Learning Mixtures of Linear Dynamical Systems

Yanxi Chen [1] **H. Vincent Poor** [1]

Abstract

We study the problem of learning a mixture of multiple linear dynamical systems (LDSs) from unlabeled short sample trajectories, each generated by one of the LDS models. Despite the wide applicability of mixture models for time-series data, learning algorithms that come with end-to-end performance guarantees are largely absent from existing literature. There are multiple sources of technical challenges, including but not limited to (1) the presence of latent variables (i.e. the unknown labels of trajectories); (2) the possibility that the sample trajectories might have lengths much smaller than the dimension d of the LDS models; and (3) the complicated temporal dependence inherent to time-series data. To tackle these challenges, we develop a two-stage meta-algorithm, which is guaranteed to efficiently recover each ground-truth LDS model up to error $\widetilde{O}(\sqrt{d/T})$, where T is the total sample size. We validate our theoretical studies with numerical experiments, confirming the efficacy of the proposed algorithm.

1. Introduction

Imagine that we are asked to learn multiple linear dynamical systems (LDSs) from a mixture of unlabeled sample trajectories — namely, each sample trajectory is generated by one of the LDSs of interest, but we have no idea which system it is. To set the stage and facilitate discussion, recall that in a classical LDS, one might observe a sample trajectory $\{\boldsymbol{x}_t\}_{0\le t\le T}$ generated by an LDS obeying $\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}_t$, where $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ determines the system dynamics in the noiseless case, and $\{\boldsymbol{w}_t\}_{t\ge 0}$ denote independent zero-mean noise vectors with covariance $\mathsf{cov}(\boldsymbol{w}_t) = \boldsymbol{W} \succ \boldsymbol{0}$. The mixed LDSs setting considered herein extends classical LDSs by allowing for mixed measurements as described below; see Figure 1 for a visualization of the scenario.

- *Multiple linear systems*. Suppose that there are K different LDSs as represented by $\{(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})\}_{1\le k\le K}$, where $\boldsymbol{A}^{(k)} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{W}^{(k)} \in \mathbb{R}^{d\times d}$ represent the state transition matrix and noise covariance matrix of the k-th LDS, respectively. Here and throughout, we shall refer to $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})$ as the system matrix for the k-th LDS. We only assume that $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}) \neq (\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$ for any $k \neq \ell$, whereas $\boldsymbol{A}^{(k)} = \boldsymbol{A}^{(\ell)}$ or $\boldsymbol{W}^{(k)} = \boldsymbol{W}^{(\ell)}$ is allowed.

- *Mixed sample trajectories*. We collect a total number of M unlabeled trajectories from these LDSs. More specifically, the m-th sample trajectory is drawn from one of the LDSs in the following manner: set $(\boldsymbol{A}, \boldsymbol{W})$ to be $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})$ for some $1 \le k \le K$, and generate a trajectory of length T_m obeying

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}_t, \quad \text{where the } \boldsymbol{w}_t\text{'s are i.i.d.,}$$
$$\mathbb{E}[\boldsymbol{w}_t] = \boldsymbol{0},\ \mathsf{cov}(\boldsymbol{w}_t) = \boldsymbol{W} \succ \boldsymbol{0}.$$

  Note, however, that the label k associated with each sample trajectory is a latent variable not revealed to us, resulting in a mixture of unlabeled trajectories. The current paper focuses on the case where the length of each trajectory is somewhat short, making it infeasible to estimate the system matrix from a single trajectory.

- *Goal*. The aim is to jointly learn the system matrices $\{(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})\}_{1\le k\le K}$ from the mixture of sample trajectories. In particular, we seek to accomplish this task in a sample-efficient manner, where the total sample size is defined to be the aggregate trajectory length $\sum_{m=1}^{M} T_m$.

Motivations. The mixed LDSs setting described above is motivated by many real-world scenarios where a single time-series model is insufficient to capture the complex and heterogeneous patterns in temporal data. For instance, in psychology (Bulteel et al., 2016), researchers collect multiple time-series trajectories (e.g. depression-related symptoms over a period of time) from different patients. Fitting this data with multi-modal LDSs (instead of a single model) not only achieves better fitting performance, but also helps

[1]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA. Correspondence to: Yanxi Chen <yanxic@princeton.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

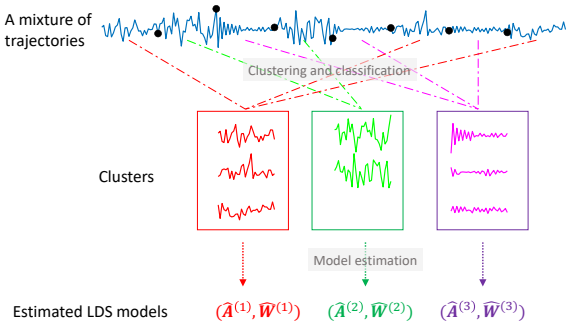

Figure 1. A high-level visualization of the mixed LDSs formulation, and the algorithmic idea of combining clustering, classification, and model estimation. Here, we consider the special case where the multiple short trajectories come from the segments of a single continuous trajectory. The black dots within the continuous trajectory represent the time steps when the latent variable (i.e. the unknown label) changes.

identify subgroups of the persons, which further inspires interpretations of the results and tailored treatments for patients from different subgroups. Another example concerns an automobile sensor dataset, which consists of a single continuous (but possibly time-varying) trajectory of measurements from the sensors (Hallac et al., 2017). Through segmentation of the trajectory, clustering of the short pieces, and learning within each cluster, one can discover, and obtain meaningful interpretations for, a small number of key driving modes (such as "driving straight", "slowing down", "turning"). See Section 4 for a longer list of applications.

Challenges. While there is no shortage of potential applications, a mixture of LDSs is far more challenging to learn compared to the classical setting with a single LDS. In particular, the presence of the *latent variables*, i.e. the unknown labels of the sample trajectories, significantly complicates matters. One straightforward idea is to learn a coarse model for each trajectory, followed by proper clustering of these coarse models (to be utilized for refined model estimation); however, this idea becomes infeasible in the high-dimensional setting unless all trajectories are sufficiently long. Another popular approach is to alternate between model estimation and clustering of trajectories, based on, say, the expectation-maximization (EM) algorithm; unfortunately, there is no theoretical support for such EM-type algorithms, and we cannot preclude the possibilities that the algorithms get stuck at undesired local optima. The lack of theoretical guarantees in prior literature motivates us to come up with algorithms that enjoy provable performance guarantees. Finally, the present paper is also inspired by the recent progress in *meta-learning for mixed linear regression* (Kong et al., 2020b;a), where the goal is to learn multiple linear models from a mixture of independent samples; note, however, that the temporal dependence underlying time-series data in our case poses substantial challenges and calls for the development of new algorithmic ideas.

Main contributions. In this work, we take an important step towards guaranteed learning of mixed LDSs, focusing on algorithm design that comes with end-to-end theoretical guarantees. In particular, we propose a two-stage meta-algorithm to tackle the challenge of mixed LDSs:

1. *Coarse estimation*: perform a coarse-level clustering of the unlabeled sample trajectories (assisted by dimension reduction), and compute initial model estimation for each cluster;

2. *Refinement*: classify additional trajectories (and add each of them to the corresponding cluster) based on the above coarse model estimates, followed by refined model estimation with the updated clusters.

This two-stage meta approach, as well as the specific methods for individual steps, will be elucidated in Section 2.

Encouragingly, assuming that the noise vectors $\{\boldsymbol{w}_{m,t}\}$ are independent Gaussian random vectors, the proposed two-stage algorithm is not only computationally efficient, but also guaranteed to succeed in the presence of a polynomial sample size. Informally, our algorithm achieves exact clustering/classification of the sample trajectories as well as faithful model estimation, under the following conditions (with a focus on the dependency on the dimension d): (1) each short trajectory length T_m is allowed to be much smaller than d; (2) the total trajectory length of each stage is linear in d (up to logarithmic factors); (3) to achieve a final model estimation error $\epsilon \to 0$ (in the spectral norm), it suffices to have a total trajectory length of order d/ϵ^2 for each LDS model. See Section 3 (in particular, Corollary 3.7) for the precise statements of our main results, which will also be validated numerically.

It is worth noting that, although we focus on mixed LDSs for concreteness, we will make clear that the proposed modular algorithm is fairly flexible and can be adapted to learning mixtures of other time-series models, as long as certain technical conditions are satisfied; see Remark 2.3 at the end of Section 2 for a detailed discussion.

Notation. Throughout this paper, vectors and matrices are represented by boldface letters. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we define $(\boldsymbol{x})_i \in \mathbb{R}$ as the i-th entry of $\boldsymbol{x}$, and $\|\boldsymbol{x}\|_2$ as its ℓ_2 norm; for a matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, we define $(\boldsymbol{X})_i \in \mathbb{R}^n$ as the transpose of the i-th row of $\boldsymbol{X}$, and $\|\boldsymbol{X}\|$ (resp. $\|\boldsymbol{X}\|_{\mathrm{F}}$) as its spectral (resp. Frobenius) norm. For a symmetric matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, we denote its maximal (resp. minimal) eigenvalue as $\lambda_{\mathsf{max}}(\boldsymbol{X})$ (resp. $\lambda_{\mathsf{min}}(\boldsymbol{X})$); if in addition $\boldsymbol{X}$ is positive definite, we denote its condition number as $\kappa(\boldsymbol{X}) :=$

Algorithm 1 A two-stage algorithm for mixed LDSs

Input: M short trajectories $\{\boldsymbol{X}_m\}_{1\le m\le M}$ (where $\boldsymbol{X}_m = \{\boldsymbol{x}_{m,t}\}_{0\le t\le T_m}$); parameters τ, G.
`// Stage 1: coarse estimation.`
Run Algorithm 2 with $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\mathsf{subspace}}}$ to obtain subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}_{1\le i\le d}$.
Run Algorithm 3 with $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\mathsf{clustering}}}$, $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}_{1\le i\le d}$, τ, G, to obtain clusters $\{\mathcal{C}_k\}_{1\le k\le K}$.
Run Algorithm 4 with $\{\mathcal{C}_k\}_{1\le k\le K}$ to obtain coarse models $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$.
`// Stage 2: refinement.`
Run Algorithm 5 with $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\mathsf{classification}}}$, $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$, $\{\mathcal{C}_k\}_{1\le k\le K}$, to update clusters $\{\mathcal{C}_k\}$.
Run Algorithm 4 with $\{\mathcal{C}_k\}_{1\le k\le K}$ to obtain refined models $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$.
Output: final model estimation $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$ and clusters $\{\mathcal{C}_k\}_{1\le k\le K}$.

$\lambda_{\mathsf{max}}(\boldsymbol{X})/\lambda_{\mathsf{min}}(\boldsymbol{X}) \ge 1$. If $\boldsymbol{A} = [A_{ij}]_{1\le i\le m, 1\le j\le n}$ and $\boldsymbol{B} = [B_{ij}]_{1\le i\le m, 1\le j\le n}$ are matrices of the same dimension, we denote by $\langle \boldsymbol{A}, \boldsymbol{B}\rangle \coloneqq \sum_{i=1}^m \sum_{j=1}^n A_{ij}B_{ij}$ their inner product. Given vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ where $n < d$, let $\mathsf{span}\{\boldsymbol{x}_i, 1 \le i \le N\} \in \mathbb{R}^{d\times n}$ represent the subspace spanned by these vectors. Let $\boldsymbol{I}_d$ be the $d\times d$ identity matrix.

We always use the superscript "(k)" to indicate "the k-th model", as in $\boldsymbol{A}^{(k)}$ and $\boldsymbol{W}^{(k)}$; this is to be distinguished from the superscript "k" without the parentheses, which simply means the power of k. For a discrete set Ω, we denote by $|\Omega|$ its cardinality. Define $\mathbb{1}(\mathcal{E})$ to be the indicator function, which takes value 1 if the event $\mathcal{E}$ happens, and 0 otherwise. Let $a_n \lesssim b_n$ indicate that $a_n \le C_0 b_n$ for all $n = 1, 2, \ldots$, where $C_0 > 0$ is some universal constant; moreover, $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$, and $a_n \asymp b_n$ indicates that $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold simultaneously.

2. Algorithms

We propose a two-stage paradigm for solving the mixed LDSs problem, as summarized in Algorithm 1. It consists of several subroutines as described in Algorithms 2–5; due to space limitation, in this section we will only illustrate the key ideas behind these subroutines, with full details deferred to Appendix A. Note that Algorithm 1 is stated in a modular manner, and one might replace certain subroutines by alternative schemes in order to handle different settings and model assumptions.

Let us first introduce some additional notation and assumptions that will be useful for our presentation, without much loss of generality. To begin with, we augment the notation for each sample trajectory with its trajectory index; that is, for each $1 \le m \le M$, the m-th trajectory — denoted by $\boldsymbol{X}_m \coloneqq \{\boldsymbol{x}_{m,t}\}_{0\le t\le T_m}$ — starts with some initial state $\boldsymbol{x}_{m,0} \in \mathbb{R}^d$, and evolves according to the k_m-th LDS for some *unknown* label $1 \le k_m \le K$ such that

$$\boldsymbol{x}_{m,t+1} = \boldsymbol{A}^{(k_m)}\boldsymbol{x}_{m,t} + \boldsymbol{w}_{m,t}, \text{ where the } \boldsymbol{w}_{m,t}\text{'s are i.i.d.,}$$
$$\mathbb{E}[\boldsymbol{w}_{m,t}] = \boldsymbol{0},\ \mathsf{cov}(\boldsymbol{w}_{m,t}) = \boldsymbol{W}^{(k_m)} \succ \boldsymbol{0} \tag{1}$$

for all $0 \le t \le T_m - 1$. Next, we divide the M sample trajectories $\{\boldsymbol{X}_m\}_{1\le m\le M}$ in hand into three disjoint subsets $\mathcal{M}_{\mathsf{subspace}}, \mathcal{M}_{\mathsf{clustering}}, \mathcal{M}_{\mathsf{classification}}$ satisfying

$$\mathcal{M}_{\mathsf{subspace}} \cup \mathcal{M}_{\mathsf{clustering}} \cup \mathcal{M}_{\mathsf{classification}} = \{1, 2, \ldots, M\},$$

where each subset of samples will be employed to perform one subroutine. We assume that all trajectories within each subset have the same length, namely,

$$T_m = \begin{cases} T_{\mathsf{subspace}} & \text{if} \quad m \in \mathcal{M}_{\mathsf{subspace}}, \\ T_{\mathsf{clustering}} & \text{if} \quad m \in \mathcal{M}_{\mathsf{clustering}}, \\ T_{\mathsf{classification}} & \text{if} \quad m \in \mathcal{M}_{\mathsf{classification}}. \end{cases} \tag{2}$$

Finally, we assume that $K \le d$, so that performing subspace estimation in Algorithm 1 will be helpful (otherwise one might simply eliminate this step). The interested readers are referred to Appendix E for discussions of some potential extensions of our algorithms (e.g. adapting to the case where the trajectories within a subset have different lengths, or the case where certain parameters are *a priori* unknown).

2.1. Preliminary facts

We first introduce some preliminary background on the autocovariance structures and mixing property of linear dynamical systems, which form the basis of our algorithms for subspace estimation and clustering of trajectories.

Stationary covariance matrices. Consider first a single LDS model $(\boldsymbol{A}, \boldsymbol{W})$, with $\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}_t$. If $\mathbb{E}[\boldsymbol{x}_t] = \boldsymbol{0}, \mathsf{cov}(\boldsymbol{x}_t) = \boldsymbol{\Gamma}$ and $\mathbb{E}[\boldsymbol{w}_t] = \boldsymbol{0}, \mathsf{cov}(\boldsymbol{w}_t) = \boldsymbol{W}$, then it follows that $\mathbb{E}[\boldsymbol{x}_{t+1}] = \boldsymbol{0}$, and $\mathsf{cov}(\boldsymbol{x}_{t+1}) = \boldsymbol{A}\cdot\mathsf{cov}(\boldsymbol{x}_t)\cdot\boldsymbol{A}^\top + \mathsf{cov}(\boldsymbol{w}_t) = \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^\top + \boldsymbol{W}$. Under certain assumption on stability, this leads to the *order-0 stationary autocovariance matrix* defined as (Kailath et al., 2000)

$$\begin{aligned}\boldsymbol{\Gamma}(\boldsymbol{A}, \boldsymbol{W}) &\coloneqq \mathbb{E}\big[\boldsymbol{x}_t{\boldsymbol{x}_t}^\top | \boldsymbol{A}, \boldsymbol{W}\big] \\ &= \boldsymbol{A}\cdot\boldsymbol{\Gamma}(\boldsymbol{A}, \boldsymbol{W})\cdot\boldsymbol{A}^\top + \boldsymbol{W} = \sum_{t=0}^{\infty} \boldsymbol{A}^t\boldsymbol{W}(\boldsymbol{A}^t)^\top, \end{aligned} \tag{3}$$

and the *order-1 stationary autocovariance matrix*

$$\boldsymbol{Y}(\boldsymbol{A}, \boldsymbol{W}) \coloneqq \mathbb{E}\big[\boldsymbol{x}_{t+1}{\boldsymbol{x}_t}^\top | \boldsymbol{A}, \boldsymbol{W}\big] = \boldsymbol{A}\cdot\boldsymbol{\Gamma}(\boldsymbol{A}, \boldsymbol{W}). \tag{4}$$

For the mixed LDSs case (1) with K models, we abbreviate

$$\boldsymbol{\Gamma}^{(k)} \coloneqq \boldsymbol{\Gamma}(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}),\ \boldsymbol{Y}^{(k)} \coloneqq \boldsymbol{Y}(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}). \quad (5)$$

Mixing. Expanding the LDS recursion $\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}_t$ with $\mathsf{cov}(\boldsymbol{w}_t) = \boldsymbol{W}$, we have

$$\boldsymbol{x}_{t+s} = \boldsymbol{A}^s \boldsymbol{x}_t + \sum_{i=1}^{s} \boldsymbol{A}^{i-1} \boldsymbol{w}_{t+s-i}, \quad s = 1, 2, \dots \quad (6)$$

If $\boldsymbol{A}$ satisfies certain assumptions regarding stability and if s is larger than certain mixing time of the LDS, then the first term on the right-hand side of (6) approaches zero, while the second term is independent of the history up to time t, with covariance close to the stationary autocovariance $\boldsymbol{\Gamma}(\boldsymbol{A}, \boldsymbol{W})$. This suggests that, for two samples within the same trajectory that are sufficiently far apart, we can treat them as being (almost) independent of each other; this simple fact shall inspire our algorithmic design and streamline statistical analysis later on. It is noteworthy that the proposed algorithms do not require prior knowledge about the mixing times of the LDS models.

2.2. Subspace estimation

Recall that the notation $(\boldsymbol{\Gamma}^{(k)})_i \in \mathbb{R}^d$ (resp. $(\boldsymbol{Y}^{(k)})_i$) represents the transpose of the i-th row of $\boldsymbol{\Gamma}^{(k)}$ (resp. $\boldsymbol{Y}^{(k)}$) defined in (5). Consider the following subspaces:

$$\boldsymbol{V}_i^\star \coloneqq \mathsf{span}\Big\{(\boldsymbol{\Gamma}^{(k)})_i, 1 \le k \le K\Big\},$$
$$\boldsymbol{U}_i^\star \coloneqq \mathsf{span}\Big\{(\boldsymbol{Y}^{(k)})_i, 1 \le k \le K\Big\}, \quad 1 \le i \le d,$$

each of which has rank at most K. We develop a spectral method to estimate these subspaces, which will in turn allow for proper dimension reduction in subsequent steps. Here, we only introduce the idea for estimating $\{\boldsymbol{U}_i^\star\}$, since the idea for $\{\boldsymbol{V}_i^\star\}$ is very similar.

We first divide $\{0, 1, \dots, T_{\mathsf{subspace}}\}$ into four segments of the same size, and denote the 2nd (resp. 4th) segment as Ω_1 (resp. Ω_2). According to the mixing property of LDS, if T_{subspace} is larger than some appropriately defined mixing time, then (1) each sample trajectory in $\mathcal{M}_{\mathsf{subspace}}$ will mix sufficiently and nearly reach stationarity (when constrained to $t \in \Omega_1 \cup \Omega_2$), (2) the samples in Ω_1 are nearly independent of those in Ω_2. Therefore, defining

$$\boldsymbol{g}_{m,i,j} \coloneqq \frac{1}{|\Omega_j|} \sum_{t \in \Omega_j} (\boldsymbol{x}_{m,t+1})_i\, \boldsymbol{x}_{m,t} \quad (7)$$

for $1 \le i \le d$ and $j \in \{1, 2\}$, it is easy to check that $\mathbb{E}[\boldsymbol{g}_{m,i,1}\boldsymbol{g}_{m,i,2}^\top] \approx \mathbb{E}[\boldsymbol{g}_{m,i,1}]\mathbb{E}[\boldsymbol{g}_{m,i,2}]^\top \approx (\boldsymbol{Y}^{(k_m)})_i (\boldsymbol{Y}^{(k_m)})_i^\top$. Taking the average over $m \in \mathcal{M}_{\mathsf{subspace}}$, we have constructed a matrix $\widehat{\boldsymbol{G}}_i$ such that $\mathbb{E}[\widehat{\boldsymbol{G}}_i] \approx \sum_{k=1}^K p^{(k)} (\boldsymbol{Y}^{(k)})_i (\boldsymbol{Y}^{(k)})_i^\top =: \boldsymbol{G}_i$, where $p^{(k)}$ denotes the fraction of sample trajectories generated by the k-th model. As a consequence, if T_{subspace} and $|\mathcal{M}_{\mathsf{subspace}}|$ are sufficiently large, then one might expect $\widehat{\boldsymbol{G}}_i$ to be a good approximation of $\boldsymbol{G}_i$, the latter of which is symmetric, has rank at most K, and has $\boldsymbol{U}_i^\star$ as its eigenspace. All this motivates us to compute the rank-K eigenspace of $\widehat{\boldsymbol{G}}_i + \widehat{\boldsymbol{G}}_i^\top$.

2.3. Clustering

In this step, we propose to construct a similarity matrix $\boldsymbol{S}$ for the sample trajectories in $\mathcal{M}_{\mathsf{clustering}}$, based on their pairwise comparisons; then we can apply any mainstream clustering algorithm (e.g. spectral clustering (Chen et al., 2021b)) to $\boldsymbol{S}$, dividing $\mathcal{M}_{\mathsf{clustering}}$ into K disjoint clusters $\{\mathcal{C}_k\}_{1 \le k \le K}$ such that the trajectories in each cluster are primarily generated by the same LDS model.

The remaining question is how to perform the pairwise comparisons; we intend to achieve this by comparing the autocovariance matrices associated with the sample trajectories. Note that, even though $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}) \neq (\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$ for $k \neq \ell$, it is indeed possible that $\boldsymbol{\Gamma}^{(k)} = \boldsymbol{\Gamma}^{(\ell)}$ or $\boldsymbol{Y}^{(k)} = \boldsymbol{Y}^{(\ell)}$. Fortunately, the following fact ensures the separation of $(\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)})$ versus $(\boldsymbol{\Gamma}^{(\ell)}, \boldsymbol{Y}^{(\ell)})$.

Fact 2.1. *If $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}) \neq (\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$, then we have either $\boldsymbol{\Gamma}^{(k)} \neq \boldsymbol{\Gamma}^{(\ell)}$ or $\boldsymbol{Y}^{(k)} \neq \boldsymbol{Y}^{(\ell)}$ (or both).*

Now, let us compare the m-th and n-th trajectories for some $m, n \in \mathcal{M}_{\mathsf{clustering}}$. In order to determine whether they have the same label (namely $k_m = k_n$), we propose to estimate the quantity $\|\boldsymbol{\Gamma}^{(k_m)} - \boldsymbol{\Gamma}^{(k_n)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k_m)} - \boldsymbol{Y}^{(k_n)}\|_{\mathrm{F}}^2$ using the data samples $\{\boldsymbol{x}_{m,t}\}$ and $\{\boldsymbol{x}_{n,t}\}$, which is expected to be small (resp. large) if $k_m = k_n$ (resp. $k_m \neq k_n$). To do so, let us divide $\{0, 1, \dots, T_{\mathsf{clustering}}\}$ evenly into four segments, and denote by Ω_1 (resp. Ω_2) the 2nd (resp. 4th) segment. Recall our earlier definition of the $\boldsymbol{g}$ vectors in (7). Assuming sufficient mixing and utilizing near independence between samples from Ω_1 and those from Ω_2, we might resort to the following statistic:

$$\mathsf{stat}_Y \coloneqq \sum_{i=1}^{d} \Big\langle \boldsymbol{U}_i{}^\top \big(\boldsymbol{g}_{m,i,1} - \boldsymbol{g}_{n,i,1}\big), \boldsymbol{U}_i{}^\top \big(\boldsymbol{g}_{m,i,2} - \boldsymbol{g}_{n,i,2}\big) \Big\rangle,$$

whose expectation is given by

$$\mathbb{E}[\mathsf{stat}_Y] \approx \sum_{i=1}^{d} \Big\| \boldsymbol{U}_i{}^\top \big((\boldsymbol{Y}^{(k_m)})_i - (\boldsymbol{Y}^{(k_n)})_i\big) \Big\|_2^2$$
$$\approx \sum_{i=1}^{d} \big\|(\boldsymbol{Y}^{(k_m)})_i - (\boldsymbol{Y}^{(k_n)})_i\big\|_2^2 = \big\|\boldsymbol{Y}^{(k_m)} - \boldsymbol{Y}^{(k_n)}\big\|_{\mathrm{F}}^2;$$

here, the second approximation holds if each subspace $\boldsymbol{U}_i$ is sufficiently close to $\boldsymbol{U}_i^\star = \mathsf{span}\{(\boldsymbol{Y}^{(j)})_i, 1 \le j \le K\}$.

The purpose of utilizing $\{\boldsymbol{U}_i\}$ is to reduce the variance of stat_Y. Similarly, we can compute another statistic stat_Γ with $\mathbb{E}[\mathsf{stat}_\Gamma] \approx \|\boldsymbol{\Gamma}^{(k_m)} - \boldsymbol{\Gamma}^{(k_n)}\|_{\mathrm{F}}^2$. Consequently, we shall declare $k_m \neq k_n$ if $\mathsf{stat}_\Gamma + \mathsf{stat}_Y$ exceeds some appropriate threshold.

Remark 2.2. For subspace estimation and clustering, we split each trajectory evenly into four parts, and only utilize the 2nd and 4th. In theory, this only affects sample complexities by a constant factor. The key benefit is that no prior knowledge of the mixing time t_{mix} is needed. In cases where t_{mix} is known/estimated, one might improve data efficiency by letting the 1st and 3rd parts have only length t_{mix}.

2.4. Model estimation

Suppose that we have obtained a good clustering accuracy in the previous step. Then, for each k, we use the samples $\{\{\boldsymbol{x}_{m,t}\}_{0\le t\le T_m}\}_{m\in\mathcal{C}_k}$ to obtain an estimate $\widehat{\boldsymbol{A}}^{(k)}$ of the state transition matrix by solving a least-squares problem (Simchowitz et al., 2018; Sarkar & Rakhlin, 2019). Finally, we can estimate the noise vector $\widehat{\boldsymbol{w}}_{m,t} \coloneqq \boldsymbol{x}_{m,t+1} - \widehat{\boldsymbol{A}}^{(k)}\boldsymbol{x}_{m,t} \approx \boldsymbol{w}_{m,t}$, and use the empirical covariance of $\{\widehat{\boldsymbol{w}}_{m,t}\}$ as our estimate of the noise covariance matrix.

2.5. Classification

In the previous steps, we have obtained initial clusters $\{\mathcal{C}_k\}_{1\le k\le K}$ and coarse model estimates $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$. With the assistance of additional sample trajectories in $\mathcal{M}_{\mathsf{classification}}$, we can augment the clusters in the following way: for each new trajectory $\{\boldsymbol{x}_t\}_{0\le t\le T}$, we infer its label as $\widehat{k} = \arg\min_k L(\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)})$, and then assign this trajectory to the $\widehat{k}$-th cluster; here, the loss function $L(\boldsymbol{A}, \boldsymbol{W}) \coloneqq T\cdot\log\det(\boldsymbol{W}) + \sum_{t=0}^{T-1}(\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t)^\top \boldsymbol{W}^{-1}(\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t)$ coincides with the negative log-likelihood under a Gaussian assumption. Once this is done, we can run Algorithm 4 again with the updated clusters $\{\mathcal{C}_k\}$ to refine our model estimates, which is exactly the last step of Algorithm 1.

Remark 2.3. While this work focuses on mixtures of LDSs, we emphasize that the general principles of Algorithm 1 are applicable under much weaker assumptions. For Algorithms 2 and 3 to work, we essentially only require that each sample trajectory satisfies a certain *mixing* property, and that the autocovariances of different models are *sufficiently separated* (and hence distinguishable). As for Algorithms 4 and 5, we essentially require a *well-specified parametric form* of the time-series models. This observation might inspire future extensions (in theory or applications) of Algorithm 1 to much broader settings.

3. Main results

3.1. Model assumptions

To streamline the theoretical analysis, we focus on the case where the trajectories are driven by *Gaussian noise*; that is, for each $1 \le m \le M, 0 \le t \le T_m$, the noise vector $\boldsymbol{w}_{m,t}$ in (1) is independently generated from the Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{W}^{(k_m)})$. Next, we assume for simplicity that the labels $\{k_m\}_{1\le m\le M}$ of the trajectories are pre-determined and fixed, although one might equivalently regard $\{k_m\}$ as being random and independent of the noise vectors $\{\boldsymbol{w}_{m,t}\}$. Moreover, while our algorithms and analysis are largely insensitive to the initial states $\{\boldsymbol{x}_{m,0}\}_{1\le m\le M}$, we focus on two canonical cases for concreteness: (i) the trajectories start at *zero* state, or (ii) they are segments of *one* continuous long trajectory. This is formalized as follows:

$$\text{Case 0:}\quad \boldsymbol{x}_{m,0} = \boldsymbol{0}, \quad 1 \le m \le M; \tag{8a}$$

$$\text{Case 1:}\quad \boldsymbol{x}_{1,0} = \boldsymbol{0}, \quad \text{and}$$
$$\boldsymbol{x}_{m+1,0} = \boldsymbol{x}_{m,T_m}, \quad 1 \le m \le M-1. \tag{8b}$$

We further define the total trajectory length as $T_{\mathsf{total}} \coloneqq \sum_{1\le m\le M} T_m = \sum_{\mathsf{o}} T_{\mathsf{total,o}}$, where

$$T_{\mathsf{total,o}} \coloneqq T_{\mathsf{o}} \cdot |\mathcal{M}_{\mathsf{o}}|,$$
$$\mathsf{o} \in \{\mathsf{subspace}, \mathsf{clustering}, \mathsf{classification}\}.$$

Additionally, we assume that each model occupies a non-degenerate fraction of the data; in other words, there exists some $0 < p_{\mathsf{min}} \le 1/K$ such that for all $1 \le k \le K$ and $\mathsf{o} \in \{\mathsf{subspace}, \mathsf{clustering}, \mathsf{classification}\}$,

$$p_{\mathsf{min}} \le p_{\mathsf{o}}^{(k)} \coloneqq \frac{1}{|\mathcal{M}_{\mathsf{o}}|} \sum_{m\in\mathcal{M}_{\mathsf{o}}} \mathbb{1}(k_m = k).$$

Finally, we make the following assumptions about the ground-truth LDS models, where we recall that the autocovariance matrices $\{\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)}\}$ have been defined in (5).

Assumption 3.1. *The LDS models $\{\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}\}_{1\le k\le K}$ satisfy the following conditions:*

1. *There exist $\kappa_A \ge 1$ and $0 \le \rho < 1$ such that for any $1 \le k \le K$, $\|(\boldsymbol{A}^{(k)})^t\| \le \kappa_A \cdot \rho^t, t = 1, 2, \ldots$;*
2. *There exist $\Gamma_{\mathsf{max}} \ge W_{\mathsf{max}} \ge W_{\mathsf{min}} > 0$ and $\kappa_{w,\mathsf{cross}} \ge \kappa_w \ge 1$ such that for any $1 \le k \le K$, (i) $\lambda_{\mathsf{max}}(\boldsymbol{\Gamma}^{(k)}) \le \Gamma_{\mathsf{max}}$, (ii) $W_{\mathsf{min}} \le \lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)}) \le \lambda_{\mathsf{max}}(\boldsymbol{W}^{(k)}) \le W_{\mathsf{max}}$, (iii) $W_{\mathsf{max}}/W_{\mathsf{min}} =: \kappa_{w,\mathsf{cross}}$, (iv) $\kappa(\boldsymbol{W}^{(k)}) = \lambda_{\mathsf{max}}(\boldsymbol{W}^{(k)})/\lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)}) \le \kappa_w$;*
3. *There exist $\Delta_{\Gamma,Y}, \Delta_{A,W} > 0$ such that for any $1 \le k < \ell \le K$,*

$$\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2 \ge \Delta_{\Gamma,Y}^2,$$
$$\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2 + \frac{\|\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} \ge \Delta_{A,W}^2.$$

The first assumption states that each state transition matrix $\boldsymbol{A}^{(k)}$ is *exponentially stable*, which is a quantified version of stability and has appeared in various forms in the literature of LDS (Kailath et al., 2000; Cohen et al., 2018); here, κ_A can be regarded as a condition number, while ρ is a contraction rate. The second assumption ensures that the noise covariance matrices $\{\boldsymbol{W}^{(k)}\}$ are well conditioned, and the autocovariance matrices $\{\boldsymbol{\Gamma}^{(k)}\}$ are bounded. The third assumption quantifies the separation between different LDS models. It is important that we consider the separation of $(\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)})$ versus $(\boldsymbol{\Gamma}^{(\ell)}, \boldsymbol{Y}^{(\ell)})$ jointly, which guarantees that $\Delta_{\Gamma,Y}$ is always strictly positive (thanks to Fact 2.1), despite the possibility of $\boldsymbol{\Gamma}^{(k)} = \boldsymbol{\Gamma}^{(\ell)}$ or $\boldsymbol{Y}^{(k)} = \boldsymbol{Y}^{(\ell)}$; the reasoning for our definition of $\Delta_{A,W}$ is similar.

Remark 3.2. Since the separation parameters $\Delta_{A,W}, \Delta_{\Gamma,Y}$ are defined with respect to the Frobenius norm, we may naturally regard them as $\Delta_{A,W} = \sqrt{d}\delta_{A,W}, \Delta_{\Gamma,Y} = \sqrt{d}\delta_{\Gamma,Y}$, where $\delta_{A,W}, \delta_{\Gamma,Y}$ are the *canonical separation parameters* (in terms of the spectral norm). For example, in the simple setting where $K = 2$, $\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(2)}$, $\boldsymbol{A}^{(1)} = 0.5\boldsymbol{I}_d$ and $\boldsymbol{A}^{(2)} = (0.5 + \delta)\boldsymbol{I}_d$ for some canonical separation parameter δ, we have $\Delta_{A,W} = \|\boldsymbol{A}^{(1)} - \boldsymbol{A}^{(2)}\|_{\mathrm{F}} = \|\delta \boldsymbol{I}_d\|_{\mathrm{F}} = \sqrt{d}\delta$. This observation will be crucial for obtaining the correct dependence on the dimension d in our subsequent analysis.

Remark 3.3. Most of the parameters in Assumption 3.1 come directly from the ground-truth LDS models $\{\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}\}$, except Γ_{max} and $\Delta_{\Gamma,Y}$. In fact, they can be bounded by $\Gamma_{\mathsf{max}} \lesssim W_{\mathsf{max}}$ and $\Delta_{\Gamma,Y} \gtrsim \Delta_{A,W}$; see Fact D.4 for the formal statements. However, the pre-factors in these bounds can be pessimistic in general cases, therefore we choose to preserve Γ_{max} and $\Delta_{\Gamma,Y}$ in our analysis.

3.2. Theoretical guarantees

We are now ready to present our end-to-end performance guarantees for Algorithm 1. Our main results for Cases 0 and 1 (defined in (8)) are summarized in Theorems 3.4 and 3.5 below, with proofs deferred to Appendix B.

Theorem 3.4 (Case 0)**.** *There exist positive constants $\{C_i\}_{1\le i\le 8}$ such that the following holds for any fixed $0 < \delta < 1/2$. Consider the model (1) under the assumptions in Sections 2 and 3.1, with a focus on Case 0 (8a). Suppose that we run Algorithm 1 with parameters τ, G that satisfy $1/8 < \tau/\Delta_{\Gamma,Y}^2 < 3/8$, $G \ge C_1\iota_1$, and data $\{\boldsymbol{X}_m\}_{1\le m\le M}$ (where $\boldsymbol{X}_m = \{\boldsymbol{x}_{m,t}\}_{0\le t\le T_m}$) that satisfies*

$$T_{\mathsf{subspace}} \ge C_2 \frac{\iota_1}{1-\rho},$$

$$T_{\mathsf{total,subspace}} \ge C_3 \frac{d}{1-\rho}\left(\left(\frac{\Gamma_{\mathsf{max}}\sqrt{d}}{\Delta_{\Gamma,Y}}\right)^4 \frac{K^2}{{p_{\mathsf{min}}}^2} + 1\right)\cdot \iota_1^4, \tag{9a}$$

$$T_{\mathsf{clustering}} \ge C_4 \frac{G}{1-\rho}\left(\frac{\Gamma_{\mathsf{max}}^2 \kappa_A^2 \sqrt{dK}}{\Delta_{\Gamma,Y}^2} + 1\right)\iota_2,$$

$$T_{\mathsf{total,clustering}} \ge C_5 \frac{d\kappa_{w,\mathsf{cross}}^2}{p_{\mathsf{min}}}\left(\frac{d}{\Delta_{A,W}^2}\frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}} + 1\right)\iota_3^2, \tag{9b}$$

$$T_{\mathsf{classification}} \ge C_6\left(\kappa_w^2 + \frac{\kappa_{w,\mathsf{cross}}^6}{\Delta_{A,W}^2}\right)\iota_1^2, \tag{9c}$$

where we define the logarithmic terms $\iota_1 := \log(\frac{d\kappa_A T_{\mathsf{total}}}{\delta})$, $\iota_2 := \log\left((\frac{\Gamma_{\mathsf{max}}}{\Delta_{\Gamma,Y}} + 2)\frac{d\kappa_A T_{\mathsf{total}}}{\delta}\right)$, $\iota_3 := \log(\frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}}\frac{d\kappa_A T_{\mathsf{total}}}{\delta})$. Then, with probability at least $1-\delta$, Algorithm 1 achieves exact clustering in Line 4 and exact classification in Line 7; moreover, there exists some permutation $\pi : \{1,\dots,K\} \to \{1,\dots,K\}$ such that the final model estimation $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$ in Line 8 obeys

$$1 \le k \le K, \quad \|\widehat{\boldsymbol{A}}^{(k)} - \boldsymbol{A}^{(\pi(k))}\| \le C_7\sqrt{\frac{d\kappa_w\iota_3}{p_{\mathsf{min}}T}},$$
$$\frac{\|\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(\pi(k))}\|}{\|\boldsymbol{W}^{(\pi(k))}\|} \le C_8\sqrt{\frac{d\iota_3}{p_{\mathsf{min}}T}}, \tag{10}$$

where $T := T_{\mathsf{total,clustering}} + T_{\mathsf{total,classification}}$.

Theorem 3.5 (Case 1)**.** *Consider the same setting of Theorem 3.4, except that we focus on Case 1 (cf. (8b)) instead. Then the same performance guarantees continue to hold, if we replace the conditions on $T_{\mathsf{total,clustering}}$ and $T_{\mathsf{classification}}$ in (9) with*

$$T_{\mathsf{total,clustering}} \ge C_5 \frac{d\kappa_{w,\mathsf{cross}}^2}{p_{\mathsf{min}}}\left(\frac{d}{\Delta_{A,W}^2}\frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}}\kappa_A^2 + 1\right)\iota_3^2, \tag{11a}$$

$$T_{\mathsf{classification}} \ge C_6\left(\kappa_w^2 + \frac{\kappa_{w,\mathsf{cross}}^6}{\Delta_{A,W}^2}\right)\iota_1^2 + \frac{1}{1-\rho}\log(2\kappa_A). \tag{11b}$$

Remark 3.6. It is worth noting that all the short trajectory lengths $\{T_\circ\}$ in the above theorems have sublinear dependence on the dimension d. In particular, the $\sqrt{dK}$ dependence in $T_{\mathsf{clustering}}$ (9) is due to variance reduction achieved by subspace projection (see Section 2.3), without which this dependence would have been linear in d.

While Theorems 3.4 and 3.5 guarantee that Algorithm 1 successfully learns a mixture of LDS models with a polynomial number of samples, these results involve many parameters and may be somewhat difficult to interpret. In the following corollary, we make some simplifications and focus on the most important parameters.

Corollary 3.7. *Consider the same setting of Theorems 3.4 and 3.5. For simplicity, suppose that the condition numbers $\kappa_A, \kappa_w, \kappa_{w,\mathsf{cross}} \asymp 1$, and the fractions of data generated by different LDS models are balanced, namely $p_{\mathsf{min}} \asymp 1/K$. Moreover, define the canonical separation parameters $\delta_{A,W} := \Delta_{A,W}/\sqrt{d}$ and $\delta_{\Gamma,Y} := \Delta_{\Gamma,Y}/\sqrt{d}$,*

as suggested by Remark 3.2. Finally, define the mixing time $t_{\mathsf{mix}} := 1/(1-\rho)$. *Then we can rewrite the sample complexities in Theorems 3.4 and 3.5 as follows (where we hide the logarithmic terms* $\{\iota_i\}_{1\le i\le 3}$*): if*

$$
\begin{aligned}
T_{\mathsf{subspace}} &\gtrsim t_{\mathsf{mix}}, \\
T_{\mathsf{total,subspace}} &\gtrsim t_{\mathsf{mix}} d \left(\left(\frac{\Gamma_{\mathsf{max}} K}{\delta_{\Gamma,Y}} \right)^4 + 1 \right), \\
T_{\mathsf{clustering}} &\gtrsim t_{\mathsf{mix}} \left(\left(\frac{\Gamma_{\mathsf{max}}}{\delta_{\Gamma,Y}} \right)^2 \sqrt{\frac{K}{d}} + 1 \right), \\
T_{\mathsf{total,clustering}} &\gtrsim K d \left(\frac{1}{\delta_{A,W}^2} \frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}} + 1 \right), \\
T_{\mathsf{classification}} &\gtrsim \begin{cases} \frac{1}{d\delta_{A,W}^2} + 1 & \textit{for Case 0}, \\ \frac{1}{d\delta_{A,W}^2} + t_{\mathsf{mix}} & \textit{for Case 1}, \end{cases} \\
T_{\mathsf{total,clustering}} + T_{\mathsf{total,classification}} &\gtrsim \frac{Kd}{\epsilon^2},
\end{aligned}
$$

then with high probability, Algorithm 1 achieves exact clustering in Line 4, exact classification in Line 7, and final model estimation errors

$$
\|\widehat{\boldsymbol{A}}^{(k)} - \boldsymbol{A}^{(\pi(k))}\| \le \epsilon, \quad \frac{\|\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(\pi(k))}\|}{\|\boldsymbol{W}^{(\pi(k))}\|} \le \epsilon.
$$

Below are some key implications of Corollary 3.7.

- *Dimension* d *and targeted error* ϵ: (1) Our algorithms allow the T_{o}'s to be much smaller than (and even decrease with) d. (2) Each $T_{\mathsf{total,o}}$ shall grow linearly with d, and it takes an order of Kd/ϵ^2 samples in total to learn K models in up to $\epsilon \to 0$ errors (in the spectral norm), which is just K times the usual parametric rate (d/ϵ^2) of estimating a single model.

- *Mixing time* t_{mix}: (1) T_{subspace} and $T_{\mathsf{clustering}}$ are linear in t_{mix}, which ensures sufficient mixing and thus facilitates our algorithms for Stage 1. In contrast, $T_{\mathsf{classification}}$ depends on t_{mix} only for Case 1, with the sole and different purpose of ensuring that the states $\{\boldsymbol{x}_{m,t}\}$ are bounded throughout (see Example D.5 for an explanation). (2) While $T_{\mathsf{total,subspace}}$ needs to grow linearly with t_{mix}, this is not required for $T_{\mathsf{total,clustering}}$ and $T_{\mathsf{total,classification}}$, because our methods for model estimation (Algorithm 4) do not rely on mixing (Simchowitz et al., 2018; Sarkar & Rakhlin, 2019).

- *Canonical separation parameters* $\delta_{A,W}, \delta_{\Gamma,Y}$: (1) $T_{\mathsf{clustering}} \gtrsim 1/\delta_{\Gamma,Y}^2$ guarantees exact clustering of the trajectories, while $T_{\mathsf{classification}} \gtrsim 1/\delta_{A,W}^2$ guarantees exact classification. (2) $T_{\mathsf{total,subspace}} \gtrsim 1/\delta_{\Gamma,Y}^4$ leads to sufficiently accurate subspaces, while $T_{\mathsf{total,clustering}} \gtrsim 1/\delta_{A,W}^2$ leads to accurate initial model estimation.[1]

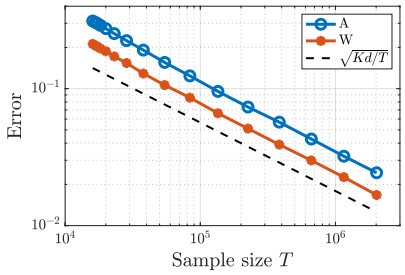

Figure 2. The model estimation errors of Algorithm 1 versus the total sample size (excluding $\mathcal{M}_{\mathsf{subspace}}$). Each curve is an average over 12 independent trials.

Remark 3.8. It is worth noting that $T_{\mathsf{clustering}} \gg T_{\mathsf{classification}}$ in Corollary 3.7, i.e. clustering requires a larger trajectory length than classification does. This justifies the benefit of the proposed two-stage procedure, compared with simply doing a large clustering of all trajectories and then one-shot model estimation.

3.3. Numerical experiments

We now validate our theoretical findings with a series of numerical experiments, confirming that Algorithm 1 successfully solves the mixed LDSs problem. In these experiments, we fix $d = 80, K = 4$; moreover, let $T_{\mathsf{subspace}} = 20$, $T_{\mathsf{clustering}} = 20$ and $T_{\mathsf{classification}} = 5$, all of which are much smaller than d. We take $|\mathcal{M}_{\mathsf{subspace}}| = 30\,d$, $|\mathcal{M}_{\mathsf{clustering}}| = 10\,d$, and vary $|\mathcal{M}_{\mathsf{classification}}|$ between $[0, 5000\,d]$. Our experiments focus on Case 1 as defined in (8b), and we generate the labels of the sample trajectories uniformly at random. The ground-truth LDS models are generated in the following manner: $\boldsymbol{A}^{(k)} = \rho \boldsymbol{R}^{(k)}$, where $\rho = 0.5$ and $\boldsymbol{R}^{(k)} \in \mathbb{R}^{d\times d}$ is a random orthogonal matrix; $\boldsymbol{W}^{(k)}$ has eigendecomposition $\boldsymbol{U}^{(k)}\boldsymbol{\Lambda}^{(k)}(\boldsymbol{U}^{(k)})^\top$, where $\boldsymbol{U}^{(k)}$ is a random orthogonal matrix, and the diagonal entries of $\boldsymbol{\Lambda}^{(k)}$ are independently drawn from the uniform distribution on $[1, 2]$.

Our experimental results are illustrated in Figure 2. Here, the horizontal axis represents the sample size $T = T_{\mathsf{total,clustering}} + T_{\mathsf{total,classification}}$ for model estimation, and the vertical axis represents the estimation errors, measured by $\max_k \|\widehat{\boldsymbol{A}}^{(k)} - \boldsymbol{A}^{(\pi(k))}\|$ (plotted in blue) and $\max_k \|\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(\pi(k))}\|/\|\boldsymbol{W}^{(\pi(k))}\|$ (plotted in orange). The results confirm our main theoretical prediction: Algorithm 1 recovers the LDS models based on a mixture of short trajectories with length $T_{\mathsf{o}} \ll d$, and achieves an error rate of $1/\sqrt{T}$. In addition, we observe in our experiments

[1] It is possible to improve the $1/\delta_{\Gamma,Y}^4$ factor in $T_{\mathsf{total,subspace}}$ to $1/\delta_{\Gamma,Y}^2$, if one is willing to pay for some extra factors of eigengaps; see Appendix B for a detailed discussion.

that the final outputs of Algorithm 1 are robust to a small number of mis-clustered or mis-classified trajectories during the intermediate steps; this is clearly an appealing property to have, especially when the T_{o}'s and $|\mathcal{M}_{\mathrm{o}}|$'s in reality are slightly smaller than what our theory requires. Additional experimental results can be found in Appendix F.

4. Related works

Meta-learning for mixed linear regression. Our work is closely related to the recent papers (Kong et al., 2020b;a) that bridge mixed linear regression (Quandt & Ramsey, 1978; Yi et al., 2014; Chen et al., 2017; Li & Liang, 2018; Chen et al., 2020; Kwon et al., 2021b; Diakonikolas & Kane, 2020; Yin et al., 2018; Pal & Mazumdar, 2020; Chen et al., 2021c; Diamandis et al., 2021) and meta-learning (Finn et al., 2017; Harrison et al., 2020; Snell et al., 2017; Du et al., 2021; Pan & Yang, 2009; Tripuraneni et al., 2020; Baxter, 2000; Maurer et al., 2016). In this setting of *meta-learning for mixed linear regression* (Kong et al., 2020b;a), one has access to multiple tasks, each containing a few independent and identically distributed samples generated by an unknown model; the inductive bias (i.e. the common structure underlying these tasks) for meta-learning is that there is only a small discrete set of ground-truth linear regression models, akin to the case of mixed linear regression. While the high-level idea of our Algorithm 1 is largely inspired by the work of (Kong et al., 2020b;a), our *detailed implementations* are substantially different, due to the lack of ideal i.i.d. assumption (among others) in our case of time-series data; in addition, we improve upon some of the analyses in (Kong et al., 2020b;a).[2]

Mixtures of time-series models and trajectories. Mixture models for time series have achieved empirical success in the study of psychology (Bulteel et al., 2016; Takano et al., 2020), neuroscience (Albert, 1991; Mezer et al., 2009), biology (Wong & Li, 2000), air pollution (D'Urso et al., 2015), economics (McCulloch & Tsay, 1994; Maharaj, 2000; Kalliovirta et al., 2016), automobile sensors (Hallac et al., 2017), and many other domains. Some specific aspects of mixture models include hypothesis testing for a pair of trajectories (Maharaj, 2000), or clustering of multiple trajectories (Liao, 2005; Aghabozorgi et al., 2015; Pathak et al., 2021; Huang et al., 2021). In addition to mixture models, other related yet different models include time-varying systems (Qu et al., 2021; Minasyan et al., 2021), systems with random parameters (Du et al., 2020), switching systems (Sun, 2006; Sarkar et al., 2019; Ansari et al., 2021), switching state-space models (Ghahramani & Hinton, 2000; Linderman et al., 2017), Markovian jump systems (Shi & Li, 2015; Zhao et al., 2019), and event-triggered systems (Sedghi et al., 2020; Schluter et al., 2020), to name just a few. There are even more related models in reinforcement learning (RL), such as latent bandit (Maillard & Mannor, 2014; Hong et al., 2020), multi-task learning (Wilson et al., 2007; Brunskill & Li, 2013; Liu et al., 2016; Sodhani et al., 2021) / meta-learning (Finn et al., 2017) / transfer-learning (Taylor & Stone, 2009; Tirinzoni et al., 2020) for RL, latent Markov decision processes (Kwon et al., 2021a; Brunskill et al., 2009; Steimle et al., 2021), and so on. What distinguishes our work from this extensive literature is that, we design algorithms and prove non-asymptotic sample complexities for model estimation, in the specific setting of mixture models that features (1) a finite set of underlying time-series models, and (2) unknown labels of the trajectories, with no probabilistic assumptions imposed on these latent variables.

Linear dynamical systems. LDS (also called vector autoregressive models) is one of the most fundamental models in system identification and optimal control (Ljung, 1998; Khalil et al., 1996). Recently, there has been a surge of studies about non-asymptotic theoretical analyses of various learning procedures for the basic LDS model (Faradonbeh et al., 2018; Simchowitz et al., 2018; Sarkar & Rakhlin, 2019), linear-quadratic regulators (Dean et al., 2020; Cohen et al., 2018; Jedra & Proutiere, 2019; Faradonbeh et al., 2020; Mania et al., 2019; Simchowitz & Foster, 2020; Fazel et al., 2018; Malik et al., 2019), and LDSs with partial observations (Oymak & Ozay, 2019; Simchowitz et al., 2019; Sarkar et al., 2021; Sun et al., 2020; Tsiamis et al., 2020; Lale et al., 2020; Zheng et al., 2021). In particular, it was only until recently that the authors of (Simchowitz et al., 2018; Sarkar & Rakhlin, 2019) proved sharp error bounds of least squares for estimating the state transition matrix of a LDS model, using a single trajectory; our analysis of Algorithm 4 is largely inspired by their techniques.

5. Discussion

This paper has developed a theoretical and algorithmic framework for learning multiple LDS models from a mixture of short, unlabeled sample trajectories. Our key contributions include a modular two-stage meta-algorithm, as well as theoretical analysis demonstrating its computational and statistical efficiency. We would like to invite the readers to contribute to this important topic by, say, further strengthening the theoretical analysis and algorithmic design. For exam-

[2] There is a concurrent preprint (Modi et al., 2021) that extends multi-task learning (Du et al., 2021; Tripuraneni et al., 2020) to time-series data, and the setting therein includes mixed LDSs as a special case. However, the authors of (Modi et al., 2021) assume oracle access to the global optimum of a non-convex optimization problem, without providing a practical algorithm that can provably find it; moreover, with the short trajectory length fixed, the estimation error bounds in that work will remain bounded away from zero, even if the number of trajectories grows to infinity. In comparison, we consider a simpler problem setting, and propose computationally efficient algorithms with better error bounds.

ple, in certain cases $T_{\mathsf{clustering}}$ can be a bottleneck compared with T_{subspace} and $T_{\mathsf{classification}}$, and thus one might hope to achieve a better dependence on $T_{\mathsf{total,clustering}}$ (e.g. allowing $T_{\mathsf{total,clustering}} \ll d$), by replacing Line 5 in Algorithm 1 with a different method (e.g. adapting Algorithm 3 of (Kong et al., 2020b) to our setting). As another example, in some practical scenarios, the data is a single continuous trajectory, and the time steps when the underlying model changes are *unknown* (Hallac et al., 2017; Harrison et al., 2020); in order to accommodate such a case, one might need to incorporate change-point detection into the learning process.

Moving beyond the current setting of mixed LDSs, we remark that there are plenty of opportunities for future studies. For instance, while our methods in Stage 1 rely on the mixing property of the LDS models, it is worth exploring whether it is feasible to handle the non-mixing case (Simchowitz et al., 2018; Sarkar & Rakhlin, 2019). Another potential direction is to consider the robustness against outliers and adversarial noise (Chen et al., 2021a; Kong et al., 2020a). One might even go further and extend the ideas to learning mixtures of other time-series models (see Remark 2.3), such as LDS with partial or nonlinear observations (Mhammedi et al., 2020), or nonlinear dynamical systems (Mania et al., 2020; Kakade et al., 2020; Foster et al., 2020). Ultimately, it would be of great importance to consider the case with controlled inputs, such as learning mixtures of linear-quadratic regulators, or latent Markov decision processes (Kwon et al., 2021a) that arises in reinforcement learning.

Acknowledgements

Y. Chen is supported in part by the ARO grant W911NF-20-1-0097, the NSF grants CCF-1907661 and IIS-1900140, and the AFOSR grant FA9550-19-1-0030. H. V. Poor is supported in part by the NSF under Grant CCF-1908308. We would like to thank Yuxin Chen and Gen Li for numerous helpful discussions.

References

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. Time-series clustering – a decade review. *Information Systems*, 53:16–38, 2015.

Albert, P. S. A two-state markov mixture model for a time series of epileptic seizure counts. *Biometrics*, pp. 1371–1381, 1991.

Ansari, A. F., Benidis, K., Kurle, R., Turkmen, A. C., Soh, H., Smola, A. J., Wang, B., and Januschowski, T. Deep explicit duration switching models for time series. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Baxter, J. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Brunskill, E. and Li, L. Sample complexity of multi-task reinforcement learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 122–131, Arlington, Virginia, USA, 2013. AUAI Press.

Brunskill, E., Leffler, B. R., Li, L., Littman, M. L., and Roy, N. Provably efficient learning with typed parametric models. *Journal of Machine Learning Research*, 10(68): 1955–1988, 2009.

Bulteel, K., Tuerlinckx, F., Brose, A., and Ceulemans, E. Clustering vector autoregressive models: Capturing qualitative differences in within-person dynamics. *Frontiers in Psychology*, 7:1540, 2016.

Chen, S., Li, J., and Song, Z. Learning mixtures of linear regressions in subexponential time via fourier moments. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 587–600, 2020.

Chen, S., Koehler, F., Moitra, A., and Yau, M. Kalman filtering with adversarial corruptions. *arXiv preprint arXiv:2111.06395*, 2021a.

Chen, Y., Yi, X., and Caramanis, C. Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates. *IEEE Transactions on Information Theory*, 64(3):1738–1766, 2017.

Chen, Y., Chi, Y., Fan, J., Ma, C., et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021b.

Chen, Y., Ma, C., Poor, H. V., and Chen, Y. Learning mixtures of low-rank models. *IEEE Transactions on Information Theory*, 67(7):4613–4636, 2021c. doi: 10.1109/TIT.2021.3065700.

Cohen, A., Hasidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. Online linear quadratic control. In *Proceedings of the International Conference on Machine Learning*, pp. 1029–1038. PMLR, 2018.

Davis, C. and Kahan, W. M. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

Diakonikolas, I. and Kane, D. M. Small covers for near-zero sets of polynomials and learning latent variable models. In *Proceedings of the 2020 IEEE 61st Annual Symposium*

on Foundations of Computer Science (FOCS), pp. 184–195. IEEE, 2020.

Diamandis, T., Eldar, Y., Fallah, A., Farnia, F., and Ozdaglar, A. A Wasserstein minimax framework for mixed linear regression. In Meila, M. and Zhang, T. (eds.), *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2697–2706. PMLR, 18–24 Jul 2021.

Du, K., Meng, Q., and Zhang, F. A q-learning algorithm for discrete-time linear-quadratic control with random parameters of unknown distribution: convergence and stabilization. *arXiv preprint arXiv:2011.04970*, 2020.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

D'Urso, P., De Giovanni, L., and Massari, R. Time series clustering by a robust autoregressive metric with application to air pollution. *Chemometrics and Intelligent Laboratory Systems*, 141:107–124, 2015.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. On adaptive linear–quadratic regulators. *Automatica*, 117: 108982, 2020.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the International Conference on Machine Learning*, pp. 1467–1476. PMLR, 2018.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning - Volume 70*, pp. 1126–1135. JMLR.org, 2017.

Foster, D., Sarkar, T., and Rakhlin, A. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pp. 851–861. PMLR, 2020.

Ghahramani, Z. and Hinton, G. E. Variational learning for switching state-space models. *Neural computation*, 12 (4):831–864, 2000.

Hallac, D., Vare, S., Boyd, S., and Leskovec, J. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 215–223, 2017.

Harrison, J., Sharma, A., Finn, C., and Pavone, M. Continuous meta-learning without tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Hong, J., Kveton, B., Zaheer, M., Chow, Y., Ahmed, A., and Boutilier, C. Latent bandits revisited. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13423–13433. Curran Associates, Inc., 2020.

Huang, L., Sudhir, K., and Vishnoi, N. Coresets for time series clustering. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jedra, Y. and Proutiere, A. Sample complexity lower bounds for linear system identification. In *Proceedings of the 2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 2676–2681. IEEE, 2019.

Kailath, T., Sayed, A. H., and Hassibi, B. *Linear Estimation*. Prentice Hall, 2000.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15312–15325. Curran Associates, Inc., 2020.

Kalliovirta, L., Meitz, M., and Saikkonen, P. Gaussian mixture vector autoregression. *Journal of Econometrics*, 192(2):485–498, 2016.

Khalil, I. S., Doyle, J., and Glover, K. *Robust and Optimal Control*. Prentice Hall, New Jersey, 1996.

Kong, W., Somani, R., Kakade, S., and Oh, S. Robust meta-learning for mixed linear regression with small batches. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4683–4696. Curran Associates, Inc., 2020a.

Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. Meta-learning for mixed linear regression. In *Proceedings of the International Conference on Machine Learning*, pp. 5394–5404. PMLR, 2020b.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Rl for latent MDPs: Regret guarantees and a lower bound. *arXiv preprint arXiv:2102.04939*, 2021a.

Kwon, J., Ho, N., and Caramanis, C. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1405–1413. PMLR, 2021b.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. Logarithmic regret bound in partially observable linear dynamical systems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20876–20888. Curran Associates, Inc., 2020.

Li, Y. and Liang, Y. Learning mixtures of linear regressions with nearly optimal complexity. In *Proceedings of the Conference On Learning Theory*, pp. 1125–1144, 2018.

Liao, T. W. Clustering of time series data – a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

Liberzon, D. *Switching in Systems and Control*. Springer Science & Business Media, 2003.

Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., and Paninski, L. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial Intelligence and Statistics*, pp. 914–922. PMLR, 2017.

Liu, Y., Guo, Z., and Brunskill, E. Pac continuous state online multitask reinforcement learning with identification. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 438–446, 2016.

Ljung, L. System identification. In *Signal analysis and prediction*, pp. 163–173. Springer, 1998.

Lugosi, G. and Mendelson, S. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Maharaj, E. A. Cluster of time series. *Journal of Classification*, 17(2):297–314, 2000.

Maillard, O.-A. and Mannor, S. Latent bandits. In *Proceedings of the International Conference on Machine Learning*, pp. 136–144. PMLR, 2014.

Malekzadeh, M., Clegg, R. G., Cavallaro, A., and Haddadi, H. Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, IoTDI '19, pp. 49–58. ACM, 2019.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P., and Wainwright, M. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 2916–2925. PMLR, 2019.

Mania, H., Tu, S., and Recht, B. Certainty equivalence is efficient for linear quadratic control. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 10154–10164, 2019.

Mania, H., Jordan, M. I., and Recht, B. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Maurer, A., Pontil, M., and Romera-Paredes, B. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

McCulloch, R. E. and Tsay, R. S. Statistical analysis of economic time series via markov switching models. *Journal of Time Series Analysis*, 15(5):523–539, 1994.

Mezer, A., Yovel, Y., Pasternak, O., Gorfine, T., and Assaf, Y. Cluster analysis of resting-state fmri time series. *Neuroimage*, 45(4):1117–1125, 2009.

Mhammedi, Z., Foster, D. J., Simchowitz, M., Misra, D., Sun, W., Krishnamurthy, A., Rakhlin, A., and Langford, J. Learning the linear quadratic regulator from nonlinear observations. *arXiv preprint arXiv:2010.03799*, 2020.

Minasyan, E., Gradu, P., Simchowitz, M., and Hazan, E. Online control of unknown time-varying dynamical systems. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Modi, A., Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Joint learning of linear time-invariant dynamical systems. *arXiv preprint arXiv:2112.10955*, 2021.

Oymak, S. and Ozay, N. Non-asymptotic identification of LTI systems from a single trajectory. In *Proceedings of the 2019 American Control Conference (ACC)*, pp. 5655–5661. IEEE, 2019.

Pal, S. and Mazumdar, A. Recovery of sparse signals from a mixture of linear samples. In *Proceedings of the International Conference on Machine Learning*, pp. 7466–7475. PMLR, 2020.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22 (10):1345–1359, 2009.

Pathak, R., Sen, R., Rao, N., Erichson, N. B., Jordan, M. I., and Dhillon, I. S. Cluster-and-conquer: A framework for time-series forecasting. *arXiv preprint arXiv:2110.14011*, 2021.

Qu, G., Shi, Y., Lale, S., Anandkumar, A., and Wierman, A. Stable online control of linear time-varying systems. In *Learning for Dynamics and Control*, pp. 742–753. PMLR, 2021.

Quandt, R. E. and Ramsey, J. B. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):730–738, 1978.

Sarkar, T. and Rakhlin, A. Near optimal finite time identification of arbitrary linear dynamical systems. In *Proceedings of the International Conference on Machine Learning*, pp. 5610–5618. PMLR, 2019.

Sarkar, T., Rakhlin, A., and Dahleh, M. Nonparametric system identification of stochastic switched linear systems. In *Proceedings of the 2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 3623–3628. IEEE, 2019.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. Finite time LTI system identification. *Journal of Machine Learning Research*, 22(26):1–61, 2021.

Schluter, H., Solowjow, F., and Trimpe, S. Event-triggered learning for linear quadratic control. *IEEE Transactions on Automatic Control*, 2020.

Sedghi, L., Ijaz, Z., Witheephanich, K., Pesch, D., et al. Machine learning in event-triggered control: Recent advances and open issues. *arXiv preprint arXiv:2009.12783*, 2020.

Shi, P. and Li, F. A survey on markovian jump systems: modeling and design. *International Journal of Control, Automation and Systems*, 13(1):1–16, 2015.

Simchowitz, M. and Foster, D. Naive exploration is optimal for online lqr. In *Proceedings of the International Conference on Machine Learning*, pp. 8937–8948. PMLR, 2020.

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

Simchowitz, M., Boczar, R., and Recht, B. Learning linear dynamical systems with semi-parametric least squares. In *Proceedings of the Conference on Learning Theory*, pp. 2714–2802. PMLR, 2019.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Sodhani, S., Zhang, A., and Pineau, J. Multi-task reinforcement learning with context-based representations. In Meila, M. and Zhang, T. (eds.), *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9767–9779. PMLR, 18–24 Jul 2021.

Steimle, L. N., Kaufman, D. L., and Denton, B. T. Multi-model markov decision processes. *IISE Transactions*, pp. 1–16, 2021.

Sun, Y., Oymak, S., and Fazel, M. Finite sample system identification: Optimal rates and the role of regularization. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 16–25. PMLR, 10–11 Jun 2020.

Sun, Z. *Switched Linear Systems: Control and Design*. Springer Science & Business Media, 2006.

Takano, K., Stefanovic, M., Rosenkranz, T., and Ehring, T. Clustering individuals on limited features of a vector autoregressive model. *Multivariate Behavioral Research*, pp. 1–19, 2020.

Taylor, M. E. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Tirinzoni, A., Poiani, R., and Restelli, M. Sequential transfer in reinforcement learning with a generative model. In *Proceedings of the International Conference on Machine Learning*, pp. 9481–9492. PMLR, 2020.

Tripuraneni, N., Jordan, M. I., and Jin, C. On the theory of transfer learning: The importance of task diversity. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tsiamis, A., Matni, N., and Pappas, G. Sample complexity of kalman filtering for unknown systems. In *Learning for Dynamics and Control*, pp. 435–444. PMLR, 2020.

Vershynin, R. *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

Wilson, A., Fern, A., Ray, S., and Tadepalli, P. Multi-task reinforcement learning: a hierarchical Bayesian approach. In *Proceedings of the International Conference on Machine Learning*, pp. 1015–1022, 2007.

Wong, C. S. and Li, W. K. On a mixture autoregressive model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):95–115, 2000.

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

Yin, D., Pedarsani, R., Chen, Y., and Ramchandran, K. Learning mixtures of sparse linear regressions using sparse graph codes. *IEEE Transactions on Information Theory*, 65(3):1430–1451, 2018.

Zhao, P., Kang, Y., and Zhao, Y.-B. A brief tutorial and survey on markovian jump systems: Stability and control. *IEEE Systems, Man, and Cybernetics Magazine*, 5(2): 37–C3, 2019.

Zheng, Y., Furieri, L., Kamgarpour, M., and Li, N. Sample complexity of linear quadratic gaussian (lqg) control for output feedback systems. In *Learning for Dynamics and Control*, pp. 559–570. PMLR, 2021.

Table 1. A list of notation and parameters. In the subscripts of $\mathcal{M}_{\mathsf{o}}, T_{\mathsf{o}}, T_{\mathsf{total,o}}$, the symbol o takes value in $\{\mathsf{subspace}, \mathsf{clustering}, \mathsf{classification}\}$.

Notation	Explanation
d	State dimension
K	Number of LDS models
k_m	The unknown label (latent variable) of the m-th trajectory
$\mathcal{M}_{\mathsf{o}}$	Subsets of M trajectories, $\{1,\dots,M\} = \mathcal{M}_{\mathsf{subspace}} \cup \mathcal{M}_{\mathsf{clustering}} \cup \mathcal{M}_{\mathsf{classification}}$
p_{min}	A lower bound for the fraction of trajectories generated by each model
T_{o}	Short trajectory length for $\mathcal{M}_{\mathsf{o}}$
$T_{\mathsf{total,o}}$	Total trajectory length, $T_{\mathsf{total,o}} = T_{\mathsf{o}} \cdot \lvert\mathcal{M}_{\mathsf{o}}\rvert$
$\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}$	State transition matrix and noise covariance matrix of the k-th LDS model
$\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)}$	Order-0 and order-1 stationary autocovariance matrices of the k-th LDS model
κ_A, ρ	$\lVert(\boldsymbol{A}^{(k)})^t\rVert \le \kappa_A \cdot \rho^t, \quad t = 1, 2, \dots$
$\Delta_{\Gamma,Y}, \Delta_{A,W}$	Model separation parameters (see Assumption 3.1)
$W_{\mathsf{min}}, W_{\mathsf{max}}$	$W_{\mathsf{min}} \le \lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)}) \le \lambda_{\mathsf{max}}(\boldsymbol{W}^{(k)}) \le W_{\mathsf{max}}, \quad 1 \le k \le K$
$\kappa_{w,\mathsf{cross}}, \kappa_w$	$\kappa_{w,\mathsf{cross}} = W_{\mathsf{max}}/W_{\mathsf{min}}; \quad \kappa(\boldsymbol{W}^{(k)}) \le \kappa_w, \quad 1 \le k \le K$
Γ_{max}	$\lVert\boldsymbol{\Gamma}^{(k)}\rVert \le \Gamma_{\mathsf{max}}, \quad 1 \le k \le K$

This appendix is organized as follows. Appendix A complements Section 2, providing full details of the proposed Algorithms 2 – 5. Appendix B includes detailed theoretical results for the algorithms, with proofs postponed to Appendix C; it also provides a proof for our main results in Section 3. Appendix D collects some miscellaneous results. In Appendix E we discuss on some potential extensions of our methods. Finally, Appendix F includes additional experimental results. For the readers' convenience, we include Table 1 for a quick reference to the key notation and parameters used in our analysis.

Additional notation. Let $\mathsf{vec}(\cdot)$ denote the vectorization of a matrix. For two matrices $\boldsymbol{A}, \boldsymbol{B}$, let $\boldsymbol{A} \otimes \boldsymbol{B}$ denote their Kronecker product. Given a sequence of real numbers $\{x_i\}_{1 \le i \le N}$, we denote its median as $\mathsf{median}\{x_i, 1 \le i \le N\}$. We shall also let $\mathsf{poly}(n)$ denote some polynomial in n of a constant degree. For a positive integer n, we denote $[n] := \{1, 2, \dots, n\}$.

A. Detailed Algorithms

A.1. Subspace estimation

Procedure. Recall that the notation $(\boldsymbol{\Gamma}^{(k)})_i \in \mathbb{R}^d$ (resp. $(\boldsymbol{Y}^{(k)})_i$) represents the transpose of the i-th row of $\boldsymbol{\Gamma}^{(k)}$ (resp. $\boldsymbol{Y}^{(k)}$) defined in (5). With this set of notation in place, let us define the following subspaces:

$$\boldsymbol{V}_i^\star := \mathsf{span}\Big\{(\boldsymbol{\Gamma}^{(k)})_i, 1 \le k \le K\Big\}, \quad \boldsymbol{U}_i^\star := \mathsf{span}\Big\{(\boldsymbol{Y}^{(k)})_i, 1 \le k \le K\Big\}, \quad 1 \le i \le d. \tag{12}$$

It is easily seen from the construction that each of these subspaces has rank at most K.

As it turns out, the collection of $2d$ subspaces defined in (12) provides crucial low-dimensional information about the linear dynamical systems of interest. This motivates us to develop a data-driven method to estimate these subspaces, which will in turn allow for proper dimension reduction in subsequent steps. Towards this end, we propose to employ a spectral method for subspace estimation using sample trajectories in $\mathcal{M}_{\mathsf{subspace}}$:

(i) divide $\{0, 1, \dots, T_{\mathsf{subspace}}\}$ into four segments of the same size, and denote the 2nd (resp. 4th) segment as Ω_1 (resp. Ω_2);

(ii) for each $m \in \mathcal{M}_{\mathsf{subspace}}, 1 \le i \le d, j \in \{1,2\}$, compute

$$\boldsymbol{h}_{m,i,j} \coloneqq \frac{1}{|\Omega_j|} \sum_{t \in \Omega_j} (\boldsymbol{x}_{m,t})_i \, \boldsymbol{x}_{m,t}, \quad \boldsymbol{g}_{m,i,j} \coloneqq \frac{1}{|\Omega_j|} \sum_{t \in \Omega_j} (\boldsymbol{x}_{m,t+1})_i \, \boldsymbol{x}_{m,t}, ; \tag{13}$$

(iii) for each $1 \le i \le d$, compute the following matrices

$$\widehat{\boldsymbol{H}}_i \coloneqq \frac{1}{|\mathcal{M}_{\mathsf{subspace}}|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} \boldsymbol{h}_{m,i,1} \boldsymbol{h}_{m,i,2}^\top, \quad \widehat{\boldsymbol{G}}_i \coloneqq \frac{1}{|\mathcal{M}_{\mathsf{subspace}}|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} \boldsymbol{g}_{m,i,1} \boldsymbol{g}_{m,i,2}^\top, \tag{14}$$

and let $\boldsymbol{V}_i \in \mathbb{R}^{d \times K}$ (resp. $\boldsymbol{U}_i$) be the top-K eigenspace of $\widehat{\boldsymbol{H}}_i + \widehat{\boldsymbol{H}}_i^\top$ (resp. $\widehat{\boldsymbol{G}}_i + \widehat{\boldsymbol{G}}_i^\top$).

The output $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}_{1 \le i \le d}$ will then serve as our estimate of $\{\boldsymbol{V}_i^\star, \boldsymbol{U}_i^\star\}_{1 \le i \le d}$. This spectral approach is summarized in Algorithm 2.

Rationale. According to the mixing property of LDS, if T_{subspace} is larger than some appropriately defined mixing time, then each sample trajectory in $\mathcal{M}_{\mathsf{subspace}}$ will mix sufficiently and nearly reach stationarity (when constrained to $t \in \Omega_1 \cup \Omega_2$). In this case, it is easy to check that

$$\mathbb{E}\big[\boldsymbol{h}_{m,i,j}\big] \approx (\boldsymbol{\Gamma}^{(k_m)})_i, \quad \mathbb{E}\big[\boldsymbol{g}_{m,i,j}\big] \approx (\boldsymbol{Y}^{(k_m)})_i, \quad 1 \le i \le d, \quad j \in \{1,2\}.$$

Moreover, the samples in Ω_1 are nearly independent of those in Ω_2 as long as the spacing between them exceeds the mixing time, and therefore,

$$\mathbb{E}\Big[\widehat{\boldsymbol{H}}_i\Big] \approx \frac{1}{|\mathcal{M}_{\mathsf{subspace}}|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} (\boldsymbol{\Gamma}^{(k_m)})_i (\boldsymbol{\Gamma}^{(k_m)})_i^\top = \sum_{k=1}^K p_{\mathsf{subspace}}^{(k)} (\boldsymbol{\Gamma}^{(k)})_i (\boldsymbol{\Gamma}^{(k)})_i{}^\top =: \boldsymbol{H}_i, \tag{15a}$$

$$\mathbb{E}\Big[\widehat{\boldsymbol{G}}_i\Big] \approx \frac{1}{|\mathcal{M}_{\mathsf{subspace}}|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} (\boldsymbol{Y}^{(k_m)})_i (\boldsymbol{Y}^{(k_m)})_i^\top = \sum_{k=1}^K p_{\mathsf{subspace}}^{(k)} (\boldsymbol{Y}^{(k)})_i (\boldsymbol{Y}^{(k)})_i{}^\top =: \boldsymbol{G}_i, \tag{15b}$$

where $p_{\mathsf{subspace}}^{(k)}$ denotes the fraction of sample trajectories generated by the k-th model, namely,

$$p_{\mathsf{subspace}}^{(k)} \coloneqq \frac{1}{|\mathcal{M}_{\mathsf{subspace}}|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} \mathbb{1}(k_m = k), \quad 1 \le k \le K. \tag{16}$$

As a consequence, if T_{subspace} and $|\mathcal{M}_{\mathsf{subspace}}|$ are both sufficiently large, then one might expect $\widehat{\boldsymbol{H}}_i$ (resp. $\widehat{\boldsymbol{G}}_i$) to be a reasonably good approximation of $\boldsymbol{H}_i$ (resp. $\boldsymbol{G}_i$), the latter of which has rank at most K and has $\boldsymbol{V}_i^\star$ (resp. $\boldsymbol{U}_i^\star$) as its eigenspace. All this motivates us to compute the rank-K eigenspaces of $\widehat{\boldsymbol{H}}_i + \widehat{\boldsymbol{H}}_i^\top$ and $\widehat{\boldsymbol{G}}_i + \widehat{\boldsymbol{G}}_i^\top$ in Algorithm 2.

A.2. Clustering

This step seeks to divide the sample trajectories in $\mathcal{M}_{\mathsf{clustering}}$ into K clusters (albeit not perfectly), such that the trajectories in each cluster are primarily generated by the same LDS model. We intend to achieve this by performing pairwise comparisons of the autocovariance matrices associated with the sample trajectories.

Key observation. Even though $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}) \neq (\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$ for $k \neq \ell$, it is indeed possible that $\boldsymbol{\Gamma}^{(k)} = \boldsymbol{\Gamma}^{(\ell)}$ or $\boldsymbol{Y}^{(k)} = \boldsymbol{Y}^{(\ell)}$. Therefore, in order to differentiate sample trajectories generated by different systems based on $\boldsymbol{\Gamma}(\boldsymbol{A}, \boldsymbol{W})$ and $\boldsymbol{Y}(\boldsymbol{A}, \boldsymbol{W})$, it is important to ensure separation of $(\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)})$ and $(\boldsymbol{\Gamma}^{(\ell)}, \boldsymbol{Y}^{(\ell)})$ when $k \neq l$, which can be guaranteed by the following fact.

Fact A.1. *If $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}) \neq (\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$, then we have either $\boldsymbol{\Gamma}^{(k)} \neq \boldsymbol{\Gamma}^{(\ell)}$ or $\boldsymbol{Y}^{(k)} \neq \boldsymbol{Y}^{(\ell)}$ (or both).*

Proof. First, observe that the definitions of $\{\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)}\}$ in (5) suggest that we can recover $\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}$ from $\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)}$ as follows:

$$\boldsymbol{A}^{(k)} = \boldsymbol{Y}^{(k)} \boldsymbol{\Gamma}^{(k)^{-1}}, \; \boldsymbol{W}^{(k)} = \boldsymbol{\Gamma}^{(k)} - \boldsymbol{A}^{(k)} \boldsymbol{\Gamma}^{(k)} \boldsymbol{A}^{(k)^\top}. \tag{17}$$

Algorithm 2 Subspace estimation

Input: short trajectories $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\mathsf{subspace}}}$, where $\boldsymbol{X}_m = \{\boldsymbol{x}_{m,t}\}_{0\le t\le T_{\mathsf{subspace}}}$.
Let $N \leftarrow \lfloor T_{\mathsf{subspace}}/4 \rfloor$, and $\Omega_1 \leftarrow \{N+j, 1\le j\le N\}, \Omega_2 \leftarrow \{3N+j, 1\le j\le N\}$.
for $(m,i,j) \in \mathcal{M}_{\mathsf{subspace}} \times [d] \times [2]$ **do**
  Compute

$$\boldsymbol{h}_{m,i,j} \leftarrow \frac{1}{|\Omega_j|}\sum_{t\in\Omega_j} (\boldsymbol{x}_{m,t})_i\, \boldsymbol{x}_{m,t}, \quad \boldsymbol{g}_{m,i,j} \leftarrow \frac{1}{|\Omega_j|}\sum_{t\in\Omega_j} (\boldsymbol{x}_{m,t+1})_i\, \boldsymbol{x}_{m,t}.$$

end for
for $i = 1,\ldots,d$ **do**
  Compute

$$\widehat{\boldsymbol{H}}_i \leftarrow \frac{1}{|\mathcal{M}_{\mathsf{subspace}}|}\sum_{m\in\mathcal{M}_{\mathsf{subspace}}} \boldsymbol{h}_{m,i,1}\boldsymbol{h}_{m,i,2}^\top, \quad \widehat{\boldsymbol{G}}_i \leftarrow \frac{1}{|\mathcal{M}_{\mathsf{subspace}}|}\sum_{m\in\mathcal{M}_{\mathsf{subspace}}} \boldsymbol{g}_{m,i,1}\boldsymbol{g}_{m,i,2}^\top,$$

  Let $\boldsymbol{V}_i \in \mathbb{R}^{d\times K}$ (resp. $\boldsymbol{U}_i$) be the top-K eigenspace of $\widehat{\boldsymbol{H}}_i + \widehat{\boldsymbol{H}}_i^\top$ (resp. $\widehat{\boldsymbol{G}}_i + \widehat{\boldsymbol{G}}_i^\top$).
end for
Output: subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}_{1\le i\le d}$.

With this, we can prove the fact by contradiction. Suppose instead that $\boldsymbol{\Gamma}^{(k)} = \boldsymbol{\Gamma}^{(\ell)}$ and $\boldsymbol{Y}^{(k)} = \boldsymbol{Y}^{(\ell)}$, then (17) yields

$$\begin{aligned}
\boldsymbol{A}^{(k)} &= \boldsymbol{Y}^{(k)}{\boldsymbol{\Gamma}^{(k)}}^{-1} = \boldsymbol{Y}^{(\ell)}{\boldsymbol{\Gamma}^{(\ell)}}^{-1} = \boldsymbol{A}^{(\ell)}, \quad \text{and} \\
\boldsymbol{W}^{(k)} &= \boldsymbol{\Gamma}^{(k)} - \boldsymbol{A}^{(k)}\boldsymbol{\Gamma}^{(k)}{\boldsymbol{A}^{(k)}}^\top = \boldsymbol{\Gamma}^{(\ell)} - \boldsymbol{A}^{(\ell)}\boldsymbol{\Gamma}^{(\ell)}{\boldsymbol{A}^{(\ell)}}^\top = \boldsymbol{W}^{(\ell)},
\end{aligned}$$

which is contradictory to the assumption that $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}) \neq (\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$. □

Idea. Let us compare a pair of sample trajectories $\{\boldsymbol{x}_t\}_{0\le t\le T_{\mathsf{clustering}}}$ and $\{\boldsymbol{z}_t\}_{0\le t\le T_{\mathsf{clustering}}}$, where $\{\boldsymbol{x}_t\}$ is generated by the system $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})$ and $\{\boldsymbol{z}_t\}$ by the system $(\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$. In order to determine whether $k=\ell$, we propose to estimate the quantity $\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2$ using the data samples $\{\boldsymbol{x}_t\}$ and $\{\boldsymbol{z}_t\}$, which is expected to be small (resp. large) if $k=\ell$ (resp. $k\neq\ell$). To do so, let us divide $\{0,1,\ldots,T_{\mathsf{clustering}}\}$ evenly into four segments, and denote by Ω_1 (resp. Ω_2) the 2nd (resp. 4th) segment, akin to what we have done in the previous step. Observe that

$$\boldsymbol{U}_i{}^\top \mathbb{E}\Big[\big((\boldsymbol{x}_{t+1})_i\boldsymbol{x}_t - (\boldsymbol{z}_{t+1})_i\boldsymbol{z}_t\big)\Big] \approx \boldsymbol{U}_i{}^\top\big((\boldsymbol{Y}^{(k)})_i - (\boldsymbol{Y}^{(\ell)})_i\big)$$

for all $1\le i\le d$ and $t\in\Omega_1\cup\Omega_2$. Assuming sufficient mixing and utilizing (near) statistical independence due to sample splitting, we might resort to the following statistic

$$\mathsf{stat}_Y \coloneqq \sum_{i=1}^d \Big\langle \boldsymbol{U}_i{}^\top \frac{1}{|\Omega_1|}\sum_{t\in\Omega_1}\big((\boldsymbol{x}_{t+1})_i\boldsymbol{x}_t - (\boldsymbol{z}_{t+1})_i\boldsymbol{z}_t\big), \boldsymbol{U}_i{}^\top \frac{1}{|\Omega_2|}\sum_{t\in\Omega_2}\big((\boldsymbol{x}_{t+1})_i\boldsymbol{x}_t - (\boldsymbol{z}_{t+1})_i\boldsymbol{z}_t\big)\Big\rangle, \tag{19}$$

whose expectation is given by

$$\begin{aligned}
\mathbb{E}[\mathsf{stat}_Y] &\approx \sum_{i=1}^d \Big\langle \boldsymbol{U}_i{}^\top\big((\boldsymbol{Y}^{(k)})_i - (\boldsymbol{Y}^{(\ell)})_i\big), \boldsymbol{U}_i{}^\top\big((\boldsymbol{Y}^{(k)})_i - (\boldsymbol{Y}^{(\ell)})_i\big)\Big\rangle \\
&= \sum_{i=1}^d \Big\|\boldsymbol{U}_i{}^\top\big((\boldsymbol{Y}^{(k)})_i - (\boldsymbol{Y}^{(\ell)})_i\big)\Big\|_2^2 \approx \sum_{i=1}^d \big\|(\boldsymbol{Y}^{(k)})_i - (\boldsymbol{Y}^{(\ell)})_i\big\|_2^2 = \big\|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\big\|_{\mathrm{F}}^2;
\end{aligned}$$

here, the first approximation is due to the near independence between samples from Ω_1 and those from Ω_2, whereas the second approximation holds if each subspace $\boldsymbol{U}_i$ is sufficiently close to $\boldsymbol{U}_i^\star = \mathsf{span}\{(\boldsymbol{Y}^{(j)})_i, 1\le j\le K\}$. The purpose of

Algorithm 3 Clustering

Input: short trajectories $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\mathsf{clustering}}}$, where $\boldsymbol{X}_m = \{\boldsymbol{x}_{m,t}\}_{0\le t\le T_{\mathsf{clustering}}}$; subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}_{1\le i\le d}$; testing threshold τ; number of copies G.
Let $N \leftarrow \lfloor T_{\mathsf{clustering}}/4G \rfloor$, and

$$\Omega_{g,1} \leftarrow \Big\{(4g-3)N + j, 1 \le j \le N\Big\}, \quad \Omega_{g,2} \leftarrow \Big\{(4g-1)N + j, 1 \le j \le N\Big\}, \quad 1 \le g \le G.$$

for $(m, i, g, j) \in \mathcal{M}_{\mathsf{clustering}} \times [d] \times [G] \times [2]$ **do**
  Compute

$$\boldsymbol{h}_{m,i,g,j} \leftarrow \frac{1}{|\Omega_{g,j}|} \sum_{t\in\Omega_{g,j}} (\boldsymbol{x}_{m,t})_i \, \boldsymbol{x}_{m,t}, \quad \boldsymbol{g}_{m,i,g,j} \leftarrow \frac{1}{|\Omega_{g,j}|} \sum_{t\in\Omega_{g,j}} (\boldsymbol{x}_{m,t+1})_i \, \boldsymbol{x}_{m,t}.$$

end for
`// Compute the similarity matrix` $\boldsymbol{S}$**:**
for $(m, n) \in \mathcal{M}_{\mathsf{clustering}} \times \mathcal{M}_{\mathsf{clustering}}$ **do**
  for $g = 1, \ldots, G$ **do**
    Compute

$$\mathsf{stat}_{\Gamma,g} \leftarrow \sum_{i=1}^{d} \Big\langle \boldsymbol{V}_i{}^\top \big(\boldsymbol{h}_{m,i,g,1} - \boldsymbol{h}_{n,i,g,1}\big), \boldsymbol{V}_i{}^\top \big(\boldsymbol{h}_{m,i,g,2} - \boldsymbol{h}_{n,i,g,2}\big) \Big\rangle, \tag{18a}$$

$$\mathsf{stat}_{Y,g} \leftarrow \sum_{i=1}^{d} \Big\langle \boldsymbol{U}_i{}^\top \big(\boldsymbol{g}_{m,i,g,1} - \boldsymbol{g}_{n,i,g,1}\big), \boldsymbol{U}_i{}^\top \big(\boldsymbol{g}_{m,i,g,2} - \boldsymbol{g}_{n,i,g,2}\big) \Big\rangle, \tag{18b}$$

  end for
  Set $S_{m,n} \leftarrow \mathbb{1}\Big(\mathsf{median}\big\{\mathsf{stat}_{\Gamma,g}, 1 \le g \le G\big\} + \mathsf{median}\big\{\mathsf{stat}_{Y,g}, 1 \le g \le G\big\} \le \tau\Big)$.
end for
Divide $\mathcal{M}_{\mathsf{clustering}}$ into K clusters $\{\mathcal{C}_k\}_{1\le k\le K}$ according to $\{S_{m,n}\}_{m,n\in\mathcal{M}_{\mathsf{clustering}}}$.
Output: clusters $\{\mathcal{C}_k\}_{1\le k\le K}$.

utilizing $\{\boldsymbol{U}_i\}$ is to reduce the variance of stat_Y. Similarly, we can compute another statistic (by replacing $\{\boldsymbol{U}_i\}$ with $\{\boldsymbol{V}_i\}$ and $(\boldsymbol{x}_{t+1})_i, (\boldsymbol{z}_{t+1})_i$ with $(\boldsymbol{x}_t)_i, (\boldsymbol{z}_t)_i$) as follows:

$$\mathsf{stat}_\Gamma \coloneqq \sum_{i=1}^{d} \Big\langle \boldsymbol{V}_i{}^\top \frac{1}{|\Omega_1|} \sum_{t\in\Omega_1} \big((\boldsymbol{x}_t)_i \boldsymbol{x}_t - (\boldsymbol{z}_t)_i \boldsymbol{z}_t\big), \boldsymbol{V}_i{}^\top \frac{1}{|\Omega_2|} \sum_{t\in\Omega_2} \big((\boldsymbol{x}_t)_i \boldsymbol{x}_t - (\boldsymbol{z}_t)_i \boldsymbol{z}_t\big) \Big\rangle, \tag{20}$$

which has expectation

$$\mathbb{E}[\mathsf{stat}_\Gamma] \approx \sum_{i=1}^{d} \|\boldsymbol{V}_i{}^\top ((\boldsymbol{\Gamma}^{(k)})_i - (\boldsymbol{\Gamma}^{(\ell)})_i)\|_2^2 \approx \|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2.$$

Consequently, we shall declare $k \neq \ell$ if $\mathsf{stat}_\Gamma + \mathsf{stat}_Y$ exceeds some appropriate threshold τ.

Procedure. We are now positioned to describe the proposed clustering procedure. We first compute the statistics stat_Γ and stat_Y for each pair of sample trajectories in $\mathcal{M}_{\mathsf{clustering}}$ by means of the method described above, and then construct a similarity matrix $\boldsymbol{S}$, in a way that $S_{m,n}$ is set to 0 if $\mathsf{stat}_\Gamma + \mathsf{stat}_Y$ (computed for the m-th and n-th trajectories) is larger than a threshold τ, or set to 1 otherwise. In order to enhance the robustness of these statistics, we divide $\{0, 1, \ldots, T_{\mathsf{clustering}}\}$ into $4G$ (instead of 4) segments, compute G copies of stat_Γ and stat_Y, and take the medians of these values. Next, we apply a mainstream clustering algorithm (e.g. spectral clustering (Chen et al., 2021b)) to the similarity matrix $\boldsymbol{S}$, and divide $\mathcal{M}_{\mathsf{clustering}}$ into K disjoint clusters $\{\mathcal{C}_k\}_{1\le k\le K}$. The complete clustering procedure is provided in Algorithm 3. The threshold τ shall be chosen to be on the order of $\min_{1\le k<\ell\le K}\{\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2\}$ (which is strictly positive due to Fact 2.1), and it suffices to set the number of copies G to be on some logarithmic order. Our choice of these parameters are specified in Theorem 3.4.

Algorithm 4 Least squares and covariance estimation

Input: clusters $\{\mathcal{C}_k\}_{1\le k\le K}$.
for $k = 1, \dots, K$ **do**
  Compute

$$\widehat{\boldsymbol{A}}^{(k)} \leftarrow \Big(\sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-1} \boldsymbol{x}_{m,t+1}\boldsymbol{x}_{m,t}{}^\top\Big)\Big(\sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-1} \boldsymbol{x}_{m,t}\boldsymbol{x}_{m,t}{}^\top\Big)^{-1}, \quad \text{and} \tag{22a}$$

$$\widehat{\boldsymbol{W}}^{(k)} \leftarrow \frac{1}{\sum_{m\in\mathcal{C}_k} T_m}\sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-1} \widehat{\boldsymbol{w}}_{m,t}\widehat{\boldsymbol{w}}_{m,t}^\top, \quad \text{where} \quad \widehat{\boldsymbol{w}}_{m,t} = \boldsymbol{x}_{m,t+1} - \widehat{\boldsymbol{A}}^{(k)}\boldsymbol{x}_{m,t}. \tag{22b}$$

end for
Output: estimated models $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$.

A.3. Model estimation

Suppose that we have obtained reasonably good clustering accuracy in the previous step, namely for each $1 \le k \le K$, the cluster $\mathcal{C}_k$ output by Algorithm 3 contains mostly trajectories generated by the same model $(\boldsymbol{A}^{(\pi(k))}, \boldsymbol{W}^{(\pi(k))})$, with π representing some permutation function of $\{1, \dots, K\}$. We propose to obtain a coarse model estimation and covariance estimation, as summarized in Algorithm 4. More specifically, for each k, we use the samples $\{\{\boldsymbol{x}_{m,t}\}_{0\le t\le T_m}\}_{m\in\mathcal{C}_k}$ to obtain an estimate of $\boldsymbol{A}^{(\pi(k))}$ by solving the following least-squares problem (Simchowitz et al., 2018; Sarkar & Rakhlin, 2019)

$$\widehat{\boldsymbol{A}}^{(k)} := \arg\min_{\boldsymbol{A}} \sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-1} \|\boldsymbol{x}_{m,t+1} - \boldsymbol{A}\boldsymbol{x}_{m,t}\|_2^2, \tag{21}$$

whose closed-form solution is given in (22a). Next, we can use $\widehat{\boldsymbol{A}}^{(k)}$ to estimate the noise vector

$$\widehat{\boldsymbol{w}}_{m,t} := \boldsymbol{x}_{m,t+1} - \widehat{\boldsymbol{A}}^{(k)}\boldsymbol{x}_{m,t} \approx \boldsymbol{w}_{m,t},$$

and finally estimate the noise covariance $\boldsymbol{W}^{(\pi(k))}$ with the empirical covariance of $\{\widehat{\boldsymbol{w}}_{m,t}\}$, as shown in (22b).

A.4. Classification

Procedure. In the previous steps, we have obtained initial clusters $\{\mathcal{C}_k\}_{1\le k\le K}$ and coarse model estimates $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$. With the assistance of additional sample trajectories $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\text{classification}}}$, we can infer their latent labels and assign them to the corresponding clusters; the procedure is stated in Algorithm 5 and will be explained shortly. Once this is done, we can run Algorithm 4 again with the updated clusters $\{\mathcal{C}_k\}$ to refine our model estimates, which is exactly the last step of Algorithm 1.

Rationale. The strategy of inferring labels in Algorithm 5 can be derived from the maximum likelihood estimator, under the assumption that the noise vectors $\{\boldsymbol{w}_{m,t}\}$ follow Gaussian distributions. Note, however, that even in the absence of Gaussian assumptions, Algorithm 5 remains effective in principle. To see this, consider a short trajectory $\{\boldsymbol{x}_t\}_{0\le t\le T}$ generated by model $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})$, i.e. $\boldsymbol{x}_{t+1} = \boldsymbol{A}^{(k)}\boldsymbol{x}_t + \boldsymbol{w}_t$ where $\mathbb{E}[\boldsymbol{w}_t] = \boldsymbol{0}, \mathsf{cov}(\boldsymbol{w}_t) = \boldsymbol{W}^{(k)}$. Let us define the following loss function

$$L(\boldsymbol{A}, \boldsymbol{W}) := T\cdot\log\det(\boldsymbol{W}) + \sum_{t=0}^{T-1}(\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t)^\top\boldsymbol{W}^{-1}(\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t). \tag{23}$$

With some elementary calculation, we can easily check that for any incorrect label $\ell \ne k$, it holds that $\mathbb{E}[L(\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)}) - L(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})] > 0$, with the proviso that $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}) \ne (\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$ and $\{\boldsymbol{x}_t\}$ are non-degenerate in some sense; in other words, the correct model $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})$ achieves the minimal expected loss (which, due to the quadratic form of the loss function, depends solely on the first and second moments of the distributions of $\{\boldsymbol{w}_t\}$, as well as the initial state $\boldsymbol{x}_0$). This justifies the proposed procedure for inferring unknown labels in Algorithm 5, provided that T_m is large enough and that the estimated LDS models are sufficiently reliable.

Algorithm 5 Classification

Input: short trajectories $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\mathsf{classification}}}$, where $\boldsymbol{X}_m = \{\boldsymbol{x}_{m,t}\}_{0\le t\le T_m}$; coarse models $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$; clusters $\{\mathcal{C}_k\}_{1\le k\le K}$.
for $m \in \mathcal{M}_{\mathsf{classification}}$ **do**
  Infer the label of the m-th trajectory by

$$\widehat{k}_m \leftarrow \arg\min_{\ell} \left\{ T_m \cdot \log\det(\widehat{\boldsymbol{W}}^{(\ell)}) + \sum_{t=0}^{T_m-1} (\boldsymbol{x}_{m,t+1} - \widehat{\boldsymbol{A}}^{(\ell)}\boldsymbol{x}_{m,t})^\top (\widehat{\boldsymbol{W}}^{(\ell)})^{-1} (\boldsymbol{x}_{m,t+1} - \widehat{\boldsymbol{A}}^{(\ell)}\boldsymbol{x}_{m,t}) \right\},$$

  then add m to cluster $\mathcal{C}_{\widehat{k}_m}$.
end for
Output: updated clusters $\{\mathcal{C}_k\}_{1\le k\le K}$.

B. Detailed analysis

This section provides detailed, modular theoretical results for the performance of Algorithms 2–5, and concludes with a proof for the main theorems in Section 3.2 (i.e. the performance guarantees of Algorithm 1).

Subspace estimation. The following theorem provides upper bounds on the errors of subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}$ output by Algorithm 2, assuming sufficient mixing of the short trajectories.

Theorem B.1. *Consider the model (1) under the assumptions in Sections 2 and 3.1. There exist some universal constants $C_1, C_2, C_3 > 0$ such that the following holds. Suppose that we run Algorithm 2 with data $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\mathsf{subspace}}}$ obeying*

$$T_{\mathsf{subspace}} \ge C_1 \cdot t_{\mathsf{mix}}, \quad T_{\mathsf{total,subspace}} \ge C_2 \cdot t_{\mathsf{mix}} d \cdot \log\frac{T_{\mathsf{total}} d}{\delta}, \quad \textit{where} \quad t_{\mathsf{mix}} \coloneqq \frac{1}{1-\rho}\cdot\log\left(\frac{d\kappa_A T_{\mathsf{total}}}{\delta}\right).$$

Then with probability at least $1-\delta$, Algorithm 2 outputs $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}_{1\le i\le d}$ satisfying the following: for all $1\le k\le K$ and $1\le i\le d$,

$$\begin{aligned}&\max\left\{\|(\boldsymbol{\Gamma}^{(k)})_i - \boldsymbol{V}_i\boldsymbol{V}_i{}^\top(\boldsymbol{\Gamma}^{(k)})_i\|_2, \|(\boldsymbol{Y}^{(k)})_i - \boldsymbol{U}_i\boldsymbol{U}_i{}^\top(\boldsymbol{Y}^{(k)})_i\|_2\right\}\\ &\quad\le C_3\cdot\Gamma_{\mathsf{max}}\left(\frac{K}{p_{\mathsf{min}}}\right)^{1/2}\left(\frac{t_{\mathsf{mix}} d}{T_{\mathsf{total,subspace}}}\log^3\frac{T_{\mathsf{total}} d}{\delta}\right)^{1/4}.\end{aligned} \tag{24}$$

Our proof (deferred to Appendix C.2) includes a novel perturbation analysis; the resulted error bound (24) has a $1/T_{\mathsf{total,subspace}}^{1/4}$ dependence and is gap-free (i.e. independent of the eigenvalue gaps of the ground-truth low-rank matrices, which can be arbitrarily close to zero in the worst case). It is possible to adapt the existing perturbation results in (Kong et al., 2020b;a) to our setting (which we include in Lemma C.3 in the appendix for completeness); however, one of them is dependent on the eigenvalue gaps, while the other one incurs a worse $1/T_{\mathsf{total,subspace}}^{1/6}$ dependence. It would be interesting future work to investigate whether a gap-free bound with a $1/T_{\mathsf{total,subspace}}^{1/2}$ dependence is possible.

Clustering. Our next theorem shows that Algorithm 3 achieves exact clustering of $\mathcal{M}_{\mathsf{clustering}}$, if $T_{\mathsf{clustering}}$ is sufficiently large and subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}$ are accurate. The proof is deferred to Appendix C.3.

Theorem B.2. *Consider the model (1) under the assumptions in Sections 2 and 3.1. There exist universal constants $C_1, C_2, c_3 > 0$ such that the following holds. Suppose that we run Algorithm 3 with data $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\mathsf{clustering}}}$, independent subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}_{1\le i\le d}$ and parameters τ, G that satisfy the following:*

- *The threshold τ obeys $1/8 < \tau/\Delta_{\Gamma,Y}^2 < 3/8$;*
- *The short trajectory length $T_{\mathsf{clustering}} = NG$, where*

$$G \ge C_1\cdot\log\frac{|\mathcal{M}_{\mathsf{clustering}}|}{\delta}, \quad N \ge C_2\left(\frac{\Gamma_{\mathsf{max}}^2\kappa_A^2\sqrt{dK}}{\Delta_{\Gamma,Y}^2}+1\right)\frac{1}{1-\rho}\log\left(\left(\frac{\Gamma_{\mathsf{max}}}{\Delta_{\Gamma,Y}}+2\right)\frac{d\kappa_A T_{\mathsf{total}}}{\delta}\right);$$

- *The subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}_{1\le i\le d}$ satisfy that, for all $1 \le i \le d$ and $1 \le k \le K$,*

$$\max\left\{\|(\boldsymbol{\Gamma}^{(k)})_i - \boldsymbol{V}_i\boldsymbol{V}_i^\top(\boldsymbol{\Gamma}^{(k)})_i\|_2, \|(\boldsymbol{Y}^{(k)})_i - \boldsymbol{U}_i\boldsymbol{U}_i^\top(\boldsymbol{Y}^{(k)})_i\|_2\right\} \le c_3\frac{\Delta_{\Gamma,Y}}{\sqrt{d}}. \tag{25}$$

Then with probability at least $1-\delta$, Algorithm 3 achieves exact clustering: for all $m_1, m_2 \in \mathcal{M}_{\mathsf{clustering}}$, $S_{m_1,m_2} = 1$ if and only if the m_1-th and m_2-th trajectories are generated by the same model, i.e. they have the same label $k_{m_1} = k_{m_2}$.

Least squares and covariance estimation. The next result controls the model estimation errors of Algorithm 4, under the assumption that every cluster is pure.

Theorem B.3. *Consider the model (1) under the assumptions in Section 3.1. There exist universal constants $C_1, C_2, C_3 > 0$ such that the following holds. Let $\{\mathcal{C}_k\}_{1\le k\le K}$ be subsets of $\mathcal{M}_{\mathsf{clustering}} \cup \mathcal{M}_{\mathsf{classification}}$ such that for all $1 \le k \le K$, $\mathcal{C}_k$ contains only short trajectories generated by model $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})$, namely $k_m = k$ for all $m \in \mathcal{C}_k$. Suppose that for all $m \in \mathcal{M}_{\mathsf{clustering}} \cup \mathcal{M}_{\mathsf{classification}}$, $T_m \ge 4$, and for all $1 \le k \le K$,*

$$T_{\mathsf{total}}^{(k)} := \sum_{m\in\mathcal{C}_k} T_m \ge C_1 d\kappa_w^2\iota, \quad \textit{where} \quad \iota := \log\Big(\frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}}\frac{d\kappa_A T_{\mathsf{total}}}{\delta}\Big).$$

Let $\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}$ be computed by (22) in Algorithm 4. Then with probability at least $1-\delta$, one has

$$\|\widehat{\boldsymbol{A}}^{(k)} - \boldsymbol{A}^{(k)}\| \le C_2\sqrt{\frac{d\kappa_w\iota}{T_{\mathsf{total}}^{(k)}}}, \quad \frac{\|\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(k)}\|}{\|\boldsymbol{W}^{(k)}\|} \le C_3\sqrt{\frac{d\iota}{T_{\mathsf{total}}^{(k)}}}, \quad 1 \le k \le K.$$

Our proof (postponed to Appendix C.4) is based on the techniques of (Simchowitz et al., 2018; Sarkar & Rakhlin, 2019), but with two major differences. First, the authors of (Simchowitz et al., 2018; Sarkar & Rakhlin, 2019) consider the setting where ordinary least squares is applied to a single continuous trajectory generated by a single LDS model; this is not the case for our setting, and thus our proof and results are different from theirs. Second, the noise covariance matrix $\boldsymbol{W}$ is assumed to be $\sigma^2\boldsymbol{I}_d$ in (Simchowitz et al., 2018; Sarkar & Rakhlin, 2019), while in our case, $\{\boldsymbol{W}^{(k)}\}_{1\le k\le K}$ are unknown and need to be estimated.

Classification. Our last theorem shows that Algorithm 5 correctly classifies all trajectories in $\mathcal{M}_{\mathsf{classification}}$, as long as the coarse models are sufficiently accurate and the short trajectory lengths are large enough; these conditions are slightly different for Cases 0 and 1 defined in (8). See Appendix C.5 for the proof.

Theorem B.4. *Consider the model (1) under the assumptions in Section 3.1. There exist universal constants $c_1, c_2, C_3 > 0$ such that the following holds. Suppose that we run Algorithm 5 with data $\{\boldsymbol{X}_m\}_{m\in\mathcal{M}_{\mathsf{classification}}}$ and independent coarse models $\{\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}\}_{1\le k\le K}$ satisfying $\|\widehat{\boldsymbol{A}}^{(k)} - \boldsymbol{A}^{(k)}\| \le \epsilon_A, \|\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(k)}\| \le \epsilon_W$ for all k. Then with probability at least $1-\delta$, Algorithm 5 correctly classifies all trajectories in $\mathcal{M}_{\mathsf{classification}}$, provided that*

$$\textit{For Case 0:} \quad \epsilon_A \le c_1\Delta_{A,W}\sqrt{\frac{W_{\mathsf{min}}}{\Gamma_{\mathsf{max}}\kappa_{w,\mathsf{cross}}(d+\iota)}}, \quad \frac{\epsilon_W}{W_{\mathsf{min}}} \le c_2\cdot\min\Big\{1, \frac{\Delta_{A,W}}{\sqrt{\kappa_{w,\mathsf{cross}}d}}\Big\}, \tag{26a}$$

$$T_m \ge C_3\Big(\kappa_w^2 + \frac{\kappa_{w,\mathsf{cross}}^6}{\Delta_{A,W}^2}\Big)\iota^2, \quad m \in \mathcal{M}_{\mathsf{classification}}; \tag{26b}$$

$$\textit{For Case 1:} \quad \epsilon_A \le c_1\Delta_{A,W}\sqrt{\frac{W_{\mathsf{min}}}{\Gamma_{\mathsf{max}}\kappa_{w,\mathsf{cross}}\kappa_A^2(d+\iota)}}, \quad \frac{\epsilon_W}{W_{\mathsf{min}}} \le c_2\cdot\min\Big\{1, \frac{\Delta_{A,W}}{\sqrt{\kappa_{w,\mathsf{cross}}d}}\Big\}, \tag{26c}$$

$$T_m \ge C_3\Big(\kappa_w^2 + \frac{\kappa_{w,\mathsf{cross}}^6}{\Delta_{A,W}^2}\Big)\iota^2 + \frac{1}{1-\rho}\log(2\kappa_A), \quad m \in \mathcal{M}_{\mathsf{classification}}, \tag{26d}$$

where $\iota := \log\frac{T_{\mathsf{total}}}{\delta}$ is a logarithmic term.

Proof of Theorems 3.4 and 3.5. Our main theorems are direct implications of the above guarantees for the individual steps.

- *Stage 1*: To begin with, according to Theorem B.1, if the condition (9a) on T_{subspace} and $T_{\mathsf{total,subspace}}$ hold, then the subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}$ output by Line 3 of Algorithm 1 are guaranteed to satisfy the error bounds (25) required by Theorem B.2. This together with the condition (9b) on $T_{\mathsf{clustering}}$ ensures exact clustering in Line 4.
- *Stage 2*: Based on this, if we further know that $T_{\mathsf{total,clustering}}$ obeys condition (9b) for Case 0 or (11a) for Case 1, then Theorem B.3 tells us that the coarse model estimation in Line 5 satisfies the error bounds (26a) or (26c) required by Theorem B.4. Together with the assumption (9c) or (11b) on $T_{\mathsf{classification}}$, this guarantees exact classification in Line 7 of Algorithm 1, for either Case 0 or 1. At the end, the final model estimation errors (10) follow immediately from Theorem B.3.

Note that all the statements above are high-probability guarantees; it suffices to take the union bound over these steps, so that the performance guarantees of Theorems 3.4 and 3.5 hold with probability at least $1-\delta$. This finishes the proof of our main theorems.

Remark B.5. The step of subspace estimation in Algorithm 1 is non-essential and optional; it allows for a smaller $T_{\mathsf{clustering}}$, but comes at the price of complicating the overall algorithm. For practitioners who prefer a simpler algorithm, they might simply remove this step (i.e. Line 3 of Algorithm 1), and replace the rank-K subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}$ with $\boldsymbol{I}_d$ (i.e. no dimensionality reduction for the clustering step). The theoretical guarantees continue to hold with minor modification: in Corollary 3.7, one simply needs to remove the conditions on $T_{\mathsf{subspace}}, T_{\mathsf{total,subspace}}$, and in the condition for $T_{\mathsf{clustering}}$, replace the factor $\sqrt{K/d}$ (where K is due to dimensionality reduction) with $\sqrt{d/d}=1$. This is one example regarding how our modular algorithms and theoretical analysis can be easily modified and adapted to accommodate different situations.

C. Proofs for Appendix B

This section starts with some preliminaries about linear dynamical systems that will be helpful later. Then, it provides the main proofs for the theorems in Appendix B.

C.1. Preliminaries

Truncation of autocovariance. Recall the notation $\boldsymbol{\Gamma}^{(k)} = \sum_{i=0}^{\infty} \boldsymbol{A}^{(k)^i} \boldsymbol{W}^{(k)} (\boldsymbol{A}^{(k)^i})^\top$ from (3) and (5). We add a subscript t to represent its t-step truncation:

$$\boldsymbol{\Gamma}_t^{(k)} \coloneqq \sum_{i=0}^{t-1} \boldsymbol{A}^{(k)^i} \boldsymbol{W}^{(k)} (\boldsymbol{A}^{(k)^i})^\top. \tag{27}$$

Also recall the assumption of exponential stability in Assumption 3.1, namely $\|(\boldsymbol{A}^{(k)})^t\| \le \kappa_A \rho^t$. As a result, $\boldsymbol{\Gamma}_t^{(k)}$ is close to $\boldsymbol{\Gamma}^{(k)}$:

$$\begin{aligned} \mathbf{0} \preccurlyeq \boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}_t^{(k)} &= \sum_{i=t}^{\infty} \boldsymbol{A}^{(k)^i} \boldsymbol{W}^{(k)} (\boldsymbol{A}^{(k)^i})^\top = \boldsymbol{A}^{(k)^t} \boldsymbol{\Gamma}^{(k)} (\boldsymbol{A}^{(k)^t})^\top, \\ \|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}_t^{(k)}\| &\le \|\boldsymbol{\Gamma}^{(k)}\| \|\boldsymbol{A}^{(k)^t}\|^2 \le \|\boldsymbol{\Gamma}^{(k)}\| \kappa_A^2 \rho^{2t} \le \Gamma_{\mathsf{max}} \kappa_A^2 \rho^{2t}. \end{aligned} \tag{28}$$

Moreover, let $\boldsymbol{Y}_t^{(k)} \coloneqq \boldsymbol{A}^{(k)} \boldsymbol{\Gamma}_t^{(k)}$, then $\boldsymbol{Y}_t^{(k)}$ is also close to $\boldsymbol{Y}^{(k)}$:

$$\|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}_t^{(k)}\| \le \|\boldsymbol{A}^{(k)}\| \|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}_t^{(k)}\| \le \Gamma_{\mathsf{max}} \kappa_A^3 \rho^{2t}. \tag{29}$$

"Independent version" of states. Given some mixing time t_{mix}, we define $\widetilde{\boldsymbol{x}}_{m,t}(t_{\mathsf{mix}})$ as the $(t_{\mathsf{mix}}-1)$-step approximation of $\boldsymbol{x}_{m,t}$:

$$\widetilde{\boldsymbol{x}}_{m,t} = \widetilde{\boldsymbol{x}}_{m,t}(t_{\mathsf{mix}}) \coloneqq \sum_{i=0}^{t_{\mathsf{mix}}-2} (\boldsymbol{A}^{(k_m)})^i \boldsymbol{w}_{m,t-i-1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k_m)}), \quad t_{\mathsf{mix}} \le t \le T_m, \quad 1 \le m \le M. \tag{30}$$

Since $\widetilde{\boldsymbol{x}}_{m,t}$ consists of only the most recent noise vectors, it is independent of the history up to $\boldsymbol{x}_{m,t-t_{\mathsf{mix}}+1}$. Our proofs for Stage 1 will rely on this "independent version" $\{\widetilde{\boldsymbol{x}}_{m,t}\}$ of states $\{\boldsymbol{x}_{m,t}\}$; the basic idea is that, for an appropriately chosen t_{mix}, a trajectory of length T can be regarded as a collection of T/t_{mix} independent samples. We will often use the notation $\widetilde{\cdot}$ to represent the "independent version" of other variables as well.

Boundedness of states. The following lemma provides upper bounds for $\{\|\widetilde{\boldsymbol{x}}_{m,t}\|_2\}$ and $\{\|\boldsymbol{x}_{m,t}\|_2\}$. This will help to control the effects of mixing errors and model estimation errors in the analyses later.

Lemma C.1 (Bounded states)**.** *Consider the model (1) under the assumptions in Sections 2 and 3.1. Fix any $t_{\mathsf{mix}} \geq 3$. Then with probability at least $1-\delta$, we have $\|\widetilde{\boldsymbol{x}}_{m,t}\|_2 \leq C_0\sqrt{\Gamma_{\mathsf{max}}(d+\log(T_{\mathsf{total}}/\delta))}$ for all $1 \leq m \leq M, t_{\mathsf{mix}} \leq t \leq T_m$, where $C_0 > 0$ is some universal constant; moreover, for both cases of initial states defined in (8), all states $\{\{\boldsymbol{x}_{m,t}\}_{0\leq t\leq T_m}\}_{1\leq m\leq M}$ are bounded throughout:*

- *Case 0: with probability at least $1-\delta$, we have $\|\boldsymbol{x}_{m,t}\|_2 \leq C_0\sqrt{\Gamma_{\mathsf{max}}(d+\log(T_{\mathsf{total}}/\delta))}$ for all m, t.*
- *Case 1: suppose that $t_{\mathsf{mix}} \geq \frac{1}{1-\rho}\log(\sqrt{2}\kappa_A)$, and $T_m \geq t_{\mathsf{mix}}$ for all m, then with probability at least $1-\delta$, we have $\|\boldsymbol{x}_{m,t}\|_2 \leq 3C_0\kappa_A\sqrt{\Gamma_{\mathsf{max}}(d+\log(T_{\mathsf{total}}/\delta))}$ for all m, t, and $\|\boldsymbol{x}_{m,t}\|_2 \leq 2C_0\sqrt{\Gamma_{\mathsf{max}}(d+\log(T_{\mathsf{total}}/\delta))}$ for all $t \geq t_{\mathsf{mix}}$ or $t = 0$.*

Proof. First, recall from Corollary 7.3.3 of (Vershynin, 2018) that, if random vector $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, then for all $u \geq 0$, we have $\mathbb{P}(\|\boldsymbol{a}\|_2 \geq 2\sqrt{d}+u) \leq 2\exp(-cu^2)$. Since $\widetilde{\boldsymbol{x}}_{m,t} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}^{(k_m)}_{t_{\mathsf{mix}}-1})$, where $\boldsymbol{\Gamma}^{(k_m)}_{t_{\mathsf{mix}}-1} \preccurlyeq \boldsymbol{\Gamma}^{(k_m)} \preccurlyeq \Gamma_{\mathsf{max}}\boldsymbol{I}_d$, we have $\mathbb{P}(\|\widetilde{\boldsymbol{x}}_{m,t}\|_2 \geq \sqrt{\Gamma_{\mathsf{max}}}(2\sqrt{d}+u)) \leq 2\exp(-cu^2)$. Taking the union bound, we have

$$\begin{aligned}
&\mathbb{P}\Big(\text{there exists } m,t \text{ such that } \|\widetilde{\boldsymbol{x}}_{m,t}\|_2 \geq \sqrt{\Gamma_{\mathsf{max}}}(2\sqrt{d}+u)\Big) \\
&\quad\leq \sum_{m=1}^{M}\sum_{t=t_{\mathsf{mix}}}^{T_m}\mathbb{P}\Big(\|\widetilde{\boldsymbol{x}}_{m,t}\|_2 \geq \sqrt{\Gamma_{\mathsf{max}}}(2\sqrt{d}+u)\Big) \leq 2T_{\mathsf{total}}\exp(-cu^2) \leq \delta,
\end{aligned}$$

where the last inequality holds if we pick $u \geq \sqrt{\frac{1}{c}\log\frac{2T_{\mathsf{total}}}{\delta}}$. This finishes the proof of the first claim in the lemma. Next, we prove the boundedness of $\{\|\boldsymbol{x}_{m,t}\|_2\}$.

- *Case 0*: It is easy to check that $\boldsymbol{x}_{m,t} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}^{(k_m)}_t)$, where $\boldsymbol{\Gamma}^{(k_m)}_t \preccurlyeq \boldsymbol{\Gamma}^{(k_m)} \preccurlyeq \Gamma_{\mathsf{max}}\boldsymbol{I}_d$. The boundedness of $\{\|\boldsymbol{x}_{m,t}\|_2\}$ can be proved by a similar argument as before, which we omit for brevity.
- *Case 1*: Define $\xi_{m,t} := \boldsymbol{x}_{m,t} - (\boldsymbol{A}^{(k_m)})^t\boldsymbol{x}_{m,0} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}^{(k_m)}_t)$. By a similar argument as before, we have with probability at least $1-\delta$, $\|\xi_{m,t}\|_2 \leq C_0\sqrt{\Gamma_{\mathsf{max}}(d+\log(T_{\mathsf{total}}/\delta))}$ for all m, t. Morever, for any $t \geq t_{\mathsf{mix}} \geq \frac{1}{1-\rho}\log(\sqrt{2}\kappa_A)$, we have $\|(\boldsymbol{A}^{(k_m)})^t\| \leq \kappa_A\rho^t \leq 1/2$. With this in place, we have

$$\begin{aligned}
&\boldsymbol{x}_{m+1,0} = \boldsymbol{x}_{m,T_m} = (\boldsymbol{A}^{(k_m)})^{T_m}\boldsymbol{x}_{m,0} + \xi_{m,T_m}, \quad \text{and thus} \\
&\|\boldsymbol{x}_{m+1,0}\| \leq \|(\boldsymbol{A}^{(k_m)})^{T_m}\|\|\boldsymbol{x}_{m,0}\|_2 + \|\xi_{m,T_m}\|_2 \leq \frac{1}{2}\|\boldsymbol{x}_{m,0}\|_2 + C_0\sqrt{\Gamma_{\mathsf{max}}(d+\log\frac{T_{\mathsf{total}}}{\delta})}.
\end{aligned}$$

  Recall the assumption that $\boldsymbol{x}_{1,0} = \boldsymbol{0}$; by induction, we have $\|\boldsymbol{x}_{m,0}\|_2 \leq 2C_0\sqrt{\Gamma_{\mathsf{max}}(d+\log(T_{\mathsf{total}}/\delta))}$ for all $1 \leq m \leq M$. Now, we have for all m, t,

$$\begin{aligned}
\|\boldsymbol{x}_{m,t}\|_2 &\leq \|(\boldsymbol{A}^{(k_m)})^t\boldsymbol{x}_{m,0}\|_2 + \|\xi_{m,t}\|_2 \\
&\leq \kappa_A\|\boldsymbol{x}_{m,0}\|_2 + C_0\sqrt{\Gamma_{\mathsf{max}}(d+\log\frac{T_{\mathsf{total}}}{\delta})} \leq 3C_0\kappa_A\sqrt{\Gamma_{\mathsf{max}}(d+\log\frac{T_{\mathsf{total}}}{\delta})};
\end{aligned}$$

  moreover, for $t \geq t_{\mathsf{mix}}$, since $\|(\boldsymbol{A}^{(k_m)})^t\| \leq 1/2$, we obtain a better bound

$$\begin{aligned}
\|\boldsymbol{x}_{m,t}\|_2 &\leq \|(\boldsymbol{A}^{(k_m)})^t\boldsymbol{x}_{m,0}\|_2 + \|\xi_{m,t}\|_2 \\
&\leq \frac{1}{2}\|\boldsymbol{x}_{m,0}\|_2 + C_0\sqrt{\Gamma_{\mathsf{max}}(d+\log\frac{T_{\mathsf{total}}}{\delta})} \leq 2C_0\sqrt{\Gamma_{\mathsf{max}}(d+\log\frac{T_{\mathsf{total}}}{\delta})}.
\end{aligned}$$

This shows the boundedness of $\{\|\boldsymbol{x}_{m,t}\|_2\}$ and completes our proof of the lemma. □

C.2. Proof of Theorem B.1

Theorem B.1 is an immediate consequence of Lemmas C.2 and C.3 below. The former shows the concentration of $\widehat{\boldsymbol{H}}_i, \widehat{\boldsymbol{G}}_i$ around the targeted low-rank matrices $\boldsymbol{H}_i, \boldsymbol{G}_i$, while the latter is a result of perturbation analysis.

Lemma C.2. *Under the setting of Theorem B.1, with probability at least $1-\delta$, we have for all $1 \le i \le d$,*

$$\max\Big\{\|\widehat{\boldsymbol{H}}_i - \boldsymbol{H}_i\|, \|\widehat{\boldsymbol{G}}_i - \boldsymbol{G}_i\|\Big\} \lesssim \Gamma_{\mathsf{max}}^2 \sqrt{\frac{t_{\mathsf{mix}} d}{T_{\mathsf{total,subspace}}} \log^3 \frac{T_{\mathsf{total}} d}{\delta}}.$$

Lemma C.3. *Consider the matrix $\boldsymbol{M}_\star = \sum_{k=1}^K p^{(k)} \boldsymbol{y}^{(k)} \boldsymbol{y}^{(k)\top} \in \mathbb{R}^{d\times d}$, where $0 < p^{(k)} < 1, \sum_{k=1}^K p^{(k)} = 1$, and $\boldsymbol{y}^{(k)} \in \mathbb{R}^d$. Let $\boldsymbol{M}$ be symmetric and satisfy $\|\boldsymbol{M} - \boldsymbol{M}_\star\| \le \epsilon$, and $\boldsymbol{U} \in \mathbb{R}^{d\times K}$ be the top-K eigenspace of $\boldsymbol{M}$. Then we have*

$$\sum_{k=1}^K p^{(k)} \|\boldsymbol{y}^{(k)} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{y}^{(k)}\|_2^2 \le 2K\epsilon, \tag{31}$$

and for all $1 \le k \le K$, it holds that

$$\|\boldsymbol{y}^{(k)} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{y}^{(k)}\|_2 \le \min\left\{\left(\frac{2K\epsilon}{p^{(k)}}\right)^{1/2}, \frac{2\epsilon}{\lambda_{\mathsf{min}}(\boldsymbol{M})}\|\boldsymbol{y}^{(k)}\|_2, \sqrt{2}\left(\frac{\epsilon}{p^{(k)}}\|\boldsymbol{y}^{(k)}\|_2\right)^{1/3}\right\}. \tag{32}$$

For our main analyses in Sections 3.2 and B, we choose the first term on the right-hand side of (32).

C.2.1. Proof of Lemma C.2

We first analyze the idealized case with i.i.d. samples; then we make a connection between this i.i.d. case and the actual case of mixed LDSs, by utilizing the mixing property of linear dynamical systems. We prove the result of Lemma C.2 for $\|\widehat{\boldsymbol{G}}_i - \boldsymbol{G}_i\|$ only, since the analysis for $\|\widehat{\boldsymbol{H}}_i - \boldsymbol{H}_i\|$ is mostly the same (and simpler in fact).

Step 1: the idealized i.i.d. case. With some abuse of notation, suppose that for all $1 \le m \le M$, we have for some $k_m \in \{1, \dots, K\}$,

$$\begin{aligned}
&\boldsymbol{x}_{m,t}, \boldsymbol{z}_{m,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{\Gamma}}^{(k_m)}), \quad \boldsymbol{w}_{m,t}, \boldsymbol{v}_{m,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}^{(k_m)}),\\
&\boldsymbol{x}_{m,t}{}' = \boldsymbol{A}^{(k_m)} \boldsymbol{x}_{m,t} + \boldsymbol{w}_{m,t}, \quad \boldsymbol{z}_{m,t}{}' = \boldsymbol{A}^{(k_m)} \boldsymbol{z}_{m,t} + \boldsymbol{v}_{m,t}, \quad 1 \le t \le N,
\end{aligned}$$

where for all $1 \le k \le K$, it holds that $\boldsymbol{W}^{(k)}, \widetilde{\boldsymbol{\Gamma}}^{(k)} \preccurlyeq \boldsymbol{\Gamma}^{(k)} \preccurlyeq \Gamma_{\mathsf{max}} \boldsymbol{I}_d$. Notice that $\mathsf{cov}(\boldsymbol{x}_{m,t}{}') = \boldsymbol{A}^{(k_m)} \widetilde{\boldsymbol{\Gamma}}^{(k_m)} (\boldsymbol{A}^{(k_m)})^\top + \boldsymbol{W}^{(k_m)} \preccurlyeq \boldsymbol{A}^{(k_m)} \boldsymbol{\Gamma}^{(k_m)} (\boldsymbol{A}^{(k_m)})^\top + \boldsymbol{W}^{(k_m)} = \boldsymbol{\Gamma}^{(k_m)} \preccurlyeq \Gamma_{\mathsf{max}} \boldsymbol{I}_d$. Consider the i.i.d. version of matrix $\widehat{\boldsymbol{G}}_i$ and its expectation $\boldsymbol{G}_i$ defined as follows:

$$\widehat{\boldsymbol{G}}_i \coloneqq \frac{1}{MN} \sum_{m=1}^M \sum_{t=1}^N \Big(\big(\boldsymbol{x}_{m,t}{}'\big)_i \boldsymbol{x}_{m,t}\Big)\Big(\big(\boldsymbol{z}_{m,t}{}'\big)_i \boldsymbol{z}_{m,t}\Big)^\top, \quad \boldsymbol{G}_i = \sum_{k=1}^K p^{(k)} (\widetilde{\boldsymbol{Y}}^{(k)})_i (\widetilde{\boldsymbol{Y}}^{(k)})_i^\top,$$

where $(\widetilde{\boldsymbol{Y}}^{(k)})_i$ is the transpose of the i-th row of $\widetilde{\boldsymbol{Y}}^{(k)} \coloneqq \boldsymbol{A}^{(k)} \widetilde{\boldsymbol{\Gamma}}^{(k)}$. For this i.i.d. setting, we claim that, if the i.i.d. sample size MN satisfies $MN \gtrsim d \cdot \log(MNd/\delta)$, then with probability at least $1-\delta$,

$$\|\widehat{\boldsymbol{G}}_i - \boldsymbol{G}_i\| \lesssim \Gamma_{\mathsf{max}}^2 \sqrt{\frac{d}{MN} \log^3 \frac{MNd}{\delta}}, \quad 1 \le i \le d. \tag{33}$$

This claim can be proved by a standard covering argument with truncation; we will provide a proof later for completeness.

Step 2: back to the actual case of mixed LDSs. Now we turn to the analysis of $\widehat{\boldsymbol{G}}_i$ defined in (14) versus its expectation $\boldsymbol{G}_i$ defined in (15b), for the mixed LDSs setting. We first show that $\widehat{\boldsymbol{G}}_i$ can be writte as a *weighted average* of some matrices, each of which can be further decomposed into an i.i.d. *part* (as in Step 1) plus a *negligible mixing error term*. Then we analyze each term in the decomposition, and finally put pieces together to show that $\widehat{\boldsymbol{G}}_i \approx \boldsymbol{G}_i$.

Step 2.1: decomposition of the index set $\Omega_1 \times \Omega_2$. Recall the definition of index sets Ω_1, Ω_2 in Algorithm 2. Denote the first index of Ω_1 (resp. Ω_2) as $\tau_1 + 1$ (resp. $\tau_2 + 1$), and let $\Delta \coloneqq \tau_2 - \tau_1$ be their distance. Also denote $N \coloneqq |\Omega_1| = |\Omega_2| \asymp T_{\mathsf{subspace}}$. For any $t \in \Omega_1$ and $1 \le j \le N$, define

$$s_j(t) \coloneqq \mathsf{Cycle}(t + \Delta + j; \Omega_2) = \begin{cases} t + \Delta + j & \text{if} \quad t + \Delta + j \le \tau_2 + N, \\ t + \Delta + j - N & \text{otherwise,} \end{cases}$$

where $\mathsf{Cycle}(i; \Omega)$ represents the cyclic indexing of value i on set Ω. Then we have

$$\Omega_1 \times \Omega_2 = \Big\{(t_1, t_2), t_1 \in \Omega_1, t_2 \in \Omega_2\Big\} = \cup_{j=1}^{N} \Big\{\big(t, s_j(t)\big), t \in \Omega_1\Big\}.$$

We further define

$$\mathcal{S}_\tau \coloneqq \{\tau_1 + \tau + f \cdot t_{\mathsf{mix}} : f \ge 0, \tau + f \cdot t_{\mathsf{mix}} \le N\}, \quad 1 \le \tau \le t_{\mathsf{mix}},$$

so that $\Omega_1 = \cup_{\tau=1}^{t_{\mathsf{mix}}} \mathcal{S}_\tau$. Notice that for each τ, the elements of $\mathcal{S}_\tau$ are at least t_{mix} far apart. Putting together, we have

$$\Omega_1 \times \Omega_2 = \cup_{j=1}^{N} \cup_{\tau=1}^{t_{\mathsf{mix}}} \Big\{\big(t, s_j(t)\big), t \in \mathcal{S}_\tau\Big\}. \tag{34}$$

Step 2.2: decomposition of $\widehat{\boldsymbol{G}}_i$. In the remaining proof, we denote $\boldsymbol{x}_{m,t}{}' \coloneqq \boldsymbol{x}_{m,t+1}$ for notational consistency. Using the decomposition (34) of $\Omega_1 \times \Omega_2$, we can rewrite $\widehat{\boldsymbol{G}}_i$ defined in (14) as a weighted average of $N t_{\mathsf{mix}}$ matrices:

$$\widehat{\boldsymbol{G}}_i = \frac{1}{|\mathcal{M}_{\mathsf{subspace}}|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} \frac{1}{N^2} \sum_{(t_1, t_2) \in \Omega_1 \times \Omega_2} (\boldsymbol{x}'_{m,t_1})_i \, \boldsymbol{x}_{m,t_1} \cdot (\boldsymbol{x}'_{m,t_2})_i \, \boldsymbol{x}_{m,t_2}{}^\top = \sum_{j=1}^{N} \sum_{\tau=1}^{t_{\mathsf{mix}}} \frac{|\mathcal{S}_\tau|}{N^2} \cdot \boldsymbol{F}_{i,j,\tau},$$

where

$$\boldsymbol{F}_{i,j,\tau} \coloneqq \frac{1}{|\mathcal{M}_{\mathsf{subspace}}| \cdot |\mathcal{S}_\tau|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} \sum_{t \in \mathcal{S}_\tau} (\boldsymbol{x}_{m,t}{}')_i \, \boldsymbol{x}_{m,t} \cdot (\boldsymbol{x}'_{m,s_j(t)})_i \, \boldsymbol{x}_{m,s_j(t)}{}^\top. \tag{35}$$

Recalling the definition of $\widetilde{\boldsymbol{x}}_{m,t}$ in (30) and $\boldsymbol{\Gamma}_t^{(k)}$ in (27), we have

$$\boldsymbol{x}_{m,t} = \widetilde{\boldsymbol{x}}_{m,t} + \underbrace{(\boldsymbol{A}^{(k_m)})^{t_{\mathsf{mix}}-1} \boldsymbol{x}_{m,t-t_{\mathsf{mix}}+1}}_{=:\boldsymbol{\delta}_{m,t}} = \widetilde{\boldsymbol{x}}_{m,t} + \boldsymbol{\delta}_{m,t},$$

where

$$\|\boldsymbol{\delta}_{m,t}\|_2 \le \|(\boldsymbol{A}^{(k_m)})^{t_{\mathsf{mix}}-1}\| \cdot \|\boldsymbol{x}_{m,t-t_{\mathsf{mix}}+1}\|_2 \le \kappa_A \rho^{t_{\mathsf{mix}}-1} \|\boldsymbol{x}_{m,t-t_{\mathsf{mix}}+1}\|_2, \tag{36}$$

and $\widetilde{\boldsymbol{x}}_{m,t} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)})$ is independent of $\boldsymbol{\delta}_{m,t}$. Moreover,

$$\boldsymbol{x}_{m,t}{}' = \boldsymbol{A}^{(k_m)} \boldsymbol{x}_{m,t} + \boldsymbol{w}_{m,t} = \widetilde{\boldsymbol{x}}'_{m,t} + \boldsymbol{A}^{(k_m)} \boldsymbol{\delta}_{m,t}, \quad \text{where} \quad \widetilde{\boldsymbol{x}}'_{m,t} \coloneqq \boldsymbol{A}^{(k_m)} \widetilde{\boldsymbol{x}}_{m,t} + \boldsymbol{w}_{m,t}.$$

We can rewrite $\boldsymbol{x}_{m,s_j(t)} = \widetilde{\boldsymbol{x}}_{m,s_j(t)} + \boldsymbol{\delta}_{m,s_j(t)}$ and $\boldsymbol{x}'_{m,s_j(t)} = \widetilde{\boldsymbol{x}}'_{m,s_j(t)} + \boldsymbol{A}^{(k_m)} \boldsymbol{\delta}_{m,s_j(t)}$ in a similar manner. Putting this back to (35), one has

$$\begin{aligned}
\boldsymbol{F}_{i,j,\tau} &= \frac{1}{|\mathcal{M}_{\mathsf{subspace}}| \cdot |\mathcal{S}_\tau|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} \sum_{t \in \mathcal{S}_\tau} (\boldsymbol{x}_{m,t}{}')_i \, \boldsymbol{x}_{m,t} \cdot (\boldsymbol{x}'_{m,s_j(t)})_i \, \boldsymbol{x}_{m,s_j(t)}{}^\top \\
&= \frac{1}{|\mathcal{M}_{\mathsf{subspace}}| \cdot |\mathcal{S}_\tau|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} \sum_{t \in \mathcal{S}_\tau} \\
&\qquad (\widetilde{\boldsymbol{x}}'_{m,t} + \boldsymbol{A}^{(k_m)} \boldsymbol{\delta}_{m,t})_i \, (\widetilde{\boldsymbol{x}}_{m,t} + \boldsymbol{\delta}_{m,t}) \cdot (\widetilde{\boldsymbol{x}}'_{m,s_j(t)} + \boldsymbol{A}^{(k_m)} \boldsymbol{\delta}_{m,s_j(t)})_i \, (\widetilde{\boldsymbol{x}}_{m,s_j(t)} + \boldsymbol{\delta}_{m,s_j(t)})^\top \\
&= \underbrace{\frac{1}{|\mathcal{M}_{\mathsf{subspace}}| \cdot |\mathcal{S}_\tau|} \sum_{m \in \mathcal{M}_{\mathsf{subspace}}} \sum_{t \in \mathcal{S}_\tau} (\widetilde{\boldsymbol{x}}'_{m,t})_i \, \widetilde{\boldsymbol{x}}_{m,t} \cdot (\widetilde{\boldsymbol{x}}'_{m,s_j(t)})_i \, \widetilde{\boldsymbol{x}}_{m,s_j(t)}{}^\top}_{=:\widetilde{\boldsymbol{F}}_{i,j,\tau}} + \boldsymbol{\Delta}_{i,j,\tau},
\end{aligned} \tag{37}$$

where $\boldsymbol{\Delta}_{i,j,\tau}$ contains all the $\{\boldsymbol{\delta}_{m,t}\}$ terms in the expansion. The key observation here is that, by our definition of index set $\mathcal{S}_\tau$, the $\{\widetilde{\boldsymbol{x}}_{m,t}\}$ terms in $\widetilde{\boldsymbol{F}}_{i,j,\tau}$ are independent, and thus we can utilize our earlier analysis of the i.i.d. case in Step 1 to study $\widetilde{\boldsymbol{F}}_{i,j,\tau}$.

Step 2.3: analysis for each term of the decomposition. Towards showing $\widehat{\boldsymbol{G}}_i \approx \boldsymbol{G}_i$, we prove in the following that $\widetilde{\boldsymbol{F}}_{i,j,\tau}$ concentrates around its expectation $\widetilde{\boldsymbol{G}}_i$, which is close to $\boldsymbol{G}_i$; moreover, the error term $\boldsymbol{\Delta}_{i,j,\tau}$ becomes exponentially small as t_{mix} grows. More specifically, we have the following:

- Recall the notation $\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1} = \boldsymbol{A}^{(k)}\boldsymbol{\Gamma}^{(k)}_{t_{\mathsf{mix}}-1}$. It holds that

$$\mathbb{E}[\widetilde{\boldsymbol{F}}_{i,j,\tau}] = \widetilde{\boldsymbol{G}}_i \coloneqq \sum_{k=1}^{K} p^{(k)} (\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1})_i (\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1})_i^\top.$$

  According to our result (33) for the i.i.d. case, for fixed j, τ, we have with probability at least $1-\delta$,

$$\begin{aligned} 1 \le i \le d, \quad \|\widetilde{\boldsymbol{F}}_{i,j,\tau} - \widetilde{\boldsymbol{G}}_i\| &\lesssim \Gamma^2_{\mathsf{max}} \sqrt{\frac{d}{|\mathcal{M}_{\mathsf{subspace}}| \cdot |\mathcal{S}_\tau|} \log^3 \frac{|\mathcal{M}_{\mathsf{subspace}}| \cdot |\mathcal{S}_\tau| d}{\delta}} \\ &\lesssim \Gamma^2_{\mathsf{max}} \sqrt{\frac{t_{\mathsf{mix}} d}{T_{\mathsf{total,subspace}}} \log^3 \frac{T_{\mathsf{total,subspace}} d}{\delta}}. \end{aligned} \tag{38}$$

- Recall from (29) that $\|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1}\| \le \Gamma_{\mathsf{max}} \kappa_A^3 \rho^{2(t_{\mathsf{mix}}-1)}$. Therefore,

$$\begin{aligned} &\big\|(\boldsymbol{Y}^{(k)})_i (\boldsymbol{Y}^{(k)})_i^\top - (\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1})_i (\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1})_i^\top\big\| \\ &\quad \le \big(\|(\boldsymbol{Y}^{(k)})_i\|_2 + \|(\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1})_i\|_2\big) \|(\boldsymbol{Y}^{(k)})_i - (\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1})_i\|_2 \\ &\quad \le \big(2\|\boldsymbol{Y}^{(k)}\| + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1}\|\big) \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1}\| \\ &\quad \le \big(2\Gamma_{\mathsf{max}}\kappa_A + \Gamma_{\mathsf{max}}\kappa_A^3 \rho^{2(t_{\mathsf{mix}}-1)}\big) \cdot \Gamma_{\mathsf{max}} \kappa_A^3 \rho^{2(t_{\mathsf{mix}}-1)} \\ &\quad = \big(2 + \kappa_A^2 \rho^{2(t_{\mathsf{mix}}-1)}\big) \cdot \Gamma^2_{\mathsf{max}} \kappa_A^4 \rho^{2(t_{\mathsf{mix}}-1)} \le 3\Gamma^2_{\mathsf{max}} \kappa_A^4 \rho^{2(t_{\mathsf{mix}}-1)}, \end{aligned}$$

  where we use the mild assumption that $t_{\mathsf{mix}} \ge 1 + \frac{\log \kappa_A}{1-\rho}$, and the fact that $\|\boldsymbol{Y}^{(k)}\|, \|\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1}\| \le \Gamma_{\mathsf{max}}\kappa_A$. Consequently,

$$\|\boldsymbol{G}_i - \widetilde{\boldsymbol{G}}_i\| \le \sum_{k=1}^{K} p^{(k)} \Big\|(\boldsymbol{Y}^{(k)})_i (\boldsymbol{Y}^{(k)})_i^\top - (\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1})_i (\boldsymbol{Y}^{(k)}_{t_{\mathsf{mix}}-1})_i^\top\Big\| \le 3\Gamma^2_{\mathsf{max}} \kappa_A^4 \rho^{2(t_{\mathsf{mix}}-1)}. \tag{39}$$

- By Lemma C.1, if $t_{\mathsf{mix}} \gtrsim \frac{1}{1-\rho}\log(2\kappa_A)$, then we have with probability at least $1-\delta$, all the $\boldsymbol{x}_{m,t}$'s and $\widetilde{\boldsymbol{x}}_{m,t}$'s involved in the definition of $\boldsymbol{\Delta}_{i,j,\tau}$ in (37) have ℓ_2 norm bounded by $\sqrt{\Gamma_{\mathsf{max}}}\mathsf{poly}(d, \kappa_A, \log(T_{\mathsf{total}}/\delta))$. This together with the upper bound on $\|\boldsymbol{\delta}_{m,t}\|_2$ in (36) implies that for all i, j, τ, it holds that

$$\|\boldsymbol{\Delta}_{i,j,\tau}\| \le \Gamma^2_{\mathsf{max}} \cdot \mathsf{poly}\Big(d, \kappa_A, \log \frac{T_{\mathsf{total}}}{\delta}\Big) \cdot \rho^{t_{\mathsf{mix}}-1}. \tag{40}$$

Step 2.4: putting pieces together. With (38), (39) and (40) in place and taking the union bound, we have with probability at least $1-\delta$, for all $1 \le i \le d$,

$$\begin{aligned} \|\widehat{\boldsymbol{G}}_i - \boldsymbol{G}_i\| = \Big\|\sum_{j=1}^{N}\sum_{\tau=1}^{t_{\mathsf{mix}}} \frac{|\mathcal{S}_\tau|}{N^2} \cdot \boldsymbol{F}_{i,j,\tau} - \boldsymbol{G}_i\Big\| &\le \max_{j,\tau} \|\widetilde{\boldsymbol{F}}_{i,j,\tau} - \widetilde{\boldsymbol{G}}_i\| + \|\widetilde{\boldsymbol{G}}_i - \boldsymbol{G}_i\| + \max_{j,\tau}\|\boldsymbol{\Delta}_{i,j,\tau}\| \\ &\lesssim \Gamma^2_{\mathsf{max}} \sqrt{\frac{t_{\mathsf{mix}} d}{T_{\mathsf{total,subspace}}} \log^3 \frac{T_{\mathsf{total,subspace}} d}{\delta}} + \Gamma^2_{\mathsf{max}} \kappa_A^4 \rho^{2(t_{\mathsf{mix}}-1)} + \Gamma^2_{\mathsf{max}} \cdot \mathsf{poly}\Big(d, \kappa_A, \log \frac{dT_{\mathsf{total}}}{\delta}\Big) \cdot \rho^{t_{\mathsf{mix}}-1} \\ &\lesssim \Gamma^2_{\mathsf{max}} \sqrt{\frac{t_{\mathsf{mix}} d}{T_{\mathsf{total,subspace}}} \log^3 \frac{T_{\mathsf{total}} d}{\delta}}, \end{aligned}$$

where the last inequality holds if $t_{\mathsf{mix}} \gtrsim \frac{1}{1-\rho}\log\big(\frac{d\kappa_A T_{\mathsf{total}}}{\delta}\big)$. This finishes the proof of Lemma C.2.

Proof of (33). Define the truncating operator

$$\mathsf{Trunc}(x; D) \coloneqq x \cdot \mathbb{1}(|x| \le D), \qquad x \in \mathbb{R}, \quad D \ge 0.$$

Consider the following truncated version of $\widehat{\boldsymbol{G}}_i$:

$$\widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}} \coloneqq \frac{1}{MN} \sum_{m=1}^{M} \sum_{t=1}^{N} \Big(\mathsf{Trunc}\Big(\big(\boldsymbol{x}_{m,t}{}'\big)_i; D_0 \Big) \boldsymbol{x}_{m,t} \Big) \Big(\mathsf{Trunc}\Big(\big(\boldsymbol{z}_{m,t}{}'\big)_i; D_0 \Big) \boldsymbol{z}_{m,t} \Big)^\top$$

(the truncating level D_0 will be specified later), and let $\boldsymbol{E}_i^{\mathsf{Trunc}} \coloneqq \mathbb{E}\big[\widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}}\big]$ be its expectation. In the following, we first show that $\widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}}$ concentrates around $\boldsymbol{E}_i^{\mathsf{Trunc}}$, and then prove that $\boldsymbol{E}_i^{\mathsf{Trunc}} \approx \boldsymbol{G}_i$.

- By a standard covering argument, we have

$$\|\widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}} - \boldsymbol{E}_i^{\mathsf{Trunc}}\| = \sup_{\boldsymbol{u},\boldsymbol{v} \in \mathcal{S}^{d-1}} \boldsymbol{u}^\top \Big(\widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}} - \boldsymbol{E}_i^{\mathsf{Trunc}} \Big) \boldsymbol{v} \le 4 \sup_{\boldsymbol{u},\boldsymbol{v} \in \mathcal{N}_{1/8}} \boldsymbol{u}^\top \Big(\widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}} - \boldsymbol{E}_i^{\mathsf{Trunc}} \Big) \boldsymbol{v},$$

  where $\mathcal{N}_{1/8}$ denotes the $1/8$-covering of the unit sphere $\mathcal{S}^{d-1}$ and has cardinality $|\mathcal{N}_{1/8}| \le 32^d$. For fixed $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{N}_{1/8}$, one has

$$\boldsymbol{u}^\top \widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}} \boldsymbol{v} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{t=1}^{N} \mathsf{Trunc}\Big(\big(\boldsymbol{x}_{m,t}{}'\big)_i; D_0 \Big) \boldsymbol{u}^\top \boldsymbol{x}_{m,t} \cdot \mathsf{Trunc}\Big(\big(\boldsymbol{z}_{m,t}{}'\big)_i; D_0 \Big) \boldsymbol{v}^\top \boldsymbol{z}_{m,t},$$

  where (cf. Chapter 2 of (Vershynin, 2018) for the definitions of subgaussian norm $\|\cdot\|_{\psi_2}$ and subexponential norm $\|\cdot\|_{\psi_1}$)

$$\Big| \mathsf{Trunc}\Big(\big(\boldsymbol{x}_{m,t}{}'\big)_i; D_0 \Big) \Big|, \Big| \mathsf{Trunc}\Big(\big(\boldsymbol{z}_{m,t}{}'\big)_i; D_0 \Big) \Big| \le D_0, \quad \|\boldsymbol{u}^\top \boldsymbol{x}_{m,t}\|_{\psi_2}, \|\boldsymbol{v}^\top \boldsymbol{z}_{m,t}\|_{\psi_2} \lesssim \sqrt{\Gamma_{\mathsf{max}}}.$$

  Hence

$$\Big\| \mathsf{Trunc}\Big(\big(\boldsymbol{x}_{m,t}{}'\big)_i; D_0 \Big) \boldsymbol{u}^\top \boldsymbol{x}_{m,t} \cdot \mathsf{Trunc}\Big(\big(\boldsymbol{z}_{m,t}{}'\big)_i; D_0 \Big) \boldsymbol{v}^\top \boldsymbol{z}_{m,t} \Big\|_{\psi_1} \lesssim D_0^2 \Gamma_{\mathsf{max}},$$

  and by Bernstein's inequality (Corollary 2.8.3 of (Vershynin, 2018)), we have

$$\mathbb{P}\bigg(\Big| \boldsymbol{u}^\top \Big(\widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}} - \boldsymbol{E}_i^{\mathsf{Trunc}} \Big) \boldsymbol{v} \Big| \ge \tau \bigg) \le 2 \exp\bigg(-c_0 MN \Big(\frac{\tau}{D_0^2 \Gamma_{\mathsf{max}}} \Big)^2 \bigg)$$

  for all $0 \le \tau \le D_0^2 \Gamma_{\mathsf{max}}$. Taking the union bound over $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{N}_{1/8}$, we have with probability at least $1 - \delta/2$,

$$\|\widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}} - \boldsymbol{E}_i^{\mathsf{Trunc}}\| \lesssim D_0^2 \Gamma_{\mathsf{max}} \sqrt{\frac{d + \log \frac{1}{\delta}}{MN}}, \quad \text{provided that} \quad MN \gtrsim d + \log \frac{1}{\delta}.$$

- Note that

$$\|\boldsymbol{G}_i - \boldsymbol{E}_i^{\mathsf{Trunc}}\| = \bigg\| \sum_{k=1}^{K} p^{(k)} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{z}_t \sim \mathcal{N}(0, \widetilde{\boldsymbol{\Gamma}}^{(k)})} \Big[\Big((\boldsymbol{x}_t{}')_i (\boldsymbol{z}_t{}')_i - \mathsf{Trunc}\big((\boldsymbol{x}_t{}')_i\big) \mathsf{Trunc}\big((\boldsymbol{z}_t{}')_i\big) \Big) \boldsymbol{x}_t \boldsymbol{z}_t{}^\top \Big] \bigg\|,$$

  where for each k,

$$\begin{aligned} &\bigg\| \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{z}_t \sim \mathcal{N}(0, \widetilde{\boldsymbol{\Gamma}}^{(k)})} \Big[\Big((\boldsymbol{x}_t{}')_i (\boldsymbol{z}_t{}')_i - \mathsf{Trunc}\big((\boldsymbol{x}_t{}')_i\big) \mathsf{Trunc}\big((\boldsymbol{z}_t{}')_i\big) \Big) \boldsymbol{x}_t \boldsymbol{z}_t{}^\top \Big] \bigg\| \\ &\quad = \sup_{\boldsymbol{u},\boldsymbol{v} \in \mathcal{S}^{d-1}} \Big| \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{z}_t \sim \mathcal{N}(0, \widetilde{\boldsymbol{\Gamma}}^{(k)})} \Big[(\boldsymbol{x}_t{}')_i (\boldsymbol{z}_t{}')_i \boldsymbol{u}^\top \boldsymbol{x}_t \boldsymbol{v}^\top \boldsymbol{z}_t \cdot \Big(1 - \mathbb{1}\big(|(\boldsymbol{x}_t{}')_i| \le D_0, |(\boldsymbol{z}_t{}')_i| \le D_0 \big) \Big) \Big] \Big| \end{aligned}$$

$$\leq \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{S}^{d-1}} \sqrt{\mathbb{E}\Big((\boldsymbol{x}_t')_i(\boldsymbol{z}_t')_i\boldsymbol{u}^\top\boldsymbol{x}_t\boldsymbol{v}^\top\boldsymbol{z}_t\Big)^2}\sqrt{\mathbb{E}\Big(1-\mathbb{1}\big(|(\boldsymbol{x}_t')_i|\leq D_0, |(\boldsymbol{z}_t')_i|\leq D_0\big)\Big)^2}$$
$$\lesssim \Gamma_{\mathsf{max}}^2\sqrt{\mathbb{P}(|\mathcal{N}(0,\Gamma_{\mathsf{max}})|>D_0)}$$
$$\lesssim \Gamma_{\mathsf{max}}^2\exp\left(-c_1\frac{D_0^2}{\Gamma_{\mathsf{max}}}\right).$$

Finally, notice that if the truncating level D_0 is sufficiently large, then we have $\widehat{\boldsymbol{G}}_i = \widehat{\boldsymbol{G}}_i^{\mathsf{Trunc}}$ with high probability. More formally, we have shown that for fixed $1\leq i\leq d$, if $MN\gtrsim d+\log(1/\delta)$, then

$$\mathbb{P}\left(\|\widehat{\boldsymbol{G}}_i-\boldsymbol{G}_i\|\lesssim D_0^2\Gamma_{\mathsf{max}}\sqrt{\frac{d+\log\frac{1}{\delta}}{MN}}+\Gamma_{\mathsf{max}}^2\exp\left(-c_1\frac{D_0^2}{\Gamma_{\mathsf{max}}}\right)\right)$$
$$\geq 1-\frac{\delta}{2}-\sum_{m,t}\mathbb{P}\Big(|(\boldsymbol{x}_{m,t}')_i|>D_0 \text{ or } |(\boldsymbol{z}_{m,t}')_i|>D_0\Big).$$

If we pick the truncating level $D_0\asymp\sqrt{\Gamma_{\mathsf{max}}\log(MN/\delta)}$, then it is easy to check that with probability at least $1-\delta$,

$$\|\widehat{\boldsymbol{G}}_i-\boldsymbol{G}_i\|\lesssim D_0^2\Gamma_{\mathsf{max}}\sqrt{\frac{d+\log\frac{1}{\delta}}{MN}}\lesssim\Gamma_{\mathsf{max}}^2\sqrt{\frac{d}{MN}\log^3\frac{MN}{\delta}}.$$

Taking the union bound over $1\leq i\leq d$ finishes our proof of (33). □

C.2.2. Proof of Lemma C.3

Define $\boldsymbol{\Delta}\coloneqq\boldsymbol{M}-\boldsymbol{M}_\star$, and denote the eigendecomposition of $\boldsymbol{M}_\star$ and $\boldsymbol{M}$ as $\boldsymbol{M}_\star=\boldsymbol{U}_\star\boldsymbol{\Lambda}_\star\boldsymbol{U}_\star^\top$ and $\boldsymbol{M}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top+\boldsymbol{U}_\perp\boldsymbol{\Lambda}_\perp\boldsymbol{U}_\perp^\top$, where diagonal matrix $\boldsymbol{\Lambda}$ (resp. $\boldsymbol{\Lambda}_\star$) contains the top-$K$ eigenvalues of $\boldsymbol{M}$ (resp. $\boldsymbol{M}_\star$). Then we have

$$\boldsymbol{\Lambda}=\boldsymbol{U}^\top\boldsymbol{M}\boldsymbol{U}=\boldsymbol{U}^\top\boldsymbol{M}_\star\boldsymbol{U}+\boldsymbol{U}^\top\boldsymbol{\Delta}\boldsymbol{U}=\sum_{k=1}^K p^{(k)}\boldsymbol{U}^\top\boldsymbol{y}^{(k)}{\boldsymbol{y}^{(k)}}^\top\boldsymbol{U}+\boldsymbol{U}^\top\boldsymbol{\Delta}\boldsymbol{U},$$
$$\boldsymbol{\Lambda}_\star=\boldsymbol{U}_\star^\top\boldsymbol{M}_\star\boldsymbol{U}_\star=\sum_{k=1}^K p^{(k)}\boldsymbol{U}_\star^\top\boldsymbol{y}^{(k)}{\boldsymbol{y}^{(k)}}^\top\boldsymbol{U}_\star.$$

Substracting these two equations gives

$$\sum_{k=1}^K p^{(k)}\Big(\boldsymbol{U}_\star^\top\boldsymbol{y}^{(k)}{\boldsymbol{y}^{(k)}}^\top\boldsymbol{U}_\star-\boldsymbol{U}^\top\boldsymbol{y}^{(k)}{\boldsymbol{y}^{(k)}}^\top\boldsymbol{U}\Big)=\boldsymbol{\Lambda}_\star-\boldsymbol{\Lambda}+\boldsymbol{U}^\top\boldsymbol{\Delta}\boldsymbol{U}.$$

Taking the trace of both sides, we get

$$\sum_{k=1}^K p^{(k)}\Big(\|\boldsymbol{U}_\star^\top\boldsymbol{y}^{(k)}\|_2^2-\|\boldsymbol{U}^\top\boldsymbol{y}^{(k)}\|_2^2\Big)=\mathsf{Tr}\big(\boldsymbol{\Lambda}_\star-\boldsymbol{\Lambda}\big)+\mathsf{Tr}\big(\boldsymbol{U}^\top\boldsymbol{\Delta}\boldsymbol{U}\big).$$

On the left-hand side,

$$\|\boldsymbol{U}_\star^\top\boldsymbol{y}^{(k)}\|_2^2-\|\boldsymbol{U}^\top\boldsymbol{y}^{(k)}\|_2^2=\|\boldsymbol{y}^{(k)}\|_2^2-\|\boldsymbol{U}^\top\boldsymbol{y}^{(k)}\|_2^2=\|\boldsymbol{y}^{(k)}-\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{y}^{(k)}\|_2^2\geq 0,$$

while on the right-hand side,

$$\mathsf{Tr}(\boldsymbol{\Lambda}_\star-\boldsymbol{\Lambda})=\sum_{k=1}^K\Big(\lambda_k(\boldsymbol{\Lambda}_\star)-\lambda_k(\boldsymbol{\Lambda})\Big)\overset{\text{(i)}}{\leq}K\|\boldsymbol{\Delta}\|\leq K\epsilon,\quad \mathsf{Tr}(\boldsymbol{U}^\top\boldsymbol{\Delta}\boldsymbol{U})\leq\|\boldsymbol{\Delta}\|\cdot\mathsf{Tr}(\boldsymbol{U}^\top\boldsymbol{U})=\|\boldsymbol{\Delta}\|\cdot\mathsf{Tr}(\boldsymbol{I}_K)\leq K\epsilon,$$

where (i) follows from Weyl's inequality. Putting things together, we have proved (31), which immediately leads to the first upper bound in (32). The second upper bound in (32) follows from a simple application of Davis-Kahan's $\sin\Theta$ theorem (Davis & Kahan, 1970), and the third term is due to Lemma A.11 of (Kong et al., 2020b); we skip the details for brevity.

C.3. Proof of Theorem B.2

Our proof follows the three steps below:

1. Consider the idealized i.i.d. case, and characterize the expectations and variances of the testing statistics computed by Algorithm 3;
2. Go back to the actual case of mixed LDSs, and analyze one copy of $\mathsf{stat}_{\Gamma,g}$ or $\mathsf{stat}_{Y,g}$ defined in (18) for some fixed $1 \le g \le G$, by decomposing it into an i.i.d. part plus a negligible mixing error term;
3. Analyze $\mathsf{median}\{\mathsf{stat}_{\Gamma,g}, 1 \le g \le G\}$ and $\mathsf{median}\{\mathsf{stat}_{Y,g}, 1 \le g \le G\}$, and prove the correct testing for each pair of trajectories, which implies that Algorithm 3 achieves exact clustering.

Step 1: the idealized i.i.d. case. Recall the definition of stat_Y in (19) when we first introduce our method for clustering. For notational convenience, we drop the subscript in stat_Y, and replace $\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}$ with $\boldsymbol{x}_t', \boldsymbol{z}_t'$; then, with some elementary linear algebra, (19) can be rewritten as

$$\begin{aligned}\mathsf{stat} &= \sum_{i=1}^{d} \Big\langle \boldsymbol{U}_i^\top \frac{1}{|\Omega_1|} \sum_{t\in\Omega_1} \big((\boldsymbol{x}_t')_i \boldsymbol{x}_t - (\boldsymbol{z}_t')_i \boldsymbol{z}_t\big), \boldsymbol{U}_i^\top \frac{1}{|\Omega_2|} \sum_{t\in\Omega_2} \big((\boldsymbol{x}_t')_i \boldsymbol{x}_t - (\boldsymbol{z}_t')_i \boldsymbol{z}_t\big) \Big\rangle \\ &= \Big\langle \boldsymbol{U}^\top \frac{1}{|\Omega_1|} \sum_{t\in\Omega_1} \mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big), \boldsymbol{U}^\top \frac{1}{|\Omega_2|} \sum_{t\in\Omega_2} \mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big) \Big\rangle,\end{aligned}$$

where we define a large orthonormal matrix

$$\boldsymbol{U} := \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{U}_d \end{bmatrix} \in \mathbb{R}^{d^2 \times dK}, \tag{41}$$

In this step, we consider the idealized i.i.d. case:

$$\begin{aligned} t \in \Omega_1 \cup \Omega_2, \quad \boldsymbol{x}_t &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{\Gamma}}^{(k)}), \quad \boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{W}}^{(k)}), \quad \boldsymbol{x}_t' = \widetilde{\boldsymbol{A}}^{(k)} \boldsymbol{x}_t + \boldsymbol{w}_t, \\ \boldsymbol{z}_t &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{\Gamma}}^{(l)}), \quad \boldsymbol{v}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \widetilde{\boldsymbol{W}}^{(l)}), \quad \boldsymbol{z}_t' = \widetilde{\boldsymbol{A}}^{(l)} \boldsymbol{z}_t + \boldsymbol{v}_t, \end{aligned}$$

where $\widetilde{\boldsymbol{\Gamma}}^{(k)}, \widetilde{\boldsymbol{\Gamma}}^{(l)}, \widetilde{\boldsymbol{W}}^{(k)}, \widetilde{\boldsymbol{W}}^{(l)}$ are $d \times d$ covariance matrices, and $\widetilde{\boldsymbol{A}}^{(k)}, \widetilde{\boldsymbol{A}}^{(l)}$ are $d \times d$ state transition matrix. Our goal is to characterize the expectation and variance of stat in this i.i.d. case.

Before we present the results, we need some additional notation. First, let $\{\boldsymbol{e}_i\}_{1\le i\le d}$ be the canonical basis of $\mathbb{R}^d$, and define

$$\begin{aligned} \widetilde{\boldsymbol{Y}}^{(k)} &:= \widetilde{\boldsymbol{A}}^{(k)} \widetilde{\boldsymbol{\Gamma}}^{(k)}, \\ \boldsymbol{\Sigma}^{(k)} &:= \Big(\widetilde{\boldsymbol{A}}^{(k)} \otimes \boldsymbol{I}_d\Big)\Big((\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2} \otimes (\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2}\Big)\Big(\boldsymbol{I}_{d^2} + \boldsymbol{P}\Big)\Big((\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2} \otimes (\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2}\Big)\Big((\widetilde{\boldsymbol{A}}^{(k)})^\top \otimes \boldsymbol{I}_d\Big), \end{aligned}$$

where $\boldsymbol{P} \in \mathbb{R}^{d^2 \times d^2}$ is a symmetric permutation matrix, whose (i,j)-th block is $\boldsymbol{e}_j \boldsymbol{e}_i^\top \in \mathbb{R}^{d\times d}, 1 \le i,j \le d$. Let $\widetilde{\boldsymbol{Y}}^{(l)}, \boldsymbol{\Sigma}^{(\ell)}$ be defined similarly, with $\widetilde{\boldsymbol{A}}^{(k)}, \widetilde{\boldsymbol{\Gamma}}^{(k)}$ replaced by $\widetilde{\boldsymbol{A}}^{(l)}, \widetilde{\boldsymbol{\Gamma}}^{(l)}$. Moreover, define

$$\boldsymbol{\mu}_{k,l} := \boldsymbol{U}^\top \mathsf{vec}\Big((\widetilde{\boldsymbol{Y}}^{(k)} - \widetilde{\boldsymbol{Y}}^{(l)})^\top\Big) \in \mathbb{R}^{dK}, \tag{42a}$$

$$\boldsymbol{\Sigma}_{k,l} := \boldsymbol{U}^\top \big(\boldsymbol{\Sigma}^{(k)} + \widetilde{\boldsymbol{W}}^{(k)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(k)} + \boldsymbol{\Sigma}^{(\ell)} + \widetilde{\boldsymbol{W}}^{(l)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(l)}\big)\boldsymbol{U} \in \mathbb{R}^{dK \times dK}. \tag{42b}$$

Now we are ready to present our results for the i.i.d. case. The first lemma below gives a precise characterization of $\mathbb{E}[\mathsf{stat}]$ and $\mathsf{var}(\mathsf{stat})$, in terms of $\boldsymbol{\mu}_{k,l}$ and $\boldsymbol{\Sigma}_{k,l}$; the second lemma provides some upper and lower bounds, which will be handy for our later analyses.

Lemma C.4. *Denote $N = \min\{|\Omega_1|, |\Omega_2|\}$. For the i.i.d. case just described, one has*

$$\mathbb{E}[\mathsf{stat}] = \|\boldsymbol{\mu}_{k,l}\|_2^2, \quad \mathsf{var}(\mathsf{stat}) \le \frac{1}{N^2}\mathsf{Tr}(\boldsymbol{\Sigma}_{k,l}^2) + \frac{2}{N}\boldsymbol{\mu}_{k,l}^\top \boldsymbol{\Sigma}_{k,l}\boldsymbol{\mu}_{k,l},$$

and the inequality becomes an equality if $|\Omega_1| = |\Omega_2| = N$.

Lemma C.5. *Consider the same setting of Lemma C.4. Furthermore, suppose that*

$$\widetilde{\boldsymbol{\Gamma}}^{(k)}, \widetilde{\boldsymbol{\Gamma}}^{(l)} \preccurlyeq \Gamma_{\mathsf{max}}\boldsymbol{I}_d, \quad \widetilde{\boldsymbol{W}}^{(k)}, \widetilde{\boldsymbol{W}}^{(l)} \preccurlyeq W_{\mathsf{max}}\boldsymbol{I}_d, \quad \|\widetilde{\boldsymbol{A}}^{(k)}\|, \|\widetilde{\boldsymbol{A}}^{(l)}\| \le \kappa_A$$

for some $0 < W_{\mathsf{max}} \le \Gamma_{\mathsf{max}}$ and $\kappa_A \ge 1$. Then the following holds true.

- *(Upper bound on expectation) It holds that*

$$\mathbb{E}[\mathsf{stat}] = \|\boldsymbol{\mu}_{k,l}\|_2^2 \le \|\widetilde{\boldsymbol{Y}}^{(k)} - \widetilde{\boldsymbol{Y}}^{(l)}\|_{\mathrm{F}}^2. \tag{43}$$

- *(Lower bound on expectation) If $\widetilde{\boldsymbol{Y}}^{(k)} \neq \widetilde{\boldsymbol{Y}}^{(l)}$ and subspaces $\{\boldsymbol{U}_i\}_{1\le i\le d}$ satisfy*

$$1 \le i \le d, \quad \max\Big\{\|(\widetilde{\boldsymbol{Y}}^{(k)})_i - \boldsymbol{U}_i\boldsymbol{U}_i{}^\top(\widetilde{\boldsymbol{Y}}^{(k)})_i\|_2, \|(\widetilde{\boldsymbol{Y}}^{(l)})_i - \boldsymbol{U}_i\boldsymbol{U}_i{}^\top(\widetilde{\boldsymbol{Y}}^{(l)})_i\|_2\Big\} \le \epsilon, \tag{44}$$

  for some $\epsilon \ge 0$, then we have

$$\mathbb{E}[\mathsf{stat}] = \|\boldsymbol{\mu}_{k,l}\|_2^2 \ge \big\|\widetilde{\boldsymbol{Y}}^{(k)} - \widetilde{\boldsymbol{Y}}^{(l)}\big\|_{\mathrm{F}}^2 - 4\epsilon\sum_{i=1}^{d}\big\|(\widetilde{\boldsymbol{Y}}^{(k)})_i + (\widetilde{\boldsymbol{Y}}^{(l)})_i\big\|_2.$$

- *(Upper bound on variance) The matrix $\boldsymbol{\Sigma}_{k,l}$ is symmetric and satisfies*

$$\boldsymbol{0} \preccurlyeq \boldsymbol{\Sigma}_{k,l} \preccurlyeq 6\Gamma_{\mathsf{max}}^2\kappa_A^2\boldsymbol{I}_{dK};$$

  this, together with the earlier upper bound (43) on $\|\boldsymbol{\mu}_{k,l}\|_2^2$, implies that

$$\mathsf{var}(\mathsf{stat}) \le \frac{1}{N^2}\mathsf{Tr}(\boldsymbol{\Sigma}_{k,l}^2) + \frac{2}{N}\boldsymbol{\mu}_{k,l}^\top\boldsymbol{\Sigma}_{k,l}\boldsymbol{\mu}_{k,l} \lesssim \left(\frac{\Gamma_{\mathsf{max}}^2\kappa_A^2}{N}\right)^2 dK + \frac{\Gamma_{\mathsf{max}}^2\kappa_A^2}{N}\|\widetilde{\boldsymbol{Y}}^{(k)} - \widetilde{\boldsymbol{Y}}^{(l)}\|_{\mathrm{F}}^2.$$

Step 2: one copy of $\mathsf{stat}_{Y,g}, \mathsf{stat}_{\Gamma,g}$ for a fixed g. Now we turn back to the mixed LDSs setting and prove Theorem B.2. Let us focus on the testing of one pair of short trajectories $\{\boldsymbol{x}_{m_1,t}\}, \{\boldsymbol{x}_{m_2,t}\}$ for some $m_1, m_2 \in \mathcal{M}_{\mathsf{clustering}}$, $m_1 \neq m_2$. For notational consistency, in this proof we rewrite these two trajectories as $\{\boldsymbol{x}_t\}$ and $\{\boldsymbol{z}_t\}$, their labels k_{m_1}, k_{m_2} as k, ℓ, and the trajectory length $T_{\mathsf{clustering}}$ as T, respectively. Also denote $\boldsymbol{x}_t' \coloneqq \boldsymbol{x}_{t+1}$ and $\boldsymbol{z}_t' \coloneqq \boldsymbol{z}_{t+1}$. Recall the definition of $\{\mathsf{stat}_{Y,g}\}_{1\le g\le G}$ in (18b); for now, we consider one specific element and ignore the subscript g. Recalling the definition of $\boldsymbol{U} \in \mathbb{R}^{d^2\times dK}$ in (41), we have

$$\begin{aligned}
\mathsf{stat}_Y &= \sum_{i=1}^{d}\Big\langle\frac{1}{N}\sum_{t\in\Omega_1}\boldsymbol{U}_i{}^\top\big((\boldsymbol{x}_t')_i\boldsymbol{x}_t - (\boldsymbol{z}_t')_i\boldsymbol{z}_t\big), \frac{1}{N}\sum_{t\in\Omega_2}\boldsymbol{U}_i{}^\top\big((\boldsymbol{x}_t')_i\boldsymbol{x}_t - (\boldsymbol{z}_t')_i\boldsymbol{z}_t\big)\Big\rangle \\
&= \Big\langle\boldsymbol{U}^\top\frac{1}{N}\sum_{t\in\Omega_1}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big), \boldsymbol{U}^\top\frac{1}{N}\sum_{t\in\Omega_2}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big)\Big\rangle,
\end{aligned}$$

where $N = \lfloor T/4G\rfloor = |\Omega_1| = |\Omega_2|$.

In the following, we show how to decompose stat_Y into an i.i.d. term plus a negligible mixing error term, and then analyze each component of this decomposition; finally, we put pieces together to give a characterization of stat_Y, or $\{\mathsf{stat}_{Y,g}\}_{1\le g\le G}$ when we put the subscript g back in at the end of this step.

Step 2.1: decomposition of stat_Y. Define $\mathcal{S}^{1,\tau_1} \coloneqq \{t_1+\tau_1+f\cdot t_{\mathsf{mix}} : f\geq 0, \tau_1+f\cdot t_{\mathsf{mix}}\leq |\Omega_1|\}$, where t_1+1 is the first index of Ω_1, and the mixing time t_{mix} will be specified later; define $\mathcal{S}^{2,\tau_2}$ similarly. Note that for each τ_1, the elements of $\mathcal{S}^{1,\tau_1}$ are at least t_{mix} far apart; moreover, we have $\Omega_1 = \cup_{\tau_1=1}^{t_{\mathsf{mix}}}\mathcal{S}^{1,\tau_1}, \Omega_2 = \cup_{\tau_2=1}^{t_{\mathsf{mix}}}\mathcal{S}^{2,\tau_2}$, and thus

$$\frac{1}{N}\sum_{t\in\Omega_1}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big) = \sum_{\tau_1=1}^{t_{\mathsf{mix}}}\frac{|\mathcal{S}^{1,\tau_1}|}{N}\cdot\frac{1}{|\mathcal{S}^{1,\tau_1}|}\sum_{t\in\mathcal{S}^{1,\tau_1}}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big),$$
$$\frac{1}{N}\sum_{t\in\Omega_2}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big) = \sum_{\tau_2=1}^{t_{\mathsf{mix}}}\frac{|\mathcal{S}^{2,\tau_2}|}{N}\cdot\frac{1}{|\mathcal{S}^{2,\tau_2}|}\sum_{t\in\mathcal{S}^{2,\tau_2}}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big).$$

Therefore, we can rewrite stat_Y as a weighted average

$$\begin{aligned}
\mathsf{stat}_Y &= \sum_{\tau_1=1}^{t_{\mathsf{mix}}}\sum_{\tau_2=1}^{t_{\mathsf{mix}}} w^{\tau_1,\tau_2}\cdot\mathsf{stat}_Y^{\tau_1,\tau_2}, \quad \text{where}\\
w^{\tau_1,\tau_2} &\coloneqq \frac{|\mathcal{S}^{1,\tau_1}||\mathcal{S}^{2,\tau_2}|}{N^2}, \quad \sum_{\tau_1=1}^{t_{\mathsf{mix}}}\sum_{\tau_2=1}^{t_{\mathsf{mix}}} w^{\tau_1,\tau_2} = 1, \quad \text{and}\\
\mathsf{stat}_Y^{\tau_1,\tau_2} &\coloneqq \Big\langle \boldsymbol{U}^\top\frac{1}{|\mathcal{S}^{1,\tau_1}|}\sum_{t\in\mathcal{S}^{1,\tau_1}}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big), \boldsymbol{U}^\top\frac{1}{|\mathcal{S}^{2,\tau_2}|}\sum_{t\in\mathcal{S}^{2,\tau_2}}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top - \boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big)\Big\rangle. \qquad (45)
\end{aligned}$$

We can further decompose $\mathsf{stat}_Y^{\tau_1,\tau_2}$ into an i.i.d. term plus a small error term. To do this, recalling the definition of $\widetilde{\boldsymbol{x}}_{m,t}$ in (30) and dropping the subscript m, we have

$$\begin{aligned}
\boldsymbol{x}_t &= \underbrace{\boldsymbol{A}^{(k)^{t_{\mathsf{mix}}-1}}\boldsymbol{x}_{t-t_{\mathsf{mix}}+1}}_{=:\boldsymbol{\delta}_{x,t}} + \widetilde{\boldsymbol{x}}_t = \boldsymbol{\delta}_{x,t} + \widetilde{\boldsymbol{x}}_t,\\
\boldsymbol{x}_t' &= \boldsymbol{A}^{(k)}\boldsymbol{x}_t + \boldsymbol{w}_t = \boldsymbol{A}^{(k)}\boldsymbol{\delta}_{x,t} + \underbrace{(\boldsymbol{A}^{(k)}\widetilde{\boldsymbol{x}}_t + \boldsymbol{w}_t)}_{=:\widetilde{\boldsymbol{x}}_t'} = \boldsymbol{A}^{(k)}\boldsymbol{\delta}_{x,t} + \widetilde{\boldsymbol{x}}_t',
\end{aligned}$$

where $\widetilde{\boldsymbol{x}}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)})$. Similarly, we rewrite $\boldsymbol{z}_t = \boldsymbol{\delta}_{z,t} + \widetilde{\boldsymbol{z}}_t, \boldsymbol{z}_t' = \boldsymbol{A}^{(\ell)}\boldsymbol{\delta}_{z,t} + \widetilde{\boldsymbol{z}}_t'$. Plugging these into the right-hand side of (45) and expanding it, one has

$$\begin{aligned}
\mathsf{stat}_Y^{\tau_1,\tau_2} &= \widetilde{\mathsf{stat}}_Y^{\tau_1,\tau_2} + \Delta_Y^{\tau_1,\tau_2}, \quad \text{where}\\
\widetilde{\mathsf{stat}}_Y^{\tau_1,\tau_2} &\coloneqq \Big\langle \boldsymbol{U}^\top\frac{1}{|\mathcal{S}^{1,\tau_1}|}\sum_{t\in\mathcal{S}^{1,\tau_1}}\mathsf{vec}\big(\widetilde{\boldsymbol{x}}_t(\widetilde{\boldsymbol{x}}_t')^\top - \widetilde{\boldsymbol{z}}_t(\widetilde{\boldsymbol{z}}_t')^\top\big), \boldsymbol{U}^\top\frac{1}{|\mathcal{S}^{2,\tau_2}|}\sum_{t\in\mathcal{S}^{2,\tau_2}}\mathsf{vec}\big(\widetilde{\boldsymbol{x}}_t(\widetilde{\boldsymbol{x}}_t')^\top - \widetilde{\boldsymbol{z}}_t(\widetilde{\boldsymbol{z}}_t')^\top\big)\Big\rangle,
\end{aligned}$$

and $\Delta_Y^{\tau_1,\tau_2}$ involves $\{\boldsymbol{\delta}_{x,t}, \boldsymbol{\delta}_{z,t}\}$ terms.

Step 2.2: analysis of each component. First, notice that the $\{\widetilde{\boldsymbol{x}}_t, \widetilde{\boldsymbol{z}}_t\}$ terms in the definition of $\widetilde{\mathsf{stat}}_Y^{\tau_1,\tau_2}$ are independent; this suggests that we can characterize $\widetilde{\mathsf{stat}}_Y^{\tau_1,\tau_2}$ by applying our earlier analysis for the i.i.d. case in Step 1. Second, $\Delta_Y^{\tau_1,\tau_2}$ involves $\{\boldsymbol{\delta}_{x,t}, \boldsymbol{\delta}_{z,t}\}$ terms, which in turn involve $\boldsymbol{A}^{(k)^{t_{\mathsf{mix}}-1}}, \boldsymbol{A}^{(\ell)^{t_{\mathsf{mix}}-1}}$ and thus will be exponentially small as t_{mix} increases, thanks to Assumption 3.1. More formally, we have the following:

- Applying Lemmas C.4 and C.5 with $(\widetilde{\boldsymbol{\Gamma}}^{(k)}, \widetilde{\boldsymbol{\Gamma}}^{(l)}) = (\boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)}, \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(\ell)})$, $(\widetilde{\boldsymbol{W}}^{(k)}, \widetilde{\boldsymbol{W}}^{(l)}) = (\boldsymbol{W}^{(k)}, \boldsymbol{W}^{(\ell)})$ and $(\widetilde{\boldsymbol{A}}^{(k)}, \widetilde{\boldsymbol{A}}^{(l)}) = (\boldsymbol{A}^{(k)}, \boldsymbol{A}^{(\ell)})$, we have

$$\begin{aligned}
\mathbb{E}\big[\widetilde{\mathsf{stat}}_Y^{\tau_1,\tau_2}\big] &= \big\|\boldsymbol{U}^\top\mathsf{vec}((\boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(\ell)})^\top)\big\|_2^2,\\
\mathsf{var}\big(\widetilde{\mathsf{stat}}_Y^{\tau_1,\tau_2}\big) &\lesssim \left(\frac{\Gamma_{\mathsf{max}}^2\kappa_A^2}{\widetilde{N}}\right)^2 dK + \frac{\Gamma_{\mathsf{max}}^2\kappa_A^2}{\widetilde{N}}\big\|\boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(\ell)}\big\|_{\mathrm{F}}^2,
\end{aligned}$$

  where $\widetilde{N} \coloneqq \min\{|\mathcal{S}^{1,\tau_1}|, |\mathcal{S}^{2,\tau_2}|\} \asymp N/t_{\mathsf{mix}}$.

- By Lemma C.1, with probability at least $1-\delta$, all $\{\boldsymbol{x}_t, \widetilde{\boldsymbol{x}}_t, \boldsymbol{z}_t, \widetilde{\boldsymbol{z}}_t\}$ terms involved in the definition of $\Delta_Y^{\tau_1,\tau_2}$ has ℓ_2 norm bounded by $\sqrt{\Gamma_{\mathsf{max}}}\mathsf{poly}(d, \kappa_A, \log(T_{\mathsf{total}}/\delta))$. This implies that

$$\big|\Delta_Y^{\tau_1,\tau_2}\big| \le \Gamma_{\mathsf{max}}^2 \cdot \mathsf{poly}\Big(d, \kappa_A, \log\frac{T_{\mathsf{total}}}{\delta}\Big) \cdot \rho^{t_{\mathsf{mix}}-1}.$$

Step 2.3: putting pieces together. Putting the subscript g back in, we have already shown that

$$\begin{aligned}\mathsf{stat}_{Y,g} &= \sum_{\tau_1=1}^{t_{\mathsf{mix}}}\sum_{\tau_2=1}^{t_{\mathsf{mix}}} w^{\tau_1,\tau_2}\cdot \mathsf{stat}_{Y,g}^{\tau_1,\tau_2} \\ &= \underbrace{\sum_{\tau_1=1}^{t_{\mathsf{mix}}}\sum_{\tau_2=1}^{t_{\mathsf{mix}}} w^{\tau_1,\tau_2}\cdot \widetilde{\mathsf{stat}}_{Y,g}^{\tau_1,\tau_2}}_{=:\widetilde{\mathsf{stat}}_{Y,g}} + \underbrace{\sum_{\tau_1=1}^{t_{\mathsf{mix}}}\sum_{\tau_2=1}^{t_{\mathsf{mix}}} w^{\tau_1,\tau_2}\cdot \Delta_{Y,g}^{\tau_1,\tau_2}}_{=:\Delta_{Y,g}} = \widetilde{\mathsf{stat}}_{Y,g} + \Delta_{Y,g},\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\big[\widetilde{\mathsf{stat}}_{Y,g}\big] &= \big\|\boldsymbol{U}^\top \mathsf{vec}((\boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(\ell)})^\top)\big\|_2^2, \\ \mathsf{var}\big(\widetilde{\mathsf{stat}}_{Y,g}\big) &\le \max_{\tau_1,\tau_2}\mathsf{var}(\widetilde{\mathsf{stat}}_{Y,g}^{\tau_1,\tau_2}) \lesssim \left(\frac{\Gamma_{\mathsf{max}}^2\kappa_A^2}{\widetilde{N}}\right)^2 dK + \frac{\Gamma_{\mathsf{max}}^2\kappa_A^2}{\widetilde{N}}\big\|\boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(\ell)}\big\|_{\mathrm{F}}^2, \\ \big|\Delta_{Y,g}\big| &\le \max_{\tau_1,\tau_2}|\Delta_{Y,g}^{\tau_1,\tau_2}| \le \Gamma_{\mathsf{max}}^2\cdot\mathsf{poly}\Big(d,\kappa_A,\log\frac{T_{\mathsf{total}}}{\delta}\Big)\cdot\rho^{t_{\mathsf{mix}}-1}.\end{aligned}$$

So far in Step 2, we have focused on the analysis of $\mathsf{stat}_{Y,g}$. We can easily adapt the argument to study $\mathsf{stat}_{\Gamma,g}$ as well; the major difference is that, in Step 2.2, we should apply Lemmas C.4 and C.5 with $(\widetilde{\boldsymbol{\Gamma}}^{(k)}, \widetilde{\boldsymbol{\Gamma}}^{(l)}) = (\boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)}, \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(\ell)})$, $(\widetilde{\boldsymbol{W}}^{(k)}, \widetilde{\boldsymbol{W}}^{(l)}) = (\boldsymbol{0}, \boldsymbol{0})$, $(\widetilde{\boldsymbol{A}}^{(k)}, \widetilde{\boldsymbol{A}}^{(l)}) = (\boldsymbol{I}_d, \boldsymbol{I}_d)$, and subspaces $\{\boldsymbol{U}_i\}$ replaced by $\{\boldsymbol{V}_i\}$ instead. The final result is that, for all $1 \le g \le G$,

$$\mathsf{stat}_{\Gamma,g} = \widetilde{\mathsf{stat}}_{\Gamma,g} + \Delta_{\Gamma,g},$$

where

$$\begin{aligned}\mathbb{E}\big[\widetilde{\mathsf{stat}}_{\Gamma,g}\big] &= \big\|\boldsymbol{V}^\top \mathsf{vec}((\boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(\ell)})^\top)\big\|_2^2, \\ \mathsf{var}\big(\widetilde{\mathsf{stat}}_{\Gamma,g}\big) &\lesssim \left(\frac{\Gamma_{\mathsf{max}}^2\kappa_A^2}{\widetilde{N}}\right)^2 dK + \frac{\Gamma_{\mathsf{max}}^2\kappa_A^2}{\widetilde{N}}\big\|\boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(\ell)}\big\|_{\mathrm{F}}^2, \\ \big|\Delta_{\Gamma,g}\big| &\le \Gamma_{\mathsf{max}}^2\cdot\mathsf{poly}\Big(d,\kappa_A,\log\frac{T_{\mathsf{total}}}{\delta}\Big)\cdot\rho^{t_{\mathsf{mix}}-1}.\end{aligned}$$

Step 3: analysis of the medians, and final results. Recall the following standard result on the concentration of medians (or median-of-means in general; see Theorem 2 of (Lugosi & Mendelson, 2019)).

Proposition C.6 (Concentration of medians)**.** *Let $X_1, \dots X_G$ be i.i.d. random variables with mean μ and variance σ^2. Then we have $|\mathsf{median}\{X_g, 1 \le g \le G\} - \mu| \le 2\sigma$ with probability at least $1 - e^{-c_0 G}$ for some constant c_0.*

Notice that by construction, $\{\widetilde{\mathsf{stat}}_{Y,g}\}_{1\le g\le G}$ are i.i.d. (and so are $\{\widetilde{\mathsf{stat}}_{\Gamma,g}\}_{1\le g\le G}$). Applying Proposition C.6 to our case, we know that if $G \gtrsim \log(1/\delta)$, then with probability at least $1-\delta$, the following holds:

- If $k = \ell$, i.e. the two trajectories are generated by the same LDS model, then

$$\begin{aligned}&\mathsf{median}\{\mathsf{stat}_{\Gamma,g}, 1\le g\le G\} + \mathsf{median}\{\mathsf{stat}_{Y,g}, 1\le g\le G\} \\ &\le \mathsf{median}\{\widetilde{\mathsf{stat}}_{\Gamma,g}\} + \mathsf{median}\{\widetilde{\mathsf{stat}}_{Y,g}\} + \max_g|\Delta_{\Gamma,g}| + \max_g|\Delta_{Y,g}| \\ &\le 2\sqrt{\mathsf{var}(\widetilde{\mathsf{stat}}_{\Gamma,g})} + 2\sqrt{\mathsf{var}(\widetilde{\mathsf{stat}}_{Y,g})} + \Gamma_{\mathsf{max}}^2\cdot\mathsf{poly}\Big(d,\kappa_A,\log\frac{T_{\mathsf{total}}}{\delta}\Big)\cdot\rho^{t_{\mathsf{mix}}-1} \\ &\lesssim \frac{\Gamma_{\mathsf{max}}^2\kappa_A^2\sqrt{dK}}{\widetilde{N}} + \Gamma_{\mathsf{max}}^2\cdot\mathsf{poly}\Big(d,\kappa_A,\log\frac{T_{\mathsf{total}}}{\delta}\Big)\cdot\rho^{t_{\mathsf{mix}}-1};\end{aligned} \tag{46}$$

- On the other hand, if $k \neq \ell$, then

$$\begin{aligned}
&\mathsf{median}\{\mathsf{stat}_{\Gamma,g}, 1 \le g \le G\} + \mathsf{median}\{\mathsf{stat}_{Y,g}, 1 \le g \le G\} \\
&\quad \ge \mathsf{median}\{\widetilde{\mathsf{stat}}_{\Gamma,g}\} + \mathsf{median}\{\widetilde{\mathsf{stat}}_{Y,g}\} - \Big(\max_g |\Delta_{\Gamma,g}| + \max_g |\Delta_{Y,g}| \Big) \\
&\quad \ge \mathbb{E}[\widetilde{\mathsf{stat}}_{\Gamma,g}] + \mathbb{E}[\widetilde{\mathsf{stat}}_{Y,g}] - 2\Big(\sqrt{\mathsf{var}(\widetilde{\mathsf{stat}}_{\Gamma,g})} + \sqrt{\mathsf{var}(\widetilde{\mathsf{stat}}_{Y,g})} \Big) - \Gamma_{\mathsf{max}}^2 \cdot \mathsf{poly}\Big(d, \kappa_A, \log \frac{T_{\mathsf{total}}}{\delta}\Big) \cdot \rho^{t_{\mathsf{mix}}-1} \\
&\quad \ge \Big\| \boldsymbol{V}^\top \mathsf{vec}\big((\boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(\ell)})^\top\big) \Big\|_2^2 + \Big\| \boldsymbol{U}^\top \mathsf{vec}\big((\boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(\ell)})^\top\big) \Big\|_2^2 \\
&\qquad - C_0 \Bigg(\frac{\Gamma_{\mathsf{max}}^2 \kappa_A^2 \sqrt{dK}}{\widetilde{N}} + \sqrt{\frac{\Gamma_{\mathsf{max}}^2 \kappa_A^2}{\widetilde{N}}} \Big(\|\boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(\ell)}\|_{\mathrm{F}} + \|\boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(\ell)}\|_{\mathrm{F}} \Big) \Bigg) \\
&\qquad - \Gamma_{\mathsf{max}}^2 \cdot \mathsf{poly}\Big(d, \kappa_A, \log \frac{T_{\mathsf{total}}}{\delta}\Big) \cdot \rho^{t_{\mathsf{mix}}-1}. \qquad (47)
\end{aligned}$$

We need to further simplify the result (47) for the $k \neq \ell$ case. According to (28) and (29), we have

$$\begin{aligned}
\|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(k)}\|_{\mathrm{F}} &\le \Gamma_{\mathsf{max}} \sqrt{d} \kappa_A^3 \rho^{2(t_{\mathsf{mix}}-1)} =: \epsilon_{\mathsf{mix}}, \\
\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)}\|_{\mathrm{F}} &\le \Gamma_{\mathsf{max}} \sqrt{d} \kappa_A^2 \rho^{2(t_{\mathsf{mix}}-1)} \le \epsilon_{\mathsf{mix}},
\end{aligned}$$

which implies that

$$\begin{aligned}
\|\boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(\ell)}\|_{\mathrm{F}} &\le \|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}} + 2\epsilon_{\mathsf{mix}}, \\
\Big\| \boldsymbol{V}^\top \mathsf{vec}\big((\boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{\Gamma}_{t_{\mathsf{mix}}-1}^{(\ell)})^\top\big) \Big\|_2^2 &\ge \max\Big\{ \|\boldsymbol{V}^\top \mathsf{vec}\big((\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)})^\top\big)\|_2 - 2\epsilon_{\mathsf{mix}}, 0 \Big\}^2 \\
&\ge \Big\| \boldsymbol{V}^\top \mathsf{vec}\big((\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)})^\top\big) \Big\|_2^2 - 4\epsilon_{\mathsf{mix}} \Big\| \boldsymbol{V}^\top \mathsf{vec}\big((\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)})^\top\big) \Big\|_2 \\
&\ge \Big\| \boldsymbol{V}^\top \mathsf{vec}\big((\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)})^\top\big) \Big\|_2^2 - 4\epsilon_{\mathsf{mix}} \|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}.
\end{aligned}$$

We can do a similar analysis for $\|\boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(\ell)}\|_{\mathrm{F}}$ and $\|\boldsymbol{U}^\top \mathsf{vec}((\boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(k)} - \boldsymbol{Y}_{t_{\mathsf{mix}}-1}^{(\ell)})^\top)\|_2^2$. Moreover, we claim (and prove later) that if the subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}$ satisfy the condition (25) in the theorem, then

$$\Big\| \boldsymbol{V}^\top \mathsf{vec}((\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)})^\top) \Big\|_2^2 + \Big\| \boldsymbol{U}^\top \mathsf{vec}((\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)})^\top) \Big\|_2^2 \ge \frac{1}{2} \Big(\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2 \Big). \qquad (48)$$

Putting these back to (47), we have for the $k \neq \ell$ case,

$$\begin{aligned}
&\mathsf{median}\{\mathsf{stat}_{\Gamma,g}, 1 \le g \le G\} + \mathsf{median}\{\mathsf{stat}_{Y,g}, 1 \le g \le G\} \\
&\quad \ge \Big\| \boldsymbol{V}^\top \mathsf{vec}\big((\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)})^\top\big) \Big\|_2^2 + \Big\| \boldsymbol{U}^\top \mathsf{vec}\big((\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)})^\top\big) \Big\|_2^2 \\
&\qquad - 4\epsilon_{\mathsf{mix}} \Big(\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}} + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}} \Big) \\
&\qquad - C_0 \Bigg(\frac{\Gamma_{\mathsf{max}}^2 \kappa_A^2 \sqrt{dK}}{\widetilde{N}} + \sqrt{\frac{\Gamma_{\mathsf{max}}^2 \kappa_A^2}{\widetilde{N}}} \Big(\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}} + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}} + 4\epsilon_{\mathsf{mix}} \Big) \Bigg) \\
&\qquad - \Gamma_{\mathsf{max}}^2 \cdot \mathsf{poly}\Big(d, \kappa_A, \log \frac{T_{\mathsf{total}}}{\delta}\Big) \cdot \rho^{t_{\mathsf{mix}}-1} \\
&\quad \overset{\text{(i)}}{\ge} \frac{1}{2} \Big(\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2 \Big) - 0.01 \Big(\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2 \Big) \\
&\qquad - C_1 \Bigg(\frac{\Gamma_{\mathsf{max}}^2 \kappa_A^2 \sqrt{dK}}{\widetilde{N}} + \sqrt{\frac{\Gamma_{\mathsf{max}}^2 \kappa_A^2}{\widetilde{N}}} \sqrt{\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2} \Bigg) \\
&\qquad - \Gamma_{\mathsf{max}}^2 \cdot \mathsf{poly}\Big(d, \kappa_A, \log \frac{T_{\mathsf{total}}}{\delta}\Big) \cdot \rho^{t_{\mathsf{mix}}-1}
\end{aligned}$$

$$\overset{\text{(ii)}}{\geq} 0.48\Big(\|\mathbf{\Gamma}^{(k)}-\mathbf{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2+\|\boldsymbol{Y}^{(k)}-\boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2\Big) - C_1\bigg(\frac{\Gamma_{\mathsf{max}}^2\kappa_A^2\sqrt{dK}}{\widetilde{N}}+\sqrt{\frac{\Gamma_{\mathsf{max}}^2\kappa_A^2}{\widetilde{N}}}\sqrt{\|\mathbf{\Gamma}^{(k)}-\mathbf{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2+\|\boldsymbol{Y}^{(k)}-\boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2}\bigg), \tag{49}$$

where (i) holds if $t_{\mathsf{mix}} \gtrsim \frac{1}{1-\rho}\log((\frac{\Gamma_{\mathsf{max}}}{\Delta_{\Gamma,Y}}+2)d\kappa_A)$ so that $\epsilon_{\mathsf{mix}} \leq 10^{-3}\Delta_{\Gamma,Y}$, and (ii) holds if $t_{\mathsf{mix}} \gtrsim \frac{1}{1-\rho}\log((\frac{\Gamma_{\mathsf{max}}}{\Delta_{\Gamma,Y}}+2)\frac{d\kappa_A T_{\mathsf{total}}}{\delta})$ so that $\Gamma_{\mathsf{max}}^2\cdot\mathsf{poly}(d,\kappa_A,\log\frac{T_{\mathsf{total}}}{\delta})\cdot\rho^{t_{\mathsf{mix}}-1}\leq 10^{-3}\Delta_{\Gamma,Y}^2$.

Putting (46) and (49) together, we can finally check that, if it further holds that $\widetilde{N} \asymp N/t_{\mathsf{mix}} \gtrsim \frac{\Gamma_{\mathsf{max}}^2\kappa_A^2\sqrt{dK}}{\Delta_{\Gamma,Y}^2}+1$, then we have with probability at least $1-\delta$,

$$\mathsf{median}\{\mathsf{stat}_{\Gamma,g}\}+\mathsf{median}\{\mathsf{stat}_{Y,g}\}\begin{cases}\leq \frac{1}{8}\Delta_{\Gamma,Y}^2 & \text{if} \quad k=\ell,\\ \geq \frac{3}{8}\big(\|\mathbf{\Gamma}^{(k)}-\mathbf{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2+\|\boldsymbol{Y}^{(k)}-\boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2\big)\geq\frac{3}{8}\Delta_{\Gamma,Y}^2 & \text{if} \quad k\neq\ell.\end{cases}$$

This together with our choice of testing threshold $\tau\in[\Delta_{\Gamma,Y}^2/8, 3\Delta_{\Gamma,Y}^2/8]$ in Algorithm 3 implies correct testing of the two trajectories $\{\boldsymbol{x}_t\},\{\boldsymbol{z}_t\}$. Finally, taking the union bound over all pairs of trajectories in $\mathcal{M}_{\mathsf{clustering}}$ leads to correct pairwise testing, which in turn implies exact clustering of $\mathcal{M}_{\mathsf{clustering}}$; this completes our proof of Theorem (B.2).

C.3.1. Proof of Lemma C.4

We first assume $|\Omega_1|=|\Omega_2|=N$ for simplicity. Recall that

$$\mathsf{stat}=\Big\langle\underbrace{\boldsymbol{U}^\top\frac{1}{|\Omega_1|}\sum_{t\in\Omega_1}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top-\boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big)}_{=:\boldsymbol{a}},\underbrace{\boldsymbol{U}^\top\frac{1}{|\Omega_2|}\sum_{t\in\Omega_2}\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top-\boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big)}_{=:\boldsymbol{b}}\Big\rangle=\langle\boldsymbol{a},\boldsymbol{b}\rangle,$$

where $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^{dK}$ are i.i.d., and

$$\mathbb{E}[\boldsymbol{a}]=\mathbb{E}\Big[\boldsymbol{U}^\top\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top-\boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big)\Big]=\boldsymbol{U}^\top\mathsf{vec}\Big((\widetilde{\boldsymbol{Y}}^{(k)}-\widetilde{\boldsymbol{Y}}^{(l)})^\top\Big)=\boldsymbol{\mu}_{k,l}.$$

Therefore, we have the expectation

$$\mathbb{E}[\mathsf{stat}]=\big\langle\mathbb{E}[\boldsymbol{a}],\mathbb{E}[\boldsymbol{b}]\big\rangle=\|\boldsymbol{\mu}_{k,l}\|_2^2.$$

It remains to compute the variance $\mathsf{var}(\mathsf{stat})=\mathbb{E}[\mathsf{stat}^2]-\mathbb{E}[\mathsf{stat}]^2$, where

$$\mathbb{E}[\mathsf{stat}^2]=\mathbb{E}\big[(\boldsymbol{a}^\top\boldsymbol{b})^2\big]=\mathsf{Tr}\big(\mathbb{E}[\boldsymbol{b}\boldsymbol{b}^\top]\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top]\big)=\mathsf{Tr}\big(\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top]^2\big). \tag{50}$$

Here $\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top]=\mathbb{E}[\boldsymbol{a}]\mathbb{E}[\boldsymbol{a}]^\top+\mathsf{cov}(\boldsymbol{a})$, and since $\boldsymbol{a}$ is an empirical average of N i.i.d. random vectors, we have

$$\mathsf{cov}(\boldsymbol{a})=\frac{1}{N}\mathsf{cov}(\boldsymbol{f}),\quad\text{where}\quad\boldsymbol{f}:=\boldsymbol{U}^\top\mathsf{vec}\big(\boldsymbol{x}_t(\boldsymbol{x}_t')^\top-\boldsymbol{z}_t(\boldsymbol{z}_t')^\top\big)\in\mathbb{R}^{dK}.$$

For now, we claim that

$$\mathsf{cov}(\boldsymbol{f})=\boldsymbol{U}^\top(\mathbf{\Sigma}^{(k)}+\widetilde{\boldsymbol{W}}^{(k)}\otimes\widetilde{\mathbf{\Gamma}}^{(k)}+\mathbf{\Sigma}^{(\ell)}+\widetilde{\boldsymbol{W}}^{(l)}\otimes\widetilde{\mathbf{\Gamma}}^{(l)})\boldsymbol{U}=\mathbf{\Sigma}_{k,l}, \tag{51}$$

which will be proved soon later. Putting these back to (50), one has

$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top]=\mathsf{cov}(\boldsymbol{a})+\mathbb{E}[\boldsymbol{a}]\mathbb{E}[\boldsymbol{a}]^\top=\frac{1}{N}\mathbf{\Sigma}_{k,l}+\boldsymbol{\mu}_{k,l}\boldsymbol{\mu}_{k,l}^\top,$$
$$\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top]^2=\frac{1}{N^2}\mathbf{\Sigma}_{k,l}^2+\frac{1}{N}(\mathbf{\Sigma}_{k,l}\boldsymbol{\mu}_{k,l}\boldsymbol{\mu}_{k,l}^\top+\boldsymbol{\mu}_{k,l}\boldsymbol{\mu}_{k,l}^\top\mathbf{\Sigma}_{k,l})+\|\boldsymbol{\mu}_{k,l}\|_2^2\boldsymbol{\mu}_{k,l}\boldsymbol{\mu}_{k,l}^\top,$$

and finally

$$\mathsf{var}(\mathsf{stat})=\mathbb{E}[\mathsf{stat}^2]-\mathbb{E}[\mathsf{stat}]^2=\mathsf{Tr}(\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top]^2)-\|\boldsymbol{\mu}_{k,l}\|_2^4=\frac{1}{N^2}\mathsf{Tr}(\mathbf{\Sigma}_{k,l}^2)+\frac{2}{N}\boldsymbol{\mu}_{k,l}^\top\mathbf{\Sigma}_{k,l}\boldsymbol{\mu}_{k,l},$$

which completes our calculation of the variance for the case of $|\Omega_1|=|\Omega_2|=N$. For the more general case where (without loss of generality) $|\Omega_1|=N\leq|\Omega_2|$, we simply need to modify the equation (50) to an inequality $\mathbb{E}[\mathsf{stat}^2]=\mathsf{Tr}(\mathbb{E}[\boldsymbol{b}\boldsymbol{b}^\top]\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top])\leq\mathsf{Tr}(\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top]^2)$, and the remaining analysis is the same. This finishes the proof of Lemma C.4.

Proof of (51). For notational simplicity, we drop the subscripts t in the definition of $\boldsymbol{f}$. Then we have $\boldsymbol{f} = \boldsymbol{U}^\top \mathsf{vec}\big(\boldsymbol{x}(\boldsymbol{x}')^\top - \boldsymbol{z}(\boldsymbol{z}')^\top\big)$, and hence

$$\mathsf{cov}(\boldsymbol{f}) = \boldsymbol{U}^\top \mathsf{cov}\Big(\mathsf{vec}\big(\boldsymbol{x}(\boldsymbol{x}') - \boldsymbol{z}(\boldsymbol{z}')^\top\big)\Big)\boldsymbol{U} = \boldsymbol{U}^\top \Big(\mathsf{cov}\big(\mathsf{vec}(\boldsymbol{x}(\boldsymbol{x}')^\top)\big) + \mathsf{cov}\big(\mathsf{vec}(\boldsymbol{z}(\boldsymbol{z}')^\top)\big)\Big)\boldsymbol{U}, \tag{52}$$

where the second equality uses the independence between $(\boldsymbol{x}, \boldsymbol{x}')$ and $(\boldsymbol{z}, \boldsymbol{z}')$.

Let us focus on $\mathsf{cov}\big(\mathsf{vec}(\boldsymbol{x}(\boldsymbol{x}')^\top)\big)$. For notational simplicity, for now we rewrite $\widetilde{\boldsymbol{A}}^{(k)}, \widetilde{\boldsymbol{W}}^{(k)}, \widetilde{\boldsymbol{\Gamma}}^{(k)}$ as $\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{\Gamma}$. Define $\boldsymbol{y} \coloneqq \boldsymbol{\Gamma}^{-1/2}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, and recall that $\boldsymbol{x}' = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ where $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W})$. Using the fact that $\mathsf{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\top \otimes \boldsymbol{A})\mathsf{vec}(\boldsymbol{B})$ for any matrices of compatible shape, we have

$$\begin{aligned}
\boldsymbol{g} &\coloneqq \mathsf{vec}\Big(\boldsymbol{x}\big(\boldsymbol{x}'\big)^\top\Big) = \mathsf{vec}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{A}^\top) + \mathsf{vec}(\boldsymbol{x}\boldsymbol{w}^\top) \\
&= (\boldsymbol{A} \otimes \boldsymbol{I}_d)\mathsf{vec}(\boldsymbol{x}\boldsymbol{x}^\top) + \mathsf{vec}(\boldsymbol{x}\boldsymbol{w}^\top) = (\boldsymbol{A} \otimes \boldsymbol{I}_d)\mathsf{vec}(\boldsymbol{\Gamma}^{1/2}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{\Gamma}^{1/2}) + \mathsf{vec}(\boldsymbol{x}\boldsymbol{w}^\top) \\
&= \underbrace{(\boldsymbol{A} \otimes \boldsymbol{I}_d)(\boldsymbol{\Gamma}^{1/2} \otimes \boldsymbol{\Gamma}^{1/2})\mathsf{vec}(\boldsymbol{y}\boldsymbol{y}^\top)}_{=:\boldsymbol{g}_1} + \underbrace{\mathsf{vec}(\boldsymbol{x}\boldsymbol{w}^\top)}_{=:\boldsymbol{g}_2} = \boldsymbol{g}_1 + \boldsymbol{g}_2.
\end{aligned}$$

Note that $\mathbb{E}[\boldsymbol{g}_2] = \boldsymbol{0}$, $\mathbb{E}[\boldsymbol{g}] = \mathbb{E}[\boldsymbol{g}_1]$, and $\mathbb{E}[\boldsymbol{g}_1\boldsymbol{g}_2^\top] = \boldsymbol{0}$. Hence

$$\mathsf{cov}(\boldsymbol{g}) = \mathbb{E}[\boldsymbol{g}\boldsymbol{g}^\top] - \mathbb{E}[\boldsymbol{g}]\mathbb{E}[\boldsymbol{g}]^\top = \mathsf{cov}(\boldsymbol{g}_1) + \mathbb{E}[\boldsymbol{g}_2\boldsymbol{g}_2^\top]. \tag{53}$$

For the second term on the right-hand side, we have

$$\boldsymbol{g}_2 = \mathsf{vec}(\boldsymbol{x}\boldsymbol{w}^\top) = \begin{bmatrix} w_1\boldsymbol{x} \\ \vdots \\ w_d\boldsymbol{x} \end{bmatrix}, \quad \mathbb{E}[\boldsymbol{g}_2\boldsymbol{g}_2^\top] = \mathbb{E}\Big[[w_i w_j \boldsymbol{x}\boldsymbol{x}^\top]_{1\le i,j\le d}\Big] = \boldsymbol{W} \otimes \boldsymbol{\Gamma};$$

as for the first term, we claim (and prove soon later) that

$$\mathsf{cov}\Big(\mathsf{vec}(\boldsymbol{y}\boldsymbol{y}^\top)\Big) = \boldsymbol{I}_{d^2} + \boldsymbol{P}, \tag{54}$$

which implies that

$$\begin{aligned}
\mathsf{cov}(\boldsymbol{g}_1) &= \mathsf{cov}\Big((\boldsymbol{A} \otimes \boldsymbol{I}_d)(\boldsymbol{\Gamma}^{1/2} \otimes \boldsymbol{\Gamma}^{1/2})\mathsf{vec}(\boldsymbol{y}\boldsymbol{y}^\top)\Big) \\
&= (\boldsymbol{A} \otimes \boldsymbol{I}_d)(\boldsymbol{\Gamma}^{1/2} \otimes \boldsymbol{\Gamma}^{1/2})\mathsf{cov}\Big(\mathsf{vec}(\boldsymbol{y}\boldsymbol{y}^\top)\Big)(\boldsymbol{\Gamma}^{1/2} \otimes \boldsymbol{\Gamma}^{1/2})(\boldsymbol{A}^\top \otimes \boldsymbol{I}_d) \\
&= (\boldsymbol{A} \otimes \boldsymbol{I}_d)(\boldsymbol{\Gamma}^{1/2} \otimes \boldsymbol{\Gamma}^{1/2})(\boldsymbol{I}_{d^2} + \boldsymbol{P})(\boldsymbol{\Gamma}^{1/2} \otimes \boldsymbol{\Gamma}^{1/2})(\boldsymbol{A}^\top \otimes \boldsymbol{I}_d).
\end{aligned}$$

Putting these back to (53), one has

$$\mathsf{cov}\big(\mathsf{vec}(\boldsymbol{x}(\boldsymbol{x}')^\top)\big) = \mathsf{cov}(\boldsymbol{g}) = (\boldsymbol{A} \otimes \boldsymbol{I}_d)(\boldsymbol{\Gamma}^{1/2} \otimes \boldsymbol{\Gamma}^{1/2})(\boldsymbol{I}_{d^2} + \boldsymbol{P})(\boldsymbol{\Gamma}^{1/2} \otimes \boldsymbol{\Gamma}^{1/2})(\boldsymbol{A}^\top \otimes \boldsymbol{I}_d) + \boldsymbol{W} \otimes \boldsymbol{\Gamma},$$

which is equal to $\boldsymbol{\Sigma}^{(k)} + \widetilde{\boldsymbol{W}}^{(k)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(k)}$ if we return to the original notation of $\widetilde{\boldsymbol{A}}^{(k)}, \widetilde{\boldsymbol{W}}^{(k)}, \widetilde{\boldsymbol{\Gamma}}^{(k)}$. By a similar analysis, we can show that $\mathsf{cov}\big(\mathsf{vec}(\boldsymbol{z}(\boldsymbol{z}')^\top)\big) = \boldsymbol{\Sigma}^{(\ell)} + \widetilde{\boldsymbol{W}}^{(l)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(l)}$. Putting these back to (52) finishes our calculation of $\mathsf{cov}(\boldsymbol{f})$.

Finally, it remains to prove (54). Denote

$$\boldsymbol{u} \coloneqq \mathsf{vec}(\boldsymbol{y}\boldsymbol{y}^\top) = \begin{bmatrix} y_1\boldsymbol{y} \\ \vdots \\ y_d\boldsymbol{y} \end{bmatrix}.$$

Then $\mathbb{E}[\boldsymbol{u}] = \mathsf{vec}(\boldsymbol{I}_d)$, and thus $\mathbb{E}[\boldsymbol{u}]\mathbb{E}[\boldsymbol{u}]^\top = [\boldsymbol{e}_i\boldsymbol{e}_j^\top]_{1\le i,j\le d}$. Next, consider

$$\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] = \mathbb{E}\Big[[y_i y_j \boldsymbol{y}\boldsymbol{y}^\top]_{1\le i,j\le d}\Big].$$

- For $i = j$,
$$\mathbb{E}\big[y_i^2 y_k y_\ell\big] = \begin{cases} 3 & \text{if} \quad k = \ell = i, \\ 1 & \text{if} \quad k = \ell \neq i, \\ 0 & \text{if} \quad k \neq \ell, \end{cases}$$
and hence $\mathbb{E}[y_i^2 \boldsymbol{y}\boldsymbol{y}^\top] = \boldsymbol{I}_d + 2\boldsymbol{e}_i\boldsymbol{e}_i^\top$.
- For $i \neq j$,
$$\mathbb{E}\big[y_i y_j y_k y_\ell\big] = \begin{cases} 1 & \text{if} \quad k = i, \ell = j \quad \text{or} \quad k = j, \ell = i, \\ 0 & \text{otherwise}, \end{cases}$$
and hence $\mathbb{E}[y_i y_j \boldsymbol{y}\boldsymbol{y}^\top] = \boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top$.

Putting together, the (i, j)-th $d \times d$ block of $\mathsf{cov}(\boldsymbol{u}) = \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] - \mathbb{E}[\boldsymbol{u}]\mathbb{E}[\boldsymbol{u}]^\top$ is equal to $\boldsymbol{I}_d + \boldsymbol{e}_i\boldsymbol{e}_i^\top$ if $i = j$, and $\boldsymbol{e}_j\boldsymbol{e}_i^\top$ if $i \neq j$. In other words, $\mathsf{cov}(\mathsf{vec}(\boldsymbol{y}\boldsymbol{y}^\top)) = \mathsf{cov}(\boldsymbol{u}) = \boldsymbol{I}_{d^2} + \boldsymbol{P}$, where $\boldsymbol{P} = [\boldsymbol{e}_j\boldsymbol{e}_i^\top]_{1 \le i,j \le d}$ is a symmetric permutation matrix; this completes our proof of (54). □

C.3.2. Proof of Lemma C.5

First, it holds that

$$\|\boldsymbol{\mu}_{k,l}\|_2^2 = \sum_{i=1}^d \Big\|\boldsymbol{U}_i{}^\top\big((\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\big)\Big\|_2^2 \le \sum_{i=1}^d \Big\|(\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\Big\|_2^2 = \|\widetilde{\boldsymbol{Y}}^{(k)} - \widetilde{\boldsymbol{Y}}^{(l)}\|_{\mathrm{F}}^2,$$

which gives the upper bound on $\mathbb{E}[\mathsf{stat}] = \|\boldsymbol{\mu}_{k,l}\|_2^2$. For the lower bound, the triangle inequality tells us that

$$\Big\|\boldsymbol{U}_i{}^\top\big((\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\big)\Big\|_2 = \Big\|\boldsymbol{U}_i\boldsymbol{U}_i{}^\top(\widetilde{\boldsymbol{Y}}^{(k)})_i - \boldsymbol{U}_i\boldsymbol{U}_i{}^\top(\widetilde{\boldsymbol{Y}}^{(l)})_i\Big\|_2 \ge \max\Big\{\|(\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\|_2 - 2\epsilon, 0\Big\},$$

which implies that

$$\Big\|\boldsymbol{U}_i{}^\top\big((\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\big)\Big\|_2^2 \ge \|(\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\|_2^2 - 4\epsilon\|(\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\|_2,$$

and hence

$$\begin{aligned}\|\boldsymbol{\mu}_{k,l}\|_2^2 &= \sum_{i=1}^d \Big\|\boldsymbol{U}_i{}^\top\big((\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\big)\Big\|_2^2 \ge \sum_{i=1}^d \|(\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\|_2^2 - 4\epsilon\sum_{i=1}^d \|(\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\|_2 \\ &= \|\widetilde{\boldsymbol{Y}}^{(k)} - \widetilde{\boldsymbol{Y}}^{(l)}\|_{\mathrm{F}}^2 - 4\epsilon\sum_{i=1}^d \|(\widetilde{\boldsymbol{Y}}^{(k)})_i - (\widetilde{\boldsymbol{Y}}^{(l)})_i\|_2.\end{aligned}$$

It remains to upper bound $\boldsymbol{\Sigma}_{k,l}$. Recall the definition

$$\boldsymbol{\Sigma}_{k,l} = \boldsymbol{U}^\top\big(\boldsymbol{\Sigma}^{(k)} + \widetilde{\boldsymbol{W}}^{(k)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(k)} + \boldsymbol{\Sigma}^{(\ell)} + \widetilde{\boldsymbol{W}}^{(l)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(l)}\big)\boldsymbol{U}.$$

We will utilize the following basic facts: (1) for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with eigenvalues $\{\lambda_i\}$ and $\{\mu_j\}$ respectively, their Kronecker product $A \otimes B$ has eigenvalues $\{\lambda_i\mu_j\}$; (2) For matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ of compatible shapes, it holds that $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C}) \otimes (\boldsymbol{B}\boldsymbol{D})$. These imply that

$$\boldsymbol{0} \preccurlyeq \widetilde{\boldsymbol{W}}^{(k)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(k)} \preccurlyeq \|\widetilde{\boldsymbol{W}}^{(k)}\|\|\widetilde{\boldsymbol{\Gamma}}^{(k)}\|\boldsymbol{I}_{d^2} \preccurlyeq W_{\mathsf{max}}\Gamma_{\mathsf{max}}\boldsymbol{I}_{d^2},$$

and

$$\begin{aligned}\boldsymbol{0} \preccurlyeq \boldsymbol{\Sigma}^{(k)} &= \Big(\widetilde{\boldsymbol{A}}^{(k)} \otimes \boldsymbol{I}_d\Big)\Big((\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2} \otimes (\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2}\Big)\Big(\boldsymbol{I}_{d^2} + \boldsymbol{P}\Big)\Big((\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2} \otimes (\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2}\Big)\Big((\widetilde{\boldsymbol{A}}^{(k)})^\top \otimes \boldsymbol{I}_d\Big) \\ &\preccurlyeq 2\Big(\widetilde{\boldsymbol{A}}^{(k)} \otimes \boldsymbol{I}_d\Big)\Big((\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2} \otimes (\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2}\Big)\Big((\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2} \otimes (\widetilde{\boldsymbol{\Gamma}}^{(k)})^{1/2}\Big)\Big((\widetilde{\boldsymbol{A}}^{(k)})^\top \otimes \boldsymbol{I}_d\Big)\end{aligned}$$

$$
\begin{aligned}
&= 2\Big(\widetilde{\boldsymbol{A}}^{(k)} \otimes \boldsymbol{I}_d\Big)\Big(\widetilde{\boldsymbol{\Gamma}}^{(k)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(k)}\Big)\Big((\widetilde{\boldsymbol{A}}^{(k)})^\top \otimes \boldsymbol{I}_d\Big) \\
&\preccurlyeq 2\Gamma_{\mathsf{max}}^2\Big(\widetilde{\boldsymbol{A}}^{(k)} \otimes \boldsymbol{I}_d\Big)\Big((\widetilde{\boldsymbol{A}}^{(k)})^\top \otimes \boldsymbol{I}_d\Big) \\
&= 2\Gamma_{\mathsf{max}}^2\Big(\widetilde{\boldsymbol{A}}^{(k)}(\widetilde{\boldsymbol{A}}^{(k)})^\top\Big) \otimes \boldsymbol{I}_d \preccurlyeq 2\Gamma_{\mathsf{max}}^2\kappa_A^2\boldsymbol{I}_{d^2}.
\end{aligned}
$$

Using the conditions $W_{\mathsf{max}} \le \Gamma_{\mathsf{max}}$ and $\kappa_A \ge 1$, we have

$$
\boldsymbol{\Sigma}^{(k)} + \widetilde{\boldsymbol{W}}^{(k)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(k)} \preccurlyeq (W_{\mathsf{max}}\Gamma_{\mathsf{max}} + 2\Gamma_{\mathsf{max}}^2\kappa_A^2)\boldsymbol{I}_{d^2} \preccurlyeq 3\Gamma_{\mathsf{max}}^2\kappa_A^2\boldsymbol{I}_{d^2}.
$$

We can upper bound $\boldsymbol{\Sigma}^{(\ell)} + \widetilde{\boldsymbol{W}}^{(l)} \otimes \widetilde{\boldsymbol{\Gamma}}^{(l)}$ by the same analysis. As a result,

$$
\boldsymbol{\Sigma}_{k,l} \preccurlyeq \boldsymbol{U}^\top(6\Gamma_{\mathsf{max}}^2\kappa_A^2\boldsymbol{I}_{d^2})\boldsymbol{U} = 6\Gamma_{\mathsf{max}}^2\kappa_A^2\boldsymbol{U}^\top\boldsymbol{U} = 6\Gamma_{\mathsf{max}}^2\kappa_A^2\boldsymbol{I}_{dK},
$$

which finishes the proof of Lemma C.5.

C.3.3. Proof of (48)

Let ϵ be the right-hand side of the condition (25) on the subspaces. Then, applying the second point of Lemma C.5 (with $\widetilde{\boldsymbol{Y}}^{(k)} = \boldsymbol{Y}^{(k)}, \widetilde{\boldsymbol{Y}}^{(l)} = \boldsymbol{Y}^{(\ell)}$) tells us that

$$
\begin{aligned}
\Big\|\boldsymbol{U}^\top\mathsf{vec}\big((\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)})^\top\big)\Big\|_2^2 &\ge \big\|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\big\|_{\mathrm{F}}^2 - 4\epsilon\sum_{i=1}^d \big\|(\boldsymbol{Y}^{(k)})_i - (\boldsymbol{Y}^{(\ell)})_i\big\|_2 \\
&\ge \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2 - 4\epsilon\sqrt{d}\|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}},
\end{aligned}
$$

where the last line follows from the Cauchy-Schwarz inequality:

$$
\sum_{i=1}^d \big\|(\boldsymbol{Y}^{(k)})_i - (\boldsymbol{Y}^{(\ell)})_i\big\|_2 \le \sqrt{d\sum_{i=1}^d \big\|(\boldsymbol{Y}^{(k)})_i - (\boldsymbol{Y}^{(\ell)})_i\big\|_2^2} = \sqrt{d}\big\|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\big\|_{\mathrm{F}}.
$$

We can lower bound $\|\boldsymbol{V}^\top\mathsf{vec}((\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)})^\top)\|_2^2$ similarly. Putting pieces together, we have

$$
\begin{aligned}
&\Big\|\boldsymbol{V}^\top\mathsf{vec}\big((\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)})^\top\big)\Big\|_2^2 + \Big\|\boldsymbol{U}^\top\mathsf{vec}\big((\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)})^\top\big)\Big\|_2^2 \\
&\quad\ge \|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2 - 4\epsilon\sqrt{d}\big(\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}} + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}\big) \\
&\quad\ge \frac{1}{2}\Big(\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2\Big),
\end{aligned}
$$

where the last inequality is due to the assumption $\epsilon \lesssim \Delta_{\Gamma,Y}/\sqrt{d}$. This completes our proof of (48).

C.4. Proof of Theorem B.3

It suffices to prove the error bounds for one specific value of k, and then take the union bound over $1 \le k \le K$. For notational convenience, in this proof we rewrite $T_{\mathsf{total}}^{(k)}, \boldsymbol{A}^{(k)}, \widehat{\boldsymbol{A}}^{(k)}, \boldsymbol{W}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}$ as $T, \boldsymbol{A}, \widehat{\boldsymbol{A}}, \boldsymbol{W}, \widehat{\boldsymbol{W}}$, respectively. We will investigate the close-form solution $\widehat{\boldsymbol{A}}$, and prepare ourselves with a self-normalized concentration bound; this will be helpful in finally proving the error bounds for $\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|$ and $\|\widehat{\boldsymbol{W}} - \boldsymbol{W}\|$.

Step 1: preparation. Recall the least-squares solution

$$
\widehat{\boldsymbol{A}} = \Big(\sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-1} \boldsymbol{x}_{m,t+1}\boldsymbol{x}_{m,t}{}^\top\Big)\Big(\sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-1} \boldsymbol{x}_{m,t}\boldsymbol{x}_{m,t}{}^\top\Big)^{-1}.
$$

Using the notation

$$
\boldsymbol{X} \coloneqq \begin{bmatrix} \vdots \\ \boldsymbol{x}_{m,t}{}^\top \\ \vdots \end{bmatrix}_{0\le t\le T_m-1, m\in\mathcal{C}_k} \in \mathbb{R}^{T\times d}, \quad \boldsymbol{X}_+ \coloneqq \begin{bmatrix} \vdots \\ \boldsymbol{x}_{m,t+1}{}^\top \\ \vdots \end{bmatrix}, \quad \boldsymbol{N} \coloneqq \begin{bmatrix} \vdots \\ \boldsymbol{w}_{m,t}{}^\top \\ \vdots \end{bmatrix},
$$

we have

$$\boldsymbol{X}_+^\top = \boldsymbol{A}\boldsymbol{X}^\top + \boldsymbol{N}^\top, \quad \widehat{\boldsymbol{A}} = \boldsymbol{X}_+^\top \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \boldsymbol{A} + \boldsymbol{N}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1},$$

namely

$$\boldsymbol{\Delta}_A \coloneqq \widehat{\boldsymbol{A}} - \boldsymbol{A} = \boldsymbol{N}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{55}$$

We will utilize the following matrix form of self-normalized concentration; see Lemma C.4 of (Kakade et al., 2020).

Lemma C.7 (Self-normalized concentration, matrix form)**.** *Consider filtrations $\{\mathcal{F}_t\}$, and random vectors $\{\boldsymbol{x}_t, \boldsymbol{z}_t\}$ satisfying $\boldsymbol{x}_t \in \mathcal{F}_{t-1}$ and $\boldsymbol{z}_t|\mathcal{F}_{t-1} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Let $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ be a fixed, symmetric, positive definite matrix, and denote $\overline{\boldsymbol{V}}_T \coloneqq \sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t^\top + \boldsymbol{V}$. Then with probability at least $1-\delta$,*

$$\Big\|(\overline{\boldsymbol{V}}_T)^{-1/2}\Big(\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{z}_t^\top\Big)\Big\| \lesssim \sqrt{d + \log\frac{1}{\delta} + \log\frac{\det(\overline{\boldsymbol{V}}_T)}{\det(\boldsymbol{V})}}.$$

Step 2: estimation error of $\widehat{\boldsymbol{A}}$. Let us rewrite (55) as

$$\boldsymbol{\Delta}_A^\top = (\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}\cdot(\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}\boldsymbol{X}^\top\boldsymbol{N}. \tag{56}$$

It is obvious that $\boldsymbol{X}^\top\boldsymbol{X}$ plays a crucial role. Recall from Lemma C.1 that with probability at least $1-\delta$, $\|\boldsymbol{x}_{m,t}\|_2 \le D_{\mathsf{vec}}$ for some $D_{\mathsf{vec}} \lesssim \sqrt{\Gamma_{\mathsf{max}}}\cdot\mathsf{poly}(d, \kappa_A, \log(T_{\mathsf{total}}/\delta))$. Then trivially we have the upper bound

$$\boldsymbol{X}^\top\boldsymbol{X} \preccurlyeq D_{\mathsf{vec}}^2 T\cdot\boldsymbol{I}_d =: \boldsymbol{V}_{\mathsf{up}}.$$

For a lower bound, we claim (and prove later) that with probability at least $1-\delta$,

$$\boldsymbol{X}^\top\boldsymbol{X} \succcurlyeq \frac{1}{5}T\cdot\boldsymbol{W} =: \boldsymbol{V}_{\mathsf{lb}}, \quad \text{provided that} \quad T \gtrsim \kappa_w^2 d\cdot\log\Big(\frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}}\frac{\kappa_A d T_{\mathsf{total}}}{\delta}\Big). \tag{57}$$

Now we are ready to control $\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\| = \|\boldsymbol{\Delta}_A\|$. From (56), we have

$$\|\boldsymbol{\Delta}_A\| \le \|(\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}\|\cdot\|(\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}\boldsymbol{X}^\top\boldsymbol{N}\|.$$

First, the lower bound (57) on $\boldsymbol{X}^\top\boldsymbol{X}$ tells us that $\|(\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}\| \lesssim 1/\sqrt{T\cdot\lambda_{\mathsf{min}}(\boldsymbol{W})}$. Moreover, applying Lemma C.7 with $\boldsymbol{V} = \boldsymbol{V}_{\mathsf{lb}}$, one has with probability at least $1-\delta$,

$$\begin{aligned}\|(\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}\boldsymbol{X}^\top\boldsymbol{N}\| &\lesssim \|(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{V})^{-1/2}\boldsymbol{X}^\top\boldsymbol{N}\| \lesssim \|\boldsymbol{W}^{1/2}\|\sqrt{d + \log\frac{1}{\delta} + \log\frac{\det(\boldsymbol{V}_{\mathsf{up}}+\boldsymbol{V}_{\mathsf{lb}})}{\det(\boldsymbol{V}_{\mathsf{lb}})}} \\ &\lesssim \sqrt{\|\boldsymbol{W}\|}\sqrt{d\cdot\log\big(\frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}}\frac{d\kappa_A T_{\mathsf{total}}}{\delta}\big)}.\end{aligned} \tag{58}$$

Putting these together, we have

$$\|\boldsymbol{\Delta}_A\| \lesssim \frac{1}{\sqrt{T\cdot\lambda_{\mathsf{min}}(\boldsymbol{W})}}\cdot\sqrt{\|\boldsymbol{W}\|}\sqrt{d\cdot\log\big(\frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}}\frac{d\kappa_A T_{\mathsf{total}}}{\delta}\big)} \lesssim \sqrt{\frac{d\cdot\kappa_w}{T}\log\big(\frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}}\frac{d\kappa_A T_{\mathsf{total}}}{\delta}\big)},$$

which proves our upper bound for $\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|$ in the theorem.

Step 3: estimation error of $\widehat{\boldsymbol{W}}$. By the definition of $\widehat{\boldsymbol{w}}_{m,t} = \boldsymbol{x}_{m,t+1} - \widehat{\boldsymbol{A}}\boldsymbol{x}_{m,t} = \boldsymbol{w}_{m,t} - \boldsymbol{\Delta}_A\boldsymbol{x}_{m,t}$, we have

$$\widehat{\boldsymbol{N}} \coloneqq \begin{bmatrix}\vdots \\ \widehat{\boldsymbol{w}}_{m,t}^\top \\ \vdots\end{bmatrix} \in \mathbb{R}^{T\times d},$$

$$\widehat{\boldsymbol{N}}^\top = \boldsymbol{N}^\top - \boldsymbol{\Delta}_A\boldsymbol{X}^\top = \boldsymbol{N}^\top - \boldsymbol{N}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{N}^\top\big(\boldsymbol{I}_T - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big).$$

Notice that $\boldsymbol{I}_T - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is a symmetric projection matrix. Therefore,

$$\widehat{\boldsymbol{W}} = \frac{1}{T}\widehat{\boldsymbol{N}}^\top\widehat{\boldsymbol{N}} = \frac{1}{T}\boldsymbol{N}^\top\big(\boldsymbol{I}_T - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\boldsymbol{N},$$

and thus

$$\widehat{\boldsymbol{W}} - \boldsymbol{W} = \Big(\frac{1}{T}\boldsymbol{N}^\top\boldsymbol{N} - \boldsymbol{W}\Big) - \frac{1}{T}\boldsymbol{N}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{N}. \tag{59}$$

The first term on the right-hand side can be controlled by a standard result for covariance estimation (see Proposition D.1): with probability at least $1-\delta$,

$$\Big\|\frac{1}{T}\boldsymbol{N}^\top\boldsymbol{N} - \boldsymbol{W}\Big\| \lesssim \|\boldsymbol{W}\|\sqrt{\frac{d+\log\frac{1}{\delta}}{T}}.$$

As for the second term, we rewrite it as

$$\frac{1}{T}\boldsymbol{N}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{N} = \frac{1}{T}\Big((\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}\boldsymbol{X}^\top\boldsymbol{N}\Big)^\top\Big((\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}\boldsymbol{X}^\top\boldsymbol{N}\Big).$$

Applying our earlier self-normalized concentration bound (58), we have

$$\Big\|\frac{1}{T}\boldsymbol{N}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{N}\Big\| \lesssim \frac{\|\boldsymbol{W}\|d\cdot\log(\frac{\Gamma_{\max}}{W_{\min}}\frac{d\kappa_A T_{\text{total}}}{\delta})}{T}.$$

Putting these back to (59), we have

$$\begin{aligned}\|\widehat{\boldsymbol{W}} - \boldsymbol{W}\| &\le \|\frac{1}{T}\boldsymbol{N}^\top\boldsymbol{N} - \boldsymbol{W}\| + \|\frac{1}{T}\boldsymbol{N}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{N}\| \\ &\lesssim \|\boldsymbol{W}\|\sqrt{\frac{d+\log\frac{1}{\delta}}{T}} + \|\boldsymbol{W}\|\frac{d\cdot\log(\frac{\Gamma_{\max}}{W_{\min}}\frac{d\kappa_A T_{\text{total}}}{\delta})}{T} \lesssim \|\boldsymbol{W}\|\sqrt{\frac{d\cdot\log(\frac{\Gamma_{\max}}{W_{\min}}\frac{d\kappa_A T_{\text{total}}}{\delta})}{T}},\end{aligned}$$

where the last inequality uses $T \gtrsim d\cdot\log(\frac{\Gamma_{\max}}{W_{\min}}\frac{d\kappa_A T_{\text{total}}}{\delta})$. This finishes our proof of Theorem B.3.

C.4.1. Proof of (57)

We start with the following decomposition:

$$\begin{aligned}\boldsymbol{X}^\top\boldsymbol{X} &= \sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-1}\boldsymbol{x}_{m,t}\boldsymbol{x}_{m,t}^\top \succcurlyeq \sum_{m\in\mathcal{C}_k}\sum_{1\le t\le T_m-1}\boldsymbol{x}_{m,t}\boldsymbol{x}_{m,t}^\top \\ &= \sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-2}\boldsymbol{x}_{m,t+1}\boldsymbol{x}_{m,t+1}^\top = \sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-2}(\boldsymbol{A}\boldsymbol{x}_{m,t}+\boldsymbol{w}_{m,t})(\boldsymbol{A}\boldsymbol{x}_{m,t}+\boldsymbol{w}_{m,t})^\top \\ &= \underbrace{\sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-2}\boldsymbol{w}_{m,t}\boldsymbol{w}_{m,t}^\top}_{:=\boldsymbol{P}} \\ &\quad + \underbrace{\sum_{m\in\mathcal{C}_k}\sum_{0\le t\le T_m-2}\Big(\boldsymbol{A}\boldsymbol{x}_{m,t}\boldsymbol{x}_{m,t}^\top\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{x}_{m,t}\boldsymbol{w}_{m,t}^\top + \boldsymbol{w}_{m,t}\boldsymbol{x}_{m,t}^\top\boldsymbol{A}^\top\Big)}_{=:\boldsymbol{Q}} \\ &= \boldsymbol{P}+\boldsymbol{Q}.\end{aligned} \tag{60}$$

Lower bound for $\boldsymbol{P}$. By Proposition D.1, we have with probability at least $1-\delta$,

$$\|\frac{1}{T-|\mathcal{C}_k|}\boldsymbol{P} - \boldsymbol{W}\| \lesssim \|\boldsymbol{W}\|\sqrt{\frac{d+\log\frac{1}{\delta}}{T}} \lesssim \lambda_{\min}(\boldsymbol{W})\cdot\boldsymbol{I}_d, \quad \text{provided that} \quad T \gtrsim \kappa_w^2(d+\log\frac{1}{\delta}).$$

As a result, we have $\frac{1}{T-|\mathcal{C}_k|}\boldsymbol{P} \succcurlyeq \frac{1}{2}\boldsymbol{W}$, which implies $\boldsymbol{P} \succcurlyeq \frac{1}{4}T\cdot\boldsymbol{W}$.

Lower bound for $\boldsymbol{Q}$. Let $\mathcal{N}_\epsilon$ be an ϵ-net of the unit sphere $\mathcal{S}^{d-1}$ (where the value of $0 < \epsilon < 1$ will be specified later), and let π be the projection onto $\mathcal{N}_\epsilon$. Recall the standard result that $|\mathcal{N}_\epsilon| \le (9/\epsilon)^d$. Moreover, for any $\boldsymbol{v} \in \mathcal{S}^{d-1}$, denote $\boldsymbol{\Delta}_{\boldsymbol{v}} \coloneqq \boldsymbol{v} - \pi(\boldsymbol{v})$, which satisfies $\|\boldsymbol{\Delta}_{\boldsymbol{v}}\|_2 \le \epsilon$. Then we have

$$\begin{aligned}\lambda_{\mathsf{min}}(\boldsymbol{Q}) &= \inf_{\boldsymbol{v}\in\mathcal{S}^{d-1}} \boldsymbol{v}^\top \boldsymbol{Q}\boldsymbol{v} = \inf_{\boldsymbol{v}\in\mathcal{S}^{d-1}} (\pi(\boldsymbol{v}) + \boldsymbol{\Delta}_{\boldsymbol{v}})^\top \boldsymbol{Q}(\pi(\boldsymbol{v}) + \boldsymbol{\Delta}_{\boldsymbol{v}}) \\ &\ge \inf_{\boldsymbol{v}\in\mathcal{S}^{d-1}} \pi(\boldsymbol{v})^\top \boldsymbol{Q}\pi(\boldsymbol{v}) - (2\epsilon + \epsilon^2)\|\boldsymbol{Q}\| \ge \inf_{\boldsymbol{v}\in\mathcal{N}_\epsilon} \boldsymbol{v}^\top \boldsymbol{Q}\boldsymbol{v} - 3\epsilon\|\boldsymbol{Q}\|. \end{aligned} \tag{61}$$

For $\|\boldsymbol{Q}\|$, we simply use a crude upper bound, based on the boundedness of $\{\|\boldsymbol{x}_{m,t}\|_2\}$ (Lemma C.1): with probability at least $1-\delta$, one has

$$\|\boldsymbol{Q}\| \le \sum_{m\in\mathcal{C}_k} \sum_{0\le t\le T_m-2} \Big(\|\boldsymbol{A}\boldsymbol{x}_{m,t}\|_2^2 + 2\|\boldsymbol{A}\boldsymbol{x}_{m,t}\|_2\|\boldsymbol{w}_{m,t}\|_2\Big) \lesssim \Gamma_{\mathsf{max}} \cdot \mathsf{poly}\Big(\kappa_A, d, T_{\mathsf{total}}, \frac{1}{\delta}\Big). \tag{62}$$

Next, we lower bound $\inf_{\boldsymbol{v}\in\mathcal{N}_\epsilon} \boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v}$. First, consider a fixed $\boldsymbol{v} \in \mathcal{N}_\epsilon$; denoting $\boldsymbol{y}_{m,t} \coloneqq \boldsymbol{A}\boldsymbol{x}_{m,t}$ and $\boldsymbol{u}_{m,t} \coloneqq \boldsymbol{W}^{-1/2}\boldsymbol{w}_{m,t} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, we have

$$\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v} = \sum_{m\in\mathcal{C}_k} \sum_{0\le t\le T_m-2} (\boldsymbol{v}^\top\boldsymbol{y}_{m,t})^2 + 2\sum_{m\in\mathcal{C}_k} \sum_{0\le t\le T_m-2} \boldsymbol{v}^\top\boldsymbol{y}_{m,t}\cdot\boldsymbol{u}_{m,t}^\top\boldsymbol{W}^{1/2}\boldsymbol{v},$$

where $\boldsymbol{u}_{m,t}^\top\boldsymbol{W}^{1/2}\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{v}^\top\boldsymbol{W}\boldsymbol{v})$. Lemma D.2 (the scalar version of self-normalized concentration) tells us that, for any fixed $\lambda > 0$, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{m\in\mathcal{C}_k} \sum_{0\le t\le T_m-2} \boldsymbol{v}^\top\boldsymbol{y}_{m,t}\cdot\boldsymbol{u}_{m,t}^\top\boldsymbol{W}^{1/2}\boldsymbol{v} &\ge -\sqrt{\boldsymbol{v}^\top\boldsymbol{W}\boldsymbol{v}}\Big(\frac{\lambda}{2}\sum_{m\in\mathcal{C}_k} \sum_{0\le t\le T_m-2} (\boldsymbol{v}^\top\boldsymbol{y}_{m,t})^2 + \frac{1}{\lambda}\log\frac{1}{\delta}\Big) \\ &\ge -\sqrt{\|\boldsymbol{W}\|}\Big(\frac{\lambda}{2}\sum_{m\in\mathcal{C}_k} \sum_{0\le t\le T_m-2} (\boldsymbol{v}^\top\boldsymbol{y}_{m,t})^2 + \frac{1}{\lambda}\log\frac{1}{\delta}\Big).\end{aligned}$$

Replacing δ with $\delta/(9/\epsilon)^d$ and taking the union bound, we have with probability at least $1-\delta$, for any $\boldsymbol{v} \in \mathcal{N}_\epsilon$,

$$\begin{aligned}\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v} &\ge \sum_{m\in\mathcal{C}_k} \sum_{0\le t\le T_m-2} (\boldsymbol{v}^\top\boldsymbol{y}_{m,t})^2 - \sqrt{\|\boldsymbol{W}\|}\Big(\lambda\sum_{m\in\mathcal{C}_k} \sum_{0\le t\le T_m-2} (\boldsymbol{v}^\top\boldsymbol{y}_{m,t})^2 + \frac{2}{\lambda}\big(d\cdot\log\frac{9}{\epsilon} + \log\frac{1}{\delta}\big)\Big) \\ &= \big(1 - \sqrt{\|\boldsymbol{W}\|}\lambda\big)\sum_{m\in\mathcal{C}_k} \sum_{0\le t\le T_m-2} (\boldsymbol{v}^\top\boldsymbol{y}_{m,t})^2 - \sqrt{\|\boldsymbol{W}\|}\,\frac{2}{\lambda}\Big(d\cdot\log\frac{9}{\epsilon} + \log\frac{1}{\delta}\Big).\end{aligned}$$

With the choice of $\lambda = 1/\sqrt{\|\boldsymbol{W}\|}$, this implies

$$\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v} \ge -2\|\boldsymbol{W}\|(d\cdot\log\frac{9}{\epsilon} + \log\frac{1}{\delta}), \quad \text{for all} \quad \boldsymbol{v} \in \mathcal{N}_\epsilon. \tag{63}$$

Putting (62) and (63) back to (61), we have with probability at least $1-\delta$,

$$\lambda_{\mathsf{min}}(\boldsymbol{Q}) \ge \inf_{\boldsymbol{v}\in\mathcal{N}_\epsilon} \boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v} - 3\epsilon\|\boldsymbol{Q}\| \ge -C_0\Big(\|\boldsymbol{W}\|\big(d\cdot\log\frac{9}{\epsilon} + \log\frac{1}{\delta}\big) + \epsilon\,\Gamma_{\mathsf{max}}\cdot\mathsf{poly}\big(\kappa_A, d, T_{\mathsf{total}}, \frac{1}{\delta}\big)\Big)$$

for some universal constant $C_0 > 0$.

Putting things together. Recall the decomposition $\boldsymbol{X}^\top\boldsymbol{X} \succcurlyeq \boldsymbol{P} + \boldsymbol{Q}$ in (60). We have already shown that if $T \gtrsim \kappa_w^2(d + \log\frac{1}{\delta})$, then with probability at least $1-\delta$,

$$\boldsymbol{X}^\top\boldsymbol{X} \succcurlyeq \boldsymbol{P} + \boldsymbol{Q} \succcurlyeq \frac{1}{4}T\cdot\boldsymbol{W} - C_0\Big(\|\boldsymbol{W}\|\big(d\cdot\log\frac{9}{\epsilon} + \log\frac{1}{\delta}\big) + \epsilon\,\Gamma_{\mathsf{max}}\cdot\mathsf{poly}\big(\kappa_A, d, T_{\mathsf{total}}, \frac{1}{\delta}\big)\Big)\boldsymbol{I}_d.$$

It is easy to check that, if we further choose

$$\epsilon \asymp \frac{1}{\mathsf{poly}(\kappa_A, d, T_{\mathsf{total}}, \frac{1}{\delta}, \frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}})}, \quad T \gtrsim \kappa_w^2 d\cdot\log\Big(\frac{\Gamma_{\mathsf{max}}}{W_{\mathsf{min}}}\frac{\kappa_A d T_{\mathsf{total}}}{\delta}\Big),$$

then we have $\boldsymbol{X}^\top\boldsymbol{X} \succcurlyeq \frac{1}{5}T\cdot\boldsymbol{W}$, which finishes the proof of (57).

C.5. Proof of Theorem B.4

In this proof, we show the correct classification of one short trajectory $\{\boldsymbol{x}_{m,t}\}_{0\le t\le T_m}$ (with true label k_m) for some $m \in \mathcal{M}_{\mathsf{classification}}$; then it suffices to take the union bound to prove the correct classification of all trajectories in $\mathcal{M}_{\mathsf{classification}}$. For notational simplicity, we drop the subscript m and rewrite $\{\boldsymbol{x}_{m,t}\}$, T_m, k_m as $\{\boldsymbol{x}_t\}$, T, k, respectively. The basic idea of this proof is to show that

$$L(\widehat{\boldsymbol{A}}^{(\ell)}, \widehat{\boldsymbol{W}}^{(\ell)}) > L(\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}) \tag{64}$$

for any incorrect label $\ell \neq k$, where L is the loss function defined in (23) and used by Algorithm 5 for classification. Our proof below is simply a sequence of arguments that finally transform (64) into a sufficient condition in terms of the coarse model errors ϵ_A, ϵ_W and short trajectory length T.

Before we proceed, we record a few basic facts that will be useful later. First, the assumption $\|\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(k)}\| \le \epsilon_W \le 0.1 W_{\mathsf{min}}$ implies that $\lambda_{\mathsf{min}}(\widehat{\boldsymbol{W}}^{(k)}) \ge 0.9\lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)})$, and $\widehat{\boldsymbol{W}}^{(k)}$ is well conditioned with $\kappa(\widehat{\boldsymbol{W}}^{(k)}) \lesssim \kappa(\boldsymbol{W}^{(k)}) \le \kappa_w$. Morever, by Lemma C.1, with probability at least $1-\delta$, we have for all $0 \le t \le T$,

- In Case 0, $\|\boldsymbol{x}_t\|_2 \le D_x \lesssim \sqrt{\Gamma_{\mathsf{max}}(d + \log\frac{T_{\mathsf{total}}}{\delta})}$, provided that $T \gtrsim 1$;
- In Case 1, $\|\boldsymbol{x}_t\|_2 \le D_x \lesssim \kappa_A\sqrt{\Gamma_{\mathsf{max}}(d + \log\frac{T_{\mathsf{total}}}{\delta})}$, provided that $T \gtrsim \frac{1}{1-\rho}\log(2\kappa_A)$.

Now we are ready to state our proof. Throughout our analyses, we will make some intermediate claims, whose proofs will be deferred to the end of this subsection.

Step 1: a sufficient condition for correct classification. In the following, we prove that for a fixed $\ell \neq k$, the condition (64) holds with high probability; at the end of the proof, we simply take the union bound over $\ell \neq k$. Using $\boldsymbol{x}_{t+1} = \boldsymbol{A}^{(k)}\boldsymbol{x}_t + \boldsymbol{w}_t$ where $\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}^{(k)})$, we can rewrite the loss function L as

$$\begin{aligned} L(\boldsymbol{A}, \boldsymbol{W}) = T\cdot\log\det(\boldsymbol{W}) + \sum_{t=0}^{T-1} \boldsymbol{w}_t{}^\top \boldsymbol{W}^{-1}\boldsymbol{w}_t \\ + \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top(\boldsymbol{A}^{(k)} - \boldsymbol{A})^\top \boldsymbol{W}^{-1}(\boldsymbol{A}^{(k)} - \boldsymbol{A})\boldsymbol{x}_t + 2\sum_{t=0}^{T-1} \boldsymbol{w}_t{}^\top \boldsymbol{W}^{-1}(\boldsymbol{A}^{(k)} - \boldsymbol{A})\boldsymbol{x}_t. \end{aligned}$$

After some basic calculation, (64) can be equivalently written as

$$\begin{aligned} &L(\widehat{\boldsymbol{A}}^{(\ell)}, \widehat{\boldsymbol{W}}^{(\ell)}) - L(\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}) = \mathrm{(A)} + \mathrm{(B)} - \mathrm{(C)} > 0, \\ \text{where}\quad &\mathrm{(A)} \coloneqq T\cdot\Big(\log\det(\widehat{\boldsymbol{W}}^{(\ell)}) - \log\det(\widehat{\boldsymbol{W}}^{(k)})\Big) + \sum_{t=0}^{T-1} \boldsymbol{w}_t{}^\top\Big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} - (\widehat{\boldsymbol{W}}^{(k)})^{-1}\Big)\boldsymbol{w}_t \\ &\mathrm{(B)} \coloneqq \Big(\sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})^\top(\widehat{\boldsymbol{W}}^{(\ell)})^{-1}(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})\boldsymbol{x}_t + 2\sum_{t=0}^{T-1} \boldsymbol{w}_t{}^\top(\widehat{\boldsymbol{W}}^{(\ell)})^{-1}(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})\boldsymbol{x}_t\Big) \\ &\mathrm{(C)} \coloneqq \Big(\sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(k)})^\top(\widehat{\boldsymbol{W}}^{(k)})^{-1}(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(k)})\boldsymbol{x}_t + 2\sum_{t=0}^{T-1} \boldsymbol{w}_t{}^\top(\widehat{\boldsymbol{W}}^{(k)})^{-1}(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(k)})\boldsymbol{x}_t\Big) \end{aligned}$$

Step 2: a lower bound for $\mathrm{(A)} + \mathrm{(B)} - \mathrm{(C)}$**.** Intuitively, we expect that $\mathrm{(A)} + \mathrm{(B)}$ should be large because the LDS models $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})$ and $(\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$ are well separated, while $\mathrm{(C)}$ should be small if $(\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)}) \approx (\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})$. More formally, we claim that the following holds for some universal constants $C_1, C_2, C_3 > 0$:

- With probability at least $1-\delta$,

$$\begin{aligned} \mathrm{(A)} \ge T\cdot\Big[\log\det(\widehat{\boldsymbol{W}}^{(\ell)}) - \log\det(\widehat{\boldsymbol{W}}^{(k)}) + \mathsf{Tr}\Big((\boldsymbol{W}^{(k)})^{1/2}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} - (\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)(\boldsymbol{W}^{(k)})^{1/2}\Big)\Big] \\ - C_1\sqrt{T}\Big\|(\boldsymbol{W}^{(k)})^{1/2}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} - (\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)(\boldsymbol{W}^{(k)})^{1/2}\Big\|_{\mathrm{F}}\log\frac{1}{\delta}; \end{aligned} \tag{65}$$

- With probability at least $1-\delta$,

$$\text{(B)} \geq C_2\bigg(T\frac{\|\boldsymbol{A}^{(k)}-\widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} - \kappa_w\kappa_{w,\mathsf{cross}}\log\frac{1}{\delta}\bigg), \tag{66}$$

  provided that $T \gtrsim \kappa_w^2\log^2(1/\delta)$;

- With probability at least $1-\delta$,

$$\text{(C)} \leq C_3\bigg(T\frac{D_x^2\|\boldsymbol{A}^{(k)}-\widehat{\boldsymbol{A}}^{(k)}\|^2}{W_{\mathsf{min}}} + \kappa_w\log\frac{1}{\delta}\bigg), \tag{67}$$

  provided that $T \gtrsim 1$ (under Case 0) or $T \gtrsim \frac{1}{1-\rho}\log(2\kappa_A)$ (under Case 1).

Putting these together, we have with probability at least $1-\delta$,

$$\begin{aligned}\text{(A)}+\text{(B)}-\text{(C)} \geq C_4\cdot T\cdot\bigg[&\log\frac{\det(\widehat{\boldsymbol{W}}^{(\ell)})}{\det(\widehat{\boldsymbol{W}}^{(k)})} + \mathsf{Tr}\Big(\boldsymbol{W}^{(k)}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1}-(\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)\Big)\\ &+\frac{\|\boldsymbol{A}^{(k)}-\widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} - \frac{D_x^2\|\boldsymbol{A}^{(k)}-\widehat{\boldsymbol{A}}^{(k)}\|^2}{W_{\mathsf{min}}}\bigg]\\ &- C_5\bigg[\sqrt{T}\Big\|(\boldsymbol{W}^{(k)})^{1/2}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1}-(\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)(\boldsymbol{W}^{(k)})^{1/2}\Big\|_{\mathrm{F}} + \kappa_w\kappa_{w,\mathsf{cross}}\bigg]\log\frac{1}{\delta}\end{aligned}$$

for some universal constants $C_4, C_5 > 0$, provided that $T \gtrsim \kappa_w^2\log^2\frac{1}{\delta}$ (under Case 0), or $T \gtrsim \kappa_w^2\log^2\frac{1}{\delta} + \frac{1}{1-\rho}\log(2\kappa_A)$ (under Case 1). Now we have a lower bound of (A) + (B) − (C) as an order-T term minus a low-order term. Therefore, to show (A) + (B) − (C) > 0, it suffices to prove that (a) the leading factor of order-T term is positive and large, and (b) the low-order term is negligible compared with the order-T term. More specifically, under the assumption that the coarse models $\{\widehat{\boldsymbol{A}}^{(j)}, \widehat{\boldsymbol{W}}^{(j)}\}$ satisfy $\|\widehat{\boldsymbol{A}}^{(j)}-\boldsymbol{A}^{(j)}\| \leq \epsilon_A$, $\|\widehat{\boldsymbol{W}}^{(j)}-\boldsymbol{W}^{(j)}\| \leq \epsilon_W \leq 0.1W_{\mathsf{min}}$ for all $1 \leq j \leq K$, we make the following claims:

- (Order-T term is large.) Define the leading factor $\widehat{D}_{k,\ell}$ of the order-T term and a related parameter $D_{k,\ell}$ as follows:

$$\begin{aligned}\widehat{D}_{k,\ell} &\coloneqq \log\frac{\det(\widehat{\boldsymbol{W}}^{(\ell)})}{\det(\widehat{\boldsymbol{W}}^{(k)})} + \mathsf{Tr}\Big(\boldsymbol{W}^{(k)}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1}-(\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)\Big) + \frac{\|\boldsymbol{A}^{(k)}-\widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} - \frac{D_x^2\|\boldsymbol{A}^{(k)}-\widehat{\boldsymbol{A}}^{(k)}\|^2}{W_{\mathsf{min}}},\\ D_{k,\ell} &\coloneqq \frac{1}{\kappa_{w,\mathsf{cross}}}\bigg(\frac{\|\boldsymbol{W}^{(k)}-\boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} + \|\boldsymbol{A}^{(k)}-\boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2\bigg) \gtrsim \frac{\Delta_{A,W}^2}{\kappa_{w,\mathsf{cross}}},\end{aligned} \tag{68}$$

  where the last inequality follows from Assumption 3.1. They are related in the sense that

$$D_{k,\ell} \lesssim \widetilde{D}_{k,\ell} \coloneqq \log\frac{\det(\boldsymbol{W}^{(\ell)})}{\det(\boldsymbol{W}^{(k)})} + \mathsf{Tr}\Big(\boldsymbol{W}^{(k)}\big((\boldsymbol{W}^{(\ell)})^{-1}-(\boldsymbol{W}^{(k)})^{-1}\big)\Big) + \frac{\|\boldsymbol{A}^{(k)}-\boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}}, \tag{69}$$

  where $\widetilde{D}_{k,\ell}$ is defined in the same way as $\widehat{D}_{k,\ell}$, except that the coarse models are replaced with the accurate ones; moreover, we have

$$D_{k,\ell} \lesssim \widehat{D}_{k,\ell}, \quad \text{provided that} \quad \epsilon_A \lesssim \sqrt{\frac{W_{\mathsf{min}}D_{k,\ell}}{D_x^2}}, \quad \epsilon_W \lesssim W_{\mathsf{min}}\sqrt{\frac{D_{k,\ell}}{d}}. \tag{70}$$

- (Low-order term is negligible.) With (70) in place, we have

$$\begin{aligned}&\text{(A)}+\text{(B)}-\text{(C)}\\ &\geq C_6T\cdot D_{k,\ell} - C_5\bigg[\sqrt{T}\Big\|(\boldsymbol{W}^{(k)})^{1/2}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1}-(\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)(\boldsymbol{W}^{(k)})^{1/2}\Big\|_{\mathrm{F}} + \kappa_w\kappa_{w,\mathsf{cross}}\bigg]\log\frac{1}{\delta}.\end{aligned} \tag{71}$$

  We claim that

$$\text{if} \quad T \gtrsim \bigg(\frac{\kappa_{w,\mathsf{cross}}^5}{D_{k,\ell}}+1\bigg)\log^2\frac{1}{\delta}, \quad \text{then} \quad \text{(A)}+\text{(B)}-\text{(C)} \geq C_7T\cdot D_{k,\ell} > 0. \tag{72}$$

Step 3: putting things together. So far, we have proved that for a short trajectory generated by $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)})$ and for a fixed $\ell \neq k$, it holds with probability at least $1-\delta$ that $L(\widehat{\boldsymbol{A}}^{(\ell)}, \widehat{\boldsymbol{W}}^{(\ell)}) > L(\widehat{\boldsymbol{A}}^{(k)}, \widehat{\boldsymbol{W}}^{(k)})$, provided that

$$\epsilon_A \lesssim \sqrt{\frac{W_{\mathsf{min}} D_{k,\ell}}{D_x^2}}, \quad \epsilon_W \lesssim W_{\mathsf{min}} \cdot \min\Big\{1, \sqrt{\frac{D_{k,\ell}}{d}}\Big\}, \quad T \gtrsim \begin{cases} \Big(\kappa_w^2 + \frac{\kappa_{w,\mathsf{cross}}^5}{D_{k,\ell}}\Big)\log^2\frac{1}{\delta} & \text{for Case 0,} \\ \Big(\kappa_w^2 + \frac{\kappa_{w,\mathsf{cross}}^5}{D_{k,\ell}}\Big)\log^2\frac{1}{\delta} + \frac{1}{1-\rho}\log(2\kappa_A) & \text{for Case 1.} \end{cases}$$

Plugging in the relation $D_{k,\ell} \gtrsim \Delta_{A,W}^2/\kappa_{w,\mathsf{cross}}$ and $D_x \lesssim \sqrt{\Gamma_{\mathsf{max}}(d + \log\frac{T_{\mathsf{total}}}{\delta})}$ (for Case 0) or $D_x \lesssim \kappa_A\sqrt{\Gamma_{\mathsf{max}}(d + \log\frac{T_{\mathsf{total}}}{\delta})}$ (for Case 1), the above conditions become

$$\begin{aligned} \text{For Case 0:} \quad & \epsilon_A \lesssim \sqrt{\frac{W_{\mathsf{min}}\Delta_{A,W}^2}{\Gamma_{\mathsf{max}}\kappa_{w,\mathsf{cross}}(d+\log\frac{T_{\mathsf{total}}}{\delta})}}, \quad \epsilon_W \lesssim W_{\mathsf{min}} \cdot \min\Big\{1, \frac{\Delta_{A,W}}{\sqrt{\kappa_{w,\mathsf{cross}} d}}\Big\}, \\ & T \gtrsim \Big(\kappa_w^2 + \frac{\kappa_{w,\mathsf{cross}}^6}{\Delta_{A,W}^2}\Big)\log^2\frac{1}{\delta}; \\ \text{For Case 1:} \quad & \epsilon_A \lesssim \sqrt{\frac{W_{\mathsf{min}}\Delta_{A,W}^2}{\Gamma_{\mathsf{max}}\kappa_{w,\mathsf{cross}}\kappa_A^2(d+\log\frac{T_{\mathsf{total}}}{\delta})}}, \quad \epsilon_W \lesssim W_{\mathsf{min}} \cdot \min\Big\{1, \frac{\Delta_{A,W}}{\sqrt{\kappa_{w,\mathsf{cross}} d}}\Big\}, \\ & T \gtrsim \Big(\kappa_w^2 + \frac{\kappa_{w,\mathsf{cross}}^6}{\Delta_{A,W}^2}\Big)\log^2\frac{1}{\delta} + \frac{1}{1-\rho}\log(2\kappa_A). \end{aligned}$$

Finally, taking the union bound over all $\ell \neq k$ as well as over all trajectories in $\mathcal{M}_{\mathsf{classification}}$ finishes the proof of Theorem B.4.

C.5.1. Proof of (65).

Since $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}^{(k)})$, we have

$$\sum_{t=0}^{T-1} \boldsymbol{w}_t{}^\top\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} - (\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)\boldsymbol{w}_t = \boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z},$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{Td})$, and $\boldsymbol{M} \in \mathbb{R}^{Td\times Td}$ is a block-diagonal matrix with $\boldsymbol{Q} := (\boldsymbol{W}^{(k)})^{1/2}((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} - (\widehat{\boldsymbol{W}}^{(k)})^{-1})(\boldsymbol{W}^{(k)})^{1/2} \in \mathbb{R}^{d\times d}$ as its diagonal blocks. Therefore, by the Hanson-Wright inequality (Theorem 6.2.1 of (Vershynin, 2018)), we have

$$\mathbb{P}\bigg(\Big|\boldsymbol{z}^\top\boldsymbol{M}\boldsymbol{z} - \mathbb{E}[\boldsymbol{z}^\top\boldsymbol{M}\boldsymbol{z}]\Big| \geq u\bigg) \leq 2\exp\bigg(-c\min\Big\{\frac{u^2}{\|\boldsymbol{M}\|_{\mathrm{F}}^2}, \frac{u}{\|\boldsymbol{M}\|}\Big\}\bigg),$$

where $\|\boldsymbol{M}\|_{\mathrm{F}}^2 = T\|\boldsymbol{Q}\|_{\mathrm{F}}^2$, $\|\boldsymbol{M}\| = \|\boldsymbol{Q}\|$, and $\mathbb{E}[\boldsymbol{z}^\top\boldsymbol{M}\boldsymbol{z}] = \mathsf{Tr}(\boldsymbol{M}) = T\cdot\mathsf{Tr}(\boldsymbol{Q})$. Choosing $u \gtrsim \sqrt{T}\|\boldsymbol{Q}\|_{\mathrm{F}}\log\frac{1}{\delta}$, we have with probability at least $1-\delta$, $|\boldsymbol{z}^\top\boldsymbol{M}\boldsymbol{z} - T\cdot\mathsf{Tr}(\boldsymbol{Q})| \leq C_1\sqrt{T}\|\boldsymbol{Q}\|_{\mathrm{F}}\log\frac{1}{\delta}$, which immediately leads to our lower bound (65) for (A).

C.5.2. Proof of (66).

Denote $\boldsymbol{u}_t = (\boldsymbol{W}^{(k)})^{-1/2}\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and $\boldsymbol{y}_t = (\boldsymbol{W}^{(k)})^{1/2}(\widehat{\boldsymbol{W}}^{(\ell)})^{-1}(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})\boldsymbol{x}_t$. Then we have

$$\text{(B)} = \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})^\top(\widehat{\boldsymbol{W}}^{(\ell)})^{-1}(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})\boldsymbol{x}_t + 2\sum_{t=0}^{T-1}\boldsymbol{u}_t{}^\top\boldsymbol{y}_t.$$

By Lemma D.2, we have with probability at least $1-\delta$, $\sum_{t=0}^{T-1}\boldsymbol{u}_t{}^\top\boldsymbol{y}_t \geq -(\frac{\lambda}{2}\sum_{t=0}^{T-1}\|\boldsymbol{y}_t\|_2^2 + \frac{1}{\lambda}\log\frac{2}{\delta})$ for any fixed $\lambda > 0$. This implies that

$$\text{(B)} \geq \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})^\top(\widehat{\boldsymbol{W}}^{(\ell)})^{-1}(\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})\boldsymbol{x}_t - \lambda\sum_{t=0}^{T-1}\boldsymbol{y}_t{}^\top\boldsymbol{y}_t - \frac{2}{\lambda}\log\frac{2}{\delta}$$

$$= \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top (\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})^\top \Big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} - \lambda \cdot (\widehat{\boldsymbol{W}}^{(\ell)})^{-1} \boldsymbol{W}^{(k)} (\widehat{\boldsymbol{W}}^{(\ell)})^{-1} \Big) (\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}) \boldsymbol{x}_t - \frac{2}{\lambda} \log \frac{2}{\delta}.$$

Choosing $\lambda = 0.05/(\kappa_w \kappa_{w,\mathsf{cross}})$, we have $\lambda \cdot (\widehat{\boldsymbol{W}}^{(\ell)})^{-1} \boldsymbol{W}^{(k)} (\widehat{\boldsymbol{W}}^{(\ell)})^{-1} \preccurlyeq 0.1 (\widehat{\boldsymbol{W}}^{(\ell)})^{-1}$, and thus

$$\begin{aligned}
\text{(B)} &\geq 0.9 \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top (\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})^\top (\widehat{\boldsymbol{W}}^{(\ell)})^{-1} (\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}) \boldsymbol{x}_t - 40 \kappa_w \kappa_{w,\mathsf{cross}} \log \frac{2}{\delta} \\
&\geq 0.9 \lambda_{\mathsf{min}}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} \big) \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top (\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)})^\top (\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}) \boldsymbol{x}_t - 40 \kappa_w \kappa_{w,\mathsf{cross}} \log \frac{2}{\delta}. \qquad (73)
\end{aligned}$$

Now it remains to lower bound $\sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta} \boldsymbol{x}_t$, where $\boldsymbol{\Delta} \coloneqq \boldsymbol{\Delta}_A^\top \boldsymbol{\Delta}_A$ and $\boldsymbol{\Delta}_A \coloneqq \boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}$. Since $\boldsymbol{\Delta} \succcurlyeq \boldsymbol{0}$, we have

$$\begin{aligned}
\sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta} \boldsymbol{x}_t &\geq \sum_{t=0}^{T-2} \boldsymbol{x}_{t+1}{}^\top \boldsymbol{\Delta} \boldsymbol{x}_{t+1} = \sum_{t=0}^{T-2} (\boldsymbol{A}^{(k)} \boldsymbol{x}_t + \boldsymbol{w}_t)^\top \boldsymbol{\Delta} (\boldsymbol{A}^{(k)} \boldsymbol{x}_t + \boldsymbol{w}_t) \\
&= \underbrace{\sum_{t=0}^{T-2} \boldsymbol{w}_t{}^\top \boldsymbol{\Delta} \boldsymbol{w}_t}_{\text{(i)}} + \underbrace{\sum_{t=0}^{T-2} \boldsymbol{x}_t{}^\top \boldsymbol{A}^{(k)}{}^\top \boldsymbol{\Delta} \boldsymbol{A}^{(k)} \boldsymbol{x}_t + 2 \sum_{t=0}^{T-2} \boldsymbol{w}_t{}^\top \boldsymbol{\Delta} \boldsymbol{A}^{(k)} \boldsymbol{x}_t}_{\text{(ii)}}. \qquad (74)
\end{aligned}$$

We can lower bound (i) by the Hanson-Wright inequality, similar to our previous proof of (65); the result is that, with probability at least $1 - \delta$, one has

$$\text{(i)} \geq T \cdot \mathsf{Tr}\Big((\boldsymbol{W}^{(k)})^{1/2} \boldsymbol{\Delta} (\boldsymbol{W}^{(k)})^{1/2} \Big) - C_0 \sqrt{T} \big\| (\boldsymbol{W}^{(k)})^{1/2} \boldsymbol{\Delta} (\boldsymbol{W}^{(k)})^{1/2} \big\|_{\mathrm{F}} \log \frac{1}{\delta}.$$

To lower bound (ii), we apply Lemma D.2, which shows that with probability at least $1 - \delta$,

$$\Big| \sum_{t=0}^{T-2} \boldsymbol{w}_t{}^\top \boldsymbol{\Delta} \boldsymbol{A}^{(k)} \boldsymbol{x}_t \Big| \leq \frac{\lambda}{2} \sum_{t=0}^{T-2} \boldsymbol{x}_t{}^\top (\boldsymbol{A}^{(k)})^\top \boldsymbol{\Delta} \boldsymbol{W}^{(k)} \boldsymbol{\Delta} \boldsymbol{A}^{(k)} \boldsymbol{x}_t + \frac{1}{\lambda} \log \frac{2}{\delta},$$

for any fixed $\lambda > 0$, hence

$$\text{(ii)} \geq \sum_{t=0}^{T-2} \boldsymbol{x}_t{}^\top \boldsymbol{A}^{(k)\top} (\boldsymbol{\Delta} - \lambda \cdot \boldsymbol{\Delta} \boldsymbol{W}^{(k)} \boldsymbol{\Delta}) \boldsymbol{A}^{(k)} \boldsymbol{x}_t - \frac{2}{\lambda} \log \frac{2}{\delta}.$$

Recall $\boldsymbol{\Delta} = \boldsymbol{\Delta}_A^\top \boldsymbol{\Delta}_A$, and thus $\boldsymbol{\Delta} - \lambda \cdot \boldsymbol{\Delta} \boldsymbol{W}^{(k)} \boldsymbol{\Delta} = \boldsymbol{\Delta}_A^\top (\boldsymbol{I}_d - \lambda \cdot \boldsymbol{\Delta}_A \boldsymbol{W}^{(k)} \boldsymbol{\Delta}_A^\top) \boldsymbol{\Delta}_A \succcurlyeq \boldsymbol{0}$ if we choose $\lambda = 1/(\|\boldsymbol{W}^{(k)}\| \|\boldsymbol{\Delta}_A\|^2) = 1/(\|\boldsymbol{W}^{(k)}\| \|\boldsymbol{\Delta}\|)$; this implies that

$$\text{(ii)} \geq -\frac{2}{\lambda} \log \frac{2}{\delta} = -2 \|\boldsymbol{W}^{(k)}\| \|\boldsymbol{\Delta}\| \log \frac{2}{\delta}.$$

Putting these back to (74), we have

$$\begin{aligned}
\sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta} \boldsymbol{x}_t &\geq \text{(i)} + \text{(ii)} \\
&\geq T \cdot \mathsf{Tr}\Big((\boldsymbol{W}^{(k)})^{1/2} \boldsymbol{\Delta} (\boldsymbol{W}^{(k)})^{1/2} \Big) - C_0 \sqrt{T} \big\| (\boldsymbol{W}^{(k)})^{1/2} \boldsymbol{\Delta} (\boldsymbol{W}^{(k)})^{1/2} \big\|_{\mathrm{F}} \log \frac{1}{\delta} - 2 \|\boldsymbol{W}^{(k)}\| \|\boldsymbol{\Delta}\| \log \frac{2}{\delta} \\
&\geq T \cdot \lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)}) \mathsf{Tr}(\boldsymbol{\Delta}_A^\top \boldsymbol{\Delta}_A) - C_0 \sqrt{T} \|\boldsymbol{W}^{(k)}\| \|\boldsymbol{\Delta}_A^\top \boldsymbol{\Delta}_A\|_{\mathrm{F}} \log \frac{1}{\delta} - 2 \|\boldsymbol{W}^{(k)}\| \|\boldsymbol{\Delta}_A^\top \boldsymbol{\Delta}_A\| \log \frac{2}{\delta} \\
&\geq T \cdot \lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)}) \|\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2 \cdot \left(1 - \frac{C_0}{\sqrt{T}} \kappa(\boldsymbol{W}^{(k)}) \log \frac{1}{\delta} - \frac{2}{T} \kappa(\boldsymbol{W}^{(k)}) \log \frac{2}{\delta} \right) \\
&\geq 0.9 T \cdot \lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)}) \|\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last inequality holds if $T \gtrsim \kappa_w^2 \log^2 \frac{1}{\delta}$.

Going back to (73), we have

$$\begin{aligned}
\text{(B)} &\geq 0.9\lambda_{\mathsf{min}}((\widehat{\boldsymbol{W}}^{(\ell)})^{-1}) \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta} \boldsymbol{x}_t - 40\kappa_w \kappa_{w,\mathsf{cross}} \log \frac{2}{\delta} \\
&\geq 0.81T \cdot \lambda_{\mathsf{min}}((\widehat{\boldsymbol{W}}^{(\ell)})^{-1}) \lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)}) \|\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2 - 40\kappa_w \kappa_{w,\mathsf{cross}} \log \frac{2}{\delta} \\
&\geq 0.7T \frac{\|\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} - 40\kappa_w \kappa_{w,\mathsf{cross}} \log \frac{2}{\delta},
\end{aligned}$$

which finishes the proof of (66).

C.5.3. Proof of (67).

Denote $\boldsymbol{\Delta} = \boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(k)}$, $\boldsymbol{u}_t = (\boldsymbol{W}^{(k)})^{-1/2} \boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and $\boldsymbol{y}_t = (\boldsymbol{W}^{(k)})^{1/2} (\widehat{\boldsymbol{W}}^{(k)})^{-1} \boldsymbol{\Delta} \boldsymbol{x}_t$. Then one has

$$\text{(C)} = \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta}^\top (\widehat{\boldsymbol{W}}^{(k)})^{-1} \boldsymbol{\Delta} \boldsymbol{x}_t + 2 \sum_{t=0}^{T-1} \boldsymbol{w}_t{}^\top (\widehat{\boldsymbol{W}}^{(k)})^{-1} \boldsymbol{\Delta} \boldsymbol{x}_t = \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta}^\top (\widehat{\boldsymbol{W}}^{(k)})^{-1} \boldsymbol{\Delta} \boldsymbol{x}_t + 2 \sum_{t=1}^{T} \boldsymbol{u}_t{}^\top \boldsymbol{y}_t.$$

By Lemma D.2, we have with probability at least $1-\delta$, $\sum_{t=0}^{T-1} \boldsymbol{u}_t{}^\top \boldsymbol{y}_t \leq \frac{\lambda}{2} \sum_{t=0}^{T-1} \|\boldsymbol{y}_t\|_2^2 + \frac{1}{\lambda} \log \frac{2}{\delta}$ for any fixed $\lambda > 0$. This implies that

$$\begin{aligned}
\text{(C)} &\leq \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta}^\top (\widehat{\boldsymbol{W}}^{(k)})^{-1} \boldsymbol{\Delta} \boldsymbol{x}_t + \lambda \sum_{t=0}^{T-1} \boldsymbol{y}_t{}^\top \boldsymbol{y}_t + \frac{2}{\lambda} \log \frac{2}{\delta} \\
&= \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta}^\top (\widehat{\boldsymbol{W}}^{(k)})^{-1} \boldsymbol{\Delta} \boldsymbol{x}_t + \lambda \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta}^\top (\widehat{\boldsymbol{W}}^{(k)})^{-1} \boldsymbol{W}^{(k)} (\widehat{\boldsymbol{W}}^{(k)})^{-1} \boldsymbol{\Delta} \boldsymbol{x}_t + \frac{2}{\lambda} \log \frac{2}{\delta} \\
&= \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta}^\top \Big((\widehat{\boldsymbol{W}}^{(k)})^{-1} + \lambda \cdot (\widehat{\boldsymbol{W}}^{(k)})^{-1} \boldsymbol{W}^{(k)} (\widehat{\boldsymbol{W}}^{(k)})^{-1} \Big) \boldsymbol{\Delta} \boldsymbol{x}_t + \frac{2}{\lambda} \log \frac{2}{\delta} \\
&\leq 1.5 \bigg(\frac{1}{\lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)})} + \frac{\lambda \cdot \|\boldsymbol{W}^{(k)}\|}{\lambda_{\mathsf{min}}(\boldsymbol{W}^{(k)})^2} \bigg) \sum_{t=0}^{T-1} \boldsymbol{x}_t{}^\top \boldsymbol{\Delta}^\top \boldsymbol{\Delta} \boldsymbol{x}_t + \frac{2}{\lambda} \log \frac{2}{\delta}.
\end{aligned}$$

Choosing $\lambda \asymp 1/\kappa_w$ and recalling $\|\boldsymbol{x}_t\|_2 \leq D_x$, we have $\text{(C)} \lesssim \frac{1}{W_{\mathsf{min}}} T \cdot D_x^2 \|\boldsymbol{\Delta}\|^2 + \kappa_w \log \frac{1}{\delta}$, which finishes the proof of (67).

C.5.4. Proof of (69).

Denote $\boldsymbol{\Delta} \coloneqq \boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}$. Then the right-hand side of (69) becomes

$$\begin{aligned}
\widetilde{D}_{k,\ell} &= \log\det \boldsymbol{W}^{(\ell)} - \log\det \boldsymbol{W}^{(k)} + \mathsf{Tr}\Big(\boldsymbol{W}^{(k)} (\boldsymbol{W}^{(\ell)})^{-1} - \boldsymbol{I}_d \Big) + \frac{\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} \\
&= \log\det \boldsymbol{W}^{(\ell)} - \log\det(\boldsymbol{W}^{(\ell)} + \boldsymbol{\Delta}) + \mathsf{Tr}\Big((\boldsymbol{W}^{(\ell)} + \boldsymbol{\Delta})(\boldsymbol{W}^{(\ell)})^{-1} - \boldsymbol{I}_d \Big) + \frac{\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} \\
&= \log\det \boldsymbol{W}^{(\ell)} - \log\det \Big((\boldsymbol{W}^{(\ell)})^{1/2} \big(\boldsymbol{I}_d + (\boldsymbol{W}^{(\ell)})^{-1/2} \boldsymbol{\Delta} (\boldsymbol{W}^{(\ell)})^{-1/2} \big) (\boldsymbol{W}^{(\ell)})^{1/2} \Big) \\
&\quad + \mathsf{Tr}\Big((\boldsymbol{W}^{(\ell)})^{-1/2} \boldsymbol{\Delta} (\boldsymbol{W}^{(\ell)})^{-1/2} \Big) + \frac{\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} \\
&= \mathsf{Tr}(\boldsymbol{X}) - \log\det(\boldsymbol{I}_d + \boldsymbol{X}) + \frac{\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}},
\end{aligned}$$

where we define $\boldsymbol{X} \coloneqq (\boldsymbol{W}^{(\ell)})^{-1/2} \boldsymbol{\Delta} (\boldsymbol{W}^{(\ell)})^{-1/2}$. Notice that $\boldsymbol{X}$ is symmetric and satisfies

$$\boldsymbol{X} + \boldsymbol{I}_d = (\boldsymbol{W}^{(\ell)})^{-1/2} \boldsymbol{W}^{(k)} (\boldsymbol{W}^{(\ell)})^{-1/2} \succ \boldsymbol{0},$$

$$\|\boldsymbol{X}\| \le \big\|(\boldsymbol{W}^{(\ell)})^{-1/2}\big\|^2 \big\|\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}\big\| \le \frac{2W_{\mathsf{max}}}{W_{\mathsf{min}}} = 2\kappa_{w,\mathsf{cross}},$$
$$\|\boldsymbol{X}\|_{\mathrm{F}}^2 = \big\|(\boldsymbol{W}^{(\ell)})^{-1/2}\boldsymbol{\Delta}(\boldsymbol{W}^{(\ell)})^{-1/2}\big\|_{\mathrm{F}}^2 \ge \frac{\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2}.$$

Therefore, by Lemma D.3, we have

$$\mathsf{Tr}(\boldsymbol{X}) - \log\det(\boldsymbol{I}_d + \boldsymbol{X}) \ge \frac{\|\boldsymbol{X}\|_{\mathrm{F}}^2}{6\kappa_{w,\mathsf{cross}}} \ge \frac{\|\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}^2}{6\kappa_{w,\mathsf{cross}}W_{\mathsf{max}}^2},$$

and thus

$$\widetilde{D}_{k,\ell} = \mathsf{Tr}(\boldsymbol{X}) - \log\det(\boldsymbol{I}_d + \boldsymbol{X}) + \frac{\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} \ge \frac{\|\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}^2}{6\kappa_{w,\mathsf{cross}}W_{\mathsf{max}}^2} + \frac{\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} \asymp D_{k,\ell},$$

which finishes the proof of (69).

C.5.5. Proof of (70).

Recall the definition

$$\widehat{D}_{k,\ell} = \log\frac{\det(\widehat{\boldsymbol{W}}^{(\ell)})}{\det(\widehat{\boldsymbol{W}}^{(k)})} + \mathsf{Tr}\Big(\boldsymbol{W}^{(k)}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} - (\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)\Big) + \frac{\|\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} - \frac{D_x^2\|\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(k)}\|^2}{W_{\mathsf{min}}}.$$

First, we have

$$\begin{aligned}
&\log\frac{\det(\widehat{\boldsymbol{W}}^{(\ell)})}{\det(\widehat{\boldsymbol{W}}^{(k)})} + \mathsf{Tr}\Big(\boldsymbol{W}^{(k)}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} - (\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)\Big) \\
&\quad= \underbrace{\Big[\log\frac{\det(\widehat{\boldsymbol{W}}^{(\ell)})}{\det(\boldsymbol{W}^{(k)})} + \mathsf{Tr}\Big(\boldsymbol{W}^{(k)}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1} - (\boldsymbol{W}^{(k)})^{-1}\big)\Big)\Big]}_{\text{(i)}} \\
&\quad\quad- \underbrace{\Big[\log\frac{\det(\widehat{\boldsymbol{W}}^{(k)})}{\det(\boldsymbol{W}^{(k)})} + \mathsf{Tr}\Big(\boldsymbol{W}^{(k)}\big((\widehat{\boldsymbol{W}}^{(k)})^{-1} - (\boldsymbol{W}^{(k)})^{-1}\big)\Big)\Big]}_{\text{(ii)}}.
\end{aligned}$$

We can lower bound (i) by the same idea of our earlier proof for (69), except that we replace $\boldsymbol{W}^{(\ell)}$ in that proof with $\widehat{\boldsymbol{W}}^{(\ell)}$; this gives us

$$\text{(i)} \gtrsim \frac{\|\boldsymbol{W}^{(k)} - \widehat{\boldsymbol{W}}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}W_{\mathsf{max}}^2}.$$

As for (ii), applying Lemma D.3 with $\boldsymbol{X} = (\boldsymbol{W}^{(k)})^{-1/2}(\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(k)})(\boldsymbol{W}^{(k)})^{-1/2}$, one has

$$\text{(ii)} = \mathsf{Tr}(\boldsymbol{X}) - \log\det(\boldsymbol{X} + \boldsymbol{I}_d) \le \|\boldsymbol{X}\|_{\mathrm{F}}^2 \le \frac{{\epsilon_W}^2 d}{W_{\mathsf{min}}^2}.$$

Putting things together, we have

$$\begin{aligned}
\widehat{D}_{k,\ell} &= \text{(i)} - \text{(ii)} + \frac{\|\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} - \frac{D_x^2\|\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(k)}\|^2}{W_{\mathsf{min}}} \\
&\ge \underbrace{C_1\frac{1}{\kappa_{w,\mathsf{cross}}}\left(\frac{\|\boldsymbol{W}^{(k)} - \widehat{\boldsymbol{W}}^{(\ell)}\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} + \|\boldsymbol{A}^{(k)} - \widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2\right)}_{\text{(iii)}} - \underbrace{C_2\left(\frac{{\epsilon_W}^2 d}{W_{\mathsf{min}}^2} + \frac{D_x^2{\epsilon_A}^2}{W_{\mathsf{min}}}\right)}_{\text{(iv)}}
\end{aligned} \tag{75}$$

If ϵ_W, ϵ_A satisfy (70), then (iv) $\le c_0 D_{k,\ell}$ for some sufficiently small constant $c_0 > 0$. As for (iii), according to the definition of $D_{k,\ell}$ in (68), there are two possible cases:

- If $\frac{1}{\kappa_{w,\mathsf{cross}}}\frac{\|\boldsymbol{W}^{(k)}-\boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} \geq \frac{D_{k,\ell}}{2}$, then it is easy to check that (70) implies that

$$\frac{\epsilon_W\sqrt{d}}{W_{\mathsf{max}}} \leq \frac{1}{4}\frac{\|\boldsymbol{W}^{(k)}-\boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}}{W_{\mathsf{max}}},$$

and hence

$$\begin{aligned}\text{(iii)} &\gtrsim \frac{1}{\kappa_{w,\mathsf{cross}}}\frac{\|\boldsymbol{W}^{(k)}-\widehat{\boldsymbol{W}}^{(\ell)}\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} \geq \frac{1}{\kappa_{w,\mathsf{cross}}}\left(\frac{\|\boldsymbol{W}^{(k)}-\boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}}{W_{\mathsf{max}}}-\frac{\epsilon_W\sqrt{d}}{W_{\mathsf{max}}}\right)^2\\ &\gtrsim \frac{1}{\kappa_{w,\mathsf{cross}}}\frac{\|\boldsymbol{W}^{(k)}-\boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} \gtrsim D_{k,\ell}.\end{aligned}$$

- On the other hand, if $\frac{1}{\kappa_{w,\mathsf{cross}}}\|\boldsymbol{A}^{(k)}-\boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2 \geq \frac{D_{k,\ell}}{2}$, then one can check that (70) implies that

$$\epsilon_A\sqrt{d} \leq \frac{1}{4}\|\boldsymbol{A}^{(k)}-\boldsymbol{A}^{(\ell)}\|_{\mathrm{F}},$$

and hence

$$\text{(iii)} \gtrsim \frac{\|\boldsymbol{A}^{(k)}-\widehat{\boldsymbol{A}}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} \geq \frac{(\|\boldsymbol{A}^{(k)}-\boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}-\epsilon_A\sqrt{d})^2}{\kappa_{w,\mathsf{cross}}} \gtrsim \frac{\|\boldsymbol{A}^{(k)}-\boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2}{\kappa_{w,\mathsf{cross}}} \gtrsim D_{k,\ell}.$$

In sum, it is always guaranteed that $\text{(iii)} \gtrsim D_{k,\ell}$. Going back to (75), we claim that $\widehat{D}_{k,\ell} \geq \text{(iii)} - \text{(iv)} \gtrsim D_{k,\ell}$ as long as ϵ_W and ϵ_A satisfy (70), which finishes the proof of (70).

C.5.6. Proof of (72).

Recall the lower bound (71) for (A) + (B) − (C). We want to show that the low-order term is dominated by the order-T term, namely $T \cdot D_{k,\ell}$. First, if $T \gtrsim \frac{\kappa_w\kappa_{w,\mathsf{cross}}}{D_{k,\ell}}\log\frac{1}{\delta}$, then $\kappa_w\kappa_{w,\mathsf{cross}}\log\frac{1}{\delta} \lesssim T \cdot D_{k,\ell}$. Next, we have

$$\begin{aligned}&\left\|(\boldsymbol{W}^{(k)})^{1/2}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1}-(\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)(\boldsymbol{W}^{(k)})^{1/2}\right\|_{\mathrm{F}}\\ &\quad\leq \left\|\boldsymbol{W}^{(k)}\right\|\cdot\left\|(\widehat{\boldsymbol{W}}^{(\ell)})^{-1}-(\widehat{\boldsymbol{W}}^{(k)})^{-1}\right\|_{\mathrm{F}} = \left\|\boldsymbol{W}^{(k)}\right\|\cdot\left\|(\widehat{\boldsymbol{W}}^{(\ell)})^{-1}(\widehat{\boldsymbol{W}}^{(k)}-\widehat{\boldsymbol{W}}^{(\ell)})(\widehat{\boldsymbol{W}}^{(k)})^{-1}\right\|_{\mathrm{F}}\\ &\quad\leq \left\|\boldsymbol{W}^{(k)}\right\|\cdot\left\|(\widehat{\boldsymbol{W}}^{(\ell)})^{-1}\right\|\cdot\left\|(\widehat{\boldsymbol{W}}^{(k)})^{-1}\right\|\cdot\left(\|\boldsymbol{W}^{(k)}-\boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}+2\sqrt{d}\epsilon_W\right)\\ &\quad\leq \frac{2W_{\mathsf{max}}}{W_{\mathsf{min}}^2}\left(\|\boldsymbol{W}^{(k)}-\boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}+2\sqrt{d}\epsilon_W\right) = \frac{2\kappa_{w,\mathsf{cross}}}{W_{\mathsf{min}}}\left(\|\boldsymbol{W}^{(k)}-\boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}+2\sqrt{d}\epsilon_W\right).\end{aligned}$$

Notice that the definition (68) of $D_{k,\ell}$ implies that $\|\boldsymbol{W}^{(k)}-\boldsymbol{W}^{(\ell)}\|_{\mathrm{F}} \leq \sqrt{\kappa_{w,\mathsf{cross}}W_{\mathsf{max}}^2 D_{k,\ell}}$, and thus

$$\left\|(\boldsymbol{W}^{(k)})^{1/2}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1}-(\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)(\boldsymbol{W}^{(k)})^{1/2}\right\|_{\mathrm{F}} \lesssim \frac{\kappa_{w,\mathsf{cross}}}{W_{\mathsf{min}}}\left(\sqrt{\kappa_{w,\mathsf{cross}}W_{\mathsf{max}}^2 D_{k,\ell}}+\sqrt{d}\epsilon_W\right).$$

Now it is easy to checked that

$$\begin{aligned}\text{if}\quad & T \gtrsim \left(\frac{\kappa_{w,\mathsf{cross}}^5}{D_{k,\ell}}+\left(\frac{\kappa_{w,\mathsf{cross}}\sqrt{d}\epsilon_W}{W_{\mathsf{min}}D_{k,\ell}}\right)^2\right)\log^2\frac{1}{\delta},\\ \text{then}\quad & \sqrt{T}\left\|(\boldsymbol{W}^{(k)})^{1/2}\big((\widehat{\boldsymbol{W}}^{(\ell)})^{-1}-(\widehat{\boldsymbol{W}}^{(k)})^{-1}\big)(\boldsymbol{W}^{(k)})^{1/2}\right\|_{\mathrm{F}}\log\frac{1}{\delta} \lesssim T\cdot D_{k,\ell}.\end{aligned}$$

Due to our assumption (70) on ϵ_W, we have $(\frac{\kappa_{w,\mathsf{cross}}\sqrt{d}\epsilon_W}{W_{\mathsf{min}}D_{k,\ell}})^2 \lesssim \frac{\kappa_{w,\mathsf{cross}}^2}{D_{k,\ell}} \leq \frac{\kappa_{w,\mathsf{cross}}^5}{D_{k,\ell}}$, and thus it suffices to have $T \gtrsim (\frac{\kappa_{w,\mathsf{cross}}^5}{D_{k,\ell}}+1)\log^2\frac{1}{\delta}$, which finishes the proof of (72).

D. Miscellaneous results

Proposition D.1. *Consider* $\boldsymbol{a}_t, \boldsymbol{b}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d), 1 \le t \le N$, *where* $N \gtrsim d + \log(1/\delta)$. *Then with probability at least* $1-\delta$,

$$\left\| \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{a}_t \boldsymbol{a}_t^\top - \boldsymbol{I}_d \right\| \lesssim \sqrt{\frac{d + \log \frac{1}{\delta}}{N}}, \quad \left\| \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{a}_t \boldsymbol{b}_t^\top \right\| \lesssim \sqrt{\frac{d + \log \frac{1}{\delta}}{N}}.$$

The proof follows from a standard covering argument (cf. (Vershynin, 2018)), which we skip for brevity.

Lemma D.2 (Self-normalized concentration, scalar version)**.** *Suppose that random vectors* $\{\boldsymbol{u}_t, \boldsymbol{y}_t\}_{1 \le t \le T}$ *and filtrations* $\{\mathcal{F}_t\}_{0 \le t \le T-1}$ *satisfy* $\mathcal{F}_t = \sigma(\boldsymbol{u}_i, 1 \le i \le t)$, $\boldsymbol{y}_t \in \mathcal{F}_{t-1}$, *and* $\boldsymbol{u}_t | \mathcal{F}_{t-1} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. *Then for any fixed* $\lambda > 0$, *with probability at least* $1-\delta$, *we have*

$$\left| \sum_{t=1}^{T} \boldsymbol{u}_t{}^\top \boldsymbol{y}_t \right| < \frac{\lambda}{2} \sum_{t=1}^{T} \|\boldsymbol{y}_t\|_2^2 + \frac{1}{\lambda} \log \frac{2}{\delta}.$$

Proof. Using basic properties of the Gaussian distribution, we have for all fixed $\lambda \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\sum_{t=1}^{T} \big(\lambda \cdot \boldsymbol{u}_t{}^\top \boldsymbol{y}_t - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_t\|_2^2 \big) \Big) \Big] \\
&\quad = \mathbb{E}\Big[\exp\Big(\sum_{t=1}^{T-1} \big(\lambda \cdot \boldsymbol{u}_t{}^\top \boldsymbol{y}_t - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_t\|_2^2 \big) \Big) \cdot \underbrace{\mathbb{E}\Big[\exp\big(\lambda \cdot \boldsymbol{u}_T^\top \boldsymbol{y}_T - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_T\|_2^2 \big) | \mathcal{F}_{T-1} \Big]}_{\le 1} \Big] \\
&\quad \le \mathbb{E}\Big[\exp\Big(\sum_{t=1}^{T-1} \big(\lambda \cdot \boldsymbol{u}_t{}^\top \boldsymbol{y}_t - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_t\|_2^2 \big) \Big) \Big].
\end{aligned}$$

Continuing this expansion leads to the result $\mathbb{E}[\exp(\sum_{t=1}^{T} (\lambda \cdot \boldsymbol{u}_t{}^\top \boldsymbol{y}_t - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_t\|_2^2))] \le 1$. Now, letting $z = \log(2/\delta)$ and using Markov's inequality, we have

$$\begin{aligned}
\mathbb{P}\Big(\sum_{t=1}^{T} \big(\lambda \cdot \boldsymbol{u}_t{}^\top \boldsymbol{y}_t - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_t\|_2^2 \big) \ge z \Big) &= \mathbb{P}\Big(\exp\Big(\sum_{t=1}^{T} \big(\lambda \cdot \boldsymbol{u}_t{}^\top \boldsymbol{y}_t - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_t\|_2^2 \big) \Big) \ge \exp(z) \Big) \\
&\le \exp(-z) \cdot \mathbb{E}\Big[\exp\Big(\sum_{t=1}^{T} \big(\lambda \cdot \boldsymbol{u}_t{}^\top \boldsymbol{y}_t - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_t\|_2^2 \big) \Big) \Big] \le \exp(-z) = \frac{\delta}{2}.
\end{aligned}$$

In other words, with probability at least $1 - \delta/2$, we have $\sum_{t=1}^{T} (\lambda \cdot \boldsymbol{u}_t{}^\top \boldsymbol{y}_t - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_t\|_2^2) < z = \log(2/\delta)$, which implies $\sum_{t=1}^{T} \boldsymbol{u}_t{}^\top \boldsymbol{y}_t < \frac{\lambda}{2} \sum_{t=1}^{T} \|\boldsymbol{y}_t\|_2^2 + \frac{1}{\lambda} \log \frac{2}{\delta}$ if $\lambda > 0$. By a similar argument (but with λ replaced by $-\lambda$), we have with probability at least $1 - \delta/2$, $\sum_{t=1}^{T} (-\lambda \cdot \boldsymbol{u}_t{}^\top \boldsymbol{y}_t - \frac{1}{2}\lambda^2 \|\boldsymbol{y}_t\|_2^2) < \log(2/\delta)$, which implies $\sum_{t=1}^{T} \boldsymbol{u}_t{}^\top \boldsymbol{y}_t > -(\frac{\lambda}{2} \sum_{t=1}^{T} \|\boldsymbol{y}_t\|_2^2 + \frac{1}{\lambda} \log \frac{2}{\delta})$ if $\lambda > 0$. Finally, taking the union bound finishes the proof of the lemma. □

Lemma D.3. *Consider a symmetric matrix* $\boldsymbol{X} \in \mathbb{R}^{d \times d}$ *that satisfies* $\boldsymbol{I}_d + \boldsymbol{X} \succ \boldsymbol{0}$. *If* $\|\boldsymbol{X}\| \le B$ *for some* $B \ge 2$, *then we have*

$$\mathsf{Tr}(\boldsymbol{X}) - \log\det(\boldsymbol{I}_d + \boldsymbol{X}) \ge \frac{\|\boldsymbol{X}\|_{\mathrm{F}}^2}{3B}.$$

On the other hand, if $\boldsymbol{X} \succcurlyeq -\frac{1}{2}\boldsymbol{I}_d$, *then we have*

$$\mathsf{Tr}(\boldsymbol{X}) - \log\det(\boldsymbol{I}_d + \boldsymbol{X}) \le \|\boldsymbol{X}\|_{\mathrm{F}}^2.$$

Proof. Denote $\{\lambda_i\}_{1 \le i \le d}$ as the eigenvalues of $\boldsymbol{X}$, which satisfies $\lambda_i > -1$ for all $1 \le i \le d$. Then

$$\mathsf{Tr}(\boldsymbol{X}) - \log\det(\boldsymbol{I}_d + \boldsymbol{X}) = \sum_{i=1}^{d} \lambda_i - \log \prod_{i=1}^{d} (1 + \lambda_i) = \sum_{i=1}^{d} \Big(\lambda_i - \log(1 + \lambda_i) \Big).$$

It can be checked (via elementary calculus) that, for all $-1 < \lambda \le B$ where $B \ge 2$, it holds that $\lambda - \log(1+\lambda) \ge \lambda^2/(3B)$. Therefore,

$$\mathsf{Tr}(\boldsymbol{X}) - \log\det(\boldsymbol{I}_d + \boldsymbol{X}) \ge \sum_{i=1}^{d} \frac{\lambda_i^2}{3B} = \frac{\|\boldsymbol{X}\|_{\mathrm{F}}^2}{3B},$$

which completes the proof of our first claim. Similarly, it can be checked that, for all $\lambda \ge -1/2$, one has $\lambda - \log(1+\lambda) \le \lambda^2$, which implies that

$$\mathsf{Tr}(\boldsymbol{X}) - \log\det(\boldsymbol{I}_d + \boldsymbol{X}) \le \sum_{i=1}^{d} \lambda_i^2 = \|\boldsymbol{X}\|_{\mathrm{F}}^2;$$

this completes the proof of our second claim. □

Fact D.4. *In the setting of Section 3.1, it holds that* $\Gamma_{\mathsf{max}} \le W_{\mathsf{max}}\kappa_A^2/(1-\rho)$ *and* $\Delta_{\Gamma,Y} \ge \Delta_{A,W} \cdot W_{\mathsf{max}} W_{\mathsf{min}}/(4\kappa_A^2 \Gamma_{\mathsf{max}})$.

Proof. First, consider $\boldsymbol{\Gamma} = \sum_{i=0}^{\infty} \boldsymbol{A}^i \boldsymbol{W} (\boldsymbol{A}^i)^\top$, where $\|\boldsymbol{W}\| \le W_{\mathsf{max}}$ and $\|\boldsymbol{A}^i\| \le \kappa_A \rho^i$. Then we have $\|\boldsymbol{\Gamma}\| \le \sum_{i=0}^{\infty} \|\boldsymbol{A}^i\|^2 \|\boldsymbol{W}\| \le W_{\mathsf{max}}\kappa_A^2 \sum_{i=0}^{\infty} \rho^{2i} \le W_{\mathsf{max}}\kappa_A^2/(1-\rho)$, which proves our upper bound for Γ_{max}.

Next, let us turn to $\Delta_{\Gamma,Y}$. Our proof below can be viewed as a quantitative version of the earlier proof for Fact 2.1. We will show that, if the autocovariance matrices between the k-th and the ℓ-th models are close (in Frobenius norm), then the models themselves should also be close; our lower bound for $\Delta_{\Gamma,Y}$ in terms of $\Delta_{A,W}$ then follows from contraposition.

Consider two LDS models $(\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}) \ne (\boldsymbol{A}^{(\ell)}, \boldsymbol{W}^{(\ell)})$ and their autocovariance matrics $(\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)}), (\boldsymbol{\Gamma}^{(\ell)}, \boldsymbol{Y}^{(\ell)})$. Recall from (17) that $\boldsymbol{A}^{(k)} = \boldsymbol{Y}^{(k)} {\boldsymbol{\Gamma}^{(k)}}^{-1}$ and $\boldsymbol{W}^{(k)} = \boldsymbol{\Gamma}^{(k)} - \boldsymbol{A}^{(k)} \boldsymbol{\Gamma}^{(k)} {\boldsymbol{A}^{(k)}}^\top = \boldsymbol{\Gamma}^{(k)} - \boldsymbol{A}^{(k)} {\boldsymbol{Y}^{(k)}}^\top$; $\boldsymbol{A}^{(\ell)}$ and $\boldsymbol{W}^{(\ell)}$ can be expressed similarly.

- First, regarding $\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}$, one has

$$\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)} = \boldsymbol{Y}^{(k)} {\boldsymbol{\Gamma}^{(k)}}^{-1} - \boldsymbol{Y}^{(\ell)} {\boldsymbol{\Gamma}^{(\ell)}}^{-1} = (\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}) {\boldsymbol{\Gamma}^{(k)}}^{-1} + \boldsymbol{Y}^{(\ell)} ({\boldsymbol{\Gamma}^{(k)}}^{-1} - {\boldsymbol{\Gamma}^{(\ell)}}^{-1}),$$

  where

$$\boldsymbol{Y}^{(\ell)} ({\boldsymbol{\Gamma}^{(k)}}^{-1} - {\boldsymbol{\Gamma}^{(\ell)}}^{-1}) = \boldsymbol{Y}^{(\ell)} {\boldsymbol{\Gamma}^{(\ell)}}^{-1} (\boldsymbol{\Gamma}^{(\ell)} - \boldsymbol{\Gamma}^{(k)}) {\boldsymbol{\Gamma}^{(k)}}^{-1} = \boldsymbol{A}^{(\ell)} (\boldsymbol{\Gamma}^{(\ell)} - \boldsymbol{\Gamma}^{(k)}) {\boldsymbol{\Gamma}^{(k)}}^{-1}.$$

  Therefore, we have

$$\begin{aligned} \|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}} &\le \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}} \|{\boldsymbol{\Gamma}^{(k)}}^{-1}\| + \|\boldsymbol{A}^{(\ell)}\| \|\boldsymbol{\Gamma}^{(\ell)} - \boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}} \|{\boldsymbol{\Gamma}^{(k)}}^{-1}\| \\ &\le \frac{1}{W_{\mathsf{min}}} \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}} + \frac{\kappa_A}{W_{\mathsf{min}}} \|\boldsymbol{\Gamma}^{(\ell)} - \boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}}, \end{aligned} \tag{76}$$

  where the last line follows from $\boldsymbol{\Gamma}^{(k)} \succcurlyeq \boldsymbol{W}^{(k)} \succcurlyeq W_{\mathsf{min}} \boldsymbol{I}_d$ and $\|\boldsymbol{A}^{(\ell)}\| \le \kappa_A \rho \le \kappa_A$.

- Next, we turn to the analysis of $\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}$, which satisfies

$$\begin{aligned} \boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)} &= (\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}) - (\boldsymbol{A}^{(k)} {\boldsymbol{Y}^{(k)}}^\top - \boldsymbol{A}^{(\ell)} {\boldsymbol{Y}^{(\ell)}}^\top), \\ \|\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}\|_{\mathrm{F}} &\le \|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}} + \|\boldsymbol{A}^{(k)} {\boldsymbol{Y}^{(k)}}^\top - \boldsymbol{A}^{(\ell)} {\boldsymbol{Y}^{(\ell)}}^\top\|_{\mathrm{F}}. \end{aligned}$$

  Notice that

$$\begin{aligned} \|\boldsymbol{A}^{(k)} {\boldsymbol{Y}^{(k)}}^\top - \boldsymbol{A}^{(\ell)} {\boldsymbol{Y}^{(\ell)}}^\top\|_{\mathrm{F}} &= \|\boldsymbol{A}^{(k)} (\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)})^\top + (\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}) {\boldsymbol{Y}^{(\ell)}}^\top\|_{\mathrm{F}} \\ &\le \|\boldsymbol{A}^{(k)} (\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)})^\top\|_{\mathrm{F}} + \|(\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}) {\boldsymbol{Y}^{(\ell)}}^\top\|_{\mathrm{F}} \\ &\le \kappa_A \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}} + \kappa_A \Gamma_{\mathsf{max}} \|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}, \end{aligned}$$

  where the last line is due to $\|\boldsymbol{Y}^{(\ell)}\| = \|\boldsymbol{A}^{(\ell)} \boldsymbol{\Gamma}^{(\ell)}\| \le \|\boldsymbol{A}^{(\ell)}\| \cdot \|\boldsymbol{\Gamma}^{(\ell)}\| \le \kappa_A \Gamma_{\mathsf{max}}$. Therefore, we have

$$\|\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}\|_{\mathrm{F}} \le \|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}} + \kappa_A \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}} + \kappa_A \Gamma_{\mathsf{max}} \|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}. \tag{77}$$

For notational simplify, denote $\boldsymbol{\Delta}_A \coloneqq \boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}$, $\boldsymbol{\Delta}_W \coloneqq \boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}$, $\boldsymbol{\Delta}_\Gamma \coloneqq \boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}$, $\Delta_Y \coloneqq \boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}$. From (77), one has

$$\begin{aligned}
\|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2 + \frac{\|\boldsymbol{\Delta}_W\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} &\le \|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2 + \frac{3}{W_{\mathsf{max}}^2}\Big(\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2 + \kappa_A^2\|\Delta_Y\|_{\mathrm{F}}^2 + \kappa_A^2\Gamma_{\mathsf{max}}^2\|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2\Big) \\
&\le \frac{3}{W_{\mathsf{max}}^2}\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2 + \frac{3\kappa_A^2}{W_{\mathsf{max}}^2}\|\Delta_Y\|_{\mathrm{F}}^2 + \Big(1 + \frac{3\kappa_A^2\Gamma_{\mathsf{max}}^2}{W_{\mathsf{max}}^2}\Big)\|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2 \\
&\le \frac{3}{W_{\mathsf{max}}^2}\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2 + \frac{3\kappa_A^2}{W_{\mathsf{max}}^2}\|\Delta_Y\|_{\mathrm{F}}^2 + \frac{4\kappa_A^2\Gamma_{\mathsf{max}}^2}{W_{\mathsf{max}}^2}\|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last line follows from $\kappa_A^2\Gamma_{\mathsf{max}}^2/W_{\mathsf{max}}^2 \ge 1$. Morever, (76) tells us that $\|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2 \le \frac{2}{W_{\mathsf{min}}^2}\|\Delta_Y\|_{\mathrm{F}}^2 + \frac{2\kappa_A^2}{W_{\mathsf{min}}^2}\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2$. Putting together, we have

$$\begin{aligned}
\|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2 + \frac{\|\boldsymbol{\Delta}_W\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} &\le \frac{3}{W_{\mathsf{max}}^2}\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2 + \frac{3\kappa_A^2}{W_{\mathsf{max}}^2}\|\Delta_Y\|_{\mathrm{F}}^2 + \frac{4\kappa_A^2\Gamma_{\mathsf{max}}^2}{W_{\mathsf{max}}^2}\|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2 \\
&\le \frac{3}{W_{\mathsf{max}}^2}\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2 + \frac{3\kappa_A^2}{W_{\mathsf{max}}^2}\|\Delta_Y\|_{\mathrm{F}}^2 + \frac{4\kappa_A^2\Gamma_{\mathsf{max}}^2}{W_{\mathsf{max}}^2}\Big(\frac{2}{W_{\mathsf{min}}^2}\|\Delta_Y\|_{\mathrm{F}}^2 + \frac{2\kappa_A^2}{W_{\mathsf{min}}^2}\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2\Big) \\
&= \Big(\frac{3}{W_{\mathsf{max}}^2} + \frac{4\kappa_A^2\Gamma_{\mathsf{max}}^2}{W_{\mathsf{max}}^2}\frac{2\kappa_A^2}{W_{\mathsf{min}}^2}\Big)\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2 + \Big(\frac{3\kappa_A^2}{W_{\mathsf{max}}^2} + \frac{4\kappa_A^2\Gamma_{\mathsf{max}}^2}{W_{\mathsf{max}}^2}\frac{2}{W_{\mathsf{min}}^2}\Big)\|\Delta_Y\|_{\mathrm{F}}^2 \\
&\le \frac{11\kappa_A^4\Gamma_{\mathsf{max}}^2}{W_{\mathsf{max}}^2 W_{\mathsf{min}}^2}\Big(\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2 + \|\Delta_Y\|_{\mathrm{F}}^2\Big).
\end{aligned}$$

In sum, we have just shown that, if $\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2 + \|\Delta_Y\|_{\mathrm{F}}^2 < \Delta_{\Gamma,Y}^2$, then $\|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2 + \frac{\|\boldsymbol{\Delta}_W\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} < \frac{11\kappa_A^4\Gamma_{\mathsf{max}}^2}{W_{\mathsf{max}}^2 W_{\mathsf{min}}^2}\Delta_{\Gamma,Y}^2$. Equivalently (by contraposition), if $\|\boldsymbol{\Delta}_A\|_{\mathrm{F}}^2 + \frac{\|\boldsymbol{\Delta}_W\|_{\mathrm{F}}^2}{W_{\mathsf{max}}^2} \ge \Delta_{A,W}^2$, then $\|\boldsymbol{\Delta}_\Gamma\|_{\mathrm{F}}^2 + \|\Delta_Y\|_{\mathrm{F}}^2 \ge \frac{W_{\mathsf{max}}^2 W_{\mathsf{min}}^2}{11\kappa_A^4\Gamma_{\mathsf{max}}^2}\Delta_{A,W}^2$. This proves our lower bound for $\Delta_{\Gamma,Y}$ in terms of $\Delta_{A,W}$. □

Example D.5. *It has been known that in our Case 1 (i.e. a single continuous trajectory), the quick switching of multiple LDS models may lead to exponentially large states, even if each individual model is stable (Liberzon, 2003). We give a quick example for completeness. Consider*

$$\boldsymbol{A}^{(k)} = 0.99\begin{bmatrix} 0 & 2 \\ \frac{1}{2} & 0 \end{bmatrix}, \quad \boldsymbol{A}^{(\ell)} = 0.99\begin{bmatrix} 0 & 3 \\ \frac{1}{3} & 0 \end{bmatrix},$$

both satisfying the stability condition in Assumption 3.1 with $\rho = 0.99 < 1$ and hence $t_{\mathsf{mix}} \asymp 1/(1-\rho) = 100$. Suppose that each short trajectory has only a length of 2, and the m-th (resp. $(m+1)$-th) trajectory has label $k_m = \ell$ (resp. $k_{m+1} = k$). Then $\boldsymbol{x}_{m+2,0}$ is equal to $\boldsymbol{A}^{(k)}\boldsymbol{A}^{(\ell)}\boldsymbol{x}_{m,0}$ plus a mean-zero noise term, where

$$\boldsymbol{A}^{(k)}\boldsymbol{A}^{(\ell)} = 0.99^2\begin{bmatrix} 0 & 2 \\ \frac{1}{2} & 0 \end{bmatrix}\begin{bmatrix} 0 & 3 \\ \frac{1}{3} & 0 \end{bmatrix} = 0.99^2\begin{bmatrix} \frac{2}{3} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}$$

has spectral radius $0.99^2 \cdot 3/2 > 1$; this will cause the exponential explosion of the states.

E. Extensions of Algorithm 1

Different trajectory lengths. Recall that in Section 2, we assume that all short trajectories within each subset of data $\mathcal{M}_{\mathsf{o}}$ have the same length $T_m = T_{\mathsf{o}}$. If this is not the case, we can easily modify our algorithms in the following ways:

- For subspace estimation, the easiest way to handle different T_m's is to simply *truncate* the trajectories in $\mathcal{M}_{\mathsf{subspace}}$ so that they have the same length $T_{\mathsf{subspace}} = \min_{m\in\mathcal{M}_{\mathsf{subspace}}} T_m$, and then apply Algorithm 2 without modification. However, this might waste many samples when some trajectories of $\mathcal{M}_{\mathsf{subspace}}$ are much longer than others; one way to resolve this is to *manually divide* the longer trajectories into shorter segments of comparable lengths, before doing truncation. A more refined method is to modify Algorithm 2 itself, by *re-defining the index sets* Ω_1, Ω_2 separately for each trajectory; moreover, in the definition of $\widehat{\boldsymbol{H}}_i$ and $\widehat{\boldsymbol{G}}_i$, one might consider assigning larger weights to longer trajectories, instead of using the uniform weight $1/|\mathcal{M}_{\mathsf{subspace}}|$.

- For clustering (or pairwise testing) of $\mathcal{M}_{\text{clustering}}$, we can handle various T_m's similarly, by either truncating each pair of trajectories to the same length, or modifying Algorithm 3 itself (via re-defining the index sets $\{\Omega_{g,1}, \Omega_{g,2}\}_{1\le g\le G}$ separately for each trajectory).
- Our methods for model estimation and classification, namely Algorithms 4 and 5, are already adaptive to different T_m's in $\mathcal{M}_{\text{clustering}}$ and $\mathcal{M}_{\text{classification}}$, and hence need no modification.

Unknown parameters. Next, we show how to handle the case when certain parameters are unknown to the algorithms:

- In Algorithm 2, we set the dimension of the output subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}$ to be K (the number of models). If K is unknown, we might instead examine the eigenvalues of $\widehat{\boldsymbol{H}}_i + \widehat{\boldsymbol{H}}_i^\top$ and $\widehat{\boldsymbol{G}}_i + \widehat{\boldsymbol{G}}_i^\top$, and pick the subspace dimension that covers most of the energy in the eigenvalues.
- In Algorithm 3, we need to know the separation parameter $\Delta_{\Gamma,Y}$ (in order to choose the testing threshold τ appropriately) and the number of models K (for clustering). If either $\Delta_{\Gamma,Y}$ or K is unknown, we might instead try different values of threshold τ, and pick the one that (after permutation) makes the similarity matrix $\boldsymbol{S}$ block-diagonal with as few blocks as possible.

F. Additional experiments

F.1. Synthetic experiments: clustering and classification

First, we take a closer look at the performance of our clustering method (Algorithm 3) through synthetic experiments. We set the parameters $d = 40, K = 2, \rho = 0.5, \delta = 0.12$. The LDS models are generated by $\boldsymbol{A}^{(k)} = (\rho \pm \delta)\boldsymbol{R}$ and $\boldsymbol{W}^{(k)} = \boldsymbol{I}_d$, where $\boldsymbol{R}$ is a random orthogonal matrix. We let $|\mathcal{M}_{\text{clustering}}| = 5d$, and vary $T_{\text{clustering}} \in [10, 60]$. We run our clustering method on the dataset $\mathcal{M}_{\text{clustering}}$, either with or without the assistance of subspace estimation (Algorithm 2) and dimensionality reduction. For the former case, we use the same dataset $\mathcal{M}_{\text{clustering}}$ for subspace estimation, without sample splitting, which is closer to practice; for the latter, we simply replace the subspaces $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}$ with $\boldsymbol{I}_d$. The numerical results are illustrated in Figure 3 (left), confirming that (1) in both cases, the clustering error decreases as $T_{\text{clustering}}$ increases, and (2) subspace estimation and dimensionality reduction significantly improve the clustering accuracy.

Next, we examine the performance of our classification method (Algorithm 5) in the same setting as above. We first obtain a coarse model estimation by running Stage 1 of Algorithm 1 on the dataset $\mathcal{M}_{\text{clustering}}$, with $|\mathcal{M}_{\text{clustering}}| = 10d$ and $T_{\text{clustering}} = 30$. Then, we run classification on the dataset $\mathcal{M}_{\text{classification}}$, with varying $T_{\text{classification}} \in [4, 50]$. The numerical results are included in Figure 3 (right), showing that the classification error rapidly decreases to zero as $T_{\text{classification}}$ grows.

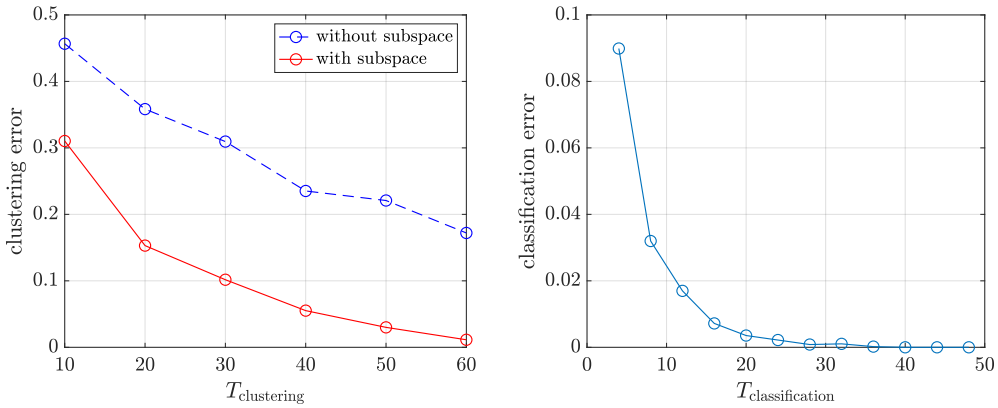

Figure 3. **Left:** mis-clustering rate versus $T_{\text{clustering}}$. **Right:** mis-classification rate versus $T_{\text{classification}}$.

F.2. Real-data experiments: MotionSense

To show the practical relevance of the proposed algorithms, we work with the MotionSense dataset (Malekzadeh et al., 2019). This datasets consists of multivariate time series of dimension $d = 12$, collected (at a rate of 50Hz) by accelerometer

and gyroscope sensors on a mobile phone while a person performs various activities, such as "jogging", "walking", "sitting", and so on. In our experiments, we break the data into 8-second short trajectories, and treat the human activities as latent variables. Figure 4 (left) illustrates what the data looks like. Notice that the time series do not satisfy the mixing property assumed in our theory, but are rather periodic instead.

As a preliminary attempt to apply our algorithms in the real world, we show that the proposed clustering method (which is one of the most crucial step in our overall approach), without any modification, works reasonably well even for this dataset. To be concrete, we apply Algorithm 3 (without dimensionality reduction, i.e. $\{\boldsymbol{V}_i, \boldsymbol{U}_i\}$ are set to $\boldsymbol{I}_d$) to a mixture of 12 "jogging" and 12 "walking" trajectories. Figure 4 (right) shows the resulted *distance matrix*, which is defined in the same way as Line 11 of Algorithm 3, but without thresholding. Its clear block structure confirms that, with an appropriate choice of threshold τ, Algorithm 3 will return an accurate/exact clustering of the mixed trajectories. These results are strong indication that the proposed algorithms in this work might generalize to much broader settings than what our current theory suggests, and we hope that this will inspire further extensions and applications of the proposed methods.

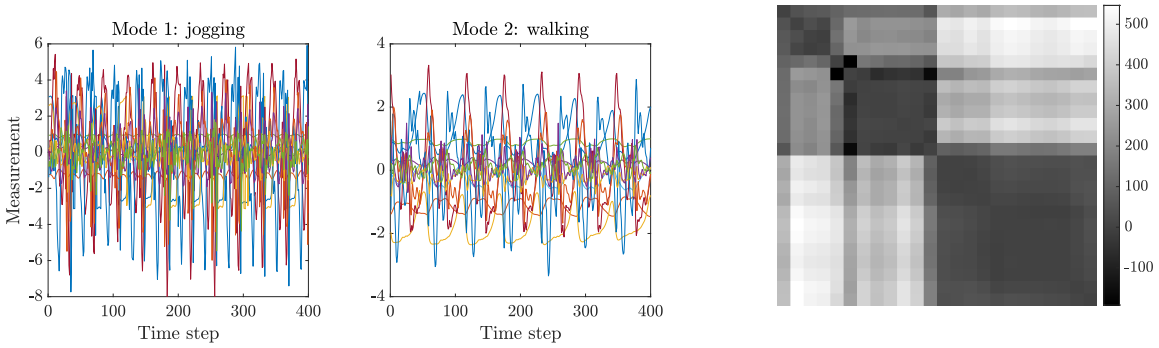

Figure 4. **Left:** examples of "jogging" and "walking" trajectories from the MotionSense dataset. **Right:** the distance matrix constructed by Algorithm 3 for 12 "jogging" and 12 "walking" trajectories.