Learning Low-dimensional Latent Dynamics from High-dimensional Observations: Non-asymptotics and Lower Bounds

Yuyang Zhang [1] **Shahriar Talebi** [1] **Na Li** [1]

Abstract

In this paper, we focus on learning a linear time-invariant (LTI) model with low-dimensional latent variables but high-dimensional observations. We provide an algorithm that recovers the high-dimensional features, i.e. column space of the observer, embeds the data into low dimensions and learns the low-dimensional model parameters. Our algorithm enjoys a sample complexity guarantee of order $\tilde{\mathcal{O}}(n/\epsilon^2)$, where n is the observation dimension. We further establish a fundamental lower bound indicating this complexity bound is optimal up to logarithmic factors and dimension-independent constants. We show that this inevitable linear factor of n is due to the learning error of the observer's column space in the presence of high-dimensional noise. Extending our results, we consider a meta-learning problem inspired by various real-world applications, where the observer column space can be collectively learned from datasets of multiple LTI systems. An end-to-end algorithm is then proposed, facilitating learning LTI systems from a meta-dataset which breaks the sample complexity lower bound in certain scenarios.

1. Introduction

Analyzing high-dimensional time series data is essential for numerous real-world applications in finance (Mudassir et al., 2020), economics (Maliar & Maliar, 2015; Masini et al., 2023) and biology (Churchland et al., 2012; Hajnal et al., 2023; Xia et al., 2021; Gallego et al., 2020; Stringer et al., 2019). High-dimensional time series observations often find succinct representation through a set of low-dimensional latent variables. In this paper, the focus is to learn the low-dimensional dynamics capturing the very essence of the time series, which becomes useful in various down-stream tasks like prediction and inference (Churchland et al., 2012; Mudassir et al., 2020; Pandarinath et al., 2018).

Popular techniques for such analysis range from linear models, such as linear time-invariant (LTI) systems (Sikander & Prasad, 2015; Bui-Thanh et al., 2008; Hespanha, 2018) and auto-regressive models (Dong et al., 2022; Poloni & Sbrana, 2019; Qin, 2022), to more complex nonlinear models, exemplified by recurrent neural networks (Yu et al., 2021; Sussillo & Barak, 2013; Medsker & Jain, 1999). Herein, we focus on LTI models as they are interpretable and require much less computational power in many real-world applications (Gallego et al., 2020; Churchland et al., 2012; Dong et al., 2022).

Specifically, we learn partially observed LTI systems in the following form

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad y_t = Cx_t + \eta_t, \qquad (1)$$

where $y_t \in \mathbb{R}^n$ are *high-dimensional* observations, $x_t \in \mathbb{R}^r$ are low-dimensional latent variables with $n \gg r$, $u_t \in \mathbb{R}^m$ are inputs and w_t, η_t are process and observation noise, respectively. Matrices A and B are the system parameters for the dynamics and inputs, respectively, and C can be seen as the observer, mapping latent variables to *noisy observations*. Recently, a contemporary perspective has revitalized well-established system identification (SYSID) algorithms for similar setups (Shirani Faradonbeh et al., 2018; Sarkar & Rakhlin, 2019; Oymak & Ozay, 2019; Sarkar et al., 2021; Zheng & Li, 2021). Without any structural assumptions, they have established finite time convergence results with a rate of $\tilde{\mathcal{O}}(\sqrt{n \cdot \text{poly}\,(r,m)\,/T})$, leading to a sample complexity of $\tilde{\mathcal{O}}(n \cdot \text{poly}\,(r,m)\,/\epsilon^2)$ for an ϵ-well approximation of model parameters. Such results are not satisfying for LTI systems with a large observation dimension n. One alternative way to directly applying the SYSID algorithm is to initially project the high-dimensional data to lower dimensions before conducting further analysis (Saul et al.; Churchland et al., 2012), either for efficiency or for interpretability. Moreover, these projections may often consist of "meta information" that are shared across different tasks. This is consistent with many scenarios in meta-learning settings where we deal with more complex datasets collected

[1]SEAS, Harvard University, Cambridge, USA. Correspondence to: Yuyang Zhang <yuyangzhang@g.harvard.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

from similar observers but possibly with different latent dynamics. The above facts motivate us to investigate the provable efficiency of this type of projection method and its generalization ability.

Contributions. In this paper, we study learning LTI systems with *high-dimensional noisy observations*. We adopt a two-stage procedure to first extract high-dimensional features, i.e. column space of the observer. We subsequently perform standard SYSID on the resulting low-dim data and recover the original model parameters. We establish sample complexity for this bifold procedure that scales as $\tilde{\mathcal{O}}([n + \text{poly}\,(r, m)]/\epsilon^2)$ which further reduces to only $\tilde{\mathcal{O}}(\text{poly}\,(r, m)\,/\epsilon^2)$ in the absence of observation noise. This then naturally leads to the following question:

Can this linear dependency on observer dimension n be possibly improved by some carefully designed algorithm?

We show that the linear dependency on n is unavoidable in the presence of observation noise–as shown in our lower bound result Theorem 4.1, requiring at least $\mathcal{O}(n/\epsilon^2)$ samples from our learning problem. Unfortunately, observation noises always exist for real-world applications, as the sensors always involve uncertainties. The lower bound also indicates that the sample complexity of the proposed algorithm is optimal up to some logarithmic factors and dimension-independent constants.

Additionally, our pragmatic solution–namely, the separation of learning the high-dimensional features (i.e. the observer column space) and learning the rest of the model parameters–can be extended to meta-learning setups by collecting meta-data from a set of dynamical systems that share the same observer model. By adopting a "leave-one-out" strategy for statistical consistency, we show that as long as the meta-data is *collectively* large in order of $\tilde{\mathcal{O}}(n/\epsilon^2)$, we can successfully obtain an ϵ-well approximation of all the systems parameters involved in the meta-data (Theorem 5.1). We finally note that such metadata with the same observer is common in real-world applications (Hajnal et al., 2023; Xia et al., 2021; Marks & Goard, 2021; Gallego et al., 2020). One example is in neuroscience where neuron activities are measured over a long period by the same set of electrodes or imaging devices. Although the subject may perform different tasks or demonstrate behavior corresponding to different latent dynamics, the observer model (i.e. the electrodes or the imaging device) remains the same, resulting in the metadata considered here.

To summarize our contributions: Firstly, we provide Column Space Projection-based SYSID (Col-SYSID) (Algorithm 1 in Section 3) that learns LTI systems through high-dimensional observations with complexity $\tilde{\mathcal{O}}(n/\epsilon^2)$ (Theorem 3.4 in Section 3). As compared with existing algorithms, ours reduces the polynomial dependency on the latent dimension r and the input dimension m. We then establish a sample complexity lower bound for this problem, indicating the optimality of the above result (Theorem 4.1 in Section 4). Our algorithmic idea is further extended to a meta-learning setting, where we have access to datasets from multiple similar systems. We provide an end-to-end framework for handling this meta-dataset and learning all included systems (Algorithm 3 in Section 5). With the help of the meta-dataset, we break the sample complexity lower bound in certain scenarios (Theorem 5.1 in Section 5).

Before proceeding, we set the following notations throughout the paper.

Notations: Without further explanation, let δ be any probability in $(0, 1/e)$. We use $\text{poly}\,(\cdot)$ to denote polynomial and logarithmic dependences. We use $\lesssim$ and $\gtrsim$ to hide absolute constants. We use $\tilde{\mathcal{O}}(\cdot)$ to hide all problem-related constants, absolute constants and logrithmic factors. Let $[N]$ be the set of integers $\{1, \ldots, N\}$. Let $\mathcal{M}_{[N]}$ denote the set $\{\mathcal{M}_n\}_{n\in[N]}$ and let $\mathcal{M}_{-n}$ denote the set $\{\mathcal{M}_{n'}\}_{n'\in[N]\setminus\{n\}}$ when the full set $[N]$ is clear from the context. For any matrix $M \in \mathbb{R}^{m\times n}$, let $\sigma_1(M) \geq \sigma_2(M) \geq \cdots \geq \sigma_{\max\{m,n\}}(M)$ be its singular values, and let $\sigma_{\min}(M)$ be the *minimum non-zero* singular value. Let $\|M\|$ denote its operator norm, and $M^\intercal$ denote its transpose. Let $\text{col}(M)$ be the column space of M and let Φ_M denote any orthonormal matrix whose columns form a basis of $\text{col}(M)$. For any orthonormal matrix Φ, we use $\Phi^\perp$ to denote the matrix such that $\begin{bmatrix}\Phi & \Phi^\perp\end{bmatrix}$ is unitary, and we refer to $\Phi^\perp$ as the orthogonal complement of Φ. For any *positive semi-definite matrix* $\Sigma \in \mathbb{R}^{n\times n}$, we slightly abuse the notation and let $\mathcal{N}(0, \Sigma)$ denote the distribution of $\Sigma^{1/2}x$, where $x \sim \mathcal{N}(0, I_n)$ follows the standard Gaussian distribution.

1.1. Other Related Works

Other Linear Models with Low-Dimensional Dynamics. Recently, there are lines of research on autoregressive models with low-dimensional representations (Qin, 2022; Dong et al., 2022; Poloni & Sbrana, 2019). Indeed, we may also be able to reconstruct low-dimensional dynamics from the model with the learned representations. However, sample complexity results are not currently available for such models. Another related line of research is on dynamic factor models (Hallin et al., 2023; Anderson et al., 2022; Breitung & Eickmeier, 2006). Anderson et al. (2022) only provides asymptotic convergence guarantees. Finite time convergence is indeed developed in Hallin et al. (2023). However, their results can not imply dependence on the system dimensions and are only about the convergence rates. Moreover, certain assumptions might not be easily verified in our setting.

General SYSID Algorithms. The most relevant sample complexity results are provided in recent SYSID literature for learning partially observed LTI systems (Zheng & Li, 2021; Lee, 2022; Djehiche & Mazhar, 2022; Sarkar et al., 2021; Oymak & Ozay, 2019). The last four papers focus on systems with stable latent dynamics, while the last paper extends to unstable latent dynamics. Unfortunately, all algorithms have suboptimal sample complexities. There also exist a line of work providing sample complexity for learning fully observed systems (Sarkar & Rakhlin, 2019; Shirani Faradonbeh et al., 2018), whose analysis techniques is applicable to unstable dynamics. This paper focuses on stable latent dynamics, and we leave for future directions to extend the results to unstable latent dynamics.

Existing SYSID Lower Bounds. There exist literature providing lower bounds for learning partially observed LTI systems. However, their results are not directly comparable to ours in our setting. The most relevant papers are Sun et al. (2023); Fattahi (2021). The former provides a lower bound depending on r/n, which is tailored for systems with $r \gg n$. The latter provides a lower bound proportional to $1/\epsilon^4$. The lower bound follows a slower rate than ours, though being logarithmic in the system dimensions. Sun et al. (2022) provides lower bounds defined for noise-free settings. Mao et al. (2021) provides lower bounds for systems with $C = I$. Bakshi et al. (2023) provides lower bounds for learning almost uncontrollable and unobservable systems. Other related works include Djehiche & Mazhar (2021); Jedra & Proutiere (2023); Simchowitz et al. (2018), which develop lower bound for fully observed systems.

General Subspace Learning Algorithms. To learn the observer column space, several existing literature may be related. Vaswani et al. (2018); Balzano et al. (2018); Candès et al. (2011); Blanchard et al. (2007) focus on analysis for PCA subspace learning. However, they assume that the dataset is i.i.d. sampled, which is not the case for dynamic systems. Dynamic factor model techniques are also related (Hallin et al., 2023). As discussed previously, the results can not imply the dependence on the system dimensions. Other ideas include Deng et al. (2020), whose sample complexity remains an open question.

2. Preliminaries and The Problem Setup

Consider linear dynamical systems in the form of Equation (1) with $x_0 = 0$ (for simplicity), process noises $w_t \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_w)$ and isotropic observation noises $\eta_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_\eta^2 I)$. Here Σ_w and $\sigma_\eta^2 I$ are *positive semi-definite matrices.* We denote such systems by $\mathcal{M} = (r, n, m, A, B, C, \Sigma_w, \sigma_\eta^2 I)$. In standard system identification setup, given the input-output trajectory data from the system, the objective is to learn system parameters up to the well-known similarity transformation class, i.e., to learn tuple $(\widehat{A}, \widehat{B}, \widehat{C})$ such that for some invertible matrix S,[1]

$$\widehat{A} = S^{-1}AS, \quad \widehat{B} = S^{-1}B, \quad \widehat{C} = CS.$$

Furthermore, in order to ensure this learning problem is well-posed, we assume (A, B) is controllable and (A, C) is observable. This is often referred to as a *minimal realization* (a state-space description of minimal size that explains the given input-output data (Schutter, 2000)). The corresponding system $\mathcal{M}$ is then called a minimal system.

High-dimensional System Identification Problem (`HD-SYSID`): Consider learning minimal system $\mathcal{M} = (r, n, m, A, B, C, \Sigma_w, \sigma_\eta^2 I)$ with *high-dimensional observations* and *low-dimensional latent states and inputs*; namely, $r, m \ll n$. Here, covariances Σ_w, $\sigma_\eta^2 I$ are *positive semi-definite covariances*. For $k = 1, 2$, we choose input trajectories $\mathcal{U}_k = \{u_{k,t}\}_{t=0}^{T_k - 1}$ and get corresponding observations $\mathcal{Y}_k = \{y_{k,t}\}_{t=0}^{T_k}$. Here, every input $u_{k,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_u)$ is sampled independently with *positive definite covariance* $\Sigma_u \succ 0$.[2] With the two datasets $\mathcal{D}_1 = \mathcal{Y}_1 \cup \mathcal{U}_1$ and $\mathcal{D}_2 = \mathcal{Y}_2 \cup \mathcal{U}_2$, our objective is to output approximate system matrices $\left(\widehat{A}, \widehat{B}, \widehat{C}\right)$ such that with high probability,

$$\max\left\{\left\|S^{-1}AS - \widehat{A}\right\|, \left\|S^{-1}B - \widehat{B}\right\|, \left\|CS - \widehat{C}\right\|\right\} \leq \epsilon$$

for some invertible matrix S.

In `HD-SYSID` problem, we specifically consider systems with $r, m \ll n$ and isotropic observation noise. We select inputs with positive definite covariance Σ_u so that the system is fully excited and therefore learnable. The problem setup can be extended to the case where we have access to $K \geq 2$ independent data trajectories. Our proposed approach and theoretical analysis in the following sections can also be readily adapted.

3. SYSID with Column Space Projection

Before diving into the details, we first state the necessary assumption and definition for `HD-SYSID`.

Assumption 3.1. There exist constants $\psi_A \geq 1$ and $\rho_A \in (0, 1)$, which are independent of system dimensions, such that

$$\left\|A^i\right\| \leq \psi_A \left(\rho_A\right)^{i-1}, \quad \forall i \in \mathbb{N}.$$

[1] Here, S represents a change of basis for the latent variables resulting in a modification of the system representations from (A, B, C) to $(S^{-1}AS, S^{-1}B, CS)$–yet describing exactly the same input-output behavior. See Section 4.4 in (Hespanha, 2018).

[2] Our approach can be easily extended to a weaker assumption on Σ_u: $(A, (\Sigma_w + B\Sigma_u B^\intercal)^{1/2})$ being controllable. Here we assume $\Sigma_u \succ 0$ to keep the results clean and interpretable.

The existence of ψ_A and ρ_A is a standard assumption (Oymak & Ozay, 2019) and is guaranteed by Lemma 6 in Talebi et al. (2023) as long as the spectral radius of A is smaller than 1. However, the constants may be dependent on system dimensions in the literature. Here we assume ψ_A and ρ_A to be independent of system dimensions.

Definition 3.2 (System Identification Oracle (`Sys-Oracle`))**.** Consider a system $\mathcal{M} = (r, n, m, A, B, C, \Sigma_w, \Sigma_\eta)$ with a minimal realization (A, B, C) and arbitrary dimensions (r, m, n). Assume $\mathcal{M}$ satisfy Assumption 3.1. Given an input-output trajectory data sequence $\mathcal{D} = \mathcal{Y} \cup \mathcal{U}$ where $\mathcal{U} = \{u_t\}_{t=0}^{T-1}$ with $u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_u)$ and $\mathcal{Y} = \{y_t\}_{t=0}^{T-1}$ is the corresponding output, the `Sys-Oracle` outputs approximation $\left(\widehat{A}, \widehat{B}, \widehat{C}\right)$ such that with probability at least $1 - \delta$,

$$\max\left\{\left\|S^{-1}AS - \widehat{A}\right\|, \left\|S^{-1}B - \widehat{B}\right\|, \left\|CS - \widehat{C}\right\|\right\} \leq \frac{\text{poly}\,(r, n, m)}{\sqrt{T}} \cdot \kappa_2,$$

for any $T \geq \kappa_1 \cdot \text{poly}\,(r, n, m)$ and some invertible matrix S. Here $\kappa_1 = \kappa_1\,(\mathcal{M}, \mathcal{U}, \delta)\,, \kappa_2 = \kappa_2\,(\mathcal{M}, \mathcal{U}, \delta)$ are problem-related constants independent of system dimensions modulo logarithmic factors.

The `Sys-Oracle` defined above is applicable to all minimal systems with arbitrary dimensions (r, n, m), which does not necessarily lie in the regime where $n \gg r, m$. It represents standard system identification algorithms, including the celebrated Ho-Kalman algorithm (Sarkar et al., 2021; Oymak & Ozay, 2019). The constraints of such existing algorithms are captured by Assumption 3.1. These algorithms also require the inputs and noises to be sampled independently and fully exciting, which is captured by the definition of system $\mathcal{M}$ and inputs $\mathcal{U}$ in `Sys-Oracle`. In Appendix D, we will provide an example of such an oracle for the completeness of this paper.

3.1. The Algorithm

Our proposed algorithm consists of two components. Firstly in Line 2, using the first data trajectory, we approximate the column space of the high-dimensional observer C and get matrix $\widehat{\Phi}_C \in \mathbb{R}^{n \times \text{rank}(C)}$. The columns of $\widehat{\Phi}_C$ form an orthonormal basis of the space. This approximation is accomplished by the Column Space Approximation Subroutine (`Col-Approx`), which essentially calculates the covariance of the observations $\{y_{1,t}\}_{t=0}^{T_1}$ and extracts the eigenspace of the large eigenvalues. The details of the subroutine is postponed to Section 3.3.

Secondly, in Lines 3-5, we learn the system parameters with the second data trajectory. With the help of $\widehat{\Phi}_C$, we can project the high-dimensional observations onto lower dimensions, i.e. we let $\tilde{y}_{2,t} = \widehat{\Phi}_C{}^\intercal y_{2,t}$. This projected sequence of observations satisfies the following dynamics

$$x_{t+1} = Ax_t + Bu_t + w_t,$$
$$\tilde{y}_{2,t} = \widehat{\Phi}_C{}^\intercal C x_t + \widehat{\Phi}_C{}^\intercal \eta_t.$$

that is generated by an equivalent system $\widehat{\mathcal{M}}$ denoted by

$$\widehat{\mathcal{M}} = (r, \text{rank}(\widehat{\Phi}_C), m, A, B, \widehat{\Phi}_C{}^\intercal C, \Sigma_w, \sigma_\eta^2 I). \quad (2)$$

Remark 3.3 (Why two trajectories?). . The reason for us to switch to the second trajectory for parameter learning is to *ensure the independence between $\widehat{\Phi}_C$ and the second trajectory,* which will ensure that $\widehat{\Phi}_C{}^\intercal \eta_t$ is still independent of other system variables. This is critical for the application of `Sys-Oracle`. Note that if the algorithm is equipped with any `Sys-Oracle` that handles dependent noise, one can easily simplify Algorithm 1 to a single trajectory. Developing such `Sys-Oracle` is an interesting future direction and ideas from (Tian et al., 2023; Simchowitz et al., 2018) may be related. □

Lastly, we then feed this low-dimensional dataset to the `Sys-Oracle` for learning the corresponding low-dimensional parameters $\left(\widehat{A}, \widehat{B}, \tilde{C}\right) \approx \left(A, B, \widehat{\Phi}_C{}^\intercal C\right)$. Finally, we recover C from $\tilde{C}$ through $\widehat{\Phi}_C$. If $\widehat{\Phi}_C$ approximates the column space of C well, then $\widehat{\Phi}_C \widehat{\Phi}_C{}^\intercal$ should also be close to $\Phi_C \Phi_C{}^\intercal$, which is the projection onto the column space of C. Therefore, it is intuitive to expect that $\widehat{\Phi}_C \widehat{\Phi}_C{}^\intercal C \approx \Phi_C \Phi_C{}^\intercal C = C$.

Algorithm 1 Column Space Projection-based SYSID (Col-SYSID)

1: **Inputs:** Data $\mathcal{Y}_1, \mathcal{D}_2$; Noise covariance Σ_η; Subroutines `Col-Approx`, `Sys-Oracle`;
2: Approximate the observer column space:
$$\widehat{\Phi}_C \leftarrow \texttt{Col-Approx}\,(\mathcal{Y}_1)$$
3: Project dataset $\mathcal{D}_2 = \{y_{2,t}\}_{t=0}^{T_2} \cup \{u_{2,t}\}_{t=0}^{T_2-1}$ onto the column space:
$$\tilde{\mathcal{D}}_2 \leftarrow \left\{\widehat{\Phi}_C{}^\intercal y_{2,t}\right\}_{t=0}^{T_2} \cup \{u_{2,t}\}_{t=0}^{T_2-1}$$
4: Identify low-dimensional parameters:
$$\widehat{A}, \widehat{B}, \tilde{C} \leftarrow \texttt{Sys-Oracle}\left(\tilde{\mathcal{D}}_2\right)$$
5: Recover the high-dimensional observer:
$$\widehat{C} \leftarrow \widehat{\Phi}_C \tilde{C}$$
6: **Outputs:** $\left(\widehat{A}, \widehat{B}, \widehat{C}\right)$

3.2. Sample Complexity of the Algorithm

We now provide the following sample complexity result for Algorithm 1.

Theorem 3.4. *Consider system $\mathcal{M}$ and datasets $\mathcal{D}_1 = \mathcal{U}_1 \cup \mathcal{Y}_1, \mathcal{D}_2 = \mathcal{U}_2 \cup \mathcal{Y}_2$ (with lengths T_1, T_2 respectively) in* `HD-SYSID`*. Suppose $\mathcal{M}$ satisfies Assumption 3.1. If*

$$T_1 \gtrsim \kappa_3 \cdot n^2 r^2, \quad T_2 \geq \kappa_1 \cdot poly\,(r, m)\,,$$

then $(\widehat{A}, \widehat{B}, \widehat{C})$ from Algorithm 1 satisfy the following for some invertible matrix S with probability at least $1-\delta$

$$\max\left\{\left\|S^{-1}AS - \widehat{A}\right\|, \left\|S^{-1}B - \widehat{B}\right\|, \left\|CS - \widehat{C}\right\|\right\} \lesssim \kappa_4 \cdot \sqrt{\frac{n}{T_1}}\left\|\widehat{C}\right\| + \kappa_2 \cdot \sqrt{\frac{poly\,(r,m)}{T_2}}.$$

Here $\kappa_1 = \kappa_1(\widehat{\mathcal{M}}, \mathcal{U}_2, \delta), \kappa_2 = \kappa_2(\widehat{\mathcal{M}}, \mathcal{U}_2, \delta), \kappa_3 = \kappa_3\left(\mathcal{M}, \mathcal{U}_{[2]}, \delta\right)$ and $\kappa_4 = \kappa_4\left(\mathcal{M}, \mathcal{U}_1, \delta\right)$ are all problem-related constants independent of system dimensions modulo logarithmic factors with $\widehat{\mathcal{M}}$ defined in Equation (2). Also, κ_1 and κ_2 are defined in Definition 3.2, while the definitions of κ_3, κ_4 are summarized in Theorem A.1 in the appendix.

The above error bound consists of two terms. The first term, also the dominating term, is the error of observer column space approximation and the second term is due to learning of the rest of the system. Together, the error bound directly translates to a sample complexity of $\tilde{\mathcal{O}}([n+\text{poly}\,(r,m)]/\epsilon^2)$. This complexity is equivalent to $\tilde{\mathcal{O}}(n/\epsilon^2)$ in our setting with $r, m \ll n$.

Remark 3.5. Here the first term indicates an interesting insight: *the norm of our approximated observation matrix may affect the performance of the algorithm*. Therefore, in practice, one might want to choose `Sys-Oracle` that outputs $\tilde{C}$ with a reasonable norm (line 4 of Algorithm 1). This ensures $\left\|\widehat{C}\right\|$ is not too large because $\left\|\widehat{C}\right\| \leq \left\|\widehat{\Phi}_C\right\|\left\|\tilde{C}\right\| = \left\|\tilde{C}\right\|$. □

For more concrete results, we instantiate Algorithm 1 with Ho-Kalman algorithm (Sarkar et al., 2021; Oymak & Ozay, 2019), Algorithm 4 in Appendix D as `Sys-Oracle`. Then the above theorem leads to the following corollary.

Corollary 3.6. *Consider the special minimal system $\mathcal{M} = (r, n, m, A, B, C, \sigma_w^2 I, \sigma_\eta^2 I)$ and datasets $\mathcal{D}_1 = \mathcal{U}_1 \cup \mathcal{Y}_1, \mathcal{D}_2 = \mathcal{U}_2 \cup \mathcal{Y}_2$ (with length T_1, T_2 respectively), where the inputs are sampled independently from $\mathcal{N}(0, \sigma_u^2 I)$. Suppose $\mathcal{M}$ satisfies Assumption 3.1. If*

$$T_1 \gtrsim \tilde{\kappa}_3 \cdot n^2 r^2, \quad T_2 \geq \tilde{\kappa}_1 \cdot r^3(r+m),$$

then $(\widehat{A}, \widehat{B}, \widehat{C})$ from Algorithm 1 satisfy the following for some invertible matrix S with probability at least $1-\delta$

$$\max\left\{\left\|S^{-1}AS - \widehat{A}\right\|, \left\|S^{-1}B - \widehat{B}\right\|, \left\|CS - \widehat{C}\right\|\right\} \lesssim \tilde{\kappa}_4 \cdot \sqrt{\frac{n}{T_1}}\left\|\widehat{C}\right\| + \tilde{\kappa}_2 \cdot \sqrt{\frac{r^5(r+m)}{T_2}}.$$

Here $\tilde{\kappa}_1, \tilde{\kappa}_2, \tilde{\kappa}_3, \tilde{\kappa}_4$ are all problem-related constants only dependent of system $\mathcal{M}$ and inputs $\mathcal{U}_1 \cup \mathcal{U}_2$ (as compared to κ_1 and κ_2 in Theorem 3.4 that depend on $\widehat{\mathcal{M}}$). They are also independent of system dimensions modulo log factors. Detailed definitions of the constants are listed in Corollary D.2.

Based on the above corollary, we now compare the error of our algorithm instantiated with the example oracle, denoted by Δ_1, with the error of the example oracle when it is directly applied to the dataset, denoted by Δ_2. For a fair comparison, we consider Δ_2 as the error with datasets generated by inputs $\mathcal{U} = \{u_t\}_{t=0}^{T_1+T_2-1}$ of length $T_1 + T_2$. Here the inputs are sampled independently from $\mathcal{N}(0, \sigma_u^2 I)$.

The error of our algorithm is upper bounded by

$$\Delta_1 \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{n}{T_1}} + \sqrt{\frac{r^5(r+m)}{T_2}}\right).$$

The above error directly translates to the sample complexity of $\tilde{\mathcal{O}}([n + r^5(r+m)]/\epsilon^2) = \tilde{\mathcal{O}}(n/\epsilon^2)$ when $r, m \ll n$. On the other hand, from Corollary D.1

$$\Delta_2 \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{r^5(r+n+m)}{T_1+T_2}}\right).$$

And this gives a sample complexity of order $\tilde{\mathcal{O}}([nr^5 + r^5(r+m)]/\epsilon^2) = \tilde{\mathcal{O}}(nr^5/\epsilon^2)$ when $r, m \ll n$. Therefore, in the regime where $r, m \ll n$, it is clear that our algorithm enjoys a better sample complexity as compared to directly applying the example oracle.

3.3. Proof Ideas of Theorem 3.4

Although conceptually simple, justifying this tight sample complexity result is hard. The difficulty lies in ensuring the accuracy of the observer column space approximation $\widehat{\Phi}_C$, which is learned mainly by PCA on the given data (detailed in the following algorithm).

Algorithm 2 `Col-Approx`

1: **Input:** Observations $\mathcal{Y} = \{y_t\}_{t=0}^{T}$;
2: Calculating data covariance

$$\Sigma_y \leftarrow \sum_{t=0}^{T} y_t y_t^\intercal$$

3: Estimating observer rank

$$\widehat{r}_c \leftarrow \arg\max_i \left(\sigma_i(\Sigma_y) - \sigma_{i+1}(\Sigma_y) > T^{3/4}\right)$$

4: Estimating column space

$$\widehat{\Phi}_C \leftarrow \text{ first } \widehat{r}_c \text{ eigenvectors of } \Sigma_y$$

5: **Output:** $\widehat{\Phi}_C$

Existing analysis of PCA (Candès et al., 2011; Vaswani et al., 2018; Chen & Storey, 2015), matrix factorization (Gribonval et al., 2015) and other subspace learning techniques (Tripuraneni et al., 2021; Zhang et al., 2023) do not apply to our setting. This is because these methods assume i.i.d. data samples, while our dataset $\mathcal{D}_1$ consists of correlated data points generated by a dynamical system. We need to apply martingale tools (Abbasi-yadkori et al., 2011; Sarkar & Rakhlin, 2019) for our analysis.

Moreover, the noises in the dynamical systems accumulate across time steps, leading to large noise terms in the analysis. Directly applying the aforementioned tools leads to loose sample complexity bounds. Since we are learning a subspace embedded in $\mathbb{R}^n$, it is reasonable to expect a sample complexity of $\mathcal{O}(n/\epsilon^2)$ for an ϵ-good approximation of the subspace [3]. We achieve this result by developing our own subspace perturbation lemma specifically tailored to this setting. The above ideas on analyzing `Col-Approx` translate to the following lemma.

Lemma 3.7. *Consider system $\mathcal{M}$ and dataset $\mathcal{D}_1 = \mathcal{U}_1 \cup \mathcal{Y}_1$ in `HD-SYSID`. Suppose $\mathcal{M}$ satisfies Assumption 3.1. If*

$$T_1 \gtrsim \kappa_5 \cdot n^2 r^2,$$

then $\widehat{\Phi}_C = \texttt{Col-Approx}(\mathcal{Y}_1, \Sigma_\eta)$ satisfies the following with probability at least $1-\delta$

$$\textit{rank}(\widehat{\Phi}_C) = \textit{rank}(C), \quad \left\|\widehat{\Phi}_C^{\perp\intercal} \Phi_C\right\| \lesssim \kappa_4 \cdot \sqrt{\frac{n}{T_1}}.$$

Here $\kappa_4 = \kappa_4(\mathcal{M}, \mathcal{U}_1, \delta)$ and $\kappa_5 = \kappa_5(\mathcal{M}, \mathcal{U}_1, \delta)$ are both problem-related constants independent of system dimensions modulo logarithmic factors (details in Lemma A.2).

Recall Φ_C denotes the orthonormal matrix whose columns form a basis of $\text{col}(C)$ and $\widehat{\Phi}_C^\perp$ denote the matrix such that $\begin{bmatrix}\widehat{\Phi}_C^\perp & \widehat{\Phi}_C\end{bmatrix}$ is unitary.

Proof Sketch of Lemma 3.7. $\widehat{\Phi}_C$ is constructed from the eigenvectors of matrix $\Sigma_y = \sum_{t=0}^{T_1} y_{1,t} y_{1,t}^\intercal$. We decompose it as follows

$$\sum_{t=0}^{T_1} y_{1,t} y_{1,t}^\intercal = \underbrace{C\left(\sum_{t=0}^{T_1} x_{1,t} x_{1,t}^\intercal\right) C^\intercal}_{\Sigma_C} + \sigma_\eta^2 (T_1+1) I + \underbrace{\sum_{t=0}^{T_1} \left(\eta_{1,t}\eta_{1,t}^\intercal - \sigma_\eta^2 I\right)}_{\Delta_2} + \underbrace{\sum_{t=0}^{T_1} \left(C x_{1,t} \eta_{1,t}^\intercal + \eta_{1,t} x_{1,t}^\intercal C^\intercal\right)}_{\Delta_1 + \Delta_1^\intercal}.$$

Here Σ_C contains the information on the observer column space Φ_C, while Δ_1 and Δ_2 are noise terms. The rest of the proof is decomposed into three steps. In the first step, we upper bound the noise terms Δ_1, Δ_2 and lower bound the latent state covariance $\sum_{t=0}^{T_1} x_{1,t} x_{1,t}$. In the second step, we show that the eigenvectors in the column space of C has eigenvalues much larger then the other eigenvalues. This gap eigenvalue gap enables us to identify the dimension of the column space of C. Finally, in the third step, we apply our own subspace perturbation result to bound the accuracy of the approximated column space.

Step 1. With system (A, B) being controllable, the inputs injected fully excite the latent dynamics, leading to the following lower bound on $\sum_{t=0}^{T_1} x_{1,t} x_{1,t}$:

$$\sum_{t=0}^{T_1} x_{1,t} x_{1,t} \succ \tilde{\mathcal{O}}(T_1) I.$$

The upper bound on Δ_2, i.e. $\|\Delta_2\| \le \tilde{\mathcal{O}}(\sqrt{T_1})$, follows from standard guassian concentration arguments. To upper bound Δ_1, we apply the martingale tools from previous works (Abbasi-yadkori et al., 2011; Sarkar & Rakhlin, 2019) and get the following relative error bounds

$$\left\|(\Sigma_C + T_1 I)^{-\frac{1}{2}} \Delta_1\right\| \le \tilde{\mathcal{O}}(\sqrt{T_1}). \tag{3}$$

Step 2. Since the perturbations are small, standard eigenvalue perturbation bounds gives

$$\sigma_{\text{rank}(C)}(\Sigma_y) \ge \tilde{\mathcal{O}}(T_1), \quad \sigma_{\text{rank}(C)+1}(\Sigma_y) \le \tilde{\mathcal{O}}(\sqrt{T_1}).$$

This eigenvalue gap as large as $\tilde{\mathcal{O}}(T_1)$ makes it easy to determine the rank of C. Therefore, with high probability, $\text{rank}(\widehat{\Phi}_C) = \text{rank}(C)$.

Step 3. Finally, we bound the accuracy of the eigenspace of $\Sigma_y = \Sigma_C + \Delta_2 + (\Delta_1 + \Delta_1^\intercal)$. Here Δ_1 is the so-called relative perturbation, because we bound its norm by comparing it with the information $(\Sigma_C + T_1 I)^{\frac{1}{2}}$ (in Equation (3)). Since $\left\|(\Sigma_C + T_1 I)^{\frac{1}{2}}\right\|$ can be large, directly applying standard subspace perturbation results leads to sub-optimal error bound. For this, we develop a specific relative subspace perturbation bound, adapted from classic eigenspace perturbation results. It ensures both Δ_2 and the relative noise Δ_1 will not perturb the eigenspace of Σ_y too much. This finishes the proof. □

With the above lemma, the proof of Theorem 3.4 can then be decomposed into two steps. *In the first step*, we guarantee the feasibiliy to learn the low-dimensional system parameters $(A, B, \widehat{\Phi}_C^\intercal C)$ from the projected dataset (Line 3 in Algorithm 1). Since $\widehat{\Phi}_C$ is accurate enough, the projected dataset almost preserves all the information. This ensures that the equivalent system generating the projected

[3] Other intuitions for this complexity comes from standard covariance concentration results for i.i.d. data (Corollary 2.1 of (Rudelson, 1999)). In these results, $\mathcal{O}(n/\epsilon^2)$ samples are required to learn an ϵ-good covariance, which directly translates to a good column space estimation of the covariance.

dataset ($\widehat{\mathcal{M}}$ in Equation (2)) is observable and controllable. Therefore, `Sys-Oracle` outputs accurate estimation of the low-dimensional system parameters. *In the second step*, we analyze the errors of the recovered high-dimensional parameters $(\widehat{A}, \widehat{B}, \widehat{\Phi}_C\tilde{C})$, where $\tilde{C}$ is our estimation of $\widehat{\Phi}_C{}^\intercal C$. The errors on $\widehat{A}$ and $\widehat{B}$ are automatically bounded by the definition of `Sys-Oracle`. The error on the recovered high-dimensional observer $\widehat{C} = \widehat{\Phi}_C\tilde{C}$ can be decomposed as follows

$$\begin{aligned}\left\|CS - \widehat{C}\right\| &\le \left\|CS - \widehat{\Phi}_C\widehat{\Phi}_C{}^\intercal CS\right\| + \left\|\widehat{\Phi}_C\widehat{\Phi}_C{}^\intercal CS - \widehat{\Phi}_C\tilde{C}\right\| \\ &= \left\|\widehat{\Phi}_C^\perp\left(\widehat{\Phi}_C^\perp\right)^\intercal CS\right\| + \left\|\widehat{\Phi}_C{}^\intercal CS - \tilde{C}\right\|.\end{aligned}$$

Here the first term is the column space approximation error and the second term is the error learning the low-dimensional observer $\tilde{C} \approx \widehat{\Phi}_C{}^\intercal C$. Combining the above inequality with Lemma 3.7 and Definition 3.2, along with some algebraic arguments, proves the theorem.

4. Lower Bounds

Now that Algorithm 1 accomplishes `HD-SYSID` with sample complexity $\tilde{\mathcal{O}}(n/\epsilon^2)$, one natural question is: *Is this linear dependency on observer dimension n unavoidable?* In this section, we provide the following theorem to show that this dependency is necessary.

While we present this lower bound for learning LTI systems with a single input-output data trajectory, it can be extended to multiple trajectories. The more general version is deferred to the appendix (Theorem B.2).

Theorem 4.1. *Suppose $n \ge 2$, and choose positive scalars $\delta \le \frac{1}{2}$ and $\epsilon \le 0.6$. Consider the class of minimal systems $\mathcal{M} = (r, n, m, A, B, C, \Sigma_w, \Sigma_\eta)$ with different A,B,C matrices. All parameters except A, B, C are fixed and known. Moreover, $r < n$ and Σ_η is positive definite. Let $\mathcal{D} = \{y_t\}_{t=0}^{T} \cup \{u_t\}_{t=0}^{T-1} \cup \{\text{all known parameters}\}$ denote the associated single trajectory dataset. Here the input u_t satisfy: 1). u_t is independently sampled; 2). $\mathbb{E}(u_t) = 0$. Consider any estimator $\widehat{f}$ mapping $\mathcal{D}$ to $(\widehat{A}(\mathcal{D}), \widehat{B}(\mathcal{D}), \widehat{C}(\mathcal{D})) \in \mathbb{R}^{r\times r} \times \mathbb{R}^{r\times m} \times \mathbb{R}^{n\times r}$. If*

$$T < \frac{\phi_\eta(1-2\delta)\log 1.4}{50\left(\psi_w + \psi_u\right)} \cdot \frac{n + \log\frac{1}{2\delta}}{\epsilon^2},$$

there exists a system $\mathcal{M}_0 = (r, n, m, A_0, B_0, C_0, \Sigma_w, \Sigma_\eta)$ with dataset $\mathcal{D}$ such that

$$\mathbb{P}\left\{\left\|C_0B_0 - \widehat{C}\left(\mathcal{D}\right)\widehat{B}\left(\mathcal{D}\right)\right\| \ge \epsilon\right\} \ge \delta.$$

Here $\mathbb{P}$ denotes the distribution of $\mathcal{D}$ generated by system $\mathcal{M}_0$. Related constants are defined as follows

$$\begin{aligned}\phi_\eta &\coloneqq \sigma_{\min}(\Sigma_\eta), \quad \psi_w \coloneqq \sigma_1(\Sigma_w), \\ \psi_u &\coloneqq \max_{t\in[0,T-1]} \sigma_1\left(\mathbb{E}(u_t u_t{}^\intercal)\right).\end{aligned}$$

The above theorem indicates that with less than $\mathcal{O}(n/\epsilon^2)$ samples, any estimator has a constant probability to fail learning the product CB. This error lower bound on $\widehat{C}(\mathcal{D})\widehat{B}(\mathcal{D})$ can be translated to corresponding errors on

$$\max\left\{\left\|S^{-1}B - \widehat{B}\left(\mathcal{D}\right)\right\|, \left\|CS - \widehat{C}\left(\mathcal{D}\right)\right\|\right\} \ge \epsilon,$$

under mild conditions. This indicates that the estimator fails to learn either B or C well. (The proof details are provided in Corollary B.3). Another perspective to understand this error on $\widehat{C}(\mathcal{D})\widehat{B}(\mathcal{D})$ is as follows. In the celebrated Ho-Kalman algorithm for `HD-SYSID`, the estimation error the Hankel operator is lower bounded by the estimation error of $\widehat{C}(\mathcal{D})\widehat{B}(\mathcal{D})$. Therefore, our lower bound indicates the failure of estimating the Hankel operator, which prevents the algorithm from giving satisfying outputs.

The above results then give a $\mathcal{O}(n/\epsilon^2)$ sample complexity lower bound on `Sys-Oracle`. Therefore, the sample complexity of Algorithm 1, i.e. $\tilde{\mathcal{O}}(n/\epsilon^2)$, is *optimal* up to logarithmic and problem-related constants.

We defer the proof of Theorem 4.1 to Appendix B. But we would like to briefly discuss the reason behind this unavoidable dependency on n. Therein, we construct a set of system instances with different observation matrices such that their column spaces are "relatively close" but "quantifiably distinct". In fact, we take advantage of the fact that in high dimensional space $\mathbb{R}^n$, there are too many distinct subspaces that are close to each other. And therefore, any estimator needs correspondingly many samples, i.e. $\mathcal{O}(n/\epsilon^2)$, to distinguish the true column space from many other candidates. This difficulty is indeed due to the high dimensional observation noise.

In the following section, we introduce a more general problem pertinent to high-dimensional "meta-datasets." While this lower bound still has implications even for this general problem, we show how to confine its implied difficulty on the total length of the meta-dataset and not on its personalized portions.

5. Meta SYSID

Thus far, we have introduced Algorithm 1 for `HD-SYSID` that provides near-optimal dependency on system dimensions utilizing the idea of low-dimensional embedding. Now we will investigate how this algorithmic idea facilitates learning from so-called meta-datasets (or metadata)—multiple datasets from different systems sharing the same observer. As shown in (Hajnal et al., 2023; Xia et al., 2021; Marks & Goard, 2021; Gallego et al., 2020), this is a real-world setting of significant importance.

Following our analysis in Section 3.3, we observe that the $\tilde{\mathcal{O}}(n/\epsilon^2)$ sample complexity arises from the error of column space $\widehat{\Phi}_C$. Modulo this error, the sample complexity

reduces vastly to $\tilde{\mathcal{O}}(\text{poly}\,(r,m)\,/\epsilon^2)$. This motivates us to again separate the learning of the observer column space Φ_C from the learning of personalized system parameters. Utilizing multiple datasets from different systems, we can collectively learn the observer column space much more accurately. Subsequently, equipped with this accurate column space approximation, there is hope to learn every single system with much fewer samples. To formalize the above idea, we introduce the following "High-dimensional Meta System Identification Problem (`Meta-SYSID`)."

High-dimensional Meta System Identification Problem (`Meta-SYSID`): Consider the identification of K minimal systems $\mathcal{M}_k = (r, n, m, A_k, B_k, C, \Sigma_{w,k}, \sigma^2_{\eta,k} I), k \in [K]$ with the same dimensions ($r, m \ll n$) and observation matrix C. Here covariances $\{\Sigma_{w,k}, \sigma^2_{\eta,k} I\}_{k\in[K]}$ are *positive semi-definite matrices*. For every system $\mathcal{M}_k$, we choose an input sequence $\mathcal{U}_k = \{u_{k,t}\}_{t=0}^{T_k-1}$ and get observations $\mathcal{Y}_k = \{y_{k,t}\}_{t=0}^{T_k}$. Here every input $u_{k,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_{u,k})$ is sampled independently from both the system variables and other inputs with *positive definite covariance* $\Sigma_{u,k}$. This single trajectory dataset is denoted by $\mathcal{D}_k = \mathcal{Y}_k \cup \mathcal{U}_k$. With the K datasets $\bigcup_{k\in[K]} \mathcal{D}_k$, our objective is to learn all system parameters well. Namely, we aim to identify $\{\widehat{A}_k, \widehat{B}_k, \widehat{C}_k\}_{k=1}^K$ such that, for every $k \in [K]$, the following holds for some invertible matrix S_k with high probability

$$\max\left\{\left\|S_k^{-1} A_k S_k - \widehat{A}_k\right\|, \left\|S_k^{-1} B_k - \widehat{B}_k\right\|, \left\|C S_k - \widehat{C}_k\right\|\right\} \le \epsilon.$$

5.1. The Meta-Learning Algorithm and Its Sample Complexity

Algorithm 3 Meta Space Projection-based SYSID (Meta-SYSID)

1: **Inputs:** Meta Datasets $\mathcal{D}_{[K]} = \mathcal{Y}_{[K]} \cup \mathcal{U}_{[K]}$; Subroutines `Col-Approx`, `Sys-Oracle`;
2: **for** $k \in [K]$ **do**
3: Leave one out and approximate observer column space
$$\widehat{\Phi}_{C,k} \leftarrow \texttt{Col-Approx}\,(\mathcal{Y}_{-k})$$
4: Project dataset $\mathcal{D}_k = \{y_{k,t}\}_{t=0}^{T_k} \cup \{u_{k,t}\}_{t=0}^{T_k-1}$ onto the column space:
$$\tilde{\mathcal{D}}_k \leftarrow \{\widehat{\Phi}_{C,k}{}^\intercal y_{k,t}\}_{t=0}^{T_k} \cup \{u_{k,t}\}_{t=0}^{T_k-1}$$
5: Identify low-dimensional parameters:
$$\widehat{A}_k, \widehat{B}_k, \tilde{C}_k \leftarrow \texttt{Sys-Oracle}\left(\tilde{\mathcal{D}}_k\right)$$
6: Recover the high-dimensional observer:
$$\widehat{C}_k \leftarrow \widehat{\Phi}_{C,k} \tilde{C}_k$$
7: **end for**
8: **Outputs:** $\{\widehat{A}_k, \widehat{B}_k, \widehat{C}_k\}_{k\in[K]}$

Algorithm 3 follows similar steps as Algorithm 1. For every system k, we approximate the column space, learn the low-dimensional system parameters, and finally project them back into the high-dimension. And similarly, to ensure the independence between $\widehat{\Phi}_{C,k}$ and $\tilde{\mathcal{D}}_k$, the dataset $\tilde{\mathcal{D}}_k$ itself is left out in the first step. This independence is critical for the application of the `Sys-Oracle` subroutine.

Now we provide the theoretical guarantee for the above algorithm.

Theorem 5.1. *Consider systems $\mathcal{M}_{[K]}$ and datasets $\mathcal{D}_{[K]} = \mathcal{U}_{[K]} \cup \mathcal{Y}_{[K]}$ in `Meta-SYSID`. Suppose $\mathcal{M}_{[K]}$ satisfy Assumption 3.1. Then for any fixed system $k_0 \in [K]$, if $\{T_k\}_{k\in[K]}$ and $T_{-k_0} := \sum_{k\neq k_0} T_k$ satisfy*

$$T_{k_0} \ge \kappa_1 \cdot \textit{poly}\,(r,m)\,, \quad T_{-k_0} \gtrsim \kappa_3 \cdot n^2 r^2 \log^8 K,$$

then $(\widehat{A}_{k_0}, \widehat{B}_{k_0}, \widehat{C}_{k_0})$ from Algorithm 3 satisfy the following for some invertible matrix S with probability at least $1-\delta$

$$\max\left\{\left\|S^{-1} A_{k_0} S - \widehat{A}_{k_0}\right\|, \left\|S^{-1} B_{k_0} - \widehat{B}_{k_0}\right\|, \left\|C_{k_0} S - \widehat{C}_{k_0}\right\|\right\}$$
$$\lesssim \kappa_2 \cdot \sqrt{\frac{\textit{poly}\,(r,m)}{T_{k_0}}} + \kappa_4 \cdot \sqrt{\frac{n}{T_{-k_0}}} \left\|\widehat{C}_{k_0}\right\|.$$

Here $\kappa_1 = \kappa_1(\widehat{\mathcal{M}}_{k_0}, \mathcal{U}_{k_0}, \delta), \kappa_2 = \kappa_2(\widehat{\mathcal{M}}_{k_0}, \mathcal{U}_{k_0}, \delta), \kappa_3 = \kappa_3\left(\mathcal{M}_{[K]}, \mathcal{U}_{[K]}, \delta\right), \kappa_4 = \kappa_4\left(\mathcal{M}_{-k_0}, \mathcal{U}_{-k_0}, \delta\right)$ are all problem-related constants independent of system dimensions modulo logarithmic factors. κ_1, κ_2 are defined in Definition 3.2, while the definitions of κ_3, κ_4 are summarized in Theorem C.1. $\widehat{\mathcal{M}}_{k_0}$ is define as follows

$$\widehat{\mathcal{M}}_{k_0} = (r, \textit{rank}(\widehat{\Phi}_{C,k_0}), m, A_{k_0}, B_{k_0}, \widehat{\Phi}_{C,k_0}{}^\intercal C, \Sigma_w, \sigma^2_{\eta,k} I).$$

In the above result, we only require $T_{k_0} = \tilde{\mathcal{O}}(\text{poly}\,(r,m)\,/\epsilon^2)$ for an ϵ accurate approximation when $T_{-k_0} = \tilde{\mathcal{O}}(n/\epsilon^2)$. Namely, as long as we have enough metadata, the number of samples needed from single systems are vastly reduced and *is independent of the observer dimension n*. In real-world applications when we have large, i.e. $\tilde{\mathcal{O}}(n/\epsilon^2)$, dataset from a single system, the above result significantly helps for few-shot learning of other similar systems. Also, whenever we have access to numerous, i.e. $\tilde{\mathcal{O}}(n)$ similar systems, the samples required from every system is also independent of n.

6. Simulations

We simulate our algorithm for a set of simple systems where $r = m = 1$, $\Sigma_w = 0$ and $\sigma_\eta = 1$. We use $A = 0.9, B = 1$ and randomly sample C with orthonormal columns. We choose inputs with covariance $\Sigma_u = 0.1I$. The choice of

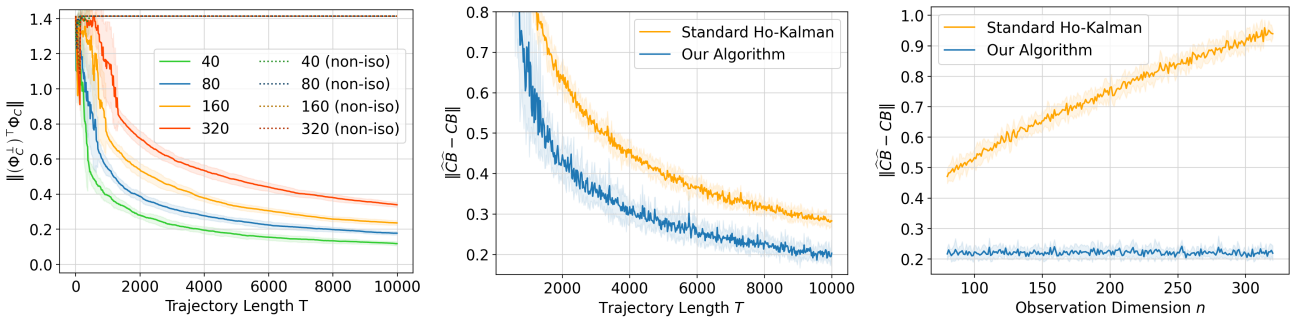

Figure 1. Left: Error for Column Space Approximation (`Col-Approx`, Algorithm 2). Center: Error for `HD-SYSID` (Algorithm 1) and Standard Ho-Kalman with $n = 160$. Right: Error for `Meta-SYSID` (Algorithm 3) and Standard Ho-Kalman for $n \in [80, 320]$.

n will be clear in the context. The results are reported in Figure 5.1.

We first simulate `Col-Approx` (Algorithm 2) with $n = 40, 80, 160, 320$ and a single trajectory data of length $T = 10000$ separately (Section 5.1, left). For isotropic observation noises as in `HD-SYSID`, i.e. $\Sigma_\eta = \sigma_\eta^2 I$, the approximated column space converges to the true column space as is shown by the solid lines. Moreover, the algorithm is simulated for non-isotropic observation noise where the covariance is set to be $\Phi_C \Phi_C{}^\intercal$. As shown in by the dotted curves, the algorithm fail to converge. This is because non-isotropic noise perturbs the eigenspace of the observation covariance Σ_y too much so that the information on Φ_C is drowned.

Next, we simulate Algorithm 1 (Our Algorithm) with $T = 10000$ and $n = 160$ (Section 5.1, center). We plot the error, i.e. $\|\widehat{C}\widehat{B} - CB\|$, during the learning process. As a comparison, error for the standard Ho-Kalman (Oymak & Ozay, 2019) (Naive Algorithm) is also included. It is clear that our Algorithm 3 outperforms the standard Ho-Kalman.

Finally, we consider the `Meta-SYSID` setting for $n \in [80, 320]$ with a meta-dataset from $\lfloor n/40 \rfloor$ randomly sampled systems. Without loss of generality, the error $\|\widehat{C}\widehat{B} - CB\|$ for learning the first system is plotted. For Algorithm 3 (Our Algorithm), it is clear that the learning error doesn't grow as the observation dimension increases. This is because our algorithm utilizes information from the meta-dataset to learn the observer column space, which constitutes the major challenge of system learning. However, the standard Ho-Kalman algorithm learns every system independently, and therefore the error is influenced by the observation dimension.

7. Conclusions and Future Directions

In conclusion, our focus has been on learning a linear time-invariant (LTI) model characterized by low-dimensional latent variables and high-dimensional observations. The introduced Col-SYSID Algorithm serves as a solution with a commendable complexity of $\tilde{\mathcal{O}}(n/\epsilon^2)$. Our analysis also delves into the fundamental limitations of this problem, establishing a sample complexity lower bound that essentially underscores the optimality of our proposed algorithm. Extending the scope of our results, we address a meta-learning setting where datasets from multiple analogous systems are available. This leads to the Meta-SYSID algorithm, an end-to-end framework adept at managing the meta-dataset and effectively learning all included systems.

While our current work lays a solid foundation, future directions could explore extensions to non-linear settings, or investigate adaptive approaches to handle varying or completely unknown observation noise. Additionally, incorporating real-world applications and practical considerations could further enrich the utility of our results.

Impact Statement

The goal of this work is to advance theory in high-dimension machine learning, specifically in latent space space learning. There might be minor potential societal consequences, but none of those are considered as necessary to be specifically highlighted here.

Acknowledgments

This work is supported by NSF AI institute 2112085, NSF ECCS 2328241, and NIH R01LM014465.

References

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://papers.nips.cc/paper_files/paper/2011/hash/e1d5be1c7f2f456670de3d53c7b54f4a-Abstract.

html.

Anderson, B. D. O., Deistler, M., and Lippi, M. Linear System Challenges of Dynamic Factor Models. *Econometrics*, 10(4):35, December 2022. ISSN 2225-1146. doi: 10.3390/econometrics10040035. URL https://www.mdpi.com/2225-1146/10/4/35. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Bakshi, A., Liu, A., Moitra, A., and Yau, M. A New Approach to Learning Linear Dynamical Systems. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, pp. 335–348, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 978-1-4503-9913-5. doi: 10.1145/3564246.3585247. URL https://dl.acm.org/doi/10.1145/3564246.3585247.

Balzano, L., Chi, Y., and Lu, Y. M. Streaming PCA and Subspace Tracking: The Missing Data Case. *Proceedings of the IEEE*, 106(8):1293–1310, August 2018. ISSN 1558-2256. doi: 10.1109/JPROC.2018.2847041. URL https://ieeexplore.ieee.org/document/8417980. Conference Name: Proceedings of the IEEE.

Blanchard, G., Bousquet, O., and Zwald, L. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294, March 2007. ISSN 1573-0565. doi: 10.1007/s10994-006-6895-9. URL https://doi.org/10.1007/s10994-006-6895-9.

Boucheron, S., Lugosi, G., and Massart, a. P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, New York, March 2013. ISBN 978-0-19-953525-5.

Breitung, J. and Eickmeier, S. Dynamic factor models. *Allgemeines Statistisches Archiv*, 90(1):27–42, March 2006. ISSN 1614-0176. doi: 10.1007/s10182-006-0219-z. URL https://doi.org/10.1007/s10182-006-0219-z.

Bui-Thanh, T., Willcox, K., and Ghattas, O. Model Reduction for Large-Scale Systems with High-Dimensional Parametric Input Space. *SIAM Journal on Scientific Computing*, 30(6):3270–3288, January 2008. ISSN 1064-8275. doi: 10.1137/070694855. URL https://epubs.siam.org/doi/abs/10.1137/070694855. Publisher: Society for Industrial and Applied Mathematics.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL https://dl.acm.org/doi/10.1145/1970392.1970395.

Chen, X. and Storey, J. D. Consistent Estimation of Low-Dimensional Latent Structure in High-Dimensional Data, October 2015. URL http://arxiv.org/abs/1510.03497. arXiv:1510.03497 [stat].

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, July 2012. ISSN 1476-4687. doi: 10.1038/nature11129. URL https://www.nature.com/articles/nature11129. Number: 7405 Publisher: Nature Publishing Group.

Davis, C. and Kahan, W. M. The Rotation of Eigenvectors by a Perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. ISSN 0036-1429. URL https://www.jstor.org/stable/2949580. Publisher: Society for Industrial and Applied Mathematics.

Deng, H., Chen, W., Shen, Q., Ma, A. J., Yuen, P. C., and Feng, G. Invariant subspace learning for time series data based on dynamic time warping distance. *Pattern Recognition*, 102:107210, June 2020. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107210. URL https://www.sciencedirect.com/science/article/pii/S0031320320300169.

Djehiche, B. and Mazhar, O. Non asymptotic estimation lower bounds for LTI state space models with Cram\'er-Rao and van Trees, September 2021. URL http://arxiv.org/abs/2109.08582. arXiv:2109.08582 [math, stat].

Djehiche, B. and Mazhar, O. Efficient learning of hidden state LTI state space models of unknown order, February 2022. URL http://arxiv.org/abs/2202.01625. arXiv:2202.01625 [math, stat].

Dong, Y., Qin, S. J., and Boyd, S. P. Extracting a low-dimensional predictable time series. *Optimization and Engineering*, 23(2):1189–1214, June 2022. ISSN 1389-4420, 1573-2924. doi: 10.1007/s11081-021-09643-x. URL https://link.springer.com/10.1007/s11081-021-09643-x.

Fattahi, S. Learning Partially Observed Linear Dynamical Systems from Logarithmic Number of Samples. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, pp. 60–72. PMLR, May 2021. URL https://proceedings.mlr.press/v144/fattahi21a.html. ISSN: 2640-3498.

Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A., and Miller, L. E. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature Neuroscience*, 23(2):260–270, February 2020. ISSN 1546-1726. doi: 10.1038/

s41593-019-0555-4. URL https://www.nature.com/articles/s41593-019-0555-4. Number: 2 Publisher: Nature Publishing Group.

Gribonval, R., Jenatton, R., Bach, F., Kleinsteuber, M., and Seibert, M. Sample Complexity of Dictionary Learning and Other Matrix Factorizations. *IEEE Transactions on Information Theory*, 6(61): 3469–3486, 2015. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2015.2424238. URL https://www.infona.pl//resource/bwmeta1.element.ieee-art-000007088631.

Hajnal, M. A., Tran, D., Einstein, M., Martelo, M. V., Safaryan, K., Polack, P.-O., Golshani, P., and Orbán, G. Continuous multiplexed population representations of task context in the mouse primary visual cortex. *Nature Communications*, 14(1):6687, October 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-42441-w. URL https://www.nature.com/articles/s41467-023-42441-w. Number: 1 Publisher: Nature Publishing Group.

Hallin, M., Nisol, G., and Tavakoli, S. Factor models for high-dimensional functional time series I: Representation results. *Journal of Time Series Analysis*, 44(5-6):578–600, 2023. ISSN 1467-9892. doi: 10.1111/jtsa.12676. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jtsa.12676. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jtsa.12676.

Hespanha, J. P. *Linear Systems Theory: Second Edition*. Princeton University Press, February 2018. ISBN 978-1-4008-9008-8.

Jedra, Y. and Proutiere, A. Finite-Time Identification of Linear Systems: Fundamental Limits and Optimal Algorithms. *IEEE Transactions on Automatic Control*, 68(5):2805–2820, May 2023. ISSN 1558-2523. doi: 10.1109/TAC.2022.3221705. URL https://ieeexplore.ieee.org/document/9946382. Conference Name: IEEE Transactions on Automatic Control.

Lee, H. Improved rates for prediction and identification of partially observed linear dynamical systems, March 2022. URL http://arxiv.org/abs/2011.10006. arXiv:2011.10006 [cs, eess, math, stat].

Li, R.-C. Relative Perturbation Theory: II. Eigenspace and Singular Subspace Variations. *SIAM Journal on Matrix Analysis and Applications*, 20(2):471–492, June 1998. ISSN 0895-4798. doi: 10.1137/S0895479896298506. URL https://epubs.siam.org/doi/abs/10.1137/S0895479896298506. Publisher: Society for Industrial and Applied Mathematics.

Maliar, L. and Maliar, S. Merging simulation and projection approaches to solve high-dimensional problems with an application to a new Keynesian model. *Quantitative Economics*, 6(1):1–47, 2015. ISSN 1759-7331. doi: 10.3982/QE364. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/QE364. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE364.

Mao, Y., Hovakimyan, N., Voulgaris, P., and Sha, L. Finite-Time Model Inference From A Single Noisy Trajectory, January 2021. URL http://arxiv.org/abs/2010.06616. arXiv:2010.06616 [cs, eess].

Marks, T. D. and Goard, M. J. Stimulus-dependent representational drift in primary visual cortex. *Nature Communications*, 12(1):5169, August 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25436-3. URL https://www.nature.com/articles/s41467-021-25436-3. Number: 1 Publisher: Nature Publishing Group.

Masini, R. P., Medeiros, M. C., and Mendes, E. F. Machine learning advances for time series forecasting. *Journal of Economic Surveys*, 37(1):76–111, 2023. ISSN 1467-6419. doi: 10.1111/joes.12429. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/joes.12429. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joes.12429.

Medsker, L. and Jain, L. C. *Recurrent Neural Networks: Design and Applications*. CRC Press, December 1999. ISBN 978-1-4200-4917-6. Google-Books-ID: ME1SAkN0PyMC.

Mudassir, M., Bennbaia, S., Unal, D., and Hammoudeh, M. Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. *Neural Computing and Applications*, July 2020. ISSN 1433-3058. doi: 10.1007/s00521-020-05129-6. URL https://doi.org/10.1007/s00521-020-05129-6.

Oymak, S. and Ozay, N. Non-asymptotic Identification of LTI Systems from a Single Trajectory, February 2019. URL http://arxiv.org/abs/1806.05722. arXiv:1806.05722 [cs, math, stat].

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., Trautmann, E. M., Kaufman, M. T., Ryu, S. I., Hochberg, L. R., Henderson, J. M., Shenoy, K. V., Abbott, L. F., and Sussillo, D. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815, October 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0109-9. URL https://www.nature.com/articles/s41592-018-0109-9. Number: 10 Publisher: Nature Publishing Group.

Poloni, F. and Sbrana, G. Closed-form results for vector moving average models with a univariate estimation approach. *Econometrics and Statistics*, 10:27–52, April 2019. ISSN 2452-3062. doi: 10.1016/j.ecosta.2018.06.003. URL https://www.sciencedirect.com/science/article/pii/S2452306218300327.

Qin, S. J. Latent vector autoregressive modeling and feature analysis of high dimensional and noisy data from dynamic systems. *AIChE Journal*, 68:e17703, June 2022. ISSN 0001-1541. doi: 10.1002/aic.17703. URL https://ui.adsabs.harvard.edu/abs/2022AIChE..68E7703Q. ADS Bibcode: 2022AIChE..68E7703Q.

Rudelson, M. Random Vectors in the Isotropic Position. *Journal of Functional Analysis*, 164(1):60–72, May 1999. ISSN 0022-1236. doi: 10.1006/jfan.1998.3384. URL https://www.sciencedirect.com/science/article/pii/S0022123698933845.

Sarkar, T. and Rakhlin, A. Near optimal finite time identification of arbitrary linear dynamical systems. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5610–5618. PMLR, May 2019. URL https://proceedings.mlr.press/v97/sarkar19a.html. ISSN: 2640-3498.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. Finite time LTI system identification. *The Journal of Machine Learning Research*, 22(1):26:1186–26:1246, January 2021. ISSN 1532-4435.

Saul, L. K., Labs, T., Ave, P., Park, F., and Roweis, S. T. An Introduction to Locally Linear Embedding.

Schutter, B. D. Minimal state-space realization in linear system theory: an overview. *Journal of Computational and Applied Mathematics*, 121(1):331–354, September 2000. ISSN 0377-0427. doi: 10.1016/S0377-0427(00)00341-1. URL https://www.sciencedirect.com/science/article/pii/S0377042700003411.

Shirani Faradonbeh, M. K., Tewari, A., and Michailidis, G. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, October 2018. ISSN 0005-1098. doi: 10.1016/j.automatica.2018.07.008. URL https://www.sciencedirect.com/science/article/pii/S0005109818303546.

Sikander, A. and Prasad, R. Linear time-invariant system reduction using a mixed methods approach. *Applied Mathematical Modelling*, 39(16):4848–4858, August 2015. ISSN 0307-904X. doi: 10.1016/j.apm.2015.04.014. URL https://www.sciencedirect.com/science/article/pii/S0307904X15002504.

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 31st Conference On Learning Theory*, pp. 439–473. PMLR, July 2018. URL https://proceedings.mlr.press/v75/simchowitz18a.html. ISSN: 2640-3498.

Stewart, G. W. G. W. *Matrix perturbation theory*. Computer science and scientific computing. Academic Press, Boston, 1990. ISBN 0-12-670230-6.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature*, 571 (7765):361–365, July 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1346-5.

Sun, S., Li, J., and Mo, Y. Finite Time Performance Analysis of MIMO Systems Identification, October 2023. URL http://arxiv.org/abs/2310.11790. arXiv:2310.11790 [cs, eess].

Sun, Y., Oymak, S., and Fazel, M. Finite Sample Identification of Low-Order LTI Systems via Nuclear Norm Regularization. *IEEE Open Journal of Control Systems*, 1: 237–254, 2022. ISSN 2694-085X. doi: 10.1109/OJCSYS.2022.3200015. URL https://ieeexplore.ieee.org/document/9870857. Conference Name: IEEE Open Journal of Control Systems.

Sussillo, D. and Barak, O. Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation*, 25(3):626–649, March 2013. ISSN 0899-7667. doi: 10.1162/NECO_a_00409. URL https://doi.org/10.1162/NECO_a_00409.

Talebi, S., Taghvaei, A., and Mesbahi, M. Data-driven Optimal Filtering for Linear Systems with Unknown Noise Covariances, October 2023. URL http://arxiv.org/abs/2305.17836. arXiv:2305.17836 [cs, eess, math].

Tian, Y., Zhang, K., Tedrake, R., and Sra, S. Toward Understanding Latent Model Learning in MuZero: A Case Study in Linear Quadratic Gaussian Control. July 2023. URL https://openreview.net/forum?id=r9YZ357Trz.

Tripuraneni, N., Jin, C., and Jordan, M. Provable Meta-Learning of Linear Representations. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10434–10443. PMLR, July 2021. URL https://proceedings.mlr.press/v139/tripuraneni21a.html. ISSN: 2640-3498.

Vaswani, N., Bouwmans, T., Javed, S., and Narayanamurthy, P. Robust Subspace Learning: Robust PCA, Robust Subspace Tracking, and Robust Subspace Recovery. *IEEE Signal Processing Magazine*, 35:32–55, July 2018. ISSN 1053-5888. doi: 10.1109/MSP.2018.2826566. URL https://ui.adsabs.harvard.edu/abs/2018ISPM...35d..32V. ADS Bibcode: 2018ISPM...35d..32V.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://www.cambridge.org/core/books/highdimensional-probability/797C466DA29743D2C8213493BD2D2102.

Xia, J., Marks, T. D., Goard, M. J., and Wessel, R. Stable representation of a naturalistic movie emerges from episodic activity with gain variability. *Nature Communications*, 12(1):5170, August 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25437-2. URL https://www.nature.com/articles/s41467-021-25437-2. Number: 1 Publisher: Nature Publishing Group.

Yu, W., Kim, I. Y., and Mechefske, C. Analysis of different RNN autoencoder variants for time series classification and machine prognostics. *Mechanical Systems and Signal Processing*, 149:107322, February 2021. ISSN 0888-3270. doi: 10.1016/j.ymssp.2020.107322. URL https://www.sciencedirect.com/science/article/pii/S0888327020307081.

Zhang, T. T. C. K., Toso, L. F., Anderson, J., and Matni, N. Meta-Learning Operators to Optimality from Multi-Task Non-IID Data, August 2023. URL http://arxiv.org/abs/2308.04428. arXiv:2308.04428 [cs, eess, stat].

Zheng, Y. and Li, N. Non-Asymptotic Identification of Linear Dynamical Systems Using Multiple Trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, November 2021. ISSN 2475-1456. doi: 10.1109/LCSYS.2020.3042924. URL https://ieeexplore.ieee.org/document/9284539. Conference Name: IEEE Control Systems Letters.

A. Upper Bounds for HD-SYSID — Proof of Theorem 3.4

Recall the setting in HD-SYSID. We consider system $\mathcal{M} = (r, n, m, A, B, C, \Sigma_w, \sigma_\eta^2 I)$ and inputs $\mathcal{U}_1 = \{u_{1,t}\}_{t=0}^{T_1-1}$, $\mathcal{U}_2 = \{u_{2,t}\}_{t=0}^{T_2-1}$ sampled independently from $\mathcal{N}(0, \Sigma_u)$. To simplify future analysis, we define the following notations

$$\begin{aligned}
&\psi_C = \sigma_1(C), \quad \psi_\eta = \sigma_1(\sigma_\eta^2 I), \quad \psi_w = \sigma_1(\Sigma_w + B\Sigma_u B^\intercal), \\
&\phi_C = \sigma_{\min}(C), \quad \phi_O = \sigma_{\min}\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{r-1} \end{bmatrix}\right), \quad \phi_R = \sigma_{\min}\left(\begin{bmatrix} B & AB & \dots & A^{r-1}B \end{bmatrix}\right), \quad \phi_u = \sigma_{\min}(\Sigma_u).
\end{aligned} \tag{4}$$

Here we assume all ψ's satisfy $\psi \geq 1$, otherwise we define ψ to be $\max\{1, \sigma_1(\cdot)\}$. Similarly, we assume all ϕ's satisfy $\phi \leq 1$, otherwise we define ϕ to be $\min\{1, \sigma_{\min}(\cdot)\}$.

Recall that auxiliary system $\widehat{\mathcal{M}}$ is defined as follows with $\widehat{\Phi}_C$ being the approximated observer column space:

$$\begin{aligned}
x_{t+1} &= Ax_t + Bu_t + w_t, \\
y_t &= \widehat{\Phi}_C{}^\intercal C x_t + \widehat{\Phi}_C{}^\intercal \eta_t.
\end{aligned} \tag{5}$$

Now we are ready to restate Theorem 3.4 in full details.

Theorem A.1 (Theorem 3.4 Restated)**.** *Consider $\mathcal{M}$, datasets $\mathcal{D}_1 = \mathcal{U}_1 \cup \mathcal{Y}_1, \mathcal{D}_2 = \mathcal{U}_2 \cup \mathcal{Y}_2$ in HD-SYSID and constants defined above. Suppose system $\mathcal{M}$ satisfies Assumption 3.1 with constants ψ_A and ρ_A. If T_1 and T_2 satisfy*

$$T_1 \gtrsim \kappa_3 \cdot n^2 r^3, \quad T_2 \geq \kappa_1 \cdot \mathit{poly}(r, m), \tag{6}$$

then $(\widehat{A}, \widehat{B}, \widehat{C})$ from Algorithm 1 satisfy the following for some invertible matrix S with probability at least $1 - \delta$

$$\max\left\{\left\|S^{-1}AS - \widehat{A}\right\|, \left\|S^{-1}B - \widehat{B}\right\|, \left\|CS - \widehat{C}\right\|\right\} \lesssim \kappa_4 \cdot \sqrt{\frac{n}{T_1}}\left\|\widehat{C}\right\| + \kappa_2 \cdot \sqrt{\frac{\mathit{poly}(r, m)}{T_2}}. \tag{7}$$

Here $\kappa_1 = \kappa_1(\widehat{\mathcal{M}}, \mathcal{U}_2, \delta)$ and $\kappa_2 = \kappa_2(\widehat{\mathcal{M}}, \mathcal{U}_2, \delta)$ are defined in Definition 3.2. $\kappa_3 = \kappa_3\left(\mathcal{M}, \mathcal{U}_{[2]}, \delta\right)$, $\kappa_4 = \kappa_4\left(\mathcal{M}, \mathcal{U}_1, \delta\right)$ are detailed below. All of them are problem-related constants independent of system dimensions modulo logarithmic factors.

$$\begin{aligned}
\kappa_3\left(\mathcal{M}, \mathcal{U}_{[2]}, \delta\right) &= \max\left\{\kappa_4^2 \frac{\psi_A^2 \psi_C^2}{(1-\rho_A^2)\phi_O^2}, \left(\frac{\psi_\eta^2 \psi_C^2 \psi_w \psi_A^2}{(1-\rho_A^2)\phi_u^2 \phi_C^4 \phi_R^4}\right)^2 \log^2\left(\frac{\psi_C^2 \psi_w \psi_A^4}{1-\rho_A^2} r \log\frac{r}{\delta}\right) \log^4\left(\frac{r}{\delta}\right)\right\}, \\
\kappa_4\left(\mathcal{M}, \mathcal{U}_1, \delta\right) &= \frac{\psi_\eta}{\phi_u \phi_C^2 \phi_R^2}\sqrt{\log\frac{1}{\delta}}.
\end{aligned} \tag{8}$$

Proof. Based on Equation (6), T_1 satisfies the condition of Lemma A.2. We apply Lemma A.2 on $(\mathcal{D}_1, \Sigma_\eta)$ and get the following with probability at least $1 - \frac{\delta}{2}$

$$\left\|\widehat{\Phi}_C^\perp{}^\intercal \Phi_C\right\| \lesssim \kappa_4\sqrt{\frac{n}{T_1}} \coloneqq \Delta_\Phi. \tag{9}$$

The system generating dataset $\tilde{\mathcal{D}}_2$ (line 3 in Algorithm 1), denoted by $\widehat{\mathcal{M}}$, is rewritten as in Equation (5).

Step 1. We first show that $\widehat{\mathcal{M}}$ is still a minimal system. The controllability directly comes from the fact that $R = \begin{bmatrix} B & AB & \dots & A^{r-1}B \end{bmatrix}$ is full row rank, because system $\mathcal{M}$ is minimal. On the other hand, for the observability matrix

O, we know that

$$\begin{aligned}
\text{rank}(O) &= \text{rank}\left(\begin{bmatrix} \widehat{\Phi}_C{}^\intercal C \\ \widehat{\Phi}_C{}^\intercal CA \\ \vdots \\ \widehat{\Phi}_C{}^\intercal CA^{r-1} \end{bmatrix}\right) \\
&\geq \text{rank}\left(\text{diag}\left(\widehat{\Phi}_C, \ldots, \widehat{\Phi}_C\right)\begin{bmatrix} \widehat{\Phi}_C{}^\intercal C \\ \widehat{\Phi}_C{}^\intercal CA \\ \vdots \\ \widehat{\Phi}_C{}^\intercal CA^{r-1} \end{bmatrix}\right) = \text{rank}\left(\begin{bmatrix} \widehat{\Phi}_C\widehat{\Phi}_C{}^\intercal C \\ \widehat{\Phi}_C\widehat{\Phi}_C{}^\intercal CA \\ \vdots \\ \widehat{\Phi}_C\widehat{\Phi}_C{}^\intercal CA^{r-1} \end{bmatrix}\right) \\
&= \text{rank}\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{r-1} \end{bmatrix} - \begin{bmatrix} \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal C \\ \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA \\ \vdots \\ \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA^{r-1} \end{bmatrix}\right).
\end{aligned} \tag{10}$$

Consider the second term. We first observe that the following holds for all $i \in [r-1]$

$$\begin{aligned}
\left\|\widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA^i\right\| &= \left\|\widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal\Phi_C\Phi_C{}^\intercal CA^i\right\| \leq \left\|\widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal\Phi_C\right\| \left\|\Phi_C{}^\intercal CA^i\right\| \\
&= \left\|\widehat{\Phi}_C^\perp{}^\intercal\Phi_C\right\| \left\|CA^i\right\| \\
&\leq \Delta_\Phi\psi_A\psi_C\rho_A^{i-1}.
\end{aligned} \tag{11}$$

Moreover, for $i = 0$, we know

$$\left\|\widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal C\right\| \leq \left\|\widehat{\Phi}_C^\perp{}^\intercal\Phi_C\right\| \|C\| \leq \Delta_\Phi\psi_C. \tag{12}$$

Therefore,

$$\begin{aligned}
\left\|\begin{bmatrix} \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal C \\ \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA \\ \vdots \\ \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA^{r-1} \end{bmatrix}\right\| &= \sqrt{\left\|\sum_{i=0}^{r-1} A^i{}^\intercal C^\intercal\widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA^i\right\|} \\
&\leq \sqrt{\sum_{i=0}^{r-1}\left\|\widehat{\Phi}_C^\perp{}^\intercal CA^i\right\|^2} \\
&\leq \sqrt{(\Delta_\Phi\psi_C)^2 + \sum_{i=1}^{r-1}\left(\Delta_\Phi\psi_A\psi_C\rho_A^{i-1}\right)^2} \\
&\leq 2\frac{\psi_A\psi_C}{\sqrt{1-\rho_A^2}}\Delta_\Phi
\end{aligned} \tag{13}$$

From Equation (6), we know that $T_1 \gtrsim \kappa_4^2 \frac{\psi_A^2 \psi_C^2}{(1-\rho_A^2)\phi_O^2} n$. Combining Theorem 1 in (Stewart, 1990) gives

$$\begin{aligned}
&\sigma_r\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{r-1} \end{bmatrix} - \begin{bmatrix} \widehat{\Phi}_C^{\perp}\widehat{\Phi}_C^{\perp\intercal} C \\ \widehat{\Phi}_C^{\perp}\widehat{\Phi}_C^{\perp\intercal} CA \\ \vdots \\ \widehat{\Phi}_C^{\perp}\widehat{\Phi}_C^{\perp\intercal} CA^{r-1} \end{bmatrix}\right) \\
&\geq \sigma_r\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{r-1} \end{bmatrix}\right) - \left\| \begin{bmatrix} \widehat{\Phi}_C^{\perp}\widehat{\Phi}_C^{\perp\intercal} C \\ \widehat{\Phi}_C^{\perp}\widehat{\Phi}_C^{\perp\intercal} CA \\ \vdots \\ \widehat{\Phi}_C^{\perp}\widehat{\Phi}_C^{\perp\intercal} CA^{r-1} \end{bmatrix} \right\| \\
&\geq \frac{\phi_O}{2} > 0.
\end{aligned} \tag{14}$$

This implies that

$$r \geq \text{rank}(O) \geq r. \tag{15}$$

Namely, the system is observable. Since $\left(A, B, \widehat{\Phi}_C{}^{\intercal} C\right)$ is controllable and observable, we conclude that $\widehat{\mathcal{M}}$ is minimal.

Step 2. We now apply `Sys-Oracle` on this minimal system. Recall the dynamics of $\widehat{\mathcal{M}}$:

$$\begin{aligned}
x_{t+1} &= Ax_t + Bu_t + w_t, \\
y_t &= \widehat{\Phi}_C{}^{\intercal} Cx_t + \widehat{\Phi}_C{}^{\intercal} \eta_t.
\end{aligned} \tag{16}$$

Since $\widehat{\Phi}_C$ is independent of the second trajectory, $\{\widehat{\Phi}_C \eta_{2,t}\}_{t=0}^{T_2}$ are i.i.d. noises independent of other variables of the second trajectory. Moreover, $\widehat{\mathcal{M}}$ satisfy Assumption 3.1. Therefore, we can apply `Sys-Oracle`. Let $r_c = \text{rank}(C)$. With $T_2 \geq \kappa_1 \text{poly}\,(r, r_c, m) = \kappa_1 \text{poly}\,(r, m)$, outputs $\widehat{A}, \widehat{B}, \tilde{C}$ satisfy the following for some invertible matrix S with probability at least $1 - \frac{\delta}{2}$

$$\max\left\{\left\|S^{-1}AS - \widehat{A}\right\|, \left\|S^{-1}B - \widehat{B}\right\|, \left\|\widehat{\Phi}_C{}^{\intercal} CS - \tilde{C}\right\|\right\} \leq \kappa_2 \sqrt{\frac{\text{poly}\,(r, r_c, m)}{T_2}} =: \Delta_{\widehat{\mathcal{M}}}. \tag{17}$$

Therefore, our final approximation for C satisfies

$$\begin{aligned}
\left\|CS - \widehat{C}\right\| &= \left\|CS - \widehat{\Phi}_C\widehat{\Phi}_C{}^{\intercal} CS + \widehat{\Phi}_C\widehat{\Phi}_C{}^{\intercal} CS - \widehat{\Phi}_C\tilde{C}\right\| \\
&\leq \left\|CS - \widehat{\Phi}_C\widehat{\Phi}_C{}^{\intercal} CS\right\| + \left\|\widehat{\Phi}_C\widehat{\Phi}_C{}^{\intercal} CS - \widehat{\Phi}_C\tilde{C}\right\| \\
&\leq \left\|\widehat{\Phi}_C^{\perp}\left(\widehat{\Phi}_C^{\perp}\right)^{\intercal} CS\right\| + \left\|\widehat{\Phi}_C{}^{\intercal} CS - \tilde{C}\right\| \\
&\leq \left\|\widehat{\Phi}_C^{\perp}\left(\widehat{\Phi}_C^{\perp}\right)^{\intercal} CS\right\| + \Delta_{\widehat{\mathcal{M}}}.
\end{aligned} \tag{18}$$

For the first term, we know that

$$\begin{aligned}
\left\|\widehat{\Phi}_C^{\perp}\left(\widehat{\Phi}_C^{\perp}\right)^{\intercal} CS\right\| &= \left\|\widehat{\Phi}_C^{\perp}\left(\widehat{\Phi}_C^{\perp}\right)^{\intercal} \Phi_C\Phi_C{}^{\intercal} CS\right\| = \left\|\left(\widehat{\Phi}_C^{\perp}\right)^{\intercal} \Phi_C\Phi_C{}^{\intercal} CS\right\| \\
&\leq \left\|\left(\widehat{\Phi}_C^{\perp}\right)^{\intercal} \Phi_C\right\| \|\Phi_C{}^{\intercal}\| \|CS\| \\
&\leq \Delta_\Phi \|CS\|.
\end{aligned} \tag{19}$$

Therefore,

$$\left\|CS - \widehat{C}\right\| \leq \Delta_\Phi \|CS\| + \Delta_{\widehat{\mathcal{M}}}. \tag{20}$$

To get the upper bound w.r.t. $\left\|\widehat{C}\right\|$, we notice that

$$\|CS\| \leq \left\|\widehat{C}\right\| + \left\|CS - \widehat{C}\right\| \leq \left\|\widehat{C}\right\| + \Delta_{\widehat{\mathcal{M}}} + \Delta_\Phi \|CS\|. \tag{21}$$

Rearraging the terms gives

$$\|CS\| \leq \frac{\left\|\widehat{C}\right\| + \Delta_{\widehat{\mathcal{M}}}}{1 - \Delta_\Phi}. \tag{22}$$

Substituting back gives the following with probability at least $1 - \delta$

$$\begin{aligned}
\left\|CS - \widehat{C}\right\| &\leq \frac{1}{1-\Delta_\Phi}\Delta_{\widehat{\mathcal{M}}} + \frac{\Delta_\Phi}{1-\Delta_\Phi}\left\|\widehat{C}\right\| \leq 2\Delta_{\widehat{\mathcal{M}}} + \frac{\Delta_\Phi}{1-\Delta_\Phi}\left\|\widehat{C}\right\| \\
&\overset{(i)}{\leq} 2\Delta_{\widehat{\mathcal{M}}} + 2\Delta_\Phi\left\|\widehat{C}\right\| \\
&\lesssim \kappa_2 \cdot \sqrt{\frac{\text{poly}\,(r, r_c, m)}{T_2}} + \kappa_4 \cdot \sqrt{\frac{n}{T_1}}\left\|\widehat{C}\right\| \\
&\leq \kappa_2 \cdot \sqrt{\frac{\text{poly}\,(r, m)}{T_2}} + \kappa_4 \cdot \sqrt{\frac{n}{T_1}}\left\|\widehat{C}\right\|
\end{aligned} \tag{23}$$

Here (i) is because $\Delta_\Phi \leq 1/2$ due to Equation (6). Finally, we conclude that

$$\max\left\{\left\|S^{-1}AS - \widehat{A}\right\|, \left\|S^{-1}B - \widehat{B}\right\|, \left\|CS - \widehat{C}\right\|\right\} \lesssim \kappa_2 \cdot \sqrt{\frac{\text{poly}\,(r, m)}{T_2}} + \kappa_4 \cdot \sqrt{\frac{n}{T_1}}\left\|\widehat{C}\right\|. \tag{24}$$

□

A.1. Upper Bounds for `Col-Approx`

The theoretical guarantee for `Col-Approx` is presented in the following lemma.

Lemma A.2 (Lemma 3.7 Restated)**.** *Consider system $\mathcal{M}$, dataset $\mathcal{D}_1 = \mathcal{U}_1 \cup \mathcal{Y}_1$ in `HD-SYSID` and constants defined at the beginning of Appendix A. Suppose $\mathcal{M}$ satisfies Assumption 3.1 with constants ψ_A and ρ_A. If*

$$T_1 \gtrsim \underbrace{\left(\frac{\psi_\eta^2\psi_C^2\psi_w\psi_A^2}{(1-\rho_A^2)\phi_u^2\phi_C^4\phi_R^4}\right)^2 \log^2\left(\frac{\psi_C^2\psi_w\psi_A^4}{1-\rho_A^2} r\log\frac{r}{\delta}\right)\log^4(\frac{r}{\delta})}_{\kappa_5(\mathcal{M},\mathcal{U}_1,\delta)} \cdot n^2 r^3, \tag{25}$$

then $\widehat{\Phi}_C = \texttt{Col-Approx}(\mathcal{Y}_1, \Sigma_\eta)$ satisfies the following with probability at least $1 - \delta$

$$\widehat{r}_c = \textit{rank}(C), \quad \left\|\widehat{\Phi}_C^{\perp\intercal}\Phi_C\right\| \lesssim \underbrace{\frac{\psi_\eta}{\phi_u\phi_C^2\phi_R^2}\sqrt{\log\frac{1}{\delta}}}_{\kappa_4(\mathcal{M},\mathcal{U}_1,\delta)} \cdot \sqrt{\frac{n}{T_1}}. \tag{26}$$

Proof. For simplicity, we omit all subscript 1 for the rest of this section. From the system dynamics, we know that

$$\begin{aligned}
\Sigma_y &= \sum_{t=0}^{T} y_t y_t^\intercal = \sum_{t=0}^{T} Cx_t x_t^\intercal C^\intercal + \sum_{t=0}^{T}\eta_t\eta_t^\intercal + \sum_{t=0}^{T}(Cx_t\eta_t^\intercal + \eta_t x_t^\intercal C^\intercal) \\
&= \sum_{t=0}^{T} Cx_t x_t^\intercal C^\intercal + \sum_{t=0}^{T}\left(\eta_t\eta_t^\intercal - \sigma_\eta^2 I\right) + \sum_{t=0}^{T}(Cx_t\eta_t^\intercal + \eta_t x_t^\intercal C^\intercal) + (T+1)\sigma_\eta^2 I
\end{aligned} \tag{27}$$

Here the first term is the information on $\text{col}(C)$, while the second and third terms are noises. For the rest of the proof, we first upper bound norms of the noise terms (*step 1*). With this, we show that $\widehat{r}_c = \text{rank}(C)$ with high probability (*step 2*). We then apply our subspace perturbation result to upper bound the influence of the noises on the eigenspace of the first term (*step 3*).

Step 1: Noise Norm Upper Bounds. Define $r_c = \text{rank}(C)$. Notice that we can write $C = \Phi_C \alpha$, where $\Phi_C \in \mathbb{R}^{n \times r_c}$ consists of orthonormal columns that form a basis of $\text{col}(C)$ and $\alpha \in \mathbb{R}^{r_c \times r}$ is a full row rank matrix. It is then clear that

$$\sigma_{\min}(\alpha) = \sigma_{\min}(C) \geq \phi_C, \quad \sigma_1(\alpha) = \sigma_1(C) \leq \psi_C. \tag{28}$$

Let $\Sigma_C = \sum_{t=0}^{T} C x_t x_t^\intercal C^\intercal$, $\bar{\Sigma}_C = \Sigma_C + TI$ and $\Sigma_\alpha = \sum_{t=0}^{T} \alpha x_t x_t^\intercal \alpha^\intercal$. From the definitions, it is clear that $\Sigma_C = \Phi_C \Sigma_\alpha \Phi_C^\intercal$ and $\bar{\Sigma}_C = \Phi_C \Sigma_\alpha \Phi_C^\intercal + TI$. Then from Lemma A.3, Lemma A.4, Lemma A.5, and Lemma A.6, we have the following for $T \gtrsim \frac{\psi_w^2 \psi_C^4 \psi_A^4}{\phi_u^2 \phi_C^4 \phi_R^4} r^3 \log \frac{r}{\delta} \cdot \tilde{U}^2$ (from Equation 25) with probability at least $1 - \delta$

$$\begin{aligned}
&\frac{\phi_u \phi_C^2 \phi_R^2}{8} TI \preceq \Sigma_\alpha \precsim \frac{\psi_C^2 \psi_w \psi_A^2}{1 - \rho_A^2} rT \log \frac{1}{\delta} I, \\
&\left\| \left(\bar{\Sigma}_C\right)^{-\frac{1}{2}} \sum_{t=0}^{T} C x_t \eta_t^\intercal \right\| \lesssim \tilde{U} \sqrt{\psi_\eta n \log \frac{1}{\delta}}, \\
&\left\| \sum_{t=0}^{T} \left(\eta_t \eta_t^\intercal - \sigma_\eta^2 I\right) \right\| \lesssim \psi_\eta \sqrt{nT \log \frac{1}{\delta}}.
\end{aligned} \tag{29}$$

Here $\tilde{U} = \sqrt{\log\left(\frac{\psi_C^2 \psi_w \psi_A^4}{1 - \rho_A^2} r \log \frac{r}{\delta}\right)}$.

Step 2: Order Estimation Guarantee. Consider the following matrix

$$\Sigma_y = \Sigma_C + \underbrace{\sum_{t=0}^{T} \left(\eta_t \eta_t^\intercal - \sigma_\eta^2 I\right) + \sum_{t=0}^{T} \left(C x_t \eta_t^\intercal + \eta_t C^\intercal x_t^\intercal\right)}_{\Delta} + (T+1)\sigma_\eta^2 I. \tag{30}$$

The inequalities of Step 1 imply

$$\begin{aligned}
\|\Delta\| &\leq \left\| \sum_{t=0}^{T} \left(\eta_t \eta_t^\intercal - \sigma_\eta^2 I\right) \right\| + 2 \left\| \sum_{t=0}^{T} C x_t \eta_t^\intercal \right\| \\
&\lesssim \psi_\eta \sqrt{nT \log \frac{1}{\delta}} + \tilde{U} \sqrt{\psi_\eta n \log \frac{1}{\delta}} \left\| \left(\bar{\Sigma}_C\right)^{\frac{1}{2}} \right\| \\
&= \psi_\eta \sqrt{nT \log \frac{1}{\delta}} + \tilde{U} \sqrt{\psi_\eta n \log \frac{1}{\delta}} \left(\left\| \left(TI + \Sigma_\alpha\right)^{\frac{1}{2}} \right\| \right) \\
&\lesssim \psi_\eta \sqrt{nT \log \frac{1}{\delta}} + \tilde{U} \sqrt{\psi_\eta n \log \frac{1}{\delta}} \sqrt{\frac{\psi_C^2 \psi_w \psi_A^2}{1 - \rho_A^2} rT \log \frac{1}{\delta} + T} \\
&\lesssim \tilde{U} \psi_\eta \sqrt{n \log \frac{1}{\delta}} \sqrt{\frac{\psi_C^2 \psi_w \psi_A^2}{1 - \rho_A^2} rT \log \frac{1}{\delta}} \\
&= \frac{\sqrt{\psi_\eta^2 \psi_C^2 \psi_w \psi_A^2}}{\sqrt{1 - \rho_A^2}} \tilde{U} \log\left(\frac{1}{\delta}\right) \sqrt{nrT}
\end{aligned} \tag{31}$$

Therefore, for each $i \in [r_c]$, we have the following for some positive constant c_1:

$$\begin{aligned}\sigma_i\left(\Sigma_y\right) - (T+1)\sigma_\eta^2 &\geq \sigma_i(\Sigma_C) - \|\Delta\| \geq \sigma_{\min}(\Sigma_\alpha) - \|\Delta\| \\ &\geq \frac{\phi_u\phi_C^2\phi_R^2}{8}T - c_1\frac{\sqrt{\psi_\eta^2\psi_C^2\psi_w\psi_A^2}}{\sqrt{1-\rho_A^2}}\tilde{U}\log\left(\frac{1}{\delta}\right)\sqrt{nrT} \\ &\geq \frac{\phi_u\phi_C^2\phi_R^2}{16}T.\end{aligned} \tag{32}$$

Here the first inequality holds according to Theorem 1 in (Stewart, 1990) and the last line is because $T \gtrsim \frac{\psi_\eta^2\psi_C^2\psi_w\psi_A^2}{(1-\rho_A^2)\phi_u^2\phi_C^4\phi_R^4}\log\left(\frac{\psi_C^2\psi_w\psi_A^4}{1-\rho_A^2}r\log\frac{r}{\delta}\right)\log^2(\frac{1}{\delta})\cdot nr$ (from Equation 25). For $i \in [r_c+1, n]$,

$$\sigma_i(\Sigma_y) - (T+1)\sigma_\eta^2 \leq \|\Delta\| \leq c_1\frac{\sqrt{\psi_\eta^2\psi_C^2\psi_w\psi_A^2}}{\sqrt{1-\rho_A^2}}\tilde{U}\log\left(\frac{1}{\delta}\right)\sqrt{nrT}. \tag{33}$$

Based on the above three inequalities and Equation 25, we conclude that with probability at least $1-\delta$, the following hold

$$\begin{aligned}\sigma_i(\Sigma_y) - \sigma_j(\Sigma_y) &\leq 2c_1\frac{\sqrt{\psi_\eta^2\psi_C^2\psi_w\psi_A^2}}{\sqrt{1-\rho_A^2}}\tilde{U}\log\left(\frac{1}{\delta}\right)\sqrt{nrT} \\ &< T^{3/4}, \quad \forall i < j \in [r_c+1, n] \\ \sigma_{r_c}(\Sigma_y) - \sigma_{r_c+1}(\Sigma_y) &\geq \frac{\phi_u\phi_C^2\phi_R^2}{16}T - c_1\frac{\sqrt{\psi_\eta^2\psi_C^2\psi_w\psi_A^2}}{\sqrt{1-\rho_A^2}}\tilde{U}\log\left(\frac{1}{\delta}\right)\sqrt{nrT} \geq \frac{\phi_u\phi_C^2\phi_R^2}{32}T \\ &> T^{3/4}.\end{aligned}$$

The above is because $T \gtrsim \left(\frac{\psi_\eta^2\psi_C^2\psi_w\psi_A^2}{(1-\rho_A^2)\phi_u^2\phi_C^4\phi_R^4}\right)^2\log^2\left(\frac{\psi_C^2\psi_w\psi_A^4}{1-\rho_A^2}r\log\frac{r}{\delta}\right)\log^4(\frac{1}{\delta})\cdot n^2r^2$. Therefore, from the definition of $\widehat{r}_c$, we know that $\widehat{r}_c = r_c$ with probability at least $1-\delta$.

Step 3: Column Space Estimation Guarantee. With $\widehat{r}_c = r_c$, now we try to apply our subspace perturbation result, i.e. Lemma A.7, on matrix $\Sigma_y - \sigma_\eta^2(T+1)I + TI$. Notice that this matrix has exactly the same eigenspace as Σ_y and therefore the eigenspace of its first r_c eigenvectors is $\widehat{\Phi}_C$ (line 9 in Algorithm 2). This matrix can be decomposed as

$$\begin{aligned}\Sigma_y - \sigma_\eta^2(T+1)I + TI &= \Sigma_C + \sum_{t=0}^{T}\left(\eta_t\eta_t^\intercal - \sigma_\eta^2 I\right) + \sum_{t=0}^{T}\left(Cx_t\eta_t^\intercal + \eta_t C^\intercal x_t^\intercal\right) + TI. \\ &= \underbrace{\bar{\Sigma}_C}_{M \text{ in Lemma A.7}} + \underbrace{\sum_{t=0}^{T}\left(\eta_t\eta_t^\intercal - \sigma_\eta^2 I\right)}_{\Delta_2 \text{ in Lemma A.7}} + \underbrace{\sum_{t=0}^{T}\left(Cx_t\eta_t^\intercal + \eta_t C^\intercal x_t^\intercal\right)}_{\Delta_1+\Delta_1^\intercal \text{ in Lemma A.7}}.\end{aligned} \tag{34}$$

For matrix $\bar{\Sigma}_C = \Phi_C\Sigma_\alpha\Phi_C{}^\intercal + TI = \Phi_C\left(\Sigma_\alpha + TI\right)\Phi_C{}^\intercal + T\Phi_C^\perp\Phi_C^{\perp\intercal}$, it is clear that its SVD can be written as

$$\bar{\Sigma}_C = \begin{bmatrix}\Phi_C & \Phi_C^\perp\end{bmatrix}\begin{bmatrix}\Lambda_1 & \\ & TI\end{bmatrix}\begin{bmatrix}\Phi_C{}^\intercal \\ \Phi_C^{\perp\intercal}\end{bmatrix}, \quad \Lambda_1 = \operatorname{diag}\left(\sigma_1(\Sigma_\alpha) + T, \ldots, \sigma_{\min}(\Sigma_\alpha) + T\right). \tag{35}$$

where Φ_C is an orthonormal basis of $\operatorname{col}(C)$. Then we conclude that the following holds for some large enough positive constant c_3

$$\begin{aligned}&\sigma_1\left(\bar{\Sigma}_C\right) \leq T + \sigma_1(\Sigma_\alpha) \leq c_3\left(1 + \frac{\psi_C^2\psi_w\psi_A^2}{1-\rho_A^2}r\log\frac{1}{\delta}\right)T \leq 2c_3\left(\frac{\psi_C^2\psi_w\psi_A^2}{1-\rho_A^2}r\log\frac{1}{\delta}\right)T, \\ &\sigma_{r_c}\left(\bar{\Sigma}_C\right) - \sigma_{r_c+1}(\bar{\Sigma}_C) = \sigma_{\min}(\Sigma_\alpha) \geq \frac{\phi_u\phi_C^2\phi_R^2}{32}T, \quad \sigma_{\min}(\bar{\Sigma}_C) = T.\end{aligned} \tag{36}$$

Now we are ready to apply Lemma A.7 on $\bar{\Sigma}_C$ with the following constants for c_4, c_5 large enough

$$\begin{aligned}
&\alpha = c_4\tilde{U}\sqrt{\psi_\eta n \log\frac{1}{\delta}}, \quad \beta = c_5\psi_\eta\sqrt{nT\log\frac{1}{\delta}},\\
&\delta_M = \frac{\phi_u\phi_C^2\phi_R^2}{32}T, \quad \psi_M = 2c_3\frac{\psi_C^2\psi_w\psi_A^2}{1-\rho_A^2}rT\log\frac{1}{\delta}, \quad \phi_M = T, \quad \sigma_1(\Lambda_2) = T.
\end{aligned} \tag{37}$$

Again, $\tilde{U} = \sqrt{\log\left(\frac{\psi_C^2\psi_w\psi_A^4}{1-\rho_A^2}r\log\frac{r}{\delta}\right)}$. From Equation 25, it is clear that $\sqrt{\phi_M} \gtrsim \alpha$ and $\delta_M \gtrsim \alpha\sqrt{\psi_M} + \beta$. Therefore,

$$\begin{aligned}
\left\|\widehat{\Phi}_C^{\perp\intercal}\Phi_C\right\| = \left\|\widehat{\Phi}_C^{\perp\intercal}\tilde{\Phi}_C\right\| &\lesssim \frac{\psi_\eta}{\phi_u\phi_C^2\phi_R^2}\sqrt{\frac{n}{T}\log\frac{1}{\delta}} + \tilde{U}\frac{\sqrt{\psi_\eta}}{\phi_u\phi_C^2\phi_R^2}\sqrt{\frac{n}{T}\log\frac{1}{\delta}} + \tilde{U}^2\frac{\psi_\eta\sqrt{\psi_C^2\psi_w\psi_A^2}}{\phi_u\phi_C^2\phi_R^2\sqrt{1-\rho_A^2}}\frac{n\sqrt{r}}{T}\log^2\frac{1}{\delta}\\
&\lesssim \frac{\psi_\eta}{\phi_u\phi_C^2\phi_R^2}\sqrt{\frac{n}{T}\log\frac{1}{\delta}} + \tilde{U}^2\frac{\psi_\eta\sqrt{\psi_C^2\psi_w\psi_A^2}}{\phi_u\phi_C^2\phi_R^2\sqrt{1-\rho_A^2}}\frac{n\sqrt{r}}{T}\log^2\frac{1}{\delta}\\
&\lesssim \frac{\psi_\eta}{\phi_u\phi_C^2\phi_R^2}\sqrt{\frac{n}{T}\log\frac{1}{\delta}}.
\end{aligned} \tag{38}$$

□

A.1.1. Supporting Details

Lemma A.3. *Consider the same setting as Lemma A.2. Consider full row rank matrices $\{\alpha \in \mathbb{R}^{a\times r}\}_{k\in[K]}$ for any positive integer $a \le r$. Let*

$$\Sigma_\alpha = \sum_{t\in[T]}\alpha x_t x_t^\intercal\alpha^\intercal, \quad \phi_\alpha = \min\{\sigma_{\min}(\alpha), 1\}, \quad \psi_\alpha = \max\{\sigma_1(\alpha), 1\}. \tag{39}$$

If $T \gtrsim \frac{\psi_w^2\psi_\alpha^4\psi_A^4}{\phi_u^2\phi_\alpha^4\phi_R^4}r^3\log\frac{r}{\delta}\cdot\log\left(\frac{\psi_\alpha^2\psi_w\psi_A^4}{1-\rho_A^2}r\log\frac{r}{\delta}\right)$, then the following holds with probability at least $1-\delta$

$$\Sigma_\alpha \succeq \frac{\phi_u\phi_\alpha^2\phi_R^2}{8}TI. \tag{40}$$

Proof. For further analysis, we let $\tilde{w}_t \coloneqq Bu_t + w_t \sim \mathcal{N}(0, \Sigma_{\tilde{w}})$ with $\Sigma_{\tilde{w}} \coloneqq B\Sigma_u B^\intercal + \Sigma_w$ and we can rewrite the system dynamics as follows

$$\begin{aligned}
x_{t+1} &= Ax_t + Bu_t + w_t = Ax_t + \tilde{w}_t,\\
y_t &= Cx_t + \eta_t.
\end{aligned} \tag{41}$$

For simplicity, we define $\Sigma_{\alpha,\tau} \coloneqq \alpha A^\tau\left(\sum_{t=1}^{T-\tau}x_\tau x_\tau^\intercal\right)(\alpha A^\tau)^\intercal$. It is then clear that

$$\begin{aligned}
\Sigma_\alpha = \Sigma_{\alpha,0} &= \sum_{t=1}^{T}\alpha x_t x_t^\intercal\alpha^\intercal\\
&= \sum_{t=1}^{T}\alpha\left(Ax_{t-1}x_{t-1}^\intercal A^\intercal + \tilde{w}_{t-1}\tilde{w}_{t-1}^\intercal + Ax_{t-1}\tilde{w}_{t-1}^\intercal + \tilde{w}_{t-1}x_{t-1}^\intercal A^\intercal\right)\alpha^\intercal\\
&= \Sigma_{\alpha,1} + \sum_{t=0}^{T-1}\alpha\tilde{w}_t\tilde{w}_t^\intercal\alpha^\intercal + \sum_{t=0}^{T-1}\left(\alpha Ax_t(\alpha\tilde{w}_t)^\intercal + (\alpha\tilde{w}_t)x_t^\intercal A^\intercal\alpha^\intercal\right).
\end{aligned}$$

By Lemma A.4 and A.5, the following events hold with probability at least $1-\delta/r$,

$$\begin{aligned}
\left\|\sum_{t=0}^{T-1}\alpha\tilde{w}_t\tilde{w}_t^\intercal\alpha^\intercal - T\cdot\alpha\Sigma_{\tilde{w}}\alpha^\intercal\right\| &\lesssim \psi_w\psi_\alpha^2\sqrt{rT\log\frac{2r}{\delta}},\\
\left\|(\Sigma_{\alpha,1}+TI)^{-\frac{1}{2}}\sum_{t=0}^{T-1}\alpha Ax_t(\alpha\tilde{w}_t)^\intercal\right\| &\lesssim \tilde{U}\sqrt{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}},
\end{aligned} \tag{42}$$

with $\tilde{U} = \sqrt{\log\left(\frac{\psi_\alpha^2 \psi_w \psi_A^4}{1-\rho_A^2} r \log\frac{2r}{\delta}\right)}$. From the first inequality, it is clear that the following holds for some large enough positive constant c_1

$$\sum_{t=0}^{T-1} \alpha\tilde{w}_t\tilde{w}_t{}^\intercal\alpha^\intercal \succeq T\cdot\alpha\Sigma_{\tilde{w}}\alpha^\intercal - c_1\psi_w\psi_\alpha^2\sqrt{rT\log\frac{2r}{\delta}}I. \tag{43}$$

From the second inequality, we have the following

$$\begin{aligned}
&\left\|(\Sigma_{\alpha,1}+TI)^{-\frac{1}{2}}\sum_{t=0}^{T-1}\alpha A x_t\left(\alpha\tilde{w}_t\right)^\intercal(\Sigma_{\alpha,1}+TI)^{-\frac{1}{2}}\right\| \\
\lesssim\ &\tilde{U}\sqrt{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}\left\|(\Sigma_{\alpha,1}+TI)^{-\frac{1}{2}}\right\| \\
\lesssim\ &\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{T}}.
\end{aligned} \tag{44}$$

This implies that the following holds for large enough positive constant c_2

$$\sum_{t=0}^{T-1}\left(\alpha A x_t\left(\alpha\tilde{w}_t\right)^\intercal + \left(\alpha\tilde{w}_t\right)x_t{}^\intercal A^\intercal\alpha^\intercal\right) \succeq -c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{T}}\left(\Sigma_{\alpha,1}+TI\right). \tag{45}$$

Plugging back into Equation A.1.1 gives the following for some positive constant c_3

$$\begin{aligned}
\Sigma_{\alpha,0} &\succeq \Sigma_{\alpha,1} + T\alpha\Sigma_{\tilde{w}}\alpha^\intercal - c_1\psi_w\psi_\alpha^2\sqrt{rT\log\frac{2r}{\delta}}I - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{T}}\left(\Sigma_{\alpha,1}+TI\right) \\
&\succeq \left(1-c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{T}}\right)\Sigma_{\alpha,1} + T\alpha\Sigma_{\tilde{w}}\alpha^\intercal \\
&\qquad - c_1\psi_w\psi_\alpha^2\sqrt{rT\log\frac{2r}{\delta}}I - c_2\tilde{U}\sqrt{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}TI} \\
&\succeq \left(1-c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{T}}\right)\Sigma_{\alpha,1} + T\alpha\Sigma_{\tilde{w}}\alpha^\intercal - c_3\psi_w\psi_\alpha^2\tilde{U}\sqrt{rT\log\frac{2r}{\delta}}I
\end{aligned} \tag{46}$$

Similarly, we expand $\Sigma_{\alpha,1}, \Sigma_{\alpha,2}, \cdots, \Sigma_{\alpha,r-1}$ and have the following with probability at least $1-\delta$

$$\begin{aligned}
\Sigma_{\alpha,0} &\succeq \left(T\alpha\Sigma_{\tilde{w}}\alpha^\intercal - c_3\psi_w\psi_\alpha^2\tilde{U}\sqrt{rT\log\frac{2r}{\delta}}I\right) + \left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{T}}\right)\Sigma_{\alpha,1} \\
&\succeq \left(T\alpha\Sigma_{\tilde{w}}\alpha^\intercal - c_3\psi_w\psi_\alpha^2\tilde{U}\sqrt{rT\log\frac{2r}{\delta}}I\right) \\
&\quad + \left((T-1)(\alpha A)\Sigma_{\tilde{w}}(\alpha A)^\intercal - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\sqrt{r(T-1)\log\frac{2r}{\delta}}I\right)\cdot\left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{T}}\right) \\
&\quad + \left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{T}}\right)\left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\psi_A^2\log\frac{2r}{\delta}}{T-1}}\right)\Sigma_{\alpha,2} \\
&\succeq \cdots \succeq \left((T-r+1)\sum_{i=0}^{r-1}\left(\alpha A^i\right)\Sigma_{\tilde{w}}\left(\alpha A^i\right)^\intercal - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{rT\log\frac{2r}{\delta}}I\right)\left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\psi_A^2\log\frac{2r}{\delta}}{T-r+1}}\right)^{r-1} \\
&\quad + \left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\psi_A^2\log\frac{2r}{\delta}}{T-r+1}}\right)^{r}\Sigma_{\alpha,r}
\end{aligned}$$

The above inequality is further simplified as follows

$$\begin{aligned}
\Sigma_{\alpha,0} &\overset{(i)}{\succeq} \left(\phi_u(T-r+1)\sum_{i=0}^{r-1}\left(\alpha A^i\right)BB^\intercal\left(\alpha A^i\right)^\intercal - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{rT\log\frac{2r}{\delta}}I\right)\left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\psi_A^2\log\frac{2r}{\delta}}{T-r+1}}\right)^{r} \\
&\overset{(ii)}{\succeq} \left(\phi_u(T-r+1)\alpha\left(\sum_{i=0}^{r-1}A^iB(A^iB)^\intercal\right)\alpha^\intercal - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{rT\log\frac{2r}{\delta}}I\right)\left(1 - \frac{1}{2r}\right)^{r} \\
&\overset{(iii)}{\succeq} \frac{1}{2}\left(\phi_u(T-r+1)\alpha\left(\sum_{i=0}^{r-1}A^iB(A^iB)^\intercal\right)\alpha^\intercal - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{rT\log\frac{2r}{\delta}}I\right) \\
&\succeq \frac{1}{2}\left(\phi_u\phi_\alpha^2\phi_R^2(T-r+1) - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{rT\log\frac{2r}{\delta}}\right)I \\
&\overset{(iv)}{\succeq} \frac{\phi_u\phi_\alpha^2\phi_R^2}{8}TI
\end{aligned}$$

Here (i) is because $\Sigma_{\alpha,r}\succeq 0$ and $\Sigma_{\tilde{w}}\succeq B\Sigma_u B^\intercal\succeq\phi_u BB^\intercal$, (ii) is because $T\gtrsim\psi_w\psi_\alpha^2\psi_A^2 r^3\log\frac{r}{\delta}\cdot\log\left(\frac{\psi_\alpha^2\psi_w\psi_A^4}{1-\rho_A^2}r\log\frac{r}{\delta}\right)$, (iii) is because $(1-\frac{1}{2r})^r\geq\frac{1}{2}$ for all positive integers and (iv) is because $T\gtrsim\frac{\psi_w^2\psi_\alpha^4\psi_A^4}{\phi_u^2\phi_\alpha^4\phi_R^4}r^3\log\frac{r}{\delta}\cdot\log\left(\frac{\psi_\alpha^2\psi_w\psi_A^4}{1-\rho_A^2}r\log\frac{r}{\delta}\right)$. □

Lemma A.4. *Consider the same setting as Lemma A.2. For any positive integer a, let $\{\zeta_t\in\mathbb{R}^a\}_{t=0}^T$ be a sequence of i.i.d Gaussian vectors from $\mathcal{N}(0,\Sigma_\zeta)$ such that ζ_t is independent of x_t. Define $\psi_\zeta=\sigma_1(\Sigma_\zeta)$. We make the following defition for any matrix $P\in\mathbb{R}^{b\times r}$ for positive integer b*

$$\bar{\Sigma}_P=\sum_{t\in[T]}Px_tx_t^\intercal P^\intercal+TI,\quad \psi_P=\sigma_1(P). \tag{47}$$

Then for any δ, the following holds with probability at least $1-\delta$,

$$\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}}\sum_{t\in[T]}Px_t\zeta_t^\intercal\right\|\lesssim\tilde{U}\sqrt{\max\{r,a\}\psi_\zeta\log\frac{1}{\delta}}. \tag{48}$$

Here $\tilde{U}=\sqrt{\log\left(\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2}r\log\frac{1}{\delta}\right)}$.

Proof. From Lemma A.9, we know that the following holds for some vector $v \in \mathcal{S}^{a-1}$

$$\mathbb{P}\left(\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}} \sum_{t\in[T]} P x_t \zeta_t{}^\intercal\right\| > z\right) \leq 5^a \mathbb{P}\left(\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}} \sum_{t\in[T]} P x_t \zeta_t{}^\intercal v\right\| > \frac{z}{2}\right). \tag{49}$$

Notice that $\zeta_t{}^\intercal v$ are independent Gaussian variables from distribution $\mathcal{N}(0, v^\intercal \Sigma_\zeta v)$ for all $t \in [T]$, which is $c_1\sqrt{\psi_\zeta}$-subGaussian for some positive constant c_1. Then applying Theorem 1 in (Abbasi-yadkori et al., 2011) on sequence $\{Px_t\}_{t\in[T]}$ and sequence $\{\zeta_t{}^\intercal v\}_{t\in[T]}$, gives the following inequality

$$\mathbb{P}\left(\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}} \sum_{t\in[T]} P x_t \zeta_t{}^\intercal v\right\| > \sqrt{2c_1^2\psi_\zeta \log\left(\frac{\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(TI)^{-\frac{1}{2}}}{\delta}\right)}\right) \leq \delta. \tag{50}$$

Substituting the above result back gives the following inequality

$$\mathbb{P}\left(\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}} \sum_{t\in[T]} P x_t \zeta_t\right\| > \sqrt{8c_1^2\psi_\zeta \log\left(\frac{\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(TI)^{-\frac{1}{2}}}{\delta}\right)}\right) \leq 5^a\delta, \tag{51}$$

which implies the following inequality holds with probability at least $1-\delta/2$

$$\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}} \sum_{t\in[T]} P x_t \zeta_t\right\| \leq \sqrt{8c_1^2\psi_\zeta \log\left(\frac{2\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(TI)^{-\frac{1}{2}}}{\delta}\right) + 8c_1^2\psi_\zeta a\log 5}. \tag{52}$$

Now consider $\bar{\Sigma}_P = \sum_{t\in[T]} P x_t x_t{}^\intercal P^\intercal + TI$. Then

$$\det(TI) = T^b, \quad \det(\bar{\Sigma}_P) = T^{b-r}\prod_{i=1}^{r}\left(T + \lambda_i\left(\sum_{t\in[T]} P x_t x_t{}^\intercal P^\intercal\right)\right). \tag{53}$$

From Lemma A.6, we have the following with probability at least $1-\delta/2$

$$\left\|\sum_{t\in[T]} x_t x_t{}^\intercal\right\| \lesssim \frac{\psi_w\psi_A^2 r}{1-\rho_A^2} T\log\frac{2}{\delta} \lesssim \frac{\psi_w\psi_A^2 r}{1-\rho_A^2} T\log\frac{1}{\delta}. \tag{54}$$

Therefore,

$$\begin{aligned}\lambda_i\left(\sum_{t\in[T]} P x_t x_t{}^\intercal P^\intercal\right) &\leq \left\|\sum_{t\in[T]} x_t x_t{}^\intercal\right\| \|PP^\intercal\| \\ &\lesssim \frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2} rT\log\frac{1}{\delta}\end{aligned} \tag{55}$$

Substituting back gives the following for some positive constant c_2

$$\begin{aligned}\det(\bar{\Sigma}_P) &= T^{b-r}\prod_{i=1}^{r}\left(T + \lambda_i\left(\sum_{t\in[T]} P x_t x_t{}^\intercal P^\intercal\right)\right) \\ &\leq T^b\left(1 + c_2\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2} r\log\frac{1}{\delta}\right)^r,\end{aligned} \tag{56}$$

which gives

$$\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(TI)^{-\frac{1}{2}} \leq \left(1 + c_2\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2} r\log\frac{1}{\delta}\right)^{\frac{r}{2}}. \tag{57}$$

Finally, with a union bound over above events, we get the following with probability at least $1-\delta$ from Equation 52 and 56

$$\begin{aligned}
&\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}}\sum_{t\in[T]}Px_t\zeta_t\right\| \\
&\leq\sqrt{8c_1^2\psi_\zeta\log\left(\frac{2\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(TI)^{-\frac{1}{2}}}{\delta}\right)+8c_1^2\psi_\zeta a\log 5} \\
&\leq\sqrt{4c_1^2r\psi_\zeta\log\left(2+2c_2\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2}r\log\frac{1}{\delta}\right)+8c_1^2\psi_\zeta\log\frac{1}{\delta}+8c_1^2\psi_\zeta a\log 5} \\
&\lesssim\sqrt{\max\{r,a\}\psi_\zeta\log\frac{1}{\delta}}\cdot\sqrt{\log\left(\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2}r\log\frac{1}{\delta}\right)}.
\end{aligned} \tag{58}$$

This completes the proof. □

Lemma A.5. *For any positive integer a, let $\{w_t\in\mathbb{R}^a\}_{t\in[T]}$ be independent Gaussian vectors from $w_t\sim\mathcal{N}(0,\Sigma_{w,t})$. Let*

$$\psi_w=\max_{t\in[T]}\sigma_1(\Sigma_{w,t}),\quad \phi_w=\min_{t\in[T]}\sigma_{\min}(\Sigma_{w,t}). \tag{59}$$

Then with probability at least $1-\delta$,

$$\left\|\sum_{t=1}^T w_tw_t^\intercal-\sum_{t=1}^T\Sigma_{w,t}\right\|\lesssim\psi_w\sqrt{aT\log\frac{1}{\delta}}. \tag{60}$$

Proof. From Lemma A.9, we know that there exist $x,y\in\mathcal{S}^{a-1}$ s.t.

$$\mathbb{P}\left(\left\|\sum_{t=1}^T w_tw_t^\intercal-\sum_{t=1}^T\Sigma_{w,t}\right\|>z\right)\leq 5^{2a}\mathbb{P}\left(\sum_{t=1}^T\left[(y^\intercal w_t)(x^\intercal w_t)-y^\intercal\Sigma_{w,t}x\right]>\frac{1}{4}z\right) \tag{61}$$

For any x,y, following the definition of w_t, we have

$$\|(y^\intercal w_t)(x^\intercal w_t)-y^\intercal\Sigma_{w,t}x\|_{\psi_1}\lesssim\|(y^\intercal w_t)(x^\intercal w_t)\|_{\psi_1}\leq\|y^\intercal w_t\|_{\psi_2}\|x^\intercal w_t\|_{\psi_2}\lesssim\psi_w. \tag{62}$$

Here the first and second inequalities are from Lemma 2.6.8 and Lemma 2.7.7 in (Vershynin, 2018). Directly applying Bernstein's inequality on $\{(y^\intercal w_t)(x^\intercal w_t)-y^\intercal\Sigma_{w,t}x\}_{t=1}^T$ (subexponential random variables) gives the following for some positive constant c_1 with probability at least $1-\delta$

$$\left\|\sum_{t=1}^T\left[(y^\intercal w_t)(x^\intercal w_t)-y^\intercal\Sigma_{w,t}x\right]\right\|\leq c_1\psi_w\sqrt{T\log(\frac{2}{\delta})}, \tag{63}$$

which is equivalent to

$$5^{2a}\mathbb{P}\left(\left\|\sum_{t=1}^T\left[(y^\intercal w_t)(x^\intercal w_t)-y^\intercal\Sigma_{w,t}x\right]\right\|>c_1\psi_w\sqrt{T\log(\frac{2}{\delta})}\right)\leq 5^{2a}\delta. \tag{64}$$

Letting $\delta'=5^{2a}\delta$ gives

$$5^{2a}\mathbb{P}\left(\left\|\sum_{t=1}^T\left[(y^\intercal w_t)(x^\intercal w_t)-y^\intercal\Sigma_{w,t}x\right]\right\|>c_1\psi_w\sqrt{2aT\log(\frac{10}{\delta'})}\right)\leq\delta'. \tag{65}$$

Substituting back into Equation 61 gives

$$\mathbb{P}\left(\left\|\sum_{t=1}^T w_tw_t^\intercal-\sum_{t=1}^T\Sigma_{w,t}\right\|>4c_1\psi_w\sqrt{2aT\log(\frac{10}{\delta'})}\right)\leq\delta'. \tag{66}$$

This finishes the first part of the proof.

□

Lemma A.6. *Consider the same setting as Lemma A.2. For any δ, with probability at least $1-\delta$,*

$$\left\|\sum_{t=0}^{T} x_t x_t^\intercal\right\| \lesssim \frac{\psi_w \psi_A^2 r}{1-\rho_A^2} T \log\frac{1}{\delta}. \tag{67}$$

Proof. Directly applying Proposition 8.4 in (Sarkar & Rakhlin, 2019) gives the following with probability at least $1-\delta$

$$\left\|\sum_{t=0}^{T} x_t x_t^\intercal\right\| \lesssim \psi_w \mathrm{tr}\left(\sum_{t=0}^{T-1}\Gamma_t(A)\right)\log\frac{1}{\delta} \le \psi_w r\left\|\sum_{t=0}^{T-1}\Gamma_t(A)\right\|\log\frac{1}{\delta}. \tag{68}$$

From Assumption 3.1, we know that

$$\|\Gamma_t(A)\| = \left\|\sum_{\tau=0}^{t} A^\tau A^{\tau\intercal}\right\| \preceq \sum_{\tau=0}^{t}\|A^\tau\|^2 \le 1 + \frac{\psi_A^2}{1-\rho_A^2} \lesssim \frac{\psi_A^2}{1-\rho_A^2}. \tag{69}$$

Substituting back gives

$$\left\|\sum_{t=0}^{T} x_t x_t^\intercal\right\| \lesssim \frac{\psi_w \psi_A^2 r}{1-\rho_A^2} T \log\frac{1}{\delta}. \tag{70}$$

□

A.1.2. Other Lemmas

Lemma A.7. *Given perturbation matrices Δ_1, Δ_2 and p.d square matrix M of the same dimension. Suppose they satisfy the following for some positive constant $\alpha \in [0,1)$ and $\beta \ge 0$*

$$\left\|M^{-\frac{1}{2}}\Delta_1\right\| \le \alpha < \sqrt{\phi_M}, \quad \|\Delta_2\| \le \beta. \tag{71}$$

Denote the SVD of the matrices as follows

$$M = \begin{bmatrix}U_1 & U_2\end{bmatrix}\begin{bmatrix}\Lambda_1 & \\ & \Lambda_2\end{bmatrix}\begin{bmatrix}U_1^\intercal \\ U_2^\intercal\end{bmatrix}, \quad \tilde{M} := M + (\Delta_1 + \Delta_1^\intercal) + \Delta_2 = \begin{bmatrix}\tilde{U}_1 & \tilde{U}_2\end{bmatrix}\begin{bmatrix}\tilde{\Lambda}_1 & \\ & \tilde{\Lambda}_2\end{bmatrix}\begin{bmatrix}\tilde{U}_1^\intercal \\ \tilde{U}_2^\intercal\end{bmatrix}, \tag{72}$$

and define the following constants

$$\psi_M = \sigma_1(M), \quad \phi_M = \sigma_{\min}(M), \quad 0 < \delta_M \le \sigma_{\min}(\Lambda_1) - \sigma_1(\Lambda_2). \tag{73}$$

Then we have

$$\left\|\tilde{U}_2^\intercal U_1\right\| \le \frac{\alpha^2+\beta}{\delta_M - 4\alpha\sqrt{\psi_M} - 3\alpha^2 - \beta} + \frac{\alpha}{\sqrt{\phi_M}}\left(1 + \frac{2-\alpha/\sqrt{\phi_M}}{1-\alpha/\sqrt{\phi_M}}\frac{\sigma_1(\Lambda_2) + 2\alpha\sqrt{\psi_M} + \alpha^2}{\delta_M - 2\alpha\sqrt{\psi_M} - \alpha^2}\right) \tag{74}$$

In the case where $\sqrt{\phi_M} \ge C\alpha$, $\delta_M \ge C\left(\alpha\sqrt{\psi_M} + \beta\right)$ for large enough C, the above result is further simplified to

$$\left\|\tilde{U}_2^\intercal U_1\right\| \lesssim \frac{\beta}{\delta_M} + \frac{\alpha\sigma_1(\Lambda_2)/\sqrt{\phi_M}}{\delta_M} + \frac{\alpha^2\sqrt{\psi_M/\phi_M}}{\delta_M}. \tag{75}$$

Remark A.8. This result is adapted from previous subspace relative perturbation results (Theorem 3.2 in (Li, 1998)). It provides tighter bound for matrices with ϕ_M is close to ψ_M, as compared to the standard Davis-Kahan $\sin\Theta$ theorem. To be more specific, Davis-Kahan gives the bound of order $\frac{\alpha\sqrt{\psi_M}}{\delta_M}$. Here our result features the bound $\frac{\alpha\sigma_1(\Lambda_2)/\sqrt{\phi_M}}{\delta_M}$.

Proof. Define

$$\widehat{M} = M + (\Delta_1 + \Delta_1^\intercal) + \Delta_1^\intercal M^{-1}\Delta_1 = \begin{bmatrix}\widehat{U}_1 & \widehat{U}_2\end{bmatrix}\begin{bmatrix}\widehat{\Lambda}_1 & \\ & \widehat{\Lambda}_2\end{bmatrix}\begin{bmatrix}\widehat{U}_1^\intercal \\ \widehat{U}_2^\intercal\end{bmatrix}. \tag{76}$$

From its definition, we immediately have

$$\begin{aligned}\left\|\widehat{M}-M\right\| &\leq 2\left\|\Delta_1\right\| + \left\|\Delta_1{}^\intercal M^{-1}\Delta_1\right\| \\ &\overset{(i)}{\leq} 2\alpha\sqrt{\psi_M} + \left\|M^{-\frac{1}{2}}\Delta_1\right\|^2 \\ &\leq 2\alpha\sqrt{\psi_M} + \alpha^2.\end{aligned} \tag{77}$$

Here the first term in (i) is because

$$\alpha \geq \left\|M^{-\frac{1}{2}}\Delta_1\right\| \geq \sigma_{\min}\left(M^{-\frac{1}{2}}\right)\left\|\Delta_1\right\| = \frac{1}{\sqrt{\psi_M}}\left\|\Delta_1\right\|. \tag{78}$$

And the second term is because $\|A^\intercal A\| = \|A\|^2$. Then from Theorem 1 in (Stewart, 1990), we know that

$$\sigma_1\left(\widehat{\Lambda}_2\right) \leq \sigma_1\left(\Lambda_2\right) + \left\|\widehat{M}-M\right\| \leq \sigma_1\left(\Lambda_2\right) + 2\alpha\sqrt{\psi_M} + \alpha^2. \tag{79}$$

which in turn gives

$$\widehat{\delta} \coloneqq \sigma_{\min}(\Lambda_1) - \sigma_1(\widehat{\Lambda}_2) \geq \delta_M - 2\alpha\sqrt{\psi_M} - \alpha^2. \tag{80}$$

For the rest of the proof, we upper bound $\left\|\widehat{U}_2{}^\intercal U_1\right\|$, $\left\|\tilde{U}_2{}^\intercal \widehat{U}_1\right\|$ and then derives the desired result. To upper bound $\left\|\widehat{U}_2{}^\intercal U_1\right\|$, we let $D = M^{-1}\Delta_1$ with

$$\begin{aligned}&\|D\| \leq \left\|M^{-\frac{1}{2}}\right\|\left\|M^{-\frac{1}{2}}\Delta_1\right\| \leq \frac{\alpha}{\sqrt{\phi_M}} < 1, \\ &\left\|(I+D)^{-1}\right\| = \frac{1}{\sigma_{\min}\left(I+D\right)} \overset{(i)}{\leq} \frac{1}{1-\sigma_1(D)} \leq \frac{1}{1-\alpha/\sqrt{\phi_M}}.\end{aligned} \tag{81}$$

Here (i) is because $|\sigma_{\min}(I+D) - \sigma_{\min}(I)| \leq \sigma_1(D)$ from Thereom 1 in (Stewart, 1990). And the definition of D gives

$$\widehat{M} = \left(I+\Delta_1{}^\intercal M^{-1}\right) M \left(I+M^{-1}\Delta_1\right) = \left(I+D^\intercal\right) M \left(I+D\right). \tag{82}$$

Now from Equation 80 we apply Theorem 3.2 in (Li, 1998) gives

$$\begin{aligned}\left\|\widehat{U}_2{}^\intercal U_1\right\| &\leq \left\|I-(I+D^\intercal)\right\| + \left(\sigma_1\left(\Lambda_2\right) + 2\alpha\sqrt{\psi_M} + \alpha^2\right)\frac{\left\|(I+D)^\intercal - (I+D)^{-1}\right\|}{\widehat{\delta}} \\ &\leq \frac{\alpha}{\sqrt{\phi_M}} + \frac{\sigma_1\left(\Lambda_2\right) + 2\alpha\sqrt{\psi_M} + \alpha^2}{\widehat{\delta}}\left(\left\|(I+D)^\intercal - I\right\| + \left\|I-(I+D)^{-1}\right\|\right) \\ &\leq \frac{\alpha}{\sqrt{\phi_M}} + \frac{\sigma_1\left(\Lambda_2\right) + 2\alpha\sqrt{\psi_M} + \alpha^2}{\widehat{\delta}}\left(\frac{\alpha}{\sqrt{\phi_M}} + \left\|(I+D)^{-1}\right\|\left\|I+D-I\right\|\right) \\ &\leq \frac{\alpha}{\sqrt{\phi_M}} + \frac{\sigma_1\left(\Lambda_2\right) + 2\alpha\sqrt{\psi_M} + \alpha^2}{\widehat{\delta}}\left(\frac{\alpha}{\sqrt{\phi_M}} + \frac{1}{1-\alpha/\sqrt{\phi_M}}\frac{\alpha}{\sqrt{\phi_M}}\right) \\ &= \frac{\alpha}{\sqrt{\phi_M}}\left(1 + \frac{2-\alpha/\sqrt{\phi_M}}{1-\alpha/\sqrt{\phi_M}}\frac{\sigma_1\left(\Lambda_2\right) + 2\alpha\sqrt{\psi_M} + \alpha^2}{\widehat{\delta}}\right).\end{aligned} \tag{83}$$

We now upper bound $\left\|\tilde{U}_2{}^\intercal \widehat{U}_1\right\|$. We start from upper bounding $\left\|\widehat{M}-\tilde{M}\right\|$

$$\left\|\widehat{M}-\tilde{M}\right\| = \left\|\Delta_1{}^\intercal M^{-1}\Delta_1 - \Delta_2\right\| \leq \left\|\Delta_1{}^\intercal M^{-1}\Delta_1\right\| + \left\|\Delta_2\right\| \leq \alpha^2 + \beta. \tag{84}$$

Moreover, we know that

$$\begin{aligned}
&\sigma_{\min}(\widehat{\Lambda}_1) - \sigma_1(\tilde{\Lambda}_2) \\
&= \sigma_{\min}(\widehat{\Lambda}_1) - \sigma_1(\widehat{\Lambda}_2) + \sigma_1(\widehat{\Lambda}_2) - \sigma_1(\tilde{\Lambda}_2) \\
&\geq \sigma_{\min}(\widehat{\Lambda}_1) - \sigma_1(\widehat{\Lambda}_2) - \alpha^2 - \beta \\
&= \sigma_{\min}(\widehat{\Lambda}_1) - \sigma_{\min}(\Lambda_1) + \sigma_{\min}(\Lambda_1) - \sigma_1(\Lambda_2) + \sigma_1(\Lambda_2) - \sigma_1(\widehat{\Lambda}_2) - \alpha^2 - \beta \\
&\geq -2\left\|\widehat{M} - M\right\| - \alpha^2 - \beta + \delta_M \\
&\geq -4\alpha\sqrt{\psi_M} - 2\alpha^2 - \alpha^2 - \beta + \delta_M.
\end{aligned} \tag{85}$$

Here the first and second inequalities are due to Stewart (1990, Theorem 1) . Directly applying the Davis-Kahan $\sin\theta$ Theorem in (Davis & Kahan, 1970) gives

$$\left\|\tilde{U}_2{}^\intercal \widehat{U}_1\right\| \leq \frac{\left\|\widehat{M} - \tilde{M}\right\|}{\sigma_{\min}(\widehat{\Lambda}_1) - \sigma_1(\tilde{\Lambda}_2)} \leq \frac{\alpha^2 + \beta}{-4\alpha\sqrt{\psi_M} - 3\alpha^2 - \beta + \delta_M}. \tag{86}$$

Finally, we conclude that

$$\begin{aligned}
\left\|\tilde{U}_2{}^\intercal U_1\right\| &= \left\|\tilde{U}_2{}^\intercal \left(\widehat{U}_1\widehat{U}_1{}^\intercal + \widehat{U}_2\widehat{U}_2{}^\intercal\right) U_1\right\| \leq \left\|\tilde{U}_2{}^\intercal \widehat{U}_1\right\| + \left\|\widehat{U}_2{}^\intercal U_1\right\| \\
&\leq \frac{\alpha^2 + \beta}{\delta_M - 4\alpha\sqrt{\psi_M} - 3\alpha^2 - \beta} + \frac{\alpha}{\sqrt{\phi_M}}\left(1 + \frac{2 - \alpha/\sqrt{\phi_M}}{1 - \alpha/\sqrt{\phi_M}} \frac{\sigma_1(\Lambda_2) + 2\alpha\sqrt{\psi_M} + \alpha^2}{\widehat{\delta}}\right) \\
&\leq \frac{\alpha^2 + \beta}{\delta_M - 4\alpha\sqrt{\psi_M} - 3\alpha^2 - \beta} + \frac{\alpha}{\sqrt{\phi_M}}\left(1 + \frac{2 - \alpha/\sqrt{\phi_M}}{1 - \alpha/\sqrt{\phi_M}} \frac{\sigma_1(\Lambda_2) + 2\alpha\sqrt{\psi_M} + \alpha^2}{\delta_M - 2\alpha\sqrt{\psi_M} - \alpha^2}\right).
\end{aligned} \tag{87}$$

□

Lemma A.9. *For any positive integer a, b, let $M \in \mathbb{R}^{a\times b}$ be any random matrix. There exist $x \in \mathcal{S}^{b-1}$ and $y \in \mathcal{S}^{a-1}$ such that for all $\varepsilon < 1$*

$$\mathbb{P}(\|M\| > z) \leq \left(1 + \frac{2}{\varepsilon}\right)^{a+b} \mathbb{P}\left(\|y^\intercal M x\| > (1-\varepsilon)^2 z\right). \tag{88}$$

Proof. From Proposition 8.1 of (Sarkar & Rakhlin, 2019), we know that there exist an $x \in \mathcal{S}^{b-1}$ s.t. the following holds for any $\varepsilon < 1$

$$\mathbb{P}(\|M\| > z) \leq \left(1 + \frac{2}{\varepsilon}\right)^{b} \mathbb{P}\left(\|Mx\| > (1-\varepsilon)z\right) = \left(1 + \frac{2}{\varepsilon}\right)^{b} \mathbb{P}\left(\|x^\intercal M^\intercal\| > (1-\varepsilon)z\right). \tag{89}$$

Applying the proposition again on $x^\intercal M^\intercal$ gives the following for some vector $y \in \mathcal{S}^{a-1}$

$$\mathbb{P}\left(\|x^\intercal M^\intercal\| > (1-\varepsilon)z\right) \leq \left(1 + \frac{2}{\varepsilon}\right)^{a} \mathbb{P}\left(\|x^\intercal M y\| > (1-\varepsilon)^2 z\right). \tag{90}$$

Therefore, we have

$$\mathbb{P}(\|M\| > z) \leq \left(1 + \frac{2}{\varepsilon}\right)^{a+b} \mathbb{P}\left(\|x^\intercal M^\intercal y\| > (1-\varepsilon)^2 z\right) \tag{91}$$

□

B. Lower Bounds for `HD-SYSID` — Proof of Theorem B.2

An essential part for the proof is the Birge's Inequality stated below.

Theorem B.1. *Let $\{\mathbb{P}_i\}_{i\in[N]}$ be probability distributions on $(\Omega, \mathcal{F})$. Let $\{F_i\}_{i\in[N]} \in \mathcal{F}$ be **pairwise disjoint events**. If $\delta = \min_{i\in[N]} \mathbb{P}_i(F_i) \geq \frac{1}{2}$ (the success rate), then*

$$(2\delta - 1)\log\left(\frac{\delta}{1-\delta}(N-1)\right) \leq \frac{1}{N-1}\sum_{i\in[N]} KL(\mathbb{P}_i||\mathbb{P}_1). \tag{92}$$

Proof. From the Birge's inequality (Boucheron et al., 2013), we directly have

$$\delta\log\left(\delta/\frac{1-\delta}{N-1}\right) + (1-\delta)\log\left((1-\delta)/(1-\frac{1-\delta}{N-1})\right) \leq \frac{1}{N-1}\sum_{i\in[N]} \mathrm{KL}(\mathbb{P}_i||\mathbb{P}_1), \tag{93}$$

On the other hand, the following holds for any $N \geq 3$

$$\begin{aligned}
&\delta\log\left(\delta/\frac{1-\delta}{N-1}\right) + (1-\delta)\log\left((1-\delta)/(1-\frac{1-\delta}{N-1})\right) \\
\overset{(i)}{\geq}\ &\delta\log\left(\delta/\frac{1-\delta}{N-1}\right) + (1-\delta)\log\left((1-\delta)/((N-1)\delta)\right) \\
=\ &(2\delta-1)\log\left(\frac{\delta}{1-\delta}(N-1)\right)
\end{aligned} \tag{94}$$

Here (i) holds because $1 - \frac{1-\delta}{N-1} \leq 1 \leq 2\delta \leq (N-1)\delta$. For $N = 2$, the above inequality naturally holds. Combining the two cases finishes the proof. □

Now we prove Theorem 4.1 by proving the following more general version that holds for multiple trajectories.

Theorem B.2. *Suppose $n \geq 2$, $K \geq 1$, and choose positive scalars $\delta \leq \frac{1}{2}$ and $\epsilon \leq 0.6$. Consider the class of minimal systems $\mathcal{M} = (r, n, m, A, B, C, \Sigma_w, \Sigma_\eta)$ with different A,B,C matrices. All other parameters are fixed and known. Moreover, $r < n$ and Σ_η is positive definite. For every $k \in [K]$, let $\mathcal{D}_k = \{y_{k,t}\}_{t=0}^{T_k} \cup \{u_{k,t}\}_{t=0}^{T_k-1} \cup \{$all known parameters$\}$ denote the associated single trajectory dataset. Here the inputs $u_{k,t}$ satisfy: 1). $u_{k,t}$ is sampled independently;2). $\mathbb{E}(u_{k,t}) = 0$. Consider any estimator $\widehat{f}$ mapping the datasets $\mathcal{D} \coloneqq \cup_{k\in[K]}\mathcal{D}_k$ to $(\widehat{A}(\mathcal{D}), \widehat{B}(\mathcal{D}), \widehat{C}(\mathcal{D})) \in \mathbb{R}^{r\times r} \times \mathbb{R}^{r\times m} \times \mathbb{R}^{n\times r}$. If*

$$\mathcal{T} \coloneqq \sum_{k\in[K]} T_k < \frac{\phi_\eta(1-2\delta)\log 1.4}{50\,(\psi_w + \psi_u)} \cdot \frac{n + \log\frac{1}{2\delta}}{\epsilon^2}, \tag{95}$$

there exists a system $\mathcal{M}_0 = (r, n, m, A_0, B_0, C_0, \Sigma_w, \Sigma_\eta)$ with dataset $\mathcal{D}$ such that

$$\mathbb{P}\left\{\left\|\widehat{C}\,(\mathcal{D})\,\widehat{B}\,(\mathcal{D}) - C_0B_0\right\| \geq \epsilon\right\} \geq \delta. \tag{96}$$

Here $\mathbb{P}$ denotes the distribution of $\mathcal{D}$ generated by system $\mathcal{M}_0$. Related constants are defined as follows

$$\begin{aligned}
&\phi_\eta \coloneqq \sigma_{\min}(\Sigma_\eta), \quad \psi_w \coloneqq \sigma_1(\Sigma_w), \\
&\psi_u \coloneqq \max_{k,t} \sigma_{\max}\left(\mathbb{E}(u_{k,t}{u_{k,t}}^\intercal)\right).
\end{aligned}$$

Proof. Fix a constant $\epsilon < 0.1$ and a $\frac{1}{2}$-packing of the intersection of the unit ball in $\mathbb{R}^n$, denoted by $\tilde{\mathcal{P}}$. Now let $\mathcal{P} = \tilde{\mathcal{P}}\backslash\{(-1, 0, \ldots, 0)^\intercal\}$. We know $|\mathcal{P}| \geq 2^n - 1$ and we let $\mathcal{P} = \{p_i\}_{i=1}^{|\mathcal{P}|}$. Now consider the following set of matrices $\{A_i, B_i, C_i\}_{i=1}^{|\mathcal{P}|}$ where

$$\begin{aligned}
A_i &= \epsilon\left(E_{r,1} + \sum_{j\in[r-1]} E_{j,j+1}\right) \in \mathbb{R}^{r\times r}, \quad B_i = E_{1,1} \in \mathbb{R}^{r\times m}, \\
C_i &= 4\epsilon\left(p_i{e_1}^\intercal + \sum_{j\in[r]} E_{j,j}\right) \in \mathbb{R}^{n\times r}.
\end{aligned}$$

Step 1: We first prove that all systems are both observable and controllable, and therefore a minimal realizations. Firstly, notice that $\text{rank}(C_i) = r$. Therefore the observability matrix is full column rank, implying that the system is observable. For controllability, we first show by induction that for any interger a satisfying $1 \leq a < r$, the following holds

$$A_i^a = \epsilon^a \left(\sum_{j=1}^{a} E_{r-a+j,j} + \sum_{j=1}^{r-a} E_{j,j+a} \right). \tag{97}$$

The base case where $a = 1$ is easily verified from the definition of A_i. Now suppose the induction hypothesis holds for all positive integers $a \leq a_0$ for integer $a_0 \in [1, r-2]$. Then for $a = a_0 + 1 < r$, we anow that

$$\begin{aligned}
A_i^{a_0+1} &= A_i^{a_0} A_i = \epsilon^{a+1} \left(\sum_{j_0=1}^{a_0} E_{r-a_0+j_0,j_0} + \sum_{j_0=1}^{r-a_0} E_{j_0,j_0+a_0} \right) \left(E_{r,1} + \sum_{j\in[r-1]} E_{j,j+1} \right) \\
&= \epsilon^{a+1} \left(\sum_{j_0=1}^{a_0} E_{r-a_0+j_0,j_0} E_{r,1} + \sum_{j_0=1}^{a_0} E_{r-a_0+j_0,j_0} \sum_{j\in[r-1]} E_{j,j+1} \right. \\
&\qquad \left. + \sum_{j_0=1}^{r-a_0} E_{j_0,j_0+a_0} E_{r,1} + \sum_{j_0=1}^{r-a_0} E_{j_0,j_0+a_0} \sum_{j\in[r-1]} E_{j,j+1} \right) \\
&\overset{(i)}{=} \epsilon^{a+1} \left(0 + \sum_{j_0=1}^{a_0} E_{r-a_0+j_0,j_0+1} + E_{r-a_0,1} + \sum_{j_0=1}^{r-a_0} E_{j_0,j_0+a_0+1} \right) \\
&= \epsilon^{a+1} \left(\sum_{j_0=1}^{a_0} E_{r-(a_0+1)+(j_0+1),j_0+1} + E_{r-(a_0+1)+1,1} + \sum_{j_0=1}^{r-(a_0+1)} E_{j_0,j_0+a_0+1} \right) \\
&= \epsilon^{a+1} \left(\sum_{j_0=1}^{a_0+1} E_{r-(a_0+1)+j_0,j_0} + \sum_{j_0=1}^{r-(a_0+1)} E_{j_0,j_0+a_0+1} \right).
\end{aligned} \tag{98}$$

Here, again, (i) is because $E_{a_1,b_1} E_{a_2,b_2} = E_{a_1,b_2}$ if $b_1 = a_2$. Otherwise the product is zero. This finishes the induction. Now with A_i^a for positive integer $a < r$, we directly have

$$\begin{aligned}
(A_i^a B_i)_1 &= \epsilon^a \left(\left(\sum_{j=1}^{a} E_{r-a+j,j} + \sum_{j=1}^{r-a} E_{j,j+a} \right) \sum_{j\in\min\{r,m\}} E_{j,j} \right)_1 \\
&= \epsilon^a \left(\left(\sum_{j=1}^{a} E_{r-a+j,j} + \sum_{j=1}^{r-a} E_{j,j+a} \right) E_{1,1} \right)_1 \\
&= \epsilon^a E_{r-a+1,1}.
\end{aligned} \tag{99}$$

Here for any matrix M, we use $(M)_1$ to denote its first column. Therefore, we have

$$\begin{aligned}
&\text{rank}\left(\begin{bmatrix} B & AB & \cdots & A^{r-1}B \end{bmatrix} \right) \\
\geq\ &\text{rank}\left(\begin{bmatrix} (B)_1 & (AB)_1 & \cdots & (A^{r-1}B)_1 \end{bmatrix} \right) \\
=\ &\text{rank}\left(\begin{bmatrix} e_1 & \epsilon e_r & \epsilon^2 e_{r-1} \cdots & \epsilon^{r-1} e_2 \end{bmatrix} \right) \\
=\ &r.
\end{aligned} \tag{100}$$

Therefore, the controllability matrix is full row rank, implying that the system is controllable. Therefore, for all i, the matrices $\{A_i, B_i, C_i\}$ represents a minimal system.

Step 2: Now we show learning at least one of those minimal realizations is hard. Consider any estimator $\widehat{f}$ mapping $\mathcal{D}$ to $\left(\widehat{A}(\mathcal{D}), \widehat{B}(\mathcal{D}), \widehat{C}(\mathcal{D}) \right) \in \mathbb{R}^{r\times r} \times \mathbb{R}^{r\times m} \times \mathbb{R}^{n\times r}$.

We now define events $\{F_i\}_{i\in[|\mathcal{P}|]}$ as follows:

$$F_i = \left\{\mathcal{D} : \left\|\widehat{C}\left(\mathcal{D}\right)\widehat{B}\left(\mathcal{D}\right) - C_iB_i\right\| < \epsilon\right\} \tag{101}$$

Under our construction, for any $i, j \in [|\mathcal{P}|]$, notice that

$$C_iB_i = 4\epsilon\left(p_ie_1{}^\intercal + \sum_{j\in[r]}E_{j,j}\right)E_{1,1} = 4\epsilon p_ie_1{}^\intercal + 4\epsilon E_{1,1}. \tag{102}$$

Therefore, it is clear that

$$\|C_iB_i - C_jB_j\| = 4\epsilon\left\|(p_i - p_j)e_1{}^\intercal\right\| = 4\epsilon\left\|p_i - p_j\right\|_2 \geq 2\epsilon. \tag{103}$$

Therefore,

$$\begin{aligned}&\left\|\widehat{C}\left(\mathcal{D}\right)\widehat{B}\left(\mathcal{D}\right) - C_iB_i\right\| + \left\|\widehat{C}\left(\mathcal{D}\right)\widehat{B}\left(\mathcal{D}\right) - C_jB_j\right\|\\ \geq\ &\|C_iB_i - C_jB_j\|\\ \geq\ &2\epsilon.\end{aligned} \tag{104}$$

Namely, all events $\{F_i\}_{i\in[|\mathcal{P}|]}$ are pairwise disjoint.

For any $k \in [K]$, let u^k, x^k, y^k denote the corresponding trajectories of the inputs, latent variables and observations. Moreover, let $u = \cup_{k\in[K]}u^k$, $y = \cup_{k\in[K]}y^k$ and $x = \cup_{k\in[K]}x^k$ denote the collection of all trajectories. For any $i \in [|\mathcal{P}|]$, let $\mathbb{P}_i^{UXY}$ denote the joint distribution of $\{u, x, y\}$ generated by system $\mathcal{M}_i = (r, n, m, A_i, B_i, C_i, \Sigma_w, \Sigma_\eta)$. Let $\mathbb{P}_i^{UY}$, $\mathbb{P}_i^{UX}$ be the marginal distribution of $\{y, u\}$ and $\{x, u\}$. Similarly, define $\mathbb{P}_i^{X|U}$, $\mathbb{P}_i^{X|UY}$ and $\mathbb{P}_i^{Y|UX}$ to be the conditional distributions. For the rest of the proof, we apply Theorem B.1 on the events $\{F_i\}_{i\in[|\mathcal{P}|]}$ and distributions $\{\mathbb{P}_i^{UY}\}_{i\in[|\mathcal{P}|]}$.

We first upper bound the KL-divergence between distributions $\{\mathbb{P}_i^{UY}\}_{i=1}^{|\mathcal{P}|}$. To do this, we start from showing that $\mathrm{KL}(\mathbb{P}_i^{UY}||\mathbb{P}_j^{UY}) \leq \mathrm{KL}(\mathbb{P}_i^{UXY}||\mathbb{P}_j^{UXY})$.

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_i^{UXY}||\mathbb{P}_j^{UXY}) =\ &\int\mathbb{P}_i^{UXY}(u,x,y)\log\frac{\mathbb{P}_i^{UXY}(u,x,y)}{\mathbb{P}_j^{UXY}(u,x,y)}\mathrm{d}u\mathrm{d}x\mathrm{d}y\\ =\ &\int\mathbb{P}_i^{UY}(u,y)\mathbb{P}_i^{X|U=u,Y=y}(x)\log\frac{\mathbb{P}_i^{UY}(u,y)\mathbb{P}_i^{X|U=u,Y=y}(x)}{\mathbb{P}_j^{UY}(u,y)\mathbb{P}_j^{X|U=u,Y=y}(x)}\mathrm{d}u\mathrm{d}x\mathrm{d}y\\ =\ &\int\mathbb{P}_i^{UY}(u,y)\mathbb{P}_i^{X|U=u,Y=y}(x)\log\frac{\mathbb{P}_i^{UY}(u,y)}{\mathbb{P}_j^{UY}(u,y)}\mathrm{d}u\mathrm{d}x\mathrm{d}y\\ &+\int\mathbb{P}_i^{UY}(u,y)\mathbb{P}_i^{X|U=u,Y=y}(x)\log\frac{\mathbb{P}_i^{X|U=u,Y=y}(x)}{\mathbb{P}_j^{X|U=u,Y=y}(x)}\mathrm{d}u\mathrm{d}x\mathrm{d}y\\ =\ &\int\mathbb{P}_i^{UY}(u,y)\left(\int\mathbb{P}_i^{X|U=u,Y=y}(x)\mathrm{d}x\right)\log\frac{\mathbb{P}_i^{UY}(u,y)}{\mathbb{P}_j^{UY}(u,y)}\mathrm{d}u\mathrm{d}y\\ &+\int\mathbb{P}_i^{UY}(u,y)\mathrm{KL}\left(\mathbb{P}_i^{X|U=u,Y=y}||\mathbb{P}_j^{X|U=u,Y=y}\right)\mathrm{d}u\mathrm{d}y\\ \overset{(i)}{\geq}\ &\int\mathbb{P}_i^{UY}(u,y)\log\frac{\mathbb{P}_i^{UY}(u,y)}{\mathbb{P}_j^{UY}(u,y)}\mathrm{d}u\mathrm{d}y\\ =\ &\mathrm{KL}(\mathbb{P}_i^{UY}||\mathbb{P}_j^{UY}).\end{aligned} \tag{105}$$

Here (i) holds because KL-divergence (the second term) is non-negative. Therefore, we only need to upper bound $\mathrm{KL}(\mathbb{P}_i^{UXY}||\mathbb{P}_j^{UXY})$. Notice that

$$
\begin{aligned}
\log \frac{\mathbb{P}_i^{UXY}(u,x,y)}{\mathbb{P}_j^{UXY}(u,x,y)} &= \log \frac{\mathbb{P}_i^{U}(u)\mathbb{P}_i^{X|U=u}(x)\mathbb{P}_i^{Y|U=u,X=x}(y)}{\mathbb{P}_j^{U}(u)\mathbb{P}_j^{X|U=u}(x)\mathbb{P}_j^{Y|U=u,X=x}(y)} \\
= \log \frac{\prod_{k=1,t=0}^{K,T_k-1}\mathbb{P}_i(u_{k,t})}{\prod_{k=1,t=0}^{K,T_k-1}\mathbb{P}_j(u_{k,t})} &+ \log \frac{\prod_{k=1,t=1}^{K,T_k}\mathbb{P}_i(x_{k,t}|x_{k,0:t-1},u_{k,0:t-1})}{\prod_{k=1,t=1}^{K,T_k}\mathbb{P}_j(x_{k,t}|x_{k,0:t-1},u_{k,0:t-1})} \\
&+ \log \frac{\prod_{k=1,t=0}^{K,T_k}\mathbb{P}_i(y_{k,t}|x_{k,0:t},u_{k,0:t})}{\prod_{k=1,t=0}^{K,T_k}\mathbb{P}_j(y_{k,t}|x_{k,0:t},u_{k,0:t})} \\
= \sum_{k=1,t=0}^{K,T_k-1}\log \frac{\mathbb{P}_i(u_{k,t})}{\mathbb{P}_j(u_{k,t})} &+ \sum_{k=1,t=1}^{K,T_k}\log \frac{\mathbb{P}_i(x_{k,t}|x_{k,t-1},u_{k,t-1})}{\mathbb{P}_j(x_{k,t}|x_{k,t-1},u_{k,t-1})} \\
&+ \sum_{k=1,t=0}^{K,T_k}\log \frac{\mathbb{P}_i(y_{k,t}|x_{k,t},u_{k,t})}{\mathbb{P}_j(y_{k,t}|x_{k,t},u_{k,t})} \\
= \sum_{k=1,t=0}^{K,T_k}\log &\frac{\mathbb{P}_i(y_{k,t}|x_{k,t},u_{k,t})}{\mathbb{P}_j(y_{k,t}|x_{k,t},u_{k,t})}.
\end{aligned} \tag{106}
$$

Here the last line is because the distribution of u and x doesn't depend on the i or j. This is holds for x because for any i, A_i and B_i take the same value. Therefore,

$$
\begin{aligned}
&\mathrm{KL}(\mathbb{P}_i^{UXY}||\mathbb{P}_j^{UXY}) \\
&= \int \mathbb{P}_i^{UXY}(u,x,y)\log\frac{\mathbb{P}_i^{UXY}(u,x,y)}{\mathbb{P}_j^{UXY}(u,x,y)}\mathrm{d}x\mathrm{d}y \\
&= \int \mathbb{P}_i^{UXY}(u,x,y)\left(\sum_{k=1,t=0}^{K,T_k}\log\frac{\mathbb{P}_i(y_{k,t}|x_{k,t},u_{k,t})}{\mathbb{P}_j(y_{k,t}|x_{k,t},u_{k,t})}\right)\mathrm{d}u\mathrm{d}x\mathrm{d}y \\
&= \sum_{k=1,t=0}^{K,T_k}\mathbb{E}_i\left(\log\frac{\mathbb{P}_i(y_{k,t}|x_{k,t},u_{k,t})}{\mathbb{P}_j(y_{k,t}|x_{k,t},u_{k,t})}\right) \\
&\overset{(i)}{=} \sum_{k=1,t=0}^{K,T_k}\mathbb{E}_i\left(\mathbf{KL}\left(\mathbb{P}_i^{Y}(\cdot|x_{k,t},u_{k,t})||\mathbb{P}_j^{Y}(\cdot|x_{k,t},u_{k,t})\right)\right) \\
&\overset{(ii)}{=} \frac{1}{2}\sum_{k=1,t=0}^{K,T_k}\mathbb{E}_i\left(x_{k,t}{}^{\intercal}(C_i-C_j)^{\intercal}\Sigma_\eta^{-1}(C_i-C_j)x_{k,t}\right)
\end{aligned} \tag{107}
$$

Here in (i) follows from the definition of KL-divergence and by taking the expectation of x_t (or y_t in the second term). And (ii) follows from the fact that $P_i^X = P_j^X$ (this is because $A_i = A_j$ and $B_i = B_j$) and from the KL-divergence between two gaussians. We further simplify the expression as follows

$$
\begin{aligned}
\mathrm{KL}(\mathbb{P}_i^{UXY}||\mathbb{P}_j^{UXY}) &= \frac{1}{2}\sum_{k=1,t=0}^{K,T_k}\mathbb{E}_i\left(x_{k,t}{}^{\intercal}(C_i-C_j)^{\intercal}\Sigma_\eta^{-1}(C_i-C_j)x_{k,t}\right) \\
&\le \frac{1}{2\phi_\eta}\sum_{k=1,t=0}^{K,T_k}\mathbb{E}_i\left(x_{k,t}{}^{\intercal}(C_i-C_j)^{\intercal}\Sigma_\eta^{-1}(C_i-C_j)x_{k,t}\right) \\
&\le \frac{1}{2\phi_\eta}\sum_{k=1,t=0}^{K,T_k}\mathbb{E}_i\left(x_{k,t}{}^{\intercal}(C_i-C_j)^{\intercal}(C_i-C_j)x_{k,t}\right) \\
&= \frac{1}{2\phi_\eta}\sum_{k=1,t=0}^{K,T_k}\mathrm{tr}\left((C_i-C_j)\mathbb{E}_i\left(x_{k,t}x_{k,t}{}^{\intercal}\right)(C_i-C_j)^{\intercal}\right)
\end{aligned} \tag{108}
$$

Now we consider term $\mathbb{E}_i\left(x_{k,t}x_{k,t}{}^\intercal\right)$. Notice that $x_{k,t}=\sum_{\tau=0}^{t-1}A_i^\tau\left(B_iu_{k,t-1-\tau}+w_{k,t-1-\tau}\right)$, we have

$$\begin{aligned}
\mathbb{E}_i(x_{k,t}x_{k,t}{}^\intercal) &= \mathbb{E}_i\left[\sum_{\tau=0}^{t-1}A_i^\tau\left(B_iu_{k,t-1-\tau}+w_{k,t-1-\tau}\right)\sum_{\tau=0}^{t-1}\left(B_iu_{k,t-1-\tau}+w_{k,t-1-\tau}\right)^\intercal(A_i^\tau)^\intercal\right]\\
&= \sum_{\tau_1,\tau_2=0}^{t-1}A_i^{\tau_1}\mathbb{E}_i\left[\left(B_iu_{k,t-1-\tau_1}+w_{k,t-1-\tau_1}\right)\left(B_iu_{k,t-1-\tau_2}+w_{k,t-1-\tau_2}\right)^\intercal\right](A_i^{\tau_2})^\intercal\\
&\overset{(i)}{=} \sum_{\tau=0}^{t-1}A_i^{\tau}\mathbb{E}_i\left[\left(B_iu_{k,t-1-\tau}+w_{k,t-1-\tau}\right)\left(B_iu_{k,t-1-\tau}+w_{k,t-1-\tau}\right)^\intercal\right](A_i^{\tau})^\intercal\\
&= \sum_{\tau=0}^{t-1}A_i^\tau\left(B_i\mathbb{E}(u_{k,t}u_{k,t}{}^\intercal)B_i{}^\intercal+\Sigma_w\right)(A_i^\tau)^\intercal\\
&\preceq \left(\psi_w+\|B\|^2\psi_u\right)\Gamma_{t-1}(A)\\
&= \left(\psi_w+\psi_u\right)\Gamma_{t-1}(A).
\end{aligned}\tag{109}$$

Here (i) is because u_{τ_1},w_{τ_1} are zero-mean and independent of u_{τ_2},w_{τ_2} for $\forall\tau_1\neq\tau_2$. And the last line is because $B_i=E_{1,1}$. Substituting back gives

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}_i^{UXY}||\mathbb{P}_j^{UXY})\\
\leq\ &\frac{1}{2\phi_\eta}\sum_{k=1,t=0}^{K,T_k}\mathrm{tr}\left((C_i-C_j)\mathbb{E}_i\left(x_{k,t}x_{k,t}{}^\intercal\right)(C_i-C_j)^\intercal\right)\\
\leq\ &\frac{\psi_w+\psi_u}{2\phi_\eta}\sum_{k=1,t=1}^{K,T_k}\mathrm{tr}\left((C_i-C_j)\Gamma_{t-1}(A)(C_i-C_j)^\intercal\right).
\end{aligned}\tag{110}$$

Here the last line is because $x_0=0$. Since

$$\|A\|=\epsilon\left\|E_{r,1}+\sum_{j\in[r-1]}E_{j,j+1}\right\|=\epsilon.\tag{111}$$

This is because $E_{r,1}+\sum_{j\in[r-1]}E_{j,j+1}$ is just the identity matrix after some column permutation. We conclude that

$$\Gamma_t(A_i)=\sum_{\tau=0}^{t}A_i^t(A_i^t)^\intercal\preceq\sum_{\tau=0}^{t}\epsilon^{2t}I\preceq\frac{1}{1-\epsilon^2}I.\tag{112}$$

Substituting back gives

$$\mathrm{KL}(\mathbb{P}_i^{UXY}||\mathbb{P}_j^{UXY})\leq\frac{\psi_w+\psi_u}{2(1-\epsilon^2)\phi_\eta}\sum_{k=1,t=0}^{K,T_k-1}\mathrm{tr}\left((C_i-C_j)(C_i-C_j)^\intercal\right)\tag{113}$$

From the definitions of C_i and C_j, we notice that

$$C_i-C_j=4\epsilon(p_i-p_j)e_1{}^\intercal.\tag{114}$$

Then

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_i^{UXY}||\mathbb{P}_j^{UXY}) &\leq \frac{8\left(\psi_w+\psi_u\right)\epsilon^2}{(1-\epsilon^2)\phi_\eta}\sum_{k=1,t=0}^{K,T_k-1}\mathrm{tr}\left(\left(p_ie_1{}^\intercal-p_je_1{}^\intercal\right)\left(p_ie_1{}^\intercal-p_je_1{}^\intercal\right)^\intercal\right)\\
&= \frac{8\left(\psi_w+\psi_u\right)\epsilon^2}{(1-\epsilon^2)\phi_\eta}\sum_{k=1,t=0}^{K,T_k-1}\|p_ie_1{}^\intercal-p_je_1{}^\intercal\|_{\mathrm{F}}^2\\
&= \frac{8\left(\psi_w+\psi_u\right)\epsilon^2}{(1-\epsilon^2)\phi_\eta}\sum_{k=1,t=0}^{K,T_k-1}\|p_i-p_j\|_2^2
\end{aligned}\tag{115}$$

Since p_i and p_j lie in the unit ball, we know that $\|p_i - p_j\| \leq 2$. Therefore,

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_i^{UXY} || \mathbb{P}_j^{UXY}) &\leq \frac{32\left(\psi_w + \psi_u\right)\epsilon^2}{(1-\epsilon^2)\phi_\eta} \sum_{k=1,t=0}^{K,T_k-1} 1 = \frac{32\left(\psi_w + \psi_u\right)\epsilon^2}{(1-\epsilon^2)\phi_\eta} \sum_{k=1}^{K} T_k \\ &= \frac{32\left(\psi_w + \psi_u\right)\epsilon^2}{(1-\epsilon^2)\phi_\eta} \mathcal{T}.\end{aligned} \tag{116}$$

Here the last line is due to the definition of $\mathcal{T}$.

Now applying Theorem B.1 on distributions $\{\mathbb{P}_i^{UY}\}_{i\in[|\mathcal{P}|]}$ and events $\{F_i\}_{i\in[|\mathcal{P}|]}$ gives the following result. If $\min_{i\in[|\mathcal{P}|]} \mathbb{P}_i^{UY}(F_i) \geq 1-\delta$, then the following holds

$$(1-2\delta)\log\left(\frac{1-\delta}{\delta}(|\mathcal{P}|-1)\right) \leq \frac{1}{N-1}\sum_{i\in[|\mathcal{P}|]} \mathrm{KL}(\mathbb{P}_i^{UY} || \mathbb{P}_1^{UY}) \leq \sup_{i\in[\mathcal{P}]} \mathrm{KL}(\mathbb{P}_i^{UY} || \mathbb{P}_1^{UY}). \tag{117}$$

For the LHS, since $|\mathcal{P}| \geq 2^n - 1$, we have $|\mathcal{P}| - 2 \geq 1.4^n$. Therefore,

$$\begin{aligned}\text{LHS} &= (1-2\delta)\log\left(\frac{1-\delta}{\delta}(|\mathcal{P}|-1)\right) \\ &\geq (1-2\delta)\log\left(\frac{1-\delta}{\delta}1.4^n\right) \\ &\geq (1-2\delta)n\log 1.4 + (1-2\delta)\log\left(\frac{1-\delta}{\delta}\right) \\ &\geq (1-2\delta)n\log 1.4 + (1-2\delta)\log\frac{1}{2\delta}\end{aligned} \tag{118}$$

For the RHS, we get the following from Equation 116

$$\text{RHS} \leq \sup_{i\in[\mathcal{P}]} \mathrm{KL}(\mathbb{P}_i^{UXY} || \mathbb{P}_1^{UXY}) \leq \frac{32\left(\psi_w + \psi_u\right)\epsilon^2}{(1-\epsilon^2)\phi_\eta}\mathcal{T}. \tag{119}$$

Therefore, we have the following for any $\epsilon \leq 0.6$

$$\mathcal{T} \geq \frac{(1-2\delta)(1-\epsilon^2)\phi_\eta}{32\left(\psi_w + \psi_u\right)} \cdot \frac{n + \log\frac{1}{2\delta}}{\epsilon^2} \geq \frac{\phi_\eta(1-2\delta)\log 1.4}{50\left(\psi_w + \psi_u\right)} \cdot \frac{n + \log\frac{1}{2\delta}}{\epsilon^2}. \tag{120}$$

The last inequality is because $\epsilon \leq 0.6$. Namely, if $\min_{i\in[|\mathcal{P}|]} \mathbb{P}_i^{UY}(F_i) \geq 1-\delta$, then $\mathcal{T} \geq \frac{\phi_\eta(1-2\delta)\log 1.4}{50(\psi_w+\psi_u)} \cdot \frac{n+\log\frac{1}{2\delta}}{\epsilon^2}$. In other words, if $\mathcal{T} < \frac{\phi_\eta(1-2\delta)\log 1.4}{50(\psi_w+\psi_u)} \cdot \frac{n+\log\frac{1}{2\delta}}{\epsilon^2}$, there exists at least one $i \in [|\mathcal{P}|]$ such that $\mathbb{P}_i^{UY}(F_i) < 1-\delta$. For this i, we then know that

$$\mathbb{P}_i^{UY}(\bar{F}_i) \geq \delta. \tag{121}$$

where $\bar{F}_i$ is the complementary event of F_i.

□

Corollary B.3. *Consider the same setting as Theorem 4.1. Moreover, assume* $\left\|\widehat{B}(\mathcal{D})\right\|, \left\|\widehat{C}(\mathcal{D})\right\| \leq \bar{U}$, *i.e., the estimator outputs bounded approximations. Let* ϵ *be any scalar such that* $\epsilon \leq \min\{0.2/\bar{U}, \left\|\widehat{B}(\mathcal{D})\right\| \left\|\widehat{C}(\mathcal{D})\right\| /\bar{U}\}$. *Then if*

$$\mathcal{T} < \frac{\phi_\eta(1-2\delta)\log 1.4}{450\bar{U}^2\left(\psi_w + \psi_u\right)} \cdot \frac{n + \log\frac{1}{2\delta}}{\epsilon^2} \tag{122}$$

there exists a minimal system $\mathcal{M}_0 = (r, n, m, A_0, B_0, C_0, \Sigma_w, \Sigma_\eta)$ *with dataset* $\mathcal{D}$ *such that*

$$\mathbb{P}\left\{\max\left\{\left\|\widehat{C}\left(\mathcal{D}\right) - C_0 S\right\|, \left\|\widehat{B}\left(\mathcal{D}\right) - S^{-1}B_0\right\|\right\} \geq \epsilon\right\} \geq \delta. \tag{123}$$

Proof. Let $\epsilon' = 3\bar{U}\epsilon$. For all $\epsilon \leq 0.2/\bar{U}$, we know that $\epsilon' \leq 0.6$. Therefore, applying Theorem 4.1 with ϵ' gives the following result. If $T < \frac{\phi_\eta(1-2\delta)\log 1.4}{450(\psi_w+\psi_u)\bar{U}^2} \cdot \frac{n+\log\frac{1}{2\delta}}{\epsilon^2}$, there exists $\mathcal{M}_0 = (r, n, m, A_0, B_0, C_0, \Sigma_w, \Sigma_\eta)$ such that

$$\mathbb{P}\left\{\left\|\widehat{C}\left(\mathcal{D}\right)\widehat{B}\left(\mathcal{D}\right) - C_0B_0\right\| \geq 3\bar{U}\epsilon\right\} \geq \delta. \tag{124}$$

Since $\epsilon \leq \frac{\left\|\widehat{B}(\mathcal{D})\right\|\left\|\widehat{C}(\mathcal{D})\right\|}{\bar{U}}$, we know that $3\bar{U}\epsilon \leq 3\left\|\widehat{B}(\mathcal{D})\right\|\left\|\widehat{C}(\mathcal{D})\right\|$. Therefore, applying Lemma B.4 gives

$$\begin{aligned}
&\left\|\widehat{C}\left(\mathcal{D}\right)\widehat{B}\left(\mathcal{D}\right) - C_0B_0\right\| \geq 3\bar{U}\epsilon \\
\Rightarrow\ &\left\|\widehat{C}(\mathcal{D}) - C_0S\right\| \geq \frac{3\bar{U}\epsilon}{3\left\|\widehat{B}\right\|} \geq \epsilon \quad \text{or} \quad \left\|\widehat{B}(\mathcal{D}) - S^{-1}B_0\right\| \geq \frac{\epsilon}{3\left\|\widehat{C}\right\|} \geq \epsilon.
\end{aligned} \tag{125}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\max\left\{\left\|\widehat{C}\left(\mathcal{D}\right) - C_0S\right\|, \left\|\widehat{B}\left(\mathcal{D}\right) - S^{-1}B_0\right\|\right\} \geq \epsilon\right\} \\
\geq\ &\mathbb{P}\left\{\left\|\widehat{C}\left(\mathcal{D}\right)\widehat{B}\left(\mathcal{D}\right) - C_0B_0\right\| \geq 3\bar{U}\epsilon\right\} \geq \delta. \\
\geq\ &\delta.
\end{aligned} \tag{126}$$

□

Lemma B.4. *Fix any invertible matrix S. Suppose $\left\|\widehat{C}\widehat{B} - CB\right\| \geq \epsilon$ for $\epsilon \leq 3\left\|\widehat{B}\right\|\left\|\widehat{C}\right\|$. Then either*

$$\left\|\widehat{C} - CS\right\| \geq \frac{\epsilon}{3\left\|\widehat{B}\right\|} \quad \textit{or} \quad \left\|\widehat{B} - S^{-1}B\right\| \geq \frac{\epsilon}{3\left\|\widehat{C}\right\|}. \tag{127}$$

Remark B.5. In other words, for any similarity transformation, we can not learn either $\tilde{B}$ or $\tilde{C}$ well.

Proof. We show this claim by contradiction. For any invertible matrix S, suppose

$$\left\|\widehat{C} - CS\right\| < \frac{\epsilon}{3\left\|\widehat{B}\right\|} \quad \text{and} \quad \left\|\widehat{B} - S^{-1}B\right\| < \frac{\epsilon}{3\left\|\widehat{C}\right\|}. \tag{128}$$

Then for any $\epsilon \leq \|B\|\,\|C\|$, we know that

$$\begin{aligned}
&\left\|\widehat{C}\widehat{B} - CB\right\| \\
\leq\ &\left\|\widehat{C}\widehat{B} - CS\widehat{B}\right\| + \left\|CS\widehat{B} - CSS^{-1}B\right\| \\
\leq\ &\left\|\widehat{C} - CS\right\|\left\|\widehat{B}\right\| + \|CS\|\left\|\widehat{B} - S^{-1}B\right\| \\
\leq\ &\frac{\epsilon}{3} + \left(\left\|CS - \widehat{C}\right\| + \left\|\widehat{C}\right\|\right)\frac{\epsilon}{3\left\|\widehat{C}\right\|} \\
\leq\ &\frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}\frac{\epsilon}{3\left\|\widehat{B}\right\|\left\|\widehat{C}\right\|} \\
\leq\ &\epsilon,
\end{aligned} \tag{129}$$

which causes a contradiction. □

C. Upper Bounds for `Meta-SYSID` — Proof of Theorem 5.1

For this section, we consider a slightly more general setting than `Meta-SYSID` and prove the theorem in this setting. Consider K minimal systems $\mathcal{M}_k = \left(r_k, n, m_k, A_k, B_k, C_k, \Sigma_{w,k}, \sigma_{\eta,k}^2 I\right)$ with $r_k, m_k \ll n (\forall k \in [K])$ and the same observer column space. Namely,

$$\mathrm{col}(C_1) = \mathrm{col}(C_2) = \cdots = \mathrm{col}(C_K) \tag{130}$$

For every system $\mathcal{M}_k$, we choose an independent input sequence $\mathcal{U}_k = \{u_{k,t}\}_{t=0}^{T_k-1}$ with $u_{k,t} \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_{u,k})$, and we observe $\mathcal{Y}_k = \{y_{k,t}\}_{t=0}^{T_k}$. This single trajectory dataset by $\mathcal{D}_k = \mathcal{Y}_k \cup \mathcal{U}_k$.

For the above systems, we define the following notations

$$\begin{aligned}
&r = \max_{k\in[K]} r_k, \quad m = \max_{k\in[K]} m_k, \quad \phi_u = \min_{k\in[K]} \sigma_{\min}(\Sigma_u) \\
&\psi_C = \max_{k\in[K]} \sigma_1\left(C_k\right), \quad \psi_\eta = \max_{k\in[K]} \sigma_{\eta,k}^2, \quad \psi_w = \max_{k\in[K]} \sigma_1(\Sigma_{w,k} + B_k \Sigma_{u,k} B_k{}^\intercal), \\
&\phi_C = \min_{k\in[K]} \sigma_{\min}(C_k), \quad \phi_O = \min_{k\in[K]} \sigma_{\min}\begin{pmatrix}\begin{bmatrix} C_k \\ C_k A_k \\ \vdots \\ C_k A_k^{r-1} \end{bmatrix}\end{pmatrix}, \quad \phi_R = \min_{k\in[K]} \sigma_{\min}\left(\begin{bmatrix} B_k & A_k B_k & \dots & A_k^{r-1} B_k \end{bmatrix}\right)
\end{aligned} \tag{131}$$

For $\widehat{\Phi}_{C,k}$ from Algorithm 3, we define the following auxiliary system $\widehat{\mathcal{M}}_k$ (the projected version of system $\mathcal{M}_k$) which will be useful for further analysis.

$$\begin{aligned}
x_{k,t+1} &= A_k x_{k,t} + B_k u_{k,t} + w_{k,t}, \\
\tilde{y}_{k,t} &= \widehat{\Phi}_{C,k}{}^\intercal C_k x_{k,t} + \widehat{\Phi}_{C,k}{}^\intercal \eta_{k,t}.
\end{aligned} \tag{132}$$

Now we are ready to restate Theorem 5.1 in full details.

Theorem C.1 (Theorem 5.1 Restated for the More General Setting)**.** *Consider the systems $\mathcal{M}_{[K]}$, datasets $\mathcal{D}_{[K]} = \mathcal{Y}_{[K]} \cup \mathcal{U}_{[K]}$ and constants defined above. Let $\mathcal{M}_{[K]}$ satisfy Assumption 3.1 with constants ψ_A and ρ_A. Fix any system $k_0 \in [K]$. If T_{k_0} and $T_{-k_0} \coloneqq \sum_{k\neq k_0} T_k$ satisfy*

$$T_{-k_0} \gtrsim \kappa_3 \cdot n^2 r^3, \quad T_{k_0} \geq \kappa_1 \cdot \mathit{poly}\left(r, m\right), \tag{133}$$

then $\left(\widehat{A}_{k_0}, \widehat{B}_{k_0}, \widehat{C}_{k_0}\right)$ from Algorithm 3 satisfy the following for some invertible matrix S with probability at least $1-\delta$

$$\begin{aligned}
&\max\left\{\left\|S^{-1} A_{k_0} S - \widehat{A}_{k_0}\right\|, \left\|S^{-1} B_{k_0} - \widehat{B}_{k_0}\right\|, \left\|C_{k_0} S - \widehat{C}_{k_0}\right\|\right\} \\
&\lesssim \kappa_4 \cdot \sqrt{\frac{n}{T_{-k_0}}} \left\|\widehat{C}_{k_0}\right\| + \kappa_2 \cdot \sqrt{\frac{\mathit{poly}\left(r, m\right)}{T_{k_0}}}.
\end{aligned} \tag{134}$$

Here $\kappa_1 = \kappa_1(\widehat{\mathcal{M}}_{k_0}, \mathcal{U}_{k_0}, \delta)$ and $\kappa_2 = \kappa_2(\widehat{\mathcal{M}}_{k_0}, \mathcal{U}_{k_0}, \delta)$ are defined in Definition 3.2. $\kappa_3 = \kappa_3\left(\mathcal{M}_{[K]}, \mathcal{U}_{[K]}, \delta\right)$, $\kappa_4 = \kappa_4\left(\mathcal{M}_{-k_0}, \mathcal{U}_{-k_0}, \delta\right)$ are detailed below. All of them are problem-related constants independent of system dimensions modulo logarithmic factors.

$$\begin{aligned}
\kappa_3\left(\mathcal{M}_{[K]}, \mathcal{U}_{[K]}, \delta\right) &= \max\left\{\kappa_4^2 \frac{\psi_A^2 \psi_C^2}{(1-\rho_A^2)\phi_O^2}, \left(\frac{\psi_\eta^2 \psi_C^2 \psi_w \psi_A^2}{(1-\rho_A^2)\phi_u^2 \phi_C^4 \phi_R^4}\right)^2 \log^2\left(\frac{\psi_C^2 \psi_w \psi_A^4}{1-\rho_A^2} r \log\frac{Kr}{\delta}\right) \log^4\left(\frac{Kr}{\delta}\right)\right\}, \\
\kappa_4\left(\mathcal{M}_{-k_0}, \mathcal{U}_{-k_0}, \delta\right) &= \frac{\psi_\eta}{\phi_u \phi_C^2 \phi_R^2} \sqrt{\log\frac{1}{\delta}}.
\end{aligned} \tag{135}$$

Proof of Theorem C.1. For clarity of the proof, we abbreviate all subscripts k_0. Based on Equation (133), T_{-k_0} satisfies the condition of Lemma C.2. We apply Lemma C.2 on $(\mathcal{D}_{-k_0}, \Sigma_{\eta,-k_0})$ and get the following with probability at least $1-\frac{\delta}{2}$

$$\left\|\widehat{\Phi}_C^{\perp\intercal}\Phi_C\right\| \lesssim \kappa_4\sqrt{\frac{n}{T_{-k_0}}} \coloneqq \Delta_\Phi. \tag{136}$$

The system generating dataset $\tilde{\mathcal{D}}$, denoted by $\widehat{\mathcal{M}}$, is rewritten in Equation (132). Following exactly the same derivation as in Step 1 of proof of Theorem A.1, we know that $\widehat{\mathcal{M}}$ is minimal. Since $\widehat{\Phi}_C$ is independent of the trajectory from $\mathcal{M}$, the noises $\{\widehat{\Phi}_C{}^\intercal\eta_t\}_{t=0}^T$ are sampled independently of other variables of this trajectory. Moreover, $\widehat{\mathcal{M}}$ satisfy Assumption 3.1. Therefore, we can apply `Sys-Oracle`. Following exactly the same derivation as in Step 2 of proofs of Theorem A.1, we get the following for some invertible matrix S with probability at least $1-\delta$

$$\max\left\{\left\|S^{-1}AS-\widehat{A}\right\|,\left\|S^{-1}B-\widehat{B}\right\|,\left\|CS-\widehat{C}\right\|\right\} \lesssim \kappa_2\cdot\sqrt{\frac{\text{poly}\,(r,m)}{T}}+\kappa_4\cdot\sqrt{\frac{n}{T_{-k_0}}}\left\|\widehat{C}\right\|. \tag{137}$$

□

C.1. Upper Bounds for `Col-Approx`

The theoretical guarantee for `Col-Approx` with multiple trajectories in Algorithm 3 is presented in the following lemma.

Lemma C.2. *Fix any positive integer K. Consider K systems $\mathcal{M}_{[K]}$ and their datasets $\mathcal{D}_{[K]}=\mathcal{Y}_{[K]}\cup\mathcal{U}_{[K]}$ satisfying the description in Appendix C. Suppose the systems satisfy Assumption 3.1 with constants ψ_A and ρ_A. With the notations defined in Appendix C, if $\mathcal{T}\coloneqq\sum_{k\in[K]}T_k$ satisfies the following inequality*

$$\mathcal{T}\gtrsim\underbrace{\left(\frac{\psi_\eta^2\psi_C^2\psi_w\psi_A^2}{(1-\rho_A^2)\phi_u^2\phi_C^4\phi_R^4}\right)^2\log^2\left(\frac{\psi_C^2\psi_w\psi_A^4}{1-\rho_A^2}r\log\frac{Kr}{\delta}\right)\log^4(\frac{Kr}{\delta})}_{\kappa_5\left(\mathcal{M}_{[K]},\mathcal{U}_{[K]},\delta\right)}\cdot n^2r^3, \tag{138}$$

then $\widehat{\Phi}_C=$ `Col-Approx`$\left(\mathcal{Y}_{[K]}\right)$ (Algorithm 2) satisfies the following with probability at least $1-\delta$

$$\widehat{r}_c=\textit{rank}(C_1)=\cdots=\textit{rank}(C_K),\quad \left\|\widehat{\Phi}_C^{\perp\intercal}\Phi_C\right\|\lesssim\underbrace{\frac{\psi_\eta}{\phi_u\phi_C^2\phi_R^2}\sqrt{\log\frac{1}{\delta}}}_{\kappa_4\left(\mathcal{M}_{[K]},\mathcal{U}_{[K]},\delta\right)}\cdot\sqrt{\frac{n}{\mathcal{T}}}. \tag{139}$$

Proof. From the system dynamics, we know that

$$\begin{aligned}\Sigma_y&=\sum_{k\in[K]}\sum_{t=0}^{T_k}y_{k,t}y_{k,t}{}^\intercal=\sum_{k\in[K]}\sum_{t=0}^{T_k}C_kx_{k,t}x_{k,t}{}^\intercal C_k{}^\intercal+\sum_{k\in[K]}\sum_{t=0}^{T_k}\eta_{k,t}\eta_{k,t}{}^\intercal+\sum_{k\in[K]}\sum_{t=0}^{T_k}(C_kx_{k,t}\eta_{k,t}{}^\intercal+\eta_{k,t}C_k{}^\intercal x_{k,t}{}^\intercal)\\&=\sum_{k\in[K]}\sum_{t=0}^{T_k}C_kx_{k,t}x_{k,t}{}^\intercal C_k{}^\intercal+\sum_{k\in[K]}\sum_{t=0}^{T_k}(\eta_{k,t}\eta_{k,t}{}^\intercal-\sigma_{\eta,k}^2I)+\sum_{k\in[K]}\sum_{t=0}^{T_k}(C_kx_{k,t}\eta_{k,t}{}^\intercal+\eta_{k,t}C_k{}^\intercal x_{k,t}{}^\intercal)\\&\quad+\sum_{k\in[K]}(T_k+1)\sigma_{\eta,k}^2I\end{aligned} \tag{140}$$

Here the first term is the information on the shared column space of $\{C_k\}_{k\in[K]}$, while the second and third terms are only noise. For the rest of the proof, we first upper bound the norms of the noises (*step 1*). With this, we show that $\widehat{r}_c=\text{rank}(C_1)$ with high probability (*step 2*). Then we apply our subspace perturbation result to upper bound the influence of the noises on the eigenspace of the first term (*step 3*).

Step 1: Noise Norm Upper Bounds. With $\text{rank}(C_1) = r_c$ and $\text{col}(C_1) = \cdots = \text{col}(C_K)$, we can rewrite $C_k = \Phi_C \alpha_k$ where an orthonormal basis of $\text{col}(C_1)$ forms the columns of $\Phi_C \in \mathbb{R}^{n \times r_c}$ and $\alpha_k \in \mathbb{R}^{r_c \times r}$ is a full row rank matrix. It is then clear that

$$\sigma_{\min}(\alpha_k) = \sigma_{\min}(C_k) \geq \phi_C, \quad \sigma_1(\alpha_k) = \sigma_1(C_k) \leq \psi_C, \quad \forall k \in [K]. \tag{141}$$

Let $\Sigma_C = \sum_{k\in[K]} \sum_{t=0}^{T_k} C_k x_{k,t} x_{k,t}^\intercal C_k^\intercal$, $\bar{\Sigma}_C = \Sigma_C + \mathcal{T} I$ and $\Sigma_\alpha = \sum_{k\in[K]} \sum_{t=0}^{T_k} \alpha_k x_{k,t} x_{k,t}^\intercal \alpha_k^\intercal$. We then have $\Sigma_C = \Phi_C \Sigma_\alpha \Phi_C^\intercal$ and $\bar{\Sigma}_C = \Phi_C \Sigma_\alpha \Phi_C^\intercal + \mathcal{T} I$. Then from Lemma C.3, Lemma C.4, Lemma A.5, and Lemma A.6, we have the following with probability at least $1 - \delta$

$$\begin{aligned}
&\frac{\phi_u \phi_C^2 \phi_R^2}{8} \mathcal{T} I \preceq \Sigma_\alpha \precsim \frac{\psi_C^2 \psi_w \psi_A^2}{1 - \rho_A^2} r \mathcal{T} \log \frac{K}{\delta} I, \\
&\left\| \left(\bar{\Sigma}_C\right)^{-\frac{1}{2}} \sum_{k\in[K]} \sum_{t=0}^{T_k} C_k x_{k,t} \eta_{k,t}^\intercal \right\| \lesssim \tilde{U} \sqrt{\psi_\eta n \log \frac{1}{\delta}}, \\
&\left\| \sum_{k\in[K]} \sum_{t=0}^{T_k} \left(\eta_{k,t} \eta_{k,t}^\intercal - \sigma_{\eta,k}^2 I\right) \right\| \lesssim \psi_\eta \sqrt{n \mathcal{T} \log \frac{1}{\delta}},
\end{aligned} \tag{142}$$

with $\tilde{U} = \sqrt{\log\left(\frac{\psi_\alpha^2 \psi_w \psi_A^4}{1-\rho_A^2} r \log \frac{2Kr}{\delta}\right)}$.

Step 2: Order Estimation Guarantee.

$$\Sigma_y = \Sigma_C + \underbrace{\sum_{k\in[K]} \sum_{t=0}^{T_k} \left(\eta_{k,t} \eta_{k,t}^\intercal - \sigma_{\eta,k}^2 I\right) + \sum_{k\in[K]} \sum_{t=0}^{T_k} \left(C_k x_{k,t} \eta_{k,t}^\intercal + \eta_{k,t} C_k^\intercal x_{k,t}^\intercal\right)}_{\Delta} + \sum_{k\in[K]} (T_k + 1) \sigma_{\eta,k}^2 I. \tag{143}$$

The inequalities of Step 1 imply

$$\begin{aligned}
\|\Delta\| &\leq \left\| \sum_{k\in[K]} \sum_{t=0}^{T_k} \left(\eta_{k,t} \eta_{k,t}^\intercal - \sigma_{\eta,k}^2 I\right) \right\| + 2 \left\| \sum_{k\in[K]} \sum_{t=0}^{T_k} C_k x_{k,t} \eta_{k,t}^\intercal \right\| \\
&\lesssim \psi_\eta \sqrt{n \mathcal{T} \log \frac{1}{\delta}} + \tilde{U} \sqrt{\psi_\eta n \log \frac{1}{\delta}} \left\| \left(\bar{\Sigma}_C\right)^{\frac{1}{2}} \right\| \\
&= \psi_\eta \sqrt{n \mathcal{T} \log \frac{1}{\delta}} + \tilde{U} \sqrt{\psi_\eta n \log \frac{1}{\delta}} \left(\left\| \left(\mathcal{T} I + \Sigma_\alpha\right)^{\frac{1}{2}} \right\| \right) \\
&\lesssim \psi_\eta \sqrt{n \mathcal{T} \log \frac{1}{\delta}} + \tilde{U} \sqrt{\psi_\eta n \log \frac{1}{\delta}} \sqrt{\frac{\psi_C^2 \psi_w \psi_A^2}{1 - \rho_A^2} r \mathcal{T} \log \frac{K}{\delta} + \mathcal{T}} \\
&\lesssim \tilde{U} \psi_\eta \sqrt{n \log \frac{1}{\delta}} \sqrt{\frac{\psi_C^2 \psi_w \psi_A^2}{1 - \rho_A^2} r \mathcal{T} \log \frac{K}{\delta}} \\
&= \sqrt{\frac{\psi_\eta^2 \psi_C^2 \psi_w \psi_A^2}{1 - \rho_A^2}} \tilde{U} \log \frac{K}{\delta} \sqrt{n r \mathcal{T}}
\end{aligned} \tag{144}$$

Therefore, for each $i \in [r_c]$ we have the following for some positive constant c_1

$$\begin{aligned}
\sigma_i\left(\Sigma_y\right) - \sum_{k\in[K]} (T_k + 1) \sigma_{\eta,k}^2 &\geq \sigma_i(\Sigma_C) - \|\Delta\| \geq \sigma_{\min}(\Sigma_\alpha) - \|\Delta\| \\
&\geq \frac{\phi_u \phi_C^2 \phi_R^2}{8} \mathcal{T} - c_1 \sqrt{\frac{\psi_\eta^2 \psi_C^2 \psi_w \psi_A^2}{1 - \rho_A^2}} \tilde{U} \log \frac{K}{\delta} \sqrt{n r \mathcal{T}} \\
&\geq \frac{\phi_u \phi_C^2 \phi_R^2}{16} \mathcal{T}.
\end{aligned} \tag{145}$$

Here the first inequality holds according to Theorem 1 in (Stewart, 1990) and the last line is because of Equation 138. For $i \in [r_c+1, n]$,

$$\sigma_i(\Sigma_y) - \sum_{k\in[K]} (T_k+1)\sigma_{\eta,k}^2 \le \|\Delta\| \le c_1 \sqrt{\frac{\psi_\eta^2 \psi_C^2 \psi_w \psi_A^2}{1-\rho_A^2}} \tilde{U} \log\frac{K}{\delta} \sqrt{nr\mathcal{T}}. \tag{146}$$

Based on the above three inequalities and Equation 138, we conclude that with probability at least $1-\delta$, the following hold

$$\begin{aligned}
\sigma_i(\Sigma_y) - \sigma_j(\Sigma_y) &\le 2c_1 \sqrt{\frac{\psi_\eta^2 \psi_C^2 \psi_w \psi_A^2}{1-\rho_A^2}} \tilde{U} \log\frac{K}{\delta} \sqrt{nr\mathcal{T}} \\
&< \mathcal{T}^{3/4}, \quad \forall i < j \in [r_c], \\
\sigma_{r_c}(\Sigma_y) - \sigma_{r_c+1}(\Sigma_y) &\ge \frac{\phi_u \phi_C^2 \phi_R^2}{16} \mathcal{T} - c_1 \sqrt{\frac{\psi_\eta^2 \psi_C^2 \psi_w \psi_A^2}{1-\rho_A^2}} \tilde{U} \log\frac{K}{\delta} \sqrt{nr\mathcal{T}} \\
&> \mathcal{T}^{3/4}.
\end{aligned} \tag{147}$$

Therefore, from the definition of $\widehat{r}_c$, we know that $\widehat{r}_c = r_c$ with probability at least $1-\delta$.

Step 3: Column Space Estimation Guarantee. With $\widehat{r}_c = r_c$, now we try to apply our subspace perturbation result, i.e. Lemma A.7, on matrix $\Sigma_y - \sum_{k\in[K]}(T_k+1)\sigma_{\eta,k}^2 I + \mathcal{T}I$. Notice that this matrix has exactly the same eigenspace as Σ_y and therefore the eigenspace of its first r_c eigenvectors is $\widehat{\Phi}_C$ (line 9 in Algorithm 2). This matrix can be decomposed as

$$\begin{aligned}
&\Sigma_y - \sum_{k\in[K]} (T_k+1)\sigma_{\eta,k}^2 I + \mathcal{T}I \\
=& \underbrace{\bar{\Sigma}_C}_{M \text{ in Lemma A.7}} + \underbrace{\sum_{k\in[K]} \sum_{t=0}^{T_k} \left(\eta_{k,t}\eta_{k,t}^\intercal - \sigma_{\eta,k}^2 I\right)}_{\Delta_2 \text{ in Lemma A.7}} + \underbrace{\sum_{k\in[K]} \sum_{t=0}^{T_k} \left(C_k x_{k,t} \eta_{k,t}^\intercal + \eta_{k,t} C_k^\intercal x_{k,t}^\intercal\right)}_{\Delta_1+\Delta_1^\intercal \text{ in Lemma A.7}}.
\end{aligned} \tag{148}$$

For matrix $\bar{\Sigma}_C = \Phi_C \Sigma_\alpha \Phi_C^\intercal + \mathcal{T}I = \Phi_C\left(\Sigma_\alpha + \mathcal{T}I\right)\Phi_C^\intercal + \mathcal{T}\Phi_C^\perp \Phi_C^{\perp\intercal}$, it is clear that its SVD can be written as

$$\bar{\Sigma}_C = \begin{bmatrix}\tilde{\Phi}_C & \Phi_C^\perp\end{bmatrix} \begin{bmatrix}\Lambda_1 & \\ & \mathcal{T}I\end{bmatrix} \begin{bmatrix}\tilde{\Phi}_C^\intercal \\ \Phi_C^{\perp\intercal}\end{bmatrix}, \quad \Lambda_1 = \text{diag}\left(\sigma_1(\Sigma_\alpha)+\mathcal{T}, \ldots, \sigma_{\min}(\Sigma_\alpha)+\mathcal{T}\right). \tag{149}$$

where $\tilde{\Phi}_C$ is an orthonormal basis of $\text{col}(C_1)$. Then we conclude that the following hold for some large enough positive constant c_3

$$\begin{aligned}
&\sigma_1\left(\bar{\Sigma}_C\right) \le \mathcal{T} + \sigma_1(\Sigma_\alpha) \le c_3\left(1 + \frac{\psi_C^2 \psi_w \psi_A^2}{1-\rho_A^2} r \log\frac{K}{\delta}\right)\mathcal{T}, \\
&\sigma_{r_c}\left(\bar{\Sigma}_C\right) - \sigma_{r_c+1}(\bar{\Sigma}_C) = \sigma_{\min}(\Sigma_\alpha) \ge \frac{\phi_u \phi_C^2 \phi_R^2}{32}\mathcal{T}, \quad \sigma_{\min}(\bar{\Sigma}_C) = \mathcal{T}.
\end{aligned} \tag{150}$$

Now we are ready to apply Lemma A.7 on $\bar{\Sigma}_C$ with the following for positive constants c_4, c_5 large enough

$$\begin{aligned}
&\alpha = c_4 \tilde{U}\sqrt{\psi_\eta n \log\frac{1}{\delta}}, \quad \beta = c_5 \psi_\eta \sqrt{n\mathcal{T}\log\frac{1}{\delta}}, \\
&\delta_M = \frac{\phi_u \phi_C^2 \phi_R^2}{32}\mathcal{T}, \quad \psi_M = 2c_3 \frac{\psi_C^2 \psi_w \psi_A^2}{1-\rho_A^2} r\mathcal{T}\log\frac{K}{\delta}, \quad \phi_M = \mathcal{T}, \quad \sigma_1(\Lambda_2) = \mathcal{T}.
\end{aligned} \tag{151}$$

Again, $\tilde{U} = \sqrt{\log\left(\frac{\psi_\alpha^2\psi_w\psi_A^4}{1-\rho_A^2} r \log\frac{2Kr}{\delta}\right)}$. From Equation 138, it is clear that $\sqrt{\phi_M} \gtrsim \alpha$ and $\delta_M \gtrsim \alpha\sqrt{\psi_M} + \beta$. Therefore,

$$\begin{aligned}
\left\|\widehat{\Phi}_C^{\perp\intercal}\Phi_C\right\| = \left\|\widehat{\Phi}_C^{\perp\intercal}\tilde{\Phi}_C\right\| &\lesssim \frac{\psi_\eta}{\phi_u\phi_C^2\phi_R^2}\sqrt{\frac{n}{\mathcal{T}}\log\frac{1}{\delta}} + \tilde{U}\frac{\sqrt{\psi_\eta}}{\phi_u\phi_C^2\phi_R^2}\sqrt{\frac{n}{\mathcal{T}}\log\frac{1}{\delta}} + \tilde{U}^2\frac{\psi_\eta\sqrt{\psi_C^2\psi_w\psi_A^2}}{\phi_u\phi_C^2\phi_R^2\sqrt{1-\rho_A^2}}\frac{n\sqrt{r}}{\mathcal{T}}\log^2\frac{K}{\delta} \\
&\lesssim \frac{\psi_\eta}{\phi_u\phi_C^2\phi_R^2}\sqrt{\frac{n}{\mathcal{T}}\log\frac{1}{\delta}} + \tilde{U}^2\frac{\psi_\eta\sqrt{\psi_C^2\psi_w\psi_A^2}}{\phi_u\phi_C^2\phi_R^2\sqrt{1-\rho_A^2}}\frac{n\sqrt{r}}{\mathcal{T}}\log^2\frac{K}{\delta} \\
&\lesssim \frac{\psi_\eta}{\phi_u\phi_C^2\phi_R^2}\sqrt{\frac{n}{\mathcal{T}}\log\frac{1}{\delta}}.
\end{aligned} \tag{152}$$

□

C.1.1. Supporting Details

Lemma C.3. *Consider the same setting as Theorem C.2. Consider full row rank matrices $\{\alpha_k \in \mathbb{R}^{a\times r_k}\}_{k\in[K]}$ for any positive integer $a \le \min_k r_k$. Let*

$$\Sigma_\alpha = \sum_{k\in[K],t\in[T_k]} \alpha_k x_{k,t} x_{k,t}^\intercal \alpha_k^\intercal, \quad \phi_\alpha = \min\left\{\min_{k\in[K]}\{\sigma_a(\alpha_k)\}, 1\right\}, \quad \psi_\alpha = \max\left\{\max_{k\in[K]}\{\sigma_1(\alpha_k)\}, 1\right\}. \tag{153}$$

Then the following holds with probability at least $1-\delta$ if $\mathcal{T} \gtrsim \frac{\psi_w^2\psi_\alpha^4\psi_A^2}{\phi_w^2\phi_u^2\phi_\alpha^4\phi_R^4} Kr^3 \log\frac{r}{\delta} \cdot \log\left(\frac{\psi_\alpha^2\psi_w\psi_A^4}{1-\rho_A^2} r\log\frac{2Kr}{\delta}\right)$,

$$\Sigma_\alpha \succeq \frac{\phi_u\phi_\alpha^2\phi_R^2}{8}\mathcal{T}I. \tag{154}$$

Proof. We first define set $\mathcal{K} = \{k : T_k \ge r\}$. For further analysis, we let $\tilde{w}_{k,t} \coloneqq B_k u_{k,t} + w_{k,t} \sim \mathcal{N}(0, \Sigma_{\tilde{w},k})$ with $\Sigma_{\tilde{w},k} \coloneqq B_k\Sigma_{u,k}B_k^\intercal + \Sigma_{w,k}$ and we can rewrite the system dynamics as follows

$$\begin{aligned}
x_{k,t+1} &= A_k x_{k,t} + \tilde{w}_{k,t} = A_k x_{k,t} + B_k u_{k,t} + w_{k,t}, \\
y_{k,t} &= C_k x_{k,t} + \eta_{k,t}.
\end{aligned} \tag{155}$$

Then we know that $\sigma_1(\Sigma_{\tilde{w},k}) \le \psi_w$. For simplicity, we define $\Sigma_{\alpha,\tau} \coloneqq \sum_{k\in\mathcal{K}}\sum_{t=1}^{T_k-\tau}(\alpha_k A_k^\tau)(x_{k,\tau}x_{k,\tau}^\intercal)(\alpha_k A_k^\tau)^\intercal$. It is then clear that

$$\begin{aligned}
\Sigma_\alpha \succeq \Sigma_{\alpha,0} &= \sum_{k\in\mathcal{K}}\sum_{t=1}^{T_k}\alpha_k x_{k,t}x_{k,t}^\intercal\alpha_k^\intercal \\
&= \sum_{k\in\mathcal{K}}\sum_{t=1}^{T_k}\alpha_k\left(A_k x_{k,t-1}x_{k,t-1}^\intercal A_k^\intercal + \tilde{w}_{k,t-1}\tilde{w}_{k,t-1}^\intercal + A_k x_{k,t-1}\tilde{w}_{k,t-1}^\intercal + \tilde{w}_{k,t-1}x_{k,t-1}^\intercal A_k^\intercal\right)\alpha_k^\intercal \\
&= \Sigma_{\alpha,1} + \sum_{k\in\mathcal{K}}\sum_{t=0}^{T_k-1}\alpha_k\tilde{w}_{k,t}\tilde{w}_{k,t}^\intercal\alpha_k^\intercal + \sum_{k\in\mathcal{K}}\sum_{t=0}^{T_k-1}\left(\alpha_k A_k x_{k,t}\left(\alpha_k\tilde{w}_{k,t}\right)^\intercal + \left(\alpha_k\tilde{w}_{k,t}\right)x_{k,t}^\intercal A_k^\intercal\alpha_k^\intercal\right).
\end{aligned} \tag{156}$$

By Lemma C.4 and A.5, the following events hold with probability at least $1-\delta/r$,

$$\begin{aligned}
\left\|\sum_{k\in\mathcal{K}}\sum_{t=0}^{T_k-1}\alpha_k\tilde{w}_{k,t}\tilde{w}_{k,t}^\intercal\alpha_k^\intercal - \sum_{k\in\mathcal{K}}T_k\alpha_k\Sigma_{\tilde{w},k}\alpha_k^\intercal\right\| &\lesssim \psi_w\psi_\alpha^2\sqrt{r\mathcal{T}\log\frac{2r}{\delta}}, \\
\left\|(\Sigma_{\alpha,1}+\mathcal{T}I)^{-\frac{1}{2}}\sum_{k\in\mathcal{K}}\sum_{t=0}^{T_k-1}\alpha_k A_k x_{k,t}\left(\alpha_k\tilde{w}_{k,t}\right)^\intercal\right\| &\lesssim \tilde{U}\sqrt{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}},
\end{aligned} \tag{157}$$

with $\tilde{U} = \sqrt{\log\left(\frac{\psi_\alpha^2\psi_w\psi_A^4}{1-\rho_A^2} r \log\frac{2Kr}{\delta}\right)}$. From the first inequality, it is clear that the following holds for some large enough positive constant c_1

$$\sum_{k\in\mathcal{K}}\sum_{t=0}^{T_k-1}\alpha_k\tilde{w}_{k,t}\tilde{w}_{k,t}{}^\intercal\alpha_k{}^\intercal \succeq \sum_{k\in\mathcal{K}} T_k\alpha_k\Sigma_{\tilde{w},k}\alpha_k{}^\intercal - c_1\psi_w\psi_\alpha^2\sqrt{r\mathcal{T}\log\frac{2r}{\delta}}I \tag{158}$$

We have the following from the second inequality

$$\begin{aligned}
&\left\|(\Sigma_{\alpha,1}+\mathcal{T}I)^{-\frac{1}{2}}\sum_{k\in\mathcal{K}}\sum_{t=0}^{T_k-1}\alpha_k A_k x_{k,t}\left(\alpha_k\tilde{w}_{k,t}\right)^\intercal(\Sigma_{\alpha,1}+\mathcal{T}I)^{-\frac{1}{2}}\right\| \\
\lesssim\ &\tilde{U}\sqrt{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}\left\|(\Sigma_{\alpha,1}+\mathcal{T}I)^{-\frac{1}{2}}\right\| \\
\lesssim\ &\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{\mathcal{T}}}.
\end{aligned} \tag{159}$$

This implies the following for some positive constant c_2

$$\sum_{k\in\mathcal{K}}\sum_{t=0}^{T_k-1}\left(\alpha_k A_k x_{k,t}\left(\alpha_k\tilde{w}_{k,t}\right)^\intercal + \left(\alpha_k\tilde{w}_{k,t}\right)x_{k,t}{}^\intercal A_k{}^\intercal\alpha_k{}^\intercal\right) \succeq -c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{\mathcal{T}}}\left(\Sigma_{\alpha,1}+\mathcal{T}I\right). \tag{160}$$

Plugging back into Equation 156 gives the following for some positive constant c_3

$$\begin{aligned}
\Sigma_{\alpha,0} &\succeq \Sigma_{\alpha,1} + \sum_{k\in\mathcal{K}} T_k\alpha_k\Sigma_{\tilde{w},k}\alpha_k{}^\intercal - c_1\psi_w\psi_\alpha^2\sqrt{r\mathcal{T}\log\frac{2r}{\delta}}I - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{\mathcal{T}}}\left(\Sigma_{\alpha,1}+\mathcal{T}I\right) \\
&\succeq \left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\log\frac{2r}{\delta}}{\mathcal{T}}}\right)\Sigma_{\alpha,1} + \left(\sum_{k\in\mathcal{K}} T_k\alpha_k\Sigma_{\tilde{w},k}\alpha_k{}^\intercal - c_3\psi_w\psi_\alpha^2\tilde{U}\sqrt{r\mathcal{T}\log\frac{2r}{\delta}}I\right)
\end{aligned} \tag{161}$$

Similarly, we expand $\Sigma_{\alpha,1}, \Sigma_{\alpha,2}, \cdots, \Sigma_{\alpha,r-1}$ and have the following with probability at least $1-\delta$

$$\begin{aligned}
\Sigma_{\alpha,0} \succeq\ &\left(\sum_{k\in\mathcal{K}}(T_k-r+1)\sum_{i=0}^{r-1}\left(\alpha_k A_k^i\right)\Sigma_{\tilde{w},k}\left(\alpha_k A_k^i\right)^\intercal - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{r(\mathcal{T}-Kr+K)\log\frac{2r}{\delta}}I\right) \\
&\cdot\left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\psi_A^2\log\frac{2r}{\delta}}{\mathcal{T}-Kr+K}}\right)^{r-1} + \left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\psi_A^2\log\frac{2r}{\delta}}{\mathcal{T}-Kr+K}}\right)^{r}\Sigma_{\alpha,r}
\end{aligned}$$

The above inequality is further simplified as follows

$$\begin{aligned}
\Sigma_{\alpha,0} &\overset{(i)}{\succeq} \left(\phi_u \sum_{k\in\mathcal{K}}(T_k - r + 1)\sum_{i=0}^{r-1}\left(\alpha_k A_k^i\right) B_k B_k{}^\intercal \left(\alpha_k A_k^i\right)^\intercal - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{r\mathcal{T}\log\frac{2r}{\delta}}I\right) \\
&\quad\cdot\left(1 - c_2\tilde{U}\sqrt{\frac{r\psi_w\psi_\alpha^2\psi_A^2\log\frac{2r}{\delta}}{\mathcal{T}-Kr+K}}\right)^r \\
&\overset{(ii)}{\succeq} \left(\phi_u\sum_{k\in\mathcal{K}}(T_k - r)\alpha_k\left(\sum_{i=0}^{r-1}A_k^i B_k (A_k^i B_k)^\intercal\right)\alpha_k{}^\intercal - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{r\mathcal{T}\log\frac{2r}{\delta}}I\right)\left(1-\frac{1}{2r}\right)^r \\
&\overset{(iii)}{\succeq} \frac{1}{2}\left(\phi_u\sum_{k\in\mathcal{K}}(T_k - r)\alpha_k\left(\sum_{i=0}^{r-1}A_k^i B_k (A_k^i B_k)^\intercal\right)\alpha_k{}^\intercal - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{r\mathcal{T}\log\frac{2r}{\delta}}I\right) \\
&\overset{(iv)}{\succeq} \frac{1}{2}\left(\phi_u\phi_\alpha^2\phi_R^2(\mathcal{T}-2Kr) - c_3\psi_w\psi_\alpha^2\psi_A^2\tilde{U}\cdot r\sqrt{r\mathcal{T}\log\frac{2r}{\delta}}\right)I \\
&\overset{(v)}{\succeq} \frac{\phi_u\phi_\alpha^2\phi_R^2}{8}\mathcal{T}I
\end{aligned}$$

Here (i) is because $\Sigma_{\alpha,r}\succeq 0$ and $\Sigma_{\tilde{w},k}\succeq B_k\Sigma_{u,k}B_k{}^\intercal\succeq\phi_u B_k B_k{}^\intercal$, (ii) is because $\mathcal{T}\gtrsim\psi_w\psi_\alpha^2\psi_A^2 Kr^3\log\frac{r}{\delta}\cdot\log\left(\frac{\psi_\alpha^2\psi_w\psi_A^4}{1-\rho_A^2}r\log\frac{Kr}{\delta}\right)$, (iii) is because $(1-\frac{1}{2r})^r\geq\frac{1}{2}$ for all positive integers, (iv) is because $\sum_{k\in\mathcal{K}}(T_k-r)\geq -Kr+\sum_{k\in\mathcal{K}}T_k = -Kr+\mathcal{T}-\sum_{k\notin\mathcal{K}}T_k\geq -Kr+\mathcal{T}-Kr$, and (v) is because $\mathcal{T}\gtrsim\frac{\psi_w^2\psi_\alpha^4\psi_A^4}{\phi_u^2\phi_\alpha^4\phi_R^4}Kr^3\log\frac{r}{\delta}\cdot\log\left(\frac{\psi_\alpha^2\psi_w\psi_A^4}{1-\rho_A^2}r\log\frac{Kr}{\delta}\right)$. □

Lemma C.4. *Consider the same setting as Theorem C.2. For any positive integer a, let $\{\zeta_{k,t}\in\mathbb{R}^a\}_{t=0}^{T_k}$ be a sequence of i.i.d Gaussian vectors from $\mathcal{N}(0,\Sigma_{\zeta,k})$ such that $\zeta_{k,t}$ is independent of $x_{k,t}$. Define $\psi_\zeta=\max_k\sigma_1(\Sigma_{\zeta,k})$. We make the following definition for any $b\times r$ matrix $P_{[K]}$ for any positive integer b*

$$\bar{\Sigma}_P = \sum_{k\in[K],t\in[T_k]} P_k x_{k,t}x_{k,t}{}^\intercal P_k{}^\intercal + \mathcal{T}I,\quad \psi_P=\max_{k\in[K]}\sigma_1(P_k). \tag{162}$$

Then the following holds with probability at least $1-\delta$,

$$\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}}\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}\zeta_{k,t}{}^\intercal\right\|\lesssim\tilde{U}\sqrt{\max\{r,a\}\psi_\zeta\log\frac{1}{\delta}}. \tag{163}$$

Here $\tilde{U}=\sqrt{\log\left(\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2}r\log\frac{K}{\delta}\right)}$.

Proof. From Lemma A.9, we know that the following holds for some vector $v\in\mathcal{S}^{a-1}$

$$\mathbb{P}\left(\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}}\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}\zeta_{k,t}{}^\intercal\right\|>z\right)\leq 5^a\mathbb{P}\left(\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}}\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}\zeta_{k,t}{}^\intercal v\right\|>\frac{z}{2}\right). \tag{164}$$

Notice that $\zeta_{k,t}{}^\intercal v$ are independent Gaussian varaibles from distribution $\mathcal{N}(0,v^\intercal\Sigma_{\zeta,k}v)$ for all $k\in[K],t\in[T_k]$, which is $c_1\sqrt{\psi_\zeta}$-subGaussian for some positive constant c_1. Then applying Theorem 1 in (Abbasi-yadkori et al., 2011) on sequence $\{P_k x_{k,t}\}_{t,k}$ and sequence $\{\zeta_{k,t}{}^\intercal v\}_{t,k}$, gives the following inequality

$$\mathbb{P}\left(\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}}\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}\zeta_{k,t}{}^\intercal v\right\|>\sqrt{2c_1^2\psi_\zeta\log\left(\frac{\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(\mathcal{T}I)^{-\frac{1}{2}}}{\delta}\right)}\right)\leq\delta. \tag{165}$$

Substituting the above result back gives the following inequality

$$\mathbb{P}\left(\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}}\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}\zeta_{k,t}\right\| > \sqrt{8c_1^2\psi_\zeta\log\left(\frac{\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(\mathcal{T}I)^{-\frac{1}{2}}}{\delta}\right)}\right)\leq 5^a\delta, \tag{166}$$

which implies the following inequality holds with probability at least $1-\frac{\delta}{2}$

$$\left\|\left(\bar{\Sigma}_P\right)^{-\frac{1}{2}}\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}\zeta_{k,t}\right\| \leq \sqrt{8c_1^2\psi_\zeta\log\left(\frac{2\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(\mathcal{T}I)^{-\frac{1}{2}}}{\delta}\right)+8c_1^2\psi_\zeta a\log 5}. \tag{167}$$

Now consider $\bar{\Sigma}_P=\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}x_{k,t}{}^\intercal P_k{}^\intercal+\mathcal{T}I$. Then

$$\det(\mathcal{T}I)=(\mathcal{T})^b,\quad \det(\bar{\Sigma}_P)=(\mathcal{T})^{b-r}\prod_{i=1}^{r}\left(\mathcal{T}+\lambda_i\left(\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}x_{k,t}{}^\intercal P_k{}^\intercal\right)\right). \tag{168}$$

From Lemma A.6, we have the following with probability at least $1-\delta/(2K)$ for every $k\in[K]$

$$\left\|\sum_{t\in[T_k]}x_{k,t}x_{k,t}{}^\intercal\right\|\lesssim\frac{\psi_w\psi_A^2 r}{1-\rho_A^2}T_k\log\frac{2K}{\delta}\lesssim\frac{\psi_w\psi_A^2 r}{1-\rho_A^2}T_k\log\frac{K}{\delta}. \tag{169}$$

Therefore,

$$\begin{aligned}
&\lambda_i\left(\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}x_{k,t}{}^\intercal P_k{}^\intercal\right)\\
&\leq\sum_{k\in[K]}\left\|\sum_{t\in[T_k]}P_k x_{k,t}x_{k,t}{}^\intercal P_k{}^\intercal\right\|\\
&\leq\sum_{k\in[K]}\left\|\sum_{t\in[T_k]}x_{k,t}x_{k,t}{}^\intercal\right\|\|P_k P_k{}^\intercal\|\\
&\lesssim\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2}r\mathcal{T}\log\frac{K}{\delta}
\end{aligned} \tag{170}$$

Substituting back gives the following for some positive constant c_2

$$\begin{aligned}
\det(\bar{\Sigma}_P)&=\mathcal{T}^{b-r}\prod_{i=1}^{r}\left(\mathcal{T}+\lambda_i\left(\sum_{k\in[K],t\in[T_k]}P_k x_{k,t}x_{k,t}{}^\intercal P_k{}^\intercal\right)\right)\\
&\leq\mathcal{T}^b\left(1+c_2\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2}r\log\frac{K}{\delta}\right)^r,
\end{aligned} \tag{171}$$

which gives

$$\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(\mathcal{T}I)^{-\frac{1}{2}}\leq\left(1+c_2\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2}r\log\frac{K}{\delta}\right)^{\frac{r}{2}}. \tag{172}$$

Finally, with a union bound over all above events, we get the following with probability at least $1-\delta$ from Equation 167

and 171

$$\begin{aligned}
&\left\| \left(\bar{\Sigma}_P\right)^{-\frac{1}{2}} \sum_{k\in[K], t\in[T_k]} P_k x_{k,t} \zeta_{k,t} \right\| \\
&\leq \sqrt{8c_1^2\psi_\zeta \log\left(\frac{2\det(\bar{\Sigma}_P)^{\frac{1}{2}}\det(\mathcal{T}I)^{-\frac{1}{2}}}{\delta}\right) + 8c_1^2\psi_\zeta a\log 5} \\
&\leq \sqrt{4c_1^2 r\psi_\zeta \log\left(2 + 2c_2\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2} r\log\frac{K}{\delta}\right) + 8c_1^2\psi_\zeta\log\frac{1}{\delta} + 8c_1^2\psi_\zeta a\log 5} \\
&\lesssim \sqrt{\max\{r,a\}\psi_\zeta\log\frac{1}{\delta}} \cdot \sqrt{\log\left(\frac{\psi_P^2\psi_w\psi_A^2}{1-\rho_A^2} r\log\frac{K}{\delta}\right)}.
\end{aligned} \tag{173}$$

This completes the proof. □

D. Example `Sys-Oracle` — The Ho-Kalman Algorithm (Oymak & Ozay, 2019)

We now introduce an example oracle for systems with *isotropic noise covariances*, $\mathcal{M} = (r, n, m, A, B, C, \sigma_w^2 I, \sigma_\eta^2 I)$. Before going into details, we define necessary notations. Let $\mathcal{G}_d(\mathcal{M})$ and $\mathcal{H}_d(\mathcal{M})$ as follows for any positive integer d

$$\begin{aligned}
\mathcal{G}_d(\mathcal{M}) &= \begin{bmatrix} CB & CAB & \cdots & CA^{2d-1}B \end{bmatrix} \in \mathbb{R}^{n\times 2dm}, \\
\mathcal{H}_d(\mathcal{M}) &= \begin{bmatrix} CB & CAB & \cdots & CA^{d-1}B & CA^dB \\ CAB & CA^2B & \cdots & CA^dB & CA^{d+1}B \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ CA^{d-1}B & CA^dB & \cdots & CA^{2d-2}B & CA^{2d-1}B \end{bmatrix} \in \mathbb{R}^{dn\times(d+1)m}.
\end{aligned} \tag{174}$$

Moreover, let $\mathcal{H}_d^-(\mathcal{M}) \in \mathbb{R}^{dn\times dm}$ and $\mathcal{H}_d^+(\mathcal{M}) \in \mathbb{R}^{dn\times dm}$ be the first and last dm columns of $\mathcal{H}_d(\mathcal{M})$, respectively. Now we define all necessary constants for the above $\mathcal{M}$

$$\begin{aligned}
&\phi_{\mathcal{H}}(\delta) = \sigma_{\min}\left(\mathcal{H}_d^-(\mathcal{M})\right), \quad \psi_{\mathcal{H}}(\delta) = \sigma_1\left(\mathcal{H}_d(\mathcal{M})\right), \quad d = \max\left\{r, \lceil\log\frac{1}{\delta}\rceil\right\}, \\
&\psi_B = \sigma_1(B), \quad \phi_B = \sigma_r(B), \quad \psi_C = \sigma_1(C), \quad \phi_C = \sigma_{\min}(C)
\end{aligned} \tag{175}$$

Here we assume all ψ's satisfy $\psi \geq 1$, otherwise we define ψ to be $\max\{1, \sigma_1(\cdot)\}$. Similarly, we assume all ϕ's satisfy $\phi \leq 1$, otherwise we define ϕ to be $\min\{1, \sigma_{\min}(\cdot)\}$. The algorithm is summarized as follows. We note that this algorithm is not as general as we defined in `Sys-Oracle`. To be more specific, it only works with identity noise covariances and requires to know the dimension of the latent variables.

Algorithm 4 Ho-Kalman

1: **Input:** Latent Variable Dimension r. Dataset $\{y_t\}_{t=0}^{T}, \{u_t\}_{t=0}^{T-1}$. Failure Probability δ

2: Estimating markov parameters $\widehat{\mathcal{G}} = \begin{bmatrix} \widehat{CB} & \widehat{CAB} & \cdots & \widehat{CA^{2d-1}B} \end{bmatrix}$

$$d \leftarrow \max\left\{r, \lceil \log \frac{1}{\delta} \rceil\right\}, \quad \widehat{\mathcal{G}} \leftarrow \underset{\mathcal{G} \in \mathbb{R}^{n \times 2dm}}{\arg\min} \sum_{t=2d}^{T} \left\| y_t - \mathcal{G} \begin{bmatrix} u_{t-1} \\ u_{t-2} \\ \vdots \\ u_{t-2d} \end{bmatrix} \right\|^2$$

3: Constructing Hankel matrices $\widehat{\mathcal{H}}^-, \widehat{\mathcal{H}}^+$

$$\widehat{\mathcal{H}}^- = \begin{bmatrix} \widehat{CB} & \widehat{CAB} & \cdots & \widehat{CA^{d-1}B} \\ \widehat{CAB} & \widehat{CA^2B} & \cdots & \widehat{CA^dB} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{CA^{d-1}B} & \widehat{CA^dB} & \cdots & \widehat{CA^{2d-2}B} \end{bmatrix}, \quad \widehat{\mathcal{H}}^+ = \begin{bmatrix} \widehat{CAB} & \widehat{CA^2B} & \cdots & \widehat{CA^dB} \\ \widehat{CA^2B} & \widehat{CA^3B} & \cdots & \widehat{CA^{d+1}B} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{CA^dB} & \widehat{CA^{d+1}B} & \cdots & \widehat{CA^{2d-1}B} \end{bmatrix}$$

4: Truncated (rank-r) SVD on $\widehat{\mathcal{H}}^-$

$$U_r, \Sigma_r, V_r \leftarrow \text{Top } r \text{ singular vectors \& values of } \widehat{\mathcal{H}}^-$$

5: Estimating system parameters

$$\begin{aligned} \widehat{A} &\leftarrow \left(U_r \Sigma_r^{1/2}\right)^{\dagger} \widehat{\mathcal{H}}^+ \left(\Sigma_r^{1/2} V_r{}^{\intercal}\right)^{\dagger}, \\ \widehat{B} &\leftarrow \text{First } m \text{ columns of } \Sigma_r^{1/2} V_r{}^{\intercal}, \\ \widehat{C} &\leftarrow \text{First } n \text{ rows of } U_r \Sigma_r^{1/2} \end{aligned}$$

For the paper to be self-contained, a theoretical guarantee of the above algorithm is provided.

Corollary D.1. *Consider $\mathcal{M} = (r, n, m, A, B, C, \sigma_w^2 I, \sigma_\eta^2 I)$ with any r, n, m, independent inputs $u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2 I)$ and notations in Equations (174) and (175). Suppose $\mathcal{M}$ satisfy Assumption 3.1 with constants ψ_A and ρ_A. Consider single trajectory dataset $\mathcal{Y} = \{y_t\}_{t=0}^{T}, \mathcal{U} = \{u_t\}_{t=0}^{T-1}$ from the system. If*

$$T \gtrsim \underbrace{\frac{\sigma_\eta^2 + (\sigma_w^2 + \sigma_u^2 \psi_B^2)\psi_C^2 \psi_A^6 / (1-\rho_A^2)^2}{\phi_{\mathcal{H}}^2 \sigma_u^2} \log^5(r(r+n+m)) \log^2 T \log^9 \frac{1}{\delta}}_{\kappa_1(\mathcal{M}, \mathcal{U}, \delta) \text{ in Definition 3.2}} \cdot r^3 (r+n+m). \tag{176}$$

then with probability at least $1-\delta$, Algorithm 4 outputs $\left(\widehat{A}, \widehat{B}, \widehat{C}\right)$ s.t. there exists an invertible matrix S

$$\begin{aligned} &\max\left\{\left\|S^{-1}AS - \widehat{A}\right\|, \left\|S^{-1}B - \widehat{B}\right\|, \left\|CS - \widehat{C}\right\|\right\} \\ &\lesssim \underbrace{\frac{\psi_{\mathcal{H}}}{\phi_{\mathcal{H}}^2} \frac{\sigma_\eta + \psi_A^3 \psi_C \sqrt{(\sigma_w^2 + \sigma_u^2 \psi_B^2)/(1-\rho_A^2)}}{\sigma_u} \sqrt{\log^5(r(r+n+m)) \log^2 T \log^9 \frac{1}{\delta}}}_{\kappa_2(\mathcal{M}, \mathcal{U}, \delta) \text{ in Definition 3.2}} \cdot \sqrt{\frac{r^5(r+n+m)}{T}}. \end{aligned} \tag{177}$$

Here $\phi_{\mathcal{H}} = \phi_{\mathcal{H}}(\delta)$, $\psi_{\mathcal{H}} = \psi_{\mathcal{H}}(\delta)$.

Proof. Proof of this Theorem is mainly the applications of Theorems in (Oymak & Ozay, 2019). Let $q := r + n + m$ and recall $d = \max\left\{r, \lceil \log \frac{1}{\delta} \rceil\right\}$ in Algorithm 4. For simplicity, let $\mathcal{G} = \mathcal{G}_d(\mathcal{M})$.

Step 1: $\widehat{\mathcal{G}}$ Estimation Guarantee. From Equation (176), we know that $T \gtrsim dn/(1-\rho_A^{2d+1}) \geq n(2d+1)/(1-\rho_A^{2d+1})$ and $T \gtrsim dn\log^4(dn)\log^2(T) \gtrsim (2d+1)q\log^2((4d+2)q)\log^2(2Tq)$. We then apply Theorem 3.1 of (Oymak & Ozay, 2019) and get

$$\begin{aligned}\left\|\widehat{\mathcal{G}}-\mathcal{G}\right\| \lesssim \frac{1}{\sigma_u}\left(\sigma_\eta + \sigma_w\sqrt{\left\|C\left(\sum_{i=0}^{2d-1}A^iA^{i\intercal}\right)C^\intercal\right\|} + \psi_A\|C\|\left\|A^{2d}\right\|\sqrt{\frac{(2d+1)\|\Gamma_\infty\|}{1-\rho_A^{4d+2}}}\right) \\ \cdot\sqrt{\frac{(2d+1)q\log^2((4d+2)q)\log^2(2Tq)}{T}},\end{aligned} \tag{178}$$

Here $\Gamma_\infty = \sigma_w^2\sum_{i=0}^\infty A^iA^{i\intercal} + \sigma_u^2\sum_{i=0}^\infty A^iBB^\intercal A^{i\intercal}$. From Assumption 3.1, it is clear that

$$\begin{aligned}\left\|C\left(\sum_{i=0}^{2d-1}A^iA^{i\intercal}\right)C^\intercal\right\| \leq \psi_C^2\left\|\sum_{i=0}^{2d-1}A^iA^{i\intercal}\right\| \leq \psi_C^2\sum_{i=0}^{2d-1}\left\|A^i\right\|^2 \lesssim \frac{\psi_C^2\psi_A^2}{1-\rho_A^2}, \\ \|\Gamma_\infty\| \leq \left(\sigma_w^2+\sigma_u^2\psi_B^2\right)\sum_{i=0}^\infty\left\|A^i\right\|^2 \lesssim \frac{\left(\sigma_w^2+\sigma_u^2\psi_B^2\right)\psi_A^2}{1-\rho_A^2}.\end{aligned} \tag{179}$$

Substituting back gives

$$\begin{aligned}\left\|\widehat{\mathcal{G}}-\mathcal{G}\right\| &\lesssim \frac{1}{\sigma_u}\left(\sigma_\eta + \sigma_w\sqrt{\frac{\psi_C^2\psi_A^2}{1-\rho_A^2}} + \psi_A^2\psi_C\rho_A^{2d-1}\sqrt{\frac{d}{1-\rho_A^{4d+2}}\frac{\left(\sigma_w^2+\sigma_u^2\psi_B^2\right)\psi_A^2}{1-\rho_A^2}\log\frac{24d}{\delta}}\right)\sqrt{\frac{dq\log^4(dq)\log^2 T}{T}} \\ &\lesssim \frac{1}{\sigma_u}\left(\sigma_\eta + \sigma_w\sqrt{\frac{\psi_C^2\psi_A^2}{1-\rho_A^2}} + \psi_A^2\psi_C\sqrt{\frac{\left(\sigma_w^2+\sigma_u^2\psi_B^2\right)\psi_A^2}{(1-\rho_A^2)^2}d\log\frac{d}{\delta}}\right)\sqrt{\frac{dq\log^4(dq)\log^2 T\log\frac{1}{\delta}}{T}} \\ &\overset{(i)}{\lesssim} \frac{1}{\sigma_u}\left(\sigma_\eta + \psi_A^3\psi_C\sqrt{\frac{\left(\sigma_w^2+\sigma_u^2\psi_B^2\right)}{(1-\rho_A^2)^2}}\right)\sqrt{\frac{d^2q\log^5(dq)\log^2 T\log\frac{1}{\delta}}{T}}. \\ &\lesssim \frac{1}{\sigma_u}\left(\sigma_\eta + \psi_A^3\psi_C\sqrt{\frac{\left(\sigma_w^2+\sigma_u^2\psi_B^2\right)}{(1-\rho_A^2)^2}}\right)\sqrt{\frac{r^2q\log^5(rq)\log^2 T\log^8\frac{1}{\delta}}{T}}.\end{aligned} \tag{180}$$

Here (i) is because $d \lesssim r\log(\frac{1}{\delta})$ and $\log(dq) \lesssim \log(rq)\log(\frac{1}{\delta})$. The above inequalities hold with probability at least

$$1-\left(2\exp(-(2d+1)q) + 3(2Tm)^{-\log(2Tm)\log^2((4d+2)m)} + (2d+1)\left(\exp(-100(2d+1)q) + 2\exp(-100n\log\frac{24d}{\delta})\right)\right)$$ [4]

We further simplify the probability as follows

$$\begin{aligned}&1-\left(2\exp(-(2d+1)q) + 3(2Tm)^{-\log(2Tm)\log^2((4d+2)m)} + (2d+1)\left(\exp(-100(2d+1)q) + 2\exp(-100n\log\frac{24d}{\delta})\right)\right) \\ &\overset{(i)}{\geq} 1-\frac{\delta}{4}-\left(2\exp(-2dq) + (2d+1)\left(\exp(-100dq) + 2\exp(-100n\log\frac{24d}{\delta})\right)\right) \\ &\overset{(ii)}{\geq} 1-\frac{\delta}{4}-\frac{\delta}{4}-\left((2d+1)\left(\exp(-100dq) + 2\exp(-100n\log\frac{24d}{\delta})\right)\right) \\ &\overset{(iii)}{\geq} 1-\frac{\delta}{4}-\frac{\delta}{4}-\frac{\delta}{4}-(4d+2)\exp\left(-100n\log\frac{24d}{\delta}\right) \\ &\geq 1-\frac{\delta}{4}-\frac{\delta}{4}-\frac{\delta}{4}-\frac{4d+2}{24d}\delta \\ &\geq 1-\delta.\end{aligned} \tag{181}$$

[4] The last term is because we use $\tau' = \tau\log(\frac{24d}{\delta})$ in Corollary D.3 of (Oymak & Ozay, 2019).

Here (i) is because $3(2Tm)^{-\log(2Tm)\log^2(4dm)} \le 3(2\cdot\frac{1}{\delta})^{-7} \le \frac{\delta}{4}$, where the first inequality is because $\log^2(4dm) \ge \log^2(16)$. (ii) is because $d \ge \log\frac{1}{\delta}$ and $2\exp(-2dq) \le 2\exp(-6\log(1/\delta)) \le 2\delta^6 \le \delta/4$ for $\delta \in (0, e^{-1})$ and (iii) is because $(2d+1)\exp(-100dq) \le (2d+1)\exp(-2d-1)\exp(-90dq) \le \exp(-270d) \le \delta/4$ for $\delta \in (0, e^{-1})$.

Step 2: Parameter Recovery Guarantee. From the above result, we get the following for some constant c_1

$$\sqrt{d}\left\|\widehat{\mathcal{G}}-\mathcal{G}\right\| \le \frac{c_1}{\sigma_u}\left(\sigma_\eta + \psi_A^3\psi_C\sqrt{\frac{(\sigma_w^2+\sigma_u^2\psi_B^2)}{(1-\rho_A^2)^2}}\right)\sqrt{\frac{r^2 q\log^5(rq)\log^2 T\log^8\frac{1}{\delta}}{T}}\cdot\sqrt{r\log\frac{1}{\delta}} \le \frac{\phi_{\mathcal{H}}}{4}. \tag{182}$$

Here the last inequality is because of Equation 176. Then we apply Corollary 5.4[5] from (Oymak & Ozay, 2019) and get the following for some balanced realization $\bar{A},\bar{B},\bar{C}$ and some unitary matrix U

$$\begin{aligned}
&\max\left\{\left\|\bar{A}-U^\intercal\widehat{A}U\right\|,\left\|\bar{B}-U^\intercal\widehat{B}\right\|,\left\|\bar{C}-\widehat{C}U\right\|\right\}\\
\lesssim\ &\frac{\psi_{\mathcal{H}}}{\phi_{\mathcal{H}}^2}\sqrt{r^2 d}\left\|\widehat{\mathcal{G}}-\mathcal{G}\right\|\\
\lesssim\ &\frac{\psi_{\mathcal{H}}}{\phi_{\mathcal{H}}^2}\frac{\sigma_\eta+\psi_A^3\psi_C\sqrt{(\sigma_w^2+\sigma_u^2\psi_B^2)/(1-\rho_A^2)^2}}{\sigma_u}\sqrt{\frac{r^5 q\log^5(rq)\log^2 T\log^9\frac{1}{\delta}}{T}}.
\end{aligned} \tag{183}$$

Since $\bar{A},\bar{B},\bar{C}$ is a balanced realization, there exist some invertible matrix S s.t.

$$\begin{aligned}
&\max\left\{\left\|S^{-1}AS-\widehat{A}\right\|,\left\|S^{-1}B-\widehat{B}\right\|,\left\|CS-\widehat{C}\right\|\right\}\\
\lesssim\ &\frac{\psi_{\mathcal{H}}}{\phi_{\mathcal{H}}^2}\frac{\sigma_\eta+\psi_A^3\psi_C\sqrt{(\sigma_w^2+\sigma_u^2\psi_B^2)/(1-\rho_A^2)^2}}{\sigma_u}\sqrt{\frac{r^5 q\log^5(rq)\log^2 T\log^9\frac{1}{\delta}}{T}}.
\end{aligned} \tag{184}$$

□

D.1. Upper Bounds for `HD-SYSID` with Oracle (Algorithm 4)

Corollary D.2 (Corollary 3.6 Restated). *Consider $\mathcal{M}=(r,n,m,A,B,C,\sigma_w^2 I,\sigma_\eta^2 I)$ and datasets $\mathcal{D}_1=\mathcal{U}_1\cup\mathcal{Y}_1,\mathcal{D}_2=\mathcal{U}_2\cup\mathcal{Y}_2$ (with length T_1,T_2 respectively) where the inputs are sampled independently from $\mathcal{N}(0,\sigma_u^2 I)$. Consider constants defined for $\mathcal{M}$ in Appendix A and in Equation (175). Suppose $\mathcal{M}$ satisfies Assumption 3.1 with constants ψ_A and ρ_A. If*

$$T_1 \gtrsim \tilde{\kappa}_3\cdot n^2 r^2,\quad T_2 \gtrsim \tilde{\kappa}_1\cdot r^3(r+m), \tag{185}$$

then $(\widehat{A},\widehat{B},\widehat{C})$ from Algorithm 1 satisfy the following for some invertible matrix S with probability at least $1-\delta$

$$\max\left\{\left\|S^{-1}AS-\widehat{A}\right\|,\left\|S^{-1}B-\widehat{B}\right\|,\left\|CS-\widehat{C}\right\|\right\} \lesssim \tilde{\kappa}_4\cdot\sqrt{\frac{n}{T_1}}\left\|\widehat{C}\right\|+\tilde{\kappa}_2\cdot\sqrt{\frac{r^5(r+m)}{T_2}}. \tag{186}$$

$\kappa_{[4]}$ are detailed below. All of them are problem-related constants independent of system dimensions modulo log factors.

$$\begin{aligned}
\tilde{\kappa}_1 &= \frac{\psi_\eta+\psi_w\psi_C^2\psi_A^6/(1-\rho_A^2)^2}{\phi_{\mathcal{H}}^2\sigma_u^2}\log^5(r(r+m))\log^2 T\log^9\frac{1}{\delta},\\
\tilde{\kappa}_2 &= \frac{\psi_{\mathcal{H}}}{\phi_{\mathcal{H}}^2}\frac{\sqrt{\psi_\eta}+\psi_A^3\psi_C\sqrt{\psi_w^2/(1-\rho_A^2)}}{\sigma_u}\sqrt{\log^5(r(r+m))\log^2 T\log^9\frac{1}{\delta}},\\
\tilde{\kappa}_3 &= \frac{1}{\phi_{\mathcal{H}}^2}\max\left\{\tilde{\kappa}_4^2\frac{\psi_A^4\psi_B^2\psi_C^2}{(1-\rho_A^2)\phi_O^2},\ \left(\frac{\psi_w\psi_\eta\psi_C^2\psi_A^2}{\phi_u\phi_C^2\phi_R^2(1-\rho_A^2)}\right)^4\log^4(r)\log^8\left(\frac{1}{\delta}\right)\right\},\\
\tilde{\kappa}_4 &= \frac{\psi_\eta}{\phi_u\phi_C^2\phi_R^2}\sqrt{\log\frac{1}{\delta}}.
\end{aligned} \tag{187}$$

[5] The following inequality can be obtained by substituting all $r\left\|L-\widehat{L}\right\|$ by $\frac{r^2\left\|L-\widehat{L}\right\|^2}{\sigma_{\min}(L)}$ in the original result. This is valid because the original proof used $\frac{2}{\sqrt{2}-1}\frac{\left\|L-\widehat{L}\right\|_F^2}{\sigma_{\min}(L)} \le 5r\left\|L-\widehat{L}\right\|$ in Lemma B.1 of (Oymak & Ozay, 2019)

Here $\phi_{\mathcal{H}} = \phi_{\mathcal{H}}(\delta)$ and $\psi_{\mathcal{H}} = \psi_{\mathcal{H}}(\delta)$ are defined in Equation (175).

Proof. From Equation (185), $T_1 \gtrsim \tilde{\kappa}_3 n^2 r^2$, which satisfies the condition of Lemma A.2. We apply Lemma A.2 on $(\mathcal{D}_1, \Sigma_\eta)$ and get the following with probability at least $1 - \frac{\delta}{2}$ for some constant c_1

$$\left\| \widehat{\Phi}_C^{\perp\intercal} \Phi_C \right\| \le c_1 \tilde{\kappa}_4 \sqrt{\frac{n}{T_1}} \le \frac{\phi_{\mathcal{H}}(1-\rho_A^2)}{8\psi_A^2 \psi_B \psi_C}. \tag{188}$$

Then from Lemma D.3 we know $\kappa_1\left(\widehat{\mathcal{M}}, \mathcal{U}_2, \delta\right) \le 4\kappa_1\left(\mathcal{M}, \mathcal{U}_2, \delta\right), \quad \kappa_2\left(\widehat{\mathcal{M}}, \mathcal{U}_2, \delta\right) \le 4\kappa_2\left(\mathcal{M}, \mathcal{U}_2, \delta\right)$. Here $\widehat{\mathcal{M}} = (r, \text{rank}(\widehat{\Phi}_C{}^\intercal C), m, A, B, \widehat{\Phi}_C{}^\intercal C, \sigma_w^2 I, \sigma_\eta^2 I)$. Moreover, from their definitions, we know that $\kappa_1(\mathcal{M}, \mathcal{U}_2, \delta) \le \tilde{\kappa}_1$ and $\kappa_2(\mathcal{M}, \mathcal{U}_2, \delta) \le \tilde{\kappa}_2$. Combining with Equation (185) gives

$$\begin{aligned} T_2 &\gtrsim \tilde{\kappa}_1 r^3 (r+m) \ge \kappa_1\left(\mathcal{M}, \mathcal{U}_2, \delta\right) r^3 (r+m) \\ &\gtrsim \kappa_1\left(\widehat{\mathcal{M}}, \mathcal{U}_2, \delta\right) r^3 (r + \text{rank}(\widehat{\Phi}_C{}^\intercal C) + m). \end{aligned} \tag{189}$$

From the proof of Theorem 3.4 (step 1), we know that we can apply Algorithm 4 on $\widehat{\mathcal{M}}$. Applying Algorithm 4 gives the following with probability at least $1 - \frac{\delta}{2}$ from Corollary D.1

$$\begin{aligned} &\max\left\{ \left\| S^{-1} A S - \widehat{A} \right\|, \left\| S^{-1} B - \widehat{B} \right\|, \left\| \widehat{\Phi}_C{}^\intercal C S - \tilde{C} \right\| \right\} \\ \lesssim\ & \kappa_2(\widehat{\mathcal{M}}, \mathcal{U}_2, \delta) \cdot \sqrt{\frac{r^5 (r + \text{rank}(\widehat{\Phi}_C{}^\intercal C) + m)}{T_2}} \\ \lesssim\ & \kappa_2(\mathcal{M}, \mathcal{U}_2, \delta) \cdot \sqrt{\frac{r^5 (r+m)}{T_2}} \\ \lesssim\ & \tilde{\kappa}_2 \cdot \sqrt{\frac{r^5 (r+m)}{T_2}} \end{aligned} \tag{190}$$

Then from the proof of Theorem 3.4 (step 2), we know that

$$\begin{aligned} &\max\left\{ \left\| S^{-1} A S - \widehat{A} \right\|, \left\| S^{-1} B - \widehat{B} \right\|, \left\| C S - \widehat{C} \right\| \right\} \\ \lesssim\ & \tilde{\kappa}_2 \cdot \sqrt{\frac{r^5 (r+m)}{T_2}} + \tilde{\kappa}_4 \cdot \sqrt{\frac{n}{T_1}} \left\| \widehat{C} \right\|. \end{aligned} \tag{191}$$

□

D.2. Other Lemmas for κ_1 and κ_2

Lemma D.3. *Fix any δ. Consider system $\mathcal{M} = (r, n, m, A, B, C, \sigma_w^2 I, \sigma_\eta^2)$ and independent inputs $\mathcal{U} = \{u_t\}_{t=0}^{T-1}$ with $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$. Let $\widehat{\mathcal{M}} = (r, \text{rank}(\widehat{\Phi}_C{}^\intercal C), m, A, B, \widehat{\Phi}_C{}^\intercal C, \sigma_w^2 I, \sigma_\eta^2 I)$, where $\widehat{\Phi}_C$ is a orthonormal matrix satisfying*

$$\left\| \left(\widehat{\Phi}_C^{\perp}\right)^\intercal \Phi_C \right\| \le \Delta_\Phi \le \frac{\phi_{\mathcal{H}}(1-\rho_A^2)}{8\psi_A^2 \psi_B \psi_C}. \tag{192}$$

Let $\kappa_1(\cdot, \mathcal{U}, \delta), \kappa_2(\cdot, \mathcal{U}, \delta)$ be defined as in Equations (176) and (177). Then we have

$$\kappa_1\left(\widehat{\mathcal{M}}, \mathcal{U}, \delta\right) \le 4\kappa_1\left(\mathcal{M}, \mathcal{U}, \delta\right), \quad \kappa_2\left(\widehat{\mathcal{M}}, \mathcal{U}, \delta\right) \le 4\kappa_2\left(\mathcal{M}, \mathcal{U}, \delta\right) \tag{193}$$

Proof. For system $\mathcal{M}$, we list all related parameters as follows[6].

$$\sigma_w^2, \sigma_\eta^2, \rho_A, \psi_A, \psi_B, \psi_C, \phi_{\mathcal{H}}, \psi_{\mathcal{H}}. \tag{194}$$

[6]Here we omit σ_u since it doesn't change between the two systems.

For the projected system $\widehat{\mathcal{M}}$, we denote corresponding parameters as

$$\widehat{\sigma}_w^2, \widehat{\sigma}_\eta^2, \widehat{\rho}_A, \widehat{\psi}_A, \widehat{\psi}_B, \widehat{\psi}_C, \widehat{\phi}_{\mathcal{H}}, \widehat{\psi}_{\mathcal{H}}. \tag{195}$$

It is clear that $\widehat{\sigma}_w^2, \widehat{\sigma}_\eta^2, \widehat{\rho}_A, \widehat{\psi}_A, \widehat{\psi}_B$ remain unchanged. We only need to consider parameters $\widehat{\psi}_C, \widehat{\phi}_{\mathcal{H}}, \widehat{\psi}_{\mathcal{H}}$. We know that

$$\widehat{\psi}_C = \left\|\widehat{\Phi}_C{}^\intercal C\right\| \leq \left\|\widehat{\Phi}_C\right\| \|C\| = \psi_C. \tag{196}$$

Moreover, for $d = \max\{r, \lceil \frac{1}{\delta} \rceil\}$, we have

$$\left\|\mathcal{H}_d^+(\widehat{\mathcal{M}})\right\| = \left\|\text{diag}\left(\widehat{\Phi}_C{}^\intercal, \ldots, \widehat{\Phi}_C{}^\intercal\right)\mathcal{H}_d^+(\mathcal{M})\right\| \leq \left\|\widehat{\Phi}_C\right\| \left\|\mathcal{H}_d^+(\mathcal{M})\right\| = \psi_{\mathcal{H}}. \tag{197}$$

Therefore, $\widehat{\psi}_{\mathcal{H}} \leq \psi_{\mathcal{H}}$. For $\widehat{\phi}_{\mathcal{H}}$, we know from Lemma D.4 that $\widehat{\phi}_{\mathcal{H}} \geq \phi_{\mathcal{H}}/2$ for $\Delta_\Phi \leq \frac{\phi_{\mathcal{H}}(1-\rho_A^2)}{8\psi_A^2\psi_B\psi_C}$.

These bounds imply that

$$\kappa_2\left(\mathcal{M}, \mathcal{U}, \delta\right) \geq \frac{1}{4}\kappa_2\left(\widehat{\mathcal{M}}, \mathcal{U}, \delta\right), \quad \kappa_1\left(\mathcal{M}, \mathcal{U}, \delta\right) \geq \frac{1}{4}\kappa_1\left(\widehat{\mathcal{M}}, \mathcal{U}, \delta\right). \tag{198}$$

□

Lemma D.4. *Consider the same setting as Lemma D.3. Consider constants $\phi_{\mathcal{H}}$ and $\widehat{\phi}_{\mathcal{H}}$ of $\mathcal{M}$ and $\widehat{\mathcal{M}}$. Then if $\Delta_\Phi \leq \frac{\phi_{\mathcal{H}}(1-\rho_A^2)}{8\psi_A^2\psi_B\psi_C}$, then $\widehat{\phi}_{\mathcal{H}} \geq \frac{\phi_{\mathcal{H}}}{2}$.*

Proof. Recall that $d =$ From the definition of $\widehat{\phi}_{\mathcal{H}}$, we know that

$$\begin{aligned}
&\sigma_{\min}\left(\text{diag}\left(\widehat{\Phi}_C{}^\intercal, \ldots \widehat{\Phi}_C{}^\intercal\right)\mathcal{H}_d^-(\mathcal{M})\right) = \sigma_{\min}\left(\text{diag}\left(\widehat{\Phi}_C\widehat{\Phi}_C{}^\intercal, \ldots, \widehat{\Phi}_C\widehat{\Phi}_C{}^\intercal\right)\mathcal{H}_d^-(\mathcal{M})\right) \\
&= \sigma_{\min}\left(\mathcal{H}_d^-(\mathcal{M}) - \text{diag}\left(\widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal, \ldots, \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal\right)\mathcal{H}_d^-(\mathcal{M})\right) \\
&\geq \sigma_{\min}\left(\mathcal{H}_d^-(\mathcal{M})\right) - \sigma_1\left(\text{diag}\left(\widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal, \ldots, \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal\right)\mathcal{H}_d^-(\mathcal{M})\right)
\end{aligned} \tag{199}$$

Notice that

$$\text{diag}\left(\widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal, \ldots, \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal\right)\mathcal{H}_d^-(\mathcal{M}) = \begin{bmatrix} \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal C \\ \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA \\ \vdots \\ \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA^{d-1} \end{bmatrix}\begin{bmatrix} B & AB & \cdots & A^{d-1}B \end{bmatrix} \tag{200}$$

From Equation 13, we directly know the following

$$\left\|\begin{bmatrix} \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal C \\ \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA \\ \vdots \\ \widehat{\Phi}_C^\perp\widehat{\Phi}_C^\perp{}^\intercal CA^{d-1} \end{bmatrix}\right\| \leq \frac{2\psi_A\psi_C}{\sqrt{1-\rho_A^2}}\Delta_\Phi. \tag{201}$$

Moreover,

$$\left\|\begin{bmatrix} B & AB & \cdots A^{d-1}B \end{bmatrix}\right\| \leq \sqrt{\sum_{i=0}^{d-1}\|A^iBB^\intercal A^{i\intercal}\|} \leq \frac{2\psi_A\psi_B}{\sqrt{1-\rho_A^2}}. \tag{202}$$

Substituting back gives

$$\widehat{\phi}_{\mathcal{H}} \geq \phi_{\mathcal{H}} - 4\frac{\psi_A^2\psi_B\psi_C}{(1-\rho_A^2)}\Delta_\Phi. \tag{203}$$

Since $\Delta_\Phi \leq \frac{\phi_{\mathcal{H}}(1-\rho_A^2)}{8\psi_A^2\psi_B\psi_C}$, we know $\widehat{\phi}_{\mathcal{H}} \geq \phi_{\mathcal{H}}/2$. □