Convergence of Online Identification for Mixed Linear Regression with Two Components

Yujing Liu, Zhixin Liu, *Member, IEEE* and Lei Guo, *Fellow, IEEE*

***Abstract*—This paper investigates the online identification and data clustering problems for mixed linear regression (MLR) model with two components, including the symmetric MLR, and the asymmetric MLR with the balanced mixture. Two corresponding new online identification algorithms are introduced based on the expectation-maximization (EM) and least-squares (LS) principles. It is shown that both algorithms will converge to the true parameter set for any non-zero initial value without resorting to the traditional i.i.d data assumptions. The main challenge in our investigation lies in the fact that the gradient of the likelihood function does not have a unique zero, and a key step in our analysis is to establish the stability of the corresponding differential equation in order to apply the celebrated Ljung's ODE method. It is also shown that the within-cluster error and the probability that the new data is categorized into the correct cluster are asymptotically the same as those in the case of known parameters. Finally, numerical simulations are provided to verify the effectiveness of our online algorithms.**

***Index Terms*—Mixed linear regression, online identification, non-i.i.d data, convergence, data clustering**

I. Introduction

Mixed linear regression (MLR) has been extensively studied in the fields of system identification, statistical learning, and computer science due to its convenience and effectiveness in capturing the nonlinearity of uncertain system dynamics. It was first proposed as a generalization of switching regressions [1] and has found numerous applications including trajectory clustering [2], health care analysis [3], face recognition [4], drug sensitivity prediction [5] and the relationship between genes and disease phenotype [6]. In MLR, each input-output data belongs to one of the uncertain linear regression models or sub-models but we don't know which sub-model it comes from, i.e., the label of data is unknown to us. We note that the MLR model is closely related to several models used in the investigation of control theory. For example, the piecewise affine model is widely studied (cf., [7]–[9]), where the mixture laws depend on the system states rather than on random processes as in the MLR. The bilinear systems can be closely related to MLR provided that the input signal switches over a finite set, as in the case of bang-bang control (cf., [10]). Moreover, the switched linear systems are widely used in adaptive identification and control (cf., [11]–[16]), and notably, MLR can be regarded as a class of such systems in which both the true parameter set and the data label sequence are unknown (cf., [15], [16]).

For learning and prediction of MLR, how to construct algorithms based on the observed data to identify the unknown parameters and to categorize the newly observed data into correct clusters is of fundamental importance. Due to the coupling between parameter estimates and data label estimates, the MLR identification problem is proven to be NP-hard if there is no assumption imposed on the properties of the observed data [17]. Nonetheless, it still attracts much attention from researchers in diverse fields under additional assumptions such as Gaussian and the independent and identically distributed (i.i.d) assumptions on the data. The commonly used methods include the tensor-based method (cf., [18]–[20]), the optimization-based method [21] and the expectation-maximization (EM) method [22]. In the tensor-based method [23], an efficient spectral decomposition of the observed tensor matrix is needed so that the subspace spanned by true parameters is included in the subspace spanned by eigenvectors of the tensor matrix. By grid searching with a sufficiently small grid resolution in this subspace, the exact recovery guarantee for the MLR problem is given (see e.g., [17], [24]) , but this method suffers from high sample complexity and high computational complexity. In the optimization-based method, minimizing the non-convex mean square error of the MLR problem can be converted into the optimization of some objective functions with nice properties such as convexity or smoothness [21]. However, solving the optimization problem is essentially to optimize a nuclear norm function with linear constraints [25] or to solve a mixed integer programming problem [26], both of which may lead to high computational cost. The EM algorithm [22], including E-step and M-step, is a general technique to estimate unknown parameters with hidden random variables. The E-step is used to evaluate the expectation of the log-likelihood function for the complete data set based on the current parameter estimate, while the M-step is used to update the estimate by solving the corresponding maximization problem. Compared with the other two methods, the lower computational cost of EM makes it more attractive to solve MLR problems in practice (cf., [2], [27]).

In the theoretical aspect, there has been remarkable progress

This work was supported by Natural Science Foundation of China under Grants T2293770 and 12288201.

Yujing Liu, Zhixin Liu and Lei Guo are with the State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Mathematical Science, University of Chinese Academy of Sciences, Beijing 100049, China. (E-mails: liuyujing@amss.ac.cn, lzx@amss.ac.cn, Lguo@amss.ac.cn).

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3649290

on solving the MLR identification problems by using the EM algorithm. In the symmetric MLR case, the mirror symmetry prior information can be used to simplify the design and analysis of EM algorithms (cf., [28]). For example, Balakrishnan et al. [29] studied the population EM algorithm and obtained local convergence results under the assumption that the regressor is i.i.d with a standard Gaussian distribution, and later Klusowski et al. [30] proved a larger basin of attraction for local convergence. Kwon et al. [31] established the convergence to the true parameter set of the population EM for any non-zero initial value by verifying that both the angle and distance between the estimate and the true parameter are decreasing using Stein's Gaussian lemma under the i.i.d data assumptions, and thus overcame the difficulty of non-unique optimal parameters for MLR problems. For EM algorithm with finite number of samples, convergence results in the probability sense can also be established (cf., [31]). In the case where there is no symmetric assumptions on the MLR model, the local convergence of the population EM algorithm is established by verifying the convexity of a small neighborhood of the true parameters under the i.i.d Gaussian data assumption (cf., [21], [30], [32]). Later, Zilber et al. proposed a novel EM-type algorithm in [33] and provided an upper bound on the estimation error with arbitrary initialization.

To summarize, all the above-mentioned theoretical investigations have several common features. Firstly, the regressors are required to be i.i.d Gaussian, which is hard to be satisfied in many important situations, especially in stochastic uncertain systems with feedback control [34]. Some studies are devoted to relaxing the i.i.d standard Gaussian assumption on the data, allowing the regressor to follow a Gaussian distribution but with general covariances [24] or a general continuous distribution [20]. However, while these studies relax the distribution assumption, they do not relax the independence assumption on the data. Secondly, the computational algorithms are of off-line character. In fact, the off-line population EM algorithm is used in most previous investigations, which requires infinite number of samples at each iteration. Although the EM algorithm with a finite number of samples has been proposed (cf., [29], [31]), the computational approach still remains off-line and the convergence results are derived in a high probability sense only. In contrast to off-line algorithm, the online algorithm is desirable in many practical situations, which is updated conveniently based on both the current estimate and new input-output data, without requiring storage of all the old data and with lower computational cost. Thirdly, there are few global convergence results to the true parameter set in general (cf., [30], [32], [33]). The only exception is the symmetric MLR problem where such a convergence result is established for the population EM algorithm with the i.i.d data assumption. For the asymmetric MLR problem, only local convergence around the true parameters has been obtained and there is currently no theoretical guarantee for global convergence to the true parameter set, even when adopting the population EM algorithm under i.i.d data Gaussian assumptions. How to establish such a convergence result for EM algorithms without relying on i.i.d assumptions still remains to be an open problem.

In this paper, we consider the online identification and data clustering problems for MLR with two components. The main contributions of this paper can be summarized as follows: *Firstly,* different from the existing off-line EM algorithms used in MLR identification problems (cf., [29]–[33]), we propose an online EM algorithm to estimate the unknown parameters of the symmetric MLR, which alternates between computing the probability that the new data belongs to each sub-model and updating the parameter estimates based on the current estimate and the new observation. Building on this algorithm and the least-squares (LS) principle, we then devise a two-step online identification algorithm to estimate the unknown parameters of the asymmetric MLR with the balanced mixture. *Secondly,* by adopting Ljung's ODE method [35], we transfer the convergence analysis of proposed stochastic recursive algorithms to the stability analysis of deterministic ordinary differential equations (ODEs) with multiple equilibria. Furthermore, by making efforts to establish the stability of the corresponding ODEs, for the first time, we are able to establish the global convergence of the proposed algorithms to the true parameter set without requiring the widely-used i.i.d data assumptions in the previous studies (cf., [29]–[33]). *Finally,* based on the proposed online identification algorithms, we prove that the data clustering performance, including the within-cluster error and the probability that the new data can be categorized into the correct cluster, can asymptotically achieve the same performance as those in the case where true parameters are known.

The remainder of this paper is organized as follows: Section II presents the problem formulation. In Section III, we propose our online EM algorithms. Sections IV states the main results on the convergence of parameter identification and the performance of data clustering algorithms for MLR problems. Sections V gives the proofs of the main results. Section VI provides numerical simulations to verify the effectiveness of our algorithms. Finally, we conclude the paper in Section VII.

II. Problem Formulation

A. Basic Notations

In the sequel, $v \in \mathbb{R}^d$ is a d-dimensional column vector, v^τ and $\|v\|$ are its transpose and Euclidean norm, respectively. For a matrix $A \in \mathbb{R}^{d\times d}$, $\|A\|$ is its spectral norm and $tr(A)$ is its trace. For a symmetric matrix $P \in \mathbb{R}^{d\times d}$, the maximum and minimum eigenvalues are denoted as $\lambda_{\max}(P)$ and $\lambda_{\min}(P)$, respectively. For two matrices A and B, $A > (\geq) B$ means that $A - B$ is a positive (semi)-definite matrix.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, where Ω is the sample space, the σ-algebra $\mathcal{F}$ on Ω is a family of events and P is a probability measure on $(\Omega, \mathcal{F})$. For an event $\mathcal{A} \in \mathcal{F}$, its complement $\mathcal{A}^c$ is defined by $\mathcal{A}^c = \Omega - \mathcal{A}$. The indicator function $\mathbb{I}_{\mathcal{A}}$ on Ω is defined by $\mathbb{I}_{\mathcal{A}} = 1$ if the event $\mathcal{A}$ occurs and $\mathbb{I}_{\mathcal{A}} = 0$ otherwise. If $P(\mathcal{A}) = 1$, then it is said that the event $\mathcal{A}$ occurs almost surely (a.s.). An infinite sequence of events $\{\mathcal{A}_k, k \geq 1\}$ is said to happen infinitely often (i.o.) if $\mathcal{A}_k$ happens for an infinite number of indices $k \in \{1, 2, \cdots\}$. Moreover, a sequence of random variables $\{x_k, k \geq 0\}$ is called uniformly integrable (u.i.) if

$\lim\limits_{a\to\infty}\sup_{k\geq 1}\int_{[|x_k|>a]}|x_k|dP=0$. We use $\mathbb{E}[\cdot]$ to denote the mathematical expectation operator, and $\mathbb{E}[\cdot|\mathcal{F}_k]$ to represent the conditional expectation operator given $\mathcal{F}_k$, where $\{\mathcal{F}_k\}$ is a non-decreasing sequence of σ-algebras. According to convention, $x\sim F$ indicates that the random variable x obeys the distribution F and $\mathcal{N}(\mu,\sigma^2)$ is the Gaussian distribution with μ and σ^2 being the mean and the variance, respectively.

Definition 1: A sequence of random variables $\{x_k, k\geq 1\}$ is said to be asymptotically stationary if for any $\epsilon>0$, and any set $C\in\mathcal{B}^\infty$ with $\mathcal{B}^\infty$ being the Borel set of $\mathbb{R}^\infty$, there exists $K>0$ such that for all $k\geq K$,

$$|P(\{x_k,x_{k+1},\cdots\}\in C)-P(\{x_{k+1},x_{k+2},\cdots\}\in C)|\leq\epsilon.$$

It is further ergodic if $\lim\limits_{k\to\infty}\mathbb{E}\|x_k\|$ exists and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}x_k=\lim_{k\to\infty}\mathbb{E}x_k,\ \text{a.s.}$$

In the above definition, if both parameters ϵ, K can take value 0, then $\{x_k,k\geq 1\}$ is called stationary and ergodic, which is consistent with the traditional definition as in [36].

B. Problem Statement

Consider the following mixed linear regression (MLR) model consisting of two sub-models:

$$y_{k+1}=\begin{cases}\beta_1^{*\tau}\phi_k+w_{k+1}, & \text{if} \quad z_k=1,\\ \beta_2^{*\tau}\phi_k+w_{k+1}, & \text{if} \quad z_k=-1,\end{cases} \tag{1}$$

where β_1^* and β_2^* are unknown parameter vectors in $\mathbb{R}^d$ that determine the sub-models, $\phi_k\in\mathbb{R}^d$, $y_{k+1}\in\mathbb{R}$ and $w_{k+1}\in\mathbb{R}$ are the regressor vector, observation and the system noise. In addition, $z_k\in\{-1,1\}$ is a hidden variable, namely, we do not know which sub-model the data $\{\phi_k,y_{k+1}\}$ comes from.

Remark 1: It is worth mentioning that the MLR model is said to be symmetric if the true parameters satisfy $\beta_1^*=-\beta_2^*$ and the phase retrieval model [28] is such a case. The MLR model is said to be balanced if the hidden variable z_k has equal probabilities, i.e., $P(z_k=1)=P(z_k=-1)$. One can find that some real-world situations, e.g., the classical human perception of tones can be mathematically formulated as the model (1) with a balanced mixture (cf., [27], [37]).

The aim of this paper is to develop online algorithms to simultaneously estimate the true parameters β_1^* and β_2^* by using the streaming data $\{\phi_k,y_{k+1}\}_{k=1}^\infty$, and establish convergence results for the identification algorithms. Based on the estimates of β_1^* and β_2^*, we further investigate the probability and performance that the newly generated data can be categorized into the correct cluster.

III. Online EM Algorithms

In this section, we design online identification algorithms based on the likelihood method for both symmetric and asymmetric MLR models. We note that the symmetric MLR case will bring benefits for the design and theoretical analysis of the algorithm and will pave a way for the study of the asymmetric MLR case, we will therefore first consider the identification problem of the symmetric MLR problem.

A. Online EM Algorithm for Symmetric MLR Problem

The symmetric MLR model can be simplified as follows:

$$y_{k+1}=z_k\beta^{*\tau}\phi_k+w_{k+1}, \tag{2}$$

where β_1^* in (1) is denoted as β^*. The symmetric prior information simplifies the estimation of the posterior probability of which cluster the newly generated data comes from, thereby facilitating the design and analysis of the algorithm.

In order to illustrate the design principle of the identification algorithm, we take the i.i.d $\mathcal{N}(0,\sigma^2)$ noise, i.i.d regression vector and i.i.d balanced hidden variable cases with mutual independence among all three sequences just for simplicity of derivation (not actually used in our convergence analysis). Then we derive the likelihood function of model (2) with the parameter β:

$$\mathcal{L}_n(\beta)=\prod_{k=1}^{n}P(y_{k+1}|\beta,\phi_k)=\prod_{k=1}^{n}\Bigg[\frac{1}{2\sqrt{2\pi\sigma^2}}\times \Bigg[\exp\left(-\frac{(y_{k+1}-\beta^\tau\phi_k)^2}{2\sigma^2}\right)+\exp\left(-\frac{(y_{k+1}+\beta^\tau\phi_k)^2}{2\sigma^2}\right)\Bigg]\Bigg]. \tag{3}$$

With simple calculations, the gradient of the corresponding log-likelihood function with respect to β has the following expression:

$$\begin{aligned}&\mathcal{T}_n(\beta)=\nabla_\beta\log(\mathcal{L}_n)\\ =&\frac{1}{\sigma^2}\left\{\sum_{k=1}^{n}\left[-\phi_k\phi_k^\tau\beta+\phi_k y_{k+1}\tanh\left(\frac{\beta^\tau\phi_k y_{k+1}}{\sigma^2}\right)\right]\right\},\end{aligned} \tag{4}$$

where $\tanh(x)$ is the hyperbolic tangent function defined as $\tanh(x)=\frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$. We can see that (4) has multiple zeros, and it is hard to obtain the closed-form expression of maximum likelihood estimation (MLE). Hence, we adopt the EM algorithm (cf., [22]) to approximate the MLE. Denote β_t as the estimates of β^* at the time instant t. The EM algorithm is conducted according to two steps:

1) E-step: compute an auxiliary function $\mathbb{Q}_k(\beta)$, i.e., the log-likelihood for the data set $\{y_{t+1},z_t,\phi_t\}_{t=1}^k$ based on β_t:

$$\begin{aligned}\mathbb{Q}_k(\beta)=&-\frac{1}{2\sigma^2}\left\{\sum_{t=1}^{k}\Big[P(z_t=1|\phi_t,y_{t+1},\beta_t)(y_{t+1}-\beta^\tau\phi_t)^2\right.\\ &\left.+P(z_t=-1|\phi_t,y_{t+1},\beta_t)(y_{t+1}+\beta^\tau\phi_t)^2\Big]\right\}+kc,\end{aligned}$$

where $c=\log(\frac{1}{2\sqrt{2\pi}\sigma})$, the conditional probabilities of the hidden variable z_t given $\{\phi_t,y_{t+1},\beta_t\}$ are as follows:

$$P(z_t=1|\phi_t,y_{t+1},\beta_t)=\frac{\exp\left(\frac{\beta_t^\tau\phi_t y_{t+1}}{\sigma^2}\right)}{\exp\left(\frac{\beta_t^\tau\phi_t y_{t+1}}{\sigma^2}\right)+\exp\left(-\frac{\beta_t^\tau\phi_t y_{t+1}}{\sigma^2}\right)},$$

and $P(z_t=-1|\phi_t,y_{t+1},\beta_t)=1-P(z_t=1|\phi_t,y_{t+1},\beta_t)$.

2) M-step: update the estimate β_k by maximizing the function $\mathbb{Q}_k(\beta)$:

$$\begin{aligned}&\beta_{k+1}=\arg\max_{\beta}\mathbb{Q}_k(\beta)\\ =&\left(\sum_{t=1}^{k}\phi_t\phi_t^\tau\right)^{-1}\left(\sum_{t=1}^{k}\phi_t y_{t+1}\tanh\left(\frac{\beta_t^\tau\phi_t y_{t+1}}{\sigma^2}\right)\right).\end{aligned} \tag{5}$$

It is clear that the EM algorithm is a soft version of the well-known k-means algorithm [38].

Denote

$$\bar{y}_{k+1} = y_{k+1}\tanh\big(\frac{\beta_k^\tau \phi_k y_{k+1}}{\sigma^2}\big). \tag{6}$$

We can see that the equation (5) is the standard formula of LS with output $\bar{y}_{k+1}$ and the regressor ϕ_k, hence following the same way as the derivation of the recursive LS [39], we get the resulting online EM algorithm as shown in Algorithm 1. Unlike the existing off-line EM algorithms where batch data is required at each iteration (cf., [31]), the parameter estimates in Algorithm 1 will be recursively updated based on the current estimate and the new observation data.

Algorithm 1 Online EM algorithm for symmetric MLR

$$\beta_{k+1} = \beta_k + a_k P_k\phi_k\big(y_{k+1}\tanh\big(\frac{\beta_k^\tau\phi_k y_{k+1}}{\sigma^2}\big) - \beta_k^\tau\phi_k\big), \tag{7a}$$

$$P_{k+1} = P_k - a_k P_k\phi_k\phi_k^\tau P_k, \tag{7b}$$

$$a_k = \frac{1}{1+\phi_k^\tau P_k\phi_k}, \tag{7c}$$

where β_k is the estimate at time k, P_k is the adaptation gain matrix, the initial values $\beta_0 \neq 0$ and $P_0 > 0$ can be chosen arbitrarily.

B. Online EM Algorithm for Asymmetric MLR Problem

First of all, we show that the asymmetric case can be transferred to a case that can be dealt with by both the method in the symmetric case and the method of LS.

Let us denote $\theta_1^* = \frac{\beta_1^*+\beta_2^*}{2}$ and $\theta_2^* = \frac{\beta_1^*-\beta_2^*}{2}$. Clearly, the parameters β_1^* and β_2^* in (1) will be obtained once the parameters θ_1^* and θ_2^* are identified. The MLR model (1) can then be equivalently rewritten into the following model:

$$y_{k+1} = \theta_1^{*\tau}\phi_k + z_k\theta_2^{*\tau}\phi_k + w_{k+1}. \tag{8}$$

Under the balanced assumption on the hidden variable z_k, and the mutual independence assumption among z_k, ϕ_k and w_{k+1} for each $k \geq 0$, we have $\mathbb{E}\left[z_k\theta_2^{*\tau}\phi_k|\phi_k\right] = 0$, and then

$$\mathbb{E}\left[y_{k+1}|\phi_k\right] = \theta_1^{*\tau}\phi_k, \tag{9}$$

which is actually a linear regression model and can be estimated by the LS algorithm.

Thus, we propose a two-step identification algorithm to estimate the parameters θ_1^* and θ_2^*. Firstly, by (9), the parameter θ_1^* is estimated using the LS algorithm. Secondly, by replacing the unknown parameter θ_1^* in (8) by its LS estimate given in the first step, we can then estimate the unknown parameter θ_2^* in the same way as that for the symmetric case stated above. The whole algorithm is summarized in Algorithm 2.

Algorithm 2 Online EM algorithm for asymmetric MLR

#Step 1: LS estimation $\theta_{k,1}$ of θ_1^*

$$\theta_{k+1,1} = \theta_{k,1} + a_k P_k\phi_k\left(y_{k+1} - \theta_{k,1}^\tau\phi_k\right),$$

$$P_{k+1} = P_k - a_k P_k\phi_k\phi_k^\tau P_k, \quad a_k = \frac{1}{1+\phi_k^\tau P_k\phi_k},$$

#Step 2: EM estimation $\theta_{k,2}$ of θ_2^*

$$\theta_{k+1,2} = \theta_{k,2} + a_k P_k\phi_k\big[m_{k+1}\tanh\big(\frac{\theta_{k,2}^\tau\phi_k m_{k+1}}{\sigma^2}\big) - \theta_{k,2}^\tau\phi_k\big],$$

$$m_{k+1} = y_{k+1} - \theta_{k,1}^\tau\phi_k,$$

$$\beta_{k+1,1} = \theta_{k+1,1} + \theta_{k+1,2}, \quad \beta_{k+1,2} = \theta_{k+1,1} - \theta_{k+1,2},$$

where $\beta_{k,1}$ and $\beta_{k,2}$ are the estimates at time k, P_k is the adaptation gain matrix, the initial values $\theta_{0,1}, \theta_{0,2} \neq 0$, $P_0 > 0$ can be chosen arbitrarily.

Remark 2: The differences between Algorithms 1-2 and the classical stochastic gradient descent (SGD) algorithm and other similar approaches in adaptive identification including LS (cf., [39]) are as follows: 1) Rather than the mean-squares loss typically used in SGD and other similar approaches in adaptive identification, Algorithm 1 is designed to optimize the log-likelihood function of the observed data based on the probability density function (p.d.f) of the noise, resulting in a smoother gradient function and simpler analysis. Besides, different from the existing algorithms (e.g., the EM algorithms) that directly minimize the original non-convex log-likelihood or the mean-squares functions, Algorithm 2 operates in an online two-step manner by minimizing two coupled criterion functions, thus exhibiting global convergence to the true parameter set; 2) Algorithms 1-2 employ the adaptation gain matrices, rather than the same step size for each coordinate in SGD, can potentially accelerate the convergence rate. As for other similar approaches in adaptive identification, one class of most commonly-used algorithms is the adaptive Newton-like algorithms constructed by using the adaptation Hessian matrices. However, the adaptation Hessian matrices here may not be positive-definite due to the coupling with parameter estimates, whereas Algorithms 1-2 ensure the positive-definiteness of adaptation gain matrices by adopting the EM principle based on the posterior probability estimation of the hidden variable.

IV. Main Results

In this section, we give the main results concerning the convergence of parameter estimates and the data clustering performance of Algorithms 1 and 2, respectively.

A. Performance of Algorithm 1

To establish a rigorous theory on the performance of the identification algorithms, we need to introduce some assumptions on the hidden variable z_k, the noise w_{k+1}, and the regressor ϕ_k.

Assumption 1: The sequence of hidden variables $\{z_k\}$ is i.i.d with distribution $P(z_k = 1) = p \in (0,1)$ and $P(z_k = -1) = 1-p$. In addition, z_k is independent of ϕ_k for $k \geq 0$.

Remark 3: As far as we know, Assumption 1 on the mixing weights in the symmetric MLR case is the weakest one compared to most existing works that assume balanced ($p = \frac{1}{2}$) [31] or unbalanced mixtures with additional constraints [33]. Besides, Assumption 1 is also adopted by [40] in the noiseless case.

Assumption 2: The sequence of noises $\{w_{k+1}\}$ is i.i.d with Gaussian distribution $\mathcal{N}(0,\sigma^2)$. In addition, w_{k+1} is independent of $\{z_t\}_{t\leq k}$ and $\{\phi_t\}_{t\leq k}$ for $k\geq 0$.

Assumption 3: The regressor sequence $\{\phi_k\}$ is asymptotically stationary and ergodic and $\{\|\phi_k\|^4\}$ is uniformly integrable. In addition, its p.d.f $g_k(x)$ satisfies

$$\lim_{k\to\infty} g_k(x) = \bar{g}(x) \in \mathcal{G} = \Big\{g(x) : g(x) \text{ is a function of } \|x\| \text{ for } x\in\mathbb{R}^d, \int xx^\tau g(x)dx > 0\Big\}. \tag{10}$$

Remark 4: We remark that the set $\mathcal{G}$ of probability density functions include many familiar distributions, such as the uniform distribution on the sphere, the isotropic Gaussian distribution, the Logistic distribution, the polynomial distribution and the probability density functions introduced in [41].

Remark 5: The requirement that $g(x)$ is a function of $\|x\|$ means that it has the rotation-invariant property. This property is assumed for simplicity of presentation and can be further relaxed. For example, if the asymptotically stationary density function of ϕ_k is Gaussian with zero mean and non-unit covariance matrix $\Sigma > 0$, then $g(x)$ will have the form $g_0(\Sigma^{-1/2}x)$, which does not satisfy the rotation-invariant property although the standard normal density $g_0(x)$ does. In this case, let

$$\widehat{\phi}_k = \bar{R}_k^{-\frac{1}{2}}\phi_k, \bar{R}_k = \bar{R}_{k-1} + \frac{1}{k}\left(\phi_k\phi_k^\tau - \bar{R}_{k-1}\right).$$

From the fact that $\bar{R}_k \to \Sigma$ almost surely as $k\to\infty$, it is easy to obtain that $\widehat{\phi}_k$ converges in distribution to a standard normal random variable. Then by replacing ϕ_k with $\widehat{\phi}_k$ in Algorithm 1, it can be transferred to the rotation invariant case in Assumption 3 and our main results in Theorems 1-4 still hold.

Remark 6: Assumption 3 is weaker than that in most existing literature where the regressor $\{\phi_k\}$ is required to be i.i.d with a standard Gaussian distribution (cf., [29]–[31]). Assumption 3 can be satisfied in many cases, for example, when $\{\phi_k\}$ is generated by the following standard stochastic linear dynamical system excited or driven by a white noise signal:

$$\phi_{k+1} = A\phi_k + e_{k+1},$$

where the matrix A is stable, and e_k is i.i.d with $\mathcal{N}(0,I)$.

Based on the above assumptions, we give the convergence result for parameter identification and data clustering performance of Algorithm 1 as follows:

1) Convergence of Algorithm 1: We give the following main theorem on the convergence of the identification Algorithm 1:

Theorem 1: Let Assumptions 1-3 be satisfied. Then for any initial values $\beta_0 \neq 0$ and $P_0 > 0$, the estimate β_k generated by Algorithm 1 will converge to a limit point that belongs to the set $\{\beta^*, -\beta^*\}$ almost surely.

Remark 7: Note that the convergence property provided in Theorem 1 is of local nature in the sense that the limit point of β_k may depend on the initial value β_0 of Algorithm 1. A somewhat surprising fact is that this local convergence property is sufficient for guaranteeing the global optimality of the data clustering performance asymptotically, as will be rigorously shown in Theorem 2 below. This is reminiscent of the well-known self-tuning regulators in adaptive control, where the control performance can still achieve its optimal value even though the parameter estimates may not converge to the true parameter values (cf., [42]–[44]).

2) Clustering Performance of Algorithm 1: For the new data $\{\phi_k, y_{k+1}\}$, denote its corresponding cluster as $\mathcal{I}_k^* = 1$ if $z_k = 1$, and $\mathcal{I}_k^* = 2$ if $z_k = -1$. Based on the estimate β_k generated by Algorithm 1, we can online categorize $\{\phi_k, y_{k+1}\}$ to the corresponding cluster $\mathcal{I}_k \in \{1,2\}$ according to the criterion:

$$\mathcal{I}_k = \arg\min_{i=1,2}\{(y_{k+1} + (-1)^i\beta_k^\tau\phi_k)^2\}. \tag{11}$$

To evaluate the clustering performance for the within-cluster errors, a commonly-used evaluation index (cf., [33]) is defined as follows:

$$J_n = \frac{1}{n}\sum_{k=1}^{n}(y_{k+1} + (-1)^{\mathcal{I}_k}\beta_k^\tau\phi_k)^2. \tag{12}$$

Our purpose is to provide a lower bound to the probability that $\{y_{k+1}, \phi_k\}$ can be categorized into the correct cluster, and an upper bound of the within-cluster error. The main result on the performance of data clustering is stated as follows:

Theorem 2: Let Assumptions 1-3 be satisfied. Then the probability that the new data $\{\phi_k, y_{k+1}\}$ is categorized into the correct cluster is bounded from below by

$$\lim_{k\to\infty} P(\mathcal{I}_k = \mathcal{I}_k^*) \geq 1 - \mathbb{E}\left[\exp\left(-\frac{(\beta^{*\tau}\phi)^2}{2\sigma^2}\right)\right], \tag{13}$$

and the within-cluster error (12) satisfies

$$\lim_{n\to\infty} J_n = \sigma^2 + 4\mathbb{E}\left[\eta(\phi)\right] \leq \sigma^2, \tag{14}$$

with

$$\eta(\phi) = (\beta^{*\tau}\phi)^2\Phi\left(-\frac{|\beta^{*\tau}\phi|}{\sigma}\right) - \sigma|\beta^{*\tau}\phi|\Phi'\left(-\frac{|\beta^{*\tau}\phi|}{\sigma}\right) \leq 0,$$

where ϕ is a random vector with p.d.f $\bar{g} \in \mathcal{G}$ being the asymptotic stationary p.d.f of ϕ_k, $\Phi(x)$ and $\Phi'(x)$ are the standard Gaussian distribution function and density function, respectively.

Remark 8: From the proof of Theorem 2, one can find that the bounds given in (13) and (14) are actually the same bounds as in the case where the true parameter β^* is known. It can also be seen that the data clustering performance is positively related to the signal-to-noise ratio $\frac{|\beta^{*\tau}\phi|}{\sigma}$. Specifically, as the noise variance σ^2 tends to zero, the lower bound to the probability of correct categorization will converge to 1 and the upper bound of the within-cluster error will approach 0.

It goes without saying that given a specific form of the density function $\bar{g}$, one can obtain a more explicit bound concerning the probability that the new data is categorized into the correct cluster.

Example 1: Let conditions of Theorem 2 be satisfied and $\bar{g}$ be the p.d.f of Gaussian distribution $\mathcal{N}(0,\Sigma)$ with $\Sigma > 0$. Then with the estimate β_k generated by Algorithm 1, we have

$$\lim_{k\to\infty} P(\mathcal{I}_k = \mathcal{I}_k^*) \geq 1 - \frac{1}{\sqrt{1 + \frac{\beta^{*\tau}\Sigma\beta^*}{\sigma^2}}}. \tag{15}$$

The proof of (15) is provided in Appendix I.

B. Performance of Algorithm 2

In this subsection, we first give the convergence results of Algorithm 2 and then present the corresponding data clustering performance. For this purpose, the noise $\{w_{k+1}\}$ and the regressor $\{\phi_k\}$ are still assumed to obey Assumptions 2 and 3 in Section IV-A, while the sequence of hidden variables $\{z_k\}$ is assumed to satisfy the following assumption:

Assumption 1′: The sequence of hidden variables $\{z_k\}$ is i.i.d with balanced distribution $P(z_k = 1) = P(z_k = -1) = \frac{1}{2}$. In addition, z_k is independent of ϕ_k for each $k \geq 0$.

1) Convergence of Algorithm 2: Based on the convergence theory presented in Section IV-A and the celebrated convergence property of the LS, we can obtain the following main result on the convergence of Algorithm 2:

Theorem 3: Let Assumptions $1'$ and 2-3 be satisfied. Then for any initial values $\theta_{0,1}$, $\theta_{0,2} \neq 0$ and $P_0 > 0$, the estimate $(\beta_{k,1}, \beta_{k,2})$ by Algorithm 2 will converge to a limit point that belongs to the set $\{(\beta_1^*, \beta_2^*), (\beta_2^*, \beta_1^*)\}$ almost surely.

2) Clustering Performance of Algorithm 2: Similar to the analysis in Section IV-A, based on the estimates $\beta_{k,1}$ and $\beta_{k,2}$ generated by Algorithm 2, we can also online categorize $\{\phi_k, y_{k+1}\}$ to the corresponding cluster $\mathcal{I}_k' \in \{1, 2\}$ according to the following criterion:

$$\mathcal{I}_k' = \arg\min_{i=1,2}\{(y_{k+1} - \beta_{k,i}^\tau \phi_k)^2\}. \tag{16}$$

Moreover, the corresponding within-cluster error is defined as

$$J_n' = \frac{1}{n}\sum_{k=1}^{n}(y_{k+1} - \beta_{k,\mathcal{I}_k'}^\tau \phi_k)^2. \tag{17}$$

In the following theorem, we give an asymptotic lower bound to the probability that $\{y_{k+1}, \phi_k\}$ can be categorized correctly and an upper bound of the within-cluster error:

Theorem 4: Let Assumptions $1'$ and 2-3 be satisfied. Then the probability that the new data $\{\phi_k, y_{k+1}\}$ is categorized into the correct cluster is bounded from below by

$$\lim_{k\to\infty} P(\mathcal{I}_k' = \mathcal{I}_k^*) \geq 1 - \mathbb{E}\left[\exp\left(-\frac{((\beta_1^* - \beta_2^*)^\tau \phi)^2}{8\sigma^2}\right)\right],$$

and the within-cluster error (17) satisfies

$$\lim_{n\to\infty} J_n' = \sigma^2 + \mathbb{E}\left[\eta(\phi)\right] \leq \sigma^2,$$

where $\eta(\phi) = ((\beta_1^* - \beta_2^*)^\tau \phi)^2 \Phi\left(-\frac{|(\beta_1^*-\beta_2^*)^\tau \phi|}{2\sigma}\right) - 2\sigma|(\beta_1^* - \beta_2^*)^\tau \phi|\Phi'\left(-\frac{|(\beta_1^*-\beta_2^*)^\tau \phi|}{2\sigma}\right)$, ϕ, $\Phi(x)$ and $\Phi'(x)$ are defined in Theorem 2.

Similar to Example 1, if $\bar{g}$ has a specific form, one can also obtain the following concrete result:

Example 2: Let conditions of Theorem 4 be satisfied and $\bar{g}$ be defined in Example 1. Then with the estimates $\beta_{k,1}$ and $\beta_{k,2}$ generated by Algorithm 2, we have

$$\lim_{k\to\infty} P(\mathcal{I}_k' = \mathcal{I}_k^*) \geq 1 - \frac{1}{\sqrt{1 + \frac{(\beta_1^*-\beta_2^*)^\tau \Sigma(\beta_1^*-\beta_2^*)}{4\sigma^2}}}.$$

V. Proofs of the Main Results

In this section, we provide proofs of Theorems 1-4.

A. Proof of Theorem 1

The celebrated Ljung's ODE method [35] provides a general analytical technique for recursive algorithms by establishing the relationship between the asymptotic behavior of the recursive algorithms and the stability of the corresponding ODEs.

Denote $R_{k+1} = \frac{1}{k}\left[P_0^{-1} + \sum_{t=0}^{k} \phi_t \phi_t^\tau\right]$. From (7) and the matrix inverse formula [45], it follows that $a_k P_k \phi_k = P_{k+1}\phi_k = \frac{1}{k} R_{k+1}^{-1}\phi_k$. Thus Algorithm 1 can be rewritten in the following equivalent form:

$$\beta_{k+1} = \beta_k + \frac{1}{k} Q_1(x_k, \phi_k, y_{k+1}), \tag{18a}$$

$$R_{k+1} = R_k + \frac{1}{k} Q_2(x_k, \phi_k, y_{k+1}), \tag{18b}$$

with $Q_1(x_k, \phi_k, y_{k+1}) = R_{k+1}^{-1}\phi_k\big(y_{k+1}\tanh\left(\frac{\beta_k^\tau \phi_k y_{k+1}}{\sigma^2}\right) - \beta_k^\tau \phi_k\big)$, $Q_2(x_k, \phi_k, y_{k+1}) = \phi_k\phi_k^\tau - R_k$, and $x_k = \begin{bmatrix} \beta_k^\tau & \mathrm{vec}^\tau(R_k) \end{bmatrix}^\tau$, where $\mathrm{vec}(\cdot)$ denotes the operator by stacking the columns of a matrix on top of one another. Then x_k evolves according to the following form:

$$x_{k+1} = x_k + \frac{1}{k} Q(x_k, \phi_k, y_{k+1}), \tag{19}$$

where

$$Q(x_k, \phi_k, y_{k+1}) = \begin{bmatrix} Q_1(x_k, \phi_k, y_{k+1}) \\ \mathrm{vec}\left(Q_2(x_k, \phi_k, y_{k+1})\right) \end{bmatrix}. \tag{20}$$

In order to analyze (19), we introduce the following ODEs:

$$\frac{d}{dt}\beta(t) = R^{-1}(t) f(\beta(t)), \tag{21a}$$

$$\frac{d}{dt}R(t) = G - R(t), \tag{21b}$$

where $f(\beta(t)) = \lim\limits_{k\to\infty} \mathbb{E}\big[\phi_k\big(y_{k+1}\tanh\left(\frac{\beta^\tau(t)\phi_k y_{k+1}}{\sigma^2}\right) - \beta^\tau(t)\phi_k\big)\big]$ and $G = \lim\limits_{k\to\infty} \mathbb{E}\left[\phi_k \phi_k^\tau\right]$.

The main results of Ljung's ODE method can be restated in the following proposition, which plays an important role in our analysis:

Proposition 1: [35] Let D be an open and connected subset of $\mathbb{R}^{d+d^2}$ and D_s be a compact subset of D such that the trajectory of (21) starting in D_s remains in D_s for $t > 0$. Assume also that there is an invariant set $D_c \subset D_s$ of (21) such that its attraction domain $D_A \supset D_s$. Then $x_k \to D_c$ as $k \to \infty$ almost surely, provided that the following conditions are satisfied:

B1) The function $Q(x, \phi, y)$ defined in (20) is locally Lipschitz continuous for $x \in D$ with fixed ϕ and y, that is, for $x_i \in \mathcal{U}(x, \rho(x))$ with $\rho(x) > 0$,

$$\|Q(x_1, \phi, y) - Q(x_2, \phi, y)\| < \mathcal{R}(x, \phi, y, \rho(x))\|x_1 - x_2\|,$$

where $x = \begin{bmatrix} \beta^\tau & \mathrm{vec}^\tau(R) \end{bmatrix}^\tau$, and $\mathcal{U}(x, \rho(x))$ is the $\rho(x)$-neighborhood of x, i.e., $\mathcal{U}(x, \rho(x)) = \{\bar{x} : \|x - \bar{x}\| < \rho(x)\}$.

B2) $\frac{1}{n}\sum\limits_{k=1}^{n} \mathcal{R}(x, \phi_k, y_{k+1}, \rho(x))$ converges to a finite limit for any $x \in D$ as $n \to \infty$.

B3) $\lim\limits_{k\to\infty} \mathbb{E}\left[Q(x, \phi_k, y_{k+1})\right]$ exists for $x \in D$ and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} Q(x, \phi_k, y_{k+1}) = \lim_{k\to\infty}\mathbb{E}\left[Q(x, \phi_k, y_{k+1})\right]. \tag{22}$$

B4) There exists a positive constant L such that the following events happen i.o. with probability 1:

$$x_k \in D_s \text{ and } \|\phi_k\| \le L.$$

In order to establish the convergence property of Algorithm 1, we will verify all the conditions in Proposition 1. We first verify Conditions B1)-B3) in Lemmas 1-2 with their proofs given in Appendix I.

Lemma 1: Under Assumptions 1-3, $\{d_k \triangleq [z_k\ \phi_k^\tau\ w_{k+1}]^\tau, k \ge 1\}$ is an asymptotically stationary and ergodic process with bounded fourth moment. In addition, any measurable function of d_k is also an asymptotically stationary and ergodic stochastic process.

Lemma 2: Consider Algorithm 1 subject to Assumptions 1-3. Then Conditions B1)-B3) in Proposition 1 are satisfied in the open set $D = \{x : R > 0\}$ with $x = \left[\begin{array}{cc} \beta^\tau & \operatorname{vec}^\tau(R) \end{array}\right]^\tau$.

For the verification of the remaining conditions required by Proposition 1, i.e., the stability analysis of ODEs (21) and Condition B4), we provide some useful lemmas (Lemmas 3-5) with their proofs given in Appendix I.

Lemma 3: For a random variable $y \sim p\mathcal{N}(a, \sigma^2) + (1-p)\mathcal{N}(-a, \sigma^2)$ with $p \in [0, 1]$ and a being a constant, we have $\mathbb{E}\left[y \tanh\left(\frac{ay}{\sigma^2}\right)\right] = a$.

Lemma 4: [46] For a random variable $y \sim \mathcal{N}(a, \sigma^2)$ with constants a and $\hat{a}$ satisfying $a\hat{a} \ge 0$, we have

$$\mathbb{E}\left[\frac{\hat{a}y}{\sigma^2}\tanh'\left(\frac{\hat{a}y}{\sigma^2}\right)\right] \ge 0,$$
$$\mathbb{E}\left[\tanh\left(\frac{\hat{a}y}{\sigma^2}\right)\right] \ge 1 - \exp\left(-\frac{|a|\min(|a|, |\hat{a}|)}{2\sigma^2}\right),$$

where $\tanh'(\cdot)$ is the derivative function of $\tanh(\cdot)$.

Lemma 5: For $x > 0$ and $c > 0$, the function

$$F(c, x) = \int_{-\infty}^{\infty} f(c, x, w)\exp\left(-\frac{w^2}{2\sigma^2}\right)dw \tag{23}$$

is increasing with respect to x, where $f(c, x, w) = (w + x)\tanh\left(\frac{c(w+x)}{\sigma^2}\right) + (w - x)\tanh\left(\frac{c(w-x)}{\sigma^2}\right)$.

Based on the above lemmas, we now proceed to verify the remaining conditions required by Proposition 1 through four steps. To be specific, we will establish the stability analysis of ODEs (21) in Steps 1-3 and then verify Condition B4) in Step 4.

Proof of Theorem 1: Without loss of generality, we focus on the analysis for the case of $\beta^* \ne 0$ and then give some additional explanations for the case of $\beta^* = 0$.

We now investigate the properties of ODEs (21). Specifically, we prove the following assertion:

the ODEs (21) has the invariant set $D_c = \{x : \beta \in \{\beta^*, -\beta^*\} \cup D_{us}, R = G\}$ with the domain of attraction $D_A = \{x : \|R - G\| < \varepsilon\}$. (24)

where $x = [\beta^\tau\ \operatorname{vec}^\tau(R)]^\tau$, $D_{us} = \{0, c^*v, -c^*v\}$, v is a unit vector orthogonal to β^*, c^* and ε are positive constants to be determined later. We prove the assertion (24) by establishing the stability properties of (21) on three subsets $D_{A,1} = \{x : \beta^\tau\beta^* > 0, \|R - G\| < \varepsilon\}$, $D_{A,2} = \{x : \beta^\tau\beta^* < 0, \|R - G\| < \varepsilon\}$ and $D_{A,3} = \{x : \beta^\tau\beta^* = 0, \|R - G\| < \varepsilon\}$ of the domain of attraction D_A. For the case of $x(0) \in D_{A,1}$, we show that $x(t) \in D_{A,1}$ for $t \ge 0$ in Steps 1-2 below and then establish the corresponding stability result in Step 3 below.

From (21b), we have

$$R(t) = G + e^{-t}(R(0) - G). \tag{25}$$

By (21) and the fact that the asymptotically stationary p.d.f of ϕ_k has the rotation invariant property, we have $G = c_0 I$ with c_0 being a positive constant [41]. From (25), it follows that $\|R(t) - G\| \le \|R(0) - G\|$. So it suffices to show $\beta^\tau(t)\beta^* > 0$ when proving $x(t) \in D_{A,1}$ for $t \ge 0$. For this, we derive the lower and upper bounds of $\|\beta(t)\|$ in (21a) in Step 1 below.

Step 1: Boundedness of $\|\beta(t)\|$ generated by (21a).

For the convenience of analysis, we choose a set of standard orthogonal basis $\{v_1(t), \cdots, v_d(t)\}$, where $v_1(t) = \frac{\beta(t)}{\|\beta(t)\|}$ and $v_2(t)$ belongs to span$\{\beta(t), \beta^*\}$. Then $\beta(t)$, β^* and $f(\beta(t))$ defined in (21) can be written as follows:

$$\begin{aligned} &\beta(t) = \sum\nolimits_{i=1}^{d} b_i(t)v_i(t),\ \beta^* = \sum\nolimits_{i=1}^{d} b_i^*(t)v_i(t), \\ &f(\beta(t)) = \sum\nolimits_{i=1}^{d} h_i(\beta(t))v_i(t), \end{aligned} \tag{26}$$

where $b_1(t) = \|\beta(t)\|$, $b_i(t) \equiv 0 (i > 1)$, $b_1^*(t) = \beta^{*\tau}v_1(t)$, $b_2^*(t) = \beta^{*\tau}v_2(t)$, $b_i^*(t) \equiv 0 (i > 2)$ and $h_i(\beta(t)) = v_i^\tau(t)f(\beta(t))$. Note that there may exist a time instant t_0 such that $v_2(t) = v_1(t)$ for $t \ge t_0$. For such a case, the orthogonal basis defined above degenerates, but the analysis in this part still holds. Besides, $v_1(t)$ is differentiable whenever $\|\beta(t)\| \ne 0$ since $\beta(t)$ is continuously differentiable, and we will show below $\|\beta(t)\| > 0$ for $t \ge 0$. Moreover, we give some illustrations on properties of $f(\beta(t))$. By (21) and Assumptions 2-3, we have

$$f(\beta(t)) = \mathbb{E}\left[\phi\left(y\tanh\left(\frac{\beta^\tau(t)\phi y}{\sigma^2}\right) - \beta^\tau(t)\phi\right)\right], \tag{27}$$

where ϕ is a random vector with p.d.f $\bar{g} \in \mathcal{G}$ being the asymptotic stationary p.d.f of ϕ_k and the random variable y given ϕ obeys the distribution $\mathcal{N}(\beta^{*\tau}\phi, \sigma^2)$ by Assumption 2 and Lemma 3. Denote $a_i(t) = v_i^\tau(t)\phi$, $i \in [d]$, we have

$$\phi = \sum\nolimits_{i=1}^{d} a_i(t)v_i(t). \tag{28}$$

By Assumption 3, we know that the p.d.f $\bar{g}(\phi)$ of ϕ has the rotation-invariant property. Thus from (28), it follows that the p.d.f of $a(t) = [a_1(t), \cdots, a_d(t)]^\tau$ equals $\bar{g}(a(t))$, which is an even function in $a_i(t)$ and also the marginal p.d.f of $a_i(t)$ is an even function in $a_i(t)$ for $i \in [d]$. So we have $\mathbb{E}[a_1(t)a_i(t)] = 0, i \in [d]\backslash\{1\}$ and $\mathbb{E}[a_1^3(t)a_2(t)] = 0$. Moreover, by the definition of G in (21) and $G = c_0 I$ in (25), we have $\mathbb{E}[\phi\phi^\tau] = c_0 I$, thus $\mathbb{E}[a_i^2(t)] = c_0$ for $i \in [d]$. Besides, by assumptions that $\{\|\phi_k\|^4\}$ is u.i. and the p.d.f of ϕ has the rotation-invariant property in Assumption 3, we can obtain that $\mathbb{E}[a_i^4(t)] = c_1, i \in [d]$ with a positive constant c_1. These properties will be used in the following analysis without citations.

We now prove that $b_1(t) = \|\beta(t)\|$ has a positive lower bound for $t \ge 0$. From $x(0) \in D_{A,1}$, it suffices to prove that there exists a constant $b_l > 0$ such that

$$\frac{db_1(t)}{dt} > 0, \text{ if } 0 < b_1(t) < b_l. \tag{29}$$

By the fact $v_i^\tau(t)v_i(t)\equiv 1$ for $i\in[d]$, we have

$$v_i^\tau(t)\frac{dv_i(t)}{dt}\equiv 0. \tag{30}$$

Thus from (21a), (26) and (30), we have

$$\begin{aligned}\frac{db_1(t)}{dt}=&\frac{d[\beta^\tau(t)v_1(t)]}{dt}\\ =&v_1^\tau(t)\frac{d\beta(t)}{dt}+b_1(t)v_1^\tau(t)\frac{dv_1(t)}{dt}=v_1^\tau(t)R^{-1}(t)f(\beta(t))\\ =&c_0^{-1}h_1(\beta(t))+v_1^\tau(t)(R^{-1}(t)-c_0^{-1}I)f(\beta(t))\triangleq S(\beta(t)).\end{aligned} \tag{31}$$

By (26) and (28), we have $\beta^\tau(t)\phi=a_1(t)b_1(t)$ and $\beta^{*\tau}\phi=a_1(t)b_1^*(t)+a_2(t)b_2^*(t)$, thus by (27), we have for $i\in[d]$,

$$h_i(\beta(t))=\mathbb{E}\big[a_i(t)\big(y\tanh\big(\frac{a_1(t)b_1(t)y}{\sigma^2}\big)-a_1(t)b_1(t)\big)\big], \tag{32}$$

where y obeys the distribution $\mathcal{N}(a_1(t)b_1^*(t)+a_2(t)b_2^*(t),\sigma^2)$ given $a(t)$. Then by $\tanh(0)=0$, we have $h_i(\beta(t))=0$ and $f(\beta(t))=0$ for $b_1(t)=0$, thus $S(\beta(t))=0$ for $b_1(t)=0$. Hence, by (31) and the mean-value theorem, we obtain

$$S(\beta(t))=(b_1(t)-0)\frac{dS(\beta(t))}{db_1(t)}\bigg|_{b_1(t)=\zeta(t)}, \tag{33}$$

where $\zeta(t)\in[0,b_1(t)]$. To prove (29), i.e., $S(\beta(t))>0$ if $0<b_1(t)<b_l$, we proceed to show

$$\frac{dS(\beta(t))}{db_1(t)}\bigg|_{b_1(t)=0}>0. \tag{34}$$

By (26), the fact $h_i(\beta(t))=0$ and $f(\beta(t))=0$ when $b_1(t)=0$, we have

$$\begin{aligned}\frac{dS(\beta(t))}{db_1(t)}\bigg|_{b_1(t)=0}=&c_0^{-1}\frac{dh_1(\beta(t))}{db_1(t)}\bigg|_{b_1(t)=0}+v_1^\tau(t)\\ &(R^{-1}(t)-c_0^{-1}I)\bigg(\sum_{i=1}^{d}\frac{v_i(t)dh_i(\beta(t))}{db_1(t)}\bigg)\bigg|_{b_1(t)=0}.\end{aligned} \tag{35}$$

To analyze (35), we consider $\frac{dh_i(\beta(t))}{db_1(t)}\big|_{b_1(t)=0}, i\in[d]$ term by term. Firstly, by Assumptions 1-3, the definitions of y in (27) and $a_1(t)$ in (28), we have $\mathbb{E}[a_1^2(t)]<\infty$, $\mathbb{E}[a_1^2(t)y^2]<\infty$. Then by (32), $\tanh'(0)=1$, $\|b_1^*(t)\|^2+\|b_2^*(t)\|^2=\|\beta^*\|^2$ and the properties of $a(t)$, we obtain

$$\begin{aligned}&\frac{dh_1(\beta(t))}{db_1(t)}\bigg|_{b_1(t)=0}=\mathbb{E}\left[\frac{a_1^2(t)y^2}{\sigma^2}-a_1^2(t)\right]\\ =&\mathbb{E}\left[a_1^2(t)\frac{\sigma^2+(a_1(t)b_1^*(t)+a_2(t)b_2^*(t))^2}{\sigma^2}-a_1^2(t)\right]\\ =&\mathbb{E}\left[\frac{a_1^4(t)b_1^{*2}(t)+a_1^2(t)a_2^2(t)b_2^{*2}(t)}{\sigma^2}\right]>0.\end{aligned} \tag{36}$$

Secondly, by (32), we have

$$h_2(\beta(t))=\mathbb{E}\big[a_2(t)y\tanh\big(\frac{a_1(t)b_1(t)y}{\sigma^2}\big)\big]. \tag{37}$$

Similar to (36), with simple calculations, it follows that

$$\frac{dh_2(\beta(t))}{db_1(t)}\bigg|_{b_1(t)=0}=2\mathbb{E}\left[\frac{a_1^2(t)a_2^2(t)}{\sigma^2}\right]b_1^*(t)b_2^*(t). \tag{38}$$

Thirdly, by (32) and the fact that the marginal p.d.f of $a_i(t)$ is even, it is not difficult to obtain that for $i>2$,

$$h_i(\beta(t))\equiv 0,\ \text{and}\ \frac{dh_i(\beta(t))}{db_1(t)}\bigg|_{b_1(t)=0}=0. \tag{39}$$

Choose $\varepsilon=\frac{1}{4}c_0$ in $D_{A,1}$, we have $-\frac{1}{5c_0}I<R^{-1}(t)-c_0^{-1}I<\frac{1}{3c_0}I$ and then we have $v_1^\tau(t)(R^{-1}(t)-c_0^{-1}I)v_1(t)\geq-\frac{1}{5c_0}$, $v_1^\tau(t)(R^{-1}(t)-c_0^{-1}I)v_2(t)\geq-\frac{1}{3c_0}$. Thus by substituting (36), (38), (39) into (35), and using Schwarz inequality, we can obtain

$$\begin{aligned}\frac{dS(\beta(t))}{db_1(t)}\bigg|_{b_1(t)=0}>&\frac{4\left[\mathbb{E}[a_1^4(t)]b_1^{*2}(t)+\mathbb{E}[a_1^2(t)a_2^2(t)]b_2^{*2}(t)\right]}{5c_0\sigma^2}\\ -&\frac{2\mathbb{E}[a_1^2(t)a_2^2(t)]b_1^*(t)b_2^*(t)}{3c_0\sigma^2}\geq\frac{\|\beta^*\|^2\mathbb{E}\left[a_1^2(t)a_2^2(t)\right]}{3c_0\sigma^2}>0,\end{aligned}$$

where the facts $\mathbb{E}[a_1^4(t)]=\mathbb{E}[a_2^4(t)]$ and $\mathbb{E}\left[a_1^2(t)a_2^2(t)\right]>0$ derived from the rotation-invariant property of the p.d.f of ϕ in Assumption 3 are used. Thus (34) is obtained. From the continuity of $\frac{dS(\beta(t))}{db_1(t)}$ at $b_1(t)=0$, it follows that there exists a positive constant b_l such that $\frac{dS(\beta(t))}{db_1(t)}\big|_{b_1(t)=\zeta(t)}>0$ if $0<\zeta(t)\leq b_1(t)\leq b_l$. Then by (33), we have $S(\beta(t))>0$ if $0<b_1(t)\leq b_l$, thus (29) holds. Therefore, we can obtain that $b_1(t)$ has the following positive lower bound:

$$b_1(t)\geq\underline{b}=\min\{b_l,b_1(0)\}. \tag{40}$$

We now derive an upper bound of $b_1(t)=\|\beta(t)\|$. By (32), $|\tanh(\cdot)|\leq 1$ and Assumptions 2-3, we have

$$h_1(\beta(t))\leq\mathbb{E}[|a_1(t)y|]-b_1(t)\mathbb{E}[a_1^2(t)]=p_0-c_0b_1(t), \tag{41}$$

where $p_0=\mathbb{E}\left[|a_1(t)y|\right]<\infty$. By (37), (39) and (41), we have

$$\|f(\beta(t))\|\leq\|h_1(\beta(t))+h_2(\beta(t))\|\leq p_1+c_0b_1(t), \tag{42}$$

where $p_1=p_0+\mathbb{E}[|a_2(t)y|]<\infty$. Thus by (31), we have

$$\frac{db_1(t)}{dt}\leq c_0^{-1}(p_0-c_0b_1(t))+\frac{1}{3}c_0^{-1}(p_1+c_0b_1(t))\leq 0, \tag{43}$$

if $b_1(t)\geq\frac{3p_0+p_1}{2c_0}$. Thus $b_1(t)$ has the following upper bound:

$$b_1(t)\leq\bar{b}=\max\{(3p_0+p_1)/(2c_0),b_1(0)\}. \tag{44}$$

Step 2: Proof of $\beta^\tau(t)\beta^*>0$ for all $t\geq 0$.

For this purpose, by $\beta^\tau(t)\beta^*=b_1(t)b_1^*(t)$ and $b_1(t)>0$ derived in Step 1, we only need to prove that $b_1^*(t)>0$ for $t\geq 0$. For this purpose, by the fact that $b_1^*(0)>0$ for $x(0)\in D_{A,1}$, it suffices to show that

$$\frac{db_1^*(t)}{dt}\geq 0,\ \text{if}\ b_1^*(t)>0. \tag{45}$$

By (26), we have $b_1^{*2}(t)+b_2^{*2}(t)\equiv\|\beta^*\|^2$ and then

$$b_1^*(t)\frac{db_1^*(t)}{dt}=-b_2^*(t)\frac{db_2^*(t)}{dt}. \tag{46}$$

In order to prove (45), we first analyze the properties of $\frac{db_2^*(t)}{dt}$ for $0\leq|b_2^*(t)|<\|\beta^*\|$. By $b_2(t)\equiv 0$ in (26), it follows that $\frac{db_2(t)}{dt}\equiv 0$, then by (21) and (26), we have

$$\begin{aligned}\frac{db_2(t)}{dt}=&\frac{d[\beta^\tau(t)v_2(t)]}{dt}=\beta^\tau(t)\frac{dv_2(t)}{dt}+v_2^\tau(t)\frac{d\beta(t)}{dt}\\ =&b_1(t)v_1^\tau(t)\frac{dv_2(t)}{dt}+v_2^\tau(t)R^{-1}(t)f(\beta(t))\equiv 0.\end{aligned} \tag{47}$$

Thus by (30) and (47), we have

$$\begin{aligned}\frac{db_2^*(t)}{dt} &= \beta^{*\tau}\frac{dv_2(t)}{dt} = b_1^*(t)v_1^\tau(t)\frac{dv_2(t)}{dt}\\ =&-\frac{b_1^*(t)}{b_1(t)}v_2^\tau(t)R^{-1}(t)f(\beta(t)) = -\frac{b_1^*(t)}{b_1(t)}c_0^{-1}h_2(\beta(t))\\ &-\frac{b_1^*(t)}{b_1(t)}v_2^\tau(t)(R^{-1}(t)-c_0^{-1}I)f(\beta(t)).\end{aligned} \tag{48}$$

We analyze the right-hand-side (RHS) of (48) term by term. For the first term, let us denote the marginal p.d.f of $(a_1(t), a_2(t))$ as follows:

$$\begin{aligned}&\tilde{g}(a_1(t), a_2(t))\\ &= \int_{\mathbb{R}}\cdots\int_{\mathbb{R}}\bar{g}(a_1(t), \cdots, a_d(t))da_3(t)\cdots da_d(t),\end{aligned} \tag{49}$$

where $\bar{g}$ is the p.d.f of the random vector ϕ defined in (27). Since $\tanh(z)$ is odd and $z\tanh(z)$ is even, by (37), we have

$$\begin{aligned}h_2(\beta(t)) =& \mathbb{E}\Big[a_2(t)\big[a_1(t)b_1^*(t)+a_2(t)b_2^*(t)+w\big]\\ &\tanh\Big(\frac{a_1(t)b_1(t)\big[a_1(t)b_1^*(t)+a_2(t)b_2^*(t)+w\big]}{\sigma^2}\Big)\Big]\\ =&\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}}a_2(t)\big[a_1(t)b_1^*(t)+a_2(t)b_2^*(t)+w\big]\\ &\tanh\Big(\frac{a_1(t)b_1(t)\big[a_1(t)b_1^*(t)+a_2(t)b_2^*(t)+w\big]}{\sigma^2}\Big)\\ &\quad\tilde{g}(a_1(t), a_2(t))\exp\Big(-\frac{w^2}{2\sigma^2}\Big)dwda_2(t)da_1(t) \qquad (50)\\ =&\frac{1}{\sqrt{2\pi}\sigma}\int_{a_1(t)>0}\int_{a_2(t)>0}\big\{I(a_1(t), a_2(t))\\ &+I(-a_1(t), a_2(t))+I(-a_1(t), -a_2(t))\\ &+I(a_1(t), -a_2(t))\big\}\tilde{g}(a_1(t), a_2(t))da_2(t)da_1(t)\\ =&\frac{1}{\sqrt{2\pi}\sigma}\int_{a_1(t)>0}\int_{a_2(t)>0}a_2(t)\big\{F\big(c(t), |a_1(t)b_1^*(t)\\ &+a_2(t)b_2^*(t)|\big)-F\big(c(t), |a_1(t)b_1^*(t)-a_2(t)b_2^*(t)|\big)\big\}\\ &\qquad\tilde{g}(a_1(t), a_2(t))da_2(t)da_1(t),\end{aligned}$$

where $I(a_1(t), a_2(t)) = a_2(t)\int_{\mathbb{R}}[a_1(t)b_1^*(t)+a_2(t)b_2^*(t)+w]\tanh\left(\frac{a_1(t)b_1(t)[a_1(t)b_1^*(t)+a_2(t)b_2^*(t)+w]}{\sigma^2}\right)\exp\left(-\frac{w^2}{2\sigma^2}\right)dw$, the function $F(c(t), x(t))$ is defined in Lemma 5 with $c(t) = a_1(t)b_1(t) > 0$ (since $b_1(t) > 0$ by (40)). Besides, for positive $a_1(t)$, $a_2(t)$ and $b_1^*(t)$, we know that the term $|a_1(t)b_1^*(t)+a_2(t)b_2^*(t)| - |a_1(t)b_1^*(t)-a_2(t)b_2^*(t)|$ has the same sign as $b_2^*(t)$, thus by Lemma 5, it follows that $F\big(c(t), |a_1(t)b_1^*(t)+a_2(t)b_2^*(t)|\big) - F\big(c(t), |a_1(t)b_1^*(t)-a_2(t)b_2^*(t)|\big)$ has the same sign as $b_2^*(t)$. Therefore, by (50), we have for positive $b_1^*(t)$,

$$\begin{aligned}&h_2(\beta(t)) \geq 0 \text{ if } 0 \leq b_2^*(t) < \|\beta^*\|,\\ &h_2(\beta(t)) \leq 0 \text{ if } -\|\beta^*\| < b_2^*(t) \leq 0,\end{aligned} \tag{51}$$

where the equality holds if and only if $b_2^*(t) = 0$. Choose $\varepsilon = \frac{c_0^{-1}|h_2(\beta(0))|}{p_1+c_0\bar{b}}$, by (42), (44) and (48), we obtain that $\frac{db_2^*(t)}{dt}$ has an opposite sign with $b_2^*(t)$ at $t = 0$. Besides, from (25), it follows that $\|R^{-1}(t)-c_0^{-1}I\| = O(e^{-t})$. Then by (48), (50), (51), we can derive that for positive $b_1^*(t)$,

$$\begin{aligned}&\frac{db_2^*(t)}{dt} \leq 0 \text{ if } 0 \leq b_2^*(t) < \|\beta^*\|,\\ &\frac{db_2^*(t)}{dt} \geq 0 \text{ if } -\|\beta^*\| < b_2^*(t) \leq 0,\end{aligned} \tag{52}$$

where the equality holds if and only if $b_2^*(t) = 0$ (The detailed proof of (52) is provided in Appendix I). Therefore, by (46) and (52), (45) is proved. Moreover, by taking the Lyapunov function as $b_2^{*2}(t)$ and using the Lasalle invariance principle, we obtain

$$\lim_{t\to\infty} b_2^*(t) = 0, \text{ and } |b_2^*(t)| \leq |b_2^*(0)|. \tag{53}$$

Thus by (53) and $b_1^*(t)^2 + b_2^*(t)^2 \equiv \|\beta^*\|^2$, we have

$$b_1^*(0) \leq b_1^*(t) \leq \|\beta^*\|. \tag{54}$$

Combining all the above analysis, and letting

$$\varepsilon \triangleq \min\left\{\frac{1}{4}c_0, \frac{c_0^{-1}|h_2(\beta(0))|}{dp_0+c_0\bar{b}}\right\}, \tag{55}$$

then by (40) and (54), we obtain $\beta^\tau(t)\beta^* = b_1(t)b_1^*(t) > \underline{b}b_1^*(0) > 0$ for all $t \geq 0$.

Step 3: Analysis of the Lyapunov function.

We first establish the stability properties of $\beta(t)$ in (21a) for the case of $x(0) \in D_{A,1}$ using the Lyapunov method.

By the definition of $D_{A,1}$, (25) and (55), we know that $R(t) \geq \frac{3}{4}c_0I, t \geq 0$. Now we consider the following Lyapunov function:

$$V(\beta(t)) = \frac{1}{2}\tilde{\beta}^\tau(t)R(t)\tilde{\beta}(t),$$

where $\tilde{\beta}(t) = \beta(t) - \beta^*$. Then we have the following derivative of V along the trajectories (21):

$$\frac{dV(\beta(t), R(t))}{dt} = \tilde{\beta}^\tau(t)f(\beta(t)) + \frac{1}{2}\tilde{\beta}^\tau(t)\left(G-R(t)\right)\tilde{\beta}(t). \tag{56}$$

For the first term on the RHS of (56), by (26), we have

$$\tilde{\beta}^\tau(t)f(\beta(t)) = \tilde{b}_1(t)h_1(\beta(t)) - b_2^*(t)h_2(\beta(t)), \tag{57}$$

where $\tilde{b}_1(t) = b_1(t) - b_1^*(t)$. With simple calculations, we have by (32),

$$\tilde{b}_1(t)h_1(\beta(t)) = \tilde{b}_1(t)(L_1(\beta(t)) + L_2(\beta(t))), \tag{58}$$

where

$$\begin{aligned}L_1(\beta(t)) =& \mathbb{E}\big[a_1(t)[a_1(t)b_1^*(t)+a_2(t)b_2^*(t)+w]\times\\ &\tanh\big(\frac{a_1(t)b_1(t)[a_1(t)b_1^*(t)+a_2(t)b_2^*(t)+w]}{\sigma^2}\big) - a_1(t)\times\\ &[a_1(t)b_1^*(t)+w]\tanh\big(\frac{a_1(t)b_1(t)[a_1(t)b_1^*(t)+w]}{\sigma^2}\big)\big],\end{aligned}$$

and

$$\begin{aligned}L_2(\beta(t)) =& \mathbb{E}\big[a_1(t)[a_1(t)b_1^*(t)+w]\times\\ &\tanh\big(\frac{a_1(t)b_1(t)[a_1(t)b_1^*(t)+w]}{\sigma^2}\big) - a_1^2(t)b_1(t)\big],\end{aligned}$$

with $w \sim \mathcal{N}(0,\sigma^2)$ given $a_1(t)$ and $a_2(t)$. By mean-value theorem, on the one hand, we have

$$\begin{aligned}\tilde{b}_1(t)L_1(\beta(t)) = \tilde{b}_1(t)b_2^*(t)\mathbb{E}\big[a_1(t)a_2(t)\big\{\tanh(l_1(t))\\ + l_1(t)\tanh'(l_1(t))\big\}\big] \le 1.2c_0|\tilde{b}_1(t)||b_2^*(t)|,\end{aligned} \tag{59}$$

where $l_1(t) = \frac{a_1(t)b_1(t)(a_1(t)b_1(t)+a_2(t)\zeta_1(t)+w)}{\sigma^2}$, $\zeta_1(t)$ is between 0 and $b_2^*(t)$, and the last inequality holds by $|\tanh(z) + z\tanh'(z)| \le 1.2$ and Schwarz inequality. On the other hand, by Lemma 3 and Assumption 2, we have $a_1(t)b_1(t) = \mathbb{E}\big[[a_1(t)b_1(t)+w]\tanh\big(\frac{a_1(t)b_1(t)[a_1(t)b_1(t)+w]}{\sigma^2}\big)\big|a_1(t)\big]$, then it follows that

$$\begin{aligned}L_2(\beta(t)) = \mathbb{E}\big[a_1(t)[a_1(t)b_1^*(t)+w]\times\\ \tanh\big(\frac{a_1(t)b_1(t)[a_1(t)b_1^*(t)+w]}{\sigma^2}\big) - a_1(t)\times\\ [a_1(t)b_1(t)+w]\tanh\big(\frac{a_1(t)b_1(t)[a_1(t)b_1(t)+w]}{\sigma^2}\big)\big]\\ = -\tilde{b}_1(t)\mathbb{E}\big[a_1^2(t)\big\{\tanh(l_2(t)) + l_2(t)\tanh'(l_2(t))\big\}\big],\end{aligned}$$

where $l_2(t) = \frac{a_1(t)b_1(t)[a_1(t)\zeta_2(t)+w]}{\sigma^2}$, $\zeta_2(t)$ is between $b_1^*(t)$ and $b_1(t)$. By (40) and (54), we have $\min(\zeta_2^2(t), \zeta_2(t)b_1(t)) \ge \min(b_1^{*2}(0), \underline{b}^2) \triangleq \underline{c} > 0$. By taking $\hat{a} = a_1(t)b_1(t)$ and $a = a_1(t)\zeta_2(t)$ in Lemma 4, we can derive that

$$\tilde{b}_1(t)L_2(\beta(t)) \le -C\tilde{b}_1^2(t), \tag{60}$$

where $C = \mathbb{E}\big[a_1^2(t)\big(1-\exp\big(-\frac{\underline{c}a_1^2(t)}{2\sigma^2}\big)\big)\big]$ is a positive constant by Assumption 3. Then by (56)-(60), $\tilde{\beta}^2(t) = \tilde{b}_1^2(t) + b_2^{*2}(t)$ from (26), and $G = c_0I$, we have

$$\begin{aligned}\frac{d}{dt}V(\beta(t),R(t)) \le -C\tilde{b}_1^2(t) + r_1(t)\\ \le -Cc_0^{-1}V(\beta(t),R(t)) + r_2(t),\end{aligned} \tag{61}$$

where $r_1(t) = 1.2c_0|\tilde{b}_1(t)||b_2^*(t)| - b_2^*(t)h_2(\beta(t)) + \frac{1}{2}\tilde{\beta}^\tau(t)\,(G-R(t))\,\tilde{\beta}(t)$ and $r_2(t) = r_1(t) + Cb_2^{*2}(t) + Cc_0^{-1}\tilde{\beta}^\tau(t)\,(G-R(t))\,\tilde{\beta}(t)$. By (25) and (53), we have $\lim_{t\to\infty} r_2(t) = 0$. Hence, by Lemma 8 and the comparison principle [47], we obtain $\lim_{t\to\infty} V(\beta(t),R(t)) = 0$, then from the positive-definiteness property of $R(t)$, we have $\lim_{t\to\infty}\beta(t) = \beta^*$. Therefore, we obtain that $D_{c,1} = \{x : \beta = \beta^*, R = G\}$ is the invariant set with domain of attraction $D_{A,1}$.

We now provide the additional analysis for the other two cases of $x(0) \in D_{A,2}$ and $x(0) \in D_{A,3}$.

For the case of $x(0) \in D_{A,2}$, similar to the analysis for the case $x(0) \in D_{A,1}$, we can obtain that $D_{c,2} = \{x : \beta = -\beta^*, R = G\}$ is the invariant set with the domain of attraction $D_{A,2}$ by choosing the Lyapunov function as $\frac{1}{2}\tilde{\beta}^\tau(t)R(t)\tilde{\beta}(t)$ with $\tilde{\beta}(t) = \beta(t) + \beta^*$.

For the case of $x(0) \in D_{A,3}$, we have $b_1^*(0) = 0$ and $b_2^*(0) = \|\beta^*\|$. Besides, by (46), we have $\frac{db_2^*(t)}{dt} \equiv 0$ if $b_1^*(t) = 0$ and $b_2^*(t) = \|\beta^*\|$. Thus we know $b_2^*(t) \equiv \|\beta^*\|$ and $b_1^*(t) \equiv 0$, i.e., $x(t) \in D_{A,3}$ for all $t \ge 0$. Furthermore, by (26)-(27), it follows that

$$\begin{aligned}h_1(\beta(t)) = \mathbb{E}\big[a_1(t)[a_2(t)\|\beta^*\| + w]\times\\ \tanh\big(\frac{a_1(t)b_1(t)[a_2(t)\|\beta^*\| + w]}{\sigma^2}\big)\big] - \mathbb{E}[a_1^2(t)b_1(t)]\end{aligned} \tag{62}$$

Besides, by (50), we have $h_2(\beta(t)) = 0$, and (39) still holds. Thus by (25), (31), (41) and (62), we have

$$\frac{db_1(t)}{dt} = (c_0^{-1} + O(e^{-t}))h_1(\beta(t)). \tag{63}$$

By (62), we have that if $b_1(t) = 0$, then $h_1(\beta(t)) = 0$ and $\frac{dh_1(\beta(t))}{db_1(t)} > 0$, and if $b_1(t)$ is positive and large, then $h_1(\beta(t)) < 0$ and $\frac{dh_1(\beta(t))}{db_1(t)} < 0$. Thus it is clear that there exists a constant $c^* > 0$ such that $h_1(\beta(t)) < 0$ if $b_1(t) \ge c^*$ and $h_1(\beta(t)) > 0$ if $0 < b_1(t) < c^*$. So by Lemma 8, we have $\lim_{t\to\infty} b_1(t) = c^*$ if $b_1(0) > 0$. Similarly, we can get $\lim_{t\to\infty} b_1(t) = -c^*$ if $b_1(0) < 0$, and $b_1(t) \equiv 0$ if $b_1(0) = 0$. Therefore, we obtain that $D_{c,3} = \{x : \beta \in D_{us}, R = G\}$ is the invariant set with domain of attraction $D_{A,3}$, where D_{us} is defined in (24).

The above analysis shows that the assertion (24) holds. By Proposition 1, the remaining proof concerns the compact set D_s. Below, we give an explicit expression for D_s:

$$D_s = \{x : \|\beta\| \le \max\{m_0, \bar{b}\}, \|R - G\| \le \varepsilon_1\}, \tag{64}$$

where $m_0 = \sqrt{2\|\beta^*\|^2 + 2c_0^{-1}\sigma^2}$, $0 < \varepsilon_1 < \varepsilon$, $\bar{b}$ and ε are defined in (44) and (55), respectively. From $\|\beta(t)\| = b_1(t)$, (25) and (44), it is clear that the trajectory of (21) that starts in D_s remains in D_s for $t > 0$.

Step 4: Convergence of the sequence $\{\beta_k\}$

For the remaining proof, we need to verify Condition B4) in Proposition 1. By Lemma 1, it follows that $\{\phi_k\}$ is bounded i.o. with probability 1. We only need to prove that the event $\{x_k \in D_s, k \ge 0\}$ happens i.o. with probability 1.

We now analyze the properties of R_{k+1} and β_k, respectively.

Firstly, by (18) and Assumption 3, we have

$$\lim_{k\to\infty} R_{k+1} = \lim_{k\to\infty}\frac{1}{k}\sum_{t=1}^{k}\phi_t\phi_t^\tau = G, \tag{65}$$

where $G = c_0I$. Then for any $\varepsilon > 0$, the event $\|R_{k+1} - G\| \le \varepsilon$ happens i.o. with probability 1.

Secondly, we show that

$$\{\|\beta_{k+1}\| \le m_0, k \ge 0\} \text{ happens i.o. with probability 1.} \tag{66}$$

Let us denote $\Psi_k = [\phi_0^\tau\ \cdots\ \phi_k^\tau]^\tau$ and $Y_{k+1} = [\bar{y}_1\ \cdots\ \bar{y}_{k+1}]^\tau$ with $\bar{y}_{k+1}$ defined in (6). By (65), we have

$$\frac{1}{k}\Psi_k^\tau\Psi_k = \frac{1}{k}\sum_{t=0}^{k}\phi_t\phi_t^\tau \xrightarrow{k\to\infty} G,$$

then by Lemma 7, we have for sufficiently large k,

$$Y_{k+1}^\tau\Psi_k(\Psi_k^\tau\Psi_k)^{-1}\Psi_k^\tau Y_{k+1} \le Y_{k+1}^\tau Y_{k+1}. \tag{67}$$

By Algorithm 1, it is evident that

$$\beta_{k+1} = P_{k+1}\sum_{t=0}^{k}\phi_t\bar{y}_{t+1} = (P_0^{-1} + \Psi_k^\tau\Psi_k)^{-1}\Psi_k^\tau Y_{k+1},$$

thus by (67), it follows that

$$\begin{aligned}&\|\beta_{k+1}\|\\ \le&\|(P_0^{-1} + \Psi_k^\tau\Psi_k)^{-\frac{1}{2}}\|\|(P_0^{-1} + \Psi_k^\tau\Psi_k)^{-\frac{1}{2}}\Psi_k^\tau Y_{k+1}\|\\ \le&\|(\Psi_k^\tau\Psi_k)^{-\frac{1}{2}}\|\|Y_{k+1}\| = \|(\frac{1}{k}\Psi_k^\tau\Psi_k)^{-\frac{1}{2}}\|\|\frac{1}{\sqrt{k}}Y_{k+1}\|.\end{aligned} \tag{68}$$

By the definition of $\bar{y}_{k+1}$ in (6), model (2), Assumptions 1-3 and the fact $|\tanh(\cdot)| \leq 1$, we have

$$\frac{1}{k}\|Y_{k+1}\|^2 = \frac{1}{k}\sum_{t=0}^{k}\bar{y}_{t+1}^2 \leq \frac{2}{k}\sum_{t=0}^{k}[(\beta^{*\tau}\phi_t)^2 + w_{t+1}^2] \xrightarrow{k\to\infty} 2c_0\|\beta^*\|^2 + 2\sigma^2.$$

Thus we have

$$\limsup_{k\to\infty}\|\beta_{k+1}\| \leq m_0, \text{ a.s.,} \tag{69}$$

and (66) holds. Therefore, by (65) and (66), we have $\{x_k \in D_s\}$ happens i.o. with probability 1 and Condition B4) is verified. Then by Proposition 1, we have $x_k \to D_c$ as $k \to \infty$ almost surely and D_c is defined in (24).

We now prove that β_k converges to a limit point β^* or $-\beta^*$.

We first show that β_k will not converge to the unstable point set D_{us} of D_c defined in (24). To establish this, it suffices to verify that $\lim\limits_{k\to\infty}\beta_k^\tau\beta^* \neq 0$. By Algorithm 1, we can see that β_k is a rational function of random variables $\{z_0, \phi_0, w_1, \cdots, z_{k-1}, \phi_{k-1}, w_k\}$, which are jointly absolutely continuous with respect to Lebesgue measure by Assumptions 1-3. From the results in [48], it follows that β_k is also absolutely continuous with respect to Lebesgue measure. Then with the initial value of Algorithm 1 satisfying that $\beta_0 \neq 0$, we have for any $k \geq 1$,

$$P(\beta_k^\tau\beta^* \neq 0) = 1. \tag{70}$$

By (70) and the fact that the points in D_{us} are saddle points, we can derive $\beta_k^\tau\beta^* \nrightarrow 0$ (with its detailed proof in Appendix I). Thus β_k will converge to the set $\{\beta^*, -\beta^*\}$ almost surely.

We then prove that β_k converges to a limit point. Denote

$$\mathcal{F}_k = \sigma\{\phi_t, z_t, w_t, t \leq k\} \tag{71}$$

and

$$e_{k+1} = y_{k+1}\tanh\big(\frac{\beta_k^\tau\phi_k y_{k+1}}{\sigma^2}\big) - \beta_k^\tau\phi_k. \tag{72}$$

By model (2) and Assumption 2, we have

$$\mathbb{E}\left[e_{k+1}^2|\mathcal{F}_k\right] \leq 3[(\beta^{*\tau}\phi_k)^2 + (\beta_k^\tau\phi_k)^2] + 3\sigma^2. \tag{73}$$

From (7b) and (65), it follows that $\frac{1}{k}P_k^{-1} = \frac{1}{k}[\sum_{t=1}^{k-1}\phi_t\phi_t + P_0^{-1}] \to G > 0$, thus we have $\|P_k\| = O(\frac{1}{k})$ and then $\|P_{k+1}-P_k\| = \|P_{k+1}(P_{k+1}^{-1}-P_k^{-1})P_k\| = \|P_{k+1}\phi_k\phi_k^\tau P_k\| = O(\frac{\|\phi_k\|^2}{k^2})$. Hence, by (7b), we have

$$a_k\|P_k\phi_k\|^2 = tr(P_{k+1} - P_k) = O(\frac{\|\phi_k\|^2}{k^2}).$$

Moreover, by the assumption that $\{\|\phi_k\|^4\}$ is u.i. in Assumption 3, we have $\sup_{k\geq 1}\mathbb{E}\left[\|\phi_k\|^4\right] < \infty$. Thus by (7a), (69), (73), we obtain

$$\begin{aligned}&\sum_{k=1}^{\infty}\mathbb{E}[\|a_kP_k\phi_ke_{k+1}\|^2] = \sum_{k=1}^{\infty}\mathbb{E}[\mathbb{E}[\|a_kP_k\phi_ke_{k+1}\|^2|\mathcal{F}_k]]\\ \leq&3\sum_{k=1}^{\infty}\mathbb{E}\left[a_k\|P_k\phi_k\|^2[(\beta^{*\tau}\phi_k)^2 + (\beta_k^\tau\phi_k)^2 + \sigma^2]\right]\\ =&O(\sum_{k=1}^{\infty}\frac{\mathbb{E}\left[\|\phi_k\|^4\right]}{k^2}) < \infty.\end{aligned} \tag{74}$$

Since for any sequence of random variables Z_k, $\sum_{k=1}^{\infty}\mathbb{E}[|Z_k|] < \infty$ implies $\sum_{k=1}^{\infty}Z_k < \infty$ [36], we have that $\sum_{k=1}^{\infty}\|a_kP_k\phi_ke_{k+1}\|^2$ converges a.s. Thus by (7a) and (72), we have

$$\sum_{k=1}^{\infty}\|\beta_{k+1} - \beta_k\|^2 = \sum_{k=1}^{\infty}\|a_kP_k\phi_ke_{k+1}\|^2 < \infty. \tag{75}$$

Hence, $\lim\limits_{k\to\infty}\|\beta_{k+1} - \beta_k\|^2 = 0$, which means that β_k cannot jump from a small neighborhood of β^* to a small neighborhood of $-\beta^*$ infinite times. Consequently, β_k will converge to a limit point which is either β^* or $-\beta^*$ almost surely.

We now proceed to show that our algorithm can handle the case $\beta^* = 0$. At this time, $b_1^*(t) = b_2^*(t) \equiv 0$. Then by the facts $x\tanh(x) \leq x^2$, we know $h_1(\beta(t)) \leq 0$ if $b_1(t) \geq 0$ and $h_1(\beta(t)) \geq 0$ if $b_1(t) \leq 0$, where the equality holds if and only if $b_1(t) = 0$. So by taking the Lyapunov function as $b_1^2(t)$, and using (31) and Lemma 8, we can obtain $\lim\limits_{t\to\infty}b_1(t) = 0$. Moreover, Condition B4) holds for D_s in (64) for $\beta^* = 0$. Therefore, by Proposition 1, we have $\beta_k \to 0$ as $k \to \infty$ almost surely.

Therefore, we complete the proof of Theorem 1. ■

B. Proof of Theorem 2

Firstly, we prove the inequality (13). Without loss of generality, we assume that $\{\phi_k, y_{k+1}\}$ is generated by the sub-model $y_{k+1} = \beta^{*\tau}\phi_k + w_{k+1}$. We now show that if $\lim\limits_{k\to\infty}\beta_k = \beta^*$, then (13) holds. By (11), (71) and the definition of $\mathcal{I}_k^*$, the probability that $\{\phi_k, y_{k+1}\}$ is categorized correctly can be calculated as follows:

$$\begin{aligned}&P(\mathcal{I}_k = \mathcal{I}_k^*|\mathcal{F}_k)\\ =&P\big((y_{k+1} - \beta_k^\tau\phi_k)^2 \leq (y_{k+1} + \beta_k^\tau\phi_k)^2|\mathcal{F}_k\big)\\ =&P\big(\beta_k^\tau\phi_k(\beta^{*\tau}\phi_k + w_{k+1}) \geq 0|\mathcal{F}_k\big)\\ =&\Phi\left(|\beta^{*\tau}\phi_k|/\sigma\right) + [2\Phi\left(-|\beta^{*\tau}\phi_k|/\sigma\right) - 1]\mathbb{I}_{\{\beta_k^\tau\phi_k\phi_k^\tau\beta^*<0\}}.\end{aligned}$$

Thus by the convergence of β_k and Assumption 3, we have

$$\begin{aligned}&\lim_{k\to\infty}P(\mathcal{I}_k = \mathcal{I}_k^*) = \lim_{k\to\infty}\mathbb{E}\left[P(\mathcal{I}_k = \mathcal{I}_k^*|\mathcal{F}_k)\right]\\ \geq&\lim_{k\to\infty}\left[\mathbb{E}\left[\Phi\left(|\beta^{*\tau}\phi_k|/\sigma\right)\right] - 2\mathbb{E}[|\tilde{\beta}_k^\tau\phi_k|/\sigma]\right]\\ =&1 - \lim_{k\to\infty}\mathbb{E}\left[\Phi\left(-|\beta^{*\tau}\phi_k|/\sigma\right)\right]\\ \geq&1 - \mathbb{E}\left[\exp\left(-(\beta^{*\tau}\phi)^2/(2\sigma^2)\right)\right].\end{aligned} \tag{76}$$

If $\lim\limits_{k\to\infty}\beta_k = -\beta^*$, we can obtain the same result by a similar analysis as that of (76). Hence, the inequality (13) is obtained.

Secondly, we prove the inequality (14). We now show that if $\lim\limits_{k\to\infty}\beta_k = \beta^*$, (14) holds. Denote

$$\begin{aligned}&\mathcal{A}_{k,1} = \{\omega : y_{k+1} = \beta^{*\tau}\phi_k + w_{k+1}\},\\ &\mathcal{A}_{k,2} = \{\omega : y_{k+1} = -\beta^{*\tau}\phi_k + w_{k+1}\},\\ &\mathcal{A}_{k,3} = \{\omega : (y_{k+1} + \beta_k^\tau\phi_k)^2 \leq (y_{k+1} - \beta_k^\tau\phi_k)^2\} \cap \mathcal{A}_{k,1},\\ &\mathcal{A}_{k,4} = \{\omega : (y_{k+1} - \beta_k^\tau\phi_k)^2 \leq (y_{k+1} + \beta_k^\tau\phi_k)^2\} \cap \mathcal{A}_{k,2},\end{aligned}$$

where $\mathcal{A}_{k,1}$, $\mathcal{A}_{k,2}$ denote the events that the data $\{\phi_k, y_{k+1}\}$ is generated by these two sub-models, $\mathcal{A}_{k,3}$, $\mathcal{A}_{k,4}$ represent the events that the data $\{\phi_k, y_{k+1}\}$ is categorized into the wrong cluster, and thus $\mathcal{A}_{k,1} - \mathcal{A}_{k,3}$, $\mathcal{A}_{k,2} - \mathcal{A}_{k,4}$ are the events

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3649290

that the data is categorized into the correct cluster. Then the evaluation index (12) can be written as follows:

$$\begin{aligned} J_n =&\frac{1}{n}\sum_{k=1}^{n}(y_{k+1}-\beta_k^\tau\phi_k)^2\left[\mathbb{I}_{\{\mathcal{A}_{k,1}-\mathcal{A}_{k,3}\}}+\mathbb{I}_{\mathcal{A}_{k,4}}\right]\\ &+\frac{1}{n}\sum_{k=1}^{n}(y_{k+1}+\beta_k^\tau\phi_k)^2\left[\mathbb{I}_{\{\mathcal{A}_{k,2}-\mathcal{A}_{k,4}\}}+\mathbb{I}_{\mathcal{A}_{k,3}}\right]\\ =&L_{n,1}+L_{n,2}+L_{n,3}, \end{aligned} \tag{77}$$

where $L_{n,1}=\frac{1}{n}\sum_{k=1}^n\left\{(y_{k+1}-\beta_k^\tau\phi_k)^2\mathbb{I}_{\mathcal{A}_{k,1}}+(y_{k+1}+\beta_k^\tau\phi_k)^2\mathbb{I}_{\mathcal{A}_{k,2}}\right\}$, $L_{n,2}=\frac{1}{n}\sum_{k=1}^n\left\{(y_{k+1}+\beta_k^\tau\phi_k)^2-(y_{k+1}-\beta_k^\tau\phi_k)^2\right\}\mathbb{I}_{\mathcal{A}_{k,3}}$ and $L_{n,3}=\frac{1}{n}\sum_{k=1}^n\left\{(y_{k+1}-\beta_k^\tau\phi_k)^2-(y_{k+1}+\beta_k^\tau\phi_k)^2\right\}\mathbb{I}_{\mathcal{A}_{k,4}}$. We now analyze the RHS of (77) term by term. For the term $L_{n,1}$, we have the following expression:

$$\begin{aligned} L_{n,1}=&\frac{1}{n}\sum_{k=1}^{n}(\tilde\beta_k^\tau\phi_k)^2-\frac{2}{n}\sum_{k=1}^{n}\tilde\beta_k^\tau\phi_k w_{k+1}\mathbb{I}_{\mathcal{A}_{k,1}}\\ &+\frac{2}{n}\sum_{k=1}^{n}\tilde\beta_k^\tau\phi_k w_{k+1}\mathbb{I}_{\mathcal{A}_{k,2}}+\frac{1}{n}\sum_{k=1}^{n}w_{k+1}^2, \end{aligned}$$

where $\tilde\beta_k=\beta_k-\beta^*$. By Assumption 2, $\lim\limits_{n\to\infty}\frac{1}{n}\sum_{k=1}^n w_{k+1}^2=\sigma^2$. By $\lim\limits_{k\to\infty}\tilde\beta_k=0$ and the average boundedness of $\|\phi_k\|^2$ and $\|\phi_k w_{k+1}\|$ from Lemma 1, we obtain

$$\lim_{n\to\infty}L_{n,1}=\sigma^2,\ \text{a.s.} \tag{78}$$

For the term $L_{n,2}$, let us denote $\mathcal{A}_1=\{(\beta^{*\tau}\phi)^2+\beta^{*\tau}\phi w\le 0\}$ and $\mathcal{A}_2=\{y=\beta^{*\tau}\phi+w\}$, by Assumptions 1-3, we have $P\left(\mathcal{A}_1\cap\mathcal{A}_2|\phi\right)=P\left(\mathcal{A}_1|\phi\right)P\left(\mathcal{A}_2|\phi\right)=p\Phi\big(-\frac{|\beta^{*\tau}\phi|}{\sigma}\big)$. Besides, from Lemma 1, $[\beta^{*\tau}\phi_k\phi_k^\tau\beta^*+\beta^{*\tau}\phi_k w_{k+1}]\,\mathbb{I}_{\mathcal{A}_1\cap\mathcal{A}_2}$ is asymptotically stationary and ergodic. Thus by Assumptions 2-3 and (76), we obtain

$$\begin{aligned} &\lim_{n\to\infty}L_{n,2}=\lim_{n\to\infty}\frac{4}{n}\sum_{k=1}^{n}\left[\beta_k^\tau\phi_k\phi_k^\tau\beta^*+\beta^{*\tau}\phi_k w_{k+1}\right]\mathbb{I}_{\mathcal{A}_{k,3}}\\ =&4\mathbb{E}\left[(\beta^{*\tau}\phi)^2\mathbb{E}\left[\mathbb{I}_{\mathcal{A}_1\cap\mathcal{A}_2}|\phi\right]\right]+4\mathbb{E}\left[\beta^{*\tau}\phi\mathbb{E}\left[w\mathbb{I}_{\mathcal{A}_1\cap\mathcal{A}_2}|\phi\right]\right]\\ =&4\mathbb{E}\big[(\beta^{*\tau}\phi)^2P\left(\mathcal{A}_1\cap\mathcal{A}_2|\phi\right)\big]\\ &+4p\mathbb{E}\left[\frac{|\beta^{*\tau}\phi|}{\sqrt{2\pi}\sigma}\int_{-\infty}^{-|\beta^{*\tau}\phi|}w\exp\left(-\frac{w^2}{2\sigma^2}\right)dw\right]\\ =&4p\mathbb{E}\left[\eta(\phi)\right]. \end{aligned} \tag{79}$$

Similarly, for the term $L_{n,3}$, we have

$$\lim_{n\to\infty}L_{n,3}=4(1-p)\mathbb{E}\left[\eta(\phi)\right]. \tag{80}$$

Furthermore, since for any positive constant a, $\int_{-\infty}^{-a}a\exp\big(-\frac{x^2}{2\sigma^2}\big)dx\le\sigma^2\exp\big(-\frac{a^2}{2\sigma^2}\big)$, we have $\eta(\phi)\le 0$. Thus by (77)-(80), (14) is obtained. Similarly, if $\lim\limits_{k\to\infty}\beta_k=-\beta^*$, we also have (14). Therefore, we complete the proof. ■

C. Proofs of Theorem 3 and Theorem 4

Denote $x_k=[\theta_{k,1}^\tau\ \theta_{k,2}^\tau\ \text{vec}^\tau(R_k)]^\tau$, similar to (19), it is not difficult to obtain that x_k evolves according to the following dynamical systems:

$$x_{k+1}=x_k+\frac{1}{k}Q(x_k,\phi_k,y_{k+1}), \tag{81}$$

where $Q(x_k,\phi_k,y_{k+1})$ is determined via Algorithm 2. In order to analyze (81) using the ODE method, we introduce the corresponding ODEs as follows:

$$\frac{d}{dt}\theta_1(t)=R^{-1}(t)f_1(\theta_1(t)), \tag{82a}$$

$$\frac{d}{dt}\theta_2(t)=R^{-1}(t)f_2(\theta(t)), \tag{82b}$$

$$\frac{d}{dt}R(t)=G-R(t), \tag{82c}$$

where $f_1(\theta_1(t))=\lim\limits_{k\to\infty}\mathbb{E}\left[\phi_k\left(y_{k+1}-\theta_1^\tau(t)\phi_k\right)\right]$, $f_2(\theta(t))=\lim\limits_{k\to\infty}\mathbb{E}\big[\phi_k\big((y_{k+1}-\theta_1^\tau(t)\phi_k)\tanh\Big(\frac{\theta_2^\tau(t)\phi_k(y_{k+1}-\theta_1^\tau(t)\phi_k)}{\sigma^2}\Big)-\theta_2^\tau(t)\phi_k\big)\big]$, $\theta(t)=(\theta_1(t),\theta_2(t))$ and $G=\lim\limits_{k\to\infty}\mathbb{E}\left[\phi_k\phi_k^\tau\right]$.

Before giving the proof of Theorem 3, a related lemma is given as follows:

Lemma 6: Under Assumptions 1′ and 2-3, Conditions B1)-B3) in Proposition 1 are all satisfied in the open area $D=\{x:R>0\}$, where $x=[\theta_1^\tau\ \theta_2^\tau\ \text{vec}^\tau(R)]^\tau$.

Proof of Theorem 3: We will analyze the convergence of $\theta_{k,1}$ and $\theta_{k,2}$ separately by verifying all conditions of Proposition 1. By Lemma 6, it remains to prove the requirements on the trajectories of ODEs (82) and Condition B4).

Step 1: Convergence analysis of the sequence $\{\theta_{k,1}\}$.

It is clear that the trajectory generated by (82c) is the same as that of the ODE (21b), which evolves according to (25). We now establish the stability result of (82a). For this, let us construct the Lyapunov function $V_1(\theta_1(t),R(t))=\frac{1}{2}\tilde\theta_1^\tau(t)R(t)\tilde\theta_1(t)$ with $\tilde\theta_1(t)=\theta_1(t)-\theta_1^*$. Then we have

$$\frac{dV_1(\theta_1(t),R(t))}{dt}=\tilde\theta_1^\tau(t)f_1(\theta_1(t))+\frac{1}{2}\tilde\theta_1^\tau(t)[G-R(t)]\tilde\theta_1(t). \tag{83}$$

For the first term on the RHS of (83), by (9), (82), we have

$$\begin{aligned} &\tilde\theta_1^\tau(t)f_1(\theta_1(t))=\lim_{k\to\infty}\mathbb{E}\big[\tilde\theta_1^\tau(t)\phi_k\big(y_{k+1}-\theta_1^\tau(t)\phi_k\big)\big]\\ =&\lim_{k\to\infty}\mathbb{E}\left[\tilde\theta_1^\tau(t)\phi_k\big(\mathbb{E}\left[y_{k+1}|\phi_k\right]-\theta_1^\tau(t)\phi_k\big)\right]\\ =&-\tilde\theta_1^\tau(t)\lim_{k\to\infty}\mathbb{E}\left[\phi_k\phi_k^\tau\right]\tilde\theta_1(t)=-c_0\|\tilde\theta_1(t)\|^2, \end{aligned} \tag{84}$$

where the last equality holds by Assumption 3 and c_0 is a positive constant defined in (25). Then it follows that

$$\frac{d}{dt}V_1(\theta_1(t),R(t))=-2V_1(\theta_1(t),R(t))+r(t),$$

where $r(t)=\frac{1}{2}\tilde\theta_1^\tau(t)\left(R(t)-G\right)\tilde\theta_1(t)$ tends to 0 by (25). Thus by Lemma 8, we have $\lim\limits_{t\to\infty}V_1(\theta_1(t),R(t))=0$. From the positive-definiteness property of $R(t)$, it follows that $\lim\limits_{t\to\infty}\theta_1(t)=\theta_1^*$. Therefore, we obtain that the ODE (82a) has the invariant set $D_{c,1}'=\{[\theta_1^{*\tau}\ \text{vec}^\tau(G)]^\tau\}$ with the domain of attraction $D'=\{v=[\theta^\tau\ \text{vec}^\tau(R)]^\tau:R>0\}$.

Denote $D_s'=\{v=[\theta^\tau\ \text{vec}^\tau(R)]^\tau:\|\theta\|\le m_1,\varepsilon_1 I\le R\le\varepsilon_2 I\}$ with $m_1=\sqrt{3c_0(\|\beta_1^*\|^2+\|\beta_2^*\|^2)+3\sigma^2}$ and $0<\varepsilon_1<\varepsilon_2$. It is clear that the trajectories of (82) that starts in D_s' remains in D_s'. Besides, similar to (69), we have $\limsup_{k\to\infty}\|\theta_{k+1,1}\|\le m_1$, a.s. Thus Condition B4) is verified. By Proposition 1, we have

$$\lim_{k\to\infty}\theta_{k,1}=\theta_1^*,\ \text{a.s.} \tag{85}$$

Step 2: Convergence analysis of the sequence $\{\theta_{k,2}\}$. The proof is similar to that of Theorem 1 and we only need to establish the stability of the ODEs (82b). We just provide the analysis for $\theta_2(0) \in D_{A,1}$, and omit the analysis for $\theta_2(0) \in D_{A,2}$ and $\theta_2(0) \in D_{A,3}$.

For this, consider the Lyapunov function $V_2(\theta_2(t), R(t)) = \frac{1}{2}\tilde{\theta}_2(t)^\tau R(t)\tilde{\theta}_2(t)$ with $\tilde{\theta}_2(t) = \theta_2(t) - \theta_2^*$. Then we have

$$\frac{dV_2(\theta_2(t), R(t))}{dt} = \tilde{\theta}_2^\tau(t) f_2(\theta(t)) + \frac{1}{2}\tilde{\theta}_2^\tau(t)[G - R(t)]\tilde{\theta}_2(t). \tag{86}$$

We now analyze the first term on the RHS of (86). Denote $m_{k+1}^* = y_{k+1} - \theta_1^{*\tau}\phi_k$, then by (8), we have $m_{k+1}^* = z_k\theta_2^{*\tau}\phi_k + w_{k+1}$. Moreover, by (82), we obtain

$$f_2(\theta(t)) = h_1(\theta(t)) + h_2(\theta(t)), \tag{87}$$

where $h_1(\theta(t)) = \lim\limits_{k\to\infty} \mathbb{E}\big[\phi_k\big(m_{k+1}^* \tanh\big(\frac{\theta_2^\tau(t)\phi_k m_{k+1}^*}{\sigma^2}\big) - \theta_2^\tau(t)\phi_k\big)\big]$ and $h_2(\theta(t)) = f_2(\theta(t)) - h_1(\theta(t))$. Since the analysis of $\tilde{\theta}_2^\tau(t)h_1(\theta(t))$ is the same as that of $\tilde{\beta}^\tau(t)f(\beta(t))$ in (56), from (61) in Theorem 1, we have

$$\tilde{\theta}_2^\tau(t)h_1(\theta(t)) \leq -C\|\tilde{\theta}_2(t)\|^2 + Cb_2^{*2}(t) + 1.2\|\tilde{\theta}_2(t)\||b_2^*(t)|, \tag{88}$$

where C is a positive constant defined in (60) and the term $b_2^*(t)$ will tend to zero. Let $\tilde{\theta}_{k,1} = \theta_{k,1} - \theta_1^*$. By mean-value theorem, Schwarz inequality and the fact that $|\tanh(z) + z\tanh'(z)| \leq 1.2$, we obtain that

$$\begin{aligned} &\tilde{\theta}_2^\tau(t)h_2(\theta(t)) = \tilde{\theta}_2^\tau(t)\left(f_2(\theta(t)) - h_1(\theta(t))\right) \\ =&\lim_{k\to\infty} \mathbb{E}\big[\tilde{\theta}_2^\tau(t)\phi_k\phi_k^\tau\tilde{\theta}_{k,1}\big[\tanh(\bar{l}_k(t)) + \bar{l}_k(t)\tanh'(\bar{l}_k(t))\big]\big] \\ \leq& c_2\|\tilde{\theta}_2(t)\| \lim_{k\to\infty}\left(\mathbb{E}\|\tilde{\theta}_{k,1}\|^2\right)^{1/2}, \end{aligned} \tag{89}$$

where $\bar{l}_k(t) = \frac{\theta_2^\tau(t)\phi_k(y_{k+1} - \zeta_k(t))}{\sigma^2}$, $\zeta_k(t)$ is between $\theta_1^{*\tau}\phi_k$ and $\theta_1^\tau(t)\phi_k$, and $c_2 = 1.2\sqrt{\mathbb{E}\|\phi\|^4}$. Then by (86)-(89) and $G = c_0 I$, we obtain

$$\frac{dV_2(\theta_2(t), R(t))}{dt} \leq -Cc_0^{-1}V_2(\theta_2(t), R(t)) + r(t), \tag{90}$$

where $r(t) = Cb_2^{*2}(t) + 1.2\|\tilde{\theta}_2(t)\||b_2^*(t)| + c_2\|\tilde{\theta}_2(t)\| \lim\limits_{k\to\infty}(\mathbb{E}\|\tilde{\theta}_{k,1}\|^2)^{1/2} + (\frac{1}{2} - Cc_0^{-1})\tilde{\theta}_2^\tau(t)(G - R(t))\tilde{\theta}_2(t)$. By (25), (85), and the fact $b_2^*(t) \to 0$, we have $\lim\limits_{t\to\infty} r(t) = 0$. Thus by Lemma 8, it follows that $\lim\limits_{t\to\infty} V_2(\theta_2(t), R(t)) = 0$, and by the positiveness-definite property of $R(t)$ from (25), we have $\lim\limits_{t\to\infty}\theta_2(t) = \theta_2^*$, and the assertion (24) (replacing β^* with θ_2^*) holds.

Then by Proposition 1, we have

$$\lim_{k\to\infty}\theta_{k,2} = \theta_2^*, \text{ or } \lim_{k\to\infty}\theta_{k,2} = -\theta_2^*, \text{ a.s.} \tag{91}$$

Thus, the results of Theorem 3 can be obtained. ■

Proof of Theorem 4: The proof is similar to the way used in Theorem 2, which is omitted. ■

VI. Simulation Results

In this section, we conduct simulations for the asymmetric MLR problem to verify the effectiveness of our algorithm. Consider the data $\{\phi_k, y_{k+1}\}_{k=1}^\infty$ generated by the following dynamical model:

$$\begin{aligned} y_{k+1} &= \beta_1^{*\tau}\phi_k\mathbb{I}_{\{z_k=1\}} + \beta_2^{*\tau}\phi_k\mathbb{I}_{\{z_k=-1\}} + w_{k+1}, \\ \phi_{k+1} &= 0.5\phi_k + e_{k+1}, \end{aligned}$$

where $\phi_k \in \mathbb{R}^{10}$, z_k is i.i.d with $P(z_k = 1) = P(z_k = -1) = 0.5$, $e_{k+1} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_{10})$, $w_{k+1} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$. It is clear that the regressor $\{\phi_k\}$ is dependent, and all assumptions in Theorem 3 are satisfied.

Firstly, we conduct Algorithm 2 to estimate the unknown parameters β_1^* and β_2^*. The estimation error is defined by $\tilde{\beta}_{k,i} = \beta_{k,i} - \beta^*_{\bar{\mathcal{I}}_{k,i}}, (i = 1, 2)$ with $\bar{\mathcal{I}}_{k,i} = \arg\min_{j=1,2}\{\|\beta_{k,i} - \beta_j^*\|\}$. From Fig.1, one can see that estimation errors $\tilde{\beta}_{k,i} (i = 1, 2)$ tend to zero along the time k, and the within-cluster error (17) also decreases to the noise variance σ^2 along the time k. Moreover, we note that $\bar{\mathcal{I}}_{k,i}, (i = 1, 2)$ are convergent, i.e., $(\beta_{k,1}, \beta_{k,2})$ will converge to a limit point belonging to the set $\{(\beta_1^*, \beta_2^*), (\beta_2^*, \beta_1^*)\}$, which demonstrates the effectiveness of our algorithm.

Secondly, we compare the performance of Algorithm 2 with the population EM algorithm, which is employed in most investigations for the MLR problem. The population EM with the finite number of samples [29] is executed as follows:

E-step:

$$\alpha_{k,t}^i = \frac{\exp\left(-\frac{(y_{k+1} - \beta_{t,i}^\tau\phi_k)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(y_{k+1} - \beta_{t,1}^\tau\phi_k)^2}{2\sigma^2}\right) + \exp\left(-\frac{(y_{k+1} - \beta_{t,2}^\tau\phi_k)^2}{2\sigma^2}\right)},$$

M-step:

$$\beta_{t+1,i} = \left(\frac{1}{n}\sum_{k=1}^n \alpha_{k,t}^i\phi_k\phi_k^\tau\right)^{-1}\left(\frac{1}{n}\sum_{k=1}^n \alpha_{k,t}^i\phi_k y_{k+1}\right),$$

where $i = 1, 2$, $\beta_{t,i}$ is the estimate of β_i^* at the iteration step t, $\alpha_{k,t}^i$ is the conditional probability of $\{\phi_k, y_{k+1}\}$ belongs to the i-th sub-model based on the current estimate $\beta_{t,i}$, $\{\phi_k, y_{k+1}\}_{k=1}^n$ is a collection of n samples, and n is often chosen sufficiently large to approximate the population EM.

In our simulation of the population EM algorithm, we choose the number of samples n to be 5000 and the total iteration step T to be 20. Both Algorithm 2 and the population EM algorithm are initialized with the same values. Specifically, for $i = 1, 2$ and $j = 1, \cdots, 10$, $\beta_{0,i}^j$ is sampled from a uniform distribution $U(\beta_i^{*j} - \kappa, \beta_i^{*j} + \kappa)$, where $\beta_{0,i}^j$ and β_i^{*j} are the j-th element of $\beta_{0,i}$ and β_i^*, respectively. It can be seen that the parameter κ measures the distance between the initial values and the true parameter set. For each simulation with a given κ in $[0, 20]$, we run 500 independent realizations and plot the convergence probability of the algorithms, i.e., the proportion of 500 simulations that converge to true parameters, about the parameter κ in Fig.2. From the simulation results, we see that the convergence probability of our algorithm does not depend on the parameter κ, while the convergence probability of the population EM algorithm will decrease to zero as κ increases. The results show that the estimates generated by Algorithm 2 will converge to the set $\{(\beta_1^*, \beta_2^*), (\beta_2^*, \beta_1^*)\}$ for any non-zero initial values, while the population EM algorithm does not.

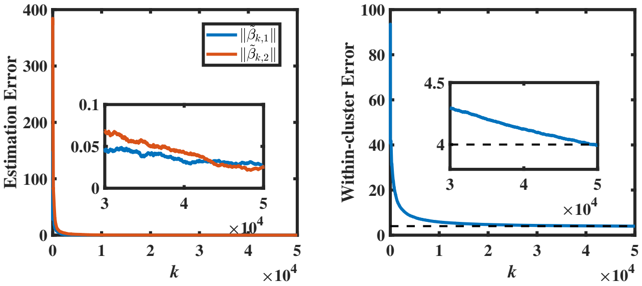

Fig. 1. Estimation error and clustering performance under Algorithm 2.

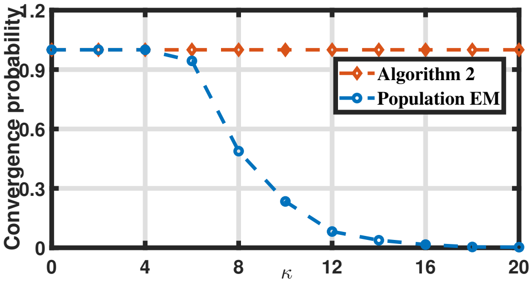

Fig. 2. Comparison of convergence probabilities of Algorithm 2 and the population EM algorithm.

VII. Conclusion

In this paper, we investigated the online identification and data clustering problems of two classes of MLRs. For the symmetric MLR problem, we proposed an online identification algorithm based on the EM principle, and for the first time, established the convergence to the true parameter set for any non-zero initial value without imposing i.i.d data assumption. For the asymmetric MLR problem, we designed a two-step online identification algorithm that separately estimates two parts of the model and obtained the corresponding global convergence to the true parameter set for the first time. Furthermore, based on the parameter estimates, we showed that the data clustering performance is asymptotically the same as the case where the true parameters are known. We note that it is difficult to identify the MLR model with more than two components or with two unbalanced components since the log-likelihood criterion function becomes a more complicated non-convex one which does not possess the nice property used in the symmetric MLR where the local maxima are the true parameters of the sub-models. For the general MLR, we are only able to establish a local convergence result [49], and more research efforts are called for in future investigations. Besides, there are several interesting problems that need to be investigated, for example, how to relax the asymptotically stationary and ergodic assumptions on the regressors, and how to address the adaptive control problem of MLR models.

Appendix I

Lemma 7: [50] Suppose that Ψ and Y are any $n \times l$ and $n \times r$-dimensional matrices, respectively and $\Psi^\tau \Psi$ is invertible. Then we have $Y^\tau \Psi(\Psi^\tau \Psi)^{-1} \Psi^\tau Y \leq Y^\tau Y$.

Lemma 8: [47] Consider the following ODE:

$$\frac{dV(t)}{dt} = -V(t) + r(t),$$

where $\lim\limits_{t\to\infty} r(t) = 0$, then we have $\lim\limits_{t\to\infty} V(t) = 0$.

A. Proof of Some Equalities and Lemmas

Proof of (15): By Assumption 3 and the property of Gaussian distribution, we have that $\beta^{*\tau}\phi \sim \mathcal{N}(0, \beta^{*\tau}\Sigma\beta^*)$. By simple calculations, we have $\mathbb{E}\big[\exp\big(-\frac{(\beta^{*\tau}\phi)^2}{2\sigma^2}\big)\big] = \frac{\sigma}{\sqrt{\sigma^2+\beta^{*\tau}\Sigma\beta^*}}$.Thus (15) holds. ■

Proof of Lemma 1: For the first statement, by Assumptions 1-3, from the stationary and independent properties among z_k, ϕ_k and w_{k+1}, it follows that $\{d_k\}$ approaches a stationary process asymptotically with bounded fourth moment. Besides, $\{d_k\}$ is ergodic since each element is ergodic. For the second statement of this lemma, it is not difficult to obtain that there exists a sequence of shift transformations $\{T_k\}$ such that $\lim_{k\to\infty} T_k = T$ and T is measure-preserving. Then following the proof-line (accompanied with a measure-preserving transformation) of the result that any measurable function of a stationary ergodic stochastic process is stationary and ergodic [36], we obtain the desired result. ■

Proof of Lemma 2: Let us first verify Condition B1). For $x \in D$ and $x_i \in \mathcal{U}(x, \rho(x))$ $(i = 1, 2)$, by (19), we have

$$\|Q_1(x_1, \phi, y) - Q_1(x_2, \phi, y)\| \leq \mathcal{R}_1(x, \phi, y, \rho)\|x_1 - x_2\|, \tag{92}$$

where

$$\begin{aligned}\mathcal{R}_1 &= \sup_{\mathcal{U}(x,\rho)} \left[\Big\|\frac{\partial Q_1(x,\phi,y)}{\partial \beta}\Big\| + \Big\|\frac{\partial Q_1(x,\phi,y)}{\partial R}\Big\|\right] \\ &\leq \sup_{\mathcal{U}(x,\rho)} \left[\frac{\|\phi\|^2}{\lambda_{\min}(R)}\left(\frac{y^2}{\sigma^2}+1\right) + \frac{\|\phi\||y| + \|\phi\|^2\|\beta\|}{\lambda^2_{\min}(R)}\right]\end{aligned}$$

with $\rho(x)$ being sufficiently small such that any point $\bar{x} = [\ \bar{\beta}^\tau \quad \text{vec}^\tau(\bar{R})\ ]^\tau$ in the area $\mathcal{U}(x, \rho(x))$ has the property that $\lambda_{\min}(\bar{R}) > 0$ by the perturbation theorem [51], and the second inequality holds by $\frac{\partial R^{-1}}{\partial R} = -R^{-1} \otimes R^{-1}$ [52],

$$\left\|\frac{\partial R^{-1}}{\partial R}\right\| = \|R^{-1} \otimes R^{-1}\| = \frac{1}{\lambda^2_{\min}(R)},$$

where $\otimes$ is the Kronecker product for matrices. Thus by (20) and (92), we have

$$\|Q(x_1, \phi, y) - Q(x_2, \phi, y)\| \leq \mathcal{R}(x, \phi, y)\|x_1 - x_2\|,$$

where $\mathcal{R} = \mathcal{R}_1 + 1$. Thus for fixed ϕ and y, $Q(x, \phi, y)$ is locally Lipschitz continuous with respect to $x \in D$, and Condition B1) is satisfied.

For the verification of Condition B2), since $\mathcal{R}$ is defined as the Lipschitz constant, we only need to verify its upper bound to satisfy B2), which is quite obvious because $\|\phi_k\| y_{k+1}$, $\|\phi_k\|^2 y^2_{k+1}$ and $\|\phi_k\|^2$ are all asymptotically stationary and ergodic by Lemma 1 and model (2), and the supremum over $\rho(x)$ only concerns with β and R and is independent of the time instant k.

The verification of Condition B3) is straightforward by Lemma 1 and model (2), and the details are omitted. This completes the proof of Lemma 2. ■

Proof of Lemma 3: From the facts that $z\tanh(z)$ is an even function in z and $\mathbb{E}_{y\sim\frac{1}{2}\mathcal{N}(a,\sigma^2)+\frac{1}{2}\mathcal{N}(-a,\sigma^2)}\left[y\tanh\left(\frac{ay}{\sigma^2}\right)\right]=a$, we can easily obtain the desired results. ■

Proof of Lemma 5: From the facts that $\tanh(z)$ and $z\tanh'(z)$ are bounded for $z\in\mathbb{R}$, it follows that $\frac{\partial f(c,x,w)}{\partial x}$ is bounded. Then we have that

$$\begin{aligned}\frac{dF(c,x)}{dx}&=\int_{-\infty}^{\infty}\frac{\partial f(c,x,w)}{\partial x}\exp\big(-\frac{w^2}{2\sigma^2}\big)dw\\ =&\int_{-\infty}^{\infty}\Big[\tanh\big(\frac{c(w+x)}{\sigma^2}\big)-\tanh\big(\frac{c(w-x)}{\sigma^2}\big)\Big]\\ &\exp\big(-\frac{w^2}{2\sigma^2}\big)dw+\int_{-\infty}^{\infty}\big[\frac{c(w+x)}{\sigma^2}\tanh'\big(\frac{c(w+x)}{\sigma^2}\big)\\ &\quad-\frac{c(w-x)}{\sigma^2}\tanh'\big(\frac{c(w-x)}{\sigma^2}\big)\big]\exp\big(-\frac{w^2}{2\sigma^2}\big)dw\\ \triangleq&L_1(c,x)+L_2(c,x).\end{aligned}$$

Since $\tanh(z)$ is odd and increasing, we have for $x>0$, $L_1(c,x)=2\int_0^{\infty}\big[\tanh\big(\frac{c(w+x)}{\sigma^2}\big)-\tanh\big(\frac{c(w-x)}{\sigma^2}\big)\big]\exp\big(-\frac{w^2}{2\sigma^2}\big)dw>0$. Moreover, for $x>0$ and $w>0$, we have $\exp\big(-\frac{(w-x)^2}{2\sigma^2}\big)>\exp\big(-\frac{(w+x)^2}{2\sigma^2}\big)$. By this inequality and the fact $z\tanh'(z)$ is an odd function in z, it follows that $L_2(c,x)=2\int_0^{\infty}\frac{cw}{\sigma^2}\tanh'\big(\frac{cw}{\sigma^2}\big)\big[\exp\big(-\frac{(w-x)^2}{2\sigma^2}\big)-\exp\big(-\frac{(w+x)^2}{2\sigma^2}\big)\big]dw>0$. Lemma 5 thus is proven. ■

Proof of (52): We just provide the proof for the first part of (52), i.e., $\frac{db_2^*(t)}{dt}\le 0$ if $0\le b_2^*(t)<\|\beta^*\|$, and the second part can be obtained by following a similar way. By (25), (42) and (55), we have $\|v_2(t)(R^{-1}(t)-c_0^{-1}I)f(\beta(t))\|\le e^{-t}\frac{1}{3}c_0^{-1}(p_1+c_0\bar{b})\triangleq\bar{c}_2e^{-t}$. Denote $\bar{S}(\beta(t))=c_0^{-1}h_2(\beta(t))-\bar{c}_2e^{-t}$. By (48), we have $\frac{db_2^*(t)}{dt}\le-\frac{b_1^*(t)}{b_1(t)}\bar{S}(\beta(t))$. To prove $\frac{db_2^*(t)}{dt}\le 0$, it suffices to show that $\bar{S}(\beta(t))\ge 0$ for $t\ge 0$. From Lemma 5 and its proof, it is clear that there exists a positive constant $\bar{m}_1$ such that for any $x>0$, $c>0$, we have $0<\frac{dF(c,x)}{dx}\le\bar{m}_1$. Then by (50), we have

$$\begin{aligned}\frac{dh_2(\beta(t))}{db_2^*(t)}\le&\frac{2}{\sqrt{2\pi}\sigma}\int_{a_1(t)>0}\int_{a_2(t)>0}a_2^2(t)\\ &\sup_{x(t)}\left|\frac{dF\big(c(t),x(t)\big)}{dx(t)}\right|\tilde{g}(a_1(t),a_2(t))da_2(t)da_1(t)\le\bar{c}_1,\end{aligned}\tag{93}$$

where $\bar{c}_1=\frac{c_0\bar{m}_1}{2\sqrt{2\pi}\sigma}$. Besides, we have $|\frac{dh_2(\beta(t))}{dt}|=|\frac{dh_2(\beta(t))}{db_2^*(t)}\times\frac{db_2^*(t)}{dt}|\le\bar{c}_1\frac{\|\beta^*\|}{\underline{b}}|\bar{S}(\beta(t))|$. Denote $\bar{c}_3=c_0^{-1}\bar{c}_1\frac{\|\beta^*\|}{\underline{b}}$, we have

$$\frac{d\bar{S}(\beta(t))}{dt}=c_0^{-1}\frac{dh_2(\beta(t))}{dt}+\bar{c}_2e^{-t}\ge-\bar{c}_3|\bar{S}(\beta(t))|+\bar{c}_2e^{-t}.$$

Thus we can obtain $\frac{d\bar{S}(\beta(t))}{dt}\big|_{\bar{S}(\beta(t))=0}\ge\bar{c}_2e^{-t}\ge 0$. Using the equality $\bar{S}(\beta(0))>0$ derived from (55), it is clear that $\bar{S}(\beta(t))\ge 0$ for all $t\ge 0$. ■

Proof of Lemma 6: The lemma can be obtained similarly to Lemma 2, and proof details are omitted here. ■

Proof of $\beta_k^{\tau}\beta^*\nrightarrow 0$: We prove $\beta_k^{\tau}\beta^*\nrightarrow 0$ by contradiction. Firstly, for $\bar{x}=[\bar{\beta}^{\tau}\ \mathrm{vec}^{\tau}(\bar{R})]^{\tau}\in D_A$, from (29), (31) and (55), we have that if $0<|\bar{\beta}^{\tau}\beta^*|<b_l$ and $\|\bar{R}-c_0I\|\le\varepsilon$, then there exists a positive constant α such that $\bar{\beta}^{\tau}\beta^*\beta^{*\tau}\bar{R}^{-1}f(\bar{\beta})>\alpha|\bar{\beta}^{\tau}\beta^*|>0$. For any integer $n>0$ and any $\Delta>0$, we define $m(n,\Delta)=\max\{m:\sum_{i=n}^{m}\frac{1}{i}\le\Delta\}$. If $\beta_k^{\tau}\beta^*\to 0$, then by (70), for sufficiently large n, there exist positive constants δ_1, δ_2 and δ_3 such that $\delta_1<|\beta_n^{\tau}\beta^*|<\delta_2<b_l$ and $\|\beta_m-\beta_n\|<\delta_3$ for $m\in[n,m(n,\Delta)]$. It is clear that δ_3 can be sufficiently small when Δ is sufficiently small. Besides, by (65), we have $\|R_n-c_0I\|\le\varepsilon$ for large n, and then we have $\beta_n^{\tau}\beta^*\beta^{*\tau}R_n^{-1}f(\beta_n)>\delta_1\alpha$ for large n. Secondly, from (18) and (21), we have $\beta_{m(n,\Delta)+1}=\beta_n+\sum_{i=n}^{m(n,\Delta)}\frac{1}{i}Q_1(x_i,\phi_i,y_{i+1})=\beta_n+\sum_{i=n}^{m(n,\Delta)}\frac{1}{i}R_n^{-1}f(\beta_n)+L_1(n,\Delta,x_n)+L_2(n,\Delta,x_n)$, where $L_1(n,\Delta,x_n)=\sum_{i=n}^{m(n,\Delta)}\frac{1}{i}[Q_1(x_n,\phi_i,y_{i+1})-R_n^{-1}f(\beta_n)]$ and $L_2(n,\Delta,x_n)=\sum_{i=n}^{m(n,\Delta)}\frac{1}{i}[Q_1(x_i,\phi_i,y_{i+1})-Q_1(x_n,\phi_i,y_{i+1})]$. By Lemma 2 and the boundedness of x_n in (65) and (66), we have $\frac{1}{\Delta}L_1(n,\Delta,x_n)\to 0$ as $n\to\infty$, and $|L_2(n,\Delta,x_n)|\le\mathcal{R}_1\Delta\max_{m\in[n,m(n,\Delta)]}\{|x_m-x_n|\}\le\mathcal{R}_1\delta_3\Delta$. Thus it is evident that $|\beta_{m(n,\Delta)+1}^{\tau}\beta^*|\ge|\beta_n^{\tau}\beta^*|+\frac{1}{2}\delta_1\alpha\Delta-o(\Delta)-\mathcal{R}_1\delta_3\Delta>|\beta_n^{\tau}\beta^*|>\delta_1$ for sufficiently small Δ and large n. Thus there exists a subsequence $|\beta_{n_k}^{\tau}\beta^*|$ with a positive lower bound $\frac{\delta_1}{2}$, which contradicts with $\beta_k^{\tau}\beta^*\to 0$. This completes the proof. ■

REFERENCES

[1] R. E. Quandt and J. B. Ramsey, “Estimating mixtures of normal distributions and switching regressions,” *Journal of the American Statistical Association*, vol. 73, no. 364, pp. 730–738, 1978.

[2] S. Gaffney and P. Smyth, “Trajectory clustering with mixtures of regression models,” in *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 63–72.

[3] P. Deb and A. M. Holmes, “Estimates of use and costs of behavioural health care: A comparison of standard and finite mixture models,” *Health Economics*, vol. 9, no. 6, pp. 475–489, 2000.

[4] X. Chai, S. Shan, X. Chen, and W. Gao, “Locally linear regression for pose-invariant face recognition,” *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1716–1725, 2007.

[5] Q. Li, R. Shi, and F. Liang, “Drug sensitivity prediction with high-dimensional mixture regression,” *PLoS One*, vol. 14, no. 2, p. e0212108, 2019.

[6] W. Chang, C. Wan, Y. Zang, C. Zhang, and S. Cao, “Supervised clustering of high-dimensional data using regularized mixture modeling,” *Briefings in Bioinformatics*, vol. 22, no. 4, p. bbaa291, 2021.

[7] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino, “A bounded-error approach to piecewise affine system identification,” *IEEE Transactions on Automatic Control*, vol. 50, no. 10, pp. 1567–1580, 2005.

[8] H. Nakada, K. Takaba, and T. Katayama, “Identification of piecewise affine systems based on statistical clustering technique,” *Automatica*, vol. 41, no. 5, pp. 905–913, 2005.

[9] S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, and R. Vidal, “Identification of hybrid systems a tutorial,” *European Journal of Control*, vol. 13, no. 2-3, pp. 242–260, 2007.

[10] P. M. Pardalos and V. A. Yatsenko, *Optimization and Control of Bilinear Systems: Theory, Algorithms, and Applications*. Springer Science & Business Media, 2010.

[11] F. Xue and L. Guo, “Necessary and sufficient conditions for adaptive stablizability of jump linear systems,” *Communications in Information and Systems*, vol. 1, no. 2, pp. 205–224, 2001.

[12] B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, “Strong consistency and rate of convergence of switched least squares system identification for autonomous markov jump linear systems,” *IEEE Transactions on Automatic Control*, vol. 69, no. 6, pp. 3952–3959, 2024.

[13] T. Sarkar, A. Rakhlin, and M. Dahleh, “Nonparametric system identification of stochastic switched linear systems,” in *IEEE 58th Conference on Decision and Control*, 2019, pp. 3623–3628.

[14] Y. Sattar, Z. Du, D. A. Tarzanagh, S. Oymak, L. Balzano, and N. Ozay, “Certainty equivalent quadratic control for Markov jump systems,” in *American Control Conference*, 2022, pp. 2871–2878.

[15] N. Ozay, C. Lagoa, and M. Sznaier, “Set membership identification of switched linear systems with known number of subsystems,” *Automatica*, vol. 51, pp. 180–191, 2015.

[16] A. Moradvandi, R. E. Lindeboom, E. Abraham, and B. De Schutter, "Models and methods for hybrid system identification: A systematic survey," in *IFAC-PapersOnLine*, vol. 56, no. 2, 2023, pp. 95–107.

[17] X. Yi, C. Caramanis, and S. Sanghavi, "Alternating minimization for mixed linear regression," in *International Conference on Machine Learning*, 2014, pp. 613–621.

[18] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *Journal of Machine Learning Research*, vol. 15, pp. 2773–2832, 2014.

[19] A. T. Chaganty and P. Liang, "Spectral experts for estimating mixtures of linear regressions," in *International Conference on Machine Learning*, 2013, pp. 1040–1048.

[20] H. Sedghi, M. Janzamin, and A. Anandkumar, "Provable tensor methods for learning mixtures of generalized linear models," in *Artificial Intelligence and Statistics*, 2016, pp. 1223–1231.

[21] K. Zhong, P. Jain, and I. S. Dhillon, "Mixed linear regression with multiple components," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[23] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[24] Y. Li and Y. Liang, "Learning mixtures of linear regressions with nearly optimal complexity," in *Conference on Learning Theory*, 2018, pp. 1125–1144.

[25] Y. Chen, X. Yi, and C. Caramanis, "A convex formulation for mixed regression with two components: Minimax optimal rates," in *Conference on Learning Theory*, 2014, pp. 560–604.

[26] T. Wang and I. C. Paschalidis, "Convergence of parameter estimates for regularized mixed linear regression models," in *IEEE 58th Conference on Decision and Control*, 2019, pp. 3664–3669.

[27] E. A. Cohen, *The Influence of Nonharmonic Partials on Tone Perception*. Stanford University, 1980.

[28] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[29] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *The Annals of Statistics*, vol. 45, no. 1, pp. 77–120, 2017.

[30] J. M. Klusowski, D. Yang, and W. Brinda, "Estimating the coefficients of a mixture of two linear regressions by expectation maximization," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3515–3524, 2019.

[31] J. Kwon, W. Qian, Y. Chen, C. Caramanis, D. Davis, and N. Ho, "Global optimality of the EM algorithm for mixtures of two-component linear regressions," *IEEE Transactions on Information Theory*, 2024.

[32] J. Kwon and C. Caramanis, "EM converges for a mixture of many linear regressions," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 1727–1736.

[33] P. Zilber and B. Nadler, "Imbalanced mixed linear regression," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[34] L. Guo, "Feedback and uncertainty: Some basic problems and results," *Annual Reviews in Control*, vol. 49, pp. 27–36, 2020.

[35] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Transactions on Automatic Control*, vol. 22, no. 4, pp. 551–575, 1977.

[36] W. F. Stout, *Almost Sure Convergence*. Academic Press, 1974.

[37] R. D. De Veaux, "Mixtures of linear regressions," *Computational Statistics & Data Analysis*, vol. 8, no. 3, pp. 227–245, 1989.

[38] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[39] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*. Birkhsträsuser, 1991.

[40] Z. Luo and A. Hashemi, "Structural properties, cycloid trajectories and non-asymptotic guarantees of EM algorithm for mixed linear regression," *arXiv preprint arXiv:2511.04937*, 2025.

[41] K. W. Fang, *Symmetric Multivariate and Related Distributions*. CRC Press, 2018.

[42] K. J. Åström and B. Wittenmark, "On self-tuning regulators," *Automatica*, vol. 9, no. 2, pp. 185–199, 1973.

[43] A. Becker, P. Kumar, and C.-Z. Wei, "Adaptive control with the stochastic approximation algorithm: Geometry and convergence," *IEEE Transactions on Automatic Control*, vol. 30, no. 4, pp. 330–338, 1985.

[44] L. Guo and H.-F. Chen, "The Åström-Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers," *IEEE Transactions on Automatic Control*, vol. 36, no. 7, pp. 802–812, 1991.

[45] G. Zielke, "Inversion of modified symmetric matrices," *Journal of the Association for Computing Machinery*, vol. 15, no. 3, p. 402–408, 1968.

[46] C. Daskalakis, C. Tzamos, and M. Zampetakis, "Ten steps of EM suffice for mixtures of two gaussians," in *Conference on Learning Theory*, 2017, pp. 704–710.

[47] W. Hahn, *Stability of Motion*. Springer, 1967.

[48] S. Meyn and P. Chines, "The zero divisor problem of multivariable stochastic adaptive control," *Systems & Control Letters*, vol. 6, no. 4, pp. 235–238, 1985.

[49] Y. Liu, Z. Liu, and L. Guo, "Convergence of online learning algorithm for a mixture of multiple linear regressions," in *International Conference on Machine Learning*, vol. 235, 2024, pp. 31 516–31 540.

[50] D. Huang and L. Guo, "Estimation of nonstationary ARMAX models based on the Hannan-Rissanen method," *The Annals of Statistics*, vol. 18, no. 4, pp. 1729–1756, 1990.

[51] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.

[52] J. E. Gentle, *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer Cham, 2008.

Yujing Liu received the B.E. degree in automation from Wuhan University in 2019, and the Ph.D. degree in system analysis and integration from Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS) in 2024. She is currently a Post-Doctoral Research Associate in AMSS, CAS.

Her research interests include the identification of stochastic mixed models, and distributed adaptive control of stochastic systems.

Zhixin Liu received the B.S. degree in mathematics from Shandong University in 2002, and the Ph.D. degree in control theory from Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS) in 2007. She is currently a Professor of AMSS, CAS, and the director of Key Laboratory of Systems and Control, CAS. She had visiting positions with the KTH Royal Institute of Technology, Stockholm, Sweden, University of New South Wales, Kensington, Australia, and University of Maryland, College Park, MD, USA. She is a coauthor of the SIGEST Paper in SIREV in 2014, and the co-recipient of the Best Theoretical Paper Award at the 13rd World Congress on Intelligent Control and Automation (WCICA). Her current research interests include multiagent systems, distributed control, and distributed estimation and filtering.

Lei Guo (M'88-SM'96-F'99) received his B.S. degree in mathematics from Shandong University in 1982 and Ph.D. degree in control theory from the Chinese Academy of Sciences (CAS) in 1987. He is currently a professor of the Academy of Mathematics and Systems Science, CAS.

He is a fellow of IEEE, a member of CAS, a foreign member of the Royal Swedish Academy of Engineering Sciences, and a fellow of the International Federation of Automatic Control (IFAC). He was awarded an honorary doctorate by Royal Institute of Technology (KTH, Sweden) in 2014, and the Hendrik W. Bode Lecture Prize by the IEEE Control Systems Society in 2019. He has served as a Council Member of IFAC, General Co-Chair of the 48th IEEE-CDC, President of China Society for Industrial and Applied Mathematics (CSIAM), and is on the editorial boards of a number of academic journals in systems and control.

His research interests include stochastic systems, adaptive identification, adaptive control, adaptive filtering, adaptive game theory, control of uncertain nonlinear systems, feedback capability, multi-agent systems, and game-based control systems.