Convergence of Expectation-Maximization Algorithm with Mixed-Integer Optimization

Joseph, Geethu

DOI
10.1109/LSP.2024.3393352

Publication date
2024

Document Version
Final published version

Published in
IEEE Signal Processing Letters

Convergence of Expectation-Maximization Algorithm With Mixed-Integer Optimization

Geethu Joseph, *Member, IEEE*

***Abstract*—The convergence of expectation-maximization (EM)-based algorithms typically requires continuity of the likelihood function with respect to all the unknown parameters (optimization variables). The requirement is not met when parameters comprise both discrete and continuous variables, making the convergence analysis nontrivial. This paper introduces a set of conditions that ensure the convergence of a specific class of EM algorithms that estimate a mixture of discrete and continuous parameters. Our results offer a new analysis technique for iterative algorithms that solve mixed-integer non-linear optimization problems. As a concrete example, we prove the convergence of an existing EM-based sparse Bayesian learning algorithm that estimates the state of a linear dynamical system with jointly sparse inputs and bursty missing observations. Our results establish that the algorithm converges to the set of stationary points of the maximum likelihood cost with respect to the continuous optimization variables.**

***Index Terms*—Discrete non-linear optimization, global convergence theorem, sparse Bayesian learning, bursty missing data.**

I. Introduction

THE Expectation-Maximization (EM) algorithm is a general technique for maximum likelihood or maximum a posteriori estimation [1], [2]. It is a crucial ingredient in well-known algorithms like Baum-Welch [3], inside-outside [4], sparse Bayesian learning (SBL) [5], and their numerous variants. EM's popularity is due to its simplicity, stability (monotonic increase in likelihood), and convergence guarantees for many statistical problems. The convergence analysis of EM, presented in [6], establishes conditions under which EM converges to a stationary point of the likelihood function. The literature also offers convergence analyses for specific cases, such as EM for Gaussian mixtures [7] and EM with squared iterative methods [8]. However, these analyses assume that the likelihood function is continuous in all unknown hyperparameters, implying that the parameters belong to an open set and are all continuous. This assumption may not hold in general. For example, the EM-based SBL framework [9] is used to estimate sparse state of a linear dynamical with missing observations, where the unknown state vector belongs to an open set, while the unknown missing data status is discrete (either missing or not missing). Some motivating applications for this problem include identifying missing data indices in occlusions due to nonlinear energy harvesting or environmental factors in wireless networks (motion tracking [10], network traffic reconstruction [11], localization refinement [12], urban traffic sensing improvement [13], and structural health monitoring [14], [15]), satellite imaging systems [16], [17], and downlink channel estimation via feedback through a bursty channel [9]. Despite demonstrating good empirical performance of the EM algorithm [9], its convergence becomes nontrivial due to discrete parameters, posing a challenge to existing EM convergence results. Insipred by this setting, we study the convergence of the EM algorithm with both continuous and discrete hyperparameters.

The contributions of this paper are twofold. First, we relax the assumption that the set of unknowns estimated by EM is purely continuous, allowing for a general set that comprises both continuous and discrete parameters. We derive mild conditions ensuring the convergence of EM to a stationary point of the likelihood function. Notably, when the unknowns belong to an open set, our results reduce to those of [6]. Second, we apply these results to establish the convergence of the EM-based SBL algorithm presented in [9].

II. Statistical Model and Convergence Result

Consider the statistical model which generates observations $\boldsymbol{Y}$, unobserved latent data $\boldsymbol{X}$, and unknown parameters $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$. We assume that a part of the parameter is continuous and the other part is discrete and finite, i.e., $\boldsymbol{\Theta} \subseteq \mathbb{G} \times \mathbb{H}$ and

$$\boldsymbol{\theta}^* = \begin{bmatrix} \boldsymbol{\gamma}^\mathsf{T} & \boldsymbol{\alpha}^\mathsf{T} \end{bmatrix}^\mathsf{T}, \tag{1}$$

where $\boldsymbol{\gamma} \in \mathbb{G}$ and $\boldsymbol{\alpha} \in \mathbb{H}$. Here, $\mathbb{G} \subseteq \mathbb{R}^N$ is a set without isolated points, and $\mathbb{H} \subseteq \mathbb{R}^K$ is a finite (countable) set. For the ease of exposition, we use $\boldsymbol{\theta}$ and $(\boldsymbol{\gamma}, \boldsymbol{\alpha})$ interchangeably.

Let $p(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta}^*)$ be the joint distribution of data $(\boldsymbol{Y}, \boldsymbol{X})$ conditioned on the parameter $\boldsymbol{\theta}^*$. Then, the maximum likelihood (ML) estimate of the unknown $\boldsymbol{\theta}^*$ is

$$\arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}), \tag{2}$$

where $L(\boldsymbol{\theta})$ is the likelihood function given by

$$L(\boldsymbol{\theta}) = \log p(\boldsymbol{Y}; \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{X}} \{ \log p(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta}) \}. \tag{3}$$

If the above optimization is not tractable, we can use the EM algorithm to solve (2). The EM algorithm is an iterative algorithm, with each iteration comprising an expectation step (E-step) and a maximization step (M-step). Let $\boldsymbol{\theta}^{(r)} = [\boldsymbol{\gamma}^{(r)\mathsf{T}} \quad \boldsymbol{\alpha}^{(r)\mathsf{T}}]^\mathsf{T}$ be the EM iterate in the $(r-1)$th iteration. In the rth iteration, the E-step computes the expected log-likelihood of $\boldsymbol{\theta}$ with respect to the distribution of $\boldsymbol{X}$ conditioned on the current iterate $\boldsymbol{\theta}^{(r)}$ and observations $\boldsymbol{Y}$,

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) = \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\theta}^{(r)}} \{ \log p(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta}) \}. \tag{4}$$

Manuscript received 31 January 2024; revised 7 April 2024; accepted 17 April 2024. Date of publication 24 April 2024; date of current version 2 May 2024. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Guang Hua.

The author is with the Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: g.joseph@tudelft.nl).

Digital Object Identifier 10.1109/LSP.2024.3393352

The M-step maximizes this function with respect to $\boldsymbol{\theta}$ to obtain the new iterate $\boldsymbol{\theta}^{(r+1)}$, i.e.,

$$\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right). \tag{5}$$

So, the r-th EM iteration can be summarized as a mapping $\mathcal{G}:\boldsymbol{\Theta}\to\boldsymbol{\Theta}$, i.e.,

$$\boldsymbol{\theta}^{(r+1)} = \mathcal{G}\left(\boldsymbol{\gamma}^{(r)},\boldsymbol{\alpha}^{(r)}\right) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}\right). \tag{6}$$

Next, we present the main result of the section, which provides a list of sufficient conditions for the EM algorithm to converge to a stationary point of the ML cost function in (2).

Theorem 1: Let $\{\boldsymbol{\theta}^{(r)}\}_{r=0}^{\infty}$ be the sequence generated by the EM algorithm, as summarized in (6), to solve the ML optimization problem in (2). Assume the following conditions,

1) There exists a constant $C\in\mathbb{R}$ such that $L(\boldsymbol{\theta}^{(0)})\le C$, for any $\boldsymbol{\theta}^{(0)}\in\boldsymbol{\Theta}$.
2) The level set $\{\boldsymbol{\theta}: L(\boldsymbol{\theta})\ge L(\boldsymbol{\theta}^{(r)})\}$ is compact, for any integer $r>0$.
3) For any $\boldsymbol{\alpha}_*\in\mathbb{H}$, the iteration mapping $\mathcal{G}(\boldsymbol{\gamma},\boldsymbol{\alpha}_*)$ in (6) is closed at all values of $[\boldsymbol{\gamma}^{\mathsf{T}} \quad \boldsymbol{\alpha}_*^{\mathsf{T}}]^{\mathsf{T}}\in\boldsymbol{\Theta}$.
4) The function $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)})$ is a continuous function of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^{(r)}$, for a fixed value of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^{(r)}$, where $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(r)}$ take the form (1) and (6), respectively.

Then, the sequence of iterates $\{\boldsymbol{\theta}^{(r)}\}_{r=1}^{\infty}$ converges to a subset of $\mathcal{S}_*$ over which $L(\boldsymbol{\theta})$ is a constant. Here,

$$\mathcal{S}_* = \{\boldsymbol{\theta}\in\boldsymbol{\Theta}: \boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\gamma}\in\mathbb{R}^N & \boldsymbol{\alpha}\in\mathbb{H}\end{bmatrix} \text{ and } \nabla_{\boldsymbol{\gamma}} L(\boldsymbol{\theta}) = \mathbf{0}\}, \tag{7}$$

where $\nabla_{\boldsymbol{\gamma}}$ denote the gradient with respect to $\boldsymbol{\gamma}$.

Proof: See Appendix A. □

Here, Conditions 1 and 2 refer to the likelihood, and Conditions 3 and 4 are linked to the iterate update procedure. Conditions 1, 2, and 4 are similar to (8), (6), and (7) in [6], respectively. Condition 3 is required for the global convergence theorem [18] to hold. Further, our analysis generalizes the EM convergence result in [6]. Using a similar proof technique, we can extend the other results in [6] to our setting (for example, the results on generalized EM). Furthermore, if the stationary points of the likelihood cost are isolated, the algorithm converges to a single point. Also, we assume that $\mathbb{H}$ is finite, such as any bounded subset of integers or rational numbers. There are estimation problems where this assumption holds, and we give an example in the next section.

III. Analysis of Kalman SBL for State Estimation of a Linear System With Missing Outputs

In this section, we prove convergence guarantees to the EM-based SBL algorithm in [9] using Theorem 1. The algorithm aims to estimate the states of a linear dynamical system with jointly sparse inputs and missing observations at unknown time instants [9]. Specifically, we consider a discrete-time linear dynamical system given by

$$\boldsymbol{x}_k = \boldsymbol{D}\boldsymbol{x}_{k-1} + \boldsymbol{u}_k \quad \text{and} \quad \boldsymbol{y}_k = \boldsymbol{\alpha}_k^*\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{w}_k. \tag{8}$$

Here, $\boldsymbol{x}_k\in\mathbb{R}^n$, $\boldsymbol{u}_k\in\mathbb{R}^n$, and $\boldsymbol{y}_k\in\mathbb{R}^m$ are the state, input, and observation at time k, respectively. Also, $\boldsymbol{w}_k\in\mathbb{R}^m$ is the zero-mean Gaussian distributed measurement noise at time k whose variance is σ^2. Also, $\boldsymbol{D}\in\mathbb{R}^{n\times n}$ is the state transition matrix, and $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ is the output matrix. The inputs are jointly sparse, and the initial state $\boldsymbol{x}_0=\mathbf{0}\in\mathbb{R}^n$. Further, $\boldsymbol{\alpha}_k^*\in\{0,1\}$ represents whether the signal part $\boldsymbol{A}\boldsymbol{x}_k$ is missing or not in the observation $\boldsymbol{y}_k$. The missing data indicator $\boldsymbol{\alpha}_k^*$ for $k=1,2,\ldots$ follows a hidden Markov model: for any integer $k>0$ and $i,j\in\{0,1\}$, we have $\mathbb{P}\{\boldsymbol{\alpha}_k^*=i|\boldsymbol{\alpha}_{k-1}^*=j\}=p_j$, if $i=j$. For a given integer value of $K<\infty$, the algorithm aims to estimate the state matrix $\boldsymbol{X}$ using the output matrix $\boldsymbol{Y}$ when $\boldsymbol{\alpha}^*$ is unknown, where we define

$$\boldsymbol{X} \triangleq \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_K\end{bmatrix}\in\mathbb{R}^{N\times K} \tag{9}$$

$$\boldsymbol{Y} \triangleq \begin{bmatrix}\boldsymbol{y}_1 & \boldsymbol{y}_2 & \ldots & \boldsymbol{y}_K\end{bmatrix}\in\mathbb{R}^{m\times K} \tag{10}$$

$$\boldsymbol{\alpha}^* \triangleq \begin{bmatrix}\boldsymbol{\alpha}_1^* & \boldsymbol{\alpha}_2^* & \ldots & \boldsymbol{\alpha}_K^*\end{bmatrix}^{\mathsf{T}}\in\{0,1\}^K. \tag{11}$$

The SBL framework assumes a fictitious Gaussian prior on the sparse vectors with a common matrix with $\boldsymbol{\gamma}^*\in\mathbb{R}_+^N$ along the diagonal, i.e., $p(\boldsymbol{x}_k|\boldsymbol{\gamma}^*)=\mathcal{N}(\mathbf{0},\mathrm{diag}\{\boldsymbol{\gamma}^*\})$. Then, we jointly estimate $\boldsymbol{\theta}^*=[\boldsymbol{\gamma}^* \quad \boldsymbol{\alpha}^*]\in\mathbb{R}_+^N\times\{0,1\}^K$. The resulting Q function in the M-step is

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(r)}) = \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\theta}^{(r)}}\{\log\left[p\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\alpha}\right)p\left(\boldsymbol{\alpha}\right)p\left(\boldsymbol{X},\boldsymbol{\gamma}\right)\right]\} \tag{12}$$

$$= \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\theta}^{(r)}}\{\log p\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\alpha}\right)\} + \log p\left(\boldsymbol{\alpha}\right) + \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\theta}^{(r)}}\{\log p\left(\boldsymbol{X},\boldsymbol{\gamma}\right)\}. \tag{13}$$

So, the optimization problem in the M-step is separable in the $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. Since the optimization is separable, the mapping $\mathcal{G}$ in (6) can be decomposed as follows:

$$\begin{bmatrix}\boldsymbol{\gamma}^{(r+1)}\\ \boldsymbol{\alpha}^{(r+1)}\end{bmatrix}\in\mathcal{G}(\boldsymbol{\gamma}^{(r)},\boldsymbol{\alpha}^{(r)}) = \begin{bmatrix}\mathcal{G}_{\gamma}\left(\boldsymbol{\gamma}^{(r)},\boldsymbol{\alpha}^{(r)}\right)\subset\mathbb{R}_+^n\\ \mathcal{G}_{\alpha}\left(\boldsymbol{\gamma}^{(r)},\boldsymbol{\alpha}^{(r)}\right)\subset\mathbb{H}\end{bmatrix}. \tag{14}$$

The resulting algorithm uses the Kalman smoothing to compute $\mathcal{G}_{\gamma}(\boldsymbol{\gamma}^{(r)},\boldsymbol{\alpha}^{(r)})$ and Viterbi algorithm to $\mathcal{G}_{\alpha}(\boldsymbol{\gamma}^{(r)},\boldsymbol{\alpha}^{(r)})$.

Further, the set $\mathcal{G}_{\gamma}(\boldsymbol{\gamma}^{(r)},\boldsymbol{\alpha}^{(r)})$ is a singleton set whereas $\mathcal{G}_{\alpha}(\boldsymbol{\gamma}^{(r)},\boldsymbol{\alpha}^{(r)})$ need not be. Also, for a given $\boldsymbol{\alpha}^{(r)}$, the mapping $\mathcal{G}(\boldsymbol{\gamma}^{(r)},\boldsymbol{\alpha}^{(r)})$ is continuous in $\boldsymbol{\gamma}^{(r)}$. Please refer to [9] for more details. Using the above properties of the algorithm, we derive the convergence results.

We start with the optimization problem (2) that the algorithm solves. To compute $L(\boldsymbol{\theta})=\log p(\boldsymbol{Y};\boldsymbol{\theta})$, we note that $\boldsymbol{Y}$ is Gaussian distributed with zero mean. Given $\boldsymbol{\theta}$, we derive $\boldsymbol{y}_k=\boldsymbol{\alpha}_k\boldsymbol{A}\sum_{j=1}^k\boldsymbol{D}^{k-j}\boldsymbol{u}_j+\boldsymbol{w}_k$ from (8). Subsequently, the covariance matrix $\boldsymbol{R}_Y(\boldsymbol{\theta})\in\mathbb{R}^{Km\times Km}$ of the vectorized version of $\boldsymbol{Y}$ is

$$\boldsymbol{R}_Y(\boldsymbol{\theta}) = \left(\mathrm{diag}\{\boldsymbol{\alpha}\}\otimes\boldsymbol{A}\right)\tilde{\boldsymbol{D}} \times\left(\boldsymbol{I}\otimes\mathrm{diag}\{\boldsymbol{\gamma}\}\right)\tilde{\boldsymbol{D}}^{\mathsf{T}}\left(\mathrm{diag}\{\boldsymbol{\alpha}\}\otimes\boldsymbol{A}^{\mathsf{T}}\right), \tag{15}$$

where $\mathrm{diag}\{\cdot\}$ represents a diagonal matrix with entries of the argument vector along the diagonal and $\tilde{\boldsymbol{D}}\in\mathbb{R}^{KN\times KN}$ as

$$\tilde{\boldsymbol{D}} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0} & \ldots & \boldsymbol{0}\\ \boldsymbol{D} & \boldsymbol{I} & \ldots & \boldsymbol{0}\\ \vdots & & \ddots & \\ \boldsymbol{D}^{K-1} & \boldsymbol{D}^{K-2} & \ldots & \boldsymbol{I}\end{bmatrix}. \tag{16}$$

By simplifying (2) using (15), the objective function of an optimization problem equivalent to (2) reduces to

$$L(\boldsymbol{\theta}) = -\frac{Km}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{R}_Y(\boldsymbol{\theta})+\sigma^2\boldsymbol{I}\right| - \frac{1}{2}\boldsymbol{y}^{\mathsf{T}}\left(\boldsymbol{R}_Y(\boldsymbol{\theta})+\sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}. \tag{17}$$

With the objective function defined, we are now ready to state the convergence result.

Theorem 2: Suppose that the noise variance $\sigma^2 > 0$. Let the iterates generated by the Bayesian state estimation algorithm in [9] be $\boldsymbol{\theta}^{(r)} \in \mathbb{R}^N \times \mathbb{H}$. Then, the sequence of iterates $\{\boldsymbol{\theta}^{(r)}\}_{r=1}^{\infty}$ converges to a subset of $\mathcal{S}_*$ over which $L(\boldsymbol{\theta})$ is a constant. Here,

$$\mathcal{S}_* = \{\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\gamma} \in \mathbb{R}^N & \boldsymbol{\alpha} \in \mathbb{H}\end{bmatrix} : \nabla_{\boldsymbol{\gamma}} L(\boldsymbol{\theta}) = \mathbf{0}\}, \tag{18}$$

where $\nabla_{\boldsymbol{\gamma}}$ denote the gradient with respect to $\boldsymbol{\gamma}$.

Proof: See Appendix B. □

Hence, our generalized result in Theorem 1 guarantees that the SBL variant in [9] with discrete parameters converges to a stationary point of the ML cost function.

IV. Conclusion

We derived the conditions for the convergence of the EM algorithm with discrete unknown parameters. As an illustration, we demonstrated the convergence of the EM-based SBL algorithm outlined in [9], proving its convergence to the set of stationary points of the maximum likelihood cost. Extending the results to the generalized class of Majorization-Minimization algorithms is an interesting future work.

Appendix A
Proof of Theorem 1

Our proof is adapted from the proofs of EM algorithm convergence in [6] and Zangwill's convergence theorem [18]. Our proof relies on some properties of the algorithm iterates as listed by the following preliminary lemmas:

Lemma 3 ([19, Theorem 8.1]): The EM algorithm formulation guarantees the following from (2):

$$L(\boldsymbol{\theta}) = Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) + g\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right), \tag{19}$$

where Q is defined in (4) and we define

$$g\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}\right) = \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\theta}^{(r)}}\{-\log p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\theta})\}. \tag{20}$$

Lemma 4 (Gibbs' inequality [20]): For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the function g defined in (20) satisfies $g(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) \geq g(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})$.

Lemma 5: The sequence of function values $\{L(\boldsymbol{\theta}^{(r)}),\ r \geq 1\}$ defined in (2) is monotonically non-decreasing and convergent.

Proof: From (6), in every iteration r,

$$Q\left(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}\right) \geq Q\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}\right). \tag{21}$$

From Lemmas 3, 4, and (21), $L(\boldsymbol{\theta}^{(r+1)}) - L(\boldsymbol{\theta}^{(r)}) \geq 0$. Consequently, the sequence $\{L(\boldsymbol{\theta}^{(r)}), r \geq 1\}$ is monotonically non-decreasing, and is bounded from above by Assumption 1. Thus, it convergences to a single point. □

Lemma 6: If $\boldsymbol{\theta}^{(r)} \notin \mathcal{S}^*$ for some $r > 0$, then we have the relation $L(\boldsymbol{\theta}^{(r+1}; \boldsymbol{\theta}^{(r)}) > L(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})$.

Proof: Using Lemma 3, we get

$$\nabla_{\boldsymbol{\gamma}} Q\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}\right) + \nabla_{\boldsymbol{\gamma}} g\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}\right) \neq \mathbf{0}. \tag{22}$$

Also, from Lemma 4, we have

$$g\left(\begin{bmatrix}\boldsymbol{\gamma}^{\mathsf{T}} & \boldsymbol{\alpha}^{(r)\mathsf{T}}\end{bmatrix}^{\mathsf{T}}; \boldsymbol{\theta}^{(r)}\right) \geq g\left(\begin{bmatrix}\boldsymbol{\gamma}^{(r)\mathsf{T}} & \boldsymbol{\alpha}^{(r)\mathsf{T}}\end{bmatrix}^{\mathsf{T}}; \boldsymbol{\theta}^{(r)}\right). \tag{23}$$

Hence, $\nabla_{\boldsymbol{\gamma}} g(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}) = \mathbf{0}$, and as a result, from (22), we have $\nabla_{\boldsymbol{\gamma}} Q(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}) \neq \mathbf{0}$. So, $\boldsymbol{\theta}^{(r)}$ is not a local $Q(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})$. Therefore, from the definition of the M-step update in (6), we conclude that

$$Q\left(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}\right) > Q\left(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)}\right). \tag{24}$$

Now, we arrive at the desired result by Lemmas 3 and 4. □

Proof of Theorem 1

Let $\boldsymbol{\theta}_* = \begin{bmatrix}\boldsymbol{\gamma}_*^{\mathsf{T}} & \boldsymbol{\alpha}_*^{\mathsf{T}} \in \mathbb{H}\end{bmatrix}^{\mathsf{T}} \in \boldsymbol{\Theta}$ be a limit point of the sequence $\{\boldsymbol{\theta}^{(r)}, r \geq 1\}$. From Lemma 5 and Assumption 2, we know that there exists a subsequence $\{\boldsymbol{\theta}^{(r_j)}, j \geq 1\}$ of $\{\boldsymbol{\theta}^{(r)}, r \geq 1\}$ such that

$$\lim_{j\to\infty} \boldsymbol{\theta}^{(r_j)} = \boldsymbol{\theta}_* = \begin{bmatrix}\boldsymbol{\gamma}_*^{\mathsf{T}} & \boldsymbol{\alpha}_*^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}. \tag{25}$$

We next construct another subsequence $\{\boldsymbol{\theta}^{(r_j+1)},\ j \geq 1\}$, which also belongs a compact set due to Assumption 2. Hence, the new sequence contains a convergent subsequence $\{\boldsymbol{\theta}^{(r_{j_l}+1)},\ l \geq 1\}$ for which there exists $\hat{\boldsymbol{\theta}}$ such that

$$\lim_{l\to\infty} \boldsymbol{\theta}^{(r_{j_l}+1)} = \hat{\boldsymbol{\theta}}. \tag{26}$$

From (6), we get

$$\boldsymbol{\theta}^{(r_{j_l}+1)} \in \mathcal{G}\left(\boldsymbol{\gamma}^{(r_{j_l})}, \boldsymbol{\alpha}^{(r_{j_l})}\right). \tag{27}$$

Here, by construction, $\{\boldsymbol{\alpha}^{(r_{j_l})}, l \geq 1\}$ is a subsequence of the convergent sequence $\{\boldsymbol{\alpha}^{(r)}, r \geq 1\}$. So it converges to $\boldsymbol{\alpha}_*$, and since the subsequence belongs to a finite set, there exists $L > 0$ such that

$$\boldsymbol{\theta}^{(r_{j_l}+1)} \in \mathcal{G}(\boldsymbol{\gamma}^{(r_{j_l})}, \boldsymbol{\alpha}_*),\quad \forall\, l > L. \tag{28}$$

Therefore, by Assumption 3 and (26), we arrive at

$$\hat{\boldsymbol{\theta}} \in \mathcal{G}\left(\lim_{l\to\infty} \boldsymbol{\gamma}^{(r_{j_l})}, \boldsymbol{\alpha}_*\right) = \mathcal{G}(\boldsymbol{\gamma}_*, \boldsymbol{\alpha}_*), \tag{29}$$

due to (25), where $\mathcal{G}(\boldsymbol{\gamma}_*, \boldsymbol{\alpha}_*)$ is the next iterate of the EM algorithm if the current iterate is $\boldsymbol{\theta}_*$. The relation (29) implies

$$L(\boldsymbol{\theta}_*) = L(\hat{\boldsymbol{\theta}}) = L\left(\mathcal{G}(\boldsymbol{\gamma}_*, \boldsymbol{\alpha}_*)\right), \tag{30}$$

due to the convergence of the sequence $\{L(\boldsymbol{\theta}^{(r)}), r \geq 1\}$ as ensured by Lemma 5.

Furthermore, by Lemma 6, if $\boldsymbol{\theta}_* \notin \mathcal{S}^*$, then, $L(\boldsymbol{\theta}_*) > L(\mathcal{G}(\boldsymbol{\gamma}_*, \boldsymbol{\alpha}_*))$. Hence, (30) holds only if $\boldsymbol{\theta}_* \in \mathcal{S}^*$. Finally, by Lemma 5, $L(\boldsymbol{\theta})$ is a constant over the subset of $\mathcal{S}^*$ to which the iterates converge, and the proof is completed.

Appendix B
Proof of Theorem 2

The proof verifies the four assumptions of Theorem 2 hold for the algorithm. We need the following supporting lemmas.

Lemma 7 ([21, Theorem 2.11]): Let $\mathcal{S}$ be an unbounded subset of $\mathbb{R}^n$, and $f : \mathcal{S} \to \mathbb{R}$ is a continuous function, then f is

said to be coercive (i.e., $\lim_{\|\boldsymbol{s}\|\to\infty} f(\boldsymbol{s}) = \infty$) if and only if all of its level sets are compact.

Lemma 8 ([22, Theorem 5.19, 5.20]): The determinant and inverse of a matrix are continuous in its elements.

Lemma 9: The function $L([\boldsymbol{\gamma}^\mathsf{T} \quad \boldsymbol{\alpha}^\mathsf{T}]^\mathsf{T})$ is continuous and coercive with respect to $\boldsymbol{\gamma}$.

Proof: By Lemma 8, the determinant and inverse of a matrix are continuous in its elements. Therefore, from (15) and (17), $L([\boldsymbol{\gamma}^\mathsf{T} \quad \boldsymbol{\alpha}^\mathsf{T}]^\mathsf{T})$ is a continuous function of $\boldsymbol{\gamma}$. Further, $\|\boldsymbol{\gamma}\|$ goes to ∞ if and only if at least one entry of $\boldsymbol{\gamma}$ goes to ∞. As a result, from (15), we get

$$\lim_{\|\boldsymbol{\gamma}\|\to\infty} \boldsymbol{y}^\mathsf{T} \left(\boldsymbol{R}_Y\boldsymbol{\theta} + \sigma^2\boldsymbol{I}\right)^{-1} \boldsymbol{y} = \boldsymbol{0}. \tag{31}$$

This relation leads to the following,

$$\lim_{\|\boldsymbol{\gamma}\|\to\infty} L([\boldsymbol{\gamma}^\mathsf{T} \quad \boldsymbol{\alpha}^\mathsf{T}]^\mathsf{T}) = \lim_{\|\boldsymbol{\gamma}\|\to\infty} \log\left|\boldsymbol{R}_Y\boldsymbol{\theta} + \sigma^2\boldsymbol{I}\right| = \infty. \tag{32}$$

Hence, we conclude that $L([\boldsymbol{\gamma}^\mathsf{T} \quad \boldsymbol{\alpha}^\mathsf{T}]^\mathsf{T})$ is a coercive function of $\boldsymbol{\gamma}$, and the proof is complete. □

Lemma 10 ([23, Theorem 4.2.1]): Let $(\mathcal{X}, \tau)$ be a topological space and the functions $q_1, q_2, q_3, \ldots, q_p : \mathcal{X} \to \mathbb{R}$ be continuous, for some finite $p > 0$. Then, the function $q_{\max}$ defined as

$$q_{\max} = \max\{q_1, q_2, \ldots, q_p\} \tag{33}$$

is continuous.

Proof of Theorem 2

We start with verifying the first assumption of Theorem 1. To this end, we notice that covariance matrix $\boldsymbol{R}_Y(\boldsymbol{\theta})$ is positive semidefinite, and thus, we have

$$\log\left|\boldsymbol{R}_Y(\boldsymbol{\theta}) + \sigma^2\boldsymbol{I}\right| \geq \log\left|\sigma^2\boldsymbol{I}\right| \tag{34}$$

$$\boldsymbol{y}^\mathsf{T} \left(\boldsymbol{R}_Y(\boldsymbol{\theta}) + \sigma^2\boldsymbol{I}\right)^{-1} \boldsymbol{y} > 0, \tag{35}$$

for any $\boldsymbol{y} \in \mathbb{R}^{Km}$. Therefore, (17) leads to the following:

$$L(\boldsymbol{\theta}) \leq -\frac{Km}{2}\log\left(2\pi\sigma^2\right), \tag{36}$$

and Assumption 1 holds.

To verify the second assumption of Theorem 2, we note that $\mathbb{H}$ is a compact set, and it is sufficient to show that $\boldsymbol{\gamma}^{(r)}$ belongs to a compact set. For this, we define

$$\mathcal{S} \triangleq \Big\{\boldsymbol{\gamma} \in \mathbb{R}^n : \ \exists\, \boldsymbol{\alpha} \in \mathbb{H} \text{ such that } L\left(\left[\boldsymbol{\gamma}^\mathsf{T} \quad \boldsymbol{\alpha}^\mathsf{T}\right]^\mathsf{T}\right) \geq L\left(\boldsymbol{\theta}^{(0)}\right)\Big\}. \tag{37}$$

Also, by Lemma 5, $L(\boldsymbol{\theta}^{(r)}) \leq L(\boldsymbol{\theta}^{(0)})$, for any integer $r > 0$. Therefore, $\boldsymbol{\gamma}^{(r)} \in \mathcal{S}$. Consequently, it is enough to show that $\mathcal{S}$ is a compact set. To this end, we rewrite $\mathcal{S}$ as a finite union of level sets of L,

$$\mathcal{S} = \bigcup_{\boldsymbol{\alpha}\in\mathbb{H}} \left\{\boldsymbol{\gamma} \in \mathbb{R}^n : \ L\left(\left[\boldsymbol{\gamma}^\mathsf{T} \quad \boldsymbol{\alpha}^\mathsf{T}\right]^\mathsf{T}\right) \geq L\left(\boldsymbol{\theta}^{(0)}\right)\right\}. \tag{38}$$

Since a finite union of compact sets is compact, we need to show that the level sets of $L([\boldsymbol{\gamma}^\mathsf{T} \quad \boldsymbol{\alpha}^\mathsf{T}]^\mathsf{T})$ for a fixed value of $\boldsymbol{\alpha}$ are compact. Invoking Lemma 7, these level sets are compact if $L([\boldsymbol{\gamma}^\mathsf{T} \quad \boldsymbol{\alpha}^\mathsf{T}]^\mathsf{T})$ is continuous and coercive with respect to $\boldsymbol{\gamma}$. Then, by Lemma 9, Assumption 2 holds.

To check the third assumption of Theorem 2, we verify if the following holds:

$$\lim_{r\to\infty} \mathcal{G}\left(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*\right) = \mathcal{G}\left(\lim_{r\to\infty} \boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*\right), \tag{39}$$

when the limits exist. For this, we recall from (14) that $\mathcal{G}_\gamma(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*)$ is a singleton set and continuous in $\boldsymbol{\gamma}^{(r)}$. So,

$$\lim_{r\to\infty} \mathcal{G}_\gamma\left(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*\right) = \mathcal{G}_\gamma\left(\lim_{r\to\infty} \boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*\right). \tag{40}$$

We next complete the proof by establishing that the limit $\lim_{r\to\infty} \mathcal{G}_\alpha(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*) = \mathcal{G}_\alpha(\lim_{r\to\infty} \boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*)$. For this, we consider the sequence $\{Q_{\max}(\boldsymbol{\gamma}^{(r)})\}_{r=1}^\infty$ where

$$Q_{\max}\left(\boldsymbol{\gamma}^{(r)}\right) = \max_{\boldsymbol{\alpha}\in\{0,1\}^K} q(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}), \tag{41}$$

where $q(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}) = Q([\boldsymbol{\gamma}^{(r)\mathsf{T}} \quad \boldsymbol{\alpha}^\mathsf{T}]^\mathsf{T}; [\boldsymbol{\gamma}^{(r)\mathsf{T}} \quad \boldsymbol{\alpha}_*^\mathsf{T}]^\mathsf{T})$ from (6). We notice that $\{0,1\}^K$ is a finite set. Also, since $p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta})$ and $p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta}^{(r)})$ are Gaussian, $q(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha})$ is a continuous function of $\boldsymbol{\gamma}^{(r)}$. Then, invoking Lemma 10, we obtain that $Q_{\max}(\boldsymbol{\gamma}^{(r)})$ is a continuous function of $\boldsymbol{\gamma}^{(r)}$. Therefore, with $\mathbb{1}$ being the indicator function, we derive that for any $\boldsymbol{\alpha}$,

$$\begin{aligned}
&\lim_{r\to\infty} \mathbb{1}\left\{\boldsymbol{\alpha} \in \mathcal{G}_\alpha(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*)\right\} \\
&\quad= \lim_{r\to\infty} \mathbb{1}\left\{q(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}) = Q_{\max}(\boldsymbol{\gamma}^{(r)})\right\} &(42)\\
&\quad= \mathbb{1}\left\{q\left(\lim_{r\to\infty} \boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}\right) = Q_{\max}\left(\lim_{r\to\infty} \boldsymbol{\gamma}^{(r)}\right)\right\} &(43)\\
&\quad= \mathbb{1}\left\{\boldsymbol{\alpha} \in \mathcal{G}_\alpha\left(\lim_{r\to\infty} \boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*\right)\right\}, &(44)
\end{aligned}$$

where (43) uses the continuity of q and $Q_{\max}$. As a result,

$$\begin{aligned}
&\lim_{r\to\infty} \mathcal{G}_\alpha(\boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*) \\
&= \left\{\boldsymbol{\alpha} \in \{0,1\}^K : \mathbb{1}\left\{\boldsymbol{\alpha} \in \mathcal{G}_\alpha\left(\lim_{r\to\infty} \boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*\right)\right\} = 1\right\} &(45)\\
&= \mathcal{G}_\alpha\left(\lim_{r\to\infty} \boldsymbol{\gamma}^{(r)}, \boldsymbol{\alpha}_*\right). &(46)
\end{aligned}$$

Hence, Assumption 3 holds.

Finally, we verify the fourth assumption of Theorem 1. The dependence of $\boldsymbol{\gamma}^{(r)}$ and $\boldsymbol{\gamma}$ on Q is via distributions $p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta})$ and $p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta}^{(r)})$, which are Gaussian. The function is computed using Kalman smoothing that involves only continuous functions of $\text{diag}\{\boldsymbol{\gamma}\}$ and $\text{diag}\{\boldsymbol{\gamma}^{(r)}\}$. Thus, Assumption 4 holds, and the proof is complete.

REFERENCES

[1] Z. Ghahramani, "Function approximation via density estimation using an EM approach," *MIT Cognit. Comput. Sci. Tech. Rep.*, vol. 9304, 1993.

[2] Z. Ghahramani and S. Roweis, "Learning nonlinear dynamical systems using an EM algorithm," in *Proc. 11th Int. Conf. Neural Inf. Process. Syst.*, 1998, pp. 431–437.

[3] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. statist.*, vol. 41, no. 1, pp. 164–171, Feb. 1970.

[4] K. Lari and S. J. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm," *Comput. Speech Lang.*, vol. 4, no. 1, pp. 35–56, Jan. 1990.

[5] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[6] C. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, pp. 95–103, Mar. 1983.
[7] L. Xu and M. I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixtures," *Neural Comput.*, vol. 8, no. 1, pp. 129–151, Jan. 1996.
[8] R. Varadhan and C. Roland, "Simple and globally convergent methods for accelerating the convergence of any EM algorithm," *Scand. J. Stat.*, vol. 35, no. 2, pp. 335–353, Jan. 2008.
[9] G. Joseph and P. K. Varshney, "State estimation of linear systems with sparse inputs and Markov-modulated missing outputs," in *Proc. Eur. Signal Process. Conf.*, 2022, pp. 837–841.
[10] H. Song, T. Liu, X. Luo, and G. Wang, "Feedback based sparse recovery for motion tracking in RF sensor networks," in *Proc. IEEE Int. Conf. Netw. Architecture Storage*, 2011, pp. 203–207.
[11] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices (extended version)," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, pp. 662–676, Jun. 2012.
[12] S. Rallapalli, L. Qiu, Y. Zhang, and Y.-C. Chen, "Exploiting temporal stability and low-rank structure for localization in mobile networks," in *Proc. Int. Conf. Mobile Comput. Netw.*, 2010, pp. 161–172.
[13] Z. Li, Y. Zhu, H. Zhu, and M. Li, "Compressive sensing approach to urban traffic sensing," in *Proc. Int. Conf. Distrib. Comput. Syst.*, 2011, pp. 889–898.
[14] S. Ji, Y. Sun, and J. Shen, "A method of data recovery based on compressive sensing in wireless structural health monitoring," *Math. Problem Eng.*, vol. 2014, pp. 1–9, Jan. 2014.
[15] V. S. G. Thadikemalla and A. S. Gandhi, "A simple and efficient data loss recovery technique for SHM applications," *Smart Mater. Struct.*, vol. 20, no. 1, pp. 35–42, Jul. 2017.
[16] M. Carlavan, L. Blanc-Féraud, M. Antonini, C. Thiebaut, C. Latry, and Y. Bobichon, "A satellite imaging chain based on the compressed sensing technique," in *Proc. On-Board Payload Data Compress Workshop*, 2012.
[17] S. Scalise, H. Ernst, and G. Harles, "Measurement and modeling of the land mobile satellite channel at Ku-band," *IEEE Trans. Veh. Technol.*, vol. 57, no. 2, pp. 693–703, Mar. 2008.
[18] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*. Upper Saddle River, NJ: Prentice-Hall, Englewood Cliffs, 1969.
[19] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 2019.
[20] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
[21] V. Arnăutu and P. Neittaanmäki, *Optimal Control From Theory to Computer Programs*. Berlin, Germany: Springer, 2003, vol. 111.
[22] J. R. Schott, *Matrix Analysis for Statistics*. Hoboken, NJ, USA: Wiley, 2016.
[23] R. S. Strichartz, *The Way of Analysis*. Burlington, MA, USA: Jones & Bartlett Learn., 2000.