Adaptive Metropolis with online relabeling

R. Bardenet, O. Cappé, Gersende Fort, Balázs Kégl

To cite this version:

R. Bardenet, O. Cappé, Gersende Fort, Balázs Kégl. Adaptive Metropolis with online relabeling. Fifteenth International Conference on Artificial Intelligence and Statistics, Apr 2012, La Palma, Spain. pp.91-99. in2p3-00698479

HAL Id: in2p3-00698479
https://hal.in2p3.fr/in2p3-00698479

Submitted on 16 May 2012

Adaptive Metropolis with Online Relabeling

Rémi Bardenet
LAL & LRI, University Paris-Sud
91898 Orsay, France
bardenet@lri.fr

Olivier Cappé
LTCI, Telecom ParisTech & CNRS
46, rue Barrault, 75013 Paris, France
cappe@telecom-paristech.fr

Gersende Fort
LTCI, Telecom ParisTech & CNRS
46, rue Barrault, 75013 Paris (France)
gfort@telecom-paristech.fr

Balázs Kégl
LAL & LRI, University Paris-Sud & CNRS
91898 Orsay, France
balazs.kegl@gmail.com

Abstract

We propose a novel adaptive MCMC algorithm named AMOR (Adaptive Metropolis with Online Relabeling) for efficiently simulating from permutation-invariant targets occurring in, for example, Bayesian analysis of mixture models. An important feature of the algorithm is to tie the adaptation of the proposal distribution to the choice of a particular restriction of the target to a domain where label switching cannot occur. The algorithm relies on a stochastic approximation procedure for which we exhibit a Lyapunov function that formally defines the criterion used for selecting the relabeling rule. This criterion reveals an interesting connection with the problem of optimal quantifier design in vector quantization which was only implicit in previous works on the label switching problem. In benchmark examples, the algorithm turns out to be fast-converging and efficient at selecting meaningful non-trivial relabeling rules to allow accurate parameter inference.

1 Introduction

Adaptive Metropolis (AM) [1] is a powerful recent algorithmic tool in numerical Bayesian data analysis. AM builds on a well-known Markov Chain Monte Carlo (MCMC) algorithm but optimizes the rate of convergence to the target distribution by automatically tuning the design parameters of the algorithm on the fly. In the case of AM, the adaptive parameter tuning rule relies on the availability of an online estimate of the covariance of the target distribution. AM is considered a pioneering contribution in *adaptive MCMC* methodology, and it represents a significant step towards developing self-tuning, turn-key sampling algorithms which are necessary for using MCMC approaches in high throughput data analysis. The optimality criterion considered in AM relies on theoretical results derived when the target distribution is multivariate Gaussian [2, 3]. When the target is multi-modal or concentrated around a non-linear manifold, the algorithm still applies but its optimality is no longer guaranteed, and its practical performance is often suboptimal.

An important case where AM usually fails is when the target is the posterior distribution in Bayesian inference on a mixture model. In this case the mixture likelihood is invariant to permuting some of the mixture components, and the chosen prior often does not favor any permutation either. Although we usually have "nice" genuinely unimodal posteriors of the parameters of *well-identified* components, the posterior is highly multimodal with exponentially many peaks, one for each permutation of the components. Running a well-mixing MCMC on such a model results in useless marginal estimates for the parameters of the individual components due to the confusion between all possible labelings. This phenomenon is called *label switching*. Several approaches have been proposed to deal with this problem, usually in a *post-processing* step *after* the posterior sample has been produced [4, 5, 6, 7, 8, 9, 10, 11]. They all aim to solve the identifiability problem in order to produce meaningful marginal posteriors for the mixture parameters. Running vanilla AM on such a model has two pitfalls. If the chain does not mix well, we get stuck in one of the modes. The adaptive proposal can then be efficient but it is

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

unclear what the empirical distribution of the chain is and it is quite likely that the results of inference will be ultimately biased with respect to the posterior distribution. On the other hand, if the MCMC chain does switch between components, the online sample covariance estimate will be too broad, resulting in poor adaptive proposals and slow convergence. Note that the latter situation is usually prevalent when using trans-dimensional samplers based on the Reversible Jump approach [4]: in this case, the dimension-changing moves force label switching and running an embedded AM algorithm is, in our experience, almost useless.

To solve these challenging issues, we develop an Adaptive Metropolis algorithm with Online Relabeling (AMOR). The main idea is to combine the online relabeling strategy of [5] with the AM algorithm of [1]. This results in a doubly adaptive algorithm that uses the sample statistics both for *(i)* adapting its proposal and for *(ii)* redefining the region onto which it constrains the chain by attempting relabeling in each iteration. We prove two important facts about the AMOR algorithm. First, we show that when the adaptation is frozen, the target of the algorithm is indeed a restriction of the original target distribution to one of its symmetric “modes”. Thus, except for the automated selection of a particular relabeling strategy, the proposed algorithm does not modify the target distribution, a feature that is missing in some of the procedures commonly used to remedy label-switching. We also establish the existence of a Lyapunov function for AMOR, that is, a quantitative criterion which defines the set of possible limiting relabeling rules. This allows us to combine relabeling with adaptive proposals, something that is impossible in the usual setup where relabeling is done on the MCMC outcome in a post-processing step. The proofs of these results - to be found in the supplementary material - also unveil interesting connections with the problem of optimal quantifier design in vector quantization for which we prove an extension of known results regarding quantification using Mahalanobis divergence.

The rest of the paper is organized as follows. In Section 2 we formalize the problem and motivate AMOR on a real-world example. In Section 3, we derive the new online relabeling procedure based on AM and present our main convergence results. We show experimental results in Section 4 and conclude in Section 5.

2 Adaptive Metropolis and label switching

In this section we briefly provide some more background regarding AM and the label switching problem. Readers familiar with these notions may skip this section.

We consider using MCMC sampling to explore the posterior distribution

$$\pi(x) \triangleq p(x|\mathbf{y}) \propto p(\mathbf{y}|x)p(x)$$

of the parameters $x \in \mathcal{X}$ given the observation $\mathbf{y}$. The posterior $\pi(x)$, whose normalization constant is usually unknown, can be explored by running a Markov chain (X_t) with stationary distribution π. In this context, π is also said to be the *target distribution* of the MCMC chain. The Symmetric Random Walk Metropolis algorithm (SRWM; [12]; corresponding to the blue steps in Figure 2 in Section 3) is one of the most popular techniques for simulating such a chain (X_t). In SRWM the user has to provide a symmetric proposal kernel that will be used to propose a new sample $\tilde{X}$ given the previous sample X_{t-1}. When the posterior is a distribution over a continuous space $\mathcal{X} = \mathbb{R}^d$, the most common proposal kernel is a multivariate Gaussian $\mathcal{N}\big(\,\cdot\,|X_{t-1}, \Sigma\big)$.

The goal of *Adaptive Metropolis* (AM) (corresponding to the blue and green steps in Figure 2) is to automatically calibrate the design parameter Σ of SRWM. When the target $\pi(x)$ is multivariate Gaussian with covariance Σ_π, the optimal choice of Σ is of the order of $(2.38)^2\Sigma_\pi/d$ [2, 3]. In practice, Σ_π is unknown thus motivating the use of an estimate of the covariance of the posterior based on samples $(X_1, \ldots, X_{t-1})$ generated so far. From a theoretical point of view, the conditional distribution of X_t given the past then depends on the whole past, rendering the analysis of AM more challenging. The convergence of AM has been recently addressed under quite general conditions (see, e.g., [13, 14] and references therein).

The optimal choice for Σ is appropriate only when the target distribution is strongly unimodal [3]. If the data $\mathbf{y}$ is assumed to be drawn independently from a mixture model, its likelihood is of the form

$$p(\mathbf{y}|x) = \prod_i \sum_{m=1}^{M} \alpha_m f(y_i|\phi^{(m)}),$$

where $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$, $\phi^{(m)}$ denotes the n-dimensional parameter vector of the mth component, and the parameter space is a subset of $(\mathbb{R}^+ \times \mathbb{R}^n)^M$. The likelihood $p(\mathbf{y}|x)$ in this case is invariant under any permutation of the mixture components. If the prior $p(x)$ is *exchangeable*, that is, it also does not favor a particular permutation, then the posterior $\pi(x)$ inherits the permutation invariance.

To illustrate the challenges of inference in this model, we present an example motivated by a signal processing problem of the water Cherenkov signals of the Pierre Auger Observatory [15]. Figure 1(a)-1(b) display the MCMC sample corresponding to a single run of AM on an exponential mixture where the rates are known: only the location parameters and the mixture weights are estimated. A flat prior is taken. The red variable gets stuck in one of the mixture components, whereas the blue, green, and brown variables visit all the three remaining components. Marginal estimates computed for the blue, green, and brown variables are then mostly identical as seen on Figure 1(b). In addition, the shaded ellipses, depicting the marginal posterior covariances of each component's parameters, indicate that the resulting empirical covariance estimate is very broad, resulting in poor efficiency of the adaptive algorithm.

Several approaches have been proposed to deal with the label switching problem. The first solution consists in modifying the prior in order to make it select a single permutation of the variables, introducing an *identifiability constraint* [4]. This solution is known to cause artificial biases with respect to the posterior by not respecting its topology [7]. An effort has then been made to adapt to the estimated posterior surface through the design of relabeling algorithms [6, 7] that process the MCMC sample *after* the completion of the simulation run. These techniques look for a permutation P_t of each individual sample point X_t so as to minimize a posterior-based criterion depending on the *whole* chain history. [5] proposed an *online* version of the relabeling procedure in which the simulation of each X_t is followed by a permutation P_t of its components. The permutation P_t is chosen to minimize a user-defined criterion that depends only on the *past* history of the chain up to time t. The major advantage of this online approach is that it is compatible with our objective of solving label switching "on the fly" in order to optimize AM for permutation-invariant models.

In both batch and online relabeling algorithms, inference is carried out by using *relabeled* samples. Since the permutation steps in the MCMC procedure modify the distribution of the chain (X_t), the target distribution is no longer the posterior $\pi(x)$. [8] showed that, empirically, relabeling induces the learning of an appropriate identifiability constraint, but the existence of a target distribution and its relation with the original target $\pi(x)$ has been an open problem, meaning that these relabeling techniques have remained mostly heuristics. [11] recently proved a convergence result for a non-adaptive, batch, identifiability constraint-based relabeling procedure, where the constraint however depends on both the unrelabeled sample and the user. Our ultimate goal is to prove a similar but sample- and user-independent result, tightening adaptivity of the MCMC procedure with adaptivity of the identifiability constraint.

3 An adaptive online relabeling algorithm

From now on, we consider the general case of MCMC sampling from a target distribution with density $\pi(x)$ on $\mathcal{X} \subseteq \mathbb{R}^d$ with respect to (w.r.t.) the Lebesgue measure, with $d = qM$. Let $\mathcal{P}$ be a finite group of $d \times d$ block permutation matrices, indexed by some permutations ν of $\{1, ..., M\}$, acting on $\mathcal{X}$ as follows: for $x = (x_1, ..., x_M) \in \mathbb{R}^d$, $P_\nu x = (x_{\nu^{-1}(1)}, ..., x_{\nu^{-1}(M)})$. As an example, let us take $d = 6$ dimensions divided in three blocks of size 2. This would correspond to having a mixture model with $M = 3$ components and $q = 2$ parameters per component. The 6×6 matrix associated to the permutation ν that sends 1 onto 2, 2 onto 3 and 3 onto 1 is

$$P_\nu = \begin{pmatrix} 0 & 0 & I_2 \\ I_2 & 0 & 0 \\ 0 & I_2 & 0 \end{pmatrix},$$

and for $x = (1, 2, ..., 6)^T$, $P_\nu x = (5, 6, 1, 2, 3, 4)^T$.

Let us now assume that π is invariant under the action of $\mathcal{P}$, i.e. $\pi(Px) = \pi(x)$ for any $P \in \mathcal{P}$. Our goal is to isolate a single mode out of the many identically repeated symmetric modes of the posterior. Formally, we are interested in restricting the target π to a *fundamental domain* $\mathcal{D}$ of the action of $\mathcal{P}$, that is, finding a subset $\mathcal{D} \subset \mathcal{X}$ which is minimal for the inclusion and for which $\{Px : x \in \mathcal{D}, P \in \mathcal{P}\} = \mathcal{X}$, up to a set of Lebesgue measure zero. Following [5, 6], we will select $\mathcal{D}$ so that the sample looks as Gaussian as possible, since we want to select a single mode of the symmetric modes of the target π. For this purpose, the domain $\mathcal{D}$ will be defined through the minimization of a Mahalanobis-type criterion.

For a $d \times d$ invertible covariance matrix Σ, let

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2} L_{(\mu,\Sigma)}(x)\right)$$

be the density of a Gaussian distribution on $\mathbb{R}^d$ with mean μ and covariance matrix Σ, where

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

3.1 Derivation of the algorithm

Let $\mathcal{C}_d^+$ be the set of real $d \times d$ symmetric positive definite matrices, and let $\theta \in \mathbb{R}^d \times \mathcal{C}_d^+$. θ will later on be taken to be the concatenation of the mean and covariance of the chain (X_t), but for now, it is fixed to an arbitrary value. AMOR combines two actions: *(i)* sample a chain with target proportional to $\pi \mathbb{1}_{\mathcal{D}}$ where $\mathcal{D}$ is a fundamental domain of the action of $\mathcal{P}$ and *(ii)* learn the domain $\mathcal{D}$ on the fly (see Figure 2).

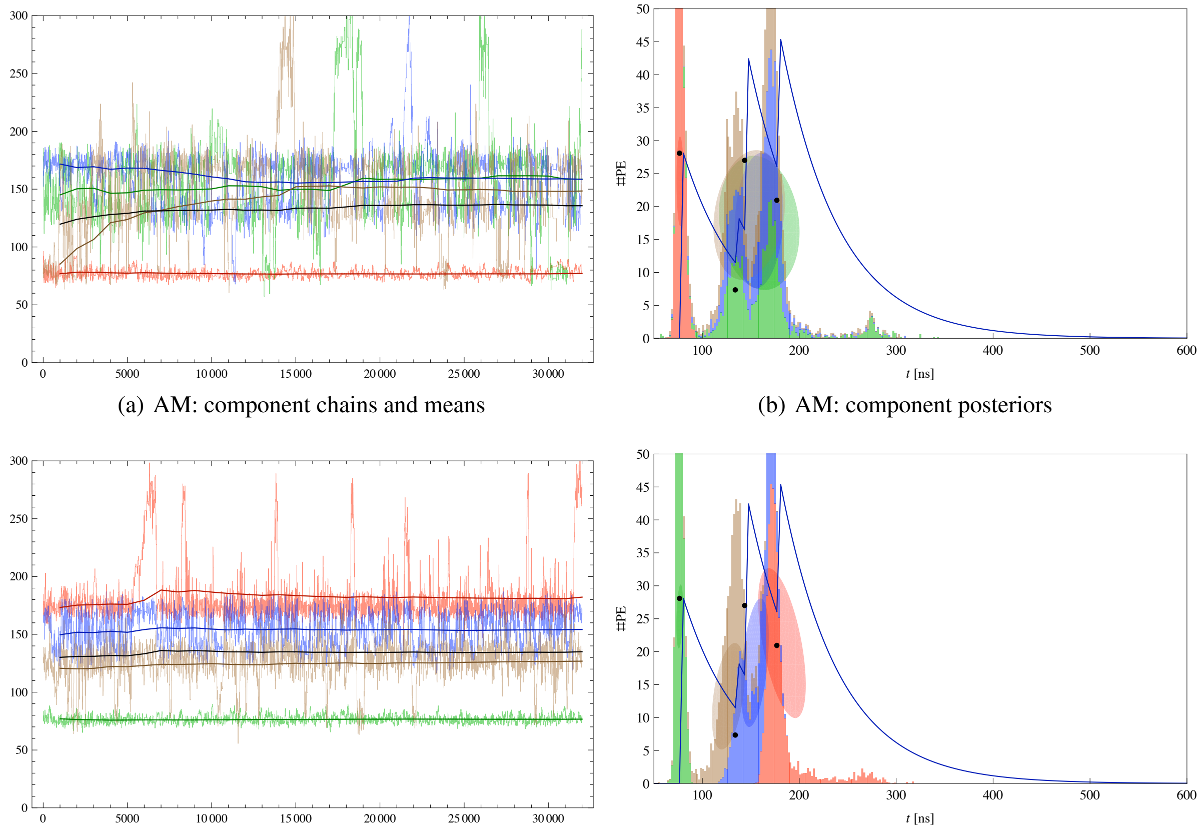
First assume that the adaptation is frozen, that is, consider AMOR when steps 12 and 13 in the pseudocode in Figure 2 are removed. In this case, we prove in Proposition 1 that our algorithm is a MCMC sampler with target distribution

$$\pi_\theta(x) = Z_\theta^{-1} \pi(x) \mathbb{1}_{V_\theta}(x), \quad \text{where } Z_\theta = \int_{V_\theta} \pi(x) dx, \tag{1}$$

and V_θ is defined by

$$V_\theta = \{x : L_\theta(x) = \min_{P \in \mathcal{P}} L_\theta(Px)\}.$$

In other words, $\pi_\theta(x)$ is the initial target $\pi(x)$ restricted to the nearest-neighbor (Voronoï) cell V_θ defined by the distortion measure L_θ.



(a) AM: component chains and means

(b) AM: component posteriors

(c) AMOR: component chains and means

(d) AMOR: component posteriors

Figure 1: The results of AM (top row) and AMOR (bottom row) algorithms on a mixture of exponential components. The right panels show the parameters of the four components (black dots: the x-coordinates are the location parameters $\phi^{(m)}$ and the y-coordinates are the mixture weights $\alpha^{(m)}$), the (unnormalized) mixture distribution (blue curve), and the marginal posteriors of the four location parameters (colored histograms). Colored ellipses are $\exp(1/2)$-level sets of Gaussian distributions: the means are the Bayesian estimates for the location and weight parameters of each component, and the covariance is the marginal posterior covariance of each location/weight couple. The left panels show the four chains of the location parameters $\phi^{(m)}$ (light colors), the running means (dark colors), and the mean of the running means (black curve). The AM algorithm shows heavy label switching among the three rightmost components whereas the AMOR algorithm separates the components nicely.

Proposition 1. *Let $\theta = (\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{C}_d^+$ and $c > 0$. Define a sequence $\{X_t, t \geq 0\}$ as prescribed by Figure 2: set $X_0 \in V_\theta$ and for $t \geq 0$,* (i) *sample $\tilde{X} \sim \mathcal{N}(\cdot|X_t, c\Sigma)$;* (ii) *conditionally on $\tilde{X}$, draw $\tilde{P}$ uniformly over the set* $\operatorname{argmin}_{P\in\mathcal{P}} L_\theta(P\tilde{X})$*;* (iii) *set $X_{t+1} = \tilde{P}\tilde{X}$ with probability $\alpha(X_t, \tilde{P}\tilde{X})$ and $X_{t+1} = X_t$ otherwise, where the acceptance ratio α is given by*

$$\alpha(x, \tilde{x}) = 1 \wedge \frac{\pi(\tilde{x})}{\pi(x)} \frac{\sum_{Q\in\mathcal{P}} \mathcal{N}(Qx|\tilde{x}, c\Sigma)}{\sum_{Q\in\mathcal{P}} \mathcal{N}(Q\tilde{x}|x, c\Sigma)} . \qquad (2)$$

Then $\{X_t, t \geq 0\}$ is a Metropolis-Hastings Markov chain with invariant distribution π_θ.

Proof We first prove by induction that $X_t \in V_\theta$ for any $t \geq 0$. This holds true for $t = 0$. Assume that $X_t \in V_\theta$. By construction, $\tilde{P}\tilde{X} \in V_\theta$; therefore, $X_{t+1} \in V_\theta$.

Recall that for a permutation matrix P, $P^{-1} = P^T$. Observe that given the current state X_t, $\tilde{P}\tilde{X}$ is sampled under the proposal distribution

$$\tilde{q}_\theta(x_t, x) = \sum_{P\in\mathcal{P}} p_\theta(P|P^T x) \mathcal{N}(P^T x|x_t, c\Sigma) ,$$

where the conditional distribution $p_\theta(\cdot|X)$ is the uniform distribution over $\operatorname{argmin}_{P\in\mathcal{P}} L_\theta(PX)$. As $\mathcal{P}$ is a group, $p_\theta(\cdot|PX) = p_\theta(\cdot|X)$ for any $P \in \mathcal{P}$. Furthermore, for any $x \in V_\theta$, the support of the distribution $\tilde{q}_\theta(x, \cdot)$ is in V_θ. This implies that for any $x, x' \in V_\theta$,

$$\tilde{q}_\theta(x, x') \propto \sum_{P\in\mathcal{P}} \mathcal{N}(Px'|x, c\Sigma) .$$

```
AMOR$\big(\pi(x), X_0, T, \mu_0, \Sigma_0, c\big)$
 1    $\mathcal{S} \leftarrow \emptyset$
 2    for $t \leftarrow 1$ to $T$
 3        $\Sigma \leftarrow c\Sigma_{t-1}$ ▷ scaled adaptive covariance
 4        $\tilde{X} \sim \mathcal{N}\big(\cdot\,|X_{t-1}, \Sigma\big)$    ▷ proposal
 5        $\tilde{P} \sim \arg\min_{P\in\mathcal{P}} L_{(\mu_{t-1},\Sigma_{t-1})}\big(P\tilde{X}\big)$    ▷ pick an optimal permutation
 6        $\tilde{X} \leftarrow \tilde{P}\tilde{X}$    ▷ permute
 7        if $\dfrac{\pi(X)\sum_P \mathcal{N}\big(PX_{t-1}|X, \Sigma\big)}{\pi(X_{t-1})\sum_P \mathcal{N}\big(PX|X_{t-1}, \Sigma\big)} > \mathcal{U}[0,1]$ then
 8            $X_t \leftarrow X$    ▷ accept
 9        else
10            $X_t \leftarrow X_{t-1}$    ▷ reject
11        $\mathcal{S} \leftarrow \mathcal{S} \cup \{X_t\}$    ▷ update posterior sample
12        $\mu_t \leftarrow \mu_{t-1} + \frac{1}{t}\big(X_t - \mu_{t-1}\big)$    ▷ update running mean and covariance
13        $\Sigma_t \leftarrow \Sigma_{t-1} + \frac{1}{t}\big((X_t - \mu_{t-1})\,(X_t - \mu_{t-1})^\intercal - \Sigma_{t-1}\big)$
14    return $\mathcal{S}$
```

Figure 2: The pseudocode of the AMOR algorithm. The steps of the classical SRWM algorithm are in blue, the adaptive MH algorithm adds the green steps, and the new online relabeling steps are in red. Notice the adaptation of both the proposal (line 4) and the selection mechanism through the dependence of $L_{(\mu,\Sigma)}$ on (μ, Σ). Note that for practical reasons, a small εI_d is often added to the covariance matrix in line 3, but [16] recently confirmed that core AM does not lead to degenerate covariances. Note also that line 5 is usually a simple assignment of the optimal permutation. In case of ties, we draw uniformly from the finite set of optimal permutations.

Therefore, for any $x, x' \in V_\theta$,

$$\begin{aligned}\alpha(x, x') &= 1 \wedge \frac{\pi_\theta(x)}{\pi_\theta(x)} \frac{\sum_{P\in\mathcal{P}} \mathcal{N}(Px|x', c\Sigma)}{\sum_{P\in\mathcal{P}} \mathcal{N}(Px'|x, c\Sigma)} \\ &= 1 \wedge \frac{\pi_\theta(x')}{\pi_\theta(x)} \frac{\tilde{q}_\theta(x', x)}{\tilde{q}_\theta(x, x')}. \quad \square\end{aligned}$$

V_θ is a fundamental domain of the action of $\mathcal{P}$ on $\mathcal{X}$. It contains only one copy of each genuine mode of π and it is sufficient to consider the restriction π_θ of π to V_θ: $\forall x \in \mathcal{X}, \exists P \in \mathcal{P}$ and $y \in V_\theta$ s.t. $x = Py$ and $\pi(x) = \pi(y)$ (by the permutation invariance of π). The sets $(PV_\theta)_{P\in\mathcal{P}}$, where $PV_\theta = \{Px : x \in V_\theta\}$, cover $\mathcal{X}$. If θ is further taken such that no P exists with $\mu = P\mu$ or $P\Sigma P^T = \Sigma$, then the sets $(PV_\theta)_{P\in\mathcal{P}}$ are pairwise disjoint. Proofs of these claims can be found in the supplementary material.

To sum up, we have a family of *cells* $(PV_\theta)_{P\in\mathcal{P}}$ that are permutations of each other, and where each cell PV_θ is the support of a permuted copy of the landscape of π (Figure 3). In view of the definition of V_θ and L_θ, forcing the chain to stay in V_θ means that we want to make the sample look “as unimodal as possible”.

The second step now is to find a convenient θ in such a way that MH is optimized: based on [3], we want to choose the covariance matrix Σ such that Σ is proportional to the covariance matrix of π_θ. This implies that $\theta = (\mu, \Sigma)$ solves the fixed point equations

$$\mu = \int x\,\pi_\theta(x)dx, \quad \Sigma = \int (x-\mu)(x-\mu)^T \pi_\theta(x)dx. \tag{3}$$

To achieve this goal, the *MH-online relabeling* (steps 3 to 11 in Figure 2) is combined with the adaptive steps 12 and 13. AMOR can thus be described as follows: let $\{K_\theta, \theta \in \mathbb{R}^d \times \mathcal{C}_d^+\}$ be the family of MH kernels described by Proposition 1. The posterior sample $\mathcal{S}$ in AMOR is obtained by running adaptive MCMC: given an initial value $\theta_0 = (\mu_0, \Sigma_0)$, define by induction $\{(X_t, \theta_t), t \geq 0\}$ as follows:

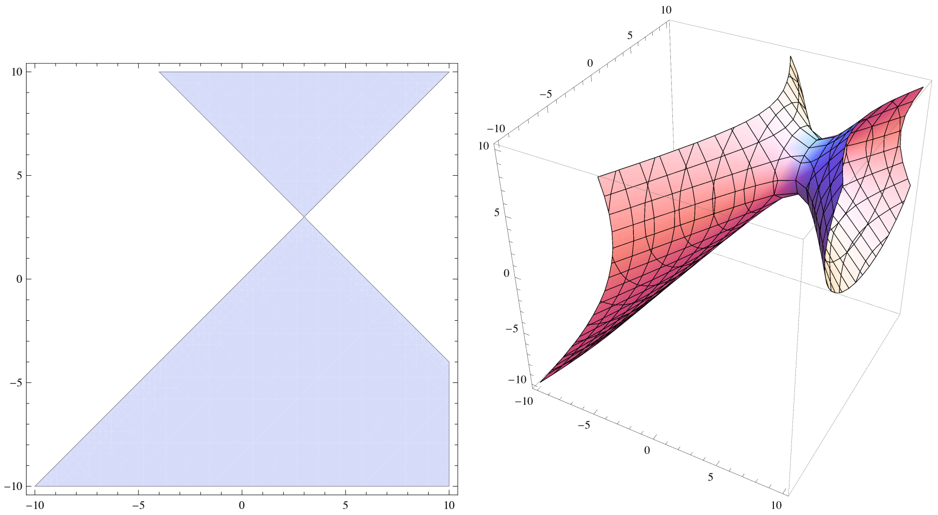
- Sample $X_t \sim K_{\theta_{t-1}}(X_{t-1}, \cdot)$,
- Update the parameter:

$$\theta_t = \theta_{t-1} + \frac{1}{t} H(X_t, \theta_{t-1}), \tag{4}$$

where

$$H(x, \theta) = (x - \mu, (x-\mu)(x-\mu)^T - \Sigma). \tag{5}$$

Recent results on convergence of adaptive MCMC show that when this algorithm converges, there exists $\theta_\star \in$



(a) 2D case

(b) 3D case

Figure 3: Examples of tessellations $(PV_\theta)_{P \in \mathcal{P}}$. (a) The plane is always cut into two parts in a "butterfly" shape. The central intersection point can move along the first diagonal, eventually until ∞. (b) Borders between two cells in 3D are quadratic hypersurfaces.

$\mathbb{R}^d \times \mathcal{C}_d^+$ such that the distribution of the posterior sample $\{X_t, t \geq 0\}$ converges to $\pi_{\theta_\star}$ and $n^{-1}\sum_{t=1}^n f(X_t) - \pi_{\theta_\star}(f)$ converges a.s. to zero for a wide family of functions f [14]. Therefore, the empirical mean and covariance of $\{X_t, t \geq 0\}$ converges to the expectation and covariance matrix of $\pi_{\theta_\star}$. Hence $\theta_\star$ solves the fixed point equations (3). In Section 3.2 we will discuss sufficient conditions for the convergence of AMOR.

Allowing only scalar matrices ($\Sigma_t = \lambda_t I$) for selection and using fixed-covariance proposals, our relabeling mechanism coincides with the one of [5]. Beside the obvious generalization to full covariance matrices and the adaptivity of the proposal, the breakthrough of our algorithm lies in its modification of the acceptance ratio that allows to identify the limiting distribution of the sample $\{X_t, t \geq 0\}$ and to prove limit theorems (e.g., ergodicity and law of large numbers).

3.2 Convergence analysis

As discussed in section 3.1, AMOR is an adaptive MCMC algorithm such that for any $\theta \in \mathbb{R}^d \times \mathcal{C}_d^+$, K_θ has its own invariant distribution π_θ. Sufficient conditions for the convergence of such algorithms have been recently derived [14]. Three main conditions are required. The first condition is that adaptation has to be diminishing; this means that some suitable distance between two consecutive kernels $K_{\theta_{t-1}}$ and K_{θ_t} vanishes. When adaptation relies on stochastic approximation as in the case of AMOR, this condition is easy to check [14, 13]. The second condition is the *containment condition*: roughly speaking, a kind of uniform-in-θ ergodic behavior of the kernels $\{K_\theta, \theta \in \mathbb{R}^d \times \mathcal{C}_d^+\}$ is required. As shown in [14, 13], this property is related to the stability of the random sequence $\{\theta_t, t \geq 0\}$. The third condition is to control the weak convergence of the random invariant distributions $\{\pi_{\theta_t}, t \geq 0\}$ to the limiting target $\pi_{\theta_\star}$. In the case of AMOR, this amounts to control both the stability and the almost sure convergence of the sequence $\{\theta_t, t \geq 0\}$.

Therefore, a key ingredient for the proof of the convergence of AMOR is to show the stability and the convergence of the stochastic approximation sequence $\{\theta_t, t \geq 0\}$. Existing results cannot apply since the draws are neither i.i.d. nor Markovian. However, since the draws X_t are obtained by applying a new kernel at each iteration, our idea is to adapt the existing results for the Markovian case. While the full proof of convergence of AMOR is out of the scope of this paper, we address here one of its most important aspects which unveils a novel relation between relabeling and vector quantization.

The main tool for the stability and convergence of stochastic approximation algorithms is the existence of a Lyapunov function [17]. The space $\mathbb{R}^d \times \mathcal{C}_d^+$ is endowed with the scalar product $\langle \theta_1, \theta_2 \rangle = \mu_1^T \mu_2 + \mathrm{Trace}(\Sigma_1 \Sigma_2)$. A continuously differentiable function $w : \mathbb{R}^d \times \mathcal{C}_d^+ \to \mathbb{R}^+$ is a Lyapunov function for the mean field h if *(i)* for any $\theta \in \mathbb{R}^d \times \mathcal{C}_d^+$, $\langle \nabla w(\theta), h(\theta) \rangle \leq 0$; *(ii)* the level sets $\{w \leq C\}$, $C > 0$, are compact subsets of $\mathbb{R}^d \times \mathcal{C}_d^+$. When $w(\mathcal{F})$ has empty interior, where $\mathcal{F} = \{\theta \in \mathbb{R}^d \times \mathcal{C}_d^+ : \langle \nabla w(\theta), h(\theta) \rangle = 0\}$, then the sequence $\{\theta_t, t \geq 0\}$ defined by (4) converges to $\mathcal{F}$ [18]. Proposition 2 shows that the function w given by $w(\theta) = -\int \log \mathcal{N}(x|\theta)\, \pi_\theta(x) dx$ is a natural candidate for the Lyapunov function. It is also proven that $\mathcal{F}$ is the set of points (μ, Σ) satisfying the fixed point equation (3).

Proposition 2. *Let us consider the AMOR algorithm (see Figure 2) with quadratic loss*

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

Define the mean field h on $\mathbb{R}^d \times \mathcal{C}_d^+$ by

$$\begin{aligned} h(\theta) &= \mathbb{E}\left[H(X,\theta)\right] \\ &= \left(\mu_{\pi_\theta} - \mu, (\mu_{\pi_\theta} - \mu)(\mu_{\pi_\theta} - \mu)^T + \Sigma_{\pi_\theta} - \Sigma\right), \end{aligned}$$

and let

$$\Theta = \{\theta \in \mathbb{R}^d \times \mathcal{C}_d^+ : \forall P \in \mathcal{P}, \Sigma \neq P\Sigma P^T \text{ or } \mu \neq P\mu\}.$$

Then $\langle \nabla w(\theta), h(\theta)\rangle \leq 0$ for any $\theta \in \Theta$ and $\langle \nabla w(\theta), h(\theta)\rangle = 0$ iff $\theta = (\mu, \Sigma)$ solves the fixed point equation (3).

The proof of Proposition 2 is given in the supplementary material (Corollary 1). We also show that for any $\theta \in \Theta$, $w(\theta)$ is, up to an additive constant, the Kullback-Leibler divergence between $\mathcal{N}(\cdot|\theta)$ and π_θ. Equivalently $w(\theta)$ is the Kullback-Leibler divergence between the original *unrelabeled* posterior π and a mixture of Gaussians, whose component parameters are images of each other through the symmetries that leave π invariant. Finally, Proposition 3 establishes that when $\theta \in \Theta$, $w(\theta)$ is a distortion measure in vector quantization [19], which is the key property for the proof of Proposition 2, and unveils a novel link between clustering and relabeling. Notice that $w(\theta)$ is exactly a distortion measure, as the first term of the right-hand side of Proposition 3 is constant over the set $\{P\Sigma P^T, P \in \mathcal{P}\}$.

Proposition 3. *For any $\theta \in \Theta$,*

$$w(\theta) = \frac{1}{2}\ln\det(\Sigma) + \frac{1}{2}\int \min_{P\in\mathcal{P}} L_{(P\mu, P\Sigma P^T)}(x)\,\pi(x)dx.$$

4 Experiments

We benchmarked the AMOR algorithm on two Bayesian inference tasks. The first one is the problem of estimating the nine parameters $\psi = (\alpha_i, \mu_i, \sigma_i)_{i=1,2,3}$ of a mixture of three one-dimensional Gaussians $\sum_{i=1}^3 \alpha_i \mathcal{N}(.|\mu_i, \sigma_i)$, taking wide flat priors over each parameter. We compared four algorithms: *(i)* an SRWM with an ordering constraint on the three means $\mu_1 \leq \mu_2 \leq \mu_3$, *(ii)* the original online relabeling algorithm of [5], *(iii)* the same online algorithm with modified acceptance ratio according to (2) (henceforth denoted as MC for Modified Celeux), and *(iv)* the AMOR algorithm. To quantify the performance after T iterations, we first selected the permutation of the running posterior mean components $(\hat{\mu}_i^{(T)})_{i=1,2,3}$ which minimized the sum of the ℓ_2 errors on the three estimates of the means $\mu_i, i = 1,2,3$, and we considered the latter sum taken at this best permutation of the posterior mean:

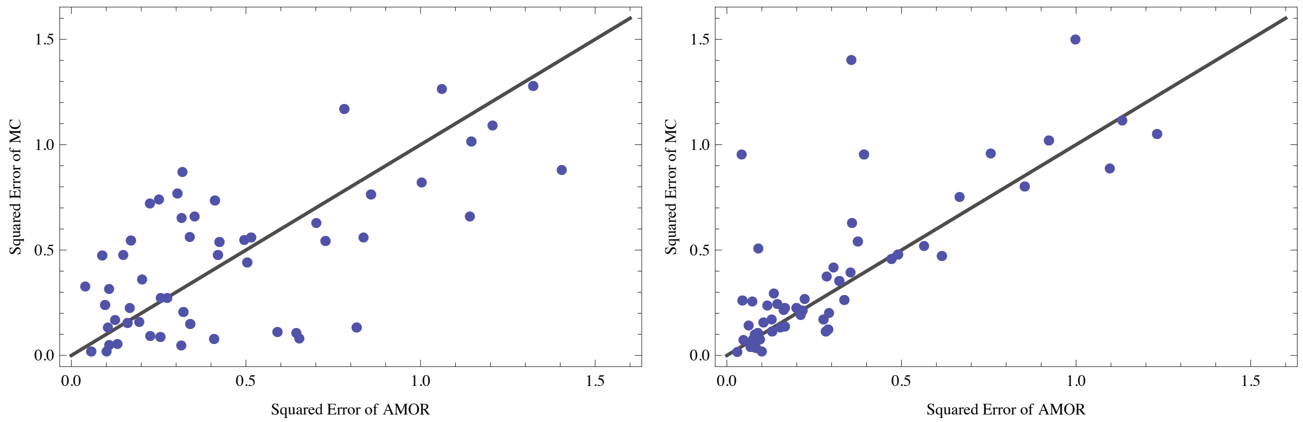
$$S_T = \arg\min_{\tau\in\mathfrak{S}_3} \sum_{i=1}^3 (\hat{\mu}_{\tau(i)}^{(T)} - \mu_i)^2.$$

We repeated this experiment 100 times on 100 different datasets coming from parameters generated as follows: draw $(\alpha_i) \sim \mathcal{D}(1)$, $\mu_i \sim \mathcal{U}_{(0,1)}$ i.i.d., and $\sigma_i \sim \mathcal{U}_{(0,0.05)}$ i.i.d. This choice of generative distribution ensured a reasonable number of datasets containing overlapping Gaussians, thus provoking switching. Figure 5(a) depicts the performance measure S_T averaged over the 100 datasets of this 9D experiment, versus time T. We use this averaging as a way to estimate the expected performance measure on a class of problems given by the generative distribution. AMOR significantly outperforms other approaches as it converges faster on average and to a better solution. As expected, imposing an ordering constraint on the means (RWM+OC) reveals a poor strategy leading to artificial additional bias. Note finally that the modification of the acceptance ratio did not increase drastically the performance of the Online Relabeling algorithm of [5] ("Original RWM+OR" vs. "Modified RWM+OR"), which is not a surprise since the additional factor in the ratio (2) is often close to 1. Figure 4 provides insight into how the two best methods (AMOR and MC) behaved after $T_1 = 1$K and $T_2 = 30$K iterations, presenting scatter plots of performances $S_{1,000}$ and $S_{30,000}$. Each point corresponds to one of the 100 datasets. Clearly, starting from a rather random distribution of the errors, AMOR took the advantage after 30k iterations, while a few cases were still better treated by MC.

We also performed a higher dimensional experiment to further investigate the comparison between AMOR and CC. This time, the goal was to estimate the three means of a 10-dimensional Gaussian mixture $\sum_{i=1}^3 1/3\,\mathcal{N}(.|\mu_i, 0.1 I_{10})$. Again, 100 datasets of 100 points each were generated with $\mu_i \sim \mathcal{U}_{(0,1)}$ i.i.d. Again, as seen Figure 5(b), AMOR stabilizes earlier and selects a better region of $\mathbb{R}^{30}$ than MC, thus illustrating the interest of combining adaptive selection and proposal mechanisms.

5 Conclusion

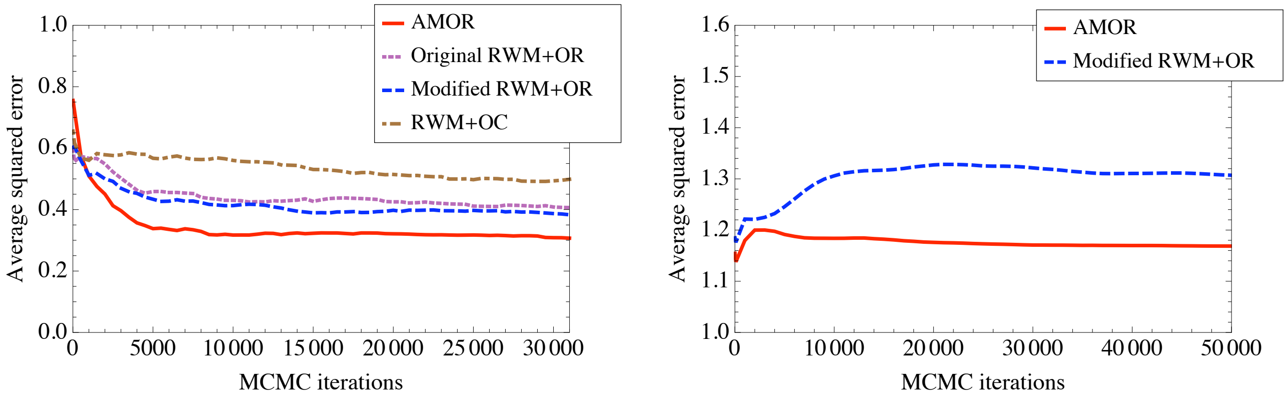
We derived AMOR, a novel MCMC algorithm for simulating from a permutation invariant target. AMOR combines Adaptive Metropolis with Online Relabeling. We justified its use by showing that it relies on a stochastic approximation procedure for which we exhibited a Lyapunov function. Experiments show that AMOR outperforms existing approaches for Bayesian inference in Gaussian mixture models. Besides, the theoretical framework of its analysis is appealing because *(i)* it generalizes previous approaches



(a) $S_{1,000}$ in the 9D experiments

(b) $S_{30,000}$ in the 9D experiments

Figure 4: Experimental comparison of the performance measure S for AMOR (x-axis) versus Modified Celeux (y-axis) in the 9D experiment.



(a) 9D experiments

(b) 30D experiments

Figure 5: Experimental comparison of several relabeling approaches. The plots show the performance measure S_T vs. T averaged over 100 datasets drawn from a common generative model.

[5], *(ii)* it paves the way towards future work on the asymptotic behavior of relabeling algorithms and convergence of the samples $\{X_t, t \geq 0\}$, and *(iii)* the Lyapunov function we derived exhibits an elegant relation with vector quantization techniques.

However, these features come at a price: future work should try to remove the need to sweep over all permutations in $\mathcal{P}$, which is prohibitive when using AMOR with large $|\mathcal{P}|$. Future work could also study variants of AMOR, using, e.g., a *soft clustering* selection mechanism, replacing the indicator appearing in the definition of π_θ by a *logistic* indicator, or equivalently relabel each new Gaussian sample in AMOR by sampling from a multinomial distribution over permutations conditionally on the history statistics, a technique similar in spirit to *probabilistic* relabeling algorithms developed in [9] and [10].

Acknowledgements

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

References

[1] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

[2] G. Roberts, A. Gelman, and W. Gilks. Weak convergence of optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120, 1997.

[3] Roberts G. O. and Rosenthal J. S. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

[4] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59(4):731–792, 1997.

[5] G. Celeux. Bayesian inference for mixtures: The label-switching problem. In R. Payne and P. Green, editors, *COMPSTAT 98*. Physica-Verlag, 1998.

[6] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, 62:795–809, 2000.

[7] J.M. Marin, K. Mengersen, and C.P. Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of Statisics*, 25, 2004.

[8] A. Jasra, C. C. Holmes, and D. A. Stephens. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statistical Science*, 20(1):50–67, 2005.

[9] A. Jasra. *Bayesian inference for mixture models via Monte Carlo*. PhD thesis, Imperial College London, 2005.

[10] M. Sperrin, T. Jaki, and E. Wit. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20:357–366, 2010.

[11] P. Papastamoulis and G. Iliopoulos. An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distribution. *Journal of Computational and Graphical Statistics*, 19:313–331, 2010.

[12] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[13] E. Saksman and M. Vihola. On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. *Annals of Applied Probability*, 20:2178–2203, 2010.

[14] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Annals of Statistics*, 2011 (to appear).

[15] Pierre Auger Collaboration. Pierre Auger project design report. Technical report, Pierre Auger Observatory, 1997.

[16] M. Vihola. Can the adaptive Metropolis algorithm collapse without the covariance lower bound? *Electronic Journal of Probability*, 16:45–75, 2011.

[17] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[18] C. Andrieu, E. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 44:283–312, 2005.

[19] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer-Verlag, 2000.