Bernoulli **21**(3), 2015, 1304–1340
DOI: 10.3150/13-BEJ578

Adaptive MCMC with online relabeling

RÉMI BARDENET[1], OLIVIER CAPPÉ[2,*], GERSENDE FORT[2,**] and BALÁZS KÉGL[1,3]

[1] *Laboratoire de Recherche en Informatique, Université Paris-Sud XI, rue Noetzlin, 91190 Gif-sur-Yvette, France. E-mail: remi.bardenet@gmail.com*
[2] *LTCI, Telecom ParisTech & CNRS, 37 rue Dareau, 75013 Paris, France.*
E-mail: [*] cappe@telecom-paristech.fr; [**] gfort@telecom-paristech.fr
[3] *CNRS, Laboratoire de l'Accélérateur Linéaire, Université Paris-Sud XI, 91898 Orsay, France.*
E-mail: balazs.kegl@gmail.com

When targeting a distribution that is *artificially* invariant under some permutations, Markov chain Monte Carlo (MCMC) algorithms face the *label-switching* problem, rendering marginal inference particularly cumbersome. Such a situation arises, for example, in the Bayesian analysis of finite mixture models. Adaptive MCMC algorithms such as adaptive Metropolis (AM), which self-calibrates its proposal distribution using an online estimate of the covariance matrix of the target, are no exception. To address the label-switching issue, *relabeling* algorithms associate a permutation to each MCMC sample, trying to obtain reasonable marginals. In the case of adaptive Metropolis (*Bernoulli* **7** (2001) 223–242), an *online* relabeling strategy is required. This paper is devoted to the AMOR algorithm, a provably consistent variant of AM that can cope with the label-switching problem. The idea is to nest relabeling steps within the MCMC algorithm based on the estimation of a *single covariance matrix* that is used *both* for adapting the covariance of the proposal distribution in the Metropolis algorithm step *and* for online relabeling. We compare the behavior of AMOR to similar relabeling methods. In the case of compactly supported target distributions, we prove a strong law of large numbers for AMOR and its ergodicity. These are the first results on the consistency of an online relabeling algorithm to our knowledge. The proof underlines latent relations between relabeling and vector quantization.

Keywords: adaptive Markov chain Monte Carlo; label-switching; stochastic approximation; vector quantization

1. Introduction

Markov chain Monte Carlo (MCMC) is a generic approach for exploring complex probability distributions based on sampling [24]. It has become the *de facto* standard tool in many applications of Bayesian inference. However, a very common situation in which MCMC algorithms face serious difficulties is when the target posterior distribution is known to be invariant under some permutations (or block permutations) of the variables. In that case, the difficulties are both computational, as most often the MCMC algorithm fails to validly visit all the modes of the posterior, and inferential, in particular rendering marginal posterior inference about the individual variables particularly cumbersome [10]. In the literature, this latter difficulty is usually referred to as the *label switching problem* [32]. The most well-known example of this situation is when performing Bayesian inference in a mixture model. In this case, the mixture likelihood is invariant to permuting the mixture components and, most often, the prior itself does not favor any specific ordering

1350-7265 © 2015 ISI/BS

of the mixture components [9,17–19,22,31,32]. Another important example arises in signal processing with additive decomposition models. In this case, the observed signal is represented as the superposition of exchangeable signals, and the main goal is to recover the individual signals or their parameters. In addition, often the number of signals also has to be determined [7,29,30]. It was observed empirically that when the dimension of the model is not known, the reversible jump sampler [23] makes it easier to visit the multiple modes corresponding to the permutations but, of course, marginal inference becomes harder due to the additional difficulty of associating components between models of varying dimension.

In this contribution, we address the label switching problem in the generic case where no useful external information on the target is known. This corresponds, for instance, to a posterior distribution when neither the likelihood is assumed to have a specific form, nor the prior is chosen to have conjugacy properties, which forbids the use of Gibbs sampling or other specialized sampling strategies. We assume, however, that the target is known to be invariant under some permutations of the parameters. This framework is typical, for instance, in experimental physics applications where the likelihood computation is commonly deferred to a *black-box* numerical code. In those cases, one cannot assume anything about the structure of the posterior or its conditional distributions, except that they should be invariant to some permutations of the parameters. We also restrict ourselves to the case where the dimension of the model is finite and known so the parameters of the model are $\mathbb{R}^d$-valued for some fixed and finite d.

Following [4], an *adaptive MCMC* algorithm is an algorithm which, given a family of MCMC transition kernels $(P_\theta)_{\theta\in\Theta}$ on a space $\mathbb{X}$, produces a $(\mathbb{X}\times\Theta)$-valued process $((X_n,\theta_n))_{n\geq 0}$ such that the conditional distribution of the sample X_{n+1} given the past is $P_{\theta_n}(X_n,\cdot)$. In practice, adaptive MCMC are MCMC algorithms that can self-calibrate their internal parameters along the iterations in order to reach decent performance without (or with almost no) knowledge about the target distribution, eliminating the grueling step of tuning the proposals. Adaptive MCMC has been an active field of research in the last ten years, following the pioneering contribution of [16] – see [3] as well as the other papers in the same special issue of *Statistics and Computing*, along with [2,4,28]. Adaptive Metropolis (hereafter AM; [16]) and its variants aim at identifying the unknown covariance structure of the target distribution along the run of a random walk Metropolis–Hastings algorithm with a multivariate Gaussian proposal. The rationale behind this approach is based on scaling results which suggest that, when d tends to $+\infty$, the chain correlation is minimized when the covariance matrix used in the proposal distribution matches, up to a constant that depends on the dimension, the covariance matrix of the target, for a large class of unimodal target distributions with independent marginals [25,26]. AM thus progressively adapts, using a stochastic approximation scheme, the covariance of the proposal distribution to the estimated covariance of the target.

It has been empirically observed in [5], and we provide further evidence of this fact below in Section 2.2, that the efficiency of AM can be greatly impaired when label switching occurs. The reason for such a difficulty is obvious: if label switching occurs, the estimated covariance matrix no longer corresponds to the local shape of the modes of the posterior and so the exploration can be far from optimal. In Section 2.2, we also provide some empirical evidence that off-the-shelf solutions to the label-switching problem, such as imposing identifiability constraints or post-processing the simulated sample, are not fully satisfactory. A key difficulty here is that most of the approaches proposed in the literature are based on post-processing of the simulated trajectories *after* the MCMC algorithm has been fully run [17–19,22,30–32]. Unfortunately, in the case of adaptive MCMC, post-processing cannot solve the improper exploration issue described above. On the other hand, online relabeling algorithms [10,12,23] often require manual tuning based on, for example, prior knowledge on the location of the redundant modes of the target. Without such manual tuning they often yield poor samplers, as we will show it in Section 2.2.

Our main purpose in this paper is to provide a provably consistent variant of AM that can cope with the label-switching problem. In [5], we proposed an adaptive Metropolis algorithm with online relabeling, called AMOR, based on the original idea of [9]. The idea is to nest relabeling steps within the MCMC algorithm based on the estimation of a *single covariance matrix* that is used *both* for adapting the covariance of the proposal distribution used in the Metropolis algorithm step *and* for online relabeling. Contrary to [9], the AMOR algorithm also corrects for the relabelings using a modified acceptance ratio. Similarly to [9], though, AMOR requires to loop over all possible relabelings of proposed points, which limits the method in practice to applications with a relatively small number of permutations. Modifications and heuristics that address this issue are out of the scope of this paper.

In Section 2.2, we provide empirical evidence that the coupling established in AMOR between the criterion used for relabeling and the estimation of the covariance of the local modes of the posterior is beneficial to avoid the distortion of the marginal distributions. Furthermore, the example considered in Section 2.2 also demonstrates that the AMOR algorithm samples from nontrivial identifiable restrictions of the posterior distribution, that is, truncations of the posterior on regions where the posterior marginals are distinct but from which the complete posterior can be recovered by permutation. The study of the convergence of AMOR in Section 3 reveals an interesting connection with the problem of optimal probabilistic quantization [14], which was implicit in earlier works on label switching. It was observed previously by [21] that some adjustments to the usual theory of stochastic approximation are necessary to analyze online optimal quantification due to the presence of points where the mean field of the algorithm is not differentiable. To circumvent this difficulty, we introduce the stable AMOR algorithm, a novel variant of the AMOR algorithm that avoids these problematic points of the parameter space. Finally, we establish consistency results for the stable AMOR algorithm, showing that it indeed asymptotically provides samples distributed under a suitably defined restriction of the posterior distribution in which the parameters are marginally identifiable.

The paper is organized as follows. In Section 2, we describe the stable AMOR algorithm and compare it with alternative approaches on an illustrative example. In Section 3, we address the convergence of the algorithm. The detailed proofs are provided in the Appendix.

2. The stable AMOR algorithm

In this section, we introduce the stable AMOR algorithm and illustrate its performance on an artificial example.

Algorithm 1

STABLEAMOR$(\pi(\cdot), X_0, T, \theta_0 = (\mu_0, \Sigma_0), c, (\gamma_t)_{t\geq 0}, \alpha, (\mathcal{K}_\psi)_{\psi\geq 0})$

```
$\mathcal{S} \leftarrow \emptyset$
$\psi \leftarrow 0$    ▷ Projection counter
for $t \leftarrow 1$ to $T$
    $\Sigma \leftarrow c\Sigma_{t-1}$ ▷ scaled adaptive covariance
    $\tilde{X} \sim \mathcal{N}(\cdot|X_{t-1}, \Sigma)$    ▷ proposal
    $\tilde{P} \sim \arg\min_{P\in\mathcal{P}} L_{\theta_{t-1}}(P\tilde{X})$    ▷ pick an optimal permutation
    $\tilde{X} \leftarrow \tilde{P}\tilde{X}$    ▷ permute
    if $\frac{\pi(\tilde{X})\sum_P \mathcal{N}(PX_{t-1}|\tilde{X}, \Sigma)}{\pi(X_{t-1})\sum_P \mathcal{N}(P\tilde{X}|X_{t-1}, \Sigma)} > \mathcal{U}[0, 1]$ then
        $X_t \leftarrow \tilde{X}$    ▷ accept
    else
        $X_t \leftarrow X_{t-1}$    ▷ reject
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{X_t\}$    ▷ update posterior sample
    $\mu_t \leftarrow \mu_{t-1} + \gamma_t(X_t - \mu_{t-1}) + \alpha\gamma_t \,\mathrm{Pen}_{t-1,1}$
    $\Sigma_t \leftarrow \Sigma_{t-1} + \gamma_t((X_t - \mu_{t-1})(X_t - \mu_{t-1})^\mathsf{T} - \Sigma_{t-1}) + \alpha\gamma_t \,\mathrm{Pen}_{t-1,2}$
    if $(\mu_t, \Sigma_t) \notin \mathcal{K}_\psi$ then
        $(\mu_t, \Sigma_t) \leftarrow (\mu_0, \Sigma_0)$    ▷ Project back to $\mathcal{K}_0$
        $\psi \leftarrow \psi + 1$    ▷ Increment projection counter
    $\theta_t \leftarrow (\mu_t, \Sigma_t)$.
return $\mathcal{S}$
```

2.1. The algorithm

Let π be a density with respect to (w.r.t.) the Lebesgue measure on $\mathbb{R}^d$ which is invariant to the action of a finite group $\mathcal{P}$ of permutation matrices, that is,

$$\forall x \in \mathbb{R}^d, \forall P \in \mathcal{P}, \qquad \pi(x) = \pi(Px).$$

Denote by $\mathcal{C}_d^+$ the set of $d \times d$ real positive definite matrices. For $\theta = (\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathcal{C}_d^+$, define $L_\theta : \mathbb{R}^d \to \mathbb{R}_+$ by

$$L_\theta(x) = (x - \mu)^T \Sigma^{-1}(x - \mu), \tag{2.1}$$

and let $\mathcal{N}(\cdot|\mu, \Sigma)$ denote the Gaussian density with mean μ and covariance matrix Σ.

Let $\Theta \subseteq \mathbb{R}^d \times \mathcal{C}_d^+$ and $(\mathcal{K}_q)_{q\in\mathbb{N}}$ be an increasing sequence of compact subsets of Θ such that $\bigcup_{q\in\mathbb{N}} \mathcal{K}_q = \Theta$.

Algorithm 1 describes the pseudocode of stable AMOR [5]. Choose $\theta_0 \in \mathcal{K}_0$.

To explain the proposal mechanism of stable AMOR, let μ_{t-1} and Σ_{t-1} denote the sample mean and the sample covariance matrix, respectively, at the end of iteration $t - 1$, and let $\theta_{t-1} = (\mu_{t-1}, \Sigma_{t-1})$. Let us also $\mathcal{S}$ denote the MCMC sample at the end of iteration $t - 1$. At iteration t, a point $\tilde{X}$ is first drawn from a Gaussian centered at the previous state X_{t-1} and with covariance $c\Sigma_{t-1}$, where c implements the optimal scaling results in [25,26] discussed in Section 1 (steps 4 and 5). Then in steps 6 and 7, $\tilde{X}$ is replaced by $\tilde{P}\tilde{X}$, where $\tilde{P}$ is a uniform draw over the permutations in $\arg\min_P L_{\theta_{t-1}}(P\tilde{X})$ that minimize the relabeling criterion (2.1).[1] This relabeling step makes the augmented sample $S \cup \{\tilde{P}\tilde{X}\}$ look as Gaussian as possible among all augmented sets $S \cup \{P\tilde{X}\}$, $P \in \mathcal{P}$. Formally, it can be seen as a projection onto the Voronoi cell $V_{\theta_{t-1}}$, where

$$V_\theta = \bigl\{x \in \mathbb{X}/L_\theta(x) \leq L_\theta(Px), \forall P \in \mathcal{P}\bigr\}. \tag{2.2}$$

Then, in steps 8 to 11, the candidate $\tilde{P}\tilde{X}$ is accepted or rejected according to the usual Metropolis–Hastings rule. The sample mean and covariance are adapted according to a Stochastic Approximation (SA) scheme in steps 13 and 14; $\alpha \in [0, \infty)$ and $\mathrm{Pen}_{t,i}$ is a penalty term used to drive the parameters $\theta_t = (\mu_t, \Sigma_t)$ toward the set of interest Θ. In Section 3, we will give examples of parameter set Θ and penalty terms $\mathrm{Pen}_{t,i}$. $(\gamma_t)_{t\geq 1}$ is a sequence of nonnegative steps, usually set according to a polynomial decay $\gamma_t \sim \gamma_\star t^{-\beta}$ for some $\beta \in (1/2, 1]$. Finally, steps 15 to 17 are a truncation mechanism with random varying bounds to make the SA algorithm stable. In SA procedures, such a step is a way to make the paths $(\theta_t)_{t\geq 0}$ bounded with probability one, which is a required property to prove the convergence of these procedures (see, e.g., [11]). We will provide in Section 3 sufficient conditions implying that the number of random truncations is finite along almost all paths $(\theta_t)_{t\geq 0}$, thus implying that after a finite number of iterations, everything happens as if steps 15 to 17 were omitted. In practice, it is often reported in the literature that SA is stable even when these stabilization steps are omitted.

Stable AMOR is a doubly adaptive MCMC algorithm since it is adaptive both in its *proposal* and *relabeling* mechanisms. This means that, besides the proposal distribution, its target also changes with the number of iterations. In Section 3, we will prove that, at each iteration t, AMOR implements a random walk Metropolis–Hastings kernel with stationary distribution $\pi_\theta \propto \pi \mathbb{1}_{V_\theta}$.

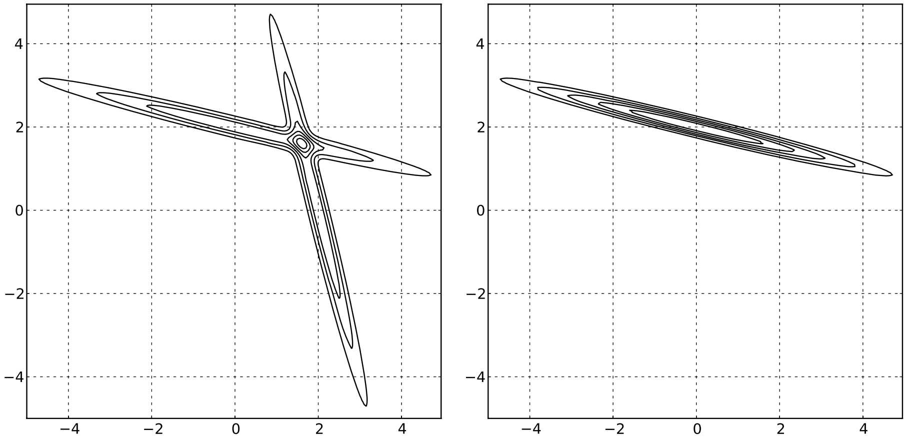
2.2. An illustrative example

In this section, we consider an artificial target aimed at illustrating the gap in performance between the stable AMOR algorithm and other common approaches to the label switching problem, which are compatible with adaptive MCMC. Consider the two-dimensional p.d.f. π depicted in Figure 1(a), which satisfies $\pi(x) = \pi(Px)$ for $P \in \mathcal{P}$, where

$$\mathcal{P} = \left\{\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}, \begin{pmatrix}0 & 1\\1 & 0\end{pmatrix}\right\}.$$

The density π is a mixture of two densities with equal weights obtained by superposing the Gaussian p.d.f. π_{SEED} represented in Figure 1(b) with a symmetrized version of itself. This artificial target does not correspond to the posterior distribution in an actual inference problem. In particular, although π itself is a mixture, it is not the posterior distribution of the parameters of any specific mixture model. Nevertheless, it is relevant because it is permutation invariant and

[1] Step 6 usually boils down to selecting the permutation $\tilde{P}$ that minimizes $L_{\theta_{t-1}}$. In case of ties, however, $\tilde{P}$ should be drawn uniformly over the set on which the minimum is achieved.



(a) π (b) π_{SEED}

Figure 1. Panel (a) shows the target distribution π used in Section 2.2, obtained by symmetrizing the Gaussian π_{SEED} shown in panel (b). π_{SEED} has mean $(0, 2)$ and covariance matrix with diagonal $(16, 1)$ and nondiagonal terms equal to -0.975.

the desired solution of the label switching problem is well defined: we know that, under suitable relabeling, we can obtain univariate near-Gaussian marginals for both coordinates by recovering the marginals of the two-dimensional Gaussian π_{SEED} in Figure 1(b). In spite of its simplicity, this example is challenging because the two marginals of π_{SEED} have similar means (0 and 2) and one has large variance, which makes them hard to separate. Given the modest dimension of the problem, we fix the number of MCMC iterations to 20 000, of which 4000 are discarded as burn-in. For each algorithm, we assess the quality of the relabeling strategy by looking at the corresponding restriction π' of the target π, and we assess the efficiency of the sampling by plotting the autocorrelation function of each sample and comparing the sample histograms with the marginals of π'.

The results obtained when applying AM, without any relabeling, are shown in Figure 2. The marginal posteriors are sampled quite well (Figures 2(c) and 2(d)) and the covariance of the joint sample (indicated by a thick ellipse Figure 2(a)) is almost symmetric. This is not surprising: the joint distribution, although severely non-Gaussian, is unimodal, and the number of iterations is large enough for AM to explore both the original seed π_{SEED} and its symmetric version by frequent label switching. On the other hand, the covariance of the joint distribution π (Figure 1(a)) is broader than the covariance of the seed π_{SEED} (Figure 1(b)). This results in poor adaptive proposals and slow mixing as indicated by the slight differences between the marginals and the sample marginals, and by the autocorrelation function of the first component of the sample in Figure 2(b). The reference (dashed line) is the autocorrelation function of an MCMC chain with optimal covariance (proportional to the covariance of the target) targeting the single Gaussian π_{SEED} (Figure 1(b)).

We now consider a modified version of AM with online relabeling obtained by simply ordering the variables, meaning that after each proposal $x = (x_1, x_2)$, the components of the proposed

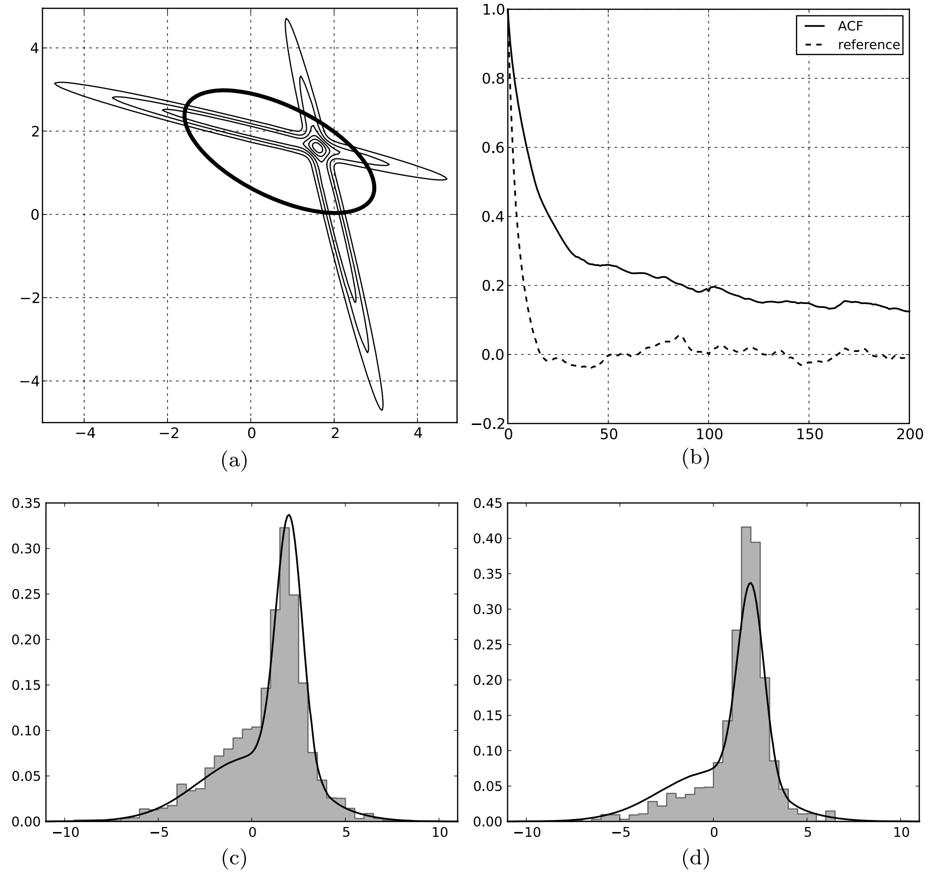


Figure 2. Results of vanilla AM on the two-dimensional target π of Figure 1. The rest of the caption is the same for Figures 3 to 5. On panel (a), level lines of π are depicted in thin black lines; a thick ellipse centered at the empirical mean μ_T of the sample $\mathcal{S}$ indicates the set $\{x : (x - \mu_T)^T \Sigma_T^{-1}(x - \mu_T) = 1\}$, where Σ_T is the sample covariance. When appropriate, the region of the space selected by (the last iteration of) the algorithm corresponds to the unshaded background while the region not selected is shaded. On panel (b), the autocorrelation function (ACF) of the first component of $\mathcal{S}$ is plotted as a solid line. The dashed line indicates the ACF obtained when sampling from the seed Gaussian π_{SEED} of Figure 1(b) using a random walk Metropolis algorithm with an optimally tuned covariance matrix. Panels (c) and (d) display the histograms of the two marginal samples. The solid curves are the marginals of π in this figure. In Figures 3 to 5, they are the marginals of π restricted to the unshaded region selected by the algorithms.

point are permuted so that $x_1 \le x_2$. This strategy is known as *imposing an identifiability constraint*. It is known to perform badly when the constraint does not respect the topology of the target [19]. The results of this approach on our illustrative example are shown in Figure 3. The unshaded triangle in Figure 3 shows that this time the sample is restricted to a sub-region of $\mathbb{R}^2$ where the components are identifiable. Unfortunately, the marginals of π restricted to the

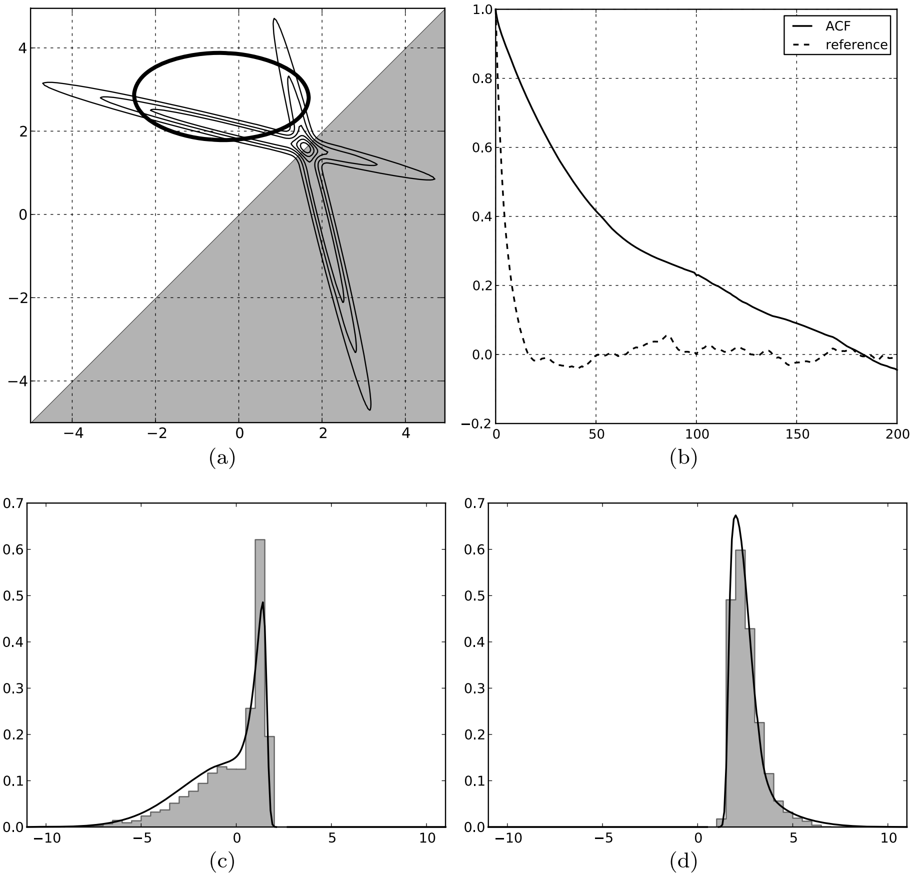


Figure 3. Results of AM with online ordering constraint. For details about the plots, see the caption of Figure 2.

unshaded triangle in Figures 3(c) and 3(d) are even more highly skewed than the marginals of the full joint distribution π. In addition, sampling from the restricted distribution π' is not easier than before indicated by the autocorrelation function in Figure 3(b).

Next, we consider the approach introduced by Celeux in [9]. Celeux's algorithm builds on a nonadaptive random-walk Metropolis, where online relabeling is performed in the following way: when a point $x = (x^{(1)}, x^{(2)})$ is proposed at time t, it is relabeled by

$$\begin{aligned} x \leftarrow \arg\min \Biggl\{ & \begin{pmatrix} x^{(1)} - \mu_t^{(1)} \\ x^{(2)} - \mu_t^{(2)} \end{pmatrix}^T D_t^{-1} \begin{pmatrix} x^{(1)} - \mu_t^{(1)} \\ x^{(2)} - \mu_t^{(2)} \end{pmatrix}, \\ & \begin{pmatrix} x^{(2)} - \mu_t^{(1)} \\ x^{(1)} - \mu_t^{(2)} \end{pmatrix}^T D_t^{-1} \begin{pmatrix} x^{(2)} - \mu_t^{(1)} \\ x^{(1)} - \mu_t^{(2)} \end{pmatrix} \Biggr\}, \end{aligned} \tag{2.3}$$

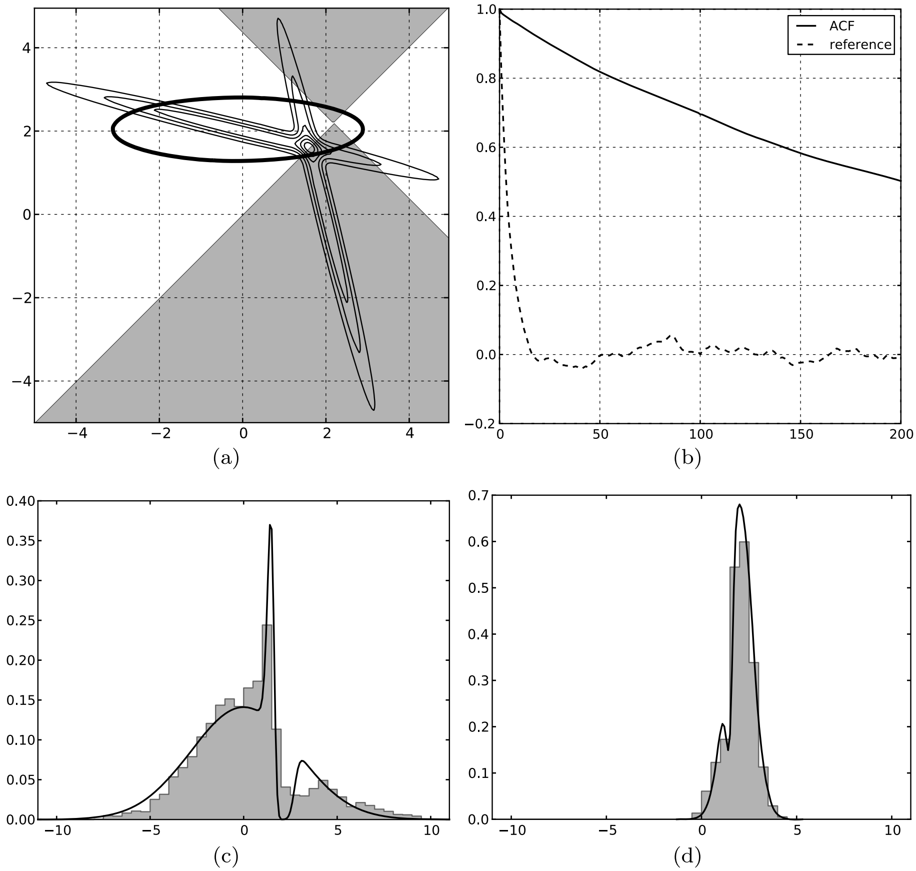


Figure 4. Results of Celeux's algorithm. For details about the plots, see the caption of Figure 2.

where $\mu_t = (\mu_t^{(1)}, \mu_t^{(2)})$ is the empirical mean of the current sample $x_{1:t} = x_1, \dots, x_t$ and D_t is the diagonal matrix containing the empirical variances of the coordinates of $x_{1:t}$ on its diagonal. Formally, this relabeling rule is equivalent to steps 6 and 7 of Algorithm 1, but with all nondiagonal elements of Σ equal to zero. The results of Celeux's algorithm are shown in Figure 4. It is hard to determine precisely the formal target of the algorithm. In particular, given the non-isotropic shape of the target, we used a non-isotropic Gaussian proposal with diagonal covariance matrix, and while the preservation of the detailed balance condition then requires incorporating a term into the acceptance ratio to account for the relabeling, it is absent in this approach. It is still possible that the algorithm is *approximately* sampling from the restriction π' of π to this unshaded area in Figure 4 (which represents the relabeling rule implemented at the end of the run) in a certain sense. The histograms in Figures 4(c) and 4(d) *are* in agreement with the solid line marginals. Certainly, there are no formal guarantees that this should happen. On the other hand, in Section 3 we can prove the corresponding claim for the stable AMOR algorithm.

This relabeling strategy seems to recover π_{SEED} better than the mere ordering of coordinates as suggested by the marginal plots in Figures 4(c) and 4(d) which are less skewed and now roughly centered at the correct values (0 and 2, respectively). However, using a diagonal covariance D_t also generates some distortion which results in a severely non-Gaussian, bimodal marginal in Figure 4(c). Because of these imperfections and due to the uncorrelated proposal, the autocorrelation in Figure 4(b) indicates, again, a much less efficient sampling than in the case of an optimal Metropolis chain targeting π_{SEED}.

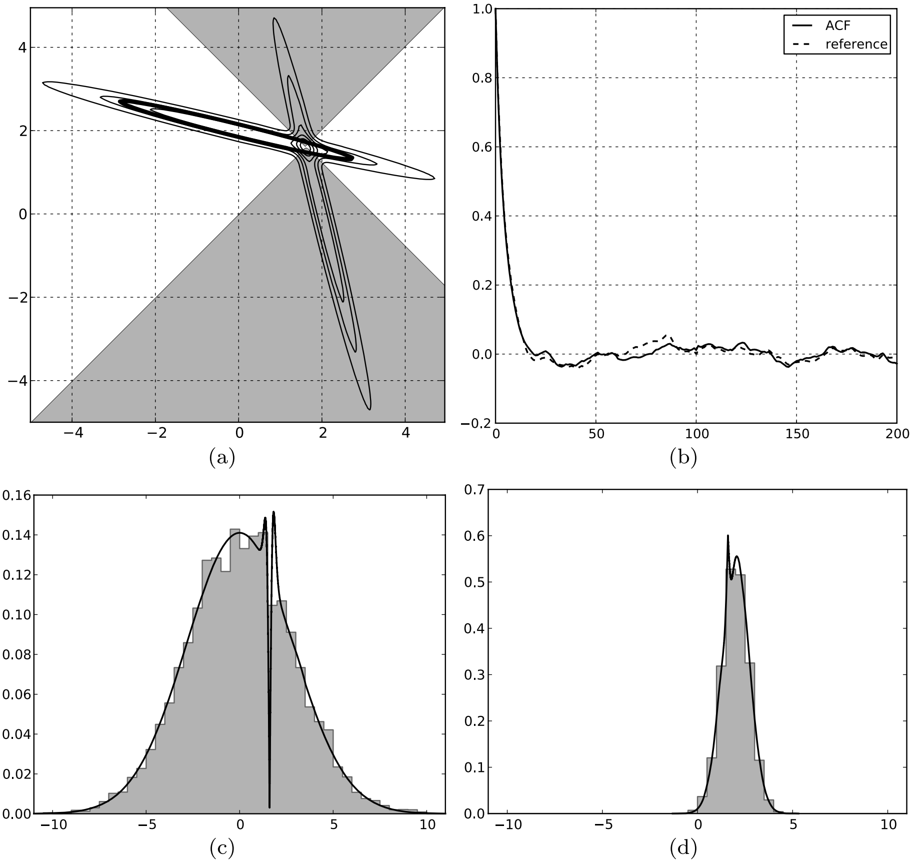
The significance of Celeux's algorithm is that its adaptive relabeling rule (2.3) makes it possible to resolve the permutation invariance problem in a nontrivial way which appears to be more adapted to the true geometry of the target. It is still not perfect, and, as suggested by [32], one should replace the diagonal covariance matrix in (2.3) by the full covariance matrix of the sample. However, [32] explored this idea only as a post-processing approach. A severe difficulty in this context is the computational cost: if T denotes the number of drawn samples and p is the number of permutations to which π is invariant, the required post-processing is a combinatorial problem with p^T possible relabelings. This eventually led [32] to consider a more tractable alternative instead. More importantly in our context, we have seen above (e.g., in Figure 2) that running an adaptive MCMC on the full permutation-invariant target may result in a poor mixing performance. To achieve both relevant relabeling and efficient adaptivity, the key idea of stable AMOR is to link the covariance of the proposal distribution and the covariance used for relabeling, which are proportional to each other in stable AMOR.

Figure 5 displays the results obtained using stable AMOR on our running example. Stable AMOR does separate $\mathbb{R}^2$ in two regions that respect the topology of the target much more closely than the approaches examined previously. Figure 5(a) indicates that the relabeled target is as Gaussian as possible among all partitionings based on a quadratic criterion of the form (2.1). The marginal histograms in Figures 5(c) and 5(d) now look almost Gaussian. They closely match the marginals of both the restricted distribution π' and the seed distribution π_{SEED} in Figure 1(b). Furthermore, the autocorrelation function of stable AMOR (Figure 5(b)) is as good as the reference autocorrelation function corresponding to an optimally tuned random walk Metropolis–Hastings algorithm targeting the seed Gaussian π_{SEED} in Figure 1(b). This perfect adaptation is possible because the sample covariance now matches the covariance of the target restricted to the unshaded region of the plane (Figure 5(a)).

On this example, the stable AMOR algorithm thus automatically achieves, without any tuning, a satisfactory result that cannot be obtained with any of the methods examined previously. Further examples of the behavior of stable AMOR are given in the supplemental article [6]. We are now ready to prove our main result which shows that, under suitable conditions, stable AMOR indeed asymptotically samples from the target distribution restricted to a region on which the marginals are identifiable, and that the sample mean and covariance converge to the corresponding moments of the restricted target.

3. Convergence results

We prove the convergence of stable AMOR under the following condition on π.



(a) (b)

(c) (d)

Figure 5. Results of stable AMOR. For details about the plots, see the caption of Figure 2.

Assumption 1. *π is a density w.r.t. the Lebesgue measure on $\mathbb{R}^d$, which is bounded and with compact support $\mathbb{X}$, and which is invariant to permutations in the group $\mathcal{P}$:*

$$\forall x \in \mathbb{X}, \forall P \in \mathcal{P}, \qquad \pi(Px) = \pi(x).$$

This section is organized as follows. We first describe which version of the stable AMOR algorithm we consider, and we show that it is an adaptive MCMC algorithm. We then characterize the limiting behavior of the sequence $(\theta_t)_{t\geq 0}$ (see Theorem 3.2) and address a strong law of large numbers for the samples $(X_t)_{t\geq 0}$, as well as the ergodicity of the sampler (see Theorems 3.3 and 3.4). All proofs are given in the Appendix.

We are interested in finding a subset V_θ of $\mathbb{X}$ of the form (2.2) such that the cells $(PV_\theta)_{P\in\mathcal{P}}$ cover $\mathbb{X}$. We will also ask that for any $P, Q \in \mathcal{P}$, $P \neq Q$, the Lebesgue measure of $PV_\theta \cap QV_\theta$

is null. Therefore, we choose the parameter set Θ as follows (see Lemma A.1 in the Appendix):

$$\Theta = \{(\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{C}_d^+ / \forall P \in \mathcal{P}^*, \Sigma^{-1}\mu \neq P\Sigma^{-1}\mu\}, \tag{3.1}$$

where $\mathcal{P}^* = \mathcal{P} \setminus \{\mathrm{Id}\}$. The set $\mathbb{R}^d \times \mathcal{C}_d^+$ is endowed with the scalar product $\langle (a, A), (b, B)\rangle = a^T b + \operatorname{Trace}(A^T B)$. We will use the same notation $\|\cdot\|$ for the norm induced by this scalar product, for the Euclidean norm on $\mathbb{R}^d$, and for the norm $\|A\| = \operatorname{Trace}(A^T A)^{1/2}$ on $d \times d$ real matrices.

Since we want to drive the parameter toward the set Θ, we address the convergence of the stable AMOR when $\alpha > 0$ and the penalty term is given by

$$\mathrm{Pen}_{t,1} = -\sum_{P \in \mathcal{P}^*} \frac{1}{\|(I-P)\Sigma_t^{-1}\mu_t\|^4} U_P \Sigma_t^{-1}\mu_t, \tag{3.2}$$

$$\mathrm{Pen}_{t,2} = \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I-P)\Sigma_t^{-1}\mu_t\|^4} \bigl(\mu_t\mu_t^T \Sigma_t^{-1} U_P + U_P \Sigma_t^{-1}\mu_t\mu_t^T\bigr), \tag{3.3}$$

where $U_P = (I-P)^T(I-P)$. For the stabilization step, we consider the sequence of compact sets $(\mathcal{K}_{\delta_q})_{q \geq 0}$ where

$$\mathcal{K}_\delta = \Bigl\{(\mu, \Sigma) \in \Theta \colon \inf_{P \in \mathcal{P}^*} \bigl\|(I-P)\Sigma^{-1}\mu\bigr\| \geq \delta\Bigr\}, \tag{3.4}$$

and $(\delta_q)_{q \geq 0}$ is any decreasing positive sequence such that $\lim_{q\to\infty} \delta_q = 0$ and $\mathcal{K}_{\delta_0}$ is not empty.

Stable AMOR can be cast into the family of adaptive MCMC algorithms, in which the updating rule of the design parameter relies on a stochastic approximation scheme. Adaptive MCMC can be described as follows: given a family of transition kernels $(P_\theta)_{\theta \in \Theta}$, the algorithm produces a $(\mathbb{X} \times \Theta)$-valued process $((X_t, \theta_t))_{t \geq 0}$ such that the conditional distribution of X_t given its past history $X_1, \ldots, X_{t-1}$ is given by the transition kernel $P_{\theta_{t-1}}(X_{t-1}, \cdot)$. This algorithm is designed so that when t tends to infinity, the distribution of X_t converges to the invariant distribution of the kernel P_{θ_t}. Sufficient conditions for the convergence of such adaptive procedures were recently proposed by [13,27]. In particular, [27] provided sufficient conditions in terms of the so-called containment condition and diminishing adaptation. Furthermore, [13] showed that when each transition kernel P_θ has its own invariant distribution π_θ, an additional condition on the convergence of these distributions is also required. We prove below that in our settings, each transition kernel of stable AMOR has its own invariant distribution; and this additional condition is satisfied as soon as $(\theta_t)_{t \geq 0}$ converges almost surely. In order to establish this property, we will resort to convergence results for stochastic approximation algorithms.

As a preliminary step for the convergence of stable AMOR, the stability and the convergence of the design parameter sequence $(\theta_t)_{t \geq 0}$ is established. Sufficient conditions for the convergence of stochastic approximation procedures rely on the existence of a (sufficiently regular) Lyapunov function on Θ, on the behavior of the mean field at the boundary of the parameter set Θ, and on the magnitude of the step-size sequence $(\gamma_t)_{t \geq 0}$.

The compactness assumption (Assumption 1) makes it simpler to analyze the limiting behavior of the algorithm. The noncompact case is far more technical and will not be addressed in this paper; see, e.g., [13] (respectively [1], Section 3) for examples of convergence of adaptive MCMC (respectively a stochastic approximation procedure) when the support of π is not compact (respectively when the controlled Markov chain dynamics is not compactly supported).

Let us prove that stable AMOR is an adaptive MCMC algorithm. For any $\theta \in \Theta$, define the transition kernel P_θ on $(\mathbb{X}, \mathcal{X})$ by

$$P_\theta(x, A) = \int_{A \cap V_\theta} \alpha_\theta(x, y) q_\theta(x, y)\,\mathrm{d}y + \mathbb{1}_A(x) \int_{V_\theta} \bigl(1 - \alpha_\theta(x, z)\bigr) q_\theta(x, z)\,\mathrm{d}z, \tag{3.5}$$

where V_θ is given by (2.2),

$$\alpha_\theta(x, y) = 1 \wedge \frac{\pi(y) q_\theta(y, x)}{\pi(x) q_\theta(x, y)} \tag{3.6}$$

and

$$q_\theta(x, y) = \sum_{P \in \mathcal{P}} \mathcal{N}(Py | x, c\Sigma). \tag{3.7}$$

For $\theta \in \Theta$, define also

$$\pi_\theta = |\mathcal{P}| \mathbb{1}_{V_\theta} \pi. \tag{3.8}$$

The following proposition shows that $q_\theta(x, \cdot)$ is a density on V_θ and, the distribution π_θ given by (3.8) is invariant for the transition kernel P_θ. It also establishes that stable AMOR is an adaptive MCMC algorithm: given (X_{t-1}, θ_{t-1}), X_t is obtained by one iteration of a random-walk Metropolis–Hastings algorithm with proposal $q_{\theta_{t-1}}$ and invariant distribution $\pi_{\theta_{t-1}}$.

Proposition 3.1. *Under Assumption 1, the following assertions hold*:

(1) *For any $\theta \in \Theta$ and $x \in \mathbb{X}$, $\int_{V_\theta} q_\theta(x, y)\,\mathrm{d}y = 1$.*
(2) *For any $\theta \in \Theta$, $\pi_\theta P_\theta = \pi_\theta$ and for any $x \in V_\theta$, $P_\theta(x, V_\theta) = 1$.*
(3) *Let $(\theta_t, X_t)_{t \geq 0}$ be given by Algorithm 1. Conditionally on $\sigma(X_0, \theta_0, X_1, \theta_1, \ldots, X_{t-1}, \theta_{t-1})$, the distribution of X_t is $P_{\theta_{t-1}}(X_{t-1}, \cdot)$.*

Note that the proof of Proposition 3.1 is independent of the update scheme of $(\theta_t)_{t \geq 0}$, which makes the proposition valid whatever the choice of $\alpha\,\mathrm{Pen}_{t,i}$.

Denote by $\mathcal{S}_d$ the set of $d \times d$ symmetric real matrices. Let $\alpha > 0$ be fixed and define $H : \mathbb{X} \times \Theta \to \mathbb{R}^d \times \mathcal{S}_d$ by

$$H(x, \theta) = \bigl(H_\mu(x, \theta), H_\Sigma(x, \theta)\bigr) \tag{3.9}$$

where

$$\begin{aligned} H_\mu(x, \theta) &= x - \mu - \alpha \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^4} U_P \Sigma^{-1} \mu, \\ H_\Sigma(x, \theta) &= (x - \mu)(x - \mu)^T - \Sigma \\ &\quad + \alpha \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^4} \bigl(\mu\mu^T \Sigma^{-1} U_P + U_P \Sigma^{-1} \mu\mu^T\bigr). \end{aligned}$$

Let

$$\mu_{\pi_\theta} = \int x\pi_\theta(x)\,\mathrm{d}x, \tag{3.10}$$

$$\Sigma_{\pi_\theta} = \int (x-\mu_{\pi_\theta})(x-\mu_{\pi_\theta})^T\pi_\theta(x)\,\mathrm{d}x, \tag{3.11}$$

be the expectation and covariance matrix of π_θ, respectively. Define the *mean field* $h:\Theta \to \mathbb{R}^d \times \mathcal{S}_d$ by

$$h(\theta) = \bigl(h_\mu(\theta), h_\Sigma(\theta)\bigr), \tag{3.12}$$

where

$$\begin{aligned} h_\mu(\theta) &= \mu_{\pi_\theta} - \mu - \alpha \sum_{P\in\mathcal{P}^*} \frac{1}{\|(I-P)\Sigma^{-1}\mu\|^4} U_P\Sigma^{-1}\mu, \\ h_\Sigma(\theta) &= \Sigma_{\pi_\theta} - \Sigma + (\mu_{\pi_\theta}-\mu)(\mu_{\pi_\theta}-\mu)^T \\ &\quad + \alpha \sum_{P\in\mathcal{P}^*} \frac{1}{\|(I-P)\Sigma^{-1}\mu\|^4}\bigl(\mu\mu^T\Sigma^{-1}U_P + U_P\Sigma^{-1}\mu\mu^T\bigr). \end{aligned}$$

The key ingredient for the proof of the convergence of the sequence $(\theta_t)_{t\geq 0}$ is the existence of a Lyapunov function w for the mean field h: we prove in the Appendix (see Lemma A.2) that the function $w:\Theta \to \mathbb{R}_+$, defined by

$$w(\theta) = -\int \log\mathcal{N}(x|\theta)\pi_\theta(x)\,\mathrm{d}x + \frac{\alpha}{2}\sum_{P\in\mathcal{P}^*}\frac{1}{\|(I-P)\Sigma^{-1}\mu\|^2}, \tag{3.13}$$

is continuously differentiable on Θ and satisfies $\langle\nabla w, h\rangle \leq 0$. In addition, $\langle\nabla w(\theta), h(\theta)\rangle = 0$ if and only if θ is in the set

$$\mathcal{L} = \bigl\{\theta\in\Theta : h(\theta) = 0\bigr\} = \bigl\{\theta\in\Theta : \nabla w(\theta) = 0\bigr\}. \tag{3.14}$$

The convergence of the sequence $(\theta_t)_{t\geq 0}$ is proved by verifying the sufficient conditions for the convergence of the stochastic approximation for Lyapunov stable dynamics given in [1]. The first step is to prove that the sequence is bounded with probability one: we prove that, almost surely, the number of projections ψ is finite so that the projection mechanism (steps 15 to 17 in Algorithm 1) never occurs after a (random) finite number of iterations. We then prove the convergence of the stable sequence. To achieve that goal, following the same lines as in [1], we make the following assumption.

Assumption 2. *Let $\mathcal{L}$ be given by* (3.14). *There exists $M_\star > 0$ such that $\mathcal{L} \subset \{\theta : w(\theta) \leq M_\star\}$, and $w(\mathcal{L})$ has an empty interior.*

For $x \in \mathbb{R}^d$ and $A \subset \mathbb{R}^d$, define $\mathrm{d}(x, A) = \inf_{a\in A}\|x-a\|$. The following result is proved in the Appendix.

Theorem 3.2. *Let $\beta \in (1/2, 1]$ and $\gamma_\star > 0$. Let $(\theta_t)_{t\geq 0}$ be the sequence produced by Algorithm 1 with $\alpha > 0$, the penalty term given by (3.2) and (3.3), the compact sets $\mathcal{K}_\delta$ given by (3.4) and $\gamma_t \sim \gamma_\star t^{-\beta}$ when $t \to +\infty$. Under Assumptions 1 and 2,*

(1) *The sequence $(\theta_t)_{t\geq 0}$ is stable: almost surely, there exist $M > 0$ and $t_\star > 0$ such that for any $t \geq t_\star$, $\theta_t \in \{\theta \in \Theta : w(\theta) \leq M\}$. In addition, the number of projections is finite almost surely.*

(2) *Almost surely, $(w(\theta_t))_t$ converges to $w^\star \in w(\mathcal{L})$ and $\limsup_t \mathrm{d}(\theta_t, \mathcal{L}_{w^\star}) \to 0$ where $\mathcal{L}_{w^\star} = \{\theta \in \mathcal{L}, w(\theta) = w^\star\}$.*

Theorem 3.2 states the convergence of $(\theta_t)_{t\geq 0}$ to the set $\mathcal{L}$ of the zeros of h; note that this set neither depends on the initial values (θ_0, X_0) nor on other design parameters. In our experiments, we always observed pointwise convergence. This is a hint that, in practice, $\mathcal{L}$ does not contain accumulation points. We now state a strong law of large numbers for the samples $(X_t)_{t\geq 0}$.

Theorem 3.3. *Let $\beta \in (1/2, 1]$, $\gamma_\star > 0$, and $\theta^\star \in \mathcal{L}$. Let $(X_t, \theta_t)_{t\geq 0}$ be the sequence generated by Algorithm 1 with $\alpha > 0$, the penalty term given by (3.2) and (3.3), the compact sets $\mathcal{K}_\delta$ given by (3.4) and $\gamma_t \sim \gamma_\star t^{-\beta}$ when $t \to +\infty$. Under Assumptions 1 and 2, on the set $\{\lim_t \theta_t = \theta^\star\}$, almost surely,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \pi_{\theta^\star}(f),$$

for any bounded function f.

It is easily checked (by using Lemma A.1) that, when the function f is invariant to permutations in the group $\mathcal{P}$, $\pi_\theta(f) = \pi(f)$ for any $\theta \in \Theta$. A careful reading of the proof of this theorem (see the remark in Section A.6) shows that for such a function f, when the sequence $(\theta_t)_{t\geq 0}$ is stable but does not necessarily converge, it holds, almost surely,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \pi(f).$$

Finally, Theorem 3.4 yields the ergodicity of stable AMOR.

Theorem 3.4. *Let $\beta \in (1/2, 1]$, $\gamma_\star > 0$, and $\theta^\star \in \mathcal{L}$. Let $(X_t, \theta_t)_{t\geq 0}$ be the sequence generated by Algorithm 1 with $\gamma_t \sim \gamma_\star t^{-\beta}$ when $t \to +\infty$. Under Assumptions 1 and 2,*

$$\lim_{t\to\infty} \sup_{\|f\|_\infty \leq 1} \left| \mathbb{E}\left[f(X_t) \mathbb{1}_{\lim_q \theta_q = \theta^\star}\right] - \pi_{\theta^\star}(f) \mathbb{P}\Big(\lim_q \theta_q = \theta^\star\Big) \right| = 0.$$

Here again, a careful reading of the proof shows that when f is invariant to permutations in the group $\mathcal{P}$, we have (see the Remark in Section A.7)

$$\lim_{t\to\infty} \left| \mathbb{E}\left[f(X_t)\right] - \pi(f) \right| = 0.$$

The expression (3.13) of w provides insight into the links between relabeling and vector quantization [14]. The first term is similar to a distortion measure in vector quantization as noted in [5]. It can also be seen as the cross-entropy between π_θ and a Gaussian with parameters θ. The second term in (3.13) is similar to a barrier penalty in continuous optimization [8]. From this perspective, Algorithm 1 can be seen as a constrained optimization procedure that minimizes the cross-entropy. In that sense, if $\theta^\star$ denotes a solution to this optimization problem, the *relabeled target* $\pi_{\theta^\star} \propto \mathbb{1}_{V_{\theta^\star}} \pi$ is the restriction of π to one of its symmetric modes $V_{\theta^\star}$ that looks as Gaussian as possible among all such restrictions.

Vector quantization algorithms have already been investigated using stochastic approximation tools [21]. However, stability was guaranteed in previous work by making strong assumptions on the trajectories of the process $(\theta_t)_{t\geq 0}$, such as in [21], Theorem 32; see also [21], Results 33–37 and Remark 38. These assumptions ensure that (θ_t) stays asymptotically away from sets where the function used elsewhere as a Lyapunov function is not differentiable. In this paper, we adopt a different strategy by introducing the modifications of the stable AMOR algorithm and adding a barrier term in the definition of our Lyapunov function (3.13) that penalizes these sets. One of the contributions of this paper is to show that this penalization strategy leads to a stable algorithm, without requiring any strong assumption on (θ_t).

4. Conclusion

We illustrated stable AMOR, an adaptive Metropolis algorithm with online relabeling and proved that a strong law of large numbers holds for this sampler. The stable version of AMOR, given in Algorithm 1, coincides with AMOR (proposed in [5]) when the penalty coefficient α is set to zero and no reprojection is performed. In practice, we observed that stable AMOR is very robust to the choice of α. Figure 6 illustrates this robustness on the toy example of Section 2.2.

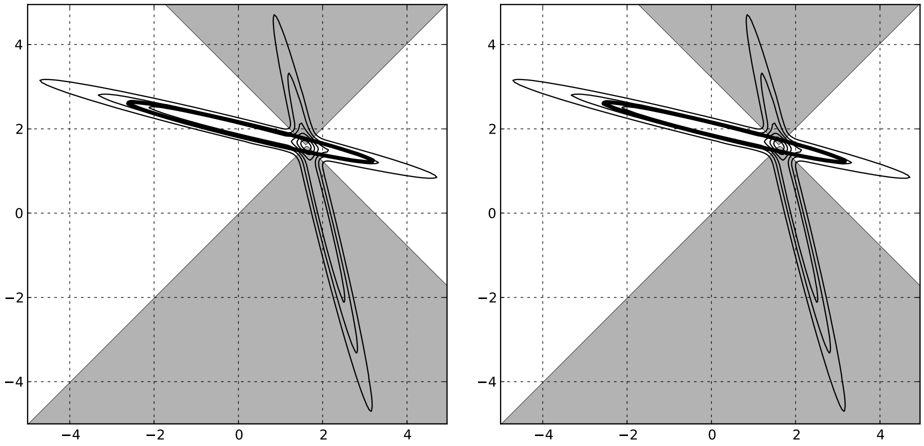


Figure 6. Results of stable AMOR on the toy example of Section 2.2, with $\delta_q = 10^{-2}2^{-q}$, and $\alpha = 10^{-3}$ (left) and $\alpha = 1$ (right).

Our algorithm adapts both its proposal and its target on the fly, which makes it a turn-key algorithm. Our results lead to a sound characterization of the target of stable AMOR that does not depend on the initialization of the algorithm nor on the user. This is the first theoretical analysis of an online relabeling algorithm to our knowledge. The proof further shows how relabeling is related to vector quantization. Unlike previous work on stochastic approximation schemes for vector quantization, we make no strong assumptions on the trajectories of the process considered, rather, we ensure that the appropriate constraint is satisfied by introducing penalization directly into the stochastic approximation framework.

We now examine possible directions for future work. First, following our analysis in Section 3, the question of the control of the convergence of stable AMOR arises, and proving a central limit theorem would be a natural next step. Second, the online nature of stable AMOR makes it cheaper than its post-processing counterpart, but it still requires to sweep over all elements of $\mathcal{P}$ at each iteration. This is prohibitive in problems with large $|\mathcal{P}|$, such as additive models with a large number of components. In future work, we will concentrate on algorithmic modifications to reduce this cost, potentially inspired by *probabilistic* relabeling algorithms [17,31], while conserving our theoretical results. Third, we are interested in extending stable AMOR to trans-dimensional problems, such as mixtures with an unknown number of components. Reversible jump MCMC (RJMCMC; [15]) also suffers from label-switching and inferential difficulties. We will study algorithms that combine RJMCMC and stable AMOR.

Appendix: Proofs

Throughout the proof, let $\Delta_\pi > 0$ be such that

$$x \in \mathbb{X} \Rightarrow \|x\| \leq \Delta_\pi. \tag{A.1}$$

For any function $f : D \to \mathbb{R}$, we will denote by $\|f\|_\infty = \sup_{x\in D} |f(x)|$.

A.1. Preliminary results

We restate (with a slight adaptation) Lemma 1 of the supplementary material from [5] that we will use extensively.

Lemma A.1. *Let $\theta \in \Theta$.*

(1) *The sets $\{PV_\theta, P \in \mathcal{P}\}$ cover $\mathbb{X}$, and for any $P, Q \in \mathcal{P}$ such that $P \neq Q$, the Lebesgue measure of $PV_\theta \cap QV_\theta$ is zero.*

(2) *Let λ be a measure on $(\mathbb{X}, \mathcal{X})$ with a density w.r.t. the Lebesgue measure. Furthermore, let λ be such that for any $A \in \mathcal{X}$ and $\mathcal{P} \in \mathcal{P}$, $\lambda(PA) = \lambda(A)$. Then $\lambda(V_\theta) = \lambda(\mathbb{X})/|\mathcal{P}|$.*

Proof. The proof is along the lines of Lemma 1 of the supplementary material in [5], and it is thus omitted. It can be found in the supplemental article to the present paper [6]. □

A.2. Proof of Proposition 3.1

(1) By the definition (3.1) of Θ and Lemma A.1, $\forall\theta \in \Theta, x \in \mathbb{X}$, it holds that

$$\int_{V_\theta} q_\theta(x,y)\,\mathrm{d}y = \sum_{P\in\mathcal{P}} \int_{V_\theta} \mathcal{N}(Py|x,c\Sigma)\,\mathrm{d}y = 1.$$

(2) Let $(X_t)_{t\geq 0}$ and $(\theta_t)_{t\geq 0}$ be the random processes defined by Algorithm 1. Let $\mathcal{F}_t = \sigma(X_0,\theta_0,\dots,X_t,\theta_t)$. We prove that for any measurable positive function f,

$$\mathbb{E}\bigl[f(X_t)|\mathcal{F}_{t-1}\bigr] = \int f(x_t) P_{\theta_{t-1}}(X_{t-1},x_t)\,\mathrm{d}x_t, \qquad \text{w.p.1.}$$

Let f be measurable and positive. Let $(\tilde{P},\tilde{X})$ be the r.v. defined by steps 5 and 6. Let U be a uniform r.v. independent of $\sigma(X_0,\theta_0,\dots,X_{t-1},\theta_{t-1},\tilde{P},\tilde{X})$. By construction, it holds that

$$\begin{aligned}\mathbb{E}\bigl[f(X_t)|\mathcal{F}_{t-1}\bigr] &= \mathbb{E}\bigl[f(\tilde{P}\tilde{X})\bigl(1-\alpha_{\theta_{t-1}}(X_{t-1},\tilde{P}\tilde{X})\bigr)|\mathcal{F}_{t-1}\bigr] \\ &\quad + f(X_{t-1})\mathbb{E}\bigl[\bigl(1-\alpha_{\theta_{t-1}}(X_{t-1},\tilde{P}\tilde{X})\bigr)|\mathcal{F}_{t-1}\bigr].\end{aligned} \tag{A.2}$$

Now note that the projection mechanism (steps 15 to 17 of Algorithm 1) guarantees that $\theta_{t-1} \in \Theta$ with probability 1. By Lemma A.1, $\theta\in\Theta$ implies $\mathbb{X} = \bigcup_P (PV_\theta)$ and

$$\forall P,Q\in\mathcal{P} \quad \text{such that} \quad P\neq Q, \mathrm{Leb}(PV_\theta \cap QV_\theta) = 0.$$

Thus, for any measurable and bounded function $\varphi:\mathbb{X}\times\Theta\to\mathbb{R}$, we have

$$\int_{\mathbb{X}} \varphi(x,\theta)\,\mathrm{d}x = \sum_{Q\in\mathcal{P}} \int_{QV_\theta\cap(\cup_{R\neq Q} RV_\theta)^c} \varphi(x,\theta)\,\mathrm{d}x.$$

Applying this decomposition to the second term in the RHS of (A.2) yields

$$\begin{aligned}&\mathbb{E}\bigl[f(\tilde{P}\tilde{X})\mathbb{1}_{U\leq\alpha_{\theta_{t-1}}(X_{t-1},\tilde{P}\tilde{x})}|\mathcal{F}_{t-1}\bigr] \\ &\quad = \sum_{P\in\mathcal{P}} \int h(Px)\frac{1}{N(x,\theta_{t-1})}\mathbb{1}_{V_{\theta_{t-1}}}(Px)\mathcal{N}(x|X_{t-1},c\Sigma_{t-1})\,\mathrm{d}x \\ &\quad = \sum_{P,Q\in\mathcal{P}} \int_{QV_{\theta_{t-1}}\cap(\bigcup_{R\neq Q} RV_{\theta_{t-1}})^c} h(Px)\frac{1}{N(x,\theta_{t-1})}\mathbb{1}_{V_{\theta_{t-1}}}(Px)\mathcal{N}(x|X_{t-1},c\Sigma_{t-1})\,\mathrm{d}x,\end{aligned}$$

where $N(x,\theta) = |\{Q\in\mathcal{P}/Qx\in V_\theta\}|$. Using Lemma A.1 again,

$$\theta\in\Theta, \qquad x\notin \bigcup_{P\neq Q}(PV_\theta\cap QV_\theta) \Rightarrow N(x,\theta)=1,$$

and thus

$$\mathbb{E}\big[f(\tilde{P}\tilde{X})\mathbb{1}_{U\le\alpha_{\theta_{t-1}}(X_{t-1},\tilde{P}\tilde{X})}|\mathcal{F}_{t-1}\big] = \sum_{P\in\mathcal{P}}\int h(y)\mathbb{1}_{V_{\theta_{t-1}}}(y)\mathcal{N}\big(P^{-1}y|X_{t-1},c\Sigma_{t-1}\big)\,\mathrm{d}y$$

$$= \int_{V_{\theta_{t-1}}} h(y)q_{\theta_{t-1}}(X_{t-1},y)\,\mathrm{d}y,$$

where in the last step we used the fact that $\mathcal{P}$ is a group. Similarly,

$$\mathbb{E}\big[\big(1-\alpha_{\theta_{t-1}}(X_{t-1},\tilde{P}\tilde{X})\big)|X_0,\theta_0,\dots,X_{t-1},\theta_{t-1}\big]$$

$$= \int_{V_{\theta_{t-1}}} \big(1-\alpha_{\theta_{t-1}}(X_{t-1},y)\big)q_{\theta_{t-1}}(X_{t-1},y)\,\mathrm{d}y;$$

and this concludes the proof.

(3) This proof amounts to check the classical detailed balance condition [24], and it is thus omitted. It is included in the supplemental article [6].

A.3. The Lyapunov function

Lemma A.2 establishes the existence of a Lyapunov function for the mean field h given by (3.12).

Lemma A.2. *Under Assumption 1, the mean field h is continuous on Θ, the function w defined by (3.13) is $\mathcal{C}^1$ on Θ and*

(1) $\nabla_\mu w(\theta) = -\Sigma^{-1}h_\mu(\theta)$ *and* $\nabla_\Sigma w(\theta) = -\frac{1}{2}\Sigma^{-1}h_\Sigma(\theta)\Sigma^{-1}$.

(2) $\langle\nabla w(\theta), h(\theta)\rangle \le 0$ *on* Θ *and* $\langle\nabla w(\theta), h(\theta)\rangle = 0$ *iff* $\theta\in\mathcal{L}$.

(3) *For any* $M > 0$, *the level set*

$$\mathcal{W}_M = \big\{\theta\in\Theta : w(\theta)\le M\big\} \tag{A.3}$$

is a compact subset of Θ, *and there exist* $\delta_1, \delta_2 > 0$ *such that*

$$\inf_{\theta\in\mathcal{W}_M}\inf_{P\in\mathcal{P}^*}\big\|(I-P)\Sigma^{-1}\mu\big\| \ge \delta_1 \tag{A.4a}$$

and

$$\inf_{\theta\in\mathcal{W}_M}\lambda_{\min}(\Sigma)\ge\delta_2, \tag{A.4b}$$

where $\lambda_{\min}(\Sigma)$ *denotes the minimal eigenvalue of the real symmetric matrix* Σ.

Remark A.3. As a consequence of Lemma A.2, observe that for any $M > 0$, there exists $\delta > 0$ such that $\mathcal{W}_M \subseteq \mathcal{K}_\delta$, where $\mathcal{K}_\delta$ is defined in (3.4).

Proof. *(Continuity of h.)* This proof is a straightforward application of Lebesgue's dominated convergence theorem, and it is thus omitted. It is included in the supplemental article, a link to which can be found at the end of this paper [6].

(w is C^1 on Θ.) It is shown in [5], Proposition 3 of the supplementary material, that the first term in the RHS of (3.13) is continuously differentiable on Θ. Since $\|(I-P)\Sigma^{-1}\mu\| \neq 0$ for any $P \in \mathcal{P}^*$ and $(\mu, \Sigma) \in \Theta$, the second term in the RHS of (3.13) is continuously differentiable on Θ. By [5], Proposition 3 of the supplementary material, it holds for any $\theta = (\mu, \Sigma) \in \Theta$ that

$$\begin{aligned}
\nabla_\mu w(\theta) &= -\Sigma^{-1}(\mu_{\pi_\theta} - \mu) + \alpha \sum_P \frac{1}{\|(I-P)\Sigma^{-1}\mu\|^4} \Sigma^{-1} U_P \Sigma^{-1} \mu \\
&= -\Sigma^{-1} h_\mu(\theta), \\
\nabla_\Sigma w(\theta) &= -\frac{1}{2} \Sigma^{-1}\bigl(\Sigma_{\pi_\theta} - \Sigma + (\mu - \mu_{\pi_\theta})(\mu - \mu_{\pi_\theta})^T\bigr)\Sigma^{-1} \\
&\quad - \frac{\alpha}{2} \sum_P \frac{1}{\|(I-P)\Sigma^{-1}\mu\|^4} \Sigma^{-1}\bigl(\mu\mu^T \Sigma^{-1} U_P\bigr)\Sigma^{-1} + U_P \Sigma^{-1} \mu\mu^T \\
&= -\frac{1}{2}\Sigma^{-1} h_\Sigma(\theta)\Sigma^{-1}.
\end{aligned}$$

Hence, upon noting that $h_\Sigma(\theta)$ and Σ^{-1} are symmetric,

$$\begin{aligned}
\bigl\langle \nabla w(\theta), h(\theta) \bigr\rangle &= -h_\mu(\theta)^T \Sigma^{-1} h_\mu(\theta) - \tfrac{1}{2}\operatorname{Trace}\bigl(\Sigma^{-1} h_\Sigma(\theta) \Sigma^{-1} h_\Sigma(\theta)\bigr) \\
&= -h_\mu(\theta)^T \Sigma^{-1} h_\mu(\theta) - \tfrac{1}{2}\operatorname{Trace}\bigl(\Sigma^{-1/2} h_\Sigma(\theta) \Sigma^{-1} h_\Sigma(\theta) \Sigma^{-1/2}\bigr).
\end{aligned}$$

The first term of the RHS is negative since $\Sigma \in \mathcal{C}_d^+$ and the second term is negative since $(A, B) \mapsto \operatorname{Trace}(A^T B)$ is a scalar product. Therefore, $\langle \nabla w(\theta), h(\theta) \rangle \leq 0$ with equality if and only if f $\theta \in \mathcal{L}$.

($\mathcal{W}_M$ is compact.) We prove (A.4a). By the definition (3.13) of w, for any $\theta \in \mathcal{W}_M$, we have

$$-\int \log \mathcal{N}(x|\theta) \pi_\theta(x) \, \mathrm{d}x + \frac{\alpha}{2} \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I-P)\Sigma^{-1}\mu\|^2} \leq M.$$

In particular, the first term in the LHS is a cross-entropy, and it is thus nonnegative (alternatively, see [5], Proposition 1 of the supplementary material). Consequently, for any $\theta \in \mathcal{W}_M$, we have

$$\sum_{P \in \mathcal{P}^*} \frac{1}{\|(I-P)\Sigma^{-1}\mu\|^2} \leq \frac{2M}{\alpha}.$$

This yields $\|(I-P)\Sigma^{-1}\mu\|^2 \geq \frac{\alpha}{2M}$ for any $P \in \mathcal{P}^*$, thus concluding the proof of (A.4a).

We now prove (A.4b). Let $\theta = (\mu, \Sigma) \in \mathcal{W}_M$. Denote by $(\lambda_i(\Sigma))_{i \leq d}$ the eigenvalues of Σ. Since Σ is symmetric, there exist $d \times d$ matrices Q_θ, Λ_θ such that $\Sigma = Q_\theta \Lambda_\theta Q_\theta^T$, Q_θ is orthogonal, and $\Lambda_\theta = \operatorname{diag}(\lambda_i(\Sigma))$. Then

$$\begin{aligned}2M &\geq 2w(\theta) \geq -2\int \log \mathcal{N}(x|\theta)\pi_\theta(x)\,\mathrm{d}x \\ &= d\log(2\pi) + \log\det\Sigma + (\mu_{\pi_\theta} - \mu)^T\Sigma^{-1}(\mu_{\pi_\theta} - \mu) + \operatorname{Trace}\bigl(\Sigma^{-1}\Sigma_{\pi_\theta}\bigr) \qquad \text{(A.5)} \\ &\geq \sum_{i=1}^d \log\lambda_i(\theta) + 0 + \operatorname{Trace}\bigl(\Sigma^{-1}\Sigma_{\pi_\theta}\bigr).\end{aligned}$$

Set $b_i(\theta) = (Q_\theta^T\Sigma_{\pi_\theta}Q_\theta)_{ii}$. Then

$$\operatorname{Trace}\bigl(\Sigma^{-1}\Sigma_{\pi_\theta}\bigr) = \operatorname{Trace}\bigl(Q_\theta\Lambda_\theta^{-1}Q_\theta^T\Sigma_{\pi_\theta}\bigr) = \operatorname{Trace}\bigl(Q_\theta^T\Sigma_{\pi_\theta}Q_\theta\Lambda_\theta^{-1}\bigr) = \sum_{i=1}^d \frac{b_i(\theta)}{\lambda_i(\theta)}. \qquad \text{(A.6)}$$

Therefore, for any $\theta \in \mathcal{W}_M$,

$$\sum_{i=1}^d \log\lambda_i(\theta) + \frac{b_i(\theta)}{\lambda_i(\theta)} \leq 2M. \qquad \text{(A.7)}$$

We now prove that for any i, $\inf_{\mathcal{W}_M} b_i > 0$. This property, combined with (A.7), will conclude the proof of (A.4b). Let $\varepsilon > 0$ be such that $2^d\varepsilon\|\pi\|_\infty\Delta_\pi^{d-1} < |\mathcal{P}|$, and for $v \in \{x \in \mathbb{R}^d : \|x\| = 1\}$, let

$$B_\varepsilon^v(\theta) = \bigl\{x \in \operatorname{Supp}(\pi) \cap V_\theta : \bigl|\langle x - \mu_{\pi_\theta}, v\rangle\bigr| \leq \varepsilon\bigr\}. \qquad \text{(A.8)}$$

Note that by Assumption 1,

$$\pi\bigl(B_\varepsilon^v(\theta)\bigr) \leq \|\pi\|_\infty \operatorname{Leb}\bigl(B_\varepsilon^v(\theta)\bigr) \leq 2^d\varepsilon\|\pi\|_\infty\Delta_\pi^{d-1}.$$

Then, by definition of ε,

$$\pi\bigl(V_\theta \setminus B_\varepsilon^v(\theta)\bigr) \geq |\mathcal{P}| - 2^d\varepsilon\|\pi\|_\infty\Delta_\pi^{d-1} > 0. \qquad \text{(A.9)}$$

Now, if (e_i) denotes the canonical basis of $\mathbb{R}^d$, then

$$\begin{aligned}b_i(\theta) &= |\mathcal{P}|e_i^TQ_\theta^T\left(\int_{V_\theta}(x - \mu_{\pi_\theta})(x - \mu_{\pi_\theta})^T\pi(x)\,\mathrm{d}x\right)Q_\theta e_i \\ &= |\mathcal{P}|\int_{V_\theta}(Q_\theta e_i)^T(x - \mu_{\pi_\theta})(x - \mu_{\pi_\theta})^TQ_\theta e_i\pi(x)\,\mathrm{d}x \\ &= |\mathcal{P}|\int_{V_\theta}\langle x - \mu_{\pi_\theta}, Q_\theta e_i\rangle^2\pi(x)\,\mathrm{d}x \qquad \text{(A.10)} \\ &\geq |\mathcal{P}|\int_{V_\theta\setminus B_\varepsilon^{Q_\theta e_i}(\theta)}\langle x - \mu_{\pi_\theta}, Q_\theta e_i\rangle^2\pi(x)\,\mathrm{d}x \\ &\geq \varepsilon^2|\mathcal{P}|\pi\bigl(V_\theta \setminus B_\varepsilon^{Q_\theta e_i}(\theta)\bigr),\end{aligned}$$

where the last inequality follows from the definition (A.8) of $B_\varepsilon^{Q_\theta e_i}(\theta)$. Thus, by (A.9), $b_i(\theta)$ is bounded away from zero on $\mathcal{W}_M$.

As w is continuous on Θ, $\{\theta \in \Theta, w(\theta) \leq M\}$ is closed. From (A.4b), (A.5) and Assumption 1, $\mu \mapsto (\mu_{\pi_\theta} - \mu)^T \Sigma^{-1} (\mu_{\pi_\theta} - \mu)$ is bounded on $\mathcal{W}_M$. In addition, (A.5), (A.6) and (A.10) imply that $\Sigma \mapsto \log \det \Sigma$ is bounded on $\mathcal{W}_M$. These properties combined with (A.4b) imply that $\mathcal{W}_M$ is bounded. Hence, $\mathcal{W}_M$ is compact. □

A.4. Regularity in θ of the Poisson solution

Lemma A.4.

(1) *For any $M > 0$, there exists $\rho \in (0, 1)$ such that for any $x \in \mathbb{X}$ and any $\theta \in \mathcal{W}_M$, $\|P_\theta^n(x, \cdot) - \pi_\theta\|_{\mathrm{TV}} \leq 2(1 - \rho)^n$.*

(2) *Under Assumption 1, for any $\theta \in \Theta$, there exists a solution $\hat{H}_\theta$ of the Poisson equation, that is, $\hat{H}_\theta - P_\theta \hat{H}_\theta = H(\cdot, \theta) - \pi_\theta H(\cdot, \theta)$. Furthermore, for any $M > 0$,*

$$\sup_{\theta \in \mathcal{W}_M} \sup_{x \in \mathbb{X}} \bigl|\hat{H}_\theta(x)\bigr| < \infty. \tag{A.11}$$

Proof. (1) It is sufficient to prove that there exists $\rho \in (0, 1)$ such that for any $x \in \mathbb{X}$ and $\theta \in \mathcal{W}_M$, $P_\theta(x, \cdot) \geq \rho \pi_\theta$ (see, e.g., [20], Theorem 16.2.4). By (3.5), for any $x \in \mathbb{X}$ and $A \in \mathcal{X}$, $P_\theta(x, A) \geq \int_{A \cap V_\theta} \alpha_\theta(x, y) q_\theta(x, y) \, \mathrm{d}y$. By Lemma A.2, there exists $a > 0$ such that for any $(\mu, \Sigma) \in \mathcal{W}_M$, any $m, z \in \mathbb{X}$, and any $P \in \mathcal{P}$, we have $\mathcal{N}(Pz | m, \Sigma) \geq a$. Thus, for any $\theta \in \mathcal{W}_M$ and $y \in V_\theta$, it holds that

$$\alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) \geq a |\mathcal{P}| \left(1 \wedge \frac{\pi(y)}{\pi(x)}\right) \mathbb{1}_{V_\theta}(y) \geq \frac{a}{\|\pi\|_\infty} \pi_\theta(y). \tag{A.12}$$

Thus, we have $P_\theta(x, \cdot) \geq \rho \pi_\theta$ for any $x \in \mathbb{X}$ and $\theta \in \mathcal{W}_M$ with $\rho = a / \|\pi\|_\infty$.

(2) By item (1),

$$\hat{H}_\theta(x) = \sum_n P_\theta^n \bigl(H(x, \theta) - \pi_\theta\bigl(H(\cdot, \theta)\bigr)\bigr)$$

exists and solves the Poisson equation. (A.11) trivially follows from item (1). □

Lemma A.5. *Let $M > 0$ and $\kappa \in (0, 1/2)$. Under Assumption 1, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \Theta$, it holds that*

$$\mathrm{Leb}(V_\theta \setminus V_{\theta'}) \leq C \bigl\|\theta - \theta'\bigr\|^{1 - 2\kappa}, \tag{A.13}$$

where $\mathrm{Leb}(A)$ *denotes the Lebesgue measure of the set A.*

Proof. We prove that there exist $\bar{C}, \bar{h} > 0$, such that for any $\theta \in \mathcal{W}_M$ and any $\theta' \in \Theta$ such that $\|\theta - \theta'\| \leq \bar{h}$, $\mathrm{Leb}(V_\theta \setminus V_{\theta'}) \leq \bar{C} \|\theta - \theta'\|^{1 - 2\kappa}$. Note that since $V_\theta \subset \mathbb{X}$ and since $\mathbb{X}$ is bounded, there exists $\check{C} > 0$ such that $\mathrm{Leb}(V_\theta \setminus V_{\theta'}) \leq \check{C}$. Therefore, (A.13) holds with $C = \bar{C} \vee \check{C} / \bar{h}^{1 - 2\kappa}$.

By Lemma A.2, w is uniformly continuous on $\mathcal{W}_{M+1}$, and there exists $h_0 > 0$ small enough for which

$$\left[\theta \in \mathcal{W}_M, \theta' \in \Theta, \left\|\theta - \theta'\right\| < h_0\right] \Rightarrow \forall u \in [0,1], \qquad \theta + u\left(\theta' - \theta\right) \in \mathcal{W}_{M+1}. \tag{A.14}$$

Let $\bar{h} \le h_0$. Let $\theta = (\mu, \Sigma) \in \mathcal{W}_M$ and $\theta' \neq \theta$ such that $\|\theta - \theta'\| \le \bar{h}$.

By definition of the set V_ϑ, for any $x \in V_\theta \setminus V_{\theta'}$, there exists $P \in \mathcal{P}^*$ such that $L_{\theta'}(x) - L_{\theta'}(P^T x) > 0$ and $L_\theta(x) - L_\theta(P^T x) \le 0$. Since $\vartheta \mapsto L_\vartheta(x) - L_\vartheta(P^T x)$ is continuous on $\mathcal{W}_{M+1}$, there exists $u \in [0,1]$ depending on x, θ, θ', and P such that $L_{\theta+u(\theta'-\theta)}(x) - L_{\theta+u(\theta'-\theta)}(P^T x) = 0$. Therefore,

$$V_\theta \setminus V_{\theta'} \subset \bigcup_{P \in \mathcal{P}^*} \mathcal{V}_P,$$

where

$$\mathcal{V}_P = \bigcup_{u \in [0,1]} \mathcal{Z}\left(L_{\theta+u(\theta'-\theta)}(\cdot) - L_{\theta+u(\theta'-\theta)}\left(P^T \cdot\right)\right) \cap \mathbb{X}; \tag{A.15}$$

and $\mathcal{Z}(f)$ denotes the zeros of the function f. The proof proceeds by showing that for any $P \in \mathcal{P}^*$, $\mathcal{V}_P$ is included in a measurable set with measure $O(\|\theta - \theta'\|^{1-2\kappa})$.

Let $P \in \mathcal{P}^*$. Let $B(0, \Delta_\pi) = \{y \in \mathbb{R}^d : \|y\| \le \Delta_\pi\}$, where Δ_π is defined by A.1. For any $x \in B(0, \Delta_\pi)$, define

$$\begin{aligned} l_\theta(x) &= 2\mu^T \Sigma^{-1}\left(I - P^T\right)x, \\ q_\theta(x) &= x^T\left(\Sigma^{-1} - P\Sigma^{-1}P^T\right)x, \\ \mathsf{B}_{\theta,\theta'} &= \left\{x \in B(0, \Delta_\pi) : \left|l_\theta(x)\right| \le \left\|\theta - \theta'\right\|^\kappa\right\}. \end{aligned}$$

Denote by $\mathbb{S}$ the unit sphere $\{x \in \mathbb{R}^d / \|x\| = 1\}$. Let $u \in [0,1]$ and $tv \in \mathcal{Z}(L_{\theta+u(\theta'-\theta)}(\cdot) - L_{\theta+u(\theta'-\theta)}(P^T \cdot)) \cap \mathbb{X}$ where $t \in [0, \Delta_\pi]$ and $v \in \mathbb{S}$. Upon noting that for any $\vartheta \in \mathcal{W}_{M+1}$,

$$L_\vartheta(tv) - L_\vartheta\left(tP^T v\right) = t\left(q_\vartheta(v)t - l_\vartheta(v)\right), \tag{A.16}$$

we consider several cases:

(i) $tv \in \mathsf{B}_{\theta,\theta'}$.

(ii) $tv \notin \mathsf{B}_{\theta,\theta'}$ and $q_{\theta+u(\theta'-\theta)}(v) = 0$. Then, by (A.16), $l_{\theta+u(\theta'-\theta)}(tv) = 0$ which implies that $tv \in \mathsf{B}_{\theta,\theta'}$. This yields a contradiction.

(iii) $tv \notin \mathsf{B}_{\theta,\theta'}$ and $q_{\theta+u(\theta'-\theta)}(v) \neq 0$. Then $t \neq 0$ and, by (A.16),

$$t = \frac{l_{\theta+u(\theta'-\theta)}(v)}{q_{\theta+u(\theta'-\theta)}(v)}. \tag{A.17}$$

Since we assumed $t \in [0, \Delta_\pi]$, this ratio is positive. In order to characterize the point tv, additional notations are required. First, note that by Lemma A.2, there exists $C_1 > 0$ such that for any $\tilde{\theta} = (\tilde{\mu}, \tilde{\Sigma}) \in \mathcal{W}_{M+1}$,

$$\|\tilde{\theta} - \theta\| \leq h_0 \Rightarrow \left\| \tilde{\Sigma}^{-1} - \Sigma^{-1} \right\| \leq C_1 \|\tilde{\Sigma} - \Sigma\|.$$

Thus, there exists $C_2 > 0$ such that for any $\tilde{\theta} \in \mathcal{W}_{M+1}$, $\|\tilde{\theta} - \theta\| \leq h_0$, and for any $x \in B(0, \Delta_\pi)$,

$$\begin{aligned} \left| l_{\tilde{\theta}}(x) - l_\theta(x) \right| &= 2 \left| \mu^T \left[\tilde{\Sigma}^{-1} - \Sigma^{-1} \right] \left(I - P^T \right) x + (\tilde{\mu} - \mu)^T \tilde{\Sigma}^{-1} \left(I - P^T \right) x \right| \\ &\leq C_2 \|\tilde{\theta} - \theta\|. \end{aligned} \tag{A.18}$$

Note that since $x, \mu \in B(0, \Delta_\pi)$, C_2 does not depend on x and θ. Similarly, there exists $C_3 > 0$ such that for $x \in B(0, \Delta_\pi)$ and $\tilde{\theta} \in \mathcal{W}_{M+1}$ satisfying $\|\tilde{\theta} - \theta\| \leq h_0$,

$$\left| q_{\tilde{\theta}}(x) - q_\theta(x) \right| \leq C_3 \|\tilde{\theta} - \theta\|. \tag{A.19}$$

We can assume without loss of generality that $\bar{h}$ is small enough so that

$$\left\| \theta - \theta' \right\| \leq \bar{h} \Rightarrow \left\| \theta - \theta' \right\|^\kappa - (C_2 + 2C_3\Delta_\pi) \left\| \theta - \theta' \right\| \geq \tfrac{1}{2} \left\| \theta - \theta' \right\|^\kappa. \tag{A.20}$$

We now distinguish three subcases.

(a) $v \in \mathsf{B}_{\theta,\theta'}$.

(b) $v \notin \mathsf{B}_{\theta,\theta'}$ and $q_\theta(v) \neq 0$. Since $t \in [0, \Delta_\pi]$, (A.17) implies that $|q_{\theta+u(\theta'-\theta)}(v)| \geq |l_{\theta+u(\theta'-\theta)}(v)|/\Delta_\pi$. Since $v \notin \mathsf{B}_{\theta,\theta'}$, $|l_\theta(v)| \geq \|\theta - \theta'\|^\kappa$ and by using (A.18),

$$|l_{\theta+u(\theta'-\theta)}| \geq \left| l_\theta(v) \right| - \left| l_{\theta+u(\theta'-\theta)} - l_\theta(v) \right| \geq \left\| \theta - \theta' \right\|^\kappa - C_2 \left\| \theta - \theta' \right\|.$$

Hence, it holds that $|q_{\theta+u(\theta'-\theta)}(v)| \geq (\|\theta - \theta'\|^\kappa - C_2\|\theta - \theta'\|)/\Delta_\pi$, and, by (A.19), we have $|q_\theta(v)| \geq |q_{\theta+u(\theta'-\theta)}(v)| - C_3\|\theta - \theta'\|$. These inequalities together with (A.18) and (A.20) lead to

$$\left| t - \frac{l_\theta(v)}{q_\theta(v)} \right| = \left| \frac{l_{\theta+u(\theta'-\theta)}(v)}{q_{\theta+u(\theta'-\theta)}(v)} - \frac{l_\theta(v)}{q_\theta(v)} \right| \leq C_4 \left\| \theta - \theta' \right\|^{1-2\kappa},$$

for some $C_4 > 0$.

(c) $v \notin \mathsf{B}_{\theta,\theta'}$ and $q_\theta(v) = 0$. Then by (A.18) and (A.19),

$$t \geq \frac{\|\theta - \theta'\|^\kappa - C_2\|\theta - \theta'\|}{C_3\|\theta - \theta'\|} \geq 2\Delta_\pi,$$

which is in contradiction with the assumption that $t \leq \Delta_\pi$.

As a conclusion, we have just proved that $\mathcal{V}_P$ is included in the union of three sets defined by $\mathsf{B}_{\theta,\theta'}$ (case (i)), by $\{tv : t \in [0, \Delta_\pi], v \in \mathbb{S} \cap \mathsf{B}_{\theta,\theta'}\}$ (case (iii)(a)), and by

$$\left\{ tv : v \in \mathbb{S}, v \notin \mathsf{B}_{\theta,\theta'}, q_\theta(v) \neq 0, 0 \leq t \leq \Delta_\pi, \left| t - \frac{l_\theta(v)}{q_\theta(v)} \right| \leq C_4 \left\| \theta - \theta' \right\|^{1-2\kappa} \right\}$$

(case (iii)(c)). This concludes the first step.

The second step consists in computing an upper bound for the Lebesgue measure of each of these three sets. For simplifying the presentation, we detail the case $d = 2$ and use polar coordinates (ρ, ϕ); the argument remains valid when $d > 2$ using generalized spherical coordinates. Define $t_\theta(\phi) = l_\theta(\mathrm{e}^{\mathrm{i}\phi})/q_\theta(\mathrm{e}^{\mathrm{i}\phi})$. Rephrasing the conclusion of the first step, we have $\mathcal{V}_P \subset \bigcup_{\ell=1}^{3} \mathcal{V}_P^{(\ell)}$ with

$$\begin{aligned}
\mathcal{V}_P^{(1)} &= \mathsf{B}_{\theta,\theta'},\\
\mathcal{V}_P^{(2)} &= \{(\rho,\phi)/\rho \in [0, \Delta_\pi], \mathrm{e}^{\mathrm{i}\phi} \in \mathsf{B}_{\theta,\theta'}\},\\
\mathcal{V}_P^{(3)} &= \{(\rho,\phi)/\mathrm{e}^{\mathrm{i}\phi} \notin \mathsf{B}_{\theta,\theta'}, q_\theta(\mathrm{e}^{\mathrm{i}\phi}) \neq 0, 0 \le \rho \le \Delta_\pi, |\rho - t_\theta(\phi)| \le C_4 \|\theta - \theta'\|^{1-2\kappa}\}.
\end{aligned}$$

These sets are Borel sets. By definition of $\mathcal{W}_M$, l_θ is not identically zero, and thus

$$\mathrm{Leb}(\mathcal{V}_P^{(1)}) = \mathrm{Leb}(\mathsf{B}_{\theta,\theta'}) \le 2\Delta_\pi \frac{\|\theta - \theta'\|^{1-2\kappa}}{\|2\mu^t \Sigma^{-1}(I - P^T)\|} \le C_5 \|\theta - \theta'\|^{1-2\kappa}$$

for some $C_5 > 0$ as a consequence of Lemma A.2. For $\mathcal{V}_P^{(2)}$, note that it is upper bounded by the reunion of the two circular sectors in bold lines in Figure 7. This area is easily bounded by the

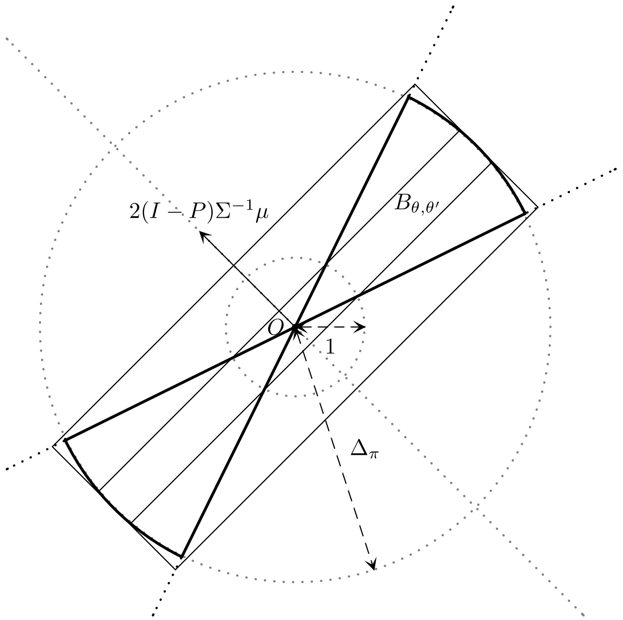


Figure 7. Bounding the measure of the set $\mathcal{V}_P^{(2)}$.

area of the outer rectangle, which is proportional to $\|\theta-\theta'\|^{1-2\kappa}$. Finally,

$$\operatorname{Leb}\bigl(\mathcal{V}_P^{(3)}\bigr)=\int_0^{2\pi}\left[\frac{\rho^2}{2}\right]_{0\vee(t_\theta(\phi)-C_4\|\theta-\theta'\|^{1-2\kappa})}^{\Delta_\pi\wedge(t_\theta(\phi)+C_4\|\theta-\theta'\|^{1-2\kappa})}\mathbb{1}_{q_\theta(\mathrm{e}^{\mathrm{i}\phi})\neq 0}\,\mathrm{d}\phi.$$

We can assume without loss of generality that $\bar{h}$ is small enough so that $2C_4\bar{h}^{1-2\kappa}<\Delta_\pi$. Therefore, we can partition $[0,2\pi]=\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}$, where

$$\begin{aligned}
\mathcal{A}&=\bigl\{\phi\in[0,2\pi]/t_\theta(\phi)-C_4\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}\geq 0\text{ and }t_\theta(\phi)+C_4\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}\leq\Delta_\pi\bigr\},\\
\mathcal{B}&=\bigl\{\phi\in[0,2\pi]/t_\theta(\phi)-C_4\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}\geq 0\text{ and }t_\theta(\phi)+C_4\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}\geq\Delta_\pi\bigr\},\\
\mathcal{C}&=\bigl\{\phi\in[0,2\pi]/t_\theta(\phi)-C_4\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}\leq 0\text{ and }0\leq t_\theta(\phi)+C_4\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}\leq\Delta_\pi\bigr\}.
\end{aligned}$$

This yields

$$\begin{aligned}
\operatorname{Leb}\bigl(\mathcal{V}_P^{(3)}\bigr)&\leq 2C_4\int_{\mathcal{A}}t_\theta(\phi)\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}\,\mathrm{d}\phi+\frac{1}{2}\int_{\mathcal{B}}\bigl(\Delta_\pi^2-\bigl(t_\theta(\phi)-C_4\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}\bigr)^2\bigr)\,\mathrm{d}\phi\\
&\quad+\frac{1}{2}\int_{\mathcal{C}}\bigl(t_\theta(\phi)+C_4\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}\bigr)^2\,\mathrm{d}\phi \qquad\text{(A.21)}\\
&\leq C_6\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}, \qquad\text{(A.22)}
\end{aligned}$$

for some $C_6>0$, since on $\mathcal{A}$, $0\leq t_\theta(\phi)\leq\Delta_\pi$, on $\mathcal{B}$, $(t_\theta(\phi)-C_4\|\theta-\theta'\|^{1-2\kappa})^2\geq(\Delta_\pi-2C_4\|\theta-\theta'\|^{1-2\kappa})^2$, and on $\mathcal{C}$, $|t_\theta(\phi)|\leq C_4\|\theta-\theta'\|^{1-2\kappa}$.

This concludes the proof. □

Lemma A.6 (Regularity in θ of the invariant distribution π_θ). *Let $M>0$ and $\kappa\in(0,1/2)$. Under Assumption 1, there exists $C>0$ such that for any $\theta\in\mathcal{W}_M$ and $\theta'\in\Theta$,*

$$\|\pi_\theta-\pi_{\theta'}\|_{\mathrm{TV}}\leq C\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}.$$

Proof. By definition of the total variation,

$$\|\pi_\theta-\pi_{\theta'}\|_{\mathrm{TV}}\leq|\mathcal{P}|\bigl(\pi(V_\theta\setminus V_{\theta'})+\pi(V_{\theta'}\setminus V_\theta)\bigr).$$

Since

$$V_{\theta'}\setminus V_\theta=V_\theta\setminus(V_\theta\cap V_{\theta'}),\qquad V_\theta\setminus V_{\theta'}=V_\theta\setminus(V_\theta\cap V_{\theta'}),$$

it holds that

$$\pi(V_{\theta'}\setminus V_\theta)=\frac{1}{|\mathcal{P}|}-\pi(V_\theta\cap V_{\theta'})=\pi(V_\theta\setminus V_{\theta'}),$$

where we used Lemma A.1. Then, by Assumption 1 and Lemma A.5, there exists $C>0$ such that for any $\theta\in\mathcal{W}_M$ and $\theta'\in\Theta$,

$$\|\pi_\theta-\pi_{\theta'}\|_{\mathrm{TV}}\leq 2\|\pi\|_\infty\operatorname{Leb}(V_\theta\setminus V_{\theta'})\leq C\bigl\|\theta-\theta'\bigr\|^{1-2\kappa}.$$ □

Lemma A.7 (Regularity in θ of the kernels P_θ). *Let $M > 0$ and $\kappa \in (0, 1/2)$. Under Assumption 1, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \mathcal{W}_{M+1}$,*

$$\left\| P_\theta(x, \cdot) - P_{\theta'}(x, \cdot)\right\|_{\mathrm{TV}} \le C \left\|\theta - \theta'\right\|^{1-2\kappa}.$$

Proof. From the definition of the transition kernel P_θ, we have

$$\begin{aligned}
&\left|P_\theta f(x) - P_{\theta'} f(x)\right| \\
&\quad\le \left|\int f(y)\bigl(\alpha_\theta(x,y)q_\theta(x,y)\mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x,y)q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y)\bigr)\,\mathrm{d}y\right| \\
&\qquad + \left|f(x)\right|\left|\int \bigl(\alpha_{\theta'}(x,y)q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y) - \alpha_\theta(x,y)q_\theta(x,y)\mathbb{1}_{V_\theta}(y)\bigr)\,\mathrm{d}y\right| \qquad \text{(A.23)} \\
&\quad\le 2\|f\|_\infty \int \left|\alpha_\theta(x,y)q_\theta(x,y)\mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x,y)q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y)\right|\,\mathrm{d}y \\
&\quad= 2\|f\|_\infty \sum_{i=1}^{4} \Delta^i_{\theta,\theta'}(x),
\end{aligned}$$

where

$$\begin{aligned}
\Delta^1_{\theta,\theta'}(x) &= \int_{\mathcal{A}_\theta(x)\cap\mathcal{A}_{\theta'}(x)} \left|\alpha_\theta(x,y)q_\theta(x,y)\mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x,y)q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y)\right|\,\mathrm{d}y, \\
\Delta^2_{\theta,\theta'}(x) &= \int_{\mathcal{R}_\theta(x)\cap\mathcal{R}_{\theta'}(x)} \left|\alpha_\theta(x,y)q_\theta(x,y)\mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x,y)q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y)\right|\,\mathrm{d}y, \\
\Delta^3_{\theta,\theta'}(x) &= \int_{\mathcal{A}_\theta(x)\cap\mathcal{R}_{\theta'}(x)} \left|\alpha_\theta(x,y)q_\theta(x,y)\mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x,y)q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y)\right|\,\mathrm{d}y, \\
\Delta^4_{\theta,\theta'}(x) &= \int_{\mathcal{R}_\theta(x)\cap\mathcal{A}_{\theta'}(x)} \left|\alpha_\theta(x,y)q_\theta(x,y)\mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x,y)q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y)\right|\,\mathrm{d}y
\end{aligned}$$

and

$$\mathcal{A}_\theta(x) = \bigl\{y : \alpha_\theta(x,y) = 1\bigr\}, \qquad \mathcal{R}_\theta(x) = \bigl\{y : \alpha_\theta(x,y) < 1\bigr\}.$$

We now upper bound each term

$$\begin{aligned}
\Delta^1_{\theta,\theta'}(x) &= \int_{\mathcal{A}_\theta(x)\cap\mathcal{A}_{\theta'}(x)} \left|\sum_{Q\in\mathcal{P}} \bigl(\mathbb{1}_{V_\theta}(y)\mathcal{N}(Qy|x,\Sigma) - \mathbb{1}_{V_{\theta'}}(y)\mathcal{N}\bigl(Qy|x,\Sigma'\bigr)\bigr)\right|\,\mathrm{d}y \\
&\le \int \left|\mathbb{1}_{V_\theta}(y) - \mathbb{1}_{V_{\theta'}}(y)\right| \sum_{Q\in\mathcal{P}} \mathcal{N}(Qy|x,\Sigma) \qquad \text{(A.24)} \\
&\qquad + \mathbb{1}_{V_{\theta'}}(y) \sum_{Q\in\mathcal{P}} \left|\mathcal{N}(Qy|x,\Sigma) - \mathcal{N}\bigl(Qy|x,\Sigma'\bigr)\right|\,\mathrm{d}y.
\end{aligned}$$

By Lemma A.2, there exist $a, b > 0$ such that for any $\theta \in \mathcal{W}_{M+1}$, $m, z \in \mathbb{X}$, and $Q \in \mathcal{P}$, we have

$$a \leq \mathcal{N}(Qz|m, c\Sigma) \leq b, \tag{A.25}$$

so that the first term in the RHS of (A.24) is bounded by

$$\begin{aligned}\int \left|\mathbb{1}_{V_\theta}(y) - \mathbb{1}_{V_{\theta'}}(y)\right| \sum_{Q \in \mathcal{P}} \mathcal{N}(Qy|x, \Sigma)\,\mathrm{d}y &\leq |\mathcal{P}|b \int \left|\mathbb{1}_{V_\theta}(y) - \mathbb{1}_{V_{\theta'}}(y)\right| \mathrm{d}y \\ &= |\mathcal{P}|b \int \left(\mathbb{1}_{V_\theta \setminus V_{\theta'}}(y) + \mathbb{1}_{V_{\theta'} \setminus V_\theta}(y)\right) \mathrm{d}y \\ &\leq C \left\|\theta - \theta'\right\|^{1-2\kappa},\end{aligned}$$

where we used Lemma A.5. Let us now consider the second term of the right-hand side of (A.24). Using the uniform continuity of w on $\mathcal{W}_{M+1}$ (see Lemma A.2), there exists $\bar{h}$ small enough such that

$$\theta \in \mathcal{W}_M, \qquad \|h\| < \bar{h} \Rightarrow \theta + h \in \mathcal{W}_{M+1}. \tag{A.26}$$

For any $\theta \in \mathcal{W}_M$, $\theta' \in \mathcal{W}_{M+1}$ such that $\|\theta - \theta'\| \geq \bar{h}$, there exists C_1 such that

$$\sum_{Q \in \mathcal{P}} \left|\mathcal{N}(Qy|x, \Sigma) - \mathcal{N}\left(Qy|x, \Sigma'\right)\right| \mathrm{d}y \leq C_1 \left\|\theta - \theta'\right\|^{1-2\kappa}.$$

Assume now that $\theta \in \mathcal{W}_M$, $\theta' \in \mathcal{W}_{M+1}$ and $\|\theta - \theta'\| < \bar{h}$. Denote by

$$\Sigma_t = (1-t)\Sigma + t\Sigma'. \tag{A.27}$$

By (A.26) and (A.4b), Σ_t^{-1} exists and $\sup_{t \leq 1, \theta \in \mathcal{W}_M, \theta' \in \mathcal{W}_{M+1}} \|\Sigma_t^{-1}\| < \infty$. We can then write

$$\begin{aligned}\left|\mathcal{N}(Qy|x, \Sigma) - \mathcal{N}\left(Qy|x, \Sigma'\right)\right| &= \int_0^1 \mathcal{N}(Qy|x, \Sigma_t) \left|\frac{\mathrm{d}}{\mathrm{d}t} \log \mathcal{N}(Qy|x, \Sigma_t)\right| \mathrm{d}t \\ &\leq b \int_0^1 \left|\frac{\mathrm{d}}{\mathrm{d}t} \log \mathcal{N}(Qy|x, \Sigma_t)\right| \mathrm{d}t.\end{aligned} \tag{A.28}$$

In addition, by Assumption 1, there exists C_2 such that

$$\left|\frac{\mathrm{d}}{\mathrm{d}t} \log \mathcal{N}(Qy|x, \Sigma_t)\right| = \left|(x - Qy)^T \Sigma_t^{-1} \left(\Sigma' - \Sigma\right) \Sigma_t^{-1} (x - Qy)\right| \leq C_2 \left\|\theta - \theta'\right\|. \tag{A.29}$$

We thus have proved that

$$\left[\theta \in \mathcal{W}_M, \theta' \in \mathcal{W}_{M+1}, \left\|\theta - \theta'\right\| < \bar{h}\right] \Rightarrow \left|\mathcal{N}(Qy|x, \Sigma) - \mathcal{N}\left(Qy|x, \Sigma'\right)\right| \leq C \left\|\theta - \theta'\right\|.$$

Therefore, it is established that $\|\Delta^1_{\theta,\theta'}\|_\infty \leq C\|\theta - \theta'\|^{1-2\kappa}$.

Let us consider the second term $\Delta^2_{\theta,\theta'}(x)$ in the RHS of (A.23). Note first that if $x \in \mathbb{X}$ and $y \in \mathcal{R}_\theta(x) \cap \mathcal{R}_{\theta'}(x)$, then by (A.25), $\pi(y)/\pi(x) \le b/a$, so

$$\begin{aligned}\Delta^2_{\theta,\theta'}(x) &= \int_{\mathcal{R}_\theta(x)\cap\mathcal{R}_{\theta'}(x)} \frac{\pi(y)}{\pi(x)} \left| \sum_{Q\in\mathcal{P}} \bigl(\mathbb{1}_{V_\theta}(y)\mathcal{N}(Qx|y,\Sigma) - \mathbb{1}_{V_{\theta'}}(y)\mathcal{N}\bigl(Qx|y,\Sigma'\bigr)\bigr)\right| \mathrm{d}y \\ &\le \frac{b}{a}\int_{\mathcal{R}_\theta(x)\cap\mathcal{R}_{\theta'}(x)} \left| \sum_{Q\in\mathcal{P}} \bigl(\mathbb{1}_{V_\theta}(y)\mathcal{N}(Qx|y,\Sigma) - \mathbb{1}_{V_{\theta'}}(y)\mathcal{N}\bigl(Qx|y,\Sigma'\bigr)\bigr)\right| \mathrm{d}y.\end{aligned}$$

Therefore, repeating the above discussion for the bound of $\Delta^1_{\theta,\theta'}(x)$, it is established that $\|\Delta^2_{\theta,\theta'}\|_\infty \le C\|\theta-\theta'\|^{1-2\kappa}$.

To deal with $\Delta^3_{\theta,\theta'}(x)$, first observe that there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$, $\theta' \in \mathcal{W}_{M+1}$, and $x, y \in \mathbb{X}$, we have

$$\left| \frac{q_\theta(y,x)}{q_\theta(x,y)} - \frac{q_{\theta'}(y,x)}{q_{\theta'}(x,y)} \right| \le C\bigl\|\theta-\theta'\bigr\|, \tag{A.30}$$

because of (3.7), (A.25) and the above discussion for the upper bound of $\Delta^1_{\theta,\theta'}(x)$. Now let $y \in \mathcal{A}_\theta(x) \cap \mathcal{R}_{\theta'}(x)$, then we have

$$\frac{\pi(y)q_{\theta'}(y,x)}{\pi(x)q_{\theta'}(x,y)} \le 1 \le \frac{\pi(y)q_\theta(y,x)}{\pi(x)q_\theta(x,y)},$$

which, combined with (A.30), yields

$$1 - C\frac{\pi(y)}{\pi(x)}\bigl\|\theta-\theta'\bigr\| \le \frac{\pi(y)q_{\theta'}(y,x)}{\pi(x)q_{\theta'}(x,y)} \le 1.$$

Thus,

$$\begin{aligned}\Delta^3_{\theta,\theta'}(x) &= \int_{\mathcal{A}_\theta(x)\cap\mathcal{R}_{\theta'}(x)} \left| q_\theta(x,y)\mathbb{1}_{V_\theta}(y) - \frac{\pi(y)q_{\theta'}(y,x)}{\pi(x)q_{\theta'}(x,y)} q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y)\right| \mathrm{d}y \\ &\le \int \biggl(\bigl| q_\theta(x,y)\mathbb{1}_{V_\theta}(y) - q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y)\bigr| \\ &\qquad \vee \cdots \vee \biggl| q_\theta(x,y)\mathbb{1}_{V_\theta}(y) - q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y) \\ &\qquad\qquad + C\frac{\pi(y)}{\pi(x)}\bigl\|\theta-\theta'\bigr\| q_{\theta'}(x,y)\mathbb{1}_{V_{\theta'}}(y)\biggr|\biggr) \mathrm{d}y.\end{aligned}$$

Therefore, it is established that $\|\Delta^3_{\theta,\theta'}\|_\infty \le C\|\theta-\theta'\|^{1-2\kappa}$.

The upper bound of $\Delta^4_{\theta,\theta'}(x)$ is similar, and thus its proof is omitted. □

Lemma A.8 (Regularity in θ of the solution of the Poisson equation). *Let $M > 0$ and $\kappa \in (0, 1/2)$. Under Assumption 1, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \mathcal{W}_{M+1}$,*

$$\|P_\theta \hat{H}_\theta - P_{\theta'} \hat{H}_{\theta'}\|_\infty \leq C \|\theta - \theta'\|^{1-2\kappa}.$$

Proof. We recall the following result, proved in [13], Lemma 5.5, page 24: there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$, $\theta' \in \mathcal{W}_{M+1}$, and $x \in \mathbb{X}$,

$$\begin{aligned} &\|P_\theta \hat{H}_\theta - P_{\theta'} \hat{H}_{\theta'}\|_\infty \\ &\quad \leq C \| H(\cdot, \theta) - H(\cdot, \theta') \|_\infty \qquad\qquad (A.31) \\ &\qquad + C \sup_{\theta \in \mathcal{W}_M} \| H(\cdot, \theta) \|_\infty \Big\{ \|\pi_\theta - \pi_{\theta'}\|_{\mathrm{TV}} + \sup_{x \in \mathbb{X}} \| P_\theta(x, \cdot) - P_{\theta'}(x, \cdot) \|_{\mathrm{TV}} \Big\}. \end{aligned}$$

Here, $\sup_{\theta \in \mathcal{W}_M} \|H(\cdot, \theta)\|_\infty$ is finite by Lemma A.2. Now, by Lemma A.2 again, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \mathcal{W}_{M+1}$,

$$\| H(\cdot, \theta) - H(\cdot, \theta') \|_\infty \leq C \|\theta - \theta'\|.$$

The upper bounds for the two last terms in the RHS of (A.31) result from Lemmas A.6 and A.7, respectively. □

A.5. Proof of Theorem 3.2

We start by proving two lemmas.

Lemma A.9. *Let $(\gamma_t)_{t>0}$ be a sequence such that $\sum_t \gamma_t^2 < \infty$, $\sum_t |\gamma_{t+1} - \gamma_t| < \infty$, and $\sum_t \gamma_t^{2(1-\kappa)} < \infty$ for some $\kappa \in (0, 1/2)$. Denote by ψ_t the value of the projection counter at the end of iteration t, in Algorithm 1. Let $(\theta_t, X_t)_{t\geq 0}$ be the sequence generated by Algorithm 1. Under Assumptions 1 and 2, for any $M > 0$,*

$$\lim_{L \to +\infty} \sup_{\ell \geq 1} \left\| \left(\prod_{k=L}^{L+\ell} \mathbb{1}_{\theta_k \in \mathcal{W}_M} \mathbb{1}_{\psi_{k+1} = \psi_k} \right) \sum_{k=L}^{L+\ell} \gamma_{k+1} \big(H(X_{k+1}, \theta_k) - h(\theta_k) \big) \right\| = 0 \qquad w.p.1, \quad (A.32)$$

where H, h, w and $\mathcal{W}_M$ are given by (3.9), (3.12), (3.13) and (A.3), respectively.

Proof. The proof is adapted from Theorem 2.7 in [13], and it is thus omitted. It can be found in the supplemental article [6]. □

Lemma A.10. *Let $M \in (0, M_\star)$ and set*

$$\Gamma_{M_\star}^M = \{ \theta \in \Theta : M_\star \leq w(\theta) \leq M \}, \qquad \iota = \inf_{\theta \in \Gamma_{M_\star}^M} |\langle \nabla w(\theta), h(\theta) \rangle|.$$

Under Assumptions 1 *and* 2, *there exist* $\delta \in (0, \iota)$ *and* $\lambda, \beta > 0$ *such that*

(A) $u \in \mathcal{W}_{M_\star}, 0 \le \gamma \le \lambda, \|\xi\| \le \beta \Rightarrow w(u + \gamma h(u) + \gamma \xi) \le M$, *and*
(B) $u \in \Gamma^M_{M_\star}, 0 \le \gamma \le \lambda, \|\xi\| \le \beta \Rightarrow w(u + \gamma h(u) + \gamma \xi) < w(u) - \gamma \delta$.

Proof. The proof is adapted from Lemma 2.1 in [1], and it is thus omitted. It can be found in the supplemental article [6]. □

Proof of item (1) in Theorem 3.2. Let $M > M_\star$, let q (depending on M) be such that (see Remark A.3)

$$\mathcal{W}_M \subset \mathcal{W}_{M+2} \subseteq \mathcal{K}_{\delta_q}, \tag{A.33}$$

and let $\theta_0 \in \mathcal{W}_M$. Let λ, β be given by Lemma A.10. By Lemma A.2, w and h are uniformly continuous on $\mathcal{W}_{M+1}$, and there exists $\eta > 0$ such that

$$x \in \mathcal{W}_M, \qquad \|x - y\| < \eta \Rightarrow \left|w(x) - w(y)\right| < 1 \quad \text{and} \quad \left\|h(x) - h(y)\right\| < \beta. \tag{A.34}$$

By Lemma A.9, there exists an almost surely finite r.v. N such that w.p.1.,

$$n \ge N \Rightarrow \gamma_n \Bigl(1 + \sup_{x \in \mathbb{X}, \theta \in \mathcal{W}_M} \left\|H(x, \theta)\right\|\Bigr) < \lambda \wedge \eta \tag{A.35}$$

and

$$\sup_{\ell \ge 1} \left(\prod_{i=N}^{N+\ell} \mathbb{1}_{\theta_i \in \mathcal{W}_{M+1}} \mathbb{1}_{\psi_{i+1} = \psi_i}\right) \left\|\sum_{i=N}^{N+\ell} \gamma_{i+1}\bigl(H(X_{i+1}, \theta_i) - h(\theta_i)\bigr)\right\| < \eta. \tag{A.36}$$

The proof is by contradiction. Denote by ψ_t the number of projections at the end of iteration t. We assume that $\mathbb{P}(\lim_t \psi_t = +\infty) > 0$. We can assume without loss of generality that

$$w(\theta_N) \le M, \qquad \psi_N \ge q$$

on the set $\{\lim_t \psi_t = +\infty\}$. Define the sequence $(\theta'_{N+k})_{k \ge 0}$ as

$$\theta'_N = \theta_N \quad \text{and} \quad \theta'_{N+k+1} = \theta'_{N+k} + \gamma_{N+k+1} h(\theta_{N+k}).$$

We prove by induction on k that for any $k \ge 0$, on the set $\{\lim_t \psi_t = +\infty\}$,

$$\theta'_{N+k} \in \mathcal{W}_M, \qquad \theta_{N+k} \in \mathcal{W}_{M+1}, \qquad \left\|\theta'_{N+k} - \theta_{N+k}\right\| < \eta, \qquad \psi_{N+k+1} = \psi_{N+k}.$$

The case $k = 0$ is trivial since $\theta'_N = \theta_N \in \mathcal{W}_M$ and by using (A.34), (A.35) and (A.33) on the set $\{\lim_t \psi_t = +\infty\}$. Assume this property holds for $k \in \{0, 1, \dots, \ell\}$. Then we have

$$\theta'_{N+\ell+1} = \theta'_{N+\ell} + \gamma_{N+\ell+1} h\bigl(\theta'_{N+\ell}\bigr) + \gamma_{N+\ell+1}\bigl(h(\theta_{N+\ell}) - h\bigl(\theta'_{N+\ell}\bigr)\bigr).$$

Since $\|\theta'_{N+\ell} - \theta_{N+\ell}\| < \eta$ and $\theta'_{N+\ell}$ is in $\mathcal{W}_M$, we have $\|h(\theta'_{N+\ell}) - h(\theta_{N+\ell})\| < \beta$. Since $\gamma_{N+\ell+1} < \lambda$ by (A.35), we can apply Lemma A.10 to obtain $\theta'_{N+\ell+1} \in \mathcal{W}_M$. In addition,

$$\begin{aligned}\theta'_{N+\ell+1} - \theta_{N+\ell+1} &= \sum_{i=N}^{N+\ell} \gamma_{i+1}\big(H(X_{i+1}, \theta_i) - h(\theta_i)\big)\mathbb{1}_{\psi_{i+1}=\psi_i} \\ &\quad + \sum_{i=N}^{N+\ell} \big(\gamma_{i+1} h(\theta_i) + \theta_i - \theta_0\big)\mathbb{1}_{\psi_{i+1}\neq\psi_i} \\ &= \left(\prod_{i=N}^{N+\ell} \mathbb{1}_{\theta_i \in \mathcal{W}_{M+1}}\right) \sum_{i=N}^{N+\ell} \gamma_{i+1}\big(H(X_{i+1}, \theta_i) - h(\theta_i)\big)\mathbb{1}_{\psi_{i+1}=\psi_i},\end{aligned}$$

where we used the induction assumption in the last equality. From (A.34) and (A.36), this yields $\|\theta'_{N+\ell+1} - \theta_{N+\ell+1}\| < \eta$ and $w(\theta_{N+\ell+1}) \leq M + 1$. Finally by (A.34), equations (A.35) and (A.33) imply that on the set $\{\lim_t \psi_t = +\infty\}$

$$\theta_{N+\ell} + \gamma_{N+\ell+1} H(X_{N+\ell+1}, \theta_{N+\ell}) \in \mathcal{W}_{M+2} \subset \mathcal{K}_{\psi_{N+\ell}},$$

that is, $\psi_{N+\ell+1} = \psi_{N+\ell}$. This concludes the induction.

As a consequence of this induction, we have $\psi_{N+\ell} = \psi_N$ for any $\ell \geq 0$ on the set $\{\lim_t \psi_t = +\infty\}$ which is a contradiction.

Proof of item (2) *in Theorem* 3.2. The proof is along the same lines as the proof of Theorem 2.3 of [1], page 5, and is thus omitted.

A.6. Proof of Theorem 3.3

The proof consists in checking the conditions of [13], Corollary 2.8. Let f be a measurable bounded function.

By Lemma A.4, (i) there exists a measurable function $\hat{f}_\theta$ such that $\hat{f}_\theta - P_\theta \hat{f}_\theta = f - \pi_\theta f$; and (ii) for any compact set $\mathcal{W}_M$, there exists L (depending upon M) such that

$$\forall \theta \in \mathcal{W}_M, x \in \mathbb{X}, \qquad \big|\hat{f}_\theta(x)\big| \leq L.$$

By Theorem 3.2, $\mathbb{P}(\Omega_M) \uparrow 1$ when M tends to infinity where

$$\Omega_M = \bigcap_{t\geq 0} \{\theta_t \in \mathcal{W}_M\}. \tag{A.37}$$

Therefore, in order to apply [13], Corollary 2.8, we only have to prove that almost surely,

$$\sum_k k^{-1} \sup_{x\in\mathbb{X}} \big\| P_{\theta_k}(x, \cdot) - P_{\theta_{k-1}}(x, \cdot)\big\|_{\mathrm{TV}} \mathbb{1}_{\Omega_M} < \infty, \tag{A.38}$$

$$\lim_t \pi_{\theta_t}(f)\mathbb{1}_{\Omega_M} = \pi_{\theta^\star}(f)\mathbb{1}_{\Omega_M}. \tag{A.39}$$

By Lemma A.7, there exists C and $\kappa \in (0, 1/2)$ such that

$$\sup_{x\in\mathbb{X}} \left\| P_{\theta_k}(x,\cdot) - P_{\theta_{k-1}}(x,\cdot) \right\|_{\mathrm{TV}} \mathbb{1}_{\Omega_M} \le C \|\theta_k - \theta_{k-1}\|^{1-2\kappa}.$$

In addition, by Theorem 3.2, there exists a random variable K, almost surely finite, such that for any $k \ge K$,

$$\|\theta_k - \theta_{k-1}\| \mathbb{1}_{\Omega_M} \le \gamma_k \sup_{\theta\in\mathcal{W}_M, x\in\mathbb{X}} \left|H(x,\theta)\right|.$$

This yields

$$\sum_{k\ge K} k^{-1} \sup_{x\in\mathbb{X}} \left\| P_{\theta_k}(x,\cdot) - P_{\theta_{k-1}}(x,\cdot) \right\|_{\mathrm{TV}} \mathbb{1}_{\Omega_M} \le C \sum_{k\ge K} k^{-1} \gamma_k^{1-2\kappa},$$

for some constant $C > 0$. This concludes the proof of (A.38). The limit (A.39) is a consequence of Lemma A.6.

Remark. Note that in the proof above we use that the number of random truncations is finite almost surely (when claiming that $\lim_M \mathbb{P}(\Omega_M) \uparrow 1$) but only use the convergence of the sequence $(\theta_t)_{t\ge 0}$ in order to establish (A.39). When f is such that $\pi_\theta(f) = \pi(f)$ for any $\theta \in \Theta$ (for example when f is symmetric with respect to permutations), then (A.39) holds even if $(\theta_t)_{t\ge 0}$ does not converge.

A.7. Proof of Theorem 3.4

Let f be a measurable function such that $\|f\|_\infty \le 1$ and set

$$I_t(f) = \left|\mathbb{E}\left[f(X_t)\mathbb{1}_B\right] - \pi_{\theta^\star}(f)\mathbb{P}(B)\right| = \left|\mathbb{E}\left[\left(f(X_t) - \pi_{\theta^\star}(f)\right)\mathbb{1}_B\right]\right|,$$

where $B = \{\lim_q \theta_q = \theta_\star\}$. Let $\varepsilon > 0$. We prove that there exists T_ε such that for all $t \ge T_\varepsilon$, $\sup_{\{f:\|f\|_\infty \le 1\}} I_t(f) \le 4\varepsilon$. Choose $\kappa \in (0, 1/2)$ and $\delta > 0$ such that

$$C_{M_\star+1}\delta^{1-2\kappa} \le \varepsilon, \tag{A.40}$$

where $M_\star$ and $C_{M_\star}$ are defined in Assumption 2 and in Lemma A.6, respectively. Choose r_ε such that

$$2(1-\rho_{M_\star+1})^{r_\varepsilon} \le \varepsilon, \tag{A.41}$$

where $\rho_{M_\star+1}$ is defined in Lemma A.4. By uniform continuity of w on $\mathcal{W}_{M_\star+2}$, assume finally δ is small enough that

$$\theta \in \mathcal{W}_{M_\star+1}, \theta' \in \Theta, \qquad \left\|\theta - \theta'\right\| \le \delta \Rightarrow \left|w(\theta) - w\left(\theta'\right)\right| \le \frac{1}{r_\varepsilon + 1}. \tag{A.42}$$

There exists T_ε^1 such that for any $t \geq T_\varepsilon^1$,

$$\mathbb{P}\Big(\big\|\theta_{t-r_\varepsilon} - \theta^\star\big\| \leq \delta, \lim_q \theta_q = \theta^\star\Big) \leq \varepsilon/2.$$

Hence, for any $t \geq T_\varepsilon^1$, $I_t(f) \leq \sum_{i=1}^3 I_t^i(f) + \varepsilon$, where

$$I_t^1(f) = \big|\mathbb{E}\big[\big(f(X_t) - P_{\theta_{t-r_\varepsilon}}^{r_\varepsilon} f(X_{t-r_\varepsilon})\big)\mathbb{1}_{\|\theta_{t-r_\varepsilon}-\theta^\star\|\leq\delta}\big]\big|, \tag{A.43}$$

$$I_t^2(f) = \big|\mathbb{E}\big[\big(P_{\theta_{t-r_\varepsilon}}^{r_\varepsilon} f(X_{t-r_\varepsilon}) - \pi_{\theta_{t-r_\varepsilon}}(f)\big)\mathbb{1}_{\|\theta_{t-r_\varepsilon}-\theta^\star\|\leq\delta}\big]\big|, \tag{A.44}$$

$$I_t^3(f) = \big|\mathbb{E}\big[\big(\pi_{\theta_{t-r_\varepsilon}}(f) - \pi_{\theta^\star}(f)\big)\mathbb{1}_{\|\theta_{t-r_\varepsilon}-\theta^\star\|\leq\delta}\big]\big|. \tag{A.45}$$

We first upper bound $I_t^1(f)$. For $\theta, \theta' \in \Theta$, let

$$D\big(\theta, \theta'\big) = \sup_{x\in\mathbb{X}} \big\| P_\theta(x, \cdot) - P_{\theta'}(x, \cdot)\big\|_{\mathrm{TV}}.$$

Applying [4], Proposition 1.3.1, it comes for any $t \geq T_\varepsilon^1$,

$$\begin{aligned} I_t^1 &\leq \mathbb{E}\left[2 \wedge \sum_{j=1}^{r_\varepsilon-1} D(\theta_{t-r_\varepsilon+j}, \theta_{t-r_\varepsilon})\mathbb{1}_{\|\theta_{t-r_\varepsilon}-\theta^\star\|\leq\delta}\right] \\ &\leq \mathbb{E}\left[2 \wedge \sum_{j=1}^{r_\varepsilon-1} (r_\varepsilon - j) D(\theta_{t-r_\varepsilon+j}, \theta_{t-r_\varepsilon+j-1})\mathbb{1}_{\|\theta_{t-r_\varepsilon}-\theta^\star\|\leq\delta}\right], \end{aligned}$$

where we used that for any $q, \ell > 0$ $D(\theta_{q+\ell}, \theta_q) \leq \sum_{j=1}^\ell D(\theta_{q+j}, \theta_{q+j-1})$. By Theorem 3.2, the random iteration number τ_ψ where the last projection occurs in Algorithm 1 is finite with probability one. Let then M_ε be such that $2\mathbb{P}(\tau_\psi \geq M_\varepsilon) \leq \varepsilon/2$, so that

$$I_t^1(f) \leq \mathbb{E}\left[2 \wedge \sum_{j=1}^{r_\varepsilon-1} (r_\varepsilon - j) D(\theta_{t-r_\varepsilon+j}, \theta_{t-r_\varepsilon+j-1})\mathbb{1}_{\|\theta_{t-r_\varepsilon}-\theta^\star\|\leq\delta}\mathbb{1}_{\tau_\psi\leq M_\varepsilon}\right] + \frac{\varepsilon}{2}.$$

Let now $T_\varepsilon^2 \geq T_\varepsilon^1 \vee (M_\varepsilon + r_\varepsilon)$ be such that

$$t \geq T_\varepsilon^2 \Rightarrow \gamma_t \sup_{x\in\mathbb{X}, \theta\in\mathcal{W}_{M_\star+2}} \big\|H(x, \theta)\big\| \leq \delta.$$

Then, by recurrence and using (A.42), we obtain that on $\{\|\theta_{t-r_\varepsilon} - \theta_\star\| \leq \delta\}$, $\theta_{t-r_\varepsilon+j} \in \mathcal{W}_{M_\star+1}$ for all $0 \leq j \leq r_\varepsilon$. By Lemma A.7, this yields for any $t \geq T_\varepsilon^2$

$$I_t^1(f) \leq C_{M_\star+1}\Big[\sup_{x\in\mathbb{X}, \theta\in\mathcal{W}_{M_\star+2}} \big\|H(x, \theta)\big\|\Big]^{1-2\kappa} \sum_{j=1}^{r_\varepsilon-1} (r_\varepsilon - j)\gamma_{t-r_\varepsilon+j}^{1-2\kappa} + \frac{\varepsilon}{2},$$

and there exists $T_\varepsilon^3 \geq T_\varepsilon^2$ such that $t \geq T_\varepsilon^3 \Rightarrow \sup_{\{f:\|f\|_\infty \leq 1\}} I_t^1(f) \leq \varepsilon$.

We now consider $I_t^2(f)$; it holds

$$I_t^2 \leq \mathbb{E}\big[\big\| P_{\theta_{t-r_\varepsilon}}^{r_\varepsilon}(X_{t-r_\varepsilon}, \cdot) - \pi_{\theta_{t-r_\varepsilon}} \big\|_{\mathrm{TV}} \mathbb{1}_{\|\theta_{t-r_\varepsilon} - \theta^\star\| \leq \delta}\big].$$

By (A.42), $\|\theta_{t-r_\varepsilon} - \theta^\star\| \leq \delta \Rightarrow \theta_{t-r_\varepsilon} \in \mathcal{W}_{M_\star+1}$ and thus, applying Lemma A.4 and (A.41)

$$\sup_{\{f:\|f\|_\infty \leq 1\}} I_t^2(f) \leq 2(1 - \rho_{M_\star+1})^{r_\varepsilon} \leq \varepsilon.$$

The derivation of the upper bound of I_t^3 is similar to that of I_t^2, with Lemma A.4 replaced by Lemma A.6 and uses (A.40). Details are omitted.

Remark. The proof above can be easily adapted (details are omitted) to address the case when (i) $(\theta_t)_{t\geq 0}$ is stable but does not necessarily converges, and (ii) the function f is bounded and satisfies $\pi_\theta(f) = \pi(f)$ for any $\theta \in \Theta$. The main ingredients for this extension are to replace $\mathbb{1}_B$ with the constant function $\mathbb{1}$, and to replace the set $\{\|\theta_{t-r_\varepsilon} - \theta^\star\| \leq \delta\}$ with $\{\theta_{t-r_\varepsilon} \in \mathcal{W}_{M_\star}\}$. Since the sequence is stable, $\lim_M \mathbb{P}(\Omega_M) \uparrow 1$ where Ω_M is given by (A.37). $M_\star$ is chosen so that $\mathbb{E}[|f(X_t) - \pi(f)|\mathbb{1}_{\Omega_{M_\star}}] \leq \varepsilon$. We then obtain, for such a function f,

$$\lim_{t\to\infty} \mathbb{E}\big[f(X_t)\big] = \pi(f).$$

Acknowledgements

This work was supported by the ANR-2010-COSI-002 grant of the French National Research Agency.

Supplementary Material

Long version of the paper (DOI: 10.3150/13-BEJ578SUPP; .pdf). This long version of the paper features an additional evaluated method for Section 2.2 (AM with posterior reordering), examples of the behavior of AMOR on a nonlinear symmetrized unimodal distribution and on a genuinely bimodal distribution, and complete proofs.

References

[1] Andrieu, C., Moulines, É. and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44** 283–312. MR2177157

[2] Andrieu, C. and Robert, C.P. (2011). Controlled Markov chain Monte Carlo methods for optimal sampling. Technical Report 125, Cahiers du Ceremade, Université Paris Dauphine.

[3] Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. MR2461882

[4] Atchadé, Y., Fort, G., Moulines, E. and Priouret, P. (2011). Adaptive Markov chain Monte Carlo: Theory and methods. In *Bayesian Time Series Models* 32–51. Cambridge: Cambridge Univ. Press. MR2894232

[5] Bardenet, R., Cappé, O., Fort, G. and Kégl, B. (2012). Adaptive Metropolis with online relabeling. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*). *JMLR Workshop and Conference Proceedings* **22** 91–99. Microtome Publishing.

[6] Bardenet, R., Cappé, O., Fort, G. and Kégl, B. (2014). Supplement to "Adaptive MCMC with online relabeling." DOI:10.3150/13-BEJ578SUPP.

[7] Bardenet, R. and Kégl, B. (2012). An adaptive Monte-Carlo Markov chain algorithm for inference from mixture signals. *J. Phys. Conf. Ser.* **368** 012044.

[8] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge Univ. Press. MR2061575

[9] Celeux, G. (1998). Bayesian inference for mixtures: The label-switching problem. In *Computational Statistics Symposium* (*COMPSTAT*) 227–232. Berlin: Springer.

[10] Celeux, G., Hurn, M. and Robert, C.P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95** 957–970. MR1804450

[11] Chen, H.-F. (2002). *Stochastic Approximation and Its Applications*. *Nonconvex Optimization and Its Applications* **64**. Dordrecht: Kluwer Academic. MR1942427

[12] Cron, A.J. and West, M. (2011). Efficient classification-based relabeling in mixture models. *Amer. Statist.* **65** 16–20. MR2899648

[13] Fort, G., Moulines, E. and Priouret, P. (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.* **39** 3262–3289. MR3012408

[14] Graf, S. and Luschgy, H. (2000). *Foundations of Quantization for Probability Distributions*. *Lecture Notes in Math.* **1730**. Berlin: Springer. MR1764176

[15] Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

[16] Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

[17] Jasra, A. (2005). Bayesian inference for mixture models via Monte Carlo. Ph.D. thesis, Imperial College, London, UK.

[18] Jasra, A., Holmes, C.C. and Stephens, D.A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20** 50–67. MR2182987

[19] Marin, J.-M., Mengersen, K. and Robert, C.P. (2005). Bayesian modelling and inference on mixtures of distributions. In *Bayesian Thinking*: *Modeling and Computation*. *Handbook of Statist.* **25** 459–507. Amsterdam: Elsevier. MR2490536

[20] Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. *Communications and Control Engineering Series*. London: Springer. MR1287609

[21] Pagès, G. (1998). A space quantization method for numerical integration. *J. Comput. Appl. Math.* **89** 1–38. MR1625987

[22] Papastamoulis, P. and Iliopoulos, G. (2010). An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *J. Comput. Graph. Statist.* **19** 313–331. MR2758306

[23] Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213

[24] Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. *Springer Texts in Statistics*. New York: Springer. MR2080278

[25] Roberts, G.O., Gelman, A. and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120. MR1428751

[26] Roberts, G.O. and Rosenthal, J.S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist*. *Sci*. **16** 351–367. MR1888450

[27] Roberts, G.O. and Rosenthal, J.S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J*. *Appl*. *Probab*. **44** 458–475. MR2340211

[28] Roberts, G.O. and Rosenthal, J.S. (2009). Examples of adaptive MCMC. *J*. *Comput*. *Graph*. *Statist*. **18** 349–367. MR2749836

[29] Roodaki, A. (2012). Signal decompositions using trans-dimensional Bayesian methods. Ph.D. thesis, Supélec, Gif-sur-Yvette, France.

[30] Roodaki, A., Bect, J. and Fleury, G. (2012). Summarizing posterior distributions in signal decomposition problems when the number of components is unknown. In *IEEE International Conference on Acoustics*, *Speech*, *Signal Processing* (*ICASSP*) 3873–3876. Berlin: Springer.

[31] Sperrin, M., Jaki, T. and Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Stat*. *Comput*. **20** 357–366. MR2725393

[32] Stephens, M. (2000). Dealing with label switching in mixture models. *J*. *R*. *Stat*. *Soc*. *Ser*. *B Stat*. *Methodol*. **62** 795–809. MR1796293

Received October 2012 and revised October 2013