2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Learning Personalized Models with Clustered System Identification

Leonardo F. Toso, Han Wang, and James Anderson

Abstract— **We address the problem of learning linear system models by observing multiple trajectories from systems with differing dynamics. This framework encompasses a collaborative scenario where several systems seeking to estimate their dynamics are partitioned into clusters according to system similarity. Thus, the systems within the same cluster can benefit from the observations made by the others. Considering this framework, we present an algorithm where each system alternately estimates its cluster identity and performs an estimation of its dynamics. This is then aggregated to update the model of each cluster. We show that under mild assumptions, our algorithm correctly estimates the cluster identities and achieves an ε-approximate solution with a sample complexity that scales inversely with the number of systems in the cluster, thus facilitating a more efficient and personalized system identification.**

I. Introduction

System identification is the data-driven process of estimating a dynamic model of a system based on observations of the system trajectories. It plays a crucial role in aiding our understanding of complex systems and is a fundamental problem in numerous fields, including time-series analysis, control theory, robotics, and reinforcement learning [1], [2]. The effective utilization of available data is pivotal in obtaining an accurate model estimate with a measure of uncertainty quantification. Traditional system identification, methods [2] have focused on asymptotic analysis, which, although insightful, is restrictive when dealing with small to medium sized data sets. Motivated by this, and the fact that data generation is often costly and time consuming, modern approaches focus on developing sample complexity bounds (i.e., non-asymptotic convergence analysis).

Results on the estimation of both fully [3], [4], [5] and partially [6], [7], [8], [9], [10] observed LTI systems have demonstrated that a more precise characterization of error bounds is essential for designing efficient and robust control systems [4], [8], [11]. These studies provide non-asymptotic bounds that are functions of the number of observed trajectories (see Table 1 of [10] for a summary of the bounds).

A recent body of work has begun to formalize methods for improving sample efficiency by considering data (or models generated from data) from multiple systems [12], [13], [14], [15], [16], [17], [18]. Leveraging data from similar systems provides a promising approach although clarifying the effect of the heterogeneity in the systems and their environments is crucial. The aforementioned work have demonstrated that the benefit of collaboration typically reduces the sample complexity by a factor of the number of collaborators, when compared to the single-agent setting where each system estimate its dynamics from its own observations.

However, the approaches discussed in [12], [13], [14] compute a common estimation for all participants, thereby restricting the ability to obtain personalized estimations. Furthermore, the sample complexity bounds achieved in those studies are subject to an unavoidable heterogeneity bias that cannot be controlled by the number of trajectories or systems, thus leading to an estimation error that scales with the measure of heterogeneity among the considered systems. Specifically in [12], [13], [14] the error of the system identification process is shown to be of order $\mathcal{O}(\frac{1}{\sqrt{N}} + \varepsilon_{\text{het}})$ where ε_{het} characterizes the worst case heterogeneity and N is the number of trajectories across all systems.

Personalization in collaborative settings aims to provide tailored solutions (e.g. model estimates) to individual agents with distinct objectives, while enabling inter-agent collaboration (e.g. model sharing). This encompasses diverse topics such as representation learning [17], [19], [20], [21] and clustering [22], both widely studied in machine learning and data analysis. The present work address the aforementioned challenges by leveraging clustering techniques to achieve personalized model estimations. The idea is simple: cluster systems into groups that have identical system dynamics, and then apply collaborative learning algorithms to the clusters in order to improve sample complexity (by reducing the heterogeneity induced error ε_{het}) and achieve personalization even for heterogeneous settings.

Recent work on clustered federated learning that includes [23], [24], [25] have shown the potential of clustering techniques to collaboratively train models in heterogeneous settings with non-i.i.d. data. Building upon this success, this paper aims to apply clustering to the system identification problem, which poses unique challenges due to the dynamical nature of the system that results in non-i.i.d. data. This is in contrast to the linear regression and model training settings explored in the aforementioned work. Further details on these challenges are discussed later.

Specifically, we investigate the scenario where we have M dynamical systems, with each of them belonging to one of K different system types (which we refer to as a "cluster"). Which cluster a system belongs to is not initially disclosed. Our objective is to simultaneously identify the correct cluster identities for each of the M systems and obtain a system model by collaboratively learning with the systems in the same cluster. Our approach can lead to significant reductions

This material is based upon work supported in part by DoE under grant DE-SC0022234 and NSF awards 2144634 & 2231350. The authors are with the Department of Electrical Engineering, Columbia University in the City of New York, New York, NY, 10027, USA. Email: {lt2879, hw2786, james.anderson}@columbia.edu.

in the amount of data required to accurately estimate the system models, as illustrated in the following theorem.

Theorem 1: (main result, informal) Suppose the K system types are sufficiently different, and we observe the same number of trajectories from each system. Then, for a given cluster, with high probability, the estimation error between the learned and ground truth model is bounded by:

$$\begin{matrix}\text{estimation}\\ \text{error}\end{matrix} \lesssim \frac{1}{\sqrt{\#\text{ systems} \times \#\text{ trajectories}}} + \begin{matrix}\text{misclass.}\\ \text{rate}\end{matrix},$$

with

$$\begin{matrix}\text{misclass.}\\ \text{rate}\end{matrix} \lesssim \exp(-\#\text{ trajectories} \times \text{misclass. const.}).$$

where #systems denotes the number of systems in the cluster, and #trajectories represents the number of trajectories observed by each of them.

The first term captures the error in learning the system dynamics from systems' observations within the same cluster. It shows what one would hope; as the number of systems and observations increase, the error decreases. However, this speedup does not come for free. The second term is the penalty paid for assigning one of the M systems to one of the incorrect K clusters. One of the main results from our work is to show that *both terms* can be controlled by adjusting the number of observed trajectories. Moreover, the misclassification rate is dominated by the first term, thus leading to a an approximate sample complexity that is scale inversely with the number of system within the cluster. This is in stark contrast to [12], [13], [14] which is where the heterogeneity introduces a bias ε which is not a function of the number of systems or the volume of data at our disposal. Our work shows that by controlling both sources of error, our approach can accurately estimate the system dynamics with fewer samples, when compared to the single agent case, and provides better estimation in heterogeneous settings when compared to [12], [13], [14].

Contributions: This is the first work to introduce clustering in order to provide sample complexity gains to the collaborative system identification problem. We derive an upper bound on the estimation error (Theorem 2) that decomposes into two terms (as shown above), where each term can be controlled by adjusting the number of observed trajectories. We offer theoretical guarantees on the probability of cluster identity misclassification (Lemma 1) and thus convergence (Corollary 1). We show that under a mild assumption on the number of observed trajectories, our approach correctly estimates the cluster identities, with high probability. Moreover, we show that our method achieves an improved convergence rate when compared to the single-agent system identification process. In contrast to the federated setting [14], [15] and that of [12], [13], we are able to provide personalized models as opposed to a single generic model, thus expanding the use cases for collaborative system identification.

A. Notation

Given a matrix $G \in \mathbb{R}^{m\times n}$, the Frobenius norm of G is denoted by $\|G\|_F = \sqrt{Tr(GG^\top)}$. $\|G\| = \sigma_{\max}(G)$, where $\sigma_{\max}(G)$ is the largest singular value of G. Consider a symmetric matrix Σ, $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote its minimum and maximum eigenvalues, respectively. For systems, we use superscript (i) to denote the system index and subscript t for time. For models, subscript denotes the cluster identity, and superscript (r) is the iteration counter.

II. Problem Formulation and Algorithm

Consider M linear time-invariant (LTI) systems

$$x_{t+1}^{(i)} = A^{(i)}x_t^{(i)} + B^{(i)}u_t^{(i)} + w_t^{(i)}, \;\; t = 0,1,\ldots,T-1 \qquad (1)$$

where $x_t^{(i)} \in \mathbb{R}^{n_x}$, $u_t^{(i)} \in \mathbb{R}^{n_u}$ and $w_t^{(i)} \in \mathbb{R}^{n_x}$ are the state, input, and process noise at time t, for system $i \in [M]$. We assume that $\{u_t^{(i)}\}_{t=1}^{T-1}, \{w_t^{(i)}\}_{t=1}^{T-1}$ are random vectors distributed according to $u_t^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_{u,i}^2 I_{n_u}\right)$ and $w_t^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_{w,i}^2 I_{n_x}\right)$. Furthermore, it is assumed that $x_0^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_{x,i}^2 I_{n_x}\right)$.

We consider the setting where we have access to M datasets corresponding to observed system trajectories. Each of the datasets is generated by one of K different systems. We consider the case where $K \ll M$. We will from now on refer to the K types of different systems as "clusters", which we label as $\mathscr{C}_1, \ldots, \mathscr{C}_K$. We denote (A_j, B_j) as the ground truth system matrices of cluster $j \in [K]$. That is, $A^{(i)} = A_j$, and $B^{(i)} = B_j$, for any $i \in \mathscr{C}_j$. Note that due to the noise in model (1), two datasets generated by cluster $\mathscr{C}_j$ will be different.

The state-input pair of a single trajectory $\{x_t^{(i)}, u_t^{(i)}\}$ of system $i \in \mathscr{C}_j$ is referred to as *rollout*. We consider the setting where multiple rollouts of length T are collected and stored as $\left\{x_{l,t}^{(i)}, u_{l,t}^{(i)}\right\}_{t=0}^{T-1}$, for $l = 1, \ldots N_i$, with l denoting the l-th rollout and t the t-th time-step of the corresponding rollout. Thus, for any system $i \in \mathscr{C}_j$ and cluster $j \in [K]$, the system dynamics is described by:

$$x_{l,t+1}^{(i)} = \Theta_j z_{l,t}^{(i)} + w_{l,t}^{(i)} \quad \forall\; 1 \le l \le N_i \text{ and } 0 \le t \le T-1, \;\; (2)$$

where $z_{l,t}^{(i),\top} \triangleq \begin{bmatrix} x_{l,t}^{(i),\top} & u_{l,t}^{(i),\top} \end{bmatrix} \in \mathbb{R}^{n_x+n_u}$ corresponds to the augmented state-input pair of system $i \in \mathscr{C}_j$ over rollout l at time t, and $\Theta_j \triangleq [A_j \;\; B_j]$ denotes the concatenation of the ground truth system matrices A_j and B_j. The state update $x_{l,t+1}^{(i)}$ can be expanded recursively as follows:

$$x_{l,t}^{(i)} = G_t^{(i)} \begin{bmatrix} u_{l,0}^{(i)} \\ \vdots \\ u_{l,t-1}^{(i)} \end{bmatrix} + F_t^{(i)} \begin{bmatrix} w_{l,0}^{(i)} \\ \vdots \\ w_{l,t-1}^{(i)} \end{bmatrix} + A_j^t x_{l,0}^{(i)},$$

where, $G_t^{(i)} \triangleq \begin{bmatrix} A_j^{t-1}B_j & A_j^{t-2}B_j & \cdots & B_j \end{bmatrix}$ and $F_t \triangleq \begin{bmatrix} A_j^{t-1} & A_j^{t-2} & \cdots & I_{n_x} \end{bmatrix}$ for all $t \ge 1$.

The state-input pair $z_{l,t}^{(i)}$ is distributed according to a Gaussian distribution with zero mean and covariance matrix $\Sigma_t^{(i)}$, where,

$$\Sigma_0^{(i)} \triangleq \begin{bmatrix} \sigma_{x,i}^2 I_{n_x} & 0 \\ 0 & \sigma_{u,i}^2 I_{n_u} \end{bmatrix} \succ 0, \quad \text{for } t = 0,$$

and

$$\Sigma_t^{(i)} \triangleq \begin{bmatrix} \sigma_{u,i}^2 G_t^{(i)} G_t^{(i),\top} + \sigma_{w,i}^2 F_t^{(i)} F_t^{(i),\top} + \sigma_{x,i}^2 A_j^t (A_j^t)^\top & 0 \\ 0 & \sigma_{u,i}^2 I_{n_u} \end{bmatrix},$$

for all $t \geq 1$ and $i \in \mathscr{C}_j$, $\forall j \in [K]$, as detailed in [14].

Next, we define the offline batch matrices for each system $i \in \mathscr{C}_j$, $\forall j \in [K]$. For a single rollout l, the data is concatenated according to $X_l^{(i)} = \begin{bmatrix} x_{l,T}^{(i)} & \cdots & x_{l,1}^{(i)} \end{bmatrix} \in \mathbb{R}^{n_x \times T}$, $Z_l^{(i)} = \begin{bmatrix} z_{l,T-1}^{(i)} & \cdots & z_{l,0}^{(i)} \end{bmatrix} \in \mathbb{R}^{(n_x+n_u)\times T}$, and $W_l^{(i)} = \begin{bmatrix} w_{l,T-1}^{(i)} & \cdots & w_{l,0}^{(i)} \end{bmatrix} \in \mathbb{R}^{n_x \times T}$. This is then further stacked to construct the batch matrices $X^{(i)} = \begin{bmatrix} X_1^{(i)} & \dots & X_{N_i}^{(i)} \end{bmatrix} \in \mathbb{R}^{n_x \times N_i T}$, $Z^{(i)} = \begin{bmatrix} Z_1^{(i)} & \cdots & Z_{N_i}^{(i)} \end{bmatrix} \in \mathbb{R}^{(n_x+n_u)\times N_i T}$, and $W^{(i)} = \begin{bmatrix} W_1^{(i)} & \cdots & W_{N_i}^{(i)} \end{bmatrix} \in \mathbb{R}^{n_x \times N_i T}$. Therefore, for each system $i \in \mathscr{C}_j$, $\forall j \in [K]$, its state, input, noise, and model parameters are related according to

$$X^{(i)} = \Theta_j Z^{(i)} + W^{(i)}, \tag{3}$$

where each column of $Z^{(i)}$ and $W^{(i)}$ are sampled according to Gaussian distributions with zero means and covariance matrices $\Sigma_t^{(i)}$, $\sigma_{w,i}^2 I_{n_x}$, respectively. With that said, we are now able to introduce the clustered system identification problem.

Problem 1: We consider M dynamical systems as in (1) that are equipped with batch matrices $X^{(i)}, Z^{(i)}$, and $W^{(i)}$. Each system $i \in [M]$ is associated with its own cost function $C^{(i)}(\Theta) = \|X^{(i)} - \Theta Z^{(i)}\|_F^2$, and is unaware of its cluster identity. We aim to estimate the systems' cluster identities $\widehat{\mathscr{C}}_1, \dots, \widehat{\mathscr{C}}_K$ and use it to estimate a model $\widehat{\Theta}_j = [\widehat{A}_j \ \widehat{B}_j]$ which is close to the ground truth Θ_j, $\forall j \in [K]$.

To obtain a faster and more accurate estimation, we frame the system identification problem in the setting where systems within the same cluster can leverage data from each other. Further in this paper, we provide theoretical guarantees to support these statements.

The problem described above can be framed into an alternating optimization problem, as the actual cluster identity of each system (i.e., $\mathscr{C}_1, \dots, \mathscr{C}_K$) is not disclosed to the systems in advance. Therefore, our objective is twofold: firstly, we aim to classify the correct cluster identities of the systems by employing the Mean Square Error (MSE) as the clustering criterion, with the resulting output being the cluster estimation (CE); secondly, we use that estimation to identify the model dynamics of each cluster with a model estimation (ME) step. Next, we introduce our clustered system identification algorithm to solve this problem.

The initial step of Algorithm 1 involves the initialization of the number of clusters and the provision of an initial guess for the dynamics of each cluster. Subsequently, the algorithm iterates from line 2 to 11, during which each system estimates its corresponding cluster identity and stores this information in the form of a one-hot encoding vector denoted by e_i. The one-hot encoding vector comprises K elements, with one in the position of the estimated cluster identity and zero elsewhere. After the estimation of the cluster identity, the cluster model is updated by performing a single gradient descent iteration in line 10, with the gradient being the average of the gradients of each individual system's cost function that belongs to the cluster.

Algorithm 1 `Clustered System Identification`

1: **Initialization:** number of clusters K, step-size η_j, and model initialization $\widehat{\Theta}_j^{(0)}$ $\forall j \in [K]$,
2: **for** each iteration $r = 0, 1, \dots, R-1$ **do**
3: The systems receive the models $\{\widehat{\Theta}_1^{(r)}, \dots, \widehat{\Theta}_K^{(r)}\}$, $\forall j \in [K]$,
4: **Cluster estimation (CE):**
5: **for** each system $i \in [M]$
6: $\hat{j} = \text{argmin}_{j\in[K]} \|X^{(i)} - \widehat{\Theta}_j^{(r)} Z^{(i)}\|_F^2$,
7: define $e_i = \{e_{i,j}\}_{j=1}^K$ with $e_{i,j} = \mathbb{1}\{j = \hat{j}\}$,
8: **end for**
9: **Model estimation (ME):**
10: $\widehat{\Theta}_j^{(r+1)} = \widehat{\Theta}_j^{(r)} + \frac{2\eta_j}{\sum_{i\in[M]} e_{i,j}} \sum_{i\in[M]} e_{i,j} (X^{(i)} - \widehat{\Theta}_j^{(r)} Z^{(i)}) Z^{(i),\top}$ for all $j \in [K]$
11: **end for**
12: **Return** $\widehat{\Theta}_j^{(R)}$ for all $j \in [K]$.

Remark 1: Note that Algorithm 1 is an alternating minimization algorithm, where it performs an iterative clustering step followed by a model estimation process. Prior to the start of collaboration, each system $i \in [M]$ collects data and stores it in batch matrices $X^{(i)}, Z^{(i)}$, and $W^{(i)}$. Moreover, it is worth noting that Algorithm 1 uses the same batch matrices for both cluster identity and model estimation.

The following definitions and assumptions are required in order to analyze Algorithm 1. Subsequently, we provide the intuition behind them.

Definition 1: The minimum and maximum separation between the clusters are defined as

$$\Delta_{\min} \triangleq \min_{j \neq j'} \|\Theta_j - \Theta_{j'}\| \quad \text{and} \quad \Delta_{\max} \triangleq \max_{j \neq j'} \|\Theta_j - \Theta_{j'}\|$$

respectively.

We define $\rho^{(i)} \triangleq \frac{\Delta_{\min}^2}{\sigma_{w,i}^2}$ as the signal-to-noise ratio $\forall i \in [M]$.

Assumption 1: The initial model estimate $\widehat{\Theta}_j^{(0)}$ satisfy $\|\widehat{\Theta}_j^{(0)} - \Theta_j\| \leq \left(\frac{1}{2} - \alpha^{(0)}\right) \Delta_{\min}, \forall j \in [K]$, where $0 < \alpha^{(0)} < \frac{1}{2}$.

Assumption 2: For any fixed and small δ, the number of trajectories satisfies $N_i n_x \gtrsim \left(\frac{\rho^{(i)} \|\Sigma_t^{(i)}\| + \sqrt{n_x}}{\alpha^{(0)} \rho^{(i)} \|\Sigma_t^{(i)}\|}\right)^2 \log(\frac{MT}{\delta})$, for all $i \in [M]$. We also assume that $\Delta_{\min} \gtrsim 1 + \Delta_{\max} \sum_{i\in[M]} \sum_{t=0}^{T-1} \exp\left(-c N_i n_x \left(\frac{\alpha^{(0)} \rho^{(i)} \|\Sigma_t^{(i)}\|}{\rho^{(i)} \|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right)$ for some constant c.

Assumption 1 implies that the initial guess for the model estimates is superior to a random initialization. This assumption is standard for alternating minimization algorithms, particularly for learning mixture models [26]. The condition on the number of trajectories in Assumption 2 is a common requirement in the concentration bound analysis. This is used to guarantee that the cluster estimation procedure of

Algorithm 1 correctly estimate the cluster identities, with high probability. Note that this is a mild assumption since for well-behaved systems where $\Sigma_t^{(i)}$ is well conditioned, $N_i n_x$ is typically in the same or superior to the order of $\log\left(\frac{MT}{\delta}\right)$. The condition on $\Delta_{\min}$ in Assumption 2 is to ensure that any two clusters are well-separated. This is a standard assumption in the literature of clustering [27], [28]. Similar assumptions are exploited in [23] in the context of the linear regression problem.

III. Theoretical Guarantees

We begin our analysis by examining a single iteration of Algorithm 1. For simplicity, we omit the superscript r that denotes the iteration counter. Let us assume that we have the current estimated model $\widehat{\Theta}_j$ for all clusters $j \in [K]$ at a given iteration, such that $\|\widehat{\Theta}_j - \Theta_j\| \leq \left(\frac{1}{2} - \alpha\right)\Delta_{\min}$ for all $j \in [K]$, with $0 < \alpha < \frac{1}{2}$.

A. Probability of Cluster Identity Misclassification

Consider a system $i \in [M]$ within cluster $\mathscr{C}_j$. Let $\mathscr{M}_i^{j,j'}$ be the event in which system i is inaccurately classified as belonging to cluster $\mathscr{C}_{j'}$. The event when system i is *correctly* classified is denoted as $\mathscr{M}_i^{j,j}$. The following lemma provides an upper bound on the probability of misclassification.

Lemma 1: Suppose that $i \in \mathscr{C}_j$. There exist universal constants c_1 and c_2, such that for any $j' \neq j$,

$$\mathbb{P}\left\{\mathscr{M}_i^{j,j'}\right\} \leq c_1 \sum_{t=0}^{T-1} \exp\left(-c_2 N_i n_x \left(\frac{\alpha \rho^{(i)} \|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right).$$

We prove Lemma 1 in Appendix A. By combining Lemma 1 with the condition on $N_i n_x$ from Assumption 2, our algorithm can ensure that the probability of misclassifying system i to cluster $\mathscr{C}_{j'}$ is at most δ, where δ can be arbitrarily small. Moreover, it is noteworthy that if we assume the data $X^{(i)}$, $Z^{(i)}$, and $W^{(i)}$ to be i.i.d. with $T = 1$ and $n_x = 1$, and the columns of $Z^{(i)}$ to have an identity covariance matrix, we can recover the probability of misclassification in the linear regression problem, as discussed in [23].

B. Convergence Analysis

We now examine the convergence of Algorithm 1. The theorem below is a single-iteration convergence analysis of our algorithm. Here we assume that, at a given iteration, an estimation $\widehat{\Theta}_j$ is obtained, which closely approximates the true model Θ_j, i.e., $\|\widehat{\Theta}_j - \Theta_j\| \leq \left(\frac{1}{2} - \alpha\right)\Delta_{\min}$, $\forall j \in [K]$ and $0 < \alpha < \frac{1}{2}$. We demonstrate that $\widehat{\Theta}_j$ converges to Θ_j up to a small bias.

Theorem 2: For any fixed $0 < \delta < 1$, with $N_i \geq \max\left\{8(n_x + n_u) + 16\log\frac{2MT}{\delta}, (4n_x + 2n_u)\log\frac{MT}{\delta}\right\}$, $\forall i \in [M]$, and selected step-size $\eta_j = \frac{|\mathscr{C}_j|}{\lambda_{\min}\left(\sum_{i\in\mathscr{C}_j} N_i \sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)}$, with probability at least $1 - 3\delta$, it holds that,

$$\|\widehat{\Theta}_j^+ - \Theta_j\| \leq \frac{1}{2}\|\widehat{\Theta}_j - \Theta_j\| + \bar{c}_0 \times \frac{1}{\sqrt{\sum_{i\in\widehat{\mathscr{C}}_j} N_i}} + \bar{c}_1 \Delta_{\max} \sum_{i\in[M]}\sum_{t=0}^{T-1} \exp\left(-\bar{c}_2 N_i n_x \left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right), \tag{4}$$

for all $j \in [K]$, where $\bar{c}_0$, $\bar{c}_1$, $\bar{c}_2 > 0$ are problem dependent constants.

The proof of Theorem 2 is detailed in Appendix B. This theorem provides an upper bound for the estimation error per iteration of our algorithm. Specifically, this bound consists of three terms. The first term is a contraction term that decreases to zero as the number of iterations increases. The second term is a constant error that decreases as the total number of observed trajectories by the systems within the cluster increases. The final term is the misclassification rate, which decays exponentially with the number of observed trajectories.

Note that although our setting is different from [23], which leads to a different estimation error expression, our per-iteration estimation error is also composed of a contractive term added to a constant error that can be controlled by the amount of data (i.e., the number of observed trajectories). We proceed to show the convergence of our algorithm by demonstrating that $\alpha^{(r)}$ is non-decreasing throughout iterations and using Assumptions 1 and 2 to show that $\|\widehat{\Theta}_j^{(r+1)} - \Theta_j\| \leq \|\widehat{\Theta}_j^{(r)} - \Theta_j\|$ for all $r \in [R]$.

Therefore, equipped with the aforementioned result, the following corollary characterizes the convergence of Algorithm 1 by providing the number of iterations required to attain a certain small and near optimal error ε, i.e., $\|\widehat{\Theta}_j^{(R)} - \Theta_j\| \leq \varepsilon$, for all clusters $j \in [K]$.

Corollary 1: Frame the hypotheses of Theorem 2 and Assumptions 1 and 2. Select the step-size as $\eta_j = \frac{|\mathscr{C}_j|}{\lambda_{\min}\left(\sum_{i\in\mathscr{C}_j} N_i \sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)}$ for all $j \in [K]$. Then, after $R \geq 2 + \log(\frac{\Delta_{\min}}{4\varepsilon})$ parallel iterations, we have $\|\widehat{\Theta}_j^{(R)} - \Theta_j\| \leq \varepsilon$, with

$$\varepsilon = \tilde{c}_0 \times \frac{1}{\sqrt{\sum_{i\in\mathscr{C}_j} N_i}} + \tilde{c}_1 \Delta_{\max} \sum_{i\in[M]}\sum_{t=0}^{T-1} \exp\left(-\tilde{c}_2 N_i n_x \left(\frac{\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right), \tag{5}$$

for all $j \in [K]$, where $\tilde{c}_0$, $\tilde{c}_1$, $\tilde{c}_2 > 0$ are problem dependent constants.

The proof of this corollary can be found in [29] (extended version of this paper). Our proof builds upon similar arguments as in [23], which considers the linear regression setting. To establish the non-decreasing property of $\alpha^{(r)}$ for all $r \in [R]$ and a decrease in the additive error term over the iterations, we rely on Assumptions 1 and 2. Furthermore, we demonstrate that our algorithm achieves a sufficiently large value of $\alpha^{(r)} \geq \frac{1}{4}$ after only a small number of iterations $R \geq 2$. This indicates that after a suitable number of iterations, our Algorithm 1 produces an estimation error that scales down with the number of systems within the cluster, and is independent of the initial closeness parameter $\alpha^{(0)}$.

This corollary highlights the benefits of collaboration. It demonstrates that the estimation error scales inversely with the number of agents within a cluster, implying that as the number of systems in the cluster increases, this error decreases. This leads to a smaller error when compared to

the single agent setting, where each system estimates its dynamics using *only* its own observations.

Importantly, the presented error bound differs from that of [14]. Here the misclassification rate exponentially decays with the number of observed trajectories, whereas the heterogeneity bias ε_{het} in [14] cannot be controlled by the number of trajectories. This indicates that under heterogeneous settings where the systems are significantly different, our clustering-based approach outperforms [14] by providing better control over the sources of error. However, it is worth mentioning that when the systems are similar and personalization is not required, the approaches introduced in [12], [13], [14] may be more favorable as their error bounds scale down with the total number of systems and do not necessitate a clustering step.

IV. Numerical Results

The following simulations[1] illustrate the efficiency of Algorithm 1. Our analysis considers $M = 50$ systems, each described by an LTI model as in (1) where $K = 3$ clusters and the number of systems in each cluster is $|\mathscr{C}_1| = 10$, $|\mathscr{C}_2| = 24$, and $|\mathscr{C}_3| = 16$. The systems matrices for each cluster are described as follows:

$$A_1 \qquad\qquad A_2 \qquad\qquad A_3$$

$$\begin{bmatrix} 0.5 & 0.3 & 0.1 \\ 0.0 & 0.2 & 0.0 \\ 0.1 & 0.0 & 0.3 \end{bmatrix}, \begin{bmatrix} -0.3 & 0.0 & 0.0 \\ 0.1 & 0.4 & 0.0 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}, \begin{bmatrix} -0.1 & 0.1 & 0.1 \\ 0.1 & 0.15 & 0.1 \\ 0.1 & 0.0 & 0.2 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 1 & 0.5 \\ 0.1 & 1 \\ 0.75 & 1.5 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 & 0.5 \\ 0.1 & 1 \\ 0.75 & 1.5 \end{bmatrix}, \quad B_3 = \begin{bmatrix} 0.8 & 0.1 \\ 0.1 & 1.5 \\ 0.4 & 0.8 \end{bmatrix},$$

where the initial state, input, and process noise standard deviations, for each cluster, are set to $\sigma_{x,i} = \sigma_{u,i} = \sigma_{w,i} = 0.11$, $\forall i \in \mathscr{C}_1$, $\sigma_{x,i} = \sigma_{u,i} = \sigma_{w,i} = 0.12$, $\forall i \in \mathscr{C}_2$, and $\sigma_{x,i} = \sigma_{u,i} = \sigma_{w,i} = 0.05$, $\forall i \in \mathscr{C}_3$. We consider the same number of trajectories $N_i = 100$ for all $i \in [M]$. Moreover, the trajectory length is set to $T = 50$. We use a fixed step-size $\eta_j = 10^{-3}$, $\forall j \in [3]$. For each iteration r, the estimation error $e_r^{(j)}$ is defined as the spectral norm distance between the estimated model $\widehat{\Theta}_j^{(r)}$ and the ground truth model Θ_j, i.e., $e_r^{(j)} = \|\widehat{\Theta}_j^{(r)} - \Theta_j\|$, for all clusters $j \in [K]$.

Figure 1 depicts the estimation error $e_r^{(j)}$ as a function of the number of iterations r for all the three considered clusters. The top plots compare the performance of Algorithm 1 with and without the clustering procedure (i.e., line 5 of Algorithm 1). These plots reveals that the estimation error decreases significantly when systems with the same model are clustered and cooperate to estimate their dynamics. Conversely, in the absence of clustering, the significant heterogeneity level across the systems leads to a poor common estimation, resulting in a large estimation error and unpersonalized solutions. This confirms our theoretical results, showing that the misclassification rate in (5) outperforms the heterogeneity constant of [12], [13], [14], when dealing with heterogeneous settings.

[1]Code: `https://github.com/jd-anderson/cluster-sysID`

The plots on the bottom of Figure 1 demonstrates the benefits of collaboration among systems to learn their dynamics. This shows that the estimation error is considerably reduced when multiple systems within the same cluster (i.e., $|\mathscr{C}_1| = 10$, $|\mathscr{C}_2| = 24$, and $|\mathscr{C}_3| = 16$) leverage the data from each other to identify their dynamics, compared to the case where a single system estimate its dynamics by using its own observations. This also confirms our theoretical results, where the statistical error in (5) scales down with the number of systems in the cluster, thus highlighting the benefit of collaboration in improving estimation accuracy in a multi-system setting. Our extended version of this paper [29] also includes the plot for the number of misclassification as a function of iteration count, showing the effect of the number of observed trajectories on the misclassification rate.

V. Conclusions and Future Work

We presented an approach to address the system identification problem through the use of clustering. Our method involves partitioning different systems that observe multiple trajectories into disjoint clusters based on the similarity of their dynamics. This approach enjoys an improved convergence rate that scales inversely with the number of systems in the cluster, along with an additive misclassification rate that has been shown to be negligible under mild assumptions. Our approach enables systems within the same cluster to learn their dynamics more efficiently. Future work will involve extending the proposed formulation to online system identification and proposing an adaptive clustering approach that eliminates the necessity for the warm initialization and well-separated clusters assumptions.

References

[1] K. J. Åström and P. Eykhoff, "System identification—a survey," *Automatica*, vol. 7, no. 2, pp. 123–162, 1971.

[2] L. Ljung, "System identification," in *Signal analysis and prediction*. Springer, 1998, pp. 163–173.

[3] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5610–5618.

[4] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.

[5] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conference On Learning Theory*. PMLR, 2018, pp. 439–473.

[6] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," in *2019 American control conference (ACC)*. IEEE, 2019, pp. 5655–5661.

[7] Y. Sun, S. Oymak, and M. Fazel, "Finite sample system identification: Optimal rates and the role of regularization," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 16–25.

[8] S. Tu, R. Boczar, A. Packard, and B. Recht, "Non-asymptotic analysis of robust control from coarse-grained identification," *arXiv preprint arXiv:1707.04791*, 2017.

[9] M. Simchowitz, R. Boczar, and B. Recht, "Learning linear dynamical systems with semi-parametric least squares," in *Conference on Learning Theory*. PMLR, 2019, pp. 2714–2802.

[10] Y. Zheng and N. Li, "Non-asymptotic identification of linear dynamical systems using multiple trajectories," *IEEE Control Systems Letters*, vol. 5, no. 5, pp. 1693–1698, 2020.

[11] K. Zhou, J. Doyle, and K. Glover, *Robust and optimal control*. Prentice hall, 1996.

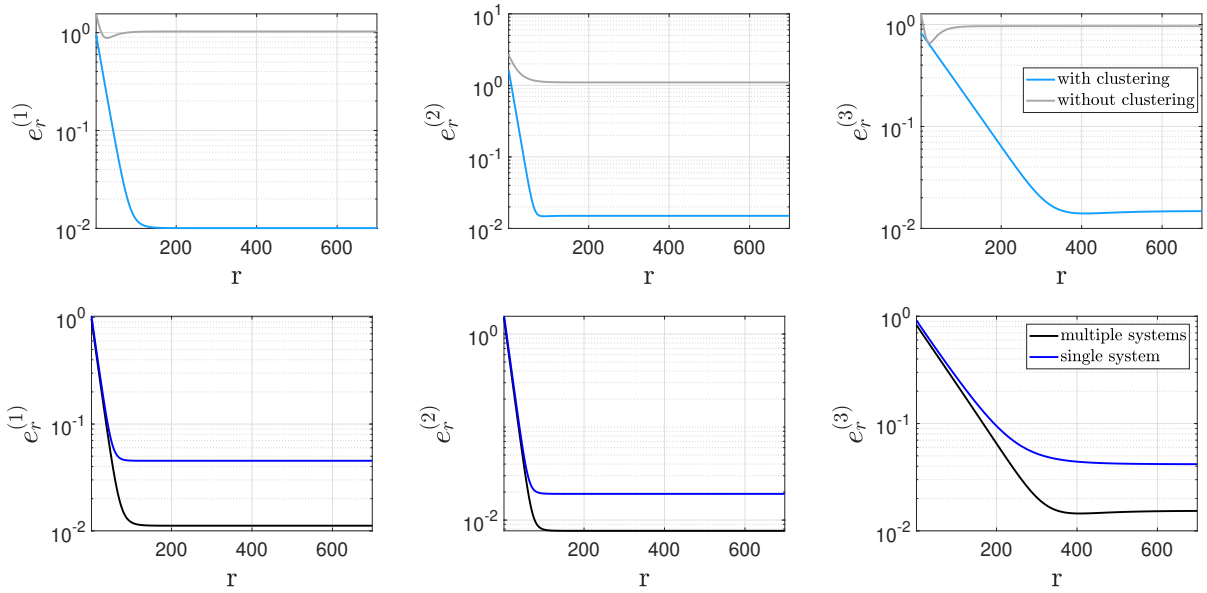

Fig. 1. Estimation error as a function of iteration count. The plot on the top considers Algorithm 1 with and without clustering, whereas the bottom consider the single and multiple agents settings.

[12] L. Xin, L. Ye, G. Chiu, and S. Sundaram, "Identifying the Dynamics of a System by Leveraging Data from Similar Systems," *arXiv preprint arXiv:2204.05446*, 2022.

[13] ——, "Learning Dynamical Systems by Leveraging Data from Similar Systems," *arXiv preprint arXiv:2302.04344*, 2023.

[14] H. Wang, L. F. Toso, and J. Anderson, "Fedsysid: A federated approach to sample-efficient system identification," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 1308–1320.

[15] Y. Chen, A. M. Ospina, F. Pasqualetti, and E. Dall'Anese, "Multi-Task System Identification of Similar Linear Time-Invariant Dynamical Systems," *arXiv preprint arXiv:2301.01430*, 2023.

[16] T. T. Zhang, K. Kang, B. D. Lee, C. Tomlin, S. Levine, S. Tu, and N. Matni, "Multi-Task Imitation Learning for Linear Dynamical Systems," *arXiv preprint arXiv:2212.00186*, 2022.

[17] T. T. Zhang, L. F. Toso, J. Anderson, and N. Matni, "Meta-Learning Operators to Optimality from Multi-Task Non-IID Data," *arXiv preprint arXiv:2308.04428*, 2023.

[18] H. Wang, L. F. Toso, A. Mitra, and J. Anderson, "Model-free Learning with Heterogeneous Dynamical Systems: A Federated LQR Approach," *arXiv preprint arXiv:2308.11743*, 2023.

[19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[20] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[21] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2089–2099.

[22] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[23] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 586–19 597, 2020.

[24] A. Ghosh, A. Mazumdar *et al.*, "An Improved Algorithm for Clustered Federated Learning," *arXiv preprint arXiv:2210.11538*, 2022.

[25] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 8, pp. 3710–3722, 2020.

[26] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *Ann. Statist.*, vol. 45, pp. 77–120, 2017.

[27] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[28] A. Kumar and R. Kannan, "Clustering with spectral norm and the k-means algorithm," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 299–308.

[29] L. F. Toso, H. Wang, and J. Anderson, "Learning Personalized Models with Clustered System Identification," *arXiv preprint arXiv:2304.01395*, 2023.

[30] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[31] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[32] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019, vol. 48.

APPENDIX

A. Proof of Lemma 1

Without loss of generality, we analyze only the first cluster $\mathscr{M}_i^{1,j}$ for some $j \neq 1$. By definition, we have

$$\mathscr{M}_i^{1,j} = \left\{ \|X^{(i)} - \widehat{\Theta}_j Z^{(i)}\|_F^2 \leq \|X^{(i)} - \widehat{\Theta}_1 Z^{(i)}\|_F^2 \right\}$$

where the batch matrices $X^{(i)}, Z^{(i)}$ and $W^{(i)}$ are related according to $X^{(i)} = \Theta_1 Z^{(i)} + W^{(i)}$. Note that $z_{l,t}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathscr{N}\left(0, \Sigma_t^{(i)}\right)$ and $w_{l,t}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathscr{N}\left(0, \sigma_{w,i}^2 I_{n_x}\right)$ are independent across trajectories (i.e., the columns of $Z^{(i)}$ and $W^{(i)}$ are independent). Then, the probability $\mathbb{P}\left\{\mathscr{M}_i^{1,j}\right\}$ can be expressed by

$$= \mathbb{P}\left\{ \left\| (\Theta_1 - \widehat{\Theta}_1) Z^{(i)} + W^{(i)} \right\|_F^2 \geq \left\| (\Theta_1 - \widehat{\Theta}_j) Z^{(i)} + W^{(i)} \right\|_F^2 \right\}$$

$$= \mathbb{P}\left\{ \sum_{t=0}^{T-1} \sum_{l=1}^{N_i} m_{l,t}^{(i),\top} m_{l,t}^{(i)} \geq \sum_{t=0}^{T-1} \sum_{l=1}^{N_i} n_{l,t}^{(i),\top} n_{l,t}^{(i)} \right\},$$

where $m_{l,t}^{(i)} = (\Theta_1 - \widehat{\Theta}_1)z_{l,t}^{(i)} + w_{l,t}^{(i)} \sim \mathscr{N}\left(0, \bar{\Sigma}_t^{(i)}\right)$, $n_{l,t}^{(i)} = (\Theta_1 - \widehat{\Theta}_j)z_{l,t}^{(i)} + w_{l,t}^{(i)} \sim \mathscr{N}\left(0, \tilde{\Sigma}_t^{(i)}\right)$, with

$$\bar{\Sigma}_t^{(i)} = (\Theta_1 - \widehat{\Theta}_1)\Sigma_t^{(i)}(\Theta_1 - \widehat{\Theta}_1)^\top + \sigma_{w,i}^2 I_{n_x},$$
$$\tilde{\Sigma}_t^{(i)} = (\Theta_1 - \widehat{\Theta}_j)\Sigma_t^{(i)}(\Theta_1 - \widehat{\Theta}_j)^\top + \sigma_{w,i}^2 I_{n_x}.$$

Therefore, we obtain

$$\mathbb{P}\left\{\mathscr{M}_i^{1,j}\right\} = \mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top}\bar{\Sigma}_t^{(i)}v_{l,t}^{(i)} \geq \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} u_{l,t}^{(i),\top}\tilde{\Sigma}_t^{(i)}u_{l,t}^{(i)}\right\},$$

with $m_{l,t}^{(i)} = (\bar{\Sigma}_t^{(i)})^{\frac{1}{2}}v_{l,t}^{(i)}$ and $n_{l,t}^{(i)} = (\bar{\Sigma}_t^{(i)})^{\frac{1}{2}}u_{l,t}^{(i)}$ for some standard normal random vectors $v_{l,t}^{(i)}$, $u_{l,t}^{(i)} \sim \mathscr{N}(0, I_{n_x})$. Then, the above expression can be rewritten as follows

$$\begin{aligned}\mathbb{P}\left\{\mathscr{M}_i^{1,j}\right\} &= \mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top}\bar{\Sigma}_t^{(i)}v_{l,t}^{(i)} \geq \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} \|\tilde{\Sigma}_t^{(i)}\|u_{l,t}^{(i),\top}u_{l,t}^{(i)}\right\} \\ &= \mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top}\bar{\Sigma}_t^{(i)}v_{l,t}^{(i)} \geq \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} c_t^{(i)}u_{l,t}^{(i),\top}u_{l,t}^{(i)}\right\}\end{aligned}$$

with $c_t^{(i)} = \|\Theta_1 - \widehat{\Theta}_j\|^2\|\Sigma_t^{(i)}\| + \sigma_{w,i}^2\sqrt{n_x}$, which implies

$$\begin{aligned}\mathbb{P}\left\{\mathscr{M}_i^{1,j}\right\} &= \mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top}\bar{\Sigma}_t^{(i)}v_{l,t}^{(i)} \geq \sum_{t=0}^{T-1}\sum_{l=1}^{N_i} c_t^{(i)}u_{l,t}^{(i),\top}u_{l,t}^{(i)}\right\} \\ &\leq \mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{l=1}^{N_i} c_t^{(i)}u_{l,t}^{(i),\top}u_{l,t}^{(i)} \leq \bar{t}\right\} + \mathbb{P}\left\{\sum_{t=0}^{T-1}\sum_{l=1}^{N_i} v_{l,t}^{(i),\top}\bar{\Sigma}_t^{(i)}v_{l,t}^{(i)} > \bar{t}\right\},\end{aligned}$$

for any $\bar{t} \geq 0$. Therefore , by using $v_{l,t}^{(i),\top}\bar{\Sigma}_t^{(i)}v_{l,t}^{(i)} \leq d_t^{(i)}v_{l,t}^{(i),\top}v_{l,t}^{(i)}$ with $d_t^{(i)} = \|\Theta_1 - \widehat{\Theta}_1\|^2\|\Sigma_t^{(i)}\| + \sigma_{w,i}^2\sqrt{n_x}$ we obtain

$$\mathbb{P}\left\{\mathscr{M}_i^{1,j}\right\} \leq \mathbb{P}\left\{\sum_{t=0}^{T-1} c_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} + \mathbb{P}\left\{\sum_{t=0}^{T-1} d_t^{(i)}V_t^{(i)} > \bar{t}\right\},$$

where $V_t^{(i)}$ are standard Chi-squared distributions with $N_i n_x$ degrees of freedom, for all $t \in \{0,1,\ldots,T-1\}$. Moreover, by using Definition 1 and Assumption 1,

$$\mathbb{P}\left\{\mathscr{M}_i^{1,j}\right\} \leq \mathbb{P}\left\{\sum_{t=0}^{T-1} f_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} + \mathbb{P}\left\{\sum_{t=0}^{T-1} g_t^{(i)}V_t^{(i)} > \bar{t}\right\},$$

with $f_t^{(i)} = (\frac{1}{2}+\alpha)^2\Delta_{\min}^2\|\Sigma_t^{(i)}\| + \sigma_{w,i}^2\sqrt{n_x}$ and $g_t^{(i)} = (\frac{1}{2}-\alpha)^2\Delta_{\min}^2\|\Sigma_t^{(i)}\| + \sigma_{w,i}^2\sqrt{n_x}$, since $c_t^{(i)} = \|\Theta_1 - \widehat{\Theta}_j\|^2\|\Sigma_t^{(i)}\| + \sigma_{w,i}^2\sqrt{n_x} \geq (\frac{1}{2}+\alpha)^2\Delta_{\min}^2\|\Sigma_t^{(i)}\| + \sigma_{w,i}^2\sqrt{n_x}$, with $\|\Theta_j - \widehat{\Theta}_1\| \geq \|\Theta_j - \Theta_1\| - \|\widehat{\Theta}_j - \Theta_j\| = (\frac{1}{2}+\alpha)\Delta_{\min}$ and $d_t^{(i)} = \|\Theta_1 - \widehat{\Theta}_1\|^2\|\Sigma_t^{(i)}\| + \sigma_{w,i}^2\sqrt{n_x} \leq (\frac{1}{2}-\alpha)^2\Delta_{\min}^2\|\Sigma_t^{(i)}\| + \sigma_{w,i}^2\sqrt{n_x}$, where $\|\Theta_1 - \widehat{\Theta}_1\| \leq (\frac{1}{2}-\alpha)\Delta_{\min}$ according to Assumption 1. Therefore, to characterize the above tail bounds, we can exploit well-established concentration inequalities as detailed in [30], [31]. To this end, we use union bound to write

$$\mathbb{P}\left\{\mathscr{M}_i^{1,j}\right\} \leq \sum_{t=0}^{T-1}\mathbb{P}\left\{f_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} + \mathbb{P}\left\{g_t^{(i)}V_t^{(i)} > \bar{t}\right\},$$

where $\mathbb{P}\left\{f_t^{(i)}V_t^{(i)} \leq \bar{t}\right\}$ can be rewritten as follows

$$\mathbb{P}\left\{f_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} = \mathbb{P}\left\{V_t^{(i)} \leq \frac{4\bar{t}}{\sigma_{w,i}^2\sqrt{n_x}\left((1+2\alpha)^2\rho^{(i)}\frac{\|\Sigma_t^{(i)}\|}{\sqrt{n_x}} + 4\right)}\right\},$$

thus, by choosing $\bar{t} = N_i n_x\left((\frac{1}{4}+\alpha^2)\Delta_{\min}^2\|\Sigma_t^{(i)}\| + \sigma_{w,i}^2\sqrt{n_x}\right)$ we obtain

$$\mathbb{P}\left\{f_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} = \mathbb{P}\left\{\frac{V_t^{(i)}}{N_i n_x} - 1 \leq \frac{-4\alpha\|\Sigma_t^{(i)}\|}{(1+2\alpha)^2\rho^{(i)}\|\Sigma_t^{(i)}\| + 4\sqrt{n_x}}\right\},$$

as per the concentration of standard Chi-squared distributions in [32], it is established that there exist universal constants c_1 and c_2, such that

$$\mathbb{P}\left\{f_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} \leq c_1\exp\left(-c_2 N_i n_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right). \tag{6}$$

Similarly, $\mathbb{P}\left\{g_t^{(i)}V_t^{(i)} > \bar{t}\right\}$ can be rewritten as follows

$$\mathbb{P}\left\{g_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} = \mathbb{P}\left\{\frac{V_t^{(i)}}{N_i n_x} - 1 \leq \frac{4\alpha\|\Sigma_t^{(i)}\|}{(1-2\alpha)^2\rho^{(i)}\|\Sigma_t^{(i)}\| + 4\sqrt{n_x}}\right\},$$

and by the concentration of Chi-squared distribution

$$\mathbb{P}\left\{g_t^{(i)}V_t^{(i)} \leq \bar{t}\right\} \leq c_3\exp\left(-c_4 N_i n_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right), \tag{7}$$

where the proof is completed by combining (6) and (7) to obtain

$$\mathbb{P}\left\{\mathscr{M}_i^{1,j}\right\} \leq c_1\sum_{t=0}^{T-1}\exp\left(-c_2 N_i n_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right).$$

B. Proof of Theorem 2

Without loss of generality, we analyze only the first cluster. Recall that the model is updated as follows:

$$\widehat{\Theta}_1^+ = \frac{1}{|\widehat{\mathscr{C}_1}|}\sum_{i\in\widehat{\mathscr{C}_1}}\tilde{\Theta}_i = \frac{1}{|\widehat{\mathscr{C}_1}|}\sum_{i\in\widehat{\mathscr{C}_1}\cap\mathscr{S}_1}\tilde{\Theta}_i + \frac{1}{|\widehat{\mathscr{C}_1}|}\sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{S}_1}}\tilde{\Theta}_i \tag{8}$$

with $\tilde{\Theta}_i = \widehat{\Theta}_1 + 2\eta_1(X^{(i)} - \widehat{\Theta}_1 Z^{(i)})Z^{(i),\top}$. Here $\widehat{\mathscr{C}_1}\cap\mathscr{C}_1$ corresponds to the set of systems correctly classified to the first cluster and $\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}$ represents the set of systems that are misclassified to the first cluster, with $\overline{\mathscr{C}_1}$ denoting the complement of $\mathscr{C}_1$. The above expression can be rewritten as follows

$$\begin{aligned}\widehat{\Theta}_1^+ = \widehat{\Theta}_1 &+ \frac{2\eta_1}{|\widehat{\mathscr{C}_1}|}\sum_{i\in\widehat{\mathscr{C}_1}\cap\mathscr{C}_1}(X^{(i)} - \widehat{\Theta}_1 Z^{(i)})Z^{(i),\top} \\ &+ \frac{2\eta_1}{|\widehat{\mathscr{C}_1}|}\sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}}(X^{(i)} - \widehat{\Theta}_1 Z^{(i)})Z^{(i),\top},\end{aligned}$$

where $X^{(i)} = \Theta_1 Z^{(i)} + W^{(i)}$ for $i \in \widehat{\mathscr{C}_1}\cap\mathscr{C}_1$, and $X^{(i)} = \Theta_j Z^{(i)} + W^{(i)}$ for $i \in \widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}$, with $j \neq 1 \in [K]$. Therefore,

by manipulating the above expression, we have

$$\widehat{\Theta}_1^+ - \Theta_1 = (\widehat{\Theta}_1 - \Theta_1)\left(I - \frac{2\eta_1}{|\widehat{\mathscr{C}_1}|}\sum_{i\in\widehat{\mathscr{C}_1}} Z^{(i)}Z^{(i),\top}\right) + \frac{2\eta_1}{|\widehat{\mathscr{C}_1}|}\sum_{i\in\widehat{\mathscr{C}_1}} W^{(i)}Z^{(i),\top} + (\Theta_j - \Theta_1)\frac{2\eta_1}{|\widehat{\mathscr{C}_1}|}|\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}|\sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}} Z^{(i)}Z^{(i),\top},$$

then we obtain $\|\widehat{\Theta}_1^+ - \Theta_1\| \le \|\mathscr{H}_1\| + \|\mathscr{H}_2\|$ with,

$$\|\mathscr{H}_1\| = \|\widehat{\Theta}_1 - \Theta_1\|\left\|I - \frac{2\eta_1}{|\widehat{\mathscr{C}_1}|}ZZ^\top\right\| + \frac{2\eta_1}{|\widehat{\mathscr{C}_1}|}\sum_{i\in\widehat{\mathscr{C}_1}}\|WZ^\top\|,$$

$$\|\mathscr{H}_2\| = \|\Theta_j - \Theta_1\|\frac{2\eta_1}{|\widehat{\mathscr{C}_1}|}|\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}|\|\bar{Z}\bar{Z}^\top\|.$$

We now concatenate the batch matrices $Z^{(i)}, W^{(i)}$ of the systems classified to the first cluster in $Z \in \mathbb{R}^{(n_x+n_u)\times N_i T|\widehat{\mathscr{C}_1}|}$ and $W \in \mathbb{R}^{n_x \times N_i T|\widehat{\mathscr{C}_1}|}$, and similarly the batch matrices $Z^{(i)}$ of the systems incorrectly classified to the first cluster are concatenated in $\bar{Z} \in \mathbb{R}^{(n_x+n_u)\times N_i T|\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}|}$. We proceed with our analysis by controlling both terms separately. To upper bound the first term, we introduce the following auxiliary results.

Proposition 1: [14, Proposition 8] For any fixed $0 < \delta < 1$, let $N_i \ge (4n_x + 2n_u)\log\frac{T|\widehat{\mathscr{C}_1}|}{\delta}$. It holds, with probability at least $1-\delta$, that

$$\left\|WZ^\top\right\| \le 4\sigma_{w,i}\sqrt{N_i(2n_x+n_u)\log\frac{9|\widehat{\mathscr{C}_1}|T}{\delta}}\sum_{t=0}^{T-1}\left\|(\Sigma_t^{(i)})^{\frac{1}{2}}\right\|. \tag{9}$$

Proposition 2: (Adapted from [14, Proposition 6]) For any fixed $0 < \delta < 1$, let $N_i \ge 8(n_x+n_u) + 16\log\frac{2|\widehat{\mathscr{C}_1}|T}{\delta}$. It holds, with probability at least $1-\delta$, that

$$ZZ^\top \succeq \frac{1}{4}\sum_{i\in\widehat{\mathscr{C}_1}} N_i \sum_{t=0}^{T-1}\Sigma_t^{(i)}, \tag{10}$$

$$\|\bar{Z}\bar{Z}^\top\| \le \frac{9}{4}\sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}}\sum_{t=0}^{T-1} N_i\left\|\Sigma_t^{(i)}\right\|. \tag{11}$$

Proof: Expression (10) follows direct from Proposition 6 in [14]. For expression (11), we can first write

$$\|\bar{Z}\bar{Z}^\top\| = \left\|\sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}}\sum_{l=1}^{N_i}\sum_{t=0}^{T-1} z_{l,t}^{(i)}z_{l,t}^{(i),\top}\right\| \le \sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}}\left\|\sum_{l=1}^{N_i}\sum_{t=0}^{T-1} z_{l,t}^{(i)}z_{l,t}^{(i),\top}\right\|$$

where $\chi_{l,t}^{(i)} = (\Sigma_t^{(i)})^{-\frac{1}{2}} z_{l,t}^{(i)}$ for any fixed l,t, and i, where $\chi_{l,t}^i \overset{\text{i.i.d.}}{\sim} \mathscr{N}(0, I_{n_x+n_u})$, for all $l \in \{1,2,\ldots,N_i\}$, we obtain

$$\|\bar{Z}\bar{Z}^\top\| \le \sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}}\sum_{t=0}^{T-1}\|\Sigma_t^{(i)}\|\left\|\sum_{l=1}^{N_i}\chi_{l,t}^{(i)}\chi_{l,t}^{(i),\top}\right\|,$$

thus, by using Proposition 6 of [14], with probability $1-\frac{\delta}{T}$, we have $\left\|\sum_{l=1}^{N_i}\chi_{l,t}^{(i)}\chi_{l,t}^{(i),\top}\right\| \le \frac{9N_i}{4}$, which implies

$$\|\bar{Z}\bar{Z}^\top\| \le \frac{9}{4}\sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}}\sum_{t=0}^{T-1} N_i\|\Sigma_t^{(i)}\|.$$

■

Therefore, with probability $1-2\delta$, we have

$$\|\mathscr{H}_1\| \le \|\widehat{\Theta}_1 - \Theta_1\|\left(1 - \frac{\eta_1}{2|\widehat{\mathscr{C}_1}|}\lambda_{\min}\left(\sum_{i\in\widehat{\mathscr{C}_1}} N_i\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)\right) + \frac{8\eta_1}{|\widehat{\mathscr{C}_1}|}\sum_{i\in\widehat{\mathscr{C}_1}}\sigma_{w,i}\sqrt{N_i(2n_x+n_u)\log\frac{9|\widehat{\mathscr{C}_1}|T}{\delta}}\sum_{t=0}^{T-1}\left\|(\Sigma_t^{(i)})^{\frac{1}{2}}\right\|,$$

and by selecting $\eta_1 = \frac{|\widehat{\mathscr{C}_1}|}{\lambda_{\min}\left(\sum_{i\in\widehat{\mathscr{C}_1}} N_i\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)}$, we obtain

$$\begin{aligned}\|\mathscr{H}_1\| &\le \frac{1}{2}\|\widehat{\Theta}_1 - \Theta_1\| + \frac{8\sum_{i\in\widehat{\mathscr{C}_1}}\sigma_{w,i}\sqrt{N_i(2n_x+n_u)\log\frac{9|\widehat{\mathscr{C}_1}|T}{\delta}}\sum_{t=0}^{T-1}\left\|(\Sigma_t^{(i)})^{\frac{1}{2}}\right\|}{\lambda_{\min}\left(\sum_{i\in\widehat{\mathscr{C}_1}} N_i\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)} \\ &\le \frac{1}{2}\|\widehat{\Theta}_1 - \Theta_1\| + \bar{c}_0 \times \frac{1}{\sqrt{\sum_{i\in\widehat{\mathscr{C}_1}} N_i}},\end{aligned} \tag{12}$$

with $\bar{c}_0 := \frac{8\sqrt{(2n_x+n_u)\log\frac{9|\widehat{\mathscr{C}_1}|T}{\delta}}\sqrt{\sum_{i\in\widehat{\mathscr{C}_1}}\sigma_{w,i}^2\left(\sum_{t=0}^{T-1}\left\|(\Sigma_t^{(i)})^{\frac{1}{2}}\right\|\right)^2}}{\min_{i\in\widehat{\mathscr{C}_1}}\lambda_{\min}\left(\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)}$ and $N_i \ge \max\{8(n_x+n_u)+16\log\frac{2|\widehat{\mathscr{C}_1}|T}{\delta},\ (4n_x+2n_u)\log\frac{|\widehat{\mathscr{C}_1}|T}{\delta}\}$, $\forall i \in \widehat{\mathscr{C}_1}$. To control $\|\mathscr{H}_2\|$, we use the Definition 1 to write

$$\|\mathscr{H}_2\| \le \Delta_{\max}|\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}|\frac{9\sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}} N_i\sum_{t=0}^{T-1}\|\Sigma_t^{(i)}\|}{2\lambda_{\min}\left(\sum_{i\in\widehat{\mathscr{C}_1}} N_i\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)},$$

which implies $\|\mathscr{H}_2\| \le c_5\Delta_{\max}|\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}|$ by using Jensen and Cauchy-Schwartz inequalities in the denominator and numerator, respectively. Here, we denote $c_5 := \frac{9\sum_{i\in\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}}\sum_{t=0}^{T-1}\|\Sigma_t^{(i)}\|}{2\min_{i\in\widehat{\mathscr{C}_1}}\left(\sum_{t=0}^{T-1}\Sigma_t^{(i)}\right)}$.

To control $|\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}|$ we can use Lemma 1 to write

$$\mathbb{E}\left[|\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}|\right] \le c_6\sum_{i\in[M]}\sum_{t=0}^{T-1}\exp\left(-c_7 N_i n_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right)$$

which yields

$$\mathbb{P}\left\{|\widehat{\mathscr{C}_1}\cap\overline{\mathscr{C}_1}| \le c_6\sum_{i\in[M]}\sum_{t=0}^{T-1}\exp\left(-\frac{c_7}{2}N_i n_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right)\right\} \ge 1 - \sum_{i\in[M]}\sum_{t=0}^{T-1}\exp\left(-\frac{c_7}{2}N_i n_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right) \ge 1-\delta,$$

by using Markov's inequality and Assumption 2 with $N_i n_x \ge c\left(\frac{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}\right)^2\log(\frac{MT}{\delta})$, for some large enough constant c such that $\frac{1}{c} < c_7$. Thus, we obtain

$$\|\mathscr{H}_2\| \le \bar{c}_1\Delta_{\max}\sum_{i\in[M]}\sum_{t=0}^{T-1}\exp\left(-\bar{c}_2 N_i n_x\left(\frac{\alpha\rho^{(i)}\|\Sigma_t^{(i)}\|}{\rho^{(i)}\|\Sigma_t^{(i)}\| + \sqrt{n_x}}\right)^2\right), \tag{13}$$

with probability at least $1-\delta$. The proof is completed by combining (12) and (13).