Statistica Sinica **20** (2010), 209-233

THE WANG-LANDAU ALGORITHM IN GENERAL STATE SPACES: APPLICATIONS AND CONVERGENCE ANALYSIS

Yves F. Atchadé and Jun S. Liu

University of Michigan and Harvard University

Abstract: The Wang-Landau algorithm (Wang and Landau (2001)) is a recent Monte Carlo method that has generated much interest in the Physics literature due to some spectacular simulation performances. The objective of this paper is two-fold. First, we show that the algorithm can be naturally extended to more general state spaces and used to improve on Markov Chain Monte Carlo schemes of more interest in Statistics. In a second part, we study asymptotic behaviors of the algorithm. We show that with an appropriate choice of the step-size, the algorithm is consistent and a strong law of large numbers holds under some fairly mild conditions. We have also shown by simulations the potential advantage of the WL algorithm for problems in Bayesian inference.

Key words and phrases: Adaptive MCMC, geometric ergodicity, Monte Carlo methods, multicanonical sampling, stochastic approximation, trans-dimensional MCMC, Wang-Landau algorithm.

1. Introduction

Although the idea of Monte Carlo computation has been around for more than a century, its first real scientific use occurred during the World War II when the first generation computer became available. Nick Metropolis coined the name "Monte Carlo" for the method when he was at Los Alamos National Labs and it quick evolved to an active research area due to the active involvements of leading physicists in the Labs. Ever since then, physicists have been at the forefront of the methodological research in the field. One of their latest additions is an algorithm proposed by F. Wang and D. P. Landau (Wang and Landau (2001)). The Wang-Landau (WL) algorithm has been successfully applied to some complex sampling problems in physics. The algorithm is closely related to *multicanonical sampling*, a method due to B. A. Berg and T. Neuhaus (Berg and Neuhaus (1992)). Briefly, if π is the probability measure of interest, the idea behind multicanonical sampling is to obtain an importance sampling distribution by partitioning the state space along the energy function $(-\log \pi(x))$ and re-weighting

appropriately each component of the partition so that the modified distribution π^* spends equal amount of time in each component, i.e., uniform in the energy space. The method is often criticized for the difficulty involved in computing the weights. The main contribution of the WL algorithm is in proposing an efficient algorithm that *simultaneously* computes the balancing weights and samples from the re-weighted distribution.

The objective of this paper is to take a more probabilistic look at the WL algorithm and to explore its potential for Monte Carlo simulation problems of more direct interest to statisticians. We achieve this goal by proposing a general state space version of the algorithm. Then we show that the WL algorithm offers an effective strategy to improve on simulated tempering and trans-dimensional MCMC.

From a probabilistic standpoint, the WL algorithm is an interesting example of adaptive Markov Chain Monte Carlo (MCMC). Adaptive MCMC is an approach to Monte Carlo simulation where the transition kernel of the algorithm is sequentially adjusted over time in order to achieve some prescribed optimality. Some early work on the subject includes Gilks, Roberts and Sahu (1998), Andrieu and Robert (2001), and Haario, Saksman and Tamminen (2001). See also Brockwell and Kadane (2005), and Mira and Sargent (2003). Early theoretical analysis includes (Atchadé and Rosenthal (2005), Andrieu and Moulines (2006), Andrieu and Atchade (2007), and Rosenthal and Roberts (2007)). We take a similar path-wise approach to analyze the WL algorithm. The analysis of the WL algorithm is not a straightforward application of the theory in the aforementioned papers because of the specific adaptive control involved. The key point is the stability of the algorithm. We say that the WL algorithm is *stable* if no component of the partition receives infinitely more visits than any other component as $n \to \infty$. On a stable sample path, and under appropriate conditions, we show that the WL algorithm learns the optimal weights and satisfies a strong law of large numbers (Theorem 4.1). In the specific cases of multicanonical sampling and simulated tempering, which includes the original the WL algorithm, we show that the algorithm is stable and that the aforementioned limit results hold.

It came to our attention after the first draft of this paper that a similar extension of the WL algorithm has been proposed independently by F. Liang and coworkers (see e.g., Liang, Liu and Carroll (2007)). Their approach differs from the WL approach in that these authors took a more classic approach based on stochastic approximation with step-sizes set deterministically.

Our proposed generalization to the WL algorithm is presented in Section 2. Some particular cases are discussed in Section 3. The theoretical analysis is discussed in Section 4, but the proofs are postponed to Section 6 to facilitate the flow of ideas.

2. The Wang-Landau Algorithm

In multicanonical sampling, we are given a state space $\mathcal{X}$ and a probability measure π. $\mathcal{X}$ is then partitioned as $\mathcal{X} = \cup\mathcal{X}_i$, where $\mathcal{X}_i \cap \mathcal{X}_j = \phi$ and π is re-weighted in each component $\mathcal{X}_i$. An abstract way to do the same and much more is the following. We start with $(\mathcal{X}_i, \mathcal{B}_i, \lambda_i)$ $i = 1, \ldots, d$, a finite family of measure spaces where λ_i is a σ-finite measure. We introduce the union space $\mathcal{X} = \bigcup_{i=1}^{d} \mathcal{X}_i \times \{i\}$. We equip $\mathcal{X}$ with the $\sigma-$algebra $\mathcal{B}$ generated by $\{(A_i, i),\ i \in \{1, \ldots, d\},\ A_i \in \mathcal{B}_i\}$ and the measure λ satisfying $\lambda(A, i) = \lambda_i(A)\mathbf{1}_{\mathcal{B}_i}(A)$. Let $h_i : \mathcal{X}_i \to \mathbb{R}$ be a non-negative measurable function and define $\theta^*(i) = \int_{\mathcal{X}_i} h_i(x)\lambda_i(dx)/Z$ where $Z = \sum_{i=1}^{d} \int_{\mathcal{X}_i} h_i(x)\lambda_i(dx)$. We assume that $\theta^*(i) > 0$ for all $i = 1, \ldots, d$, and consider the following probability measure on $(\mathcal{X}, \mathcal{B})$:

$$\pi^*(dx, i) \propto \frac{h_i(x)}{\theta^*(i)} \mathbf{1}_{\mathcal{X}_i}(x)\lambda_i(dx). \tag{2.1}$$

Our objective is to sample from π^*. The problem of sampling from such a distribution arises in a number of different Monte Carlo strategies. For example, and as explained above, if π is a probability measure of interest on some space $(\mathcal{X}, \mathcal{B}, \lambda)$, we can partition $\mathcal{X}$ along the energy function $-\log(\pi)$ and re-weight π by $\pi(\mathcal{X}_i)$ in each component $\mathcal{X}_i$. The sampling problem then becomes of the form (2.1). This powerful strategy appeared first in the Physics literature as *multicanonical sampling* (Berg and Neuhaus (1992)). This is discussed in some more details is Section 3.1.

Sampling from (2.1) also arises naturally when optimizing the *simulated tempering* algorithm (Marinari and Parisi (1992), and Geyer and Thompson (1995)). In simulated tempering, the states space $\mathcal{X}$ is not partitioned but, instead, some auxiliary distributions $\pi_2, \ldots, \pi_d$ are introduced (take $\pi_1 = \pi$). These distributions are chosen close to π but easier to sample from. For good performances, one typically imposes that all the distributions have the same weight. Taking each probability space $(\mathcal{X}, \mathcal{B}, \pi_i)$ as a component in the formalism above leads to a sampling problem of the form (2.1). Multicanonical sampling and simulated tempering have been combined in Atchadé and Liu (2006) giving an algorithm which can also be framed as (2.1). Sampling from (2.1) can also be an efficient strategy to improve on trans-dimensional MCMC samplers for Bayesian inference with model uncertainty. This is detailed in Section 3.3.

The main obstacle in sampling from π^* is that the normalizing constants θ^* are not known. The contribution of the Wang-Landau algorithm (Wang and Landau (2001)) is an efficient algorithm that simultaneously estimates θ^* and sample from π^*. The algorithm was introduced in a discrete setting with the π^*

being uniform in i. In this work we extend the algorithm to general state spaces and to arbitrary probability measures. To carry on the discussion in our general framework, we introduce the family of probability measures $\{\pi_\theta,\ \theta \in (0,\infty)^d\}$ on $(\mathcal{X}, \mathcal{B}, \lambda)$ defined by:

$$\pi_\theta(dx, i) \propto \frac{h_i(x)}{\theta(i)} \mathbf{1}_{\mathcal{X}_i}(x)\lambda_i(dx). \tag{2.2}$$

We assume that for all $\theta \in (0,\infty)^d$, we have at our disposal a transition kernel P_θ on $(\mathcal{X}, \mathcal{B})$ with invariant distribution π_θ. Note that π_θ and P_θ remain unchanged if we multiply the vector θ by a positive constant. How to build such Markov chain P_θ typically depends on the particular instance of the algorithm. We give some examples later.

The structure of the WL algorithm is as follows. We start out with some initial value $(X_0, I_0) \in \mathcal{X}$, $\phi_0 \in (0,\infty)^d$, and set $\theta_0(i) = \phi_0(i)/\sum_{j=1}^d \phi_0(j)$, $i = 1, \ldots, d$. Here θ_0 serves as an initial guess of θ^*. At iteration $n+1$, we generate (X_{n+1}, I_{n+1}) by sampling from $P_{\phi_n}(X_n, I_n; \cdot)$ and update ϕ_n to ϕ_{n+1}, which is used to form $\theta_{n+1}(i) = \phi_{n+1}(i)/\sum_j \phi_{n+1}(j)$. The updating rule for ϕ_n is fairly simple. For $i \in \{1, \ldots, d\}$, if $X_{n+1} \in \mathcal{X}_i$ (equivalently, if $I_{n+1} = i$), then $\phi_{n+1}(i) = \phi_n(i)(1+\rho)$ for some $\rho > 0$; otherwise $\phi_{n+1}(i) = \phi_n(i)$. This leads to a first version of the WL algorithm.

Algorithm 2.1. (The Wang-Landau algorithm I). *Let $\{\rho_n\}$ be a sequence of decreasing positive numbers. Let $(X_0, I_0) \in \mathcal{X}$ be given. Let $\phi_0 \in \mathbb{R}^d$ be such that $\phi_0(i) > 0$, and set $\theta_0(i) = \phi_0(i)/\sum_j \phi_0(j)$, $i = 1, \ldots, d$. At some time $n \geq 0$, given $(X_n, I_n) \in \mathcal{X}$, $\phi_n \in \mathbb{R}^d$, $\theta_n \in \mathbb{R}^d$:*

(i) *sample $(X_{n+1}, I_{n+1}) \sim P_{\theta_n}(X_n, I_n; \cdot)$;*

(ii) *for $i = 1, \ldots, d$, set $\phi_{n+1}(i) = \phi_n(i)(1 + \rho_n \mathbf{1}_{\{I_{n+1}=i\}})$ and $\theta_{n+1}(i) = \phi_{n+1}(i)/\sum_j \phi_{n+1}(j)$.*

It remains to choose the sequence $\{\rho_n\}$. As we show below, $\{\theta_n\}$ as defined by Algorithm 2.1 is a stochastic approximation process driven by $\{(X_n, I_n)\}$. The general guidelines in the literature to choose $\{\rho_n\}$ are: $\rho_n > 0$, $\sum \rho_n = \infty$, and $\sum \rho_n^{1+\varepsilon} < \infty$ for some $\varepsilon > 0$, often $\varepsilon = 1$. The typical choice is $\rho_n \propto n^{-1}$. In practice, more careful choices are often necessary for good performance. To the best of our knowledge, there is no general, satisfactory way of choosing the step-size in stochastic approximation. Interestingly, Wang-Landau came up with a clever, adaptive way of choosing $\{\rho_n\}$ that works very well in practice. We describe their approach next, again in more probabilistic terms.

Let $v_{n,k}(i)$ denote the proportion of visits to $\mathcal{X}_i \times \{i\}$ between times $n+1$ and k. That is, $v_{k,n}(i) = 0$ for $k \leq n$ and for $k \geq n+1$, $v_{n,k}(i) = [1/(k -$

$n)] \sum_{j=n+1}^{k} \mathbf{1}\{I_j = i\}$. Let $c \in (0,1)$ be a parameter to be specified by the user. We introduce two additional random sequences $\{\kappa_n\}$ and $\{a_n\}$. Initially, $\kappa_0 = 0$. For $n \geq 1$, define

$$\kappa_n = \inf\left\{k > \kappa_{n-1} : \max_{1\leq i\leq d}\left|v_{\kappa_{n-1},k}(i) - \frac{1}{d}\right| \leq \frac{c}{d}\right\}, \tag{2.3}$$

with the usual convention that $\inf \emptyset = \infty$. We need another sequence $\{\gamma_n\}$ of positive decreasing numbers, representing "stepsizes". Then, $\{a_n\}$ represents the index of the element of the sequence $\{\gamma_n\}$ used at time n : $a_0 = 0$, if $k = \kappa_j$ for some $j \geq 1$, then $a_k = a_{k-1} + 1$, otherwise $a_k = a_{k-1}$. In words, we start Algorithm 2.1 with a step-size equal to γ_0 and keep using it until time κ_1 when all the components are visited equally well. Only then do we change the step-size to γ_1 and keep it constant until time κ_2 etc... Combining this with Algorithm 2.1, we get the following.

Algorithm 2.2. (The Wang-Landau algorithm II). *Let $\{\gamma_n\}$ be a sequence of decreasing positive numbers. Let $(X_0, I_0) \in \mathcal{X}$ be given. Set $a_0 = 0$, $\kappa = 0$, $c \in (0,1)$, $\phi_0 \in \mathbb{R}^d$ such that $\phi_0(i) > 0$, and $\theta_0(i) = \phi_0(i)/\sum_j \phi_0(j)$, $i = 1, \ldots, d$. At some time $n \geq 0$, given $(X_n, I_n) \in \mathcal{X}$, $\phi_n \in \mathbb{R}^d$, $\theta_n \in \mathbb{R}^d$, a_n and κ:*

(i) *sample $(X_{n+1}, I_{n+1}) \sim P_{\theta_n}(X_n, I_n; \cdot)$;*

(ii) *for $i = 1, \ldots, d$, set $\phi_{n+1}(i) = \phi_n(i)(1 + \gamma_{a_n} \mathbf{1}_{\{I_{n+1}=i\}})$ and $\theta_{n+1}(i) = \phi_{n+1}(i) / \sum_j \phi_{n+1}(j)$;*

(iii) *if $\max_i |v_{\kappa,n+1}(i) - (1/d)| \leq c/d$ then set $\kappa = n + 1$ and $a_{n+1} = a_n + 1$, otherwise $a_{n+1} = a_n$.*

Remark 2.1.

1. The performances of Algorithm 2.2 depends very much on the choice of $\{\gamma_n\}$ and c. We show that the choice $\gamma_n \propto n^{-1}$ guaranties the convergence of the algorithm but, in practice, this type of step-size can be overly slow. The user might then consider $\gamma_n \propto a^{-n}$ $(a > 1)$ originally proposed by Wang-Landau, but we were not able to obtain the convergence of the algorithm for summable step-sizes. A good compromise is to start the sampler with $\gamma_n \propto a^{-n}$ until $\gamma_n < \varepsilon$ (e.g., $\varepsilon = 10^{-5}$) and then switch to $\gamma_n = n^{-1}$. There is a bias-variance trade-off involved in the choice of c. For c close to 0, Algorithm 2.2 will have a low bias (in estimating θ^*) but a high variance. Such values of c are more suitable for $\gamma_n \propto a^{-n}$ whereas for larger values of c, the bias will be high with a low variance. For $c = d - 1$, we get a standard stochastic approximation algorithm with deterministic step-size for which a step-size that sums to ∞ is necessary for convergence. We found empirically from our simulations that c in the range $0.4 - 0.2$ yields reasonably good samplers.

2. In the actual implementation of the algorithm, it is not necessary to re-normalize ϕ_n into θ_n as in (ii). In fact, for computational stability, we recommend carrying out the recursion on a logarithmic scale: $\log \phi_n(i) = \log \phi_{n-1}(i) + \log(1 + \gamma_{a_{n-1}})\mathbf{1}_{\{I_n=i\}}$.
3. Another interesting feature of Algorithm 2.2 is that Step (iii) can serve as a stopping rule: we stop the simulation when γ_{a_n} get smaller than some pre-specified value.
4. Under some regularity conditions, if $f : \mathcal{X} \to \mathbb{R}$ issome function of interest that is π-integrable and $\{(X_n, I_n, \theta_n)\}$ is as described in Algorithm 2.2, we will show below that

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k, I_k) \to \pi^*(f), \ a.s. \ \text{ as } \ n \to \infty.$$

If we denote by π_i the distribution on $\mathcal{X}_i$ with density with respect to λ_i proportional to $h_i(x)\mathbf{1}_{\mathcal{X}_i}(x)$, we can estimate integrals with respect to π_i as well:

$$\frac{\sum_{k=1}^{n} f(X_k, I_k)\mathbf{1}_{\mathcal{X}_i}(X_k)}{\sum_{k=1}^{n} \mathbf{1}_{\mathcal{X}_i}(X_k)} \to \pi_i(f(\cdot, i)), \ a.s. \ \text{ as } \ n \to \infty.$$

Now if we denote by π the distribution on $\mathcal{X}$ whose density with respect to λ is proportional to $h_i(x)$ on $\mathcal{X}_i$, the ratio $\pi(x, i)/\pi^*(x, i)$ is $d\theta^*(i)$ and integrals with respect to π can also computed by importance sampling:

$$\frac{d}{n}\sum_{k=1}^{n} f(X_k, I_k)\theta_k(I_k) \to \pi(f), \ a.s. \ \text{ as } \ n \to \infty.$$

Various methods for recycling Monte Carlo samples can be implemented as well. All these results follow from Theorem 4.1.

3. Some Applications

In this section, we detail briefly some applications of the general algorithm to multicanonical sampling, simulated tempering, and trans-dimensional MCMC.

3.1. Multicanonical sampling

Multicanonical sampling is a powerful algorithm proposed by Berg and Neuhaus (1992). It holds the potential of improving on mixing times of classical MCMC algorithms. It fits naturally in the framework above, but the implementation can be tedious. Assume that we want to sample from a probability measure $\pi(dx) \propto h(x)\lambda(dx)$ on some probability space $(\Sigma, \mathcal{A}, \lambda)$. We use the energy function $E(x) = -\log(h(x))$ to build a d-component partition $(\mathcal{X}_i)_i$ of Σ, $\mathcal{X}_i = \{x \in \Sigma : E_{i-1} < E(x) \leq E_i\}$, where $-\infty \leq E_0 < E_1 < \cdots < E_d \leq \infty$

are predefined values. Let $\theta^*(i) = \pi(\mathcal{X}_i)$ and assume $\theta^*(i) > 0$. As above, we introduce the union space $\mathcal{X} = \bigcup \mathcal{X}_i \times \{i\}$. The idea of multicanonical sampling is to sample from

$$\pi^*(dx, i) \propto \frac{h(x)}{\theta^*(i)} \mathbf{1}_{\mathcal{X}_i}(x)\lambda(dx),$$

which is of the form (2.1). There is a simpler formulation of the algorithm. Since the component of the partition to which a point x belongs can be obtained from x itself, multicanonical sampling is equivalent to sampling from π^* on $(\Sigma, \mathcal{A})$ given by

$$\pi^*(dx) \propto \sum_{i=1}^{d} \frac{h(x)}{\theta^*(i)} \mathbf{1}_{\mathcal{X}_i}(x)\lambda(dx), \tag{3.1}$$

and the union space formalism is not needed. After sampling from π^* in (3.1), a straightforward importance sampling estimate allows one to recover π. The algorithm tries to break the barriers in the energy landscape of the distribution by re-weighting each component $\mathcal{X}_i$. Clearly, the success depends heavily on a good choice of the energy rings $E_0, \ldots, E_d$. This typically requires some prior information on π or some pilot simulations. We point out that, although the energy function E is a natural candidate to utilize to partition the space, the idea can be extended to other functions.

In the description of multicanonical sampling given above, taking Σ as a discrete space, π the uniform distribution on $\mathcal{X}$, and $\mathcal{X}_e = \{x \in \Sigma : E(x) = e\}$, $e \in \{e \in \mathbb{R} : E(x) = e \text{ for some } x \in \Sigma\}$ yields the Wang-Landau algorithm of (Wang and Landau (2001)).

3.2. Simulated tempering

The method can be applied to the simulated tempering of Marinari and Parisi (1992) and Geyer and Thompson (1995) by taking $(\mathcal{X}_i, \mathcal{B}_i, \lambda_i) \equiv (\mathcal{X}_1, \mathcal{B}_1, \lambda_1)$ and $h_i = h^{1/t_i}$, $1 = t_1 < \cdots < t_d$. Simulated tempering is a well-known Monte Carlo strategy for sampling from difficult target distributions. Assume that the distribution of interest is $\pi_1(dx) \propto h(x)\lambda(dx)$. Typically for large temperature t, $h^{1/t}$ is a more well-behaved distribution for which faster mixing Markov chains can be built. In simulated tempering, we try to take advantage of these faster mixing chains by targeting the distribution

$$\pi_\theta(dx, i) = \frac{h^{1/t_i}(x)}{\theta(i)} \mathbf{1}_{\mathcal{X}_i}(x)\lambda_1(dx), \tag{3.2}$$

on the union space $(\mathcal{X}, \mathcal{B}, \lambda)$. A MCMC sampler P_θ with invariant distribution π_θ is readily designed. Typically P_θ takes the form

$$P_\theta((x, i); A \times \{j\}) = B_{x,\theta}(i, j)P^{[j]}(x, A), \tag{3.3}$$

where $B_{x,\theta}$ is a transition kernel on $\{1,\dots,d\}$ with invariant distribution $(h_j(x)/\theta(j))(\sum_i h_i(x)/\ \theta(i))^{-1}$ and $P^{[i]}$ a transition kernel on $(\mathcal{X}_1,\mathcal{B}_1)$ with invariant distribution proportional to $h_i(x)\lambda(dx)$. Typically, one takes $B_{x,\theta}(i,j) = (h_i(x)/\theta(i))(\sum_i h_i(x)/\theta(i))^{-1}$. Another common choice is to take $B_{x,\theta}$ as a Metropolis-Kernel on $\{1,\dots,d\}$ with proposal $q(i,j)$. By standard importance sampling techniques, we can convert samples from the higher temperature distributions to estimate π_1. The method holds for any $\theta \in \mathbb{R}^d$, $\theta(i) > 0$, but the choice of θ can significantly impact the efficiency. Heuristically it seems that, to improve on mixing, we need a θ that allows samples from fast converging distributions (but close to π); but since π_1 is the distribution of interest, for statistical efficiency, we need a θ that favors π_1. One easy way to resolve this trade-off is to choose θ such that all the distributions are equally visited. For this we need to choose $\theta(i) = \theta^*(i) \propto \int h_i(x)\lambda_1(dx)$ and sample from

$$\pi^*(dx,i) \propto \frac{h^{1/t_i}(x)}{\theta^*(i)} \mathbf{1}_{\mathcal{X}_i}(x)\lambda_1(dx).$$

This can be done with Algorithm 2.2.

Example 1. We compare a plain simulated tempering with weight $\theta(i) \equiv 1$ and the Wang-Landau simulated tempering described above for sampling from a multimodal bivariate Gaussian mixture distribution. The target distribution given below was taken from Liang and Wong (2001)

$$\pi(x) = \frac{1}{2\pi\sigma^2} \sum_{i=1}^{20} \omega_i \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu_i)'(x-\mu_i)\right\}, \tag{3.4}$$

where $\sigma = 0.1$ and $\omega_i \equiv 0.05$. The μ_i's are listed in Table 1. The distribution is highly multimodal and it is clear that a plain Random Walk Metropolis algorithm for this distribution does not mix in a reasonable time. Simulated tempering can be particularly efficient in such situations. We compare two strategies: a plain simulated tempering where $\theta(i) \equiv 1$ in (3.2) and the WL adaptation of simulated tempering as described above. We use the temperature scale $1 < 7.7 < 31.6 < 100$.

Table 2 presents the mean squared errors (MSEs) of the two methods in estimating the first two moments of the two components of π. We can see that the WL version is about three-four times more efficient than the plain version in terms of MSE. The estimates are based on 30 independent replications of the samplers. We ran each sampler for $100,000$ iterations. In applying Algorithm 2.2, we used $\gamma_n = 1/n$ and $c = 0.3$.

Table 1. The 20 means of the two-dimensional Gaussian mixture.

i	μ_{i1}	μ_{i2}	i	μ_{i1}	μ_{i2}	i	μ_{i1}	μ_{i2}	i	μ_{i1}	μ_{i2}
1	2.18	5.76	6	3.25	3.47	11	5.41	2.65	16	4.93	1.50
2	8.67	9.59	7	1.70	0.50	12	2.70	7.88	17	1.83	0.09
3	4.24	8.48	8	4.59	5.60	13	4.98	3.70	18	2.26	0.31
4	8.41	1.68	9	6.91	5.81	14	1.14	2.39	19	5.54	6.86
5	3.93	8.82	10	6.87	5.40	15	8.33	9.50	20	1.69	8.11

Table 2. Mean Square Errors of the plain and the WL simulated tempering algorithms. Based on 30 independent replications with 100,000 iterations of each sampler.

	$\mathrm{E}(X_1)$	$\mathrm{E}(X_2)$	$\mathrm{E}(X_1^2)$	$\mathrm{E}(X_2^2)$
Plain ST	0.113	0.132	11.201	12.501
WL-ST	0.029	0.041	2.818	4.023
Ratio	3.89	3.25	3.97	3.11

3.3. Application to trans-dimensional MCMC

It is often the case in Statistics that many alternative models are considered for the same data. One is then interested in issues like model comparison, model selection, and averaging. Let $f(Data|k, x_k)$ be the likelihood of model k with parameter x_k. Assume that we have a finite number d of models and that $x_k \in (\mathcal{X}_k, \mathcal{B}_k, \lambda_k)$. Let $\mathcal{X} = \bigcup_{i=1}^{d} \mathcal{X}_i \times \{i\}$ be the union space equipped as above with the σ-algebra $\mathcal{B}$ and the σ-finite measure λ. $(\mathcal{X}, \mathcal{B}, \lambda)$ is the natural space to consider when dealing both with model uncertainty and parameter estimation. In the Bayesian framework, a prior density (with respect to λ) $p(x_k, k)$ in $(\mathcal{X}, \mathcal{B})$ is specified for (x_k, k). The posterior distribution of (x_k, k) is therefore $\pi(x_k, k) \propto h_k(x_k) = f(Data|k, x_k)p(x_k, k)$. In this framework, one is often interested in the Bayes factor of model i to model j defined as $B_{ij} := \theta^*(i)p(j)/(\theta^*(j)p(i))$, where $\theta^*(i) \propto \int_{\mathcal{X}_i} \pi(x_i, i)\lambda_i(dx_i)$ and $p(i) = \int p(x_i, i)\lambda_i(dx_i)$. Trans-dimensional MCMC is a set of specialized MCMC algorithms to sample from distributions like π defined on spaces of variable dimensions. The reversible-jump algorithm of Green (Green (1995)) is the most popular such sampler.

In the spirit of the WL algorithm, an alternative to sampling directly from π is to sample from the distribution

$$\pi^*(dx_i, i) \propto \frac{h_i(x_i)}{\theta^*(i)} \mathbf{1}_{\mathcal{X}_i}(x)\lambda_i(dx_i). \tag{3.5}$$

By such re-weighting, we give the same posterior weight to all the models. The WL algorithm then offers an effective strategy to sample from π^* and we recover π by importance sampling. This strategy can improve on the mixing of the sampler

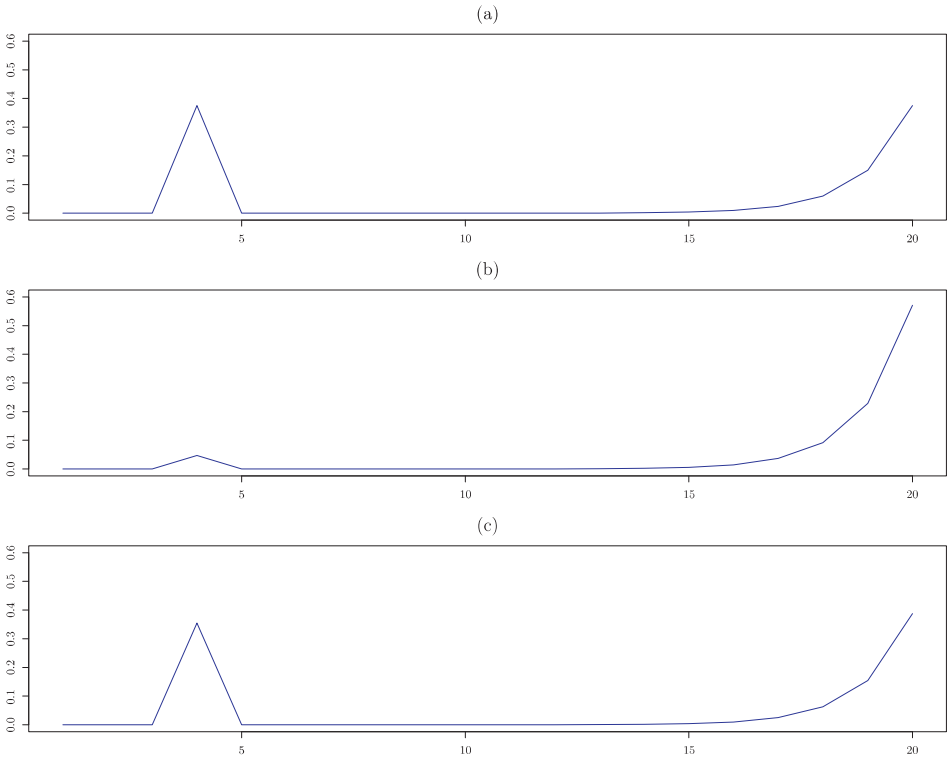

Figure 1. Marginal posterior distribution of models. (a) estimate from the plain WL-RJMCMC; (b) estimate from the RJMCMC; (c) true posterior distribution. Estimates were based on 10×10^6 iterations for the plain RJMCMC and 2×10^6 iterations for the WL-RJMCMC.

Example 2. We set $\mathcal{X}_i = \mathbb{R}^i$ for $i = 1, \ldots, 20$, and consider the following rather trivial trans-dimensional target distribution:

$$\pi(x_i, i) \propto a_i^{-1} e^{-|x_i|^2/2},$$

where we let $a_i = 1$ for $i \neq 4$, and $a_4 = (2\pi)^{-16/2}$. In this distribution, $x_i \in \mathbb{R}^i$, the i-dimensional Euclidean space, and $\pi(x_i, i)$ restricted to $\mathbb{R}^i$ is proportional to the standard normal distribution. We are interested in the marginal distribution $p(i)$ of i. This distribution, as shown in Figure 1, is bimodal with modes at 4 and 20.

We pretend that this distribution is intractable and sample from it using a Birth-and-Death Reversible-Jump MCMC. For the fixed-dimensional move, we use a Random Walk Metropolis kernel with a Gaussian proposal with covariance matrix $\sigma_p I_i$, $\sigma_p = 0.1$. We implement a Birth-and-Death move for the trans-dimensional jump. Given (x, i), we randomly select $j \in \{i - 1, i + 1\}$ with

respective probability $\omega_{i,i-1}, \omega_{i,i+1}$. We choose $\omega_{i,i+1} = 1/2$ with the usual correction at the boundaries. If $j = i+1$, we propose $y = (x_i, u)$ where $u \sim N(0, \sigma^2)$ with $\sigma = 0.1$. We accept (y, j) with probability $\min(1, A)$, where

$$A = \frac{\pi(y,j)}{\pi(x,i)} \frac{\omega_{ji}}{\omega_{ij}} \sqrt{2\pi}\sigma e^{u^2/2\sigma^2}.$$

Similarly if $j = i - 1$, we write $x = (y, u')$ with $u' \in \mathbb{R}$ and propose (y, j). This value is then accepted with probability $\min(1, A)$ with

$$A = \frac{\pi(y,j)}{\pi(x,i)} \frac{\omega_{ji}}{\omega_{ij}} \frac{1}{\sqrt{2\pi}\sigma} e^{-u'^2/2\sigma^2}.$$

This vanilla RJMCMC sampler fails to sample from π. Depending on its starting point, the sampler typically found one of the two modes and got stuck there even after 10 millions iterations (see Figure 2a). In contrast, the WL algorithm provided a reasonable estimate of the distribution in only 2 millions iterations. In this example, computationally, each WL iteration step costs roughly that for 1.2 step of the vanilla sampler. For the WL approach we used $c = 0.4$ and $\gamma_n = 2^{-n}$ until 10^{-4} before switching to $\gamma_n = n^{-1}$.

4. Some Theoretical Results

We look at some theoretical aspects of the algorithm. We investigate the convergence of θ_n to θ^* and a strong law of large numbers for $\{(X_n, I_n)\}$. The difficulty comes in proving that the algorithm is *stable* in the following sense.

Definition 4.1. Let $v_n(i)$ be the occupation measure of $\mathcal{X}_i$ by time $n : v_n(i) = (1/n)\sum_{k=1}^n \mathbf{1}_{\{I_k=i\}}$. The Wang-Landau algorithm is said to be stable if

$$\max_{i,j} \limsup_{n\to\infty} n(v_n(i) - v_n(j)) < \infty, \ a.s.. \tag{4.1}$$

This is an essential property of the algorithm. When the algorithm is *stable*, we show that all the stopping times κ_l defined in Algorithm 2.2 are finite and the step-size γ_{a_n} will gradually converges to 0. Moreover (Theorem 4.1 below) on a path where the algorithm is stable, $\theta_n \to \theta^*$ and a strong law of large numbers holds for $(1/n)\sum_{k=1}^n f(X_k, I_k)$. Then, we derive some verifiable conditions under which the algorithm is shown to be stable. To maintain the flow of ideas, some of the proofs are postponed to Section 6.

4.1. Ergodicity

Let $\Theta = \{\theta \in \mathbb{R}^d : \sum_{i=1}^d \theta(i) = 1, \ \theta(i) \in (0,1), \ i = 1, \ldots, d\}$ and $((X_0, I_0), \theta_0) \in \mathcal{X} \times \Theta$ be the initial state of the algorithm. This initial state

is considered fixed but arbitrary. Let $\{\gamma_n\}$ be the step-size sequence. Let Pr be the distribution of the process $\{X_n, I_n, \theta_n\}$ started at (X_0, I_0, θ_0) with step-size sequence $\{\gamma_n\}$, and let E be the expectation with respect to Pr. To simplify the notation, we omit explicit mention of the dependence of Pr on (X_0, I_0, θ_0) and $\{\gamma_n\}$. All statements made almost surely are with respect to Pr. For $\theta \in \mathbb{R}^d$, $|\theta|$ denotes the Euclidean norm of θ. For any $\varepsilon \in (0, \varepsilon_*)$, where $\varepsilon_* = \min_i \theta^*(i)$, let $\Theta_\varepsilon = \{\theta \in \Theta,\ \theta(i) \geq \varepsilon,\ i = 1, \ldots, d\}$. Our main assumption is that the family $\{P_\theta, \theta \in \Theta_\varepsilon\}$ is Lipschitz and uniformly V-ergodic. See e.g., Andrieu and Moulines (2006) for some examples of MCMC samplers where these assumptions hold. Before stating them, we need some notation. For any functions $f : \mathcal{X} \to \mathbb{R}$ and $W : \mathcal{X} \to [1, \infty)$ we take $|f|_W := \sup_{x \in \mathcal{X}}[(|f(x)|)/(W(x))]$, and introduce the set of W-bounded functions $L_W := \{f \text{ meas.},\ f : \mathcal{X} \to \mathbb{R}, |f|_W < \infty\}$. A transition kernel on $(\mathcal{X}, \mathcal{B})$ operates on measurable real-valued functions f as $Pf(x) = \int P(x, dy) f(y)$, and the product of two transition kernels P_1 and P_2 is the transition kernel $P_1P_2(x, A) = \int P_1(x, dy)P_2(y, A)$. For two transition kernels P_1 and P_2, $\|\!| P_1 - P_2 |\!\|_W$, the W-distance between P_1 and P_2 is

$$\|\!| P_1 - P_2 |\!\|_W := \sup_{|f|_W \leq 1} |P_1 f - P_2 f|_W \,.$$

(**A1**) There is a measurable function $V : \mathcal{X} \longrightarrow [1, \infty)$, a set $C \subset \mathcal{X}$, and a probability measure ν on $(\mathcal{X}, \mathcal{B})$ such that $\nu(C) > 0$, with the following property. For all $\varepsilon \in (0, \varepsilon_*)$, we can find constants $\lambda_\varepsilon \in (0, 1)$, $b_\varepsilon \in [0, \infty)$, $\beta_\varepsilon \in (0, 1]$ and integer $n_{0,\varepsilon}$ such that:

$$\inf_{\theta \in \Theta_\varepsilon} P_\theta^{n_{0,\varepsilon}}(x, A) \geq \beta_\varepsilon \nu(A) \mathbf{1}_C(x), \quad x \in \mathcal{X},\ A \in \mathcal{B}; \tag{4.2}$$

$$\sup_{\theta \in \Theta_\varepsilon} P_\theta V(x) \leq \lambda_\varepsilon V(x) + b_\varepsilon \mathbf{1}_C(x), \quad x \in \mathcal{X}. \tag{4.3}$$

The inequality (4.3) of (A1) is the so-called drift condition and (4.2) is the so-called minorization condition.

(**A2**) For all $\alpha \in [0, 1]$, for all $\varepsilon \in (0, \varepsilon_*)$, there exists $K = K(\alpha, \varepsilon) < \infty$ such that, for all θ, $\theta' \in \Theta_\varepsilon$,

$$\|\!| P_\theta - P_{\theta'} |\!\|_{V^\alpha} \leq K \left|\theta - \theta'\right|, \tag{4.4}$$

where V is defined in (A1).

(**A3**) $\{\gamma_n\}$ is non-increasing, $\gamma_n > 0$, $\sum \gamma_n = \infty$, and $\gamma_n = O(n^{-1})$ as $n \to \infty$.

For $B > 0$, we introduce the stopping time:

$$\tau(B) := \inf\Big\{k \geq 0 : \max_{i,j} k(v_k(i) - v_k(j)) > B\Big\}, \tag{4.5}$$

with the usual convention that $\inf \emptyset = \infty$. $\tau(B)$ is the first time that one component accumulates B visits or more than some other component. Thus, Definition 4.1 is precisely equivalent to $\tau(B) = \infty$ for some B. With this in mind, the next results says essentially that if the algorithm is *stable*, and (A1-3) hold, then it is ergodic.

Theorem 4.1. *Let* (A1−3) *hold and* $B > 0$ *be given. Then*

(i)

$$|\theta_n - \theta^*|\, \mathbf{1}_{\{\tau(B)>n\}} \to 0, \ a.s. \ \text{ as } \ n \to \infty. \tag{4.6}$$

(ii) *For any function* $f \in L_{V^{1/2}}$, *writing* $\bar{f} = f - \pi^*(f)$, *we have*

$$\frac{1}{n}\sum_{k=1}^{n} \bar{f}(X_k, I_k)\mathbf{1}_{\{\tau(B)>k-1\}} \to 0, \quad a.s. \ \text{ as } \ n \to \infty. \tag{4.7}$$

Proof. See Section 6.2.

4.2. Checking (A1–2)

We impose the drift condition and the minorization condition (A1) uniformly for $\theta \in \Theta_\varepsilon$, not uniformly for $\theta \in \Theta$. This is an important point. Indeed, and as we see now, a drift and minorization condition uniformly-in-θ for $\theta \in \Theta_\varepsilon$ is almost always true as soon as one P_θ satisfies these conditions (Proposition 4.1), whereas a minorization and drift condition uniformly-in-θ for $\theta \in \Theta$ is almost never true.

Indeed, suppose that each P_θ is a Metropolis-Hastings kernel with invariant distribution π_θ and proposal kernel Q. That is $P_\theta f(x,i) = M_\theta f(x,i) + r_\theta(x,i)f(x,i)$, where

$$M_\theta f(x,i) = \sum_{j=1}^{d} \int_{\mathcal{X}_j} \min\left(1, \frac{\theta(i)}{\theta(j)} R(y,j;x,i)\right) f(y,j) Q(x,i;dy,j),$$

$$r_\theta(x,i) = 1 - \sum_{j=1}^{d} \int_{\mathcal{X}_j} \min\left(1, \frac{\theta(i)}{\theta(j)} R(y,j;x,i)\right) Q(x,i;dy,j),$$

and $R(x,i;y,j)$ the Radon-Nikodym density of $\pi(dx,i)Q(x,i;dy,j)$ with respect to $\pi(dy,j)Q(y,j;dx,i)$.

With $\theta = (1, \ldots, 1)$, we use π (resp. P and M) to denote π_θ (resp. P_θ and M_θ). The next result states that we only need to check that P is geometrically ergodic to obtain (A1-2).

Proposition 4.1. *Suppose there is a measurable function $V : \mathcal{X} \longrightarrow [1, \infty)$, a set $C \subset \mathcal{X}$, a probability measure ν on $(\mathcal{X}, \mathcal{B})$ such that $\nu(C) > 0$, and constants $\lambda \in (0, 1)$, $b \in [0, \infty)$, $\beta \in (0, 1]$ and finite integer n_0 such that $M^{n_0}(x, A) \geq \beta\nu(A)\mathbf{1}_C(x)$, $x \in \mathcal{X}$, $A \in \mathcal{B}$, and $PV(x) \leq \lambda V(x) + b\mathbf{1}_C(x)$, $x \in \mathcal{X}$. Then* (A1−2) *hold.*

Proof. See Section 6.1.

4.3. Stability

Theorem 4.1 asserts that under (A1−3), the WL algorithm converges to the right limit on *stable* paths. This raises the question of checking the stability of the algorithm, difficult in general. The next theorem gives some easily checked conditions under which the algorithm is stable.

Theorem 4.2. *The WL algorithm is stable under either of the following two conditions.*

(a) *There exist $\varepsilon \in (0, 1)$, $K \in (0, \infty)$ and integer $n_0 \geq 0$ such that for any $i, j \in \{1, \ldots, d\}$ and $\theta \in \mathbb{R}^d$, $\theta(i)/\theta(j) > K$ implies that $P_\theta^{n_0}((x, i), \mathcal{X}_j \times \{j\}) \geq \varepsilon$ for all $x \in \mathcal{X}_i$.*

(b) *There exists $\varepsilon \in (0, 1)$, such that for any $j \in \{1, \ldots, d\}$ and $\theta \in \mathbb{R}^d$, $\theta(j) \leq \min_{1 \leq i \leq d} \theta(i)$ implies $P_\theta((x, i), \mathcal{X}_j \times \{j\}) \geq \varepsilon$ for all $x \notin \mathcal{X}_i$ and all $i \neq j$.*

Proof. See Section 6.3.

4.4. Application to multicanonical sampling

Consider the multicanonical sampling of Section 3.1. Suppose that P_θ is the *Independence-Sampler* with proposal distribution $Q(dx) = q(x)\lambda(dx)$ and invariant distribution $\pi_\theta(dx) \propto \sum_{i=1}^d (h(x)/\theta(i))\mathbf{1}_{\mathcal{X}_i}(x)\lambda(dx)$. Assume that the function $\omega(x) \propto h(x)/q(x)$ is bounded with supremum ω_0. Then clearly, for all $x \in \mathcal{X}_i$,

$$P_\theta(x, \mathcal{X}_j) \geq \min\left(1, \frac{\theta(i)}{\theta(j)}\right) \int_{\mathcal{X}_j} \min\left(1, \frac{\omega(y)}{\omega_0}\right) Q(dy) \geq \varepsilon_j,$$

as soon as $\theta(i) \geq \theta(j)$, taking $\varepsilon_j = \int_{\mathcal{X}_j} \min(1, [(\omega(y))/(\omega_0)])Q(dy) > 0$ (since $\pi(\mathcal{X}_i) > 0$). Thus for any $i, j \in \{1, \ldots, d\}$, $i \rightsquigarrow j$ and, by Theorem 4.2, the WL algorithm is *stable*. Now, since ω is bounded, each P_θ satisfies a drift condition and a minorization condition, which implies (A1−2) by Proposition 4.1.

Similarly, if P_θ is a Random Walk Metropolis with proposal kernel $q(y-x)$ we have

$$P_\theta(x, \mathcal{X}_j) \geq \min\left(1, \frac{\theta(i)}{\theta(j)}\right) \int_{\mathcal{X}_j} \min\left(1, \frac{\pi(y)}{\pi(x)}\right) q(y-x)dy.$$

It follows that if $\mathcal{X}$ is compact and π, q are positive and continuous, then $i \rightsquigarrow j$ for all i, j. Under the same assumption, (A1−2) also hold.

Corollary 4.1. *In the case of multicanonical sampling of Section* 3.1*, assume one of*

(i) *P_θ is an independent-Metropolis sampler with proposal distribution $Q(dx) = q(x)\lambda(dx)$ and $\omega \propto h/q$ is bounded;*

(ii) *P_θ is a RWM sampler with proposal $q(y-x)$; $\mathcal{X}$ is compact, and π and q are positive and continuous.*

Then the algorithm is stable and, under (A3),

$$|\theta_n - \theta^*| \to 0; \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \pi^*(f) \quad a.s. \quad \text{as} \quad n \to \infty$$

for any bounded measurable function f.

4.5. Application to simulated tempering

Theorems 4.1. and 4.2. can also be applied to the WL version of simulated tempering as described in Section 3.2. We consider the simulated tempering algorithm with kernel $P_\theta((x,i); A \times \{j\}) = B_{x,\theta}(i,j)P^{[j]}(x,A)$ where $B_{x,\theta}(i,j) \propto h_j(x)/\theta(j)$ and $P^{[j]}$ is a transition kernel (not necessarily Metorpolis-Hastings) with invariant distribution h_j. The following corollary is easily proved and omitted.

Corollary 4.2. *Suppose that $\mathcal{X}_1$ is compact, h positive and continuous, and $\{\gamma_n\}$ satisfies* (A3). *Suppose that there exist $\varepsilon > 0$, a probability measure ν and an integer n_0 such that $\nu(\mathcal{X}_i) > 0$ and $(P^{[i]})^{n_0}(x,A) \geq \varepsilon\nu(A)$, $i \in \{1, \ldots, d\}$. Then*

$$|\theta_n - \theta^*| \to 0; \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n} f(X_k, I_k) \to \pi^*(f) \quad a.s. \quad \text{as} \quad n \to \infty$$

for any bounded measurable function f.

5. Discussion and Open Problems

In this paper, we propose an extension of the WL algorithm to general state spaces. The WL algorithm differs from other adaptive Markov Chain Monte

Carlo algorithms based on stochastic approximation by the adaptive nature of its step-size. We have shown through examples that the algorithm can be used effectively to improve on simulated tempering and trans-dimensional MCMC algorithms. We have also studied the asymptotic behavior of the WL algorithm. We have shown that on stable sample paths, and with an appropriate step-size, θ_n converges to θ^* and a strong law of large numbers holds. Finally, when the state space is compact we have shown that, in most cases, the algorithm is stable and the aforementioned limit results apply.

Two main questions have remained unanswered. The first concerns the stability of the algorithm in unbounded state spaces. Second, in order to exploit the full potential of the algorithm, a more precise understanding of its efficiency is needed. In particular, we need to understand how the rate of convergence and the asymptotic variances of the algorithm are related to the parameter c and step-size γ_n. We are currently investigating some of these questions.

6. Proofs

The techniques used here can be found in various forms in Benveniste, Métivier and Priouret (1990), Delyon (1996), and Andrieu, Moulines and Priouret (2005). Throughout, C_ε denotes a generic constant whose value can be different from one equation to another. The key result in the proof of Theorem 4.1 is Lemma 6.6, which states that the weighted sum of the noise process in the stochastic approximation followed by $\{\theta_n\}$ is summable.

6.1. Proof of Proposition 4.1

Lemma 6.1. *Under the conditions of Proposition* 4.1, (A2) *holds.*

Proof. Let $\varepsilon \in (0, \varepsilon^*)$ and $\alpha \in (0, 1]$. For $\theta \in \Theta_\varepsilon$, $|f| \le V^\alpha$, and $(x, i) \in \mathcal{X}$, we have $P_\theta f(x,i) = M_\theta f(x,i) + f(x,i)(1 - M_\theta \mathbf{e}(x,i))$, where $\mathbf{e}(x,i) \equiv 1$, and

$$M_\theta f(x,i) = \sum_{j=1}^{d} \int_{\mathcal{X}_j} \min\left(1, \frac{\theta(i)}{\theta(j)} r(y,j;x,i)\right) f(y,j) Q(x,i;dy,j).$$

As a consequence, for $\theta_2, \theta_1 \in \Theta_\varepsilon$, $\|P_{\theta_2} - P_{\theta_1}\|_{V^\alpha} \le \|M_{\theta_2} - M_{\theta_1}\|_{V^\alpha} + \|M_{\theta_2} - M_{\theta_1}\|_{TV}$, where $\|P\|_{TV} = \|P\|_V$ with $V \equiv 1$. For $(x,i) \in \mathcal{X}$, the function $\theta \to M_\theta f(x,i)$ is differentiable and:

$$\sum_{j=1}^{d} \left| \frac{\partial}{\partial \theta_j} M_\theta f(x,i) \right| \le C_\varepsilon \sum_{j=1}^{d} M\,|f|\,(x,i).$$

(A2) then follows by the Mean Value Theorem and the drift condition on P.

Next we show that for any $\varepsilon \in (0, \varepsilon^*)$, the family $(P_\theta)_{\theta\in\Theta_\varepsilon})$ satisfies a uniform (in θ) drift condition.

Lemma 6.2. *Under the conditions of Proposition* 4.1, (A1) *holds.*

Proof. The minorization is immediate. Since $\min(1, ab) \geq \min(1, a)\min(1, b)$ for $a, b > 0$, $P_\theta(x, i; A) \geq M_\theta(x, i; A) \geq (1/\varepsilon)M_0(x, i; A) \geq \beta/\varepsilon\nu(A)\mathbf{1}_C(x, i)$.

The argument to show the uniform drift is fairly simple, and we only sketch it here. Let $B_{\theta,r} = \Theta_\varepsilon \cap B(\theta, r)$ the open ball of Θ_ε with center $\theta \in \Theta_\varepsilon$ and radius $r > 0$. It follows from Lemma 6.1 that by choosing $r > 0$ small enough, if P_θ satisfies a drift condition toward a small set C, then the family $\{P_{\theta'},\ \theta' \in B_{\theta,r}\}$ satisfies a uniform (in θ') drift condition toward C. Therefore, starting from P, we can find an open coverage of Θ_ε by the $B_{\theta,r}$ such that, on each such $B_{\theta,r}$, a uniform drift toward C holds. Since Θ_ε is compact it admits a finite coverage by the $B_{\theta,r}$ and, using the maximum of the constants of the drift condition toward C for each such ball, we get a uniform drift toward C for $\{P_\theta,\ \theta \in \Theta_\varepsilon\}$.

6.2. Proof of Theorem 4.1.

We start with some preliminary remarks. For any $\varepsilon \in (0, \varepsilon_*)$ and $\alpha \in (0, 1]$, it is known from Markov chain theory that (A1) implies the existence of $C_{\varepsilon,\alpha} < \infty$, $\rho_{\varepsilon,\alpha} \in (0, 1)$ and b_ε as in (A1), such that

$$\sup_{\theta\in\Theta_\varepsilon} \|\!| P_\theta^n - \pi_\theta |\!\|_{V^\alpha} \leq C_{\varepsilon,\alpha}\rho_{\varepsilon,\alpha}^n, \tag{6.1}$$

$$\sup_{\theta\in\Theta_\varepsilon} \pi_\theta(V) \leq b_\varepsilon. \tag{6.2}$$

For a proof, see e.g., Baxendale (2005) and the references therein. Define $\xi(\varepsilon) = \inf\{k \geq 0 : \theta_k \notin \Theta_\varepsilon\}$. An easy calculation using (A1) gives that

$$\begin{aligned}\mathrm{E}\left[V(X_n, I_n)\mathbf{1}_{\{\xi(\varepsilon)>n\}}\right] &= \mathrm{E}\left[V(X_n, I_n)\mathbf{1}_{\Theta_\varepsilon}(\theta_0)\cdots\mathbf{1}_{\Theta_\varepsilon}(\theta_n)\right] \\ &\leq \lambda_\varepsilon^n V(X_0, I_0) + \frac{b_\varepsilon}{(1-\lambda_\varepsilon)}\end{aligned}$$

from which we deduce that

$$\sup_n \mathrm{E}\left(V(X_n, I_n)\mathbf{1}_{\{\xi(\varepsilon)>n\}}\right) < \infty. \tag{6.3}$$

The proof of the theorem is based on some nice properties of solutions of the so-called Poisson equation. These solutions allow us to obtain a martingale approximation to the process $\sum_{k=1}^n f(X_k, I_k)$. For $f \in L_{V^\alpha}$, $\alpha \in (0, 1]$ and $\theta \in \Theta_\varepsilon$, let

$$h_\theta = \sum_{k=0}^{\infty} P_\theta^k\Big(f - \pi_\theta(f)\Big). \tag{6.4}$$

h_θ solves the Poisson equation $f - \pi_\theta(f) = h_\theta - P_\theta h_\theta$. By (A1), h_θ exists and $h_\theta \in L_{V^\alpha}$. For $f \in L_{V^\alpha}$ and $\theta, \theta' \in \Theta_\varepsilon$, writing $\bar{f}_\theta = f - \pi_\theta(f)$, we have

$$\begin{aligned}
|\pi_\theta(f) - \pi_{\theta'}(f)| &= \left|\pi_\theta\left[P_\theta^k \bar{f}_{\theta'}\right]\right| \\
&= \left|\pi_\theta\left[P_{\theta'}^k(\bar{f}_{\theta'})\right] + \sum_{j=1}^{k} \pi_\theta\left[P_\theta^{k-j}(P_\theta - P_{\theta'})P_{\theta'}^{j-1}(\bar{f}_{\theta'})\right]\right| \\
&\le C_\varepsilon\left(\rho_\varepsilon^k + \left|\theta - \theta'\right| \sum_{j=1}^{k} \rho_\varepsilon^{j-1}\right),
\end{aligned}$$

using (A1−2), from which we deduce that there exists a finite constant C_ε such that

$$\sup_{f \in L_{V^\alpha}} |\pi_\theta(f) - \pi_{\theta'}(f)| \le C_\varepsilon \left|\theta - \theta'\right|. \tag{6.5}$$

The constant C_ε is not necessarily the same from one equation to the other. Similarly, with $\bar{P}_\theta$ as the operator $P_\theta - \pi_\theta$, we have

$$\begin{aligned}
\left|P_\theta^k \bar{f}_\theta - P_{\theta'}^k \bar{f}_{\theta'}\right| &= \left|\bar{P}_\theta^k f - \bar{P}_{\theta'}^k f\right| \\
&= \left|\sum_{j=1}^{k} \bar{P}_\theta^{k-j}(P_\theta - P_{\theta'})\bar{P}_{\theta'}^{j-1} f\right| \\
&\le C_\varepsilon \left|\theta - \theta'\right| k \rho_\varepsilon^{k-1} V^\alpha,
\end{aligned}$$

using (A1−2). This, together with (6.5), implies the existence of a finite constant C_ε such that for all $\alpha \in (0, 1]$, $\theta, \theta' \in \Theta_\varepsilon$,

$$|h_\theta - h_{\theta'}|_{V^\alpha} + |P_\theta h_\theta - P_{\theta'} h_{\theta'}|_{V^\alpha} \le C_\varepsilon \left|\theta - \theta'\right|. \tag{6.6}$$

In our analysis, we mainly see $\{\theta_n\}$ as a stochastic approximation sequence. The recursion on $\{\theta_n\}$ has

$$\begin{aligned}
\theta_{n+1}(i) &= \frac{\phi_{n+1}(i)}{\sum_{e=1}^{d} \phi_{n+1}(e)} \\
&= \frac{\phi_n(i) + \gamma_{a_n}\phi_n(i)\mathbf{1}_{\{I_{n+1}=i\}}}{\sum_{e=1}^{d} \phi_n(e) + \gamma_{a_n}\phi_n(I_{n+1})} \\
&= \theta_n(i)\frac{1 + \gamma_{a_n}\mathbf{1}_{\{I_{n+1}=i\}}}{1 + \gamma_{a_n}\theta_n(I_{n+1})} \\
&= \theta_n(i) + \gamma_{a_n} H_i(\theta_n, I_{n+1}) + \gamma_{a_n}^2 r_{i,n}(\theta_n, I_{n+1}),
\end{aligned} \tag{6.7}$$

where $H_i(\theta, I) = \theta(i)(\mathbf{1}_{\{I=i\}} - \theta(I))$ and $r_{i,n}(\theta, I) = -\theta(i)\theta(I)[(\mathbf{1}_{\{I=i\}} - \theta(I))/(1 + \gamma_{a_n}\theta(I))]$.

The mean field function $h_i(\theta) = \pi_\theta(H_i)$ is

$$h_i(\theta) = \frac{\theta^*(i) - \theta(i)}{\sum_{j=1}^d \theta^*(j)/\theta(j)}. \tag{6.8}$$

Lemma 6.3. *Under* (A3), *let* $B > 0$ *be given and* $c_0 = 2Bd/c$. *Then on* $\{\tau(B) > n\}$, $a_n \geq \lfloor n/c_0 \rfloor$. *Moreover* $\sum \gamma_{a_n} = \infty$, *and* $\sum \gamma_{a_n}^2 \mathbf{1}_{\{\tau(B)>n\}} < \infty$ *almost surely.*

Proof. On $\{\tau(B) > n\}$, for any $k < k' \leq n$, and for any $i \in \{1, \ldots, d\}$, $(v_{k,k'}(i) - 1/d) \leq 2B/(k' - k)$. Therefore if $lc_0 \leq n < \tau(B)$, $\kappa_l \leq n$. That is, $a_n \geq \lfloor n/c_0 \rfloor$ on $\{\tau(B) > n\}$.

Since $\{\gamma_n\}$ is non-increasing and $a_n \leq n$, $\sum \gamma_{a_n} \geq \sum \gamma_n = \infty$. On the other hand $\sum \gamma_{a_n}^2 \mathbf{1}_{\{\tau(B)>n\}} \leq Cc_0 \sum n^{-2} < \infty$.

For any $\varepsilon > 0$, we introduce the stopping time $\xi(\varepsilon) := \inf\{k \geq 0 : \theta_k \notin \Theta_\varepsilon\}$. We need the following result. The proof is omitted.

Lemma 6.4. *Let* $\{\gamma_n,\ n \geq 0\}$ *be a non-increasing sequence of positive number and* $\{v_n, n \geq 0\}$ *be a sequence of numbers such that* $|\sum_{k=0}^N v_k| \leq B$ *for all* $N \geq 0$. *Then* $|\sum_{k=0}^N \gamma_k v_k| \leq \gamma_0 B$ *for all* $N \geq 0$.

The following lemma relates $\tau(B)$ and $\xi(\varepsilon)$.

Lemma 6.5. *Under* (A3), *for any* $B > 0$ *we can find* $\varepsilon \in (0, \varepsilon^*)$ *such that* $\tau(B) \leq \xi(\varepsilon)$.

Proof. Take $\varepsilon = (1 + (d-1)e^{B\gamma_0})^{-1} > 0$. Without any loss, we assume that $\varepsilon < \varepsilon*$ and $\varepsilon \leq \min_i \theta_0(i)$. We need to show that $\min_i \theta_n(i) > \varepsilon$ for all $n < \tau(B)$. But $\theta_n(i) = (1 + \sum_{j \neq i}[(\phi_n(j))/(\phi_n(i))])^{-1}$. It is thus enough to show that $\phi_n(j)/\phi_n(i) \leq e^{B\gamma_0}$ for all $i \neq j$ and for any $n < \tau(B)$. But

$$\frac{\phi_n(j)}{\phi_n(i)} = \exp\left(\sum_{p=0}^\infty \gamma_p \Big(N_{\kappa_p, n\wedge\kappa_{p+1}}(j) - N_{\kappa_p, n\wedge\kappa_{p+1}}(i)\Big)\right),$$

where $N_{l,m}(i) = 0$ if $m \leq l$ and $N_{l,m}(i) = \sum_{q=l+1}^m \mathbf{1}_{\mathcal{X}_i}(X_q)$ otherwise ($N_{l,m}(i)$ is the number of visits to $\mathcal{X}_i$ from time $l+1$ to m). For any $n < \tau(B)$, $|\sum_{p=0}^P (N_{\kappa_p, n\wedge\kappa_{p+1}}(i) - N_{\kappa_p, n\wedge\kappa_{p+1}}(j))| \leq B$ for all $P \geq 0$. Lemma 6.4 thus implies that $\phi_n(i)/\phi_n(j) \leq e^{\gamma_0 B}$.

Lemma 6.6. *Under* (A1–3), *let* $B > 0$ *be given. Let* $\{\gamma'_n\}$ *be a sequence that satisfies* (A3) *such that* $\sum \gamma_{\lfloor n/a \rfloor} \gamma'_{\lfloor n/a \rfloor} < \infty$ *for all* $a > 0$. *For* $\theta \in \Theta$, *Let* $H_\theta : \mathcal{X} \to \mathbb{R}$ *be a measurable function such that* $H_\theta \in L_{V^{1/2}}$. *Then*

$$\sum_{k=0}^\infty \gamma'_{a_k} \mathbf{1}_{\{\tau(B)>k\}} \Big[H_{\theta_k}(X_{k+1}, I_{k+1}) - \pi_{\theta_k}(H_{\theta_k})\Big] < \infty, \quad a.s.. \tag{6.9}$$

Proof. Let $\varepsilon > 0$ as in Lemma 6.5. From (A1) there is a function $h_\theta \in L_{V^{1/2}}$ that solves the Poisson equation $h_\theta - P_\theta h_\theta = H_\theta - \pi_\theta(H_\theta)$ for all $\theta \in \Theta_\varepsilon$. Using this we can write

$$\begin{aligned}
&\sum_{k=0}^{n} \gamma'_{a_k} \mathbf{1}_{\{\tau(B)>k\}} \Big[H_{\theta_k}(X_{k+1}, I_{k+1}) - \pi_{\theta_k}(H_{\theta_k}) \Big] \\
&\quad = \sum_{k=0}^{n} \gamma'_{a_k} \mathbf{1}_{\{\tau(B)>k\}} \Big(U^{(1)}_{k+1} + U^{(2)}_{k+1} + U^{(3)}_{k+1} \Big),
\end{aligned}$$

where

$$\begin{aligned}
U^{(1)}_{k+1} &= h_{\theta_k}(X_{k+1}, I_{k+1}) - P_{\theta_k} h_{\theta_k}(X_k, I_k), \\
U^{(2)}_{k+1} &= P_{\theta_k} h_{\theta_k}(X_k, I_k) - P_{\theta_{k+1}} h_{\theta_{k+1}}(X_{k+1}, I_{k+1}), \\
U^{(3)}_{k+1} &= P_{\theta_{k+1}} h_{\theta_{k+1}}(X_{k+1}, I_{k+1}) - P_{\theta_k} h_{\theta_k}(X_{k+1}, I_{k+1}).
\end{aligned}$$

Clearly $M_n = \sum_{k=0}^{n} \gamma'_{a_k} \mathbf{1}_{\{\tau(B)>k\}} U^{(1)}_{k+1}$ is a martingale; using Lemma 6.3, (6.3) and since h_θ and $P_\theta h_\theta \in L_{V^{1/2}}$, we have

$$\begin{aligned}
\mathrm{E}\,(M_n^2) &\le C_\varepsilon \sum_{k=0}^{n} \mathrm{E}\left[(\gamma'_{a_k})^2 \mathbf{1}_{\{\tau(B)>k\}} V(X_{k+1}, I_{k+1}) \right] \\
&\le C_\varepsilon \sum_{k=0}^{n} (\gamma'_{\lfloor k/c_0 \rfloor})^2 \mathrm{E}\left[\mathbf{1}_{\{\tau(B)>k\}} V(X_{k+1}, I_{k+1}) \right] \\
&\le C_\varepsilon \sum_{k=0}^{\infty} (\gamma'_{\lfloor k/c_0 \rfloor})^2 < \infty.
\end{aligned}$$

By Doob's Convergence Theorem for martingales, $\sum_{k=0}^{\infty} \gamma'_{a_k} \mathbf{1}_{\{\tau(B)>k\}} U^{(1)}_{k+1}$ is finite a.s.

On $\{\tau(B) = l\}$ $(l < \infty)$, $\sum_{k=0}^{\infty} \gamma'_{a_k} \mathbf{1}_{\{\tau(B)>k\}} U^{(2)}_{k+1} = \sum_{k=0}^{l-1} \gamma'_{a_k} U^{(2)}_{k+1}$, which is finite almost surely. On $\{\tau(B) = \infty\}$, we can write

$$\begin{aligned}
&\mathbf{1}_{\{\tau(B)=\infty\}} \sum_{k=0}^{n} \gamma'_{a_k} \mathbf{1}_{\{\tau(B)>k\}} U^{(2)}_{k+1} \\
&\quad = \gamma'_{a_0} P_{\theta_0} h_{\theta_0}(X_0, I_0) - \gamma'_{a_n} \mathbf{1}_{\{\tau(B)=\infty\}} P_{\theta_{n+1}} h_{\theta_{n+1}}(X_{n+1}, I_{n+1}) \\
&\qquad + \sum_{k=0}^{n-1} \Big(\gamma'_{a_{k+1}} - \gamma'_{a_k} \Big) \mathbf{1}_{\{\tau(B)=\infty\}} P_{\theta_{k+1}} h_{\theta_{k+1}}(X_{k+1}, I_{k+1}).
\end{aligned}$$

Now $\mathrm{E}\left[(\gamma'_{a_n}\mathbf{1}_{\{\tau(B)=\infty\}}P_{\theta_{n+1}}h_{\theta_{n+1}}(X_{n+1},I_{n+1}))^2\right] \le C_\varepsilon(\gamma'_{\lfloor n/c_0\rfloor})^2$ and $\sum(\gamma'_{\lfloor n/c_0\rfloor})^2 < \infty$, thus $\gamma'_{a_n}\mathbf{1}_{\{\tau(B)=\infty\}}P_{\theta_{n+1}}h_{\theta_{n+1}}(X_{n+1},I_{n+1})$ converges a.s. to 0, and

$$\begin{aligned}&\sum_{k=0}^{n-1}\left|(\gamma'_{a_{k+1}}-\gamma'_{a_k})\mathbf{1}_{\{\tau(B)=\infty\}}P_{\theta_{k+1}}h_{\theta_{k+1}}(X_{k+1},I_{k+1})\right|\\ &\quad\le C_\varepsilon\sum_{k=0}^{\infty}(\gamma'_{a_k}-\gamma'_{a_{k+1}})\mathbf{1}_{\{\tau(B)=\infty\}}V^{1/2}(X_{k+1},I_{k+1}).\end{aligned}$$

Since a_k only changes at the stopping times κ_i, we have

$$\begin{aligned}&\mathrm{E}\left[\sum_{k=0}^{\infty}(\gamma'_{a_k}-\gamma'_{a_{k+1}})\mathbf{1}_{\{\tau(B)=\infty\}}V^{1/2}(X_{k+1},I_{k+1})\right]\\ &\quad=\mathrm{E}\left[\sum_{k=1}^{\infty}(\gamma'_{k-1}-\gamma'_k)\mathbf{1}_{\{\tau(B)=\infty\}}V^{1/2}(X_{\kappa_k+1},I_{\kappa_k+1})\right]\\ &\quad=\sum_{k=1}^{\infty}(\gamma'_{k-1}-\gamma'_k)\mathrm{E}\left[\mathbf{1}_{\{\tau(B)=\infty\}}V^{1/2}(X_{\kappa_k+1},I_{\kappa_k+1})\right]\\ &\quad\le C_\varepsilon\sum_{k=1}^{\infty}\left(\gamma'_{k-1}-\gamma'_k\right)\\ &\quad\le C_\varepsilon\gamma'_0.\end{aligned}$$

By Lebesgue's Dominated Convergence Theorem, we can conclude that

$$\mathrm{E}\left[\left|\sum_{k=0}^{\infty}(\gamma'_{a_{k+1}}-\gamma'_{a_k})\mathbf{1}_{\{\tau(B)=\infty\}}P_{\theta_{k+1}}h_{\theta_{k+1}}(X_{k+1},I_{k+1})\right|\right]<\infty.$$

This is sufficient to conclude that $\sum_{k=0}^{\infty}\gamma'_{a_k}\mathbf{1}_{\{\tau(B)>k\}}U^{(2)}_{k+1}$ is finite almost surely.

Using (6.6);

$$\begin{aligned}\sum_{k=0}^{n}\gamma'_{a_k}\mathbf{1}_{\{\tau(B)>k\}}\left|U^{(3)}_{k+1}\right| &\le C_\varepsilon\sum_{k=0}^{\infty}\gamma'_{a_k}\gamma_{a_k}\mathbf{1}_{\{\tau(B)>k\}}V^{1/2}(X_{k+1},I_{k+1})\\ &\le C_\varepsilon\sum_{k=0}^{\infty}\gamma'_{\lfloor n/c_0\rfloor}\gamma_{\lfloor n/c_0\rfloor}\mathbf{1}_{\{\tau(B)>k\}}V^{1/2}(X_{k+1},I_{k+1}),\end{aligned}$$

and

$$E\left[\sum_{k=0}^{\infty}\gamma'_{\lfloor n/c_0\rfloor}\gamma_{\lfloor n/c_0\rfloor}\mathbf{1}_{\{\tau(B)>k\}}V^{1/2}(X_{k+1},I_{k+1})\right]\le C_\varepsilon\sum_{k=0}^{\infty}\gamma'_{\lfloor n/c_0\rfloor}\gamma_{\lfloor n/c_0\rfloor}<\infty$$

by (6.3). With Lebesgue's Dominated Convergence Theorem, we deduce that

$$\mathrm{E}\left[\sum_{k=0}^{\infty}\gamma'_{a_k}\mathbf{1}_{\{\tau(B)>k\}}\left|U_{k+1}^{(3)}\right|\right] \le C_\varepsilon\sum_{k=0}^{\infty}\gamma'_{\lfloor n/c_0\rfloor}\gamma_{\lfloor n/c_0\rfloor} < \infty,$$

which implies that $\sum_{k=0}^{n}\gamma'_{a_k}\mathbf{1}_{\{\tau(B)>k\}}U_{k+1}^{(3)}$ converges almost surely to a finite limit. This completes the proof of the lemma.

We are now in position to prove Theorem 4.1. We start with (i).

Proposition 6.1. *Under* (A1–3), *let* $B>0$ *be given. Then* $|\theta_n-\theta^*|\,\mathbf{1}_{\{\tau(B)>n\}} \to 0$ *with probability one as* $n\to\infty$.

Proof. The idea of the proof is borrowed from Delyon (1996). We saw in (6.7) that $\theta_{n+1}=\theta_n+\gamma_{a_n}H(\theta_n,I_{n+1})+\gamma_{a_n}^2 r_n$, where $H=(H_1,\ldots,H_d)$, $r_n=(r_{n,1},\ldots,r_{n,d})$, $H_i(\theta,I)=\theta(i)(\mathbf{1}_{\{I=i\}}-\theta(I))$ and $r_{i,n}=-\theta(i)\theta(I)[(\mathbf{1}_{\{I=i\}}-\theta(I))/(1+\gamma_{a_n}\theta(I))]$. We note that $|H|\le 1$ and $|r_n|\le 1$. Let $\varepsilon\in(0,\varepsilon^*)$ be such that $\tau(B)\le\xi(\varepsilon)$ (Lemma 6.5). We recall that the mean field function of the recursion is $h_i(\theta)=(\theta^*(i)-\theta(i))/\sum_{j=1}^{d}[(\theta^*(j))/(\theta(j))]$. We introduce $\theta'_n=\theta_n+\sum_{j=n}^{\infty}\gamma_{a_j}\mathbf{1}_{\{\tau(B)>j\}}[H(\theta_j,I_{j+1})-h(\theta_j)]$. From Lemma 6.6, $|\theta'_n-\theta_n|\to 0$ almost surely. $\{\theta'_n\}$ satisfies the recursion

$$\theta'_{n+1}=\theta'_n+\gamma_{a_n}h(\theta_n)+\gamma_{a_n}\mathbf{1}_{\{\tau(B)\le n\}}\Big[H(\theta_n,I_{n+1})-h(\theta_n)\Big]+\gamma_{a_n}^2 r_n.$$

We can then deduce that

$$\begin{aligned}
&\left|\theta'_{n+1}-\theta^*\right|^2\mathbf{1}_{\{\tau(B)>n\}}\\
&\quad=\left|\theta'_n-\theta^*\right|^2\mathbf{1}_{\{\tau(B)>n\}}+2\gamma_{a_n}\left\langle\theta'_n-\theta^*,h(\theta_n)\right\rangle\mathbf{1}_{\{\tau(B)>n\}}+\gamma_{a_n}\mathbf{1}_{\{\tau(B)>n\}}r'_n\\
&\quad\le(1-2\varepsilon\gamma_{a_n})\left|\theta'_n-\theta^*\right|^2\mathbf{1}_{\{\tau(B)>n\}}+\gamma_{a_n}\mathbf{1}_{\{\tau(B)>n\}}\Big(r'_n+\left\langle\theta'_n-\theta^*,h(\theta_n)-h(\theta'_n)\right\rangle\Big),
\end{aligned}$$

where $r'_n\to 0$ almost surely as $n\to\infty$. Since θ_n remains in the compact Θ_ε, h is continuous and since $|\theta'_n-\theta_n|\to 0$, it follows that $\langle\theta'_n-\theta^*,h(\theta_n)-h(\theta'_n)\rangle\to 0$. We can summarize the situation like this. Writing $U_n=|\theta'_n-\theta^*|^2\mathbf{1}_{\{\tau(B)>n\}}$, we have

$$U_{n+1}\le(1-2\varepsilon\gamma_{a_n})U_n+\gamma_{a_n}r''_n, \tag{6.10}$$

where $r''_n\to 0$ as $n\to\infty$. This implies that $U_n\to 0$ which, given Lemma 6.6, proves the Proposition.

To see why $U_n\to 0$, let $\delta>0$ be given. Take $n_0>0$ such that for $n\ge n_0$, $\left|r''_n\right|\le 2\varepsilon\delta$ and $(1-2\varepsilon\gamma_{a_n})\mathbf{1}_{\{\tau(B)>n\}}>0$. Then for $n\ge n_0$, $(U_{n+1}-\delta)\le(1-2\varepsilon\gamma_{a_n})(U_n-\delta)+2\varepsilon\gamma_{a_n}(r''_n/2\varepsilon-\delta)\le(1-2\varepsilon\gamma_{a_n})(U_n-\delta)$, which implies that $\limsup(U_n-\delta)\le 0$ and, since $\delta>0$ is arbitrary, we conclude that $\lim U_n=0$.

Proposition 6.2. *Under* (A1−3), *let* $B > 0$ *be given. For any function* $f \in L_{V^{1/2}}$, *denoting* $\bar{f} = f - \pi^*(f)$, *we have:*

$$\frac{1}{n}\sum_{k=1}^{n} \bar{f}(X_k, I_k)\mathbf{1}_{\{\tau(B)>k-1\}} \to 0, \quad a.s. \text{ as } n \to \infty. \tag{6.11}$$

Proof. In view of Proposition 6.1 and (6.5), we only need show that

$$\frac{1}{n}\sum_{k=1}^{n} \Big(f(X_k, I_k) - \pi_{\theta_{k-1}}(f)\Big)\mathbf{1}_{\{\tau(B)>k-1\}} \to 0 \quad a.s.. \tag{6.12}$$

Kronecker's Lemma applied to (6.9) of Lemma 6.6 with $\gamma'_n = 1/n$, $H_\theta = f$ yields (6.12).

6.3. Proof of Theorem 4.2

Proof. Assume that **[a]** hold. Define $\alpha_k = (1+\gamma_0)^{(1+n_0)k}$ and suppose that $\limsup_{n\to\infty} n(v_n(i) - v_n(j)) = \infty$. This implies the existence of an increasing sequence of integers $\{n_k,\ k \geq 1\}$ such that $n_k(v_{n_k}(i) - v_{n_k}(j)) > \alpha_k$ and $(X_{n_k}, I_{n_k}) \in \mathcal{X}_i \times \{i\}$, but $(X_{n_k+n_0}, I_{n_k+n_0}) \notin \mathcal{X}_j \times \{j\}$ for all $k \geq 1$. Clearly, $n_k(v_{n_k}(i) - v_{n_k}(j)) > \alpha_k$ implies that $\theta_{n_k}(i)/\theta_{n_k}(j)$ converges to $+\infty$. But, since i leads to j, we can then find $\varepsilon > 0$ and k_0 such that, for $k \geq k_0$,

$$\begin{aligned}\Pr\Big[&(X_{n_k+n_0}, I_{n_k+n_0}) \notin \mathcal{X}_j \times \{j\} | \mathcal{F}_{n_k}, (X_{n_k}, I_{n_k}) \in \mathcal{X}_i \times \{i\},\\ &n_k(v_{n_k}(i) - v_{n_k}(j)) > \alpha_k\Big] \leq (1-\varepsilon).\end{aligned}$$

Thus $\Pr(\limsup_{n\to\infty} n(v_n(i) - v_n(j)) = \infty) \leq \lim_{k\to\infty}(1-\varepsilon)^k = 0$.

Assume that **[b]** hold. Define $\alpha_k = (1+\gamma_0)^{2k}$ and suppose that $\limsup_{n\to\infty} n(\max_i v_n(i) - \min_j v_n(j)) = \infty$. Then we can find $i_0 \in \{1, \ldots, d\}$ and an increasing sequence of integers $\{n_k,\ k \geq 1\}$ such that $\min_j v_n(j) = v_n(i_0)$, $n_k(\max_j v_{n_k}(j) - v_{n_k}(i_0)) > \alpha_k$, $(X_{n_k}, I_{n_k}) \notin \mathcal{X}_{i_0} \times \{i_0\}$ and $(X_{n_k+1}, I_{n_k+1}) \notin \mathcal{X}_{i_0} \times \{i_0\}$ for all $k \geq 1$. Then we can proceed as above and conclude.

Acknowledgements

The authors are grateful to Christophe Andrieu, David P. Landau and Eric Moulines for very helpful discussions. We also thank David P. Sanders and the referees for comments which helped to improve the paper. This work is partly supported by the National Science Foundation grant DMS 0244638 and by a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada.

References

Andrieu, C. and Atchade, Y. F. (2007). On the efficiency of adaptive MCMC algorithms. *Electron. Comm. Probab.* **12**, 336-349.

Andrieu, C. and Robert, C. P. (2001). Controlled MCMC for optimal sampling. Technical report, Université Paris Dauphine, Ceremade 0125.

Andrieu, C. and Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.* **16**, 1462-1505.

Andrieu, C., Moulines, É. and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44**, 283-312 (electronic).

Atchadé, Y. F. and Liu, J. S. (2006). Discussion of "Equi-energy sampler" by Kou, Zhou and Wong. *Ann. Statist.* **34**, 1620-1628.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov Chain Monte Carlo algorithm. *Bernoulli* **11**, 815-828.

Baxendale, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.* **15**, 700-738.

Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic approximations*. Applications of Mathematics. Springer, Paris-New York,

Berg, B. A. and Neuhaus, T. (1992). Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.* **68**,

Brockwell, A. E. and Kadane, J. B. (2005). Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *J. Comput. Graph. Statist.* **14**, 436-458.

Delyon, B. (1996). General results on the convergence of stochastic algorithms. *IEEE Trans. Automat. Control* **41**, 1245-1255.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov Chain Monte Carlo with applications to pedigree analysis. *J. Amer. Statist. Assoc.* **90**, 909-920.

Gilks, W. R., Roberts, G. O. and Sahu, S. K. (1998). Adaptive Markov Chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.* **93**, 1045-1054.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**, 223-242.

Liang, F. and Wong, W. H. (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Amer. Statist. Assoc.* **96**, 653-666.

Liang, F., Liu, C. and Carroll, R. J. (2007). Stochastic approximation in Monte Carlo computation. *J. Amer. Statist. Assoc.* **102**, 305-320.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo schemes. *Europhys. Lett.* **19**, 451-458.

Mira, A. and Sargent, D. J. (2003). A new strategy for speeding Markov Chain Monte Carlo algorithms. *Stat. Methods Appl.* **12**, 49-60.

Rosenthal, J. S. and Roberts, G. O. (2007). Coupling and ergodicity of adaptive MCMC. *J. Appl. Probab.* **44**, 458-475.

Wang, F. and Landau, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86**, 2050-2053.

Department of Statistics, University of Michigan, Anu Arbor, MI48109, U.S.A.
E-mail: yvesa@umich.edu

Department of Statistics, Harvard University, Cambridge, MA02138, U.S.A.
E-mail: jliu@stat.harvard.edu

(Received June 2007; accepted December 2008)