Adaptive Equi-Energy Sampler : Convergence and Illustration

Amandine Schreck, Gersende Fort, Eric Moulines

Adaptive Equi-Energy Sampler: Convergence and Illustration

Amandine SCHRECK, Gersende FORT and Eric MOULINES

February 4, 2013

Abstract

Markov chain Monte Carlo (MCMC) methods allow to sample a distribution known up to a multiplicative constant. Classical MCMC samplers are known to have very poor mixing properties when sampling multimodal distributions. The Equi-Energy sampler is an interacting MCMC sampler proposed by Kou, Zhou and Wong in 2006 to sample difficult multimodal distributions. This algorithm runs several chains at different temperatures in parallel, and allow lower-tempered chains to jump to a state from a higher-tempered chain having an energy 'close' to that of the current state. A major drawback of this algorithm is that it depends on many design parameters and thus, requires a significant effort to tune these parameters.

In this paper, we introduce an Adaptive Equi-Energy (AEE) sampler which automates the choice of the selection mecanism when jumping onto a state of the higher-temperature chain. We prove the ergodicity and a strong law of large numbers for AEE, and for the original Equi-Energy sampler as well. Finally, we apply our algorithm to motif sampling in DNA sequences.

Keywords: interacting Markov chain Monte Carlo, adaptive sampler, equi-energy sampler, ergodicity, law of large numbers, motif sampling.

Author's email addresses: first-name.last-name@telecom-paristech.fr

This work is partially supported by the French National Research Agency, under the program ANR-08-BLAN-0218 BigMC.

1 Introduction

Markov Chain Monte Carlo (MCMC) methods are well-known tools for sampling a target distribution π known up to a multiplicative constant. MCMC algorithms sample π by constructing a Markov chain admitting π as unique invariant distribution. A canonical example is the the Metropolis-Hastings algorithm [27, 20]: given the current value X_n of the chain $\{X_j,\, j \geq 0\}$, it consists in proposing a move Y_{n+1} under a proposal distribution $Q(X_n, \cdot)$. This move is then accepted with probability

$$\alpha_n = 1 \wedge \pi(Y_{n+1})Q(Y_{n+1}, X_n)/[\pi(X_n)Q(X_n, Y_{n+1})] \,,$$

where $a \wedge b$ stands for $\min(a, b)$; otherwise, $X_{n+1} = X_n$.

It is known that the efficiency of MCMC methods depends upon the choice of the proposal distribution [31]. For example, when sampling multi-modal distributions, a Metropolis-Hastings algorithm with $Q(X_n, \cdot)$ equal to a Gaussian distribution centered in X_n tends to be stuck in one of the modes. So the convergence of such an algorithm will be slow, and the target distribution will not be correctly approximated unless a huge number of points is sampled.

Efficient implementations of MCMC rely on a strong expertise of the user in order to choose a proposal kernel and, more generally, design parameters adapted to the target π.

This is the reason why *adaptive* and *interacting* MCMC methods have been introduced. Adaptive MCMC methods consist in choosing, at each iteration, a transition kernel P_θ among a family $\{P_\theta, \theta \in \Theta\}$ of kernels with invariant distribution π: the conditional distribution of X_{n+1} given the past is $P_{\theta_n}(X_n, \cdot)$ where the parameter θ_n is chosen according to the past values of the chain $\{X_n, n \geq 0\}$. From the pioneering Adaptive Metropolis algorithm of [19], many adaptive MCMC have been proposed and successfully applied (see the survey papers by [5], [31], [6] for example).

Interacting MCMC methods rely on the (parallel) construction of a family of processes with distinct stationary distributions; the key behind these techniques is to allow interactions when sampling these different processes. At least one of these processes has π as stationary distribution. The stationary distributions of the auxiliary processes are chosen in such a way that they have nice convergence properties, hoping that the process under study will inherit them. For example, in order to sample multi-modal distributions, a solution is to draw auxiliary processes with target distributions equal - up to the normalizing constant - to tempered versions π^{1/T_i}, $T_i > 1$. This solution is the basis of the parallel *tempering algorithm* [18], where the states of two parallel chains are allowed to swap. Following this tempering idea, different interacting MCMC algorithms have been proposed and studied so far [1, 11, 14, 13].

The *Equi-Energy sampler* of Kou, Zhou and Wong [22] is an example of such interacting MCMC algorithms. K processes are sampled in parallel, with target distributions (proportional to) π^{β_k}, $1 = \beta_K > \beta_{K-1} > \cdots > \beta_1$. The first chain $Y^{(1)} = \{Y_n^{(1)}, n \geq 0\}$ is usually a Markov chain; then $Y^{(k)}$ is built from $Y^{(k-1)}$ as follows: with a fixed probability ε, the current state $Y_n^{(k)}$ is allowed to jump onto a past state of the auxiliary chain $\{Y_\ell^{(k-1)}, \ell \leq n\}$, and with probability $(1-\varepsilon)$, $Y_n^{(k)}$ is obtained using a "local" MCMC move (such as a random walk Metropolis step or a Metropolis-adjusted Langevin step). This mechanism includes the computation of an acceptance ratio so that the chain $Y^{(k)}$ will have π^{β_k} as target density. As the acceptance probability of such a jump could be very low, only jumps toward selected past values of $Y^{(k-1)}$, namely those with an *energy* close to

that of the current state $Y_n^{(k)}$, are allowed. This selection step allows higher acceptance rates of the jump, and a faster convergence of the algorithm is expected.

The Equi-Energy sampler has many design parameters: the interacting probability ε, the number K of parallel chains, the temperatures $T_k = 1/\beta_k$, $k \in \{1, \ldots, K\}$ and the selection function. It is known that all of these design parameters play a role on the efficiency of the algorithm. [22] suggest some values for all these parameters, designed for practical implementation and based on empirical results on some simple models. [3] discuss the choice of the interacting probability ε in similar contexts; [8] discuss the choice of the temperatures T_k of the chains for the Parallel Tempering algorithm. Recently, an algorithm combining parallel tempering with equi-energy moves have been proposed by [10].

In this paper, we discuss the choice of the energy rings and the selection function, when the jump probability ε, the number K of auxiliary processes and the temperatures are fixed. We introduce a new algorithm, called *Adaptive Equi-Energy sampler* in which the selection function is defined adaptively based on the past history of the sampler. We also address the convergence properties of this new sampler.

Different kinds of convergence of adaptive MCMC methods have been addressed in the literature: convergence of the marginals, the law of large numbers (LLN) and central limit theorems (CLT) for additive functionals (see e.g. [29] for convergence of the marginals and weak LLN of general adaptive MCMC, [4] or [34] for LLN and CLT for adaptive Metropolis algorithms, [16] and [17] for convergence of the marginals, LLN and CLT for general adaptive MCMC algorithms - see also the survey paper by [6]).

There are quite few analysis of the convergence of interacting MCMC samplers. The original proof of the convergence of the Equi-Energy sampler in [22] (resp. [7]) contains a serious gap, mentioned in [7] (resp. [2]). [3] established a strong LLN of a simplified version of the Equi-Energy sampler, in which the number of levels is set to $K = 2$ and the proposal during the interaction step are drawn uniformly at random in the past of the auxiliary process. Finally, Fort, Moulines and Priouret [16] established the convergence of the marginals and a strong LLN for the same simplified version of the Equi-Energy sampler (with no selection) but have removed the limitations on the number of parallel chains.

The paper addresses the convergence of an interacting MCMC sampler in which the proposal are selected from energy rings which are constructed adaptively at each levels. In this paper, we obtain the convergence of the marginals and a strong LLN of a smooth version of the Equi-Energy sampler and its adaptive variant. We illustrate our results in several difficult scenarios such as sampling mixture models with “well-separated" modes and motif sampling in biological sequences. The paper is organized as follows: in Section 2, we derive our algorithm and set the notations that are used throughout the paper. The convergence results are presented in Section 3. Finally, Section 4 is devoted to the application to motif sampling in biological sequences. The proofs of the results are postponed to the Appendix.

2 Presentation of the algorithm

2.1 Notations

Let $(\mathbf{X}, \mathcal{X})$ be a measurable Polish state space and P be a Markov transition kernel on $(\mathbf{X}, \mathcal{X})$. P operates on bounded functions f on $\mathbf{X}$ and on finite positive measures μ on $\mathcal{X}$:

$$Pf(x) = \int P(x, \mathrm{d}y) f(y), \qquad \mu P(A) = \int \mu(dx) P(x, A) \ .$$

The n-iterated transition kernel P^n, $n \geq 0$ is defined by:

$$P^n(x, A) = \int P^{n-1}(x, \mathrm{d}y) P(y, A) = \int P(x, \mathrm{d}y) P^{n-1}(y, A) \ ;$$

by convention, $P^0(x, A)$ is the identity kernel. For a function $V : \mathbf{X} \to [1, +\infty[$, we denote by $|f|_V$ the V-norm of a function $f : \mathbf{X} \to \mathbb{R}$:

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

If $V = 1$, this norm is the usual uniform norm. Let $\mathcal{L}_V = \{f : \mathbf{X} \to \mathbb{R}, |f|_V < +\infty\}$. We also define the V-distance between two probability measures μ_1 and μ_2 by:

$$\|\mu_1 - \mu_2\|_V = \sup_{f, |f|_V \leq 1} |\mu_1(f) - \mu_2(f)| \ .$$

When $V = 1$, the V-distance is the total-variation distance and will be denoted by $\|\mu_1 - \mu_2\|_{TV}$.

Let $(\Theta, \mathcal{T})$ be a measurable space, and $\{P_\theta, \theta \in \Theta\}$ be a family of Markov transition kernels; Θ can be finite or infinite dimensional. It is assumed that for all $A \in \mathcal{X}$, $(x, \theta) \to P_\theta(x, A)$ is $(\mathcal{X} \otimes \mathcal{T} | \mathcal{B}([0, 1]))$-measurable, where $\mathcal{B}([0, 1])$ denotes the Borel σ-field on $[0, 1]$.

2.2 The Equi-Energy sampler

Let π be the probability density of the target distribution with respect to a dominating measure μ on $(\mathbf{X}, \mathcal{X})$. In many applications, π is known up to a multiplicative constant; therefore, we will denote by π_u the (unnormalized) density.

We denote by P the Metropolis-Hastings kernel with proposal density kernel q and invariant distribution π defined by:

$$P(x, A) = \int_A r(x, y) q(x, y) \mu(\mathrm{d}y) + \mathbf{1}_A(x) \int (1 - r(x, y)) q(x, y) \mu(\mathrm{d}y) \ ,$$

where $(x, y) \mapsto r(x, y)$ is the acceptance ratio given by

$$r(x, y) = 1 \wedge \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)} \ .$$

The Equi-Energy (EE) sampler proposed by [22] exploits the fact that it is often easier to sample a tempered version π^β, $0 < \beta < 1$, of the target distribution than π itself. This is why the algorithm

relies on an auxiliary process $\{Y_n, n \geq 0\}$, run independently from $\{X_n\}$ and admitting π^β as stationary distribution (up to a normalizing constant). This mechanism can be repeated yielding to a multi-stages Equi-Energy sampler.

We denote by K the number of processes run in parallel. Let $\varepsilon \in (0,1)$. Choose K temperatures $T_1 > \cdots > T_K = 1$ and set $\beta_k = 1/T_k$; and K MCMC kernels $\{P^{(k)}, 1 \leq k \leq K\}$ such that $\pi^{\beta_k} P^{(k)} = \pi^{\beta_k}$. K processes $Y^{(k)} = \{Y_n^{(k)}, n \geq 0\}$, $1 \leq k \leq K$, are defined by induction on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The first auxiliary process $Y^{(1)}$ is a Markov chain, with $P^{(1)}$ as transition kernel. Given the auxiliary process $Y^{(k-1)}$ up to time n, $\{Y_m^{(k-1)}, m \leq n\}$, and the current state $Y_n^{(k)}$ of the process of level k, the Equi-Energy sampler draws $Y_{n+1}^{(k)}$ as follows:

- (Metropolis-Hastings step) with probability $1-\varepsilon$, $Y_{n+1}^{(k)} \sim P^{(k)}(Y_n^{(k)}, \cdot)$.
- (equi-energy step) with probability ε, the algorithm selects a state Z_{n+1} from the auxiliary process having an energy close to that of the current state. An acceptance-rejection ratio is then computed and if accepted, $Y_{n+1}^{(k)} = Z_{n+1}$; otherwise, $Y_{n+1}^{(k)} = Y_n^{(k)}$.

In practice, [22] only apply the equi-energy step when there is at least one point in each ring. In [22], the distance between the energy of two states is defined as follows. Consider an increasing sequence of positive real numbers

$$\xi_0 = 0 < \xi_1 < \cdots < \xi_S = +\infty \, . \tag{1}$$

If the energies of two states x and y belong to the same energy ring, i.e. if there exists $1 \leq \ell \leq S$ such that $\xi_{\ell-1} \leq \pi_u(x), \pi_u(y) < \xi_\ell$, then the two states are said to have "close energy". The choice of the energy rings is most often a difficult task. As shown in Figure 3[right], the Equi-Energy sampler is inefficient when the energy rings are not appropriately defined. The efficiency of the sampler is increased when the variation of π_u in each ring is small enough so that the equi-energy move is accepted with high probability.

2.3 The Adaptive Equi-Energy sampler

We propose to modify the Equi-Energy sampler by adapting the energy rings "on the fly", based on the history of the algorithm. Our new algorithm, so called *Adaptive Equi-Energy* sampler (AEE) is similar to the Equi-Energy sampler of [22] except for the equi-energy step, which relies on adaptive boundaries of the rings. For the definition of the process $Y^{(k)}$, $k \geq 2$, adaptive boundaries computed from the process $Y^{(k-1)}$ are used.

For a distribution θ in Θ, denote by $\xi_{\theta,\ell}$, $\ell \in \{1, \cdots, S-1\}$ the bounds of the rings, computed from r.v. with distribution θ; by convention, $\xi_{\theta,0} = 0 \leq \xi_{\theta,1} \leq \cdots \leq \xi_{\theta,S-1} \leq \xi_{\theta,S} = +\infty$. Define the associated energy rings $H_{\theta,\ell} = [\xi_{\theta,\ell-1}, \xi_{\theta,\ell})$ for $\ell \in \{1, \cdots, S\}$. We consider selection functions $g_\theta(x,y)$ of the form

$$g_\theta(x,y) = \sum_{\ell=1}^{S} h_{\theta,\ell}(x) h_{\theta,\ell}(y) \, , \quad h_{\theta,\ell}(x) = (1 - d(\pi_u(x), H_{\theta,\ell}))_+ \, , \tag{2}$$

where $d(\pi_u(x), H_{\theta,\ell})$ measures the distance between $\pi_u(x)$ and the ring $H_{\theta,\ell}$. By convention $h_{\theta,\ell} = 0$ if $H_{\theta,\ell} = \emptyset$. We finally introduce a set of selection kernels $\{K_\theta^{(k)}, \theta \in \Theta\}$ for all $k \in \{2, \cdots, K\}$

defined by

$$K_\theta^{(k)}(x,A) = \int_A \alpha_\theta^{(k)}(x,y)\frac{g_\theta(x,y)\theta(\mathrm{d}y)}{\int g_\theta(x,z)\theta(\mathrm{d}z)} + \mathbf{1}_A(x)\int\{1-\alpha_\theta^{(k)}(x,y)\}\frac{g_\theta(x,y)\theta(\mathrm{d}y)}{\int g_\theta(x,z)\theta(\mathrm{d}z)}\ , \tag{3}$$

where

$$\alpha_\theta^{(k)}(x,y) = 1 \wedge \left(\frac{\pi^{\beta_k-\beta_{k-1}}(y)\int g_\theta(x,z)\theta(\mathrm{d}z)}{\pi^{\beta_k-\beta_{k-1}}(x)\int g_\theta(y,z)\theta(\mathrm{d}z)}\right)\ . \tag{4}$$

$K_\theta^{(k)}$ is associated to the equi-energy step when defining $Y^{(k)}$: a draw under the selection kernel proportional to $g_\theta(x,y)\theta(\mathrm{d}y)$ is combined with an acceptance-rejection step. The acceptance-rejection step is defined so that when $\theta \propto \pi^{\beta_{k-1}}$, π^{β_k} is invariant for $K_\theta^{(k)}$ [22].

This equi-energy step is only allowed when each ring contains at least one point (of the auxiliary process $Y^{(k-1)}$ up to time n). We therefore introduce, for all positive integer m, the set Θ_m:

$$\Theta_m \stackrel{\mathrm{def}}{=} \left\{\theta\in\Theta : \frac{1}{m} \le \inf_x \int g_\theta(x,y)\theta(\mathrm{d}y)\right\}\ . \tag{5}$$

With these notations, AEE satisfies for any $n \ge 0$ and $k \in \{1,\cdots,K\}$,

$$\mathbb{E}[f(Y_{n+1}^{(k)})|\mathcal{F}_n^{(k)}] = \mathbb{E}[f(Y_{n+1}^{(k)})|Y_n^{(k)}, Y_m^{(k-1)}, 1\le m\le n] = P^{(k)}_{\theta_n^{(k-1)}} f(Y_n^{(k)})\ , \tag{6}$$

where $\{\mathcal{F}_n^{(k)}, n\ge 0\}$ is the filtration defined by $\mathcal{F}_n^{(k)} = \sigma\left(\left\{Y_m^{(l)}, 1\le m\le n, 1\le l\le k\right\}\right)$; the transition kernel is given by $P_\theta^{(1)} = P^{(1)}$ and for $k\ge 2$,

$$P_\theta^{(k)} = (1-\varepsilon\mathbf{1}_{\theta\in\bigcup_{m\ge1}\Theta_m})P^{(k)} + \varepsilon\mathbf{1}_{\theta\in\bigcup_{m\ge1}\Theta_m}\ K_\theta^{(k)}\ ;$$

and $\theta_n^{(k)}$ is the empirical distribution

$$\theta_n^{(k)} = \frac{1}{n}\sum_{m=1}^n \delta_{Y_m^{(k)}}\ , \qquad k\in\{1,\cdots,K\}, n\ge 1\ . \tag{7}$$

Different functions d can be chosen. For example, the function given by

$$d(\pi_u(x), H_{\theta,\ell}) = \mathbf{1}_{\mathbb{R}\setminus H_{\theta,\ell}}(\pi_u(x)) = \left\{\begin{array}{ll} 0 & \text{if } \pi_u(x)\in H_{\theta,\ell},\\ 1 & \text{otherwise}\end{array}\right. \tag{8}$$

yields to a selection function g_θ such that $g_\theta(x,y) = 1$ iff x,y are in the same energy ring and $g_\theta(x,y) = 0$ otherwise. In this case, the acceptance-rejection ratio $\alpha_\theta^{(k)}(x,y)$ is equal to $1\wedge(\pi^{\beta_k-\beta_{k-1}}(y)/\pi^{\beta_k-\beta_{k-1}}(x))$ upon noting that by definition of the proposal kernel, the points x and y are in the same energy ring. By using this "hard" distance during the equi-energy jump, all the states of the auxiliary process having their energy in the same ring as the energy of the current state are chosen with the same probability, while the other auxiliary states have no chance to be selected.

Other functions d could be chosen, such as "soft" selections of the form

$$d(\pi_u(x), H_{\theta,l}) = \frac{1}{r}\min_{y\in H_{\theta,\ell}} |\pi_u(x) - y|\ , \tag{9}$$

where $r > 0$ is fixed. With this "soft" distance, given a current state $Y_n^{(k)}$, the probability for each auxiliary state $Y_i^{(k-1)}$, $i \leq n$, to be chosen is proportional to $g_{\theta_n^{(k-1)}}(Y_n^{(k)}, Y_i^{(k-1)})$. Then, the "soft" selection function allows auxiliary states having an energy in a r-neighborhood of the energy ring of $\pi_u(Y_n^{(k)})$ to be chosen, as well as states having their energy in this ring. Nevertheless, this selection function yields an acceptance-rejection ratio $\alpha_\theta^{(k)}$ which may reveal to be quite costly to evaluate.

The asymptotic behavior of AEE will be addressed in Section 3. The intuition is that when the empirical distribution $\theta_n^{(k-1)}$ of the auxiliary process of order $k-1$ converges (in some sense) to $\theta_\star^{(k-1)}$, the process $\{Y_n^{(k)}, n \geq 0\}$ will behave (in some sense) as a Markov chain with transition kernel $P^{(k)}_{\theta_\star^{(k-1)}}$.

2.4 A toy example (I)

To highlight the interest of our algorithm, we consider toy examples: the target density π is a mixture of $\mathbb{R}^d$-valued Gaussian[1] . This model is known to be difficult, as illustrated (for example) in [6] for a random walk Metropolis-Hastings sampler (SRWM), an EE-sampler and a parallel tempering algorithm. Indeed, if the modes are well separated, a Metropolis-Hastings algorithm using only "local moves" is likely to remain trapped in one of the modes for a long-period of time. In the following, AEE is implemented with ring boundaries computed as described in Section 3.3.

Figure 1.(a) displays the target density π and the simulated one for three different algorithms (SRWM, EE and AEE) in one dimension. The histograms are obtained with 10^5 samples; for EE and AEE, the probability of interaction is $\varepsilon = 0.1$, the number of parallel chains is equal to $K = 5$ and the number of rings is $S = 5$. For the adaptive definition of the rings in AEE, we choose the "hard" selection (8) and the construction of the rings defined in Section 3.3. In the same vein, Figure 2 displays the points obtained by the three algorithms when sampling a mixture of two Gaussian distributions in two dimensions. As expected, in both figures, SRWM never explores one of the modes, while EE and AEE are far more efficient.

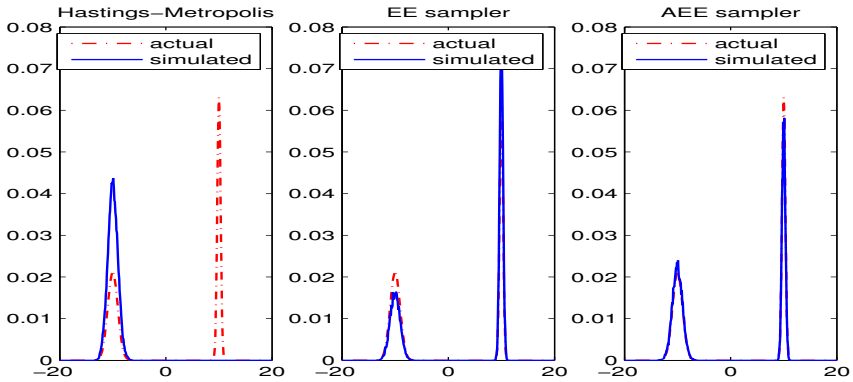


Figure 1: Comparison of SRWM (left), EE (center) and AEE (right) for a Gaussian mixture in one dimension

[1]MATLAB codes for AEE are available at the address http://perso.telecom-paristech.fr/∼schreck/index.html

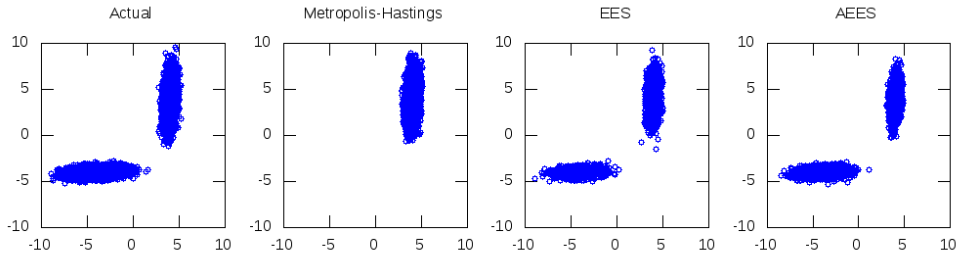


Figure 2: Comparison of the algorithms for a Gaussian mixture in two dimensions: (from left to right) the true density, SRWM, EE and AEE.

To compare EE and AEE in a more challenging situation, we consider the case of a mixture with two components in ten dimensions. We run EE and AEE with $K = 3$ parallel chains with respective temperatures $T_1 = 1, T_2 = 9, T_3 = 60$, the probability of jump ε is equal to 0.1, and the number of rings is $S = 50$. Both algorithms are initialized in one of the two modes of the distribution. For the Metropolis-Hastings step, we use a Symmetric Random Walk with Gaussian proposal; the covariance matrix of the proposal is of the form $c\ I$ where c is calibrated so that the mean acceptance rate is approximatively 0.25. Figure 3 displays, for each algorithm, the L^1-norm of the empirical mean, averaged over 10 independent trajectories, as a function of the length of the chains.

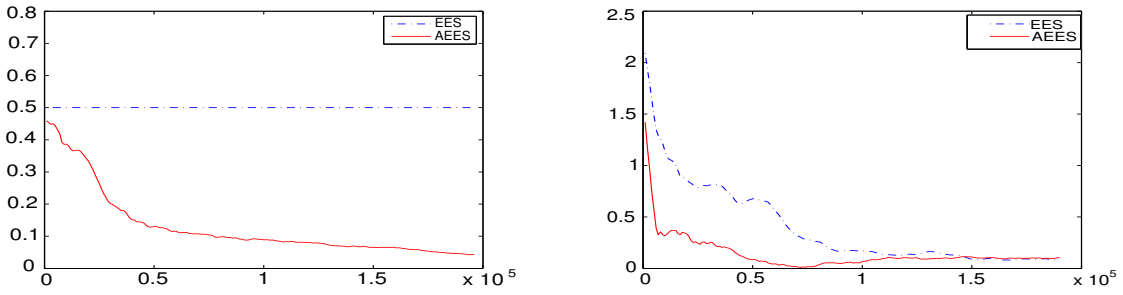
In order to show that the efficiency of EE depends crucially upon the choice of the rings, we choose a set of boundaries so that in practice, along one run of the algorithm, some of the rings are never reached. Figure 3(a) compares EE and AEE in this extreme case: even after 2×10^5 iterations, all of the equi-energy jumps are rejected for the (non-adaptive) EE, and the algorithm is trapped in one of the modes. This does not occur for AEE, and the L^1-error tends to zero as the number of iterations increases. This illustrates that our adaptive algorithm avoids the poor behaviors that EE can have when the choice of its design parameters is inappropriate.

We now run EE in a less extreme situation: we choose (fixed) energy rings so that the sampler can jump more easily than in the previous experiment between the modes. Figure 3(b) illustrates that the adaptive choice of the energy rings speeds up the convergence, as it makes the equi-energy jumps be more often accepted. To have a numerical comparison, the equi-energy jumps were accepted about ten times more often for AEE than for EE.

2.5 Toy example (II)

For a better understanding on how our algorithm behaves, Figure 4.(a) displays the evolution of the ring bounds used in the definition of $Y^{(K)}$. In this numerical application, the target density is a mixture of two Gaussian distributions in one dimension; EE and AEE are run with $K = 5$ chains, $S = 5$ rings and $\varepsilon = 0.1$, for a number of iterations varying from 0 to 10^5. As expected, the ring bounds become stable after a reasonable number of iterations. Moreover, we observed that the (non-adaptive) EE run with the rings fixed to the limiting values obtained with AEE behaves remarkably well.

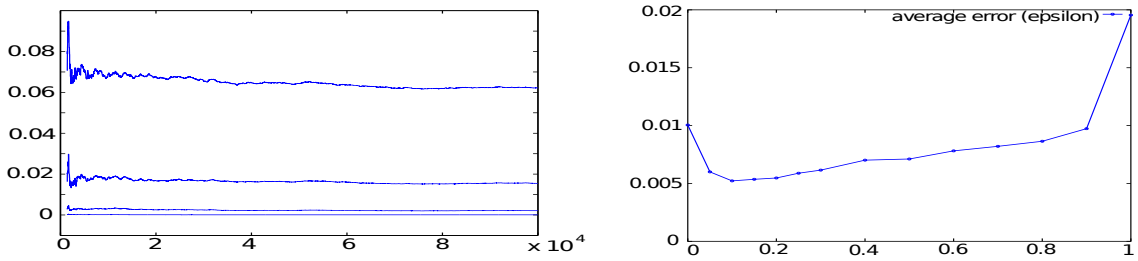
Finally, to have an idea on the role played by ε, Figure 4.(b) displays the average L^1 error of



(a) (b)

Figure 3: Error of EE (dashed line) and AEE for two different target densities in ten dimensions.

AEE for a mixture of two Gaussian distributions in one dimension, after 2×10^5 iterations and for 100 independent trajectories when ε is varying from 0 to 1. If ε is too small, AEE is not mixing well enough, and if ε is too large, the algorithm jumps easily from one mode to another but does not explore well enough each mode, which explains the 'u' shape of the curve. This experiment suggests that there exists an optimal value for ε, but to our best knowledge, the optimal choice of this design parameter is an open problem.



(a) (b)

Figure 4: (a): Evolution of the ring bounds; (b): Averaged error of AEE as a function of ε.

3 Convergence of the Adaptive Equi-Energy sampler

In this section, the convergence of the K-stages Adaptive Equi-Energy sampler is established. In order to make the proof easier, we consider the case when the distance function d in the definition of the selection function (2) is given by (9).

[16] provide sufficient conditions for the convergence of the marginals and the strong LLN (s-LLN) of interacting MCMC samplers. We use their results and show the convergence of the marginals i.e.

$$\lim_{n\to\infty} \mathbb{E}\left[f(Y_n^{(K)})\right] = \pi(f)\,,$$

for any continuous bounded functions f. Note that this implies that this limit holds for any indicator function $f = \mathbf{1}_A$ such that $\mathbb{P}(\partial A) = 0$ where ∂A denotes the boundary of A [12, Theorem 2.1]. We then establish the s-LLN: for a wide class of continuous (un)bounded functions f,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} f(Y_m^{(K)}) = \pi(f) \ , \mathbb{P} - \text{a.s.}$$

3.1 Assumptions

Our results are established for target distributions π satisfying

E1 (a) π is the density of a probability distribution on the measurable Polish space $(\mathbf{X}, \mathcal{X})$ and $\sup_{\mathbf{X}} \pi < \infty$ and for any $s \in (0, 1]$, $\int \pi^s(x)\, \mathrm{d}x < \infty$.

(b) π is continuous and positive on $\mathbf{X}$.

Usually, the user knows π up to a normalizing constant: hereafter, π_u will denote this available (unnormalized) density.

As in [16], we first introduce a set of conditions that will imply the geometric ergodicity of the kernels $P_\theta^{(k)}$, and the existence of an invariant probability measure for $P_\theta^{(k)}$ (see conditions E2). We finally introduce conditions on the boundaries of the adaptive energy rings (see conditions E3). Examples of boundaries satisfying E3 and computed from quantile estimators are given in Section 3.3 (see also [35] for stochastic approximation-based adapted boundaries).

Convergence of adaptive and interacting MCMC samplers is addressed in the literature by assuming containment conditions and diminishing adaptations (so called after [29]). Assumptions E2 is the main tool to establish a (generalized) containment condition. In our algorithm, the adaptation mechanism is due to (a) the interaction with an auxiliary process and (b) the adaption of the rings. Therefore, assumptions E2 and E3 are related to the diminishing adaptation condition (see e.g. Lemma B.6 in Section B.3).

E2 For each $k \in \{1, \ldots, K\}$:

(a) $P^{(k)}$ is a ϕ-irreducible transition kernel which is Feller on $(\mathbf{X}, \mathcal{X})$ and such that $\pi^{\beta_k} P^{(k)} = \pi^{\beta_k}$.

(b) There exist $\lambda_k \in (0, 1)$, $b_k < +\infty$ and $\tau_k \in (0, \tau_{k-1}\beta_{k-1}/\beta_k)$ such that $P^{(k)} W_k \leq \lambda_k W_k + b_k$ with

$$W_k(x) = \left(\frac{\pi^{\beta_k}(x)}{\sup_{\mathbf{X}} \pi^{\beta_k}} \right)^{-\tau_k} ; \tag{10}$$

by convention, $\tau_0\beta_0 = \beta_1$.

(c) For all $p \in (0, \sup_{\mathbf{X}} \pi)$, the sets $\{\pi \geq p\}$ are 1-small for $P^{(k)}$.

Note that by definition of τ_k and E1a, $W_{k+1} \in \mathcal{L}_{W_k}$ and $\int W_k(x)\pi^{\beta_k}(x)\mathrm{d}x < \infty$.

E2 is satisfied for example if for each k, $P^{(k)}$ is a symmetric random walk Metropolis Hastings kernel; and π is a sub-exponential target density [30, 21].

In our algorithm, $Y^{(1)}$ is a Markov chain with transition kernel $P^{(1)}$. As discussed in [28][chapters 13 and 17], E2 is sufficient to prove ergodicity and a s-LLN for $Y^{(1)}$. E2 also implies uniform W_1-moments for $Y^{(1)}$. These results, which initializes our proof by recurrence of the convergence for the process number K, is given in Proposition 3.1. Define the probability distributions

$$\theta_\star^{(k)}(\mathrm{d}x) = \frac{\pi^{\beta_k}(x)}{\int \pi^{\beta_k}(z)\mu(\mathrm{d}z)}\mu(\mathrm{d}x)\ , \qquad k \in \{1, \cdots, K\}\ . \tag{11}$$

Proposition 3.1. *Assume E1a, E2 and* $\mathbb{E}\left[W_1(Y_0^{(1)})\right] < \infty$. *Then,*

(a) For all bounded measurable functions f, $\lim_{n\to\infty} \mathbb{E}\left[f(Y_n^{(1)})\right] = \theta_\star^{(1)}(f)$.

(b) $\theta_\star^{(1)}(W_2) < +\infty$, *and for any measurable function* f *in* $\mathcal{L}_{W_1}$, $\lim_{n\to\infty} \theta_n^{(1)}(f) = \theta_\star^{(1)}(f)$ *a.s.*

(c) $\sup_n \mathbb{E}\left[W_1\left(Y_n^{(1)}\right)\right] < \infty$.

E3 (a) For any $k \in \{1, \ldots, K-1\}$, $\inf_{\ell\in\{1,\cdots,S-1\}} \int h_{\theta_\star^{(k)},\ell}(y)\ \theta_\star^{(k)}(\mathrm{d}y) > 0$.

(b) For any $k \in \{1, \ldots, K-1\}$ and $\ell \in \{1, \cdots, S-1\}$, $\lim_{n\to\infty} \left|\xi_{\theta_n^{(k)},\ell} - \xi_{\theta_\star^{(k)},\ell}\right| = 0$ w.p.1

(c) There exists $\Gamma > 0$ such that for any $k \in \{1, \ldots, K-1\}$, any $\ell \in \{1, \cdots, S-1\}$, and any $\gamma \in (0, \Gamma)$, $\limsup_n\ n^\gamma \left|\xi_{\theta_{n+1}^{(k)},\ell} - \xi_{\theta_n^{(k)},\ell}\right| < \infty$ w.p.1.

Note that by definition of $h_{\theta,\ell}$ (see (2))

$$\int h_{\theta,\ell}(y)\ \theta(\mathrm{d}y) \geq \theta(\{y : \pi_u(y) \in H_{\theta,\ell}\})\ . \tag{12}$$

Condition E3b states that the rings $\{H_{\theta_n^{(k)},\ell)}, n \geq 0\}$ converge to $H_{\theta_\star^{(k)},\ell}$ w.p.1; therefore, E3a is satisfied as soon as the limiting rings are of positive probability under the distribution of $\pi_u(Z)$ when $Z \sim \theta_\star^{(k)}$.

When the energy bounds are fixed, the conditions E3b-c are clearly satisfied and E3a holds under convenient choice of the rings. We will discuss in Section 3.3 how to check the condition E3 with adaptive energy bounds.

3.2 Convergence results

Proposition 3.2 shows that the kernels $P_\theta^{(k)}$ satisfy a geometric drift inequality and a minorization condition, with constants in the drift independent of θ for $\theta \in \Theta_m$ (Θ_m being defined in (5)). The proof is in Appendix A.1.

Proposition 3.2. *Assume E1a and E2. For all* $k \in \{1, \ldots, K\}$*:*

(a) There exist $\tilde{\lambda}_k \in (0,1)$ *and* $\tilde{b}_k < +\infty$ *such that for all* $m \geq 1$ *and any* $\theta \in \Theta_m$,

$$P_\theta^{(k)} W_k \leq \tilde{\lambda}_k W_k + \tilde{b}_k \; m\,\theta(W_k) \quad . \tag{13}$$

For all $p \in (0, \sup_{\mathbf{X}} \pi)$ *and all* $\theta \in \bigcup_m \Theta_m$, *the sets* $\{\pi \geq p\}$ *are* 1*-small for* $P_\theta^{(k)}$ *and the minorization constants depend neither upon* θ *nor on* m.

(b) For all $\theta \in \bigcup_m \Theta_m$, *there exists a probability measure* $\pi_\theta^{(k)}$ *invariant for* $P_\theta^{(k)}$. *In addition,* $\pi_\theta^{(k)}(W_k) \leq \tilde{b}_k (1-\tilde{\lambda}_k)^{-1}\; m\theta(W_k)$ *for* $\theta \in \Theta_m$.

Theorem 3.3 is proved in Section B. Theorem 3.3(a) shows that there exists $m_\star \geq 1$ such that w.p.1, for all n large enough $\theta_n^{(k)}$ belongs to some $\Theta_{m_\star}$. Note that in [2], a s-LLN for the Equi-Energy sampler is established by assuming that there exists a deterministic positive integer m such that w.p.1, $\theta_n^{(k)} \in \Theta_m$ for any n. Such a condition is quite strong since roughly speaking, it means that after n steps (even for small n), all the rings contain a number of point which is proportional to n, w.p.1. This is all the more difficult to guarantee in practice, that the rings have to be chosen prior to any exploration of π. Our approach allows to relax this strong condition.

The convergence of the marginals and the law of large numbers both require the convergence in n (k fixed) of $\{\pi_{\theta_n^{(k)}}^{(k+1)}(f), n \geq 0\}$ for some functions f. Such a convergence is addressed in Theorem 3.3(b). We will then have the main ingredients to establish the convergence results for the processes $Y^{(k)}$, $k \geq 1$.

Theorem 3.3. *Assume E1, E2, E3 and* $\mathbb{E}[W_k(Y_0^{(k)})] < \infty$ *for all* $k \in \{1, \cdots, K\}$.

(a) There exists $m_\star \geq 1$ *such that for all* $k \in \{1, \ldots, K-1\}$

$$\mathbb{P}\left(\bigcup_{q \geq 1} \bigcap_{n \geq q} \{\theta_n^{(k)} \in \Theta_{m_\star}\} \right) = 1 \; . \tag{14}$$

(b) For any $k \in \{1, \cdots, K\}$, *any* $a \in (0,1)$ *and any continuous function* $f \in \mathcal{L}_{W_k^a}$,

$$\lim_{n \to \infty} \pi_{\theta_n^{(k-1)}}^{(k)}(f) = \theta_\star^{(k)}(f) \; , \; \textit{w.p.1} \; .$$

(c) For any $k \in \{1, \cdots, K\}$ *and for all bounded continuous function* $f : \mathbf{X} \to \mathbb{R}$, $\lim_{n\to\infty} \mathbb{E}[f(Y_n^{(k)})] = \theta_\star^{(k)}(f)$.

(d) Let $a \in (0, \frac{1+\Gamma}{2} \wedge 1)$. *For any* $k \in \{1, \cdots, K\}$ *and for all continuous function* f *in* $\mathcal{L}_{W_k^a}$

$$\lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} f(Y_m^{(k)}) = \theta_\star^{(k)}(f) \qquad \mathbb{P} - a.s. \; .$$

Observe that, for the process $\{Y^{(k)}, k \in \mathbb{N}\}$, the family of functions for which the law of large numbers holds depends *(i)* upon Γ given by EE3(c) i.e. in some sense, depends upon the adaptation rate; and *(ii)* the temperature ladder. In the case τ_k can be chosen arbitrarily close to β_1/β_k for any k (see comments after [21, Theorem 4.1 and 4.3]), this family of functions only depends upon Γ and the lowest inverse temperature : it is all the more restrictive than β_1 is small.

To our best knowledge, we are the first to prove such convergence results for AEE (and EE): previous works [16, 3] consider the simpler case when there is no selection i.e. $g_\theta(x, y) = 1$.

3.3 Comments on Assumption E3

We propose to choose the adaptive boundaries $\xi_{\theta,\ell}$ as the p_ℓ-quantile of the distribution of $\pi_u(Z)$ when Z is sampled under the distribution θ. This section proves that empirical quantiles of regularly spaced orders are examples of adaptive boundaries $\xi_{\theta_n^{(k)},\ell}$ satisfying E3. Let F_θ be the cumulative distribution function (cdf) of the r.v. $\pi_u(Z)$ when $Z \sim \theta$:

$$F_\theta(x) = \int \mathbf{1}_{\{\pi_u(z) \le x\}} \theta(\mathrm{d}z) \;, \qquad x \in [0, \infty) \;.$$

We denote the quantile function associated to $\pi_u(Z)$ by:

$$F_\theta^{-1}(p) = \inf\{x \ge 0, F_\theta(x) \ge p\} \quad \forall p > 0 \;; \qquad\qquad F_\theta^{-1}(0) = 0 \;.$$

With this definition, for $0 < p_1 < \cdots < p_{S-1} < 1$, we set $\xi_{\theta,\ell} \stackrel{\text{def}}{=} F_\theta^{-1}(p_\ell)$.

With this choice of the boundaries, the condition E3a holds: by (12), E3a is satisfied because π is continuous. The conditions E3b-c require the convergence of the quantile estimators and a rate of convergence of the variation of two successive boundaries. To prove such conditions, we use an Hoeffding-type inequality.

Proposition 3.4. *Assume*

(i) The cumulative distribution function $F_{\theta_\star^{(1)}}$ where $\theta_\star^{(1)}$ is given by (11), is differentiable with positive derivative on $F_{\theta_\star^{(1)}}^{-1}((0,1))$.

(ii) there exists $\overline{W}$ such that $Y^{(1)}$ is a $\overline{W}$-uniformly ergodic Markov chain with initial distribution satisfying $\mathbb{E}\left[Y_0^{(1)}\right] < \infty$.

Then E3b-c hold with $\Gamma = 1/2$ and $K = 2$.

The proof is in Section B.5. Extensions of Proposition 3.4 to the case when $Y^{(1)}$ is not a uniformly ergodic Markov chain is, to our best knowledge, an open question. Therefore, our convergence result of AEE when the boundaries are the quantiles defined by inversion of the cdf of the auxiliary process applies to the 2-stage level and seems difficult to extend to the K-stage, $K > 2$.

We proved recently in [35] that when the quantiles are defined by a stochastic approximation procedure, the conditions E3b-c hold even under very weak conditions on the auxiliary $Y^{(k)}$, $k \ge 2$. In this case, the convergence of the K-level AEE with $K > 2$ is established.

4 Application to motif sampling in biological sequences

One of the challenges in biology is to understand how gene expression is regulated. Biologists have found that proteins called transcription factors play a role in this regulation. Indeed, transcription factors bind on special motifs of DNA and then attract or repulse the enzymes that are responsible of transcription of DNA sequences into proteins. This is the reason why finding these binding motifs is crucial. But binding motifs do not contain deterministic start and stop codons: they are only random sequences that occurs more frequently than expected under the background model.

Several methods have been proposed so far to retrieve binding motifs [36, 24, 9], which yields to a complete Bayesian model [25]. Among the Bayesian approach, one effective method is based on the Gibbs sampler [23] - it has been popularized by software programs [26, 33]. Nevertheless, as discussed in [22], it may happen that classical MCMC algorithms are inefficient for this Bayesian approach. Therefore, [22] show the interest of the Equi-Energy sampler when applied to this Bayesian inverse problem; more recently, [32] proposed a Gibbs-based algorithm for a similar model (their model differs from the following one through the assumptions on the background sequence).

We start with a description of our model for motif sampling in biological sequences - this section is close to the description in [22] but is provided to make this paper self-contained. We then apply AEE and compare it to the Interacting MCMC of [16, Section 3] (hereafter called I-MCMC), and to a Metropolis-Hastings algorithm (MH). Comparison with Gibbs-based algorithms (namely BioProspector and AlignACE) can be found in the paper of [22].

The available data is a DNA sequence, which is modeled by a background sequence in which some motifs are inserted. The background sequence is represented by a vector $\mathsf{S} = (s_1, s_2, \ldots, s_L)$ of length L. Each element s_i is a nucleotide in $\{A, C, G, T\}$; in this paper, we will choose the convention $s_i \in \{1, 2, 3, 4\}$. The length w of a motif is assumed to be known. The motif positions are collected in a vector $A = (a_1, \ldots, a_L)$, with the convention that $a_i = j$ iff the nucleotide s_i is located at position number j of a motif; and $a_i = 0$ iff s_i is not in the motif. The goal of the statistical analysis of the data S is to explore the distribution of A given the sequence S. We now introduce notations and assumptions on the model in order to define this conditional distribution.

We denote by p_0 the probability that a sub-sequence of length w of S is a motif. It is assumed that the background sequence is a Markov chain with (deterministic) transition matrix $v_0 = \{v_0(i,j)\}_{1 \le i,j \le 4}$ on $\{1, \cdots, 4\}$; and the nucleotide in a sequence are sampled from a multinomial distribution of parameter $v = \{v(i,j)\}_{1 \le i \le 4, 1 \le j \le w}$, $v(i,j)$ being the probability for the j-th element of a motif to be equal to i.

In practice, it has been observed that approximating $v_0(i,j)$ by the frequency of jumps from i to j in the (whole) sequence S is satisfying. It is assumed that the r.v. (v, p_0) are independent with prior distribution $\prod_{j=1}^{w} \chi(v(\cdot, j))$ and $\chi'(p_0)$; $\chi(v(\cdot, j))$ is a Dirichlet distribution with parameters $\iota_j = (\iota_{j,1}, \cdots, \iota_{j,4})$ and $\chi'(p_0)$ is a Beta distribution with parameters (b_1, b_2). ι_j, b_1 and b_2 are assumed to be known.

Therefore, given (v, p_0), (A, S) is a Markov chain described as follows:

- If $a_{k-1} \in \{1, \ldots, w-1\}$ then $a_k = a_{k-1} + 1$; else $\mathbb{P}(a_k = 1 | a_{k-1} \in \{0, w\}, p_0, v) = 1 - \mathbb{P}(a_k = 0 | a_{k-1} \in \{0, w\}, p_0, v) = p_0$.
- If $a_k = 0$, $s_k \sim v_0(s_{k-1}, .)$; else s_k is drawn from a Multinomial distribution with parameter $v(\cdot, a_k)$.

The chains are initialized with $\mathbb{P}(a_1 = 1|p_0) = 1 - \mathbb{P}(a_1 = 0|p_0) = p_0$; the distribution of s_1 given $a_1 = 0$ and v (resp. given $a_1 = 1$ and v) is uniform on $\{1, \cdots, 4\}$ (resp. a Multinomial distribution with parameter $v(\cdot, 1)$).

This description yields to the following conditional distribution of A given S: (up to a multiplicative constant) - see [22] for similar derivation -

$$\mathrm{P}(A|S) \propto \frac{\Gamma(N_1(A)+b_1)\Gamma(N_0(A)+b_2)}{\Gamma(N_1(A)+N_0(A)+b_1+b_2)} \prod_{i=1}^{w} \frac{\prod_{j=1}^{4}\Gamma(c_{j,i}(A)+\iota_{j,i})}{\Gamma(\sum_{\ell=1}^{4} c_{\ell,i}(A)+\iota_{\ell,i})} \cdots$$
$$\times \prod_{k=2}^{L} (\delta_{a_{k-1}+1}(a_k))^{\mathbf{1}_{a_{k-1}\in\{1,\ldots,w-1\}}} \prod_{k=2}^{L} (v_0(s_{k-1}, s_k))^{\mathbf{1}_{a_k=0}} \left(\mathbf{1}_{\{0\}}(a_1)\frac{1}{4} + \mathbf{1}_{\{1\}}(a_1)\right)$$

where

- $N_1(A) = \#\{k, a_k = 1\}$ is the number of elements of A equal to 1.
- $N_0(A) = \#\{k, a_k = 0\}$ is the number of elements of A equal to 0.
- $c_{j,i}(A) = \sum_{k=1}^{L} \mathbf{1}_{a_k=i}\mathbf{1}_{s_k=j}$ is the number of pairs (a_k, s_k) equal to (i, j).

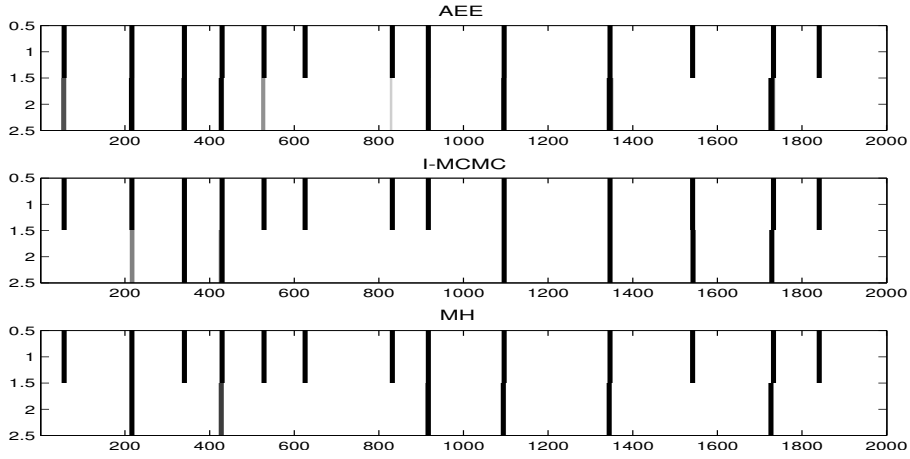


Figure 5: Results given by AEE, I-MCMC and a MH sampler

To highlight the major role of the equi-energy jumps, and the importance of the construction of the rings to make the acceptance probability of the jumps large enough, we compare AEE to I-MCMC, and to MH. The data are obtained with values of p_0, v_0 and v similar to those of [22]: $p_0 = 0.005$, $b_1 = 2$, $b_2 = 200$, $\iota_{j,i} = 1$ for all j, i, and

$$v_0 = \begin{pmatrix} 0.1 & 0.7 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \\ 0.7 & 0.1 & 0.1 & 0.1 \end{pmatrix}, \; v = \begin{pmatrix} 0.5 & 0.6 & 0.2 & 0.4 & 0.1 & 0.3 & 0.6 & 0.1 & 0.4 & 0.4 & 0.3 & 0 \\ 0 & 0.2 & 0 & 0.2 & 0.8 & 0.7 & 0 & 0.9 & 0 & 0 & 0.2 & 0.3 \\ 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0.3 & 0.5 & 0.4 & 0.1 \\ 0.5 & 0.2 & 0 & 0.4 & 0.1 & 0 & 0.4 & 0 & 0.3 & 0.1 & 0.1 & 0.6 \end{pmatrix}.$$

We sample a sequence S of length $L = 2000$ and the size of the motif is $w = 12$.

We now detail how the MH and the Metropolis-Hastings steps of AEE and I-MCMC are run. For the Metropolis-Hastings stage, the proposal distribution $p(A_n, \tilde{A}_{n+1})$ is of the form

$$p(A_n, \tilde{A}_{n+1}) = q_0(\tilde{a}_1^{n+1}) \prod_{j=1}^{L-1} q_j(\tilde{a}_j^{n+1}, \tilde{a}_{j+1}^{n+1}; A_n) \ ,$$

where we set $\tilde{A}_{n+1} = (\tilde{a}_1^{n+1}, \cdots, \tilde{a}_L^{n+1})$. The proposed state $\tilde{A}_{n+1}$ of the Metropolis-Hastings step is then sampled element by element; the distributions are designed to be close to the previous model: $\tilde{a}_{j+1}^{n+1}$ equal to $\tilde{a}_j^{n+1} + 1$ if $\tilde{a}_j^{n+1} \in \{1, \ldots, w-1\}$, and else, $\tilde{a}_{j+1}^{n+1}$ is sampled under a Bernoulli distribution of parameter

$$\frac{\hat{p}_0 \prod_{i=1}^{w} \hat{v}_{A_n}(s_{j+i-1}, i)}{\hat{p}_0 \prod_{i=1}^{w} \hat{v}_{A_n}(s_{j+i-1}, i) + (1 - \hat{p}_0) \prod_{i=1}^{w-1} v_0(s_{j+i}, s_{j+i+1})} \ ; \tag{15}$$

the replacement constant $\hat{p}_0$ is fixed by the users and $\hat{v}_{A_n}$ is given by $\hat{v}_{A_n}(s, i) \propto c_{s,i}(A_n) + c$ - where c is a value fixed by the users. $q_0(\tilde{a}_1^{n+1})$ is the Bernoulli distribution with parameter (15). Finally, the candidate $\tilde{A}_{n+1}$ is accepted with probability

$$1 \wedge \frac{\mathrm{P}(\tilde{A}_{n+1}|S)^{1/T_k}}{\mathrm{P}(A_n|S)^{1/T_k}} \frac{p(\tilde{A}_{n+1}, A_n)}{p(A_n, \tilde{A}_{n+1})} \ .$$

Figure 5 displays the results obtained by AEE, I-MCMC and a MH sampler. Each subplot displays two horizontal lines with length equal to the length of the observed DNA sequence. The upper line represents the actual localization of the motifs, and the lower line represents in gray-scale the probability for each position to be part of a motif computed by one run of each algorithm after 2000 iterations. For AEE and I-MCMC, we choose $\varepsilon = 0.1$, $K = 5$, $S = 3$. The acceptance rate of the jump for AEE was about five times higher than for I-MCMC, which confirms the interest of the rings. As expected, AEE performs better than the other algorithms: there were 13 actual motifs, and AEE retrieved 10 motifs, whereas the I-MCMC and the MH retrieved respectively 7 and 6 motifs.

5 Conclusion

As illustrated by the numerical examples, the efficiency of EE depends upon the choice of the energy rings. The adaptation we proposed improves this efficiency since it makes the probability of accepting a jump more stable. It is known that adaptation can destroy the convergence of the samplers: we proved that AEE converges under quite general conditions on the adapted bounds and these general conditions can be used to prove the convergence of AEE when applied with other adaptation strategies [35]. It is also the first convergence result for an interacting MCMC algorithm including a selection mechanism. Our sketch of proof can be a basis for the proof of other interacting MCMC such as the SIMCMC algorithm of [13], the Non-Linear MCMC algorithms described in [3, Section 3] or the PTEEM algorithm of [10].

A Results on the transition kernels $P_\theta^{(k)}$

Define

$$G_\theta(x) \stackrel{\text{def}}{=} \int g_\theta(x,z)\theta(\mathrm{d}z) \;, \qquad \tilde{\theta}(x,\mathrm{d}y) \stackrel{\text{def}}{=} \frac{g_\theta(x,y)\theta(\mathrm{d}y)}{G_\theta(x)} \;. \tag{16}$$

A.1 Proof of Proposition 3.2

The case $k=1$ is a consequence of E2 since $P_\theta^{(1)} = P^{(1)}$ for any θ so that $\pi_\theta^{(1)} \propto \pi^{\beta_1}$. We now consider the case $k \in \{2,\cdots,K\}$: in the proof below, for ease of notations we will write P, P_θ, W, λ, b and π_θ instead of $P^{(k)}$, $P_\theta^{(k)}$, W_k, λ_k, b_k and $\pi_\theta^{(k)}$.

(a) Let $m \geq 1$ and $\theta \in \Theta_m$. By definition of g_θ (see (2)) and of Θ_m (see (5)), $1/m \leq \int g_\theta(x,y)\theta(\mathrm{d}y) \leq S$. Moreover, by E2b

$$P_\theta W(x) = (1-\varepsilon)PW(x) + \varepsilon K_\theta W(x) \leq (1-\varepsilon)(\lambda W(x) + b) + \varepsilon K_\theta W(x) \;.$$

We have by (3), (10) and (16)

$$K_\theta W(x) = W(x) + \int W(y)\alpha_\theta(x,y)\left(1 - \frac{\pi^{\tau_k\beta_k}(y)}{\pi^{\tau_k\beta_k}(x)}\right)\tilde{\theta}(x,\mathrm{d}y) \;.$$

By (4),

$$K_\theta W(x) \leq W(x) + m\int_{\{y,\pi(y)\leq\pi(x)\}} W(y)\frac{\pi^{\beta_k-\beta_{k-1}}(y)}{\pi^{\beta_k-\beta_{k-1}}(x)}\left(1 - \frac{\pi^{\tau_k\beta_k}(y)}{\pi^{\tau_k\beta_k}(x)}\right) g_\theta(x,y)\theta(\mathrm{d}y) \;.$$

Defining ψ by $\psi(\sigma) = \sigma/(\sigma+1)^{(\sigma+1)/\sigma}$ gives the upper bound $\sup_{z\in[0,1]} z(1-z^\sigma) \leq \psi(\sigma)$. Hence, $K_\theta W(x) \leq W(x) + Sm\;\psi\left(\tau_k\beta_k/(\beta_k-\beta_{k-1})\right)\theta(W)$. This yields $P_\theta W(x) \leq \tilde{\lambda}W(x) + \tilde{b}m\theta(W)$ with $\tilde{\lambda} = (1-\varepsilon)\lambda + \varepsilon < 1$ and $\tilde{b} = \varepsilon S\;\psi\left(\tau_k\beta_k/(\beta_k-\beta_{k-1})\right) + (1-\varepsilon)b$. The minorization condition comes from the lower bound $P_\theta(x,A) \geq (1-\varepsilon)P(x,A)$.

(b) Let $m \geq 1$ and $\theta \in \Theta_m$. By E2a, P is φ-irreducible and so is P_θ; P_θ possesses a 1-small set and is thus aperiodic. In addition, $P_\theta W \leq (1+\tilde{\lambda})W/2 + \tilde{b}\theta(W)\mathbf{1}_{\{W\leq c\}}$, with $c \stackrel{\text{def}}{=} 2\tilde{b}m\;\theta(W)(1-\tilde{\lambda})^{-1}$ and $\{W \leq c\}$ is a 1-small set for P_θ. By [28, Chapter 15], π_θ exists and $\pi_\theta(W) \leq \tilde{b}m\theta(W)(1-\tilde{\lambda})^{-1}$.

A.2 Ergodic behavior

Lemma A.1. *Assume E1a and E2. Then for all $a \in (0,1)$, for all $m \geq 1$ and all $\theta \in \Theta_m$, there exist C_θ and $\rho_\theta \in (0,1)$ such that for all $x \in \mathbf{X}$ and any $j \geq 1$ and any $k \in \{1,\cdots,K\}$,*

$$\| \left(P_\theta^{(k)}\right)^j (x,.) - \pi_\theta^{(k)} \|_{W_k^a} \leq C_\theta\;\rho_\theta^j\;W_k^a(x) \;. \tag{17}$$

Let $k \in \{1,\cdots,K-1\}$ and assume in addition that $\lim_{n\to\infty}\theta_n^{(k)}(W_k) = \theta_\star^{(k)}(W_k)$ w.p.1. Then for any positive integer q, on the set $\bigcap_{n\geq q}\{\theta_n \in \Theta_{m_\star}\}$

$$\limsup_n \rho_{\theta_n^{(k)}} < 1, \limsup_n C_{\theta_n^{(k)}} < +\infty, \mathbb{P}-a.s. \quad . \tag{18}$$

Proof. The proof in the case $k = 1$ is a consequence of E2 and [28, Chapter 15] since $P_\theta^{(1)} = P^{(1)}$. Consider the case $k \geq 2$. Here again, the dependence upon k is omitted: P_θ, W, θ_n denote $P_\theta^{(k)}, W_k$ and $\theta_n^{(k)}$.

Proof of (17) Let $a \in (0,1)$ and set $V = W^a$. By the Jensen's inequality and Proposition 3.2, there exists $\bar{\lambda} \in (0,1)$ and $\bar{b}$ such that for any $m \geq 1$ and any $\theta \in \Theta_m$,

$$P_\theta V \leq \bar{\lambda} V + \bar{b}\, m\, \theta(W)^a \; .$$

Let $m \geq 1$ and $\theta \in \Theta_m$. By [16, Lemma 2.3.], (17) holds and there exist constants $C, \gamma > 0$ such that for any $\theta \in \Theta_m$,

$$C_\theta \vee (1-\rho_\theta)^{-1} \leq C \left(\bar{b}\; m\theta(W) \vee \delta_\theta^{-1} \vee (1-\bar{\lambda})^{-1}\right)^\gamma \; ,$$

where δ_θ is the minorizing constant of P_θ on the set $\{x : W(x) \leq 2\bar{b}m\, \theta(W)\, (1-\bar{\lambda})^{-1} - 1\}$.

Proof of (18) For all $\omega \in \bigcap_{n \geq q} \{\theta_n \in \Theta_{m_\star}\}$,

$$\limsup_n \{C_{\theta_n(\omega)} \vee (1-\rho_{\theta_n(\omega)})^{-1}\} \leq C \left(\bar{b}\, m\, \limsup_n \theta_n(W) \vee \limsup_n \delta_{\theta_n(\omega)}^{-1} \vee (1-\bar{\lambda})^{-1}\right)^\gamma \; .$$

Since $\limsup_n \theta_n(W) = \theta_\star(W) < \infty$ w.p.1, $\limsup_n \delta_{\theta_n(\omega)}^{-1} < \infty$ w.p.1. thus showing that on the set $\bigcap_{n \geq q} \{\theta_n \in \Theta_{m_\star}\}$, $\limsup_n \{C_{\theta_n(\omega)} \vee (1-\rho_{\theta_n(\omega)})^{-1}\} < \infty$. This implies (18). □

A.3 Moment conditions

Let $m_\star > 0$. Define for any positive integer q and any $k \in \{1, \cdots, K-1\}$,

$$A_{q,n}^{(k)} = \bigcap_{\ell \leq k} \bigcap_{q \leq j \leq n} \left\{\theta_j^{(\ell)} \in \Theta_{m_\star}\right\} \quad \text{if } q \leq n, \quad \text{and} \quad A_{q,n}^{(k)} = \Omega \quad \text{otherwise;}$$

by convention, $A_{q,n}^{(0)} = \Omega$ for any $q, n \geq 0$.

Lemma A.2. *Assume E1a, E2 and* $\mathbb{E}\left[W_k(Y_0^{(k)})\right] < \infty$ *for any* $k \in \{1, \cdots, K\}$. *Then for any* $k \in \{1, \cdots, K\}$,

$$\sup_{j \geq 1} \mathbb{E}\left[W_k(Y_j^{(k)}) \mathbf{1}_{A_{q,j-1}^{(k-1)}}\right] < \infty \; . \tag{19}$$

Proof. The proof is by induction on k. The case $k = 1$ is a consequence of E2 since $P_\theta^{(1)} = P^{(1)}$. Assume the property holds for $k \in \{2, \cdots, K-1\}$. In this proof, $W_{k+1}, P_\theta^{(k+1)}, \theta_n^{(k)}, Y^{(k)}, Y^{(k+1)}, P^{(k+1)}$, $K_\theta^{(k+1)}$ will be denoted by $W, P_\theta, \theta_n, Y, X, P, K_\theta$.

By (6) and Proposition 3.2 we obtain, for $j > q$

$$\begin{aligned}
\mathbb{E}\left[W(X_j)\mathbf{1}_{A_{q,j-1}^{(k)}}\right] &\leq \mathbb{E}\left[P_{\theta_{j-1}} W(X_{j-1}) \mathbf{1}_{A_{q,j-1}^{(k)}}\right] \\
&\leq \tilde{\lambda}\mathbb{E}\left[W(X_{j-1})\mathbf{1}_{A_{q,j-2}^{(k)}}\right] + \tilde{b}\, m_\star\, \mathbb{E}\left[\theta_{j-1}(W)\mathbf{1}_{A_{q,j-1}^{(k-1)}}\right] \\
&\leq \tilde{\lambda}\mathbb{E}\left[W(X_{j-1})\mathbf{1}_{A_{q,j-2}^{(k)}}\right] + \tilde{b}\, m_\star \sup_l \mathbb{E}\left[W(Y_l)\mathbf{1}_{A_{q,l-1}^{(k-1)}}\right] \; .
\end{aligned}$$

Since $W_{k+1} \in \mathcal{L}_{W_k}$, the induction assumption implies that $\sup_l \mathbb{E}\left[W(Y_l)\mathbf{1}_{A_{q,l-1}^{(k-1)}}\right] < \infty$. Iterating this inequality allows to write that for some constant C'

$$\sup_{j\geq q} \mathbb{E}\left[W(X_j)\mathbf{1}_{A_{q,j-1}^{(k)}}\right] \leq C'\, \mathbb{E}\left[W(X_q)\right] \ .$$

Finally, by definition of P_{θ_j}, either $P_{\theta_j} = P$ if $\theta_j \notin \bigcup_m \Theta_m$, or $P_{\theta_j} = (1-\epsilon)P + \epsilon K_{\theta_j}$ otherwise; note that if $\theta_j \in \bigcup_m \Theta_m$ then $\theta_j \in \Theta_{1/j}$. Since both P and P_θ for $\theta \in \bigcup_m \Theta_m$ satisfy a drift inequality (see E2 and Proposition 3.2), $\mathbb{E}\left[W(X_q)\right] < \infty$ by (6). □

B Proof of Theorem 3.3

R1 (k) There exists $m_\star > 0$ such that $\mathbb{P}\left(\bigcup_{q\geq 1}\bigcap_{n\geq q}\{\theta_n^{(k)} \in \Theta_{m_\star}\}\right) = 1$.

R2 (k) for any $a \in (0,1)$ and any continuous function $f \in \mathcal{L}_{W_k^a}$,

$$\lim_{n\to\infty} \pi^{(k)}_{\theta_n^{(k-1)}}(f) = \theta_\star^{(k)}(f) \ .$$

R3 (k) For all bounded continuous function f, $\lim_{n\to\infty} \mathbb{E}\left[f(Y_n^{(k)})\right] = \theta_\star^{(1)}(f)$.

R4 (k) $\theta_\star^{(k)}(W_{k+1}) < +\infty$, and for any $a \in (0, \frac{1+\Gamma}{2} \wedge 1)$ and any continuous function f in $\mathcal{L}_{W_k^a}$, $\theta_n^{(k)}(f) \to \theta_\star^{(k)}(f)$ a.s.

By Proposition 3.1, the conditions R3 and R4 hold for $k = 1$; R2 also holds for $k = 1$ since $\pi_\theta^{(1)} = \theta_\star^{(1)}$ for any θ. We assume that for any $j \leq k$, for $k \in \{1, \cdots, K-1\}$, the conditions R1$(j-1)$, R2(j), R3(j) and R4(j) hold. We prove that R1(k), R2$(k+1)$, R3$(k+1)$ and R4$(k+1)$ hold. To make the notations easier, the superscript k is dropped from the notations: the auxiliary process $Y^{(k)}$ will be denoted by Y, and the process $Y^{(k+1)}$ by X; $P^{(k+1)}, W_{k+1}, K_\theta^{(k+1)}, P_\theta^{(k+1)}, \alpha_\theta^{(k+1)}, \pi_\theta^{(k+1)}$ and $\theta_n^{(k)}, \theta_\star^{(k)}$ are resp. denoted by $P, W, K_\theta, P_\theta, \alpha_\theta, \pi_\theta$ and $\theta_n, \theta_\star$.

Finally, we define the V-variation of the two kernels P_θ and $P_{\theta'}$ by:

$$D_V(\theta,\theta') = \sup_{x\in\mathbf{X}} \left(\frac{\|P_\theta(x,.) - P_{\theta'}(x,.)\|_V}{V(x)}\right) \ .$$

When $V = 1$, we will simply write D.

B.1 Proof of R1(k)

The proof is prefaced with a preliminary lemma.

Lemma B.1. *For all $l \in \{1, \cdots, S-1\}$ and any θ, θ',*

$$\sup_{x\in\mathbf{X}} \left|h_{\theta,l}(x) - h_{\theta',l}(x)\right| \leq \frac{1}{r} \sup_{l\in\{1,\cdots,S-1\}} \left|\xi_{\theta,l} - \xi_{\theta',l}\right| \ .$$

Proof. Note that $|(1-a)_+ - (1-b)_+| \leq |b-a|$. Therefore, for all $x \in \mathbf{X}$:

$$|h_{\theta,l}(x) - h_{\theta',l}(x)| \leq \frac{\left|d(\pi_u(x), H_{\theta,l}) - d(\pi_u(x), H_{\theta',l})\right|}{r} .$$

This concludes the proof. □

(Proof of R1(k)) We prove there exist an integer $m_\star \geq 1$ and a positive r.v. N such that

$$\mathbb{P}(N < \infty) = 1 , \qquad \mathbb{P}\left(\bigcap_{n \geq N} \left\{ \inf_x \int g_{\theta_n}(x,y)\theta_n(\mathrm{d}y) \geq 1/m_\star \right\}\right) = 1 .$$

To that goal, we prove that with probability 1, for all n large enough,

$$\inf_x \int g_{\theta_n}(x,y)\theta_n(\mathrm{d}y) \geq \inf_{\ell \in \{1,\cdots,S-1\}} \int h_{\theta_\star,\ell}(y)\ \theta_\star(\mathrm{d}y) , \tag{20}$$

and use the assumption E3a. For all x and θ, there exists a ring index $l_{x,\theta} \in \{1,\cdots,S\}$ such that $\pi_u(x) \in H_{\theta,l_{x,\theta}}$. Upon noting that $d(\pi_u(x), H_{\theta,\ell_{x,\theta}}) = 0$, it holds

$$\liminf_n \inf_x \int g_{\theta_n}(x,y)\theta_n(\mathrm{d}y) \geq \liminf_n \inf_{l \in \{1,\cdots,S\}} \int h_{\theta_n,l}(y)\theta_n(\mathrm{d}y) .$$

We write

$$\begin{aligned}\int h_{\theta_n,l}(y)\theta_n(\mathrm{d}y) &\geq \int h_{\theta_\star,l}(y)\theta_n(\mathrm{d}y) - \int |h_{\theta_n,l}(y) - h_{\theta_\star,l}(y)|\,\theta_n(\mathrm{d}y) \\ &\geq \int h_{\theta_\star,l}(y)\theta_n(\mathrm{d}y) - \sup_{y \in \mathbf{X}} |h_{\theta_n,l}(y) - h_{\theta_\star,l}(y)| .\end{aligned}$$

By definition of $h_{\theta_\star,\ell}$, $y \mapsto h_{\theta_\star,l}(y)$ is continuous and bounded. Therefore, by R4(k), Lemma B.1 and E3b, the proof of (20) is concluded by

$$\liminf_n \int h_{\theta_n,l}(y)\theta_n(\mathrm{d}y) > \int h_{\theta_\star,l}(y)\theta_\star(\mathrm{d}y) .$$

B.2 Proof of R2$(k+1)$

First of all, observe that by definition of π_θ (see Proposition 3.2) and the expression of P_θ, $\pi_{\theta_\star} \propto \pi^{\beta_{k+1}}$. We check the conditions of [16, Theorem 2.11]. By Proposition a it is sufficient to prove that for any $q \geq 1$, $\lim_{n\to\infty} |\pi_{\theta_n}(f) - \pi_{\theta_\star}(f)| \mathbf{1}_{\bigcap_{j \geq q}\{\theta_j \in \Theta_{m_\star}\}} = 0$ w.p.1.

Case f bounded Lemma A.1 and R4(k) show that on the set $\bigcap_{j\geq q}\{\theta_j \in \Theta_{m_\star}\}$, $\limsup_n C_{\theta_n} < \infty$ and $\limsup_n (1-\rho_{\theta_n})^{-1} < \infty$ w.p.1. Equicontinuity of the class $\{P_\theta f, \theta \in \Theta_{m_\star}\}$, where f is a bounded continuous function on $\mathbf{X}$, will follow from Lemmas B.2 to B.4. Finally, the weak convergence of the transition kernels is proved in Lemma B.5.

Case f unbounded Following the same lines as in the proof of [16, Theorem 3.5], it can be proved that the above discussion for f bounded and Proposition 3.2(b) imply

$$\lim_{n\to\infty} \{\pi_{\theta_n}(f) - \pi_{\theta_\star}(f)\}\mathbf{1}_{\cap_{j\geq q}\{\theta_j \in \Theta_{m_\star}\}} = 0$$

w.p.1. for any continuous function f such that $|f|_{W^a_{k+1}} < \infty$.

Lemma B.2. *For all $\theta \in \bigcup_m \Theta_m$, and x, x', $\sup_y |g_\theta(x,y) - g_\theta(x',y)| \leq \frac{S}{r}|\pi(x) - \pi(x')|$.*

Proof. By (2),

$$|g_\theta(x,y) - g_\theta(x',y)| \leq \sum_{l=1}^{S} |h_{\theta,l}(x) - h_{\theta,l}(x')| h_{\theta,l}(y) \leq \sum_{l=1}^{S} |h_{\theta,l}(x) - h_{\theta,l}(x')| \; .$$

The proof is completed since

$$|h_{\theta,l}(x) - h_{\theta,l}(x')| \leq \frac{|d(\pi(x), H_{\theta,l}) - d(\pi(x'), H_{\theta,l})|}{r} \leq \frac{|\pi(x) - \pi(x')|}{r} \; .$$

□

Lemma B.3. *Assume E1a. For all $m \geq 1$, there exists a constant C_m such that for all $x, x', y, y' \in \mathbf{X}$ and $\theta \in \Theta_m$*

$$\left|\alpha_\theta(x,y) - \alpha_\theta(x',y)\right| \leq C_m \left[\left|\pi^{\beta_k - \beta_{k+1}}(x) - \pi^{\beta_k - \beta_{k+1}}(x')\right| + \left|\pi(x) - \pi(x')\right|\right] \; , \tag{21}$$

$$\left|\alpha_\theta(x,y) - \alpha_\theta(x,y')\right| \leq C_m \left[\left|\pi^{\beta_k - \beta_{k+1}}(y) - \pi^{\beta_k - \beta_{k+1}}(y')\right| + \left|\pi(y) - \pi(y')\right|\right] \; . \tag{22}$$

Proof. By definition of α_θ (see (4)), $\alpha_\theta(x,y) - \alpha_\theta(x',y) = (1 \wedge a) - (1 \wedge b)$, with

$$a = \frac{\pi^{\beta_{k+1}-\beta_k}(y)}{\pi^{\beta_{k+1}-\beta_k}(x)} \frac{\int g_\theta(x,z)\theta(\mathrm{d}z)}{\int g_\theta(y,z)\theta(\mathrm{d}z)} \quad \text{and} \quad b = \frac{\pi^{\beta_{k+1}-\beta_k}(y)}{\pi^{\beta_{k+1}-\beta_k}(x')} \frac{\int g_\theta(x',z)\theta(\mathrm{d}z)}{\int g_\theta(y,z)\theta(\mathrm{d}z)} \; .$$

Note that $|(1 \wedge a) - (1 \wedge b)| \leq |a - b| \, (\mathbf{1}_{a\leq 1} + \mathbf{1}_{b\leq 1, a>1})$. By symmetry, we can assume that $b \leq 1$ and this implies

$$\frac{\pi^{\beta_{k+1}-\beta_k}(y)}{\pi^{\beta_{k+1}-\beta_k}(x')} \leq \frac{\int g_\theta(y,z)\theta(\mathrm{d}z)}{\int g_\theta(x',z)\theta(\mathrm{d}z)} \leq Sm \; ,$$

since $g_\theta(x,y) \leq S$. Therefore,

$$\begin{aligned} |a-b| &= \frac{\pi^{\beta_{k+1}-\beta_k}(y)}{\int g_\theta(y,z)\theta(\mathrm{d}z)} \left| \frac{\int g_\theta(x,z)\theta(\mathrm{d}z)}{\pi^{\beta_{k+1}-\beta_k}(x)} - \frac{\int g_\theta(x',z)\theta(\mathrm{d}z)}{\pi^{\beta_{k+1}-\beta_k}(x')} \right| \\ &\leq Sm \left[\pi^{\beta_{k+1}-\beta_k}(y) \left| \pi^{\beta_k - \beta_{k+1}}(x) - \pi^{\beta_k - \beta_{k+1}}(x') \right| + m \left| \int (g_\theta(x,z) - g_\theta(x',z))\theta(\mathrm{d}z) \right| \right] \; . \end{aligned}$$

The proof of (21) is concluded by Lemma B.2. The proof of (22) is on the same lines and omitted. □

Lemma B.4. *Assume E1 and E2a. For any $m \geq 1$ and for any continuous bounded function f, the class of functions $\{P_\theta f, \theta \in \Theta_m\}$ is equicontinuous.*

Proof. Let f be a continuous function on $\mathbf{X}$, bounded by 1. Let $m \geq 1$ and $\theta \in \Theta_m$. We have

$$\begin{aligned} P_\theta f(x) - P_\theta f(x') =& (1-\varepsilon)\left(Pf(x) - Pf(x')\right) + \varepsilon\left(f(x) - f(x')\right)\left(1 - \int \alpha_\theta(x', y)\tilde{\theta}(x, \mathrm{d}y)\right) \\ &+ \varepsilon \int \left(f(y) - f(x)\right)\left(\alpha_\theta(x, y) - \alpha_\theta(x', y)\right)\tilde{\theta}(x, \mathrm{d}y) \\ &+ \varepsilon \int \alpha_\theta(x', y)(f(y) - f(x'))(\tilde{\theta}(x, \mathrm{d}y) - \tilde{\theta}(x', \mathrm{d}y)) \, , \end{aligned}$$

where $\tilde{\theta}$ is given by (16). This yields to

$$\begin{aligned} \left|P_\theta f(x) - P_\theta f(x')\right| \leq& \left|Pf(x) - Pf(x')\right| + \left|f(x) - f(x')\right| \\ &+ 2\sup_y \left|\alpha_\theta(x, y) - \alpha_\theta(x', y)\right| + 2\left\|\tilde{\theta}(x, .) - \tilde{\theta}(x', .)\right\|_{\mathrm{TV}} \, . \end{aligned}$$

We have

$$\begin{aligned} \|\tilde{\theta}(x, .) - \tilde{\theta}(x', .)\|_{\mathrm{TV}} &\leq \frac{1}{G_\theta(x)} \sup_y \left|g_\theta(x, y) - g_\theta(x', y)\right| + \frac{S}{G_\theta(x)G_\theta(x')}\left|G_\theta(x) - G_\theta(x')\right| \\ &\leq m \sup_y \left|g_\theta(x, y) - g_\theta(x', y)\right| + Sm^2 \sup_y \left|g_\theta(x, y) - g_\theta(x', y)\right| \, , \end{aligned}$$

where G_θ is given by (16). So Lemmas B.2 and B.3 imply that for all $m \geq 1$, there exists a constant C_m such that for all $\theta \in \Theta_m$:

$$\begin{aligned} \left|P_\theta f(x) - P_\theta f(x')\right| \leq& \left|Pf(x) - Pf(x')\right| + \left|f(x) - f(x')\right| \\ &+ C_m\left(|\pi(x) - \pi(x')| + |\pi^{\beta_k - \beta_{k+1}}(x) - \pi^{\beta_k - \beta_{k+1}}(x')|\right) \, . \end{aligned} \tag{23}$$

The proof is concluded since P is Feller and π is continuous. □

Lemma B.5. *Let $m \geq 1$. Assume E1, E3b and R4(k). For all $x \in \mathbf{X}$, there exists a set Ω_x such that $\mathbb{P}(\Omega_x) = 1$ and for all $\omega \in \Omega_x$ and any bounded continuous function f*

$$\lim_{n\to\infty} \left|P_{\theta_n(\omega)} f(x) - P_{\theta_\star} f(x)\right| \mathbf{1}_{\bigcap_j \{\theta_j \in \Theta_m\}} = 0 \, .$$

Proof. Following the same lines as in the proof of [16, Proposition 3.3.], it is sufficient to prove that for any $x \in \mathbf{X}$ and any bounded continuous function f, $\lim_{n\to\infty} P_{\theta_n}(f) = P_{\theta_\star}(f)$ w.p.1 on the set $\bigcap_j \{\theta_j \in \Theta_m\}$. Let f and x be fixed. We write

$$\begin{aligned} P_\theta f(x) - P_{\theta'} f(x) =& \varepsilon \int \left(\alpha_\theta(x, y) - \alpha_{\theta'}(x, y)\right)\left(f(y) - f(x)\right)\tilde{\theta}(x, \mathrm{d}y) \\ &+ \varepsilon \int \alpha_{\theta'}(x, y)\left(f(y) - f(x)\right)\left(\tilde{\theta}(x, \mathrm{d}y) - \tilde{\theta}'(x, \mathrm{d}y)\right) \, , \end{aligned} \tag{24}$$

where $\tilde{\theta}$ is given by (16). Moreover,

$$\tilde{\theta}(x,\mathrm{d}y) - \tilde{\theta}'(x,\mathrm{d}y) = \frac{g_\theta(x,y)\theta(\mathrm{d}y) - g_{\theta'}(x,y)\theta'(\mathrm{d}y)}{G_\theta(x)} + g_{\theta'}(x,y)\theta'(\mathrm{d}y)\,\frac{(G_{\theta'}(x) - G_\theta(x))}{G_\theta(x)G_{\theta'}(x)}\ .$$

This yields to

$$\begin{aligned}\varepsilon^{-1}\left(P_{\theta_n}f(x) - P_{\theta_\star}f(x)\right) &= \int \left(\alpha_{\theta_n}(x,y) - \alpha_{\theta_\star}(x,y)\right)\ \left(f(y) - f(x)\right)\tilde{\theta}_n(x,\mathrm{d}y) \\ &\quad - \int \frac{g_{\theta_\star}(x,y)}{G_{\theta_n}(x)} F(x,y)\left(\theta_\star(\mathrm{d}y) - \theta_n(\mathrm{d}y)\right) \\ &\quad - \int F(x,y)\left(\left(g_{\theta_n}(x,y) - g_{\theta_\star}(x,y)\right)\frac{\theta_n(\mathrm{d}y)}{G_{\theta_n}(x)} + g_{\theta_\star}(x,y)\theta_\star(\mathrm{d}y)\,\frac{(G_{\theta_\star}(x) - G_{\theta_n}(x))}{G_{\theta_n}(x)G_{\theta_\star}(x)}\right)\ ,\end{aligned}$$

where $F(x,y) = \alpha_{\theta_\star}(x,y)\left(f(y) - f(x)\right)$. There exists a constant C_m such that on the set $\bigcap_n\{\theta_n \in \Theta_m\}$, (see the proof of Lemma B.3 for similar upper bounds)

$$\begin{aligned}\left|\alpha_{\theta_n}(x,y) - \alpha_{\theta_\star}(x,y)\right| &\le C_m\ \left|\frac{G_{\theta_n}(x)}{G_{\theta_n}(y)} - \frac{G_{\theta_\star}(x)}{G_{\theta_\star}(y)}\right| \\ &\le m^2 S C_m\ \left(\left|G_{\theta_n}(x) - G_{\theta_\star}(x)\right| + \left|G_{\theta_n}(y) - G_{\theta_\star}(y)\right|\right)\end{aligned}$$

where $G_\theta(x)$ is defined by (16). We write by definition of the function g_θ (see (2))

$$\begin{aligned}\sup_x \left|G_{\theta_n}(x) - G_{\theta_\star}(x)\right| &\le \sup_{x,z}\left|g_{\theta_n}(x,z) - g_{\theta_\star}(x,z)\right| + \sup_x\left|\int g_{\theta_\star}(x,z)\theta_n(\mathrm{d}z) - \int g_{\theta_\star}(x,z)\theta_\star(\mathrm{d}z)\right| \\ &\le 2\sum_{l=1}^{S}\sup_z\left|h_{\theta_n,l}(z) - h_{\theta_\star,l}(z)\right| + \sum_{l=1}^{S}\left|\int h_{\theta_\star,l}(z)\theta_n(\mathrm{d}z) - \int h_{\theta_\star,l}(z)\theta_\star(\mathrm{d}z)\right|\ .\end{aligned}$$

By Lemma B.1 and E3b, the first term converges to zero w.p.1. Since $t \mapsto h_{\theta_\star,l}(t)$ is continuous, R4(k) implies that the second term tends to zero w.p.1. Therefore, on the set $\bigcap_n\{\theta_n \in \Theta_m\}$, $\sup_{x,y}\left|\alpha_{\theta_n}(x,y) - \alpha_{\theta_\star}(x,y)\right|$ converges to zero w.p.1, as well as $\sup_{x,y}\left|g_{\theta_n}(x,y) - g_{\theta_\star}(x,y)\right|$, and $\sup_x\left|G_{\theta_n}(x) - G_{\theta_\star}(x)\right|$.

Note that by Lemma B.3, $y \mapsto F(x,y)$ is bounded and continuous. Therefore, following the same lines as above, it can be proved that under R4(k) and E3b, on the set $\bigcap_n\{\theta_n \in \Theta_m\}$, $\lim_{n\to\infty}\left|\int F(x,y)\theta_n(x,\mathrm{d}y) - \int F(x,y)\theta_\star(x,\mathrm{d}y)\right| = 0$ w.p.1 □

B.3 Proof of R3(k+1)

We check the conditions of [16, Theorem 2.1]. Let f be a bounded continuous function on $\mathbf{X}$. By R2(k+1), $\lim_{n\to\infty}\pi_{\theta_n}(f) = \pi_{\theta_\star}(f) \propto \pi^{\beta_{k+1}}$ w.p.1. Let $\delta > 0$. By Proposition a, there exists $q \ge 1$ such that $\mathbb{P}(\bigcap_{n\ge q}\{\theta_n \in \Theta_{m_\star}\}) \ge 1 - \delta$. Following the same lines as in the proof of [16, Theorem 3.4], it can be proved by using Lemmas A.1, B.1 and B.6 and the condition E3c that $\lim_{n\to\infty}\mathbb{E}\left[\left(f(X_n) - \pi_{\theta_n}(f)\right)\mathbf{1}_{\bigcap_{n\ge q}\{\theta_n \in \Theta_{m_\star}\}}\right] = 0$. This concludes the proof.

Lemma B.6. *For all $m \geq 1$, there exists a constant C_m such that for any $\theta, \theta' \in \Theta_m$,*

$$D(\theta, \theta') \leq C_m \left(\|\theta - \theta'\|_{\mathrm{TV}} + \sup_{l,x} \left| h_{\theta,l}(x) - h_{\theta',l}(x) \right| \right) .$$

Proof. By definition of P_θ, for all function f bounded by 1, (24) holds. So

$$\begin{aligned} D(\theta, \theta') \leq{}& 2\varepsilon \sup_{x,y} |\alpha_\theta(x,y) - \alpha_{\theta'}(x,y)| \\ &+ 2\varepsilon S m^2 \left(\sup_{x,y} |g_\theta(x,y) - g_{\theta'}(x,y)| + \|\theta - \theta'\|_{\mathrm{TV}} + \sup_x |G_{\theta'}(x) - G_\theta(x)| \right) . \end{aligned}$$

The term $|\alpha_\theta(x,y) - \alpha_{\theta'}(x,y)|$ is equal to $|1 \wedge a - 1 \wedge b|$ with

$$a = \frac{\pi^{\beta_{k+1} - \beta_k}(y) \int g_\theta(x,z) \theta(\mathrm{d}z)}{\pi^{\beta_{k+1} - \beta_k}(x) \int g_\theta(y,z) \theta(\mathrm{d}z)} \quad \text{and} \quad b = \frac{\pi^{\beta_{k+1} - \beta_k}(y) \int g_{\theta'}(x,z) \theta'(\mathrm{d}z)}{\pi^{\beta_{k+1} - \beta_k}(x) \int g_{\theta'}(y,z) \theta'(\mathrm{d}z)} .$$

Note that $|1 \wedge a - 1 \wedge b| \leq |b - a| \left(\mathbf{1}_{\{b \leq 1, a > 1\}} + \mathbf{1}_{a \leq 1} \right)$. Therefore, for all $\theta, \theta' \in \Theta_m$,

$$\sup_{x,y} |\alpha_\theta(x,y) - \alpha_{\theta'}(x,y)| \leq S^2 m^2 \left(\sup_{x,y} |g_\theta(x,y) - g_{\theta'}(x,y)| + \|\theta - \theta'\|_{\mathrm{TV}} \right) .$$

The term $|G_{\theta'}(x) - G_\theta(x)|$ is upper bounded by

$$|G_{\theta'}(x) - G_\theta(x)| \leq \sup_{x,y} |g_\theta(x,y) - g_{\theta'}(x,y)| + S \|\theta - \theta'\|_{\mathrm{TV}} .$$

Moreover,

$$|g_\theta(x,y) - g_{\theta'}(x,y)| = \left| \sum_{l=1}^{S} [h_{\theta,l}(x) h_{\theta,l}(y) - h_{\theta',l}(x) h_{\theta',l}(y)] \right| \leq 2S \sup_{l,x} |h_{\theta,l}(x) - h_{\theta',l}(x)| .$$

This concludes the proof. □

B.4 Proof of R4($k+1$)

Let $a \in (0, \frac{1+\Gamma}{2} \wedge 1)$ and set $V = W^a$. We check the conditions of [16, Theorem 2.7]. By Proposition 3.2, condition A3 of [16] holds. By R2($k+1$), $\lim_{n \to \infty} \pi_{\theta_n}(f) = \pi_{\theta_\star}(f)$ w.p.1 for any continuous function f in $\mathcal{L}_{W^a}$. Condition A4 (resp. A5) of [16] is proved in Lemma B.7 (resp. Lemma B.8).

Lemma B.7. *Assume E1, E2, E3, R4(k), R1(j) and $\mathbb{E}[W_j(Y_0^{(j)})] < \infty$ for all $j \leq k$. Then for any $a \in (0, \frac{1+\Gamma}{2} \wedge 1)$*

$$\sum_{j \geq 1} j^{-1} (L_{\theta_j} \vee L_{\theta_{j-1}})^6 D_{W^a}(\theta_j, \theta_{j-1}) W^a(X_j) < \infty \qquad \mathbb{P} - a.s.,$$

where $L_\theta = C_\theta \vee (1 - \rho_\theta)^{-1}$.

Proof. By R1(j) for all $j \leq k$, it is sufficient to prove that for any positive integer q

$$\sum_{j\geq 1} j^{-1}(L_{\theta_j} \vee L_{\theta_{j-1}})^6 D_V(\theta_j, \theta_{j-1}) V(X_j)\ \mathbf{1}_{A_{q,j}^{(k)}} < \infty \qquad \mathbb{P}-\text{a.s.}$$

where $A_{q,j}^{(k)}$ is defined in Appendix A.3. Following the same lines as in the proof of Lemma A.2, it can be proved that $\sum_{j=1}^{q} j^{-1}(L_{\theta_j} \vee L_{\theta_{j-1}})^6 D_V(\theta_j, \theta_{j-1}) V(X_j) < \infty$ w.p.1.

By Lemma A.1 and R4(k), on the set $\bigcap_{l\geq q}\{\theta_l \in \Theta_{m_\star}\}$, $\limsup_n L_{\theta_n} < \infty$ w.p.1. Therefore, we have to prove that $\sum_{j\geq q} j^{-1} D_V(\theta_j, \theta_{j-1}) V(X_j) \mathbf{1}_{A_{q,j}^{(k)}} < \infty$ w.p.1. Following the same lines as in the proof of Lemma B.6, we obtain that on the set $A_{q,j}^{(k)}$, there exists a constant C_m such that

$$\begin{aligned} D_V(\theta_j, \theta_{j-1}) \leq C_m \left(\sup_l \left|\xi_{\theta_j,l} - \xi_{\theta_{j-1},l}\right| + \|\theta_j - \theta_{j-1}\|_{\mathrm{TV}} \right) \left(\|\theta_j\|_V + \|\theta_{j-1}\|_V\right) \\ + C_m\ \|\theta_j - \theta_{j-1}\|_V\ . \end{aligned}$$

Set s, γ, such that $s = 1 \vee (2a) < 1 + \gamma < 1 + \Gamma$. By E3c, there exists a r.v. Z finite w.p.1 such that $\mathbb{P}$-a.s.

$$|\xi_{\theta_n,l} - \xi_{\theta_{n-1},l}| + \|\theta_n - \theta_{n-1}\|_{\mathrm{TV}} \leq Z\left(\frac{1}{n^\gamma} + \frac{1}{n}\right)\ .$$

Therefore, it holds

$$\begin{aligned} \mathcal{I}_\gamma \stackrel{\text{def}}{=} \mathbb{E}\left[\left(\sum_{j\geq q} j^{-1}\left(j^{-\gamma} + j^{-1}\right)\left(\|\theta_j\|_V + \|\theta_{j-1}\|_V\right) V(X_j) \mathbf{1}_{A_{q,j}^{(k)}} \right)^{\frac{1}{s}} \right] \\ \leq \sum_{j\geq q} j^{-1/s}\left(j^{-\gamma/s} + j^{-1/s}\right) \mathbb{E}\left[\left(\|\theta_j\|_V + \|\theta_{j-1}\|_V\right)^{\frac{1}{s}} V^{\frac{1}{s}}(X_j) \mathbf{1}_{A_{q,j}^{(k)}} \right]\ . \end{aligned}$$

We have,

$$\mathcal{I}_\gamma \leq 2C(\gamma)\ \sup_j \mathbb{E}\left[\left(\|\theta_j\|_V\right)^{\frac{2}{s}} \mathbf{1}_{A_{q,j}^{(k)}} \right]^{1/2} \sup_j \mathbb{E}\left[V(X_j)^{\frac{2}{s}} \mathbf{1}_{A_{q,j}^{(k)}} \right]^{1/2}\ ,$$

where $C(\gamma) \stackrel{\text{def}}{=} \sum_{j\geq q}\left(j^{(-1-\gamma)/s} + j^{-2/s}\right)$ is finite since $2/s > 1$ and $1 + \gamma > s$. Since $V^{2/s} \leq W$, Lemma A.2 implies that $\sup_j \mathbb{E}\left[W(X_j) \mathbf{1}_{A_{q,j}^{(k)}} \right] < \infty$. In addition, since $2/s > 1$ we have, by Jensen's inequality,

$$\begin{aligned} \mathbb{E}\left[\|\theta_j\|_V^{\frac{2}{s}} \mathbf{1}_{A_{q,j}^{(k)}} \right] \leq \mathbb{E}\left[\left(\frac{1}{j}\sum_{p=1}^{j} V(Y_p) \right)^{\frac{2}{s}} \mathbf{1}_{A_{q,j}^{(k-1)}} \right] \leq \mathbb{E}\left[\frac{1}{j}\sum_{p=1}^{j} V^{\frac{2}{s}}(Y_p) \mathbf{1}_{A_{q,j}^{(k-1)}} \right] \\ \leq \sup_p \mathbb{E}\left[W(Y_p) \mathbf{1}_{A_{q,p-1}^{(k-1)}} \right] \end{aligned}$$

which is finite under Lemma A.2. Similarly, we prove that $\sum_{j\geq q} j^{-1} \|\theta_j - \theta_{j-1}\|_V V(X_j) \mathbf{1}_{A_{q,j}^{(k)}} < \infty$ w.p.1, upon noting that $\|\theta_j - \theta_{j-1}\|_V \leq j^{-1}(V(Y_j) + \theta_{j-1}(V))$. □

Lemma B.8. *Assume E1, E2, E3a-b, R4(k), R1(j) and* $\mathbb{E}[W_j(Y_0^{(j)})] < \infty$ *for all* $j \leq k$. *For any* $a \in (0,1)$,

$$\sum_{j\geq 1} j^{-1/a} L_{\theta_j}^{2/a} \; P_{\theta_j} W(X_j) < \infty \; , \qquad \mathbb{P}-a.s.$$

Proof. By R1(j) for all $j \leq k$, it is sufficient to prove that for any positive integer q

$$\sum_{j\geq 1} j^{-1/a} L_{\theta_j}^{2/a} \; P_{\theta_j} W(X_j) \; \mathbf{1}_{A_{q,j}^{(k)}} < \infty \qquad \mathbb{P}-\text{a.s.}$$

where $A_{q,j}^{(k)}$ is defined in Appendix A.3. Let $q \geq 1$. By Lemma A.1, $\sup_j L_{\theta_j} \mathbf{1}_{A_{q,j}^{(k)}} < \infty$ w.p.1; and, as in the proof of Lemma A.2, it can be proved that $\sup_j \mathbb{E}\left[P_{\theta_j} W(X_j) \mathbf{1}_{A_{q,j}^{(k)}}\right] < \infty$. The proof is concluded since $\sum_k k^{-1/a} < \infty$. □

B.5 Proof of Proposition 3.4

The proof uses a Hoeffding inequality for (non-stationary) Markov chains. The following result is proved in [15, section 5.2, theorem 17].

Proposition B.9. *Let* $(Y_k)_{k\in\mathbb{N}}$ *be a Markov chain on* $(\mathbf{X}, \mathcal{X})$, *with transition kernel* Q *and initial distribution* η. *Assume* Q *is* $\overline{W}$*-uniformly ergodic, and denote by* $\theta_\star$ *its unique invariant distribution. Then there exists a constant* K *such that for any* $t > 0$ *and for any bounded function* $f : \mathbf{X} \to \mathbb{R}$

$$\mathbb{P}\left(\sum_{i=1}^{n} f(Y_i) - n\theta_\star(f) \geq t\right) \leq K\eta(\overline{W}) \exp\left[-\frac{1}{K}\left(\frac{t^2}{n|f|_\infty^2} \wedge \frac{t}{|f|_\infty}\right)\right] \; .$$

Lemma B.10. *Assume that there exists* $\overline{W}$ *such that* $\{Y_n, n \geq 0\}$ *is a* $\overline{W}$*-uniformly ergodic Markov chain with initial distribution* η *with* $\eta(\overline{W}) < \infty$. *Let* $l \in \{1, \cdots, S-1\}$ *and* $p_l \in (0,1)$; *and set* $\xi_l = F_{\theta_\star}^{-1}(p_l)$. *For all* $\epsilon > 0$ *and any* $n \geq 1$,

$$\mathbb{P}\left(|\xi_{\theta_n,l} - \xi_l| > \epsilon\right) \leq 2K\eta(\overline{W}) \exp\left(-\frac{n}{K}\left(\delta_\epsilon^2 \wedge \delta_\epsilon\right)\right) \; ,$$

where $\delta_\epsilon = min\left\{F_{\theta_\star}(\xi_l + \epsilon) - p_l, p_l - F_{\theta_\star}(\xi_l - \epsilon)\right\}$.

Proof. Let $\epsilon > 0$. We write $\mathbb{P}\left(|\xi_{\theta_n,l} - \xi_l| > \epsilon\right) \leq \mathbb{P}\left(\xi_{\theta_n,l} \geq \xi_l + \epsilon\right) + \mathbb{P}\left(\xi_{\theta_n,l} < \xi_l - \epsilon\right)$. Since $F_{\theta_n}(x) \leq t$ iff $x \leq F_{\theta_n}^{-1}(t)$,

$$\begin{aligned}\mathbb{P}\left(\xi_{\theta_n,l} \geq \xi_l + \epsilon\right) &= \mathbb{P}\left(F_{\theta_n}^{-1}(p_l) \geq \xi_l + \epsilon\right) = \mathbb{P}\left(p_l \geq F_{\theta_n}(\xi_l + \epsilon)\right) \\ &= \mathbb{P}\left(\sum_{k=1}^{n} \mathbf{1}_{\{\pi_u(Y_k) > \xi_l + \epsilon\}} \geq n(1-p_l)\right) \; .\end{aligned}$$

Proposition B.9 is then applied with $f(x) = \mathbf{1}_{\{\pi_u(x) > \xi_l + \epsilon\}}$. As

$$\theta_\star(f) = \int \mathbf{1}_{\{\pi_u(x) > \xi_l + \epsilon\}} \theta_\star(dx) = 1 - F_{\theta_\star}(\xi_l + \epsilon) \; ,$$

we obtain

$$\mathbb{P}\left(\xi_{\theta_n,l} \geq \xi_l + \epsilon\right) = \mathbb{P}\left(\sum_{k=1}^{n} f(Y_k) - n\theta_\star(f) \geq n\left(F_{\theta_\star}(\xi_l+\epsilon) - p_l\right)\right)$$
$$\leq K\eta(\overline{W}) \exp\left(-\frac{n}{K}\left[\left(F_{\theta_\star}(\xi_l+\epsilon)-p_l\right)^2 \wedge \left(F_{\theta_\star}(\xi_l+\epsilon)-p_l\right)\right]\right) .$$

for some constant K independent of n, l, ϵ. Similarly,

$$\mathbb{P}\left(\xi_{\theta_n,l} < \xi_l - \epsilon\right) \leq K\eta(\overline{W}) \exp\left(-\frac{n}{K}\left[\left(p_l - F_{\theta_\star}(\xi_l-\epsilon)\right)^2 \wedge \left(p_l - F_{\theta_\star}(\xi_l-\epsilon)\right)\right]\right) ,$$

which concludes the proof. □

Proof of Proposition 3.4 Let $f_{\theta_\star} = F'_{\theta_\star}$ and ϵ_n be defined by

$$\epsilon_n = \frac{2\sqrt{2}}{f_{\theta_\star}(\xi_l)}\sqrt{K}\sqrt{\frac{\log(n)}{n}} ,$$

where K is given by Lemma B.10. Note that under (i), $f_{\theta_\star}(\xi_l) > 0$ since $p_l \in (0,1)$. By (i), $F_{\theta_\star}$ is differentiable and we write when $n \to \infty$

$$F_{\theta_\star}(\xi_l+\epsilon_n) - p_l = F_{\theta_\star}(\xi_l+\epsilon_n) - F_{\theta_\star}(\xi_l) = f_{\theta_\star}(\xi_l)\epsilon_n + o(\epsilon_n) .$$

Hence $F_{\theta_\star}(\xi_l+\epsilon_n) - p_l \geq \sqrt{2K}\sqrt{\frac{\log(n)}{n}}$ for n large enough. Similarly, $p_l - F_{\theta_\star}(\xi_l-\epsilon_n) \geq \sqrt{2K}\sqrt{\frac{\log(n)}{n}}$ for n large enough. So when n is large enough, $nK^{-1}\left(\delta^2_{\epsilon_n} \wedge \delta_{\epsilon_n}\right) \geq 2\log(n)$ with δ_ϵ defined in Lemma B.10. By Lemma B.10, for n large enough, to

$$\mathbb{P}\left(|\xi_{\theta_n,l} - \xi_l| > \epsilon_n\right) \leq \frac{2K\eta(\bar{W})}{n^2} .$$

As $\sum_{n=1}^{\infty} \mathbb{P}\left(|\xi_{\theta_n,l} - \xi_l| > \epsilon_n\right) < \infty$, the Borel-Cantelli lemma yields $\limsup_n \epsilon_n^{-1} |\xi_{\theta_n,l} - \xi_l| < \infty$ w.p.1. This concludes the proof.

References

[1] C. Andrieu, A. Jasra, A. Doucet, and P. Del Moral. Non-linear Markov chain Monte Carlo. In *Conference Oxford sur les méthodes de Monte Carlo séquentielles*, volume 19 of *ESAIM Proc.*, pages 79–84. 2007.

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

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

[4] C. Andrieu and E. Moulines. On the ergodicity property of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006.

[5] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Stat. Comput.*, 18(4):343–373, 2008.

[6] Y. Atchadé, G. Fort, E. Moulines, and P. Priouret. *Adaptive Markov chain Monte Carlo: Theory and Methods*, chapter 2, pages 33–53. Bayesian Time Series Models, Cambridge Univ. Press, 2011.

[7] Y. F. Atchadé and J. S. Liu. Discussion of the "equi-energy sampler" by Kou, Zhou and Wong. *Ann. Statist.*, 34:1620–1628, 2006.

[8] Y. F. Atchadé, G. O. Roberts, and J. S. Rosenthal. Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Stat. Comput.*, 21(4):555–568, 2011.

[9] Timothy L. Bailey and Charles Elkan. Fitting a Mixture Model By Expectation Maximization to Discover Motifs in Biopolymers. In *Second International Conference on Intelligent Systems for Molecular Biology*, 1994.

[10] M. Baragatti, A. Grimaud, and D. Pommeret. Parallel tempering with equi-energy moves. *Stat. Comput.*, pages 1–17, 2012.

[11] B. Bercu, P. Del Moral, and A. Doucet. A Functional Central Limit Theorem for a class of Interacting Markov Chain Monte Carlo Methods. *Electron. J. Probab.*, 14:2130–2155, 2009.

[12] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York, 1968.

[13] A. Brockwell, P. Del Moral, and A. Doucet. Sequentially interacting Markov chain Monte Carlo methods. *Ann. Statist.*, 38(6):3387–3411, 2010.

[14] P. Del Moral and Arnaud Doucet. Interacting Markov Chain Monte Carlo methods for solving nonlinear measure-valued equations. *Ann. Appl. Probab.*, 20(2):593–639, 2010.

[15] R. Douc, E. Moulines, J. Olsson, and R. VanHandel. Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, 39(1):474–513, 2011.

[16] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6):3262–3289, 2012.

[17] G. Fort, E. Moulines, P. Priouret, and P. Vandekerkhove. A simple variance inequality for U-statistics of a Markov chain with applications. Accepted in Stat. Probab. Lett., 2012.

[18] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: Proc. 23rd Symposium on the Interface, Interface Foundation, Fairfax Station, VA*, pages 156–163, 1991.

[19] H. Haario, E. Saksman, and J. Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Statist.*, 14:375–395, 1999.

[20] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57:97–109, 1970.

[21] S. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stoch. Proc. Appl.*, 85(2):341–361, 2000.

[22] S. C. Kou, Q. Zhou, and W. H. Wong. Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Statist.*, 34(4):1581–1619, 2006.

[23] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[24] Charles E. Lawrence and Andrew A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Genetics*, 7(1):41–51, 1990.

[25] J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. *J. Am. Statist. Assoc.*, 90(432):1156–1170, 1995.

[26] X. Liu, Douglas L. Brutlag, and Jun S. Liu. BioProspector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-Expressed Genes. In *Pacific Symposium on Biocomputing*, pages 127–138, 2001.

[27] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[28] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, 1993.

[29] G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.*, 44(2):458–475, 2007.

[30] G.O. Roberts and R.L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

[31] J. S. Rosenthal. *MCMC Handbook*, chapter Optimal Proposal Distributions and Adaptive MCMC. Chapman & Hall/CRC Press, 2009.

[32] J.S. Rosenthal and D.B. Woodard. Convergence rate of Markov chain methods for genomic motif discovery. Under second review at Ann. Statist., 2012.

[33] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature biotechnology*, 16(10):939–945, 1998.

[34] Eero Saksman and Matti Vihola. On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. *Ann. Appl. Probab.*, 20(6):2178–2203, November 11 2010.

[35] A. Schreck, G. Fort, A. Garivier, E. Moulines, and M. Vihola. Convergence of stochastic approximation with discontinuous dynamics. Work in progress, 2012.

[36] G. D. Stormo and G. W. Hartzell. Identifying protein-binding sites from unaligned DNA fragments. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 86, pages 1183–1187, February 1989.