# Convergence of Online Learning Algorithm for a Mixture of Multiple Linear Regressions

Yujing Liu [1 2] Zhixin Liu [1 2] Lei Guo [1 2]

## Abstract

Mixed linear regression (MLR) is a powerful model to characterize nonlinear relationships among observed data while still being simple and computationally efficient. This paper investigates the online learning and data clustering problem for MLR model with an arbitrary number of sub-models and arbitrary mixing weights. Previous investigations mainly focus on offline learning algorithms, and the convergence results are established under the independent and identically distributed (i.i.d.) input data assumption. To overcome these fundamental limitations, we propose a novel online learning algorithm for parameter estimation based on the EM principle. By using Ljung's ODE method and Lyapunov stability theorem, we first establish the almost sure convergence results of the proposed algorithm without the traditional i.i.d. assumption on the input data. Furthermore, by using the stochastic Lyapunov function method, we also provide its convergence rate analysis for the first time. Finally, we analyze the performance of online data clustering based on the parameter estimates, which is asymptotically the same as that in the case of known parameters.

## 1. Introduction

Learning the input-output relationship based on the observed data is a fundamental issue in various fields including statistical learning, system identification, and computer science (Hastie et al., 2009; Bishop & Nasrabadi, 2006). The mixed linear regression (MLR) model is a powerful technique for characterizing the nonlinear input-output relationship by utilizing a combination of multiple linear sub-models (Kong et al., 2020; Diamandis et al., 2021; Pal et al., 2022). It was first proposed as a generalization of "switching regression" (Quandt, 1972), and has found broad applications in real-world scenarios including musical perception (Cohen, 1980), trajectory clustering (Gaffney & Smyth, 1999), market segmentation (Wedel & Kamakura, 2000) and healthcare analysis (Deb & Holmes, 2000). In MLR, each input-output data is generated from one of the unknown linear regression models and the label of data, i.e., which sub-model the data comes from, is also unknown. Developing algorithms to learn the unknown parameters based on the observed data and categorize the new data into the correct cluster are crucial for the learning and prediction of MLR.

**Related Works.** The learning problem for MLR has attracted much attention of researchers from various fields including statistical learning (Balakrishnan et al., 2017), computer science (Ingrassia et al., 2014) and biological analysis (Sun et al., 2022), although it is acknowledged to be NP-hard in the absence of statistical assumptions on the observed data (Yi et al., 2014). Among the existing research, tensor-based and expectation maximization (EM)-based methods are two commonly used approaches to design learning algorithms for MLR models, and their convergence results are mostly established under the i.i.d. standard Gaussian assumption on the input data.

The tensor-based method, first proposed by Anandkumar et al. (2014), transfers the parameter learning problem to an orthogonal decomposition of a symmetric tensor derived from the moments. By the tensor-based method, the exact recovery guarantee for MLR with multiple sub-models has been established under some specific structural assumptions, such as linearly independence or near orthonormality of the sub-model parameters (Chaganty & Liang, 2013; Yi et al., 2016). However, the tensor-based method suffers from high sample and computational complexity, making it difficult to apply in practice. In addition, the statistical error of the estimator exhibits a scaling behavior relative to the norm of the sub-model parameters.

The EM algorithm, consisting of E-step and M-step, is a general technique for parameter learning of latent variable models (Dempster et al., 1977). The E-step is used to com-

[1] The Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China [2] School of Mathematical Sciences, University of Chinese Academy of Sciences,Beijing, China. Correspondence to: Zhixin Liu <lzx@amss.ac.cn>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

pute the expectation of the log-likelihood function based on current estimates, and the M-step is used to update parameter estimates by maximizing the function derived from the E-step. The EM algorithm has the advantages of simple expression and low computational complexity, while exhibiting favorable performance in addressing MLR problems in practice (De Veaux, 1989; Faria & Soromenho, 2010). In the theoretical aspect, some progress has been made in the EM algorithm to solve the learning problem of MLR models under the i.i.d. standard Gaussian assumption on the input data. Specifically, for the case of MLR with two sub-models where the sub-model parameters, denoted as $\beta_1^*$ and $\beta_2^*$, satisfy the symmetric constraint $\beta_1^* = -\beta_2^*$, local convergence results have been established by using the population EM algorithm (Balakrishnan et al., 2017; Klusowski et al., 2019), and further global convergence results have been established (Kwon et al., 2019). For the asymmetric case of MLR with two sub-models, a characterization of the local domain of attraction for the convergence of EM has been obtained in the noise-less case (Yi et al., 2014). Furthermore, for the general case of MLR with multiple sub-models and unknown mixing weights, local convergence of the EM algorithm has been established (Kwon & Caramanis, 2020), and in addition, a novel EM-type algorithm with a bounded learning error guarantee has been proposed (Zilber & Nadler, 2023).

**Limitations of Existing Work.** To summarize, all the above-mentioned theoretical investigations on the MLR learning problem have several common features.

Firstly, the input data is required to be i.i.d. with a standard Gaussian distribution, which often fails to align with real-world scenarios (Li & Liang, 2018). Here are two examples. One example comes from the trajectory clustering of MLR model in real-world data sets including movements of objects in video sequences (Gaffney & Smyth, 1999) and cyclone paths (Gaffney & Smyth, 2003), where the trajectory data at the current time is affected by that at the previous time and will in turn affect the future data, therefore the i.i.d. data assumption is obviously not satisfied. The second example comes from the MLR model in learning and model prediction control of hybrid systems, where the dynamical signals are also non-i.i.d. because of the existence of the feedback control loops (Bemporad, 2023). Some efforts have been made to relax the i.i.d. standard Gaussian data assumption. For example, for the MLR learning problem, Li and Liang (2018) assumed that the input data is i.i.d. Gaussian with different covariances, while Sedghi et al. (2016) assumed that the input data is i.i.d. with a known and continuous distribution. However, both of them still require the independence assumption on the input data. Recently, Liu et al. (2023) have relaxed the independence assumption of the input data to stationary and ergodic, but they only specifically focused on the MLR with two sub-models.

Secondly, the computational algorithms, such as the population EM algorithm (Balakrishnan et al., 2017; Klusowski et al., 2019) and the EM algorithm with a finite number of samples (Kwon & Caramanis, 2020), are of off-line character. In fact, the population EM algorithm requires infinite samples to approximate specific expectations at each iteration, which makes it impractical. Although the EM algorithm with a finite number of samples can reduce the number of samples required at each iteration, it still remains off-line and also introduces a statistical error associated with the sample size. Furthermore, with the phenomenal growth in big data sets in recent years, many real-world applications (e.g., social media) require a model to incrementally learn from a non-i.i.d. stream of data. In those cases, offline algorithms like population EM are infeasible as the whole data set is needed at every E-step. In contrast to off-line algorithms, online algorithms can be updated conveniently based on both the current estimate and newly emerged input-output observation, without requiring storage of all the old data and with lower computational cost (Chen et al., 2018; Karimi et al., 2018; 2019). Due to this advantage, online learning has received extensive and increasing attention in the machine learning community, besides its necessity in adaptive signal processing and adaptive control problems. Therefore, it is valuable to design an online learning algorithm with a convergence guarantee.

The goal of this paper is to design an online learning algorithm for MLR with multiple sub-models and further establish its corresponding convergence results without the i.i.d. input data assumption.

**Challenges and Contributions.** In this paper, we put forward an online learning algorithm based on the EM principle for MLR with multiple sub-models and arbitrary mixing weights. We remark that Ljung's ODE method is a general analysis framework for the convergence of online learning algorithms (Ljung, 1977). By leveraging this method and making efforts to establish the stability of corresponding ordinary differential equations (ODEs) by the Lyapunov stability theorem (Khalil, 2002), we are able to establish the convergence result for the online parameter learning algorithm without the i.i.d. input data assumption, and also provide the convergence rate analysis. Furthermore, we analyze the online data clustering performance based on the proposed learning algorithm.

The main contributions can be summarized as follows:

- Based on the EM principle, we propose a novel online learning algorithm for the parameter estimation of MLR with multiple sub-models and arbitrary mixing weight. The algorithm alternates between computing the posterior probability of which cluster the new data belongs to based on the current parameter estimates and updating the parameter estimates using the incoming data and its corresponding posterior probability.

- We establish the convergence results of the proposed online learning algorithm without the i.i.d. input data assumption. To the best of our knowledge, this is the first convergence result of online algorithms for MLR with multiple sub-models without the i.i.d. input data assumption. Based on the convergence property of the proposed algorithm and the stochastic Lyapunov function method, we further establish the convergence rate analysis for the first time.

- Based on the parameter estimates, 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, is asymptotically the same as that in the case of known parameters.

The remainder of this paper is organized as follows. Section 2 presents the problem formulation. The main results are stated in Section 3. The proof sketch of the main results is provided in Section 4 with its details in Appendix A. A numerical example is given in Section 5. Finally, we conclude the paper in Section 6.

## 2. Problem Formulation

### 2.1. Basic Notations

In this paper, $[m]$ is used to denote the set $\{1, 2, \cdots, m\}$. The symbol $v \in \mathbb{R}^d$ represents a $d$-dimensional column vector, $v^\tau$ and $\|v\|$ denote its transpose and Euclidean norm, respectively. For a matrix $A \in \mathbb{R}^{d\times d}$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalues respectively, $\|A\|$ denotes the operator norm induced by the Euclidean norm, i.e., $(\lambda_{\max}(AA^\tau))^{1/2}$. We say that $A$ is positive (negative) semi-definite, denoted as $A \geq 0 (A \leq 0)$ if all its eigenvalues are non-negative (non-positive).

In a probability space $(\Omega, \mathcal{F}, P)$, $\Omega$ is the sample space, the collection of events is referred to as the $\sigma$-algebra $\mathcal{F}$ on $\Omega$ and $P$ is a probability measure on $(\Omega, \mathcal{F})$. For an $\mathcal{F}$-measurable set $\mathcal{A}$, its complement $\mathcal{A}^c$ is defined by $\mathcal{A}^c = \Omega - \mathcal{A}$. The indicator function $\mathbb{I}_{\mathcal{A}}$ on $\Omega$ is defined by $\mathbb{I}_{\mathcal{A}} = 1$ if the event $\mathcal{A}$ occurs and $\mathbb{I}_{\mathcal{A}} = 0$ otherwise. The event $\mathcal{A}$ is said to happen almost surely (a.s.) if $P(\mathcal{A}) = 1$. Besides, a sequence of random variables $\{x_k, k \geq 0\}$ is called uniformly intergrable (u.i.) if $\lim_{a\to\infty} \sup_{k\geq 1} \int_{[|x_k|>a]} |x_k| dP = 0$. According to the convention, the mathematical expectation operator is denoted as $E\{\cdot\}$, and the conditional mathematical expectation operator given the event $\mathcal{A}$ is denoted as $E\{\cdot|\mathcal{A}\}$. The notation $x \sim F$ indicates that the random variable $x$ follows the distribution $F$, $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with the mean $\mu$ and the variance $\sigma^2$.

We need the following definition of asymptotically stationary and ergodic in the analysis:

**Definition 2.1.** Let $\{x_k, k \geq 1\}$ be a sequence of random variables. If for any $\varepsilon > 0$ 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

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

for all $k \geq K$, then we say that the sequence $\{x_k\}$ is asymptotically stationary. Furthermore, if $\lim_{k\to\infty} E\|x_k\|$ exists and

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

then we say it is also ergodic.

*Remark* 2.2. If both parameters $\varepsilon$ and $K$ in Definition 2.1 can take the value 0, then $\{x_k, k \geq 1\}$ is called stationary and ergodic, which is consistent with the traditional definition of stationarity and ergodicity in Stout (1974).

### 2.2. Problem Statement

Let us consider the following MLR model with multiple sub-models:

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

where $z_k \in [m]$ is the latent variable, namely, we do not know which sub-model the data $\{\phi_k, y_{k+1}\}$ comes from, and $m$ is the number of sub-models. Besides, $\beta_i^*(i \in [m])$ are unknown true parameters to be estimated, $\phi_k \in \mathbb{R}^d$, $y_{k+1} \in \mathbb{R}$ and $w_{k+1} \in \mathbb{R}$ represent the regressor vector, observation vector and the system noise at the time instant $k$, respectively.

Given the streaming data $\{\phi_k, y_{k+1}\}_{k=1}^{\infty}$, we aim at proposing an online learning algorithm to estimate the true parameters $\beta_i^*(i \in [m])$, and then providing the corresponding convergence guarantees for the learning algorithm. Based on the estimates of $\beta_i^*(i \in [m])$, we further investigate the data clustering performance.

### 2.3. Assumptions

For our purpose, we introduce the following assumptions on the true parameters $\beta_i^*(i \in [m])$, the latent variable $z_k$, the noise $w_{k+1}$, and the regressor $\phi_k$:

**Assumption 2.3.** The true parameter $\beta_i^*$ is an interior point of a known and disjoint convex compact set $D_i$ for $i \in [m]$.

*Remark* 2.4. Under the structural assumption that the true parameters are linearly independent, tensor-based methods can be used to learn the true parameters $\beta_i^*$ (Chaganty & Liang, 2013; Yi et al., 2016). The tensor-based method requires an infinite number of samples at each iteration to precisely estimate parameters in MLR model, which brings high sample and computational complexity issues.

However, the tensor operations based on finite samples can help us derive a crude yet satisfactory initialization region $D_i$ around the true parameters in the permutation sense.

**Assumption 2.5.** The sequence of latent variables $\{z_k\}$ is i.i.d. with the distribution $P(z_k = i) = \pi_i^* > 0$ for $i \in [m]$ and $\sum_{i=1}^m \pi_i^* = 1$. Furthermore, $z_k$ is independent of $\phi_k$ for each $k \geq 0$.

*Remark* 2.6. Most previous investigations typically assume that the mixture probabilities of each sub-model are equal (Klusowski et al., 2019), and leverage this prior information to facilitate the convergence analysis. This paper considers a more challenging case where the proportions of each sub-model, i.e., $\pi_i^*(i \in [m])$, can be unequal and their true values can be unknown to us.

**Assumption 2.7.** The sequence of noise $\{w_{k+1}\}$ is i.i.d. with Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Besides, $w_{k+1}$ is independent of $\{z_t, t \leq k\}$ and $\{\phi_t, t \leq k\}$ for each $k \geq 0$.

*Remark* 2.8. Most previous investigations assume that the noise variance $\sigma^2$ is known (Klusowski et al., 2019; Kwon & Caramanis, 2020). In this paper, we also provide an online learning algorithm for MLR with unknown $\sigma^2$.

**Assumption 2.9.** The sequence of regressors $\{\phi_k\}$ is asymptotically stationary and ergodic with the probability density function (p.d.f) $g_k(x)$ satisfying $\lim_{k\to\infty} g_k(x) = g(x)$ and $\int xx^\tau g(x)dx > 0$, and there exists a non-random positive definite matrix $\Sigma$ such that $g(\Sigma^{-1/2}x)$ is a function of $\|\Sigma^{-1/2}x\|$ only. Additionally, the sequence $\{\|\phi_k\|^4\}$ is uniformly integrable.

*Remark* 2.10. The stationary p.d.f $g(x)$ of the regressor in Assumption 2.9 can contain a variety of distributions, such as Gaussian distribution $\mathcal{N}(0, \Sigma)$, uniform distribution, Polynomial and Laplace distributions, the rotation-invariant distributions (Qian et al., 2019) and the elliptically symmetric distributions (Fang et al., 2018).

*Remark* 2.11. Assumption 2.9 is much weaker than the i.i.d. Gaussian data assumption used in most previous works (Kwon & Caramanis, 2020). Here we give an example of $\{\phi_k\}$ that is not i.i.d. but satisfies our Assumption 2.9:

**Example 1:** Consider the following standard stochastic linear dynamical system widely used in automatic control and many other practical fields:

$$\phi_{k+1} = A\phi_k + e_{k+1}, \tag{2}$$

where the matrix $A$ is stable, and $e_k$ is i.i.d. with a Gaussian distribution. It is easy to see that the observed state or output signal sequence $\{\phi_k\}$ is stationary and ergodic, but not i.i.d. because $\phi_k$ and $\phi_{k+1}$ are strongly correlated.

### 2.4. Online Projected-EM Algorithm

In this subsection, we first construct an online learning algorithm based on the EM principle for the MLR model (1) with known noise variance $\sigma^2$ and then we also provide an online algorithm for the MLR model (1) with unknown $\sigma^2$.

To start with, let Assumptions 2.5, 2.7 and 2.9 be satisfied and the regressor be i.i.d.[1] Then for the MLR model (1) with the parameters $\beta^* = [\ \beta_1^{*\tau} \ \cdots \ \beta_m^{*\tau}\ ]^\tau$ and $\pi^* = [\ \pi_1^* \ \cdots \ \pi_m^*\ ]^\tau$, we can derive the likelihood function of the observed data $\mathbb{O}_n = \{y_1, \cdots, y_{n+1}\}$ given the input data $\mathbb{U}_n = \{\phi_1, \cdots, \phi_n\}$ as follows:

$$\begin{aligned}\mathcal{L}(\beta^*, \pi^*) &= P(\mathbb{O}_n|\mathbb{U}_n) = \prod_{k=1}^n P(y_{k+1}|\phi_k) \\ &= \prod_{k=1}^n \sum_{i=1}^m \left[\frac{\pi_i^*}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_{k+1} - \beta_i^{*\tau}\phi_k)^2}{2\sigma^2}\right)\right].\end{aligned} \tag{3}$$

Since the likelihood function is non-convex, it is hard to obtain an explicit form of the maximum likelihood estimation (MLE). Therefore, we adopt the classical EM algorithm (Dempster et al., 1977) to approximate the MLE.

Denote $\beta_k = [\beta_{k,1}^\tau \ \cdots \ \beta_{k,m}^\tau]^\tau$ and $\pi_k = [\pi_{k,1} \ \cdots \ \pi_{k,m}]^\tau$ with $\beta_{k,i}$ and $\pi_{k,i}$ being the estimates of $\beta_i^*$ and $\pi_i^*$ at the time instant $k$, respectively. The EM algorithm is conducted according to the following two steps:

1) E-step: compute the log-likelihood function $\mathcal{Q}_k$ for the complete data set $\{y_{t+1}, z_t, \phi_t\}_{t=1}^k$ as follows:

$$\begin{aligned}\mathcal{Q}_k(\beta^*, \pi^*) &= k\log(\frac{1}{\sqrt{2\pi}\sigma}) \\ &+ \sum_{i=1}^m \sum_{t=1}^k \alpha_{t,i}\left[\log(\pi_i^*) - \frac{(y_{t+1} - \beta_i^{*\tau}\phi_t)^2}{2\sigma^2}\right],\end{aligned} \tag{4}$$

where

$$\begin{aligned}\alpha_{t,i} &= P(z_t = i|\phi_t, y_{t+1}, \beta_t) \\ &= \frac{\pi_{t,i}\exp\left(-\frac{(y_{t+1} - \beta_{t,i}^\tau\phi_t)^2}{2\sigma^2}\right)}{\sum_{j=1}^m \pi_{t,j}\exp\left(-\frac{(y_{t+1} - \beta_{t,j}^\tau\phi_t)^2}{2\sigma^2}\right)}\end{aligned} \tag{5}$$

is the probability that the data $\{\phi_t, y_{t+1}\}$ belongs to the $i$-th sub-model based on the current estimate $\beta_t$ and $\pi_t$.

2) M-step: update the estimates of parameters $\beta^*$ and $\pi^*$ based on $\alpha_{k,i}$:

$$\begin{aligned}\beta_{k+1,i} &= \arg\max_{\beta_i^*} \mathcal{Q}_k(\beta^*, \pi^*) \\ &= \left(\sum_{t=1}^k \alpha_{t,i}\phi_t\phi_t^\tau\right)^{-1}\left(\sum_{t=1}^k \alpha_{t,i}\phi_t y_{t+1}\right),\end{aligned} \tag{6a}$$

$$\pi_{k+1,i} = \arg\max_{\pi_i^*, \sum_{i=1}^m \pi_i^* = 1} \mathcal{Q}_k(\beta^*, \pi^*) = \frac{1}{k}\sum_{t=1}^k \alpha_{t,i}. \tag{6b}$$

[1] Note that the i.i.d. assumption on the regressor is used only for the construction of our algorithm, which will not be used in the theoretical analysis and main results of the algorithm.

It is clear that the equation (6a) is like the standard formula of LS with the input-output data $\{\bar{\phi}_k, \bar{y}_{k+1}\}$ with $\bar{\phi}_k = \alpha_{k,i}\phi_k$ and $\bar{y}_{k+1} = \alpha_{k,i}y_{k+1}$. Thus, using the same way as that in the derivation of the recursive LS (Lai & Wei, 1982), we can obtain the online version of (6a). Besides, with simple calculations, we also obtain the online version of (6b) as $\pi_{k+1,i} = \pi_{k,i} + (\alpha_{k,i} - \pi_{k,i})/k$. The whole online EM algorithm is presented in Algorithm 1.

**Algorithm 1** Online Projected-EM Algorithm

```
1: Initialization: $\beta_{0,i}, P_{0,i} > 0, i \in [m]$.
2: At each time step $k+1$, we have data $\{\phi_k, y_{k+1}\}$.
3: Recursively calculate the estimates for $i \in [m]$:
     $\beta_{k+1,i} = \Pi_{D_i}[\beta_{k,i} + a_{k,i}\alpha_{k,i}P_{k,i}\phi_k(y_{k+1} - \beta_{k,i}^\tau\phi_k)],$
     $P_{k+1,i} = P_{k,i} - a_{k,i}\alpha_{k,i}P_{k,i}\phi_k\phi_k^\tau P_{k,i},$
     $a_{k,i} = \frac{1}{1 + \alpha_{k,i}\phi_k^\tau P_{k,i}\phi_k},$
     $\pi_{k+1,i} = \Pi_{D_{m+1}}[\pi_{k,i} + \frac{1}{k}(\alpha_{k,i} - \pi_{k,i})],$        (7)
   where $\alpha_{k,i}$ is defined in (5), $\Pi_{D_i}(\cdot), i \in [m]$ and
   $\Pi_{D_{m+1}}(\cdot)$ are the projection operators defined on the
   convex compact set $D_i = \{x | \|x - \beta_i^*\| \le d_i\}, i \in [m]$
   and $D_{m+1} = \{x | \pi_{\min} \le x \le 1\}$, respectively.
```

*Remark* 2.12. Notice that the online learning algorithm with only one sample at each iteration is susceptible to unbound stochastic noise. In order to solve this problem, we introduce a projection operator $\Pi_{D_i}$ to restrict the estimates within the given region $D_i (i \in [m+1])$. Additionally, we can choose $\pi_{\min}$ small enough such that $\pi_i^* \in D_{m+1}$ for all $i \in [m]$. As mentioned in Remark 2.4, the region $D_i$ can be obtained by the tensor-based method under linear independent assumptions on the true parameters $\beta_i^*$.

When the noise variance $\sigma^2$ is unknown, we can give its estimate $\sigma_k^2$ at the time instant $k$ as follows:

$$\sigma_{k+1}^2 = \sigma_k^2 + \frac{1}{k}\left[\left(y_{k+1} - \sum_{i=1}^{m}\alpha_{k,i}\beta_{k,i}^\tau\phi_k\right)^2 - \sigma_k^2\right].$$

Based on this, we can update the estimates $\beta_{k,i}$ and $\pi_{k,i}$, $i \in [m]$ by replacing the term $\sigma^2$ with its estimate $\sigma_k^2$ at the time instant $k$ in Algorithm 1.

## 3. Main Results

In this section, we first establish the convergence and the convergence rate results of Algorithm 1 for the MLR model with known noise variance $\sigma^2$ in Subsection 3.1, and then present the performance analysis of data clustering based on the parameter estimates in Subsection 3.2. The convergence results and the corresponding analysis of the online learning algorithm of MLR with unknown $\sigma^2$ is similar to that of MLR with known $\sigma^2$, thus we omit them for space limitations.

Before presenting the main results, we introduce the following notations for $i \in [m]$ and $j \in [m]$:

$$R_{ij}^* = \|\beta_i^* - \beta_j^*\|, \ \ R_{\min}^* = \min_{i,j,i\neq j} R_{ij}^*,$$
$$R_{\max}^* = \max_{i,j} R_{ij}^*, \ \ d_{\max} = \max_i d_i,$$

where $d_i$ is defined in Algorithm 1.

### 3.1. Convergence of Algorithm 1

We give the following main theorem on the convergence of the online learning Algorithm 1.

**Theorem 3.1.** *Under Assumptions 2.3-2.9, the estimates $\beta_{k,i}$ and $\pi_{k,i}$ ($i \in [m]$) generated by Algorithm 1 will converge locally to the true parameters, i.e.,*

$$\lim_{k\to\infty} \beta_{k,i} = \beta_i^* \ \ \textit{and} \ \ \lim_{k\to\infty} \pi_{k,i} = \pi_i^*, \ \textit{a.s.}$$

*Remark* 3.2. Firstly, the almost sure convergence result for MLR with multiple sub-models established in Theorem 3.1 is stronger than the convergence result in the high probability sense established in most prior investigations. Secondly, for the MLR learning problem, prior works (Balakrishnan et al., 2017; Kwon & Caramanis, 2020) mainly focused on the local convergence property of the population EM under the i.i.d. standard Gaussian assumption of the input data. Here we establish the convergence results for the online learning algorithm without requiring this traditional i.i.d. input data assumption. Thirdly, note that it is difficult to determine which radius of local convergence in the prior work and our work is larger, since 1) Both radii of local convergence in the analysis are of existence rather than constructiveness nature; 2) We have used a more general assumption on the regressors, which is weaker than the i.i.d. standard Gaussian assumption used in the prior work. Of course, if the regressor is assumed to be i.i.d. Gaussian in our work, we can obtain a radius of local convergence no smaller than that in the prior work, even though we utilize an online learning algorithm and the prior work adopts an algorithm in an expectation form.

*Remark* 3.3. To our knowledge, the previous investigations that need i.i.d. data assumption do not apply to the non-i.i.d. data case in the analysis of the stochastic recursive algorithm in our paper, because the analysis of the stochastic recursive algorithm can no longer be reduced to the deterministic case in the general non-i.i.d. case. To overcome this difficulty, we have adopted Ljung's ODE method (Ljung, 1977) in the field of adaptive systems, where the system signals do not satisfy the i.i.d. condition. However, to verify the conditions required by the ODE method is still challenging, which constitutes one of the main tasks of our paper.

*Remark* 3.4. Note that the likelihood function (3) is nonconvex, thus without a satisfactory initialization region, the EM algorithm, even the population EM, may not have a desired convergence property in general. The requirement on the initialization region $D_i$ ensures that only the true parameters are the stationary points in the initialization region, which is a key point to establish the convergence result in theory. From the proof of Theorem 3.1, it is evident that the signal-to-noise ratio $R^*_{\min}/\sigma$, the number of sub-models $m$ and also the smallest proportion of sub-models $\pi_{\min}$ significantly influence the requirements of $D_i$.

*Remark* 3.5. As mentioned in Remark 2.4, there are several investigations focused on the derivation of $D_i$ by utilizing the tensor-based method (Pearson, 1894; Yi et al., 2016; Chaganty & Liang, 2013), which can help us to choose an appropriate $D_i$. These investigations are under two assumptions: 1) $\{\phi_k\}$ is i.i.d. with a standard Gaussian distribution; 2) $\beta_i^*, i \in [m]$ are linearly independent, along with the following two key facts: $E\{y^2(\phi\phi^\tau - I)\} = \sum_{i=1}^m 2\pi_i^*\beta_i^*\beta_i^{*\tau}$, $E\{y^3\phi\otimes\phi\otimes\phi\} - \sum_{j=1}^d E\{y^3(e_j\otimes\phi\otimes e_j + e_j\otimes e_j\otimes \phi + \phi\otimes\ e_j\otimes\ e_j)\} = \sum_{i=1}^m 6\pi_i^*\beta_i^*\otimes\beta_i^*\otimes\beta_i^*$. Therefore, one can use the tensor-based method (Pearson, 1894; Zhong et al., 2016) to recover the parameters to a given precision with a requirement on the number of samples.

We now describe the derivation of $D_i$ under the asymptotically stationary and ergodic assumption on the regressor by the tensor-based method. There are some previous explorations on this problem. For example, under the assumption that the regressor is i.i.d. with a general continuous distribution, Sedgh et al. (2016) have constructed a novel third-order cross-moments, which can be used to obtain $D_i$ by the tensor decomposition method. The principle component analysis result established by Chen (2002) implies that under the asymptotically stationary and ergodic assumption on the regressor, one can still obtain $D_i$. By utilizing these two results, it is possible to obtain the theoretical guarantee for the tensor-based initialization technique under our assumption, which will be considered in our future work.

For the convergence rate analysis, we need the following additional but fairly general assumptions on the regressors:

**Assumption 3.6.** The sequence of regressor $\{\phi_k\}$ satisfies the following conditions:

(1) $\inf\limits_{k\geq 0} E\{\phi_k\phi_k^\tau|\mathcal{F}_{k-1}\} \triangleq \underline{c} > 0$, $\sup\limits_{k\geq 0} E\{\|\phi_k\|^4|\mathcal{F}_{k-1}\} \triangleq \bar{c} < \infty$;

(2) There exist positive constants $\gamma$, $\delta$ and $b_0$ such that for any $0 < b < b_0$ and any $k \geq 0$, we have

$$\sup_{\|\beta\|=1} P((\beta^\tau\phi_k)^2 < b|\mathcal{F}_{k-1}) \leq \gamma b^\delta. \tag{8}$$

*Remark* 3.7. The condition (8) effectively says that the conditional probability density function of $\phi_k$ given $\mathcal{F}_{k-1}$ should be free of discrete and singular components, which admits a quite large class of continuous distributions, e.g., Gaussian, uniform, etc. More discussions about (8) can be found in (Niedzwiecki & Guo, 1991). We now give two examples of Assumption 2.9 and 3.6:

**Example 2**: $\{\phi_k\}$ is i.i.d. with $E\{\|\phi_1\|^{4+\kappa}\} < \infty$ for a small constant $\kappa > 0$ and also satisfies the distribution assumption required by Assumption 2.9.

**Example 3**: $\{\phi_k\}$ is generated by the stable dynamical system (2), where $\{e_k\}$ is i.i.d. with a bounded elliptically contoured distribution (Fang et al., 2018), e.g., the uniform distribution defined on an ellipse.

Based on the convergence result in Theorem 3.1, we now provide its convergence rate in the following theorem:

**Theorem 3.8.** *Let the conditions of Theorem 3.1 and Assumption 3.6 hold. If* $\frac{\pi_{\min}R^*_{\min}\underline{c}}{\sigma\bar{c}} \geq C > 0$, *then we have*

$$\|\tilde{\beta}_{k,i}\|^2 = O\left(\frac{1}{k^{\frac{1}{2}-\eta}}\right),\ \|\tilde{\pi}_{k,i}\|^2 = O\left(\frac{1}{k^{\frac{1}{2}-\eta}}\right),\ a.s.,$$

*where* $\tilde{\beta}_{k,i} = \beta_{k,i} - \beta_i^*$, $\tilde{\pi}_{k,i} = \pi_{k,i} - \pi_i^*$, $i \in [m]$, $\eta$ *is an arbitrary small positive constant,* $\underline{c}$ *and* $\bar{c}$ *are the constants defined in Assumption 3.6, and* $C$ *is a constant related to the local convergence property in Theorem 3.1.*

### 3.2. Clustering Performance of Algorithm 1

Based on the estimate $\beta_{k,i}$, we can categorize the new data $\{\phi_k, y_{k+1}\}$ to the cluster according to the following criterion:

$$\mathcal{I}_k = \arg\min_{i\in[m]}\{(y_{k+1} - \beta_{k,i}^\tau\phi_k)^2\}. \tag{9}$$

In order to evaluate the data clustering performance, we introduce the following within-cluster error:

$$J_n = \frac{1}{n}\sum_{k=1}^n (y_{k+1} - \beta_{k,\mathcal{I}_k}^\tau\phi_k)^2, \tag{10}$$

which is widely adopted in the analysis of the MLR problem (Yi et al., 2016). Then we give a lower bound to the probability that the new data $\{\phi_k, y_{k+1}\}$ is categorized to the correct cluster and an upper bound of the within-cluster error (10) in the following theorem:

**Theorem 3.9.** *Let the conditions of Theorem 3.1 hold. Then we have*

$$\begin{aligned}&\lim_{k\to\infty} P(\{\phi_k, y_{k+1}\} \textit{ is categorized correctly})\\ &\geq 1 - E\left\{\max_{j\neq i}\exp\left(-\frac{((\beta_i^* - \beta_j^*)^\tau\phi)^2}{8\sigma^2}\right)\right\},\end{aligned} \tag{11}$$

*and*

$$\lim_{n\to\infty} J_n \leq \sigma^2 + \gamma \leq \sigma^2,\ a.s., \tag{12}$$

*where $\phi$ is a random vector with p.d.f $g$ defined in Assumption 2.9, $\gamma = \sum_{i=1}^{m} \pi_i^* \min_{j\neq i} E\left\{\gamma_{i,j}(\phi)\right\} \leq 0$ with $\gamma_{i,j}(\phi) = [(\beta_i^* - \beta_j^*)^\tau \phi]^2 \Phi\left(-|(\beta_i^* - \beta_j^*)^\tau \phi|/(2\sigma)\right) - 2\sigma|(\beta_i^* - \beta_j^*)^\tau \phi|\Phi'\left(-|(\beta_i^* - \beta_j^*)^\tau \phi|/(2\sigma)\right)$, $\Phi(x) = P(t \leq x)$ and $\Phi'(x) = E\{t\mathbb{I}_{\{t\leq x\}}\}$ with $t$ being a standard Gaussian random variable.*

*Remark* 3.10. The above theorem demonstrates that the data clustering performance is positively correlated with the signal-to-noise ratio $\frac{R^*_{\min}}{\sigma}$. Specifically, if the signal-to-noise ratio $\frac{R^*_{\min}}{\sigma}$ tends to infinity, the probability that $\{\phi_k, y_{k+1}\}$ can be categorized correctly will tend to 1 and the upper bound of the within-cluster error will tend to $\sigma^2$. In addition, from the proof of Theorem 3.9, one can see that the bounds given in (11) and (12) are actually the same as those for the case where the true parameters $\beta_i^*$ are known.

## 4. Proof of Main Results

In this section, we give the proof of Theorems 3.1, 3.8-3.9.

### 4.1. Proof of Theorem 3.1

Ljung's ODE method (Ljung, 1977) provides a general analytical technique for the convergence analysis of recursive algorithms by establishing the relationship between the convergence properties of the recursive algorithm and the stability of the corresponding ODEs.

In this subsection, we adopt this ODE method to establish the convergence results of Algorithm 1. For this purpose, we first rewrite (7) as follows:

$$x_{k+1} = \Pi_S\left[x_k + \frac{1}{k}Q(x_k, \phi_k, y_{k+1})\right], \tag{13}$$

where $x_k = [\beta_{k,1}^\tau \ \pi_{k,1} \ \cdots \ \beta_{k,m}^\tau \ \pi_{k,m} \ \text{vec}^\tau(R_{k,1}) \ \cdots \ \text{vec}^\tau(R_{k,m})]^\tau$, the projected domain $S$ is defined as

$$S = \big\{x \in \mathbb{R}^{m(d+1)+d^2} : \beta_i \in D_i, \pi_i \in D_{m+1}, x = [\beta_1^\tau \ \pi_1 \ \cdots \ \beta_m^\tau \ \pi_m \ \text{vec}^\tau(R_1) \ \cdots \ \text{vec}^\tau(R_m)]^\tau\big\},$$

the function $Q(x_k, \phi_k, y_{k+1})$ is defined as

$$\begin{aligned} Q(x_k,\phi_k, y_{k+1}) = \big[&Q_{1,1}(x_k, \phi_k, y_{k+1})^\tau \\ &Q_{1,2}(x_k, \phi_k, y_{k+1}) \cdots Q_{m,1}(x_k, \phi_k, y_{k+1})^\tau \\ &Q_{m,2}(x_k, \phi_k, y_{k+1}) \ \text{vec}^\tau(\bar{Q}_1(x_k, \phi_k, y_{k+1})) \\ &\cdots \text{vec}^\tau(\bar{Q}_m(x_k, \phi_k, y_{k+1}))\big]^\tau, \end{aligned}$$

with

$$\begin{aligned} &Q_{i,1}(x_k, \phi_k, y_{k+1}) = R_{k+1,i}^{-1}\alpha_{k,i}\phi_k(y_{k+1} - \beta_{k,i}^\tau\phi_k), \\ &Q_{i,2}(x_k, \phi_k, y_{k+1}) = \alpha_{k,i} - \pi_{k,i}, \\ &\bar{Q}_i(x_k, \phi_k, y_{k+1}) = \alpha_{k,i}\phi_k\phi_k^\tau - R_{k,i}, i \in [m], \end{aligned} \tag{14}$$

where vec$(\cdot)$ denotes the operator by stacking the columns of a matrix on top of one another. In order to analyze (13), we construct the following ODEs,

$$\frac{d}{dt}\beta_i(t) = R_i^{-1}(t) f_{i,1}(x(t)), \tag{15a}$$

$$\frac{d}{dt}\pi_i(t) = f_{i,2}(x(t)), \tag{15b}$$

$$\frac{d}{dt}R_i(t) = G_i(x(t)) - R_i(t), \tag{15c}$$

where

$$\begin{aligned} f_{i,1}(x(t)) &= \lim_{k\to\infty} E\{\alpha_{k,i}\phi_k\left(y_{k+1} - \beta_i(t)^\tau\phi_k\right)\}, \\ f_{i,2}(x(t)) &= \lim_{k\to\infty} E\{\alpha_{k,i}\} - \pi_i(t), \\ G_i(x(t)) &= \lim_{k\to\infty} E\{\alpha_{k,i}\phi_k\phi_k^\tau\}, i \in [m], \end{aligned}$$

with $x(t) = [\beta_1^\tau(t) \ \pi_1(t) \ \cdots \ \beta_m^\tau(t) \ \pi_m(t) \ \text{vec}^\tau(R_1(t)) \ \cdots \ \text{vec}^\tau(R_m(t))]^\tau$.

We first give the main results related to Ljung's ODE method in the following proposition, which plays an important role in the convergence analysis of this paper.

**Proposition 4.1.** *(Ljung, 1977) Let $S_A$ be an open, bounded and connected set in $\mathbb{R}^{d^2+(d+1)m}$ and $\bar{S}$ be its compact subset. If the ODEs (15) have an invariant set $S_c$ with domain of attraction $S_A \supset \bar{S}$, then we have $x_k \to S_c$ as $k \to \infty$ almost surely, provided that the following conditions are satisfied:*

*C1) The function $Q(x, \phi, y)$ defined in (13) is locally Lipschitz continuous for $x \in S_A$ with fixed $\phi$ and $y$, that is, for any $x_1, x_2 \in \mathcal{U}(x, \rho(x))$ 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 $\mathcal{U}(x, \rho(x))$ is the $\rho(x)$-neighborhood of $x$, i.e., $\mathcal{U}(x, \rho(x)) = \{\bar{x} : \|x - \bar{x}\| < \rho(x)\}$;*

*C2) $\frac{1}{n}\sum\limits_{k=1}^{n} \mathcal{R}(x, \phi_k, y_{k+1}, \rho(x))$ converges to a finite limit for any $x \in S_A$ as $n \to \infty$;*

*C3) $\lim\limits_{k\to\infty} E\left\{Q(x, \phi_k, y_{k+1})\right\}$ exists for $x \in S_A$ and also*

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

*where $x = \left[\beta_1^\tau \ \pi_1 \ \cdots \ \beta_m^\tau \ \pi_m \ vec^\tau(R_1) \ \cdots \ vec^\tau(R_m)\right]^\tau$.*

We prove Theorem 3.1 by verifying all the conditions required by Ljung's ODE method in Proposition 4.1. We first verify Conditions C1)-C3) related to the function $Q(x_k, \phi_k, y_{k+1})$ in the following Lemma with its proof given in Appendix A.1.

**Lemma 4.2.** *Under Assumptions 2.3-2.9, Conditions C1)-C3) in Proposition 4.1 are satisfied on the set* $S_A = \{x : \|\beta_i\| < M, \pi_i > \pi_{\min}, R_i > 0, x = [\beta_1^\tau \; \pi_1 \; \cdots \; \beta_m^\tau \; \pi_m \; vec^\tau(R_1) \; \cdots \; vec^\tau(R_m)]^\tau\}$*, where* $M$ *can be chosen sufficiently large and* $\pi_{\min} \in (0,1)$ *can be sufficiently small.*

By Lemma 4.2, the remaining key step in Ljung's ODE method is to establish the stability analysis of the ODEs (15), which is illustrated in the following Lemma. Here we let $S_A$ be sufficiently close to $\bar{S}$.

**Lemma 4.3.** *Under conditions of Theorem 3.1, the ODEs (15) has an invariant set* $D_c = \{x^*\}$ *with the domain of attraction* $S_A \supset \bar{S}$*, where*

$$x^* = [\beta_1^{*\tau} \; \pi_1^* \; \cdots \; \beta_m^{*\tau} \; \pi_m^* \; \pi_1^* vec^\tau(G) \; \cdots \; \pi_m^* vec^\tau(G)]^\tau, \tag{16}$$

*with* $G = E\{\phi\phi^\tau\}$*,* $\phi$ *is a random vector with p.d.f* $g$ *defined in Assumption 2.9, and*

$$\bar{S} = \{x \in S_A : \beta_i \in D_i, \pi_i \in D_{m+1}, i \in [m], x = [\beta_1^\tau \; \pi_1 \; \cdots \; \beta_m^\tau \; \pi_m \; vec^\tau(R_1) \; \cdots \; vec^\tau(R_m)]^\tau\}. \tag{17}$$

Here we provide a proof sketch of this lemma and give the proof details in Appendix A.1.

***Proof Sketch of Lemma 4.3.*** Denote $\tilde{\beta}_i(t) = \beta_i(t) - \beta_i^*$, and $\tilde{\pi}_i(t) = \pi_i(t) - \pi_i^*, i \in [m]$. We consider the following Lyapunov function:

$$\begin{aligned} &V(x(t)) \\ =&\frac{1}{2}\sum_{i=1}^m [\tilde{\beta}_i^\tau(t) R_i(t) \tilde{\beta}_i(t) + \tilde{\pi}_i^2(t) + \|R_i(t) - \pi_i^* G\|_F^2], \end{aligned}$$

where $G$ is defined in (16). Then we obtain the following derivative of $V(x(t))$ along the trajectories of ODEs (15):

$$\begin{aligned} \frac{dV(x(t))}{dt} &= \sum_{i=1}^m \left[\tilde{\beta}_i^\tau(t) f_{i,1}(x(t)) + \tilde{\pi}_i(t) f_{i,2}(x(t))\right. \\ &+ \frac{1}{2}\tilde{\beta}_i^\tau(t)(G_i(x(t)) - R_i(t))\tilde{\beta}_i(t) + \text{vec}(R_i(t) - \pi_i^* G)^\tau \\ &\times \left.\text{vec}(G_i(x(t)) - R_i(t))\right] \triangleq \sum_{i=1}^m J_i(x(t)). \end{aligned}$$

We proceed to show that $J_i(x(t)) \leq 0$. For this, we first define the following events for $i \in [m]$ and $j \in [m]\backslash\{i\}$,

$$\begin{aligned} &\mathcal{A} = \{\omega : |w| \leq \sigma\eta, w \sim \mathcal{N}(0, \sigma^2)\} \\ &\mathcal{A}_{i,j}(t) = \{\omega : |\tilde{\beta}_i^\tau(t)\phi| \vee |\tilde{\beta}_j^\tau(t)\phi| \leq 0.25|(\beta_j^* - \beta_i^*)^\tau\phi|\}, \\ &\mathcal{A}'_{i,j} = \{\omega : |(\beta_j^* - \beta_i^*)^\tau\phi| \geq 4\sqrt{2}\sigma\eta\}, \\ &A_i(t) = \mathcal{A} \cap (\cap_{j\neq i}\mathcal{A}_{i,j}(t)) \cap (\cap_{j\neq i}\mathcal{A}'_{i,j}), \\ &B_i = \{\omega : y = \beta_i^{*\tau}\phi + w, w \sim \mathcal{N}(0, \sigma^2)\}, \end{aligned}$$

where $a \vee b$ denotes $\max\{a, b\}$, $\eta$ is a parameter to be determined, and $\phi$ is defined in Theorem 3.9. Besides, in the following analysis, $\mathcal{A}_{i,j}(t)$ and $A_i(t)$ are abbreviated as $\mathcal{A}_{i,j}$ and $A_i$ for simplicity of expression, respectively.

By Lemmas A.1-A.2 in Appendix A.1, with simple calculations, it is not difficult to obtain that

$$\begin{aligned} &J_i(x(t)) \leq -\pi_i^* P(A_i) E\left\{\alpha_i(t)\phi\phi^\tau \middle| A_i \cap B_i\right\} \|\tilde{\beta}_i(t)\|^2 \\ &- \tilde{\pi}_i^2(t) - \|R_i(t) - \pi_i^* G\|_F^2 \\ &+ \tilde{\beta}_i^\tau(t) E\left\{\phi(\alpha_i(t) - \alpha_i^*)\left(y - \beta_i^{*\tau}\phi\right)\right\} \\ &+ \tilde{\pi}_i(t) E\left\{\alpha_i(t) - \alpha_i^*\right\} + \frac{1}{2}\tilde{\beta}_i^\tau(t)[G_i(x(t)) - R_i(t)]\tilde{\beta}_i(t) \\ &+ \text{vec}(R_i(t) - \pi_i^* G)^\tau \text{vec}(E\{(\alpha_i(t) - \alpha_i^*)\phi\phi^\tau\}), \end{aligned} \tag{18}$$

where $\alpha_i^*$ $(\alpha_i(t))$ with its explicit expression in (23)((22)) in the Appendix A.1 represents the posterior probabilities that the data $\{\phi, y\}$ belongs to the $i$-th model with true parameters $x^*$ in (16) (the parameter estimate $x(t)$ in (15)).

We now analyze the right-hand-side (RHS) of (18) term by term. On the event $A_i \cap B_i$, we have

$$\alpha_j(t) \leq [\pi_j(t)/\pi_i(t)] \exp(-\eta^2), j \neq i, \tag{19}$$

and then $\alpha_i(t) \geq 1 - \frac{1-\pi_{\min}}{\pi_{\min}} \exp(-\eta^2)$. Thus, it is easy to see that the first term on the RHS of (18) has an upper bound $-c_1\pi_i^* P(A_i)[1 - \frac{1-\pi_{\min}}{\pi_{\min}} \exp(-\eta^2)]\|\tilde{\beta}_i(t)\|^2$, where $c_1$ is a positive constant.

We then use the following three facts to bound the fourth, fifth, and last terms on the RHS of (18). First, on the event $A_i \cap B_i$, there exist small positive constant $\delta_{i,1}, \delta_{i,2}, i \in [m]$ such that $|\alpha_i(t) - \alpha_i^*| \leq \sum_{i=1}^m [\delta_{i,1}\|\tilde{\beta}_i(t)\| + \delta_{i,2}|\tilde{\pi}_i(t)|]$. Second, on the event $B_i^c$, without loss of generality, we assume the data $\{\phi, y\}$ is generated by the $j$-th ($j \neq i$) model. Then similar to (19), we can see that on the event $B_i^c \cap A_j$, both $\alpha_i(t)$ and $\alpha_i^*$ have an upper bound $\frac{1}{\pi_{\min}} \exp(-\eta^2)$. Besides, by Lemma A.2, we know that $P(A_j^c)$ has an upper bound $O\left((d_{\max} + \sigma\eta)/R_{\min}^* + \exp(-\eta^2/2)\right)$. Thus we can obtain the upper bounds of fourth, fifth and last terms on the event $B_i^c$. Third, on the event $A_i^c$, the fourth, fifth and last terms can also be bound by the probability of $A_i^c$, i.e., $O\left((d_{\max} + \sigma\eta)/R_{\min}^* + \exp(-\eta^2/2)\right)$. Based on these three facts, we can obtain that if $d_{\max}$ and signal-to-noise ratio $\frac{R_{\min}^*}{\sigma}$ satisfy the following condition: there exist positive and suitable constants $b_1$ and $b_2$ such that

$$R_{\min}^* \geq b_2 m \max\{\sigma[\log(b_1 m/\pi_{\min})]^{3/2}, d_{\max}^3, d_{\max}\},$$

then the fourth, fifth and last terms in (18) can be suppressed by the first three terms of (18).

By (15c), it is clear that the sixth term on the RHS of (18) will tend to zero. Hence we can obtain that $\frac{dV(x(t))}{dt} \leq 0$ for $x \in \bar{S}$. From LaSalle invariance principle, and the facts that $\left\{x \in \bar{S} : \frac{dV(x(t))}{dt} = 0\right\} = \{x^*\}$ and $S_A$ is sufficiently close to $\bar{S}$, we can obtain the desired result. □

***Proof of Theorem 3.1.*** By Proposition 4.1, Lemma 4.2 and Lemma 4.3, we can obtain that $\beta_{k,i} \to \beta_i^*, \pi_{k,i} \to \pi_i^*, i \in [m],$ a.s., as $k \to \infty$. Hence we complete the proof. □

### 4.2. Proof of Theorem 3.8

From Theorem 3.1, we know that the estimates $\beta_{k,i}$ and $\pi_{k,i}$ are convergent almost surely. Based on this fact, by following the Lyapunov analysis idea of the classical least-squares for linear regression models (e.g., Guo (1995)) and using the martingale convergence theorem (Chow & Teicher, 2003) to estimate the cumulative effect of random martingale difference sequence, we can obtain the desired convergence rate result in Theorem 3.8. The proof details are provided in Appendix A.2.

### 4.3. Proof of Theorem 3.9

Based on the convergence results in Theorem 3.1 and Assumption 2.5-2.9, especially the asymptotically stationary and ergodic property of $\{\phi_k\}$, it is not difficult to obtain the data clustering performance in Theorem 3.9. The proof details are provided in Appendix A.3.

## 5. Simulation Results

In this section, we conduct a simulation to illustrate the effectiveness of our algorithm.

Consider the input-output data $\{\phi_k, y_{k+1}\}_{k=1}^{\infty}$ are generated by the following dynamical system:

$$\begin{aligned} y_{k+1} &= \beta_{z_k}^{*\tau} \phi_k + w_{k+1}, \\ \phi_{k+1} &= 0.5\phi_k + e_{k+1}, \end{aligned}$$

where the true parameters $\beta_1^* = [1\ 15\ 13]^\tau$, $\beta_2^* = [-10\ -11\ -12]^\tau$ and $\beta_3^* = [3\ 4\ 3]^\tau$, the latent variable $\{z_k\}$ is i.i.d. with $P(z_k = 1) = 0.2, P(z_k = 1) = 0.5$ and $P(z_k = 3) = 0.3$, the state noise $\{e_{k+1}\}$ is i.i.d. with $\mathcal{N}(0, I)$ and the measurement noise $\{w_{k+1}\}$ is i.i.d. with $\mathcal{N}(0, 1)$. We can see that the input data $\{\phi_k\}$ is not i.i.d., and all assumptions in Theorem 3.1 are satisfied.

We first use the robust tensor method (Anandkumar et al., 2014) with $N = 5000$ samples at each iteration and a total iteration step $T = 10$ to obtain the estimate of initialization regions $D_i (i \in [m + 1])$. Then we conduct the online Algorithm 1 to estimate the unknown true parameters. The estimation error measured by $\|\beta_{k,i} - \beta_i^*\|^2$ and the clustering performance evaluated by the within-cluster error (10) are plotted in Figure 1, respectively. The convergence of estimation and clustering errors can demonstrate the effectiveness of our algorithm.

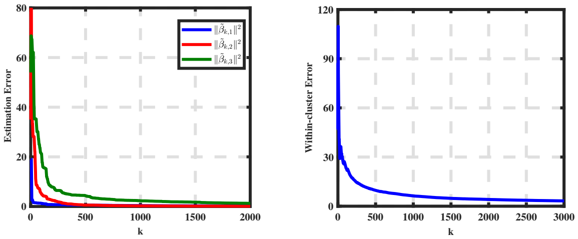

*Figure 1.* Estimation error (left) and clustering performance (right) of Algorithm 1.

## 6. Conclusion

This paper considers the parameter learning and data clustering problem for MLR with multiple sub-models and arbitrary mixing weights. To deal with the data streaming case, we propose an online learning algorithm to estimate the unknown parameters. By utilizing Ljung's ODE method, we establish the almost sure convergence results of this MLR problem without the traditional i.i.d. assumption on the input data for the first time. Based on the convergence property and using the classical stochastic Lyapunov function method, we also obtain the convergence rate analysis of the proposed algorithm for the first time. In addition, the data clustering can asymptotically achieve the same performance as the case with known parameters. Future work will consider how to relax the asymptotically stationary and ergodic assumption on the input data, and how to design algorithms with global convergence performance for the MLR problem.

## Acknowledgements

This work was supported by Natural Science Foundation of China under Grants T2293772 and 12288201, the National Key R&D Program of China under Grant 2018YFA0703800, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No.XDA27000001.

## Impact Statement

This is a basic theoretical research and we believe the ethical aspects are not applicable. For future societal consequences, our work provides a novel online learning method with a convergence guarantee, which could contribute to the understanding of general EM algorithms and applications of MLR on data clustering.

## References

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

Balakrishnan, S., Wainwright, M. J., and Yu, B. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1): 77–120, 2017.

Bemporad, A. A piecewise linear regression and classification algorithm with application to learning and model predictive control of hybrid systems. *IEEE Transactions on Automatic Control*, 68(6):3194–3209, 2023.

Bishop, C. M. and Nasrabadi, N. M. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

Chaganty, A. T. and Liang, P. Spectral experts for estimating mixtures of linear regressions. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pp. 1040–1048, Atlanta, GA, USA, 2013.

Chen, H. *Stochastic Approximation and Its Applications*. Dordrecht, The Netherlands: Kluwer, 2002.

Chen, H. and Guo, L. *Identification and stochastic adaptive control*. Springer Science & Business Media, 2012.

Chen, J., Zhu, J., Teh, Y. W., and Zhang, T. Stochastic expectation maximization with variance reduction. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, volume 31, Montréal, Canada, 2018.

Chow, Y. S. and Teicher, H. *Probability Theory: Independence, Interchangeability, Martingales*. Springer Science & Business Media, 2003.

Cohen, E. A. *The influence of nonharmonic partials on tone perception*. Stanford University, 1980.

De Veaux, R. D. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989.

Deb, P. and Holmes, A. M. Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Economics*, 9(6):475–489, 2000.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (methodological)*, 39(1):1–22, 1977.

Diamandis, T., Eldar, Y., Fallah, A., Farnia, F., and Ozdaglar, A. A wasserstein minimax framework for mixed linear regression. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pp. 2697–2706, Virtual Event, 2021.

Fang, K., Kotz, S., and Ng, K. W. *Symmetric Multivariate and Related Distributions*. CRC Press, 2018.

Faria, S. and Soromenho, G. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

Gaffney, S. and Smyth, P. Trajectory clustering with mixtures of regression models. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (SIGKDD 1990)*, pp. 63–72, San Diego, CA, USA, 1999.

Gaffney, S. and Smyth, P. Curve clustering with random effects regression mixtures. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics (AISTAT 2003)*, pp. 101–108, Valencia, Spain, 2003.

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

Guo, L. Convergence and logarithm laws of self-tuning regulators. *Automatica*, 31(3):435–450, 1995.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

Ingrassia, S., Minotti, S. C., and Punzo, A. Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis*, 71:159–182, 2014.

Karimi, B., Wai, H.-T., Moulines, E., and Lavielle, M. On the global convergence of (fast) incremental expectation maximization methods. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, volume 32, Montréal, Canada, 2018.

Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. Non-asymptotic analysis of biased stochastic approximation scheme. In *Proceedings of the 32nd Conference On Learning Theory (COLT 2019)*, pp. 1944–1974, Phoenix, AZ, USA, 2019.

Khalil, H. K. *Nonlinear systems*. Patience Hall, 2002.

Klusowski, J. M., Yang, D., and Brinda, W. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019.

Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. Meta-learning for mixed linear regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pp. 5394–5404, Virtual Event, 2020.

Kwon, J. and Caramanis, C. EM converges for a mixture of many linear regressions. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, pp. 1727–1736, Palermo, Sicily, Italy, 2020.

Kwon, J., Qian, W., Caramanis, C., Chen, Y., and Davis, D. Global convergence of the EM algorithm for mixtures of two component linear regression. In *Proceedings of the 32nd Conference on Learning Theory (COLT 2019)*, pp. 2055–2110, Phoenix, AZ, USA, 2019.

Lai, T. L. and Wei, C. Z. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

Langley, P. Crafting papers on machine learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y. and Liang, Y. Learning mixtures of linear regressions with nearly optimal complexity. In *Proceedings of the 31st Conference On Learning Theory (COLT 2018)*, pp. 1125–1144, Stockholm, Sweden, 2018.

Liu, Y., Liu, Z., and Guo, L. Global convergence of online identification for mixed linear regression. *ArXiv Preprint ArXiv:2311.18506*, 2023.

Ljung, L. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4):551–575, 1977.

Niedzwiecki, M. and Guo, L. Nonasymptotic results for finite-memory wls filters. *IEEE Transactions on Automatic Control*, 36(2):198–206, 1991.

Pal, S., Mazumdar, A., Sen, R., and Ghosh, A. On learning mixture of linear regressions in the non-realizable setting. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, pp. 17202–17220, Baltimore, Maryland, USA, 2022.

Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Qian, W., Zhang, Y., and Chen, Y. Global convergence of least squares EM for demixing two log-concave densities. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, volume 32, Vancouver, BC, Canada, 2019.

Quandt, R. E. A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67(338):306–310, 1972.

Sedghi, H., Janzamin, M., and Anandkumar, A. Provable tensor methods for learning mixtures of generalized linear models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016)*, pp. 1223–1231, Cadiz, Spain, 2016.

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

Sun, Y., Luo, Z., and Fan, X. Robust structured heterogeneity analysis approach for high-dimensional data. *Statistics in Medicine*, 41(17):3229–3259, 2022.

Wedel, M. and Kamakura, W. A. *Market Segmentation: Conceptual and Methodological Foundations*. Springer Science & Business Media, 2000.

Yi, X., Caramanis, C., and Sanghavi, S. Alternating minimization for mixed linear regression. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pp. 613–621, Beijing, China, 2014.

Yi, X., Caramanis, C., and Sanghavi, S. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *ArXiv Preprint ArXiv:1608.05749*, 2016.

Zhong, K., Jain, P., and Dhillon, I. S. Mixed linear regression with multiple components. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS 2016)*, volume 29, Barcelona, Spain, 2016.

Zilber, P. and Nadler, B. Imbalanced mixed linear regression. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, volume 36, New Orleans, USA, 2023.

# A. Proof of Main Results

In this section, we provide proof details of Theorems 3.1, 3.8-3.9.

## A.1. Proof of Theorem 3.1

In the following, we verify all the conditions in Proposition 4.1 to establish the convergence result of Algorithm 1 in Theorem 3.1. For this purpose, we first provide a related lemma on the properties of the regressor, the latent variable and the noise.

**Lemma A.1.** *Under Assumptions 2.5-2.9, $\{d_k \triangleq [\ z_k \quad \phi_k^\tau \quad w_{k+1}\ ]^\tau, k \geq 1\}$ 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.*

This lemma can be easily obtained by Assumptions 2.5-2.9 and the ergodicity theorem of the stationary process, and the proof details are omitted.

We now provide the proof of Lemma 4.2, where Conditions C1)-C3) in Proposition 4.1 are verified.

***Proof of Lemma 4.2.*** From the fact $\frac{\partial R^{-1}}{\partial R} = -R^{-1} \otimes R^{-1}$ (Gentle, 2008), we have $\left\|\frac{\partial R^{-1}}{\partial R}\right\| = \left\|R^{-1} \otimes R^{-1}\right\| = \frac{1}{\lambda_{min}^2(R)}$, where $\otimes$ is the Kronecker product for matrices. Besides, for a matrix $A \in \mathbb{R}^d$ with $q$ row partitions and $s$ column partitions, it is clear that $\|A\| \leq \|A\|_F \leq \sum_{i=1}^q \sum_{j=1}^s \|A_{ij}\|_F \leq \sqrt{d} \sum_{i=1}^q \sum_{j=1}^s \|A_{ij}\|$, where $A_{ij} (i \in [q], j \in [s])$ is the sub-matrix and $\|\cdot\|_F$ is the Frobenius norm of matrix. Then by (13) and (14), we have

$$
\begin{aligned}
&\frac{1}{\sqrt{d}}\left\|\frac{\partial Q(x,\phi,y)}{\partial x}\right\| \\
\leq&\sum_{i=1}^m\left[\sum_{j=1}^m\sum_{l=1}^2\left[\left\|\frac{\partial Q_{i,l}(x,\phi,y)}{\partial \beta_j}\right\|+\left\|\frac{\partial Q_{i,l}(x,\phi,y)}{\partial \pi_j}\right\|\right]+\left\|\frac{\partial Q_{i,1}(x,\phi,y)}{\partial R_i}\right\|\right] \\
&+\sum_{i=1}^m\left[\sum_{j=1}^m\left[\left\|\frac{\partial \bar{Q}_i(x,\phi,y)}{\partial \beta_j}\right\|+\left\|\frac{\bar{Q}_i(x,\phi,y)}{\partial \pi_j}\right\|\right]+\left\|\frac{\partial \bar{Q}_i(x,\phi,y)}{\partial R_i}\right\|\right] \\
\leq&\sum_{i=1}^m\frac{1}{\lambda_{\min}(R_i)}\Bigg\{\Bigg[\frac{\alpha_i(y-\beta_i^\tau\phi)^2\|\phi\|^2}{\sigma^2}+\alpha_i\|\phi\|^2+\sum_{j=1}^m\frac{\alpha_i\alpha_j|y-\beta_i^\tau\phi||y-\beta_j^\tau\phi|\|\phi\|^2}{\sigma^2}+\frac{\alpha_i|y-\beta_i^\tau\phi|\|\phi\|}{\pi_{\min}} \\
&+\sum_{j=1}^m\frac{\alpha_i\alpha_j|y-\beta_i^\tau\phi|\|\phi\|}{\pi_{\min}}+\frac{\alpha_i|y-\beta_i^\tau\phi|\|\phi\|}{\sigma^2}+\sum_{j=1}^m\frac{\alpha_i\alpha_j|y-\beta_j^\tau\phi|\|\phi\|}{\sigma^2}+1+\frac{\alpha_i}{\pi_{\min}}+\sum_{j=1}^m\frac{\alpha_i\alpha_j}{\pi_{\min}}\Bigg] \\
&+\frac{\alpha_i|y-\beta_i^\tau\phi|\|\phi\|}{\lambda_{\min}(R_i)}\Bigg\}+\sum_{i=1}^m\left[\frac{\alpha_i|y-\beta_i^\tau\phi|\|\phi\|}{\sigma^2}+\sum_{j=1}^m\frac{\alpha_i\alpha_j|y-\beta_j^\tau\phi|\|\phi\|}{\sigma^2}+\frac{\alpha_i}{\pi_{\min}}+\sum_{j=1}^m\frac{\alpha_i\alpha_j}{\pi_{\min}}\right]+m\|I_d\otimes I_d\| \\
\leq&\frac{1}{\bar{\lambda}_{\min}}\left[\frac{2(m+1)\|\phi\|^2(y^2+M^2\|\phi\|^2)}{\sigma^2}+\|\phi\|^2+\frac{(m+1)(\sigma^2+\pi_{\min})}{\sigma^2\pi_{\min}}\|\phi\|(|y|+M\|\phi\|)+m+\frac{m+1}{\pi_{\min}}\right] \\
&+\frac{\|\phi\|(|y|+M\|\phi\|)}{\bar{\lambda}_{\min}^2}+\frac{(m+1)(|y|+M\|\phi\|)\|\phi\|}{\sigma^2}+\frac{m+1}{\pi_{\min}}+m \\
\triangleq&\mathcal{R}(x,\phi,y,\rho(x)),
\end{aligned}
\tag{20}
$$

where $\bar{\lambda}_{\min} = \min_{i\in[m]} \lambda_{\min}(R_i)$. For $x \in S_A$, choose $\rho(x)$ sufficiently small such that any point $\bar{x}$ in the area $\mathcal{U}(x, \rho(x))$ has the property that $\bar{\lambda}_{\min} > 0$. By (20), for any fixed $\phi$ and $y$, we can obtain that $\mathcal{R}(x, \phi, y, \rho(x))$ is bounded, i.e., $Q(x, \phi, y, \rho(x))$ is locally Lipschitz continuous, thus Condition C1) is verified.

For Condition C2), we only need to verify its upper bound to satisfy C2) since $\mathcal{R}(x, \phi, y, \rho(x))$ is defined as the Lipschitz constant. It is obvious because $\|\phi_k\|$, $\|\phi_k\|y_{k+1}$, $\|\phi_k\|^2 y_{k+1}^2$ and $\|\phi_k\|^4$ are all asymptotically stationary and ergodic by Lemma A.1.

The verification of Condition C3) is straightforward by Lemma A.1 and model (1), and the details are omitted. This completes the proof of this lemma. □

We now provide the proof of the stability analysis of the ODEs (15) established in Lemma 4.3. For this, we first give some useful lemmas below.

**Lemma A.2.** *For a random variable $y \sim \sum_{i=1}^{m} \pi_i^* \mathcal{N}(a_i, \sigma^2)$ with $a_i$ and $\pi_i^* \in (0,1)$ being constants, and also $\sum_{i=1}^{m} \pi_i^* = 1$, we have*

$$E\left\{ \frac{y \exp\left(-\frac{(y-a_i)^2}{2\sigma^2}\right)}{\sum_{j=1}^{m} \pi_j^* \exp\left(-\frac{(y-a_j)^2}{2\sigma^2}\right)^2} \right\} = a_i, \quad E\left\{ \frac{\pi_i^* \exp\left(-\frac{(y-a_i)^2}{2\sigma^2}\right)}{\sum_{j=1}^{m} \pi_j^* \exp\left(-\frac{(y-a_j)^2}{2\sigma^2}\right)^2} \right\} = \pi_i^*, i \in [m].$$

*Proof.* It can be easily obtained by using the distribution of $y$. □

**Lemma A.3.** *Let $w \sim \mathcal{N}(0, \sigma^2)$ and $\phi \in \mathbb{R}^d$ be a random vector with p.d.f being $g(\phi)$ defined in Assumption 2.9. Then we have the following properties:*

*(i) $P(|w| \geq \sigma\eta) \leq 2\exp\left(-\frac{\eta^2}{2}\right)$;*

*(ii) There exist positive constants $c_0$, $c_1$ and $c_2$ such that $E\{\phi\phi^\tau\} = c_0\Sigma$ and $c_1 I \leq c_0\Sigma \leq c_2 I$, where $\Sigma$ is defined in Assumption 2.9. Besides, there exists a constant $c_3$ such that $E\{\|\phi\|^4\} < c_3$;*

*(iii) For any fixed vector $v \in \mathbb{R}^d$ and positive constant $M$, there exists a constant $c \in \left(0, \frac{\|\Sigma^{1/2}v\|}{M}\right)$ such that $P(|\phi^\tau v| \leq M) = \frac{cM}{\|\Sigma^{1/2}v\|}$;*

*(iv) For any two fixed vectors $u, v \in \mathbb{R}^d$ satisfying $\|\Sigma^{1/2}u\| \geq \|\Sigma^{1/2}v\|$, we have $P(|\phi^\tau u| \leq |\phi^\tau v|) \leq \frac{\|\Sigma^{1/2}v\|}{\|\Sigma^{1/2}u\|} \leq \frac{c_2\|v\|}{c_1\|u\|}$. Besides, we have $E\left\{\phi\phi^\tau \middle| |\phi^\tau u| \geq |\phi^\tau v|\right\} \leq 2c_2$.*

*Proof.* The property (i) can be easily obtained by the Gaussian property of $w$.

We now verify the properties (ii) and (iii). Denote $\bar{\phi} = \Sigma^{-1/2}\phi$. By Assumption 2.9, we know that the p.d.f of $\bar{\phi}$ has the rotation-invariant property and thus we have $E\{\bar{\phi}\bar{\phi}^\tau\} = c_0 I$ with $c_0$ being a positive constant (Fang et al., 2018).

The first part of (ii) can be obtained by the positive-definiteness and boundedness of $\Sigma$ and the second part of (ii) holds by the u.i. property of $\{\|\phi_k\|^4\}$. Thus the property (ii) can be verified.

Denote $\bar{\phi}^{(i)}$ as the $i$-th element of $\bar{\phi}$. From the rotation-invariant property of $\bar{\phi}$, we have

$$P(|\phi^\tau v| \leq M) = P(|\bar{\phi}^\tau(\Sigma^{1/2}v)| \leq M) = P(|\bar{\phi}^{(1)}| \|\Sigma^{1/2}v\| \leq M) = F\left(\frac{M}{\|\Sigma^{1/2}v\|}\right) - F\left(-\frac{M}{\|\Sigma^{1/2}v\|}\right) = \frac{cM}{\|\Sigma^{1/2}v\|},$$

where $F$ denotes the marginal distribution function of the p.d.f of $\bar{\phi}^{(1)}$. Thus (iii) is proved.

Moreover, by Lemmas 6 and 7 in (Yi et al., 2016) and the rotation invariant property of the p.d.f of $\bar{\phi}$, (iv) can be obtained. This completes the proof of Lemma A.3. □

We now provide the proof details of Lemma 4.3.

***Proof of Lemma 4.3.*** We now study the stability of ODEs (15). For this, consider the following Lyapunov function:

$$V(x(t)) = \frac{1}{2}\sum_{i=1}^{m}[\tilde{\beta}_i^\tau(t)R_i(t)\tilde{\beta}_i(t) + \tilde{\pi}_i^2(t) + \|R_i(t) - \pi_i^* G\|_F^2],$$

where $\tilde{\beta}_i(t) = \beta_i(t) - \beta_i^*$, $\tilde{\pi}_i(t) = \pi_i(t) - \pi_i^*$, $i \in [m]$, $G = E\{\phi\phi^\tau\}$, $\phi$ is a random vector with its p.d.f being $g$ defined in Assumption 2.9. We then have the following derivative of $V(x(t))$ along the trajectories of ODEs (15):

$$\begin{aligned}\frac{dV(x(t))}{dt} =& \sum_{i=1}^{m}\Bigg[\tilde{\beta}_i^\tau(t)f_{i,1}(x(t)) + \tilde{\pi}_i(t)f_{i,2}(x(t)) + \frac{1}{2}\tilde{\beta}_i^\tau(t)(G_i(x(t)) - R_i(t))\tilde{\beta}_i(t)\\ &+ \mathrm{vec}(R_i(t) - \pi_i^* G)^\tau \mathrm{vec}(G_i(x(t)) - R_i(t))\Bigg] \triangleq \sum_{i=1}^{m} J_i(x(t)),\end{aligned} \tag{21}$$

where $f_{i,1}(x(t)) = E\left[\alpha_i(t)\phi\left(y - \beta_i^\tau(t)\phi\right)\right]$, $f_{i,2}(x(t)) = E\left[\alpha_i(t)\right] - \pi_i(t)$, $G_i(x(t)) = E\left[\alpha_i(t)\phi\phi^\tau\right]$ with

$$\alpha_i(t) = \frac{\pi_i(t)\exp\left(-\frac{(y-\beta_i^\tau(t)\phi)^2}{2\sigma^2}\right)}{\sum_{j=1}^m \pi_j(t)\exp\left(-\frac{(y-\beta_j^\tau(t)\phi)^2}{2\sigma^2}\right)}, \tag{22}$$

and the random variable $y$ given $\phi$ obeys the distribution $\sum_{i=1}^m \pi_i^* \mathcal{N}(\beta_i^{*\tau}\phi, \sigma^2)$.

By Lemma A.2, we have $f_{i,1}(x^*) = 0$. Denote

$$\alpha_i^* = \frac{\pi_i^* \exp\left(-\frac{(y-\beta_i^{*\tau}\phi)^2}{2\sigma^2}\right)}{\sum_{j=1}^m \pi_j^* \exp\left(-\frac{(y-\beta_j^{*\tau}\phi)^2}{2\sigma^2}\right)}. \tag{23}$$

By Lemma A.2, we also have $E[a_i^*] = \pi_i^*$ and it follows that $f_{i,2}(x^*) = 0$. Thus we have

$$\begin{aligned} J_i(x(t)) = {} & \tilde{\beta}_i^\tau(t)\left(f_{i,1}(x(t)) - f_{i,1}(x^*)\right) + \tilde{\pi}_i(t)\left(f_{i,2}(x(t)) - f_{i,2}(x^*)\right) + \frac{1}{2}\tilde{\beta}_i^\tau(t)(G_i(x(t)) - R_i(t))\tilde{\beta}_i(t) \\ & + \mathrm{vec}(R_i(t) - \pi_i^* G)^\tau \mathrm{vec}(G_i(x(t)) - R_i(t)) \triangleq \sum_{j=1}^4 J_{i,j}(x(t)). \end{aligned} \tag{24}$$

We now analyze the RHS of (24) term by term. Without loss of generality, we only provide the analysis for the case $i = 1$.

**Step 1: Analysis of the term $J_{1,1}(x(t))$.** First, by the mean-value theorem, we have

$$\begin{aligned} J_{1,1}(x(t)) = {} & -E\left\{\alpha_1(t)(\tilde{\beta}_1^\tau(t)\phi)^2\right\} + E\left\{\tilde{\beta}_1^\tau(t)\phi(\alpha_1(t) - \alpha_1^*)\left(y - \beta_1^{*\tau}\phi\right)\right\} \\ = & - E\left\{\alpha_1(t)(\tilde{\beta}_1^\tau(t)\phi)^2\right\} + \frac{1}{\sigma^2}E\left\{(\tilde{\beta}_1^\tau(t)\phi)^2\alpha_{1u}(t)(1-\alpha_{1u}(t))\left(y - \beta_{1u}^\tau(t)\phi\right)\left(y - \beta_1^{*\tau}\phi\right)\right\} \\ & - \frac{1}{\sigma^2}\sum_{j=2}^m E\left\{\tilde{\beta}_1^\tau(t)\phi\phi^\tau\tilde{\beta}_j(t)\alpha_{1u}(t)\alpha_{ju}(t)(y - \beta_{ju}^\tau(t)\phi)(y - \beta_1^{*\tau}\phi)\right\} \\ & + \frac{\tilde{\pi}_1(t)}{\pi_{1u}(t)}E\left\{\tilde{\beta}_1^\tau(t)\phi\alpha_{1u}(t)(1-\alpha_{1u}(t))(y - \beta_1^{*\tau}\phi)\right\} - \sum_{j=2}^m \frac{\tilde{\pi}_j(t)}{\pi_{ju}(t)}E\left\{\tilde{\beta}_1^\tau(t)\phi\alpha_{1u}(t)\alpha_{ju}(t)(y - \beta_1^{*\tau}\phi)\right\} \end{aligned} \tag{25}$$

where $\beta_{iu}^\tau(t)\phi$ is between $\beta_i^\tau(t)\phi$ and $\beta_i^{*\tau}\phi$, $\pi_{iu}(t)$ is between $\pi_i(t)$ and $\pi_i^*$, and $\alpha_{iu}(t) = \pi_{iu}(t)\exp\left(-\frac{(y-\beta_{iu}^\tau(t)\phi)^2}{2\sigma^2}\right) / \left(\sum_{j=1}^m \pi_{ju}(t)\exp\left(-\frac{(y-\beta_{ju}^\tau(t)\phi)^2}{2\sigma^2}\right)\right)$, $i \in [m]$.

We proceed to analyze the RHS of (25) term by term. For this, we introduce the following events for $j = 2, \cdots, m$,

$$\begin{aligned} \mathcal{A}_1 &= \{\omega : |w| \le \sigma\eta, w \sim \mathcal{N}(0, \sigma^2)\}, \\ \mathcal{A}_{2,j} &= \{\omega : |\phi^\tau(\beta_1(t) - \beta_1^*)| \vee |\phi^\tau(\beta_j(t) - \beta_j^*)| \le 0.25|\phi^\tau(\beta_j^* - \beta_1^*)|\}, \\ \mathcal{A}_{3,j} &= \{\omega : |\phi^\tau(\beta_j^* - \beta_1^*)| \ge 4\sqrt{2}\sigma\eta\}. \end{aligned} \tag{26}$$

(In our analysis, events (e.g., $\mathcal{A}_{2,j}$) may be related to the variable $t$, and we do not express this relationship explicitly for simplicity of expression.) Then by Lemma A.3, we have

$$P(\mathcal{A}_1^c) \le 2\exp\left(-\frac{\eta^2}{2}\right), P(\mathcal{A}_{2,j}^c) \le \frac{8c_2 d_{\max}}{c_1 R_{\min}^*}, P(\mathcal{A}_{3,j}^c) \le \frac{4\sqrt{2}cc_0\sigma\eta}{c_1 R_{\min}^*}. \tag{27}$$

We now give the upper bound of the first term on the RHS of (25). For this, let us first denote $E_i\{\cdot\}$ as the abbreviation to $E_{y\sim\mathcal{N}(\beta_i^{*\tau}\phi,\sigma^2)}\{\cdot\}, i \in [m]$ and $y^j$ as the random variables with distribution $\mathcal{N}(\beta_j^{*\tau}\phi, \sigma^2)$ given $\phi$. Then we have

$$M_1(x(t)) \triangleq E\left\{\alpha_1(t)(\tilde{\beta}_1^\tau(t)\phi)^2\right\} = \sum_{j=1}^m \pi_j^* E_j\left\{\alpha_1(t)(\tilde{\beta}_1^\tau(t)\phi)^2\right\} \ge \pi_1^* E_1\left\{\alpha_1(t)(\tilde{\beta}_1^\tau(t)\phi)^2\right\}. \tag{28}$$

Denote $A_1 = \mathcal{A}_1 \cap \mathcal{A}_{2,2} \cdots \cap \mathcal{A}_{2,m} \cap \mathcal{A}_{3,2} \cdots \cap \mathcal{A}_{3,m}$. Then on the event $A_1$, we have

$$\begin{aligned}
&1-\alpha_1(t) = \sum_{l=2}^{m} \alpha_l(t) = \sum_{l=2}^{m} \frac{\pi_l(t)\exp\left(-\frac{(y^1-\beta_l^\tau(t)\phi)^2}{2\sigma^2}\right)}{\sum_{j=1}^{m}\pi_j(t)\exp\left(-\frac{(y^1-\beta_j^\tau(t)\phi)^2}{2\sigma^2}\right)} \\
&\leq \sum_{l=2}^{m} \frac{\pi_l(t)}{\pi_1(t)} \exp\left(\frac{(\beta_1^{*\tau}\phi-\beta_1^\tau(t)\phi+w)^2}{2\sigma^2} - \frac{(\beta_1^{*\tau}\phi-\beta_l^\tau(t)\phi+w)^2}{2\sigma^2}\right) \leq \sum_{l=2}^{m}\frac{\pi_l(t)}{\pi_1(t)}\exp(-\eta^2) \leq \delta_1\exp(-\eta^2),
\end{aligned} \tag{29}$$

where $\delta_1 = \frac{1-\pi_{\min}}{\pi_{\min}}$. Besides, by Lemma A.3 and the independence property between $\phi$ and $w$, we have

$$E_1\left\{\phi\phi^\tau|\mathcal{A}_1\right\} = E_1\left\{\phi\phi^\tau\right\} \leq c_2 I, E_1\left\{\phi\phi^\tau|\mathcal{A}_{2,j}\right\} \leq 2c_2 I, E_1\left\{\phi\phi^\tau|\mathcal{A}_{3,j}\right\} \leq c_2 I, j=2,\cdots,m, \tag{30}$$

then it follows that

$$\begin{aligned}
E_1\left\{\phi\phi^\tau \mathbb{I}_{A_1}\right\} &\leq E_1\left\{\phi\phi^\tau\mathbb{I}_{\mathcal{A}_1}\right\} + \sum_{j=2}^{m}\left(E_1\left\{\phi\phi^\tau\mathbb{I}_{\mathcal{A}_{2,j}^c}\right\} + E_1\left\{\phi\phi^\tau\mathbb{I}_{\mathcal{A}_{3,j}^c}\right\}\right) \\
&\leq \left[P(\mathcal{A}_1^c) + \sum_{j=2}^{m}\left(2P(\mathcal{A}_{2,j}^c) + P(\mathcal{A}_{3,j}^c)\right)\right] c_2 I.
\end{aligned} \tag{31}$$

By (29) and (31), we have

$$\begin{aligned}
&E_1\left\{\alpha_1(t)\phi\phi^\tau\right\} \geq E_1\left\{\alpha_1(t)\phi\phi^\tau\mathbb{I}_{A_1}\right\} \geq (1-\delta_1\exp(-\eta^2))E_1\left\{\phi\phi^\tau\mathbb{I}_{A_1}\right\} \\
\geq&(1-\delta_1\exp(-\eta^2))\left(E_1\left\{\phi\phi^\tau\right\} - E_1\left\{\phi\phi^\tau\mathbb{I}_{A_1^c}\right\}\right) \\
\geq&(1-\delta_1\exp(-\eta^2))\left(c_1 I - \left[P(\mathcal{A}_1^c) + \sum\nolimits_{j=2}^{m}\left(2P(\mathcal{A}_{2,j}^c)+P(\mathcal{A}_{3,j}^c)\right)\right]c_2 I\right).
\end{aligned} \tag{32}$$

Then by (27) and (28), we can obtain

$$M_1(x(t)) \geq \pi_1^*(1-\delta_1\exp(-\eta^2))\left(c_1 - c_2\left[P(\mathcal{A}_1^c) + \sum\nolimits_{j=2}^{m}\left(2P(\mathcal{A}_{2,j}^c)+P(\mathcal{A}_{3,j}^c)\right)\right]\right)\|\tilde{\beta}_1(t)\|^2 \geq \bar{m}\|\tilde{\beta}_1(t)\|^2, \tag{33}$$

where $\bar{m} = \pi_1^*(1-\delta_1\exp(-\eta^2))\left(c_1 - 2c_2\left[\exp\left(-\frac{\eta^2}{2}\right) + \frac{(m-1)\left(8c_2 d_{\max}+2\sqrt{2}cc_0\sigma\eta\right)}{c_1 R_{\min}^*}\right]\right)$.

We now consider the second term on the RHS of (25). For this, let us denote $M_2(x(t)) = E\left\{(\tilde{\beta}_1^\tau(t)\phi)^2\alpha_{1u}(t)(1-\alpha_{1u}(t))\left(y-\beta_{1u}^\tau(t)\phi\right)\left(y-\beta_1^{*\tau}\phi\right)\right\}$, then we have

$$M_2(x(t)) = \sum_{j=1}^{m}\pi_j^* E_j\left\{(\tilde{\beta}_1^\tau(t)\phi)^2\alpha_{1u}(t)(1-\alpha_{1u}(t))\left(y-\beta_{1u}^\tau(t)\phi\right)\left(y-\beta_1^{*\tau}\phi\right)\right\} \triangleq \sum_{j=1}^{m}\pi_j^* M_{2,j}(x(t)). \tag{34}$$

For the term $M_{2,1}(x(t))$, on the event $A_1 = \mathcal{A}_1 \cap \mathcal{A}_{2,2}\cdots\cap\mathcal{A}_{2,m}\cap\mathcal{A}_{3,2}\cdots\cap\mathcal{A}_{3,m}$, similar to (29), we have $1-\alpha_{1u}(t) \leq \delta_1\exp(-\eta^2)$. Besides, we have $\left|y^1-\beta_{1u}^\tau(t)\phi\right| \leq \frac{1}{4}\|\phi\|R_{\max}^* + \sigma\eta$. Then it follows that

$$\begin{aligned}
&L_1(x(t)) = E_1\left\{\left(\tilde{\beta}_1^\tau(t)\phi\right)^2\alpha_{1u}(t)(1-\alpha_{1u}(t))\left(y-\beta_{1u}^\tau(t)\phi\right)\left(y-\beta_1^{*\tau}\phi\right)\mathbb{I}_{A_1}\right\} \\
\leq&E_1\left\{\left(\tilde{\beta}_1^\tau(t)\phi\right)^2\delta_1\exp(-\eta^2)\left(\frac{1}{4}\|\phi\|R_{\max}^*\sigma\eta+\sigma^2\eta^2\right)\mathbb{I}_{A_1}\right\} \\
\leq&\delta_1\exp(-\eta^2)\sigma\eta\left(\frac{1}{4}R_{\max}^* c_3^{3/4} + dc_2\sigma\eta\right)\|\tilde{\beta}_1(t)\|^2.
\end{aligned} \tag{35}$$

On the event $A_1^c$, by Lemma A.3 and the independence property between $\phi$ and $w$, we have

$$\begin{aligned}
&L_2(x(t)) = E_1\left\{\left(\tilde{\beta}_1^\tau(t)\phi\right)^2\alpha_{1u}(t)(1-\alpha_{1u}(t))\left(y-\beta_{1u}^\tau(t)\phi\right)\left(y-\beta_1^{*\tau}\phi\right)\mathbb{I}_{A_1^c}\right\} \\
&\leq\left(E_1\left\{\left(d_{\max}\|\phi\|^3|w|+\|\phi\|^2w^2\right)\mathbb{I}_{\mathcal{A}_1^c}\right\} + E_1\left\{\left(d_{\max}\|\phi\|^3|w|+\|\phi\|^2w^2\right)\mathbb{I}_{\mathcal{A}_{2,2}\cdots\cap\mathcal{A}_{2,m}\cap\mathcal{A}_{3,2}\cdots\cap\mathcal{A}_{3,m}}\right\}\right)\|\tilde{\beta}_1(t)\|^2 \\
&\leq\left((\sqrt{3}\sigma^2 dc_2 + d_{\max}c_3^{3/4}\sigma)\sqrt{P(\mathcal{A}_1^c)} + (d_{\max}\sigma c_3^{3/4}+\sigma^2 dc_2)P(\mathcal{A}_{2,2}\cdots\cap\mathcal{A}_{2,m}\cap\mathcal{A}_{3,2}\cdots\cap\mathcal{A}_{3,m})\right)\|\tilde{\beta}_1(t)\|^2.
\end{aligned} \tag{36}$$

Thus from (27), (35) and (36), we have

$$M_{2,1}(x(t)) = L_1(x(t)) + L_2(x(t)) \leq m_1 \|\tilde{\beta}_1(t)\|^2, \tag{37}$$

where

$$\begin{aligned} m_1 = & \delta_1 \exp(-\eta^2)\sigma\eta \left(\frac{1}{4} R^*_{\max} c_3^{3/4} + dc_2\sigma\eta\right) + \sqrt{2}(\sqrt{3}\sigma^2 dc_2 + d_{\max} c_3^{3/4}\sigma) \exp\left(-\frac{\eta^2}{4}\right) \\ & + (d_{\max}\sigma c_3^{3/4} + \sigma^2 dc_2) \frac{(m-1)\left(8c_2 d_{\max} + 4\sqrt{2}cc_0\sigma\eta\right)}{c_1 R^*_{\min}}. \end{aligned} \tag{38}$$

Similarly, for the term $M_{2,j}(x(t))$ with $j \neq 1$, let us denote $A_1 = \mathcal{A}_1 \cap \mathcal{A}_{2,j} \cap \mathcal{A}_{3,j}$. On the event $A_1$, we have

$$\begin{aligned} \alpha_{1u}(t) = & \frac{\pi_{1u}(t) \exp\left(-\frac{(y^j - \beta_{1u}^\tau(t)\phi)^2}{2\sigma^2}\right)}{\sum_{j=1}^m \pi_{ju}(t) \exp\left(-\frac{(y^j - \beta_{ju}^\tau(t)\phi)^2}{2\sigma^2}\right)} \\ \leq & \frac{\pi_{1u}(t)}{\pi_{ju}(t)} \exp\left(\frac{(\beta_j^{*\tau}\phi - \beta_{ju}^\tau(t)\phi + w)^2}{2\sigma^2} - \frac{(\beta_j^{*\tau}\phi - \beta_{1u}^\tau(t)\phi + w)^2}{2\sigma^2}\right) \leq \frac{\pi_{1u}(t)}{\pi_{ju}(t)} \exp(-\eta^2) \leq \frac{\exp(-\eta^2)}{\pi_{\min}}, \end{aligned} \tag{39}$$

$|y^j - \beta_{1u}^\tau(t)\phi| \leq \frac{5}{4}\|\phi\| R^*_{\max} + \sigma\eta$ and $|y^j - \beta_1^{*\tau}\phi| \leq \|\phi\| R^*_{\max} + \sigma\eta$, then it follows that

$$\begin{aligned} L_1'(x(t)) = & E_j \left\{ \left(\tilde{\beta}_1^\tau(t)\phi\right)^2 \alpha_{1u}(t)(1 - \alpha_{1u}(t)) \left(y - \beta_{1u}^\tau(t)\phi\right) \left(y - \beta_1^{*\tau}\phi\right) \mathbb{I}_{A_1} \right\} \\ \leq & \frac{\exp(-\eta^2)}{\pi_{\min}} \left(\frac{5}{4} R^{*2}_{\max} E_j\left\{\|\phi\|^4 \mathbb{I}_{A_1}\right\} + \frac{9}{4} R^*_{\max}\sigma\eta E_j\left\{\|\phi\|^3 \mathbb{I}_{A_1}\right\} + \sigma^2\eta^2 E_j\left\{\|\phi\|^2 \mathbb{I}_{A_1}\right\}\right) \|\tilde{\beta}_1(t)\|^2 \\ \leq & \frac{\exp(-\eta^2)}{\pi_{\min}} \left(\frac{5}{4} R^{*2}_{\max} c_3 + \frac{9}{4} R^*_{\max}\sigma\eta c_3^{3/4} + \sigma^2\eta^2 dc_2\right) \|\tilde{\beta}_1(t)\|^2. \end{aligned} \tag{40}$$

On the event $\mathcal{A}_1^c$, we have $|y^j - \beta_{1u}^\tau(t)\phi| \leq \|\phi\|(R^*_{\max} + d_{\max}) + |w|$ and $|y^j - \beta_1^{*\tau}\phi| \leq \|\phi\| R^*_{\max} + |w|$, then it follows that

$$\begin{aligned} L_2'(x(t)) = & E_j \left\{ \left(\tilde{\beta}_1^\tau(t)\phi\right)^2 \alpha_{1u}(t)(1 - \alpha_{1u}(t)) \left(y - \beta_{1u}^\tau(t)\phi\right) \left(y - \beta_1^{*\tau}\phi\right) \mathbb{I}_{\mathcal{A}_1^c} \right\} \\ \leq & \Bigg((R^{*2}_{\max} + R^*_{\max} d_{\max}) E_j\left\{\|\phi\|^4 \mathbb{I}_{\mathcal{A}_1^c}\right\} + (2R^*_{\max} + d_{\max}) E_j\left\{\|\phi\|^3 |w| \mathbb{I}_{\mathcal{A}_1^c}\right\} + E_j\left\{\|\phi\|^2 w^2 \mathbb{I}_{\mathcal{A}_1^c}\right\}\Bigg) \|\tilde{\beta}_1(t)\|^2 \\ \leq & \Bigg((R^{*2}_{\max} + R^*_{\max} d_{\max}) c_3 P(\mathcal{A}_1^c) + (2R^*_{\max} + d_{\max}) c_3^{3/4}\sigma \sqrt{P(\mathcal{A}_1^c)} + \sqrt{3} dc_2\sigma^2 \sqrt{P(\mathcal{A}_1^c)}\Bigg) \|\tilde{\beta}_1(t)\|^2. \end{aligned} \tag{41}$$

On the event $\mathcal{A}_{2,j}^c$, we have $|y^j - \beta_{1u}^\tau(t)\phi| \leq 5d_{\max}\|\phi\| + |w|$ and $|y^j - \beta_1^{*\tau}\phi| \leq 4d_{\max}\|\phi\| + |w|$, then it follows that

$$\begin{aligned} L_3'(x(t)) = & E_j \left\{ \left(\tilde{\beta}_1^\tau(t)\phi\right)^2 \alpha_{1u}(t)(1 - \alpha_{1u}(t)) \left(y - \beta_{1u}^\tau(t)\phi\right) \left(y - \beta_1^{*\tau}\phi\right) \mathbb{I}_{\mathcal{A}_{2,j}^c} \right\} \\ \leq & \left(20 d^2_{\max} E_j\left\{\|\phi\|^4 \mathbb{I}_{\mathcal{A}_{2,j}^c}\right\} + 9d_{\max} E_j\left\{\|\phi\|^3 |w| \mathbb{I}_{\mathcal{A}_{2,j}^c}\right\} + E_j\left\{\|\phi\|^2 w^2 \mathbb{I}_{\mathcal{A}_{2,j}^c}\right\}\right) \|\tilde{\beta}_1(t)\|^2 \\ \leq & \left(20 d^2_{\max} c_3 + 9d_{\max} c_3^{3/4}\sigma + dc_2\sigma^2\right) P(\mathcal{A}_{2,j}^c) \|\tilde{\beta}_1(t)\|^2. \end{aligned} \tag{42}$$

Besides, on the event $\mathcal{A}_{2,j} \cap \mathcal{A}_{3,j}^c$, we have $|y^j - \beta_{1u}^\tau(t)\phi| \leq 5\sqrt{2}\sigma\eta + |w|$ and $|y^j - \beta_1^{*\tau}\phi| \leq 4\sqrt{2}\sigma\eta + |w|$, then it follows that

$$\begin{aligned} L_4'(x(t)) = & E_j \left\{ \left(\tilde{\beta}_1^\tau(t)\phi\right)^2 \alpha_{1u}(t)(1 - \alpha_{1u}(t)) \left(y - \beta_{1u}^\tau(t)\phi\right) \left(y - \beta_1^{*\tau}\phi\right) \mathbb{I}_{\{\mathcal{A}_{2,j} \cap \mathcal{A}_{3,j}^c\}} \right\} \\ \leq & \left(40\sigma^2\eta^2 E_j\left\{\|\phi\|^2 \mathbb{I}_{\{\mathcal{A}_{2,j}^c \cap \mathcal{A}_{3,j}^c\}}\right\} + 9\sqrt{2}\sigma\eta E_j\left\{\|\phi\|^2 |w| \mathbb{I}_{\{\mathcal{A}_{2,j}^c \cap \mathcal{A}_{3,j}^c\}}\right\} + E_j\left\{\|\phi\|^2 w^2 \mathbb{I}_{\{\mathcal{A}_{2,j}^c \cap \mathcal{A}_{3,j}^c\}}\right\}\right) \|\tilde{\beta}_1(t)\|^2 \\ \leq & \left(40\sigma^2\eta^2 + 9\sqrt{2}\sigma^2\eta + \sigma^2\right) dc_2 P(\mathcal{A}_{3,j}^c) \|\tilde{\beta}_1(t)\|^2. \end{aligned} \tag{43}$$

By (27) and (40)-(43), we can obtain

$$M_{2,j}(x(t)) = L_1'(x(t)) + L_2'(x(t)) + L_3'(x(t)) + L_4'(x(t)) \leq m_2 \|\tilde{\beta}_1(t)\|^2, \tag{44}$$

where

$$\begin{aligned} m_2 = & \frac{\exp(-\eta^2)}{\pi_{\min}} \left( \frac{5}{4} R^{*2}_{\max} c_3 + \frac{9}{4} R^*_{\max} \sigma \eta c_3^{3/4} + \sigma^2 \eta^2 d c_2 \right) + 2(R^{*2}_{\max} + R^*_{\max} d_{\max}) c_3 \exp\left(-\frac{\eta^2}{2}\right) \\ & + \sqrt{2}\Big( (2R^*_{\max} + d_{\max}) c_3^{3/4} \sigma + \sqrt{3} d c_2 \sigma^2 \Big) \exp\left(-\frac{\eta^2}{4}\right) + \Big( 20 d^2_{\max} c_3 + 9 d_{\max} c_3^{3/4} \sigma + d c_2 \sigma^2 \Big) \frac{8 c_2 d_{\max}}{c_1 R^*_{\min}} \\ & + \Big( 40 \sigma^2 \eta^2 + 9\sqrt{2} \sigma^2 \eta + \sigma^2 \Big) d c_2 \frac{4\sqrt{2} c c_0 \sigma \eta}{c_1 R^*_{\min}}. \end{aligned} \tag{45}$$

From (34), (37) and (44), it follows that

$$M_2(x(t)) \leq \pi_1^* m_1 \|\tilde{\beta}_1(t)\|^2 + (1 - \pi_1^*) m_2 \|\tilde{\beta}_1(t)\|^2, \tag{46}$$

where $m_1$ and $m_2$ are defined in (38) and (45), respectively.

We now analyze the third term on the RHS of (25). For this, let us denote $M_3(x(t)) = \sum\limits_{j=2}^{m} E\left\{ \tilde{\beta}_1^\tau(t)\phi\phi^\tau \tilde{\beta}_j(t) \alpha_{1u}(t) \alpha_{ju}(t) (y - \beta_{ju}^\tau(t)\phi)(y - \beta_1^{*\tau}\phi) \right\}$, then we have

$$M_3(x(t)) = \sum_{l=1}^{m} \pi_l^* \sum_{j=2}^{m} E_l \left\{ \tilde{\beta}_1^\tau(t)\phi\phi^\tau \tilde{\beta}_j(t) \alpha_{1u}(t) \alpha_{ju}(t) (y - \beta_{ju}^\tau(t)\phi)(y - \beta_1^{*\tau}\phi) \right\} \triangleq \sum_{l=1}^{m} \pi_l^* M_{3,l}(x(t)). \tag{47}$$

For the term $M_{3,1}(x(t))$, on the event $A_1 = \mathcal{A}_1 \cap \mathcal{A}_{2,j} \cap \mathcal{A}_{3,j}$, similar to (39), we have $\alpha_{ju}(t) \leq \frac{1}{\pi_{\min}} \exp(-\eta^2)$. Besides, we have $|y^1 - \beta_{ju}^\tau(t)\phi| \leq \frac{5}{4} R^*_{\max}\|\phi\| + \sigma\eta$, then it follows that

$$\begin{aligned} L_{j,1}(x(t)) = & E_1 \left\{ \tilde{\beta}_1^\tau(t)\phi\phi^\tau \tilde{\beta}_j(t) \alpha_{1u}(t) \alpha_{ju}(t) (y - \beta_{ju}^\tau(t)\phi)(y - \beta_1^{*\tau}\phi) \mathbb{I}_{A_1} \right\} \\ \leq & \frac{\exp(-\eta^2)}{\pi_{\min}} \left( \frac{5}{4} R^*_{\max} \sigma \eta E_1 \left\{ \|\phi\|^3 \right\} + \sigma^2 \eta^2 E_1 \left\{ \|\phi\|^2 \right\} \right) \|\tilde{\beta}_1(t)\| \|\tilde{\beta}_j(t)\| \\ \leq & \frac{\exp(-\eta^2)}{\pi_{\min}} \left( \frac{5}{4} R^*_{\max} \sigma \eta c_3^{3/4} + \sigma^2 \eta^2 d c_2 \right) \left( \frac{1}{2} \|\tilde{\beta}_1(t)\|^2 + \frac{1}{2} \|\tilde{\beta}_j(t)\|^2 \right). \end{aligned} \tag{48}$$

On the event $\mathcal{A}_1^c$, we have $|y^1 - \beta_{ju}^\tau(t)\phi| \leq (R^*_{\max} + d_{\max})\|\phi\| + |w|$, then it follows that

$$\begin{aligned} L_{j,2}(x(t)) = & E_1 \left\{ \tilde{\beta}_1^\tau(t)\phi\phi^\tau \tilde{\beta}_j(t) \alpha_{1u}(t) \alpha_{ju}(t) (y - \beta_{ju}^\tau(t)\phi)(y - \beta_1^{*\tau}\phi) \mathbb{I}_{\mathcal{A}_1^c} \right\} \\ \leq & \Big( (R^*_{\max} + d_{\max}) E_1 \left\{ \|\phi\|^3 |w| \mathbb{I}_{\mathcal{A}_1^c} \right\} + E_1 \left\{ \|\phi\|^2 w^2 \mathbb{I}_{\mathcal{A}_1^c} \right\} \Big) \|\tilde{\beta}_1(t)\| \|\tilde{\beta}_j(t)\| \\ \leq & \Big( (R^*_{\max} + d_{\max}) c_3^{3/4} \sigma + \sqrt{3} d c_2 \sigma^2 \Big) \sqrt{P(\mathcal{A}_1^c)} \left( \frac{1}{2} \|\tilde{\beta}_1(t)\|^2 + \frac{1}{2} \|\tilde{\beta}_j(t)\|^2 \right). \end{aligned} \tag{49}$$

On the event $\mathcal{A}_{2,j}^c$, we have $|y^1 - \beta_{ju}^\tau(t)\phi| \leq 5 d_{\max}\|\phi\| + |w|$, then it follows that

$$\begin{aligned} L_{j,3}(x(t)) = & E_1 \left\{ \tilde{\beta}_1^\tau(t)\phi\phi^\tau \tilde{\beta}_j(t) \alpha_{1u}(t) \alpha_{ju}(t) (y - \beta_{ju}^\tau(t)\phi)(y - \beta_1^{*\tau}\phi) \mathbb{I}_{\mathcal{A}_{2,j}^c} \right\} \\ \leq & \Big( 5 d_{\max} E_1 \left\{ \|\phi\|^3 |w| \mathbb{I}_{\mathcal{A}_{2,j}^c} \right\} + E_1 \left\{ \|\phi\|^2 w^2 \mathbb{I}_{\mathcal{A}_{2,j}^c} \right\} \Big) \|\tilde{\beta}_1(t)\| \|\tilde{\beta}_j(t)\| \\ \leq & \Big( 5 d_{\max} c_3^{3/4} \sigma + d c_2 \sigma^2 \Big) P(\mathcal{A}_{2,j}^c) \left( \frac{1}{2} \|\tilde{\beta}_1(t)\|^2 + \frac{1}{2} \|\tilde{\beta}_j(t)\|^2 \right). \end{aligned} \tag{50}$$

In addition, on the event $\mathcal{A}_{2,j}\cap\mathcal{A}_{3,j}^c$, we have $|y^1-\beta_{ju}^\tau(t)\phi|\le 5\sqrt{2}\sigma\eta+|w|$, then it follows that

$$\begin{aligned}
L_{j,4}(x(t))=&E_1\left\{\tilde{\beta}_1^\tau(t)\phi\phi^\tau\tilde{\beta}_j(t)\alpha_{1u}(t)\alpha_{ju}(t)(y-\beta_{ju}^\tau(t)\phi)(y-\beta_1^{*\tau}\phi)\mathbb{I}_{\{\mathcal{A}_{2,j}\cap\mathcal{A}_{3,j}^c\}}\right\}\\
\le&\left(5\sqrt{2}\sigma\eta E_1\left\{\|\phi\|^2|w|\mathbb{I}_{\{\mathcal{A}_{2,j}\cap\mathcal{A}_{3,j}^c\}}\right\}+E_1\left\{\|\phi\|^2w^2\mathbb{I}_{\{\mathcal{A}_{2,j}\cap\mathcal{A}_{3,j}^c\}}\right\}\right)\|\tilde{\beta}_1(t)\|\|\tilde{\beta}_j(t)\|\\
\le&\left(5\sqrt{2}\sigma^2\eta+\sigma^2\right)dc_2P(\mathcal{A}_{3,j}^c)\left(\frac{1}{2}\|\tilde{\beta}_1(t)\|^2+\frac{1}{2}\|\tilde{\beta}_j(t)\|^2\right).
\end{aligned}\tag{51}$$

From (47) and (48)-(51), we have

$$M_{3,1}(x(t))=\sum_{j=2}^m\sum_{s=1}^4L_{j,s}(x(t))\le\frac{(m-1)m_3}{2}\|\tilde{\beta}_1(t)\|^2+\frac{m_3}{2}\sum_{j=2}^m\|\tilde{\beta}_j(t)\|^2,\tag{52}$$

where

$$\begin{aligned}
m_3=&\frac{\exp(-\eta^2)}{\pi_{\min}}\left(\frac{5}{4}R_{\max}^*\sigma\eta c_3^{3/4}+\sigma^2\eta^2dc_2\right)+\sqrt{2}\big((R_{\max}^*+d_{\max})c_3^{3/4}\sigma+\sqrt{3}dc_2\sigma^2\big)\exp\left(-\frac{\eta^2}{4}\right)\\
&+\big(5d_{\max}c_3^{3/4}\sigma+dc_2\sigma^2\big)\frac{8c_2d_{\max}}{c_1R_{\min}^*}+\big(5\sqrt{2}\sigma^2\eta+\sigma^2\big)dc_2\frac{4\sqrt{2}cc_0\sigma\eta}{c_1R_{\min}^*}.
\end{aligned}\tag{53}$$

For the term $M_{3,l}(x(t)),l\neq1$ on the RHS of (47), if $j=l$, the analysis is similar to that of the term $M_{3,1}(x(t))$ and we can obtain the same result. For the term with $j\neq l$, we reconstruct the following events:

$$\begin{aligned}
&\mathcal{A}_1=\{|w|\le\sigma\eta,w\sim\mathcal{N}(0,\sigma^2)\},\\
&\mathcal{A}_{2,lj}=\{|\phi^\tau(\beta_1(t)-\beta_1^*)|\vee|\phi^\tau(\beta_j(t)-\beta_j^*)|\vee|\phi^\tau(\beta_l(t)-\beta_l^*)|\le0.25|\phi^\tau(\beta_l^*-\beta_1^*)|\vee0.25|\phi^\tau(\beta_l^*-\beta_j^*)|\},\\
&\mathcal{A}_{3,lj}=\{|\phi^\tau(\beta_l^*-\beta_1^*)|\vee|\phi^\tau(\beta_l^*-\beta_j^*)|\ge4\sqrt{2}\sigma\eta\},l=2,\cdots,m,j\neq l.
\end{aligned}\tag{54}$$

Then by Lemma A.3, we have

$$P(\mathcal{A}_1^c)\le2\exp\left(-\frac{\eta^2}{2}\right),P(\mathcal{A}_{2,lj}^c)\le\frac{12c_2d_{\max}}{c_1R_{\min}^*},P(\mathcal{A}_{3,lj}^c)\le\frac{4\sqrt{2}cc_0\sigma\eta}{c_1R_{\min}^*}.\tag{55}$$

On the event $A_1=\mathcal{A}_1\cap\mathcal{A}_{2,lj}\cap\mathcal{A}_{3,lj}$, we have $\alpha_{1u}(t)\alpha_{ju}(t)\le\frac{\exp(-\eta^2)}{\pi_{\min}}$, $|y^l-\beta_{ju}^\tau(t)\phi|\le\frac{5}{4}R_{\max}^*\|\phi\|+\sigma\eta$ and $|y^l-\beta_1^{*\tau}\phi|\le R_{\max}^*\|\phi\|+\sigma\eta$. Besides, on the event $\mathcal{A}_1^c$, we have $|y^l-\beta_{ju}^\tau(t)\phi|\le(R_{\max}^*+d_{\max})\|\phi\|+|w|$ and $|y^l-\beta_1^{*\tau}\phi|\le R_{\max}^*\|\phi\|+|w|$. On the event $\mathcal{A}_{2,lj}^c$, we have $|y^l-\beta_{ju}^\tau(t)\phi|\le5d_{\max}\|\phi\|+|w|$ and $|y^l-\beta_1^{*\tau}\phi|\le4d_{\max}\|\phi\|+|w|$. In addition, on the event $\mathcal{A}_{2,lj}\cap\mathcal{A}_{3,lj}^c$, we have $|y^l-\beta_{ju}^\tau(t)\phi|\le5\sqrt{2}\sigma\eta+|w|$ and $|y^l-\beta_1^{*\tau}\phi|\le4\sqrt{2}\sigma\eta+|w|$.

From the above facts, by using a similar analysis way as that used in (44) of $M_{2,j}(x(t))$, we can obtain

$$E_l\left\{\tilde{\beta}_1^\tau(t)\phi\phi^\tau\tilde{\beta}_j(t)\alpha_{1u}(t)\alpha_{ju}(t)(y-\beta_{ju}^\tau(t)\phi)(y-\beta_1^{*\tau}\phi)\right\}\le\frac{m_4}{2}\|\tilde{\beta}_1\|^2+\frac{m_4}{2}\|\tilde{\beta}_j\|^2,j\neq l,l\neq1,\tag{56}$$

where $m_4=\frac{\exp(-\eta^2)}{\pi_{\min}}\left(\frac{5}{4}R_{\max}^{*2}c_3+\frac{9}{4}R_{\max}^*\sigma\eta c_3^{3/4}+\sigma^2\eta^2dc_2\right)+2(R_{\max}^{*2}+R_{\max}^*d_{\max})c_3\exp\left(-\frac{\eta^2}{2}\right)+\sqrt{2}\big((2R_{\max}^*+d_{\max})c_3^{3/4}\sigma+\sqrt{3}dc_2\sigma^2\big)\exp\left(-\frac{\eta^2}{4}\right)+\left(20d_{\max}^2c_3+9d_{\max}c_3^{3/4}\sigma+dc_2\sigma^2\right)\frac{12c_2d_{\max}}{c_1R_{\min}^*}+\left(40\sigma^2\eta^2+9\sqrt{2}\sigma^2\eta+\sigma^2\right)dc_2\frac{4\sqrt{2}cc_0\sigma\eta}{c_1R_{\min}^*}$. Thus for $l\neq1$, we have

$$M_{3,l}(x(t))\le\frac{m_3}{2}\|\tilde{\beta}_1(t)\|^2+\frac{m_3}{2}\|\tilde{\beta}_l(t)\|^2+\frac{m_4}{2}\sum_{j=2,j\neq l}^m\|\tilde{\beta}_j(t)\|^2,\tag{57}$$

where $m_3$ is defined in (53). Then by (47), (52) and (57), we can obtain

$$M_3(x(t))\le\frac{((m-1)\pi_1^*+(1-\pi_1^*))m_3}{2}\|\tilde{\beta}_1(t)\|^2+\sum_{l=2}^m\frac{(\pi_1^*+\pi_l^*)m_3+(m-2)\pi_l^*m_4}{2}\|\tilde{\beta}_l(t)\|^2.\tag{58}$$

We now analyze the fourth and fifth terms on the RHS of (25). By following the similar way to that of the terms $M_2(x(t))$ in (46) and $M_3(x(t))$ in (58), we can obtain

$$\begin{aligned} &M_4(x(t)) \triangleq E\left\{\|\phi\|\alpha_{1u}(t)(1-\alpha_{1u}(t))\left(y-\beta_1^{*\tau}\phi\right)\right\} \leq \pi_1^* m_5 + (1-\pi_1^*)m_6, \\ &M_5(x(t)) \triangleq E\left\{\|\phi\|\alpha_{1u}(t)\alpha_{ju}(t)\left(y-\beta_1^{*\tau}\phi\right)\right\} \leq \pi_1^* m_7 + (1-\pi_1^*)m_6, \end{aligned} \tag{59}$$

where $m_5 = \sigma\sqrt{dc_2}\left(\delta_1\eta\exp(-\eta^2) + \sqrt{2}\exp\left(-\frac{\eta^2}{4}\right) + \frac{(m-1)\left(8c_2d_{\max}+4\sqrt{2}cc_0\sigma\eta\right)}{c_1R^*_{\min}}\right)$, $m_6 = \frac{\exp(-\eta^2)}{\pi_{\min}}(R^*_{\max}dc_2 + \sigma\eta\sqrt{dc_2}) + 2dc_2R^*_{\max}\exp\left(-\frac{\eta^2}{2}\right) + \sqrt{2dc_2}\sigma\exp\left(-\frac{\eta^2}{4}\right) + \left(\sigma\sqrt{dc_2} + 4d_{\max}dc_2\right)\frac{8c_2d_{\max}}{c_1R^*_{\min}} + \left(4\sqrt{2}\sigma\eta + \sigma\right)\frac{4\sqrt{2}cc_0\sigma\eta\sqrt{d}c_2}{c_1R^*_{\min}}$, and $m_7 = \sigma\sqrt{dc_2}\left(\frac{\exp(-\eta^2)}{\pi_{\min}}\eta + \sqrt{2}\exp\left(-\frac{\eta^2}{4}\right) + \frac{(m-1)\left(8c_2d_{\max}+4\sqrt{2}cc_0\sigma\eta\right)}{c_1R^*_{\min}}\right)$. Thus we can obtain the upper bounds of the last two terms on the RHS of (25).

By (25), (33), (46), (58) and (59), we can obtain

$$J_{1,1}(x(t)) \leq -\bar{c}_{1,1}\|\tilde{\beta}_1(t)\|^2 + \sum_{j=2}^{m}\bar{c}_{1,j}\|\tilde{\beta}_j(t)\|^2 + \bar{c}_{2,1}|\tilde{\pi}_1(t)|^2 + \sum_{j=2}^{m}\bar{c}_{2,j}|\tilde{\pi}_j(t)|^2, \tag{60}$$

where $-\bar{c}_{1,1} = -\bar{m} + \frac{1}{\sigma^2}(\pi_1^*m_1 + (1-\pi_1^*)m_2) + \frac{(m-1)\pi_1^*+(1-\pi_1^*)}{2\sigma^2}m_3 + \frac{\pi_1^*m_5+m(1-\pi_1^*)m_6+(m-1)\pi_1^*m_7}{2\pi_{\min}}$, $\bar{c}_{1,j} = \frac{(m-1)\pi_1^*+(1-\pi_1^*)}{2\sigma^2}$, $\bar{c}_{2,1} = \frac{\pi_1^*m_5+(1-\pi_1^*)m_6}{2\pi_{\min}}$ and $\bar{c}_{2,j} = \frac{\pi_1^*m_7+(1-\pi_1^*)m_6}{2\pi_{\min}}$, $j = 2, \cdots, m$.

**Step 2: Analysis of the term $J_{1,2}(x(t))$.** By the mean value theorem, we have

$$\begin{aligned} &J_{1,2}(x(t)) = \tilde{\pi}_1(t)E\left\{\alpha_1(t) - \alpha_1^*\right\} - \tilde{\pi}_1^2(t) \\ =&\frac{\tilde{\pi}_1(t)}{\sigma^2}E\big\{\tilde{\beta}_1^\tau(t)\phi\alpha_{1u}(t)(1-\alpha_{1u}(t))(y-\beta_{1u}^\tau(t)\phi)\big\} - \frac{\tilde{\pi}_1(t)}{\sigma^2}\sum_{j=2}^{m}E\big\{\tilde{\beta}_j^\tau(t)\phi\alpha_{1u}(t)\alpha_{ju}(t)(y-\beta_{ju}^\tau(t)\phi)\big\} \\ &+\frac{(\tilde{\pi}_1(t))^2}{\pi_{1u}(t)}E\left[\alpha_{1u}(t)(1-\alpha_{1u}(t))\right] - \sum_{j=2}^{m}\frac{\tilde{\pi}_1(t)\tilde{\pi}_j(t)}{\pi_{ju}(t)}E\left[\alpha_{1u}(t)\alpha_{ju}(t)\right] - \tilde{\pi}_1^2(t), \end{aligned} \tag{61}$$

where $\beta_{iu}^\tau(t)\phi$, $\pi_{iu}(t)$ and $\alpha_{iu}(t), i \in [m]$ are the same as that defined in (25).

By following a similar analysis way as that used in (60), we can obtain the upper bound of $J_{1,2}(x(t))$ as follows:

$$J_{1,2}(x(t)) \leq \sum_{j=1}^{m}\bar{c}'_{1,j}\|\tilde{\beta}_j(t)\|^2 - \bar{c}'_{2,1}|\tilde{\pi}_1(t)|^2 + \sum_{j=2}^{m}\bar{c}'_{2,j}|\tilde{\pi}_j(t)|^2, \tag{62}$$

where $\bar{c}'_{1,j} = \frac{\pi_j^*m'_1+(1-\pi_j^*)m'_2}{2\sigma^2}, j \in [m]$, $-\bar{c}'_{2,1} = -1 + \frac{m'_1+m'_2}{2\sigma^2} + \frac{(m+1)m'_3}{2\pi_{\min}}$ and $\bar{c}'_{2,j} = \frac{m'_3}{2\pi_{\min}}, j = 2, \cdots, m$. Here $m'_1 = \frac{\exp(-\eta^2)}{\pi_{\min}}\left(\frac{1}{4}R^*_{\max}dc_2 + \sigma\eta\sqrt{dc_2}\right) + 2dc_2d_{\max}\exp\left(-\frac{\eta^2}{2}\right) + \sqrt{2dc_2}\sigma\exp\left(-\frac{\eta^2}{4}\right) + (d_{\max}dc_2 + \sigma\sqrt{dc_2})\frac{(m-1)\left(8c_2d_{\max}+4\sqrt{2}cc_0\sigma\eta\right)}{c_1R^*_{\min}}$, $m'_2 = \frac{\exp(-\eta^2)}{\pi_{\min}}(\frac{5}{4}R^*_{\max}dc_2 + \sigma\eta\sqrt{dc_2}) + 2dc_2(R^*_{\max} + d_{\max})\exp\left(-\frac{\eta^2}{2}\right) + \sqrt{2dc_2}\sigma\exp\left(-\frac{\eta^2}{4}\right) + \left(\sigma\sqrt{dc_2} + 5d_{\max}dc_2\right)\frac{12c_2d_{\max}}{c_1R^*_{\min}} + \left(5\sqrt{2}\sigma\eta + \sigma\right)\frac{4\sqrt{2}cc_0\sigma\eta\sqrt{d}c_2}{c_1R^*_{\min}}$ and $m'_3 = \frac{\exp(-\eta^2)}{\pi_{\min}} + \frac{1}{2}\exp\left(-\frac{\eta^2}{2}\right) + \frac{(m-1)\left(2c_2d_{\max}+\sqrt{2}cc_0\sigma\eta\right)}{c_1R^*_{\min}}$.

**Step 3: Analysis of the terms $J_{1,3}(x(t))$ and $J_{1,4}(x(t))$.** From ODEs (15), it is clear that $\|G_1(x(t)) - R_1(t)\| \to 0$ as $t \to \infty$. By the boundedness of $\tilde{\beta}_1(t)$, we have $\lim\limits_{t\to\infty} J_{1,3}(x(t)) = 0$. Besides, by (32), we also have $R_1(t) > 0$ for all $t \geq 0$.

We now consider the term $J_{1,4}(x(t))$. First, by (15), we have

$$\begin{aligned} J_{1,4}(x(t)) &= \mathrm{vec}(R_1(t) - \pi_1^*G)^\tau\mathrm{vec}(E\{\alpha_1(t)\phi\phi^\tau - \alpha_1^*\phi\phi^\tau\}) - \|R_1(t) - \pi_1^*G\|_F^2 \\ &\leq -\frac{1}{2}\|R_1(t) - \pi_1^*G\|_F^2 + \frac{1}{2}\|E\{(a_1(t) - a_1^*)\phi\phi^\tau\}\|_F^2. \end{aligned} \tag{63}$$

Similar to (61), by the mean value theorem and following a similar analysis way as that used in (60), we can obtain the upper bound of $E\{(a_1(t)-a_1^*)\phi\phi^\tau\}$ as follows:

$$\begin{aligned}
&E\{(a_1(t)-a_1^*)\phi\phi^\tau\}\\
=&\frac{1}{\sigma^2}E\big\{\tilde{\beta}_1^\tau(t)\phi\alpha_{1u}(t)(1-\alpha_{1u}(t))(y-\beta_{1u}^\tau(t)\phi)\phi\phi^\tau\big\}-\frac{1}{\sigma^2}\sum_{j=2}^m E\big\{\tilde{\beta}_j^\tau(t)\phi\alpha_{1u}(t)\alpha_{ju}(t)(y-\beta_{ju}^\tau(t)\phi)\phi\phi^\tau\big\}\\
&+\frac{\tilde{\pi}_1(t)}{\pi_{1u}(t)}E\left[\alpha_{1u}(t)(1-\alpha_{1u}(t))\phi\phi^\tau\right]-\sum_{j=2}^m\frac{\tilde{\pi}_j(t)}{\pi_{ju}(t)}E\left[\alpha_{1u}(t)\alpha_{ju}(t)\phi\phi^\tau\right]\\
\le&\sum_{j=1}^m\bar{c}''_{1,j}\|\tilde{\beta}_j(t)\|+\sum_{j=1}^m\bar{c}''_{2,j}|\tilde{\pi}_j(t)|,
\end{aligned}\tag{64}$$

where $\bar{c}''_{1,j}=\frac{\pi_j^*m_1''+(1-\pi_j^*)m_2''}{2\sigma^2}$, $\bar{c}''_{2,1}=\frac{m_1''+m_2''}{2\sigma^2}+\frac{(m+1)m_3'}{2\pi_{\min}}$ and $\bar{c}''_{2,j}=\frac{dc_2m_3''}{2\pi_{\min}}$. Here $m_1''=\frac{\exp(-\eta^2)}{\pi_{\min}}\left(\frac{1}{4}R_{\max}^*c_3+\sigma\eta c_3^{3/4}\right)+2c_3d_{\max}\exp\left(-\frac{\eta^2}{2}\right)+\sqrt{2}c_3^{3/4}\sigma\exp\left(-\frac{\eta^2}{4}\right)+(d_{\max}c_3+\sigma c_3^{3/4})\frac{(m-1)\left(8c_2d_{\max}+4\sqrt{2}cc_0\sigma\eta\right)}{c_1R_{\min}^*}$, $m_2''=\frac{\exp(-\eta^2)}{\pi_{\min}}(\frac{5}{4}R_{\max}^*c_3+\sigma\eta c_3^{3/4})+2c_3(R_{\max}^*+d_{\max})\exp\left(-\frac{\eta^2}{2}\right)+\sqrt{2}c_3^{3/4}\sigma\exp\left(-\frac{\eta^2}{4}\right)+\left(\sigma c_3^{3/4}+5d_{\max}c_3\right)\frac{12c_2d_{\max}}{c_1R_{\min}^*}+\left(5\sqrt{2}\sigma\eta+\sigma\right)\frac{4\sqrt{2}cc_0\sigma\eta\sqrt{d}c_2}{c_1R_{\min}^*}$ and $m_3''=\frac{\exp(-\eta^2)}{\pi_{\min}}+\frac{1}{2}\exp\left(-\frac{\eta^2}{2}\right)+\frac{(m-1)\left(2c_2d_{\max}+\sqrt{2}cc_0\sigma\eta\right)}{c_1R_{\min}^*}$. Thus by (63) and (64), we can obtain that

$$J_{1,4}(x(t))\le-\frac{1}{2}\|R_1(t)-\pi_1^*G\|_F^2+dm\sum_{j=1}^m\bar{c}''^2_{1,j}\|\tilde{\beta}_j(t)\|^2+dm\sum_{j=1}^m\bar{c}''^2_{2,j}|\tilde{\pi}_j(t)|^2\tag{65}$$

By (24), (60) and (62), we can obtain

$$J_1(x(t))\le-c_{1,1}^{(1)}\|\tilde{\beta}_1(t)\|^2+\sum_{j=2}^m c_{1,j}^{(1)}\|\tilde{\beta}_j(t)\|^2-c_{2,1}^{(1)}|\tilde{\pi}_1(t)|^2+\sum_{j=2}^m c_{2,j}^{(1)}|\tilde{\pi}_j(t)|^2-\frac{1}{2}\|R_1(t)-\pi_1^*G\|_F^2,\tag{66}$$

where $c_{1,1}^{(1)}=\bar{c}_{1,1}-\bar{c}'_{1,1}-dm\bar{c}''^2_{1,1}$, $c_{1,j}^{(1)}=\bar{c}_{1,j}+\bar{c}'_{1,j}+dm\bar{c}''^2_{1,j}$, $c_{2,1}^{(1)}=\bar{c}'_{2,1}-\bar{c}_{2,1}-dm\bar{c}''^2_{2,1}$, and $c_{2,j}^{(1)}=\bar{c}_{2,j}+\bar{c}'_{2,j}+dm\bar{c}''^2_{1,j}$, $j=2,\cdots,m$.

Similarly, we can obtain the same result for $J_i(x(t))$ defined in (21) with the corresponding sequences $c_{1,j}^{(i)}$ and $c_{2,j}^{(i)}$, $i,j\in[m]$. Thus there exist sequences $\{s_{1,i}\}$ and $\{s_{2,i}\}$, $i\in[m]$ such that

$$\frac{d}{dt}V(x(t))\le-\sum_{i=1}^m s_{1,i}\|\tilde{\beta}_i(t)\|^2-\sum_{i=1}^m s_{2,i}|\tilde{\pi}_i(t)|^2+\frac{1}{2}\sum_{i=1}^m\tilde{\beta}_i^\tau(t)(G_i(x(t))-R(t))\tilde{\beta}_i(t)-\frac{1}{2}\sum_{i=1}^m\|R_i(t)-\pi_i^*G\|_F^2\tag{67}$$

holds, where $s_{1,i}=-c_{1,i}^{(i)}+\sum\limits_{j\neq i}c_{1,j}^{(i)}$ and $s_{2,i}=-c_{2,i}^{(i)}+\sum\limits_{j\neq i}c_{2,j}^{(i)}$. From the above analysis, we can obtain

$$\begin{aligned}
s_{1,i}&=-c_1^2+O\left(\frac{m\sigma}{\pi_{\min}}\left(\exp(-\eta^2)\eta^2+\exp\left(-\frac{\eta^2}{2}\right)(d_{\max}^2+d_{\max})+\frac{d_{\max}}{R_{\min}^*}+\frac{\eta^3}{R_{\min}^*}\right)\right),\\
s_{2,i}&=-1+O\left(\frac{m\sigma}{\pi_{\min}}\left(\exp(-\eta^2)\eta^2+\exp\left(-\frac{\eta^2}{2}\right)(d_{\max}^2+d_{\max})+\frac{d_{\max}}{R_{\min}^*}+\frac{\eta^3}{R_{\min}^*}\right)\right).
\end{aligned}\tag{68}$$

By (17), we can see that the region $\bar{S}$ is determined by the parameters $d_i$ of $D_i$, $i\in[m]$. Let the attraction domain $S_A$ be sufficiently close to $\bar{S}$, i.e., $S_A$ is associated with the parameter $d_i+\epsilon, i\in[m]$, where $\epsilon$ is a sufficiently small positive constant. Choose $\eta^2=\log(b_1m/\pi_{\min})$, and $b_2$ is a large constant such that for a sufficiently small $\epsilon>0$, $R_{\min}^*\ge b_2m\max\{\sigma[\log(b_1m/\pi_{\min})]^{3/2},(d_{\max}+\epsilon)^3,(d_{\max}+\epsilon)\}$, thus we have $s_{1,i}>0$ and $s_{2,i}>0$. Then we have

$$\frac{d}{dt}V(x(t))\le 0,\text{ if }x(t)\in S_A,\tag{69}$$

and the equality holds if and only if $x(t)=x^*$. By the LaSalle invariance principle, $D_c=\{x^*\}$ is the invariant set with the domain of attraction $S_A$. This completes the proof of this lemma. □

### A.2. Proof of Theorem 3.8

We now provide the proof of Theorem 3.8. For this, we first present two key lemmas in (Chen & Guo, 2012).

**Lemma A.4.** *Let $\{w_n, \mathcal{F}_n\}$ be a martingale difference sequence satisfying*

$$\sup_n E\{|w_{n+1}|^{\alpha}|\mathcal{F}_n\} < \infty, \ \textit{a.s.}, \ \alpha \in (0, 2].$$

*Then for any adapted sequence $\{f_n, \mathcal{F}_n\}$, we have*

$$\sum_{i=0}^{n} f_i w_{i+1} = O(s_n(\alpha) \log^{\frac{1}{\alpha}+\eta}(s_n^{\alpha}(\alpha) + e)), \ \textit{a.s.}, \forall \eta > 0,$$

*where $S_n^{\alpha}(\alpha) = \left(\sum\limits_{i=0}^{n} |f_i|^{\alpha}\right)$.*

**Lemma A.5.** *Let $X_1, X_2, \cdots$ be a sequence of vectors in $\mathbb{R}^d$ and $A_{n+1} = A_0 + \sum\limits_{k=1}^{n} X_k X_k^{\tau}$. Then we have*

$$\sum_{k=1}^{n} \frac{X_k^{\tau} A_k^{-1} X_k}{1 + X_k^{\tau} A_k^{-1} X_k} = O\left(\log\left(|A_n|\right)\right),$$

*where $A_0 > 0$ and $|A_n|$ is the determinant of $A_n$.*

*Proof of Theorem 3.9:* Let us construct the following stochastic Lyapunov function:

$$V_{k+1} = \sum_{i=1}^{m} \tilde{\beta}_{k+1,i}^{\tau} P_{k+1,i}^{-1} \tilde{\beta}_{k+1,i} + \sum_{i=1}^{m} (k+1)|\tilde{\pi}_{k+1,i}|^2. \tag{70}$$

Denote $\bar{\beta}_{k+1,i} = \beta_{k,i} + a_{k,i}\alpha_{k,i}P_{k,i}\phi_k(y_{k+1} - \beta_{k,i}^{\tau}\phi_k)$ and $\bar{\pi}_{k+1,i} = \pi_{k,i} + \frac{1}{k}(\alpha_{k,i} - \pi_{k,i})$. From the almost sure convergence result established in Theorem 3.1, there exists a random integer $K$ such that for all $k \geq K$, $\bar{\beta}_{k+1,i} \in D_i, (i \in [m]), \bar{\pi}_{k+1,i} \in D_{m+1}$ a.s. Besides, by Algorithm 1 and the matrix inverse formula, we have

$$P_{k+1,i}^{-1} = P_{k,i}^{-1} + \alpha_{k,i}\phi_k\phi_k^{\tau}, \ \ P_{k+1,i}\phi_k = a_{k,i}P_{k,i}\phi_k. \tag{71}$$

Thus by (7), (70) and (71), we know that for $k \geq K$,

$$\begin{aligned} V_{k+1} =& V_k - \sum_{i=1}^{m} a_{k,i}\alpha_{k,i}(\tilde{\beta}_{k,i}^{\tau}\phi_k)^2 + 2\sum_{i=1}^{m} a_{k,i}\alpha_{k,i}(\tilde{\beta}_{k,i}^{\tau}\phi_k)(y_{k+1} - \beta_i^{*\tau}\phi_k) \\ &+ \sum_{i=1}^{m} a_{k,i}\alpha_{k,i}\phi_k^{\tau}P_{k,i}\phi_k(y_{k+1} - \beta_i^{*\tau}\phi_k)^2. \end{aligned} \tag{72}$$

Summing up both sides of (72) from $k = K$ to $n$, we have

$$\begin{aligned} V_{n+1} = V_K &- \sum_{k=K}^{n}\sum_{i=1}^{m} a_{k,i}\pi_i^*(\tilde{\beta}_{k,i}^{\tau}\phi_k)^2 + \sum_{k=K}^{n}\sum_{i=1}^{m} a_{k,i}[\pi_i^* - \alpha_{k,i}](\tilde{\beta}_{k,i}^{\tau}\phi_k)^2 \\ &+ 2\sum_{k=K}^{n}\sum_{i=1}^{m} a_{k,i}\alpha_{k,i}^*(\tilde{\beta}_{k,i}^{\tau}\phi_k)(y_{k+1} - \beta_i^{*\tau}\phi_k) + 2\sum_{k=K}^{n}\sum_{i=1}^{m} a_{k,i}[\alpha_{k,i} - \alpha_{k,i}^*](\tilde{\beta}_{k,i}^{\tau}\phi_k)(y_{k+1} - \beta_i^{*\tau}\phi_k) \\ &+ \sum_{k=K}^{n}\sum_{i=1}^{m} a_{k,i}\alpha_{k,i}\phi_k^{\tau}P_{k,i}\phi_k(y_{k+1} - \beta_i^{*\tau}\phi_k)^2 \\ &- \sum_{k=K}^{n}\sum_{i=1}^{m}\left(1 + \frac{1}{k} - \frac{1}{k^2}\right)|\tilde{\pi}_{k,i}|^2 - 2\sum_{k=K}^{n}\sum_{i=1}^{m}(1 - \frac{1}{k^2})\tilde{\pi}_{k,i}(\alpha_{k,i} - \pi_i^*) + \sum_{k=K}^{n}\sum_{i=1}^{m}\frac{k+1}{k^2}(\alpha_{k,i} - \pi_{k,i})^2, \end{aligned} \tag{73}$$

where $\alpha^*_{k,i} = \left[\pi^*_i \exp\left(-\frac{(y_{t+1}-\beta_i^{*\tau}\phi_k)^2}{2\sigma^2}\right)\right] / \left[\sum_{j=1}^m \pi^*_j \exp\left(-\frac{(y_{k+1}-\beta_j^{*\tau}\phi_k)^2}{2\sigma^2}\right)\right]$.

We now analyze the RHS of (73) term by term.

For the second term on the RHS of (73), by Assumption 2.9 and the convergence of estimates, we have $a_{k,i} \to 1,$ a.s. Besides, by Assumption 3.6 (1) and Lemma A.4, we have for any $\eta > 0$,

$$\sum_{k=K}^{n}\sum_{i=1}^{m}\pi^*_i\tilde{\beta}^\tau_{k,i}[\phi_k\phi_k^\tau - E\{\phi_k\phi_k^\tau|\mathcal{F}_{k-1}\}]\tilde{\beta}_{k,i} = o(n^{\frac{1}{2}+\eta}),\ \text{a.s.} \tag{74}$$

Thus by Assumption 2.9, Assumption 3.6 and Lemma A.4, we know that

$$\begin{aligned}
&\sum_{k=K}^{n}\sum_{i=1}^{m}a_{k,i}\pi^*_i(\tilde{\beta}^\tau_{k,i}\phi_k)^2 \geq \sum_{k=K}^{n}\sum_{i=1}^{m}\pi^*_i(\tilde{\beta}^\tau_{k,i}\phi_k)^2 - o(\sum_{k=K}^{n}\sum_{i=1}^{m}\pi^*_i(\tilde{\beta}^\tau_{k,i}\phi_k)^2)\\
\geq &\sum_{k=K}^{n}\sum_{i=1}^{m}\pi^*_i\tilde{\beta}^\tau_{k,i}E\{\phi_k\phi_k^\tau|\mathcal{F}_{k-1}\}\tilde{\beta}_{k,i} - o(\sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\beta}_{k,i}\|^2) - o(n^{\frac{1}{2}+\eta})\\
\geq &\underline{c}\sum_{k=K}^{n}\sum_{i=1}^{m}\pi^*_i\|\tilde{\beta}_{k,i}\|^2 - -o(\sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\beta}_{k,i}\|^2) - o(n^{\frac{1}{2}+\eta}),\ \text{a.s.}, \forall\eta > 0.
\end{aligned} \tag{75}$$

For the third term on the RHS of (73), by Lemma A.2, we have $E\{\alpha^*_{k,i}|\mathcal{F}_k\} = \pi^*_i$. By Assumption 2.9 and Lemma A.4, we have

$$\sum_{k=K}^{n}\sum_{i=1}^{m}a_{k,i}[\pi^*_i - \alpha^*_{k,i}](\tilde{\beta}^\tau_{k,i}\phi_k)^2 = o(n^{\frac{1}{2}+\eta}) + O(1),\ \text{a.s.}, \forall\eta > 0. \tag{76}$$

Besides, by Assumption 2.9, Assumption 3.6, Lemma A.4 and the proof of Theorem 3.1, we know that

$$\begin{aligned}
&\sum_{k=K}^{n}\sum_{i=1}^{m}a_{k,i}[\pi^*_i - \alpha_{k,i}](\tilde{\beta}^\tau_{k,i}\phi_k)^2\\
=&\sum_{k=K}^{n}\sum_{i=1}^{m}E\{a_{k,i}[\alpha^*_{k,i} - \alpha_{k,i}](\tilde{\beta}^\tau_{k,i}\phi_k)^2|\mathcal{F}_{k-1}\} + \sum_{k=K}^{n}\sum_{i=1}^{m}a_{k,i}[\pi^*_i - \alpha^*_{k,i}](\tilde{\beta}^\tau_{k,i}\phi_k)^2\\
&+\sum_{k=K}^{n}\sum_{i=1}^{m}\left[a_{k,i}[\alpha^*_{k,i} - \alpha_{k,i}](\tilde{\beta}^\tau_{k,i}\phi_k)^2 - E\{a_{k,i}[\alpha^*_{k,i} - \alpha_{k,i}](\tilde{\beta}^\tau_{k,i}\phi_k)^2|\mathcal{F}_{k-1}\}\right]\\
=&o\left(\bar{c}\sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\beta}_{k,i}\|^2\right) + o(n^{\frac{1}{2}+\eta}) + O(1),\ \text{a.s.}, \forall\eta > 0.
\end{aligned} \tag{77}$$

For the fourth term on the RHS of (73), by Lemma A.2, it follows that $E\{\alpha^*_{k,i}y_{k+1}|\mathcal{F}_k\} = \beta_i^{*\tau}\phi_k$, thus we have $\{\alpha^*_{k,i}\phi_k(y_{k+1} - \beta_i^{*\tau}\phi_k), \mathcal{F}_{k-1}\}$ is a martingale difference sequence. Then by Assumption 3.6 and Lemma A.4, we can obtain

$$2\sum_{k=K}^{n}\sum_{i=1}^{m}a_{k,i}\alpha^*_{k,i}(\tilde{\beta}^\tau_{k,i}\phi_k)(y_{k+1} - \beta_i^{*\tau}\phi_k) = o\left(\sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\beta}_{k,i}\|^2\right) + o(n^{\frac{1}{2}+\eta}) + O(1),\ \text{a.s.}, \forall\eta > 0. \tag{78}$$

For the fifth term on the RHS of (73), let us denote

$$S_n = 2\sum_{k=K}^{n}\sum_{i=1}^{m}E\{a_{k,i}[\alpha_{k,i} - \alpha^*_{k,i}](\tilde{\beta}^\tau_{k,i}\phi_k)(y_{k+1} - \beta_i^{*\tau}\phi_k)|\mathcal{F}_{k-1}\}. \tag{79}$$

By Assumption 2.9, Assumption 3.6 and the proof of Theorem 3.1, we know that

$$S_n \leq (\frac{\bar{C}\bar{c}\sigma}{R^*_{\min}} + o(1))(\sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\beta}_{k,i}\|^2 + \sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\pi}_{k,i}\|^2),\ \text{a.s.}, \tag{80}$$

where $\bar{C}$ is a constant whose value can be determined by following the similar proof of Theorem 3.1. Let $\bar{C} = \frac{C}{2}$, then we have $\frac{2\bar{C}\bar{c}\sigma}{R^*_{\min}} \leq \underline{c}\pi_{\min}$. Thus by Lemma A.4, we know that

$$\begin{aligned}
&2\sum_{k=K}^{n}\sum_{i=1}^{m} a_{k,i}[\alpha_{k,i} - \alpha^*_{k,i}](\tilde{\beta}^{\tau}_{k,i}\phi_k)(y_{k+1} - \beta^{*\tau}_i\phi_k) \\
\leq&(\frac{2\bar{C}\bar{c}\sigma}{R^*_{\min}} + o(1))(\sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\beta}_{k,i}\|^2 + \sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\pi}_{k,i}\|^2) + o(n^{\frac{1}{2}+\eta}) + O(1) \\
\leq&(\underline{c}\pi_{\min} + o(1))(\sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\beta}_{k,i}\|^2 + \sum_{k=K}^{n}\sum_{i=1}^{m}\|\tilde{\pi}_{k,i}\|^2) + o(n^{\frac{1}{2}+\eta}) + O(1), \text{ a.s.}, \forall \eta > 0.
\end{aligned} \tag{81}$$

For the sixth term on the RHS of (73), by Assumption 2.9, Lemma A.5 and (71), we know that

$$\sum_{k=K}^{n}\sum_{i=1}^{m} a_{k,i}\alpha_{k,i}\phi^{\tau}_k P_{k,i}\phi_k = O(\sum_{i=1}^{m}\log|P^{-1}_{n+1,i}|) = O(\log n), \text{ a.s.} \tag{82}$$

Thus by Schwartz inequality and Assumptions 2.7 and 2.9, we have

$$\begin{aligned}
&\sum_{k=K}^{n}\sum_{i=1}^{m} a_{k,i}\alpha_{k,i}\phi^{\tau}_k P_{k,i}\phi_k(y_{k+1} - \beta^{*\tau}_i\phi_k)^2 \\
=& O\left(\sum_{k=K}^{n}\sum_{i=1}^{m} a_{k,i}\alpha_{k,i}\phi^{\tau}_k P_{k,i}\phi_k(\|\phi_k\|^2 + w^2_{k+1})\right) = O(\sqrt{n\log n}), \text{ a.s.}
\end{aligned} \tag{83}$$

Following a similar analysis way as that used for the second to sixth term on the RHS of (73), we can obtain the following upper bound of the last two terms on the RHS of (73):

$$-\sum_{k=K}^{n}\sum_{i=1}^{m}|\tilde{\pi}_{k,i}|^2 - 2\sum_{k=K}^{n}\sum_{i=1}^{m}\tilde{\pi}_{k,i}(\alpha_{k,i} - \pi^*_i) = o(n^{\frac{1}{2}+\eta}), \ \sum_{k=K}^{n}\sum_{i=1}^{m}\frac{k+1}{k^2}(\alpha_{k,i} - \pi_{k,i})^2 = O(\log n), \text{ a.s.} \tag{84}$$

Combining (73), (75), (77), (81), (83) and (84), we have for any $\eta > 0$,

$$V_{n+1} = \sum_{i=1}^{m}\tilde{\beta}^{\tau}_{n+1,i}P^{-1}_{n+1,i}\tilde{\beta}_{n+1,i} + \sum_{i=1}^{m}(n+1)|\tilde{\pi}_{n+1,i}|^2 = O(n^{\frac{1}{2}+\eta}), \text{ a.s.} \tag{85}$$

By the convergence property that $\alpha_{k,i} \to \alpha^*_{k,i}$, a.s., we have $\sum\limits_{k=1}^{n}[\alpha_{k,i} - \alpha^*_{k,i}]\phi_k\phi^{\tau}_k = o(\sum\limits_{k=1}^{n}\phi_k\phi^{\tau}_k) = o(n)$, a.s. By Assumption 2.9, the fact $E\{\alpha^*_{k,i}|\mathcal{F}_k\} = \pi^*_i$ and Lemma A.4, we have for any $\eta > 0$, $\sum\limits_{k=1}^{n}[\alpha^*_{k,i} - \pi^*_i]\phi_k\phi^{\tau}_k = o(n^{\frac{1}{2}+\eta})$, a.s. Besides, by Assumption 2.9, we also have $n = O\left(\sum\limits_{k=1}^{n}\pi^*_i\phi_k\phi^{\tau}_k\right)$, a.s. Then by

$$P^{-1}_{n+1,i} = \sum_{k=1}^{n}\alpha_{k,i}\phi_k\phi^{\tau}_k = \sum_{k=1}^{n}[\alpha_{k,i} - \alpha^*_{k,i}]\phi_k\phi^{\tau}_k + \sum_{k=1}^{n}[\alpha^*_{k,i} - \pi^*_i]\phi_k\phi^{\tau}_k + \sum_{k=1}^{n}\pi^*_i\phi_k\phi^{\tau}_k, \tag{86}$$

we know that $n = O(\lambda_{\min}(P^{-1}_{n+1,i}))$, a.s. Thus by (85), we can obtain the desired convergence rate result. □

### A.3. Proof of Theorem 3.9

Without loss of generality, we assume that $\{\phi_k, y_{k+1}\}$ is generated by the $i$-th sub-model, i.e., $y_{k+1} = \beta^{*\tau}_i\phi_k + w_{k+1}$. By Assumptions 2.3-2.4, we have $\phi_k \xrightarrow{d} \phi$ and $w_{k+1} \xrightarrow{d} w$, and $\phi_k$ and $w_{k+1}$ are independent for each $k \geq 0$. Then by Slutsky theorem (Chow & Teicher, 2003), we know that $\phi_k\phi^{\tau}_k \xrightarrow{d} \phi\phi^{\tau}$ and $\phi_k w_{k+1} \xrightarrow{d} \phi w$, where "$\xrightarrow{d}$" means the convergence in distribution. Additionally, from Theorem 3.1, we have $\beta_{k,i} \to \beta_i^*, i \in [m],$ a.s. Denote $M_{k,ij} = (\beta_{k,i} - \beta_{k,j})^\tau \phi_k(-2y_{k+1} + (\beta_{k,i} + \beta_{k,j})^\tau \phi_k)$, and $\mathcal{I}_i(k) = \arg\min_{j \neq i} M_{k,ij}$. Then by the definition of clustering criterion (9) and (87), for any $j \in [m], j \neq i$, we have

$$
\begin{aligned}
&\lim_{k\to\infty} P(\{\phi_k, y_{k+1}\} \text{ is categorized wrongly})\\
=&\lim_{k\to\infty} P\big(\min_{j\neq i}(y_{k+1} - \beta_{k,j}^\tau \phi_k)^2 \leq (y_{k+1} - \beta_{k,i}^\tau \phi_k)^2\big)\\
=&\lim_{k\to\infty} P\big(\min_{j\neq i} M_{k,ij} < 0\big)\\
=&\lim_{k\to\infty} E\left\{P(w_{k+1} > -\text{sgn}((\beta_{k,i} - \beta_{k,\mathcal{I}_i(k)})^\tau \phi_k)\left(\frac{(\beta_{k,i} + \beta_{k,\mathcal{I}_i(k)})^\tau \phi_k}{2} - \beta_i^{*\tau}\phi_k\right)|\phi_k)\right\}\\
\leq&\lim_{k\to\infty} E\left\{\max_{j\neq i} P(w_{k+1} > -\text{sgn}((\beta_{k,i} - \beta_{k,j})^\tau \phi_k)\left(\frac{(\beta_{k,i} + \beta_{k,j})^\tau \phi_k}{2} - \beta_i^{*\tau}\phi_k\right)|\phi_k)\right\}\\
=&\lim_{k\to\infty} E\left\{\max_{j\neq i} \Phi\left(-\frac{\text{sgn}((\beta_{k,i} - \beta_{k,j})^\tau)(\frac{(\beta_{k,i}+\beta_{k,j})^\tau\phi_k}{2} - \beta_i^{*\tau}\phi_k)}{\sigma}\right)\right\}\\
=&E\left\{\max_{j\neq i}\Phi\left(-\frac{|(\beta_i^* - \beta_j^*)^\tau \phi|}{2\sigma}\right)\right\} \leq E\left\{\max_{j\neq i}\exp\left(-\frac{((\beta_i^* - \beta_j^*)^\tau\phi)^2}{8\sigma^2}\right)\right\} < 1.
\end{aligned}
\tag{87}
$$

Thus we can obtain the inequality (11).

We now provide the remaining proof of Theorem 3.9. For this, denote the following events for $i \in [m]$,

$$
\begin{aligned}
\mathcal{A}_{k,i} &= \{\omega : y_{k+1} = \beta_i^{*\tau}\phi_k + w_{k+1}\},\\
\mathcal{B}_{k,ij} &= \left\{\omega : j = \arg\min_{j\in[m]}(y_{k+1} - \beta_{k,j}^\tau \phi_k)^2\right\} \cap \mathcal{A}_{k,i},
\end{aligned}
$$

where $\mathcal{A}_{k,i}$ denotes the events that the data $\{\phi_k, y_{k+1}\}$ is generated by $i$-th sub-models, $\mathcal{B}_{k,ij}$ represents the events that the data $\{\phi_k, y_{k+1}\}$ generated by the $i$-th model is categorized into the $j$-th cluster. Then the within-cluster error (10) can be rewritten as follows:

$$
\begin{aligned}
J_n =&\frac{1}{n}\sum_{k=1}^n\sum_{i=1}^m\sum_{j=1}^m (y_{k+1} - \beta_{k,i}^\tau\phi_k)^2 \mathbb{I}_{\mathcal{B}_{k,ji}}\\
=&\frac{1}{n}\sum_{k=1}^n\sum_{i=1}^m (y_{k+1} - \beta_{k,i}^\tau\phi_k)^2 \mathbb{I}_{\mathcal{A}_{k,i}} + \frac{1}{n}\sum_{k=1}^n\sum_{i=1}^m\left[\sum_{j\neq i}\left((y_{k+1} - \beta_{k,j}^\tau\phi_k)^2 - (y_{k+1} - \beta_{k,i}^\tau\phi_k)^2\right)\mathbb{I}_{\mathcal{B}_{k,ij}}\right]\\
\triangleq& L_{n,1} + L_{n,2}.
\end{aligned}
\tag{88}
$$

We now analyze the RHS of (88) term by term. As for the term $L_{n,1}$, we have

$$
L_{n,1} = \frac{1}{n}\sum_{k=1}^n\sum_{i=1}^m (\tilde{\beta}_{k,i}^\tau\phi_k)^2 \mathbb{I}_{\mathcal{A}_{k,i}} + \frac{2}{n}\sum_{k=1}^n \tilde{\beta}_{k,i}^\tau\phi_k w_{k+1}\mathbb{I}_{\mathcal{A}_{k,i}} + \frac{1}{n}\sum_{k=1}^n w_{k+1}^2,
$$

By Assumption 2.7, we know that $\lim\limits_{n\to\infty}\frac{1}{n}\sum\limits_{k=1}^n w_{k+1}^2 = \sigma^2$. By the fact that $\lim\limits_{k\to\infty}\tilde{\beta}_{k,i} = 0$ from Theorem 3.1 and the average boundedness of $\|\phi_k\|^2$ and $\|\phi_k w_{k+1}\|$ from Lemma A.1, we obtain

$$
\lim_{n\to\infty} L_{n,1} = \sigma^2, \text{ a.s.} \tag{89}
$$

For the term $L_{n,2}$, let us denote the following events:

$$\begin{aligned}
&\mathcal{A}_i = \{\omega : y = \beta_i^{*\tau}\phi + w\}, \quad \mathcal{B}_{ij} = \left\{\omega : (y - \beta_j^{*\tau}\phi)^2 \leq (y - \beta_i^{*\tau}\phi)^2\right\} \cap \mathcal{A}_i, \\
&\bar{\mathcal{B}}_{ij} = \left\{\omega : j = \arg\min_{j \in [m]} (y - \beta_j^{*\tau}\phi)^2\right\} \cap \mathcal{A}_i, \\
&\mathcal{B}^*_{k,ij} = \left\{\omega : j = \arg\min_{j \in [m]} (y_{k+1} - \beta_j^{*\tau}\phi_k)^2\right\} \cap \mathcal{A}_{k,i},
\end{aligned}$$

By Assumptions 2.5-2.9, we have $P\left(\mathcal{B}_{ij}|\phi\right) = \pi_i^* P\left((y - \beta_j^{*\tau}\phi) \leq (y - \beta_i^{*\tau}\phi)\right) = \pi_i^* \Phi\left(-\frac{|\beta_i^{*\tau}\phi - \beta_j^{*\tau}\phi|}{2\sigma}\right)$. Besides, from Lemma A.1, we can see that $\sum_{i=1}^{m}\left[\sum_{j\neq i}\left[(\beta_i^{*\tau}\phi_k - \beta_j^{*\tau}\phi_k)(\beta_i^{*\tau}\phi_k - \beta_j^{*\tau}\phi_k + 2w_{k+1})\right]\mathbb{I}_{\mathcal{B}^*_{k,ij}}\right]$ is asymptotically stationary and ergodic. Thus by Assumptions 2.7 and 2.9, and (10), we obtain

$$\begin{aligned}
&\lim_{n\to\infty} L_{n,2} = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{m}\left[\sum_{j\neq i}\left[(\beta_{k,i}^{\tau}\phi_k - \beta_{k,j}^{\tau}\phi_k)(2\beta_i^{*\tau}\phi_k + 2w_{k+1} - \beta_{k,i}^{\tau}\phi_k - \beta_{k,j}^{\tau}\phi_k)\right]\mathbb{I}_{\mathcal{B}_{k,ij}}\right] \\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{m}\left[\sum_{j\neq i}\left[(\beta_i^{*\tau}\phi_k - \beta_j^{*\tau}\phi_k)(\beta_i^{*\tau}\phi_k - \beta_j^{*\tau}\phi_k + 2w_{k+1})\right]\mathbb{I}_{\mathcal{B}^*_{k,ij}}\right] \\
&= \sum_{i=1}^{m}\left[\sum_{j\neq i} E\left\{(\beta_i^{*\tau}\phi - \beta_j^{*\tau}\phi)^2 E\{\mathbb{I}_{\bar{\mathcal{B}}_{ij}}|\phi\} + 2(\beta_i^{*\tau}\phi - \beta_j^{*\tau}\phi)E\{w\mathbb{I}_{\bar{\mathcal{B}}_{ij}}|\phi\}\right\}\right] \\
&\leq \sum_{i=1}^{m}\min_{j\neq i} E\left\{(\beta_i^{*\tau}\phi - \beta_j^{*\tau}\phi)^2 E\{\mathbb{I}_{\mathcal{B}_{ij}}|\phi\} + 2(\beta_i^{*\tau}\phi - \beta_j^{*\tau}\phi)E\{w\mathbb{I}_{\mathcal{B}_{ij}}|\phi\}\right\} \\
&= \sum_{i=1}^{m}\min_{j\neq i} E\left\{(\beta_i^{*\tau}\phi - \beta_j^{*\tau}\phi)^2 P\left(\mathcal{B}_{ij}|\phi\right) + \frac{2\pi_i^*|\beta_i^{*\tau}\phi - \beta_j^{*\tau}\phi|}{\sqrt{2\pi}\sigma}\int_{-\infty}^{-|\beta_i^{*\tau}\phi - \beta_j^{*\tau}\phi|} w\exp\left(-\frac{w^2}{2\sigma^2}\right)dw\right\} \\
&= \sum_{i=1}^{m}\pi_i^*\min_{j\neq i} E\left\{\gamma_{i,j}(\phi)\right\} \triangleq \gamma.
\end{aligned} \tag{90}$$

Furthermore, by the fact that $\int_{-\infty}^{-a} a\exp\left(-\frac{x^2}{2\sigma^2}\right)dx \leq \sigma^2\exp\left(-\frac{a^2}{2\sigma^2}\right)$ holds for any positive constant $a$, we can see that $\gamma \leq 0$. By (88)-(90), we can obtain (12). This completes the proof of Theorem 3.9. □