Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 51 (1) (2022), 239 – 252
DOI : 10.15672/hujms.899524

Research Article

Approximate verification of geometric ergodicity for multiple-step Metropolis transition kernels

David A. Spade

Department of Mathematical Sciences, University of WisconsinMilwaukee, 3200 North Cramer Street E403, Milwaukee, WI, USA

Abstract

In many applications involving discrete time Markov chains, the autocorrelation between states corresponding to nearby time points is too high to use all of these states as part of an approximate random sample from a specified target distribution. In these situations, it is common to use the output of a thinned chain, where we take samples every h steps, and h is a positive integer, in order to reduce autocorrelation. In order to justify using central limit theorems in analyses based on the output of a thinned chain, it is necessary to show that the thinned chain is geometrically ergodic. A common way to do this is to show that the chain satisfies a minorization condition and an associated drift condition. In this manuscript, we extend previous results pertaining to one-step transition kernels to handle numerical estimation of minorization and drift coefficients for h-step transition kernels for Metropolis algorithms.

Mathematics Subject Classification (2020). 50J22, 60J05, 62P99

Keywords. Markov chain Monte Carlo, geometric ergodicity, drift, minorization

1. Introduction

In many applications of statistics, Bayesian inference requires sampling from intractable probability distributions. When this problem arises, it is common to use Markov chain Monte Carlo (MCMC) methods to obtain samples from these distributions. When MCMC methods are used, it is important to have an understanding of the behavior of the underlying Markov chain. For example, if a chain is geometrically ergodic, then central limit theorems are available for inference that is carried out based on the output of the chain [18]. A problem that is often encountered is that geometric ergodicity can be a difficult property to verify analytically. A common way to show that a chain is geometrically ergodic is to demonstrate that its transition kernel satisfies a minorization condition and an associated drift condition. This problem, however, is analytically intractable in most practical settings, so often, the best that can be done is to rely on output-based convergence diagnostics, such as those presented by [6–8, 23, 24] to make a determination as to whether or not the chain has approximately reached its target distribution. These diagnostics each suffer from their own significant limitations, so Cowles and Rosenthal [3] constructed an approach based on auxiliary simulations that numerically estimates drift and minorization

Email address: spade@uwm.edu
Received: 18.03.2021; Accepted: 17.09.2021

coefficients. While this method is useful in fairly general settings, it requires dividing the state space into tiny bins and counting how many chains land in each of these bins. This process becomes prohibitively expensive even in moderate-dimensional settings, so a clear need exists for methodology that can be carried out more efficiently.

For certain MCMC algorithms, the need for more efficient methods of estimating these coefficients has been addressed. For example, Spade [20] presents an efficient method based on numerical integration for estimating minorization and drift coefficients for a version of a Metropolis algorithm, called the random-scan Metropolis (RSM) algorithm, that chooses one variable at random at each step and updates the selected variable using a Metropolis-style accept/reject decision. This method leverages a result from [4]. The clear limitation of this method is its lack of generalizability. The method is only applicable to RSM samplers and does not extend naturally to full-updating schemes. In order to address this limitation, Spade [21] exploits a result presented by [10] to construct a Monte Carlo integration-based method for approximately verifying geometric ergodicity of a full-updating random-walk Metropolis sampler. The methods described therein also use a result given by [19] to bound the mixing time of these samplers using the estimated drift and minorization coefficients. The key limitation of this method is the requirement of a symmetric density from which an increment to the current state is proposed. In other words, the method does not work well for more general Metropolis-Hastings algorithms. Spade [22] adapted the approach presented in [21] to approximately verify geometric ergodicity for Metropolis-Hastings algorithms with asymmetric proposal densities, thus providing a class of techniques for estimating drift and minorization coefficients that works by exploiting the accept/reject nature of the Metropolis-Hastings algorithm.

In many settings, a near-random sample cannot be obtained simply by taking a sequence of states from the chain. For example, if a sample of size n is desired, it is not enough simply to use the first n post-burn-in states. This is because there is often high autocorrelation between states that are observed at nearby time points. Consequently, it is common to "thin" the output of the chain. In other words, every, say h^{th} post-burn-in state is selected. If we want central limit theorems to be available for the thinned chain, then it is necessary to verify geometric ergodicity for the h-step transition kernel. This is at least as difficult to do analytically as it is for the one-step transition kernel, and aside from the Cowles and Rosenthal [3] approach, the techniques described above are ill-suited to handle h-step transition kernels. The primary goal of this manuscript is to introduce a method of approximately verifying geometric ergodicity for RWM samplers that are thinned by a factor of h efficiently and without a need to rely on the output-based convergence diagnostics. The extension of this methodology to more general h-step Metropolis-Hastings can be justified mathematically, but is technically much more difficult to do. Therefore, this will not be taken up here.

For the h-step RWM sampler, a computational technique for estimating minorization and drift coefficients is presented. While the point of the manuscript is to approximately verify geometric ergodicity of h-step RWM samplers, we will also use the estimates of the minorization and drift coefficients to give a conservative upper bound on the mixing times for each of the chains examined later in the manuscript using the Rosenthal [19] formula.

The remainder of this manuscript is organized as follows. Section 2 presents the background on Markov chains that is necessary for an understanding of the work presented in later sections. Section 3 details the proposed method of estimating minorization and drift coefficients, and Section 4 illustrates the use of this method and examines its performance in four examples. The manuscript concludes with a discussion of the implications and the limitations of the work presented herein, as well as a discussion of some open questions that will remain to be investigated.

2. Preliminaries

This section gives the background on Markov chains that is necessary for an understanding of the method presented in Section 3. Section 2.1 provides some background on the theory of general state space Markov chains, and Section 2.2 provides a full description of the RWM sampler.

2.1. Background on Markov chain convergence

Let $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m), \mathrm{P})$ be a probability space, where $\mathcal{B}(\mathbb{R}^m)$ is the σ-field consisting of Borel subsets of $\mathbb{R}^m$ and P is a probability measure. Let $(\mathsf{X}_t)_{t\geq 0}$ be an ergodic Markov chain on $\mathbb{R}^m$ with transition kernel $K : (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m)) \mapsto [0,1]$ and stationary measure π having probability density $p(\cdot)$. If h is a positive integer, then the *h-step transition kernel* $K^h(\cdot,\cdot)$ is such that for all $\mathbf{x} \in \mathbb{R}^m$, $K^h(\mathbf{x},\cdot)$ is a probability measure on $\mathcal{B}(\mathbb{R}^m)$, for all $A \in \mathcal{B}(\mathbb{R}^m)$, $K^h(\cdot, A)$ is a measurable function on $\mathbb{R}^m$, and

$$K^h(\mathbf{x}, A) = \mathrm{P}(\mathbf{X}_{t+h} \in A | \mathsf{X}_t = \mathbf{x}). \tag{2.1}$$

In order to examine geometric ergodicity, we need the total variation distance.

Definition 2.1. The *total variation distance* between $K^t(\cdot,\cdot)$ and $\pi(\cdot)$ is given by

$$\delta(K^t, \pi) = \sup_{\mathbf{x}\in\mathbb{R}^m} \sup_{A\in\mathcal{B}(\mathbb{R}^m)} \|K^t(\mathbf{x}, A) - \pi(A)\|.$$

We say that at time t, the chain $(\mathsf{X}_t)_{t\geq 0}$ has achieved ε-mixing if $\delta(K^t, \pi) \leq \varepsilon$.

Definition 2.2. The *mixing time* of $(\mathsf{X}_t)_{t\geq 0}$ for a given threshold ε is given by

$$\tau_{\mathrm{mix}}(\varepsilon) = \min\left\{t > 0 : \delta(K^t, \pi) \leq \varepsilon\right\}.$$

At this point, we are ready to provide a definition of geometric ergodicity.

Definition 2.3. The Markov chain $(\mathsf{X}_t)_{t\geq 0}$ is said to be *geometrically ergodic* if for all $t > 0$, and for some constants $C(\mathbf{x})$ and $\rho < 1$,

$$\delta(K^t, \pi) \leq C(\mathbf{x})\rho^t.$$

The property of geometric ergodicity is difficult to verify in practice, but it is well-known that, if $(\mathsf{X}_t)_{t\geq 0}$ satisfies a minorization condition and an associated drift condition, then $(\mathsf{X}_t)_{t\geq 0}$ is geometrically ergodic. Before defining these terms, we need the notion of a small set.

Definition 2.4. A set $C \in \mathcal{B}(\mathbb{R}^m)$ is called a *small set* if there exists an integer $m > 0$ and a non-trivial measure ν_m on $\mathcal{B}(\mathbb{R}^m)$ such that for all $\mathbf{x} \in C$ and for all $A \in \mathcal{B}(\mathbb{R}^m)$,

$$K^m(\mathbf{x}, A) \geq \nu_m(A).$$

Definition 2.5. A Markov chain $(\mathsf{X}_t)_{t\geq 0}$ satisfies a *minorization condition* if there exist $\varepsilon \in (0,1)$ a small set $C \in \mathcal{B}(\mathbb{R}^m)$, a positive integer k, and a probability measure $\nu(\cdot)$ such that for all $\mathbf{x} \in C$ and for all $A \in \mathcal{B}(\mathbb{R}^m)$,

$$K^k(\mathbf{x}, A) \geq \varepsilon\nu(A). \tag{2.2}$$

Definition 2.6. A Markov chain $(\mathsf{X}_t)_{t\geq 0}$ satisfies a *drift condition* if there exist constants $\lambda \in (0,1)$ and $b < \infty$, a function $V : \mathbb{R}^m \mapsto [1,\infty)$, a positive integer h, and a small set $C \in \mathcal{B}(\mathbb{R}^m)$ such that for all $\mathbf{x} \in \mathbb{R}^m$,

$$K^h V(\mathbf{x}) \leq \lambda V(\mathbf{x}) + b\mathbb{I}_C(\mathbf{x}), \tag{2.3}$$

where $K^h V(\mathbf{x}) = \mathbb{E}[V(\mathsf{X}_{t+h})|\mathbf{X}_t = \mathbf{x}]$ and the expectation is taken with respect to the h-step transition kernel.

If a Markov chain is geometrically ergodic, then central limit theorems are available for samples taken from the states of the chain [18]. Rosenthal [19] uses drift and minorization coefficients to obtain a conservative upper bound on the total variation distance between the n-step transition kernel for a geometrically ergodic Markov chain $(\mathsf{X}_t)_{t\geq 0}$ and its stationary measure.

Theorem 2.7. *[19] Assume that for a function $V : \mathbb{R}^m \mapsto [1,\infty)$, a positive integer h, and constants $\lambda \in (0,1)$ and $b < \infty$, $(\mathsf{X}_t)_{t\geq 0}$ satisfies*

$$K^h V(\mathbf{x}) \leq \lambda V(\mathbf{x}) + b\mathbb{I}_C(\mathbf{x})$$

for all $\mathbf{x} \in \mathbb{R}^m$, where $C = \{\mathbf{x} : V(\mathbf{x}) \leq d\}$ and $d > \frac{2b}{1-\lambda} - 1$. Assume that for some $\varepsilon > 0$, some probability measure $\nu(\cdot)$ on $\mathcal{B}(\mathbb{R}^m)$, and some positive integer k_0,

$$K^{hk_0}(\mathbf{x}, B) \geq \varepsilon\nu(B)$$

for all $\mathbf{x} \in C$ and for all $B \in \mathcal{B}(\mathbb{R}^m)$. Then for any $r \in (0,1)$ with $(\mathsf{X}_t)_{t\geq 0}$ beginning in the initial distribution $\mathbf{\Psi}$ and for any positive integer n,

$$\begin{aligned} \delta(K^n, \pi) &\leq (1-\varepsilon)^{\left[\frac{rn}{hk_0}\right]} + (\alpha A)^{-1}\left(\alpha^{-(1-rk_0)}A^r\right)^{\left[\frac{h}{n}\right]} \\ &\times \left(1 + \frac{b}{1-\lambda} + \mathbb{E}_{\mathbf{\Psi}}[V(\mathsf{X}_0)]\right), \text{ where} \qquad (2.4) \\ \alpha^{-1} &= \frac{1+2b+\lambda d}{1+d}, \\ A &= 1 + 2(\lambda d + b), \text{ and} \qquad (2.5) \end{aligned}$$

$[\cdot]$ denotes the greatest integer function.

2.2. The random-walk Metropolis algorithm

In this section, we describe the RWM algorithm. Let $\mathbf{x}_0$ denote the initial state of the Markov chain $(\mathsf{X}_t)_{t\geq 0}$ and assume that $\mathbf{x}_0$ falls within the support of the target density $p(\cdot)$. Given the state $\mathbf{x}_t$ at time t, X_{t+1} is obtained by first choosing an increment vector $\mathbf{y}$ from a density $q(\cdot)$ that is symmetric about $\mathbf{0}$. A proposed update $\mathbf{x}^*$ is given by $\mathbf{x}_t + \mathbf{y}$. Then $\mathsf{X}_{t+1} = \mathbf{x}^*$ with probability

$$\alpha(\mathbf{x}_t, \mathbf{x}*) = \min\left\{1, \frac{p(\mathbf{x}^*)}{p(\mathbf{x}_t)}\right\},$$

and $\mathsf{X}_{t+1} = \mathbf{x}_t$ with probability $1 - \alpha(\mathbf{x}_t, \mathbf{x}^*)$. This results in the following transition density $k(\cdot|\cdot)$ for the RWM sampler:

$$\begin{aligned} k(\mathbf{x}_{t+1}|\mathbf{x}_t) &= \alpha(\mathbf{x}_t, \mathbf{x}_t + \mathbf{y}) \\ &+ \left(\int_{\mathbb{R}^m} [1 - \alpha(\mathbf{x}_t, \mathbf{x}_t + \mathbf{y})]\, q(\mathbf{y})\, d\mathbf{y}\right) \delta_{\mathbf{x}_t}(\mathsf{X}_{t+1}), \end{aligned}$$

where $\delta_{\mathbf{x}}(\cdot)$ is the Dirac mass measure concentrated at $\mathbf{x}$. Jarner and Hansen [10] present a set of conditions that are sufficient to ensure that the target density is positive and continuous over $\mathbb{R}^m$ or some open, unbounded subset of $\mathbb{R}^m$, that the transition density

is bounded away from 0 on compact sets, and that the tails of the target density decrease quickly enough to ensure geometric ergodicity of the chain. Since we do not verify these conditions analytically in our examples, we do not state them here. However, Jarner and Hansen [10] also provide a drift function for RWM samplers that satisfy their conditions, and it will be seen in Section 4 that this drift function is suitable for the RWM samplers in each of our settings. This drift function is $V_s(\mathbf{x}) = c[p(\mathbf{x})]^{-s}$ for some $c > 0$ and for any $s \in (0, 1)$. We will use several different choices of s in the illustrative examples of Section 4.

3. Estimating drift and minorization coefficients

In this section, the technique for estimating drift and minorization coefficients is presented. In Section 3.1, we detail the estimation of the drift coefficients, and in Section 3.2, the estimation of the minorization coefficient is described.

3.1. Estimation of drift coefficients

In order to estimate drift coefficients, we first select a drift function $V(\mathbf{x})$. The process begins with the estimation of λ. Let d be some positive constant that is larger than $\frac{2b}{1-\lambda} - 1$. Since estimates of λ and b have not yet been obtained, it is best to choose a comfortably large value of d. Once estimates $\hat{\lambda}$ and $\hat{b}$ of λ and b have been found, if $\frac{2\hat{b}}{1-\hat{\lambda}} > d$, choose a larger value of d and try again. Once d is selected, choose the set $C = \{\mathbf{x} : V(\mathbf{x}) \leq d\}$ as a small set. Rosenthal [19] shows that this set is indeed small for transition kernels that satisfy a drift condition. To begin estimating λ, choose some number $N_{C'}$ of points outside of C to be initial values. The choice of $N_{C'}$ may require some experimentation, but this value should be chosen in such a way that the resulting estimate of λ is fairly stable. Once the initial states are chosen, for a given integer $h \geq 1$ that is specified by the researcher but that corresponds to the factor by which the output of the chain is to be thinned, run N_0 h-step chains from each of the initial states inside the sampled set $\hat{C}'$. Since Equation (2.3) implies that for $\mathbf{x} \notin C$, $\lambda \leq \frac{\mathbb{E}[V(\mathsf{X}_{t+h})|\mathsf{X}_t = \mathbf{x}]}{V(\mathbf{x})}$, compute the average value

$$\hat{\lambda}_{\mathbf{x}} = \frac{1}{N_0}\left(\frac{1}{V(\mathbf{x})}\sum_{i=1}^{N_0} V(\mathbf{x}_{t+h}^{(i)})\right),$$

where $\mathbf{x}_{t+h}^{(i)}$ is the ending state of the i^{th} h-step chain initialized at $\mathbf{x}$. Then a fairly conservative estimate of λ is obtained by taking

$$\hat{\lambda} = \max_{\mathbf{x} \in \hat{C}'} \hat{\lambda}_{\mathbf{x}}.$$

If $\hat{\lambda} < 1$, then this is evidence that we have a suitable drift function. Otherwise, we may need to try again with a different drift function.

Now that an estimate of λ is available, our attention turns to estimating b. To begin, choose some number N_{C_b} of points inside C from which to initialize an h-step chain. Much like $N_{C'}$, the choice of N_{C_b} may require some experimentation in order to ensure that the estimate of b is fairly stable. For a given value of $\mathbf{x}$ in the sampled set $\hat{C}_b$, run a reasonably large number, say N_1 of h-step chains. At this point, we take a brief aside to point out that, until this point, estimating b has been done in a similar fashion to the estimation of λ. However, in estimating b, an issue that can arise that we do not encounter in estimating λ but that is often a major one in estimating b for RWM samplers is dependence of the drift function on the normalizing constant of the target density. This constant cancels out in the ratio during the estimation of λ, but it does not in the estimation of b. If the drift function depends on the target density, then it will depend on the marginal density of $\mathbf{x}$ in most cases. Therefore, we need to estimate the marginal density $m(\mathbf{x})$. The simplest

method for doing this is as follows. Let $\mathcal{L}(\mathbf{x})$ denote the likelihood function based on $\mathbf{x}$. Sample some number, say M, of points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ from the prior density. Then an estimate $\hat{m}(\mathbf{x})$ of $m(\mathbf{x})$ can be obtained by taking

$$\hat{m}(\mathbf{x}) = \frac{1}{M} \sum_{j=1}^{M} \mathcal{L}(\mathbf{x}).$$

In some cases, this approach may not be feasible. In these situations, one may try one of the numerous other methods of marginal density estimation that exist in the literature, including adaptive importance sampling [15], annealed importance sampling [14], or one of several other techniques that are available [2, 9, 12].

In the remainder of this discussion of estimating b, assume that marginal density estimation is necessary. Once an estimate of the marginal density has been obtained, let $\hat{V}(\mathbf{x})$ denote the estimated drift function based on $\hat{m}(\mathbf{x})$. As the N_1 h-step chains are run from each initial state, the process proceeds in the following way. Given an initial state $\mathbf{x}_t$, for the resulting value of $\mathbf{x}_{t+h}^{(i)}$ resulting from chain i, compute $\hat{V}(\mathbf{x}_{t+h}^{(i)})$. An intermediate estimate $\hat{b}_{\mathbf{x}_t}$ given that $\mathbf{x}_t$ is the initial state is obtained partly by taking the average value of $\hat{V}(\mathbf{x}_{t+h}^{(i)})$ over the N_1 chains. However, recall that for chains that satisfy a drift condition and for $\mathbf{x}_t \in C$,

$$\mathbb{E}[V(\mathsf{X}_{t+h}) | \mathsf{X}_t = \mathbf{x}_t] \leq \lambda V(\mathbf{x}_t) + b.$$

Thus, we complete the computation of $\hat{b}_{\mathbf{x}_t}$ by taking

$$\hat{b}_{\mathbf{x}_t} = \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{V}(\mathbf{x}_{t+h}^{(i)}) - \hat{\lambda} \hat{V}(\mathbf{x}_t).$$

A final estimate $\hat{b}$ of b is obtained by taking

$$\hat{b} = \max_{t=1,2,\ldots,N_{C_b}} \hat{b}_{\mathbf{x}_t}.$$

3.2. Estimating the minorization coefficient

The estimation of the minorization coefficient for h-step transition kernels is much more complicated mathematically than it is for one-step transition densities. Cowles and Rosenthal [3] gives the following expression for ε:

$$\varepsilon = \int_{\mathbb{R}^m} \left(\inf_{\mathbf{x}_t \in C} k(\mathbf{x}_{t+h} | \mathbf{x}_t) \right) \, \mathrm{d}\mathbf{x}_{t+h}. \tag{3.1}$$

To begin, note that for an RWM sampler with increment density $q(\mathbf{y})$,

$$\begin{aligned} k(\mathbf{x}_{t+h} | \mathbf{x}_t) &= \int_{\mathbb{R}^m} \cdots \int_{\mathbb{R}^m} \prod_{i=1}^{n-1} k(\mathbf{x}_{t+i} | \mathbf{x}_{t+i-1}) \mathrm{d}\mathbf{x}_{t+h-1} \ldots \mathrm{d}\mathbf{x}_{t+1} \\ &= \int_{\mathbb{R}^m} \cdots \int_{\mathbb{R}^m} \prod_{i=1}^{n-1} k(\mathbf{x}_{t+i} | \mathbf{x}_{t+i-1}) \mathrm{d}\mathbf{y}_{h-1} \ldots \mathrm{d}\mathbf{y}_1, \end{aligned} \tag{3.2}$$

where $\mathbf{y}_i$ denotes the increment proposed at the i^{th} step of the h-step chain. Our approach is designed to estimate the integral in Equation (3.1). In order to do this, we choose some number N_C of initial states in C. We also choose some number N_2 of possible increments from $q(\mathbf{y})$. We observe here that since the RWM update depends on the increment in

such a way that the one-step transition density may be viewed as a function of $\mathbf{y}$ given the current state, we can write the integral in Equation (3.2) as

$$k(\mathbf{x}_{t+h}|\mathbf{x}_t) = \prod_{i=1}^{h-1} \int_{\mathbb{R}^m} k(\mathbf{x}_{t+i-1} + \mathbf{y}|\mathbf{x}_{t+i-1}) q(\mathbf{y}) \, \mathrm{d}\mathbf{y}.$$

Consequently, each of the integrals in the product may be viewed as the expected value of the transition density with respect to the increment density. With this in mind, in order to estimate one of these integrals, we propose each of the increments $\mathbf{y}_i$, $i = 1, \ldots, N_2$ to the current state $\mathbf{x}$, and we carry out an RWM accept/reject step. If the proposal is accepted, we store $\alpha(\mathbf{x}, \mathbf{x} + \mathbf{y}_i)$. If the proposal is rejected, we store $1 - \alpha(\mathbf{x}, \mathbf{x} + \mathbf{y})$. Once this step has been completed for each of the N_2 increments, the average of the stored values from the accept/reject step is taken as a Monte Carlo estimate of

$$\int_{\mathbb{R}^m} k(\mathbf{x} + \mathbf{y}|\mathbf{x}) q(\mathbf{y}) \, \mathrm{d}\mathbf{y}. \tag{3.3}$$

The resulting values of each accept/reject step, either $\mathbf{x}$ or $\mathbf{x} + \mathbf{y}$ are stored as the initial states for the next set of proposed updates. We compute this Monte Carlo estimate of the expected one-step transition density for each of the points in our sampled small set $\hat{C}$. Letting $\hat{\varepsilon}_{\mathbf{x}_t}^{(i)}$ denote the estimate for the point $\mathbf{x}_t \in \hat{C}$ for the i^{th} step of the chain, we use as a conservative estimate of the integral in Equation (3.3)

$$\hat{\varepsilon}^{(i)} = \min_{\mathbf{x}_t \in \hat{C}} \hat{\varepsilon}_{\mathbf{x}_t}^{(i)}.$$

This procedure is repeated for each step of the chain. We rely on the fact that, since the one-step transition density is nonnegative, the $\hat{\varepsilon}^{(i)}$ values are also nonnegative. Thus, for a finite set A of nonnegative real numbers comprising elements a_i, $i = 1, \ldots, m$,

$$\min_{i=1,\ldots,m} \prod_{i=1}^{m} a_i \geq \prod_{i=1}^{m} \min_{i=1,\ldots,m} a_i$$

to construct the estimate final estimate $\hat{\varepsilon}$ of ε. This estimate is given by

$$\hat{\varepsilon} = \prod_{i=1}^{h-1} \hat{\varepsilon}^{(i)}.$$

The processes described in Sections 3.1 and 3.2 will be seen in Section 4 to provide stable estimates of the drift and minorization coefficients without incurring a prohibitive computational cost.

4. Illustrative examples

In this section, we describe four examples that are designed to demonstrate the usefulness of the methods described in Section 3. For each example, we provide average values of the estimated drift and minorization coefficients from 50 runs of the estimation procedure, along with stability measures for these estimates. We also examine the computing time and the upper bound on the mixing time. The first of these examples deals with allele frequency, and the second involves sampling from a mixture of bivariate normal densities. The third example is designed as a more real-world example involving coronary heart disease (CHD), and the fourth example is a higher-dimensional problem involving low birth weight (LBW). All results are summarized in Table 1.

4.1. Allele frequency

In a given population, the major allele "0" occurs with probability p, and the minor allele occurs with probability $1-p$. In a set of 1,000 individual allele pairs, we aim to estimate the probability of the major allele. Let n_0 denote the number of 00 allele pairs, n_H denote the number of heterozygous (01 or 10) allele pairs, and let n_1 denote the number of 11 allele pairs. The likelihood function for p given the genetic data $\mathbf{D}$ is

$$\mathcal{L}(p|\mathbf{D}) \propto p^{n_H+2n_0}(1-p)^{n_H+2n_1}.$$

The prior density of p is the U[0,1] distribution due to a lack of prior information about p. This yields the target density

$$p(p|\mathbf{D}) \propto p^{n_H+2n_0}(1-p)^{n_H+2n_1}\mathbb{I}_{[0,1]}(p).$$

Increments to p are proposed from the N(0,0.0009) distribution. This selection gives an acceptance rate of 30.93%, which is approximately optimal for mixing [5]. We typically want an acceptance rate between 20% and 40%. The chain is thinned so that samples are taken every $h = 10$ steps. In order to estimate the drift and minorization coefficients, we use the drift function

$$\hat{V}(p) = [\hat{p}(p|\mathbf{D})]^{-0.005},$$

where $\hat{p}(\cdot|\mathbf{D})$ is the estimated target density that results from using an estimate of the marginal density. The value of d is set at 1,000. In estimating λ, we take 200 points outside of $C = \left\{p : \hat{V}(p) \leq 1,000\right\}$ and run 500 10-step chains from each of the selected initial states. The resulting estimate of λ is 0.033. To estimate b, 400 points are chosen from inside C, and 1,000 10-step chains are run from each. This leads to an estimate $\hat{b} = 31.651$ of b. We can see that $2\hat{b}/(1-\hat{\lambda})-1 = 64.455$, so the choice of d is sufficiently large. To estimate ε, 500 points are chosen from inside C, and 1,000 increments are selected from the N(0,0.0009) density. The resulting estimate $\hat{\varepsilon}$ of ε is 0.323. These estimates give an upper bound of 8,510 steps on the mixing time of the chain.

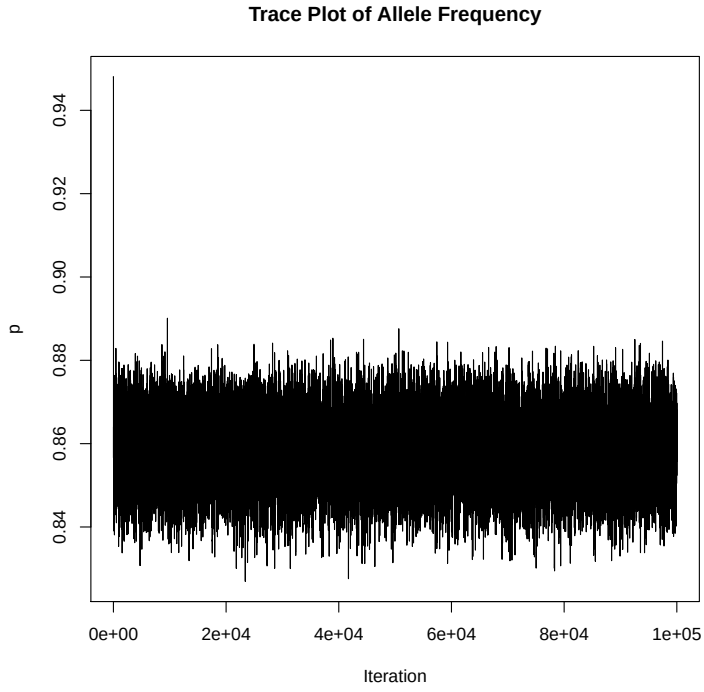

Figure 1. Trace plot of Allele Frequency Chain.

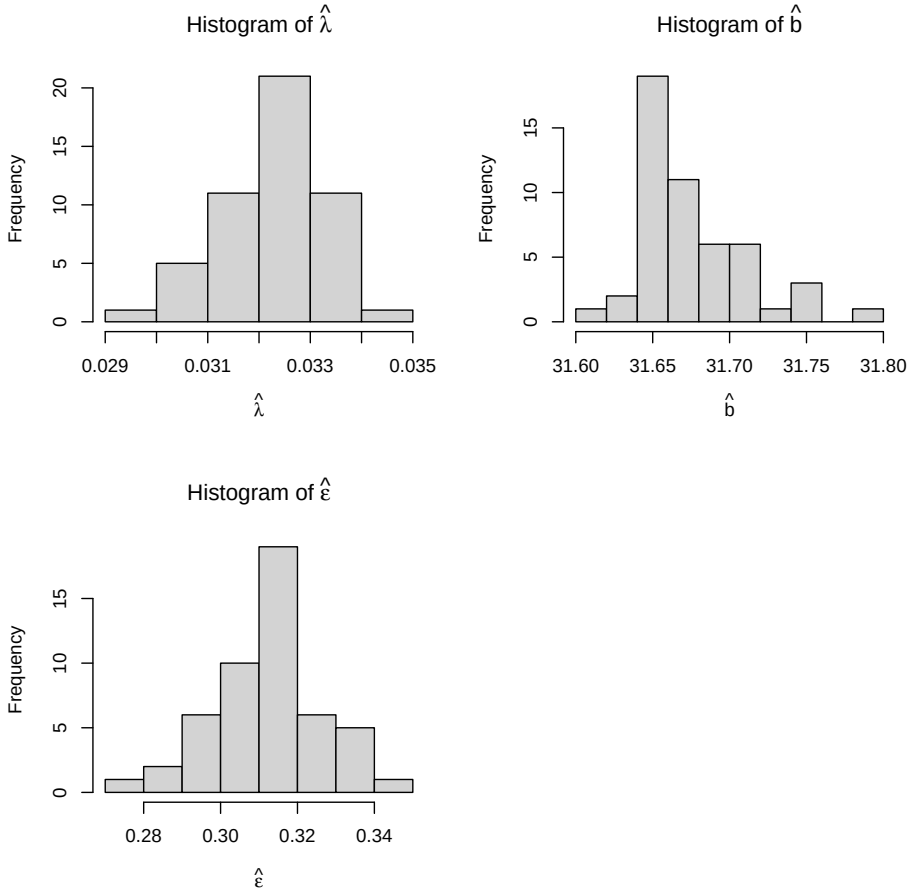

Figure 2. Histograms of $\hat{\lambda}$, $\hat{b}$, and $\hat{\varepsilon}$ for the Allele Frequency Chain.

The trace plots in Figure 1 suggest that this is a sufficient burn-in time. The process took 66.564 CPU seconds to complete. We repeated the process 50 times. The histograms of $\hat{\lambda}$, $\hat{b}$, and $\hat{\varepsilon}$ are in Figure 2. The average value of $\hat{\lambda}$ was 0.0322 with standard error 0.001, the mean value of $\hat{b}$ was 31.675 with standard error 0.005, and the mean value of $\hat{\varepsilon}$ was 0.313 with standard error 0.002. This gave an average upper bound on the mixing time of 8,534 steps with standard error 73.793 steps. For this example, our choices of tuning parameters gave estimates that are quite stable.

4.2. Mixture of bivariate normal densities

This example is a generic one in which we draw samples from a mixture of bivariate normal densities. Let $\mathbf{X}$ denote a sample. Then $\mathbf{X}$ is distributed according to a $0.5\mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ + $0.5\mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_2)$ density, where $\mathbf{\Sigma}_1 = diag(0.5, 1)$ and $\mathbf{\Sigma}_2 = diag(1, 0.5)$. Here, bivariate increments are proposed from a pair of independent uniform densities on [-1.5, 1.5]. This choice yields an acceptance rate of 27.02%. The drift function is $V(\mathbf{x}) = [p(\mathbf{x})]^{-0.05}$, where $p(\mathbf{x})$ is the target density. We thin the chain output by a factor of $h = 5$. The choice of d here is 30. In estimating λ, we choose 200 points outside the set $C = \{\mathbf{x} : V(\mathbf{x}) \leq 30\}$ and run 500 five-step chains from each. The resulting estimate of λ is $\hat{\lambda} = 0.144$. To estimate b, we use 400 initial states inside C and run 500 five-step chains from each one of them. This gave an estimate $\hat{b} = 1.018$ of b. The resulting value of $2\hat{b}/(1 - \hat{\lambda}) - 1$ is 1.378, so the choice of 30 for d is sufficiently large. Estimation of ε begins with the selection of 500 initial states inside C and 1,000 proposed increments from the increment density. The resulting estimate of ε is $\hat{\varepsilon} = 0.4375$, yielding an upper bound of 3,015 steps on the mixing time of the chain. In 50 runs of this process for this example, the average value of $\hat{\lambda}$ was 0.153 with standard error 0.004, the average value of $\hat{b}$ was 1.018 with standard error 0.0013, and the average value of $\hat{\varepsilon}$ was 0.4353 with standard error 0.0014. The average bound on the mixing time was 2,867 steps with standard error 23.305 steps. The process produces fairly stable estimates of drift and minorization coefficients. The process took an average of 31.961 CPU seconds to complete, with standard error 0.354 seconds. It is seen, then, that the estimation procedure is rather efficient in a computational sense for this example.

4.3. An example pertaining to CHD

This example examines a logistic regression model that relates the incidence of coronary heart disease to age. The data consist of information about 100 males aged 20 to 69 years. We let $\mathbf{X}$ denote the design matrix, which has dimension $100\times$ 2, where the first column is a column of ones and the second column contains ages. The presence y_i of CHD in individual i is assumed to follow a Bernoulli distribution with success probability $p_i(\boldsymbol{\beta})$, where

$$p_i(\boldsymbol{\beta}) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}}},$$

and $\mathbf{x}_i$ is the i^{th} row of $\mathbf{X}$. A priori, $\boldsymbol{\beta}$ is assumed to follow a bivariate normal distribution with mean vector $[-5, 0]^T$ and covariance matrix $\mathbf{\Sigma} = diag(0.25, 0.0625)$. This gives an acceptance rate of 27.17%. The posterior density of $\boldsymbol{\beta}$ is given by

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto \left[\prod_{i=1}^{100} \left(p_i(\boldsymbol{\beta})\right)^{y_i} \left(1 - p_i(\boldsymbol{\beta})\right)^{1-y_i}\right] e^{-2(\beta_0+5)^2 - 8\beta_1^2}.$$

In this example, the drift function is

$$\hat{V}(\boldsymbol{\beta}) = \left[\hat{p}(\boldsymbol{\beta}|\mathbf{y})\right]^{-0.04}.$$

The small set is chosen to be the set $\left\{\boldsymbol{\beta} : \hat{V}(\boldsymbol{\beta}) \leq 100\right\}$. Here, the chains are thinned by a factor of $h = 10$. To estimate λ, 200 initial states are selected from outside C, and 400 ten-step chains are run from each. The resulting estimate $\hat{\lambda}$ of λ is 0.091. To estimate b, 500 points in C are selected as initial states, and 1,000 ten-step chains are run from each. The resulting estimate of b is $\hat{b} = 2.451$. The value of $2\hat{b}/(1-\hat{\lambda}) - 1$ is 4.392. The choice of $d = 100$ is sufficiently large. In estimating λ, 500 initial states are chosen from inside C, and 1,000 $\mathbf{y}$ increments are chosen from the bivariate normal density with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is specified as above. The resulting estimate of ε is $\hat{\varepsilon} = 0.556$. This gave an upper bound of 4,010 steps on the mixing time. In the stability analysis, the mean value of $\hat{\lambda}$ is 0.093 with standard error 0.014, the mean value of $\hat{b}$ is 2.332 with standard error 0.059, and the mean value of $\hat{\varepsilon}$ is 0.564 with standard error 0.035. This gave an average bound of 3,970 steps on the mixing time with standard error 39.982 steps. The process took an average of 398.37 CPU seconds to run with standard error 12.069 seconds.

4.4. Low birth weight application

This example aims to illustrate the performance of this method in moderate dimensions. We examine a logistic regression model that connects the incidence of LBW to several predictors related to the mother. Among these are age, whether the mother smokes, and seven others. Let $\mathbf{x}_i$ denote the i^{th} row of the design matrix $\mathbf{X}$, and let $\boldsymbol{\beta}$ be a vector of regression coefficients. The response vector $\mathbf{y}$ contains 0s and 1s, where $y_i = 0$ if the i^{th} neonate weighs 2,500 grams or more, and $y_i = 1$, otherwise, where $i = 1, \ldots, 189$. The random variable Y_i follows a Bernoulli$(p_i(\boldsymbol{\beta}))$ distribution, where

$$p_i(\boldsymbol{\beta}) = \frac{e^{\mathbf{x}i\boldsymbol{\beta}}}{1 + e^{\mathbf{x}i\boldsymbol{\beta}}}.$$

A priori, $\boldsymbol{\beta}$ is assumed normal with mean vector

$$\boldsymbol{\beta}_0 = [1, 0, 0, 1, 1, 1, 0.5, 2, 0.75, 0]^T$$

and covariance matrix

$$\boldsymbol{\Sigma} = diag(0.25, 0.0625, 0.0625, \ldots, 0.0625).$$

This gives the target density

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto \left[\prod_{i=1}^{189} (p_i(\boldsymbol{\beta}))^{y_i} (1 - p_i(\boldsymbol{\beta}))^{1-y_i}\right] e^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}.$$

If $\boldsymbol{\beta}_t$ is the current state of the chain, an update $\boldsymbol{\beta}_t + \mathbf{y}$ is proposed by drawing $\mathbf{y}$ from a normal density with mean $\mathbf{0}$ and covariance matrix $diag(0.0004, 0.0001, 0.0001, \ldots, 0.0001)$. This increment density gave an acceptance rate of 25.12%. The drift function here is

$$\hat{V}(\boldsymbol{\beta}) = [\hat{p}(\boldsymbol{\beta}|\mathbf{y})]^{-0.005},$$

and the set C is chosen as $C = \left\{\boldsymbol{\beta} : \hat{V}(\boldsymbol{\beta}) \leq 1,000\right\}$. We thin the chain by a factor of $h = 20$ steps. For estimating λ, 500 initial states outside C are selected, from each of which 400 20-step chains are run. The resulting estimate of λ is $\hat{\lambda} = 0.002$. For the estimation of b, we chose 200 initial states inside C and ran 1,000 20-step chains from each. From these choices, we get an estimate $\hat{b} = 1.600$ of b. The value of $2\hat{b}/(1-\hat{\lambda}) - 1$ is 2.206, so 1,000 is a sufficiently large threshold for inclusion in C. To estimate ε, we choose 500 initial states in C and 500 increments from the increment density. The estimate of ε is $\hat{\varepsilon} = 0.232$. This gives an upper bound of 24,220 steps on the mixing time. The process took 886.893 CPU seconds to run. In the stability analysis, the mean value of $\hat{\lambda}$ was 0.0021 with standard error 0.0005, the mean value of $\hat{b}$ was 1.603 with standard error 0.0009, and the mean

value of $\hat{\varepsilon}$ was 0.241 with standard error 0.003. The average bound on the mixing time was 23,093.64 steps with standard error 310.514 steps. The process took an average of 984.750 CPU seconds to complete with standard error 88.893 seconds. In Table 1, the mean values of $\hat{\lambda}$, $\hat{b}$, and $\hat{\varepsilon}$, as well as the bound on the mixing time, are given along with their standard errors in parentheses. The AF row corresponds to the allele frequency example, the BVN row corresponds to the mixture bivariate normal example, and the CHD and LBW rows correspond to the CHD and LBW examples, respectively.

Table 1. Stability results.

	$\hat{\lambda}$	$\hat{b}$	$\hat{\varepsilon}$	**Mixing Time**
AF	0.032(0.001)	31.675(0.005)	0.313(0.005)	8,534(73.793)
BVN	0.153(0.002)	1.018(0.0013)	0.435(0.0014)	2,867(23.305)
CHD	0.093(0.014)	2.332(0.059)	0.564(0.035)	3,970(39.982)
LBW	0.0021(0.0005)	1.603(0.0009)	0.241(0.003)	23,093(310.514)

5. Conclusion

In this manuscript, we present a novel technique for approximate verification of geometric ergodicity of h-step RWM transition kernels. While it does not guarantee geometric ergodicity of these chains, it is a useful method of making determinations as to whether it would be reasonably safe to treat the chain as though it were geometrically ergodic. Herein, we demonstrate through illustrative examples that the proposed technique provides stable estimates of drift and minorization coefficients. This method is also computationally efficient in moderate dimensions. This is one property that the Cowles and Rosenthal [3] approach does have since that technique of estimating a minorization coefficient requires division of the state space into bins and that enough chains be run from each initial state to ensure adequate coverage of those bins. With that in mind, our method is certainly not without its limitations. One of these limitations is that, in many cases, it is necessary to estimate the marginal density. While in the examples that we present, this is not a difficult task, in other practical settings with complicated prior densities from which it is difficult to obtain samples, marginal density estimation can be very difficult to do. A second limitation is that our approach to estimating the minorization coefficient relies on the multiplication of h terms that lie between 0 and 1. This means that for chains with substantial autocorrelation where thinning factors are in the hundreds or thousands, estimated minorization coefficients can be very small. While this is not of concern if the goal is strictly to gather evidence for geometric ergodicity, if it is also desired to use the estimated drift and minorization coefficients to bound the mixing time, the estimated minorization coefficient may be too small to provide a practically useful bound. While this approach is able to handle moderate dimensional situations fairly easily, it will eventually come across the curse of dimensionality. As the dimension gets large, the number of initial states needs to increase accordingly in order to ensure a representative sampled set. The number of chains that need to be run from each of them will also need to increase. Without increasing both of these tuning parameters with the dimension, we lose stability in the estimation process. This leads to highly variable bounds on the mixing time. One way to reduce the size of the sets of initial states is to choose values near the boundaries of the small set that is specified.

While this method is useful in verifying geometric ergodicity of h-step RWM transition kernels, there are better methods of bounding mixing times that warrant further investigation in terms of extension to h-step transition kernels. For example, Johnson [11] uses coupled sampling paths to estimate mixing times. The method presented in that paper

works well for several classes of MCMC algorithms. Roberts and Rosenthal [17] also use couplings to establish geometric ergodicity of adaptive MCMC algorithms. Atchadé [1] uses an approximate spectral gap to bound the mixing time. The aim is to extend these results to multiple-step transition kernels. Another avenue to pursue is to see how the method presented in this manuscript might be used to approximately establish geometric ergodicity of dimension-switching algorithms such as reversible-jump MCMC. It is also important to investigate adaptations of this approach to Metropolis-Hastings samplers with asymmetric proposal densities. This is a more involved task, but the work presented here represents a useful beginning to attacking these problems.

References

[1] Y.F. Atchadé, *Approximate spectral gaps for Markov chain mixing times in high dimensions*, SIMODS **3** (3), 854-872, 2021.

[2] S. Chib, F. Nardari and N. Shephard, *Markov chain Monte Carlo methods for generalized stochastic volatility models*, J. Econometrics **108** (2), 281–316, 2002.

[3] M.K. Cowles and J.S. Rosenthal, *A simulation-based approach to convergence rates for Markov chain Monte Carlo algorithms*, Statist. Comput. **8**, 115–124, 1998.

[4] G. Fort, E. Moulines, G.O. Roberts and J.S. Rosenthal, *On the geometric ergodicity of hybrid samplers*, J. Appl. Probab. **40** (1), 123–146, 2003.

[5] A. Gelman, W.R. Gilks and G.O. Roberts, *Weak convergence and optimal scaling of random walk Metropolis algorithms*, Ann. Appl. Probab. **7** (1), 110–120, 1997.

[6] A. Gelman and D.B. Rubin, *Inference from iterative simulation using multiple sequences*, Statist. Sci. **7** (4), 457–511, 1992.

[7] J. Geweke, *Evaluating the accuracy of sampling-based approaches to calculating posterior moments*, in: Bayesian Statistics 4, Eds: J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith, Oxford University Press, 169–193, 1992.

[8] P. Heidelberger and P.D. Welch, *Simulation run length control in the presence of an initial transient*, Oper. Res. **31** (6), 1109–1144, 1983.

[9] H. Ishwaran, L.F. James and J. Sun, *Bayesian model selection in finite mixtures by marginal density decompositions*, J. Amer. Statist. Assoc. **96** (456), 1316–1332, 2001.

[10] S.F. Jarner and E. Hansen, *Geometric ergodicity of Metropolis algorithms*, Stochastic Process. Appl. **85** (2), 341–361, 2000.

[11] V.E. Johnson, *Studying convergence of Markov chain Monte Carlo algorithms using coupled sampling paths*, J. Amer. Statist. Assoc. **91** (433), 154–166, 1996.

[12] F. Liang, *Continuous contour Monte Carlo for marginal density estimation with an application to a spatial statistical model*, J. Comput. Graph. Statist. **16** (3), 608–632, 2007.

[13] S.P. Meyn and R.L. Tweedie *Markov Chains and Stochastic Stability*, 2nd ed., Springer-Verlag, London, 2005.

[14] R.M. Neal, *Annealed Importance Sampling*, Technical report, University of Toronto, Department of Statistics, 1998.

[15] M. Oh and J.O. Berger, *Adaptive importance sampling in Monte Carlo integration*, Technical report, Purdue University, Department of Statistics, 1989.

[16] G.O. Roberts, *Methods for estimating L^2 convergence of Markov chain Monte Carlo*, in: Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner, Eds: D. Berry, I. Chaloner and J. Geweke, Amsterdam, North-Holland, 373–384, 1996.

[17] G.O. Roberts and J.S. Rosenthal, *Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms*, J. Appl. Probab. **44** (2), 458-475, 2007.

[18] G.O. Roberts and R.L. Tweedie, *Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms*, Biometrika **83** (1), 95–110, 1996.
[19] J.S. Rosenthal, *Minorization conditions and convergence rates for Markov chain Monte Carlo*, J. Amer. Statist. Assoc. **90** (430), 558–566, 1995.
[20] D.A. Spade, *A computational procedure for efficient estimation of the mixing time of a random-scan Metropolis algorithm*, Stat. Comput. **26** (4), 761–781, 2016.
[21] D.A. Spade, *A computational approach to bounding the mixing time of a Metropolis-Hastings sampler*, Markov Process. Relat. Fields **26** (3), 487–516, 2020.
[22] D.A. Spade, *A Monte Carlo integration approach to estimating drift and minorization coefficients for Metropolis-Hastings samplers*, Braz. J. Probab. Stat. **35** (3), 466–483, 2021.
[23] B. Yu, *Monitoring the convergence of Markov samplers based on estimated L^1 error*, Technical report 409, University of California, Department of Statistics, 1994.
[24] B. Yu and P. Mykland, *Looking at Markov samplers through CUSUM path plots: a simple diagnostic idea*, Technical report 413, University of California, Department of Statistics, 1994.