International Journal of Statistics and Probability; Vol. 13, No. 1; March 2024
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Estimating and Calibrating Markov Chain Sample Error Variance

Yann Vestring[1] & Javad Tavakoli[1]

[1] Department of Mathematics, University of British Columbia Okanagan, Kelowna, BC, Canada

Correspondence: Yann Vestring, Department of Mathematics, University of British Columbia Okanagan, Kelowna, BC, Canada

Received: November 15, 2023 Accepted: January 29, 2024 Online Published: February 28, 2024

doi:10.5539/ijsp.v13n1p10 URL: https://doi.org/10.5539/ijsp.v13n1p10

Abstract

Markov chain Monte Carlo (MCMC) methods are a powerful and versatile tool with applications spanning a wide spectrum of fields, including Bayesian inference, computational biology, and physics. One of the key challenges in applying MCMC algorithms is to deal with estimation error. The main result in this article is a closed form, non-asymptotic solution for the sample error variance of a single MCMC estimate. Importantly, this result assumes that the state-space is finite and discrete. We demonstrate with examples how this result can help estimate and calibrate MCMC estimation error variance in the more general case, when the state-space is continuous and/or unbounded.

1. Introduction

Markov chain Monte Carlo (MCMC) is a very powerful tool for estimating and sampling from complicated high-dimensional distributions. MCMC algorithms are used in a wide spectrum of fields, including Bayesian inference (Gamerman & Lopes, 2006), computational biology (Gelman & Rubin, 1996), and physics (Binder & Heermann, 2010).

Consider an irreducible Markov chain $X : \mathcal{X} \to S$ with stationary distribution $\pi(\cdot)$. Let $g : \mathcal{X} \to \mathbb{R}$ be a measurable function. In the MCMC context, the objective is generally to estimate a finite measure of interest $\pi(g) =: \int_{x\in\mathcal{X}} \pi(x)g(x)\,dx$ based on a sample estimate $\hat{\pi}(g) =: \frac{1}{n}\sum_{t=1}^{n} g(X_t)$, where the sample path $(X_t)_{t=0}^{n}$ is obtained using simulation.

One of the biggest challenges with MCMC algorithms is to evaluate the error of the estimates. This paper seeks to provide novel tools to address this challenge by proposing approaches to estimating and calibrating $\mathbf{Var}\big(\hat{\pi}(g) - \pi(g)\big) = \mathbf{Var}\big(\hat{\pi}(g)\big)$ that should be applicable to a broad range of MCMC contexts.

2. Related Literature

A plethora of empirical diagnostic tools have been proposed in the statistical literature to analyze MCMC convergence and estimation error. Roy, 2020 provides a recent critical overview of some of the more popular approaches. In particular, they highlight that empirical diagnostics can be prone to overly conservative convergence assessments or falsely detect convergence.

The majority of the more theoretically founded diagnostics for quantifying MCMC accuracy rely on the Markov chain Central Limit Theorem (MCCLT). The MCCLT states that

$$\sqrt{n}\Big(\hat{\pi}(g) - \pi(g)\Big) \xrightarrow{\mathcal{D}} N\Big(0, \sigma^2\Big),$$

where $\sigma^2 = \lim_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big)$. The conditions under which the MCCLT applies is a well studied problem (see Jones, 2004 among others). However, these conditions can be difficult to verify, in particular for non-reversible Markov chains (see Hang Jian et al., 2022 for references on this point). In any case, there are at least two key practical challenges in invoking the MCCLT, as discussed in Robert & Casella, 2011.

First, σ^2 is generally unknown and needs to be ascertained somehow. While closed form results exist for Markov chains over finite and discrete state-spaces (Spitzner & Boucher, 2007; Trevezas & Limnios, 2009), the dominant approach in the MCMC literature is to estimate σ^2 based on simulation output using techniques such as replication and batch means (see Hang Jian et al., 2022 for more details). For example, in Hang Jian et al., 2022 the authors used simulation for various values of n to estimate $\sup_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big)$. Such approaches beg the question as to what value of n is high enough to infer asymptotic variance from sample variance. This is a hard question to answer if the convergence properties (such as the mixing time - see Levin, Peres, & Wilmer, 2009 for more details) of the Markov chain are unknown.

Second, the MCCLT only provides asymptotic guarantees, yet in practise n is always finite. This is another reason why the convergence behaviour of $\mathbf{Var}\big(\hat{\pi}(g)\big)$ needs to be understood.

While non-asymptotic bounds have been proposed in the literature (see Łatuszyski, Miasojedow, & Niemiro, 2013), we were not able to find any non-asymptotic closed form solutions for $\mathbf{Var}\big(\hat{\pi}(g)\big)$. Much of the difficulty in estimating $\mathbf{Var}\big(\hat{\pi}(g)\big)$ stems from the fact that in MCMC applications, $\mathcal{X}$ is often continuous, multidimensional, and/or unbounded. On the other hand, when $\mathcal{X}$ is discrete and finite, as we will show, $\mathbf{Var}\big(\hat{\pi}(g)\big)$ can be derived non-asymptotically by conditioning on the number of visits to each state in $\mathcal{X}$.

This article adds to the existing literature in that it i) proposes a closed form solution for $\mathbf{Var}\big(\hat{\pi}(g)\big)$ for any ergodic Markov chain with a finite and discrete state-space, ii) outlines approaches to extending this result to cases where the state-space is continuous or unbounded, and iii) demonstrates with examples how our results enable us to target a desired level of $\mathbf{Var}\big(\hat{\pi}(g)\big)$ in the context of MCMC estimation.

3. Main Results and Proofs

3.1 Main Results

Theorem 3.1. *Let $X : \mathcal{X} \to S$ be an irreducible Markov chain with finite and discrete state-space $\mathcal{X} := \{x_k : k \in [d]\}$, transition matrix P, and unique stationary distribution vector Π.*

Let $\hat{\pi}(g)$ be based on a sample path $(X_t)_{t=0}^n$ of this Markov chain, and let $\hat{P}$ be the maximum likelihood estimate of P based on this sample path. Furthermore, assume that $\hat{P}$ admits the unique stationary distribution vector $\hat{\Pi}$. In this context

$$\mathbf{Var}\Big(\hat{\pi}(g) \mid \{N_k\}_{k\in[d]}\Big) = \sum_{k=1}^{d} \frac{\gamma_k \hat{\Pi}_k^2}{N_k},$$

where $N_k = |\{t \in [n-1] : X_t = x_k\}|$ and $\{\gamma_k\}_{k\in[d]}$ is a set of scalars that depend on g, P, and a "one condition g-inverse" of P .

The definition of a one condition g-inverse is given in section 3.2. The proof of theorem 3.1 (lemmas 3.4 and 3.5) as well as the computation of $\{\gamma_k\}_{k\in[d]}$ (lemma 3.6) are given in section 3.3.

The following two corollaries follow straightforwardly from theorem 3.1.

Corollary 3.2. *By making the approximation that $\forall\, k \in [d]$, $\hat{\Pi}_k \approx N_k/n$, the above equality yields the simplified expression*

$$\mathbf{Var}\Big(\hat{\pi}(g) \mid \{N_k\}_{k\in[d]}\Big) \approx \frac{1}{n^2} \sum_{k=1}^{d} N_k \gamma_k.$$

Corollary 3.3.

$$\lim_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big) = \sum_{k=1}^{d} \gamma_k \Pi_k.$$

Corollary 3.3 follows from the fact that $\lim_{n\to\infty} \hat{\Pi}_k = \lim_{n\to\infty} \frac{N_k}{n} = \Pi_k$. We reiterate that results similar to corollary 3.3 already exist in the literature (see Spitzner & Boucher, 2007; Trevezas & Limnios, 2009).

3.2 Core Concepts Underlying Theorem 3.1

3.2.1 One condition g-inverse and Matrix Perturbation

The following definition and equation (1) are drawn from Hunter, 2005.

Definition (One condition g-inverse). *A one condition g-inverse of a matrix A is any matrix A^- such that $AA^-A = A$.*

Let $P = [p_{ij}] \in \mathbb{R}^{d\times d}$ be the transition matrix of a finite irreducible Markov chain which is assumed to have a unique steady-state distribution vector Π. Let $\tilde{P} = [\tilde{p}_{ij}] = P + E$ be the transition matrix of the perturbed Markov chain where $E = [\varepsilon_{ij}]$ is the matrix of perturbations. Notice that $\forall (i, j) \in [d]^2$, $|\varepsilon_{ij}| \leq 1$ and $\forall i \in [d]$, $\sum_{j=1}^d \varepsilon_{ij} = 0$. $\tilde{P}$ is assumed to admit the unique steady-state distribution vector $\tilde{\Pi}$. Then according to theorem 2.1 in Hunter, 2005,

$$\tilde{\Pi}^T - \Pi^T = \tilde{\Pi}^T EH, \tag{1}$$

where $H = G(I - e\Pi^T)$, $e^T = (1, ..., 1)$, I is the identity matrix, and G is a one condition g-inverse of $I - P$.

3.2.2 Computing Matrix H

An exhaustive review of the different ways to compute matrix G (and therefore matrix H) is beyond the scope of this work. For the purposes of the examples in section 4, we computed G as the so-called group inverse of $I-P$. This choice is mainly motivated by ease of computation for the transition matrix of an ergodic Markov chain (see theorem 5.2 in Meyer, 1975). Indeed, the most computationally intensive step is the calculation of the inverse of a $(d-1)^2$ principal submatrix of $I-P$.

It bears mentioning that H can also be derived based on the matrix of mean first passage times and the steady-state distribution vector (see Hunter, 2005), which can enhance interpretability.

3.2.3 The Sequence Matrix

For a sample path of length n characterised by the sequence of observed states $(X_i)_{i=0}^n$ of a discrete ergodic Markov chain with transition matrix P, recall how we defined N_i in section 3.1. For any pair of states $(i,j)\in\mathcal{X}^2$, let N_{ij} be the number of transitions from state i to state j in the sample path, or more formally $N_{ij} := |\{t\in[n-1] : (X_t, X_{t+1}) = (i,j)\}|$, $(i,j)\in[d]^2$. We will refer to $M=[N_{ij}]\in\mathbb{N}^{d\times d}$ as the sequence matrix. Also, throughout this work, we will assume that $\forall i\in[d]$, $N_i\geq 1$.

Remark. Conditional on $\{N_i \mid i\in[d]\}$, the rows of M are mutually independent and respectively follow multinomial distributions with N_i trials and unknown event probabilities $p_{i1},...,p_{id}$.

Proof. Conditional on N_i, the joint outcome of $N_{i1},...,N_{id}$ can be equated to N_i independent trials (independence follows from the Markov property) where the outcome of each trial has a categorical distribution with fixed success probabilities $p_{i1},...,p_{id}$. In other words, conditional on N_i, $N_{i1},...,N_{id}$ jointly follow a multinomial distribution with N_i trials and event probabilities $p_{i1},...,p_{id}$.

Also, $\forall i\in[d]$ let $N_{i\cdot}$ be a vector corresponding to row i of the sequence matrix M and let $P_{i\cdot}$ be a vector corresponding to row i of the unknown transition matrix P. Given known $\{N_i \mid i\in[d]\}$ and unknown $\{P_{i\cdot} \mid i\in[d]\}$, for $j\neq k$:

$$\begin{aligned}
&\mathbf{P}\Big(N_{j\cdot}=X\bigcap N_{k\cdot}=Y \mid \{N_i \mid i\in[d]\},\{P_{i\cdot} \mid i\in[d]\}\Big)\\
&=\mathbf{P}\Big(N_{j\cdot}=X \mid N_{k\cdot}=Y,\{N_i \mid i\in[d]\},\{P_{i\cdot} \mid i\in[d]\}\Big)\mathbf{P}\Big(N_{k\cdot}=Y \mid \{N_i \mid i\in[d]\},\{P_{i\cdot} \mid i\in[d]\}\Big)\\
&=\mathbf{P}\Big(N_{j\cdot}=X \mid N_j,P_{j\cdot}\Big)\mathbf{P}\Big(N_{k\cdot}=Y \mid N_k,P_{k\cdot}\Big)\\
&=\mathbf{P}\Big(N_{j\cdot}=X \mid \{N_i \mid i\in[d]\},\{P_{i\cdot} \mid i\in[d]\}\Big)\mathbf{P}\Big(N_{k\cdot}=Y \mid \{N_i \mid i\in[d]\},\{P_{i\cdot} \mid i\in[d]\}\Big).
\end{aligned}$$

□

Importantly, independence between the rows only holds conditional on the knowledge of $\{N_i \mid i\in[d]\}$. Otherwise the rows of M are of course not independent since crucially $\forall i\in[d]$, N_i depends on $\sum_{j\in[d]} N_{ji}$.

3.3 Proof of Theorem 3.1

Notice that for a sample path $(X_t)_{t=0}^n$ of an irreducible Markov chain with discrete and finite state-space $\mathcal{X}$, transition matrix P, and unique steady-state distribution vector Π, we can build the approximation $\hat{P}=[\hat{p}_{ij}]=[N_{ij}/N_i]$ of P based on the sequence matrix. Moreover, $\hat{P}$ can be viewed as a perturbation of P (and vice versa).

Hence the difference between $\hat{\pi}(g)$ and $\pi(g)$ can be expressed in terms of this perturbation of the transition matrix (lemma 3.4). This is useful because it allows us to explicitly derive $\mathbf{Var}\Big(\hat{\pi}(g)-\pi(g)\Big)$ conditional on the number of visits to each state in $\mathcal{X}$ (lemma 3.5).

Lemma 3.4. *Assume that $\hat{P}$ admits the unique steady-state distribution vector $\hat{\Pi}$. Let $E=[\varepsilon_{ij}]=\hat{P}-P$ and $H=G(I-e\Pi^T)$, where $e^T=(1,...,1)$, I is the identity matrix, and G is a one condition g-inverse of $I-P$. Then*

$$\hat{\pi}(g)-\pi(g)=\sum_{k=1}^{d}\hat{\Pi}_k\sum_{i=1}^{d}g(x_i)\sum_{l=1}^{d}h_{li}\varepsilon_{kl}.$$

Proof. Lemma 3.4 is essentially a corollary of equation (1).

$$\begin{aligned}
\hat{\pi}(g) - \pi(g) &= \sum_{k=1}^{d} \left(\hat{\Pi}_k - \Pi_k\right) g(x_k) \\
&= \left(\hat{\Pi}^T - \Pi^T\right) \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_d) \end{bmatrix} \\
&= \hat{\Pi}^T E H \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_d) \end{bmatrix} \\
&= \sum_{k=1}^{d} \hat{\Pi}_k \sum_{i=1}^{d} g(x_i) \sum_{l=1}^{d} h_{li} \varepsilon_{kl}
\end{aligned}$$

□

Lemma 3.5. *Using the same notation as for lemma 3.4,*

$$\mathbf{Var}\Big(\hat{\pi}(g) - \pi(g) \mid \{N_k\}_{k\in[d]}\Big) = \sum_{k=1}^{d} \hat{\Pi}_k^2 \sum_{i=1}^{d} \sum_{j=1}^{d} g(x_i) g(x_j) \mathbf{Cov}\Big(\sum_{l=1}^{d} h_{li}\varepsilon_{kl}, \sum_{m=1}^{d} h_{mj}\varepsilon_{km}\Big).$$

Proof. Conditional on $\{N_k\}_{k\in[d]}$,

$$\begin{aligned}
&\mathbf{Var}\Big(\hat{\pi}(g) - \pi(g)\Big) \\
&= \mathbf{Var}\left(\sum_{k=1}^{d} \hat{\Pi}_k \sum_{i=1}^{d} g(x_i) \sum_{l=1}^{d} h_{li}\varepsilon_{kl}\right) \\
&= \sum_{k=1}^{d} \sum_{k'=1}^{d} \hat{\Pi}_k \hat{\Pi}_{k'} \mathbf{Cov}\left(\sum_{i=1}^{d} g(x_i) \sum_{l=1}^{d} h_{li}\varepsilon_{kl}, \sum_{i=1}^{d} g(x_i) \sum_{l=1}^{d} h_{li}\varepsilon_{k'l}\right) \\
&= \sum_{k=1}^{d} \hat{\Pi}_k^2 \mathbf{Var}\left(\sum_{i=1}^{d} g(x_i) \sum_{l=1}^{d} h_{li}\varepsilon_{kl}\right) \\
&= \sum_{k=1}^{d} \hat{\Pi}_k^2 \mathbf{Var}\left(\sum_{i=1}^{d} g(x_i) \sum_{l=1}^{d} h_{li}\varepsilon_{kl}\right) \\
&= \sum_{k=1}^{d} \hat{\Pi}_k^2 \sum_{i=1}^{d} \sum_{j=1}^{d} g(x_i) g(x_j) \mathbf{Cov}\Big(\sum_{l=1}^{d} h_{li}\varepsilon_{kl}, \sum_{m=1}^{d} h_{mj}\varepsilon_{km}\Big)
\end{aligned}$$

The first equality follows from lemma 3.4. The third equality follows from conditional independence between the rows of the sequence matrix. □

Lemma 3.6. *Without loss of generality, assume that the row vector $P_{k.}$ has only nonzero elements. $\forall (i, j, k) \in \mathbb{N}^3$, conditional on N_k,*

$$\mathbf{Cov}\Big(\sum_{l=1}^{d} h_{li}\varepsilon_{kl}, \sum_{m=1}^{d} h_{mj}\varepsilon_{km}\Big) = \frac{1}{N_k} \hat{h}_i^T \Sigma_k \hat{h}_j,$$

where

$$\hat{h}_i = \begin{bmatrix} h_{1i} \\ \vdots \\ h_{di} \end{bmatrix} - h_{di} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\hat{h}_j = \begin{bmatrix} h_{1j} \\ \vdots \\ h_{dj} \end{bmatrix} - h_{dj} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

and

$$\Sigma_k = \begin{pmatrix} p_{k1}(1-p_{k1}) & -p_{k1}p_{k2} & \cdots & -p_{k1}p_{kd} \\ -p_{k1}p_{k2} & p_{k2}(1-p_{k2}) & \cdots & -p_{k2}p_{kd} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{k1}p_{kd} & -p_{k2}p_{kd} & \cdots & p_{kd}(1-p_{kd}) \end{pmatrix}.$$

Proof.

$$\begin{aligned}
&\mathbf{Cov}\Big(\sum_{l=1}^{d} h_{li}\varepsilon_{kl}, \sum_{m=1}^{d} h_{mj}\varepsilon_{km}\Big) \\
&= \mathbf{Cov}\Big(\sum_{l=1}^{d-1}(h_{li}-h_{di})\varepsilon_{kl}, \sum_{m=1}^{d-1}(h_{mj}-h_{dj})\varepsilon_{km}\Big) \\
&= \sum_{l=1}^{d-1}\sum_{m=1}^{d-1}(h_{li}-h_{di})(h_{mj}-h_{dj})\mathbf{Cov}\Big(\varepsilon_{kl},\varepsilon_{km}\Big) \\
&= \frac{1}{N_k^2}\sum_{l=1}^{d-1}\sum_{m=1}^{d-1}(h_{li}-h_{di})(h_{mj}-h_{dj})\mathbf{Cov}\Big(N_{kl},N_{km}\Big) \\
&= \frac{1}{N_k}\sum_{l=1}^{d-1}(h_{li}-h_{di})^2 p_{kl}(1-p_{kl}) - \frac{1}{N_k}\sum_{l=1}^{d-1}\sum_{m\neq l}(h_{li}-h_{di})(h_{mj}-h_{dj})p_{kl}p_{km} \\
&= \frac{1}{N_k}\hat{h}_i^T\Sigma_k\hat{h}_j
\end{aligned}$$

The first equality follows from the fact that $\forall\, k \in [d]$, $\sum_{l=1}^{d}\varepsilon_{kl} = 0$. The fourth equality follows from well known properties of the multinomial distribution (e.g., see chapter 2 of Rudas, 2018 for more details) □

These three lemmas collectively prove Theorem 3.1 and show how to compute $\{\gamma_k\}_{k\in[d]}$.

4. Applications and Examples

Provided P is known or can be approximated, theorem 3.1 lends itself to the computation of confidence intervals for $\pi(g)$ based on a single sample path of a Markov chain. However, when P is unknown and needs to be estimated, which is generally the case in MCMC applications, this confidence interval is not guaranteed to contain $\pi(g)$ with the prescribed level of confidence, as the variance may be underestimated.

Another application of theorem 3.1 could be to generate a sample of MCMC estimates with a target level of variance. If $\{\gamma_k\}_{k\in[d]}$ can be computed (or estimated), then corollary 3.2 lends itself to a stopping rule whereby the sample generation process is interrupted as soon as $\frac{1}{n^2}\sum_{k=1}^{d}N_k\gamma_k$ is less than or equal to the variance target. We will illustrate this application with two examples in sections 4.2 and 4.4.

To make matters more complicated, in MCMC applications $\mathcal{X}$ can often be continuous and/or unbounded. This suggests our main result is of limited use unless it is possible to extend it to cases where $\mathcal{X}$ is continuous and/or unbounded. In the next several subsections, we propose approximations to this effect which we illustrate with the examples in sections 4.2 and 4.4.

4.1 Approximation for Unbounded $\mathcal{X}$

As was previously mentioned, in MCMC applications $\mathcal{X}$ may be unbounded. However, theorem 3.1 only applies for discrete and finite $\mathcal{X}$. Hence in order to invoke theorem 3.1 when $\mathcal{X}$ is unbounded we need to find a bounded set $\hat{\mathcal{X}}$ such that $\int_{\hat{\mathcal{X}}}\pi(x)\,dx \approx 1$. In practise this can be done by identifying boundary values of $\mathcal{X}$ beyond which $\pi(x)$ becomes disproportionately small, as will be exemplified in the next subsection.

4.2 Example 1: Diffusive Non-reversible Chain

Consider the numerical example in section 6.2 in Hang Jian et al., 2022. Specifically, consider a Markov chain with state-space $\mathbb{N}$ and transition probability given by $p_{0,0} = 0.99$, $p_{0,1} = 0.01$, and $\forall\, x \geq 1$, $p_{x,x+1} = \left(\frac{x}{x+1}\right)^2$ and $p_{x,0} = 1-\left(\frac{x}{x+1}\right)^2$.

This Markov chain can be shown to have have stationary distribution

$$\pi(0) = \frac{1}{1 + \frac{0.01\pi^2}{6}}, \text{ and } \pi(x) = \frac{0.01}{x^2}\pi(0) \text{ for } x \geq 1.$$

Moreover, let

$$g(x) = \begin{cases} 0 & \text{if } x = 0, \\ x^{-1} & \text{if } x \geq 1. \end{cases}$$

Hang Jian et al., 2022 discuss this example specifically because the Markov chain is non-reversible and it is therefore non-trivial to establish a CLT and obtain a confidence interval for $\pi(g)$. They proceed to apply a theorem from their paper which provides a theoretical basis to establish a confidence interval for $\pi(g)$. Crucially, this theorem requires knowledge of a number $B > 0$ such that $\sup_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big) \leq B^2$.

In order to determine B, the authors generated samples of 1,000 values of $\hat{\pi}(g)$ for n ranging from 100 to 100,000. They obtained sample values for $n\mathbf{Var}\big(\hat{\pi}(g)\big)$ ranging from 0.0141 to 0.0181. Furthermore, their data suggests that $n\mathbf{Var}\big(\hat{\pi}(g)\big)$ probably converges to a point in the neighbourhood of 0.015. However, in the absence of any theoretical guarantees that $\sup_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big)$ is contained between 0.0141 and 0.0181, the authors understandably settled for the conservative $B = \sqrt{0.02}$.

We will now show how corollary 3.2 can be used to provide theoretically grounded evidence that $\lim_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big)$ is in the vicinity of 0.015 (which would imply that $\sup_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big)$ is also in the vicinity of 0.015).

Consider a Markov chain over state-space $[0,d]$ with transition matrix

$$\mathcal{A}_d := \begin{pmatrix} 0.99 & 0.01 & 0 & \cdots & 0 \\ 0.75 & 0 & 0.25 & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 - \left(\frac{d}{d+1}\right)^2 & 0 & 0 & \cdots & \left(\frac{d}{d+1}\right)^2 \end{pmatrix}.$$

Such a finite state-space Markov chain essentially approximates the previously described Markov chain of interest by cutting off the state space. This approximation becomes increasingly precise as d increases towards ∞. One way to see this is by observing that the first element of matrix $\mathcal{A}_d$'s steady-state distribution vector Π converges towards $\pi(0) = \frac{1}{1+\frac{0.01\pi^2}{6}} \approx 0.983817$ as d increases, as shown in the below table.

d	Π_0
5	0.984693
10	0.984277
25	0.984007
50	0.983913
75	0.983881
100	0.983865
150	0.983849

Since corollary 3.3 can not be directly applied to the Markov chain of interest (as its state-space is unbounded), our proposed strategy is to compute $\{\Pi_k\}_{k\in[0,d]}$ and $\{\gamma_k\}_{k\in[0,d]}$ based on matrix $\mathcal{A}_d$. This allows us to approximate $\lim_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big)$ for the chain of interest using corollary 3.3. The below table shows values of $\sum_{k=1}^{d} \gamma_k \Pi_k$ (which according to corollary 3.3 equals $\lim_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big)$ for the Markov chain over finite state-space $[0,d]$) for various values of d.

d	$\sum_{k=1}^{d} \gamma_k \Pi_k$
5	0.015392
10	0.015258
25	0.015213
50	0.015201
75	0.015197
100	0.015195
150	0.015194

Thus the data suggests that $\lim_{n\to\infty} n\mathbf{Var}\big(\hat{\pi}(g)\big)$ is approximately 0.0152 for the Markov chain of interest.

Next, our aim is to show how corollary 3.2 can be used to generate a sample of estimates for $\pi(g)$ with approximately a certain level of target variance. We achieve this by applying corollary 3.2 to a finite state-pace Markov chain with transition matrix $\mathcal{A}_d$, whereby $d = 3$.

Furthermore, for a sample path $(X_t)_{t=0}^n$ of the Markov chain of interest, $\forall\, k \in [0,3]$ we define

$$N_k^{'} := \begin{cases} |\{t \in [n-1] : X_t \geq 3\}| & \text{if } k = 3, \\ |\{t \in [n-1] : X_t = k\}| & \text{otherwise.} \end{cases}$$

Now consider the following numerical experiment. For a given variance target θ, we generated 1000 random sample paths for the Markov chain of interest. Importantly, we did not fix the sample path length n but instead stopped and moved on to the next sample path as soon as the following two stopping conditions were both met:

- $\forall\, k \in [0,3],\ N_k^{'} \geq 1,$
- $\frac{1}{n^2}\sum_{k=0}^{3} N_k^{'}\gamma_k \leq \theta,$

whereby $\{\gamma_k\}_{k\in[0,3]}$ was computed based on transition matrix $\mathcal{A}_d$ with $d = 3$. This yields a sample of 1000 values of $\hat{\pi}(g)$ with sample variance of roughly θ. We repeated this process for various values of θ. The respective realized sample mean, sample variance, and median sample path length are reported in the below table.

θ	Mean $\hat{\pi}(g)$	Sample $\mathbf{Var}\big(\hat{\pi}(g)\big)$	Median n
5×10^{-6}	0.01223	4.912×10^{-6}	3174
4×10^{-6}	0.01211	3.849×10^{-6}	3961
3×10^{-6}	0.01200	2.839×10^{-6}	5279
2×10^{-6}	0.01198	1.903×10^{-6}	7918
1×10^{-6}	0.01188	1.003×10^{-6}	15860
9×10^{-7}	0.01188	9.275×10^{-7}	17616
8×10^{-7}	0.01189	7.827×10^{-7}	19835
7×10^{-7}	0.01189	6.810×10^{-7}	22672
6×10^{-7}	0.01189	6.341×10^{-7}	26460

In all cases, the realized variance is very close to θ. This suggests that the stopping rule is effective despite the approximations that were made to implement it.

4.3 Approximation for Continuous $\mathcal{X}$

The idea of ”discretizing” a continuous state-space for the purposes of MCMC convergence assessment is not new. An approach to this effect was proposed in Guihenneuc-Jouyaux & Robert, 1998. The authors’ approach to discretization involves the identification of ”small sets” and renewal times (see Guihenneuc-Jouyaux & Robert, 1998 for more details). While the authors mention that there are theoretical assurances that their approach works in most MCMC setups, they admit that the need to determine ”small sets” is a difficulty. Based on the examples they give, their approach indeed seems hard to generalize to a broad range of MCMC setups. For our purposes, we suggest a simpler approach.

Let $\mathcal{X}^{'} := \{x \in \mathcal{X} : f(x) > 0\}$ and assume that $\mathcal{X}^{'}$ is bounded. Now consider a partition of $\mathcal{X}^{'}$ into d mutually disjoint and collectively exhaustive subsets $\{\mathcal{X}_i : i \in [d]\}$. More formally, $\{\mathcal{X}_i : i \in [d]\}$ satisfies:

- $\bigcup_{i\in[d]} \mathcal{X}_i = \mathcal{X}^{'}$
- $\mathcal{X}_i \bigcap \mathcal{X}_j = \emptyset$ for $i \neq j$.

$\forall\, t \in [n]$, let $W_t := \sum_{j=1}^{d} j\mathbb{1}(X_t \in \mathcal{X}_j)$. First, it is important to notice that such a partition may not be unique. Second, the sequence $(W_t)_{t\in[n]}$ does not necessarily possess the Markov property.

However, such a partition allows us to construct a finite and discrete state-space Markov chain such that conditional on $\{N_k := |\{t \in [n-1] : X_t \in \mathcal{X}_k\}|\}$, the sample estimate of the discrete state-space Markov chain is close to $\hat{\pi}(g)$. More specifically, consider a discrete and finite Markov chain which takes values over $\{\hat{x}_k : k \in [d]\}$, where $\forall\ k \in [d]\ \hat{x}_k$ is a fixed element of $\mathcal{X}_k$. Let $\{W_t\}_{t=0}^{n}$ be a sample path of this Markov chain, and let $\hat{\pi}_d(g) := \frac{1}{n}\sum_{t=1}^{n} g(W_t)$ and $\hat{N}_k := |\{t \in [n-1] : W_t = \hat{x}_k\}|$, $k \in [d]$. The following lemma will allow us to demonstrate why this partition of $\mathcal{X}$ is useful.

Lemma 4.1. *If $\hat{N}_k = N_k\ \forall\ k \in [d]$, then*

$$\mid \hat{\pi}_d(g) - \hat{\pi}(g) \mid\ \leq \max_{k\in[d]} \max_{X_t\in\mathcal{X}k} \mid g(X_t) - g(\hat{x}_k) \mid .$$

Proof.

$$\begin{aligned}
&\mid \hat{\pi}_d(g) - \hat{\pi}(g) \mid \\
&= \Big| \frac{1}{n}\sum_{t=1}^{n} g(W_t) - \frac{1}{n}\sum_{t=1}^{n} g(X_t) \Big| \\
&= \Big| \frac{1}{n}\sum_{k=1}^{d} \sum_{W_t=\hat{x}_k} g(W_t) - \frac{1}{n}\sum_{k=1}^{d} \sum_{X_t\in\mathcal{X}_k} g(X_t) \Big| \\
&= \Big| \frac{1}{n}\sum_{k=1}^{d} \Big(N_k g(\hat{x}_k) - \sum_{X_t\in\mathcal{X}_k} g(X_t) \Big) \Big| \\
&= \Big| \sum_{k=1}^{d} \frac{N_k}{n} \Big(g(\hat{x}_k) - \frac{1}{N_k} \sum_{X_t\in\mathcal{X}_k} g(X_t) \Big) \Big| \\
&\leq \sum_{k=1}^{d} \frac{N_k}{n} \max_{X_t\in\mathcal{X}_k} \mid g(X_t) - g(\hat{x}_k) \mid \\
&\leq \max_{k\in[d]} \max_{X_t\in\mathcal{X}_k} \mid g(X_t) - g(\hat{x}_k) \mid
\end{aligned}$$

□

It follows from this lemma that as long as $\{g(x) : x \in \mathcal{X}'\}$ is bounded, $\mid \hat{\pi}_d(g) - \hat{\pi}(g) \mid$ is also bounded. Moreover, since all else equal the subsets $\mathcal{X}_1, ..., \mathcal{X}_d$ should get smaller as d increases, this lemma tells us that $\mid \hat{\pi}_d(g) - \hat{\pi}(g) \mid$ should converge towards 0 as d increases (not necessarily monotonically though). In order to more formally establish convergence towards 0, we need to impose additional restrictions on $\mathcal{X}_1, ..., \mathcal{X}_d$, since there may be more than one way to construct such subsets.

Based on the intuition that if $\hat{N}_k = N_k\ \forall\ k \in [d]$, $\hat{\pi}_d(g)$ is close to $\hat{\pi}(g)$, a possible approach is to make the key approximation

$$\mathbf{Var}\Big(\hat{\pi}(g) \mid \{N_k\}_{k\in[d]}\Big) \approx \mathbf{Var}\Big(\hat{\pi}_d(g) \mid \hat{N}_k = N_k,\ \forall k \in [d]\Big), \tag{2}$$

whereby $\mathbf{Var}\Big(\hat{\pi}_d(g) \mid \{\hat{N}_k\}_{k\in[d]}\Big)$ is computed using corollary 3.2 based on a well chosen discrete Markov chain. We illustrate this strategy with an example in the next subsection.

A key question is how to chose the discrete Markov chain and compute its transition matrix. To this effect, let $q : \mathcal{X}^2 \to \mathbb{R}$ be the transition function for the continuous state-space Markov chain. Notice that when this Markov chain is stationary, $\forall\ (i, j) \in [d]^2 : i \neq j$,

$$\begin{aligned}
&\mathbf{P}\Big(X_{t+1} \in \mathcal{X}_j \mid X_t \in \mathcal{X}_i\Big) \\
&= \int_{\mathcal{X}_i} \mathbf{P}\Big(X_t = x\Big) \mathbf{P}\Big(X_{t+1} \in \mathcal{X}_j \mid X_t = x\Big)\, dx \\
&= \int_{\mathcal{X}_i} \pi(x) \int_{\mathcal{X}_j} q(x, y)\, dy\, dx.
\end{aligned}$$

For the purposes of the example in the next section, we hence considered a finite, discrete state-space Markov chain with transition matrix $P = [p_{ij}]$ where $\forall\ (i, j) \in [d]^2$

$$p_{ij} = \begin{cases} \int_{\mathcal{X}_i} \pi(x) \int_{\mathcal{X}_j} q(x,y)\, dy\, dx & \text{if } j \neq i, \\ 1 - \sum_{j \neq i} p_{ij} & \text{if } j = i. \end{cases} \tag{3}$$

In the MCMC context, $\pi(\cdot)$ is proportional to a known function $f : \mathcal{X} \to \mathbb{R}$ and $q(x, y)$ depends on the ratio $\frac{f(y)}{f(x)}$ and a carefully chosen *proposal density* function. We therefore argue that it should generally be possible to numerically estimate this double integral in the MCMC context.

4.4 Example 2: Sixmodal Target Distribution

Consider the sixmodal target distribution example discussed in Leman, Chen, & Lavine 2009 and Roy, 2020, where the distribution we seek to sample from is as follows:

$$\pi(x,y) \propto \exp\left(\frac{-x^2}{2}\right) \exp\left(-\frac{\left((\csc y)^5 - x\right)^2}{2}\right),$$

where $x \in (-3, 3)$ and $y \in (-10, 10)$. The contour plot of this distribution (known up to the normalizing constant) is given in figure 1 (which was generated using the R library ggplot2).

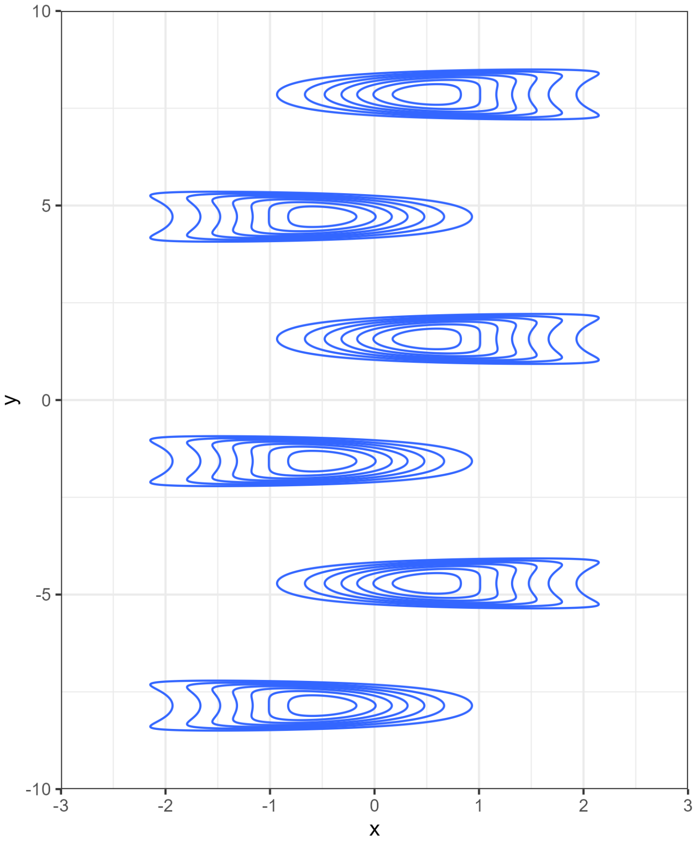

Figure 1. Contour plot of the target distribution

Roy, 2020 specifically discusses this example as a case where some common MCMC convergence diagnostic tools may mistakenly detect convergence as the Markov chain gets stuck in one of the modes.

For the purposes of demonstrating how corollaries 3.2 and 3.3 can help calibrate $\mathbf{Var}\left(\hat{\pi}(g)\right)$ and estimate $\lim_{n\to\infty} n\mathbf{Var}\left(\hat{\pi}(g)\right)$, assume that we are interested in estimating $\pi(g)$ whereby $g(x,y) = \sqrt{x^2 + y^2}$. Furthermore, assume that we seek to estimate $\pi(g)$ via a Metropolis-Hastings sampler with a uniform proposal density function over $\mathcal{X} = (-3,3) \times (-10,10)$.

To this effect, we started by partitioning the state-space $\mathcal{X}$ into subsets, as was suggested in section 4.3. Specifically, we partitioned $\mathcal{X}$ into 24 equally sized subsets as illustrated below in Figure 2 (which was generated using the R library ggplot2).

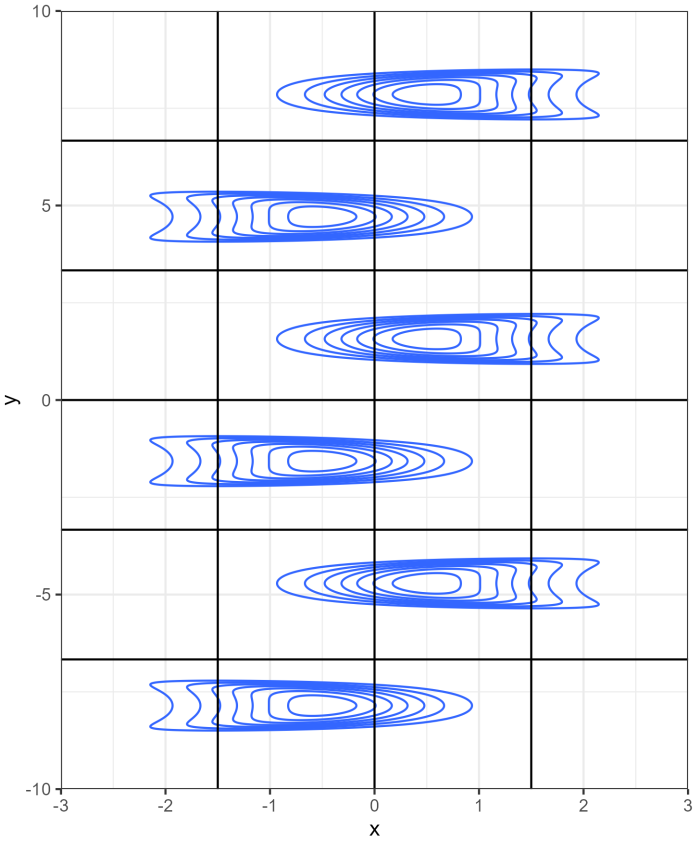

Figure 2. partition of the state-space into 24 equally sized subsets

As was explained in section 4.3, this partition allows us to introduce a discrete state-space Markov chain which approximates the dynamics of the Markov chain of interest, with the intention of using approximation (2).

To this effect, we computed the transition matrix of this discrete, finite state-space Markov chain as described in section 4.3. Specifically, we estimated the transition probabilities in equation (3) using double Riemann sums over a grid of points spanning $\mathcal{X}$ in steps of 0.1.

Taking a similar approach to section 4.2 (i.e., example 1), we then computed $\{\gamma_k\}_{k\in[24]}$ and $\{\Pi_k\}_{k\in[24]}$ based on this discrete, finite state-space Markov chain. As a first step, this yields an estimate of $\lim_{n\to\infty} n\mathbf{Var}\left(\hat{\pi}(g)\right)$ which is $\sum_{k=1}^{24} \gamma_k \Pi_k = 90.32782$.

As in section 4.2, we also used the set $\{\gamma_k\}_{k\in[24]}$ to generate samples of estimates with a given level of target variance θ. To this end, for any of the subsets $\mathcal{X}_k$, $k \in [24]$, we define $\{N_k := |\{t \in [n-1] : X_t \in \mathcal{X}_k\}|\}$, whereby $(X_t)_{t=0}^n$ is a sample path of the Markov chain of interest. This yields the stopping rule with the following two conditions:

- $\forall\, k \in [24]$, $N_k \geq 1$,
- $\frac{1}{n^2}\sum_{k=1}^{24} N_k \gamma_k \leq \theta$.

For various values of θ, we generated samples of 1000 values of $\hat{\pi}(g)$. The respective realized sample mean, sample variance, and median sample path length are reported in the below table.

θ	Mean $\hat{\pi}(g)$	Sample $\mathbf{Var}\big(\hat{\pi}(g)\big)$	Median n
10×10^{-4}	4.85649	10.48028×10^{-4}	90280
9×10^{-4}	4.85862	9.35073×10^{-4}	100341.5
8×10^{-4}	4.85851	9.07333×10^{-4}	112891
7×10^{-4}	4.85961	7.46519×10^{-4}	128986
6×10^{-4}	4.85700	6.15114×10^{-4}	150476.5
5×10^{-4}	4.85838	5.20330×10^{-4}	180559.5

Once again, we observe that the realized variance is generally quite close to θ.

5. Conclusion

Dealing with estimation error remains one of the key challenges associated with MCMC algorithms. This paper proposes new tools to address this challenge. At the cost of some initial exploration of the state-space, our approach should allow practitioners to both estimate and calibrate estimation error in a broad range of MCMC contexts.

An essential aspect requiring further refinement is with regards to how our main result can be extended to state-spaces that are continuous and/or unbounded. While our suggested approaches in section 4 should work in a broad variety of contexts, there is substantial room to formalize these approaches. Increased formality may become necessary if our results are to be applied to more complex state-spaces than those exemplified in section 4.

References

Binder, K., & Heermann, D. (2010). *Monte Carlo Simulation in Statistical Physics.* (5th ed.) Springer, Berlin.

Gamerman, D., & Lopes, H. (2006). *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference, Second Edition.* Chapman and Hall/CRC, New York.

Gelman, A., & Rubin, D. (1996). Markov chain Monte Carlo methods in biostatistics. *Statistical Methods in Medical Research, 5*(4), 339–355. https://doi.org/10.1177/096228029600500402

Guihenneuc-Jouyaux, C., & Robert, C. (1998). Discretization of Continuous Markov Chains and Markov Chain Monte Carlo Convergence Assessment. *Journal of the American Statistical Association, 93*(443), 1055–1067. https://doi.org/10.1080/01621459.1998.10473767

Hang Jian, Y., Lui, T., Lou, Z., Rosenthal, J., Shangguan, S., & Wu, Z. (2022). Markov Chain Confidence Intervals and Biases. *International Journal of Statistics and Probability, 11*(1). https://doi.org/10.5539/ijsp.v11n1p29

Hunter, J. (2005). Stationary distributions and mean first passage times of perturbed Markov chains. *Linear Algebra and its Applications, 410,* 217–243. https://doi.org/10.1016/j.laa.2005.08.005

Jones, G. (2004). On the Markov chain central limit theorem. *Probab. Surveys, 1*, 299–320. https://doi.org/10.1214/154957804100000051

Łatuszyski, K., Miasojedow, B., & Niemiro, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli, 19*(5A), 2033–2066. https://doi.org/10.3150/12-BEJ442

Leman S., Chen Y., & Lavine M. (2009). The multiset sampler. *J. Am. Stat. Assoc., 104,* 1029–1041. https://doi.org/10.1198/jasa.2009.tm08047

Levin, D., Peres, Y., & Wilmer, E. (2009). *Markov Chains and Mixing Times.* Amer. Math. Soc., Providence, RI.

Meyer, C. (1975). The Role of the Group g-inverse in the Theory of Finite Markov Chains. *SIAM Review, 17*, 443–464. https://doi.org/10.1137/1017044

Robert, C., & Casella, G. (2011). A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science, 26*, 102-115. https://doi.org/10.1214/10-STS351

Roy, V. (2020). Convergence Diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application, 7*, 387-412. https://doi.org/10.1146/annurev-statistics-031219-041300

Spitzner, D., & Boucher, T. (2007). Asymptotic variance of functionals of discrete-time Markov chains via the Drazin inverse. *Electron. Commun. Probab., 12*, 120–133. https://doi.org/10.1214/ECP.v12-1262

Trevezas, S., & Limnios, N. (2009). Variance estimation in the central limit theorem for Markov chains. *Journal of Statistical Planning and Inference, 139*, 2242–2253. https://doi.org/10.1016/j.jspi.2008.10.020

Rudas, T. (2018). *Lectures on Categorical Data Analysis.* Springer, London.

Acknowledgments

Not applicable

Authors contributions

Yann Vestring, PhD candidate, was responsible for most of the research, including the design and execution of the numerical experiments. Dr. Javad Tavakoli, as Yann Vestring's PhD supervisor, provided guidance and valuable feedback all throughout the process.

Funding

This research was supported by NSERC Discovery Grant 256233.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Canadian Center of Science and Education. The journals policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Open access

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.