Nonasymptotic bounds on the estimation error for regenerative MCMC algorithms*

Krzysztof Latuszyński, Błażej Miasojedow and Wojciech Niemiro

K. Latuszyński
Department of Statistics
University of Warwick
CV4 7AL, Coventry, UK
e-mail: latuch@gmail.com
url: http://www2.warwick.ac.uk/fac/sci/statistics/staff/research/latuszynski/

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

W. Niemiro
Faculty of Mathematics and Computer Science
Nicolaus Copernicus University
Chopina 12/18, 87-100 Toruń, Poland
e-mail: wniem@mat.uni.torun.pl

Abstract: MCMC methods are used in Bayesian statistics not only to sample from posterior distributions but also to estimate expectations. Underlying functions are most often defined on a continuous state space and can be unbounded. We consider a regenerative setting and Monte Carlo estimators based on i.i.d. blocks of a Markov chain trajectory. The main result is an inequality for the mean square error. We also consider confidence bounds. We first derive the results in terms of the asymptotic variance and then bound the asymptotic variance for both uniformly ergodic and geometrically ergodic Markov chains.

AMS 2000 subject classifications: Primary 60J10, 65C05; secondary 62L12.
Keywords and phrases: Mean sqare error, Confidence estimation, Computable bounds, Geometric drift, Asymptotic variance.

1. Introduction

Suppose that we are to estimate the expectation of a function, possibly unbounded and defined on a high dimensional space, with respect to some probability density which is known only up to a normalising constant. Such problems arise in Bayesian inference and are often solved using Markov chain Monte Carlo (MCMC) methods. The idea is to simulate a Markov chain converging to the

*Work partially supported by Polish Ministry of Science and Higher Education Grant No. N N201387234.

target distribution and take ergodic averages as estimates of the expectation. It is essential to have explicit and reliable bounds which provide information about how long the algorithms must be run to achieve a prescribed level of accuracy (c.f. [54, 25, 28]).

We consider MCMC algorithms which use independent and identically distributed random blocks of the underlying Markov chain, each block starting and ending at consecutive regeneration times. In fact, we propose a sequential version of regenerative estimator, for which the length of trajectory is "nearly fixed". This methodology is a promising alternative to both fixing the total length of the trajectory and fixing the number of regeneration cycles [45, 54, 26, 8, 9, 10]. The simulation scheme is easy to implement, provided that the regeneration times can be identified. We introduce our estimator and discuss its properties in Section 3.

The regenerative/sequential simulation scheme which we propose allows us to use directly the tools of the renewal theory and statistical sequential analysis. Our goal is to obtain quantitative bounds on the error of MCMC estimation. We aim at explicit nonasymptotic results. To this end we split the analysis into independent parts.

First, in Section 3, we derive inequalities on the mean square error (MSE) in terms of the asymptotic variance of the chain. This is obtained under very weak assumptions. We require a one step minorization condition (Assumption 2.1) and an integrability conditions that are essentially equivalent to those needed for Central Limit Theorems (CLTs) for nontrivial target functions. The proof of our main result, Theorem 3.3, depends on a classical result of Lorden [38] about the mean "excess time" for renewal processes and also on the two Wald's identities.

Next, in Section 4, we consider confidence estimation via a median trick that leads to an exponential inequality and we argue that our nonasymptotic bounds are not far off the asymptotic approximation based on the CLT.

Finally, we proceed to express the bounds in terms of computable quantities. In Section 5 we consider uniformly ergodic chains, where bounding the asymptotic variance is straightforward. Moreover, in case of a bounded target function we can compare our approach to the well known exponential inequalities for Doeblin chains. Our bound is always within a factor of at most 40β of the exponential inequality (where β is the regeneration parameter of the Doeblin chain), so it will turn out sharper for many examples of practical interest, where β is small. In Section 6 we assume the most general setting that motivates our work, namely a drift towards a small set, to replace the unknown asymptotic variance by known drift parameters. Our Assumption 6.1 is quite similar to many analogous drift conditions known in the literature, see e.g. [44, 51, 5]. For aperiodic chains Assumption 6.1 implies geometric ergodicity, but we do not need aperiodicity for our purposes. We build on some auxiliary results of [5] to derive bounds on the asymptotic variance.

The nonasymptotic confidence intervals we derive are valid in particular for unbounded target functions and Markov chains that are not uniformly ergodic. Our assumptions are comparable (in some cases identical) to those required for asymptotically valid confidence intervals (c.f. [28, 6, 17, 9]). Moreover the bounds are expressed in terms of known quantities and thus can be of interest for MCMC practitioners. In Section 8 we discuss connections with related results in literature from both applied and theoretical viewpoint.

One of the benchmarks for development of MCMC technology is the important hierarchical Bayesian model of variance components [53, 24, 29], used e.g. for small area estimation in survey sampling and in actuarial mathematics. We illustrate our theoretical results with a simple example which can be regarded as a part of this model. Since the analytic solution is known in this example, it is possible to assess the tightness of our bounds. The full model of variance components will be considered in [35] and [36].

2. Regenerative simulation

Let π be a probability distribution on a Polish space $\mathcal{X}$. Consider a Markov transition kernel P such that $\pi P = \pi$, that is π is stationary with respect to P. Assume P is π-irreducible. The regeneration/split construction of Nummelin [47] and Athreya and Ney [4] rests on the following assumption.

2.1 Assumption (Small Set)**.** *There exist a Borel set $J \subseteq \mathcal{X}$ 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 $\mathcal{X} \times \{0, 1\}$ 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

$$T_1 := \min\{n \geq 1 : \Gamma_{n-1} = 1\},$$
$$T_k := \min\{n \geq T_{k-1} : \Gamma_{n-1} = 1\}.$$

Write $\tau_k = T_k - T_{k-1}$. Unless specified otherwise, we assume that $X_0 \sim \nu(\cdot)$ and therefore $T_0 := 0$ is also a time of regeneration. Symbols $\mathbb{P}$ and $\mathbb{E}$ without subscripts will be shorthands for $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$ while initial distributions other than ν will be explicitly indicated. The random blocks

$$\Xi_k := (X_{T_{k-1}}, \ldots, X_{T_k - 1}, \tau_k)$$

for $k = 1, 2, 3, \ldots$ are i.i.d.

We assume that we can simulate the split chain (X_n, Γ_n), starting from $X_0 \sim \nu(\cdot)$. Put differently, we are able to *identify* regeneration times T_k. Mykland et al. pointed out in [45] that actual sampling from Q can be avoided. Assume that the chain X_n is generated using transition probabability P. Let $\nu(\mathrm{d}y)/P(x, \mathrm{d}y)$ denote the Radon-Nikodym derivative (in practice, the ratio of densities). Then we can recover the regeneration indicators via

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

where U_n is a sequence of i.i.d. uniform variates independent of X_n. If sampling from the renewal distribution $\nu(\cdot)$ is difficult then we can start the simulation from an arbitrary state, discard the initial part of the trajectory before the first time of regeneration and consider only blocks Ξ_k for $k = 2, 3, \ldots$, that is begin at T_1 instead of $T_0 = 0$. Thus in the regenerative scheme there is a very precise recipe for an "absolutely sufficient burn-in" time.

3. Main Theorem

Let $f : \mathcal{X} \to \mathbb{R}$ be a Borel function. The objective is to compute (estimate) the quantity

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

We assume that θ exists, i.e. $\pi(|f|) < \infty$. Regenerative estimators of θ are based on the block sums

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

Let us now introduce a sequential version of regenerative estimator. Fix n and define

$$R(n) := \min\{r : T_r \geq n\}. \tag{3.1}$$

Our basic estimator is defined as follows.

$$\hat{\theta}_{T_{R(n)}} := \frac{1}{T_{R(n)}} \sum_{i=1}^{R(n)} \Xi_k(f) = \frac{1}{T_{R(n)}} \sum_{i=0}^{T_{R(n)}-1} f(X_i). \tag{3.2}$$

In words: we stop simulation *at the first moment of regeneration past* n and compute the usual sample average. Note that we thus generate a random number of blocks. Our regenerative scheme requires only as many blocks as necessary to make the length of trajectory at least n and the "excess time" $T_{R(n)} - n$ will be shown to be small compared to n.

The result below bounds the mean square error (MSE) of the estimator defined by (3.2), (3.1) and the expected number of samples used to compute it. Let $\bar{f} := f - \pi(f)$.

3.3 Theorem. *If Assumption 2.1 holds,* $\mathbb{E}(\Xi_1(\bar{f}))^2 < \infty$ *and* $\mathbb{E}\tau_1^2 < \infty$ *then*

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

and

$$(ii) \quad \mathbb{E}\,T_{R(n)} \leq n + n_0,$$

where

$$\sigma_{\text{as}}^2(f) := \frac{\mathbb{E}(\Xi_1(\bar{f}))^2}{\mathbb{E}\tau_1}, \qquad n_0 := \frac{\mathbb{E}\tau_1^2}{\mathbb{E}\tau_1} - 1.$$

3.4 Corollary. *Under the same assumptions,*

$$\mathbb{E}\,(\hat{\theta}_{T_{R(n)}} - \theta)^2 \leq \frac{\sigma_{\text{as}}^2(f)}{n}\left(1 + \frac{n_0}{n}\right).$$

Note that the leading term $\sigma_{\text{as}}^2(f)/n$ in Corollary 3.4 is "asymptotically correct" in the sense that, under our assumptions,

$$\lim_{n\to\infty} n\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \theta\right)^2 = \sigma_{\text{as}}^2(f) \text{ and } \lim_{n\to\infty}\frac{\mathbb{E}T_{R(n)}}{n} = 1.$$

3.5 REMARK. Under Assumption 2.1, finiteness of $\mathbb{E}(\Xi(\bar{f}))^2$ is a sufficient and necessary condition for the CLT to hold for Markov chain X_n and function f. This fact is proved in [7] in a more general setting. For our purposes it is important to note that $\sigma_{\text{as}}^2(f)$ in Theorem 3.3 is indeed the *asymptotic variance* which appears in the CLT. Constant n_0 bounds the ,,*mean overshoot*" or excess length of simulations over n.

Proof of Theorem 3.3 (i). Note that

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

where $d_k := \Xi_k(f) - \theta\tau_k = \Xi_k(\bar{f})$. By the Kac theorem ([44] or [48]) we have

$$\mathbb{E}\,\Xi_k(f) = m\pi(f) = m\theta,$$

where

$$m := \mathbb{E}\tau_k = \frac{1}{\beta\pi(J)}.$$

Consequently the pairs (d_k, τ_k) are i.i.d. with $\mathbb{E}d_k = 0$ and $\mathrm{Var}d_k = m\sigma^2_{\mathrm{as}}(f)$. Since $T_{R(n)} \geq n$, it follows that

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

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

$$\mathbb{E}\left(\sum_{k=1}^{R(n)} d_k\right)^2 = \mathrm{Var}\;d_1\,\mathbb{E}R(n) = m\sigma^2_{\mathrm{as}}(f)\,\mathbb{E}R(n).$$

But 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 the claimed result follows. □

We now focus attention on bounding the “excess” or “overshoot” time

$$\Delta(n) := T_{R(n)} - n.$$

To this end, let us recall a classical result of the (discrete time) renewal theory. As before, when using symbols $\mathbb{P}$ and $\mathbb{E}$ without subscripts we refer to the chain started at the renewal distribution ν and we write $m = \mathbb{E}\tau_1$. Let $\Delta(\infty)$ be a random variable having distribution

$$\mathbb{P}\,(\Delta(\infty) = i) := \frac{1}{m}\mathbb{P}(\tau_1 > i) \quad \text{for } i = 0, 1, 2, \ldots.$$

If the distribution of τ_1 is aperiodic then it is well-known that $\Delta(n) \to \Delta(\infty)$ in distribution, as $n \to \infty$, but we will not use this fact directly. Instead, we invoke the following elegant result.

3.6 Proposition (Lorden [38])**.**

$$\mathbb{E}\,\Delta(n) \leq 2\,\mathbb{E}\,\Delta(\infty).$$

For a newer simple proof of Lorden's inequality, we refer to [11]. Proposition 3.6 gives us exacly what we need to conclude the proof of our main result.

Proof of Theorem 3.3 (ii). Write $p_i := \mathbb{P}(\tau_1 = i)$. We have

$$\begin{aligned}\mathbb{E}\,\Delta(\infty) &= \frac{1}{m}\sum_{i=1}^{\infty} i \sum_{j=i+1}^{\infty} p_j \\ &= \frac{1}{m}\sum_{j=2}^{\infty} p_j \sum_{i=1}^{j-1} i = \frac{1}{m}\sum_{j=2}^{\infty} p_j \frac{j(j-1)}{2} \\ &= \frac{1}{m}\mathbb{E}\frac{\tau_1(\tau_1-1)}{2} \\ &= \frac{1}{2m}\mathbb{E}\tau_1^2 - \frac{1}{2}.\end{aligned}$$

By the Lorden's theorem we obtain

$$\mathbb{E}\,\Delta(n) \le 2\,\mathbb{E}\,\Delta(\infty) \le \frac{1}{m}\mathbb{E}\tau_1^2 - 1,$$

which is just the desired conclusion. □

4. Confidence estimation

Although the MSE is an important quantity in its own right, it can also be used to construct estimates with fixed precision at a given level of confidence. Suppose the goal is to obtain an estimator $\hat{\theta}$ such that

$$\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) \le \alpha, \tag{4.1}$$

for given $\varepsilon > 0$ and $\alpha > 0$. Corollary 3.4 combined with the Chebyshev's inequality yields the following bound:

$$\mathbb{P}\left(|\hat{\theta}_{T_{R(n)}} - \theta| > \varepsilon\right) \le \frac{\sigma^2_{\mathrm{as}}(f)}{n\varepsilon^2}\left(1 + \frac{n_0}{n}\right). \tag{4.2}$$

If α is small then instead of using (4.2) directly, it is better to apply the so-called "median trick". This is a method introduced in 1986 in [27], later used in many papers concerned with computational complexity, eg. [21, 46] and further developed in [46]. The idea is to compute the median of independent estimates to boost the level of confidence. We simulate l independent copies of the Markov chain:

$$X_0^{(j)}, X_1^{(j)}, \dots, X_n^{(j)}, \dots \qquad (j = 1, \dots, l).$$

Let $\hat{\theta}^{(j)}$ be an estimator computed in jth repetition. The final estimate is $\hat{\theta} := \mathrm{med}(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(l)})$. We require that $\mathbb{P}(|\hat{\theta}^{(j)} - \theta| > \varepsilon) \le \delta$ $(j = 1, \dots, l)$ for some modest level of confidence $1 - \delta < 1 - \alpha$. This is ensured via Chebyshev's inequality. The well-known Chernoff's bound gives for odd l,

$$\mathbb{P}\left(|\hat{\theta} - \theta| \ge \varepsilon\right) \le \frac{1}{2}\left[4\delta(1-\delta)\right]^{l/2} = \frac{1}{2}\exp\left\{\frac{l}{2}\ln\left[4\delta(1-\delta)\right]\right\}. \tag{4.3}$$

In this way we obtain an exponential inequality for the probability of large deviations without requiring the underlying variables to be bounded or even to have a moment generating function. It is pointed out in [46] that under some assumptions there is a universally optimal choice of δ. More precisely, suppose that the bound on $\mathbb{P}(|\hat{\theta}^{(j)} - \theta| > \varepsilon)$ is of the form const$/n$ where n is the sample size used in a single repetition. Then the overall number of samples nl is the least if we choose $\delta^* \approx 0.11969$. The details are described in [46]. This method can be used in conjunction with our regenerative/sequential scheme. The right hand side of (4.2) *approximately* behaves like const$/n$. Therefore the following strategy is reasonably close to optimum. First choose n such that the right hand side of (4.2) is less than or equal to δ^*. Then choose l big enough to make the right hand side of (4.3), with $\delta = \delta^*$, less than or equal to α. Compute estimator $\hat{\theta}_{T_{R(n)}}$ repeatedly, using l independent runs of the chain. We can easily see that (4.1) holds if

$$\begin{aligned} &n \geq \frac{C_1 \sigma^2_{\text{as}}(f)}{\varepsilon^2} + n_0, \\ &l \geq C_2 \ln(2\alpha)^{-1} \quad \text{and } j \text{ is odd,} \end{aligned}$$

where $C_1 := 1/\delta^* \approx 8.3549$ and $C_2 := 2/\ln\left[4\delta^*(1-\delta^*)\right]^{-1} \approx 2.3147$ are absolute constants. By Theorem 3.3 (ii) the overall (expected) number of generated samples is

$$\mathbb{E}T_{R(n)} l \sim nl \sim C \frac{\sigma^2_{\text{as}}(f)}{\varepsilon^2} \log(2\alpha)^{-1}, \tag{4.4}$$

where $C = C_1 C_2 \approx 19.34$ and notation Left$(\alpha,\varepsilon) \sim$ Right(α,ε) means that Left/Right $\to 1$ as $\alpha, \varepsilon \to 0$. To see how tight are the bounds, let us compare (4.4) with the familiar asymptotic approximation, based on the CLT. We obtain

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

for the number of samples

$$n \sim \frac{\sigma^2_{\text{as}}(f)}{\varepsilon^2} \left[\Phi^{-1}(1-\alpha/2)\right]^2, \tag{4.5}$$

where $\hat{\theta}_n$ is a simple average over n Markov chain samples, Φ^{-1} is a quantile function of the standard normal distribution. Taking into account the fact that

$$\left[\Phi^{-1}(1-\alpha/2)\right]^2 \sim 2\log(2\alpha)^{-1}, \qquad (\alpha \to 0),$$

we arrive at the following conclusion. The right hand side of (4.4) is bigger than (4.5) roughly by a constant factor of about 10 (for small ε and α). The important difference is that (4.4) is sufficient for an *exact* confidence interval while (4.5) only for an *asymptotic* one.

5. Bounding the asymptotic variance I - uniformly ergodic chains

We are left with the task of bounding $\sigma^2_{\mathrm{as}}(f)$ and n_0, which appear in Theorem 3.3, by some computable quantities in typical situations of interest. The most important setting for applications - that of a geometrically ergodic Markov chain and unbounded target function f - is deferred to the next section. Here we start with uniformly ergodic chains, where a direct comparison of our approach to exponential inequalities [22, 32] is possible. We focus on [32] which is tight in the sense that it reduces to the Hoeffding bound when specialised to the i.i.d. case.

Uniform ergodicity of a Markov chain is equivalent to

$$P^h(x,\cdot) \geq \beta\nu(\cdot) \quad \text{for every} \quad x \in \mathcal{X} \quad \text{and some integer} \quad h \geq 1. \tag{5.1}$$

We refer to [50] or [44] for definitions of uniform and geometric ergodicity of Markov chains and further details related to these notions.

In the rest of this section we assume that $h = 1$ and hence (5.1) reduces to Assumption 2.1 with $J = \mathcal{X}$. This is the typical situation in applications. If $h > 1$ then P^h inherits the ergodic properties of P and one can use it for sampling. However, we acknowledge that if P^h is used, the identification of regeneration times can be problematic since the term $P^h(x, \mathrm{d}y)$ - needed to execute the Mykland et al. trick - will be typically intractable.

Computing n_0 in the setting of this Section is clearly trivial, since the overshoot is distributed as a geometric random variable with parameter β.

The problem of bounding the asymptotic variance under (5.1) was considered in [7]. Using results of their Section 5 with $h = 1$ and applying basic algebra we obtain

$$\sigma^2_{\mathrm{as}}(f) \leq \sigma^2\left(1 + \frac{2}{1-\sqrt{1-\beta}}\right) = \sigma^2\left(1 + 2\frac{1+\sqrt{1-\beta}}{\beta}\right) \leq 4\sigma^2/\beta, \tag{5.2}$$

where $\sigma^2 = \pi\bar{f}^2$ is the stationary variance.

With reversibility one can derive a better bound. An important class of reversible chains are Independence Metropolis-Hastings chains (see e.g. [50]) 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. [43, 3]) and translates into (5.1) with $h = 1$ and $\nu(\cdot) := \pi(\cdot)$. In this setting and using reversibility Atchadé and Perron [3] show that the spectrum of P, say $\mathcal{S}$, is contained in $[0, 1-\beta]$. For the general case of reversible chains satisfying (5.1) with $h = 1$ results of [49] lead to $\mathcal{S} \subseteq [-1+\beta, 1-\beta]$. By the spectral decomposition theorem for self adjoint operators (see e.g. [20, 30]) in both cases we have

$$\sigma^2_{\mathrm{as}}(f) \leq \int_{\mathcal{S}} \frac{1+s}{1-s} E_{f,P}(\mathrm{d}s) \leq \frac{2-\beta}{\beta}\sigma^2, \tag{5.3}$$

where $E_{f,P}$ is the spectral measure associated with f and P. The formula for $\sigma^2_{\mathrm{as}}(f)$ in (5.2) and (5.3) depends on β in an optimal way. Moreover (5.3) is sharp. To see this consider the following example.

5.4 EXAMPLE. Let $\beta \leq 1/2$ and define a Markov chain $(X_n)_{n\geq 0}$ on $\mathcal{X} = \{0, 1\}$ 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 $\sigma^2 = 1/4$. Now let us compute $\sigma^2_{\text{as}}(f)$.

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

To obtain an upper bound on the total simulation effort needed for $\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) \leq \alpha$ for our regenerative-sequential-median estimator $\hat{\theta}$, we now combine (5.2) and (5.3) with (4.4) to obtain respectively

$$(5.5) \qquad 19.34\frac{4\sigma^2}{\beta\varepsilon^2}\log(2\alpha)^{-1} \qquad \text{and} \qquad 19.34\frac{(2-\beta)\sigma^2}{\beta\varepsilon^2}\log(2\alpha)^{-1}.$$

From Section 4 and Example 5.4 we conclude that in (5.5) the form of functional dependence on all the parameters is optimal.

For f bounded let $\|f\|_{\text{sp}} := \sup_{x\in\mathcal{X}} f(x) - \inf_{x\in\mathcal{X}} f(x)$ and consider the exponential inequality for uniformly ergodic chains from [32]. For the simple average over n Markov chain samples, say $\hat{\theta}_n$, for an arbitrary starting point x, we have

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

After identifying leading terms in the resulting bound for the simulation effort required for $\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \leq \alpha$ and assuming, to facilitate comparisons, that $4\sigma^2 = \|f\|^2_{\text{sp}}$, we see that

$$(5.6) \qquad n \sim \frac{2\sigma^2}{\beta^2\varepsilon^2}\log(2\alpha)^{-1}.$$

Comparing (5.5) with (5.6) yields a ratio of 40β or 20β respectively. This in particular indicates that the dependence on β in [22, 32] probably can be improved. We note that in examples of practical interest β usually decays exponentially with dimension of $\mathcal{X}$ and our approach will often result in a lower total simulation cost. Moreover, in contrast to exponential inequalities of the classical form, our approach is valid for an unbounded target function f.

6. Bounding the asymptotic variance II - drift condition

In this Section we bound $\sigma^2_{\mathrm{as}}(f)$ and n_0 appearing in Theorem 3.3, by computable quantities under drift condition and with possibly unbounded f. Using drift conditions is a standard approach for establishing geometric ergodicity and our version is one of many equivalent drifts appearing in literature. Specifically, let J be the small set which appears in Assumption 2.1.

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

$$PV^2(x) := \int_{\mathcal{X}} P(x, \mathrm{d}y)V^2(y) \le \begin{cases} \lambda^2 V^2(x) & \text{for } x \notin J, \\ K^2 & \text{for } x \in J, \end{cases}$$

Unusual notation in the above drift condition is chosen to simplify further statements. Note that Assumption 6.1 entails

$$PV(x) \le \begin{cases} \lambda V(x) & \text{for } x \notin J, \\ K & \text{for } x \in J, \end{cases} \tag{6.2}$$

because by Jensen's inequality $PV(x) \le \sqrt{PV^2(x)}$. This simple observation is also exploited in [33] and [34, 35]. Assumptions 2.1 and 6.1 will allow us to derive explicit bounds on $\sigma^2_{\mathrm{as}}(f)$ and n_0 in terms of λ, β and K, provided that function $\bar{f}/V$ is bounded.

To simplify notation, let us write $T := \min\{n \ge 1 : \Gamma_{n-1} = 1\}$ for the first time of regeneration and $\Xi := \Xi_1$ for the first block. In contrast with the previous section, we will consider initial distributions of the chain different from ν and often equal to π, the stationary measure. The following proposition appears e.g. in [48] (for bounded g). The proof for nonnegative g is the same.

6.3 Proposition. *For $g : \mathcal{X} \to [0, \infty[$,*

$$\mathbb{E}\,\Xi(g)^2 = m\left[\mathbb{E}_\pi g(X_0)^2 + 2\sum_{n=1}^{\infty} \mathbb{E}_\pi g(X_0)g(X_n)\mathbb{I}(T > n)\right].$$

Our approach is based on the following result which is a slightly modified special case of Propositions 4.1 and 4.4 in [5], see also [39]. To make the paper reasonably self-contained we include the proof in Appendix A.

6.4 Proposition. *If* (6.2) *holds, then*

$$\mathbb{E}_x \sum_{n=1}^{T-1} V(X_n) \le \frac{\lambda(V(x) - 1)}{1 - \lambda} + \frac{K - \lambda}{\beta(1 - \lambda)} - 1.$$

Let us mention that a result similar to Theorem 6.5 can also be obtained using methods borrowed from [15], c.f. also [52], instead of [5]. Although in the cited papers the inequalities are derived for *coupling*, they could easily be modified to work in the context of *regeneration*. We will not pursue this, because Proposition 6.4 is easier to apply.

The main result in this section is the following.

6.5 Theorem.

(i) Under Assumption 2.1 and (6.2)*, constant n_0 in Theorem 3.3 satisfies*

$$n_0 \le 2\left[\frac{\lambda\pi(V)-\lambda}{1-\lambda}+\frac{K-\lambda}{\beta(1-\lambda)}-1\right].$$

(ii) If moreover $|\bar{f}(x)| \le V(x)$ then the asymptotic variance $\sigma^2_{\mathrm{as}}(f)$ satisfies

$$\sigma^2_{\mathrm{as}}(f) \le \frac{1+\lambda}{1-\lambda}\pi(V^2)+2\left[\frac{K-\lambda-\beta}{\beta(1-\lambda)}\right]\pi(V).$$

Proof. (i) We apply Proposition 6.3 to $g(x)=1$. Indeed, $\mathbb{E}T^2=\mathbb{E}\Xi(1)^2$ and

$$\begin{aligned}\mathbb{E}T^2/m &= \mathbb{E}\Xi(1)^2/m\\ &\le 1+2\mathbb{E}_\pi\sum_{n=1}^{T-1}V(X_n)\\ &\le 1+2\left[\lambda\frac{\pi(V)-1}{1-\lambda}+\frac{K-\lambda}{\beta(1-\lambda)}-1\right].\end{aligned}$$

by Proposition 6.4. The result follows because $n_0=\mathbb{E}T^2/m-1$.

(ii) By Proposition 6.3 we have

$$\begin{aligned}\sigma^2_{\mathrm{as}}(f) &= \mathbb{E}\Xi(\bar{f})^2/m \le \mathbb{E}\Xi(V)^2/m\\ &= \mathbb{E}_\pi V(X_0)^2+2\mathbb{E}_\pi\sum_{n=1}^{T-1}V(X_0)V(X_n)\end{aligned}$$

We will use Proposition 6.4 to bound the second term.

$$\begin{aligned}&\mathbb{E}_\pi\sum_{n=1}^{T-1}V(X_0)V(X_n)=\mathbb{E}_\pi V(X_0)\mathbb{E}(\sum_{n=1}^{T-1}V(X_n)|X_0)\\ &=\int_{\mathcal{X}}\pi(\mathrm{d}x)V(x)\mathbb{E}_x\sum_{n=1}^{T-1}V(X_n)\\ &\le\int_{\mathcal{X}}\pi(\mathrm{d}x)V(x)\left(\frac{\lambda(V(x)-1)}{1-\lambda}+\frac{K-\lambda}{\beta(1-\lambda)}-1\right)\\ &=\frac{\lambda}{1-\lambda}\pi(V^2)+\left[\frac{K-\lambda-\lambda\beta}{\beta(1-\lambda)}-1\right]\pi(V).\end{aligned}$$

Putting everything together, we obtain

$$\sigma^2_{\mathrm{as}}(f)\le\pi(V^2)+\frac{2\lambda}{1-\lambda}\pi(V^2)+2\left[\frac{K-\lambda-\lambda\beta}{\beta(1-\lambda)}-1\right]\pi(V),$$

which is equivalent to the desired conclusion. □

Note that in Theorem 6.5 we need only (6.2), that is the drift condition on V. Asssumption 6.1 is needed only to get a bound on $\pi(V^2)$. Indeed, it implies that $\pi V^2 = \pi P V^2 \leq \lambda^2(\pi V^2 - \pi(J)) + K^2\pi(J)$, so

$$\pi V^2 \leq \pi(J)\frac{K^2-\lambda^2}{1-\lambda^2} \leq \frac{K^2-\lambda^2}{1-\lambda^2}.$$

Analogously, (6.2) implies

$$\pi V \leq \pi(J)\frac{K-\lambda}{1-\lambda} \leq \frac{K-\lambda}{1-\lambda}.$$

Our final estimates are therefore the following.

6.6 Corollary.

(i) Under Assumptions 2.1 and 6.1,

$$n_0 \leq \frac{2}{(1-\lambda)\beta}\left[K\frac{1-\lambda(1-\beta)}{1-\lambda} - \beta\left(1+\frac{\lambda^2}{1-\lambda}\right) - \lambda\right].$$

(ii) If $\|\bar{f}\|_V := \sup_x |\bar{f}(x)|/V(x) < \infty$ *then*

$$\sigma^2_{\text{as}}(f) \leq \|\bar{f}\|^2_V \frac{K^2(2+\beta) - 2K(2\lambda+\beta) + 2\lambda^2 + 2\lambda\beta - \lambda^2\beta}{(1-\lambda)^2\beta}.$$

(iii) Moreover, $\|\bar{f}\|_V$ *can be related to* $\|f\|_V$ *by*

$$\|\bar{f}\|_V \leq \|f\|_V + \frac{\pi(J)(K-\lambda)}{(1-\lambda)\inf_{x\in\mathcal{X}} V(x)} \leq \|f\|_V + \frac{K-\lambda}{1-\lambda}.$$

Proof. To prove (iii) we compute

$$\begin{aligned} \|\bar{f}\|_V &= \sup_{x\in\mathcal{X}} \frac{|f(x)-\pi f|}{V(x)} \leq \sup_{x\in\mathcal{X}} \frac{|f(x)|+|\pi f|}{V(x)} \\ &\leq \sup_{x\in\mathcal{X}}\left(\|f\|_V + \frac{\pi V}{V(x)}\right) \leq \|f\|_V + \frac{\pi(J)(K-\lambda)}{(1-\lambda)\inf_{x\in\mathcal{X}} V(x)} \\ &\leq \|f\|_V + \frac{K-\lambda}{1-\lambda}. \end{aligned}$$

□

6.7 REMARK. In many specific examples one can obtain (with some additional effort) sharper bounds for πV, πV^2, $\|\bar{f}\|_V$ or at least bound $\pi(J)$ away from 1. However in general we assume that such bounds are not available and one will use Corollary 6.6.

7. Example

The simulation experiments described below are designed to compare the bounds proved in this paper with actual errors of MCMC estimation. Assume that $y = (y_1, \dots, y_t)$ is an i.i.d. sample from the normal distribution $\mathrm{N}(\mu, \kappa^{-1})$, where κ denotes the reciprocal of the variance. Thus we have

$$p(y|\mu,\kappa) = p(y_1,\dots,y_t|\mu,\kappa) \propto \kappa^{t/2} \exp\left[-\frac{\kappa}{2}\sum_{j=1}^{t}(y_j-\mu)^2\right].$$

The pair (μ, κ) plays the role of unknown parameter. To make things simple, let us consider "uninformative improper priors" that is assume that $p(\mu, \kappa) = p(\mu)p(\kappa) \propto \kappa^{-1}$. The posterior density is then

$$\begin{aligned} p(\mu,\kappa|y) &\propto p(y|\mu,\kappa)p(\mu,\kappa) \\ &\propto \kappa^{t/2-1}\exp\left[-\frac{\kappa t}{2}\left(s^2+(\bar{y}-\mu)^2\right)\right], \end{aligned}$$

where

$$\bar{y} = \frac{1}{t}\sum_{j=1}^{t} y_j, \quad s^2 = \frac{1}{t}\sum_{j=1}^{t}(y_j-\bar{y})^2.$$

Note that $\bar{y}$ and s^2 only determine the location and scale of the posterior. We will be using a Gibbs sampler, whose performance does not depend on scaling and location, therefore without loss of generality we can assume that $\bar{y} = 0$ and $s^2 = t$. Since $y = (y_1, \dots, y_t)$ is kept fixed, let us slightly abuse notation by using symbols $p(\kappa|\mu)$, $p(\mu|\kappa)$ and $p(\mu)$ for $p(\kappa|\mu, y)$, $p(\mu|\kappa, y)$ and $p(\mu|y)$, respectively. Now, the Gibbs sampler consists of drawing samples intermittently from both the conditionals. Start with some (μ_0, κ_0). Then, for $i = 1, 2, \dots$,

- $\kappa_i \sim \mathrm{Gamma}\left(t/2, (t/2)(s^2 + \mu_{i-1}^2)\right)$,
- $\mu_i \sim \mathrm{N}\left(0, 1/(\kappa_i t)\right)$.

If we are chiefly interested in μ then it is convenient to consider the two small steps $\mu_{i-1} \to \kappa_i \to \mu_i$ together. The transition density is

$$\begin{aligned} p(\mu_i|\mu_{i-1}) &= \int p(\mu_i|\kappa)p(\kappa|\mu_{i-1})\mathrm{d}\kappa \\ &\propto \int_0^\infty \kappa^{1/2}\exp\left[-\frac{\kappa t}{2}\mu_i^2\right] \times \\ &\qquad \times \left(s^2+\mu_{i-1}^2\right)^{t/2}\kappa^{t/2-1}\exp\left[-\frac{\kappa t}{2}\left(s^2+\mu_{i-1}^2\right)\right]\mathrm{d}\kappa \\ &= \left(s^2+\mu_{i-1}^2\right)^{t/2}\int_0^\infty \kappa^{(t-1)/2}\exp\left[-\frac{\kappa t}{2}\left(s^2+\mu_{i-1}^2+\mu_i^2\right)\right]\mathrm{d}\kappa \\ &\propto \left(s^2+\mu_{i-1}^2\right)^{t/2}\left(s^2+\mu_{i-1}^2+\mu_i^2\right)^{-(t+1)/2}. \end{aligned}$$

The proportionality constants concealed behind the $\propto$ sign depend only on t. Finally we fix scale letting $s^2 = t$ and get

$$p(\mu_i|\mu_{i-1}) \propto \left(1 + \frac{\mu_{i-1}^2}{t}\right)^{t/2} \left(1 + \frac{\mu_{i-1}^2}{t} + \frac{\mu_i^2}{t}\right)^{-(t+1)/2}. \tag{7.1}$$

If we consider the RHS of (7.1) as a function of μ_i only, we can regard the first factor as constant and write

$$p(\mu_i|\mu_{i-1}) \propto \left(1 + \left(1 + \frac{\mu_{i-1}^2}{t}\right)^{-1} \frac{\mu_i^2}{t}\right)^{-(t+1)/2}.$$

It is clear that the conditional distribution of random variable

$$\mu_i \left(1 + \frac{\mu_{i-1}^2}{t}\right)^{-1/2} \tag{7.2}$$

is t-Student distribution with t degrees of freedom. Therefore, since the t-distribution has the second moment equal to $t/(t-2)$ for $t > 2$, we infer that

$$\mathbb{E}(\mu_i^2|\mu_{i-1}) = \frac{t + \mu_{i-1}^2}{t-2}.$$

Similar computation shows that the posterior marginal density of μ satisfies

$$p(\mu) \propto \left(1 + \frac{t-1}{t}\frac{\mu^2}{t-1}\right)^{-t/2}.$$

Thus the stationary distribution of our Gibbs sampler is rescaled t-Student with $t-1$ degrees of freedom. Consequently we have

$$\mathbb{E}_\pi \mu^2 = \frac{t}{t-3}.$$

7.3 Proposition (Drift)**.** *Assume that $t \geq 4$. Let*

$$V^2(\mu) := \mu^2 + 1$$

and $J = [-a, a]$. The transition kernel of the (2-step) Gibbs sampler satisfies

$$PV^2(\mu) \leq \begin{cases} \lambda^2 V^2(\mu) & \text{for } |\mu| > a; \\ K^2 & \text{for } |\mu| \leq a, \end{cases}$$

provided that $a > \sqrt{t/(t-3)}$. The quantities λ and K are given by

$$\lambda^2 = \frac{1}{t-2}\left(\frac{2t-3}{1+a^2} + 1\right),$$

$$K^2 = 2 + \frac{a^2+2}{t-2}.$$

Moreover,

$$\pi(V^2) = \frac{2t-3}{t-3}.$$

Proof. It is enough to use the fact that

$$PV^2(\mu) = \mathbb{E}(\mu_i^2 + 1|\mu_{i-1} = \mu) = \frac{t + \mu^2}{t - 2} + 1$$

and some simple algebra. Analogously, $\pi(V^2) = \mathbb{E}_\pi \mu^2 + 1$. □

7.4 Proposition (Minorization)**.** *Let $p_{\min}$ be a subprobability density given by*

$$p_{\min}(\mu) = \begin{cases} p(\mu|a) & \text{for } |\mu| \le h(a); \\ p(\mu|0) & \text{for } |\mu| > h(a), \end{cases}$$

where $p(\cdot|\cdot)$ is the transition density given by (7.1) *and*

$$h(a) = \left\{ a^2 \left[\left(1 + \frac{a^2}{t}\right)^{t/(t+1)} - 1 \right]^{-1} - t \right\}^{1/2}.$$

Then $|\mu_{i-1}| \le a$ implies $p(\mu_i|\mu_{i-1}) \ge p_{\min}(\mu_i)$. Consequently, if we take for ν the probability measure with the normalized density $p_{\min}/\beta$ then the small set Assumption 2.1 holds for $J = [-a, a]$. Constant β is given by

$$\beta = 1 - \mathbb{P}\left(|\vartheta| \le h(a)\right) + \mathbb{P}\left(|\vartheta| \le \left(1 + \frac{a^2}{t}\right)^{-1/2} h(a)\right),$$

where ϑ is a random variable with t-Student distribution with t degrees of freedom.

Proof. The formula for $p_{\min}$ results from minimization of $p(\mu_i|\mu_{i-1})$ with respect to $\mu_{i-1} \in [-a, a]$. We use (7.1). First compute $(\mathrm{d}/\mathrm{d}\mu_{i-1}) p(\mu_i|\mu_{i-1})$ to check that the function has to attain minimum either at 0 or at a. Thus

$$p_{\min}(\mu) = \begin{cases} p(\mu|a) & \text{if } p(\mu|a) \le p(\mu|0); \\ p(\mu|0) & \text{if } p(\mu|a) > p(\mu|0). \end{cases}$$

Now it is enough to solve the inequality, say, $p(\mu|a) \le p(\mu|0)$ with respect to μ. Elementary computation shows that this inequality is fulfiled iff $\mu \le h(a)$. The formula for β follows from (7.2) and from the fact that

$$\beta = \int p_{\min}(\mu)\mathrm{d}\mu = \int_{|\mu| \le h(a)} p(\mu|a)\mathrm{d}\mu + \int_{|\mu| > h(a)} p(\mu|0)\mathrm{d}\mu.$$

□

7.5 REMARK. It is interesting to compare the asymptotic behavior of the constants in Propositions 7.3 and 7.4 for $a \to \infty$. We can immediately see that $\lambda^2 \to 1/(t-2)$ and $K^2 \sim a^2/(t-2)$. Slightly more tedious computation reveals that $h(a) \sim \text{const} \cdot a^{1/(t+1)}$ and consequently $\beta \sim \text{const} \cdot a^{-t/(t+1)}$.

The parameter of interest is the posterior mean (Bayes estimator of μ). Thus we let $f(\mu) = \mu$ and $\theta = \mathbb{E}_\pi \mu$. Note that our chain $\mu_0, \dots, \mu_i, \dots$ is a zero-mean martingale, so $\bar{f} = f$ and

$$\sigma^2_{\mathrm{as}}(f) = \mathbb{E}_\pi(f^2) = \frac{t}{t-3}.$$

Obviously we have $\|f\|_V = 1$.

In the experiments described below, $t = 50$ is kept fixed. Other experiments (not reported here) show that the value of t has little influence on the results. Table 1 illustrates inequalities in Theorem 3.3 and Corollary 3.4. The actual values of the MSE of our estimator and the mean overshoot, viz.

$$\begin{aligned} \mathrm{MSE} &:= \mathbb{E}\,(\hat{\theta}_{T_{R(n)}} - \theta)^2, \\ \mathrm{OS} &:= \mathbb{E}\,T_{R(n)} - n, \end{aligned}$$

are computed empirically, using 10000 repetitions of the experiment. They can be compared with the bounds in 3.3 and 3.4, named henceforth

$$\begin{aligned} \mathrm{BoundMSE} &:= \frac{\sigma^2_{\mathrm{as}}(f)}{n}\left(1 + \frac{n_0}{n}\right) \\ \mathrm{BoundOS} &:= n_0 = \frac{\mathbb{E}\tau_1^2}{\mathbb{E}\tau_1} - 1. \end{aligned}$$

In these formulas, we use the true value of $\sigma^2_{\mathrm{as}}(f)$, for which we have an analytical expression. Also $m = \mathbb{E}\tau_1 = \pi(J)\beta$ is computed exactly while $\mathbb{E}\tau_1^2$ is approximated via a separate (very long) series of simulations. given for two choices of the "small set" $J = [-a, a]$. We also show values of m (mean length of a regeneration cycle) and β (probability of regeneration).

n	a	MSE	BoundMSE	OS	BoundOS	m	β
10	5	0.1062	0.1087	0.1099	0.2134	1.1072	0.9032
100		0.0105	0.0107	0.1037			
1000		0.0011	0.0011	0.1073			
10	100	0.0821	0.2247	5.4768	11.1196	6.5043	0.1537
100		0.0102	0.0118	5.4871			
1000		0.0011	0.0011	5.4337			

Table 1. Actual values of the MSE and mean overshoot vs. bounds 3.3 and 3.4

Table 1 clearly shows that the inequalities in Theorem 3.3 are quite sharp. The bound on MSE, which is of primary interest, becomes almost exact for large n. The bound on the mean overshoot, which can be used to estimate the cost of the algorithm, is also very satisfactory.

We now proceed to the inequalities proved in Section 6 under the drift condition, Assumption 6.1. The final bounds in Corollary 6.6 are expressed in terms of the computable drift/minorization parameters, that is λ, K and β. We also examine how the tightness of the final bounds is influenced by replacing the true value of πV^2 by its upper bound. To this end we compute the bounds given in Theorem 6.5, using the knowledge of πV^2. In our example we compute λ, K, β and also πV^2 via Propositions 7.3 and 7.4 for different choices of $J = [-a, a]$. Parameter $t = 50$ is fixed.

Figure 1 shows how the two bounds on $\sigma^2_{\text{as}}(f)$ depend on a. The black line corresponds to the bound of Corollary 6.6 (ii) which involves only λ, K and β. The grey line gives the bound of Theorem 6.5 (ii) which assumes the knowledge of πV^2. The best values of both bounds, equal to 7.19 and 5.66, correspond to $a = 3.93$ and $a = 4.33$, respectively. The actual value of the asymptotic variance is $\sigma^2_{\text{as}}(f) = 1.064$.

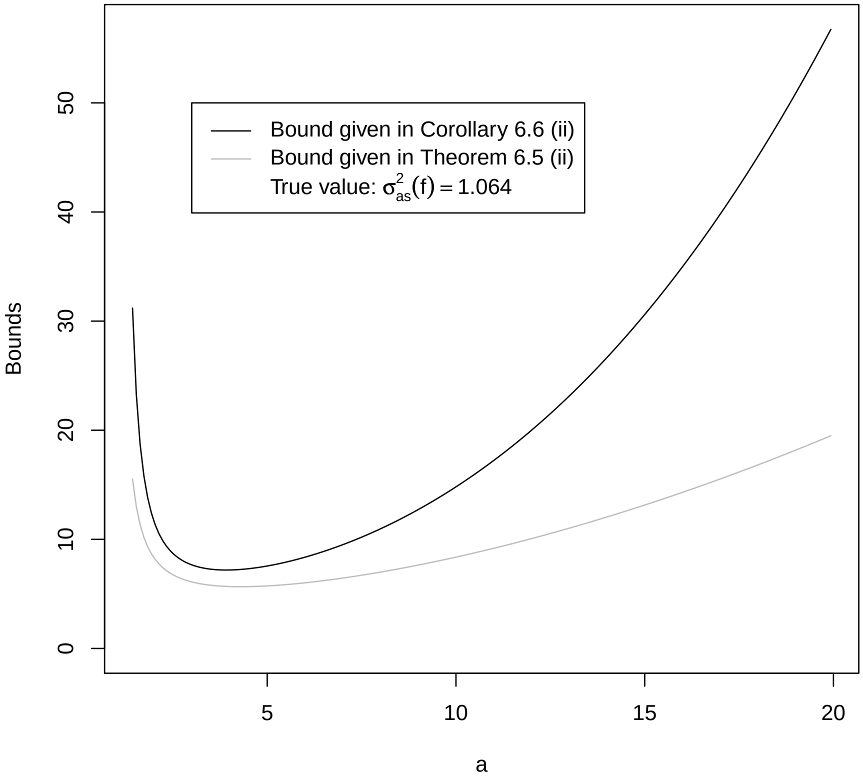

Figure 1. Bounds for the asymptotic variance $\sigma^2_{\text{as}}(f)$ as functions of a.

Figure 2 is analogous and shows two bounds on n_0. Again, the black bound involves only the drift/minorization parameters while the grey one assumes the knowlegde of πV^2. The best bounds, 2.94 and 2.50, obtain for $a = 4.73$ and $a = 4.33$, respectively.

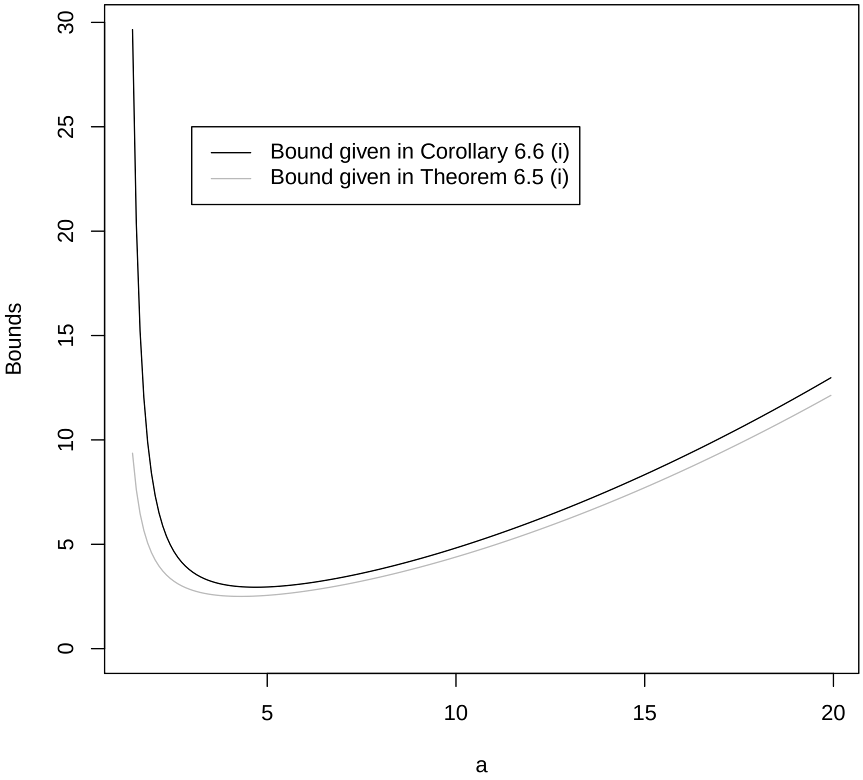

Figure 2. Bounds for n_0 as functions of a.

In contrast with the inequalities of Theorem 3.3, the bounds of Theorem 6.5 depend significantly on t, the size of sample behind the posterior distribution. In Table 2 below we summarize the best (with respect to a) bounds on $\sigma^2_{\text{as}}(f)$ for three values of t.

t	$\sigma^2_{\text{as}}(f)$	Bound 6.5 (ii)	Bound 6.6 (ii)
5	2.500	141.50	41.02
50	1.064	7.19	5.66
500	1.006	4.33	3.99

Table 2. Values of $\sigma^2_{\text{as}}(f)$ vs. bounds 6.5 and 6.6 for different values of t.

This clearly identifies the bottleneck of the approach: the bounds on $\sigma^2_{\mathrm{as}}(f)$ under drift condition in Theorem 6.5 and Corollary 6.6 can vary widely in their sharpness in specific examples. We conjecture that this may be the case in general for any bounds derived under drift conditions. Known bounds on the rate of convergence (e.g. in total variation norm) obtained under drift conditions are often very pessimistic, too (e.g. [5, 51, 29]). However, at present, drift conditions remain the main and most universal tool for proving computable bounds for Markov chains on continuous spaces. An alternative might be working with conductance but to the best of our knowledge, so far this approach has been applied successfully only to examples with compact state spaces (see e.g. [55, 42] and references therein).

8. Connections with other results

Our aim was to obtain nonasymptotic results concerning the mean square error and confidence estimation in a possibly general setting relevant for MCMC applications in Bayesian statistics. We now discuss our results in context of related work.

8.1. Related nonasymptotic results

A vast literature on nonasymptotic analysis of Markov chains is available in various settings. To place our results in this context we give a brief account, which by no means is extensive. In the case of finite state space, an approach based on the spectral decomposition was used in [2, 21, 37, 46] to derive results of related type. For bounded functions and uniformly ergodic chains on a general state space, exponential inequalities with explicit constants such as those in [22, 32] can be applied to derive confidence bounds. Comparison of the required simulation effort for the same confidence interval (Sections 4 and 5) shows that while exponential inequalities have sharper constants, our approach gives in this setting the optimal dependence on the regeneration rate β and therefore can turn out more efficient in many practical examples.

Related results come also from studying concentration of measure phenomenon for dependent random variables. For the large body of work in this area see e.g. [41], [56] and [31] (and references therein), where transportation inequalities or martingale approach have been used. These results, motivated in a more general setting, are valid for Lipschitz functions with respect to the Hamming metric. They also include expressions $\sup_{x,y\in\mathcal{X}} \|P^i(x,\cdot) - P^i(y,\cdot)\|_{\mathrm{tv}}$ and when applied to our setting, they are well suited for bounded functionals of uniformly ergodic Markov chains, but can not be applied to geometrically ergodic chains. For details we refer to the original papers and the discussion in Section 3.5 of [1].

For lazy reversible Markov chains, nonasymptotic mean square error bounds have been obtained for *bounded* target functions in [55] in a setting where explicit bounds on conductance are available. These results have been applied to approximating integrals over balls in $\mathbb{R}^d$ under some regularity conditions for the stationary measure, see [55] for details. The Markov chains considered there are in fact uniformly ergodic, however in their problem when establishing (5.1), β turns out to be exponentially small and $h > 1$, hence conductance seems to be the natural approach to make the problem tractable in high dimensions.

Tail inequalities for *bounded* functionals of Markov chains that are not uniformly ergodic were considered in [12], [1] and [16] using regeneration techniques. These results apply e.g. to geometrically or subgeometrically ergodic Markov chains, however they also involve non-explicit constants or require tractability of moment conditions of random tours between regenerations. Computing explicit bounds from these results may be possible with additional work, but we do not pursue it here.

Tail inequalities for unbounded target function f that can be applied to geometrically ergodic Markov chains have been established by Bertail and Clémençon in [10] by regenerative approach and using truncation arguments. However they involve non-explicit constants and can not be directly applied to confidence estimation.

Rates of convergence of geometrically ergodic Markov chains to their stationary distributions have been investigated in many papers. The typical setting is similar to our Section 6, i.e. one assumes a geometrical drift to a small set and a one step minorization condition. Moreover, to establish convergence rates one requires an additional condition that implies aperiodicity, which was not needed for our purposes. Most of the authors focus either on the total variation distance [50, 52, 51, 29, 53] or its weighted version [18, 5]. Such results, although of utmost theoretical importance, do not directly translate into bounds on the accuracy of estimation, because they allow us to control only the bias of estimates and the so-called burn-in time. Moreover we note, that in the drift condition setting

- convergence to stationarity is in fact not needed, we only require a bound on the asymptotic variance and on the overshoot, c.f. Section 6,
- to obtain explicit convergence rates, some version of our Proposition 6.4 is always needed (c.f. for example Section 4 of [5]) and it is in fact one of several steps required for the bound, whereas we are using Proposition 6.4 directly, avoiding other steps that could weaken the results.

8.2. Nonasymptotic vs asymptotic confidence estimation

Since nonasymptotic analysis of complicated Markov chains appears difficult, practitioners often validate MCMC estimation by a convergence diagnostics (see e.g. [13, 19] and references therein). It is however well-known that this may lead to overoptimistic conclusions, stopping the simulation far to early, and introducing bias [14, 40]. Designing asymptotic confidence intervals based on CLTs for Markov chains is often perceived as a reasonable trade-off between rigorous analysis of the algorithm and convergence heuristics and is referred to as honest MCMC estimation, c.f. [20, 25].

In what follows we argue that the nonasymptotic confidence estimation presented in the current paper requires verifying essentially the same assumptions as asymptotic confidence estimation. We also compare implementational difficulties.

Asymptotic confidence estimation for Markov chains is done e.g. by establishing Edgeworth expansions (see [8, 9]) or by applying the Glynn and Whitt sequential procedure [23] in the Markov chain context. Both methods rely heavily on strongly consistent estimation of the asymptotic variance. There has been a lot of work done recently to analyse asymptotic variance estimators for Markov chains and enable strongly consistent estimation under tractable assumptions [17, 28, 6, 26, 9, 8]. In particular we note the following.

- The most commonly used regenerative estimators (see e.g. [28, 26, 8, 9]) are known to be strongly consistent for geometrically ergodic Markov chains that satisfy a one step minorization condition and an integrability condition $\mathbb{E}_\pi|f|^{2+\delta} < \infty$ (Proposition 1 of [28]).
- Similarly, the non-overlapping and overlapping batch means estimators (see e.g. [28, 17]) are known to be strongly consistent for geometrically ergodic Markov chains that satisfy a one step minorization condition and an integrability condition $\mathbb{E}_\pi|f|^{2+\delta} < \infty$ (Proposition 4 of [6] and Theorem 2 of [17] respectively).
- Spectral variance estimators are known to be strongly consistent for geometrically ergodic Markov chains that satisfy a one step minorization condition and an integrability condition $\mathbb{E}_\pi|f|^{4+\delta} < \infty$ (Theorem 1 of [17]).

We note that geometrical ergodicity is typically established by a drift condition similar to the one used in Section 6 and the one step minorization condition usually boils down to our Assumption 2.1. As for integrability conditions, the drift condition implies $\pi V^2 < \infty$ and we require $f^2 < V^2$. Checking $E_\pi|f|^{2+\delta} < \infty$ will be typically done by ensuring $|f|^{2+\delta} < V^2$ and is therefore comparable, whereas the condition $E_\pi|f|^{4+\delta} < \infty$ for spectral variance estimation is clearly stronger.

On the algorithmic side, regenerative asymptotic variance estimators require identifying regenerations, exactly as we do, whereas batch means and spectral variance estimators do not require this.

Therefore we conclude, that if regenerations are identifiable, the price for the rigorous, nonasymptotic result is only as high as the difference between $\sigma^2_{\text{as}}(f)$ and its upper bounds e.g. those in Section 6.

Acknowledgements

Discussions with Jacek Wesołowski helped prepare an early version of this paper. The authors are also grateful to the anonymous referee for his/her constructive comments.

Appendix A: Proof of Proposition 6.4

Proof. Under (6.2) and Assumption 2.1 we are to establish

$$\mathbb{E}_x \sum_{n=1}^{T-1} V(X_n) \leq \frac{\lambda(V(x)-1)}{1-\lambda} + \frac{(K-\lambda)}{\beta(1-\lambda)} - 1.$$

The idea is to decompose the sum into shorter blocks, such that each block ends at a visit to J. Let $S := S_0 := \min\{n \geq 0 : X_n \in J\}$ and $S_j := \min\{n > S_{j-1} : X_n \in J\}$ for $j = 1, 2, \ldots$. Introduce the following notations:

$$H(x) := \mathbb{E}_x \sum_{n=0}^{S} V(X_n), \text{ for } x \in \mathcal{X},$$

$$\tilde{H} := \sup_{x \in J} \mathbb{E}_x \left(\sum_{n=1}^{S_1} V(X_n) \middle| \Gamma_0 = 0 \right) = \sup_{x \in J} \int Q(x, \mathrm{d}y) H(y).$$

Note that $H(x) = V(x)$ for $x \in J$ and that Q denotes the normalized "residual kernel".

Let us first bound $H(x)$. It is easy to check that under (6.2), for every initial distribution, $V(X_{n \wedge S})/\lambda^{n \wedge S}$ for $n = 0, 1, \ldots$ is a supermartingale with respect to $\mathcal{F}_n := \sigma(X_0, \ldots, X_n)$. Therefore $\mathbb{E}_x V(X_{n \wedge S})/\lambda^{n \wedge S} \leq V(x)$ for every $x \in \mathcal{X}$ and $n = 0, 1, \ldots$. This inequality can be multiplied by λ^n and rewiritten as follows:

$$\mathbb{E}_x V(X_S) \lambda^{n-S} \mathbb{I}(S < n) + \mathbb{E}_x V(X_n) \mathbb{I}(n \leq S) \leq \lambda^n V(x).$$

Now take a sum over $n = 0, 1, \ldots$ to obtain

$$\mathbb{E}_x V(X_S) \sum_{n=S+1}^{\infty} \lambda^{n-S} + \mathbb{E}_x \sum_{n=0}^{S} V(X_n) \leq V(x) \sum_{n=0}^{\infty} \lambda^n$$

or, equivalently,

$$\text{(A.1)} \qquad \mathbb{E}_x V(X_S) \frac{\lambda}{1-\lambda} + H(x) \leq V(x) \frac{1}{1-\lambda}.$$

Consequently, since $\mathbb{E}_x V(X_S) \geq 1$, we have for exery x,

$$\text{(A.2)} \qquad H(x) \leq \frac{V(x) - \lambda}{1-\lambda}.$$

From (6.2) we obtain $PV(x) = (1-\beta)QV(x) + \beta\nu V \leq K$ for $x \in J$, so $QV(x) \leq (K-\beta)/(1-\beta)$ and, taking into account (A.2),

$$\text{(A.3)} \qquad \tilde{H} \leq \frac{(K-\beta)/(1-\beta) - \lambda}{1-\lambda} = \frac{K - \lambda - \beta(1-\lambda)}{(1-\lambda)(1-\beta)}.$$

Recall that $T := \min\{n \geq 1 : \Gamma_{n-1} = 1\}$. For $x \in J$ we thus have

$$\begin{aligned}
&\mathbb{E}_x \sum_{n=1}^{T-1} V(X_n) \\
&= \mathbb{E}_x \sum_{j=1}^{\infty} \sum_{n=S_{j-1}+1}^{S_j} V(X_n) \mathbb{I}(\Gamma_{S_0} = \cdots = \Gamma_{S_{j-1}} = 0) \\
&= \mathbb{E}_x \sum_{j=1}^{\infty} \mathbb{E} \left(\sum_{n=S_{j-1}+1}^{S_j} V(X_n) \middle| \Gamma_{S_0} = \cdots = \Gamma_{S_{j-1}} = 0 \right) (1-\beta)^j \\
&\leq \sum_{j=1}^{\infty} \tilde{H}(1-\beta)^j \leq \frac{K-\lambda}{\beta(1-\lambda)} - 1,
\end{aligned}$$

by (A.3). For $x \notin J$ we have to add one more term at the beginning:

$$\begin{aligned}
&\mathbb{E}_x \sum_{n=1}^{T-1} V(X_n) = \mathbb{E}_x \sum_{n=1}^{S_0} V(X_n) \\
&+ \mathbb{E}_x \sum_{j=1}^{\infty} \sum_{n=S_{j-1}+1}^{S_j} V(X_n) \mathbb{I}(\Gamma_{S_0} = \cdots = \Gamma_{S_{j-1}} = 0).
\end{aligned}$$

This extra term is equal to $H(x) - V(x)$ and we can use (A.2) to bound it. Finally we obtain

$$\text{(A.4)} \qquad \mathbb{E}_x \sum_{n=1}^{T-1} V(X_n) \leq \frac{\lambda(V(x)-1)}{1-\lambda} \mathbb{I}(x \notin J) + \frac{K-\lambda}{\beta(1-\lambda)} - 1.$$

□

References

[1] R. Adamczak (2008): A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability.* 34, 1000–1034.

[2] D. Aldous (1987): On the Markov Chain Simulation Method for Uniform Combinatorial Distributions and Simulated Annealing, *Probability in the Engineering and Informational Science*, pp. 33–45.

[3] Y.F. Atchade, F. Perron (2007): On the geometric ergodicity of Metropolis-Hastings algorithms. *Statistics* 41, 77–84.

[4] 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.

[5] P.H. Baxendale (2005): Renewal Theory and Computable Convergence Rates for Geometrically Ergodic Markov Chains. *Ann. Appl. Prob.* 15, 700-738.

[6] W. Bednorz, K. Latuszyński (2007): A few Remarks on "Fixed-Width Output Analysis for Markov Chain Monte Carlo" by Jones et al. *Journal of the American Statatistical Association* 102 (480), 1485–1486.

[7] W. Bednorz, R. Latała and K. Latuszyński (2008): A Regeneration Proof of the Central Limit Theorem for Uniformly Ergodic Markov Chains. *Elect. Comm. in Probab.* 13, 85–98.

[8] P. Bertail, S. Clémençon (2004): Edgeworth expansions of suitably normalized sample mean statistics for atomic Markov chains, *Probability Theory and Related Fields*, 129, 388-414.

[9] P. Bertail, S. Clémençon (2006): Regeneration-based statistics for Harris recurrent Markov chains, pages 1–54. Number 187 in *Lecture notes in Statistics.* Springer.

[10] P. Bertail, S. Clémençon (2009): Sharp bounds for the tail of functionals of Markov chains, to appear *Probability Theory and its applications.*

[11] J.T. Chang (1994): Inequalities for the overshoot. *Ann. Appl. Probab.* 4, 1223–1233.

[12] S.J.M. Clémençon (2001): Moment and probability inequalities for sums of bounded functionals of regular Markov chains via the Nummelin splitting technique. *Statist. Probab. Lett.* 55, 227–238.

[13] M.K. Cowles, B.P. Carlin (1996): Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *J. Amer. Stat. Assoc* 91, 883–904.

[14] M.K. Cowles, G.O. Roberts, J.S. Rosenthal (1999): Possible niases induced by MCMC convergence diagnostics. *J. Stat Comput. Sim.* 64, 87–104.

[15] R. Douc, E. Moulines and J.R. Rosenthal (2004): Quantitative bounds on convergence of time-inhomogeneous Markov Chains, *Ann. Appl. Probab.* 14, 1643-1665.

[16] R. Douc, A. Guillin and E. Moulines (2008): Bounds on regeneration times and limit theorems for subgeometric Markov chains, *Ann. Inst. H. Poincar Probab. Statist.* 44, 239–257.

[17] J.M. Flegal, G.L. Jones (2009): Batch Means and Spectral Variance Estimators in Markov Chain Monte Carlo. Technical report, University of Minnesota.

[18] G. Fort (2002): Computable bounds for V-geometric ergodicity of Markov transition kernels. Preprint.

[19] A. Gelman, D.B. Rubin (1992): Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7 (4), 457–472.

[20] C. J. Geyer (1992): Practical Markov Chain Monte Carlo. *Stat. Sci.* 7 (4), 473–511.

[21] D. Gillman: A Chernoff bound for random walks on expander graphs, *SIAM J. Comput.* 27, 4, pp. 1203–1220, 1998.

[22] P.W. Glynn and D. Ormoneit (2002): Hoeffding's inequality for uniformly ergodic Markov chains, *Statist. Probab. Lett.* 56, 143–146.

[23] P.W. Glynn, W. Whitt (1992): The Asymptotic Validity of Sequential Stopping Rules for Stochastic Simulations. *The Annals of Applied Probability.* 2, 180–198.

[24] J.P. Hobert and C.J. Geyer: Geometric ergodicity of Gibbs and block Gibbs

samplers for Hierarchical Random Effects Model. *J. Multivariate Anal.* 67, 414–439.

[25] J.P. Hobert, G.L. Jones: Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo. *Statistical Science* 16(4), pp. 312–334, 2001.

[26] J.P. Hobert, G.L. Jones, B. Presnell, and J.S. Rosenthal: On the Applicability of Regenerative Simulation in Markov Chain Monte Carlo. *Biometrika* 89, pp. 731-743, 2002.

[27] M.R. Jerrum, L.G. Valiant, V.V. Vizirani: Random generation of combinatorial structures fro, a uniform distribution. *Theoretical Computer Science* 43, 169–188, 1986.

[28] G.L. Jones, M. Haran, B.S. Caffo and R. Neath (2006): Fixed-width output analysis for Markov chain Monte Carlo, *J. Amer. Statist. Association*, 101, 1537–1547

[29] G.J. Jones, J.P. Hobert: Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *Ann. Statist.* 32, pp. 784–817, 2004.

[30] C. Kipnis, S.R.S. Varadhan, (1986): Central Limit Theorem for Additive Functionals of Reversible Markov Processe and Applications to Simple Exclusions. *Commun. Math. Phys.* 104, 1–19.

[31] L. Kontorovich, K. Ramanan (2008): Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.* 36 (6), 2126–2158.

[32] 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.*

[33] K. Latuszyński (2008): Regeneration and Fixed-Width Analysis of Markov Chain Monte Carlo Algorithms. PhD Dissertation. Available at arXiv:0907.4716v1

[34] K. Latuszyński and W. Niemiro (2006): $(\varepsilon - \alpha)$-MCMC approximation under drift condition, in: Proceedings of the 6th International Workshop on Rare Event Simulation (RESIM 2006).

[35] K. Latuszyński, W. Niemiro (2008): Fixed width MCMC algorithms with application to Gibbs sampler for a hierarchical random effects model. Submitted.

[36] K. Latuszyński, W. Niemiro (2009): Nonasymptotic validity of MCMC estimators. In preparation.

[37] C.A. León, F. Perron (2004): Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.* 14, 958–970.

[38] G. Lorden: On excess over the boundary. *Ann. Math. Statist.* 41, 520–527, 1970.

[39] R.B. Lund, R.L. Tweedie (1996): Geometric convergence rates for stochastically ordered Markov chains. *Mathematics of Operations Research* 21, 182–194.

[40] P. Matthews (1993): A slowly mixing Markov chain with implications for Gibbs sampling. *Stat. Prob. Lett.* 17, 231–236.

[41] K. Marton (1996): A measure concentration inequality for contracting

Markov chains. *Geom. Funct. Anal.* 3, 556–571.

[42] P. Mathé, E. Novak (2007): Simple Monte Carlo and the Metropolis algorithm. *J. of Complexity.* 23, 673–696.

[43] K.L. Mengersen, L.R. Tweedie (1996): Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* 24, 1, 101–121.

[44] S.P. Meyn nand R.L. Tweedie: *Markov Chains and Stochastic Stability.* Springer-Verlag, 1993.

[45] P. Mykland, L. Tierney and B. Yu (1995): Regeneration in Markov chain samplers. *J. Am. Statist. Assoc.*., 90, 233–241.

[46] W. Niemiro, P. Pokarowski (2009): Fixed precision MCMC Estimation by Median of Products of Averages. *J. Appl. Probab.* 46 (2), 309–329.

[47] E. Nummelin (1978): A splitting technique for Harris recurrent Markov chains, *Z. Wahr. Verw. Geb.* 43, 309–318.

[48] E. Nummelin (2002): MC's for MCMC'ists, *International Statistical Review*, 70, 215–240.

[49] G.O. Roberts and J.S. Rosenthal (1997): Geometric ergodicity and hybrid Markov chains. *Elec. Comm. Prob.* 2 (2).

[50] G.O. Roberts and J.S. Rosenthal (2004): General state space Markov chains and MCMC algorithms. *Probability Surveys* 1, 20–71.

[51] G.O. Roberts, R.L. Tweedie: Bounds on regeneration times and convergence rates for Markov chains. *Stochatic Process. Appl.* 91, pp. 337–338, 1999.

[52] J.S. Rosenthal: Quantitative convergence rates of Markov chains: a simple account. *Elect. Comm. in Probab.* 7, 123–128, 2002.

[53] J.S. Rosenthal: Rates of convergence for Gibbs sampling for variance component models. *Ann. Statist.* 23, pp. 740–761, 1995.

[54] J.S. Rosenthal: Minorization conditions and convergence rates for Markov chains. *J. Amer. Statist. Association* 90, 558–566, 1995.

[55] D. Rudolf (2008): Explicit error bounds for lazy reversible Markov chain Monte Carlo. *J. of Complexity.* 25, 11–24.

[56] P.M. Samson (2000): Concentration of measure inequalitites for Markov chains and $\Phi-$mixing processes. *Ann. Probab.* 28, 416–461.

[57] A.J. Sinclair, M.R. Jerrum: Approximate counting, uniform generation and rapidly mixing Markov chains, *Information and Computation* 82, pp. 93–133, 1989.