Nonasymptotic bounds on the mean square error for MCMC estimates via renewal techniques

Krzysztof Łatuszyński, Błażej Miasojedow and Wojciech Niemiro

Abstract The Nummellin's split chain construction allows to decompose a Markov chain Monte Carlo (MCMC) trajectory into i.i.d. "excursions". Regenerative MCMC algorithms based on this technique use a random number of samples. They have been proposed as a promising alternative to usual fixed length simulation [25, 33, 14]. In this note we derive nonasymptotic bounds on the mean square error (MSE) of regenerative MCMC estimates via techniques of renewal theory and sequential statistics. These results are applied to costruct confidence intervals. We then focus on two cases of particular interest: chains satisfying the Doeblin condition and a geometric drift condition. Available explicit nonasymptotic results are compared for different schemes of MCMC simulation.

1 Introduction

Consider a typical MCMC setting, where π is a probability distribution on $\mathscr{X}$ and $f : \mathscr{X} \to \mathbb{R}$. The objective is to compute (estimate) the integral

Krzysztof Łatuszyński
Department of Statistics
University of Warwick
CV4 7AL, Coventry, UK, e-mail: latuch@gmail.com

Błażej Miasojedow
Institute of Applied Mathematics and Mechanics
University of Warsaw
Banacha 2, 02-097 Warszawa, Poland, e-mail: bmia@mimuw.edu.pl

Wojciech Niemiro
Faculty of Mathematics and Computer Science
Nicolaus Copernicus University
Chopina 12/18, 87-100 Toruń, Poland, e-mail: wniemiro@gmail.com

$$\theta := \pi f = \int_{\mathscr{X}} \pi(\mathrm{d}x) f(x). \tag{1}$$

Assume that direct simulation from π is intractable. Therefore one uses an ergodic Markov chain with transition kernel P and stationary distribution π to sample approximately from π. Numerous computational problems from Bayesian inference, statistical physics or combinatorial enumeration fit into this setting. We refer to [32, 30, 9] for theory and applications of MCMC.

Let $(X_n)_{n\geq 0}$ be the Markov chain in question. Typically one discards an initial part of the trajectory (called burn-in, say of length t) to reduce bias, simulates the chain for n further steps and approximates θ with an ergodic average:

$$\hat{\theta}^{\mathrm{fix}}_{t,n} = \frac{1}{n} \sum_{i=t}^{t+n-1} f(X_i). \tag{2}$$

Fixed numbers t and n are the parameters of the algorithm. Asymptotic validity of (2) is ensured by a Strong Law of Large Numbers and a Central Limit Theorem (CLT). Under appropriate regularity conditions [32, 4], it holds that

$$\sqrt{n}(\hat{\theta}^{\mathrm{fix}}_{t,n} - \theta) \to \mathscr{N}(0, \sigma^2_{\mathrm{as}}(f)), \qquad (n \to \infty), \tag{3}$$

where $\sigma^2_{\mathrm{as}}(f)$ is called the asymptotic variance. In contrast with the asymptotic theory, explicit *nonasymptotic* error bounds for $\hat{\theta}^{\mathrm{fix}}_{t,n}$ appear to be very difficult to derive in practically meaningful problems.

Regenerative simulation offers a way to get around some of the difficulties. The split chain construction introduced in [2, 28] (to be described in Section 2) allows for partitioning the trajectory $(X_n)_{n\geq 0}$ into i.i.d. random tours (excursions) between consecutive regeneration times $T_0, T_1, T_2, \ldots$. Random variables

$$\Xi_k(f) := \sum_{i=T_{k-1}}^{T_k - 1} f(X_i) \tag{4}$$

are i.i.d. for $k = 1, 2, \ldots$ ($\Xi_0(f)$ can have a different distribution). Mykland at al. in [25] suggested a practically relevant recipe for identifying $T_0, T_1, T_2, \ldots$ in simulations (formula (2) in Section 2). This resolves the burn-in problem since one can just ignore the part until the first regeneration T_0. One can also stop the simulation at a regeneration time, say T_r, and simulate r full i.i.d. tours, c.f. Section 4 of [33]. Thus one estimates θ by

$$\hat{\theta}^{\mathrm{reg}}_r := \frac{1}{T_r - T_0} \sum_{i=T_0}^{T_r - 1} f(X_i) = \frac{\sum_{k=1}^r \Xi_k(f)}{\sum_{k=1}^r \tau_k}, \tag{5}$$

where $\tau_k = T_k - T_{k-1} = \Xi_k(1)$ are the lengths of excursions. Number of tours r is fixed and the total simulation effort T_r is random. Since $\hat{\theta}^{\mathrm{reg}}_r$ involves i.i.d. random variables, classical tools seem to be sufficient to analyse its behaviour. Asymptotically, (5) is equivalent to (2) because

$$\sqrt{rm}(\hat{\theta}_r^{\text{reg}} - \theta) \to \mathscr{N}(0, \sigma_{\text{as}}^2(f)), \qquad (r \to \infty), \tag{6}$$

where $m := \mathbb{E}\,\tau_1$. Now $rm = \mathbb{E}\,(T_r - T_0)$, the expected length of trajectory, plays the role of n. However, our attempt at nonasymptotic analysis in Subsection 3.1 reveals unexpected difficulties.

In most practically relevant situations m is unknown. If m is known then instead of (5) one can use an unbiased estimator

$$\tilde{\theta}_r^{\text{unb}} := \frac{1}{rm} \sum_{k=1}^{n} \Xi_k(f), \tag{7}$$

Quite unexpectedly, (7) is *not equivalent* to (5), even in a weak asymptotic sense. The standard CLT for i.i.d. summands yields

$$\sqrt{rm}(\tilde{\theta}_r^{\text{unb}} - \theta) \to \mathscr{N}(0, \sigma_{\text{unb}}^2(f)), \qquad (r \to \infty), \tag{8}$$

where $\sigma_{\text{unb}}^2(f) := \operatorname{Var}\Xi_1(f)/m$ is in general different from $\sigma_{\text{as}}^2(f)$.

We introduce a new regenerative-sequential simulation scheme, for which better nonasymptotic results can be derived. Namely, we fix n and define

$$R(n) := \min\{r : T_r > T_0 + n\}. \tag{9}$$

The estimator is defined as follows.

$$\hat{\theta}_n^{\text{reg-seq}} := \frac{1}{T_{R(n)} - T_0} \sum_{i=T_0}^{T_{R(n)}-1} f(X_i) = \frac{\sum_{k=1}^{R(n)} \Xi_k(f)}{\sum_{k=1}^{R(n)} \tau_k},. \tag{10}$$

We thus generate a random number of tours as well as a random number of samples.

Our approach is based on inequalities for the mean square error,

$$\text{MSE} := \mathbb{E}\,(\hat{\theta} - \theta)^2. \tag{11}$$

Bounds on the MSE can be used to construct fixed precision confidence intervals. The goal is to obtain an estimator $\hat{\theta}$ which satisfies

$$\mathbb{P}(|\hat{\theta} - \theta| \le \varepsilon) \ge 1 - \alpha, \tag{12}$$

for given ε and α. We show that MSE bounds combined with the so called "median trick" lead to confidence intervals which can be better than those obtained by different methods.

The organization of the paper is the following. In Section 2 we recall the split chain construction. Nonasymptotic bounds for regenerative estimators defined by (5), (7) and (10) are derived in Section 3. Derivation of more explicit bounds which involve only computable quantities is deferred to Sections 5 and 6, where we consider classes of chains particularly important in the MCMC context. An analogous analysis of non-regenerative scheme (2) was considered in [21] and (in a different setting and using different methods) in [34].

In Section 4 we discuss the median trick [16, 27]. One runs MCMC repeatedly and computes the median of independent estimates to boost the level of confidence. The resulting confidence intervals are compared with *asymptotic* results based on the CLT.

In Section 5 we consider Doeblin chains, i.e. uniformly ergodic chains that satisfy a one step minorization condition. We compare regenerative estimators (5), (7) and (10). Moreover, we also consider a perfect sampler available for Doeblin chains, c.f. [36, 15]. We show that confidence intervals based on the median trick can outperform those obtained via exponential inequalities for a single run simulation.

In Section 6 we proceed to analyze geometrically ergodic Markov chains, assuming a drift condition towards a small set. We briefly compare regenerative schemes (5) and (10) in this setting (the unbiased estimator (7) cannot be used, because m is unknown).

2 Regenerative Simulation

We describe the setting more precisely. Let $(X_n)_{n\geq 0}$ be a Markov chain with transition kernel P on a Polish space $\mathscr{X}$ with stationary distribution π, i.e. $\pi P = \pi$. Assume P is π-irreducible. The regeneration/split construction of Nummelin [28] and Athreya and Ney [2] rests on the following assumption.

Assumption 2.1 (Small Set) *There exist a Borel set $J \subseteq \mathscr{X}$ of positive π measure, a number $\beta > 0$ and a probability measure ν such that*

$$P(x,\cdot) \geq \beta \mathbb{I}(x \in J)\nu(\cdot).$$

Under Assumption 2.1 we can define a bivariate Markov chain (X_n, Γ_n) on the space $\mathscr{X} \times \{0,1\}$ in the following way. Variable Γ_{n-1} depends only on X_{n-1} via $\mathbb{P}(\Gamma_{n-1} = 1 | X_{n-1} = x) = \beta \mathbb{I}(x \in J)$. The rule of transition from (X_{n-1}, Γ_{n-1}) to X_n is given by

$$\mathbb{P}(X_n \in A | \Gamma_{n-1} = 1, X_{n-1} = x) = \nu(A),$$
$$\mathbb{P}(X_n \in A | \Gamma_{n-1} = 0, X_{n-1} = x) = Q(x,A),$$

where Q is the normalized "residual" kernel given by

$$Q(x,\cdot) := \frac{P(x,\cdot) - \beta \mathbb{I}(x \in J)\nu(\cdot)}{1 - \beta \mathbb{I}(x \in J)}.$$

Whenever $\Gamma_{n-1} = 1$, the chain regenerates at moment n. The regeneration epochs are

$$T_0 := \min\{n : \Gamma_{n-1} = 1\},$$
$$T_k := \min\{n > T_{k-1} : \Gamma_{n-1} = 1\}.$$

Write $\tau_k = T_k - T_{k-1}$. The random tours defined by

$$\Xi_k := (X_{T_{k-1}}, \dots, X_{T_k - 1}, \tau_k), \qquad \text{where} \quad \tau_k = T_k - T_{k-1}, \tag{13}$$

are i.i.d. for $k > 0$. Without loss of generality, we assume that $X_0 \sim \nu(\cdot)$, unless stated otherwise. We therefore put $T_0 := 0$ and simplify notation. In the sequel symbols $\mathbb{P}$ and $\mathbb{E}$ without subscripts refer to the chain started at ν. If the initial distribution ξ is other than ν, it will be explicitly indicated by writing $\mathbb{P}_\xi$ and $\mathbb{E}_\xi$. Notation $m = \mathbb{E}\,\tau_1$ stands throughout the paper.

We assume that we are able to *identify* regeneration times T_k. Mykland et al. pointed out in [25] that actual sampling from Q can be avoided. We can generate the chain using transition probabability P and then recover the regeneration indicators via

$$\mathbb{P}(\Gamma_{n-1} = 1 | X_n, X_{n-1}) = \mathbb{I}(X_{n-1} \in J) \frac{\beta \nu(\mathrm{d}X_n)}{P(X_{n-1}, \mathrm{d}X_n)},$$

where $\nu(\mathrm{d}y)/P(x, \mathrm{d}y)$ denote the Radon-Nikodym derivative (in practice, the ratio of densities). Mykland's trick has been established in a number of practically relevant families (e.g. hierarchical linear models) and specific Markov chains implementations, such as block Gibbs samplers or variable-at-a-time chains, see [18, 26].

3 General results for regenerative estimators

Let $f : \mathscr{X} \to \mathbb{R}$ be a measurable function and $\theta = \pi f$. We consider block sums $\Xi_k(f)$ defined by (4). From the general Kac theorem (Theorem 10.0.1 in [24] or Equation (3.5.1) in [29]) we have

$$m = \frac{1}{\beta \pi(J)}, \qquad \mathbb{E}\,\Xi_1(f) = m \pi f = m\theta.$$

From now on we assume that $\mathbb{E}\,\Xi_1(f)^2 < \infty$ and $\mathbb{E}\,\tau_1^2 < \infty$. Let $\bar{f} := f - \pi f$ and define

$$\begin{aligned} \sigma^2_{\mathrm{as}}(f) &:= \frac{\mathbb{E}\,\Xi_1(\bar{f})^2}{m}, \\ \sigma^2_\tau &:= \frac{\operatorname{Var}\tau_1}{m} \end{aligned} \tag{14}$$

Remark 1. Under Assumption 2.1, finiteness of $\mathbb{E}\,\Xi_1(\bar{f})^2$ is a sufficient and necessary condition for the CLT to hold for Markov chain $(X_n)_{n\ge 0}$ and function f. This fact is proved in [4] in a more general setting. For our purposes it is important to note that $\sigma^2_{\mathrm{as}}(f)$ in (14) is indeed the *asymptotic variance* which appears in the CLT.

3.1 Results for $\hat{\theta}_r^{\text{reg}}$

We are to bound the estimation error which can be expressed as follows:

$$\hat{\theta}_r^{\text{reg}} - \theta = \frac{\sum_{k=1}^r \left(\Xi_k(f) - \theta\tau_k\right)}{\sum_{k=1}^r \tau_k} = \frac{\sum_{k=1}^r d_k}{T_r}. \tag{15}$$

where $d_k := \Xi_k(f) - \theta\tau_k = \Xi_k(\bar{f})$. Therefore, for any $0 < \delta < 1$,

$$\mathbb{P}(|\hat{\theta}_r^{\text{reg}} - \theta| > \varepsilon) \leq \mathbb{P}\left(\left|\sum_{k=1}^r d_k\right| > rm\varepsilon(1-\delta)\right) + \mathbb{P}\Big(T_r < rm(1-\delta)\Big).$$

Since d_k are i.i.d. with $\mathbb{E}\, d_1 = 0$ and $\text{Var} d_1 = m\sigma_{\text{as}}^2(f)$, we can use Chebyshev inequality to bound the first term above:

$$\mathbb{P}\left(\left|\sum_{k=1}^r d_k\right| > rm\varepsilon(1-\delta)\right) \leq \frac{\sigma_{\text{as}}^2(f)}{rm\varepsilon^2(1-\delta)^2} \tag{16}$$

The second term can be bounded similarly. We use the fact that τ_k are i.i.d. with $\mathbb{E}\,\tau_1 = m$ to write

$$\mathbb{P}\Big(T_r < rm(1-\delta)\Big) \leq \frac{\sigma_\tau^2}{rm\delta^2}. \tag{17}$$

We conclude the above calculation with in following Theorem.

Theorem 3.1 *Under Assumption 2.1 the following holds for every* $0 < \delta < 1$

$$\mathbb{P}(|\hat{\theta}_r^{\text{reg}} - \theta| > \varepsilon) \leq \frac{1}{rm}\left[\frac{\sigma_{\text{as}}^2(f)}{\varepsilon^2(1-\delta)^2} + \frac{\sigma_\tau^2}{m\delta^2}\right] \tag{18}$$

and is minimized by

$$\delta = \delta^* := \frac{\sigma_\tau^{2/3}}{\sigma_{\text{as}}^{2/3}(f)\varepsilon^{-2/3} + \sigma_\tau^{2/3}}.$$

Obviously, $\mathbb{E}\, T_r = rm$ is the expected length of trajectory. The main drawback of Theorem 3.1 is that the bound on the estimation error depends on m, which is typically unknown. This fact motivates our study of another estimator, $\hat{\theta}_n^{\text{reg-seq}}$, for which we can obtain much more satisfactory results. We think that derivation of better nonasymptotic bounds for $\hat{\theta}_r^{\text{reg}}$ (not involving m) is an open problem.

3.2 Results for $\tilde{\theta}_r^{\text{unb}}$

Recall that $\tilde{\theta}_r^{\text{unb}}$ can be used only when m is known and this situation is rather rare in MCMC applications. Analysis of $\tilde{\theta}_r^{\text{unb}}$ is straightforward, because it is simply a sum of i.i.d. random variables. In particular, we obtain the following.

Corollary 3.2 *Under Assumption 2.1,*

$$\mathbb{E}\,(\tilde{\theta}_r^{\text{unb}} - \theta)^2 = \frac{\sigma^2_{\text{unb}}(f)}{rm}, \qquad \mathbb{P}(|\tilde{\theta}_r^{\text{unb}} - \theta| > \varepsilon) \leq \frac{\sigma^2_{\text{unb}}(f)}{rm\,\varepsilon^2}. \tag{19}$$

Note that $\sigma^2_{\text{unb}}(f) = \operatorname{Var}\Xi_1(f)/m$ can be expressed as

$$\sigma^2_{\text{unb}}(f) = \sigma^2_{\text{as}}(f) + \theta^2\sigma^2_\tau + 2\theta\rho(\bar{f},1), \tag{20}$$

where $\rho(\bar{f},1) := \operatorname{Cov}(\Xi_1(\bar{f}),\Xi_1(1))/m$. This follows from the simple observation that $\operatorname{Var}\Xi_1(f) = \mathbb{E}\,(\Xi_1(\bar{f}) + \theta(\tau_1 - m))^2$.

3.3 Results for $\hat{\theta}_n^{\text{reg-seq}}$

The result below bounds the MSE and the expected number of samples used to compute the estimator.

Theorem 3.3 *If Assumption 2.1 holds then*

$$(i) \quad \mathbb{E}\,(\hat{\theta}_n^{\text{reg-seq}} - \theta)^2 \leq \frac{\sigma^2_{\text{as}}(f)}{n^2}\,\mathbb{E}\,T_{R(n)}$$

and

$$(ii) \quad \mathbb{E}\,T_{R(n)} \leq n + C_0,$$

where

$$C_0 := \sigma^2_\tau + m + 1.$$

Corollary 3.4 *Under Assumption 2.1,*

$$\mathbb{E}\,(\hat{\theta}_n^{\text{reg-seq}} - \theta)^2 = \frac{\sigma^2_{\text{as}}(f)}{n}\left(1 + \frac{C_0}{n}\right), \tag{21}$$

$$\mathbb{P}(|\hat{\theta}_n^{\text{reg-seq}} - \theta| > \varepsilon) \leq \frac{\sigma^2_{\text{as}}(f)}{n\varepsilon^2}\left(1 + \frac{C_0}{n}\right). \tag{22}$$

Remark 2. Note that the leading term $\sigma^2_{\text{as}}(f)/n$ in (21) is “asymptotically correct” in the sense that the standard fixed length estimator has MSE $\sim \sigma^2_{\text{as}}(f)/n$. The regenerative-sequential scheme is “close to the fixed length simulation”, besause $\lim_{n\to\infty}\mathbb{E}\,T_{R(n)}/n = 1$.

Proof (of Theorem 3.3). Just as in (15) we have

$$\hat{\theta}_n^{\text{reg-seq}} - \theta = \frac{\sum_{k=1}^{R(n)} (\Xi_k(f) - \theta \tau_k)}{\sum_{k=1}^{R(n)} \tau_k} = \frac{1}{T_{R(n)}} \sum_{k=1}^{R(n)} d_k,$$

where pairs (d_k, τ_k) are i.i.d. with $\mathbb{E}\, d_1 = 0$ and $\operatorname{Var} d_1 = m\sigma_{\text{as}}^2(f)$. Since $T_{R(n)} > n$, it follows that

$$\mathbb{E}\,(\hat{\theta}_n^{\text{reg-seq}} - \theta)^2 \leq \frac{1}{n^2} \mathbb{E} \left(\sum_{k=1}^{R(n)} d_k \right)^2.$$

Since $R(n)$ is a stopping time with respect to $\mathscr{G}_k = \sigma((d_1, \tau_1), \ldots, (d_k, \tau_k))$, we are in a position to apply the two Wald's identities. The second identity yields

$$\mathbb{E} \left(\sum_{k=1}^{R(n)} d_k \right)^2 = \operatorname{Var} d_1 \, \mathbb{E}\, R(n) = m\sigma_{\text{as}}^2(f)\, \mathbb{E}\, R(n).$$

In this expression we can replace $m\mathbb{E}\, R(n)$ by $\mathbb{E}\, T_{R(n)}$ because of the first Wald's identity:

$$\mathbb{E}\, T_{R(n)} = \mathbb{E} \sum_{k=1}^{R(n)} \tau_k = \mathbb{E}\, \tau_1 \, \mathbb{E}\, R(n) = m\mathbb{E}\, R(n)$$

and (i) follows.

We now focus attention on bounding the expectation of the "overshoot" $\Delta(n) := T_{R(n)} - n$. Since we assume that $X_0 \sim \nu$, the cumulative sums $\tau_1 = T_1 < T_2 < \ldots < T_k < \ldots$ form a (nondelayed) renewal process in discrete time. Let us invoke the following elegant theorem of Lorden [22]:

$$\mathbb{E}\, \Delta(n) \leq 2\, \mathbb{E}_\pi \tau_1. \tag{23}$$

By a classical result of the renewal theory, the distribution of τ_1 under stationarity is given by

$$\mathbb{P}_\pi(\tau_1 = i) = \frac{\mathbb{P}(\tau_1 \geq i)}{m} \quad \text{for } i = 1, 2, \ldots. \tag{24}$$

For a newer simple proof of Lorden's inequality, we refer to [6]. Equation (23) yields immediately (ii) with $C_0 = 2\mathbb{E}_\pi$. Elementary calculation (using (24) and summation by parts) shows that $2\mathbb{E}_\pi \tau_1 = \mathbb{E}\, \tau_1^2/m + 1$. Recall the definition of σ_τ^2 to complete the proof.

4 The median trick

This ingeneous method of constructing fixed precision MCMC algorithms was introduced in 1986 in [16], later used in many papers concerned with computational complexity and further developed in [27]. We run l independent copies of the Markov chain. Let $\hat{\theta}^{(j)}$ be an estimator computed in jth run. The final estimate is $\hat{\theta} := \text{med}(\hat{\theta}^{(1)},\ldots,\hat{\theta}^{(l)})$. To ensure that $\hat{\theta}$ satisfies (12), we require that $\mathbb{P}(|\hat{\theta}^{(j)}-\theta|>\varepsilon)\leq a$ $(j=1,\ldots,l)$ for some modest level of confidence $1-a<1-\alpha$. This is obtained via Chebyshev's inequality, if a bound on MSE is available. The well-known Chernoff's inequality gives for odd l,

$$\mathbb{P}(|\hat{\theta}-\theta|\geq\varepsilon)\leq\frac{1}{2}\left[4a(1-a)\right]^{l/2}=\frac{1}{2}\exp\left\{\frac{l}{2}\ln\left[4a(1-a)\right]\right\}. \tag{25}$$

It is pointed out in [27] that under some assumptions there is a universal choice of a, which nearly minimizes the overall number of samples, $a^*\approx 0.11969$. The details are described in [27].

Let us now examine how the median trick works in conjunction with regenerative MCMC. We focus on $\hat{\theta}_n^{\text{reg-seq}}$, because Corollary 3.4 gives the best available bound on MSE. We first choose n such that the right hand side of (22) is less than or equal to a^*. Then choose l big enough to make the right hand side of (25) (with $a=a^*$) less than or equal to α. Compute estimator $\hat{\theta}_n^{\text{reg-seq}}$ repeatedly, using l independent runs of the chain. We can easily see that (12) holds if

$$\begin{aligned} n &\geq \frac{C_1\sigma_{\text{as}}^2(f)}{\varepsilon^2}+C_0,\\ l &\geq C_2\ln(2\alpha)^{-1} \text{ and } j \text{ is odd}, \end{aligned}$$

where $C_1 := 1/a^*\approx 8.3549$ and $C_2 := 2/\ln\left[4a^*(1-a^*)\right]^{-1}\approx 2.3147$ are absolute constants. The overall (expected) number of generated samples is $l\mathbb{E}\,T_{R(n)}\sim nl$ as $\varepsilon\to 0$ and $n\to\infty$, by Theorem 3.3 (ii). Consequently for $\varepsilon\to 0$ the cost of the algorithm is approximately

$$nl\sim C\frac{\sigma_{\text{as}}^2(f)}{\varepsilon^2}\log(2\alpha)^{-1}, \tag{26}$$

where $C=C_1C_2\approx 19.34$. To see how tight is the obtained lower bound, let us compare (26) with the familiar asymptotic approximation, based on the CLT. Consider an estimator based on one MCMC run of length n, say $\hat{\theta}_n=\hat{\theta}_{t,n}^{\text{fix}}$ with $t=0$. From (3) we infer that

$$\lim_{\varepsilon\to 0}\mathbb{P}(|\hat{\theta}_n-\theta|>\varepsilon)=\alpha,$$

holds for

$$n\sim\frac{\sigma_{\text{as}}^2(f)}{\varepsilon^2}\left[\Phi^{-1}(1-\alpha/2)\right]^2, \tag{27}$$

where Φ^{-1} is the quantile function of the standard normal distribution. Taking into account the fact that $[\Phi^{-1}(1-\alpha/2)]^2\sim 2\log(2\alpha)^{-1}$ for $\alpha\to 0$ we arrive at the following conclusion. The right hand side of (26) is bigger than (27) roughly by a constant factor of about 10 (for small ε and α). The important difference is that (26) is sufficient for an *exact* confidence interval while (27) only for an *asymptotic* one.

5 Doeblin Chains

Assume that the transition kernel P satisfies the following Doeblin condition: there exist $\beta > 0$ and a probability measure ν such that

$$P(x,\cdot) \geq \beta \nu(\cdot) \quad \text{for every} \quad x \in \mathscr{X}. \tag{28}$$

This amounts to taking $J := \mathscr{X}$ in Assumption 2.1. Condition (28) implies that the chain is uniformly ergodic. We refer to [32] and [24] for definition of uniform ergodicity and related concepts. As a consequence of the regeneration construction, in our present setting τ_1 is distributed as a geometric random variable with parameter β and therefore

$$m = \mathbb{E}\,\tau_1 = \frac{1}{\beta} \qquad \text{and} \qquad \sigma_\tau^2 = \frac{\operatorname{Var}\tau_1}{m} = \frac{1-\beta}{\beta}.$$

Bounds on the asymptotic variance $\sigma_{\mathrm{as}}^2(f)$ under (28) are well known. Let $\sigma^2 = \pi \bar{f}^2$ be the stationary variance. Results in Section 5 of [4] imply that

$$\sigma_{\mathrm{as}}^2(f) \leq \sigma^2 \left(1 + \frac{2}{1-\sqrt{1-\beta}}\right) \leq \frac{4\sigma^2}{\beta}. \tag{29}$$

If the chain is reversible, there is a better bound. We can use the well-known formula for $\sigma_{\mathrm{as}}^2(f)$ in terms of the spectral decomposition of P (e.g. expression "C" in [11]). Results of [31] show that the spectrum of P is a subset of $[-1+\beta, 1-\beta]$. We conclude that for reversible Doeblin chains,

$$\sigma_{\mathrm{as}}^2(f) \leq \frac{2-\beta}{\beta}\sigma^2 \leq \frac{2\sigma^2}{\beta}. \tag{30}$$

An important class of reversible chains are Independence Metropolis-Hastings chains (see e.g. [32]) that are known to be uniformly ergodic if and only if the rejection probability $r(x)$ is uniformly bounded from 1 by say $1-\beta$. This is equivalent to the candidate distribution being bounded below by $\beta\pi$ (c.f. [23, 1]) and translates into (28) with $\nu = \pi$. The formula for $\sigma_{\mathrm{as}}^2(f)$ in (29) and (30) depends on β in an optimal way. Moreover (30) is sharp. To see this consider the following example.

Example 5.1 *Let $\beta \leq 1/2$ and define a Markov chain $(X_n)_{n\geq 0}$ on $\mathscr{X} = \{0,1\}$ with stationary distribution $\pi = \{1/2, 1/2\}$ and transition matrix*

$$P = \begin{bmatrix} 1-\beta/2 & \beta/2 \\ \beta/2 & 1-\beta/2 \end{bmatrix}.$$

Hence $P = \beta\pi + (1-\beta)I_2$ and $P(x,\cdot) \geq \beta\pi$. Moreover let $f(x) = x$. Thus $\sigma^2 = 1/4$. To compute $\sigma_{\mathrm{as}}^2(f)$ we use another well-known formula (expression "B" in [11]):

$$\sigma^2_{\text{as}}(f) = \sigma^2 + 2\sum_{i=1}^{\infty} \text{Cov}\{f(X_0), f(X_i)\}$$
$$= \sigma^2 + 2\sigma^2 \sum_{i=1}^{\infty}(1-\beta)^i = \frac{2-\beta}{\beta}\sigma^2.$$

To compute $\sigma^2_{\text{unb}}(f)$*, note that* $\Xi_1(f) = \mathbb{I}(X_0 = 1)\tau_1$*. Since* τ_1 *is independent of* X_0 *and* $X_0 \sim \nu = \pi$ *we obtain*

$$\sigma^2_{\text{unb}}(f) = \beta \text{Var}\,\Xi_1(f) = \beta\left[\mathbb{E}\,\text{Var}(\Xi_1(f)|X_0) + \text{Var}\,\mathbb{E}\,(\Xi_1(f)|X_0)\right]$$
$$= \frac{1-\beta}{2\beta} + \frac{1}{4\beta} = \frac{3-2\beta}{\beta}\sigma^2.$$

Interestingly, in this example $\sigma^2_{\text{unb}}(f) > \sigma^2_{\text{as}}(f)$*.*

In the setting of this Section, we will now compare upper bounds on the total simulation effort needed for different MCMC schemes to get $\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) \le \alpha$.

5.1 Regenerative-sequential estimator and the median trick

Recall that this simulation scheme consists of l MCMC runs, each of approximate length n. Substituting (29) in (26) we obtain that the expected number of samples is

$$nl \sim 19.34\frac{4\sigma^2}{\beta\varepsilon^2}\log(2\alpha)^{-1} \quad \text{and} \quad nl \sim 19.34\frac{(2-\beta)\sigma^2}{\beta\varepsilon^2}\log(2\alpha)^{-1} \tag{31}$$

(respectively in the general case and for reversible chains). Note also that in the setting of this Section we have an exact expression for the constant C_0 in Theorem 3.3. Indeed, $C_0 = 2\mathbb{E}_\pi \tau_1 = 1/\beta$.

5.2 Standard one-run average and exponential inequalty

For uniformly ergodic chains a direct comparison of our approach to exponential inequalities [10, 19] is possible. We focus on [19] which is tight in the sense that it reduces to the Hoeffding bound when specialised to the i.i.d. case. For f bounded let $\|f\|_{\text{sp}} := \sup_{x\in\mathscr{X}} f(x) - \inf_{x\in\mathscr{X}} f(x)$. Consider the simple average over n Markov chain samples, say $\hat{\theta}_n = \hat{\theta}^{\text{fix}}_{t,n}$ with $t = 0$. For an arbitrary initial distribution ξ we have

$$\mathbb{P}_\xi(|\hat{\theta}_n - \theta| > \varepsilon) \le 2\exp\left\{-\frac{n-1}{2}\left(\frac{2\beta}{\|f\|_{\text{sp}}}\varepsilon - \frac{3}{n-1}\right)^2\right\}.$$

After identifying leading terms we can see that to make the right hand side less than α we need

$$n \sim \frac{\|f\|_{\mathrm{sp}}^2}{2\beta^2\varepsilon^2}\log(\alpha/2)^{-1} \geq \frac{2\sigma^2}{\beta^2\varepsilon^2}\log(\alpha/2)^{-1}. \quad (32)$$

Comparing (31) with (32) yields a ratio of roughly 40β or 20β respectively. This in particular indicates that the dependence on β in [10, 19] probably can be improved. We note that in examples of practical interest β usually decays exponentially with dimension of $\mathscr{X}$ and using the regenerative-sequential-median scheme will often result in a lower total simulation cost. Moreover, this approach is valid for an unbounded target function f, in contrast with classical exponential inequalities.

5.3 Perfect sampler and the median trick

For Doeblin chains, if regeneration times can be identified, perfect sampling can be performed easily as a version of read-once algorithm [36]. This is due to the following observation. If condition (28) holds and $X_0 \sim \nu$ then

$$X_{T_i-1}, \quad i = 1, 2, \ldots \quad (33)$$

are i.i.d. random variables from π (see [5, 29, 15, 4] for versions of this result). Therefore from each random tour between regeneration times one can obtain a single perfect sample (by taking the state of the chain prior to regeneration) and use it for i.i.d. estimation. We define

$$\hat{\theta}_n^{\mathrm{perf}} := \frac{1}{n}\sum_{i=1}^{n} f(X_{T_i-1}). \quad (34)$$

Clearly

$$\mathbb{E}\,(\hat{\theta}_n^{\mathrm{perf}} - \theta)^2 = \frac{\sigma^2}{n} \qquad \text{and} \qquad \mathbb{P}(|\hat{\theta}_n^{\mathrm{perf}} - \theta| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}. \quad (35)$$

If we combine this prefect sampler with the median trick we obtain an algorithm with expected number of samples

$$nl \sim 19.34\frac{\sigma^2}{\beta\varepsilon^2}\log(2\alpha)^{-1}. \quad (36)$$

Comparing (31) with (32) and (35) leads to the conclusion that if one targets rigorous nonasymptotic results in the Doeblin chain setting, the approach described here outperforms other methods.

5.4 Remarks on other schemes

The bound for $\hat{\theta}_r^{\text{reg}}$ in Theorem 3.1 is clearly inferior to that for $\hat{\theta}_n^{\text{reg-seq}}$ in Corollary 3.4 Therefore we excluded the scheme based on $\hat{\theta}_r^{\text{reg}}$ from our comparisons.

As for $\tilde{\theta}_r^{\text{unb}}$, this estimator can be used in the Doeblin chains setting, because $m = 1/\beta$ is known. The bounds for $\tilde{\theta}_r^{\text{unb}}$ in Subsection 3.2 involve $\sigma^2_{\text{unb}}(f)$. Although we cannot provide a rigorous proof, we conjecture that in most practical situations we have $\sigma^2_{\text{unb}}(f) > \sigma^2_{\text{as}}(f)$, because $\rho(\bar{f},1)$ in (20) is often close to zero. If this is the case, then the bound for $\tilde{\theta}_r^{\text{unb}}$ is inferior to that for $\hat{\theta}_n^{\text{reg-seq}}$.

6 A Geometric Drift Condition

Using drift conditions is a standard approach for establishing geometric ergodicity. We refer to [32] or [24] for the definition and further details. The assumption below is the same as in [3]. Specifically, let J be the small set which appears in Assumption 2.1.

Assumption 6.1 (Drift) *There exist a function $V : \mathscr{X} \to [1,\infty[$, constants $\lambda < 1$ and $K < \infty$ such that*

$$PV(x) := \int_{\mathscr{X}} P(x,\mathrm{d}y)V(y) \leq \begin{cases} \lambda V(x) & \text{for } x \notin J, \\ K & \text{for } x \in J, \end{cases}$$

In many papers conditions similar to 6.1 have been established for realistic MCMC algorithms in statistical models of practical relevance [12, 7, 8, 13, 18, 35]. This opens the possibility of computing our bounds in these models.

Under Assumption 6.1, it is possible to bound $\sigma^2_{\text{as}}(f)$, σ^2_τ and C_0 which appear in Theorems 3.1 and 3.3, by expressions involving only λ, β and K. The following result is a minor variation of Theorem 6.5 in [20].

Theorem 6.2 *If Assumptions 2.1 and 6.1 hold and f is such that $\|\bar{f}\|_{V^{1/2}} := \sup_x |\bar{f}(x)|/V^{1/2}(x) < \infty$, then*

$$\sigma^2_{\text{as}}(f) \leq \|\bar{f}\|^2_{V^{1/2}} \left[\frac{1+\lambda^{1/2}}{1-\lambda^{1/2}}\pi(V) + \frac{2(K^{1/2}-\lambda^{1/2}-\beta(2-\lambda^{1/2}))}{\beta(1-\lambda^{1/2})}\pi(V^{1/2}) \right]$$

$$C_0 \leq \frac{\lambda^{1/2}}{1-\lambda^{1/2}}\pi(V^{1/2}) + \frac{K^{1/2}-\lambda^{1/2}-\beta}{\beta(1-\lambda^{1/2})},$$

To bound σ^2_τ we can use the obvious inequality $\sigma^2_\tau = C_0 - m - 1 \leq C_0 - 2$. Moreover, one can easily control πV^2 and πV, and further replace $\|\bar{f}\|_{V^{1/2}}$ e.g. by $\|f\|_{V^{1/2}} + (K^{1/2}-\lambda^{1/2})/(1-\lambda^{1/2})$, we refer to [20] for details.

Let us now discuss possible approaches to confidence estimation in the setting of this section. Perfect sampling is in general unavailable. For unbounded f we cannot apply exponential inequalities for the standard one-run estimate. Since m is unknown we cannot use $\tilde{\theta}_r^{\text{unb}}$. This leaves $\hat{\theta}_r^{\text{reg}}$ and $\hat{\theta}_n^{\text{reg-seq}}$ combined with the median trick. To analyse $\hat{\theta}_r^{\text{reg}}$ we can apply Theorem 3.1. Upper bounds for $\sigma^2_{\text{as}}(f)$ and σ^2_τ are available. However, in Theorem 3.1 we will also need a *lower bound* on m. Without further assumptions we can only write

$$m = \frac{1}{\pi(J)\beta} \geq \frac{1}{\beta}. \tag{37}$$

In the above analysis (37) is particularly disappointing. It multiplies the bound by an unexpected and substantial factor, as $\pi(J)$ is typically small in applications. For $\hat{\theta}_n^{\text{reg-seq}}$ we have much more satisfactory results. Theorems 3.3 and 6.2 can be used to obtain explicit bounds which do not involve m.

We note that nonasymptotic confidence intervals for MCMC estimators under drift condition have also been obtained in [21], where identification of regeneration times has not been assumed. In absence of regeneration times a different approach has been used and the bounds are typically weaker.

References

1. Y.F. Atchade, F. Perron (2007): On the geometric ergodicity of Metropolis-Hastings algorithms. *Statistics* 41, 77–84.
2. K.B. Athreya and P. Ney (1978): A new approach to the limit theory of recurrent Markov chains, *Trans. Amer. Math. Soc.* 245, 493–501.
3. P.H. Baxendale (2005): Renewal Theory and Computable Convergence Rates for Geometrically Ergodic Markov Chains. *Ann. Appl. Prob.* 15, 700-738.
4. W. Bednorz, R. Latała and K. Łatuszyński (2008): A Regeneration Proof of the Central Limit Theorem for Uniformly Ergodic Markov Chains. *Elect. Comm. in Probab.* 13, 85–98.
5. L.A. Breyer and G.O. Roberts (2001): Catalytic perfect simulation. *Methodol. Comput. Appl. Probab.* **3** 161–177.
6. J.T. Chang (1994): Inequalities for the overshoot. *Ann. Appl. Probab.* 4, 1223–1233.
7. G. Fort and E. Moulines (2000): V-subgeometric ergodicity for a Hastings–Metropolis algorithm. *Statist. Probab. Lett.* 49, 401–410.
8. G. Fort, E. Moulines, G.O. Roberts, and J.S. Rosenthal (2003): On the geometric ergodicity of hybrid samplers. *J. Appl. Probab.* 40 (1), 123-146.
9. W.R. Gilks, S. Richardson, D.J. Spiegelhalter: *Markov chain Monte Carlo in practice.* Chapman & Hall, 1998.
10. P.W. Glynn and D. Ormoneit (2002): Hoeffding's inequality for uniformly ergodic Markov chains, *Statist. Probab. Lett.* 56, 143–146.
11. O. Häggström J.S. Rosenthal (2007): On variance conditions for Markov chain CLTs. *Elect. Comm. in Probab.* 12 , 454–464.
12. J.P. Hobert and C.J. Geyer (1998): Geometric ergodicity of Gibbs and block Gibbs samplers for Hierarchical Random Effects Model. *J. Multivariate Anal.* 67, 414–439.
13. J.P. Hobert and G.L. Jones (2004): Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *Ann. Statist.* 32 (2), 784–817
14. J.P. Hobert, G.L. Jones, B. Presnell, and J.S. Rosenthal (2002): On the Applicability of Regenerative Simulation in Markov Chain Monte Carlo. *Biometrika* 89, 731-743.
15. J.P. Hobert and C.P. Robert (2004): A mixture representation of π with applications in Markov chain Monte Carlo and perfect smpling. *Ann. Appl. Probab.* **14** 1295–1305.

16. M.R. Jerrum, L.G. Valiant, V.V. Vizirani (1986): Random generation of combinatorial structures fro, a uniform distribution. *Theoretical Computer Science* 43, 169–188.
17. G.J. Jones, J.P. Hobert (2004): Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *Ann. Statist.* 32, pp. 784–817.
18. A.A. Johnson and G.L. Jones (2010): Gibbs sampling for a Bayesian hierarchical general linear model. *Electronic J. Statist.* 4, 313–333.
19. I. Kontoyiannis, L. Lastras-Montano, S.P. Meyn (2005): Relative Entropy and Exponential Deviation Bounds for General Markov Chains. *2005 IEEE International Symposium on Information Theory.*
20. K. Łatuszyński, B. Miasojedow annd W. Niemiro (2009): Nonasymptotic bounds on the estimation error for regenerative MCMC algorithms. arXiv:0907.4915v1
21. K. Łatuszyński, W. Niemiro (2011): Rigorous confidence bounds for MCMC under a geometric drift condition. *J. of Complexity* 27, 23–38.
22. G. Lorden: On excess over the boundary. *Ann. Math. Statist.* 41, 520–527, 1970.
23. K.L. Mengersen, L.R. Tweedie (1996): Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* 24, 1, 101–121.
24. S.P. Meyn and R.L. Tweedie: *Markov Chains and Stochastic Stability.* Springer-Verlag, 1993.
25. P. Mykland, L. Tierney and B. Yu (1995): Regeneration in Markov chain samplers. *J. Am. Statist. Assoc.*., 90, 233–241.
26. R. Neath, G.L. Jones (2009): Variable-at-a-time implementation of Markov chain Monte Carlo. Preprint. arXiv:0903.0664v1
27. W. Niemiro, P. Pokarowski (2009): Fixed precision MCMC Estimation by Median of Products of Averages. *J. Appl. Probab.* 46 (2), 309–329.
28. E. Nummelin (1978): A splitting technique for Harris recurrent Markov chains, *Z. Wahr. Verw. Geb.* 43, 309–318.
29. E. Nummelin (2002): MC's for MCMC'ists, *International Statistical Review*, 70, 215–240.
30. C.P. Robert and G. Casella: *Monte Carlo Statistical Methods.* Springer-Verlag, New York, 2004.
31. G.O. Roberts and J.S. Rosenthal (1997): Geometric ergodicity and hybrid Markov chains. *Elec. Comm. Prob.* 2 (2).
32. G.O. Roberts and J.S. Rosenthal (2004): General state space Markov chains and MCMC algorithms. *Probability Surveys* 1, 20–71.
33. J.S. Rosenthal (1995): Minorization conditions and convergence rates for Markov chains. *J. Amer. Statist. Association* 90, 558–566.
34. D. Rudolf (2008): Explicit error bounds for lazy reversible Markov chain Monte Carlo. *J. of Complexity.* 25, 11–24.
35. V. Roy, J.P. Hobert (2010): On Monte Carlo methods for Bayesian multivariate regression models with heavy-tailed errors. *J. Multivariate Anal.* 101, 1190–1202
36. D.B. Wilson (2000): How to couple from the past using a read-once source of randomness. *Random Structures Algorithms* 16 (1), 85–113.