Chapter 1

Adaptive Markov Chain Monte Carlo: Theory and Methods

Yves Atchadé [1], *Gersende Fort and Eric Moulines* [2], *Pierre Priouret* [3]

1.1 Introduction

Markov chain Monte Carlo (MCMC) methods allow to generate samples from an arbitrary distribution π known up to a scaling factor; see Robert and Casella (1999). The method consists in sampling a Markov chain $\{X_k, k \geq 0\}$ on a state space X with transition probability P admitting π as its unique *invariant* distribution, *i.e* $\pi P = \pi$. Such samples can be used *e.g.* to approximate $\pi\left(f\right) \stackrel{\text{def}}{=} \int_{\mathsf{X}} f\left(x\right) \pi\left(dx\right)$ for some π-integrable function $f : \mathsf{X} \to \mathbb{R}$, by

$$\frac{1}{n} \sum_{k=1}^{n} f(X_k) \; . \tag{1.1}$$

In general, the transition probability P of the Markov chain depends on some tuning parameter θ defined on some space Θ which can be either finite dimensional or infinite dimensional. The success of the MCMC procedure depends crucially upon a proper choice of θ. To illustrate, consider the standard Metropolis-Hastings (MH) algorithm. For simplicity, we assume that π has a density also denoted by π with respect to the Lebesgue measure on $\mathsf{X} = \mathbb{R}^d$ endowed with its Borel σ-field $\mathcal{X}$. Given that the chain is at x, a candidate y is sampled from a *proposal transition density* $q(x, \cdot)$ and is accepted with probability $\alpha(x, y)$ defined as

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \frac{q(y, x)}{q(x, y)} \; ,$$

where $a \wedge b \stackrel{\text{def}}{=} \min(a, b)$. Otherwise, the move is rejected and the Markov chain stays at its current location x. The transition kernel P of the associated Markov Chain is reversible with respect to π, *i.e.* for any non-negative measurable function f, $\iint \pi(dx)P(x, dy)f(x, y) = \iint \pi(dx)P(x, dy)f(y, x)$ and therefore admits π as invariant distribution. A commonly used choice for the proposal kernel is the symmetric increment random-walk leading to the Random Walk MH algorithm

[1] University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States.
[2] LTCI, CNRS - Telecom ParisTech, 46 rue Barrault, 75634 Paris Cedex 13, France.
[3] LPMA, Université P. & M. Curie- Boîte courrier 188, 75252 Paris Cedex 05, France.

(hereafter SRWM), in which $q(x, y) = q(y - x)$ for some symmetric proposal distribution q on $\mathbb{R}^d$. The SRWM is frequently used because it is easy to implement and often efficient enough even for complex high-dimensional distributions, provided that the proposal distribution is chosen properly. Finding a good proposal for a particular problem is not necessarily an easy task. A possible choice of the increment distribution q is the multivariate normal with zero-mean and covariance matrix Γ, $\mathcal{N}(0, \Gamma)$, leading to the N-SRWM algorithm. As illustrated in Figure 1.1 in the one-dimensional case $d = 1$, if the variance is either too small or too large, then the convergence rate of the N-SRWM algorithm will be slow and any inference from values drawn from the chain are likely to be unreliable[4]. Intuitively, this may be understood as follows. If the variance is too small, then almost all the proposed values are accepted, and the algorithm behaves almost as a random walk. Because the difference between two successive values are small, the algorithm visits the state space very slowly. On the contrary, if the variance is too large, the proposed moves are nearly all into low-probability area of the target distribution. These proposals are often rejected and the algorithm stays at the same place. Finding a proper scale is thus mandatory. Gelman et al. (1996) have shown that if the target and the proposal distributions are both gaussian, then an appropriate choice for covariance matrix for the N-SRWM is $\Gamma = (2.38^2/d)\Gamma_\pi$, where Γ_π is the covariance matrix of the target distribution. In a particularly large dimensional context ($d \to \infty$), (Roberts and Rosenthal, 2001, Theorem 5) prove that this choice is optimal (see also Gelman et al. (1996) and Roberts et al. (1997)). In practice this covariance matrix Γ is determined by trial and error, using several realizations of the Markov chain. This hand-tuning requires some expertise and can be time-consuming. In order to circumvent this problem, Haario et al. (1999) (see also Haario et al. (2001)) have proposed a novel algorithm, referred to as *adaptive Metropolis*, to update continuously Γ during the run, according to the past values of the simulations. This

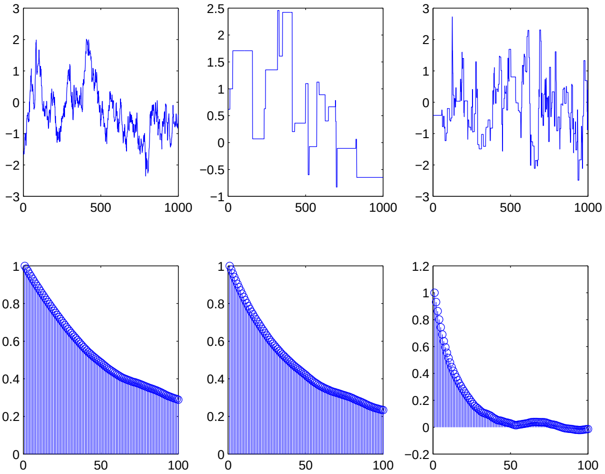

Figure 1.1: The N-SRWM in one dimension.

[4] After J. Rosenthal, this effect is referred to as the *goldilock* principle

algorithm can be summarized as follows,

$$\mu_{k+1} = \mu_k + \frac{1}{k+1}(X_{k+1} - \mu_k) \qquad k \geq 0 \tag{1.2}$$

$$\Gamma_{k+1} = \Gamma_k + \frac{1}{k+1}((X_{k+1} - \mu_k)(X_{k+1} - \mu_k)^{\mathrm{T}} - \Gamma_k) \;, \tag{1.3}$$

X_{k+1} being simulated from the Metropolis kernel with Gaussian proposal distribution $\mathcal{N}\left(0, (2.38^2/d)\Gamma_k\right)$. It was soon recognized by Andrieu and Robert (2001) in an influential technical report that such a scheme can be cast into a more general framework, referred to as *controlled MCMC*. In this framework, the transition kernel of the chain depends upon a finite-dimensional parameter θ, *i.e.* instead of having a *single* transition kernel P with stationary distribution π, we consider a parametric family $\{P_\theta, \theta \in \Theta\}$ each having π as its stationary distribution, $\pi P_\theta = \pi$ for all $\theta \in \Theta$. In a controlled MCMC, the value of the parameter θ is updated online, ideally by optimizing a criterion reflecting the sampler's performance. This problem shares some similarities with Markov decision processes, the choice of the parameter θ can be seen as an *action*. There are some difficulties however when trying to push this similarity too far, because it is not obvious how to define a proper *reward* function. Controlled MCMC is a specific example of an *internal* adaptation setting where the "parameter" θ is updated from the past history of the chain. Other examples of internal adaptation algorithms, which do not necessarily rely on a stochastic approximation step, are given in Section 1.2.2; see also Andrieu and Thoms (2008) and Rosenthal (2009). When attempting to simulate from probability measures with multiple modes or when the dimension of the state-space is large, the Markov kernels $\{P_\theta, \theta \in \Theta\}$ might mix so slowly that an *internal* adaptation strategy cannot always be expected to work. Other forms of adaptation can then be considered, using one or several *auxiliary* processes, which are run in parallel to the chain $\{X_k \;, k \geq 0\}$ targeting π. Because the target chains is adapted using some auxiliary processes, we refer to this adaptation framework as *external*. The idea of running several MCMC in parallels and making them interact has been suggested by many authors (see for example Chauveau and Vandekerkhove (1999), Chauveau and Vandekerkhove (2002), Jasra et al. (2007)). For example, one may sample an instrumental target distribution $\pi^\star$ on a product space X^N (N being the number of MCMC run in parallel) such that $\pi^\star$ admits π as a marginal. The simplest idea consists in defining the target density on the product space X^2 as the (tensor) product of two independent distributions $\pi^\star = \pi \otimes \pi_1$, where π_1 is easier to explore but related to π, and constructing a kernel allowing to *swap* the components of the chain via a so-called *exchange step*. This simple construction has been shown to improve the overall mixing of the chain, by allowing the badly mixing chain to explore the state more efficiently with the help of the auxiliary chain. It is possible to extend significantly this simple idea by allowing more general interactions between the auxiliary chains. In particular, as suggested in Kou et al. (2006), Andrieu et al. (2007) orAtchadé (2009), it is possible to make the auxiliary chains interact with the whole set of past simulations. Instead of allowing to swap only the current states of the auxiliary chains, we may for example exchange the current state of the target chain with one of the states visited by an auxiliary chain in the past. The selection of this state can be guided by sampling in the past of the auxiliary chains, with weights depending for example on the current value of the state. This class of methods are referred to as *interacting MCMC*. These chains can be cast into the framework outlined above, by allowing the parameter θ to take its value in an infinite dimensional space (θ_k might for example be an appropriately specified weighted empirical measure of the auxiliary chains at time k). The purpose of this chapter is to review adaptive MCMC methods, emphasizing the links between internal (controlled MCMC) and external (interacting MCMC) algorithms. The emphasis of this review is to evidence general techniques to construct well-behaved adaptive MCMC algorithms and to prove their convergence. This chapter complements two recent surveys on this topic by Andrieu and Thoms (2008) and Rosenthal (2009) which put more emphasis on the design of adaptive algorithms.

The paper is organized as follows. In Section 1.2.2 the general framework of controlled MCMC is introduced and some examples are given. In Section 1.2.3, interacting MCMC algorithms are presented. In Section 1.2.1, it is shown that these two types of adaptations can be cast into a common unifying framework.

In the context of adaptive MCMC, the convergence of the parameter θ_k is not the central issue; the focus is rather on the way the simulations $\{X_k, k \geq 0\}$ approximate π. The minimal requirements are that, the marginal distribution of X_k converges in an appropriate sense to the stationary distribution π, and that the sample mean $n^{-1} \sum_{k=1}^{n} f(X_k)$, for f chosen in a suitably large class of functions, converges to $\pi(f)$. In Section 1.3 we establish the convergence of the marginal distribution of $\{X_k, k \geq 0\}$ and in Section 1.4, we establish the consistency (*i.e.* a strong law of large numbers) for $n^{-1} \sum_{k=1}^{n} f(X_k)$. Finally and for pedagogical purposes, we show how to apply these results in Section 1.5 to the equi-energy sampler of Kou et al. (2006).

1.2 Algorithms

1.2.1 A general framework for adaptive MCMC

In the sequel, all the variables are defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We let $\{P_\theta, \theta \in \Theta\}$ be a parametric family of Markov kernels on $(\mathsf{X}, \mathcal{X})$. We consider a process $\{(X_n, \theta_n), n \geq 0\}$ and a filtration $\mathcal{F} = \{\mathcal{F}_n, n \geq 0\}$ such that $\{(X_n, \theta_n), n \geq 0\}$ is adapted to $\mathcal{F}$ and for each n, and any non-negative function f,

$$\mathbb{E}\left[f(X_{n+1}) \,|\, \mathcal{F}_n\right] = P_{\theta_n} f(X_n) = \int P_{\theta_n}(X_n, \mathrm{d}y) f(y) \ , \mathbb{P} - \text{a.s.} \tag{1.4}$$

1.2.2 Internal adaptive MCMC

The so-called *internal* adaptive MCMC algorithm corresponds to the case where the parameter θ_n depends on the whole history $X_0, \ldots, X_n, \theta_0, \ldots, \theta_{n-1}$ though in practice it is often the case that the pair process $\{(X_n, \theta_n), n \geq 0\}$ is Markovian. The so-called controlled MCMC algorithm is a specific class of internal adaptive algorithms. According to this scheme, the parameter $\{\theta_k\}$ is updated according to a single step of a stochastic approximation procedure,

$$\theta_{k+1} = \theta_k + \gamma_{k+1} H(\theta_k, X_k, X_{k+1}), \quad k \geq 0 \ , \tag{1.5}$$

where X_{k+1} is sampled from $P_{\theta_k}(X_k, \cdot)$. In most cases, the function H is chosen so that the adaptation is easy to implement, requiring only moderate amounts of extra computer programming, and not adding a large computational overhead (Rosenthal (2007) developed some generic adaptive software). For reasons that will become obvious below, the rate of adaptation tends to zero as the number k of iterations goes to infinity, *i.e.* $\lim_{k \to \infty} \gamma_k = 0$. On the other hand, $\sum_{k=0}^{\infty} \gamma_k = \infty$, meaning that the sum of the parameter moves can still be infinite, *i.e.* the sequence $\{\theta_k, k \geq 0\}$ may move at an infinite distance from the initial value θ_0. It is not necessarily required that the parameters $\{\theta_k, k \geq 0\}$ converge to some fixed value. An in depth description of controlled MCMC algorithms is given in Andrieu and Thoms (2008), where these algorithms are illustrated with many examples (some of which are given below). Instead of adapting the parameter θ_k continuously with smaller and smaller step sizes it is possible to adapt it by batches, which can be computationally less demanding. We may for example define an increasing sequence of time instants $\{T_j, j \geq 0\}$ where adaptation occurs, the parameter being kept constant between two such instants. In this case, the recursion becomes

$$\theta_{T_{j+1}} = \theta_{T_j} + (T_{j+1} - T_j)^{-1} \sum_{k=T_j+1}^{T_{j+1}} H(\theta_{T_j}, X_k, X_{k+1}) \;, \tag{1.6}$$

where X_{k+1} is sampled from $P_{\theta_k}(X_k, \cdot)$, $\theta_k = \theta_{T_j}$ for $k \in \{T_j + 1, \ldots, T_{j+1} - 1\}$. The rate at which the adaptation takes place is diminished by taking $\lim_{j\to\infty} T_{j+1} - T_j = +\infty$. Instead of selecting deterministic time intervals, it is also possible to adapt the parameter at time k with a probability p_k, with $\lim_{k\to\infty} p_k = 0$ (see Roberts and Rosenthal (2007)). The usefulness of this extra bit of randomness is somehow questionable. The recursions (1.5) or (1.6) is aimed at solving the equation $h(\theta) = 0$ where $\theta \mapsto h(\theta)$ is the *mean field* of the stochastic approximation procedure (see for example Benveniste et al. (1990) or Kushner and Yin (2003)),

$$h(\theta) \overset{\text{def}}{=} \int_{\mathsf{X}} H(\theta, x, x')\pi(dx)P_\theta(x, dx') \;. \tag{1.7}$$

Returning to the adaptive Metropolis example in the introduction, the parameter θ is equal to the mean and the covariance matrix of the multivariate distribution, $\theta = (\mu, \Gamma) \in \Theta = (\mathbb{R}^d, \mathcal{C}_+^d)$, where $\mathcal{C}_+^d$ is the cone of symmetric non-negative $d \times d$ matrices. The function expression of H is explicitly given in Eqs. (1.2) and (1.3). Assuming that $\int_{\mathsf{X}} |x|^2 \pi(dx) < \infty$, one can easily check that the associated mean field function is given by

$$h(\theta) = \int_{\mathsf{X}} H(\theta, x)\pi(dx) = (\mu_\pi - \mu, (\mu_\pi - \mu)(\mu_\pi - \mu)^{\mathsf{T}} + \Gamma_\pi - \Gamma)^{\mathsf{T}} \;, \tag{1.8}$$

with μ_π and Γ_π the mean and covariance of the target distribution,

$$\mu_\pi \overset{\text{def}}{=} \int_{\mathsf{X}} x \; \pi(dx) \quad \text{and} \quad \Gamma_\pi \overset{\text{def}}{=} \int_{\mathsf{X}} (x - \mu_\pi)(x - \mu_\pi)^{\mathsf{T}} \pi(dx) \;. \tag{1.9}$$

The stationary point of the algorithm is unique $\theta_\star = (\mu_\pi, \Gamma_\pi)$. Provided that the step size is appropriately chosen, a stochastic approximation procedure will typically converge toward that stationary point. The convergence is in general nevertheless not trivial to establish; see for example Andrieu et al. (2005), Andrieu and Moulines (2006). The behavior of the Adaptive Metropolis (AM) algorithm is illustrated in figure 1.2. The target distribution is a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ with dimension $d = 200$. The eigenvalues of the covariance of the target distribution are regularly spaced in the interval $[10^{-2}, 10^3]$. This is a rather challenging simulation tasks, with a rather large dispersion of the eigenvalues. The proposal distribution at step $k \geq 2d$ is given by

$$P_{\theta_k}(x, \cdot) = (1 - \beta)\mathcal{N}(x, (2.38)^2 \Gamma_k / d) + \beta \mathcal{N}(x, 0.1 * \mathrm{Id}/d) \;,$$

where Γ_k is the current estimate of the covariance matrix given in (1.3) and β is a positive constant (we take $\beta = 0.05$, as suggested in Roberts and Rosenthal

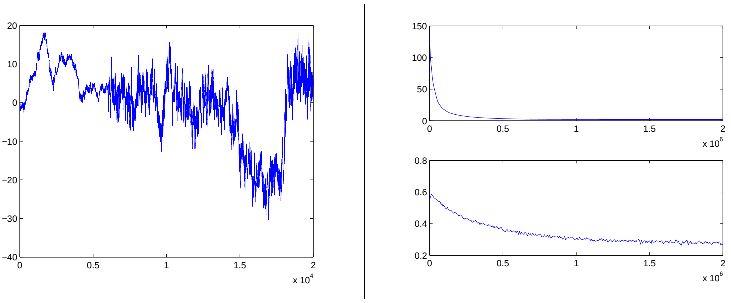

Figure 1.2: Left Panel: trace plot of the 20000 first iterations of the AM algorithm. Right Panel, Top: suboptimality factor as a function of the number of iterations; Right Panel, Bottom: mean acceptance rate as a function of the number of iterations (obtained by averaging the number of accepted moves on a sliding window of size 6000)

(2006)). The initial value for Γ_0 is $(2.38/\sqrt{d})\mathrm{Id}$. The rationale of using such β is to avoid the algorithm being stuck with a singular covariance matrix (in the original AM algorithm, Haario et al. (2001) suggested to regularize the covariance matrix by loading the diagonal; another more sophisticated solution, based on projections onto compact sets is considered in Andrieu and Moulines (2006)). Figure 1.2 displays the trace plot of the first coordinate of the chain for dimension $d = 200$ together with the suboptimality criterion introduced in Roberts and Rosenthal (2001), defined as

$$b_k \stackrel{\text{def}}{=} d \frac{\sum_{i=1}^{d} \lambda_{i,k}^{-2}}{\left(\sum_{i=1}^{d} \lambda_{i,k}^{-1}\right)^2}$$

where $\lambda_{i,k}$ are the eigenvalues of the matrix $\Gamma_k^{1/2}\Sigma^{-1/2}$. Usually we will have $b_k > 1$, and the closer b_k is to 1, the better. The criterion being optimized in AM is therefore b_k^{-1}. Another algorithm which benefits from a strong theoretical background, is the expected acceptance probability (*i.e.* the fraction of proposed moves which are accepted) of the MH algorithm for random walk Metropolis algorithms or Langevin based MH updates, Roberts et al. (1997). As shown for SRWM algorithm in one dimension, if the expected acceptance probability is (very) close to 1, this suggests that scale is too small. If this fraction is (very) close to 0, this suggests that it is too large (as in Figure 1.1). But if this fraction is far from 0 and far from 1, then we may consider that the algorithm avoids both extremes and is properly tune. The choice of a proper scale can be automated by controlling the expected acceptance ratio. For simplicity, we consider only the one-dimensional SRWM, where θ is the scale. In this case, we choose $H(x, x') \stackrel{\text{def}}{=} \{\alpha(x, x') - \alpha_\star\}$ which is associated to the mean field $h(\theta) = \int \pi(dx)P_\theta(x, dx')\alpha(x, x') - \alpha_\star$, where the acceptance ratio is $\alpha(x, x') = 1 \wedge \pi(x')/\pi(x)$. The value of $\alpha_\star$ can be set to 0.4 or 0.5 (the "optimal" value in this context is 0.44). The stationary acceptance probability has several advantages. The same idea applies in large-dimensional context. We may for example couple this approach with the AM algorithm: instead of using the asymptotic $(2.38)^2/d$ factor, we might let the algorithm determine automatically a proper scaling by controlling both the covariance matrix of the proposal distribution and the mean acceptance rate. Consider first the case where the target distribution and the proposal distribution are Gaussian with identity covariance matrix. We learn the proper scaling by targeting a mean acceptance rate $\alpha_\star = 0.234$, which is known to be optimal in a particular large dimensional context. To illustrate the behavior of the algorithm, we display in Figure 1.3 the

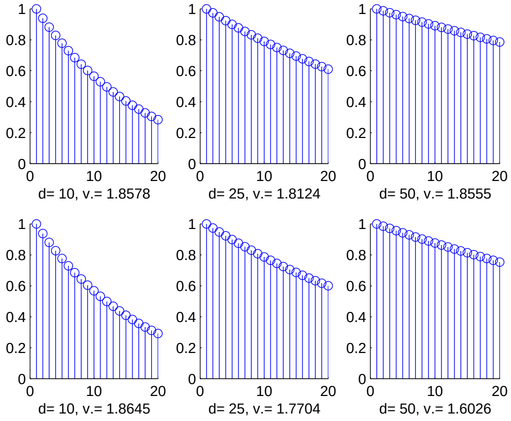

Figure 1.3: Autocorrelation function for $g(x^1, \ldots, x^d) = x^1$ and value of the integrated correlation v for different dimensions $d = 10, 25, 50$; in the top panels, the optimal scaling is used. In the bottom panels, the scaling is adapted by controlling the mean acceptance rate

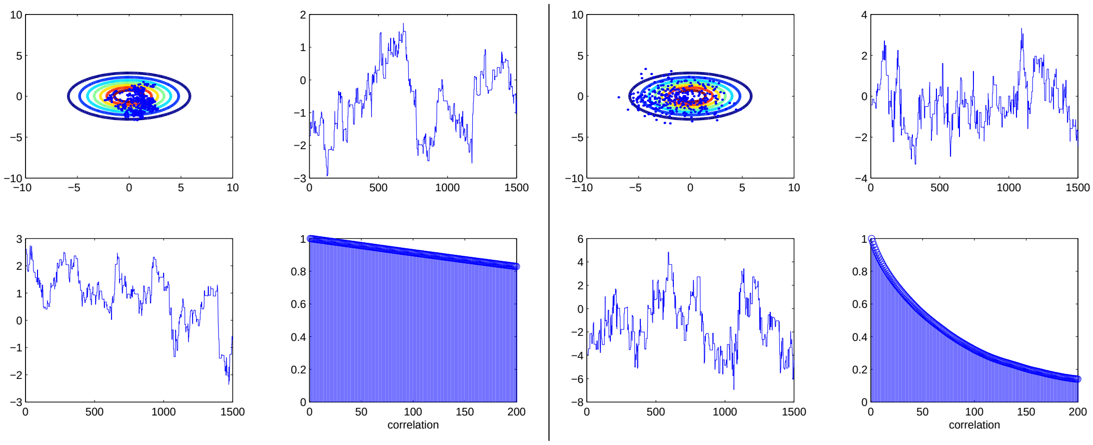

Figure 1.4: Left panel: SRWM algorithm with identity covariance matrix with optimal scaling; Right panel: Adaptive Metropolis with adapted scaling; the targeted value of the mean acceptance rate is $\alpha_\star = 0.234$

autocorrelation function $\{\mathrm{Corr}_\pi\left[g(X_0), g(X_k)\right], k \geq 0\}$ and the integrated autocorrelation time $1 + 2\sum_{k=0}^{\infty} \mathrm{Corr}_\pi(g(X_0), g(X_k))$ as a function of the dimension d for the function $g(x^1, \ldots, x^d) = x^1$ (these quantities are estimated using a single run of length 50000). In another experiment, we try to learn simultaneously the covariance and the scaling. The target distribution is zero-mean Gaussian with covariance drawn at random as above in dimension $d = 100$. Another possibility consists in using a Metropolis-within-Gibbs algorithm, where the variables are updated one at a time (in either systematic or random order), each using a MCMC algorithm with a one-dimensional proposal (see Roberts and Rosenthal (2006); Rosenthal (2009); this algorithm is also closely related to the Single-Component Adaptive Metropolis introduced in Haario et al. (2005)). It has recently been applied successfully to high-dimensional inference for statistical genetics (Turro et al., 2007). More sophisticated adaptation techniques may be used. An obvious idea is to try to make the adaptation "local", *i.e.* to adapt to the local behavior of the target density. Such techniques have been introduced to alleviate the main weakness of the adaptive Metropolis algorithm when applied to spatially non-homogeneous target which is due to the use of a single global covariance distribution for the proposal. Consider for example the case where the target density is a mixture of Gaussian distributions, $\pi = \sum_{j=1}^{p} a_j \mathcal{N}(\mu_j, \Sigma_j)$. Provided that the overlap between the components is weak, and the covariances Σ_j are widely different, then there does not exist a common proposal distribution which is well-fitted to sample in the regions surrounding each mode. This example suggests to tune the empirical covariance matrices by learning the *history* of the past simulations in different regions of the state space. To be more specific, assume that there exists a partition: $\mathsf{X} = \bigcup_{j=1}^{p} \mathsf{X}_j$. Then, according to the discussion above, it seems beneficial to use different proposal distributions in each set of the partition. We might for example use the proposal :

$$q_\theta(x; x') = \sum_{j=1}^{p} \mathbb{1}_{\mathsf{X}_j}(x) \phi(x'; x, \Gamma_j) ,$$

where $\mathbb{1}_A(x)$ is the indicator of the set A, and $\phi(x; \mu, \Gamma)$ is the density of a d-dimensional Gaussian distribution with mean μ and covariance Γ. Here, the parameter θ collects the covariances of the individual proposal distributions within each region. With such a proposal, the acceptance ratio of MH algorithm becomes:

$$\alpha_\theta(x; x') = 1 \wedge \sum_{i,j=1}^{p} \frac{\pi(x')}{\pi(x)} \frac{\phi(x; x', \Gamma_j)}{\phi(x'; x, \Gamma_i)} \mathbb{1}_{\Gamma_i}(x) \mathbb{1}_{\Gamma_j}(x') .$$

To adapt the covariance matrices $\{\Gamma_i, i = 1, \ldots, p\}$ we can for example use the update equations (1.2) and (1.3) within each region. To ensure a proper communication between the regions, it is recommended to mix the adaptive kernel with a fixed kernel. This technique is investigated in Craiu et al. (2008). Another interesting direction of research is to adapt the proposal distribution in an independent MH setting. Recall that in this case, the proposed moves do not depend on the current state, *i.e.* $q(x, x') = q(x')$ where q is some probability distribution on X. The acceptance ratio writes in such case $\alpha(x, x') = 1 \wedge (\pi(x')q(x)/\pi(x)q(x'))$. To perform adequately, the proposal q should be chosen sufficiently close to the target π. A natural idea is to choose a parametric family of distribution $\{q_\theta, \theta \in \Theta\}$, and to adapt the parameter θ from the history of the draws. This technique is of course closely related to the adaptive importance sampling idea; see for example Rubinstein and Kroese (2008). Because of their extreme flexibility and their tractability, a finite mixture of Gaussian distributions is an appealing choice for that purpose. Recall that any continuous distribution can be approximated arbitrary well by a finite mixture of normal densities with common covariance matrix; in addition, discrete mixture of Gaussians are fast to sample from, and their likelihood is easy to calculate, which is of key importance. In this case, the parameters θ are the mixing proportions and the mean and covariances of the component densities. Several approaches have been considered to fit these parameters. Other mixture from the exponential family can also be considered, such as a discrete/continuous mixture of Student's t-distribution (see Andrieu and Thoms (2008) for details). Andrieu and Moulines (2006) suggest to fit these parameters using a maximum likelihood approach or, equivalently, by maximizing the cross-entropy between the proposal distribution and the target; this algorithm shares some similarities with the so-called adaptive independence sampler developed in Keith et al. (2008). In this framework, the parameters are fitted using a sequential version of the EM algorithm (Cappe and

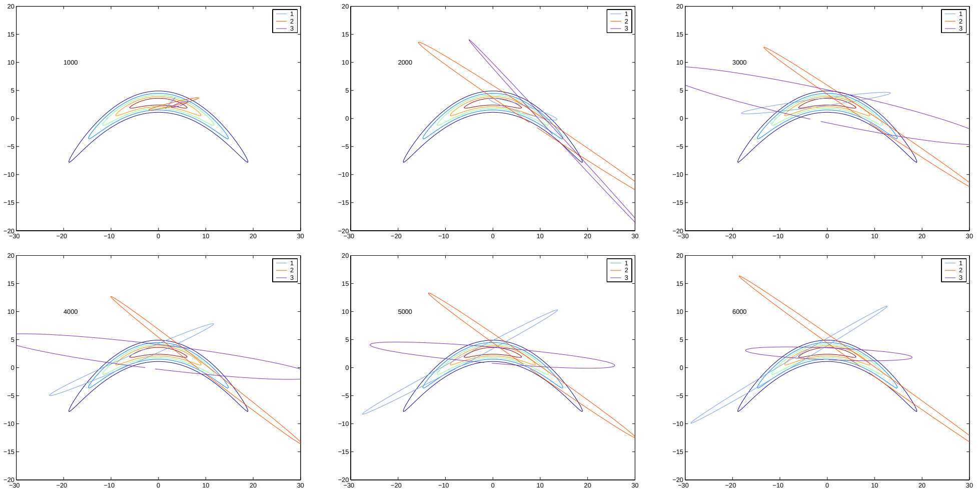

Figure 1.5: Adaptive fit of a mixture of three Gaussian distributions with arbitrary means and covariance using the maximum likelihood approach developed in Andrieu and Moulines (2006)

Moulines, 2009) (several improvements on this basic schemes are presented in Andrieu and Thoms (2008)). Giordani and Kohn (2008) proposed a principled version replacing the EM by the k-harmonic mean, an extension of the k-means algorithm that allows for soft membership. This algorithm is less sensitive to convergence to local minima; in addition degeneracies of the covariance matrices of the components can be easily prevented. An example of fit is given in Figure 1.5. There are many other possible forms of adaptive MCMC and many of these are presented in the recent surveys by Roberts and Rosenthal (2006), Andrieu and Thoms (2008) and, Rosenthal (2009).

1.2.3 External Adaptive Algorithm

The so-called *external* adaptive MCMC algorithms correspond to the case where the parameter process $\{\theta_n, n \geq 0\}$ is computed using an auxiliary process $\{Y_k\ , k \geq 0\}$ run independently from the process $\{X_k\ , k \geq 0\}$ (several auxiliary processes can be of course used). More precisely, it is assumed that the process is adapted to the natural filtration of the process $\{Y_k\ , k \geq 0\}$, meaning that for each n, θ_n is a function of the history $Y_{0:n} \stackrel{\mathrm{def}}{=} (Y_0, Y_1, \ldots, Y_n)$ of the auxiliary process. In addition, conditionally to the auxiliary process $\{Y_k\ , k \geq 0\}$, $\{X_k\ , k \geq 0\}$ is an inhomogeneous Markov Chain, for any $0 \leq k \leq n$, and any bounded function f,

$$\mathbb{E}\left[f(X_{k+1}) \,|\, X_{0:k}, Y_{0:k}\right] = P_{\theta_k} f(X_k)\ .$$

The use of an auxiliary process to learn the proposal distribution in an independent MH algorithm has been considered in Chauveau and Vandekerkhove (2001) and Chauveau and Vandekerkhove (2002). In this setting θ_n is a distribution (and is thus non parametric), and is obtained using an histogram. This approach works best in situations where the dimension of the state space d is small, otherwise, the histogram estimation becomes very unreliable (in Chauveau and Vandekerkhove (2002) only one and two-dimensional examples are considered).

(Kou et al., 2006) have introduced another form of interaction: instead of trying to learn a well fitted proposal distribution, these authors suggest to ”swap” the current state of the Markov Chain with a state sampled from the history of the auxiliary processes. A similar idea has been advocated in Atchadé (2009). In this setting, the "parameter" θ_n is also infinite dimensional, and these algorithms may be seen as a kind of *non-parametric* extensions of the controlled MCMC procedures.

All these new ideas originate from *parallel tempering* and *simulated tempering*, two influential algorithms developed in the early 1990's to speed up the convergence of MCMC algorithms (Geyer (1991); Marinari and Parisi (1992); Geyer and Thompson (1995)). In these approaches, the sampling algorithm moves progressively to the target distribution π through a sequence of "easily sampled" distributions. The idea behind *parallel tempering* algorithm by Geyer (1991) is to perform parallel Metropolis sampling at different *temperatures*. Occasionally, a *swap* between the states of two neighboring chains (two chains running at adjacent temperature levels) is proposed. The acceptance probability for the swap is computed to ensure that the joint states of all the parallel chains evolve according to the Metropolis-Hastings rule targeting the product distribution. The objective of the parallel tempering is to use the faster mixing of the high temperature chains to improve the mixing of the low temperature chains. The *simulated tempering algorithm* introduced in Marinari and Parisi (1992) exploits a similar idea but using a markedly different approach. Instead of using multiple parallel chains, this algorithm runs a single chain but augments the state of this chain by an auxiliary variable, the temperature, that is dynamically moved up or down the temperature ladder.

The Equi-Energy (EE) sampler exploits the parallel tempering idea, in the sense that the algorithm runs several chains at different temperatures, but allows for more general swaps between states of the neighboring chains. The idea is to replace an instantaneous swap by a so-called *equi-energy* swap. To avoid cumbersome notations, we assume here that there is a single auxiliary process, but it should be stressed that the EE sampler works best in presence of multiple auxiliary processes, covering a wide range of temperatures. Let π be the target density distribution on $(\mathsf{X}, \mathcal{X})$. For $\beta \in (0, 1)$ define the tempered density $\tilde{\pi} \propto \pi^\beta$. The auxiliary process $\{Y_n, n \geq 0\}$ is X-valued and is such that its marginal distribution converges as n goes to infinity to $\tilde{\pi}$. Let P be a transition kernel on $(\mathsf{X}, \mathcal{X})$ with unique invariant distribution π (in most cases, P is a MH kernel). Choose $\epsilon \in (0, 1)$, which is the probability of proposing a swap between the states of two neighboring chains. Define a partition $\mathsf{X} = \bigcup_{\ell=1}^K \mathsf{X}_\ell$, where X_ℓ are the so-called *rings* (a term linked with the particular choice of the partition in Kou et al. (2006), which is defined as the level set of the logarithm of the target distribution). At iteration n of the algorithm, two actions may be taken:

1. with probability $(1-\epsilon)$, we move the current state X_n according to the Markov kernel P,

2. with probability ϵ, we propose to swap the current state X_n with a state Z drawn from the past of the auxiliary process with weights proportional to $\{g(X_n, Y_i), i \leq n\}$, where

$$g(x, y) \stackrel{\text{def}}{=} \sum_{\ell=1}^{K} \mathbb{1}_{\mathsf{X}_\ell \times \mathsf{X}_\ell}(x, y) \; . \tag{1.10}$$

   More precisely, we propose a move Z at random within the same ring as X_n. This move is accepted with probability $\alpha(X_n, Z)$, where the acceptance probability α is defined by $\alpha : \mathsf{X} \times \mathsf{X} \to [0, 1]$ defined by

$$\alpha(x, y) \stackrel{\text{def}}{=} 1 \wedge \left(\frac{\pi(y)}{\tilde{\pi}(y)} \left[\frac{\pi(x)}{\tilde{\pi}(x)} \right]^{-1} \right) = 1 \wedge \frac{\pi^{1-\beta}(y)}{\pi^{1-\beta}(x)} \; . \tag{1.11}$$

More formally, let Θ be the set of the probability measures on $(\mathsf{X}, \mathcal{X})$. For any distribution $\theta \in \Theta$, define the Markov transition kernel

$$P_\theta(x, \cdot) \stackrel{\text{def}}{=} (1 - \epsilon_\theta(x))P(x, A) + \epsilon_\theta(x)K_\theta(x, A) \tag{1.12}$$

with

$$K_\theta(x, A) \stackrel{\text{def}}{=} \int_A \alpha(x, y)\frac{g(x, y)\theta(dy)}{\theta[g(x, \cdot)]} + \mathbb{1}_A(x)\int \{1 - \alpha(x, y)\}\,\frac{g(x, y)\theta(dy)}{\theta[g(x, \cdot)]}\ , \tag{1.13}$$

where $\theta[g(x, \cdot)] \stackrel{\text{def}}{=} \int g(x, y)\theta(dy)$ and

$$\epsilon_\theta(x) = \epsilon \mathbb{1}_{\theta[g(x,\cdot)]>0}\ . \tag{1.14}$$

The kernel $K_{\tilde{\pi}}$ can be seen as a Hastings-Metropolis kernel with proposal kernel $g(x, y)\tilde{\pi}(y)/\int g(x, y)\tilde{\pi}(dy)$ and target distribution π. Hence, $\pi P_{\tilde{\pi}} = \pi$.

Using the samples drawn from this process, a family of weighted empirical probability distributions $\{\theta_n, n \geq 0\}$ is recursively constructed as follows:

$$\theta_n \stackrel{\text{def}}{=} \frac{1}{n+1}\sum_{j=0}^{n} \delta_{Y_j} = \left(1 - \frac{1}{n+1}\right)\theta_{n-1} + \frac{1}{n+1}\delta_{Y_n}\ , \tag{1.15}$$

where, δ_y is the Dirac mass at y and by convention, $\theta_{-1} = 0$. Given the current value X_n and the sequence $\{\theta_k, k \leq n\}$, X_{n+1} is obtained by sampling the kernel $P_{\theta_n}(X_n, \cdot)$.

Figure 1.6 illustrates the effectiveness of equi-energy sampler. In that example, the target distribution is the two-dimensional Gaussian mixture introduced in (Kou et al., 2006, p. 1591-1592) Figure 1.6(a) represents independent sample points from the target distribution. This distribution has 20 well separated modes (most local modes are more than 15 standard deviations away from the nearest ones) and poses a serious challenge for sampling algorithms. We test a plain SRWM, parallel tempering and the equi-energy sampler on this problem. For parallel tempering and the equi-energy sampler, we use 5 parallel chains at inverse-temperature $\beta = 1, 0.36,\ 0.13,\ 0.05,\ 0.02$. For the equi-energy sampling, we define 5 equi-energy rings $\mathsf{X}_1 = \{x \in \mathbb{R}^2 :\ -\log \pi(x) < 2\}$, $\mathsf{X}_2 = \{x \in \mathbb{R}^2 :\ 2 \leq -\log \pi(x) < 6.3\}$, $\mathsf{X}_3 = \{x \in \mathbb{R}^2 :\ 6.3 \leq -\log \pi(x) < 20\}$, $\mathsf{X}_4 = \{x \in \mathbb{R}^2 :\ 20 \leq -\log \pi(x) < 63.2\}$ and $\mathsf{X}_5 = \{x \in \mathbb{R}^2 :\ -\log \pi(x) \geq 63.2\}$. Figure 1.6 (b) plots the first $2,000$ iterations from the EE sampler; (c) plots the first $2,000$ iterations from parallel tempering and (d) plots the first $2,000$ iterations from a plain SRWM algorithm. The plain SRWM has a very poor mixing for this example. Parallel tempering mixes better but the equi-energy sampler mixes even faster.

1.3 Convergence of the marginal distribution

There is a difficulty with either the internal or external adaptation procedures: as the parameter estimate θ_k depends on the whole past either of the process or the auxiliary process, the process $\{X_k, k \geq 0\}$ is no longer a Markov chain and classical convergence results do not hold. This may cause serious problems, as illustrated in this naive example. Let $\mathsf{X} = \{1, 2\}$ and consider for $\theta, t_1, t_2 \in \Theta = (0, 1)$, with $t_1 \neq t_2$, the following Markov transition probability matrices

$$P_\theta = \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix} \qquad \tilde{P} = \begin{bmatrix} 1-t_1 & t_1 \\ t_2 & 1-t_2 \end{bmatrix}\ .$$

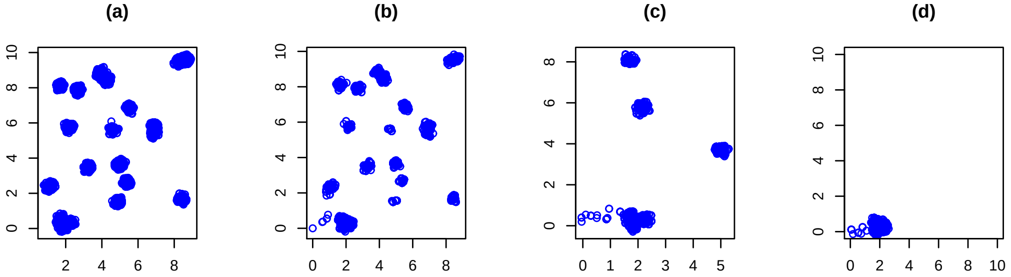

Figure 1.6: Comparison of Equi-Energy sampler (b), Parallel tempering (c) and a Plain SRWM (d). (a) represents the target distribution.

For any $\theta \in \Theta$, $\pi = (1/2, 1/2)$ satisfies $\pi P_\theta = \pi$. However if we let θ_k be a given function $\Xi : \mathsf{X} \to (0,1)$ of the current state, *i.e.* $\theta_k = \Xi(X_k)$ with $\Xi(i) = t_i$, $i \in \{1,2\}$, one defines a new Markov chain with transition probability $\tilde{P}$. Now $\tilde{P}$ has $[t_2/(t_1+t_2), t_1/(t_1+t_2)]$ as invariant distribution. Instead of improving the mixing, the adaptation has in such case destroyed the convergence to π.

The first, admittedly weak, requirement for an MCMC algorithm to be well-behaved is that the distribution of X_n converges weakly to the target distribution π as $n \to \infty$, *i.e.* for any measurable bounded function $f : \mathsf{X} \to \mathbb{R}$, $\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = \pi(f)$. When this limit holds uniformly for any bounded function $f : \mathsf{X} \to \mathbb{R}$, this convergence of the marginal $\{X_n, n \geq 0\}$ is also referred to in the literature as the “ergodicity” of the marginal $\{X_n, n \geq 0\}$ (see e.g. Roberts and Rosenthal (2007)). Convergence of the marginal distributions has been obtained for a general class of adaptive algorithm (allowing ”extra” randomization) by Roberts and Rosenthal (2007) and for a more restricted class of adaptive algorithms by Andrieu and Moulines (2006); Atchadé and Fort (To appear). The analysis of the interacting MCMC algorithm began more recently and the theory is still less developed. In this Section, we extend the results obtained in Roberts and Rosenthal (2007) to cover the case where the kernels $\{P_\theta, \theta \in \Theta\}$ do not all have the same invariant distribution π.

All the proofs of this section are postponed in Section 1.6.

1.3.1 Notations

For a bounded function f, the supremum norm is denoted by $|f|_\infty$. Let $\mathsf{b}\mathcal{X}$ be the set of measurable functions from X to $\mathbb{R}$ which are bounded by 1: $|f|_\infty \leq 1$.

More generally, for a function $V : \mathsf{X} \to [1, +\infty)$, we define the V-norm of a function $f : \mathsf{X} \to \mathbb{R}$ by $|f|_{\mathrm{V}} \overset{\text{def}}{=} \sup_{\mathsf{X}} |f| V^{-1}$. Let $\mathcal{L}_V$ be the set of the functions with finite V-norm.

For a signed measure μ on $(\mathsf{X}, \mathcal{X})$, we define the total variation norm and the V-norm by $\|\mu\|_{\mathrm{TV}} \overset{\text{def}}{=} \sup_{\{f : |f|_\infty \leq 1\}} |\mu(f)|$ and $\|\mu\|_{\mathrm{V}} \overset{\text{def}}{=} \sup_{\{f : |f|_V \leq 1\}} |\mu(f)|$. We denote by $\xrightarrow{\text{a.s.}}$ and $\xrightarrow{\mathrm{P}}$ and $\xrightarrow{\mathcal{D}}$, the a.s. convergence and the convergence in probability.

1.3.2 Main result

Consider the following assumptions

A1 For any $\theta \in \Theta$, P_θ is a transition kernel on $(\mathsf{X}, \mathcal{X})$ that possesses an unique invariant probability measure π_θ.

For any $\theta, \theta' \in \Theta$, define

$$D_{\mathrm{TV}}(\theta, \theta') \stackrel{\mathrm{def}}{=} \sup_{x \in \mathsf{X}} \| P_\theta(x, \cdot) - P_{\theta'}(x, \cdot) \|_{\mathrm{TV}} \ . \tag{1.16}$$

A2 *(Diminishing Adaptation)* $D_{\mathrm{TV}}(\theta_n, \theta_{n-1}) \xrightarrow{\mathrm{P}} 0$.

The diminishing adaptation A2 condition is fundamental. It requires that the amount of change in the parameter value at the nth iteration vanishes as $n \to \infty$. Note that it is not necessarily required that the parameter sequence $\{\theta_n, n \geq 0\}$ converges to some fixed value. Since the user controls the updating scheme, assumption A2 can be ensured by an appropriate planning of the adaptation. For example, if the algorithm adapts only at the successive time instants $\{T_n, n \geq 0\}$, (or is adapted at iteration n with probability $p(n)$), then A2 is automatically satisfied if $\lim_{n\to\infty} T_{n+1} - T_n = \infty$ (or if $\lim_{n\to\infty} p(n) = 0$).

One interesting consequence of (1.4) and A2 is that, for any integer $N \geq 1$, the conditional expectation $\mathbb{E}\left[f(X_{n+N}) \,|\, \mathcal{F}_n\right]$ behaves like $P^N_{\theta_n} f(X_n)$, *i.e.* the conditional expectation is as if the adaptation was frozen at time n during the N successive time steps.

Proposition 1.3.1. *For any integers $n, N > 0$*

$$\sup_{\{f:\, f \in \mathsf{b}\mathcal{X}\}} \left| \mathbb{E}\left[f(X_{n+N}) \,|\, \mathcal{F}_n\right] - P^N_{\theta_n} f(X_n) \right| \leq \sum_{j=1}^{N-1} \mathbb{E}\left[D_{\mathrm{TV}}(\theta_{n+j}, \theta_n) \,|\, \mathcal{F}_n \right] \ , \quad \mathbb{P}-a.s.$$

For $x \in \mathsf{X}$, $\theta \in \Theta$ and for any $\epsilon > 0$, define

$$M_\epsilon(x, \theta) \stackrel{\mathrm{def}}{=} \inf\{n \geq 0, \|P^n_\theta(x, \cdot) - \pi_\theta\|_{\mathrm{TV}} \leq \epsilon\} \ .$$

A3 *(Containment Condition)* For any $\epsilon > 0$, the sequence $\{M_\epsilon(X_n, \theta_n), n \geq 0\}$ is bounded in probability, *i.e.* $\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}(M_\epsilon(X_n, \theta_n) \geq M) = 0$.

The containment condition A3 expresses a kind of uniform-in-θ ergodicity of the kernels $\{P_\theta, \theta \in \Theta\}$. We show indeed in Proposition 1.3.4 below that a sufficient condition for A3 is a uniform-in-θ geometric drift inequality which is known to imply a uniform-in-θ ergodicity $\lim_n \sup_\theta \|P^n_\theta(x, \cdot) - \pi_\theta\|_{\mathrm{TV}} = 0$ (see e.g. Lemma 1.7.1(iii)). As discussed in Atchadé and Fort (To appear) (see also Bai (2008)), this uniform-in-θ geometric drift inequality could be weakened and replaced by a uniform-in-θ sub-geometric drift inequality. The containment condition A3 is thus satisfied for many sensible adaptive schemes. It holds in particular for adaptive SRWM and Metropolis-within-Gibbs algorithms under very general conditions (Bai et al., 2009). It is, however, possible to construct pathological counter-examples, where containment does not hold.

Theorem 1.3.2 establishes the convergence of the marginal distribution:

Theorem 1.3.2. *Assume A1, A2 and A3. Then for any $\epsilon > 0$, there exist n_0, N, $n_0 \geq N$ such that for any $n \geq n_0$*

$$\sup_{\{f: f \in \mathsf{b}\mathcal{X}\}} \left| \mathbb{E}\left[f(X_n) - \pi_{\theta_{n-N}}(f) \right] \right| \leq \epsilon \ .$$

When $\pi_\theta = \pi_\star$ for any $\theta \in \Theta$, this result shows the convergence of the marginal distribution to $\pi_\star$; see (Roberts and Rosenthal, 2007, Theorem 13). The fact that some of the conditions A1, A2 and A3 are also necessary has been discussed in the case the transition kernels $\{P_\theta, \theta \in \Theta\}$ have the same invariant probability distribution $\pi_\star$. In particular, (Bai, 2008, Theorem 4.1) shows that if the transition kernels $\{P_\theta, \theta \in \Theta\}$ are ergodic, uniformly-in-θ, and the marginal $\{X_k, k \geq 0\}$ converges to $\pi_\star$ then the containment condition holds.

Theorem 1.3.3 establishes the convergence of $\{\pi_{\theta_n}, n \geq 0\}$ under the additional assumption:

A4 There exist $\mathsf{F} \subseteq \mathsf{b}\mathcal{X}$ and a probability measure $\pi_\star$ on $(\mathsf{X}, \mathcal{X})$ such that, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{\{f:f\in\mathsf{F}\}} \mathbb{P}\left(|\pi_{\theta_n}(f) - \pi_\star(f)| \geq \epsilon\right) = 0 \; .$$

Theorem 1.3.3. *Assume A1, A2, A3 and A4. Then*

$$\lim_{n\to\infty} \sup_{\{f:f\in\mathsf{F}\}} \left|\mathbb{E}\left[f(X_n)\right] - \pi_\star\left(f\right)\right| = 0 \; .$$

1.3.3 Sufficient conditions for A1, A2, A3 and A4

To check the existence of an invariant probability measure π_θ for each transition kernel P_θ (assumption A1) and the containment condition (assumption A3), we provide conditions in terms, essentially, of a uniform-in-θ drift condition and a uniform-in-θ minorization condition. For the convergence of the random sequence of the expectations $\{\pi_{\theta_n}(f), n \geq 0\}$ to $\pi_\star(f)$ (assumption A4), we provide a sufficient condition in terms of almost-sure convergence of the kernels $\{P_{\theta_n}, n \geq 0\}$, which implies the almost-sure convergence of $\{\pi_{\theta_n}(f), n \geq 0\}$ for any fixed bounded function f.

B1 (a) For any $\theta \in \Theta$, P_θ is phi-irreducible.

(b) There exist a measurable function $V : \mathsf{X} \to [1, +\infty)$, constants $b \in (0, +\infty)$, $\lambda \in (0, 1)$ and a set $\mathcal{C} \in \mathcal{X}$ such that for any $\theta \in \Theta$,

$$P_\theta V \leq \lambda V + b \, \mathbb{1}_\mathcal{C} \; .$$

(c) The level sets of V are 1-small uniformly in θ *i.e.* for any $v \geq 1$, there exist a constant $\varepsilon > 0$ and a probability measure ν such that for any $\theta \in \Theta$, $P_\theta(x, \cdot) \geq \varepsilon \; \mathbb{1}_{\{V \leq v\}}(x) \; \nu$.

Note that in B1b, we can assume without loss of generality (and we do so) that $\sup_\mathcal{C} V < +\infty$. Note also that B1 implies that each transition kernel P_θ is aperiodic.

Proposition 1.3.4. *(i) Assume B1. For any $\theta \in \Theta$, the kernel P_θ is positive recurrent with invariant probability π_θ and $\sup_\theta \pi_\theta(V) \leq (1-\lambda)^{-1} b$.*

(ii) Under B1b, the containment condition A3 holds.

As shown in the counter-example in the introduction of this Section 1.3, the rate at which the parameter is adapted cannot be arbitrary. By Theorem 1.3.2, combining Proposition 1.3.4 with A2 yields the following result in the case $\pi_\theta = \pi_\star$ for any $\theta \in \Theta$.

Theorem 1.3.5 (Convergence of the Marginal when $\pi_\theta = \pi_\star$ for any $\theta \in \Theta$)**.** *Assume B1, A2 and $\pi_\theta = \pi_\star$ for all $\theta \in \Theta$. Then*

$$\lim_n \sup_{\{f\colon f \in \mathsf{b}\mathcal{X}\}} \left| \mathbb{E}\left[f(X_n)\right] - \pi_\star(f) \right| = 0 \; .$$

In the general case, when the invariant distributions π_θ depend upon θ, checking A4 is usually problem specific. When π_θ is available in closed-form, it might be possible to check A4 directly. For external adaptations algorithms (such as the EE sampler) such expression is not readily available (π_θ is known to exist, but computing its expression is out of reach). In this case, the only available option is to state conditions on the transition kernel:

E There exist $\theta_\star \in \Theta$ and a set $A \in \mathcal{A}$ such that $\mathbb{P}(A) = 1$ and for all $\omega \in A$, $x \in \mathsf{X}$, and $B \in \mathcal{X}$, $\lim_n P_{\theta_n(\omega)}(x, B) = P_{\theta_\star}(x, B)$.

Lemma 1.7.3 shows that this condition implies $\lim_n P^k_{\theta_n(\omega)}(x, B) = P^k_{\theta_\star}(x, B)$ for any integer k. This result, combined with the assumption B1, implies the following Proposition

Proposition 1.3.6. *Assume B1 and E. Let $\alpha \in [0, 1)$ and $f_\star : \mathsf{X} \to \mathbb{R}$ be a measurable function such that $f \in \mathcal{L}_{V^\alpha}$. Then* $\lim_n \pi_{\theta_n}(f) = \pi_{\theta_\star}(f)$*, $\mathbb{P}$-a.s.*

In E, it is assumed that the limiting kernel is of the form $P_{\theta_\star}$ for some $\theta_\star \in \Theta$. This assumption can be relaxed by imposing additional conditions on the limiting kernel; these conditions can be read from the proof of Proposition 1.3.6. Details are left to the interested reader. We thus have

Theorem 1.3.7 (Convergence of the Marginal)**.** *Assume B1, A2 and E. Then, for any bounded function f,*

$$\lim_{n\to\infty} \left| \mathbb{E}[f(X_n)] - \pi_{\theta_\star}(f) \right| = 0 \; .$$

1.4 Strong law of large numbers

In this Section, we establish a Strong Law of Large Number for adaptive MCMC. Haario et al. (2001) were the first to prove the consistency of Monte Carlo averages for the algorithm described by (1.2) and (1.3) for bounded functions, using mixingales techniques; these results have later been extended by Atchadé and Rosenthal (2005) to unbounded functions. Andrieu and Moulines (2006) have established the consistency and the asymptotic normality of $n^{-1} \sum_{k=1}^n f(X_k)$ for bounded and unbounded function for controlled MCMC algorithms associated to a stochastic approximation procedure (see Atchadé and Fort (To appear) for extensions). Roberts and Rosenthal (2007) prove a weak law of large numbers for bounded functions for general adaptive MCMC samplers. Finally, Atchadé (2009) provides a consistency result for an external adaptive sampler.

The proof is based on the so-called martingale technique (see (Meyn and Tweedie, 2009, Chapter 17)). We consider first this approach in the simple case of an homogeneous Markov chain, *i.e.* $P_\theta = P$. In this context, this technique amounts to decompose $n^{-1} \sum_{k=1}^n f(X_k) - \pi_\star(f)$ as $n^{-1} M_n(f) + R_n(f)$ where $\{M_n(f), n \geq 0\}$ is a $\mathbb{P}$-martingale (w.r.t. the natural filtration) and $R_n(f)$ is a remainder term. The martingale is shown to converge a.s. using standard results and the remainder terms is shown to converge to 0. This decomposition is not unique and different choices can be found in the literature. The most usual is based on the *Poisson equation* with forcing function f, namely $\hat{f} - P\hat{f} = f - \pi_\star(f)$. Sufficient conditions for the existence of a solution to the Poisson equation can be found in (Meyn and Tweedie, 2009, Chapter 17) (see also Lemma 1.7.1 below). In terms of $\hat{f}$, the martingale and the remainder terms may be expressed as

$$M_n(f) \overset{\text{def}}{=} \sum_{k=1}^{n} \left\{ \hat{f}(X_k) - P\hat{f}(X_{k-1}) \right\} , \qquad R_n(f) \overset{\text{def}}{=} n^{-1} \left[P\hat{f}(X_0) - P\hat{f}(X_n) \right] .$$

Proposition 1.4.1, which follows directly from (Hall and Heyde, 1980, Theorem 2.18), provides sufficient conditions for the almost-sure convergence of $n^{-1} \sum_{k=1}^{n} f(X_k)$ to $\pi_\star(f)$.

Proposition 1.4.1. *Let $\{X_k \ , k \geq 0\}$ be a Markov chain with transition kernel P and invariant distribution $\pi_\star$. Let $f : \mathsf{X} \to \mathbb{R}$ be a measurable function; assume that the Poisson equation $g - Pg = f - \pi_\star(f)$ with forcing function f possesses a solution denoted by $\hat{f}$. If $|P\hat{f}(X_0)| < +\infty$ $\mathbb{P}$-a.s. and there exists $p \in [1, 2]$ such that $\sum_k k^{-p} \mathbb{E}\left[|\hat{f}(X_k)|^p \right] < +\infty$, then $\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} f(X_k) = \pi_\star(f)$, $\mathbb{P}$-a.s.*

Provided that, $\sum_{j\geq 0} |P^j\{f - \pi_\star(f)\}(x)| < +\infty$ for any $x \in \mathsf{X}$, then $\hat{f}(x) \overset{\text{def}}{=} \sum_{j\geq 0} P^j\{f - \pi_\star(f)\}(x)$ is a solution to the Poisson equation. Lemma 1.7.1 gives sufficient conditions for this series to converge: under conditions which are essentially a Foster-Lyapunov drift inequality of the form $PV \leq \lambda V + b\mathbb{1}_{\mathcal{C}}$, for any $\beta \in [0, 1]$ and any $f \in \mathcal{L}_{V^\beta}$, a solution to the Poisson equation exists and this solution satisfies the condition $\sup_k \mathbb{E}\left[|\hat{f}(X_k)|^p \right] < +\infty$ for any p such that $\beta \leq p\beta \leq 1$. Hence, Proposition 1.4.1 with $p \in (1, 2]$ shows that a strong LLN holds for any function $f \in \mathcal{L}_{V^\beta}$, $\beta \in [0, 1)$. On the other hand, this drift condition also implies that $\pi_\star(V) < +\infty$.Therefore, the approach based on martingales yields to slightly weaker results than that obtained using the regenerative approach (see (Meyn and Tweedie, 2009, Section 17.2)).

The method based on martingales has been successfully used to prove strong LLN (and central limit theorems) for different adaptive chains (see e.g. Andrieu and Moulines (2006); Atchadé and Fort (To appear); Atchadé (2009)). A weak law of large numbers for bounded functions is also proved in Roberts and Rosenthal (2007); the proof is not based on a martingale-type transformation but relies on the coupling of the adaptive process with a 'frozen' chain with fixed transition kernels. Their convergence is established under the assumption that the kernels $\{P_\theta, \theta \in \Theta\}$ are simultaneously uniformly ergodic *i.e.* $\lim_n \sup_{x\in\mathsf{X}} \sup_{\theta\in\Theta} \|P_\theta^n(x, \cdot) - \pi_\theta\|_{\mathrm{TV}} = 0$. Using the same coupling approach, Yang (2008) also discusses the existence of a weak LLN for unbounded functions.

We develop below a general scheme of proof for strong LLN, based on the martingale approach. Let a measurable function $f : \mathsf{X} \to \mathbb{R}$. For any $\theta \in \Theta$, denote by $\hat{f}_\theta$ the solution to the Poisson equation $g - P_\theta g = f - \pi_\theta(f)$ with forcing function f. Consider the following decomposition:

$$n^{-1} \sum_{k=1}^{n} \{f(X_k) - \pi_{\theta_{k-1}}(f)\} = n^{-1} M_n(f) + \sum_{i=1}^{2} R_{n,i}(f) , \tag{1.17}$$

where

$$
\begin{aligned}
M_n(f) &\stackrel{\text{def}}{=} \sum_{k=1}^{n} \{\hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} \hat{f}_{\theta_{k-1}}(X_{k-1})\} \;, \\
R_{n,1}(f) &\stackrel{\text{def}}{=} n^{-1} \left(P_{\theta_0} \hat{f}_{\theta_0}(X_0) - P_{\theta_{n-1}} \hat{f}_{\theta_{n-1}}(X_n) \right) \;, \\
R_{n,2}(f) &\stackrel{\text{def}}{=} n^{-1} \sum_{k=1}^{n-1} \{P_{\theta_k} \hat{f}_{\theta_k}(X_k) - P_{\theta_{k-1}} \hat{f}_{\theta_{k-1}}(X_k)\} \;.
\end{aligned}
$$

$\{M_n(f), n \geq 0\}$ is a $\mathbb{P}$-martingale and $R_{n,i}(f)$, $i = 1, 2$ are remainder terms. Conditions for the almost-sure convergence to zero of $\{n^{-1} M_n(f), n \geq 0\}$ and of the residual term $\{R_{n,1}(f), n \geq 1\}$ are similar to those of Proposition 1.4.1; as discussed above they are implied by a geometric drift inequality, that is, in the adaptive case, they are implied by B1. The remaining term $R_{n,2}(f)$ requires additional attention. A (slight) strengthening of the diminishing adaptation condition is required. Define for $\theta, \theta' \in \Theta$,

$$
D_{V^\alpha}(\theta, \theta') \stackrel{\text{def}}{=} \sup_{x \in \mathsf{X}} V^{-\alpha}(x) \, \| P_\theta(x, \cdot) - P_{\theta'}(x, \cdot) \|_{\mathrm{V}^\alpha} \;.
$$

Lemma 1.4.2. *Assume B1. Let $\alpha \in (0, 1)$. Then, there exists a constant C such that for any $f \in \mathcal{L}_{V^\alpha}$ and for any $\theta, \theta' \in \Theta$,*

$$
\begin{aligned}
&\| \pi_\theta - \pi_{\theta'} \|_{\mathrm{V}^\alpha} \leq C \; D_{V^\alpha}(\theta, \theta') \;, \\
&| P_\theta \hat{f}_\theta - P_{\theta'} \hat{f}_{\theta'} |_{V^\alpha} \leq C |f|_{V^\alpha} \; D_{V^\alpha}(\theta, \theta') \;, \\
&| \hat{f}_\theta - \hat{f}_{\theta'} |_{V^\alpha} \leq C |f|_{V^\alpha} \; D_{V^\alpha}(\theta, \theta') \;.
\end{aligned}
$$

The proof is in Section 1.6. As a consequence of this Lemma, a sufficient condition for the convergence of $\{R_{n,2}(f), n \geq 0\}$ when $f \in \mathcal{L}_{V^\alpha}$ is

B2 For any $\alpha \in (0, 1)$,

$$
\sum_{k \geq 1} k^{-1} D_{V^\alpha}(\theta_k, \theta_{k-1}) \; V^\alpha(X_k) < +\infty \;, \quad \mathbb{P} - \text{a.s.}
$$

We now have all the ingredients to establish the following result

Theorem 1.4.3. *Assume B1, B2 and $\mathbb{E}[V(X_0)] < +\infty$. Then, for any $\alpha \in (0, 1)$ and $f \in \mathcal{L}_{V^\alpha}$, $\lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} \{f(X_k) - \pi_{\theta_{k-1}}(f)\} = 0$, $\mathbb{P}$-a.s.*

If $\pi_\theta = \pi_\star$ for any $\theta \in \Theta$, Theorem 1.4.3 implies the strong LLN for any functions $f \in \mathcal{L}_{V^\alpha}$, whatever $\alpha \in (0, 1)$. When $\pi_\theta \neq \pi_\star$, we have to prove the almost-sure convergence of the term $n^{-1} \sum_{k=1}^{n} \pi_{\theta_{k-1}}(f)$ to $\pi_\star(f)$ for some probability measure $\pi_\star$. This last convergence can be obtained from the condition E. Combining Proposition 1.3.6 and Theorem 1.4.3 yields

Theorem 1.4.4. *Assume B1 and E. Then, for any $\alpha \in (0, 1)$ and $f \in \mathcal{L}_{V^\alpha}$, $\lim_{n \to +\infty} n^{-1} \sum_{k=1}^{n} \pi_{\theta_{k-1}}(f) = \pi_\star(f)$, $\mathbb{P}$-a.s.*

1.5 Convergence of the Equi-Energy sampler

We study the convergence of the EE sampler described in Section 1.2.3. We prove the convergence of the marginals and a law of large numbers.

If $\{Y_n, n \geq 0\}$ is such that $n^{-1} \sum_{k=1}^{n} f(Y_k) \to \tilde{\pi}(f)$ a.s. for any bounded function f, the empirical distributions $\{\theta_n, n \geq 0\}$ converge weakly to $\tilde{\pi}$ so that, asymptotically, the dynamic of X_n is given by $P_{\tilde{\pi}}$. Since $\pi P_{\tilde{\pi}} = \pi$, it is expected that π governs the distribution of $\{X_n, n \geq 0\}$ asymptotically. By application of the results of Sections 1.3 and 1.4 this intuition can now be formalized.

1.5.1 Convergence of the marginal

We need X to be a Polish space. For simplicity, we assume that $\mathsf{X} = \mathbb{R}^d$ but our results remain true for more general state space X. In addition, it is assumed throughout this section that the rings are well-behaved, so that the closure of each ring X_ℓ coincides with the closure of its interior. Assume that

EES1 (a) π is a positive density distribution on X and $\sup_{\mathsf{X}} \pi < +\infty$.

(b) π is continuous on X.

EES2 P is a phi-irreducible transition kernel on $(\mathsf{X}, \mathcal{X})$ such that $\pi P = \pi$. The level sets $\{x \in \mathsf{X}, \pi(x) \geq p\}$ are 1-small for P, for any $p > 0$.

EES3 (a) There exist $\tau \in (0,1]$, $\lambda \in (0,1)$, $b < +\infty$ and a set $\mathcal{C} \in \mathcal{X}$ such that

$$PV \leq \lambda V + b \mathbb{1}_{\mathcal{C}} \;, \qquad \text{with} \qquad V(x) \overset{\text{def}}{=} \left(\pi(x) / \sup_{\mathsf{X}} \pi \right)^{-\tau(1-\beta)} \;.$$

(b) The probability of swapping ϵ is bounded by,

$$0 \leq \epsilon < \frac{1-\lambda}{1-\lambda+\tau(1-\tau)^{\frac{1-\tau}{\tau}}} \;. \tag{1.18}$$

If P is a symmetric random-walk Hastings-Metropolis (SRWM) kernel with a symmetric proposal distribution q which is positive on X, then $\pi P = \pi$ and P is π-irreducible (Mengersen and Tweedie, 1996, Lemma 1.1.). If in addition q is positive and continuous on X then, since π is positive and continuous on X, any compact set of X is 1-small (Mengersen and Tweedie, 1996, Lemma 1.2.). This discussion evidences that under EES1, condition EES2 is verified with P being, for example, a SRWM kernel with Gaussian proposal. Drift conditions for P being a SRWM kernel have been often discussed in the literature: conditions are both on the target distribution and on the proposal distribution. Depending upon π is decaying in the tails, the drift condition is geometric (as assumed in EES3) or sub-geometric; see e.g. Roberts and Tweedie (1996) for the geometric case and Fort and Moulines (2000) for the polynomial case. Conditions on π so that EES3 holds when P is a SRWM kernel (resp. an hybrid sampler) can be found e.g. in Roberts and Tweedie (1996) and Jarner and Hansen (2000) (resp. in Fort et al. (2003)).

Under conditions which imply that the target density π is super-exponential, Jarner and Hansen (2000) show that the functions proportional to π^{-s} for some $s \in (0,1)$ solve a Foster-Lyapunov drift inequality (Jarner and Hansen, 2000, Theorems 4.1 and 4.3). Hence, when P is a SRWM kernel, with a target density and a proposal distribution satisfying the conditions of Jarner and Hansen (2000), we can choose $\tau \in (0, 1/(1-\beta))$.

Lemma 1.5.1. *Assume EES1a and EES3. Then, for $\theta \in \Theta$, the Foster-Lyapunov condition $P_\theta V(x) \leq \tilde{\lambda} V(x) + b \mathbb{1}_{\mathcal{C}}(x)$ holds, with*

$$\tilde{\lambda} \overset{\text{def}}{=} (1-\epsilon)\lambda + \epsilon \left(1 + \tau(1-\tau)^{\frac{1-\tau}{\tau}} \right) \;.$$

The proof of Lemma 1.5.1 is in Section 1.6.

Proposition 1.5.2. *(i) Assume EES1a and EES3. Then, B1b holds and there exists a constant C such that* $\sup_n \mathbb{E}\left[V(X_n)\right] \leq C\mathbb{E}\left[V(X_0)\right]$.

(ii) Assume EES1a, EES2 and EES3. Then B1a and B1c hold.

EES4 For any bounded function $f : \mathsf{X} \to \mathbb{R}$, $\lim_n \theta_n(f) = \tilde{\pi}(f)$ $\mathbb{P}$-a.s.

Condition EES4 holds whenever $\{Y_n, n \geq 0\}$ is an ergodic Markov chain with stationary distribution $\tilde{\pi}$. A sufficient condition for ergodicity is the phi-irreducibility, the aperiodicity and the existence of an unique invariant distribution $\tilde{\pi}$ for the transition kernel of the chain $\{Y_n, n \geq 0\}$. For example, such a Markov chain $\{Y_n, n \geq 0\}$ can be obtained by running a HM sampler with invariant distribution $\tilde{\pi}$, with proposal distribution chosen in such a way that the sampler is ergodic (see *e.g.* Robert and Casella (1999)).

Proposition 1.5.3. *Assume EES1b and EES4. Then A2 and E holds.*

The proof is in Section 1.6. The second assertion really depends upon the topology of X: we assumed that $\mathsf{X} = \mathbb{R}^d$ but this condition can be weakened; details are left to the interested reader. From the above discussions, we have by application of Theorem 1.3.7:

Theorem 1.5.4. *Assume EES1 to EES4. Then for any bounded function f*

$$\lim_n \left|\mathbb{E}\left[f(X_n)\right] - \pi(f)\right| = 0 \; .$$

1.5.2 Strong law of large numbers

We have to strengthen the conditions on the auxiliary process $\{Y_n, n \geq 0\}$ as follows.

EES5 $\sup_n \mathbb{E}\left[V(Y_n)\right] < +\infty$ and for any $\alpha \in (0,1)$, $\lim_n \theta_n(V^\alpha) = \tilde{\pi}(V^\alpha)$ $\mathbb{P}$ a.s.

When $\{Y_n, n \geq 0\}$ is a Markov chain with transition kernel Q and invariant distribution $\tilde{\pi}$, a sufficient condition for EES5 is the existence of a Foster-Lyapunov drift inequality for Q. More precisely, if

> Q is phi-irreducible, aperiodic and there exist constants $\lambda' \in (0,1)$, $b' \in (0, +\infty)$, and a function W such that $QW \leq \lambda' W + b'$,
>
> $V \in \mathcal{L}_W$ and $\mathbb{E}\left[W(Y_0)\right] < +\infty$

then $\sup_n \mathbb{E}\left[W(Y_n)\right] \leq C\, \mathbb{E}\left[W(Y_0)\right]$ (see Lemma 1.7.1), and $\lim_n \theta_n(U) = \tilde{\pi}(U)$ a.s. for any function $U \in \mathcal{L}_W$ (see (Meyn and Tweedie, 2009, Theorem 17.1.7)). We thus have EES5.

Proposition 1.5.5. *Assume EES1a, EES3, EES5 and* $\mathbb{E}\left[V(X_0)\right] < +\infty$. *For any $\alpha \in (0,1)$, B2 holds.*

The proof is in Section 1.6. We now have all the ingredients to apply Theorems 1.4.3 and 1.4.4: from Propositions 1.5.2, 1.5.3 and 1.5.5, we have

Theorem 1.5.6. *Assume EES1 to EES5 and* $\mathbb{E}\left[V(X_0)\right] < +\infty$. *For any $\alpha \in (0,1)$, and any $f \in L_{V^\alpha}$,*

$$\lim_n n^{-1} \sum_{k=1}^{n} f(X_k) = \pi_\star(f) \; , \mathbb{P} - a.s.$$

We thus established that the process $\{X_n, n \geq 0\}$ possesses V-moment which are uniformly bounded (Proposition 1.5.2), its distribution converges to π (Theorem 1.5.4) and it satisfies a strong law of large numbers for a wide family of functions (Theorem 1.5.6). These results are obtained provided the auxiliary process $\{Y_n, n \geq 0\}$ also possesses uniformly bounded V-moment and satisfies a strong law of large numbers (see EES4 and EES5). Repeated applications of this analysis provide sufficient conditions for the equi-energy sampler with more than one auxiliary process - as originally described in Kou et al. (2006) - to be ergodic and to satisfy a strong law of large numbers. Details are omitted.

1.6 Proof of the main results

1.6.1 Proof of Section 1.3

Proof of Proposition 1.3.1

We prove this by induction. By (1.4), the statement is trivially true for $N = 1$ and any $n \geq 0$. Suppose that the statement is true for some $N \geq 1$ and for all $n \geq 0$. Then for $f \in \mathsf{b}\mathcal{X}$:

$$\begin{aligned}\left|\mathbb{E}\left[f(X_{n+N+1})|\mathcal{F}_n\right] - P_{\theta_n}^{N+1} f(X_n)\right| &= \left|\mathbb{E}\left[P_{\theta_{n+N}} f(X_{n+N})|\mathcal{F}_n\right] - P_{\theta_n}^{N+1} f(X_n)\right| \\ &\leq \left|\mathbb{E}\left[P_{\theta_{n+N}} f(X_{n+N}) - P_{\theta_n} f(X_{n+N})|\mathcal{F}_n\right]\right| \\ &\quad + \left|\mathbb{E}\left[\left(P_{\theta_n} f\right)(X_{n+N})|\mathcal{F}_n\right] - P_{\theta_n}^N \left(P_{\theta_n} f\right)(X_n)\right| .\end{aligned}$$

The first term on the rhs is bounded by $\mathbb{E}\left[D(\theta_{n+N}, \theta_n) \,|\, \mathcal{F}_n\right]$ and the second term is bounded by $\sum_{j=1}^{N-1} \mathbb{E}\left[D(\theta_{n+j}, \theta_n) \,|\, \mathcal{F}_n\right]$ by the induction assumption.

Proof of Theorem 1.3.2

Let $\epsilon > 0$. Since $k \mapsto \left\|P_\theta^k(x, \cdot) - \pi_\theta\right\|_{\mathrm{TV}}$ is non-increasing,

$$\sup_{k \geq M_\epsilon(x,\theta)} \left\|P_\theta^k(x, \cdot) - \pi_\theta\right\|_{\mathrm{TV}} \leq \|P_\theta^{M_\epsilon(x,\theta)}(x, \cdot) - \pi_\theta\|_{TV} \leq \epsilon \,, \tag{1.19}$$

where we used the definition of $M_\epsilon(x, \theta)$ in the last inequality. Under A3, there exist $N \in \mathbb{N}$ such that $\mathbb{P}(M_\epsilon(X_n, \theta_n) \geq N) \leq \epsilon$ for any $n \in \mathbb{N}$.

Set $\eta \overset{\mathrm{def}}{=} \epsilon/(2N^2)$. Under A2, there exists $n_{\epsilon,N}$ such that for any $n \geq n_{\epsilon,N}$, $\mathbb{P}\left(D_{\mathrm{TV}}(\theta_{n+1}, \theta_n) \geq \eta\right) \leq \eta$. Using Proposition 1.3.1, we get for $n \geq N + n_{\epsilon,N}$

$$\begin{aligned}&\sup_{\{f, |f|_\infty \leq 1\}} \mathbb{E}\left[\left|\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right] - P_{\theta_{n-N}}^N f(X_{n-N})\right|\right] \\ &\leq \sum_{j=1}^{N-1} \sum_{l=1}^{j} \mathbb{E}\left[D_{\mathrm{TV}}(\theta_{n-N+l}, \theta_{n-N+l-1})\right] \\ &\qquad\qquad \leq \epsilon/2 + \sum_{j=1}^{N-1} \sum_{l=1}^{j} \mathbb{P}\left(D_{\mathrm{TV}}(\theta_{n-N+l}, \theta_{n-N+l-1}) \geq \eta\right) \leq \epsilon \,.\end{aligned}$$

We write for $n \geq N$,

$$\begin{aligned}\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right] - \pi_{\theta_{n-N}}(f) = \mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right] - P_{\theta_{n-N}}^N f(X_{n-N}) \\ + P_{\theta_{n-N}}^N f(X_{n-N}) - \pi_{\theta_{n-N}}(f) \,,\end{aligned}$$

so that for $n \geq N + n_{\epsilon,N}$ and for any $f \in \mathsf{b}\mathcal{X}$,

$$\begin{aligned}
&\left|\mathbb{E}\left[f(X_n) - \pi_{\theta_{n-N}}(f)\right]\right| \\
&\leq \mathbb{E}\left[\left|\mathbb{E}\left[f(X_n)|\mathcal{F}_{n-N}\right] - P^N_{\theta_{n-N}} f(X_{n-N})\right|\right] + \mathbb{E}\left[\left|P^N_{\theta_{n-N}} f(X_{n-N}) - \pi_{\theta_{n-N}}(f)\right|\right] \\
&\leq \epsilon + \mathbb{E}\left[\left|P^N_{\theta_{n-N}} f(X_{n-N}) - \pi_{\theta_{n-N}}(f)\right|\right] \; .
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left[\left|P^N_{\theta_{n-N}} f(X_{n-N}) - \pi_{\theta_{n-N}}(f)\right|\right] &\leq 2\; \mathbb{P}\left(M_\epsilon(X_{n-N}, \theta_{n-N}) \geq N\right) \\
&+ \mathbb{E}\left[\left\|P^N_{\theta_{n-N}}(X_{n-N}, \cdot) - \pi_{\theta_{n-N}}(\cdot)\right\|_{\mathrm{TV}} \mathbb{1}_{\{M_\epsilon(X_{n-N},\theta_{n-N}) \leq N\}}\right] \; .
\end{aligned}$$

The rhs is upper bounded by 3ϵ, from (1.19) and the definition of N, uniformly in f for $f \in \mathsf{b}\mathcal{X}$.

Proof of Theorem 1.3.3

Let $\epsilon > 0$. Under A4, there exists n_ϵ such that

$$\sup_{n \geq n_\epsilon} \sup_{\{f : f \in \mathsf{F}\}} \mathbb{P}(|\pi_{\theta_{n-N}}(f) - \pi_\star(f)| \geq \epsilon) \leq \epsilon \; . \tag{1.20}$$

By (1.20), we have for any function $f \in \mathsf{F}$,

$$\begin{aligned}
&\sup_{n \geq n_\epsilon} \sup_{\{f : f \in \mathsf{F}\}} \mathbb{E}\left[\left|\pi_{\theta_{n-N}}(f) - \pi_\star(f)\right|\right] \\
&\qquad \leq 2 \sup_{n \geq n_\epsilon} \sup_{\{f : f \in \mathsf{F}\}} \mathbb{P}(|\pi_{\theta_{n-N}}(f) - \pi_\star(f)| \geq \epsilon) + \epsilon \leq 3\epsilon \; .
\end{aligned}$$

This is combined with Theorem 1.3.2 to prove the stated result.

Proof of Proposition 1.3.4

By (Meyn and Tweedie, 2009, Theorems 14.0.1 and 14.3.7), P_θ possesses an unique invariant probability measure π_θ and $\pi_\theta(V) \leq (1-\lambda)^{-1} b \pi_\theta(\mathcal{C}) \leq (1-\lambda)^{-1} b$.

By Lemma 1.7.1, there exist a constant C (depending upon ϵ) such that $\sup_\theta M_\epsilon(x,\theta) \leq C \ln V(x)$ and thus, a constant C' such that

$$\mathbb{P}\left(M_\epsilon(X_n, \theta_n) \geq M\right) \leq \mathbb{E}\left[C' \; \frac{V(X_n)}{M} \wedge 1\right] \leq \mathbb{E}\left[C' \; \frac{\mathbb{E}\left[V(X_n) \,|\, \mathcal{F}_0\right]}{M} \wedge 1\right] \; .$$

By iterating the drift inequality we have $\sup_n \mathbb{E}\left[V(X_n) \,|\, \mathcal{F}_0\right] \leq V(X_0) + (1-\lambda)^{-1} b < \infty$ (see (Andrieu and Moulines, 2006, Lemma 5) for a similar argument) which implies that for some constant C'',

$$\sup_n \mathbb{P}\left(M_\epsilon(X_n, \theta_n) \geq M\right) \leq \mathbb{E}\left[C'' \; \frac{V(X_0)}{M} \wedge 1\right] \; .$$

Since $V(x) < +\infty$ is finite for any x, the dominated convergence theorem yields the desired result.

Proof of Proposition 1.3.6

Let $\epsilon > 0$. We have for any $k \geq 0$ and $x \in \mathsf{X}$,

$$|\pi_{\theta_n}(f) - \pi_{\theta_\star}(f)| \leq 2 \sup_{\theta \in \Theta} \left\| \pi_\theta - P_\theta^k(x, \cdot) \right\|_{\mathrm{V}^\alpha} + \left| P_{\theta_n}^k f(x) - P_{\theta_\star}^k f(x) \right| .$$

Under B1, there exist constants $C < \infty$ and $\rho \in (0,1)$ such that $\sup_\theta \left\| P_\theta^k(x, \cdot) - \pi_\theta \right\|_{\mathrm{V}^\alpha} \leq C\rho^k \ V(x)$ - see Lemma 1.7.1. We choose k large enough so that $2C\rho^k \ V(x_\star) < \epsilon/2$. This implies that for any $\omega \in \mathcal{A}$ and any $n \geq 0$,

$$|\pi_{\theta_n(\omega)}(f) - \pi_{\theta_\star}(f)| \leq \epsilon/2 + \left| P_{\theta_n(\omega)}^k f(x_\star) - P_{\theta_\star}^k f(x_\star) \right| .$$

It remains to prove that for any $\omega \in A$, the second term in the rhs tends to zero as $n \to +\infty$, for fixed $k, f, x_\star, \theta_\star$. We have

$$\begin{aligned} P_{\theta_n}^k f(x_\star) - P_{\theta_\star}^k f(x_\star) &= \sum_{j=1}^k \left\{ P_{\theta_n}^j P_{\theta_\star}^{k-j} f(x_\star) - P_{\theta_n}^{j-1} P_{\theta_\star}^{k-j+1} f(x_\star) \right\} \\ &= \sum_{j=1}^k \int_{\mathsf{X}^2} P_{\theta_n}^{j-1}(x_\star, dy) \left\{ P_{\theta_n}(y, dz) - P_{\theta_\star}(y, dz) \right\} P_{\theta_\star}^{k-j} f(z) . \end{aligned}$$

The proof follows the same lines as (Atchadé, 2009, Lemma 4.9). We establish that for any $j \leq k$, and for any $\omega \in A$,

$$\lim_n \int_{\mathsf{X}^2} P_{\theta_n(\omega)}^{j-1}(x_\star, dy) \left\{ P_{\theta_n(\omega)}(y, dz) - P_{\theta_\star}(y, dz) \right\} P_{\theta_\star}^{k-j} f(z) = 0 , \tag{1.21}$$

which will conclude the proof. This is done under two successive applications of Lemma 1.7.2. Under E, for any $\omega \in A$, $y \in \mathsf{X}$ and $j \leq k$, Lemma 1.7.2 applied with $\mu_n \leftarrow P_{\theta_n(\omega)}(y, \cdot)$, $\mu \leftarrow P_{\theta_\star}(y, \cdot)$, $f_n = f = P_{\theta_\star}^{k-j} f$ and $V \leftarrow V^\alpha$ shows that for all $\omega \in A$, $y \in \mathsf{X}$, $j \leq k$,

$$\lim_n \int_{\mathsf{X}} \left\{ P_{\theta_n(\omega)}(y, dz) - P_{\theta_\star}(y, dz) \right\} P_{\theta_\star}^{k-j} f(z) = 0 . \tag{1.22}$$

Let $j \leq k$. By Lemma 1.7.3, for any $\omega \in A$ and $B \in \mathcal{X}$, $\lim_n P_{\theta_n(\omega)}^{j-1}(x_\star, B) = P_{\theta_\star}^{j-1}(x_\star, B)$. Then (1.22) and Lemma 1.7.2 applied with $\mu_n \leftarrow P_{\theta_n(\omega)}^{j-1}(x_\star, \cdot)$, $\mu \leftarrow P_{\theta_\star}^{j-1}(x_\star, \cdot)$, $f_n(y) \leftarrow P_{\theta_n(\omega)} P_{\theta_\star}^{k-j} f(y)$, $f \leftarrow P_{\theta_\star}^{k-j+1} f(y)$ and $V \leftarrow V^\alpha$, implies (1.21).

1.6.2 Proof of Section 1.4

Proof of Lemma 1.4.2

We provide a self-contained proof which follows the same arguments as in (Benveniste et al., 1990, Part II). Let $\alpha \in (0,1)$. The Jensen's inequality implies that $P_\theta V^\alpha \leq \lambda^\alpha V^\alpha + b^\alpha \mathbb{1}_C$. Results of Lemma 1.7.1 thus apply with V replaced by V^α.

First assertion. We have for any $n \geq 1$,

$$P_\theta^n f - P_{\theta'}^n f = \sum_{j=0}^{n-1} P_\theta^j \left(P_\theta - P_{\theta'} \right) \left(P_{\theta'}^{n-j-1} f - \pi_{\theta'}(f) \right) .$$

By Lemma 1.7.1(iii) there exist constants C and $\rho \in (0,1)$ such that for any $n \geq 0$, $x \in \mathsf{X}$, $\sup_\theta \| P_\theta^n(x, \cdot) - \pi_\theta \|_{V^\alpha} \leq C \ \rho^n \ V^\alpha(x)$. Furthermore, by iterating the drift inequality B1b, we have, for any $x \in \mathsf{X}$, $\sup_j \sup_\theta P_\theta^j V^\alpha(x) < +\infty$. This yields for any $k \geq 1$, $x_\star \in \mathsf{X}$,

$$\|\pi_\theta - \pi_{\theta'}\|_{V^\alpha} \leq \|\pi_\theta - P_\theta^k(x_\star, \cdot)\|_{V^\alpha} + \|P_\theta^k(x_\star, \cdot) - P_{\theta'}^k(x_\star, \cdot)\|_{V^\alpha} + \|P_{\theta'}^k(x_\star, \cdot) - \pi_{\theta'}\|_{V^\alpha}$$
$$\leq 2C\ \rho^k V^\alpha(x_\star) + \sup_{|h|_{V^\alpha} \leq 1} \left| \sum_{j=0}^{k-1} P_\theta^j \left(P_\theta - P_{\theta'}\right) \left(P_{\theta'}^{k-j-1} h - \pi_{\theta'}(h)\right)(x_\star) \right| .$$

The second term in the rhs is upper bounded by

$$C D_{V^\alpha}(\theta, \theta') \sum_{j=1}^{k-1} \rho^{k-j-1} \sup_\theta P_\theta^j V^\alpha(x_\star) \ \leq \frac{C}{1-\rho} D_{V^\alpha}(\theta, \theta') \sup_j \sup_\theta P_\theta^j V^\alpha(x_\star) .$$

Hence, there exists a constant C such that for any $k \geq 1$,

$$\|\pi_\theta - \pi_{\theta'}\|_{V^\alpha} \leq C \left\{ \rho^k + D_{V^\alpha}(\theta, \theta') \right\} .$$

Since this holds true for any k, this yields the desired result.

Second assertion. Under the stated assumptions $\hat{f}_\theta$ exists and is given by $\sum_{n \geq 0} P_\theta^n \{f - \pi_\theta(f)\}$. Following the same lines as in Benveniste et al. (1990) (see also (Andrieu and Moulines, 2006, Proposition 3)), we have for any $n \geq 1$,

$$P_\theta^n f - P_{\theta'}^n f$$
$$= \sum_{j=0}^{n-1} \left(P_\theta^j - \pi_\theta\right) \left(P_\theta - P_{\theta'}\right) \left(P_{\theta'}^{n-j-1} f - \pi_{\theta'}(f)\right) + \sum_{j=0}^{n-1} \left\{ \pi_\theta P_{\theta'}^{n-j-1} f - \pi_\theta P_{\theta'}^{n-j} f \right\}$$
$$= \sum_{j=0}^{n-1} \left(P_\theta^j - \pi_\theta\right) \left(P_\theta - P_{\theta'}\right) \left(P_{\theta'}^{n-j-1} f - \pi_{\theta'}(f)\right) + \pi_\theta(f) - \pi_\theta P_{\theta'}^n f ,$$

where we used that $\pi_\theta P_\theta = \pi_\theta$ in the second equality. This implies for any $n \geq 1$,

$$P_\theta^n \{f - \pi_\theta(f)\} - P_{\theta'}^n \{f - \pi_{\theta'}(f)\}$$
$$= \sum_{j=0}^{n-1} \left(P_\theta^j - \pi_\theta\right) \left(P_\theta - P_{\theta'}\right) \left(P_{\theta'}^{n-j-1} f - \pi_{\theta'}(f)\right) - \pi_\theta \{P_{\theta'}^n f - \pi_{\theta'}(f)\} .$$

Hence, by summing wrt n the previous identity,

$$P_\theta \hat{f}_\theta - P_{\theta'} \hat{f}_{\theta'} = \sum_{n \geq 1} \sum_{j=0}^{n-1} \left(P_\theta^j - \pi_\theta\right) \left(P_\theta - P_{\theta'}\right) \left(P_{\theta'}^{n-j-1} f - \pi_{\theta'}(f)\right) - \pi_\theta P_{\theta'} \hat{f}_{\theta'}$$
$$= \sum_{n \geq 1} \sum_{j=0}^{n-1} \left(P_\theta^j - \pi_\theta\right) \left(P_\theta - P_{\theta'}\right) \left(P_{\theta'}^{n-j-1} f - \pi_{\theta'}(f)\right) + \left(\pi_{\theta'} - \pi_\theta\right) P_{\theta'} \hat{f}_{\theta'} ,$$

where we used that $P_{\theta'} \hat{f}_{\theta'}$ is $\pi_{\theta'}$-integrable (see Lemma 1.7.1) and $\pi_{\theta'}(P_{\theta'} \hat{f}_{\theta'}) = 0$. Since $\sup_\theta |\hat{f}_\theta| \leq C V^\alpha$, we have

$$\left| \left(\pi_{\theta'} - \pi_\theta\right) \hat{f}_{\theta'} \right| \leq \sup_\theta |\hat{f}_\theta|_{V^\alpha} \ \|\pi_{\theta'} - \pi_\theta\|_{V^\alpha} .$$

This implies that for any $n \geq 1$ and $j \leq n-1$,

$$\begin{aligned}
&\left|\left(P_\theta^j - \pi_\theta\right)(P_\theta - P_{\theta'})\left(P_{\theta'}^{n-j-1} f - \pi_{\theta'}(f)\right)(x)\right| \\
&\leq C\ \rho^j V^\alpha(x) \left|(P_\theta - P_{\theta'})\left(P_{\theta'}^{n-j-1} f - \pi_{\theta'}(f)\right)\right|_{\mathrm{V}^\alpha} \\
&\leq C\ \rho^j V^\alpha(x)\ \{\sup_{x \in \mathsf{X}} V^{-\alpha}(x)\ \ \|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_{\mathrm{V}^\alpha}\}\ \left|P_{\theta'}^{n-j-1} f - \pi_{\theta'}(f)\right|_{\mathrm{V}^\alpha} \\
&\leq C\, |f|_{\mathrm{V}^\alpha}\ \rho^{n-1}\ D_{V^\alpha}(\theta, \theta')\ V^\alpha(x)\ .
\end{aligned}$$

This concludes the proof of the second assertion. The third assertion is based on the equality

$$P_\theta \hat{f}_\theta(x) - P_{\theta'} \hat{f}_{\theta'}(x) = \hat{f}_\theta(x) - \hat{f}_{\theta'}(x) + \pi_\theta(f) - \pi_{\theta'}(f)\ ,$$

and the two other assertions.

Proof of Theorem 1.4.3

Under B1, each transition kernel P_θ is aperiodic and $\sup_{\mathcal{C}} V < +\infty$ (see comments in Section 1.3.3). By the Jensen's inequality, we have $P_\theta V^\alpha \leq \lambda^\alpha V^\alpha + b^\alpha \mathbb{1}_{\mathcal{C}}$. Hence, by Lemma 1.7.1(ii), there exists a constant C such that for any $\theta \in \Theta$, any $x \in \mathsf{X}$ and any $f \in \mathcal{L}_{V^\alpha}$, $|\hat{f}_\theta(x)| + |P_\theta \hat{f}_\theta(x)| \leq C|f|_{V^\alpha}\ V^\alpha(x)$. By iterating the drift inequality we have $\sup_n \mathbb{E}[V(X_n)] \leq \mathbb{E}[V(X_0)] + (1-\lambda)^{-1} b$. Therefore, there exists $p \in (1,2]$ such that

$$\sum_{k \geq 1} k^{-p} \mathbb{E}\left[\left|\hat{f}_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} \hat{f}_{\theta_{k-1}}(X_{k-1})\right|^p\right] < +\infty\ .$$

As in Proposition 1.4.1, it is proved that $\lim_{n\to\infty} n^{-1} M_n(f) + R_{n,1}(f) = 0$ $\mathbb{P}$-a.s. Lemma 1.4.2 shows that, for $f \in \mathcal{L}_{V^\alpha}$,

$$|R_{n,2}(f)| \leq C|f|_{V^\alpha} n^{-1} \sum_{k=1}^{n-1} D_{V^\alpha}(\theta_k, \theta_{k-1}) V^\alpha(X_k)\ ,$$

which implies, using the Kronecker Lemma (Hall and Heyde, 1980, Section 2.6), that $R_{n,2}(f) \xrightarrow{\text{a.s.}} 0$.

1.6.3 Proof of Section 1.5

Proof of Lemma 1.5.1

Set $C_\theta(x) \overset{\text{def}}{=} \int g(x,y)\theta(dy)$. If $C_\theta(x) = 0$, then $P_\theta V = PV \leq \lambda V + b\mathbb{1}_C$ and the result follows. Assume now that $C_\theta(x) > 0$. By (1.13),

$$K_\theta(x, V) = V(x) + C_\theta^{-1}(x) \int g(x,y)\alpha(x,y)\{V(y) - V(x)\}\,\theta(dy)\ .$$

Following the same lines as in the proof of (Atchadé, 2009, Lemma 4.2), we write for $x \in \mathsf{X}_\ell$,

$$\begin{aligned}
&V^{-1}(x) \int g(x,y)\alpha(x,y)\{V(y) - V(x)\}\,\theta(dy) \\
&= \int_{\mathsf{X}_\ell} \left\{1 \wedge \frac{\pi^{1-\beta}(y)}{\pi^{1-\beta}(x)}\right\}\left\{\frac{V(y)}{V(x)} - 1\right\}\theta(dy) \\
&= \int_{A(x)\cap \mathsf{X}_\ell} \left\{\frac{V(y)}{V(x)} - 1\right\}\theta(dy) + \int_{R(x)\cap \mathsf{X}_\ell} \left(\frac{\pi(y)}{\pi(x)}\right)^{1-\beta}\left\{\frac{V(y)}{V(x)} - 1\right\}\theta(dy)
\end{aligned}$$

where $A(x) \stackrel{\text{def}}{=} \{y \in \mathsf{X}, \pi(y) \geq \pi(x)\} = \{y, \alpha(x,y) = 1\}$ is the acceptance region and $R(x) \stackrel{\text{def}}{=} \mathsf{X} \setminus A(x)$ is the rejection region. Since $V \propto \pi^{-\tau(1-\beta)}$, for $x \in \mathsf{X}_\ell$,

$$V^{-1}(x) \int g(x,y)\alpha(x,y)\left\{V(y) - V(x)\right\}\theta(dy)$$
$$\leq \int_{R(x)\cap\mathsf{X}_\ell} \left(\frac{\pi(y)}{\pi(x)}\right)^{1-\beta} \left\{\left(\frac{\pi(y)}{\pi(x)}\right)^{-\tau(1-\beta)} - 1\right\}\theta(dy) \leq \tau(1-\tau)^{\frac{1-\tau}{\tau}} C_\theta(x) ,$$

by using that, for $\tau \in (0,1]$, $\sup_{x\in[0,1]} x(x^{-\tau} - 1) \leq \tau(1-\tau)^{\frac{1-\tau}{\tau}}$. The proof follows.

Proof of Proposition 1.5.2

Lemma 1.5.1 implies B1b. By iterating the drift inequality we have the upper bound on $\sup_n \mathbb{E}\left[V(X_n)\right]$. Since P is phi-irreducible and $\epsilon < 1$, P_θ is phi-irreducible for any $\theta \in \Theta$, thus showing B1a. The level sets of V, $\{V \leq v\}$ are small for P and thus for P_θ, uniformly-in-θ, since $P_\theta(x,\cdot) \geq (1-\epsilon)P(x,\cdot)$. This yields B1c.

Proof of Proposition 1.5.3

(i) Set $C_\theta(x) \stackrel{\text{def}}{=} \int g(x,y)\theta(dy)$. Under the stated assumptions, there exists $A \in \mathcal{A}$ such that $\mathbb{P}(A) = 1$ and for any $\omega \in A$, for any $x \in \mathsf{X}_l$

$$\lim_n C_{\theta_n(\omega)}(x) = \int_{\mathsf{X}_l} \tilde{\pi}(dy) ,$$

and the rhs is positive under the assumptions on π and on the partition of X. Hence, there exists a random variable N, finite a.s. , such that for any $n \geq N$, and for any bounded function f such that $|f|_\infty \leq 1$, we have a.s.

$$\left|P_{\theta_n} f(x) - P_{\theta_{n-1}} f(x)\right| \leq 2\epsilon \left\| \frac{g(x,\cdot)\theta_n}{C_{\theta_n}(x)} - \frac{g(x,\cdot)\theta_{n-1}}{C_{\theta_{n-1}}(x)} \right\|_{\mathrm{TV}} ,$$

upon noting that by definition of α, $|\alpha(x,y)| \leq 1$. By definition of $\{\theta_n, n \geq 0\}$, we have a.s. for any $n \geq N$,

$$\sup_{x\in\mathsf{X}} \sup_{\{f: |f|_\infty \leq 1\}} \left| \frac{\theta_n(g(x,\cdot)f)}{C_{\theta_n}(x)} - \frac{\theta_{n-1}(g(x,\cdot)f)}{C_{\theta_{n-1}}(x)} \right| \leq 2 \sup_{x\in\mathsf{X}} \frac{1}{\sum_{k=1}^n g(x,Y_k)} .$$

By definition of g, $\sup_{x\in\mathsf{X}} \left(\sum_{k=1}^n g(x,Y_k)\right)^{-1} = n^{-1} \sup_{\ell\in\{1,\ldots,K\}} \left(\theta_n(\mathsf{X}_\ell)\right)^{-1}$. Hence, EES4 implies that the rhs tends to zero a.s. which concludes the proof.

(ii) Set $\theta_\star \stackrel{\text{def}}{=} \tilde{\pi}$ and $C_{\theta_\star}(x) \stackrel{\text{def}}{=} \int g(x,y)\tilde{\pi}(dy)$. As discussed above, there exists a random variable N, finite a.s. , such that for any $n \geq N$, we have a.s.

$$P_{\theta_n(\omega)}(x,B) - P_{\tilde{\pi}}(x,B) = \epsilon \int_B \left\{ \frac{\theta_n(dy)}{C_{\theta_n}(x)} - \frac{\tilde{\pi}(dy)}{C_{\theta_\star}(x)} \right\} g(x,y)\ \alpha(x,y)$$
$$- \epsilon \int_{\mathsf{X}} \left\{ \frac{\theta_n(dy)}{C_{\theta_n}(x)} - \frac{\tilde{\pi}(dy)}{C_{\theta_\star}(x)} \right\} g(x,y)\alpha(x,y) .$$

Under EES4, for any $x \in \mathsf{X}$ and $B \in \mathcal{X}$, there exists $A_{x,B} \in \mathcal{A}$ such that $\mathbb{P}(A_{x,B}) = 1$ and for any $\omega \in A_{x,B}$,

$$\lim_n P_{\theta_n(\omega)}(x,B) = P_{\theta_\star}(x,B) . \tag{1.23}$$

Since X is Polish, it admits a dense subset $\mathcal{D}$ and a countable generating algebra $\mathcal{B}$ of $\mathcal{X}$. Hence, there exists a set $A_{\mathcal{D},\mathcal{B}} \in \mathcal{A}$ such that $\mathbb{P}(A_{\mathcal{D},\mathcal{B}}) = 1$ and for any $x \in \mathcal{D}$, $B \in \mathcal{B}$, $\omega \in A_{\mathcal{D},\mathcal{B}}$, the limit (1.23) holds.

We now prove that the limit holds for any $x \in \mathcal{D}$ and $B \in \mathcal{X}$. Let $x \in \mathcal{D}$ and $B \in \mathcal{X}$. For $\epsilon > 0$, there exist $B_-, B_+ \in \mathcal{B}$ such that $B_- \subseteq B \subseteq B_+$ and

$$P_{\theta_\star}(x,B) - \epsilon \leq P_{\theta_\star}(x,B_-) \leq P_{\theta_\star}(x,B) \leq P_{\theta_\star}(x,B_+) \leq P_{\theta_\star}(x,B) + \epsilon \; .$$

By positivity of the probability $P_{\theta_n}(x,\cdot)$, we have $P_{\theta_n}(x,B_-) \leq P_{\theta_n}(x,B) \leq P_{\theta_n}(x,B_+)$. In addition, for any $\omega \in A_{\mathcal{D},\mathcal{B}}$, $\lim_n P_{\theta_n(\omega)}(x,B') = P_{\theta_\star}(x,B')$ for $B' \in \{B_-, B_+\}$. This yields

$$\begin{aligned} P_{\theta_\star}(x,B) - \epsilon \leq P_{\theta_\star}(x,B_-) &\leq \liminf_n P_{\theta_n(\omega)}(x,B) \\ &\leq \limsup_n P_{\theta_n(\omega)}(x,B) \leq P_{\theta_\star}(x,B_+) \leq P_{\theta_\star}(x,B) + \epsilon \; . \end{aligned}$$

This implies that for any $\omega \in A_{\mathcal{D},\mathcal{B}}$, $x \in \mathcal{D}$ and $B \in \mathcal{X}$, (1.23) holds. The last step is to prove that the limit holds for any $x \in \mathsf{X}$. For $x, x' \in \mathsf{X}$, we write

$$P_{\theta_n}(x,B) - P_{\theta_\star}(x,B) = \epsilon \left\{ N_{\theta_n}(x,B) - N_{\theta_\star}(x,B) - N_{\theta_n}(x,\mathsf{X}) + N_{\theta_\star}(x,\mathsf{X}) \right\}$$

where $N_\theta(x,B) \overset{\text{def}}{=} C_\theta(x)^{-1} \int \theta(dy)\, g(x,y)\alpha(x,y)\, \mathbb{1}_B(y)$. By EES1b and Lemma 1.6.1, for any $x \in \mathsf{X}$ and $k \geq 1$, there exists $x_k \in \mathcal{D}$ such that $\lim_k x_k = x$, $g(x,\cdot) = g(x_k,\cdot)$ (this is a consequence of the assumptions on the partition of X) and for any $B \in \mathcal{X}$ and $\theta \in \Theta$,

$$N_\theta(x_k,B) - \frac{1}{k} \leq N_\theta(x,B) \leq N_\theta(x_k,B) + \frac{1}{k} \; .$$

This implies that for any $\omega \in A_{\mathcal{D},\mathcal{B}}$, $B \in \mathcal{X}$ and $x \in \mathsf{X}$,

$$N_{\theta_\star}(x_k,B) - \frac{1}{k} \leq \liminf_n N_{\theta_n(\omega)}(x,B) \leq \limsup_n N_{\theta_n(\omega)}(x,B) \leq N_{\theta_\star}(x_k,B) + \frac{1}{k} \; ,$$

since $\lim_n N_{\theta_n(\omega)}(x_k,B) = N_{\theta_\star}(x_k,B)$. By Lemma 1.6.1, this yields $\lim_k N_{\theta_\star}(x_k,B) = N_{\theta_\star}(x,B)$. Hence, we proved that for any $\omega \in A_{\mathcal{D},\mathcal{B}}$, $x \in \mathsf{X}$ and $B \in \mathcal{X}$, (1.23) holds.

Proof of Proposition 1.5.5

As in the proof of Proposition 1.5.3, there exists a random variable N finite a.s. such that for any $n \geq N$, we have a.s. $\sup_{x \in \mathsf{X}} |C_{\theta_n}(x) - C_{\theta_{n-1}}(x)| = 0$ and $\inf_{x \in \mathsf{X}} C_{\theta_n}(x) > 0$. We thus have for any $n \geq N$, a.s.

$$\begin{aligned} D_{V^\alpha}(\theta_n, \theta_{n-1}) &\leq 2\epsilon \sup_{x \in \mathsf{X}} \left\| \frac{g(x,\cdot)\theta_n}{C_{\theta_n}(x)} - \frac{g(x,\cdot)\theta_{n-1}}{C_{\theta_{n-1}}(x)} \right\|_{\mathsf{V}^\alpha} \\ &\leq n^{-1} \sup_\ell \theta_n^{-1}(\mathsf{X}_\ell) \left\{ V^\alpha(Y_n) + \left((n-1)^{-1} \sum_{k=1}^{n-1} V^\alpha(Y_k) \right) \sup_\ell \theta_{n-1}^{-1}(\mathsf{X}_\ell) \right\} \; . \end{aligned}$$

Note that $\lim_n \left\{ \theta_n^{-1}(\mathsf{X}_\ell) - \tilde{\pi}^{-1}(\mathsf{X}_\ell) \right\} = 0$ and $\lim_n n^{-1} \sum_{k=1}^{n-1} V^\alpha(Y_k) = \tilde{\pi}(V^\alpha) > 0$, $\mathbb{P}$-a.s. Hence, there exists a $\mathbb{P}$-a.s. finite integer valued random variable $\tilde{N}$, and two finite random variables U_1 and U_2 such that, for all $n \geq \tilde{N} \geq N$,

$$D_{V^\alpha}(\theta_n, \theta_{n-1}) \leq n^{-1} \left(V^\alpha(Y_n) U_1 + U_2 \right) \; . \tag{1.24}$$

To check B2, we prove that $\sum_{k\geq \tilde{N}} k^{-2} V^{\alpha}(X_k) < +\infty$ and $\sum_{k\geq \tilde{N}} k^{-2} V^{\alpha}(Y_k)\ V^{\alpha}(X_k) < +\infty$ a.s. by applying Lemma 1.7.4. The first sum converges since $\sup_k \mathbb{E}\left[V(X_k)\right] \leq C\ \mathbb{E}\left[V(X_0)\right]$ (see Proposition 1.5.3). For the second sum, we apply Lemma 1.7.4 with $a = (1+\epsilon)/2$ for some $\epsilon \in (0,1)$ such that $\alpha(1+\epsilon) \leq 1$. We have by the Hölder's inequality

$$\begin{aligned}\mathbb{E}\left[V^{\alpha a}(Y_k)\ V^{\alpha a}(X_k)\right] \leq \mathbb{E}\left[V^{\alpha(1+\epsilon)}(Y_k)\right]^{1/2}\ \mathbb{E}\left[V^{\alpha(1+\epsilon)}(X_k)\right]^{1/2} \\ \leq \mathbb{E}\left[V(Y_k)\right]^{1/2}\ \mathbb{E}\left[V(X_k)\right]^{1/2} \ .\end{aligned}$$

Under the stated assumptions, $\sup_k \mathbb{E}\left[V(Y_k) + V(X_k)\right] < +\infty$ which concludes the proof of the convergence of the series. and the proof.

Lemma 1.6.1 is adapted from (Atchadé, 2009, Lemma 4.3).

Lemma 1.6.1. *For any $\theta \in \Theta, x \in \mathsf{X}$, $B \in \mathcal{X}$ such that $C_\theta(x) \overset{\text{def}}{=} \int g(x,y)\theta(dy) \in (0,+\infty)$ define*

$$N_\theta(x,B) \overset{\text{def}}{=} C_\theta(x)^{-1} \int_B \theta(dy) g(x,y) \left\{1 \wedge \frac{\pi^{1-\beta}(y)}{\pi^{1-\beta}(x)}\right\} \ .$$

For any $x, x' \in \mathsf{X}$, such that $g(x,\cdot) = g(x',\cdot)$

$$\sup_{B\in\mathcal{X}} \sup_{\theta\in\Theta} \left|N_\theta(x,B) - N_\theta(x',B)\right| \leq \left(\frac{\pi(x)\vee\pi(x')}{\pi(x)\wedge\pi(x')}\right)^{1-\beta} - 1 \ .$$

Proof. We assume that $\pi(x) \leq \pi(x')$. We have

$$\begin{aligned}&C_\theta(x)\ \left|N_\theta(x,B) - N_\theta(x',B)\right| \\ &= \left|\int \theta(dy)\ g(x,y) \left\{1 \wedge \frac{\pi^{1-\beta}(y)}{\pi^{1-\beta}(x)} - 1 \wedge \frac{\pi^{1-\beta}(y)}{\pi^{1-\beta}(x')}\right\}\right| \\ &\leq \int_{\{y, \pi(y)\leq\pi(x)\}} \theta(dy)\ g(x,y)\pi^{1-\beta}(y)\ \left(\frac{1}{\pi^{1-\beta}(x)} - \frac{1}{\pi^{1-\beta}(x')}\right) \\ &+ \int_{\{y, \pi(x)\leq\pi(y)\leq\pi(x')\}} \theta(dy)\ g(x,y)\pi^{1-\beta}(y)\ \left(\frac{1}{\pi^{1-\beta}(y)} - \frac{1}{\pi^{1-\beta}(x')}\right) \\ &\leq \left(\frac{1}{\pi^{1-\beta}(x)} - \frac{1}{\pi^{1-\beta}(x')}\right) \int_{\{y, \pi(y)\leq\pi(x')\}} \theta(dy)\ g(x,y)\pi^{1-\beta}(y) \\ &\qquad\qquad \leq \left(\frac{\pi^{1-\beta}(x')}{\pi^{1-\beta}(x)} - 1\right) C_\theta(x) \ .\end{aligned}$$

□

1.7 Technical lemmas

Lemma 1.7.1. *Let P be a phi-irreducible and aperiodic transition kernel. Assume there exist a measurable function $V : \mathsf{X} \to [1,+\infty)$, $\lambda \in (0,1)$, $b \in (0,+\infty)$ and a set $\mathcal{C}$ such that $PV \leq \lambda V + b\mathbb{1}_{\mathcal{C}}$ and $\sup_{\mathcal{C}} V < +\infty$. Assume in addition that the level sets $\{V \leq v\}$ of V are 1-small with minorizing constant ϵ_v.*

(i) Then the chain possesses an invariant distribution π such that $\pi(V) \leq (1-\lambda)^{-1}b$, In addition, $\sup_n \int \xi(dx)P^nV(x) \leq \lambda\xi(V) + (1-\lambda)^{-1}b$ for any distribution ξ on $(\mathsf{X},\mathcal{X})$.

(ii) *For any function $f \in \mathcal{L}_V$, there exists a solution $\hat{f}$ to the Poisson equation that satisfies for any $a \in (0,1)$ and any $x \in \mathsf{X}$,*

$$|\hat{f}(x)| \leq \frac{|f|_V}{a} \left(V(x) + \pi(V) + \frac{\{2b + \lambda(\sup_{\mathcal{C}} V + (1-a)^{-1} b)\}}{\epsilon_{\sup_{\mathcal{C}} V \vee (1-a)^{-1} b}} \right) .$$

(iii) *There exist constants C and $\rho \in (0,1)$ - depending upon $\sup_{\mathcal{C}} V$, b, λ and the minorizing constant $\epsilon_{\sup_{\mathcal{C}} V}$ such that*

$$\|P^n(x,\cdot) - \pi\|_{\mathrm{V}} \leq C \ \rho^n \ V(x) \ .$$

Proof. The control of $\pi(V)$ is a consequence of (Meyn and Tweedie, 2009, Theorem 14.3.7). The uniform control of $\mathbb{E}_\xi[V(X_n)]$ is obtained by iterating the drift inequality. Item (ii) is a consequence of (Fort and Moulines, 2003, Proposition 22); details of proof are omitted. (iii) is proved in Douc et al. (2004). □

Lemma 1.7.2 is established in (Atchadé, 2009, Lemma 4.8).

Lemma 1.7.2. *Let $\{\mu_n, n \geq 0\}$ be a family of probability measures on $(\mathsf{X}, \mathcal{X})$ and $\{f_n : \mathsf{X} \to \mathbb{R}, n \geq 0\}$ be a family of measurable functions. Assume that*

(i) *$\mu_n(A) \to \mu(A)$ for any $A \in \mathcal{X}$.*

(ii) *there exists $f : \mathsf{X} \to \mathbb{R}$ such that $f_n(x) \to f(x)$ for any $x \in \mathsf{X}$.*

(iii) *there exists a function $V : \mathsf{X} \to (0, +\infty)$ such that $\sup_n |f_n|_V < +\infty$, $\mu(V) < +\infty$ and $\sup_n \mu_n(V^\alpha) < +\infty$ for some $\alpha > 1$.*

Then $\lim_n \mu_n(f_n) = \mu(f)$.

Lemma 1.7.3. *Assume E. Then for any $k \geq 1$, for all $\omega \in A$, $x \in \mathsf{X}$, and $B \in \mathcal{X}$, $\lim_n P^k_{\theta_n(\omega)}(x, B) = P^k_{\theta_\star}(x, B)$.*

Proof. The proof is by induction on k. The case $k = 1$ holds by E. Assume that the property holds with k. We write

$$P^{k+1}_{\theta_n(\omega)}(x,B) - P^{k+1}_{\theta_\star}(x,B) = \int P^k_{\theta_n(\omega)}(x,dy) P_{\theta_n(\omega)}(y,B) - \int P^k_{\theta_\star}(x,dy) P_{\theta_\star}(y,B)$$

and apply Lemma 1.7.2 with $\mu_n \leftarrow P^k_{\theta_n(\omega)}(x,dy)$, $\mu \leftarrow P^k_{\theta_\star}(x,dy)$, $f_n \leftarrow P_{\theta_n(\omega)}(\cdot, B)$, $f \leftarrow P_{\theta_\star}(\cdot, B)$, $V = 1$. The induction assumption and E imply that the condition of Lemma 1.7.2 and thus, the property holds for $k+1$. This concludes the induction. □

Lemma 1.7.4. *Let $\{W_k, k \geq 1\}$ be a family of positive random variables and $\{\rho_k, k \geq 1\}$ be positive sequence. Then the series $\sum_k \rho_k W_k$ is finite $\mathbb{P}$-a.s. if there exists $a \in (0,1]$ such that (a) $\sum_k \rho_k^a < +\infty$ and (b) $\sup_k \mathbb{E}[W_k^a] < +\infty$.*

Proof. We have by sub-additivity

$$\mathbb{E}\left[\left(\sum_k \rho_k W_k\right)^a\right] \leq \sum_k \rho_k^a \ \mathbb{E}[W_k^a] \leq \left(\sup_k \mathbb{E}[W_k^a]\right) \sum_k \rho_k^a \ .$$

□

Bibliography

ANDRIEU, C., JASRA, A., DOUCET, A. and DEL MORAL, P. (2007). On nonlinear Markov chain Monte Carlo via self-interacting approximations. Tech. rep., Available at http://stats.ma.ic.ac.uk/a/aj2/public_html/.

ANDRIEU, C. and MOULINES, E. (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).

ANDRIEU, C. and ROBERT, C. (2001). Controlled MCMC for optimal sampling. Tech. Rep. Tech. Rep. 0125, Cahiers de Mathématiques du Ceremade; Université Paris-Dauphine.

ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing* **18** 343–373.

ATCHADÉ, Y. (2009). A cautionary tale on the efficiency of some adaptive Monte Carlo schemes. Tech. rep., ArXiv:0901:1378v1.

ATCHADÉ, Y. and FORT, G. (To appear). Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. *Bernoulli. ArXiv math.PR/0807.2952 (revision, Jan 2009)* .

ATCHADÉ, Y. F. and ROSENTHAL, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11** 815–828.

BAI, Y. (2008). The simultaneous drift conditions for Adaptive Markov Chain Monte Carlo algorithms. Tech. rep., Univ. of Toronto, available at http://www.probability.ca/jeff/ftpdir/yanbai2.pdf.

BAI, Y., ROBERTS, G. and ROSENTHAL, J. (2009). On the Containement condition for Adaptive Markov Chain Monte Carlo algorithms. Tech. rep., Univ. of Toronto, available at http://www.probability.ca/jeff/.

BENVENISTE, A., MÉTIVIER, M. and PRIOURET, P. (1990). *Adaptive Algorithms and Stochastic Approximations*, vol. 22. Springer-Verlag, Berlin.

CAPPE, O. and MOULINES, E. (2009). Online EM algorithm for latent data models. To appear in J. Roy. Stat. Soc., B.

Chauveau, D. and Vandekerkhove, P. (1999). Un algorithme de Hastings-Metropolis avec apprentissage séquentiel. *C. R. Acad. Sci. Paris Sér. I Math.* **329** 173–176.

Chauveau, D. and Vandekerkhove, P. (2001). Algorithmes de Hastings-Metropolis en interaction. *C. R. Acad. Sci. Paris Sér. I Math.* **333** 881–884.

Chauveau, D. and Vandekerkhove, P. (2002). Improving convergence of the Hastings-Metropolis algorithm with an adaptive proposal. *Scand. J. Statist.* **29** 13–29.

Craiu, R. V., Rosenthal, J. S. and Yang, C. (2008). Learn from thy neighbor: Parallel-chain adaptive MCMC. Tech. rep., University of Toronto, available at http://www.probability.ca/jeff/ftpdir/chao4.pdf.

Douc, R., Moulines, E. and Rosenthal, J. (2004). Quantitative bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.* **14** 1643–1665.

Fort, G. and Moulines, E. (2000). V-subgeometric ergodicity for a Hastings-Metropolis algorithm. *Stat. Probab. Lett.* **49** 401–410.

Fort, G. and Moulines, E. (2003). Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *Ann. Statist.* **31** 1220–1251.

Fort, G., Moulines, E., Roberts, G. and Rosenthal, J. (2003). On the geometric ergodicity of hybrid samplers. *J. Appl. Probab.* **40** 123–146.

Gelman, A., Roberts, G. O. and Gilks, W. R. (1996). Efficient Metropolis jumping rules. In *Bayesian statistics, 5 (Alicante, 1994)*. Oxford Sci. Publ., Oxford Univ. Press, New York.

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

Geyer, C. J. and Thompson, E. (1995). Annealing Markov chain Monte Carlo with Applications to pedigree analysis. *Journal of the American Statistical Association* **90** 909–920.

Giordani, P. and Kohn, R. (2008). Adaptive Independent Metropolis-Hastings by Fast Estimation of Mixtures of Normals.

Haario, H., Saksman, E. and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Stat.* **14** 375–395.

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

Haario, H., Saksman, E. and Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC. *Comput. Statist.* **20** 265–273.

Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and Its Application.* Academic Press, New York.

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

Jasra, A., Stephens, D. A. and Holmes, C. C. (2007). On population-based simulation for static inference. *Stat. Comput.* **17** 263–279.

Keith, J., Kroese, D. and Sofronov, G. (2008). Adaptive independence samplers. *Statistics and Computing* **18** 409–420.

Kou, S., Zhou, Q. and Wong, W. (2006). Equi-energy sampler with applications to statistical inference and statistical mechanisms (with discussion). *Ann. Statist.* **34** 1581–1619.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, vol. 35. 2nd ed. Springer, New York.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo schemes. *Europhysics letters* **19** 451–458.

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

Meyn, S. and Tweedie, R. (2009). *Markov Chains and Stochastic Stability.* Cambridge University Press. Second Edition.

Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer-Verlag, New-York.

Roberts, G., Gelman, A. and Gilks, W. (1997). Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *The Annals of Applied Probability* **7** 110–120.

Roberts, G. and Rosenthal, J. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.* **16** 351–367.

Roberts, G. and Rosenthal, J. (2006). Examples of adaptive MCMC. Tech. rep., Univ. of Toronto. To appear in J. Computational Graphical Statistics.

Roberts, G. and Tweedie, R. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–10.

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

Rosenthal, J. S. (2007). AMCMC: An R interface for adaptive MCMC. *Comput. Statist. Data Anal.* **51** 5467–5470.

Rosenthal, J. S. (2009). *MCMC Handbook*, chap. Optimal Proposal Distributions and Adaptive MCMC. Chapman & Hall/CRC Press.

Rubinstein, R. and Kroese, D. (2008). *Simulation and the Monte Carlo method.* 2nd ed. Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Turro, E., Bochkina, N., Hein, A.-M. and Richardson, S. (2007). Bgx: a bioconductor package for the bayesian integrated analysis of affymetrix genechips. *BMC Bioinformatics* **8** 439.

Yang, C. (2008). On the weak law of large number for unbounded functionals of adaptive MCMC. Tech. rep., Univ. of Toronto, available at http://www.probability.ca/jeff/ftpdir/chao2.pdf.