# Stochastic Sampling Algorithms for State Estimation of Jump Markov Linear Systems

Arnaud Doucet, Andrew Logothetis, and Vikram Krishnamurthy

***Abstract*—Jump Markov linear systems are linear systems whose parameters evolve with time according to a finite-state Markov chain. Given a set of observations, our aim is to estimate the states of the finite-state Markov chain and the continuous (in space) states of the linear system. The computational cost in computing conditional mean or maximum *a posteriori* (MAP) state estimates of the Markov chain or the state of the jump Markov linear system grows exponentially in the number of observations.**

**In this paper, we present three globally convergent algorithms based on stochastic sampling methods for state estimation of jump Markov linear systems. The cost per iteration is linear in the data length. The first proposed algorithm is a data augmentation (DA) scheme that yields conditional mean state estimates. The second proposed scheme is a stochastic annealing (SA) version of DA that computes the joint MAP sequence estimate of the finite and continuous states. Finally, a Metropolis–Hastings DA scheme based on SA is designed to yield the MAP estimate of the finite-state Markov chain is proposed. Convergence results of the three above-mentioned stochastic algorithms are obtained.**

**Computer simulations are carried out to evaluate the performances of the proposed algorithms. The problem of estimating a sparse signal developing from a neutron sensor based on a set of noisy data from a neutron sensor and the problem of narrowband interference suppression in spread spectrum code-division multiple-access (CDMA) systems are considered.**

## I. Introduction

A DISCRETE time jump Markov linear system can be modeled as

$$x_{t+1} = A(r_{t+1})x_t + B(r_{t+1})v_{t+1} + F(r_{t+1})u_{t+1}, \quad (1)$$

$$y_t = C(r_t)x_t + D(r_t)w_t + G(r_t)u_t, \quad (2)$$

where

| | |
|---|---|
| $t = 1, 2, \cdots$ | discrete time; |
| $r_t$ | unknown realization of a finite-state Markov chain with states in $\{1, 2, \cdots, s\}$; |
| $x_t$ | unknown state of the jump linear system; |
| $v_t$ and $w_t$ | uncorrelated Gaussian white noise sequences; |
| $y_t$ | observation at time $t$; |
| $u_t$ | a known exogenous input sequence; |
| $A(r_t)$, $B(r_t)$, $C(r_t)$, $D(r_t)$, $F(r_t)$, and $G(r_t)$ | time-varying matrices that evolve according to the realization of the finite-state Markov chain $r_t$. |

For notational convenience, let $\boldsymbol{y} \triangleq (y_1, \cdots, y_T)$, $\boldsymbol{x} \triangleq (x_0, \cdots, x_T)$, and $\boldsymbol{r} \triangleq (r_1, \cdots, r_T)$ denote the sequence of measurements, the states of the jump Markov linear system, and the states of the finite-state Markov chain, respectively.

Jump Markov linear systems appear in several fields in electrical engineering (see [12] and references therein), including control (e.g., hybrid systems, target tracking), signal processing (e.g., blind channel equalization), communications (e.g., interference suppression in mobile telephony), and other areas, such as econometrics and biometrics. In these fields, given the $T$ observations $y_1, \cdots, y_T$ generated by the signal model (1), (2), and assuming the parameters $(A(1), \cdots, A(s), B(1), \cdots, B(s), C(1), \cdots, C(s), D(1), \cdots, D(s),$ and $F(1), \cdots, F(s), G(1), \cdots, G(s))$ are known, it is of interest to compute the following state estimates:

- conditional mean state estimates of $r_t$ and $x_t$, namely, $\mathbb{E}\{r_t|\boldsymbol{y}\}$ and $\mathbb{E}\{x_t|\boldsymbol{y}\}$, for $t = 1, 2, \cdots, T$;
- maximum *a posteriori* (MAP) sequence estimates $r_t$ and $x_t$ defined as $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \arg\max_{\boldsymbol{x},\boldsymbol{r}} f(\boldsymbol{x}, \boldsymbol{r}|\boldsymbol{y})$ and $\hat{\boldsymbol{r}} = \arg\max_{\boldsymbol{r}} f(\boldsymbol{r}|\boldsymbol{y})$, where $f(\cdot|\cdot)$ denotes the conditional probability density (or mass) function.

Unfortunately, it is well known that exact computation of these estimates involves a prohibitive computational cost of order $s^T$, where $T$ denotes the number of measurements and $s^T$ corresponds to all possible realizations of the finite Markov chain [25]. Thus, it is necessary to consider in practice suboptimal estimation algorithms. A variety of such suboptimal algorithms have been proposed; see, for example, [7], [24], and [25]. In particular, [25] presents a truncated (approximate) maximum likelihood procedure for parameter estimation and a truncated approximation of the conditional mean state estimates. The estimates are computed using a bank of Kalman filters.

This paper presents three stochastic iterative algorithms for computing conditional mean state estimates and MAP estimates of the Markov state $r_t$ and the state of the jump Markov linear system $x_t$ in (1) and (2). These algorithms are based on the data augmentation (DA) algorithm (proposed by Tanner and Wong [21], [22]) and two originally proposed hybrid DA/stochastic

Manuscript received December 1, 1997; revised July 30, 1998 and December 28, 1998. Recommended by Associate Editor, J. C. Spall.

A. Doucet is with the Signal Processing Group, Department of Engineering, University of Cambridge, CB2 1PZ Cambridge, U.K. (e-mail: ad2@eng.cam.ac.uk).

A. Logothetis is with Automatic Control, Department of Signals, Sensors and Systems, Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: andrew.logothetis@s3.kth.se).

V. Krishnamurthy is with the CRC for Sensor Signal and Information Processing, Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3052, Australia (e-mail: v.krishnamurthy@ee.mu.oz.au).

Publisher Item Identifier S 0018-9286(00)00396-2.

annealing (SA) algorithms. The algorithms have a computational cost of $O(T)$ per iteration.

Although the DA algorithm proposed in [21] and [22] is a well-known algorithm in the statistical literature, it is rarely mentioned or applied in the engineering literature. The DA algorithm is one of the simplest Markov chain Monte Carlo (MCMC) algorithms [21], [22]. MCMC are powerful stochastic algorithms used to sample from complex multivariate probability distributions. These methods are well known in image processing because they have been introduced by Geman and Geman in 1984 [6] to simulate from the Gibbs distribution of a Markov random field. Their introduction in the early 1990's has revolutionized the field of applied statistics. The key idea behind MCMC methodology consists of sampling from a "target" complex probability distribution by simulating a Markov chain that admits as its invariant (or stationary) distribution, the "target" distribution. In this paper, we show how the DA algorithm can be used to estimate conditional mean estimates $\mathbb{E}[x_t|\mathbf{y}_1, \cdots, y_T]$ and $\mathbb{E}[r_t|\mathbf{y}_1, \cdots, y_T]$ for jump Markov linear systems. Then, we propose two new hybrid SA/DA algorithms that yield MAP sequence estimates.

In recent work [12], we have used the expectation maximization algorithm to iteratively compute MAP sequence estimates for jump Markov linear systems. Although these EM algorithms in some cases perform remarkably well, convergence to a local stationary point (maximum, minimum, or saddle point) is a major drawback. A significant advantage of the SA methods proposed in this paper is that they are asymptotically globally convergent.

*Main Results:* We now list the main results and organization of this paper:

- Section II formally presents the signal model and estimation objectives.
- In Section III, we use the DA algorithm, proposed by [21], together with the law of large numbers [20], to compute conditional mean state estimates of the finite-state Markov chain and the state of jump Markov linear system. In Theorem 3.1, we prove the convergence of the DA applied to the state-space model (1) and (2).
- In Section IV, a new SA algorithm is derived based on the DA algorithm. The algorithm yields joint MAP state sequence estimates of the finite-state Markov chain and the state of jump Markov linear system. Sufficient conditions for the convergence of the proposed algorithm to the global maxima are presented in Theorem 4.1.
- In Section V, a Metropolis–Hastings SA algorithm, based on DA, which yields the MAP state sequence estimate of the Markov chain, is derived. Sufficient conditions for the convergence of the proposed algorithm to the global maxima are presented in Theorem 5.1.
- In Section VI, the proposed algorithms are applied to two practical examples: 1) estimation of sparse signals and 2) narrowband interference suppression in spread spectrum code-division multiple access (CDMA) mobile communication systems. As detailed in Section VI, several recent papers in the communications and adaptive signal processing literature have studied these problems and developed recursive (on-line) state and parameter estimation algorithms. Our objective is to illustrate the use of stochastic sampling algorithms in these applications.

## II. Problem Formulation

In this section, we present the signal model and outline the estimation objectives.

### A. Signal Model

Let $t \in \{1, 2, \cdots\}$ denote discrete time. Let $r_t$ denote a discrete-time, time-homogeneous, $s$-state, first-order Markov chain with transition probabilities

$$p_{ij} \triangleq \Pr\{r_{t+1} = j | r_t = i\}, \qquad (i, j \in S);$$
$$S = \{1, 2, \cdots, s\}. \tag{3}$$

Denote the initial probability distribution as $p_i \triangleq \Pr\{r_1 = i\}$, for $i \in S$, such that $p_i \geq 0, \ \forall i \in S$ and $\sum_{i=1}^{S} p_i = 1$. The transition probability matrix $[p_{ij}]$, is an $s \times s$ matrix, with elements satisfying $p_{ij} \geq 0$ and $\sum_{j=1}^{s} p_{ij} = 1$, for each $i \in S$.

$s^T$ possible realizations of the Markov chain $\boldsymbol{r}$ exist. The set of all realizations of $\boldsymbol{r}$ is denoted as $\{\boldsymbol{r}(1), \boldsymbol{r}(2), \cdots, \boldsymbol{r}(s^T)\}$. The realizations $\boldsymbol{r}(l)$ for $l = 1, \cdots, s^T$ are ordered arbitrarily to simplify notation.

Let $R$ denote the set of realizations of the finite Markov chain $r_t$ of non-null prior probability; that is

$$R = \Bigg\{ \boldsymbol{r}(l) = (r_1(l), \cdots, r_T(l)); l \in \{1, \cdots, s^T\}$$
$$\text{such that } p_{r_1(l)} \prod_{t=1}^{T-1} p_{r_t(l) r_{t+1}(l)} > 0 \Bigg\}. \tag{4}$$

*Notation:* We will use $n_x$ to denote the dimension of an arbitrary vector $x$. We use $\int \varphi(\boldsymbol{r})\, d\boldsymbol{r}$ instead of $\sum_{\boldsymbol{r}(l) \in S^T} \varphi(\boldsymbol{r}(l))$ for $\varphi$: $S^T \to \mathbb{R}$, where $S^T = S \times S \times \cdots \times S$. We do not distinguish between random variables and their realizations. $f(\cdot|\cdot)$ will be used to represent distributions of both discrete and continuous random variables. Superscripts denote exponentiation, e.g., $f^{1/T(k)}$. Superscripts enclosed in brackets, e.g., $\boldsymbol{x}^{(k)}$ denotes iteration number $k$. The index $k$ denotes iteration number of the various iterative algorithms. ∘

Consider the jump Markov linear system of (1) and (2), where $x_t \in \mathbb{R}^{n_x}$ is the system state, $y_t \in \mathbb{R}^{n_y}$ is the observation at time $t$, $u_t \in \mathbb{R}^{n_u}$ is a known deterministic input, $v_t \in \mathbb{R}^{n_v}$ is a zero-mean white Gaussian noise sequence with identity covariance $\boldsymbol{I}_{n_v}$, and $B(i)B^t(i) > 0$ $(\forall i \in S)$,[1] $w_t \in \mathbb{R}^{n_w}$, is a zero-mean white Gaussian noise sequence with identity covariance $\boldsymbol{I}_{n_w}$ and $D(i)D'(i) > 0$ $(\forall i \in S)$. $A, B, C, D, F,$ and $G$ are functions of the Markov chain state $r_t$, i.e., $\{A, B, C, D, F, G\} \subset \{A(i), B(i), C(i), D(i), F(i), G(i); i \in S\}$, and they evolve according to the realization of the finite-state Markov

[1]This assumption is not restrictive as in numerous applications when $B(i)B'(i)$ is singular. The jump Markov linear system can be transformed to a new system where the noise covariance matrix is positive definite. See [10, Sec. 3.9] or [5] for details.

chain $r_t$. We assume $x_0 \sim \mathcal{N}(\hat{x}_0, P_0)$ and let $x_0$, $v_t$ and $w_t$ be mutually independent for all $t$.

*Assumption 2.1:* The model parameters $\lambda$ are assumed known where

$$\lambda \triangleq \{p_i, p_{ij}, A(i), B(i), C(i), D(i), F(i), G(i), \hat{x}_0, P_0; \quad i, j \in S\}. \tag{5}$$

*Remark 2.1:* The assumption that $x_0$ is normally distributed is easily relaxed to finite Gaussian mixture distributions. Because any regular distribution can be approximated arbitrarily closely by a finite mixture of Gaussians, this assumption is not really restrictive. The procedure for dealing with finite-mixture Gaussian distributed $x_0$ is as follows. Suppose $x_0$ satisfies $x_0 \sim \sum_{i=1}^{n} \pi_i \mathcal{N}(m_i, \Sigma_i)$, where $\sum_{i=1}^{N} \pi_i = 1$ and $\pi_i \geq 0$. Introduce the indicator variable $I_0 \in \{1, \cdots, n\}$ such that $x_0 | (I_0 = i) \sim \mathcal{N}(m_i, \Sigma_i)$. Then, conditioned upon $(I_0, r_1, \cdots, r_T)$, the system (1) and (2) is linear Gaussian, so we can use the algorithms presented below.

*Remark:* In many applications, the model parameters are known. For example, in spread spectrum communications systems (see Section VI-B) $p_{ij}$ are known *a priori* since they are a function of the spreading code. In cases in which the model parameters are not known, several algorithms for estimating these parameters are available; see, for example, [25]. Such models are also considered under a dynamic linear model framework in [26]. It is also possible in a Bayesian framework to use algorithms presented in this paper to jointly compute state and parameter estimates. An important issue beyond the scope of this paper is the identifiability of the model parameters of a jump Markov linear system.

### B. Estimation Objectives

Given the observations $\boldsymbol{y}$, assuming the model parameters $\lambda$ are exactly known, all Bayesian inference for jump Markov linear systems relies on the joint posterior distribution $f(\boldsymbol{r}(l), \boldsymbol{x}|\boldsymbol{y})$. In this paper, the following three Bayesian estimation problems are considered:

1) Conditional mean estimates of $\boldsymbol{x}$ and $\boldsymbol{r}$: compute optimal (in a mean-square sense) estimates of $\boldsymbol{x}$ and $\boldsymbol{r}$ given by $\mathbb{E}\{\boldsymbol{r}|\boldsymbol{y}\}$ and $\mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}\}$.
2) MAP estimate of $\boldsymbol{x}$ and $\boldsymbol{r}$: compute optimal (in a MAP sense) state estimates of $\boldsymbol{x}$ and $\boldsymbol{r}$ by maximizing the joint posterior distribution, i.e.,

$$(\hat{\boldsymbol{x}}, \hat{\boldsymbol{r}}) = \arg\max_{\boldsymbol{x}, \boldsymbol{r}(l)} f(\boldsymbol{x}, \boldsymbol{r}(l)|\boldsymbol{y}). \tag{6}$$

3) MAP estimate of $\boldsymbol{r}$: compute optimal (in a MAP sense) state estimates of $\boldsymbol{r}$ by maximizing the marginal posterior distribution, i.e.,

$$\hat{\boldsymbol{r}} = \arg\max_{\boldsymbol{r}(l)} f(\boldsymbol{r}(l)|\boldsymbol{y}). \tag{7}$$

## III. Conditional Mean Estimation

The aim of this section is to compute the conditional mean estimates $\mathbb{E}\{\boldsymbol{r}|\boldsymbol{y}\}$ and $\mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}\}$ for the signal model (1), (2). These estimates are "theoretically" obtained by integration with respect to the joint posterior distribution $f(\boldsymbol{r}(l), \boldsymbol{x}|\boldsymbol{y})$. If we were able to obtain $N$ (for large $N$) i.i.d. samples $\{(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)}); k = 1, \cdots, N\}$ according to the distribution $f(\boldsymbol{r}(l), \boldsymbol{x}|\boldsymbol{y})$, then using the Law of Large Numbers [20], conditional mean estimates can be computed by averaging, thus solving the state estimation problem. Unfortunately, obtaining such i.i.d. samples from the posterior distribution $f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})$ is not straightforward. Thus, alternative sampling schemes must be investigated.

### A. Data Augmentation

In this paper, we compute samples from the posterior distribution $f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})$ using MCMC methods, and in particular, we use DA algorithm. The samples are then used to compute conditional mean estimates of the states $\boldsymbol{r}$ and $\boldsymbol{x}$. The proposed conditional mean state estimator via the data augmentation algorithm is summarized in Fig. 1.

*Remark 3.1:* Theoretically speaking, the DA algorithm does not have a stopping criterion. However, a reasonable choice (see, for example, [4]) is to terminate the algorithm when $||\overline{\boldsymbol{r}}(N) - \overline{\boldsymbol{r}}(N-1)||_2$ is less than some specified tolerance limit.

*Sampling Schemes:* The DA algorithm presented in Fig. 1 requires us to compute samples from $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}^{(k-1)})$ and $f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}^{(k)})$. One possible scheme is the efficient forward filtering–backward sampling recursions introduced by Carter and Kohn [2] and independently by Früwirth-Schnatter [5]. These recursions are given in the Appendix. An alternative scheme, not investigated here, for sampling from posterior densities of Gaussian state space systems is the simulation smoother of De Jong and Shephard [3].

### B. Convergence of Data Augmentation

The DA algorithm described in Section III-A has been used in [2] for identification of linear state-space models with errors that are a mixture of normals and coefficients that can switch with time. We generalize the model in [2] to our jump Markov linear system given by (1) and (2). Our main contribution here is to prove that sampling via the DA algorithm converges uniformly geometrically fast to yield samples that mimic samples from the desired (or target) distribution. As a consequence, the Law of Large Numbers holds, which states that sample averaging will converge almost surely (a.s.) to the expected value.

From construction [see (8) and (9)], the process $\{(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)}); k \in \mathbb{N}\}$ is a homogeneous Markov chain with transition kernel given by

$$K\left(\left(\boldsymbol{r}^{(k-1)}, \boldsymbol{x}^{(k-1)}\right) = (\boldsymbol{r}(l), \boldsymbol{x}); \left(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)}\right) = (\boldsymbol{r}(m), \boldsymbol{x}')\right) = f(\boldsymbol{x}'|\boldsymbol{y}, \boldsymbol{r}(l))\, f(\boldsymbol{r}(m)|\boldsymbol{y}, \boldsymbol{x}'). \tag{10}$$

The following theorem (proven in Appendix II-A) and corollary are the main results of this section.

*Theorem 3.1—Uniform Geometric Ergodicity:* A constant $0 \leq \rho < 1$, $\rho \triangleq 1 - \sum_{\boldsymbol{r}(m) \in R} \min_{\boldsymbol{r}(l) \in R} K(\boldsymbol{r}(l); \boldsymbol{r}(m))$ exists, such that, for any initial distribution $p_0(\boldsymbol{r}, \boldsymbol{x})$ of $(\boldsymbol{r}^{(0)}, \boldsymbol{x}^{(0)})$, the

**Data Augmentation Algorithm**

1. **Initialization:** Select randomly $\left(\mathbf{r}^{(0)}, \mathbf{x}^{(0)}\right)$.
2. **Iteration:** Given $\left(\mathbf{r}^{(k-1)}, \mathbf{x}^{(k-1)}\right)$, compute $\left(\mathbf{r}^{(k)}, \mathbf{x}^{(k)}\right)$ for the $k$th iteration (for $k = 1, 2, \ldots$) as follows:
   - Simulate $\mathbf{x}^{(k)}$ from
   $$\mathbf{x}^{(k)} \sim f\left(\mathbf{x} | \mathbf{y}, \mathbf{r}^{(k-1)}\right) \tag{8}$$
   - Simulate $\mathbf{r}^{(k)}$ from
   $$\mathbf{r}^{(k)} \sim f\left(\mathbf{r} | \mathbf{y}, \mathbf{x}^{(k)}\right) \tag{9}$$
3. **State Estimation:** Compute conditional mean estimates using the empirical estimators $\overline{\mathbf{x}}(N)$ and $\overline{\mathbf{r}}(N)$ (defined in Eqs. (14) and (15)) or the mixture estimators $\hat{\mathbf{x}}(N)$ and $\hat{\mathbf{r}}(N)$ (defined in Eqs. (16) and (17)).

Fig. 1. Algorithm I: Conditional mean state estimator via the data augmentation algorithm.

state distribution of the Markov chain $(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)})$ at iteration $k$, denoted as $p_k(\boldsymbol{r}, \boldsymbol{x})$, satisfies

$$\iint |p_k(\boldsymbol{r}, \boldsymbol{x}) - f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})| \, d\boldsymbol{r} \, d\boldsymbol{x} \leq 2\rho^{k-2}. \tag{11}$$

Thus, the sequence $\{(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)}); k \in \mathbb{N}\}$ obtained from the DA sampling scheme is a Markov chain with invariant distribution $f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})$ and is ergodic. Ergodicity implies convergence of ergodic (sample) averages [23, Th. 3, p. 1717]. Uniform ergodicity implies that the Law of Large Numbers and a central limit theorem also hold [4], [18], and [23, Th. 5, p. 1717]).

*Corollary 3.1—Convergence of Ergodic Averages:* For every real-valued function $\varphi$: $S^T \times \mathbb{R}^{(T+1)n_x} \to \mathbb{R}$, let us consider the time average of the $N$ first outputs of the Markov chain $\overline{\varphi}_N \triangleq (1/N) \sum_{k=0}^{N-1} \varphi(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)})$. If $\mathbb{E}_{f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})}[|\varphi(\boldsymbol{r}, \boldsymbol{x})|] < \infty$, then, for any initial distribution,

$$\overline{\varphi}_N \to \mathbb{E}_{f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})}[\varphi(\boldsymbol{r}, \boldsymbol{x})] \qquad \text{a.s.} \tag{12}$$

If $\mathbb{E}_{f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})}[|\varphi(\boldsymbol{r}, \boldsymbol{x})|^2] < +\infty$, then a constant $\sigma(\varphi)$ exists such that the distribution of

$$\sqrt{N} \left(\overline{\varphi}_N - \mathbb{E}_{f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})}[\varphi(\boldsymbol{r}, \boldsymbol{x})]\right) \tag{13}$$

converges in distribution to a zero-mean normal distribution of variance $\sigma^2(\varphi)$.

*Remark 3.2:* Theorem 3.1 states that the sample $(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)}; k \in \mathbb{N})$ generated for large $k$ via the DA algorithm, will mimic a sample from the posterior distribution $f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})$.

*Remark 3.3:* $2\rho^{k-2}$ in (11) of Theorem 3.1, is an upper bound on the rate of convergence of the DA algorithm. Because $\rho$ is not known, we cannot *a priori* choose the number of iterations $k$ required for the DA algorithm to converge to a desired level of accuracy.

*Computing Conditional Mean Estimates:* Corollary 3.1 can straightforwardly be applied to compute the conditional mean estimates of $\boldsymbol{r}$ and $\boldsymbol{x}$ by the ergodic averages $\overline{\boldsymbol{r}}(N)$ and $\overline{\boldsymbol{x}}(N)$[2]:

$$\overline{\boldsymbol{r}}(N) \triangleq \frac{1}{N} \sum_{k=0}^{N-1} \boldsymbol{r}^{(k)} \tag{14}$$

$$\overline{\boldsymbol{x}}(N) \triangleq \frac{1}{N} \sum_{k=0}^{N-1} \boldsymbol{x}^{(k)} \tag{15}$$

where $\overline{\boldsymbol{r}}$ and $\overline{\boldsymbol{x}}$ are known as the *empirical estimators* (see [11]).

Conditional mean estimates may also be computed via the *mixture estimators* (see [11]), i.e.,

$$\hat{\boldsymbol{r}}(N) \triangleq \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}\left[\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}^{(k)}\right] \tag{16}$$

$$\hat{\boldsymbol{x}}(N) \triangleq \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}\left[\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}^{(k)}\right]. \tag{17}$$

The empirical estimator and the mixture estimator will almost surely converge to the true condition mean estimates. The question is to determine which of the two estimators yields smaller asymptotic variances.

The following proposition shows that the mixture estimator yields smaller asymptotic variances.

*Proposition 3.1:* When the Markov chain $\{(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)}); k \in \mathbb{N}\}$ is in its stationary regime, then the estimates

[2] $\overline{\boldsymbol{r}}(N)$ and $\hat{\boldsymbol{r}}(N)$ defined subsequently in (16), denote the average over $N$ iterations. They are not to be confused with $\boldsymbol{r}(\cdot)$ defined in (4).

**Simulated Annealing Data Augmentation Algorithm**

1. Initialization: Select randomly $\left(\mathbf{r}^{(0)}, \mathbf{x}^{(0)}\right)$.

2. Iteration: Given $\left(\mathbf{r}^{(k-1)}, \mathbf{x}^{(k-1)}\right)$, compute $\left(\mathbf{r}^{(k)}, \mathbf{x}^{(k)}\right)$ for the $k$th iteration (for $k = 1, 2, \ldots$) as follows:

   - Simulate $\mathbf{x}^{(k)}$ from
   $$\mathbf{x}^{(k)} \sim \overline{f}^{1/T(k)}\left(\mathbf{x}|\,\mathbf{y}, \mathbf{r}^{(k-1)}\right) \tag{21}$$
   - Simulate $\mathbf{r}^{(k)}$ from
   $$\mathbf{r}^{(k)} \sim \overline{f}^{1/T(k)}\left(\mathbf{r}|\,\mathbf{y}, \mathbf{x}^{(k)}\right) \tag{22}$$

   where $\overline{f}^{1/T(k)}\left(\mathbf{x}|\,\mathbf{y}, \mathbf{r}^{(k-1)}\right)$ and $\overline{f}^{1/T(k)}\left(\mathbf{r}|\,\mathbf{y}, \mathbf{x}^{(k)}\right)$ are defined in Eqs. (23) and (24), respectively.

Fig. 2. Algorithm II: SA–DA algorithm for computing joint MAP sequence estimates of the finite-state Markov chain $\boldsymbol{r}$ and the state of the jump linear system $\boldsymbol{x}$.

$(1/N)\sum_{k=0}^{N-1}\mathbb{E}[\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}^{(k)}]$ and $(1/N)\sum_{k=0}^{N-1}\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}^{(k)}]$ converge a.s. to $\mathbb{E}[\boldsymbol{r}|\boldsymbol{y}]$ and $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]$, respectively, and satisfy

$$\operatorname{var}\left[\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}^{(k)}\right]\right] \leq \operatorname{var}\left[\frac{1}{N}\sum_{k=0}^{N-1}\boldsymbol{r}^{(k)}\right] \tag{18}$$

$$\operatorname{var}\left[\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}^{(k)}\right]\right] \leq \operatorname{var}\left[\frac{1}{N}\sum_{k=0}^{N-1}\boldsymbol{x}^{(k)}\right]. \tag{19}$$

*Proof:* The proof can be found in [11, Th. 4.1].

In our case, $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}^{(k)}]$ and $\mathbb{E}[\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}^{(k)}]$ can be easily computed using, respectively, a Kalman smoother [1] and the forward–backward recursions of a hidden Markov model smoother [17].

## IV. Maximum *A Posteriori* State Sequence Estimation of $\boldsymbol{x}$ and $\boldsymbol{r}$

### A. Simulated Annealing Data Augmentation Scheme

SA is a numerical optimization technique that allows us to solve combinatorial optimization problems [14]. It is a stochastic algorithm, which will converge to globally optimal solutions, by randomly generating a sequence of possible solutions.

In this section, we introduce an algorithm for obtaining optimal, in the MAP sense, joint sequence estimates of the states $\boldsymbol{x}$ and $\boldsymbol{r}$ defined in (6). We propose an SA version of the DA algorithm; i.e., we build a nonhomogeneous version of the DA algorithm dependent on a deterministic so-called cooling schedule $\{T(k);\, k \in \mathbb{N}\}$, verifying

$$T(k+1) \leq T(k) \quad \text{and} \quad \lim_{k\to+\infty} T(k) = 0. \tag{20}$$

The proposed algorithm is summarized in Fig. 2.

To implement this algorithm, we sample from $\overline{f}^{1/T(k)}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}^{(k-1)})$ and $\overline{f}^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}^{(k)})$ using the efficient forward filtering–backward sampling recursions. The desired densities in (21) and (22) are defined as follows:

$$\overline{f}^{1/T(k)}\left(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}^{(k-1)}\right) \propto f^{1/T(k)}\left(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}^{(k-1)}\right) \tag{23}$$

$$\overline{f}^{1/T(k)}\left(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}^{(k)}\right) \propto f^{1/T(k)}\left(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}^{(k)}\right) \tag{24}$$

where $\propto$ denotes proportionality.

### B. Convergence of the Algorithm

Here, we obtain sufficient conditions on $T(k)$ to ensure convergence of the Markov chain $\{(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)});\, k \in \mathbb{N}\}$ to the set of global maxima. First, we note that obtaining MAP estimates of $\boldsymbol{r}$ and $\boldsymbol{x}$ is equivalent to solve a NP (nondeterministic polynomial) hard combinatorial optimization problem. Indeed (6) is equivalent to

$$(\hat{\boldsymbol{r}}, \hat{\boldsymbol{x}}) = \arg\max_{\boldsymbol{r}(l),\, \boldsymbol{x}} f(\boldsymbol{r}(l)|\boldsymbol{y}) f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}(l)). \tag{25}$$

Note, $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}(l))$ is a $n_x(T+1)$-dimensional Gaussian distribution of mean $\boldsymbol{m_l}$ and covariance $\boldsymbol{\Sigma}_l$, with a strictly positive determinant $|\boldsymbol{\Sigma}_l|$. Thus, $\hat{\boldsymbol{r}}$ is obtained as

$$\hat{\boldsymbol{r}} = \arg\max_{\boldsymbol{r}(l)\in S^T} f(\boldsymbol{r}(l)|\boldsymbol{y})|\boldsymbol{\Sigma}_l|^{-1/2}. \tag{26}$$

We denote $\mathcal{M} \subset R$ the set of these global maxima. Once $\hat{\boldsymbol{r}}$, say, $\hat{\boldsymbol{r}} = \boldsymbol{r}(l)$, has been obtained, then $\hat{\boldsymbol{x}} = \boldsymbol{m_l}$ and $\boldsymbol{m_l}$ is obtained via a Kalman smoother [1].

Our proof of convergence relies on the fact that the stochastic process $\{\boldsymbol{r}^{(k)};\, k \in \mathbb{N}\}$ is a inhomogeneous finite-state space Markov chain that converges asymptotically toward the set of global maxima $\mathcal{M}$. From the definition of our algorithm,

$\{(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)}); k \in \mathbb{N}\}$ is an inhomogeneous Markov chain with transition kernel at time $k$ given by

$$K_k\left(\left(\boldsymbol{r}^{(k-1)}, \boldsymbol{x}^{(k-1)}\right); \left(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)}\right)\right) = \overline{f}^{1/T(k)}\left(\boldsymbol{x}^{(k)}|\boldsymbol{y}, \boldsymbol{r}^{(k-1)}\right)\overline{f}^{1/T(k)}\left(\boldsymbol{r}^{(k)}|\boldsymbol{y}, \boldsymbol{x}^{(k)}\right). \quad (27)$$

Thus, $\{\boldsymbol{r}^{(k)}; k \in \mathbb{N}\}$ is an inhomogeneous finite-state space Markov chain of transition kernel

$$K_k\left(\boldsymbol{r}^{(k-1)} = \boldsymbol{r}(l); \boldsymbol{r}^{(k)} = \boldsymbol{r}(m)\right) = \int \overline{f}^{1/T(k)}(\boldsymbol{r}(m)|\boldsymbol{y}, \boldsymbol{x})\,\overline{f}^{1/T(k)}(\boldsymbol{x}|\boldsymbol{y}\boldsymbol{r}(l))\,d\boldsymbol{x}. \quad (28)$$

The main result in this section is given by the following theorem, which is proven in Appendix II-F.

*Theorem 4.1:* If $\{T(k); k \in \mathbb{N}\}$ satisfies

$$T(k) = \frac{\gamma}{\ln(k+u)} \quad (29)$$

with $\gamma \geq L$, where $L$ is defined in (81) and $u > 0$, then the Markov chain $\{\boldsymbol{r}^{(k)}; k \in \mathbb{N}\}$ is strongly ergodic. For any initial distribution, the distribution $p_k(\boldsymbol{r})$ of $\boldsymbol{r}^{(k)}$ satisfies

$$\lim_{k\to+\infty} \int |p_k(\boldsymbol{r}) - \overline{f}^{\infty}(\boldsymbol{r}|\boldsymbol{y})|\,d\boldsymbol{r} = 0 \quad (30)$$

where $\overline{f}^{\infty}(\boldsymbol{r}|\boldsymbol{y})$ is defined in (32). Furthermore, and the Markov chain the distribution $p_k(\boldsymbol{x})$ of $\boldsymbol{x}^{(k)}$ satisfies

$$\lim_{k\to+\infty} \int |p_k(\boldsymbol{x}) - f_k(\boldsymbol{x})|\,d\boldsymbol{x} = 0 \quad (31)$$

where $f_k(\boldsymbol{x}) \triangleq \int \overline{f}^{1/T(k)}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r})\overline{f}^{\infty}(\boldsymbol{r}|\boldsymbol{y})\,d\boldsymbol{r}$.

The proof of Theorem 4.1 involves the following four lemmas.

*Lemma 4.1:* $K_k((\boldsymbol{r}^{(k-1)}, \boldsymbol{x}^{(k-1)}); (\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)}))$ admits $\overline{f}^{1/T(k)}(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})$ as its invariant distribution, where $\overline{f}^{1/T(k)}(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y}) \propto f^{1/T(k)}(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})$. Thus, the marginal transition kernel $K_k(\boldsymbol{r}^{(k-1)}; \boldsymbol{r}^{(k)})$ admits $\overline{f}^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y}) \triangleq \int \overline{f}^{1/T(k)}(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})\,d\boldsymbol{x}$ as its invariant distribution.

*Proof:* See Appendix II-B.

*Lemma 4.2:* The sequence of invariant distributions $\overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})$ converges, as $k$ goes to infinity, to a distribution $\overline{f}^{\infty}(\boldsymbol{r}(l)|\boldsymbol{y})$ localized on the set $\mathcal{M}$

$$\overline{f}^{\infty}(\boldsymbol{r}(l)|\boldsymbol{y}) = \frac{|\boldsymbol{\Sigma}_l|^{1/2}}{\sum_{\boldsymbol{r}(m)\in\mathcal{M}} |\boldsymbol{\Sigma}_m|^{1/2}}\,\delta_{\mathcal{M}}(\boldsymbol{r}(l)) \quad (32)$$

where $\delta_{\mathcal{M}}(\boldsymbol{r}(l)) = 1$ if $\boldsymbol{r}(l) \in \mathcal{M}$ and $\delta_{\mathcal{M}}(\boldsymbol{r}(l)) = 0$ otherwise.

*Proof:* See Appendix II-C.

*Lemma 4.3:* The sequence of invariant discrete distributions $\overline{f}^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y})$ satisfies

$$\sum_{k=0}^{+\infty} \int |\overline{f}^{1/T(k+1)}(\boldsymbol{r}|\boldsymbol{y}) - \overline{f}^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y})|\,d\boldsymbol{r} < +\infty. \quad (33)$$

*Proof:* See Appendix II-D.

*Lemma 4.4:* $k_0 \in \mathbb{N}$ exists such that the sequence of transitions kernels $K_k$ of the inhomogeneous finite-state–space Markov chain $\{\boldsymbol{r}^{(k)}; k \in \mathbb{N}\}$ satisfies for any $(\boldsymbol{r}(l), \boldsymbol{r}(m)) \in R \times R$ and $k \geq k_0$

$$K_k(\boldsymbol{r}(l); \boldsymbol{r}(m)) \geq C\exp\left(-\frac{L}{T(k)}\right) \quad (34)$$

where $0 < C < \infty$ and $0 \leq L < \infty$ are some constants independent of $k$.

*Proof:* See Appendix II-E.

## V. Maximum *A Posteriori* State Sequence Estimate of $\boldsymbol{r}$

### A. Metropolis–Hastings/Data Augmentation Simulated Annealing Scheme

In this section, we present an SA algorithm based on DA for obtaining the MAP state sequence estimate of the finite Markov chain $\boldsymbol{r}$ defined in (7). We build a nonhomogeneous Markov chain whose transition kernel at iteration $k$ depends on a cooling schedule $\{T(k); k \in \mathbb{N}\}$, verifying

$$T(k+1) \leq T(k) \quad \text{and} \quad \lim_{k\to+\infty} T(k) = 0. \quad (35)$$

The proposed algorithm is summarized in Fig. 3.

To implement the proposed algorithm, we sample from $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}^{(k-1)})$ and $f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}^{(k)})$ using the forward filtering–backward sampling recursions. To evaluate the acceptance ratio (38), we need to evaluate $f(\boldsymbol{r}|\boldsymbol{y})$ up to a normalizing constant. We have $f(\boldsymbol{r}|\boldsymbol{y}) \propto f(\boldsymbol{r})f(\boldsymbol{y}|\boldsymbol{r})$. The first term is the prior distribution of the realization of the finite-state Markov chain. The second term is the likelihood that is evaluated using the weighted sequence of innovations given by the Kalman filter [1]. Because the Kalman filter is used to compute samples from $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r})$, the additional computational cost of evaluating $f(\boldsymbol{r}|\boldsymbol{y})$ is minimal.

### B. Convergence of the Algorithm

We obtain sufficient conditions on $T(k)$ to ensure convergence of the algorithm toward the set $\mathcal{M}^{\star} \subset R$ of global maxima of $f(\boldsymbol{r}|\boldsymbol{y})$ following the approach of Mitra *et al.* [14]. From the definition of the algorithm, $\{\boldsymbol{r}^{(k)}; k \in \mathbb{N}\}$ is an inhomogeneous finite-state Markov chain of transition kernels $\{K_k; k \in \mathbb{N}\}$

$$K_k(\boldsymbol{r}(l); \boldsymbol{r}(m)) = \alpha_k(\boldsymbol{r}(l), \boldsymbol{r}(m))K_h(\boldsymbol{r}(l); \boldsymbol{r}(m)) + \delta_{lm}\int (1 - \alpha_k(\boldsymbol{r}(l), \boldsymbol{r}))K_h(\boldsymbol{r}(l); \boldsymbol{r})\,d\boldsymbol{r} \quad (39)$$

**Metropolis-Hastings/Data Augmentation Simulated Annealing Algorithm**

1. Initialization: Select randomly $\left(\mathbf{r}^{(0)}, \mathbf{x}^{(0)}\right)$ and set $k = 1$.

2. Iteration: Given $\left(\mathbf{r}^{(k-1)}, \mathbf{x}^{(k-1)}\right)$, compute $\left(\mathbf{r}^{(k)}, \mathbf{x}^{(k)}\right)$ for the $k$th iteration as follows:

   - Simulate $\mathbf{x}^{(k)}$ from
   $$\mathbf{x}^{(k)} \sim f\left(\mathbf{x} | \mathbf{y}, \mathbf{r}^{(k-1)}\right) \tag{36}$$
   - Simulate a 'candidate' $\mathbf{r}_c$ from
   $$\mathbf{r}_c \sim f\left(\mathbf{r} | \mathbf{y}, \mathbf{x}^{(k)}\right) \tag{37}$$
   - Evaluate the acceptance ratio $\alpha\left(\mathbf{r}^{(k-1)}, \mathbf{r}_c\right)$:
   $$\alpha_k\left(\mathbf{r}^{(k-1)}, \mathbf{r}_c\right) = \min\left\{\left[\frac{f\left(\mathbf{r}_c | \mathbf{y}\right)}{f\left(\mathbf{r}^{(k-1)} | \mathbf{y}\right)}\right]^{1/T(k)-1}, 1\right\} \tag{38}$$
   - Simulate an uniform random variable $u$ on $[0, 1]$, $u \sim \mathcal{U}(0, 1)$.
   - If $u < \alpha_k\left(\mathbf{r}^{(k-1)}, \mathbf{r}_c\right)$ then set $\mathbf{r}^{(k)} = \mathbf{r}_c$, otherwise set $\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)}$ .

Fig. 3. Algorithm III: Metropolis–Hastings/DA–SA algorithm for computing the MAP estimate of the finite-state Markov chain $\boldsymbol{r}$.

where

$\delta_{lm}$ Kronecker delta;

$K_h$ associated (homogeneous) DA Markov chain kernel defined as follows:

$$K_h(\boldsymbol{r}(l); \boldsymbol{r}(m)) \triangleq \int f(\boldsymbol{r}(m)|\boldsymbol{y}, \boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}(l))\, d\boldsymbol{x}. \tag{40}$$

The following theorem is the main result of this section and is proven in Appendix II-I.

*Theorem 5.1:* If $\{T(k);\, k \in \mathbb{N}\}$ satisfies

$$T(k) = \frac{\gamma}{\ln(k+u)} \tag{41}$$

with $\gamma \geq L$, where $L$ is defined in (85), and $u > 0$, then the Markov chain $\{\boldsymbol{r}^{(k)};\, k \in \mathbb{N}\}$ is strongly ergodic. For any initial distribution, the distribution $p_k(\boldsymbol{r})$ of $\boldsymbol{r}^{(k)}$ satisfies

$$\lim_{k \to +\infty} \int \left|p_k(\boldsymbol{r}) - \overline{f}^{\infty}(\boldsymbol{r})\right| d\boldsymbol{r} = 0 \tag{42}$$

where $\overline{f}^{\infty}(\boldsymbol{r})$ is defined in (45). Furthermore, the distribution $p_k(\boldsymbol{x})$ of $\boldsymbol{x}^{(k)}$ satisfies

$$\lim_{k \to +\infty} \int \left|p_k(\boldsymbol{x}) - f_k(\boldsymbol{x})\right| d\boldsymbol{x} = 0 \tag{43}$$

where $\overline{f}^{\infty}(\boldsymbol{x}) \triangleq \int f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}) \overline{f}^{\infty}(\boldsymbol{r})\, d\boldsymbol{r}$.

The proof of Theorem 5.1 involves the following four lemmas.

*Lemma 5.1:* For any $k \in \mathbb{N}^*$, $K_k(\boldsymbol{r}(m); \boldsymbol{r}(l))$ admits $\overline{f}^{1/T(k)}(\boldsymbol{r})$ as its invariant distribution, where

$$\overline{f}^{1/T(k)}(\boldsymbol{r}) \propto f^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y}). \tag{44}$$

*Proof:* See Appendix II-G.

We obtain straightforwardly the following lemmas; see, for example, [14] and [27].

*Lemma 5.2:* The sequence of invariant distributions $\overline{f}^{1/T(k)}(\boldsymbol{r})$ [defined in (44)] converges, as $k$ goes to infinity, toward the set of global maxima $\mathcal{M}^{\star}$ of $f(\boldsymbol{r}(l)|\boldsymbol{y})$; that is,

$$\overline{f}^{\infty}(\boldsymbol{r}(l)|\boldsymbol{y}) = \frac{\delta_{\mathcal{M}^{\star}}(\boldsymbol{r}(l))}{\overline{\mathcal{M}^{\star}}} \tag{45}$$

where $\delta_{\mathcal{M}^{\star}}(\boldsymbol{r}(l)) = 1$ if $\boldsymbol{r}(l) \in \mathcal{M}^{\star}$ and $\delta_{\mathcal{M}^{\star}}(\boldsymbol{r}(l)) = 0$ otherwise. $\overline{\mathcal{M}^{\star}}$ denotes the cardinality of $\mathcal{M}^{\star}$.

*Lemma 5.3:* The sequence of invariant discrete distributions $\overline{f}^{1/T(k)}(\boldsymbol{r})$ satisfies

$$\sum_{k=0}^{+\infty} \int \left|\overline{f}^{1/T(k+1)}(\boldsymbol{r}|\boldsymbol{y}) - \overline{f}^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y})\right| d\boldsymbol{r} < +\infty.$$

*Lemma 5.4:* $k_0$ exists such that, for any $k \geq k_0$ and for any $(\boldsymbol{r}(l), \boldsymbol{r}(m)) \in R \times R$, the sequence of transitions kernels $K_k$ of the inhomogeneous finite-state–space Markov chain $\{\boldsymbol{r}^{(k)};\, k \in \mathbb{N}\}$ satisfies

$$K_k(\boldsymbol{r}(l); \boldsymbol{r}(m)) \geq C \exp\left(-\frac{L}{T(k)}\right)$$

where $0 < C < \infty$ and $0 \leq L < \infty$ are some constants independent of $k$.

*Proof:* See Appendix II-H.

## VI. Applications and Numerical Examples

Theoretically, the DA and SA based on DA algorithms require an infinite number of iterations to give the exact values of conditional mean estimates and of global maxima, respectively.

In all of our computer simulations below, the first $N_0$ iterations of the DA algorithm are discarded. The first $N_0$ iterations are assumed to correspond to the so-called "burn in period" (or the transient period prior to the convergence) of the Markov chain.[3] As in [4], the DA algorithm is iterated until the desired computed values of the ergodic averages are no longer modified.

As shown in Sections IV and V, the SA algorithms require logarithmic cooling schedules. Such schedules are too slow to be implemented in practice. As is usually done in practice [6], [14], [27], we have implemented exponential and polynomial cooling schedules; i.e., $T(k) = C\alpha^k$ with $0 < \alpha < 1$ and $T(k) = Ck^n$. Our simulations (not presented here) show that little difference exists between polynomial and exponential cooling schedules. Therefore, in the simulations presented below, only exponential cooling schedules are used.

Computer simulations were carried out to evaluate the performances of our three algorithms. Section VI-A considers the problem of estimating a sparse signal developing from a neutron sensor based on a set of noisy data. Section VI-B considers the problem of narrowband interference suppression in spread spectrum CDMA communication systems.

### A. Estimation of a Sparse Signal

In several problems related to geophysics, nuclear science, or speech processing, the signal of interest can be modeled as an autoregressive process excited by a noise that admits as marginal distribution a mixture of Gaussians [13]. We consider the following model:

$$s_t = a_1 s_{t-1} + a_2 s_{t-2} + \tilde{v}_t \tag{46}$$

$$y_t = s_t + \sigma_w w_t \tag{47}$$

where $\tilde{v}$ is the dynamic (mixture) noise process

$$\tilde{v}_t \sim \lambda \mathcal{N}(0, \sigma_1^2) + (1-\lambda)\mathcal{N}(0, \sigma_2^2), \quad w_t \sim \mathcal{N}(0, 1).$$

$\tilde{v}_t$ is often assumed to be a white noise sequence, but it could be also modeled as a first-order Markov sequence to take into account the dead time of the sensor. This model (46) and (47) can be reexpressed as the jump Markov linear system (1) and (2), where the state vector is $x_t = (s_t, s_{t-1})'$ and for all $k$, $u_t = 0$

$$A = \begin{pmatrix} a_1 & a_2 \\ 1 & 0 \end{pmatrix}, \quad C = (1 \quad 0)$$

$$D = \sigma_w, \quad F = 0, \quad G = 0,$$

[3]Methods for determining the burn-in period $N_0$ are beyond the scope of this paper.

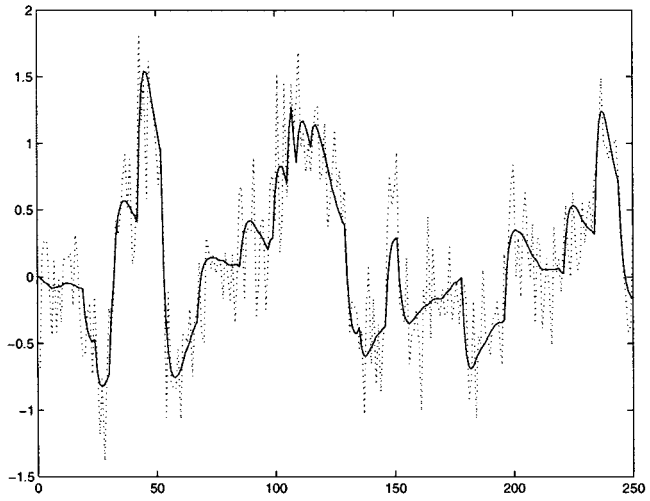


Fig. 4. Typical realization of the data: (—) true signal and $(\cdots)$ noisy observed signal.

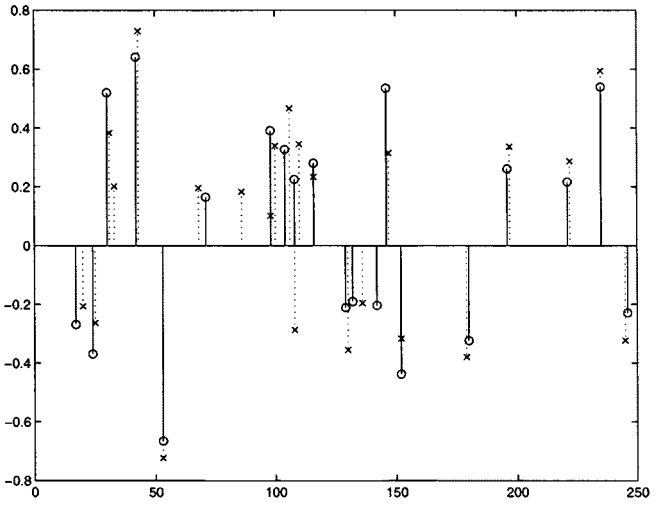


Fig. 5. True (×) and estimated (o) dynamic noise sequence. The estimates are the conditional mean estimates computed via Algorithm 1.

and $B(r_t) = (\sigma_i, 0)'$ with $i = r_t$. In the following simulations, we set the following parameters: $T = 250$, Markov noise: $p_1 = 0.1, p_2 = 0.9, p_{11} = 0.05, p_{22} = 0.894$ (so $r_t$ is in its stationary regime), $a_1 = 1.51$, $a_2 = -0.55$, $\lambda = 0.05$, $\sigma_1 = 0.30$, and $\sigma_2 = 0.01$. It models a neutron sensor in a noisy environment.

In Fig. 4, the signal $s_t$ and its noisy observations $y_t$ are displayed. We have chosen to illustrate the performance of our three algorithms by comparing the estimates of the dynamic noise $\tilde{v}_t$, which are directly obtained from the estimates of the states. The closer the estimate of the dynamic noise to the true dynamic noise, the better the performance of the algorithm [13]. In Fig. 5, we display the conditional mean estimate of the dynamic noise $\mathbb{E}\{\tilde{v}_t|\boldsymbol{y}\}$, computed using Algorithm 1. In Fig. 6, we present $\mathbb{E}\{\tilde{v}_t|\boldsymbol{y}, \boldsymbol{r}\}$, where $\boldsymbol{r}$ is computed using Algorithm 2. Finally, in Fig. 7, we present $\mathbb{E}\{\tilde{v}_t|\boldsymbol{y}, \boldsymbol{r}\}$, where $\boldsymbol{r}$ is the MAP estimate given by Algorithm 3. In all cases, the algorithms were initialized randomly.

Algorithms 1 and 3 give satisfactory results, because the signal-to-noise ratio (SNR) is quite low. Algorithm 2 underestimates the number of impulses in the signal. We must not

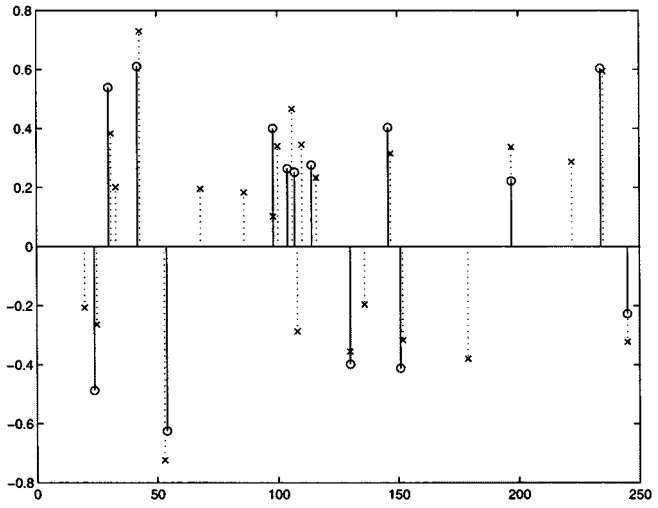


Fig. 6. True (×) and estimated (o) dynamic noise sequence. The estimates are given by $\mathbb{E}\{\bar{v}_t|\boldsymbol{y}, \boldsymbol{r}\}$, where $\boldsymbol{r}$ is computed using Algorithm 2.

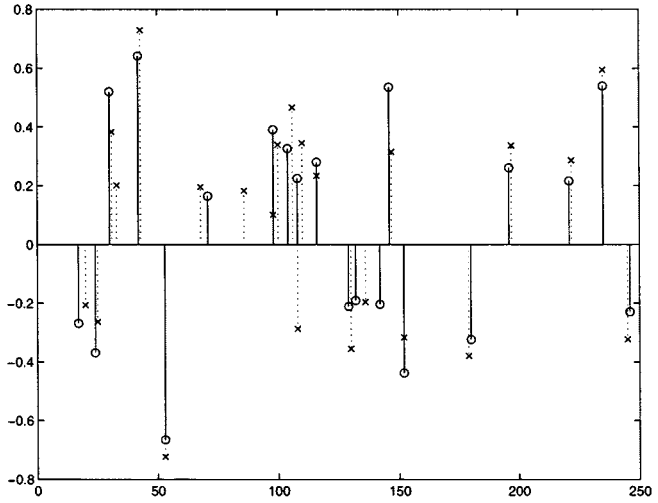


Fig. 7. True (×) and estimated (o) dynamic noise sequence. The estimates are given by $\mathbb{E}\{\bar{v}_t|\boldsymbol{y}, \boldsymbol{r}\}$, where $\boldsymbol{r}$ is computed using Algorithm 3.

be surprised by this result. Other works using these criteria in the specific field of sparse signal deconvolution report similar results [13] and adopt the criterion that we maximize using Algorithm 3.

In this example, we discard the first $N_0 = 20$ samples simulated by the DA algorithm. Then, we take into account the 150 following iterations of the DA algorithm. For the SA algorithms, we implement an exponential cooling schedule $T(k) = C\alpha^k$ with $C = 1$ and $\alpha = 0.98$, and we use 150 iterations.

### *B. Narrowband Interference Suppression in Spread Spectrum CDMA*

CDMA spread-spectrum signaling is among the most promising multiplexing technologies for cellular telecommunications services, such as personal telecommunications, mobile telephony, and indoor wireless networks. The explosive growth in cellular telephony, in conjunction with emerging new applications, has inspired significant interest in interference suppression techniques for enhancing the performance of CDMA systems. CDMA provides a means of separating the signals of multiple users transmitting simultaneously and occupying the same radio frequency (RF) bandwidth. It is well known that system performance is greatly enhanced if the receiver employs some means of suppressing narrowband interference before signal "despreading" [28].

Numerous recent papers study the problem of narrowband interference suppression in CDMA systems; see [15], [16], and [29], and the references therein. Our aim here is to examine the use of the iterative stochastic sampling algorithms proposed in the previous sections for narrowband interference suppression. Note, however, that realistic algorithms would be recursive (online).

In the papers [15], [16], [29], the following signal model is used: the sampled received signal $y_t$ consists of the spread spectrum signal $r_t$ from $N$ users, the narrowband interference $i_t$, and observation noise $w_t$, that is

$$y_t = r_t + i_t + \sigma_w w_t \tag{48}$$

where $w_t$ is a zero-mean white Gaussian process of variance 1. As in [15], [16], and [29], the narrowband interference $i_t$ is modeled as a second-order autoregressive process with both poles at $z = 0.99$; i.e.,

$$i_t = 1.98 i_{t-1} - 0.98 i_{t-2} + \sigma_e e_t \tag{49}$$

where $e_t$ is a zero-mean white Gaussian process of variance 1. The power of the received spread spectrum signal for each user was held constant with amplitudes ±1, randomly selected, and $r_t$ was binomially distributed.

The CDMA spread spectrum model (48) and (49) can be re-expressed as the jump Markov model of (1) and (2), where the state vector $x_t = (i_t \;\; i_{t-1})'$ denotes the state of the narrowband interference at times $k$ and $k-1$, $u_t = 1$ for all $k$ and

$$\begin{aligned} A &= \begin{pmatrix} 1.98 & -0.980 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} \sigma_e \\ 0 \end{pmatrix} \\ C &= (1 \;\; 0), \quad D = \sigma_w, \quad F = 0 \end{aligned} \tag{50}$$

and $G(1) = 1$, $G(2) = -1$. In the numerical examples below, we considered a single user, with $\sigma_e = 0.02$ and we compared the performance of the three proposed algorithms for increased observation noise. We compute the conditional mean estimate and the MAP estimates of the discrete sequence $r_t$. In this case, the joint MAP estimate given by Algorithm 2 and the MAP estimate given by Algorithm 3 are theoretically equal. We present the bit error rate for these algorithms. To evaluate the bit error rate from the conditional mean estimate $\mathbb{E}[r_t|\boldsymbol{y}]$, we set $\hat{r}_t = 1$ if $\overline{r}_t(N) > 0$ and $\hat{r}_t = -1$ otherwise.

The algorithms were run on 400 points, and averaged over 100 independent runs. In all cases, the algorithms were initialized randomly.

This problem is statistically easier than the problem described in Section VI-A. In this example, we discard the first $N_0 = 20$ samples simulated by the DA algorithm. Then, taking into account the 50 following iterations of the DA algorithm has appeared to be sufficient. For the SA algorithms, we implement

TABLE I
BIT ERROR RATE (IN PERCENT) OF THE THREE ALGORITHMS FOR INTERFERENCE SUPPRESION IN CDMA SYSTEMS

| $\sigma_w$ | Algorithm 1 | Algorithm 2 | Algorithm 3 |
|---|---|---|---|
| 0.5 | 3.13 | 3.51 | 3.25 |
| 0.6 | 5.88 | 6.82 | 6.48 |
| 0.7 | 8.84 | 10.23 | 10.61 |
| 0.8 | 11.89 | 13.02 | 14.90 |
| 0.9 | 14.54 | 16.12 | 17.88 |
| 1.0 | 17.29 | 18.04 | 21.12 |

an exponential cooling schedule $T(k) = C\alpha^k$ with $C = 1$ and $\alpha = 0.80$, and we use 50 iterations.

Our numerical examples presented here and other simulations (not presented here) suggest that Algorithm 1 performs the best for the CDMA narrowband interference suppression problem. The results obtained are better than those obtained using the EM algorithm [12].

## VII. CONCLUSION

In this paper, we have presented three iterative stochastic sampling algorithms to compute conditional mean estimates and MAP state estimates of jump Markov linear models. The computational cost of an iteration of each algorithm is linear in the data length. Convergence results for these algorithms toward required estimates have been obtained. A key property of the two algorithms for MAP state estimation is that they are asymptotically globally convergent. This property is in contrast to gradient type algorithms, such as the EM algorithm [12], which suffer from convergence to stationary points. Two applications (in sparse signal detection/estimation and narrowband interference suppression in CDMA communication systems) were presented to show their performances.

Future work will focus on adaptive recursive versions of the proposed algorithms in narrowband interference suppression, and multiuser detection in CDMA systems.

## APPENDIX I
## FORWARD FILTERING–BACKWARD SAMPLING RECURSIONS

### A. Sampling from $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r})$

We have the following decomposition

$$p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}) = p(x_T|\boldsymbol{y}_T, \boldsymbol{r}) \prod_{t=0}^{T-1} p(x_t|\boldsymbol{y}_t, \boldsymbol{r}, x_{t+1}) \tag{51}$$

where $\boldsymbol{y}_t \triangleq (y_1, \cdots, y_t)$, $\boldsymbol{r}_t \triangleq (r_1, \cdots, r_t)$, and $\boldsymbol{r}_0 = \boldsymbol{y}_0 = \emptyset$. Given $x_{t+1}$, $p(x_t|\boldsymbol{y}_t, \boldsymbol{r}, x_{t+1})$ is a Gaussian distribution. This decomposition suggests the following algorithm [2].

1) *Kalman Filter—Forward Filtering*: Set $m_{0|0} = \hat{x}_0$ and $P_{0|0} = P_0$. Then, for $t = 1, \cdots, T$, compute using the Kalman filter equations

$$m_{t|t-1} = A(r_t)m_{t-1|t-1} + F(r_t)u_t \tag{52}$$
$$P_{t|t-1} = A(r_t)P_{t-1|t-1}A'(r_t) + B(r_t)B'(r_t) \tag{53}$$
$$z_{t|t-1} = C(r_t)m_{t|t-1} + G(r_t)u_t \tag{54}$$
$$S_t = C(r_t)P_{t|t-1}C'(r_t) + D(r_t)D'(r_t) \tag{55}$$
$$m_{t|t} = m_{t|t-1} + P_{t|t-1}C'(r_t)S_t^{-1}(y_t - z_{t|t-1}) \tag{56}$$
$$P_{t|t} = P_{t|t-1} - P_{t|t-1}C'(r_t)S_t^{-1}C(r_t)P_{t|t-1} \tag{57}$$

and store for $t = 1, \cdots, T$, $m_{t|t-1} = \mathbb{E}\{x_t|\boldsymbol{y}_{t-1}, \boldsymbol{r}\}$, $m_{t|t} = \mathbb{E}\{x_t|\boldsymbol{y}_t, \boldsymbol{r}\}$, $P_{t|t-1} = \mathbb{E}\{(x_t - m_{t|t-1})(x_t - m_{t|t-1})'|\boldsymbol{y}_{t-1}, \boldsymbol{r}\}$, and $P_{t|t} = \mathbb{E}\{(x_t - m_{t|t})(x_t - m_{t|t})'|\boldsymbol{y}_t, \boldsymbol{r}\}$.

2) *Backward Sampling*: For $t = T, \cdots, 0$, sample $x_t$ from $\mathcal{N}(m_t^*, P_t^*)$, where $m_T^* = m_{T|T}$, $P_T^* = P_{T|T}$, and for $0 \le t < T$

$$m_t^* = m_{t|t} + P_{t|t}A'(r_{t+1})P_{t+1|t}^{-1}(x_{t+1} - m_{t+1|t}) \tag{58}$$
$$P_t^* = P_{t|t} - P_{t|t}A'(r_{t+1})P_{t+1|t}^{-1}A(r_{t+1})P_{t|t}. \tag{59}$$

### B. Sampling from $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x})$

$p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x})$ can be decomposed as follows:

$$p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}) = p(r_T|\boldsymbol{y}_T, \boldsymbol{x}_T) \prod_{t=1}^{T-1} p(r_t|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{r}_{t+1}), \tag{60}$$

where $\boldsymbol{y}_t \triangleq (y_1, \cdots, y_t)$, $\boldsymbol{x}_t \triangleq (x_0, \cdots, x_t)$. Given $r_{t+1}$, $p(r_t|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{r}_{t+1})$ is a discrete distribution. It suggests the following algorithm to sample from $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x})$.

1) *Optimal Filter—Forward Filtering*: For $t = 2, \cdots, T$, compute the optimal filter [17]. For any $i$, evaluate

$$p(r_t = i|\boldsymbol{y}_{t-1}, \boldsymbol{x}_{t-1}) = \sum_{j=1}^{s} p_{ji}p(r_{t-1} = j|\boldsymbol{y}_{t-1}, \boldsymbol{x}_{t-1}) \tag{61}$$

$$p(r_t = i|\boldsymbol{y}_t, \boldsymbol{x}_t) = \frac{p(y_t, x_t|y_{t-1}, x_{t-1}, r_t = i)p(r_t = i|\boldsymbol{y}_{t-1}, \boldsymbol{x}_{t-1})}{\sum_{j=1}^{s} p(y_t, x_t|y_{t-1}, x_{t-1}, r_t = j)p(r_t = j|\boldsymbol{y}_{t-1}, \boldsymbol{x}_{t-1})} \tag{62}$$

and store for $t = 1, \cdots, T$ and for $i = 1, \cdots, s$ $p(r_t = i|\boldsymbol{y}_{t-1}, \boldsymbol{x}_{t-1})$ and $p(r_t = i|\boldsymbol{y}_t, \boldsymbol{x}_t)$.

2) *Backward Sampling*: Sample $r_T$ from $p(r_T|\boldsymbol{y}_T, \boldsymbol{x})$. Then, for $t = T - 1, \cdots, 1$, sample $r_t$ from $p(r_t|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{r}_{t+1})$, where for $i = 1, \cdots, s$

$$p(r_t = i|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{r}_{t+1}) = \frac{p(r_{t+1}|r_t = i)p(r_t = i|\boldsymbol{y}_t, \boldsymbol{x}_t)}{\sum_{j=1}^{s} p(r_{t+1}|r_t = j)p(r_t = j|\boldsymbol{y}_t, \boldsymbol{x}_t)}. \tag{63}$$

## APPENDIX II
## PROOF OF THEOREMS AND LEMMAS

### A. Proof of Theorem 3.1

The marginal sequence $\{\boldsymbol{r}^{(k)};\ k \in \mathbb{N}\}$ is a Markov chain with transition kernel given by

$$K\left(\boldsymbol{r}^{(k-1)} = \boldsymbol{r}(l);\, \boldsymbol{r}^{(k)} = \boldsymbol{r}(m)\right) = \int f(\boldsymbol{r}(m)|\boldsymbol{y}, \boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}(l))\, d\boldsymbol{x}. \tag{64}$$

By construction [21], [22], $\{(\boldsymbol{r}^{(k)}, \boldsymbol{x}^{(k)});\ k \in \mathbb{N}\}$ admits $f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})$ as its invariant distribution. Thus, $\{\boldsymbol{r}^{(k)};\ k \in \mathbb{N}\}$ admits $f(\boldsymbol{r}|\boldsymbol{y})$ as its invariant distribution. From the model assumptions given in Section II-A and (4) in particular, the posterior distribution $f(\boldsymbol{r}|\boldsymbol{y})$ is strictly positive on $R$ and null on $S^T \backslash R$. Moreover, we have $K(\boldsymbol{r}(l); \boldsymbol{r}(m)) > 0$ for any $(\boldsymbol{r}(l), \boldsymbol{r}(m)) \in S^T \times R$ and $K(\boldsymbol{r}(l); \boldsymbol{r}(m)) = 0$ for any $(\boldsymbol{r}(l), \boldsymbol{r}(m)) \in S^T \times S^T \backslash R$; i.e., at $k = 1$, the finite-state Markov chain enters in $R$ and never leaves this set. Thus, $\{\boldsymbol{r}^{(k)};\ k \in \mathbb{N}\}$ is irreducible and aperiodic on $R$. Hence, it is uniformly ergodic. From [19, pp. 401–402], $\{\boldsymbol{r}^{(k)};\ k \in \mathbb{N}\}$ satisfies

$$\int |p_k(\boldsymbol{r}) - f(\boldsymbol{r}|\boldsymbol{y})|\, d\boldsymbol{r} \leq 2\rho^{k-1}$$

[and $\int |p_k(\boldsymbol{r}) - f(\boldsymbol{r}|\boldsymbol{y})|\, d\boldsymbol{r} \leq 2\rho^k$ if $\boldsymbol{r}^{(0)} \in R$], where $0 \leq \rho < 1$ satisfies

$$\rho = 1 - \sum_{\boldsymbol{r}(m) \in R} \min_{\boldsymbol{r}(l) \in R} K(\boldsymbol{r}(l); \boldsymbol{r}(m)).$$

Other bounds exist in the literature. Applying the *duality principle* of Robert and Diebolt [4], [18], [23], we now show that

$$\int |p_k(\boldsymbol{x}) - f(\boldsymbol{x}|\boldsymbol{y})|\, d\boldsymbol{x} \leq 2\rho^{k-2}. \tag{65}$$

Thus, the property of uniform geometric convergence of the Markov chain $\{\boldsymbol{r}^{(k)};\ k \in \mathbb{N}\}$ is "transferred" to the continuous state-space Markov chain $\{\boldsymbol{x}^{(k)};\ k \in \mathbb{N}\}$. To prove (65), first note that

$$\begin{aligned} p_k(\boldsymbol{r}(l), \boldsymbol{x}') &= f(\boldsymbol{r}(l)|\boldsymbol{x}', \boldsymbol{y}) \int p_{k-1}(\boldsymbol{r}) f(\boldsymbol{x}'|\boldsymbol{r}, \boldsymbol{y})\, d\boldsymbol{r} \\ &= f(\boldsymbol{r}(l)|\boldsymbol{x}', \boldsymbol{y}) p_k(\boldsymbol{x}'). \end{aligned} \tag{66}$$

Thus,

$$\begin{aligned} &\int |p_k(\boldsymbol{x}) - f(\boldsymbol{x}|\boldsymbol{y})|\, d\boldsymbol{x} \\ &= \int \left| \int f(\boldsymbol{x}|\boldsymbol{r}, \boldsymbol{y}) p_{k-1}(\boldsymbol{r})\, d\boldsymbol{r} - \int f(\boldsymbol{x}|\boldsymbol{r}, \boldsymbol{y}) f(\boldsymbol{r}|\boldsymbol{y})\, d\boldsymbol{r} \right| d\boldsymbol{x} \\ &\leq \iint f(\boldsymbol{x}|\boldsymbol{r}, \boldsymbol{y}) |p_{k-1}(\boldsymbol{r}) - f(\boldsymbol{r}|\boldsymbol{y})|\, d\boldsymbol{r}\, d\boldsymbol{x} \\ &= \int |p_{k-1}(\boldsymbol{r}) - f(\boldsymbol{r}|\boldsymbol{y})| d\boldsymbol{r} \\ &\leq 2\rho^{k-2}. \end{aligned} \tag{67}$$

Using (66), we now prove the uniform ergodicity of the Markov chain $\{\boldsymbol{r}^{(k)}, \boldsymbol{x}^k);\ k \in \mathbb{N}\}$

$$\begin{aligned} &\iint |p_k(\boldsymbol{r}, \boldsymbol{x}) - f(\boldsymbol{r}, \boldsymbol{x}|\boldsymbol{y})|\, d\boldsymbol{r}\, d\boldsymbol{x} \\ &= \iint |f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}) p_k(\boldsymbol{x}) - f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{y})|\, d\boldsymbol{r}\, d\boldsymbol{x} \\ &\leq \int f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}) \int |p_k(\boldsymbol{x}) - f(\boldsymbol{x}|\boldsymbol{y})|\, d\boldsymbol{x}\, d\boldsymbol{r} \\ &= \int |p_k(\boldsymbol{x}) - f(\boldsymbol{x}|\boldsymbol{y})|\, d\boldsymbol{x} \\ &\leq 2\rho^{k-2}. \end{aligned} \tag{68}$$

### B. Proof of Lemma 4.1

$$\begin{aligned} &\iint K_k\left((\boldsymbol{r}, \boldsymbol{x}); (\boldsymbol{r}', \boldsymbol{x}')\right) \overline{f}^{1/T(k)}(\boldsymbol{r}, \boldsymbol{x})\, d\boldsymbol{r}\, d\boldsymbol{x} \\ &= \iint \overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y}, \boldsymbol{r}) \overline{f}^{1/T(k)}(\boldsymbol{r}'|\boldsymbol{y}, \boldsymbol{x}') \overline{f}^{1/T(k)} \\ &\quad \cdot (\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}) \overline{f}^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y})\, d\boldsymbol{r}\, d\boldsymbol{x} \\ &= \int \overline{f}^{1/T(k)}(\boldsymbol{x}', \boldsymbol{r}|\boldsymbol{y}) \overline{f}^{1/T(k)}(\boldsymbol{r}'|\boldsymbol{y}, \boldsymbol{x}')\, d\boldsymbol{r} \\ &= \overline{f}^{1/T(k)}(\boldsymbol{r}', \boldsymbol{x}'). \end{aligned}$$

### C. Proof of Lemma 4.2

See (69) at the bottom of the next page, where $n \triangleq n_x(T+1)$. If $\boldsymbol{r}(l) \notin \mathcal{M}$, $\boldsymbol{r}(m) \in \mathcal{M}$ exists such that

$$\lim_{k \to +\infty} \left( \frac{f(\boldsymbol{r}(m)|\boldsymbol{y})\, |\boldsymbol{\Sigma}_m|^{-1/2}}{f(\boldsymbol{r}(l)|\boldsymbol{y})\, |\boldsymbol{\Sigma}_l|^{-1/2}} \right)^{1/T(k)} = +\infty \tag{70}$$

and thus

$$\lim_{k \to +\infty} \overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y}) = 0. \tag{71}$$

If $\boldsymbol{r}(l) \in \mathcal{M}$, then

$$\lim_{k \to +\infty} \overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y}) = \left( \sum_{\boldsymbol{r}(m) \in \mathcal{M}} \frac{|\boldsymbol{\Sigma}_m|^{1/2}}{|\boldsymbol{\Sigma}_l|^{1/2}} \right)^{-1}. \tag{72}$$

### D. Proof of Lemma 4.3

The proof of this result follows from arguments similar to those of Mitra *et al.* [14, Proposition 3.3, pp. 755–756]. For any $\boldsymbol{r}(l) \in R$, by differentiating with respect to $T$ the function $\overline{f}^{1/T}(\boldsymbol{r}(l)|\boldsymbol{y})$, we obtain

$$\begin{aligned} &\frac{d\overline{f}^{1/T}(\boldsymbol{r}(l)|\boldsymbol{y})}{dT} \\ &= \frac{|\boldsymbol{\Sigma}_l|^{1/2} \left( f(\boldsymbol{r}(l)|\boldsymbol{y})\, |\boldsymbol{\Sigma}_l|^{-1/2} \right)^{1/T}}{C^2(T) T^2} \sum_{\boldsymbol{r}(m) \in R} |\boldsymbol{\Sigma}_m|^{1/2} \\ &\quad \cdot \left( f(\boldsymbol{r}(m)|\boldsymbol{y})\, |\boldsymbol{\Sigma}_m|^{-1/2} \right)^{1/T} \ln \left( \frac{f(\boldsymbol{r}(m)|\boldsymbol{y})\, |\boldsymbol{\Sigma}_m|^{-1/2}}{f(\boldsymbol{r}(l)|\boldsymbol{y})\, |\boldsymbol{\Sigma}_l|^{-1/2}} \right) \end{aligned} \tag{73}$$

where

$$C(T)=\sum_{\boldsymbol{r}(m)\in R}|\boldsymbol{\Sigma}_m|^{1/2}\left(f(\boldsymbol{r}(m)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_m|^{-1/2}\right)^{1/T}. \quad (74)$$

From this result, it follows that, for any $\boldsymbol{r}(l)\in\mathcal{M}$ and for $k\geq 0$

$$\overline{f}^{1/T(k+1)}(\boldsymbol{r}(l)|\boldsymbol{y})-\overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})>0 \quad (75)$$

because each term on the right-hand side of (73) is either zero or negative; thus, $d\overline{f}^{1/T}(\boldsymbol{r}(l)|\boldsymbol{y})/dT<0$. From (20), (75) immediately follows.

Let $\mathcal{H}(l)$, for $\boldsymbol{r}(l)\in R\backslash\mathcal{M}$, denote the set of $\boldsymbol{r}(m)$ such that $f(\boldsymbol{r}(m)|\boldsymbol{y})|\boldsymbol{\Sigma}_m|^{-1/2}>f(\boldsymbol{r}(l)|\boldsymbol{y})|\boldsymbol{\Sigma}_l|^{-1/2}$; that is

$$\mathcal{H}(l)=\Big\{\boldsymbol{r}(m)\colon f(\boldsymbol{r}(m)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_m|^{-1/2}>f(\boldsymbol{r}(l)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_l|^{-1/2},\ \boldsymbol{r}(m)\in R,\boldsymbol{r}(l)\in R\backslash\mathcal{M}\Big\}. \quad (76)$$

For $\boldsymbol{r}(l)\in R\backslash\mathcal{M}$

$$\begin{aligned}&\frac{C^2(T)T^2}{\left(|\boldsymbol{\Sigma}_l|^{1/2}\left(f(\boldsymbol{r}(l)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_l|^{-1/2}\right)^{1/T}\right)^2}\frac{d\overline{f}^{1/T}(\boldsymbol{r}(l)|\boldsymbol{y})}{dT}\\&\quad=\sum_{\boldsymbol{r}(m)\in\mathcal{H}(i)}\frac{|\boldsymbol{\Sigma}_m|^{1/2}}{|\boldsymbol{\Sigma}_l|^{1/2}}\left(\frac{f(\boldsymbol{r}(m)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_m|^{-1/2}}{f(\boldsymbol{r}(l)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_l|^{-1/2}}\right)^{1/T}\\&\qquad\cdot\ln\left(\frac{f(\boldsymbol{r}(m)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_m|^{-1/2}}{f(\boldsymbol{r}(l)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_l|^{-1/2}}\right)\\&\qquad-\sum_{\boldsymbol{r}(m)\notin\mathcal{H}(i)}\frac{|\boldsymbol{\Sigma}_m|^{1/2}}{|\boldsymbol{\Sigma}_l|^{1/2}}\left(\frac{f(\boldsymbol{r}(m)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_m|^{-1/2}}{f(\boldsymbol{r}(l)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_l|^{-1/2}}\right)^{1/T}\\&\qquad\cdot\ln\left(\frac{f(\boldsymbol{r}(l)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_l|^{-1/2}}{f(\boldsymbol{r}(m)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_m|^{-1/2}}\right). \quad (77)\end{aligned}$$

Following the same arguments as in [14], we make the following observations: the first term on the right-hand side of (77) is monotonically increasing with decreasing $T$, and the second term on the right-hand side of (77) is monotonically decreasing with decreasing $T$. Furthermore, in the limit $T\to 0$, the first term tends toward $+\infty$, and the second term tends toward zero. Thus, for $T\to 0$, $d\overline{f}^{1/T}(\boldsymbol{r}(l)|\boldsymbol{y})/dT>0$. The set $R\backslash\mathcal{M}$ being finite, a $k'\in\mathbb{N}$ exists such that, for any $k>k'$

$$\overline{f}^{1/T(k+1)}(\boldsymbol{r}(l)|\boldsymbol{y})-\overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})<0,\qquad\text{if }\boldsymbol{r}(l)\in R\backslash\mathcal{M}. \quad (78)$$

Then, following Mitra *et al.* [14, Proposition 5.1, pp. 762–763], from (75) and (78) and for $k>k'$, we have

$$\begin{aligned}&\sum_{\boldsymbol{r}(l)\in R}\left|\overline{f}^{1/T(k+1)}(\boldsymbol{r}(l)|\boldsymbol{y})-\overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})\right|\\&\quad=\sum_{\boldsymbol{r}(l)\in\mathcal{M}}\left(\overline{f}^{1/T(k+1)}(\boldsymbol{r}(l)|\boldsymbol{y})-\overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})\right)\\&\qquad-\sum_{\boldsymbol{r}(l)\in R\backslash\mathcal{M}}\left(\overline{f}^{1/T(k+1)}(\boldsymbol{r}(l)|\boldsymbol{y})-\overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})\right)\\&\quad=2\sum_{\boldsymbol{r}(l)\in\mathcal{M}}\left(\overline{f}^{1/T(k+1)}(\boldsymbol{r}(l)|\boldsymbol{y})-\overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})\right).\end{aligned}$$

Hence

$$\sum_{k=k'}^{+\infty}\sum_{\boldsymbol{r}(l)\in R}\left|\overline{f}^{1/T(k+1)}(\boldsymbol{r}|\boldsymbol{y})-\overline{f}^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y})\right|\leq 2. \quad (79)$$

The result (33) immediately follows.

### *E. Proof of Lemma 4.4*

Using Bayes's rule

$$\begin{aligned}&\overline{f}^{1/T(k)}(\boldsymbol{r}(m)|\boldsymbol{y},\boldsymbol{x}')\\&\quad=\frac{\overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{r}(m))\overline{f}^{1/T(k)}(\boldsymbol{r}(m)|\boldsymbol{y})}{\int\overline{f}^{1/T(k)}(\boldsymbol{x}',\boldsymbol{r}|\boldsymbol{y})\,d\boldsymbol{r}}\\&\quad=\frac{\overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{r}(m))\overline{f}^{1/T(k)}(\boldsymbol{r}(m)|\boldsymbol{y})}{\int\overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{r})\overline{f}^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y})\,d\boldsymbol{r}}.\end{aligned}$$

$k_0\in\mathbb{N}$ exists such that, for any $(\boldsymbol{r}(m'),\boldsymbol{r}(l'))\in\mathcal{M}\times R$, $\overline{f}^{1/T(k)}(\boldsymbol{r}(m')|\boldsymbol{y})\geq\overline{f}^{1/T(k)}(\boldsymbol{r}(l')|\boldsymbol{y})$; thus, the denominator of the expression above can be bounded

$$\begin{aligned}&\int\overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{r})\overline{f}^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y})\,d\boldsymbol{r}\\&\quad\leq\overline{f}^{1/T(k)}(\boldsymbol{r}(m')|\boldsymbol{y})\int\overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{r})\,d\boldsymbol{r}\\&\quad\leq\overline{f}^{1/T(k)}(\boldsymbol{r}(m')|\boldsymbol{y})\overline{R}b_k \quad (80)\end{aligned}$$

---

$$\begin{aligned}\overline{f}^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})&=\frac{\int f^{1/T(k)}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{r}(l))\,d\boldsymbol{x}f^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})}{\iint f^{1/T(k)}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{r})f^{1/T(k)}(\boldsymbol{r}|\boldsymbol{y})\,d\boldsymbol{x}\,d\boldsymbol{r}}\\&=\frac{\left((2\pi)^{n/2}|\boldsymbol{\Sigma}_l|^{1/2}\right)^{1-1/T(k)}(T(k))^{n/2}f^{1/T(k)}(\boldsymbol{r}(l)|\boldsymbol{y})}{\sum_{\boldsymbol{r}(m)\in R}\left((2\pi)^{n/2}|\boldsymbol{\Sigma}_m|^{1/2}\right)^{1-1/T(k)}(T(k))^{n/2}f^{1/T(k)}(\boldsymbol{r}(m)|\boldsymbol{y})}\\&=\left(\sum_{\boldsymbol{r}(m)\in R}\frac{|\boldsymbol{\Sigma}_m|^{1/2}}{|\boldsymbol{\Sigma}_l|^{1/2}}\left(\frac{f(\boldsymbol{r}(m)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_m|^{-1/2}}{f(\boldsymbol{r}(l)|\boldsymbol{y})\,|\boldsymbol{\Sigma}_l|^{-1/2}}\right)^{1/T(k)}\right)^{-1} \quad (69)\end{aligned}$$

where $\overline{R}$ is the cardinality of $R$ and

$$b_k = \max_{\boldsymbol{r}(l')\in R}\left[\overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{r}) = \mathcal{N}(\boldsymbol{m}_{l'}, T(k)\boldsymbol{\Sigma}_{l'})\right] = (2\pi T(k))^{-n/2}\max_{\boldsymbol{r}(l')\in R}|\boldsymbol{\Sigma}_{l'}|^{-1/2}.$$

Thus

$$\begin{aligned}K_k(\boldsymbol{r}(l);\boldsymbol{r}(m)) &= \int \overline{f}^{1/T(k)}(\boldsymbol{r}(m)|\boldsymbol{y},\boldsymbol{x}')\overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{r}(l))\,d\boldsymbol{x}' \\ &\geq \frac{\overline{f}^{1/T(k)}(\boldsymbol{r}(m)|\boldsymbol{y})}{\overline{f}^{1/T(k)}(\boldsymbol{r}(m')|\boldsymbol{y})\overline{R}b_k}\int \overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{r}(m)) \\ &\quad\cdot \overline{f}^{1/T(k)}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{r}(l))\,d\boldsymbol{x}'.\end{aligned}$$

For any $(\boldsymbol{r}(l),\boldsymbol{r}(m))\in R\times R$

$$\int \overline{f}^{1/T(k)}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{r}(l))\overline{f}^{1/T(k)}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{r}(m))\,d\boldsymbol{x} = \int \mathcal{N}(\boldsymbol{m}_l, T(k)\boldsymbol{\Sigma}_l)\mathcal{N}(\boldsymbol{m}_m, T(k)\boldsymbol{\Sigma}_m)\,d\boldsymbol{x}.$$

We obtain

$$\begin{aligned}&\int \overline{f}^{1/T(k)}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{r}(l))\overline{f}^{1/T(k)}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{r}(m))\,d\boldsymbol{x} \\ &= \frac{|\boldsymbol{\Sigma}_{l,m}|^{1/2}}{(2\pi T(k))^{n/2}|\boldsymbol{\Sigma}_l|^{1/2}|\boldsymbol{\Sigma}_m|^{1/2}} \\ &\quad\cdot\exp\left(\frac{-1}{2T(k)}(\boldsymbol{m}_l-\boldsymbol{m}_m)'(\boldsymbol{\Sigma}_l+\boldsymbol{\Sigma}_m)^{-1}(\boldsymbol{m}_l-\boldsymbol{m}_m)\right)\end{aligned}$$

where $\boldsymbol{\Sigma}_{l,m}^{-1} \triangleq \boldsymbol{\Sigma}_l^{-1}+\boldsymbol{\Sigma}_m^{-1}$. Define the following constants:

$$C' = \overline{R}^{-1}\min_{(\boldsymbol{r}(l),\boldsymbol{r}(m))\in R\times R}\left(|\boldsymbol{\Sigma}_{l,m}|^{1/2}\Big/|\boldsymbol{\Sigma}_l\boldsymbol{\Sigma}_m|^{1/2}\right)\Big/\max_{\boldsymbol{r}(l')\in R}|\boldsymbol{\Sigma}_{l'}|^{-1/2} > 0$$

and

$$L_1 = \tfrac{1}{2}\max_{(\boldsymbol{r}(l),\boldsymbol{r}(m))\in R\times R}(\boldsymbol{m}_l-\boldsymbol{m}_m)'\cdot(\boldsymbol{\Sigma}_l+\boldsymbol{\Sigma}_m)^{-1}(\boldsymbol{m}_l-\boldsymbol{m}_m).$$

Then

$$K_k(\boldsymbol{r}(l);\boldsymbol{r}(m)) \geq C'\exp\left(-\frac{L_1}{T(k)}\right)\frac{\overline{f}^{1/T(k)}(\boldsymbol{r}(m)|\boldsymbol{y})}{\overline{f}^{1/T(k)}(\boldsymbol{r}(m')|\boldsymbol{y})}.$$

Using (69), we have

$$\frac{\overline{f}^{1/T(k)}(\boldsymbol{r}(m)|\boldsymbol{y})}{\overline{f}^{1/T(k)}(\boldsymbol{r}(m')|\boldsymbol{y})} = \left(\frac{|\boldsymbol{\Sigma}_m|}{|\boldsymbol{\Sigma}_{m'}|}\right)^{1/2}\left[\frac{f(\boldsymbol{r}(m)|\boldsymbol{y})|\boldsymbol{\Sigma}_m|^{-1/2}}{f(\boldsymbol{r}(m')|\boldsymbol{y})|\boldsymbol{\Sigma}_{m'}|^{-1/2}}\right]^{1/T(k)}.$$

Thus, with

$$C = C'\min_{(\boldsymbol{r}(l),\boldsymbol{r}(m))\in R\times R}\left(\frac{|\boldsymbol{\Sigma}_m|}{|\boldsymbol{\Sigma}_l|}\right)^{1/2}$$

$$L_2 = \max_{(\boldsymbol{r}(l),\boldsymbol{r}(m))\in R\times R}\ln\left(\frac{f(\boldsymbol{r}(m)|\boldsymbol{y})|\boldsymbol{\Sigma}_m|^{-1/2}}{f(\boldsymbol{r}(l)|\boldsymbol{y})|\boldsymbol{\Sigma}_l|^{-1/2}}\right)$$

and

$$L = L_1 + L_2 \tag{81}$$

we obtain the result.

### F. Proof of Theorem 4.1

We first prove that $\{\boldsymbol{r}^{(k)};\, k\in\mathbb{N}\}$ is weakly ergodic ([8, Definition V.1.1, p. 137], [14, Definition 4.4, p. 758]). The inhomogeneous Markov chain $\{K_k;\, k\in\mathbb{N}^*\}$ is weakly ergodic if $\sum_{k=1}^{\infty}[1-\zeta(K_k)] = +\infty$, where the ergodic coefficient $\zeta(P)$ of any Markov transition kernel (stochastic matrix) $P$ on $R$ is defined as

$$\zeta(P) = 1 - \min_{(\boldsymbol{r}(l),\boldsymbol{r}(m))\in R\times R}\sum_{\boldsymbol{r}(k)\in R}\cdot\min(P(\boldsymbol{r}(l);\boldsymbol{r}(k)),\, P(\boldsymbol{r}(m);\boldsymbol{r}(k))). \tag{82}$$

Using Lemma 4.4, the ergodic coefficient of $K_k$ is given by

$$\begin{aligned}\zeta(K_k) &= 1 - \min_{(\boldsymbol{r}(l),\boldsymbol{r}(m))\in R\times R}\sum_{\boldsymbol{r}(l')\in R}\cdot\min\left(K_k(\boldsymbol{r}(l);\boldsymbol{r}(l')),\, K_k(\boldsymbol{r}(m);\boldsymbol{r}(l'))\right) \\ &\leq 1 - \min_{(\boldsymbol{r}(l),\boldsymbol{r}(m))\in R\times R}K_k(\boldsymbol{r}(l);\boldsymbol{r}(m)) \\ &\leq 1 - C\exp\left(-\frac{L}{T(k)}\right).\end{aligned} \tag{83}$$

Thus, if $T(k) = \gamma/\ln(k+u)$

$$\sum_{k=k_0}^{+\infty}[1-\zeta(K_k)] \geq C\sum_{k=k_0}^{+\infty}\left(\frac{1}{k+u}\right)^{L/\gamma}.$$

The ergodic coefficient $\zeta(K_k)$ diverges if $\gamma \geq L$, implying the weak ergodicity of $\{\boldsymbol{r}(k);\, k\in\mathbb{N}\}$. Now, using Lemma 4.3 and [8, Theorem V.4.3, p. 160] (see also [14, Theorem 4.2, p. 759]), we obtain that $\{\boldsymbol{r}(k);\, k\in\mathbb{N}\}$ is strongly ergodic and (30) is thus satisfied.

To prove the result in (31), we make use of the following which implies a weak convergence result

$$\begin{aligned}&\int |p_{k+1}(\boldsymbol{x}) - f_{k+1}(\boldsymbol{x})|\,d\boldsymbol{x} \\ &= \int\left|\int \overline{f}^{1/T(k+1)}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{r})\left(p_k(\boldsymbol{r}) - \overline{f}^{\infty}(\boldsymbol{r}|\boldsymbol{y})\right)d\boldsymbol{r}\right|d\boldsymbol{x} \\ &\leq \iint \left|p_k(\boldsymbol{r}) - \overline{f}^{\infty}(\boldsymbol{r}|\boldsymbol{y})\right|\overline{f}^{1/T(k+1)}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{r})\,d\boldsymbol{r}\,d\boldsymbol{x} \\ &\leq \int\left|p_k(\boldsymbol{r}) - \overline{f}^{\infty}(\boldsymbol{r}|\boldsymbol{y})\right|d\boldsymbol{r}.\end{aligned} \tag{84}$$

From (30) and (84), (31) immediately follows.

### G. Proof of Lemma 5.1

Our algorithm is nothing but a Metropolis–Hastings algorithm of invariant distribution $\overline{f}^{1/T(k)}(\boldsymbol{r})$ and proposal distribution $K_h(\boldsymbol{r}(l); \boldsymbol{r}(m))$ [23]. The simple expression of the acceptance ratio (38) follows from the fact that $K_h(\boldsymbol{r}(l); \boldsymbol{r}(m))$ is in detailed balance with $f(\boldsymbol{r}(m)|\boldsymbol{y})$; i.e., $f(\boldsymbol{r}(l))K_h(\boldsymbol{r}(l); \boldsymbol{r}(m)) = f(\boldsymbol{r}(m))K_h(\boldsymbol{r}(m); \boldsymbol{r}(l))$.

### H. Proof of Lemma 5.4

We denote $\mathcal{N}$ as the set of global minima of $f(\boldsymbol{r}(m)|\boldsymbol{y})$ on $R$.

- For all $\boldsymbol{r}(l) \in R\backslash\mathcal{N}$, $\boldsymbol{r}(m) \in R$ exists such that $f(\boldsymbol{r}(m)|\boldsymbol{y}) < f(\boldsymbol{r}(l)|\boldsymbol{y})$. A finite $k_1$ exists such that, for any $k \geq k_1$, $1/T(k) \geq 1$, and for any $\boldsymbol{r}(m) \in R\backslash\{\boldsymbol{r}(l)\}$, we have

$$\begin{aligned} &K_k(\boldsymbol{r}(l); \boldsymbol{r}(m)) \\ &\quad = \alpha_k(\boldsymbol{r}(l), \boldsymbol{r}(m))K_h(\boldsymbol{r}(l); \boldsymbol{r}(m)) \\ &\quad \geq \left[\min_{\boldsymbol{r}(m)\in R} \frac{f(\boldsymbol{r}(m)|\boldsymbol{y})}{f(\boldsymbol{r}(l)|\boldsymbol{y})}\right]^{1/T(k)-1} \\ &\qquad \cdot \min_{\boldsymbol{r}(m)\in R\backslash\{\boldsymbol{r}(l)\}} K_h(\boldsymbol{r}(l); \boldsymbol{r}(m)) \\ &\quad \geq C\exp\left(-\frac{L}{T(k)}\right) \end{aligned}$$

where

$$L = \max_{(\boldsymbol{r}(l), \boldsymbol{r}(m))\in R\times R} \ln \frac{f(\boldsymbol{r}(l)|\boldsymbol{y})}{f(\boldsymbol{r}(m)|\boldsymbol{y})} \tag{85}$$

and

$$C = \exp(L) \times \min_{\boldsymbol{r}(l)\in R\backslash\mathcal{M}^*} \min_{\boldsymbol{r}(m)\in R\backslash\{\boldsymbol{r}(l)\}} K_h(\boldsymbol{r}(l); \boldsymbol{r}(m)).$$

- If $\boldsymbol{r}(m) = \boldsymbol{r}(l)$,

$$\begin{aligned} &K_k(\boldsymbol{r}(l); \boldsymbol{r}(l)) \\ &\quad = K_h(\boldsymbol{r}(l); \boldsymbol{r}(l)) + \sum_{\boldsymbol{r}(m)\in R\backslash\{\boldsymbol{r}(l)\}} \\ &\qquad \cdot (1 - \alpha_k(\boldsymbol{r}(l), \boldsymbol{r}(m)))K_h(\boldsymbol{r}(l); \boldsymbol{r}(m)) \\ &\quad = K_h(\boldsymbol{r}(l); \boldsymbol{r}(l)) + \sum_{\boldsymbol{r}(m)\in(R\backslash\{\boldsymbol{r}(l)\})/f(\boldsymbol{r}(m)|\boldsymbol{y})<f(\boldsymbol{r}(l)|\boldsymbol{y})} \\ &\qquad \cdot \left(1 - \left(\frac{f(\boldsymbol{r}(m)|\boldsymbol{y})}{f(\boldsymbol{r}(l)|\boldsymbol{y})}\right)^{1/T(k)-1}\right) K_h(\boldsymbol{r}(l); \boldsymbol{r}(m)). \end{aligned}$$

As $k$ increases, $K_k(\boldsymbol{r}(l); \boldsymbol{r}(l))$ monotonically increases as

$$\left(\frac{f(\boldsymbol{r}(m)|\boldsymbol{y})}{f(\boldsymbol{r}(l)|\boldsymbol{y})}\right)^{1/T(k)-1}$$

monotonically decreases. Thus, a finite $k_2 \geq k_1$ exists such that, for any $k \geq k_2, K_k(\boldsymbol{r}(l); \boldsymbol{r}(l)) \geq C\exp(-(L/T(k)))$.

- If $\boldsymbol{r}(l) \in \mathcal{M}^*$, then, for any $\boldsymbol{r}(m) \in R\backslash\{\boldsymbol{r}(l)\}$, we have $\alpha_k(\boldsymbol{r}(l), \boldsymbol{r}(m)) = 1$ and

$$K_k(\boldsymbol{r}(l); \boldsymbol{r}(m)) = K_h(\boldsymbol{r}(l); \boldsymbol{r}(m)) > 0.$$

Hence, a finite $k_3$ exists such that, for any $k \geq k_3$, $K_k(\boldsymbol{r}(l); \boldsymbol{r}(m)) \geq C\exp(-(L/T(k)))$. Now, choosing $k_0 = \max(k_2, k_3)$, the result follows.

### I. Proof of Theorem 5.1

The first part is similar to the proof of Theorem 4.1 and is thus omitted.

Equation (43) is obtained from the following result by taking the limit as $k$ goes to infinity

$$\begin{aligned} &\int \left|p_{k+1}(\boldsymbol{x}) - \overline{f}^{\infty}(\boldsymbol{x})\right| d\boldsymbol{x} \\ &\quad = \int \left|\int f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r})p_k(\boldsymbol{r}) - f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r})\overline{f}^{\infty}(\boldsymbol{r})\, d\boldsymbol{r}\right| d\boldsymbol{x} \\ &\quad = \int f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r}) \left|\int p_k(\boldsymbol{r}) - \overline{f}^{\infty}(\boldsymbol{r})\, d\boldsymbol{r}\right| d\boldsymbol{x} \\ &\quad \leq \int \left|p_k(\boldsymbol{r}) - \overline{f}^{\infty}(\boldsymbol{r})\right| d\boldsymbol{r} \int f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{r})\, d\boldsymbol{x} \\ &\quad \leq \int \left|p_k(\boldsymbol{r}) - \overline{f}^{\infty}(\boldsymbol{r})\right| d\boldsymbol{r}. \end{aligned}$$


## ACKNOWLEDGMENT

The authors wish to dedicate this paper to Prof. S. Beeman, DSTO, Australia—their friend, philosopher, and guide.


## REFERENCES


[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

[2] C. K. Carter and R. Kohn, "On Gibbs sampling for state space models," *Biometrika*, vol. 81, pp. 541–553, 1994.

[3] P. DeJong and N. Shephard, "The simulation smoother for time series models," *Biometrika*, vol. 82, pp. 339–350, 1995.

[4] J. Diebolt and C. P. Robert, "Estimation of finite mixture distributions through Bayesian sampling," *J. R. Stat. Soc. B*, vol. 56, pp. 363–375, 1994.

[5] S. Früwirth-Schnatter, "Data augmentation and dynamic linear models," *J. Time Ser. Anal.*, vol. 15, pp. 183–202, 1994.

[6] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images," *IEEE Trans. Pattern. Anal. Machine Intell.*, vol. PAMI-6, pp. 721–741, 1984.

[7] R. E. Helmick, W. D. Blair, and S. A. Hoffman, "Fixed-interval smoothing for Markovian switching systems," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1845–1855, June 1995.

[8] D. L. Isaacson and R. W. Madsen, *Markov Chains Theory and Applications*. New York: Wiley, 1976.

[9] V. Krishnamurthy and A. Logothetis, "Adaptive nonlinear filters for narrowband interference suppression in spread spectrum CDMA," *IEEE Trans. Commun.*, vol. 47, May 1999.

[10] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.

[11] J. S. Liu, W. H. Wong, and A. Kong, "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes," *Biometrika*, vol. 81, pp. 27–40, 1994.

[12] A. Logothetis and V. Krishnamurthy, "Expectation maximization algorithms for MAP estimation of jump Markov linear systems," IEEE Signal Processing, Aug. 1999, to be published.

[13] J. M. Mendel, *Maximum Likelihood Deconvolution*. New York: Springer, 1990.

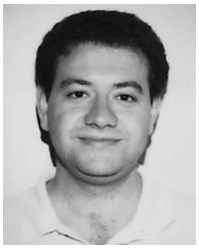
[14] D. Mitra, F. Romeo, and A. Sangiovanni-Vincetelli, "Convergence and finite-time behavior of simulated annealing," *Adv. Appl. Prob.*, vol. 18, pp. 747–771, 1986.

[15] H. V. Poor and X. Wang, "Code-aided interference suppression for DS-CDMA communications—Part I: Interference suppression capability," *IEEE Trans. Commun.*, vol. 45, pp. 1101–1111, Sept. 1997.

[16] ——, "Code-aided interference suppression for DS-CDMA communications—Part II: Parallel blind adaptive implementation," *IEEE Trans. Commun.*, vol. 45, pp. 1112–1122, Sept. 1997.
[17] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.
[18] C. P. Robert, G. Celeux, and J. Diebolt, "Bayesian estimation of hidden Markov models: A stochastic implementation," *Stat. Prob. Lett.*, vol. 16, pp. 77–83, 1993.
[19] J. S. Rosenthal, "Convergence rates of Markov chains," *SIAM Rev.*, vol. 37, pp. 387–415, 1995.
[20] A. N. Shiryaev, *Probability*. New York: Springer-Verlag, 1996.
[21] M. A. Tanner and W. A. Wong, "The calculation of posterior distributions by data augmentation," *J. Am. Stat. Assoc.*, vol. 82, pp. 528–540, 1987.
[22] M. A. Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer-Verlag, 1993.
[23] L. Tierney, "Markov chains for exploring posterior distributions," *Ann. Stat.*, vol. 22, pp. 1701–1762, 1994.
[24] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.
[25] J. K. Tugnait, "Adaptive estimation and identification for discrete systems with Markov jump parameters," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 1054–1065, 1982.
[26] M. West and J. F. Harrison, *Bayesian Forecasting and Dynamic Models*, ser. Springer Series in Statistics, 1996.
[27] P. J. Van Laarhoven and E. H. L. Arts, *Simulated Annealing: Theory and Applications*. Amsterdam, The Netherlands: Reidel, 1987.
[28] L. B. Milstein, "Interference rejection techniques in spread spectrum communications," *Proc. IEEE*, vol. 76, pp. 657–671, June 1988.
[29] L. A. Rusch and H. V. Poor, "Narrowband interference suppression in CDMA spread spectrum communications," *Trans. Commun.*, vol. 42, pp. 1969–1979, 1994.

**Arnaud Doucet** was born in Melle, France, on November 2, 1970. He received the M.S. degree from the Institut National des Télécommunications, France, in 1993, and the Ph.D. degree from the University of Paris, Orsay, France, in 1997.

He was a Visiting Researcher in the Signal Processing Group of Cambridge University, Cambridge, U.K., sponsored by a postdoctoral grant from the French Ministry of Defence. He has been a Research Associate in the same group since 1999. He was sponsored by an EPSRC grant.



**Andrew Logothetis** received the doctoral degree from the University of Melbourne, Parkville, Australia, in 1998.

He is currently a Senior Research Engineer at Bandspeed, Melbourne, Australia. In 1998, he was a Research Fellow at the Department of Signals, Sensors, and Systems, Royal Institute of Technology, Sweden. His research interests include statistical signal processing in communication systems.

**Vikram Krishnamurthy** was born in 1966. He received the bachelor's degree in electrical engineering from the University of Auckland, New Zealand, in 1988 and the doctoral degree from the Australian National University, Canberra, in 1992.

He is currently an Associate Professor and Deputy Head of the Department of Electrical Engineering, University of Melbourne, Australia. His research interests include time-series analysis, hidden Markov models, stochastic filtering theory, and statistical signal processing in communication systems.

Dr. Krishnamurthy is currently an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING and *Systems and Control Letters*.