Simulated Annealing for Maximum A Posteriori Parameter Estimation of Hidden Markov Models

Christophe Andrieu and Arnaud Doucet

Abstract—Hidden Markov models are mixture models in which the populations from one observation to the next are selected according to an unobserved finite state-space Markov chain. Given a realization of the observation process, our aim is to estimate both the parameters of the Markov chain and of the mixture model in a Bayesian framework.

In this paper, we present an original simulated annealing algorithm which, in the same way as the EM (Expectation–Maximization) algorithm, relies on data augmentation, and is based on stochastic simulation of the hidden Markov chain. This algorithm is shown to converge toward the set of Maximum A Posteriori (MAP) parameters under suitable regularity conditions.

Index Terms—Bayesian estimation, data augmentation, hidden Markov models, maximum a posteriori, simulated annealing.

I. INTRODUCTION

HIDDEN Markov models (HMM) are models in which the distribution from one observation to the next is selected according to an unobserved “hidden” finite state-space Markov chain. These models have many applications in signal processing [1], communications [2], [3], and applied statistics [5]–[9]. In several practical cases of interest, neither the parameters of the hidden Markov chain nor the parameters of the mixture model are known and have to be estimated from a finite number of observations. The most popular method used to perform Maximum-Likelihood (ML) or Penalized ML estimation of these parameters is the EM (Expectation–Maximization) algorithm [10]. When the number of observations is very large, the EM algorithm yields very satisfactory results in most applications. However, there are also well-documented shortcomings of this deterministic algorithm. For example, it can converge to local maxima or saddle points of the (penalized) log-likelihood function, and its limiting position is dependent on initialization. When small sample sizes are available, the objective function of the maximization becomes highly irregular and the EM algorithm can perform poorly. This is one of the main reasons why stochastic versions of EM, including SEM (Stochastic EM) [11] and MCEM (Monte Carlo EM) [6], [12] have been developed. In some applications, such as blind ML filter identification [4] or estimation of finite mixture distributions [13], it has been shown experimentally that these stochastic algorithms outperform the EM algorithm. However, contrary to the algorithm presented here, there is a lack of theoretical convergence results for SEM and MCEM.

In this paper, we address the problem of the estimation of these parameters in a Bayesian framework; that is, these parameters are regarded as random with appropriate prior distributions. Several Markov chain Monte Carlo (MCMC) algorithms have been proposed to estimate the posterior distribution of these parameters for models from this class [3], [5], [7], [14]. In a Bayesian framework, MCMC methods consist of building a homogeneous Markov chain whose asymptotic distribution converges to the posterior distribution of interest. These methods are particularly suitable for the estimation of conditional expectations, e.g., MMSE (Minimum Mean-Square Estimate), via sample path averages [15], [16], but rather inefficient for the estimation of the MAP values of the parameters. Indeed, these homogeneous algorithms do not focus especially on high posterior density regions. Thus a large amount of the computational burden is spent exploring regions of low posterior probability, which are of no interest when one wants the MAP estimate. Here, in order to achieve efficient computation of the MAP estimate for parameters of HMM, we design a simulated annealing algorithm which, under given conditions, converges toward the set of these estimates in the sense specified in Section IV.

In this context, “simulated annealing” means that we build an inhomogeneous Markov chain whose sequence of invariant distributions concentrates on the set of the MAP parameters. Classical simulated annealing relies on a particular case of the Metropolis algorithm: exploration of the objective function is based on a random walk [17], [18]. In a Bayesian framework, such an implementation would not make use of the statistical structure of the model. Furthermore, one of the main drawbacks of classical simulated annealing based on the Metropolis algorithm is that it is not usually theoretically valid when some of the parameters to be estimated are continuous and do not lie in a compact set [19], [20]. Here we construct an algorithm which exploits the statistical structure of the model by introducing a set of missing data, here the hidden Markov chain, in a similar way to the EM algorithm [1]. This data augmentation provides the algorithm with two interesting properties. The first is algorithmic as it allows the parameters to be easily updated. The second is theoretical as the discrete nature of the missing data set ensures convergence of the algorithm for the whole set of parameters, including continuous parameters lying in unbounded sets. Using Monte Carlo simulations, we compare this algorithm and the EM algorithm. This algorithm is shown to perform better than EM, in the sense specified in Section V, while having an equivalent computational complexity at each iteration.
The paper is organized as follows. In Section II we formally present the model and the estimation objectives. In Section III, we introduce the simulated annealing algorithm and discuss implementation issues. In Section IV, we give sufficient conditions to ensure convergence of this scheme toward the set of MAP parameters. In Section V, we demonstrate the application of this algorithm to switching autoregressions with a Markov regime and to Markov-modulated Poisson processes. In Appendix A, notation is given. In Appendix B, the forward filtering–backward sampling procedure is described. In Appendix C, the way to evaluate the likelihood function is given. Finally, the proofs of convergence are grouped in Appendix D.

II. MODEL AND PROBLEM FORMULATION

A. Model Structure

Let us formally detail the signal model. Let \( k \in \{1, 2, \cdots \} \) denote a discrete time index. Let \( r_k \) denote a discrete-time, time-homogeneous, \( s \)-state, first-order Markov chain with transition probabilities

\[
\pi_{ij} \triangleq \Pr\{r_{k+1} = j | r_k = i\}, \quad (i, j \in S); \quad S = \{1, 2, \cdots, s\}.
\]

Denote the initial probability distribution as \( \pi_0 \), such that \( \sum_{i=1}^{s} \pi_i = 1 \) and \( \pi_i \geq 0 \), for each \( i \in S \). For notational convenience, denote the initial distribution \( \pi_0 \), the \( i \)th row of the transition matrix \( \pi_0 \), and the Markov chain and observations from time 1 to \( k \)

\[
r_k \triangleq (r_1, \cdots, r_k)
\]

and

\[
y_k \triangleq (y_1, \cdots, y_k)
\]

respectively.

At each time \( k \), the observation \( y_k \) is distributed conditional upon \( y_{k-1} \), \( r_k = i \) and \( \theta_i \), the vector parameter of the distribution, as

\[
y_k | (y_{k-1}, \theta_i, r_k = i) \sim f_i(y_k | y_{k-1}, \theta_i).
\]

Unconditionally, we have

\[
P(y_k | y_{k-1}, \theta) = \begin{cases} \sum_{i=1}^{s} f_i(y_k | \theta_i) \pi_i, & \text{if } k = 1 \\ \sum_{i=1}^{s} f_i(y_k | y_{k-1}, \theta_i) \Pr(r_k = i), & \text{if } k \geq 2 \end{cases}
\]

which is a finite mixture distribution. The set of unknown parameters to estimate is \( \theta \triangleq \{\theta_i, \pi_0, \pi_{ij}, i, j \in S\} \in \Theta \subset \mathbb{R}^{\eta\theta} \), where \( \eta\theta \) is the dimension of \( \theta \).

B. Bayesian Model

In this paper, we adopt a fully Bayesian approach, that is, the parameter \( \theta \) is regarded as random with prior distribution \( P(\theta) \). Prior distributions are used, in principle, to incorporate our knowledge of the parameters. Less informative priors should be employed when such knowledge is limited [21, pp. 357–367].

\[
P(\theta) \text{ is a probability measure on } (\Theta, \mathcal{B}(\Theta)), \text{ where } \mathcal{B}(\Theta) \text{ is the Borel } \sigma\text{-algebra associated to the parameter space } \Theta.
\]

\[
P(\theta) \text{ is assumed absolutely continuous with respect to the Lebesgue measure } d\theta, \text{ and we denote by } p(\theta) \text{ the corresponding density.}
\]

The following structure for \( p(\theta) \) is assumed:

\[
p(\theta) = p(\theta_1, \cdots, \theta_s) \prod_{i=1}^{s} p(\pi_i)
\]

that is to say that the parameters of the mixture models, the initial distribution, and the rows of the transition matrix are mutually independent. We also assume that \( p(\theta) \) can be evaluated pointwise up to a proportionality factor and that \( \pi_i, i = 0, \cdots, s \) have Dirichlet distributions as priors [21, pp. 134–135], i.e.,

\[
\pi_i \sim \text{Dir}(\alpha_i), \text{ where } \alpha_i = (\alpha_{i1}, \cdots, \alpha_{is}) \text{ and } \alpha_{ij} > 0. \text{ If } \alpha_{ij} = 1 \text{ for all } (i, j) \in \{0, 1, \cdots, s\} \times S \text{ we obtain the uninformative uniform distribution commonly used in Bayesian estimation} \ [8], \ [14]. \text{ Furthermore, this distribution is a conditionally conjugate prior} \ [21, pp. 265–285], \text{ which allows efficient stochastic simulation of these parameters as required by our algorithm. Similar prior distributions are adopted in} \ [3], \ [5], \ [7], \ [14]. \text{ The prior distributions for } \theta_1, \cdots, \theta_s \text{ will depend upon the application under study. Several examples will be presented in Section V.}
\]

C. Estimation Objectives

Given the observations \( y_N \), where \( N \) is the number of available data, our objective is to obtain the MAP estimate \( \hat{\theta}_{MAP} \) of \( \theta \) defined as

\[
\hat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} p(\theta | y_N).
\]

Performing this estimation involves solving a complex global optimization problem on a general (continuous and noncompact) state space.

III. SIMULATED ANNEALING VIA DATA AUGMENTATION

We present here a simulated annealing-like algorithm which, in a similar way to the EM algorithm, relies on the introduction of the hidden Markov chain \( r_N \) to yield a so-called “augmented data set.” The finite nature of this missing data set is the crucial feature which provides the algorithm with good convergence properties.

A. Presentation of the Algorithm

We construct an inhomogeneous Markov chain

\[
\left\{ \left( \theta^{(n)}, s^{(n)}_N \right); n \in \mathbb{N} \right\}
\]

whose transition kernel at iteration \( n \) depends on a deterministic cooling schedule \( \{T(n); n \in \mathbb{N}\} \ [18] \) satisfying

\[
T(n) + 1 \leq T(n) \text{ and } \lim_{n \to +\infty} T(n) = 0.
\]

\( \theta^{(n)} \) is an estimate of the set of parameters \( \theta \) at iteration \( n \). Sufficient conditions on the cooling schedule to ensure convergence of the sequence \( \{\theta^{(n)}; n \in \mathbb{N}\} \)
toward the set of global maxima of $p(\theta | y_N)$ are given in Section IV. The algorithm proceeds as follows.

### Simulated Annealing via Data Augmentation

1) **Initialization.** Set $(\theta^{(0)}, r_N^{(0)}) \in \Theta \times S_N$ randomly.
2) **Iteration** $n, n \geq 1$
   - Sample $r_N^{(n)} \sim p(r_N | y_N, \theta^{(n-1)})$.
   - Sample $\theta' \sim p(\theta | y_N, r_N^{(n)})$.
   - Evaluate the acceptance probability
     \[
     \alpha_n (\theta^{(n-1)}, \theta') = \min \left\{ \left[ \frac{p(\theta | y_N)}{p(\theta^{(n-1)} | y_N)} \right]^{1/T(n)-1}, 1 \right\} 
     \]
     - Sample $u \sim U[0,1]$. If $u \leq \alpha_n (\theta^{(n-1)}, \theta')$ then set $\theta^{(n)} = \theta'$, otherwise set $\theta^{(n)} = \theta^{(n-1)}$.
3) Set $n = n + 1$ and go to step 2).

### B. Implementation Issues

At each iteration $n$, one must be able to sample efficiently from $p(r_N | y_N, \theta^{(n-1)})$ and $p(\theta | y_N, r_N^{(n)})$, and then evaluate an acceptance probability $\alpha_n (\theta^{(n-1)}, \theta')$.

1) **Simulation from $p(r_N | y_N, \theta)$:** $p(r_N | y_N, \theta)$ is a discrete probability distribution. However, the evaluation of this distribution would be cumbersome due to its high dimensionality for large $N$. Fortunately, it is not necessary to evaluate this distribution to sample from it. Here we use the forward filtering–backward sampling recursion recently introduced independently by Carter and Kohn [22] and Chib [5]. This efficient recursion is detailed in Appendix B. Its computational cost is $O(N^2)$, which is linear in the data length $N$.

2) **Simulation from $p(\theta | y_N, r_N)$:** From the assumptions on the prior distribution and the observation model, we have
   \[
   p(\theta | y_N, r_N) = 1 \prod_{i=1}^{s} p(\pi_i | y_N, r_N). 
   \]

Sampling from $p(\theta_1, \ldots, \theta_s | y_N, r_N)$ is problem-dependent and typically relies on classical techniques [23] which will be illustrated in Section V. Sampling from $p(\pi_i | y_N, r_N) = p(\tau_i | y_N)$ is easy as we obtain
   \[
   \pi_i | r_N \sim D_{\delta_j} (\alpha_{ij} + n_{ij1}, \ldots, \alpha_{ij} + n_{ij}) \] (8)
   \[
   \text{where } n_{ij} = \delta_{i,j} (\delta_{i,j} = 1 \text{ if } i = j \text{ and } 0 \text{ otherwise}), \text{ and } \delta_{i,j} \text{ is the total number of one-step transitions from state } i \text{ to state } j \text{ in } r_N.
   \]

3) **Evaluation of the Acceptance Probability $\alpha_n (\theta, \theta')$:**
   The acceptance probability $\alpha_n (\theta, \theta')$ can be computed because we can evaluate the following probability in a pointwise manner up to a normalizing constant:
   \[
   p(\theta | y_N) \propto p(y_N | \theta) p(\theta). 
   \]
   The second term on the right is the prior distribution which is typically easily evaluated analytically up to a proportionality factor. The first term is the likelihood for the data which is usually calculated using Baum’s forward–backward formula [1]. Here we do not make use of this recursion because, while simulating from $p(y_N | \theta; \theta)$, auxiliary quantities (denominator of (40)) are computed that can be used to evaluate the likelihood. This alternative evaluation is detailed in Appendix C.

### IV. STUDY OF THE ALGORITHM

In this section, our main result is Theorem 1, which proves that the proposed algorithm converges to the set of MAP parameters in the following sense:
   \[
   \lim_{n \to +\infty} ||\mu_n - \bar{p}^{1/T(n)}|| = 0 
   \]
   where $||\cdot||$ is a norm defined in Section IV-A. $T(n) \in \mathbb{N}$ is an appropriate logarithmic cooling schedule (see Theorem 1). $\mu_n$ is the distribution of the $n$th sample $\theta^{(n)}$ of the inhomogeneous Markov chain built from our algorithm, and $\bar{p}^{1/T(n)} (\cdot | y_N)$ ($n \in \mathbb{N}$) is the sequence of normalized versions of $p^{1/T(n)} (\cdot | y_N)$, which converges to a distribution located on the set of MAP parameters (Lemma 3).

Our proof of convergence relies on the study of the inhomogeneous Markov chain $\{\theta^{(n)}; n \in \mathbb{N}\}$. In Section IV-A, notation and hypotheses are given. In Section IV-B, we show that each transition kernel of the Markov chain $\{\theta^{(n)}; n \in \mathbb{N}\}$ is uniformly ergodic. In Section IV-C, we study the behavior of the sequence of invariant distributions of these transition kernels. Then, in Section IV-D, we show that a logarithmic cooling schedule ensures a convergence result of the algorithm toward the set of global maxima of the posterior distribution.

#### A. Notations and Assumptions

Let $\Lambda$ be the set of all probability measures and $\Xi$ the set of all Markov transition kernels on $(\Theta, B(\Theta))$. We use the following standard definitions and notations [16], [19], [24]. The total variation norm between two elements $\mu$ and $\nu$ of $\Lambda$ is defined as
   \[
   ||\mu - \nu|| \triangleq \sup_{A \in B(\Theta)} |\mu(A) - \nu(A)|. 
   \]

Let $\{\theta^{(n)}; n \in \mathbb{N}\}$ be an inhomogeneous Markov chain of initial distribution $\mu_0 \in \Lambda$ and Markov transition kernel $K_n \in \Xi$ at iteration $n$ satisfying for any $A \in B(\Theta)$
   \[
   \Pr \left( \theta^{(n)} \in A \right) \triangleq \int_A K_n (\cdot | \theta^{(n-1)}), \theta^{(n-1)}). 
   \]

We define, for $m < n$, $K^{(m,n)} \triangleq K_{m+1} K_{m+2} \cdots K_n$ so that the probability distribution $\mu_n$ of $\theta^{(n)}$ is defined by
   \[
   \mu_n = \mu_{n-1} K_n = \mu_{n-2} K_{m+1} K_{m+2} = \ldots = \mu_0 K^{(0,n)}. 
   \]
where

\[ \mu_n(\mathbf{G}^n) = \int_{\Theta} \mu_{n-1}(\mathbf{G}^{n-1}) K_n(\mathbf{G}^{n-1}, \mathbf{G}^n). \]

Let \( \tilde{\Theta} \) be the interior of \( \Theta \). We assume that the following assumptions hold.

**Assumption 1:** \( f_i(\cdot | \cdot, \theta_i) \) and \( p(\theta_i) \) are twice continuously differentiable on \( \tilde{\Theta} \) for \( i = 1, \ldots, s \).

This implies that \( p(\theta | y_N) \) is twice continuously differentiable on \( \tilde{\Theta} \).

**Assumption 2:** The set of global maxima \( \Theta_{\text{MAP}} = \{ \hat{\theta}_1, \ldots, \hat{\theta}_p \} \) is finite and is contained in \( \tilde{\Theta} \). The Hessian matrix

\[ \left[ -(\partial^2 \ln p(\theta | y_N)) / \partial \theta_i \partial \theta_j \right] \]

is nonsingular at any point of \( \Theta_{\text{MAP}} \).

**Assumption 3:** There exists \( \theta^* \in \tilde{\Theta} \) such that for any \( r_N \in S^N, p(y_N | r_N, \theta^*) > 0 \).

We denote by \( \tilde{\Theta}_{\text{MAP}} \) any element of \( \Theta_{\text{MAP}} \).

**Assumption 4:**

\[- \int_{\Theta} p(\theta | y_N) \ln[p(\theta | y_N)] d\theta < +\infty. \]

**B. Study of the Transition Kernels**

First, it is easily checked that \( \{(\theta^{(n)}), r_N^{(n)} \} ; n \in \mathbb{N} \} \) is an inhomogeneous Markov chain of transition kernel on \( (\Theta \times S^N, \mathcal{B}(\Theta) \times S^N) \) at iteration \( n \)

\[ K_n[(\theta, r_N), (\theta', r'_N)] \]

\[ \overset{\Delta}{=} \mathbb{P}_T \left[ \theta^{(n)} \in \delta \theta', r_N^{(n)} = r'_N \left| \left( \theta^{(n-1)}, r_N^{(n-1)} \right) = (\theta, r_N) \right. \right] \]

\[ = \alpha_n(\theta, \theta') p(\theta | y_N, r_N) \delta(\theta') + \delta(\theta') \]

\[ \times \int_{\Theta} (1 - \alpha_n(\theta, u)) p(u | y_N, r'_N) d\theta \]

\[ p(r'_N | y_N, \theta) \] (12)

where \( \delta(\theta') \) is the delta-Dirac point-mass measure located at \( \theta' \). Thus integrating out \( r'_N \), we see that \( \{(\theta^{(n)}); n \in \mathbb{N} \} \) is also a Markov chain on \( (\Theta, \mathcal{B}(\Theta)) \) of transition kernel \( K_n(\theta, d\theta') \) at iteration \( n \)

\[ K_n(\theta, d\theta') \overset{\Delta}{=} \alpha_n(\theta, \theta') K_n(\theta, \theta') d\theta' \]

\[ + \delta(\theta') \int_{\Theta} (1 - \alpha_n(\theta, u)) K_n(\theta, u) d\theta \] (13)

where \( K_n(\theta, \theta') \) is a homogeneous transition kernel

\[ K_n(\theta, \theta') \overset{\Delta}{=} \int_{S^N} p(\theta | y_N, r'_N) p(r'_N | y_N, \theta) d\theta' \]

the integration being with respect to the counting measure.

**Lemma 1:** \( K_n(\theta, \theta') \) satisfies detailed balance for \( p(\theta | y_N) \), i.e.,

\[ p(\theta | y_N) K_n(\theta, \theta') = p(\theta | y_N) K_n(\theta', \theta). \]

**Proposition 1:** There exists \( \varepsilon > 0 \) such that \( K_n(\theta, \theta') \) satisfies, for all \( \theta \in \Theta \)

\[ K_n(\theta, \theta') \geq \varepsilon p(\theta') \]

where \( p(\theta') \) is a probability density on \( (\Theta, \mathcal{B}(\Theta)) \).

**Lemma 2:** There exists \( n_0 \in \mathbb{N}^* \) such that for any \( n \geq n_0 \),

\[ K_n(\theta, d\theta') \]

admits as invariant density

\[ \mathbb{P}_T(\theta | y_N) \]

\[ \overset{\Delta}{=} \frac{p(\theta | y_N)}{\int_{\Theta} p(\theta | y_N) d\theta}. \]

**Proposition 2:** \( K_n(\theta, d\theta') \) is a uniformly ergodic transition kernel. There exists \( \alpha \in (0, 1), \varepsilon > 0 \), and \( n_0 \in \mathbb{N}^* \) such that for any \( n \geq n_0 \) and \( \theta \in \Theta \)

\[ K_n(\theta, d\theta') \geq \varepsilon_n p(\theta') \]

where \( \varepsilon_n = \varepsilon \alpha^{1/T(n)} \) and \( p(\theta') \) is a probability density on \( (\Theta, \mathcal{B}(\Theta)) \).

**Remark 1:** In practice one can adopt a mixture of \( n_K \geq 2 \) transition kernels

\[ \tilde{K}_n(\theta, d\theta') = \sum_{i=1}^{n_K} p_i K_i(\theta, d\theta') \]

where \( 0 < p_i < 1 \) and \( \sum_{i=1}^{n_K} p_i = 1 \) including the transition kernel \( K_n(\theta, d\theta') \) studied above as a component and, for example, another Metropolis component based on a random walk proposal distribution allowing a “local” exploration of the distribution. In this case, \( \tilde{K}_n(\theta, d\theta') \) remains uniformly ergodic [16, Proposition 3, pp. 1716].

**C. Study of the Invariant Distributions**

We now study the sequence of invariant densities \( \{\mathbb{P}_T(\theta | y_N); n \in \mathbb{N}\} \). To simplify notation, the probability distribution associated with \( \mathbb{P}_T(\theta | y_N) d\theta \) will be denoted \( \mathbb{P}_T(\theta) \).

**Lemma 3:** The sequence of invariant distributions \( \mathbb{P}_T(\theta) \) converges to a probability distribution located on the set of global maxima \( \Theta_{\text{MAP}} \)

\[ \mathbb{P}_\infty(\theta) = \frac{\alpha(\tilde{\Theta}) \delta(\theta)}{\sum_{i=1}^{n} \alpha(\tilde{\Theta}_i)} \]

(19)

where

\[ \alpha(\tilde{\Theta}_i) \overset{\Delta}{=} \left[ \det \left[ -(\partial^2 \ln p(\theta | y_N)) / \partial \theta_i \partial \theta_j \right] \right]_{\theta=\tilde{\Theta}_i}^{-1/2} \]

and

\[ Z(\beta) \overset{\Delta}{=} \int_{\Theta} \left( \frac{p(\theta)}{\mathbb{P}_T(\theta_{\text{MAP}})} \right)^{\beta} d\theta = (2\pi)^{p/2} \left[ \sum_{i=1}^{n} \alpha(\tilde{\Theta}_i) + \varepsilon(\beta) \right] \]

(20)

where \( \lim_{\beta \to +\infty} \varepsilon(\beta) = 0 \).

**Proposition 3** ([19, Theorem 3.2, pp. 871–872]): There exists \( n_0 \in \mathbb{N}^* \) such that for any \( n > n_0 \)

\[ \sum_{i=1}^{n} \left\| \mathbb{P}_T(i+1) - \mathbb{P}_T(i) \right\| \leq 2 \ln \left( \frac{Z(1/T)}{Z(1/T(n))} \right). \]
D. Convergence of the Simulated Annealing Algorithm

In this subsection, we prove that \(|\mu_n - \overline{p}^{1/T(n)}|\) goes to zero as \(n\) increases for suitable logarithmic cooling schedules.

Proposition 4: For any positive integers \(m, n\) such that \(\frac{n}{m} < \frac{1}{\epsilon}\), we have

\[
\left\| \mu_n - \overline{p}^{1/T(n)} \right\| \leq \prod_{i=m+1}^{n-m+1} (1 - \epsilon^2) + \sum_{i=m}^{n-1} \left\| \overline{p}^{1/T(i+1)} - \overline{p}^{1/T(i)} \right\| \tag{22}
\]

Lemma 4: For any \(\epsilon \in (0, 1), \gamma > 0\) and \(m_n = [n - n^{1-\epsilon^2}]\), if

\[
T(n) = (-1 + \epsilon) \ln(\alpha)/\ln(n + \gamma)
\]

(where \(\alpha \in (0, 1)\) is defined in Proposition 2) then

\[
\lim_{n \to \infty} \prod_{i=m_n}^{n} (1 - \epsilon^2) = 0. \tag{23}
\]

Lemma 5: For any \(\epsilon \in (0, 1), \gamma > 0\), and \(m_n = [n - n^{1-\epsilon^2}]\), if

\[
T(n) = (-1 + \epsilon) \ln(\alpha)/\ln(n + \gamma)
\]

then

\[
\lim_{n \to \infty} \sum_{i=m_n}^{n-1} \left\| \overline{p}^{1/T(i+1)} - \overline{p}^{1/T(i)} \right\| = 0 \tag{24}
\]

Theorem 1: For any \(\epsilon \in (0, 1)\) and \(\gamma > 0\), if

\[
T(n) = (-1 + \epsilon) \ln(\alpha)/\ln(n + \gamma)
\]

then, for any initial distribution \(\mu_0\) of \(\theta^{(0)}\), the Markov chain generated by the simulated annealing converges in the following sense:

\[
\lim_{n \to \infty} \left\| \mu_n - \overline{p}^{1/T(n)} \right\| = 0. \tag{25}
\]

Consequently, for any \(\xi \in (0, 1)\)

\[
\lim_{n \to \infty} \Pr \left( \frac{p(\theta^{(n)}) y_N}{p(\hat{\theta}_{\text{MAP}}) y_N} > 1 - \xi \right) = 1. \tag{26}
\]

Remark 2: This proof of convergence can be extended to some simulated annealing algorithms derived from (homogeneous) Markov chain Monte Carlo algorithms which are uniformly ergodic (see, for example, [16, pp. 1714–1716]).

Remark 3: As outlined in [19, pp. 873], if the set \(\hat{\Theta}_{\text{MAP}}\) is of null Lebesgue measure (which is a realistic assumption, clearly true under Assumption 2) we cannot obtain

\[
\lim_{n \to \infty} \left\| \mu_n - \overline{p}^{1/n} \right\| = 0.
\]

Note that in a discrete setup, the set of maxima is of nonnull measure with respect to the reference counting measure. This allows one to obtain such a convergence result under appropriate assumptions [17], [18], [26].

V. APPLICATIONS

A logarithmic cooling schedule is required to ensure theoretical convergence of the simulated annealing algorithm. In practice, it is well known that such cooling schedules decrease too slowly toward zero. So, as is usually done, we implement a geometric cooling schedule \(T(n) = C \gamma^n\) [18], [20]. In the following examples, we run our algorithm for 250 iterations with \(C = 1\) and \(\gamma = 0.085\) so that \(T(250) \approx 0.023\). For each example, we simulated 500 realizations of the signal and then ran 250 iterations of the SA (simulated annealing) algorithm and 250 iterations of the EM algorithm as these algorithms have an equivalent computational complexity. These algorithms have been initialized randomly with the same values of the parameters. We denote by \(\hat{\theta}_{\text{SA}}^j\) and \(\hat{\theta}_{\text{EM}}^j\), respectively, the estimates obtained using the SA algorithm and the EM algorithm for the \(j\)th realization. For the EM algorithm, the posterior probability increases at each iteration, so \(\hat{\theta}_{\text{EM}}^j\) is the value obtained at the final iteration. For the SA algorithm, this property is no longer valid and we select \(\hat{\theta}_{\text{SA}}^j\) as the estimate which maximizes the penalized log-likelihood among the 250 iterations.

A. Switching Autoregressions with a Markov Regime

We consider the problem of the MAP estimation of the parameters of a switching autoregression process with a Markov regime [5], [28], [29]. We assume that the observations are generated from the following system equation:

\[
y(k) = y(k)^t a_{y(k)} + \sigma_{y(k)} v_k, \quad \text{for} \quad k = 1, \cdots, N \tag{27}
\]

where

\[
a_{y(k)} \doteq (a, r_{y(k)}, \cdots, a, r_{y(k)}; \cdots)^t
\]

and

\[
y(k) \doteq (y_k, \cdots, y_k, \cdots)^t.
\]

\(v_k\) is Gaussian white noise of variance 1. We define

\[
\xi_i \doteq \{k; r_k = i, k = 1, \cdots, N\}
\]

and \(n_i \doteq \text{card}(\xi_i)\).

In this case, we want to estimate \(\theta = \{a_i, \sigma_{y(i)}^2, \cdots, a_i, \sigma_{y(i)}^2; i \in S\}\). We assume that

\[
p(a_1, \cdots, a_s, \sigma_{y(i)}^2, \cdots, \sigma_{y(i)}^2) = \prod_{i=1}^{s} p(a_i) p(\sigma_{y(i)}^2) . \tag{28}
\]

The AR coefficients \(a_i\) are assumed to be uniformly distributed on their domain of stability which is a reasonable assumption in numerous physical applications, i.e., \(a_i \sim \mathcal{U}(A_p)\), where \(A_p\) is the set of order-\(p\) stationary AR processes. Note, however,
that it does not imply that $y_k$ is stationary [28]. There is an identifiability problem as the posterior distribution is not modified up to a permutation of the labels of the hidden Markov chain. Hence we introduce an additional ordering constraint on the $s$ autoregressive coefficients. To simplify notation, we do not enforce this constraint later on. For the variances of dynamic noises, we set conjugate inverse Gaussian prior distributions $\sigma_i^2 \sim IG\left((\nu_0,i)/2, (\gamma_0,i)/2\right)$ with $\nu_0,i, \gamma_0,i > 0$ for all $i = 1, \ldots, s$ to ensure integrability of the posterior distribution. Note that these priors tend to the improper uninformative Jeffreys’ prior as $\nu_0,i, \gamma_0,i \to 0$.

To apply the algorithm described in Section III, one needs to be able to sample from

$$p\left(\theta_i, \sigma_i^2 \mid y_N, r_N\right) = p\left(\sigma_i^2 \mid y_N, r_N, \theta_i\right) p(\theta_i \mid y_N, r_N).$$

(29)

There are several cases to distinguish:

- If $n_i = 0$

$$p\left(\theta_i, \sigma_i^2 \mid y_N, r_N\right) = p\left(\sigma_i^2 \mid y_N, r_N\right).$$

(30)

Sampling from $p(\sigma_i^2)$ is straightforward [23]. To sample from $p(\theta_i)$, we use the efficient procedure described in [30].

- If $n_i \geq 1$, we define

$$\sum_i \tilde{\gamma} = \sum_{k \in \tilde{E}_i} (\gamma(k) \theta_i^t) \quad \tilde{m}_i = \sum_{k \in \tilde{E}_i} \gamma(k) y_k.$$

Then

$$p(\theta_i \mid y_N, r_N) \propto \left[\sum_{k \in \tilde{E}_i} (y_k - \tilde{\theta}_i \tilde{m}_i)^2 + \gamma_0,i\right] \mathcal{U}_{\tilde{A}_p}(\theta_i).$$

(31)

Two cases must be considered:

- If $n_i + \nu_0,i - p > 0$, then $p(\theta_i \mid y_N, r_N)$ is a Student-$t$ distribution restricted to the set $\tilde{A}_p$

$$\theta_i \mid (y_N, r_N) \sim T_p\left(n_i + \nu_0,i - p, \tilde{m}_i, \left(\gamma_0,i + \sum_{k \in \tilde{E}_i} y_k^2 - \tilde{m}_i \tilde{m}_i^t / n_i + \nu_0,i - p\right)\right) \mathcal{U}_{\tilde{A}_p}(\theta_i).$$

(32)

We sample from this distribution by first sampling from a Student-$t$ distribution and then performing a rejection step [23].

- If $n_i + \nu_0,i - p < 0$, we sample from $p(\theta_i \mid y_N, r_N)$ by first sampling from $\mathcal{U}_{\tilde{A}_p}(\theta_i)$ and then performing a rejection step. $p(\sigma_i^2 \mid y_N, r_N, \theta_i)$ is an inverse Gamma distribution from which one can easily sample [23]

$$\sigma_i^2 \mid (y_N, r_N, \theta_i) \sim IG\left(\frac{\nu_0,i + n_i}{2}, \frac{\sum_{k \in \tilde{E}_i} (y_k - \tilde{\theta}_i \tilde{m}_i)^2}{2}\right).$$

(34)

For the numerical simulations, the following example is taken from [28, pp. 501]. We have simulated 500 realizations $y_j^i, j = 1, \ldots, 500$ of $Y = 500$ observations of a two-state $(s = 2)$ hidden Markov chain whose parameters are specified in Tables I and II. The prior distribution we chose is defined by $\nu_0,i = 1, \gamma_0,i = 0, \alpha_{i,j} = 1$ for $i, j \in S$. Assumptions 1 and 3 are clearly true. We assume that Assumption 2 is valid. One can check that Assumption 4 is satisfied for this prior distribution.

In simulations, we obtained

$$\frac{1}{500} \sum_{j=1}^{500} \left(\ln p\left(\tilde{\theta}_i \mid y_j^i\right) - \ln p\left(\hat{\theta}_EM \mid y_j^i\right)\right) = 0.24$$

which means that the simulated annealing performs better in finding the global mode of the penalized likelihood. In Tables I and II, the empirical means, biases, and standard deviations of the estimated parameters defined as

$$\hat{\theta}_{SA} \sim \frac{1}{500} \sum_{j=1}^{500} \sigma_j \quad \text{bias} \left[\sigma_i^2\right] = \sigma_i^2 - \hat{\theta}_{EM} \left[\sigma_i^2\right],$$

and similarly for the AR coefficients and the probability transitions. This estimation problem appears easy, both in statistical and optimization terms, as the estimates obtained using the SA and the EM algorithms are similar and really good: their bias and variance with respect to the true value are very small. For such a simple estimation problem, there is no significant advantage in using a stochastic algorithm. We present in the next subsection a more complex problem.

### B. Markov-Modulated Poisson Processes

We consider here the problem of MAP parameter estimation for Markov-modulated Poisson processes [5], [27], [1], [8]. Let us assume that we observe counting data. These counting data are modeled as a Poisson process with a parameter switching according to a hidden Markov chain of unknown statistics. More formally

$$y_j^i \mid (y_{j-1}^i, r_j) \sim P(\lambda_j | \tau_i).$$

(35)

We want to estimate $\theta \equiv \{\lambda_i, \pi_i, \pi_{ij} | i, j \in S\}$. To complete the Bayesian model, we assume that the parameters $\lambda_i, i \in S,$ are
a priori independent, and have conjugate Gamma priors \( \lambda_i \sim G(\alpha_i, \beta_i) \) where \( \alpha_i > 0 \) and \( \beta_i > 0 \). There is an identifiability problem as the posterior distribution is not modified up to a permutation of the labels \( \tau_k \) of the hidden Markov chain. Hence we include an additional constraint 0 \( \leq \lambda_1 < \cdots < \lambda_S \) so that \[
p(\lambda_1, \cdots, \lambda_S) \propto \prod_{i=1}^{S} G(\lambda_i; a_i, b_i) \prod_{0 < \lambda_k < \cdots < \lambda_S} (\lambda_1, \cdots, \lambda_S).
\]

(36)

To implement the simulated annealing algorithm described in Section III, it is necessary to sample from \( p(\lambda_1, \cdots, \lambda_S | y, r_N) \) where \[
p(\lambda_1, \cdots, \lambda_S | y, r_N)
\propto \prod_{i=1}^{S} G \left( \lambda_i; a_i + \sum_{k=1}^{N} y_{k,i} \delta_{\tau_k,i}, b_i + \sum_{k=1}^{N} \delta_{\tau_k,i} \right)
\times \prod_{0 < \lambda_k < \cdots < \lambda_S} (\lambda_1, \cdots, \lambda_S).
\]

(37)

It is straightforward to sample from this distribution [23].

For the numerical simulations, the following example is taken from [8]. We have simulated 500 realizations \( y_{N-j}, j = 1, \cdots, 500 \) of \( N = 500 \) observations of a four-state \((S = 4)\) hidden Markov chain whose parameters are specified in Tables III and IV. The prior distribution we chose is quite uninformative as we set \( \alpha_i = 1, \beta_i = 10^{-3}, \alpha_{ij} = 1 \) for \( i, j \in S \). This example is known to be difficult from a statistical point of view [8]. Assumptions 1 and 3 are clearly true. Assumption 2 is supposed valid. One can check that Assumption 4 is satisfied for this prior distribution. In simulations, the following results have been obtained. For the log-posterior distribution, we obtained \[
\frac{1}{500} \sum_{j=1}^{500} \left[ \ln p \left( \hat{\theta}_{SA}^j \mid y_N \right) - \ln p \left( \hat{\theta}_{EM}^j \mid y_N \right) \right] = 2.06.
\]

It means again that the SA algorithm outperforms EM in terms of maximization. Table IV presents the empirical means, biases, and standard deviations of the estimated parameters \( \lambda_i, i \in S \) defined as

\[
\hat{\lambda}_i = \frac{1}{500} \sum_{j=1}^{500} \lambda_i^j, \quad \text{bias} [\lambda_i] = \lambda_i - \hat{\lambda}_i, \quad \text{var} [\lambda_i] = \frac{1}{500} \sum_{j=1}^{500} \left( \lambda_i^j - \lambda_i \right)^2.
\]

VI. Conclusion

In this paper, we have proposed an original simulated annealing algorithm to obtain MAP estimates of the parameters of HMM. This algorithm relies on stochastic simulation of the hidden Markov chain. Under suitable regularity assumptions, the simulation of this finite missing data set gives uniform ergodicity of the inhomogeneous transition kernel. This property is used to obtain sufficient conditions that ensure convergence of the algorithm toward the set of MAP parameters whereas the EM is only ensured to converge toward a stationary point of the posterior distribution. This algorithm has been applied to

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \lambda_3 )</th>
<th>( \lambda_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>True values</td>
<td>0.50</td>
<td>2.50</td>
<td>5.00</td>
<td>8.00</td>
</tr>
<tr>
<td>( \hat{\lambda}_{SA} \mid \text{bias} )</td>
<td>0.60/ -0.10</td>
<td>2.60/ -0.10</td>
<td>4.47/0.03</td>
<td>7.64/0.36</td>
</tr>
<tr>
<td>( \hat{\lambda}_{EM} \mid \text{bias} )</td>
<td>0.68/ -0.18</td>
<td>2.74/ -0.24</td>
<td>4.58/0.42</td>
<td>7.58/0.42</td>
</tr>
<tr>
<td>( \text{var} \hat{\lambda}_{SA} )</td>
<td>0.32</td>
<td>0.64</td>
<td>1.04</td>
<td>1.20</td>
</tr>
<tr>
<td>( \text{var} \hat{\lambda}_{EM} )</td>
<td>0.36</td>
<td>0.83</td>
<td>1.26</td>
<td>1.22</td>
</tr>
</tbody>
</table>

For this difficult statistical estimation problem, the estimates obtained using SA are significantly closer to the true values than those obtained using the EM algorithm and the variance is reduced.

TABLE II

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( a_{1,1} )</th>
<th>( a_{1,2} )</th>
<th>( a_{2,1} )</th>
<th>( a_{2,2} )</th>
<th>( \sigma_1^2 )</th>
<th>( \sigma_2^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>True values</td>
<td>1.50</td>
<td>-0.70</td>
<td>1.70</td>
<td>-0.72</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( \hat{\lambda}_{SA} \mid \text{bias} )</td>
<td>1.53/ -0.03</td>
<td>-0.71/0.01</td>
<td>1.68/0.02</td>
<td>-0.67/ -0.05</td>
<td>1.03/ -0.03</td>
<td>0.98/0.02</td>
</tr>
<tr>
<td>( \hat{\lambda}_{EM} \mid \text{bias} )</td>
<td>1.54/ -0.04</td>
<td>-0.69/ -0.01</td>
<td>1.69/0.01</td>
<td>-0.71/ -0.01</td>
<td>1.04/ -0.04</td>
<td>1.01/ -0.01</td>
</tr>
<tr>
<td>( \sqrt{\text{var} \hat{\lambda}_{SA} \mid \text{bias}} )</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
<td>0.07</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>( \sqrt{\text{var} \hat{\lambda}_{EM} \mid \text{bias}} )</td>
<td>0.07</td>
<td>0.04</td>
<td>0.05</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
</tr>
</tbody>
</table>

TABLE III

<table>
<thead>
<tr>
<th>Transition Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
</tr>
<tr>
<td>( j )</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

TABLE IV

<table>
<thead>
<tr>
<th>Poisson Process Intensities: True Values and Simulation Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>True values</td>
</tr>
<tr>
<td>( \hat{\lambda}_{SA} \mid \text{bias} )</td>
</tr>
<tr>
<td>( \hat{\lambda}_{EM} \mid \text{bias} )</td>
</tr>
<tr>
<td>( \text{var} \hat{\lambda}_{SA} )</td>
</tr>
<tr>
<td>( \text{var} \hat{\lambda}_{EM} )</td>
</tr>
</tbody>
</table>
maximum a posteriori parameter estimation for switching autoregressions with a Markov regime and for Markov-modulated Poisson processes. It might be a good alternative to the EM algorithm for complex statistical problems as it has an equivalent computational complexity at each iteration and appears in practice to be less sensitive to initialization.

APPENDIX A

See Notation at the top of this page.

APPENDIX B

FORWARD FILTERING–BACKWARD SAMPLING RECURSION

To sample from \( p(\mathbf{r}_N \mid \mathbf{y}_N, \Theta) \), note that \( p(\mathbf{r}_N \mid \mathbf{y}_N, \Theta) \) can be decomposed as follows:

\[
p(\mathbf{r}_N \mid \mathbf{y}_N, \Theta) = p(\mathbf{r}_N \mid \mathbf{y}_N, \Theta) \prod_{k=1}^{N-1} p(\mathbf{r}_k \mid \mathbf{y}_{N,k}, \Theta, \mathbf{r}_{k+1})
\]

where \( \mathbf{y}_k \equiv (y_{1k}, \cdots, y_{kk}) \). Given \( r_{k+1} \), \( p(\mathbf{r}_k \mid \mathbf{y}_{k}, \Theta, r_{k+1}) \) is a discrete distribution. This suggests the following algorithm to sample from \( p(\mathbf{r}_N \mid \mathbf{y}_N, \Theta) \) [5], [22].

1) Optimal filter (forward filtering): For \( k = 2, \cdots, N \), compute the optimal filter [31]. For any \( k \), evaluate

\[
p(r_k = i \mid \mathbf{y}_{k-1}, \Theta) = \sum_{j=1}^{s} \pi_{ij} p(r_{k-1} = j \mid \mathbf{y}_{k-1}, \Theta)
\]

and store for \( k = 1, \cdots, N \) and for \( i = 1, \cdots, s \),

\[
p(r_k = i \mid \mathbf{y}_{k-1}, \Theta), \text{ and } p(r_k = i \mid \mathbf{y}_k, \Theta).
\]
2) Backward sampling: Sample \( r_N \) from \( p(r_N | y_N, \theta) \) then for \( k = N - 1, \ldots, 1 \), sample \( r_k \) from \( p(r_k | y_{k+1}, \theta, r_{k+1}) \) where if \( r_{k+1} = l \) we have for \( i = 1, \ldots, s \)
\[
p(r_k = i | y_k, \theta, r_{k+1}) = \frac{\pi_{jl} I(r_k = i | y_{k+1}, \theta)}{\sum_{j=1}^{s} \pi_{jl} I(r_k = j | y_k, \theta)}.
\]

\[ (41) \]

**APPENDIX C**

**EVALUATION OF THE LIKELIHOOD**

The likelihood \( p(y_N) \) can be decomposed as follows:

\[
p(y_N | \theta) = p(y_N | \theta) \prod_{k=2}^{N} p(y_k | y_{k-1}, \theta)
\]

where

\[
p(y_1 | \theta) = \sum_{i=1}^{s} p(y_1, r_1 = i | \theta)
\]

\[ (42) \]

\[
p(y_k | y_{k-1}, \theta) = \sum_{i=1}^{s} f_i(y_k | y_{k-1}, \theta) p(r_k = i | y_{k-1}, \theta).
\]

\[ (43) \]

Equation (44) is evaluated when computing the forward filtering–backward sampling recursion in Appendix B. This is the denominator of (40). This evaluation of the likelihood was first proposed by Devijer [32] who suggests that it is numerically more robust than Baum's forward–backward procedure.

**APPENDIX D**

**PROOFS**

**Proof of Proposition 1:** The transition kernel \( K_N(\theta, \theta') \) satisfies

\[
K_N(\theta, \theta') = \int_{S^N} I(r_N' | y_N, \theta)p(\theta' | y_N, r_N') dr_N'
\]

\[ \geq \int_{S^N} I(r_N' | y_N, \theta) \min_{r_N} I(\theta' | y_N, r_N') dr_N'
\]

\[ \geq \mu(\theta')
\]

\[ (45) \]

where

\[
\mu(\theta') \triangleq \min_{r_N} I(\theta' | y_N, r_N').
\]

\( \mu(\theta') \) is integrable because, for any \( r_N' \)
\[
\int_{\Theta} \mu(\theta') d\theta' \leq \int_{\Theta} p(\theta' | y_N, r_N') d\theta' = 1.
\]

From Assumption 3, there exists \( \theta^* \in \Theta \) such that
\[
\min_{r_N} I(\theta^* | y_N, r_N') > 0.
\]

Indeed, as for any \( r_N, p(\theta^* | r_N') > 0 \) on \( \Theta \), then
\[
p(\theta^* | y_N, r_N') > 0
\]

for any \( r_N' \). From Assumption 1, we have that \( \mu(\theta) \) is continuous at \( \theta^* \). Since \( S^N \) is finite, then \( \mu(\theta') \) is continuous at \( \theta^* \) and

\[
\varepsilon = \int_{\Theta} \mu(\theta') d\theta' > 0.
\]

Thus \( \mu(\theta') / \varepsilon \) is a probability density on \( (\Theta, B(\Theta)) \) and (16) follows.

**Proof of Lemma 2:** Under Assumption 2, there exists \( \tilde{\theta}_{\text{MAP}} \in \Theta \) and since the “temperature” \( T(n) \) decreases toward 0, there exists \( n_0 \in \mathbb{N}^+ \) such that for any \( n \geq n_0 \)
\[
0 < \int_{\Theta} p_{1/T(n)}(\theta) p(\theta | y_N) d\theta' \leq \int_{\Theta} p(\theta | y_N) d\theta' < +\infty
\]

so that \( p_{1/T(n)}(\theta) p(\theta | y_N) \) is well defined. Then we note that \( K_N(\theta, \theta') \) is nothing but a Metropolis–Hastings algorithm with proposal distribution \( K_N(\theta, \theta') \) and invariant density \( p_{1/T(n)}(\theta) p(\theta | y_N) \) [15], [16]. Indeed, the acceptance probability of such a Metropolis–Hastings algorithm satisfies

\[
\alpha_n(\theta, \theta') = \min \left\{ \frac{p_{1/T(n)}(\theta) p(\theta | y_N) K_N(\theta, \theta')} {p_{1/T(n)}(\theta') p(\theta' | y_N) K_N(\theta, \theta')}, 1 \right\}
\]

\[ = \min \left\{ \frac{p(\theta | y_N)} {p(\theta' | y_N)}, 1 \right\}
\]

because of Lemma 1. This is expression (6).

**Proof of Proposition 2:**

\[
K_N(\theta, \theta') \geq \alpha_n(\theta, \theta') K_N(\theta, \theta') d\theta'
\]

\[ \geq \varepsilon \alpha_n(\theta, \theta') \mu(\theta') d\theta'
\]

from Proposition 1. \( T(n) \) decreasingly converging toward 0, there exists \( n_0 \in \mathbb{N}^+ \) such that for any \( n \geq n_0 \)
\[
\alpha_n(\theta, \theta') \geq \left[ \frac{p(\theta | y_N)} {p(\theta' | y_N)} \right]^{1/T(n)-1}.
\]

Then

\[
K_N(\theta, \theta') \geq \varepsilon \alpha^{1/T(n)-1} \ll_{C_n}(\theta) p(\theta') d\theta'
\]

where

\[
C_n \triangleq \left\{ \theta' \in \Theta : \frac{p(\theta | y_N)} {p(\theta' | y_N)} > \alpha \right\}.
\]

From Assumptions 1 and 3, there exists \( \theta^* \in \Theta \) such that \( \mu(\theta^*) > 0 \) and \( p(\theta) \) is continuous in \( \theta^* \). From Assumption 2, the number of global maxima is finite, thus there exists \( \alpha \in (0, 1) \) such that

\[
\varepsilon' \triangleq \int_{\Theta} \ll_{C_n}(\theta) p(\theta') d\theta' = \int_{C_n} \mu(\theta') d\theta' > 0.
\]

Denoting \( \tilde{\mu}(\theta') \triangleq \mu(\theta') \ll_{C_n}(\theta') / \varepsilon' \), we have

\[
K_N(\theta, \theta') \geq \varepsilon' \alpha^{1/T(n)-1} \tilde{\mu}(\theta') d\theta'
\]

where \( \tilde{\mu}(\theta') \) is a probability density on \( (\Theta, B(\Theta)) \). From [16, Theorem 1, pp. 1758], this implies that \( K_N(\theta, \theta') \) is \( p_{1/T(n)}(\theta) p(\theta) \)-irreducible as it is \( \tilde{\mu}(\theta') \)-irreducible, as a result of Lemma 2. Furthermore, \( K_N(\theta, \theta') \) is clearly strongly aperiodic [24, pp. 118]. Thus the Markov transition kernel
Proof of Lemma 3: This is an application of Laplace’s formula, see for example [25]. The result follows under Assumptions 1 and 2.

Proof of Proposition 3: The proof is similar to the one of Haario et al. [19]. In applying the Lebesgue-dominated convergence theorem, which allows for derivation under the integral sign, Haario et al. assume that $-\ln[p(\theta) \mid Y_N]/P(\theta)_{\text{MAP}} \mid Y_N$ is bounded above on $\Theta$. Here, instead, we use Assumption 4. First note that for $i$ sufficiently large and all $\theta \in \Theta$

$$0 \geq \ln \left( \frac{-n \theta(\theta) \mid Y_N}{P(\theta)_{\text{MAP}} \mid Y_N} \right) \exp \frac{1}{T(n)} \ln \left( \frac{-n \theta(\theta) \mid Y_N}{P(\theta)_{\text{MAP}} \mid Y_N} \right)$$

which under Assumption 4 and using the dominated convergence theorem allows us to write

$$\int \frac{1}{T(n)} \ln \left( \frac{-n \theta(\theta) \mid Y_N}{P(\theta)_{\text{MAP}} \mid Y_N} \right) \, d\theta.$$

This is the starting point of the proof of [19, Theorem 3.2.], which is then the same.

Proof of Proposition 4: We have

$$\begin{align*}
\| \mu_n - \beta I(T) \| & = \left\| \mu_n K^{(m, n)} - \beta I(T) \right\| \\
& \leq \left\| \mu_n K^{(m, n)} - \beta I(T) \right\| \\
& \quad + \left\| \beta I(T) K^{(m, n)} - \beta I(T) \right\|.
\end{align*}$$

Using the straightforward generalization to inhomogeneous Markov chains of [24, Theorem 16.2.4, pp. 392–393], we obtain from Proposition 2

$$\left\| \mu_n - \beta I(T) \right\| \leq \prod_{i=m+1}^{n} (1 - \epsilon_i)$$

and one can verify that

$$\beta I(T) K^{(m, n)} - \beta I(T) = \sum_{i=m}^{n-1} \left( \beta I(T) - \beta I(T+1) \right) K^{(i,n)}$$

Thus

$$\left\| \beta I(T) K^{(m, n)} - \beta I(T) \right\| \leq \sum_{i=m}^{n-1} \left\| \left( \beta I(T) - \beta I(T+1) \right) K^{(i,n)} \right\| \leq \sum_{i=m}^{n-1} \left\| \beta I(T) - \beta I(T+1) \right\|.$$

Proof of Lemma 4: Noting $\alpha = (1/1 + \epsilon)$

$$\sum_{i=m+1}^{n} \epsilon_i = \sum_{i=m+1}^{n} \alpha^{i/(T+1)} \sum_{i=m+1}^{n} (i + \gamma)^{-\alpha} \geq \epsilon \int_{m+1}^{n} (x + \gamma)^{\alpha} \, dx$$

$$= \frac{\epsilon}{\alpha} (n + \gamma)^{\alpha} - (n - n^{1-\alpha} + n + \gamma)^{\alpha} \geq \frac{\epsilon}{\alpha} (n + \gamma)^{\alpha} \left( 1 - \left( 1 - \alpha n^{1-\alpha} \right) \right) \quad \text{as} \quad n \to +\infty.$$

Proof of Lemma 5: From Proposition 3, we have that

$$\sum_{i=m+1}^{n} \left\| \beta I(T) - \beta I(T+1) \right\| \leq \ln \left( \frac{Z(1/T(m_n-1))}{Z(1/T(n-1))} \right)$$

and from Lemma 3

$$\lim_{n \to +\infty} \ln \left( \frac{Z(1/T(m_n-1))}{Z(1/T(n-1))} \right) = \lim_{n \to +\infty} \frac{n-1}{2} \ln \left( \frac{\ln(m_n - 1 + \gamma)}{\ln(n - 1 + \gamma)} \right) = 0.$$

Proof of Theorem 1: From Proposition 4, we have for any sequence $m_n < n$,

$$\left\| \mu_n - \beta I(T) \right\| \leq \prod_{i=m+1}^{n} (1 - \epsilon_i)$$

Choosing $m_n = \lfloor n - n^{1-\alpha} \rfloor$, then the theorem is proved using Lemmas 4 and 5. Equation (26) is straightforwardly obtained from [19, Corollary 5.4, pp. 880].

ACKNOWLEDGMENT

The authors would like to acknowledge R. A. Cook, L. Davis, S. J. Godsill, and D. Leporini for various helpful comments on a preliminary version of this work.

REFERENCES


