(ε − α)–MCMC–Approximation under Drift Condition

Krzysztof Łatuszyński*  Wojciech Niemiro†

October 10, 2006

Abstract

We assume a drift condition towards a small set and bound the mean square error of estimators obtained by taking averages along a single trajectory of a Markov chain Monte Carlo algorithm. We use these bounds to determine the length of the trajectory and burn in time that ensures desired precision of estimation with given probability. Let \( I \) be the value of interest and \( \hat{I} \) its MCMC estimate. Precisely, our lower bounds for the length of the trajectory and burn in time ensure that

\[
P( |\hat{I} - I| \leq \varepsilon ) \geq 1 - \alpha
\]

and depend only and explicitly on drift parameters, \( \varepsilon \) and \( \alpha \).

1 Introduction

An essential part of many problems in Bayesian inference is the computation of analytically intractable integral

\[
I = \int_{\mathcal{X}} f(x)\pi(x)dx,
\]

where \( f(x) \) is the target function of interest, \( \mathcal{X} \) is often a region in high-dimensional space and the probability distribution \( \pi \) over \( \mathcal{X} \) is usually known up to a normalizing constant and direct simulation from \( \pi \) is not feasible (see e.g. [4], [9]). The common approach to this problem is to simulate an ergodic Markov chain \((X_n)_{n \geq 0}\), using a transition kernel \( P \), with stationary distribution \( \pi \), which ensures the convergence in distribution of \( X_n \) to a random variable from \( \pi \). Thus, for a "large enough" \( n_0 \), \( X_n \) for \( n \geq n_0 \) can be considered as distributed as \( \pi \). Since a simple and powerful algorithm has been introduced in 1953 by Metropolis et al. in a very seminal paper [10], various sampling schemes and approximation strategies have been developed and analyzed ([12], [9], [4]) and the method is referred to as Markov chain Monte Carlo (MCMC).

There are two classical strategies to build an MCMC estimate of \( I \).

---

*Institute of Econometrics, Warsaw School of Economics; e-mail: latuch@gmail.com
†Faculty of Mathematics and Computer Science, Nicolaus Copernicus University; e-mail: wniem@mat.uni.torun.pl
• Use average along a single trajectory of the underlying Markov chain and discard the initial part to reduce bias. In this case the estimate is of the form
\[
\hat{I}_{t,n} = \frac{1}{n} \sum_{i=t}^{t+n-1} f(X_i)
\]
and $t$ is called the burn in time.

• Use average over final states of multiple independent runs of the chain i.e. for an estimate take
\[
\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} f(X_t^{(k)}),
\]
where $k$ numbers the independent runs of the chain.

The first strategy is believed to be more efficient then the latter (see [12] for discrete space analysis) and is usually the practitioners choice. Yet it is harder to analyze. The goal of this paper is to derive lower bounds for $n$ and $t$ in (1) that ensure the following condition of $(\varepsilon, \alpha)-$approximation:
\[
P(|\hat{I}_{t,n} - I| \leq \varepsilon) \geq 1 - \alpha,
\]
where $\varepsilon$ is the precision of estimation and $1 - \alpha$, the confidence level.

Results of this or related type have been obtained for discrete state space $\mathcal{X}$ and bounded target function $f$ by Aldous in [1], Gillman in [5] and recently by León and Perron in [8]. Niemiro and Pokarowski in [12] give results for relative precision estimation. For uniformly ergodic chains on continuous state space $\mathcal{X}$ and bounded function $f$ Hoeffding type inequalities are available (due to Glynn and Ormonait in [6], and an improved bound due to Meyn et al. in [7]) and can easily lead to the desired $(\varepsilon - \alpha)-$approximation. To our best knowledge there are no explicit bounds for $n$ and $t$ in more general settings, especially when $f$ is not bounded and the chain is not uniformly ergodic. A good example of the state of the art approach to dealing with this problem is presented by Jones at al. in the recent paper [3]. They suggest procedures of estimating variance of the asymptotic normal distribution for geometrically ergodic split chains and thus under the additional assumption of $E\pi |f|^{2+\delta} < \infty$ for some $\delta > 0$, they obtain a version of a sequential asymptotically valid $(\varepsilon, \alpha)-$approximation via MCMC method. The asymptotics refers to $\varepsilon$, so precisely speaking their result amounts to $P(|I - S_n| \leq \varepsilon) \rightarrow 1 - \alpha$ as $\varepsilon \rightarrow 0$, where $S_n = 1/\bar{n} \sum_{i=0}^{\bar{n}} f(X_i)$ and $\bar{n} = \bar{n}(X_0, X_1, \ldots)$ is a random stopping time.

Our approach is to assume a version of the well known drift condition towards a small set (Assumption 2.1) and give explicit lower bounds on $n$ and $t$ in terms of drift parameters defined in Assumption 2.1 and approximation parameters defined in (2).

The rest of the paper is organized as follows. In section 2 we introduce the drift condition assumption and preliminary results. In section 3 we present our main results and discuss a possible trade-off between drift parameters. In section 4 we proceed with proofs.

1.1 Notation and Basic Definitions

Throughout this paper, $\pi$ represents the probability measure of interest, defined on some measurable state space $(\mathcal{X}, \mathcal{F})$ and $f : \mathcal{X} \rightarrow \mathbb{R}$, the target function. Let $(X_n)_{n \geq 0}$ be a time homogeneous Markov chain on $(\mathcal{X}, \mathcal{F})$, by $\pi_0$ denote its
initial distribution and by $P$ denote its transition kernel. Let $I = \int_X f(x)\pi(dx)$ be the value of interest and $\hat{I}_{t,n} = \frac{1}{n} \sum_{i=t}^{t+n-1} f(X_i)$ its MCMC estimate along one walk.

For a probability measure $\mu$ and a transition kernel $Q$, by $\mu Q$ we denote a probability measure defined by $\mu Q(\cdot) := \int_X Q(x,\cdot)\mu(dx)$, furthermore if $g$ is a real-valued function on $X$ let $Qg(x) := \int_X g(y)Q(x,dy)$ and $\mu g := \int_X g(x)\mu(dx)$. We will also use $E_{\mu}g$ for $\mu g$, especially if $\mu = \delta_x$ we will write $E_xg$. For transition kernels $Q_1$ and $Q_2$, $Q_1 Q_2$ is also a transition kernel defined by $Q_1 Q_2(x,\cdot) := \int_X Q_2(y,\cdot)Q_1(x,dy)$.

Let $V : X \to [1,\infty)$ be a measurable function. For measurable function $g : X \to R$ define its $V$-norm as

$$|g|_V := \sup_{x \in X} \frac{|g(x)|}{V(x)}.$$  

To evaluate the distance between two probability measures $\mu_1$ and $\mu_2$ we use the $V$-norm distance, defined for probability measures $\mu_1$ and $\mu_2$ as

$$\|\mu_1 - \mu_2\|_V := \sup_{|g| \leq V} |\mu_1 g - \mu_2 g|.$$  

Finally for two transition kernels $Q_1$ and $Q_2$ the $V$-norm distance between $Q_1$ and $Q_2$ is defined by

$$|||Q_1 - Q_2|||_V := \|||Q_1(x,\cdot) - Q_2(x,\cdot)|||_V = \sup_{x \in X} \frac{|||Q_1(x,\cdot) - Q_2(x,\cdot)|||_V}{V(x)}.$$  

For a probability distribution $\mu$, define a transition kernel $\mu(x,\cdot) := \mu(\cdot)$, to allow for writing $|||Q - \mu|||_V$ and $|||\mu_1 - \mu_2|||_V$. Define also

$$B_V := \{ f : X \to R, |f|_V < \infty \}.$$  

Now if $|||Q_1 - Q_2|||_V < \infty$, then $Q_1 - Q_2$ is a bounded operator from $B_V$ to itself, and $|||Q_1 - Q_2|||_V$ is its operator norm. See [11] for details.

2 A Drift Condition and Preliminary Lemmas

We analyze the MCMC estimation along a single trajectory under the following assumption of a drift condition towards a small set.

**Assumption 2.1.** (A.1) Small set. There exist $C \subseteq X$, $\beta > 0$ and a probability measure $\nu$ on $X$, such that for all $x \in C$ and $A \subseteq X$

$$P(x, A) \geq \beta \nu(A).$$

(A.2) Drift. There exist a function $V : X \to [1,\infty)$ and constants $\lambda < 1$ and $K < \infty$ satisfying

$$PV(x) \leq \begin{cases} \lambda V(x), & \text{if } x \notin C, \\ K, & \text{if } x \in C. \end{cases}$$

(A.3) Aperiodicity. There exists $\tilde{\beta} > 0$ such that $\beta \nu(C) \geq \tilde{\beta}$.  


In the sequel we refer to $\beta, V(x), \lambda, K, \beta$ as drift parameters.

This type of drift condition is often assumed and widely discussed in Markov chains literature. Substantial effort has been devoted to establishing convergence rates for Markov chains under Assumption 2.1 or related assumptions. For discussion of various drift conditions and their relation see Meyn and Tweedie [11]. In the sequel we make use of recent convergence bounds obtained by Baxendale in [2].

Theorem 2.2 (Baxendale [2]). Under Assumption 2.1 $(X)_{n\geq 0}$ has a unique stationary distribution $\pi$ and $\pi V < \infty$. Moreover, there exists $\rho < 1$ depending only and explicitly on $\beta, \beta, \lambda$ and $K$ such that whenever $\rho < \gamma < 1$ there exists $M < \infty$ depending only and explicitly on $\gamma, \beta, \beta, \lambda$ and $K$ such that for all $n \geq 0$

$$||P^n - \pi||_V \leq M\gamma^n.$$  

Formulas for $\rho$ and $M$ are given in the original paper [2]. Since there are different formulas for general operators, self adjoint operators and self adjoint positive operators in both atomic and nonatomic case, we encourage the reader to look them up in [2], rather then add several superfluous pages to our paper.

To our knowledge the above-mentioned theorem gives the best available explicit constants.

Corollary 2.3. Under Assumption 2.1

$$||\pi_0 P^n - \pi||_V \leq \min\{\pi_0 V, ||\pi_0 - \pi||_V\} M\gamma^n,$$

where $M$ and $\gamma$ are such as in Theorem 2.2.

Proof. From Theorem 2.2 we have $||P^n(x, \cdot) - \pi(\cdot)||_V \leq M\gamma^n V(x)$, which yields

$$\pi_0 VM\gamma^n \geq \int_X ||P^n(x, \cdot) - \pi(\cdot)||_V \pi_0(dx) \geq \sup_{|g| \leq V} \int_X |P^n(x, \cdot)g - \pi g| \pi_0(dx) \geq \sup_{|g| \leq V} ||\pi_0 P^n g - \pi g|| = ||\pi_0 P^n - \pi||_V.$$

Now let $b_V = \inf_{x \in X} V(x)$. Since $|||\cdot||_V$ is an operator norm and $\pi$ is invariant for $P$, we have

$$||\pi_0 P^n - \pi||_V = b_V ||||\pi_0 P^n - \pi||_V = b_V |||(\pi_0 - \pi)(P^n - \pi)||_V \geq b_V |||\pi_0 - \pi||_V |||P^n - \pi||_V = ||\pi_0 - \pi||_V |||P^n - \pi||_V \geq \pi_0 - \pi||_V M\gamma^n.$$

Now we focus on the following simple but useful observation.

Lemma 2.4. If for a Markov chain $(X_n)_{n\geq 0}$ on $X$ with transition kernel $P$ Assumption 2.1 holds with parameters $\beta, V(x), \lambda, K, \beta$, it holds also with $\beta_r := \beta, V_r(x) := V(x)^{1/r}, \lambda_r := \lambda^{1/r}, K_r := K^{1/r}, \beta_r := \beta$ for every $r > 1$.

Proof. It is enough to check (A.2). For $x \notin C$ by Jensen inequality we have

$$\lambda V(x) \geq \int_X V(y) P(x, dy) \geq \left( \int_X V(y)^{1/r} P(x, dy) \right)^r$$

and hence $PV(x)^{1/r} \leq \lambda^{1/r} V(x)^{1/r}$, as claimed. Similarly for $x \in C$ we obtain $PV(x)^{1/r} \leq K^{1/r}$.

\[4\]
Bonds on

Under Assumption 2.1 we have

Corollary 2.5.

for an appropriate choice of

By

MSE

Main Result

parameters defined in Lemma 2.4.

drift condition with

Theorem 3.1.

Assume the Drift Condition 2.1 holds and

(2) and are also of independent interest.

where

and

MSE

(3) is emphasized below as a corollary.

The foregoing bound is easy to interpret: \( V |f|^2 \) should be close to \( Var \pi f \) for an appropriate choice of \( V \), moreover \( 2M_2/(1-\gamma_2) \) corresponds to the autocorrelation of the chain and the last term \( M \min \{ \pi_0 V, \| \pi_0 - \pi \| V \} / n(1-\gamma) \) is the price for nonstationarity of the initial distribution.

Theorem 3.1 is explicitly stated for \( \tilde{I}_{0,n} \), but the structure of the bound is flexible enough to cover most typical settings as indicated below.

Corollary 3.3.

In the setting of Theorem 3.1, for any appropriate choice of \( V \), the first two terms are of independent interest.

Theorem 3.1 is explicitly stated for \( \tilde{I}_{0,n} \), but the structure of the bound is flexible enough to cover most typical settings as indicated below.

Corollary 3.3.

In the setting of Theorem 3.1, for any appropriate choice of \( V \), the first two terms are of independent interest.

Theorem 3.1 is explicitly stated for \( \tilde{I}_{0,n} \), but the structure of the bound is flexible enough to cover most typical settings as indicated below.

Corollary 3.3.

In the setting of Theorem 3.1, for any appropriate choice of \( V \), the first two terms are of independent interest.
Bound (5) corresponds to the situation when a perfect sampler is available. For deterministic start without burn in and with burn in (6) and (7) should be applied respectively. 

$(\varepsilon - \alpha)-$approximation is an easy corollary of $MSE$ bonds by the Chebyshev inequality.

**Theorem 3.4** ($(\varepsilon - \alpha)-$approximation). Let

$$b = \frac{\pi V||f_c||_{V'}^2}{\varepsilon^2\alpha} \left(1 + \frac{2M_r}{1 - \gamma_r}\right), \quad \tilde{c} = \frac{4M^2V(x)||f_c||_{V'}^2}{\varepsilon^2\alpha(1 - \gamma)} \left(1 + \frac{2M_r}{1 - \gamma_r}\right),$$

$$c = \frac{M \min\{\pi_0V, \pi_0 - \pi\}||f_c||_{V'}^2}{\varepsilon^2\alpha(1 - \gamma)} \left(1 + \frac{2M_r}{1 - \gamma_r}\right),$$

$$n(t) = \frac{b + \sqrt{b^2 + 4c(t)}}{2}, \quad c(t) = \frac{M^2\gamma^4V(x)||f_c||_{V'}^2}{\varepsilon^2\alpha(1 - \gamma)} \left(1 + \frac{2M_r}{1 - \gamma_r}\right).$$

Then under Assumption 2.1,

$$P(|\tilde{I}_{0,n} - I| \leq \varepsilon) \geq 1 - \alpha, \quad \text{if} \quad X_0 \sim \pi_0, \quad n \geq \frac{b + \sqrt{b^2 + 4c}}{2}. \quad (8)$$

$$P(|\tilde{I}_n - I| \leq \varepsilon) \geq 1 - \alpha, \quad \text{if} \quad \begin{cases} X_0 \sim \delta_x, \\ t \geq (\ln \gamma)^{-1} \ln \left(\frac{16 + \sqrt{b^2 + 4c(t)}}{2\ln^2 \gamma}\right), \\ n \geq n(t). \end{cases} \quad (9)$$

And the above bounds in (9) give the minimal length of the trajectory $(t + n)$ resulting from (7).

## 4 Proofs

We now proceed to prove results from Section 3.

**Proof of Theorem 3.1.** Without loss of generality consider $f_c = f - \pi f$ instead of $f$ and assume $||f_c||_{V'} = 1$. Note that in this setting $|f_c|^2 \leq \pi V$, $MSE(\hat{I}_{0,n}) = E_{\pi_0}(\hat{I}_{0,n})^2$, and also for every $r \in [\frac{p}{2\gamma}, p],$

$$|f_c|_{V^1/r} \leq ||f_c|^p|_{V^1/r} = 1 \quad \text{and} \quad |f_c|_{V^1-1/r} \leq ||f_c|^p-p/r|_{V^1-1/r} = 1.$$ 

Obviously

$$nMSE(\hat{I}_{0,n}) = \frac{1}{n} \sum_{i=0}^{n-1} E_{\pi_0}f_c(X_i)^2 + \frac{2}{n} \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} E_{\pi_0}f_c(X_i)f_c(X_j). \quad (10)$$

We start with a bound for the first term of the right hand side of (10). Since $f_c^2(x) \leq V(x)$, we use Corollary 2.3 for $f_c^2$. Let $C = \min\{\pi_0V, \pi_0 - \pi\}$ and proceed

$$\frac{1}{n} \sum_{i=0}^{n-1} E_{\pi_0}f_c(X_i)^2 = \frac{1}{n} \sum_{i=0}^{n-1} \pi_0 f_c^2 \leq \pi f_c^2 + \frac{1}{n} \sum_{i=0}^{n-1} CM \gamma^2 \leq \pi V + \frac{CM}{n(1 - \gamma)}. \quad (11)$$
To bound the second term of the right hand side of (10) note that $|f_c| \leq V^{1/r}$ and use Corollary 2.5.

$$\frac{2}{n} \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} E_{\epsilon_n} f_c(X_i) f_c(X_j) = \frac{2}{n} \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} \pi_0 \left( P^i \left( f_c P^{j-i} f_c \right) \right)$$

$$\leq \frac{2}{n} \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} \pi_0 \left( P^i \left( |f_c| P^{j-i} f_c \right) \right)$$

$$\leq \frac{2M_r}{n} \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} \gamma_r^{j-i} \pi_0 \left( P^i \left( |f_c| V^{1/r} \right) \right)$$

$$\leq \frac{2M_r}{n(1-\gamma_r)} \sum_{i=0}^{n-2} \pi_0 \left( P^i \left( |f_c| V^{1/r} \right) \right) = \spadesuit$$

Now recall since $|f_c| \leq V^{1/r}$ and $|f_c| \leq V^{1-1/r}$, also $|f_c V^{1/r}| \leq V$ and use Corollary 2.3 for $|f_c V^{1/r}|$.

$$\spadesuit \leq \frac{2M_r}{n(1-\gamma_r)} \sum_{i=0}^{n-2} \left( \pi \left( |f_c| V^{1/r} \right) + CM \gamma^i \right) \leq \frac{2M_r}{1-\gamma_r} \left( \pi V + \frac{CM}{n(1-\gamma)} \right). \quad (12)$$

Combine (11) and (12) to obtain

$$MSE(\hat{I}_{0,n}) \leq \frac{|f_c|^2 V}{n} \left( 1 + \frac{2M_r}{1-\gamma_r} \right) \left( \pi V + \frac{CM}{n(1-\gamma)} \right).$$

\[ \square \]

**Proof of Corollary 3.3.** Only (7) needs a proof. Note $X_t \sim \delta_x P^t$. Now use Theorem 2.2 to see $\|\delta_x P^t - \pi\|_V \leq M \gamma^t V(x)$, and apply Theorem 3.1 with $\pi_0 = \delta_x P^t$.

\[ \square \]

**Proof of Theorem 3.4.** From the Chebyshev’s inequality we get

$$P(|\hat{I}_{t,n} - I| \leq \varepsilon) = 1 - P(|\hat{I}_{t,n} - I| \geq \varepsilon) \geq 1 - \frac{MSE(\hat{I}_{t,n})}{\varepsilon^2} \geq 1 - \alpha \quad \text{if} \quad MSE(\hat{I}_{t,n}) \leq \varepsilon^2 \alpha. \quad (13)$$

To prove (8) set $C = \min\{\pi_0 V, \|\pi_0 - \pi\|_V\}$, and combine (13) with (3) to get

$$n^2 - n \frac{\pi V |f_c|^2 p^{2/p} V}{\varepsilon^2 \alpha} \left( 1 + \frac{2M_r}{1-\gamma_r} \right) \frac{MC |f_c|^2 p^{2/p} V}{\varepsilon^2 \alpha (1-\gamma)} \left( 1 + \frac{2M_r}{1-\gamma_r} \right) \geq 0,$$

and hence

$$n \geq \frac{b + \sqrt{b^2 + 4c}}{2}, \quad \text{where}$$

$$b = \frac{\pi V |f_c|^2 p^{2/p} V}{\varepsilon^2 \alpha} \left( 1 + \frac{2M_r}{1-\gamma_r} \right), \quad c = \frac{MC |f_c|^2 p^{2/p} V}{\varepsilon^2 \alpha (1-\gamma)} \left( 1 + \frac{2M_r}{1-\gamma_r} \right).$$
The only difference in (9) is that now
\[ c = c(t) = \frac{M^2 \gamma^t V(x) ||f_c||^2/p}{\varepsilon^2 \alpha (1 - \gamma)} \left( 1 + \frac{2M_\varepsilon}{1 - \gamma_r} \right), \]
hence it is easy to check the best bound on \( t \) and \( n \) is such that
\[ n \geq n(t) \quad \text{and} \quad t \geq \min\{t \in \mathbb{N} : n'(t) \geq -1\}, \quad \text{where} \quad n(t) = \frac{b + \sqrt{b^2 + 4c(t)}}{2}. \]
Again by standard calculations we get
\[ t \geq (\ln \gamma)^{-1} \ln \left( \frac{16 + \sqrt{4 + b^2 \ln^2 \gamma}}{2\hat{c} \ln^2 \gamma} \right) \quad \text{and} \quad n \geq n(t), \]
where
\[ \hat{c} = \frac{4M^2 V(x)||f_c||^2/p}{\varepsilon^2 \alpha (1 - \gamma)} \left( 1 + \frac{2M_\varepsilon}{1 - \gamma_r} \right). \]
This completes the proof. \( \square \)

References
