J. Appl. Prob. **52,** 811–825 (2015)
Printed in England

ON THE CONVERGENCE RATES OF SOME ADAPTIVE MARKOV CHAIN MONTE CARLO ALGORITHMS

YVES ATCHADÉ,* *University of Michigan, Ann Arbor*
YIZAO WANG,** *University of Cincinnati*

Abstract

In this paper we study the mixing time of certain adaptive Markov chain Monte Carlo (MCMC) algorithms. Under some regularity conditions, we show that the convergence rate of importance resampling MCMC algorithms, measured in terms of the total variation distance, is $O(n^{-1})$. By means of an example, we establish that, in general, this algorithm does not converge at a faster rate. We also study the interacting tempering algorithm, a simplified version of the equi-energy sampler, and establish that its mixing time is of order $O(n^{-1/2})$.

Keywords: Adaptive Markov chain Monte Carlo; mixing time; total variation distance; importance resampling algorithm; equi-energy sampler

2010 Mathematics Subject Classification: Primary 65C05; 65C40
Secondary 60J05

1. Introduction

Constructing Markov chain Monte Carlo (MCMC) transition kernels to sample efficiently from a given distribution π, say, is a difficult task in practice as it requires a careful choice and tuning of the kernel. The development of adaptive MCMC (AMCMC) methods is partly motivated by the need of overcoming this difficulty. Instead of having a fixed Markov kernel P, at each round n an AMCMC algorithm selects a kernel $P_{\widehat{\theta}_n}$ from a family of Markov kernels $\{P_\theta\}_{\theta\in\Theta}$, where the value (parameter) $\widehat{\theta}_n$ is computed based on possibly all the samples generated up to time n, so that the transition kernel is automatically self-adapted; see, for example, the recent survey [6] and the references therein.

In this paper we investigate the convergence rates of two AMCMC algorithms: the *importance resampling MCMC* (IRMCMC) algorithm introduced by Atchadé [4], and the *equi-energy* (EE) *sampler* by Kou *et al.* [12]. The IRMCMC algorithm is also referred to as the *interacting annealing* algorithm [7]. For the EE sampler, we actually focus on a simplified version, which is sometimes referred to as the *interacting tempering* (IT) algorithm [10].

Throughoutthe paper we denote by $\{X_n\}_{n\in\mathbb{N}}$ the random process generated by either of these algorithms. Limit theorems, notably convergence of marginal distributions and the law of large numbers have been known; see, for example, [2]– [5], and [9]. Central limit theorems for such AMCMC algorithms have been considered only recently by Fort *et al.* [10] and Bercu *et al.* [7].

Received 9 January 2014; revision received 28 July 2014.

* Postal address: Department of Statistics, University of Michigan, 439 West Hall, 1085 South University, Ann Arbor, MI 48109–1107, USA.
** Postal address: Department of Mathematical Sciences, University of Cincinnati, 2815 Commons Way, Cincinnati OH 45221, USA. Email address: yizao.wang@uc.edu

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

In short, introducing the auxiliary chain makes the stochastic process no longer Markov, which raises considerable technical difficulties.

In this paper we study the *convergence rate* (or mixing time) of the IRMCMC and IT algorithms. That is, we provide upper bounds on the distances between $\mathcal{L}_{X_n}$ (the distribution of X_n) and the target distribution. Such mixing time results provide information on the burn-in time of the algorithm. There are few results in the literature on the mixing rates of the AMCMC. Andrieu and Atchadé [1] considered an AMCMC with a finite-dimensional parameter. Related results have been obtained by Schmidler and Woodard [14] and Woodard *et al.* [16] on convergence rates of AMCMC and related algorithms, although with a different point of view from ours: they focused on the lower bound in terms of the problem size, not the simulation rounds.

We show that the IRMCMC algorithm has convergence rate of order $O(n^{-1})$. In particular, we also provide a simple example, for which the convergence rate has lower bound $1/n$. We also show that for an m-tuple IRMCMC algorithm (to be defined in Section 2.4), the mixing time is within $O(n^{-1}(\log n)^{m-1})$. For the IT algorithm, under some regularity conditions, we show that the rate of convergence is $O(n^{-1/2})$ in terms of a slightly weaker norm than the total variation distance. These results do not automatically lead to a precise method for selecting burn-in periods because the constants in the derived bounds are hard to compute in most practical cases. However, from a practical viewpoint, this analysis can be viewed as a cautionary tale, suggesting that AMCMC samplers based on auxiliary chains typically require longer burn-in periods than standard, well-behaved MCMC samplers.

The rest of the paper is organized as follows. In the remainder of the introduction we provide a general description of the algorithms considered in the paper and introduce some notation. Section 2 is devoted to the IRMCMC algorithm. The convergence rate is established in Section 2.1, and for multiple IRMCMC algorithms in Section 2.4. Section 3 is devoted to the IT algorithm.

1.1. Notation

We assume that the state space $\mathcal{X}$ is a Polish space equipped with a metric d, and $\mathcal{B}$ is the associated Borel σ-algebra. In addition, $(\mathcal{X}, \mathbb{B})$ is a measure space with a reference σ-finite measure, which we denote for short by $\mathrm{d}x$. Let π and π_Y be probability measures on $(\mathcal{X}, \mathbb{B})$. We assume that π and π_Y are both absolutely continuous with respect to $\mathsf{d}x$ and with a little abuse of notation, we also use π and π_Y to denote the density, respectively. That is, we write $\pi(\mathsf{d}x) = \pi(x)\mathsf{d}x$ and similarly for π_Y. For a transition kernel Q, a measure ν and a function h, we shall write $\nu Q(\cdot) \triangleq \int \nu(\mathrm{d}z)Q(z, \cdot)$, and $Qh(\cdot) \triangleq \int Q(\cdot, \mathrm{d}z)h(z)$.

In this paper, an AMCMC algorithm is a stochastic process $\{(X_n, Y_n)\}_{n\geq 0}$ in $\mathcal{X} \times \mathcal{X}$, designed such that the main chain X_n converges to the target distribution π in a certain sense to be described precisely later. We also assume that the auxiliary chain $\{Y_n\}_{n\geq 0}$ converges to π_Y. For the two algorithms analyzed in this paper, we assume that the evolution of the auxiliary chain is independent of the main chain. The auxiliary chain is not necessarily Markov. Write $\mathcal{F}_n = \sigma(X_0, \ldots, X_n, Y_0, \ldots, Y_n)$.

We denote by $\widehat{\pi}_{Y,n}$ the empirical measure associated to the auxiliary chain $\{Y_n\}_{n\in\mathbb{N}}$ defined by $\widehat{\pi}_{Y,n}(\cdot) \triangleq (1/n)\sum_{i=1}^{n} \delta_{Y_i}(\cdot)$. For functions $f\colon \mathcal{X} \to \mathbb{R}$, we write

$$\widehat{\pi}_{Y,n}(\overline{f}) \triangleq \widehat{\pi}_{Y,n}(f) - \pi_Y(f).$$

We avoid writing $\overline{f}$ for the centered version of f, as it would be unclear with respect to which measure f is centered, especially in the setup of multiple chains. We let C denote general constants that do not depend on n, but may change from line to line.

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

2. The IRMCMC

We consider the IRMCMC method described in Atchadé [4].

Algorithm 1. (*The IRMCMC.*) Fix $\varepsilon \in (0, 1)$. Pick arbitrary $X_0 = x_0$ and $Y_0 = y_0$. Let P be an arbitrary Markov kernel with invariant distribution π. At each round n, X_n and Y_n are conditionally independent given $\mathcal{F}_{n-1}$, and (with w.p. meaning with probability)

$$X_n \mid \mathcal{F}_{n-1} \sim \begin{cases} P(X_{n-1}, \cdot) & \text{w.p. } 1-\varepsilon, \\ \widehat{\theta}_{n-1}(\cdot) & \text{w.p. } \varepsilon, \end{cases}$$

where $\widehat{\theta}_n$ is the (randomly) weighted empirical distribution defined by

$$\widehat{\theta}_n(\cdot) = \sum_{i=1}^{n} \frac{\widetilde{w}(Y_i)}{\sum_{j=1}^{n} \widetilde{w}(Y_j)} \delta_{Y_i}(\cdot) = \frac{\int_{\cdot} \widetilde{w}(z) \widehat{\pi}_{Y,n}(\mathrm{d}z)}{\int_{\mathcal{X}} \widetilde{w}(z) \widehat{\pi}_{Y,n}(\mathrm{d}z)},$$

with $\widetilde{w}(y) \propto \pi(y)/\pi_Y(y) =: w(y)$, and $\widehat{\theta}_0 = \delta_{y_0}$. Recall that π_Y is the limiting distribution of the auxiliary chain $\{Y_n\}_{n\geq 0}$. We assume that $|w|_\infty \triangleq \sup_{x\in\mathcal{X}} |w(x)| < \infty$.

For all probability measures θ on $\mathcal{X}$, we introduce

$$P_\theta(x, \cdot) = (1-\varepsilon)P(x, \cdot) + \varepsilon\theta(\cdot). \tag{1}$$

In this way, for any bounded function $f\colon \mathcal{X} \to \mathbb{R}$, $\mathbb{E}(f(X_{n+1}) \mid \mathcal{F}_n) = P_{\widehat{\theta}_n} f(X_n)$ almost surely, where $\mathbb{E}$ is the expectation.

Remark 1. The assumption on the boundedness of w is not too restrictive. Indeed, very often in practice, we have $\widetilde{\pi}$, the unnormalized density function of π as a bounded function, and set the auxiliary chain with stationary distribution $\widetilde{\pi}_Y \propto \pi_Y$ obtained by $\widetilde{\pi}_Y = \widetilde{\pi}^T$ with $T \in (0, 1)$. In this case, $\widetilde{w} = \widetilde{\pi}/\widetilde{\pi}_Y$ is bounded and thus so is w.

2.1. Convergence rate of the IRMCMC

The following equivalent representation of Algorithm 1 is useful. Let $\{Z_n\}_{n\geq 0}$ be a sequence of independent and identically distributed random variables with $\mathbb{P}(Z_1 = 1) = 1 - \mathbb{P}(Z_1 = 0) = \varepsilon$, where $\mathbb{P}$ is the probability measure. Assume that $\{Z_n\}_{n\geq 0}$ and $\{Y_n\}_{n\geq 0}$ are independent and for each $n \geq 1$, Z_n and $\mathcal{F}_{n-1}$ are independent. Then, at round n, we can introduce Z_n, and write the conditional distribution of X_n given Z_n, $\mathcal{F}_{n-1}$ as

$$X_n \mid \mathcal{F}_{n-1}, Z_n \sim \begin{cases} P(X_{n-1}, \cdot) & \text{if } Z_n = 0, \\ \widehat{\theta}_{n-1}(\cdot) & \text{if } Z_n = 1. \end{cases}$$

Define

$$\tau_0 = 0, \qquad \tau_{i+1} = \min\{k > \tau_i : Z_k = 1\}, \qquad n^* = \max\{k\colon \tau_k \leq n\}.$$

Observe that at each time $\tau_k > 0$, conditioning on $Y_0, Y_1, \dots, Y_{\tau_k - 1}, X_{\tau_k}$ is sampled from $\widehat{\theta}_{\tau_k - 1}$, independent of $X_0, \dots, X_{\tau_k - 1}$. Furthermore, $Y_0, \dots, Y_n$ are independent from $\tau_1, \dots, \tau_{n^*}$. Therefore, we first focus on

$$\eta_n \triangleq \mathbb{P}(X_{n+1} \in \cdot \mid Z_{n+1} = 1) = \mathbb{E}\widehat{\theta}_n(\cdot), \qquad n \in \mathbb{N}.$$

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

We first obtain a bound on the total variation distance $\|\eta_n - \pi\|_{\mathrm{TV}}$. Recall that, given two probability distributions μ and ν, the total variation distance $\|\mu - \nu\|_{\mathrm{TV}}$ is defined by $\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\sup_{|f|_\infty \le 1} |\mu(f) - \nu(f)|$. For convenience, write

$$B_n \triangleq |w|_\infty \sup_{|f|_\infty \le 1} \mathbb{E}\widehat{\pi}_{Y,n}(\overline{f}) + |w|_\infty^2 \sup_{|f|_\infty \le 1} \mathbb{E}(\widehat{\pi}_{Y,n}(\overline{f}))^2, \qquad n \in \mathbb{N}.$$

Recall that throughout, we assume that $|w|_\infty < \infty$.

Lemma 1. *For all $n \in \mathbb{N}$, $\|\eta_n - \pi\|_{\mathrm{TV}} \le B_n$.*

The proof of Lemma 1 is postponed to Section 2.2. Lemma 1 yields a bound on the rate of convergence of $\mathcal{L}_{X_n}$ towards π in the total variation norm, as shown in the following theorem. We set $B_0 = B_{-1} = 1$.

Theorem 1. *Consider $\{X_n\}_{n\in\mathbb{N}}$ generated from Algorithm 1. Then,*

$$\|\mathcal{L}_{X_n} - \pi\|_{\mathrm{TV}} \le \sum_{\ell=0}^{n} (1-\varepsilon)^{n-\ell} B_{\ell-1}. \tag{2}$$

Furthermore, for any bounded measurable function f,

$$\mathbb{E}\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - \pi(f))\bigg)^2 \le \frac{80\varepsilon^{-2}|f|_\infty^2}{n} + 64\varepsilon^{-2}|f|_\infty^2 + |f|_\infty^2\bigg(\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}\sqrt{B_k}\bigg)^2, \qquad n \in \mathbb{N}. \tag{3}$$

The proof of Theorem 1 is postponed to Section 2.2.

Remark 2. In Theorem 1, we do not assume any ergodicity assumption on the kernel P. In the case that P is geometrically ergodic, one can improve (2) quantitatively by bounding the term $\|\eta_k P^{n-k} - \pi\|_{\mathrm{TV}}$ more effectively. For example, if P is uniformly ergodic with rate ρ, then (2) would become $\|\mathcal{L}_{X_n} - \pi\|_{\mathrm{TV}} \le \sum_{\ell=0}^{n}[\rho(1-\varepsilon)]^{n-\ell} B_{\ell-1}$. A similar improvement can be formulated for (3). However, these improvements do not change the rate but only the constant in the corollary below. Besides, such improvements will not be easily available if P is subgeometrically ergodic.

Now, as a corollary, we obtain an upper bound on the convergence rate of the IRMCMC algorithm, under the following assumption.

Assumption 1. *There exists a finite constant C such that for all measurable functions $f\colon \mathcal{X} \to \mathbb{R}$, with $|f|_\infty \le 1$,*

$$\mathbb{E}\widehat{\pi}_{Y,n}(\overline{f}) \le \frac{C}{n}, \qquad \mathbb{E}(\widehat{\pi}_{Y,n}(\overline{f}))^2 \le \frac{C}{n}.$$

Remark 3. Since $\mathbb{E}(\widehat{\pi}_{Y,n}(\overline{f}))^2 = n^{-1}\mathbb{E}(n^{-1/2}\sum_{i=1}^{n}(f(Y_i) - \pi_Y(f)))^2$, the second equation of Assumption 1 simply requires the finiteness of asymptotic variance under $\{Y_n\}_{n\in\mathbb{N}}$ which is also a very desirable property in practice. This is a fairly mild assumption that holds for many processes with short-range dependence; see, for example, Häggström and Rosenthal [11] for further discussion when $\{Y_n\}_{n\in\mathbb{N}}$ is a Markov chain.

The first equation of Assumption 1 is also a fairly mild ergodicity assumption.

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

Corollary 1. *Consider the IRMCMC (Algorithm 1). If Assumption 1 holds then there exists a finite constant C such that*

$$\|\mathcal{L}_{X_n} - \pi\|_{\mathrm{TV}} \le \frac{C}{n}.$$

Furthermore, for any bounded measurable function f,

$$\mathbb{E}\biggl(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - \pi(f))\biggr)^2 \le C|f|_\infty^2, \qquad n \in \mathbb{N}.$$

Proof. Under Assumption 1, (2) yields

$$\begin{aligned}\|\mathcal{L}_{X_n} - \pi\|_{\mathrm{TV}} &\le \frac{C}{n}\biggl[\sum_{\ell=1}^{\lfloor n/2\rfloor}(1-\varepsilon)^{n-\ell}\frac{n}{\ell} + \sum_{\ell=\lfloor n/2\rfloor+1}^{n}(1-\varepsilon)^{n-\ell}\frac{n}{\ell}\biggr]\\ &\le \frac{C}{n}\biggl[(1-\varepsilon)^{n/2}n + \frac{2}{1-\varepsilon}\biggr].\end{aligned}$$

This proves the first conclusion. The proof of the second is staightforward and is thus omitted.

2.2. Proofs of Lemma 1 and Theorem 1

Proof of Lemma 1. Rewrite $\eta_n(f)$ as

$$\begin{aligned}\eta_n(f) &= \mathbb{E}\biggl(\sum_{j=1}^{n}\frac{w(Y_j)}{\sum_{l=1}^{n}w(Y_l)}f(Y_j)\biggr)\\ &= \mathbb{E}\biggl(\frac{1}{n}\sum_{j=1}^{n}w(Y_j)f(Y_j) + \biggl(1 - \frac{1}{n}\sum_{j=1}^{n}w(Y_j)\biggr)\sum_{j=1}^{n}\frac{w(Y_j)f(Y_j)}{\sum_{l=1}^{n}w(Y_l)}\biggr)\\ &= \mathbb{E}(\widehat{\pi}_{Y,n}(wf) - \widehat{\pi}_{Y,n}(\overline{w})\widehat{\theta}_n(f)),\end{aligned}$$

where in the third equality above, we use the fact that $\pi_Y(w) = 1$. Since $\pi(f) = \pi_Y(wf)$, $\|\eta_n - \pi\|_{\mathrm{TV}} = \sup_{|f|_\infty\le 1}\frac{1}{2}(\eta_n(f) - \pi(f))$ is bounded by

$$\begin{aligned}&\frac{1}{2}\sup_{|f|_\infty\le 1}\mathbb{E}\widehat{\pi}_{Y,n}(\overline{wf}) + \frac{1}{2}\sup_{|f|_\infty\le 1}\mathbb{E}(\widehat{\pi}_{Y,n}(\overline{w})\widehat{\theta}_n(f))\\ &\qquad\le \frac{1}{2}\sup_{|f|_\infty\le 1}\mathbb{E}\widehat{\pi}_{Y,n}(\overline{wf}) + \frac{1}{2}\sup_{|f|_\infty\le 1}\mathbb{E}(\widehat{\pi}_{Y,n}(\overline{w})\pi_Y(wf))\\ &\qquad\quad + \frac{1}{2}\sup_{|f|_\infty\le 1}\mathbb{E}(\widehat{\pi}_{Y,n}(\overline{w})(\widehat{\theta}_n(f) - \pi_Y(wf))).\end{aligned}$$

Observe that

$$\sup_{|f|_\infty\le 1}\mathbb{E}(\widehat{\pi}_{Y,n}(\overline{w})\pi_Y(wf)) = \sup_{|f|_\infty\le 1}\pi(f)\mathbb{E}\widehat{\pi}_{Y,n}(\overline{w}) \le |w|_\infty\sup_{|f|_\infty\le 1}\widehat{\pi}_{Y,n}(\overline{f})$$

and $|w|_\infty \ge 1$. Therefore,

$$\|\eta_n - \pi\|_{\mathrm{TV}} \le |w|_\infty\sup_{|f|_\infty\le 1}\mathbb{E}\widehat{\pi}_{Y,n}(\overline{f}) + \frac{1}{2}\sup_{|f|_\infty\le 1}\mathbb{E}(\widehat{\pi}_{Y,n}(\overline{w})(\widehat{\theta}_n(f) - \pi_Y(wf))). \tag{4}$$

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

By the Cauchy–Schwarz inequality,

$$\begin{aligned}\sup_{|f|_\infty \le 1} & \mathbb{E}(\widehat{\pi}_{Y,n}(\overline{w})(\widehat{\theta}_n(f) - \pi_Y(wf))) \\ &\le [\mathbb{E}(\widehat{\pi}_{Y,n}(\overline{w}))^2]^{1/2} \sup_{|f|_\infty \le 1} [\mathbb{E}(\widehat{\theta}_n(f) - \pi_Y(wf))^2]^{1/2}. \end{aligned} \tag{5}$$

The first term is bounded by $|w|_\infty \sup_{|f|_\infty \le 1}[\mathbb{E}(\widehat{\pi}_{Y,n}(\overline{f}))^2]^{1/2}$. For the second term, observe that

$$\mathbb{E}(\widehat{\theta}_n(f) - \pi_Y(wf))^2 \le 2\mathbb{E}(\widehat{\theta}_n(f) - \widehat{\pi}_{Y,n}(wf))^2 + 2\mathbb{E}(\widehat{\pi}_{Y,n}(wf) - \pi_Y(wf))^2, \tag{6}$$

and

$$\begin{aligned}\mathbb{E}(\widehat{\theta}_n(f) - \widehat{\pi}_{Y,n}(wf))^2 &= \mathbb{E}\biggl(\sum_{j=1}^{n} \frac{w(Y_j)f(Y_j)}{\sum_{l=1}^{n} w(Y_l)} - \frac{1}{n}\sum_{j=1}^{n} w(Y_j)f(Y_j)\biggr)^2 \\ &= \mathbb{E}((1 - \widehat{\pi}_{Y,n}(w))^2\widehat{\theta}_n^2(f)) \\ &\le \mathbb{E}(\pi_Y(w) - \widehat{\pi}_{Y,n}(w))^2 \\ &\le |w|_\infty^2 \sup_{|g|_\infty \le 1} \mathbb{E}(\widehat{\pi}_{Y,n}(\overline{g}))^2, \end{aligned}$$

and the above calculation holds for all $f \colon |f|_\infty \le 1$. So, (6) can be expressed as

$$\sup_{|f|_\infty \le 1} \mathbb{E}(\widehat{\theta}_n(f) - \pi_Y(wf))^2 \le 4|w|_\infty^2 \sup_{|f|_\infty \le 1} \mathbb{E}(\widehat{\pi}_{Y,n}(\overline{f}))^2. \tag{7}$$

Combining (4), (5), and the above inequality yields the desired result.

Proof of Theorem 1. We recall that τ_{n^*} is the last time k before n that the main chain is sampled from $\widehat{\theta}_{k-1}$. Now, we can write

$$\begin{aligned}\|\mathcal{L}_{X_n} - \pi\|_{\mathrm{TV}} &= \sup_{|f|_\infty \le 1} \frac{1}{2}\biggl|\sum_{k=0}^{n} \mathbb{E}(f(X_n)\mathbf{1}_{\{\tau_{n^*}=k\}}) - \pi(f)\biggr| \\ &= \sup_{|f|_\infty \le 1} \frac{1}{2}\biggl|\sum_{k=0}^{n} \mathbb{P}(\tau_{n^*} = k)[\mathbb{E}(f(X_n) \mid \tau_{n^*} = k) - \pi(f)]\biggr|, \end{aligned}$$

where $\mathbf{1}$ is the indicator function. Thus,

$$\|\mathcal{L}_{X_n} - \pi\|_{\mathrm{TV}} \le \sum_{k=0}^{n} \mathbb{P}(\tau_{n^*} = k) \sup_{|f|_\infty \le 1} \frac{1}{2}|\mathbb{E}(f(X_n) \mid \tau_{n^*} = k) - \pi(f)|. \tag{8}$$

Observe that the conditional distribution of X_n, given that $\tau_{n^*} = k \ge 1$, is $\eta_{k-1}P^{n-k}$ (set $\eta_0 = \delta_{Y_0}$). Then,

$$\begin{aligned}\sup_{|f|_\infty \le 1} \frac{1}{2}|\mathbb{E}(f(X_n) \mid \tau_{n^*} = k) - \pi(f)| &= \sup_{|f|_\infty \le 1} \frac{1}{2}|\eta_{k-1}P^{n-k}(f) - \pi(f)| \\ &= \|\eta_{k-1}P^{n-k} - \pi\|_{\mathrm{TV}}. \end{aligned}$$

By the fact that $\pi P = \pi$, we have $\|\eta_{k-1}P^{n-k} - \pi\|_{\text{TV}} \le \|\eta_{k-1} - \pi\|_{\text{TV}} \le B_{k-1}$ by Lemma 1. Also, $\mathbb{P}(\tau_{n^*} = k) = \varepsilon(1-\varepsilon)^{n-k}$ for $k = 1, \dots, n$ and $\mathbb{P}(\tau_{n^*} = 0) = (1-\varepsilon)^n$. Thus, (8) can be expressed as (2).

To establish (3), we show that the partial sum $\sum_{k=1}^n (f(X_k) - \pi(f))$ admits a well-behaved martingale approximation. For a probability measure θ on $\mathcal{X}$, define

$$\pi_\theta(A) = \varepsilon \sum_{j=0}^{\infty} (1-\varepsilon)^j (\theta P^j)(A), \qquad A \in \mathcal{B}.$$

Clearly, π_θ is a probability measure on $(\mathcal{X}, \mathcal{B})$, and one can verify that $\pi_\theta P_\theta = \pi_\theta$, and, moreover, that for any bounded measurable function f, and $n \ge 1$,

$$P_\theta^n f(x) - \pi_\theta(f) = (1-\varepsilon)^n P^n f(x) - \varepsilon \sum_{j=n}^{\infty} (1-\varepsilon)^j (\theta P^j) f. \tag{9}$$

Indeed, the $n = 1$ case follows from the definition of P_θ in (1). For $n \ge 1$, by induction, $P_\theta^{n+1} f(x) - \pi_\theta(f) = P_\theta^n(P_\theta f)(x) - \pi_\theta(P_\theta f)$ equals

$$\begin{aligned}
&(1-\varepsilon)^{n+1} P^{n+1} f(x) + (1-\varepsilon)^n \varepsilon\theta f - \varepsilon \sum_{j=n}^{\infty} (1-\varepsilon)^j (\theta P^j)[(1-\varepsilon)Pf + \varepsilon\theta f] \\
&\quad = (1-\varepsilon)^{n+1} P^{n+1} f(x) + \varepsilon \sum_{j=n+1}^{\infty} (1-\varepsilon)^j (\theta P^j) f.
\end{aligned}$$

It follows from (9) that $\|P_\theta^n(x, \cdot) - \pi_\theta\|_{\text{TV}} \le 2(1-\varepsilon)^n$, and, consequently, the function

$$g_\theta(x) = \sum_{j=0}^{\infty} (P_\theta^j f(x) - \pi_\theta(f)) \tag{10}$$

is well defined with $|g_\theta|_\infty \le 2\varepsilon^{-1}|f|_\infty$, and satisfies Poisson's equation,

$$g_\theta(x) - P_\theta g_\theta(x) = f(x) - \pi_\theta(f), \qquad x \in \mathcal{X}. \tag{11}$$

In particular, we have $f(X_k) - \pi_{\widehat{\theta}_{k-1}}(f) = g_{\widehat{\theta}_{k-1}}(X_k) - P_{\widehat{\theta}_{k-1}} g_{\widehat{\theta}_{k-1}}(X_k)$ almost surely. Using this, we write

$$\sum_{k=1}^n (f(X_k) - \pi(f)) = \sum_{k=1}^n (\pi_{\widehat{\theta}_{k-1}}(f) - \pi(f)) + \sum_{k=1}^n (f(X_k) - \pi_{\widehat{\theta}_{k-1}}(f))$$

with

$$\begin{aligned}
\sum_{k=1}^n (f(X_k) - \pi_{\widehat{\theta}_{k-1}}(f)) &= \sum_{k=1}^n (g_{\widehat{\theta}_{k-1}}(X_k) - P_{\widehat{\theta}_{k-1}} g_{\widehat{\theta}_{k-1}}(X_{k-1})) \qquad (12)\\
&\quad + \sum_{k=1}^n (P_{\widehat{\theta}_{k-1}} g_{\widehat{\theta}_{k-1}}(X_{k-1}) - P_{\widehat{\theta}_k} g_{\widehat{\theta}_k}(X_k)) \\
&\quad + \sum_{k=1}^n (P_{\widehat{\theta}_k} g_{\widehat{\theta}_k}(X_k) - P_{\widehat{\theta}_{k-1}} g_{\widehat{\theta}_{k-1}}(X_k)). \qquad (13)
\end{aligned}$$

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

From the definition of π_θ, note that we can write

$$\sum_{k=1}^{n}(\pi_{\widehat{\theta}_{k-1}}(f)-\pi(f))=\sum_{k=1}^{n}\widehat{\theta}_{k-1}(f_\varepsilon-\pi(f_\varepsilon)),$$

where $f_\varepsilon(x)=\varepsilon\sum_{j=0}^{\infty}(1-\varepsilon)^j P^j f(x)$. Thus,

$$\mathbb{E}\biggl(\sum_{k=1}^{n}(\pi_{\widehat{\theta}_{k-1}}(f)-\pi(f))\biggr)^2\le\biggl(\sum_{k=1}^{n}(\mathbb{E}\widehat{\theta}_{k-1}^2(f_\varepsilon-\pi(f_\varepsilon)))^{1/2}\biggr)^2\le|f|_\infty^2\biggl(\sum_{k=0}^{n-1}\sqrt{B_k}\biggr)^2,$$

where in the last equality, we use the fact that $\sup_{|f|_\infty\le 1}\mathbb{E}\widehat{\theta}_k^2(f-\pi(f))\le B_k$, established in (7) in the proof of Lemma 1.

We now bound the three sums on the right-hand side of (13). By (9), (10), and (11) for any probability measures θ,θ', and $x\in\mathcal{X}$,

$$P_\theta g_\theta(x)-P_{\theta'}g_{\theta'}(x)=\int(\theta'-\theta)(\mathrm{d}z)\biggl(\varepsilon\sum_{j=0}^{\infty}j(1-\varepsilon)^j P^j f(z)\biggr).$$

This implies that

$$\biggl|\sum_{k=1}^{n}(P_{\widehat{\theta}_k}g_{\widehat{\theta}_k}(X_k)-P_{\widehat{\theta}_{k-1}}g_{\widehat{\theta}_{k-1}}(X_k))\biggr|=\biggl|(\widehat{\theta}_0-\widehat{\theta}_n)\biggl(\varepsilon\sum_{j=0}^{\infty}j(1-\varepsilon)^j P^j f\biggr)\biggr|\le\frac{2(1-\varepsilon)}{\varepsilon}|f|_\infty.$$

Next, observe that

$$\begin{aligned}\biggl|\sum_{k=1}^{n}(P_{\widehat{\theta}_{k-1}}g_{\widehat{\theta}_{k-1}}(X_{k-1})-P_{\widehat{\theta}_k}g_{\widehat{\theta}_k}(X_k))\biggr|&=|P_{\widehat{\theta}_0}g_{\widehat{\theta}_0}(X_0)-P_{\widehat{\theta}_n}g_{\widehat{\theta}_n}(X_n)|\\&\le|g_{\widehat{\theta}_0}|_\infty+|g_{\widehat{\theta}_n}|_\infty\\&\le 4\varepsilon^{-1}|f|_\infty.\end{aligned}$$

Finally, we also note that $\sum_{k=1}^{n}(g_{\widehat{\theta}_{k-1}}(X_k)-P_{\widehat{\theta}_{k-1}}g_{\widehat{\theta}_{k-1}}(X_{k-1}))=:\sum_{k=1}^{n}D_k$ is a martingale with respect to $\{\mathcal{F}_n\}$, whence $\mathbb{E}(\sum_{k=1}^{n}D_k)^2=\sum_{k=1}^{n}\mathbb{E}D_k^2\le 4n\sup_\theta|g_\theta|_\infty^2\le 16\varepsilon^{-2}|f|_\infty^2 n$. Using all the above, we obtain (3).

2.3. An example on the lower bound

We provide an example where $O(n^{-1})$ is also the lower bound of the rate for both $\|\eta_n-\pi\|_{\mathrm{TV}}$ and $\|\mathcal{L}_{X_n}-\pi\|_{\mathrm{TV}}$. This shows that the rate in our upper bound in Corollary 1 is optimal.

Example 1. Consider the simple case when $\mathcal{X}=\{\pm 1\}$ and $\pi=\pi_Y$. In this case, the weight function is uniform ($w\equiv 1$). Suppose that the auxiliary chain $\{Y_n\}_{n\ge 0}$ has transition matrix

$$\boldsymbol{P}_Y=\begin{pmatrix}1-a & a\\ b & 1-b\end{pmatrix}\quad\text{with } a,b\in(0,1).$$

The corresponding Markov chain has stationary distribution $\pi_Y=(a+b)^{-1}(b,a)$ and eigenvalues $\lambda_1=1$ and $\lambda_2=1-a-b$. Suppose that $a+b\neq 1$ and the chain starts at $Y_0=-1$. By straightforward calculation, $\mathbb{P}(Y_n=-1)=a/(a+b)+b/(a+b)\lambda_2^n$, and

$$\mathbb{E}\widehat{\pi}_{Y,n}(\{-1\})-\pi_Y(\{-1\})=\biggl(\frac{a}{a+b}\biggr)\biggl(\frac{1}{n}\biggr)\biggl(\frac{\lambda_2-\lambda_2^{n+1}}{1-\lambda_2}\biggr).$$

It then follows from the definition that $\|\eta_n-\pi\|_{\mathrm{TV}}\ge C/n$.

Furthermore, in (1) set $P(x, \cdot) = \pi(\cdot)$. That is, P is the *best* kernel we can put into the algorithm, in the sense that it takes one step to arrive at the stationary distribution (although this is too ideal to be practical). Now,

$$\begin{aligned}\mathbb{P}(X_n = -1) - \pi(\{-1\}) &= (1-\varepsilon)\pi(\{-1\}) + \varepsilon\mathbb{E}\widehat{\pi}_{Y,n}(\{-1\}) - \pi(\{-1\}) \\ &= \varepsilon(\mathbb{E}\widehat{\pi}_{Y,n}(\{-1\}) - \pi_Y(\{-1\})).\end{aligned}$$

It then follows that $\|\mathcal{L}_{X_n} - \pi\|_{\mathrm{TV}} \geq C/n$.

2.4. The multiple IRMCMC

We discuss a multiple chain IRMCMC algorithm and establish a similar convergence rate as in Section 2.1 by a repeated application of Theorem 1. For $m \geq 1$ and $\ell \in \{0, \dots, m\}$, let $\pi^{(\ell)}$ be a probability measure on $\mathcal{X}$, and P_ℓ a Markov kernel with invariant distribution $\pi^{(\ell)}$ such that $\pi^{(m)} = \pi$.

Algorithm 2. (*The multiple IRMCMC.*) Choose $(X_0^{(0)}, \dots, X_0^{(m)}) = (x_0^{(0)}, \dots, x_0^{(m)})$ and fix $\varepsilon \in (0, 1)$. Given $\mathcal{F}_n = \sigma\{(X_k^{(0)}, \dots, X_k^{(m)}), 0 \leq k \leq n\}$: sample independently $X_{n+1}^{(0)} \sim P_0(X_n^{(0)}, \cdot)$, and for $1 \leq \ell \leq m$, $X_{n+1}^{(\ell)} \sim P_{\ell, \widehat{\theta}_n^{(\ell-1)}}(X_n^{(\ell)}, \cdot)$ with

$$P_{\ell,\theta}(x, \cdot) = (1-\varepsilon)P_\ell(x, \cdot) + \varepsilon\theta(\cdot), \qquad \widehat{\theta}_n^{(\ell-1)}(\cdot) = \sum_{i=1}^n \frac{w_\ell(X_i^{(\ell-1)})}{\sum_{j=1}^n w_\ell(X_j^{(\ell-1)})}\delta_{X_i^{(\ell-1)}}(\cdot),$$

with $w_\ell(x) = \pi_\ell(x)/\pi_{\ell-1}(x)$, $x \in \mathcal{X}$.

To bound $\|\mathcal{L}_{X_n^{(\ell)}} - \pi_\ell\|_{\mathrm{TV}}$, it suffices to control

$$B_n^{(\ell-1)} \triangleq \sup_{|f|_\infty \leq 1} \mathbb{E}\widehat{\pi}_{X^{(\ell-1)},n}(\overline{f}) + \sup_{|f|_\infty \leq 1} \mathbb{E}(\widehat{\pi}_{X^{(\ell-1)},n}(\overline{f}))^2, \qquad n \in \mathbb{N},$$

where this time $\widehat{\pi}_{X^{(\ell)},n}(\overline{f}) \triangleq \widehat{\pi}_{X^{(\ell)},n}(f) - \pi_\ell(f)$. In fact, it suffices to control $B_n^{(0)}$, which is the purpose of the following assumption.

Assumption 2. *As $n \to \infty$, the initial Markov chain $\{X_n^{(0)}\}_{n \geq 0}$ satisfies $B_n^{(0)} \leq C/n$.*

Theorem 2. *Consider the multiple IRMCMC (Algorithm 2) for which Assumption 2 holds and* $\max_{\ell=1,\dots,m} |w_\ell|_\infty < \infty$. *Then, for $\ell = 1, \dots, m$, there exists a finite constant C such that, for $n \geq 2$,*

$$\|\mathcal{L}_{X_n^{(\ell)}} - \pi_\ell\|_{\mathrm{TV}} \leq \frac{C(\log n)^{\ell-1}}{n},$$

and for any bounded measurable function f,

$$\mathbb{E}\bigg(\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i^{(\ell)}) - \pi_\ell(f))\bigg)^2 \leq C.$$

Proof. This follows easily from a repeated application of Theorem 1.

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

3. IT algorithm

In this section we consider the IT algorithm as follows. Recall that the auxiliary chain $\{Y_n\}_{n\geq 0}$ evolves independently from the main chain $\{X_n\}_{n\geq 0}$.

Algorithm 3. (*The IT algorithm.*) Fix $\varepsilon \in (0, 1)$. Start at $X_0 = x_0$ and $Y_0 = y_0$. At each round n, generate

$$X_n \sim \begin{cases} P(X_{n-1}, \cdot) & \text{w.p. } 1-\varepsilon, \\ K_{\widehat{\pi}_{Y,n-1}}(X_{n-1}, \cdot) & \text{w.p. } \varepsilon, \end{cases}$$

where $\widehat{\theta}_n = \widehat{\pi}_{Y,n}$ is the empirical measure associated to $\{Y_n\}_{n\geq 0}$ and K_θ is defined by

$$K_\theta(x, A) = \mathbf{1}_A(x) + \int_{\mathcal{X}} \bigg(1 \wedge \frac{\pi(z)\pi_Y(x)}{\pi(x)\pi_Y(z)}\bigg)(\mathbf{1}_A(z) - \mathbf{1}_A(x))\theta(\mathrm{d}z).$$

In other words, for all nonnegative functions $h\colon \mathcal{X} \to \mathbb{R}$ and $n \in \mathbb{N}$,

$$\mathbb{E}_x(h(X_{n+1}) \mid \mathcal{F}_n) = P_{\widehat{\pi}_{Y,n}} h(X_n) \quad \text{almost surely,}$$

where for any probability measure θ on $\mathcal{X}$, P_θ is defined as

$$P_\theta(x, A) = (1-\varepsilon)P(x, A) + \varepsilon K_\theta(x, A).$$

Recall that, we write $\pi(\mathrm{d}x) \equiv \pi(x)\mathrm{d}x$ and similarly for π_Y with a little abuse of notation, and $w(x) = \pi(x)/\pi_Y(x)$. We assume that $|w|_\infty < \infty$.

The kernel K_{π_Y} is the independent Metropolis kernel with target π and proposal π_Y. It is well known that under the assumption $|w|_\infty < \infty$ (recall Remark 1), the kernel K_{π_Y} is uniformly ergodic [13], and this property is inherited by P_{π_Y}. That is, there exists $C_0 < \infty$, $\rho \in (0, 1)$ such that

$$\|P^n_{\pi_Y}(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq C_0\rho^n, \qquad n \geq 0. \tag{14}$$

3.1. Convergence rate of the IT algorithm

We make the following assumptions.

Assumption 3. *There exists a finite universal constant C such that for any measurable function $f\colon \mathcal{X} \to \mathbb{R}$, with $|f|_\infty \leq 1$,*

$$\sup_n \mathbb{P}\bigg(\bigg|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(f(Y_j) - \pi_Y(f))\bigg| > x\bigg) \leq C\exp\bigg(-\frac{x^2}{C\sigma^2(f)}\bigg),$$

where $\sigma^2(f) \triangleq \mathrm{var}_{\pi_Y}(f)$.

Assumption 4. *The function $w\colon \mathcal{X} \to \mathbb{R}$ is continuous (with respect to the metric on $\mathcal{X}$), and*

$$\sup_{x\in\mathcal{X}} \frac{\phi(x)}{w^2(x)} < \infty, \tag{15}$$

where $\phi(x) \triangleq \pi_Y(\{z\colon w(z) \leq w(x)\})$.

Assumption 5. *The kernel P is such that if $f\colon \mathcal{X} \to \mathbb{R}$ is continuous, then Pf is also continuous.*

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

Remark 4. The deviation bound appears naturally in the proof of Assumption 3 although these types of bound are not widely available for Markov chains. A continuous-time version appeared in [8, Proposition 1.2] but extension to discrete-time Markov chains along the same arguments is apparently not straightforward.

Remark 5. Assumption 4 can be difficult to check in practice, but is not overly restrictive. For example, consider $\mathcal{X} = \mathbb{R}$ and $\pi_Y = \pi^T$ with some $T \in (0, 1)$. For the sake of simplicity, we focus on $x \in \mathbb{R}_+$ and define $\phi_+(x) \triangleq \pi_Y(\{z > 0 \colon w(z) \le w(x)\})$. Suppose that the density $\pi(x)$ decays asymptotically as $x^{-\alpha}$ for $\alpha > 1$ as $x \to \infty$. Then, $\pi_Y(x) \sim x^{-T\alpha}$ and $w(x) \sim x^{(T-1)\alpha}$. Here and below, we write $a(x) \sim b(x)$ if $\lim_{x\to\infty} a(x)/b(x) = 1$. Furthermore, assume that $T\alpha > 1$. Then, $\phi_+(x) \sim (T\alpha - 1)^{-1}x^{1-T\alpha}$ and $\phi_+(x)/w^2(x) \sim 1/(T\alpha - 1)x^{1+2\alpha-3T\alpha}$. Therefore, (15) holds if $T > (1 + 2\alpha)/(3\alpha)$.

Theorem 3. *Consider the IT algorithm described above and suppose that Assumptions 3–5 hold. Then there exists a constant C such that for all continuous bounded functions $f \colon \mathcal{X} \to \mathbb{R}$ and $n \in \mathbb{N}$,*

$$|\mathbb{E}(f(X_n) - \pi(f))| \le \frac{C|f|_\infty}{\sqrt{n}}.$$

Proof. Fix $n \ge 2$ and $1 \le q \le n$. Fix $f \colon \mathcal{X} \to \mathbb{R}$ with $|f|_\infty = 1$. Then write

$$\mathbb{E}_x f(X_n) - P^n_{\pi_Y} f(x) = \mathbb{E}_x(P^{n-q}_{\pi_Y} f(X_q) - P^n_{\pi_Y} f(x)) - \mathbb{E}_x(P^{n-q}_{\pi_Y} f(X_q) - f(X_n)).$$

For the first term, we can use (14) to obtain $|\mathbb{E}_x(P^{n-q}_{\pi_Y} f(X_q) - P^n_{\pi_Y} f(x))| \le C\rho^{n-q}$ for some finite constant C that does not depend on f. For the second term, we write

$$\begin{aligned}
\mathbb{E}_x(P^{n-q}_{\pi_Y} f(X_q) - f(X_n)) &= \mathbb{E}_x\Bigg(\sum_{j=q}^{n-1}(P^{n-j}_{\pi_Y} f(X_j) - P^{n-j-1}_{\pi_Y} f(X_{j+1}))\Bigg) \\
&= \sum_{j=q}^{n-1} \mathbb{E}_x(P^{n-j}_{\pi_Y} f(X_j) - \mathbb{E}_x(P^{n-j-1}_{\pi_Y} f(X_{j+1}) \mid \mathcal{F}_j)) \\
&= \sum_{j=q}^{n-1} \mathbb{E}_x(P^{n-j}_{\pi_Y} f(X_j) - P_{\widehat{\pi}_{Y,j}} P^{n-j-1}_{\pi_Y} f(X_j)) \\
&= \sum_{j=q}^{n-1} C_0\rho^{n-j-1}\mathbb{E}_x((P_{\pi_Y} - P_{\widehat{\pi}_{Y,j}})\zeta_{n,j}(X_j)), \qquad (16)
\end{aligned}$$

where in the last line, we write

$$\zeta_{n,j}(x) = \frac{P^{n-j-1}_{\pi_Y}(f(x) - \pi_Y(f))}{C_0\rho^{n-j-1}}, \qquad x \in \mathcal{X},$$

with C_0 and ρ chosen as in (14). As a consequence of (14), we have $|\zeta_{n,j}|_\infty \le 1$. It is also continuous by the continuity of f and Assumption 5.

To simplify the notation, for any function $g \colon \mathcal{X} \to \mathbb{R}$, define

$$H_g(x, z) \triangleq \alpha(x, z)(g(z) - g(x)), \qquad x,\ z \in \mathcal{X}, \qquad (17)$$

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

where

$$\alpha(x,z) \triangleq 1 \wedge \frac{w(z)}{w(x)}.$$

Thus, we can write

$$P_\theta g(x) - P_{\pi_Y} g(x) = \varepsilon \int H_g(x,z)(\theta(\mathrm{d}z) - \pi_Y(\mathrm{d}z)).$$

For any $g\colon \mathcal{X} \to \mathbb{R}$, we introduce the class of functions $\mathcal{F}_g \triangleq \{z \mapsto H_g(x,z)\colon x \in \mathcal{X}\}$, and the empirical process

$$\mathbb{G}_n(h) \triangleq \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (h(Y_j) - \pi_Y(h)), \qquad h \in \mathcal{F}_g.$$

Therefore, the expectation term in (16) can be written as

$$\begin{aligned}
\mathbb{E}_x((P_{\pi_Y} - P_{\widehat{\pi}_{Y,j}})\zeta_{n,j}(X_j)) &= \varepsilon \mathbb{E}_x\biggl(\int H_{\zeta_{n,j}}(X_j,z)(\pi_Y(\mathrm{d}z) - \widehat{\pi}_{Y,j}(\mathrm{d}z))\biggr) \\
&= -\varepsilon \mathbb{E}_x\biggl(\frac{1}{j}\sum_{\ell=1}^{j} H_{\zeta_{n,j}}(X_j, Y_\ell) - \int_{\mathcal{X}} H_{\zeta_{n,j}}(X_j,z)\pi_Y(\mathrm{d}z)\biggr) \\
&= -\frac{\varepsilon}{\sqrt{j}} \mathbb{E}_x(\mathbb{G}_j(H_{\zeta_{n,j}}(X_j,\cdot))),
\end{aligned}$$

whence,

$$\begin{aligned}
|\mathbb{E}_x(P_{\pi_Y}^{n-q} f(X_q) - f(X_n))| &= \biggl|\varepsilon \sum_{j=q}^{n-1} \frac{C_0 \rho^{n-j-1}}{\sqrt{j}} \mathbb{E}_x(\mathbb{G}_j(H_{\zeta_{n,j}}(X_j,\cdot)))\biggr| \\
&\le C_0 \sum_{j=q}^{n-1} \frac{\rho^{n-j-1}}{\sqrt{j}} \mathbb{E}_x\Bigl(\sup_{h \in \mathcal{F}_{\zeta_{n,j}}} |\mathbb{G}_j(h)|\Bigr).
\end{aligned}$$

In Lemma 2 below, we prove that for any continuous function $g\colon \mathcal{X} \to \mathbb{R}$ such that $|g|_\infty \le 1$, $\mathbb{E}_x(\sup_{h\in\mathcal{F}_g} |\mathbb{G}_n(h)|) \le C$ for some constant C that does not depend on n nor g. We conclude that

$$|\mathbb{E}_x(P_{\pi_Y}^{n-q} f(X_q) - f(X_n))| \le C \sum_{j=q}^{n-1} \frac{1}{\sqrt{j}} \rho^{n-j-1}.$$

Thus, for any $1 \le q \le n$,

$$|\mathbb{E}_x(f(X_n)) - \pi_Y(f)| \le C\biggl\{\rho^n + \rho^{n-q} + \varepsilon \sum_{j=q}^{n-1} \frac{\rho^{n-j-1}}{\sqrt{j}}\biggr\} \le C n^{-1/2}$$

by choosing $q = n - \lfloor -\log n / 2 \log \rho \rfloor$.

We rely on the following technical result on the auxiliary chain $\{Y_n\}_{n\ge0}$.

Lemma 2. *Suppose that Assumptions 3 and 4 hold. Then there exists a constant C such that, for all continuous functions $g\colon \mathcal{X} \to \mathbb{R}$ such that $|g|_\infty \le 1$,*

$$\sup_{n\in\mathbb{N}} \mathbb{E}_x\Bigl(\sup_{h\in\mathcal{F}_g} |\mathbb{G}_n(h)|\Bigr) \le C.$$

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

Proof. Throughout the proof $n \geq 1$ is fixed. Assumption 3 suggests the following metric on $\mathcal{F}_g$:

$$\mathsf{d}(h_1, h_2) = \sigma(h_1 - h_2) = \left(\int_{\mathcal{X}} (h_1(x) - h_2(x))^2 \pi_Y(\mathrm{d}x)\right)^{1/2},$$

which has the following properties. For $x_1, x_2 \in \mathcal{X}$, it is easy to check that

$$|H_g(x_1, z) - H_g(x_2, z)| \leq 2|\alpha(x_1, z) - \alpha(x_2, z)| + |g(x_1) - g(x_2)|. \tag{18}$$

It follows that

$$\begin{aligned}&\mathsf{d}(H_g(x_1, \cdot), H_g(x_2, \cdot))\\ &\qquad \leq \sqrt{2}|g(x_1) - g(x_2)| + 2\sqrt{2}\sqrt{\int |\alpha(x_1, z) - \alpha(x_2, z)|^2 \pi_Y(\mathrm{d}z)}.\end{aligned} \tag{19}$$

This implies that the diameter of $\mathcal{F}_g$ is bounded by $\delta(\mathcal{F}_g) = 4\sqrt{2}$. It also implies that with respect to d, the empirical process $\{\mathbb{G}_n(h), h \in \mathcal{F}_g\}$ is separable. Indeed, for $x \in \mathcal{X}$ arbitrary and $h = H_g(x, \cdot)$, using the Polish assumption, we can find a sequence $x_m \in \mathcal{X}$ (x_m belongs to a countable subset of $\mathcal{X}$) such that $x_m \to x$ as $m \to \infty$. Setting $h_m = H_g(x_m, \cdot)$, it follows from (19) and the continuity of g and w that $h_m \to h$ in $(\mathcal{F}_g, \mathsf{d})$, and from (18) it is easy to show that $\mathbb{G}_n(h_m) - \mathbb{G}_n(h) = n^{-1/2}\sum_{\ell=1}^{n}(H_g(x, Y_\ell) - H_g(x_m, Y_\ell)) + \sqrt{n}\pi_Y(H_g(x, \cdot) - H_g(x_m, \cdot)) \to 0$ as $m \to \infty$ for all realizations of $\{Y_1, \dots, Y_n\}$.

For any $h_1, h_2 \in \mathcal{F}_g$, Assumption 3 implies that for any $x > 0$,

$$\mathbb{P}_x(|\mathbb{G}_n(h_1) - \mathbb{G}_n(h_2)| > x) \leq C \exp\left(-\frac{x^2}{c\mathsf{d}^2(h_1, h_2)}\right).$$

Here, the constant C above is universal for all g such that $|g|_\infty \leq 1$. Indeed, (17) implies that for such a function g, $h \in \mathcal{F}_g$ implies $|h|_\infty \leq 2$. Then, we apply [15, Corollary 2.2.8] to conclude that for $h_{0,g} \in \mathcal{F}_g$, there exists a constant C independent of g such that

$$\mathbb{E}_x\left(\sup_{h \in \mathcal{F}_g} |\mathbb{G}_n(h)|\right) \leq \mathbb{E}_x|\mathbb{G}_n(h_{0,g})| + C\int_0^{\delta(\mathcal{F}_g)} \sqrt{1 + \log \mathsf{D}(\varepsilon, \mathcal{F}_g, \mathsf{d})}\,\mathrm{d}\varepsilon < \infty,$$

where $\mathsf{D}(\varepsilon, \mathcal{F}_g, \mathsf{d})$ is the packing number of $\mathcal{F}_g$ with respect to d. Since all elements of $\mathcal{F}_g$ have a sup-norm of at most 2, Assumption 3 implies that $\sup_{n \in \mathbb{N}} \mathbb{E}_x|\mathbb{G}_n(h_{0,g})| \leq C < \infty$, where C does not depend on g. To control the entropy number, we further bound the right-hand side of (19).

Without loss of generality, assume that $x_1, x_2 \in \mathcal{X}$ and $w(x_1) < w(x_2)$. If $w(x_1) \vee w(x_2) \leq w(z)$ then $\alpha(x_1, z) - \alpha(x_2, z) = 0$. If $w(z) \leq w(x_1)$ then

$$|\alpha(x_1, z) - \alpha(x_2, z)|^2 = \left|\frac{w(z)}{w(x_1)} - \frac{w(z)}{w(x_2)}\right|^2 \leq \frac{1}{w(x_1)^2}(w(x_2) - w(x_1))^2.$$

If $w(x_1) \leq w(z) \leq w(x_2)$ then

$$|\alpha(x_1, z) - \alpha(x_2, z)|^2 = \left|1 - \frac{w(z)}{w(x_2)}\right|^2 \leq \frac{1}{w(x_2)^2}(w(x_2) - w(x_1))^2.$$

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

Thus,

$$\int |\alpha(x_1, z) - \alpha(x_2, z)|^2 \pi_Y(\mathrm{d}z) \le \left(\frac{\phi(x_1)}{w(x_1)^2} + \frac{\phi(x_2)}{w(x_2)^2}\right)(w(x_2) - w(x_1))^2$$
$$\le C(w(x_2) - w(x_1))^2,$$

where $\phi(x) \triangleq \pi_Y(\{z\colon w(z) \le w(x)\})$, and the last inequality follows from Assumption 4. Together with (19), we conclude from this bound that there exists a constant C_0 independent of g such that

$$\mathsf{d}(H_g(x_1, \cdot), H_g(x_2, \cdot)) \le C_0(|g(x_1) - g(x_2)| + |w(x_2) - w(x_1)|). \tag{20}$$

Since $|g|_\infty \le 1$ and $w(x) \in [0, |w|_\infty]$, this implies that the ε-packing number of $(\mathcal{F}_g, \mathsf{d})$ is at most of order ε^{-2}, independent of g. A detailed proof is provided below. It follows that $\int_0^{\delta(\mathcal{F}_g)} \sqrt{1 + \log \mathsf{D}(\varepsilon, \mathcal{F}_g, \mathsf{d})}\, \mathrm{d}\varepsilon \le C \int_0^{\delta(\mathcal{F}_g)} \sqrt{1 + \log(1/\varepsilon)}\, \mathrm{d}\varepsilon < \infty$, which proves the lemma.

To complete the proof, we show that the ε-packing number of $(\mathcal{F}_g, \mathsf{d})$ is at most of order ε^{-2}, independent of g. That is, the cardinality of any ε-*separate set* is at most of order ε^{-2} (recall that a set is an ε-separate set if any two points of this set have distance larger than ε). Note that the functions in $\mathcal{F}_g$ are indexed by $x \in \mathcal{X}$.

Firstly, one can divide the set $\mathcal{X}$ into $N = \lfloor 2/\varepsilon \rfloor + 1$ disjoint subsets $S(1), \dots, S(N)$, so that for every two points x, y within the same $S(i)$, $|g(x) - g(y)| < \varepsilon$. Note that N does not depend on g. For example, consider $g^{-1}([-1, -1 + \varepsilon]), g^{-1}((-1 + \varepsilon, -1 + 2\varepsilon]), \dots, g^{-1}((-1 + (N - 1)\varepsilon, 1])$.

Secondly, for each set $S(i)$, one can again divide it into $N' = \lfloor |w|_\infty/\varepsilon \rfloor + 1$ disjoint subsets, denoted by $S(i, j)$, $j = 1, \dots, N'$, so that within each $S(i, j)$ for every two points x, y, $|w(x) - w(y)| < \varepsilon$.

Finally, $\{S(i, j)\}_{i=1,\dots,N, j=1,\dots,N'}$ forms a disjoint partition of $\mathcal{X}$. The construction and (20) requires that any $2C_0\varepsilon$-separate set contains at most one point in each $S(i, j)$. Therefore, the ε-packing number is at most of order $1/\varepsilon^2$.

Acknowledgements

We thank the anonymous referee for carefully reading our paper and for helpful suggestions. The first author was partly funded by the National Science Foundation (grant no. SES 1229261).

References

[1] Andrieu, C. and Atchadé, Y. F. (2007). On the efficiency of adaptive MCMC algorithms. *Electron. Commun. Prob.* **12,** 336–349.

[2] Andrieu, C., Jasra, A., Doucet, A. and Del Moral, P. (2008). A note on convergence of the equi-energy sampler. *Stoch. Anal. Appl.* **26,** 298–312.

[3] Andrieu, C., Jasra, A., Doucet, A. and Del Moral, P. (2011). On nonlinear Markov chain Monte Carlo. *Bernoulli* **17,** 987–1014.

[4] Atchadé, Y. F. (2009). Resampling from the past to improve on MCMC algorithms. *Far East J. Theoret. Statist.* **27,** 81–99.

[5] Atchadé, Y. F. (2010). A cautionary tale on the efficiency of some adaptive Monte Carlo schemes. *Ann. Appl. Prob.* **20,** 841–868.

[6] Atchadé, Y., Fort, G., Moulines, E. and Priouret, P. (2011). Adaptive Markov chain Monte Carlo: theory and methods. In *Bayesian Time Series Models*, Cambridge University Press, pp. 32–51.

[7] Bercu, B., Del Moral, P. and Doucet, A. (2012). Fluctuations of interacting Markov chain Monte Carlo methods. *Stoch. Process. Appl.* **122,** 1304–1331.

[8] Cattiaux, P. and Guillin, A. (2008). Deviation bounds for additive functionals of Markov processes. *ESAIM Prob. Statist.* **12,** 12–29.

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press

[9] Fort, G., Moulines, E. and Priouret, P. (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.* **39,** 3262–3289.

[10] Fort, G., Moulines, E., Priouret, P. and Vandekerkhove, P. (2014). A central limit theorem for adaptive and interacting Markov chains. *Bernoulli* **20,** 457–485.

[11] Häggström, O. and Rosenthal, J. S. (2007). On variance conditions for Markov chain CLTs. *Electron. Commun. Prob.* **12,** 454–464.

[12] Kou, S. C., Zhou, Q. and Wong, W. H. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Statist.* **34,** 1581–1652.

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

[14] Schmidler, S. C. and Woodard, D. B. (2011). Lower bounds on the convergence rates of adaptive MCMC methods. Tech. Rep., Duke University.

[15] Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

[16] Woodard, D. B., Schmidler, S. C. and Huber, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Prob.* **19,** 617–640.

https://doi.org/10.1239/jap/1445543848 Published online by Cambridge University Press