_{j\ell}\Big) g_i g_{ij}^{(n)} =$$
$$\sum_{k=1}^{m}\sum_{\ell=1}^{m} b_k b_\ell \sum_{i=1}^{m} g_i g_{ik} \sum_{j=1}^{m} g_{j\ell} g_{ij}^{(n)} = \sum_{k=1}^{m}\sum_{\ell=1}^{m} b_k b_\ell g_k \sum_{i=1}^{m} g_{ki} g_{i\ell}^{(n+1)} =$$
$$\sum_{k=1}^{m}\sum_{\ell=1}^{m} b_k b_\ell g_k g_{k\ell}^{(n+2)} = E_g(B_0 B_{n+2}).$$

Hence,

$$A_2 = \sum_{i=1}^{m} \frac{\mu_i^2 g_i}{\kappa w_i} - 2E_g(B_0 B_2) - 2\sum_{n=1}^{\infty} E_g(B_0 B_{n+2}) = \sum_{i=1}^{m} \frac{\mu_i^2 g_i}{\kappa w_i} - 2\sum_{n=1}^{\infty} E_g(B_0 B_{n+1}). \tag{9}$$

From (8) and (9) we get

$$\sigma_{IS}^2(h) - \sigma_{\hat{w}}^2(h) = A_1 + A_2 = \sum_{i=1}^{m} \frac{\mu_i^2 g_i}{\kappa w_i} + 2\sum_{n=1}^{\infty} E_g(B_0 B_n). \tag{10}$$

By the standard asymptotic theory for Markov chains $\sigma_{IS}^2(h) = E_g(B_0^2) + 2\sum_{n=1}^{\infty} E_g(B_0 B_n)$. Since $E_g(B_0^2) = Var_g\{w(X_0)h(X_0)\}$ part (b) of the theorem follows.

(c) In order to compare the two variances note first that the first term of the sum in (10) is clearly positive. Moreover, we know that $\sum_{n=2}^{\infty} E_g(B_0 B_n) \equiv \sum_{n=2}^{\infty} Cov_g(B_0, B_n) \geqslant 0$ since each residual sum of autocovariances starting from an even integer is nonnegative (see for example Geyer, 1992). Thus a sufficient condition for the variances difference to be positive is $E_g(B_0 B_1) \geqslant 0$. This expectation can be expressed as a quadratic form, namely,

$$E_g(B_0 B_1) = \sum_{i=1}^{m}\sum_{j=1}^{m} b_i b_j \tilde{g}_{ij} = \boldsymbol{b}^T \tilde{\boldsymbol{G}} \boldsymbol{b}$$

where $\tilde{g}_{ij} = g_i g_{ij}$. We will show that in the case of independence MH, the matrix $\tilde{\boldsymbol{G}}$ is nonnegative definite. Indeed, in this case,

$$\tilde{g}_{ij} = g_i g_{ij} = \frac{\sum_{k=1}^{m} \min\{q_i \pi_k, q_k \pi_i\}}{\kappa} \times \frac{\min\{q_i \pi_j, q_j \pi_i\}}{\sum_{k=1}^{m} \min\{q_i \pi_k, q_k \pi_i\}} = \kappa^{-1} \min\{q_i \pi_j, q_j \pi_i\}.$$

Consider without loss of generality that

$$q_1/\pi_1 \leqslant \cdots \leqslant q_m/\pi_m \tag{11}$$

and set $\gamma_{ij} = \kappa^{-1} q_i \pi_j$. Then clearly $\tilde{g}_{ij} = \gamma_{i\wedge j, i\vee j}$ where $i \wedge j = \min\{i, j\}$ and $i \vee j = \max\{i, j\}$ so $\tilde{\boldsymbol{G}}$ is nonnegative definite. To see that, proceed by induction to show that all its principal minors are nonnegative. Let $\tilde{\boldsymbol{G}}_{kk}$ denote the matrix consisting of the first k rows and columns of $\tilde{\boldsymbol{G}}$. Then, $|\tilde{\boldsymbol{G}}_{11}| \equiv \gamma_{1\wedge 1} = \kappa^{-1} q_1 \pi_1 > 0$. Suppose now that $|\tilde{\boldsymbol{G}}_{kk}| \geqslant 0$ for some $k \geqslant 1$. In order to prove that $|\tilde{\boldsymbol{G}}_{k+1,k+1}| \geqslant 0$, multiply the k-th row of $\tilde{\boldsymbol{G}}_{k+1,k+1}$ by π_{k+1}/π_k and subtract it from the $(k+1)$-th. Then the resulting matrix has all last row's elements equal to zero except from the last one which is $\kappa^{-1} q_{k+1}\pi_{k+1} - \kappa^{-1} q_k \pi_{k+1}^2/\pi_k = \kappa^{-1}\pi_{k+1}^2(q_{k+1}/\pi_{k+1} - q_k/\pi_k)$. By (11) this quantity is nonnegative thus, $|\tilde{\boldsymbol{G}}_{k+1,k+1}| = \kappa^{-1}\pi_{k+1}^2(q_{k+1}/\pi_{k+1} - q_k/\pi_k)|\tilde{\boldsymbol{G}}_{kk}| \geqslant 0$.

Acknowledgements

The authors wish to thank the two anonymous referees for their helpful comments and suggestions which considerably improved the paper.

References

Atchadé, Y.F. and Perron, F. (2005). Improving on the independent Metropolis– Hastings algorithm. *Statistica Sinica*, **15**, 3–18.

Billingsley, P. (1961). Statistical methods in Markov chains. *Annals of Mathematical Statistics*, **32**, 12–40.

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

Douc, R. and Robert C.P. (2011). A vanilla Rao–Blackwellisation of Metropolis-Hastings algorithms. *Annals of Statistics*, **39**, 261–277.

Geyer, C.J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473–483.

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

Jacob, P., Robert, C.P., and Smith, M. (2011). Using parallel computation to improve independent Metropolis–Hastings based estimation. *Journal of Computational and Graphical Statistics*, **20**, 616–635.

Malefaki, S. and Iliopoulos, G. (2008). On convergence of properly weighted samples to the target distribution. *Journal of Statistical Planning and Inference*, **138**, 1210–1225.

Marin, J.M. and Robert, C.P. (2010). Importance sampling methods for Bayesian discrimination between embedded models. In Frontiers of Statistical Decision Making and Bayesian Analysis (eds., M.-H. Chen, D.K. Dey, P. Müller, D. Sun, K. Ye). Chapter 14, pp.513–553.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1091.