A comparison of Rejection Sampling and Metropolis-Hastings algorithm.

Bjørn Kåre Hegstad and Henning Omre
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim
Norway

1 Introduction

When doing stochastic modeling one often face the problem of sampling from a posterior pdf on the form $\pi(\cdot) = \text{const} \cdot l(\cdot)p(\cdot)$, where the normalizing constant is unknown. The corresponding likelihood function $l(\cdot)$ tends to be complicated and computational expensive while the prior $p(\cdot)$ tends to be simpler. This work origin from stochastic reservoir characterization and history matching where calculation of the likelihood function requires a fluid flow simulation which may take hours or even days on a fast computer. The prior may also have a complicated form, but a sample from the prior is usually relatively fast obtained. Analytical forms for the posterior are usually available only in simple cases. The Metropolis-Hastings algorithm is widely used to sample posterior pdfs where the normalizing constant is unknown. It is a class of Markov chain Monte Carlo algorithm defined by Hastings (1970). The Metropolis-Hastings algorithm is iterative and hence time consuming and it is difficult to diagnose convergence. Moreover the realizations from the Markov chain are usually highly dependent. The Rejection Sampling algorithm could also be used, see Ripley (1987). This algorithm gives independent samples from $\pi(\cdot)$, but requires the normalizing constant or an upper bound for it, to be known which may be impossible in practice. Independent samples could however be convenient when estimating various quantities from the same sample from the posterior pdf.

A comparison between Metropolis-Hasting and Rejection Sampling is performed by introducing a new algorithm converging to a Markov chain Monte Carlo version of Rejection Sampling. This new algorithm does not require knowledge of the normalizing constant.

The discussion is made in a discrete, finite sample space $\Omega = \{1, 2, \ldots, m\}$. The probability measure π, and p over Ω can be interpreted in a Bayesian framework and be related as $\pi_i = w_i p_i$; $i = 1, \ldots, m$. Hence π is termed the posterior distribution, p is termed the prior distribution and w is termed the likelihood. Note that the unknown normalizing constant is included in the definition of w. In Liu (1996) w is termed importance ratio, but this is only a matter of semantics. The objective is to sample from π, usually being complex and only known up to a constant, with p a given distribution relatively simple to sample from. Hence the likelihood w will be known up to a constant.

The Metropolis-Hastings algorithm is usually applied to sample from π. In this article the special version, Metropolis-Hastings Independence Proposal (MH/IP) algorithm where the prior pdf defines the proposal function, will be considered. It is defined as follows:

Initiate

$$x_0 = k \text{ with } k \in \Omega$$

Iterate $n = 0, 1, 2, \ldots$

$$\begin{aligned} y &\sim \Pr\{Y = y\} = p_y \ : y = 1, \ldots, m \\ w_{n+1} &= \pi_y / p_y \\ x_{n+1} &\sim \begin{cases} \Pr\{X_{n+1} = y\} = \min\left\{1, \frac{w_{n+1}}{w_n}\right\} \\ \Pr\{X_{n+1} = x_n\} = 1 - \min\left\{1, \frac{w_{n+1}}{w_n}\right\} \end{cases} \end{aligned}$$

where the symbol $\sim$ is to be read 'with probability distribution according to'.

It is well known, see Hastings (1970), that this ensures that

$$\lim_{n \to \infty} \Pr\{X_n = j\} = \pi_j \ ; j = 1, \ldots, m.$$

The subsequent samples in the chain will be highly dependent regardless of the independent proposals since the acceptance probability is a function of the previous state and is not always equal one.

In this article, an alternative algorithm with the same prerequisites as the MH/IP algorithm is presented. It is termed Reducible Markov chain Monte Carlo Dynamic Rejection Sampling (RMcMC/DRS) algorithm, and it is defined as:

Initiate

$$x_0 = k \text{ with } k \in \Omega$$

Iterate $n = 0, 1, 2, \ldots$

$$\begin{aligned} y &\sim \Pr\{Y = y\} = p_y \ ; y = 1, \ldots, m \\ w_{n+1} &= \pi_y / p_y \\ x_{n+1} &= \begin{cases} \Pr\{X_{n+1} = y\} = \frac{w_{n+1}}{\max\{w_j ; j=0,1,\ldots,n+1\}} \\ \Pr\{X_{n+1} = x_n\} = 1 - \frac{w_{n+1}}{\max\{w_j ; j=0,1,\ldots,n+1\}} \end{cases} \end{aligned}$$

It is shown in Section 4 that this algorithm is not within the class of Markov chain Monte Carlo (McMC) algorithms as defined in Hastings (1970). Moreover it is shown that the algorithm converges towards a Markov chain Monte Carlo version of the Rejection Sampling (McMC/RS) algorithm. This entails that

$$\lim_{n \to \infty} \Pr\{X_n = j\} = \pi_j \ ; j = 1, \ldots, m.$$

The subsequent samples in the RMcMC/DRS algorithm is highly dependent, but it is shown that the embedded process produces independent samples from π after convergence. The definition of the embedded process is:

Definition 1 *The embedded process is defined as the accepted proposals in a Markov chain.*

Hence only the samples appearing from accepting of a new proposed state are considered. In Section 3 the McMC/RS algorithm will be discussed separately. Note, however, that this algorithm is not completely comparable to the MH/IP algorithm since the former requires the knowledge of a constant $w_0 \geq w_j = \pi_j/p_j$; $j = 1, \ldots, m$.

From the algorithms above one can observe that the values of the likelihoods, w, play a central role. Without loss of generality the sample space $\Omega = \{1, \ldots, m\}$ is ordered such that $w_1 \geq w_2 \geq \cdots \geq w_m$.

2 Metropolis-Hastings/Independent Proposal (MH/IP) algorithm

The results in this section is primarily taken from Liu (1996) hence all proofs are omitted. The MH/IP algorithm can be defined as:

Definition 2 *The MH/IP algorithm with posterior distribution π and proposal prior distribution p over the sample space $\Omega = \{1, \ldots, m\}$ is defined by the irreducible, aperiodic Markov chain $\{X_n\}$ with $X_n \in \{1, \ldots, m\}$ and $n = \{0, 1, \ldots\}$, with*

Initial state:

$$X_0 = k \text{ with } k \in \Omega$$

Transition matrix:

$$K(i,j) = \begin{cases} p_j \min\left\{1, \frac{w_j}{w_i}\right\} & \text{for } i \neq j \\ 1 - \sum\limits_{l \neq i} p_l \max\left\{1, \frac{w_l}{w_i}\right\} & \text{for } i = j \end{cases}$$

for $i, j = 1, \ldots, m$ and $w_i = \pi_i/p_i$; $i = 1, \ldots, m$

The transition matrix has the following form:

$$K(\cdot, \cdot) =$$

	1	2	3	$\cdots$	m
1	$\pi_1/w_1 + \lambda_1$	π_2/w_1	π_3/w_1	$\cdots$	π_m/w_1
2	p_1	$\pi_2/w_2 + \lambda_2$	π_3/w_2	$\cdots$	π_m/w_2
3	p_1	p_2	$\pi_3/w_3 + \lambda_3$	$\cdots$	π_m/w_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	p_1	p_2	p_3	$\cdots$	$\pi_m/w_m + \lambda_m$

with $\lambda_i = \sum\limits_{l=i}^{m} (p_l - \pi_l/w_i)$; $i = 1, \ldots, m$, being the probability of being rejected in the next iteration given that the chain is in state i. Note in particular that $\lambda_1 = 1 - \frac{1}{w_1}$ and $\lambda_m = 0$.

The characteristics of the transition matrix are given in the following theorem:

Theorem 1 *The transition matrix of the MH/IP algorithm as specified in Definition 2, has:*
Eigenvalues:

$$\lambda_0 = 1 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m-1} \geq 0$$

with

$$\lambda_i = \sum_{l=i}^{m} \left(p_l - \frac{\pi_l}{w_i} \right) \;; i = 1, \ldots, m-1$$

Eigenvectors:

$$\nu_0 = (\underbrace{1, \ldots, 1}_{m})$$

$$\nu_i = (\underbrace{0, \ldots, 0}_{i-1}, \sum_{l=i+1}^{m} \pi_l, \underbrace{-\pi_i, \ldots, -\pi_i}_{m-i}) \;; i = 1, \ldots, m-1$$

Note in particular that the second largest eigenvalue is $\lambda_1 = (1 - 1/w_1)$, hence only dependent on the largest likelihood value.

The total variation distance between two distributions is used extensively here and is defined as

Definition 3 *The total variation distance between two distributions μ and π on $\Omega = \{1, \ldots, m\}$ is defined by:*

$$\|\mu - \pi\| = \frac{1}{2} \sum_{i=1}^{m} |\mu_i - \pi_i|$$

The following lemma according to Diaconis and Hanlon (1992) is useful in the proceeding theorems:

Lemma 1 *Let $\{X_n\}$ be a reversible, irreducible, aperiodic Markov chain with $\{X_n\} \in \{1, \ldots, m\}$ and $n \in \{0, 1, \ldots\}$. The initial state is $X_0 = k$ and the transition matrix is $K(i, j)$ with π as the limiting distribution. Then the total variation distance is*

$$\|K^n(k, \cdot) - \pi(\cdot)\| \leq \frac{\lambda_1^n}{2\sqrt{\pi_k}}$$

with λ_1 being the second largest eigenvalue in absolute value.

It follows immediately from this:

Corollary 1 *The total variation distance between the distribution after n steps in the MH/IP algorithm specified in Definition 2, and the limiting distribution π is bounded by:*

$$\|K^n(k, \cdot) - \pi(\cdot)\| \leq \frac{\lambda_1^n}{2\sqrt{\pi_k}} = \frac{1}{2\sqrt{\pi_k}} \left(1 - \frac{1}{w_1} \right)^n$$

with $w_1 = \pi_1 / p_1$ being the largest likelihood value.

Hence the convergence rate of the MH/IP algorithm is dominated by the second largest eigenvalue, λ_1, being a function of the largest likelihood value only. This corresponds to the convergence bound normally used in the evaluation of McMC algorithms.

After convergence, the samples X_n from the MH/IP algorithm will be according to π. Subsequent samples will be highly dependent however. An estimate of the expected value of an arbitrary function of X has the following characteristics:

Theorem 2 *Let $X \sim \pi$ and define an arbitrary function $h(x)$ such that:*

$$\begin{aligned} E_\pi\{h(X)\} &= \mu < \infty \\ Var_\pi\{h(X)\} &= 1 \end{aligned}$$

Consider $X_1, \ldots, X_n$ as samples from the stationary pdf using MH/IP algorithm in Definition 2. Define:

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

then

$$E\{\tilde{\mu}\} = \mu$$
$$Var\{\tilde{\mu}\} \leq \frac{2}{n(1-\lambda_1)} = \frac{2w_1}{n}$$

with $w_1 = \pi_1/p_1$ being the largest likelihood value.

Hence the estimation variance of μ is dominated by the largest likelihood value. The bound $2/(n(1-\lambda_1))$ is actually valid for any reversible McMC algorithm. Since all eigenvalues in the MH/IP algorithm are larger than or equal to zero, the variance can not be smaller than for the case with independent samples corresponding to $\lambda_1 = 0$. If λ_1 could be smaller than zero, subsequent samples would be negatively correlated and variance smaller then for the independent case.

3 Markov chain Monte Carlo/Rejection Sampling (McMC/RS) algorithm

In this section a McMC version of the traditional Rejection Sampling algorithm is evaluated. Note that it is not directly comparable with the MH/IP algorithm since the former requires a constant $w_0 \geq w_1 = \pi_1/p_1$ to be known. This is in many cases not a realistic requirement. Note also that the McMC/RS algorithm is the converging process of the RMcMC/DRS algorithm discussed in the next section. This algorithm is comparable to the MH/IP algorithm. The McMC/RS algorithm has the following steps:

Initiate

$$x_0 = k \quad \text{with} \quad k \in \Omega$$

Iterate $n = 0, 1, 2, \ldots$

$$\begin{aligned} y &\sim \Pr\{Y = y\} = p_y \; ; y = 1, \ldots, m \\ w_{n+1} &= \pi_y / p_y \\ x_{n+1} &\sim \begin{cases} \Pr\{X_{n+1} = y\} = \frac{w_{n+1}}{w_0} \\ \Pr\{X_{n+1} = x_n\} = 1 - \frac{w_{n+1}}{w_0} \end{cases} \end{aligned}$$

with w_0 being a known constant such that $w_0 \geq w_1 = \pi_1/p_1$, the largest likelihood value. It will be shown that

$$\lim_{n \to \infty} \Pr\{X_n = i\} = \pi_i \; ; i = 1, \ldots, m$$

A more formal definition is:

Definition 4 *The McMC/RS algorithm with posterior distribution π and prior distribution p over the sample space $\Omega = \{1, \ldots, m\}$ is defined by the reversible, irreducible, aperiodic Markov chain $\{X_n\}$ with $X_n \in \{1, \ldots, m\}$ and $n \in \{0, 1, 2, \ldots\}$, and known parameter $w_0 \geq w_i = \pi_i/p_i \; \forall i \in \{1, \ldots, m\}$. This gives:*

Initial state:

$$X_0 = k \quad \textit{with} \quad k \in \Omega$$

Transition matrix:

$$K(i,j) = \begin{cases} p_j \frac{w_j}{w_0} & \textit{for } i \neq j \\ 1 - \sum\limits_{l \neq i} p_l \frac{w_l}{w_0} & \textit{for } i = j \end{cases}$$

for $i, j = 1, \ldots, m$ and $w_i = \pi_i/p_i \; ; i = 1, \ldots, m$.

The transition matrix has the following form:

$$K(\cdot, \cdot) =$$

	1	2	3	$\cdots$	m
1	$\pi_1/w_0 + \left(1 - \frac{1}{w_0}\right)$	π_2/w_0	π_3/w_0	$\cdots$	π_m/w_0
2	π_1/w_0	$\pi_2/w_0 + \left(1 - \frac{1}{w_0}\right)$	π_3/w_0	$\cdots$	π_m/w_0
3	π_1/w_0	π_2/w_0	$\pi_3/w_0 + \left(1 - \frac{1}{w_0}\right)$	$\cdots$	π_m/w_0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	π_1/w_0	$\cdots$	$\cdots$	$\cdots$	$\pi_m/w_0 + \left(1 - \frac{1}{w_0}\right)$

The limiting distribution is given by the following lemma:

Lemma 2 *The Markov chain $\{X_n\}$ is irreducible and aperiodic, hence the limiting distribution is identical to the stationary distribution $\mu(\cdot)$ defined by:*

$$\begin{array}{rcl} \mu(\cdot) & = & \overset{T}{\mu(\cdot)}\, K(\cdot,\cdot) \\ \mu^T\mathbf{1} & = & 1 \end{array}$$

with $\mu(\cdot)$ being a m-dimensional distribution and $\mathbf{1}$ being a m-vector of ones. This expression has one unique solution and it is trivial to show that $\pi(\cdot) = (\pi_1, \ldots, \pi_m)$ satisfy the expression.

Note also that the transition matrix can be decomposed into:

$$K(\cdot,\cdot) = K(\cdot,\cdot|\text{accept}) \cdot \Pr\{\text{accept}\} + K(\cdot,\cdot|\text{reject})\Pr\{\text{reject}\}$$

with

$$K(\cdot,\cdot|\text{accept}) = \begin{array}{c|ccccc} & 1 & 2 & 3 & \cdots & m \\ \hline 1 & \pi_1 & \pi_2 & \pi_3 & \cdots & \pi_m \\ 2 & \pi_1 & \pi_2 & \pi_3 & \cdots & \pi_m \\ 3 & \pi_1 & \pi_2 & \pi_3 & \cdots & \pi_m \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ m & \pi_1 & \pi_2 & \pi_3 & \cdots & \pi_m \end{array}$$

$$K(\cdot,\cdot|\text{reject}) = \begin{array}{c|ccccc} & 1 & 2 & 3 & \cdots & m \\ \hline 1 & 1 & 0 & 0 & \cdots & 0 \\ 2 & 0 & 1 & 0 & \cdots & 0 \\ 3 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ m & 0 & 0 & 0 & \cdots & 1 \end{array}$$

and $\Pr\{\text{accept}\} = 1/w_0$ and $\Pr\{\text{reject}\} = 1 - 1/w_0$. The terms 'accept' and 'reject' refers to acceptance or rejection of a new value in the next step.

This decomposition leads to the following lemma:

Lemma 3 *The embedded process, see Definition 1, of the Markov chain representing the McMC/RS algorithm in Definition 4 is termed $\{X_{(n)}\}$ with $X_{(n)} \in \{1, \ldots, m\}$ and $n \in \{0, 1, 2, \ldots\}$.*

The samples of the embedded process, $X_{(1)}, X_{(2)}, \ldots$ are independently, identically distributed according to π.

Proof. The embedded process is defined by the samples appearing at each acceptance. Hence, the transition matrix representative for the embedded process is $K(\cdot,\cdot|\text{accept})$. It is obvious that this transition matrix provides iid samples from π. □

A direct consequence of this is that the McMC/RS algorithm converges when the first acceptance take place. The traditional Rejection Sampling algorithm corresponds to the embedded process of the McMC/RS algorithm.

The waiting times between each acceptance, and hence each sample in the embedded process is:

Lemma 4 *Let N be the number of steps between two subsequent acceptances in the Markov chain $\{X_n\}$ as specifies in Definition 4, then*

$$Pr\{N=n\} = \left(1-\frac{1}{w_0}\right)^{n-1}\frac{1}{w_0} \quad ; n=1,2,\ldots$$

Proof. The acceptances and rejections in each step are independent with $\Pr\{\text{accept}\} = \frac{1}{w_0}$ and $\Pr\{\text{reject}\} = 1-\frac{1}{w_0}$. Hence the distribution is geometric with the appropriate parameters.

□

In this section Rejection Sampling is defined through a Markov chain and the characteristics of the transition matrix can be evaluated. The following theorem states this:

Theorem 3 *The transition matrix of the McMC/RS algorithm as specified in Definition 4 has:*

Eigenvalues:

$$\lambda_0 = 1 \geq \lambda_1 = \lambda_2 = \cdots = \lambda_{m-1} \geq 0$$

with

$$\lambda_i = 1-\frac{1}{w_0}\,; i=1,\ldots,m-1$$

Eigenvectors:

$$\nu_0 = (\underbrace{1,\ldots,1}_{m})$$

$$\nu_i = (\underbrace{0,\ldots,0}_{i-1}, \sum_{l=i+1}^{m}\pi_l, \underbrace{-\pi_i,\ldots,-\pi_i}_{m-i}) \;; i=1,\ldots,m-1$$

Proof. From the symmetry in the transition matrix K it is obvious that $\nu_0 = (1,1,\ldots,1)$ is an eigenvector with eigenvalue 1 since $(K\nu_0)^{(k)} = \sum_{i=1}^m \pi_i/w_0 + 1 - 1/w_0 = 1$ for $k = 1,\ldots,m$ where $y^{(k)}$ denotes the kth element in the vector y. Hence ν_0 is an eigenvector with corresponding eigenvalue $\lambda_0 = 1$.

Let I denote the $m\times m$-identity matrix. Since the matrix $(K-(1-1/w_0)I)$ has rank one, the determinant $det(K-(1-1/w_0)I) = 0$. Hence $1-1/w_0$ is an eigenvalue for K by definition. It is shown below that $1-1/w_0$ has $m-1$ corresponding linear independent eigenvectors, hence 1 and $1-1/w_0$ are the only eigenvalues for K.

The eigenvectors of K corresponding to $1-1/w_0$ is defined by the equation

$$K\nu = \left(1-\frac{1}{w_0}\right)\nu \tag{1}$$

Let $\nu^{(j)}$ denote the jth element of the eigenvector ν. Using K as defined in Definition 4 equation (1) gives the m equations

$$\left(1-\frac{1}{w_o}\right)\nu^{(j)}+\sum_{i=1}^{m}\frac{\pi_i}{w_o}\nu^{(i)}=\left(1-\frac{1}{w_o}\right)\nu^{(j)}\ ;j=1,2,\ldots,m$$

Since first and last term is equal this gives that ν is an eigenvector to K with corresponding eigenvalue $(1-1/w_o)$ if

$$\langle\nu,\pi\rangle=0 \tag{2}$$

where $\langle\cdot,\cdot\rangle$ denotes the l^2 inner product. There are several ways to construct $m-1$ linear independent eigenvectors orthogonal to π each giving different sets of eigenvectors. The eigenvectors found by Liu (1996) for the MH/IP algorithm fulfills condition (2). To simplify comparisons with the results in Liu (1996) these eigenvectors are chosen in the proceeding calculations.

It is trivial to demonstrate that these eigenvectors are orthogonal to π and hence fulfills (2):

$$\langle\nu_k,\pi\rangle=\left(\sum_{l=k+1}^{m}\pi_l\right)\pi_k-\pi_k\left(\pi_{l+1}+\pi_{l+2}+\cdots+\pi_{m-1}+\pi_m\right)=0$$

Hence $\nu_k=(\underbrace{0,\ldots,0}_{i-1},\sum_{l=i+1}^{m}\pi_l,\underbrace{-\pi_i,\ldots,-\pi_i}_{m-i})$; $i=1,\ldots,m-1$ are eigenvectors to K with corresponding eigenvalues $\lambda_k=(1-1/w_0)$. □

Note that $w_0\geq w_1=\pi_1/p_1$, the largest likelihood value, and that λ_1 has a minimum for $w_0=w_1$. In this case all the non-unity eigenvalues of the McMC/RS algorithm are identical to the second largest eigenvalue of the MH/IP algorithm, see Theorem 1. Moreover, the eigenvectors for the McMC/RS and the MH/IP algorithm coincides. These relations can of course be used to compare the efficiencies of the two algorithms.

When evaluating the rate of convergence in the McMC/RS algorithm it is worth to note that it converges after the first accept. In the framework of Corollary 1 one gets.

Theorem 4 *The total variation distance, see Definition 3, between the distribution after n steps in the McMC/RS algorithm specified in Definition 4, and the limiting distribution π is:*

$$\|K^n(k,\cdot)-\pi(\cdot)\|=(1-\pi_k)\left(1-\frac{1}{w_0}\right)^n=(1-\pi_k)\lambda_1^n$$

Proof.

$$\begin{aligned}&\|K^n(k,\cdot)-\pi(\cdot)\|\\ &\qquad=\|K^n(k,\cdot|A_n=0)\cdot\Pr\{A_n=0\}+K^n(k,\cdot|A_n>0)\cdot\Pr\{A_n>0\}-\pi(\cdot)\|\end{aligned}$$

with A_n being the number of accepts up to n, hence $\Pr\{A_n = 0\} = (1 - 1/w_0)^n$ and $\Pr\{A_n > 0\} = 1 - (1 - 1/w_0)^n$. This follows trivially from Lemma 4. Let $\mathbf{1}_k = (\underbrace{0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{m-k})$ and recall that the Markov chain has converged to π after the first acceptance. Then:

$$\|K^n(k, \cdot) - \pi(\cdot)\| = \left\| [\mathbf{1}_k - \pi(\cdot)] \left(1 - \frac{1}{w_0}\right)^n \right\| = (1 - \pi_k) \left(1 - \frac{1}{w_0}\right)^n = (1 - \pi_k)\, \lambda_1^n$$

with $\lambda_1 = 1 - 1/w_0$ being the value of all the non-unity eigenvalues. □

This result can be used to predict the number of steps required to obtain a certain level of convergence in the McMC/RS algorithm. The fastest convergence rate is obtained for $w_0 = w_1$, and assume for the moment that $w_0 = w_1$ is known. It is also worth noting that the exact rate of convergence can be determined, while for the MH/IP algorithm only an upper bound $(2\sqrt{\pi_k})^{-1}(1 - 1/w_1)^n$ can be found, see Corollary 1. The two algorithms have the same order of convergence in the second largest eigenvalue, but the coefficients can be very different if π_k is small. Recall, however, that the McMC/RS and MH/IP algorithms cannot be directly compared since the former requires that w_0 is known and this is not required for the latter. Note also that the general upper bound for convergence of Markov chains given in Lemma 1 can be used to evaluate the McMC/RS algorithm. This provides the upper bound $(2\sqrt{\pi_k})^{-1}\lambda_1^n = (2\sqrt{\pi_k})^{-1}(1 - 1/w_1)^n$ which is identical to the upper bound for the MH/IP algorithm.

The real advantage of the McMC/RS algorithm with respect to convergence is not that it can be predicted exactly at the point of initiation. The advantage is primarily in the convergence diagnostics since when the first acceptance occur one is ensured that the algorithm has converged. For the MH/IP algorithm it is very difficult to define reliable convergence diagnostics.

After convergence, i.e. after the first acceptance, the samples from the McMC/RS algorithm will be according to π. Subsequent samples will be highly dependent however, since the samples between acceptance steps will be identical. Recall that the embedded process provides iid samples from π.

The following lemma will be useful in the proof of the following theorem:

Lemma 5 *Let T be binomially distributed according to $B(n, q)$, then*

$$E\left\{\frac{1}{T+1}\right\} = \frac{1}{(n+1)q}\left[1 - (1-q)^{n+1}\right] \le \frac{1}{(n+1)q}$$

Proof.

$$\begin{aligned} E\left\{\frac{1}{T+1}\right\} &= \sum_{k=0}^{n} \frac{1}{k+1} \frac{n!}{(n-k)!k!} q^k (1-q)^{n-k} \\ &= \frac{1}{(n+1)q} \sum_{k=1}^{n+1} \frac{(n+1)!}{(n-k+1)!k!} q^k (1-q)^{n-k+1} = \frac{1}{(n+1)q}\left[1 - (1-q)^{n+1}\right] \end{aligned}$$

□

After convergence of the algorithm, two alternative estimators for the expected value of an arbitrary function of X can be imagined: one based on the samples of the McMC/RS algorithm and the other based on samples from the corresponding embedded process. From this follows:

Theorem 5 *Let $X \sim \pi$ and define an arbitrary function $h(x)$ such that:*

$$\begin{aligned} E_\pi\{h(X)\} &= \mu < \infty \\ Var_\pi\{h(X)\} &= 1 \end{aligned}$$

Consider a sample sequence $X_1, \ldots, X_n$ from the stationary pdf using McMC/RS algorithm in Definition 4. Define:

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

then

$$E\{\hat{\mu}\} = \mu$$

$$Var\{\hat{\mu}\} = \frac{2w_0}{n} - \frac{1}{n}\left(1 + \frac{2w_0(w_0-1)}{n}\left(1 - \left(1 - \frac{1}{w_0}\right)^n\right)\right) \leq \frac{2w_0}{n}.$$

Consider the derived sequence $X_{(0)}, X_{(1)}, \ldots, X_{(T)}$ with $X_{(0)} = X_1$ and $X_{(1)}, \ldots, X_{(T)}$ being the samples of the embedded process of $X_1, \ldots, X_n$ containing a random number of samples termed T. Define

$$\hat{\hat{\mu}} = \frac{1}{T+1} \sum_{i=0}^{T} h(X_{(i)})$$

This corresponds to the traditional Rejection Sampling estimator, and:

$$E\left\{\hat{\hat{\mu}}\right\} = \mu$$

$$Var\left\{\hat{\hat{\mu}}\right\} = \frac{w_0}{n}\left[1 - \left(1 - \frac{1}{w_0}\right)^n\right] \leq \frac{w_0}{n}$$

Proof.
(Proof for estimator $\hat{\mu}$)
According to Liu (1996) there exists a general expression for Var$\{\frac{1}{n}\sum_{i=1}^{n} h(X_i)\}$ where $X_i\,; i = 1, \ldots, n$ are samples from an irreducible, aperiodic Markov chain. Let $\lambda_i\,; i = 1, \ldots, m-1$ be the eigenvalues different from unity. Define the spectral expansion of $h(\cdot)$ to be $h(x) = \mu + \sum_{i=1}^{m-1} a_i f_i(x)$ where $f_i(x) \propto \nu_i\,; i = 1, \ldots, m-1$ and ν_i is the eigenvector corresponding

to λ_i. Then $1 = \mathrm{Var}_\pi(h(X)) = \sum_{i=1}^{m-1} a_i^2$ and

$$
\begin{aligned}
\mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right\} &= \frac{1}{n^2}\left\{n\sum_{i=1}^{m-1} a_i^2 + 2(n-1)\sum_{i=1}^{m-1}\lambda_i a_i^2 + \cdots + 2\sum_{i=1}^{m-1}\lambda_i^{n-1}a_i^2\right\} \\
&= \frac{1}{n^2}\left\{2\sum_{i=0}^{n-1}(n-i)\lambda^i - n\right\} \qquad (3) \\
&= \frac{1}{n}\left\{\frac{2}{1-\lambda} - 1 - 2\frac{\lambda(1-\lambda^n)}{n(1-\lambda)^2}\right\} \\
&= \frac{2w_0}{n} - \frac{1}{n}\left(1 + \frac{2w_0(w_0-1)}{n}\left(1-\left(1-\frac{1}{w_0}\right)^n\right)\right) \\
&\leq \frac{2w_0}{n}
\end{aligned}
$$

(Proof for estimator $\hat{\hat{\mu}}$)
Consider the estimator $\hat{\hat{\mu}}$. From Lemma 3, the samples $X_{(0)}, \ldots, X_{(T)}$ are iid samples from π, hence:

$$
E\left\{\hat{\hat{\mu}}\right\} = E\left\{\frac{1}{T+1}\sum_{i=0}^{T} h(X_{(i)})\right\} = \mu
$$

and

$$
\begin{aligned}
\mathrm{Var}\left\{\hat{\hat{\mu}}\right\} &= \mathrm{Var}\left\{\frac{1}{T+1}\sum_{i=0}^{T} h(X_{(i)})\right\} \\
&= \sum_{t=0}^{\infty}\mathrm{Var}\left\{\frac{1}{T+1}\sum_{i=0}^{T} h(X_{(i)})|T=t\right\}\Pr\{T=t\} \\
&= \sum_{t=0}^{\infty}\frac{1}{t+1}\Pr\{T=t\} \\
&= E\left\{\frac{1}{T+1}\right\}
\end{aligned}
$$

with T being the number of acceptances in the McMC/RS algorithm in the samples $2, 3, \ldots, n$, with acceptance probability $\frac{1}{w_0}$. Hence T is binomial $B(n-1, \frac{1}{w_0})$ and Lemma 5 applies. This entails:

$$
\mathrm{Var}\left\{\hat{\hat{\mu}}\right\} = \frac{w_0}{n}\left[1-\left(1-\frac{1}{w_0}\right)^n\right] \leq \frac{w_0}{n} \qquad \square
$$

Note again that all variance expressions will be minimized if the parameter w_0 in the McMC/RS algorithm is set to its smallest eligible value, i.e. $w_0 = w_1 = \pi_1/p_1$, the largest likelihood value.

Recall also that the upper bound in Theorem 2 is valid for estimators based on samples from all reversible McMC algorithms. This provides the upper bound $\mathrm{Var}\{\hat{\mu}\} \leq \frac{2}{n(1-\lambda_1)} = \frac{2w_0}{n}$, which is obviously correct.

Note that this gives that the upper bound on variance in MH/IP as given in Theorem 2, is twice the variance of the embedded process of McMC/RS.

In order to compare the alternative estimators for μ, one has:

Theorem 6 *Let $X \sim \pi$ and define an arbitrary function $h(x)$ such that:*

$$\begin{aligned} E_\pi\{h(X)\} &= \mu < \infty \\ Var_\pi\{h(X)\} &= 1 \end{aligned}$$

Consider an estimator $\tilde{\mu}$ based on samples from the MH/IP algorithm as defined and evaluated in Theorem 2.

Consider an estimator $\hat{\mu}$ based on samples from the McMC/RS algorithm as defined and evaluated in Theorem 5.

Consider an estimator $\hat{\hat{\mu}}$ based on samples from the embedded process of the McMC/RS algorithm as defined and evaluated in Theorem 5.

The following relations holds:

$$Var\{\tilde{\mu}\} < Var\{\hat{\mu}\} \tag{4}$$

$$\begin{aligned} Var\{\hat{\hat{\mu}}\} &= Var(\hat{\mu}) - \frac{w_0 - 1}{n}\left(1 - \frac{2w_0}{n}\right) + \left(1 - \frac{1}{w_0}\right)^n \left(\frac{2w_0}{n}\frac{w_0 - 1}{n} + \frac{w_0}{n}\right) \\ &< Var\{\hat{\mu}\} \end{aligned} \tag{5}$$

Proof.
(Proof of relation (4))
Consider the general expression for the variance in equation (3) Since $\lambda_i \leq \lambda_1$; $i = 2, 3, \ldots, m-1$ for MH/IP Markov chain the variance is always smaller or equal to the variance for the McMC/RS algorithm where $\lambda_i = \lambda_1$ $i = 2, 3, \ldots, m-1$.

(Proof of relation (5))
Relation (5) may be proved by demonstrating that

$$\mathrm{Var}(\hat{\mu}) - \mathrm{Var}(\hat{\hat{\mu}}) = \frac{w_0 - 1}{n}\left(1 - \frac{2w_0}{n}\right) + \left(1 - \frac{1}{w_0}\right)^n \left(\frac{2w_0}{n}\frac{w_0 - 1}{n} + \frac{w_0}{n}\right)$$

is larger or equal to zero for all pairs w_0, n. Recall that $w_0 \geq 1$ by definition. The expression is obviously zero or positive for $w_0 < n/2$ since this makes all terms positive or zero. It is more difficult to prove this analytically, but numerical calculations for $w_o \geq n/2$ indicates that this is true. The difference converges monotonically towards zero as $w_0 \to \infty$.

The reason why McMC/RS always has larger estimation variance than the corresponding embedded process can be explained as follows: The variance for the estimator using McMC/RS, $\hat{\mu}$, can be found by

$$\begin{aligned} \mathrm{Var}\,\{\hat{\mu}\} &= \mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right\} \\ &= \mathrm{Var}\left\{\frac{N_{(0)}}{n}h(X_{(0)}) + \sum_{i=1}^{T-1}\frac{N_{(i)}}{n}h(X_{(i)}) + \frac{N_{(T)}}{n}h(X_{(T)})\right\} = \mathrm{Var}\left\{\sum_{i=0}^{T} U_i h(X_{(i)})\right\} \end{aligned}$$

with $N_{(0)}$ being the number of steps up to the first acceptance; $N_{(i)}$; $i = 1, \ldots, T-1$ being the number of steps between acceptance i and $i+1$; and $N_{(T)}$ being the number of steps after acceptance T and up to n. Note that $\sum_{i=0}^{T} N_{(i)} = n$ of course, and hence $U_i = N_{(i)}/n$; $i = 0, \ldots, T$ appear as normalized, stochastic weights.

Since all $X_{(i)}$ are independent samples from π the estimation variance will increase if stochastic weights U_i ; $\sum_{i=1}^{T} U_i = 1$ are used instead of the constant weight $1/T$ for all samples. $\square$

From this, both the estimator based on the MH/IP algorithm and the estimator based on the traditional Rejection Sampling has lower or equal variance than the estimator based on the McMC/RS algorithm. The latter defines an upper bound for the two other algorithms.

It is not trivial to compare the estimators based on the MH/IP algorithm and the traditional Rejection Sampling algorithm. It is worth noting that both convergence rate and estimation variance can be determined exactly for the latter, and that it only depends on the value of the algorithm parameter w_0. Moreover, the very favorable convergence diagnostics of the Rejection Sampling algorithm, stating that after the first acceptance the algorithm has converged, make the Rejection Sampling very attractive.

The fact that the MH/IP algorithm and the McMC/RS algorithm, being the base for traditional Rejection Sampling, are not directly comparable has been emphasized many times. The latter requires the algorithm parameter w_0, most favorably chosen to be w_1, to be known. If thorough evaluation of McMC algorithms shall take place, for example a definition of the upper bound for the convergence rate, then the second largest eigenvalue λ_1, or at least an upper bound, has to be specified. This $\lambda_1 = 1 - 1/w_1$ defines uniquely the largest likelihood value which is the crucial parameter in the McMC/RS algorithm. Since serious work require the second largest eigenvalue or an upper bound to be known, the largest likelihood value or a corresponding upper bound is known, and the McMC/RS algorithm can just as well be used. The very favorable characteristics of the embedded process and the convergence diagnostics make this a very attractive choice.

In the case where neither λ_1 nor a value for w_0 can be determined, an algorithm providing an alternative to the MH/IP algorithm is presented in the following section.

4 Reducible Markov chain Monte Carlo/Dynamic Rejection Sampling (RMcMC/DRS) algorithm

A Dynamic Rejection Sampling algorithm defined through a reducible Markov chain is introduced. It is dynamic in the sense that it adapts to an optimal value for the algorithm parameter w_0 as the algorithm proceeds. Hence the value for w_1 need not be known at the time of initiation, and this makes the algorithm comparable to the MH/IP algorithm.

The steps of the RMcMC/DRS algorithm is defined in Section 1. A more formal definition is:

Definition 5 *The RMcMC/DRS algorithm with posterior distribution π and prior distribution p over the sample space $\Omega = \{1, \ldots, m\}$ is defined by the reducible Markov chain $\{X_n, W_n\}$ with $X_n \in \{1, \ldots, m\}, W_n \in \{w_1, \ldots, w_m\}$ and $n \in \{0, 1, 2, \cdots\}$, with*

Initial state:

$$(X_0, W_0) = (k, m_k) \quad with \quad k \in \Omega$$

Transition matrix:
The transition matrix K can be decomposed as $K = K_{X|W} K_W$ where K_W is a transition matrix for the process $\{W_n\}$, and is defined as

$$K_W(w_i, w_j) = \begin{cases} 0 & w_i < w_j \\ \sum_{l=i}^{m} p_l & w_i = w_j \\ p_j & w_i > w_j \end{cases}$$

with $w_i, w_j = w_1, \ldots, w_m$, and where $K_{X|W}$ is a transition matrix for the conditional process $\{X_n | W_n = w_s\}$, and defined as

$$K_{X|W}(i, j | w_s) = \begin{cases} p_j \frac{w_j}{w_s} & i \neq j \; ; s \leq i, j \leq m \\ 1 - \sum_{l \neq i} p_l \frac{w_l}{w_s} & i = j \; ; s \leq i, j \leq m \\ 0 & i, j < s \leq m \end{cases}$$

with $w_i = \pi_i / p_i \; ; i = 1, \ldots, m$

Note that the Markov chain $\{W_n\}$ is defined by $W_n = \max\{w_{x_0}, \ldots, w_{x_n}\} = \max\{W_{n-1}, w_{x_n}\}$ and that all states $w_i \; ; i = 2, \ldots, m$ are transient and state w_1 is absorbing. The Markov chain $\{X_n | W_n = w_s\}$ is of the McMC/RS type defined in Definition 4. Note also that this is not a Markov chain Monte Carlo algorithm according to the definition in Hastings (1970), since the chain is not irreducible. Limiting properties of the reducible Markov chain $\{X_n, W_n\}$ is:

Lemma 6 *The reducible Markov chain $\{X_n, W_n\}$ has the absorbing Markov chain $\{X_n | W_n = w_1\}$ which corresponds to the McMC/RS algorithm in Definition 4 with algorithm parameter w_0 set to its optimal value w_1.*

Proof. Note that

$$\Pr\{X_n, W_n = w_s\} = \Pr\{X_n | W_n = w_s\} \Pr\{W_n = w_s\} \; , s = 1, \ldots, m$$

and that

$$\lim_{n \to \infty} \Pr\{W_n = w_s\} = \begin{cases} 1 & s = 1 \\ 0 & s = 2, 3, \ldots, m \end{cases}$$

hence

$$\Pr\{X_n, W_n\} \xrightarrow{n \to \infty} \Pr\{X_n | W_n = w_1\} \qquad \square$$

Hence the samples from the RMcMC/DRS algorithm after convergence will be according to π, and the samples from the corresponding embedded process will be iid from π, see Lemma 3.

The reducible Markov chain representing the RMcMC/DRS algorithm, see Definition 5, is complicated to evaluate through its eigenvalues and eigenvectors. An upper bound for the total variation distance can be determined along the lines of Theorem 4, however:

Theorem 7 *The total variation distance, see Definition 3 between the distribution after n steps in the RMcMC/DRS algorithm specified in Definition 5, and the limiting distribution π is:*

$$\|K_X^n(k,\cdot)-\pi(\cdot)\| \le (1-p_1)^n + \pi_1\left[\left(\frac{1-p_1}{1-1/w_1}\right)^n - 1\right]\left(1-\frac{1}{w_1}\right)^n$$

with $K_X(\cdot,\cdot)$ being the transition matrix for the marginal process $\{X_n\}$.

Proof.

$$\|K_X^n(k,\cdot)-\pi(\cdot)\| = \left\|\sum_{l=1}^{m} K_X^n(k,\cdot|B_l^n)\Pr\{B_l^n\} - \pi(\cdot)\right\|$$

with B_l^n being the event that $\max\{w_{x_1},\ldots,w_{x_n}\} = w_l$. Note that $\Pr\{\bigcup_{l=2}^{m} B_l^n\} = (1-p_1)^n$ and $\Pr\{B_1^n\} = 1-(1-p_1)^n$ for $k \ne 1$. Then

$$\begin{aligned}\|K_X^n(k,\cdot)-\pi(\cdot)\| \le{}& \max_{l,m}\{\|K_X^n(k,\cdot|B_l^n)-\pi(\cdot)\|\}\Pr\left\{\bigcup_{l=2}^{m} B_l^n\right\} \\ &+ \sum_{v=1}^{n}\|K^{n-v}(1,\cdot|W=w_1)-\pi(\cdot)\|\Pr\left\{B_1^v \bigcap \left[\bigcup_{l=2}^{m} B_l^{v-1}\right]\right\}\end{aligned}$$

with the first term being the worst case whenever absorption of $\{W_n\}$ in w_1 has not occured before step n. The second term is the total variation for the McMC/RS algorithm in Theorem 4 whenever absorption of $\{W_n\}$ in w_1 occured at step v with initial state 1.

Then:

$$\begin{aligned}&\|K_X^n(k,\cdot)-\pi(\cdot)\| \\ &\quad\le \left(1-\min_{j\le k}\{\pi_j\}\right)(1-p_1)^n + \sum_{v=1}^{n}(1-\pi_1)\left(1-\frac{1}{w_1}\right)^{n-v}(1-p_1)^{v-1}p_1 \\ &\quad\le (1-p_1)^n + \pi_1\left[(1-p_1)^n - \left(1-\frac{1}{w_1}\right)^n\right] \\ &\quad= (1-p_1)^n(1+\pi_1) - \pi_1\left(1-\frac{1}{w_1}\right)^n \qquad \square\end{aligned}$$

The rate of convergence of the MH/IP algorithm is bounded by $(2\sqrt{\pi_k})^{-1}(1-1/w_1)^n$, see Corollary 1. This bound is smaller than for the RMcMC/DRS algorithm.

Note that the upper bound in Theorem 7 is expressed by the characteristics of the state with largest likelihood value only, i.e. state 1. In the convergence rate, the term $(1-p_1)^n$ is dominant. This term appear from the convergence rate of $\{W_n\}$ into the absorbing state w_1. Hence if faster convergence of $\{W_n\}$ can be obtained it will improve the algorithm considerably.

Convergence rate of $\{W_n\}$ may be improved by searching for the global maximum in the likelihood function w. This could be done before the Markov chain is initiated. If there are several maxima the search could be repeated at each nth step using the present state as starting state for the optimization procedure. One could also imagined taking some few optimization steps each time a proposed state in the Markov chain is accepted.

The reducible Markov chain representative for the RMcMC/DRS algorithm in Definition 5 converges towards the irreducible, aperiodic Markov chain representative for the McMC/RS algorithm in Definition 4, as $n \to \infty$. After convergence these two chains will be identical.

Hence the estimators for the expected value of an arbitrary function of X will coincide, and the results in Theorem 5 will be valid for the RMcMC/DRS algorithm also. The same applies for Theorem 6 making the comparison with the MH/IP algorithm.

It should be noted that the construction of the RMcMC/DRS algorithm is such that the sequence of likelihood values $w_n; n = 0, 1, 2, \ldots$ computed during the burn in period of the algorithm can be reused in the McMC/RS algorithm with algorithm parameter w_0 being the converging value of $\{W_n\}$. In cases where the computation of the likelihood is resource requiring this can be important. Let $Y_1, \ldots, Y_s$ be all the proposed states using the RMcMC/DRS algorithm. Assume that convergence is diagnosed after s proposals and set $w_0 = w_s$. Now the RMcMC/DRS has converged to McMC/RS. Let the same states $Y_1, \ldots, Y_s$ be the proposals also in the McMC/RS algorithm and accept each proposal with probability w_i/w_0 ; $i = 1, \ldots s$ according to Definition 4. Note that all the likelihoods w_i ; $i = 1, \ldots, s$ are calculated previously in the RMcMC/DRS algorithm. This defines a sequence from the stationary distribution using McMC/RS.

5 Closing remarks

A comparison of Metropolis-Hastings algorithm with independent proposal (MH/IP) and Rejection Sampling is performed by generalizing the formulation of Rejection Sampling to a Reducible Markov chain Monte Carlo algorithm (RMcMC/DRS). This new formulation defines a Markov chain whose embedded process converges to Rejection Sampling and it requires the same prior knowledge of the distributions as the MH/IP algorithm.

It is demonstrated that a smaller upper bound for estimation variance is available for Rejection Sampling than for MH/IP. The upper bound for the convergence rate is smaller for MH/IP than for RMcMC/DRS. However, if a theoretical study of convergence properties in a specific MH/IP algorithm is performed, the second largest eigenvalue λ_1 or an upper bound must be found. This entails that the largest likelihood value or a corresponding upper bound is uniquely defined and that Rejection Sampling with smaller upper bounds for estimation variance, could be applied instead.

If convergence is monitored by graphical inspection RMcMC/DRS gives a monotone sequence $\{W_n\}$ converging to the unknown constant w_1. It may be easier to diagnose convergence using this than the usual McMC graphical convergence diagnostics. The fact that Rejection Sampling provides iid samples from π could make the RMcMC/DRS algorithm preferable if iid samples are desirable.

Acknowledgment

We are grateful to Arnoldo Frigessi for inspiring discussions. This work is founded by the Norwegian Research Council.

References

Billingsley, P. (1986). *Probability and Measure* (Second ed.). John Wiley & Sons.

Diaconis, P. and P. Hanlon (1992). Eigen Analysis for Some Examples of the Metropolis Algorithm. *Contemporary Mathematics 138*, pp. 99–117.

Hastings, W. K. (1970). Monte Carlo sampling using Markov chains and their applications. *Biometrica 57*(1), pp. 97–109.

Karr, A. F. (1993). *Probability*. Springer Texts is Statistics. Springer-Verlag.

Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag.

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing 6*, pp. 113–119.

Ripley, B. D. (1987). *Stochastic Simulation*. Probability and Mathematical Statistics. John Wiley & Sons.