Joint Learning of Linear Time-Invariant Dynamical Systems

Aditya Modi [a], Mohamad Kazem Shirani Faradonbeh [b], Ambuj Tewari [c], George Michailidis [d]

[a] *Microsoft, Mountain View, CA-94043, USA*

[b] *Department of Mathematics, Southern Methodist University, Dallas, TX-75275, USA*

[c] *Department of Statistics, University of Michigan, Ann Arbor, MI-48109, USA*

[d] *Department of Statistics, University of California, Los Angeles, CA-90095, USA*

Abstract

Linear time-invariant systems are very popular models in system theory and applications. A fundamental problem in system identification that remains rather unaddressed in extant literature is to leverage commonalities amongst *related* systems to estimate their transition matrices more accurately. To address this problem, we investigate methods for jointly estimating the transition matrices of multiple systems. It is assumed that the transition matrices are *unknown* linear functions of some *unknown* shared basis matrices. We establish finite-time estimation error rates that fully reflect the roles of trajectory lengths, dimension, and number of systems under consideration. The presented results are fairly general and show the significant gains that can be achieved by pooling data across systems, in comparison to learning each system individually. Further, they are shown to be *robust* against moderate model misspecifications. To obtain the results, we develop novel techniques that are of independent interest and are applicable to similar problems. They include tightly bounding estimation errors in terms of the eigen-structures of transition matrices, establishing sharp high probability bounds for singular values of dependent random matrices, and capturing effects of misspecified transition matrices as the systems evolve over time.

Key words: Multiple Linear Systems; Data Sharing; Finite Time Identification; Autoregressive Processes; Joint Estimation.

1 Introduction

The problem of identifying the transition matrices in linear time-invariant (LTI) systems has been extensively studied in the literature [8,26,29]. Recent papers establish finite-time rates for accurately learning the dynamics in various online and offline settings [16,36,39]. Notably, existing results are established when the goal is to identify the transition matrix of *a single* system.

However, in many application areas of LTI systems, one observes state trajectories of *multiple* dynamical systems. So, in order to be able to efficiently use the full data of all state trajectories and utilize the possible commonalities the systems share, we need to estimate the transition matrices of all systems *jointly*. The range of applications is remarkably extensive, including dynamics of economic indicators in US states [35,40,42], flight dynamics of airplanes at different altitudes [6], drivers of gene expressions across related species [5,19], time series data of multiple subjects that suffer from the same disease [38,41], and commonalities among multiple subsystems in control engineering [43].

In all these settings, there are strong similarities in the dynamics of the systems, which are unknown and need to be learned from the data. Hence, it becomes of interest to develop a joint learning strategy for the system parameters, by pooling the data of the underlying systems together and learn the *unknown* similarities in their dynamics. In particular, this strategy is of extra importance in settings wherein the available data is limited, for example when the state trajectories are short or the dimensions are not small.

In general, joint learning (also referred to as multitask learning) approaches aim to study estimation methods subject to *unknown* similarities across the data gener-

⋆ This paper was not presented at any IFAC meeting.

Email addresses: admodi@umich.edu (Aditya Modi), mkshiranyf@gmail.com (Mohamad Kazem Shirani Faradonbeh), tewaria@umich.edu (Ambuj Tewari), gmichail@ucla.edu (George Michailidis).

ation mechanisms. Joint learning methods are studied in supervised learning and online settings [10,4,32,33,3]. Their theoretical analyses are obtained rely on a number of technical assumptions regarding the data, including independence, identical distributions, boundedness, richness, and isotropy.

However, for the problem of joint learning of dynamical systems, additional technical challenges are present. First, the observations are temporally dependent. Second, the number of unknown parameters is the *square* of the dimension of the system, which impacts the learning accuracy. Third, since in many applications the dynamics matrices of the underlying LTI systems might possess eigenvalues of (almost) unit magnitude, conventional approaches for dependent data (e.g., mixing) inapplicable [16,36,39]. Fourth, the spectral properties of the transition matrices play a critical role on the magnitude of the estimation errors. Technically, the state vectors of the systems can scale exponentially with the multiplicities of the eigenvalues of the transition matrices (which can be as large as the dimension). Accordingly, novel techniques are required for considering all important factors and new analytical tools are needed for establishing useful rates for estimation error. Further details and technical discussions are provided in Section 3.

We focus on a commonly used setting for joint learning that involves *two layers of uncertainties*. It lets all systems share a common basis, while coefficients of the linear combinations are *idiosyncratic* for each system. Such settings are adopted in multitask regression, linear bandits, and Markov decision processes [14,22,31,44]. From another point of view, this assumption that the system transition matrices are *unknown* linear combinations of *unknown* basis matrices can be considered as a first-order approximation for unknown non-linear dynamical systems [27,30]. Further, these compound layers of uncertainties subsume a recently studied case for mixtures of LTI systems where under additional assumptions such as exponential stability and distinguishable transition matrices, joint learning from unlabeled state trajectories outperforms individual system identification [11].

The main contributions of this work can be summarized as follows. We provide novel finite-time estimation error bounds for jointly learning multiple systems, and establish that pooling the data of state trajectories can drastically decrease the estimation error. Our analysis also presents effects of different parameters on estimation accuracy, including dimension, spectral radius, eigenvalues multiplicity, tail properties of the noise processes, and heterogeneity among the systems. Further, we study learning accuracy in the presence of model misspecifications and show that the developed joint estimator can robustly handle moderate violations of the shared structure in the dynamics matrices.

In order to obtain the results, we employ advanced techniques from random matrix theory and prove sharp concentration results for sums of multiple dependent random matrices. Then, we establish tight and simultaneous high-probability confidence bounds for the sample covariance matrices of the systems under study. The analyses precisely characterize the dependence of the presented bounds on the spectral properties of the transition matrices, condition numbers, and block-sizes in the Jordan decomposition. Further, to address the issue of temporal dependence, we extend self-normalized martingale bounds to multiple matrix-valued martingales, subject to shared structures across the systems. We also present a robustness result by showing that the error due to misspecifications can be effectively controlled.

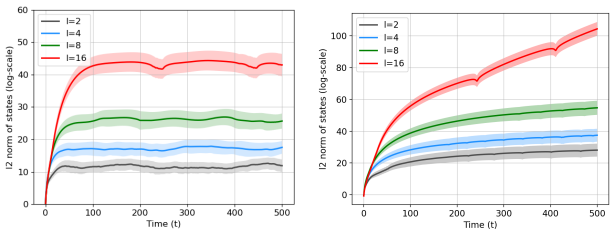
The remainder of the paper is organized as follows. The problem is formulated in Section 2. In Section 3, we describe the joint-learning procedure, study the per-system estimation error, and provide the roles of various key quantities. Then, investigation of robustness to model misspecification and the impact of violating the shared structure are discussed in Section 4. We provide numerical illustrations for joint learning in Section 5 and present the proofs of our results in the subsequent sections. Finally, the paper is concluded in Section 10.

Notation. For a matrix A, A' denotes the transpose of A. For square matrices, we use the following order of eigenvalues in terms of their magnitudes: $|\lambda_{\max}(A)| = |\lambda_1(A)| \geq |\lambda_2(A)| \geq \cdots \geq |\lambda_d(A)| = |\lambda_{\min}(A)|$. For singular values, we employ $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$. For any vector $v \in \mathbb{C}^d$, let $\|v\|_p$ denote its ℓ_p norm. We use $|||\cdot|||_{\gamma\to\beta}$ to denote the matrix operator-norm for $\beta, \gamma \in [1,\infty]$ and $A \in \mathbb{C}^{d_1\times d_2}$: $|||A|||_{\gamma\to\beta} = \sup_{v\neq\mathbf{0}} \|Av\|_\beta / \|v\|_\gamma$. When $\gamma = \beta$, we simply write $|||A|||_\beta$. For functions $f, g : \mathcal{X} \to \mathbb{R}$, we write $f \lesssim g$, if $f(x) \leq cg(x)$ for a universal constant $c > 0$. Similarly, we use $f = O(g)$ and $f = \Omega(h)$, if $0 \leq f(n) \leq c_1 g(n)$ for all $n \geq n_1$, and $0 \leq c_2 h(n) \leq f(n)$ for all $n \geq n_2$, respectively, where c_1, c_2, n_1, n_2 are large enough constants. For any two matrices of the same dimensions, we define the inner product $\langle A, B\rangle = \mathrm{tr}\,(A'B)$. Then, the Frobenius norm becomes $\|A\|_F = \sqrt{\langle A, A\rangle}$. The sigma-field generated by $X_1, X_2, \ldots, X_n$ is denoted by $\sigma(X_1, X_2, \ldots X_n)$. We denote the i-th component of the vector $x \in \mathbb{R}^d$ by $x[i]$. Finally, for $n \in \mathbb{N}$, the shorthand $[n]$ is the set $\{1, 2, \ldots, n\}$.

2 Problem Formulation

Our main goal is to study the rates of jointly learning dynamics of multiple LTI systems. Data consists of state trajectories of length T from M different systems. Specifically, for $m \in [M]$ and $t = 0, 1, \ldots, T$, let $x_m(t) \in \mathbb{R}^d$ denote the state of the m-th system, that evolves according to the Vector Auto-Regressive (VAR) process

$$x_m(t+1) = A_m x_m(t) + \eta_m(t+1). \tag{1}$$



(a) $|\lambda_1(A)| < 1$ (b) $|\lambda_1(A)| \approx 1$

Fig. 1. Logarithm of the magnitude of the state vectors vs. time, for different block-sizes in the Jordan forms of the transition matrices, which is denoted by l in (4). The exponential scaling of the state vectors with l can be seen in both plots.

Above, $A_m \in \mathbb{R}^{d\times d}$ denotes the true unknown transition matrix of the m-th system and $\eta_m(t+1)$ is a mean zero noise. For succinctness, we use Θ^* to denote the set of all M transition matrices $\{A_m\}_{m=1}^M$. The transition matrices are *related* as will be specified in Assumption 3.

Note that the above setting includes systems with longer memories. Indeed, if the states $\tilde{x}_m(t) \in \mathbb{R}^{\tilde{d}}$ obey

$$\tilde{x}_m(t) = B_{m,1}\tilde{x}_m(t-1) + \cdots + B_{m,q}\tilde{x}_m(t-q) + \eta_m(t),$$

then, by concatenating $\tilde{x}_m(t-1), \cdots, \tilde{x}_m(t-q)$ in one larger vector $x_m(t-1)$, the new state dynamics is (1), for $d = q\tilde{d}$ and $A_m = \begin{bmatrix} B_{m,1} \cdots B_{m,q-1} & B_{m,q} \\ I_{(q-1)\tilde{d}} & 0 \end{bmatrix}$.

We assume that the system states do not explode in the sense that the spectral radius of the transition matrix A_m can be *slightly* larger than one. This is required for the systems to be able to operate for a reasonable time length [25,15]. Note that this assumption still lets the state vectors grow with time, as shown in Figure 1.

Assumption 1 *For all $m \in [M]$, we have $|\lambda_1(A_m)| \leq 1 + \rho/T$, where $\rho > 0$ is a fixed constant.*

In addition to the magnitudes of the eigenvalues, further properties of the transition matrices heavily determine the temporal evolution of the systems. A very important one is the size of the largest block in the Jordan decomposition of A_m, which will be rigorously defined shortly. This quantity is denoted by l in (4). The impact of l on the state trajectories is illustrated in Figure 1, wherein we plot the *logarithm* of the magnitude of state vectors for linear systems of dimension $d = 32$. The upper plot depicts state magnitude for stable systems and for blocks of the size $l = 2, 4, 8, 16$ in the Jordan decomposition of the transition matrices. It illustrates that the state vector scales *exponentially* with l. Note that l can be as large as the system dimension d.

Moreover, the case of transition matrices with eigenvalues close to (or exactly on) the unit circle is provided in the lower panel in Figure 1. It illustrates that the state vectors grow polynomially with time, whereas the scaling with the block-size l is exponential. Therefore, in design and analysis of joint learning methods, one needs to carefully consider the effects of l and $|\lambda_1(A_m)|$.

Next, we express the probabilistic properties of the stochastic processes driving the dynamical systems. Let $\mathcal{F}_t = \sigma(x_{1:M}(0), \eta_{1:M}(1), \ldots, \eta_{1:M}(t))$ denote the filtration generated by the initial state and the sequence of noise vectors. Based on this, we adopt the following ubiquitous setting that lets the noise process $\{\eta_m(t)\}_{t=1}^\infty$ be a sub-Gaussian martingale difference sequence. Note that by definition, $\eta_m(t)$ is $\mathcal{F}_t$-measurable.

Assumption 2 *For all systems $m \in [M]$, we have $\mathbb{E}[\eta_m(t)|\mathcal{F}_{t-1}] = \mathbf{0}$ and $\mathbb{E}[\eta_m(t)\eta_m(t)'|\mathcal{F}_{t-1}] = C$. Further, $\eta_m(t)$ is sub-Gaussian; for all $\lambda \in \mathbb{R}^d$:*

$$\mathbb{E}\left[\exp\langle\lambda, \eta_m(t)\rangle\,|\mathcal{F}_{t-1}\right] \leq \exp\left(\|\lambda\|^2\sigma^2/2\right).$$

Henceforth, we denote $c^2 = \max(\sigma^2, \lambda_{\max}(C))$.

The above assumption is widely-used in the finite-sample analysis of statistical learning methods [1,17]. It includes normally distributed martingale difference sequences, for which Assumption 2 is satisfied with $\sigma^2 = \lambda_{\max}(C)$. Moreover, if the coordinates of $\eta_m(t)$ are (conditionally) independent and have sub-Gaussian distributions with constant σ_i, it suffices to let $\sigma^2 = \sum_{i=1}^d \sigma_i^2$. We let a common noise covariance matrix for the ease of expression. However, the results simply generalize to covariance matrices that vary with time and across the systems, by appropriately replacing upper- and lower-bounds of the matrices [16,36,39].

For a single system $m \in [M]$, its underlying transition matrices A_m can be *individually* learned from its own state trajectory data by using the least squares estimator [16,36]. We are interested in jointly learning the transition matrices of all M systems under the assumption that they share the following common structure.

Assumption 3 (Shared Basis) *Each transition matrix A_m can be expressed as*

$$A_m = \sum_{i=1}^{k} \beta_m^*[i] W_i^*, \tag{2}$$

where $\{W_i^*\}_{i=1}^k$ are common $d \times d$ matrices and $\beta_m^* \in \mathbb{R}^k$ contains the idiosyncratic coefficients for system m.

This assumption is commonly-used in the literature of jointly learning multiple parameters [14,44]. Intuitively, it states that each system evolves by combining the effects of k systems. These k unknown systems behind the

scene are shared by all systems $m \in [M]$, the weight of each of which is reflected by the idiosyncratic coefficients that are collected in β_m^* for system m. Thereby, the model allows for a rich heterogeneity across systems.

The main goal is to estimate $\Theta^* = \{A_m\}_{m=1}^M$ by observing $x_m(t)$ for $1 \leq m \leq M$ and $0 \leq t \leq T$. To that end, we need a reliable joint estimator that can leverage the unknown shared structure to learn from the state trajectories more accurately than individual estimations of the dynamics. Importantly, to theoretically analyze effects of all quantities on the estimation error, we encounter some challenges for joint learning of multiple systems that do *not* appear in single-system identification.

Technically, the least-squares estimate of the transition matrix of a single system admits a closed form that lets the main challenge of the analysis be concentration of the sample covariance matrix of the state vectors. However, since closed forms are not achievable for joint-estimators, learning accuracy cannot be directly analyzed. To address this, we first bound the prediction error and then use that for bounding the estimation error. To establish the former, after appropriately decomposing the joint prediction error, we study its scaling with the trajectory-length and dimension, as well as the trade-offs between the number of systems, number of basis matrices, and magnitudes of the state vectors. Then, we deconvolve the prediction error to the estimation error and the sample covariance matrices, and show useful bounds that can tightly relate the largest and smallest eigenvalues of the sample covariance matrices across all systems. Notably, this step that is not required in single-system identification is based on novel probabilistic analysis for dependent random matrices.

In the sequel, we introduce a joint estimator for utilizing the structure in Assumption 3 and analyze its accuracy. Then, in Section 4 we consider violations of the structure in (2) and establish robustness guarantees.

3 Joint Learning of LTI Systems

In this section, we propose an estimator for jointly learning the M transition matrices. Then, we establish that the estimation error decays at a significantly faster rate than competing procedures that learn each transition matrix A_m separately by using only the data trajectory of system m.

Based on the parameterization in (2), we solve for $\widehat{\mathbf{W}} = \{\widehat{W}_i\}_{i=1}^k$ and $\widehat{B} = \left[\widehat{\beta}_1 | \widehat{\beta}_2 | \cdots \widehat{\beta}_M\right] \in \mathbb{R}^{k \times M}$, as follows:

$$\widehat{\mathbf{W}}, \widehat{B} \coloneqq \underset{\mathbf{W}, B}{\operatorname{argmin}} \mathcal{L}(\Theta^*, \mathbf{W}, B), \tag{3}$$

where $\mathcal{L}(\Theta^*, \mathbf{W}, B)$ is the averaged squared loss across all M systems:

$$\frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T} \left\| x_m(t+1) - \left(\sum_{i=1}^{k} \beta_m[i] W_i \right) x_m(t) \right\|_2^2.$$

In the analysis, we assume that one can approximately find the minimizer in (3). Although the loss function in (3) is non-convex, thanks to its structure, computationally fast methods for accurately finding the minimizer are applicable. Specifically, the loss function in (3) is quadratic and the non-convexity is the bilinear dependence on $(\mathbf{W}, B)$. The optimization in (3) is of the form of explicit rank-constrained representations [9]. For such problems, it has been shown under mild conditions that gradient descent converges to a low-rank minimizer at a linear rate [46]. Moreover, it is known that methods such as stochastic gradient descent have global convergence, and these bilinear non-convexities do not lead to any spurious local minima [20]. In addition, since the loss function is biconvex in $\mathbf{W}$ and B, alternating minimization techniques converge to global optima, under standard assumptions [23]. Nonetheless, note that a near-optimal minimum for the objective function is sufficient, and we only need to estimate the product WB accurately instead of recovering both W and B. More specifically, the error of the joint estimator in (3) degrades gracefully in the presence of moderate optimization errors. For instance, suppose that the optimization problem is solved up to an error of ϵ from a global optimum. It can be shown that an additional term of magnitude $O\left(\epsilon / \lambda_{\min}(C)\right)$ arises in the estimation error, due to this optimization error. Numerical experiments in Section 5 illustrate the implementation of (3).

In the sequel, we provide key results for the joint estimator in (3) and establish the high probability decay rates of $\sum_{m=1}^M \left\| A_m - \widehat{A}_m \right\|_F^2$.

The analysis leverages high probability bounds on the sample covariance matrices of all systems, denoted by

$$\Sigma_m = \sum_{t=0}^{T-1} x_m(t) x_m(t)'.$$

For that purpose, we utilize the Jordan forms of matrices, as follows. For matrix A_m, its Jordan decomposition is $A_m = P_m^{-1} \Lambda_m P_m$, where Λ_m is a block diagonal matrix; $\Lambda_m = \operatorname{diag}(\Lambda_{m,1}, \ldots \Lambda_{m,q_m})$, and for $i = 1, \ldots q_m$, each block $\Lambda_{m,i} \in \mathbb{C}^{l_{m,i} \times l_{m,i}}$ is a Jordan matrix of the

eigenvalue $\lambda_{m,i}$. A Jordan matrix of size l for $\lambda \in \mathbb{C}$ is

$$\begin{bmatrix} \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{bmatrix} \in \mathbb{C}^{l \times l}. \tag{4}$$

Henceforth, we denote the size of each Jordan block by $l_{m,i}$, for $i = 1, \cdots, q_m$, and the size of the largest Jordan block for system m by l_m^*. Note that for *diagonalizable* matrices A_m, since Λ_m is diagonal, we have $l_m^* = 1$. Now, using this notation, we define

$$\alpha(A_m) = \begin{cases} \left\|\left\|P_m^{-1}\right\|\right\|_{\infty\to 2} \left\|P_m\right\|_\infty f(\Lambda_m) & |\lambda_{m,1}| < 1 - \frac{\rho}{T} \\ \left\|\left\|P_m^{-1}\right\|\right\|_{\infty\to 2} \left\|P_m\right\|_\infty e^{\rho+1} & ||\lambda_{m,1}| - 1| \le \frac{\rho}{T}, \end{cases} \tag{5}$$

where $\lambda_{m,1} = \lambda_1(A_m)$ and

$$f(\Lambda_m) = e^{1/|\lambda_{m,1}|} \left[\frac{l_m^* - 1}{-\log|\lambda_{m,1}|} + \frac{(l_m^* - 1)!}{(-\log|\lambda_{m,1}|)^{l_m^*}} \right].$$

The quantities in the definition of $\alpha(A_m)$ can be interpreted as follows. The term $\left\|\left\|P_m^{-1}\right\|\right\|_{\infty\to 2} \left\|P_m\right\|_\infty$ is similar to the condition number of the similarity matrix P_m in the Jordan decomposition that is used to block-diagonalize the matrix. Moreover, $f(\Lambda_m)$ for stable matrices, and $e^{\rho+1}$ for transition matrices with (almost) unit eigenvalues, capture the *long term* influences of the eigenvalues. In other words, $f(\Lambda_m)$ indicates the amount that $\eta_m(t)$ contributes to the growth of $\|x_m(s)\|$, for $s \gg t$ and $|\lambda_{m,1}| < 1 - \rho/T$. When $|\lambda| \approx 1$, $\|x_m(s)\|$ scales polynomially with the trajectory length T, since influences of the noise vectors $\eta_m(t)$ do not decay as $s - t$ grows, because of the accumulations caused by the unit eigenvalues. The exact expressions are in Theorem 1 below. Note that while $f(\Lambda_m)$ is used to obtain an analytical upper bound for the whole range $|\lambda_{m,1}| < 1 - \rho/T$, it is not tight for small values of $\lambda_{m,1}$ and tighter expressions can be obtained using the analysis in the proof of Theorem 1.

To introduce the following result, we define $\bar{b}_m$ next. First, for some $\delta_C > 0$ that will be determined later, for system m, define $\bar{b}_m = b_T(\delta_C/3) + \|x_m(0)\|_\infty$, where $b_T(\delta) = \sqrt{2\sigma^2 \log(2dMT\delta^{-1})}$. Then, we establish high probability bounds on the sample covariance matrices Σ_m with the detailed proof provided in Section 6.

Theorem 1 (Covariance matrices) *Under Assumptions 1 and 2, for each system m, let $\underline{\Sigma}_m = \underline{\lambda}_m I$ and $\bar{\Sigma}_m = \bar{\lambda}_m I$, where $\underline{\lambda}_m := 4^{-1}\lambda_{\min}(C)T$, and*

$$\bar{\lambda}_m := \begin{cases} \alpha(A_m)^2 \bar{b}_m^2 T, & \text{if } |\lambda_{m,1}| < 1 - \frac{\rho}{T}, \\ \alpha(A_m)^2 \bar{b}_m^2 T^{2l_m^*+1}, & \text{if } ||\lambda_{m,1}| - 1| \le \frac{\rho}{T}. \end{cases}$$

Then, there is T_0, such that for $m \in [M]$ and $T \ge T_0$:

$$\mathbb{P}\left[0 \prec \underline{\Sigma}_m \preceq \Sigma_m \preceq \bar{\Sigma}_m\right] \ge 1 - \delta_C. \tag{6}$$

The above two expressions for $\bar{\lambda}_m$ show that for $|\lambda_{m,1}| < 1 - \rho/T$, the largest eigenvalue of the covariance matrix grows linearly in T, whereas for $||\lambda_{m,1}| - 1| \le \rho/T$, the bounds scale exponentially with the multiplicities of the eigenvalues. Note that the bounds in Theorem 1 and the estimation error results stated hereafter require the trajectories for each system to be longer than T_0. The precise definition for T_0 can be found in the statement of Lemma 2 in Section 6.

For establishing the above, we extend existing tools for learning linear systems [1,16,36,45]. Specifically, we leverage truncation-based arguments and introduce the quantity $\alpha(A_m)$ that captures the effect of the spectral properties of the transition matrices on the magnitudes of the state trajectories. Further, we develop strategies for finding high probability bounds for largest and smallest singular values of random matrices and for studying self-normalized matrix-valued martingales.

Importantly, Theorem 1 provides a tight characterization of the sample covariance matrix for each system, in terms of the magnitudes of eigenvalues of A_m, as well as the largest block-size in the Jordan decomposition of A_m. The upper bounds show that $\bar{\lambda}_m$ grows exponentially with the dimension d, whenever $l_m^* = \Omega(d)$. Further, if A_m has eigenvalues with magnitudes close to 1, then scaling with time T can be as large as T^{2d+1}. The bounds in Theorem 1 are more general than $\operatorname{tr}\left(\sum_{t=0}^{T} A_m^t A_m'^t\right)$ that appears in some analyses [36,39], and can be used to calculate the latter term. Finally, Theorem 1 indicates that the classical framework of persistent excitation [7,21,24] is not applicable, since the lower and upper bounds of eigenvalues grow at drastically different rates.

Next, we express the joint estimation error rates.

Definition 1 *Denote $\mathcal{E}_C = \left\{0 \prec \underline{\Sigma}_m \preceq \Sigma_m \preceq \bar{\Sigma}_m\right\}$, and let $\bar{\lambda} = \max_m \bar{\lambda}_m$, $\underline{\lambda} = \min_m \underline{\lambda}_m$, $\boldsymbol{\kappa}_m = \bar{\lambda}_m/\underline{\lambda}_m$, $\boldsymbol{\kappa} = \max_m \boldsymbol{\kappa}_m$, and $\boldsymbol{\kappa}_\infty = \bar{\lambda}/\underline{\lambda}$. Note that $\boldsymbol{\kappa}_\infty > \boldsymbol{\kappa}$.*

Theorem 2 *Under Assumptions 1, 2, and 3, and for $T \ge T_0$, the estimator in (3) returns $\widehat{A}_m$ for each system $m \in [M]$, such that with probability at least $1 - \delta$, the following holds:*

$$\frac{1}{M}\sum_{m=1}^{M} \left\|\widehat{A}_m - A_m\right\|_F^2 \lesssim \frac{c^2}{\underline{\lambda}} \left(k \log \boldsymbol{\kappa}_\infty + \frac{d^2 k}{M} \log \frac{\boldsymbol{\kappa} dT}{\delta} \right).$$

The proof is provided in Section 7. By putting Theorems 1 and 2 together, the estimation error per-system [1] is

$$O\left(\frac{c^2 k \log \boldsymbol{\kappa}_\infty}{\lambda_{\min}(C)T} + \frac{c^2 d^2 k \log \frac{\boldsymbol{\kappa} dT}{\delta}}{M\lambda_{\min}(C)T}\right). \tag{7}$$

The above expression demonstrates the effects of learning the systems in a joint manner. The first term in (7) can be interpreted as the error in estimating the idiosyncratic components β_m for each system. The convergence rate is $O(k/T)$, as each β_m is a k-dimensional parameter and for each system, we have a trajectory of length T. More importantly, the second term in (7) indicates that the joint estimator in (3) effectively increases the sample size for the shared components $\{W_i\}_{i=1}^k$, by pooling the data of all systems. So, the error decays as $O(d^2k/MT)$, showing that the effective sample size for $\{W_i\}_{i=1}^k$ is MT.

In contrast, for individual learning of LTI systems, the rate is known [16,18,36,39] to be

$$\left\|\widehat{A}_m - A_m\right\|_F^2 \lesssim \frac{c^2 d^2}{\lambda_{\min}(C)T} \log \frac{\alpha(A_m)T}{\delta}.$$

Thus, the estimation error rate in (7) recovers the rate for a single system ($k = 1$), and it significantly improves for joint learning, especially when

$$k < d^2 \qquad \text{and} \qquad k < M. \tag{8}$$

Note that the above conditions are as expected. First, when $k \approx d^2$, the structure in Assumption 3 does *not* provide any commonality among the systems. That is, for $k = d^2$, the LTI systems can be totally arbitrary and Assumption 3 is automatically satisfied. This prevents reductions in the effective dimension of the unknown transition matrices, and also prevents joint learning from being any different than individual learning. Similarly, $k \approx M$ precludes all commonalities and indicates that $\{A_m\}_{m=1}^M$ are too heterogeneous to allow for any improved learning via joint estimation.

Importantly, when the largest block-size l_m^* varies significantly across the M systems, a higher degree of shared structure is needed to improve the joint estimation error for all systems. Since $\boldsymbol{\kappa}$ and $\boldsymbol{\kappa}_\infty$ depend exponentially on l_m^* (as shown in Figure 1 and Theorem 1) and l_m^* can be as large as d, we can have $\log \boldsymbol{\kappa}_\infty = \log \boldsymbol{\kappa} = \Omega(d)$. Hence, in this situation we incur an additional dimension dependence in the error of the joint estimator. Note that such effects of l_m^* are unavoidable (regardless of the employed estimator). Moreover, in this case, joint learning rates improve if $k \leq d$ and $kd \leq M$. Therefore, our analysis highlights the important effects of the large blocks in the Jordan form of the transition matrices.

The above is an inherent difference between estimating dynamics of LTI systems and learning from *independent* observations. In fact, the analysis established in this work includes stochastic matrix regressions that the data of system m consists of

$$y_m(t) = A_m x_m(t) + \eta_m(t), \tag{9}$$

wherein the regressors $x_m(t)$ are drawn from some distribution $\mathcal{D}_m$, and $y_m(t)$ is the response. Assume that $(x_m(t), y_m(t))$ are independent as m, t vary. Now, the sample covariance matrix Σ_m for each system does not depend on A_m. Hence, the error for the joint estimator is not affected by the block-sizes in the Jordan decomposition of A_m. Therefore, in this setting, joint learning always leads to improved per-system error rates, as long as the necessary conditions $k < d^2$ and $k < M$ hold.

4 Robustness to Misspecifications

In Theorem 2, we showed that Assumption 3 can be utilized for obtaining an improved estimation error, by jointly learning the M systems. Next, we consider the impacts of misspecified models on the estimation error and study robustness of the proposed joint estimator against violations of the structure in Assumption 3.

Let us first consider the deviation of the dynamics of each system $m \in [M]$ from the shared structure. Specifically, by employing the matrix D_m to denote the deviation of system m from Assumption 3, suppose that

$$A_m = \left(\sum_{i=1}^k \beta_m^*[i] W_i^*\right) + D_m. \tag{10}$$

Then, denote the *total misspecification* by $\bar{\zeta}^2 = \sum_{m=1}^M \|D_m\|_F^2$. We study the consequences of the above deviations, assuming that the same joint learning method as before is used for estimating the transition matrices.

Theorem 3 *Under Assumptions 1, 2,* (10)*, and for $T \geq T_0$, the estimator in* (3) *returns $\widehat{A}_m$ for each system $m \in [M]$, such that with probability at least $1 - \delta$, we have:*

$$\frac{1}{M}\sum_{m=1}^M \left\|\widehat{A}_m - A_m\right\|_F^2 \lesssim \frac{c^2}{\lambda}\left(k \log \boldsymbol{\kappa}_\infty + \frac{d^2 k}{M} \log \frac{\boldsymbol{\kappa} dT}{\delta}\right) + \frac{(\boldsymbol{\kappa}_\infty + 1)\,\bar{\zeta}^2}{M}. \tag{11}$$

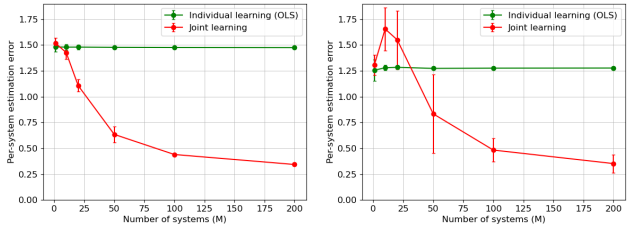
[1] In order to obtain a guarantee for the maximum error over all systems, additional assumptions on the matrix $[\beta_1^* \dots \beta_M^*]$ are required. This problem falls beyond the scope of this paper and we leave it to a future work.

The proof of Theorem 3 is provided in Section 8. In (11), we observe that the total misspecification $\bar{\zeta}^2$ imposes an additional error of $(\boldsymbol{\kappa}_\infty + 1)\bar{\zeta}^2$ for jointly learning all M system. Hence, to obtain accurate estimates, we need the total misspecification $\bar{\zeta}^2$ to be smaller than the number of systems M, as one can expect. The discussion following Theorem 2 is still applicable in the misspecified setting and indicates that in order to have accurate estimates, the number of the shared bases k must be smaller than M as well. In addition, compared to individual learning, the joint estimation error improves *despite the unknown model misspecifications*, as long as

$$\frac{\boldsymbol{\kappa}_\infty \bar{\zeta}^2}{M} \lesssim \frac{d^2}{T}.$$

This shows that when the total misspecification is proportional to the number of systems; $\bar{\zeta}^* = \Omega(M)$, we pay a constant factor proportional to $\boldsymbol{\kappa}_\infty$ on the per-system estimation error. Note that in case all systems are stable, according to Theorem 1, the maximum condition number $\boldsymbol{\kappa}_\infty$ does *not* grow with T, but it scales exponentially with l_m^*. The latter again indicates an important consequence of the largest block-sizes in Jordan decomposition that this work introduces.

Moreover, when a transition matrix A_m has eigenvalues close to or on the unit circle in the complex plane, by Theorem 1, the factor $\boldsymbol{\kappa}_\infty$ grows polynomially with T. Thus, for systems with infinite memories or accumulative behaviors, misspecifications can significantly deteriorate the benefits of joint learning. Intuitively, the reason is that effects of notably small misspecifications can accumulate over time and contaminate the whole data of state trajectories, because of the unit eigenvalues of the transition matrices A_m. Therefore, the above strong sensitivity to deviations from the shared model for systems with unit eigenvalues seems to be unavoidable.



(a) $|\lambda_1(A)| < 1$ (b) $|\lambda_1(A)| \approx 1$

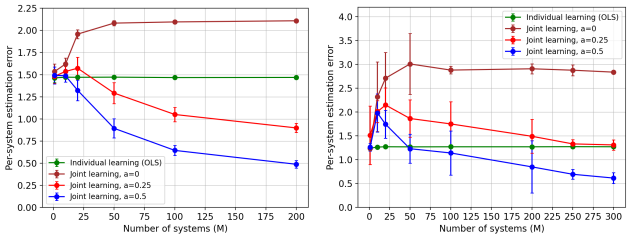
Fig. 2. Per-system estimation errors vs. the number of systems M, for the proposed joint learning method and individual least-squares estimates of the linear dynamical systems.

For example, if for the total misspecification we have $\bar{\zeta}^2 = O(M^{1-a})$, for some $a > 0$, joint estimation improves over the individual estimators, as long as $T\boldsymbol{\kappa}_\infty \lesssim M^a d^2$. Hence, when all systems are stable, the joint estimation error rate improves when the number of systems satisfies $T^{1/a} \lesssim M$. Otherwise, idiosyncrasies in system dynamics dominate the commonalities. Note that larger values of a correspond to *smaller* misspecifications. On the other hand, Theorem 3 implies that in systems with (almost) unit eigenvalues, the impact of $\bar{\zeta}^2$ is amplified. Indeed, by Theorem 1, for unit-root systems, joint learning improves over individual estimators when $d^2 M^a \gg T^{2l_m^*+2}$. That is, for benefiting from the shared structure and utilizing pooled data, the number of systems M needs to be as large as $T^{(2l_m^*+2)/a}/d^{2/a}$.

In contrast, if $\bar{\zeta}^2 = O(M^{1-a})$ for some $a > 0$, the joint estimation error for the regression problem in (9) incurs only an additive factor of $O(1/M^a)$, regardless of the largest block-sizes in the Jordan decompositions and unit-root eigenvalues. Thus, Theorem 3 further highlights the stark difference between joint learning from independent, bounded, and stationary observations, and from state trajectories of LTI systems.

5 Numerical Illustrations

We complement our theoretical analyses with a set of numerical experiments which demonstrate the benefit of jointly learning the systems. We investigate two main aspects of our theoretical results: (i) benefits of joint learning when the M systems share a common linear basis, for different values of M, and (ii) interplay of the spectral radii of the system matrices with the joint-estimation error. To that end, we compare the estimation error for the joint estimator in (3) against the ordinary least-squares (OLS) estimates of the transition matrices for each system individually. For solving (3), we use a minibatch gradient-descent-based implementation with Adam as the optimization algorithm [28]. Due to the bilinear form of the optimization objective, gradient descent methods can lead to convergence and computational issues for $\widehat{\mathbf{W}}$ and $\widehat{B}$. Although prior studies utilize regularization penalties to address this issue in some cases [44], we do not use any such regularization in our objective function in (3). Notably, our unregularized minimization exposes no convergence issue in the simulations we performed.



(a) $|\lambda_1(A)| < 1$ (b) $|\lambda_1(A)| \approx 1$

Fig. 3. Per-system estimation errors are reported vs. the number of systems M, for varying proportions of misspecified systems; M^{-a}, for $a \in \{0, 0.25, 0.5\}$.

For generating the systems, we consider settings with the number of bases $k = 10$, dimension $d = 25$, trajectory length $T = 200$, and the number of systems $M \in \{1, 10, 20, 50, 100, 200\}$. We simulate two cases:
(i) the spectral radii are in the range $[0.7, 0.9]$, and
(ii) all systems have an eigenvalue of magnitude 1.

The matrices $\{W_i\}_{i=1}^{10}$ are generated randomly, such that each entry of W_i is sampled independently from the standard normal distribution $N(0, 1)$. Using these matrices, we generate M systems by randomly generating the idiosyncratic components β_m from a standard normal distribution. For generating the state trajectories, noise vectors are isotropic Gaussian with variance 4. Additional numerical simulations using Bernoulli random matrices are provided in the full version of the paper [34].

We simulate the joint learning problem both with and without model misspecifications. For the latter, deviations from the shared structure are simulated by the components D_m, which are added randomly with probability $1/M^a$ for $a \in \{0, 0.25, 0.5\}$. The matrices D_m are generated with independent Gaussian entries of variance 0.01, leading to $\|D_m\|_F^2 \approx 6.25$ and $\bar{\zeta}^2 \approx 6.25\ M^{1-a}$, according to the dimension $d = 25$.

To report the results, for each value of M in Figure 2 (resp. Figure 3), we average the errors from 10 (resp. 20) random replicates and plot the standard deviation as the error bar. Figure 2 depicts the estimation errors for both stable and unit-root transition matrices, versus M. It can be seen that the joint estimator exhibits the expected improvement against the individual one.

More interestingly, in Figure 3(a), we observe that for stable systems, the joint estimator performs worse than the individual one, when significant violations from the shared structure occurs in all systems (i.e., $a = 0$). Note that it corroborates Theorem 3, since in this case the total misspecification $\bar{\zeta}^2$ *scales linearly* with M. However, if the proportion of systems which violate the shared structure in Assumption 3 decreases, the joint estimation error improves as expected ($a = 0.25, 0.5$).

Figure 3(b) depicts the estimation error for the joint estimator under misspecification for systems that have an eigenvalue on the unit circle in the complex plane. Our theoretical results suggest that the number of systems needs to be significantly larger in this case to circumvent the cost of misspecification in joint learning. The figure corroborates this result, wherein we observe that the joint estimation error is larger than the individual one, if all systems are misspecified (i.e., $a = 0$). Decreases in the total misspecification (i.e., $a = 0.25, 0.5$) improves the error rate for joint learning, but requires larger number of systems than the stable case.

Finally, we discuss the choice of the number of bases k for applying the joint estimator to real data. It can be handled by model selection methods such as elbow criterion and information criteria [2,37], as well as robust estimation methods in panel data and factor models [12,13]. In fact, for all $k' \geq k$, the structural assumption is satisfied and leads to similar learning rates, while $k' < k$ can lead to larger estimation errors. In Figure 4, we provide a simulation (with $T = 250, M = 50$) and report the per-system estimation error, as well as the prediction error on a validation data (which is a subset of size 50). Across all 10 runs in the experiment, we observed that if the hyperparameter k' is chosen according to the elbow criteria, the resulting number of basis models is either equal to the true value $k = 10$, or slightly larger. For misspecified models, the optimal choice of k' can vary, in the sense that large misspecifications can be added to the shared basis (i.e., $k' > k$).

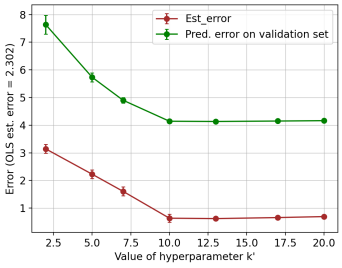


Fig. 4. Estimation and validation prediction errors versus the hyperparameter k', for the true value $k = 10$.

6 Proof of Theorem 1

In this and the following sections, we provide the detailed proofs for our results. We start by analysing the sample covariance matrix for each system which is then used to derive the estimation error rates in Theorem 2 and Theorem 3. Due to space constraints, some details of the proofs are delegated to the full version of this paper which is available online [34]. In Section 9, we provide the general probabilistic inequalities that are used throughout the proofs. Now, we prove high probability bounds for covariance matrices $\Sigma_m = \Sigma_m(T) = \sum_{t=0}^{T} x_m(t)x_m(t)'$ in Theorem 1.

6.1 Upper Bounds on Covariance Matrices

To prove an upper bound on each system covariance matrix, we use an approach for LTI systems that relies on bounding norms of exponents of matrices [16]. Using l_m^* and $\alpha(A_m)$ in (5) and $\xi_m = \left\|\left\|P_m^{-1}\right\|\right\|_{\infty\to 2} \left\|\left\|P_m\right\|\right\|_{\infty}$, the first step is to bound the sizes of all state vectors under the event $\mathcal{E}_{\text{bdd}}(\delta)$ in Proposition 7.

Proposition 1 (Bounding $\|x_m(t)\|$) *For all* $t \in$

$[T], m \in [M]$, *under the event* $\mathcal{E}_{\text{bdd}}(\delta)$, *we have:*

$$\|x_m(t)\| \leq \begin{cases} \alpha(A_m)\bar{b}_m(\delta), & \text{if} \quad |\lambda_{m,1}| < 1 - \frac{\rho}{T}, \\ \alpha(A_m)\bar{b}_m(\delta)t^{l^*_m}, & \text{if} \quad |\lambda_{m,1} - 1| \leq \frac{\rho}{T} \end{cases}.$$

where $\bar{b}_m(\delta) = (b_T(\delta) + \|x_m(0)\|_\infty)$.

Proof As before, each transition matrix A_m admits a Jordan normal form as follows: $A_m = P_m^{-1}\Lambda_m P_m$, where Λ_m is a block-diagonal matrix $\Lambda_m = \text{diag}\,(\Lambda_{m,q}, \ldots, \Lambda_{m,q})$. Each Jordan block $\Lambda_{m,i}$ is of size $l_{m,i}$. Note that for each system, the state vector satisfies:

$$\begin{aligned} x_m(t) &= \sum_{s=1}^{t} A_m^{t-s}\eta_m(s) + A_m^t x_m(0) \\ &= \sum_{s=1}^{t} P_m^{-1}\Lambda_m^{t-s}P_m\eta_m(s) + P_m^{-1}\Lambda_m^t P x_m(0). \end{aligned}$$

Now, letting $b_T(\delta)$ be the same as in Proposition 7, we can bound the ℓ_2-norm of the state vector as follows:

$$\begin{aligned} \|x_m(t)\| \leq\ & \left\|\left\|P_m^{-1}\right\|\right\|_{\infty\to 2} \left\|\left\|\sum_{s=1}^{t}\Lambda_m^{t-s}\right\|\right\|_\infty \|P_m\|_\infty\, b_T(\lambda) \\ &+ \left\|\left\|P_m^{-1}\right\|\right\|_{\infty\to 2} \left\|\left\|\Lambda_m^t\right\|\right\|_\infty \|P_m\|_\infty \|x_m(0)\|_\infty \\ &\leq \xi_m \left(\sum_{s=0}^{t}\left\|\left\|\Lambda_m^{t-s}\right\|\right\|_\infty\right)\left(b_T(\delta) + \|x_m(0)\|_\infty\right). \end{aligned}$$

For any matrix, the ℓ_∞ norm is equal to the maximum row sum. Since the powers of a Jordan matrix will follow the same block structure as the original one, we can bound the operator norm $\|\|A_m^{t-s}\|\|_\infty$ by the norm of each block. The maximum row sum for the s-th power of a Jordan block is: $\sum_{j=0}^{l-1}\binom{s}{j}\lambda^{s-j}$. Using this, we will bound the size of each state vector for the case when

(I) the spectral radius of A_m satisfies $|\lambda_1(A_m)| < 1 - \frac{\rho}{T}$,
(II) or, when $|\lambda_1(A_m) - 1| \leq \frac{\rho}{T}$, for a constant $\rho > 0$.

Case I When the Jordan block for a system matrix has eigenvalues strictly less than 1, we have:

$$\begin{aligned} \sum_{s=0}^{t}\left\|\left\|\Lambda_m^{t-s}\right\|\right\|_\infty &\leq \max_{i\in[q_m]}\sum_{s=0}^{t}\sum_{j=0}^{l_{m,i}-1}\binom{s}{j}|\lambda_{m,i}|^{s-j} \\ &\leq \sum_{s=0}^{t}\sum_{j=0}^{l^*_m-1}\frac{s^j}{j!}|\lambda_{m,1}|^{s-j} \\ &\leq \sum_{s=0}^{t}|\lambda_{m,1}|^s\, s^{l^*_m-1}\sum_{j=0}^{l^*_m-1}\frac{|\lambda_{m,1}|^{-j}}{j!} \\ &\leq e^{1/|\lambda_{m,1}|}\sum_{s=0}^{\infty}|\lambda_{m,1}|^s\, s^{l^*_m-1} \\ &\lesssim e^{1/|\lambda_{m,1}|}\left[\frac{l^*_m-1}{-\log|\lambda_{m,1}|} + \frac{(l^*_m-1)!}{(-\log|\lambda_{m,1}|)^{l^*_m}}\right]. \end{aligned}$$

Thus, for this case, each state vector can be upper bounded as $\|x_m(t)\| \leq \alpha(A_m)(b_T(\delta)+\|x_m(0)\|_\infty)$. When the matrix A_m is diagonalizable, each Jordan block is of size 1, which leads to the upper-bound $\sum_{s=0}^{t}\|\|\Lambda_m^{t-s}\|\|_\infty \leq (1-\lambda_1)^{-1}$, for all $t \geq 0$. Therefore for diagonalizable A_m, we can let $\alpha(A_m) = (1-\lambda_1)^{-1}\left\|\left\|P_m^{-1}\right\|\right\|_{\infty\to 2}\|\|P_m\|\|_\infty$.

Case II When $|\lambda_{m,1} - 1| \leq \frac{\rho}{T}$, we get $|\lambda_{m,1}|^t \leq e^\rho$, for all $t \leq T$. Therefore, since l^*_m is the largest Jordan block, we have:

$$\begin{aligned} \sum_{s=0}^{t}\left\|\left\|\Lambda_m^{t-s}\right\|\right\|_\infty &\leq \sum_{s=0}^{t}\sum_{j=0}^{l^*_m-1}\binom{s}{j}|\lambda_{m,1}|^{s-j} \leq e^\rho\sum_{s=0}^{t}\sum_{j=0}^{l^*_m-1}\binom{s}{j} \\ &\leq e^\rho\sum_{s=0}^{t}\sum_{j=0}^{l^*_m-1}s^j/j! \leq e^\rho\sum_{s=0}^{t}s^{l^*_m-1}\sum_{j=0}^{l^*_m-1}1/j! \\ &\leq e^{\rho+1}\sum_{s=0}^{t}s^{l^*_m-1} \lesssim e^{\rho+1}t^{l^*_m}. \end{aligned}$$

Therefore, the magnitude of each state vector grows polynomially with t, the exponent being at most l^*_m. For example, when A_m is diagonalizable, the Jordan block for the unit root is of size 1, givin $\sum_{s=0}^{t}\|\|\Lambda_m^{t-s}\|\|_\infty \leq e^\rho t$.

So, for systems with unit roots, the bound on each state vector is as expressed in the proposition. ■

Using the high probability upper bound on the size of each state vector, we can upper bound the covariance matrix for each system as follows:

Lemma 1 (Upper bound on Σ_m) *For all* $m \in [M]$, *the sample covariance matrix* Σ_m *of system* m *can be upper bounded under the event* $\mathcal{E}_{\text{bdd}}(\delta)$, *as follows:*

(I) When all eigenvalues of the matrix A_m are strictly less than 1 in magnitude ($|\lambda_{m,i}| < 1 - \frac{\rho}{T}$), we have

$$\lambda_{\max}(\Sigma_m) \leq \alpha(A_m)^2 \left(b_T(\delta) + \|x_m(0)\|_\infty\right)^2 T.$$

(II) When some eigenvalues of the matrix A_m are close to 1, i.e. $|\lambda_1(A_m) - 1| \leq \frac{\rho}{T}$, we have:

$$\lambda_{\max}(\Sigma_m) \leq \alpha(A_m)^2 \left(b_T(\delta) + \|x_m(0)\|_\infty\right)^2 T^{2l_{m,1}+1}.$$

Proof First note that we have:

$$\lambda_{\max}(\Sigma_m) = \left\| \sum_{t=0}^{T} x_m(t) x_m(t)' \right\|_2 \leq \sum_{t=0}^{T} \|x_m(t)\|_2^2.$$

Therefore, by Proposition 7, when all eigenvalues of A_m are strictly less than 1, we have:

$$\lambda_{\max}(\Sigma_m) \leq T\alpha(A_m)^2 \left(b_T(\delta) + \|x_m(0)\|_\infty\right)^2.$$

For the case when $1 - \frac{\rho}{T} \leq \lambda_1(A_m) \leq 1 + \frac{\rho}{T}$, we get:

$$\begin{aligned}\lambda_{\max}(\Sigma_m) &\leq \alpha(\Lambda_m)^2 \sum_{t=0}^{T} t^{2l_{m,1}} \\ &\leq \alpha(A_m)^2 \left(b_T(\delta) + \|x_m(0)\|_\infty\right)^2 T^{2l_{m,1}+1}.\end{aligned}$$

■

6.2 Lower Bound for Covariance Matrices

A lower bound result for the idiosyncratic covariance matrices can be derived using the probabilistic inequalities in the last section. We provide a detailed proof below.

Lemma 2 (Covariance lower bound.) *Define $\varkappa = \frac{d\sigma^2}{\lambda_{\min}(C)^2}$. For all $m \in [M]$, if the per-system sample size T is greater than T_0 defined as*

$$\varkappa \cdot \max\left(c_\eta \log\tfrac{18}{\delta}, 16\left(\log\left(\alpha(A)^2 \bar{b}_m(\delta)^2 + 1\right) + 2\log\tfrac{5}{\delta}\right)\right),$$

if $|\lambda_{m,1}| < 1 - \frac{\rho}{T}$, and

$$\varkappa \cdot \max\left(c_\eta \log\tfrac{18}{\delta}, 16\left(\log\left(\alpha(A)^2 \bar{b}_m(\delta)^2 T^{2l_m^*} + 1\right) + 2\log\tfrac{5}{\delta}\right)\right)$$

if $1 - \frac{\rho}{T} \leq |\lambda_{m,1}| \leq 1 + \frac{\rho}{T}$, then with probability at least $1 - 3\delta$, the sample covariance matrix Σ_m for system m can be bounded from below: $\Sigma_m(T) \succeq \frac{T\lambda_{\min}(C)}{4} I$.

Proof We bound the covariance matrix under the events $\mathcal{E}_{\text{bdd}}(\delta)$, $\mathcal{E}_\eta(\delta)$ in Propositions 7, 8, as well as the one in Proposition 10. As we consider a bound for all systems, we drop the system subscript m here. Using (1), we have:

$$\begin{aligned}\Sigma(T) \succeq\ & A\Sigma(T-1)A' + \sum_{t=1}^{T} \eta(t)\eta(t)' \\ &+ \sum_{t=0}^{T-1} \left(Ax(t)\eta(t+1)' + \eta(t+1)x(t)'A'\right)\end{aligned}$$

Since $T \geq \varkappa c_\eta \log\frac{18}{\delta} = \frac{c_\eta d\sigma^2 \log\frac{18}{\delta}}{\lambda_{\min}(C)^2}$, under the event $\mathcal{E}_\eta(\delta)$ it holds that

$$\begin{aligned}\Sigma(T) \succeq\ & A\Sigma(T-1)A' + \frac{3\lambda_{\min}(C)T}{4} \\ &+ \sum_{t=0}^{T-1} \left(Ax(t)\eta(t+1)' + \eta(t+1)x(t)'A'\right).\end{aligned}$$

Thus, for any unit vector u (i.e., on the unit sphere $\mathcal{S}^{d-1}$), we have

$$\begin{aligned}u'\Sigma(T)u \geq\ & u'A\Sigma(T-1)A'u + \frac{3\lambda_{\min}(C)T}{4} \\ &+ \sum_{t=0}^{T-1} u'\left(Ax(t)\eta(t+1)' + \eta(t+1)x(t)'A'\right)u.\end{aligned}$$

Now, by Proposition 10 with $V = T \cdot I$, we get the following result for the martingale $\sum_{t=0}^{T-1} A_m X_m(t)\eta_m(t+1)'$ and $\bar{V}_m(s) \coloneqq \sum_{t=0}^{s} A_m X_m(t) X_m(t)' A_m' + V$, with probability at least $1 - \delta$:

$$\begin{aligned}&\left\| \sum_{t=0}^{T-1} Ax(t)\eta(t+1)'u \right\| \\ &\leq \sqrt{u'A\Sigma(T-1)A'u + T} \\ &\quad \sqrt{8d\sigma^2 \log\left(\frac{5\det\left(\bar{V}_m(T-1)\right)^{1/2d} \det(TI)^{-1/2d}}{\delta^{1/d}}\right)}.\end{aligned}$$

Thus, we get:

$$\begin{aligned}&u'\Sigma(T)u \\ &\succeq u'A\Sigma(T-1)A'u - \sqrt{u'A\Sigma(T-1)A'u + T} \\ &\quad \sqrt{16d\sigma^2 \log\left(\frac{\lambda_{\max}(\bar{V}(T-1))}{T}\right) + 32d\sigma^2 \log\frac{5}{\delta}} + \frac{3\lambda_{\min}(C)T}{4}.\end{aligned}$$

Hence, we have:

$$\begin{aligned}u'\frac{\Sigma(T)}{T}u \succeq\ & u'\frac{A\Sigma(T-1)A'}{T}u + \frac{3\lambda_{\min}(C)}{4} \\ &- \sqrt{u'\frac{A\Sigma(T-1)A'}{T}u + 1}\frac{\lambda_{\min}(C)}{2} \\ \succeq\ & \frac{\lambda_{\min}(C)}{4},\end{aligned}$$

whenever T is larger than

$$\frac{16d\sigma^2}{\lambda_{\min}(C)^2}\left(\log\left(\frac{\lambda_{\max}\left(\sum_{t=0}^{T-1}AX(t)X(t)'A'\right)}{T}+1\right)+2\log\frac{5}{\delta}\right).$$

Using the upper bound analysis in Lemma 1, we show that it suffices for T to be lower bounded as

$$T \geq \frac{16d\sigma^2}{\lambda_{\min}(C)^2}\left(\log\left(\alpha(A)^2\bar{b}_m(\delta)^2+1\right)+2\log\frac{5}{\delta}\right),$$

when A is strictly stable, and as

$$T \geq \frac{16d\sigma^2}{\lambda_{\min}(C)^2}\left(\log\left(\alpha(A)^2\bar{b}_m(\delta)^2T^{2l^*}+1\right)+2\log\frac{5}{\delta}\right),$$

when $|\lambda_1(A)| \leq 1+\frac{\rho}{T}$. Since, both quantities on the RHS grow at most logarithmically with T, there exists T_0 such that it holds for all $T \geq T_0$. Combining the failure probability for all events, we get the desired result. ■

7 Proof of Theorem 2

In this section, we use the result in Theorem 1 to analyze the estimation error for the estimator in (3), under Assumption 3. For ease of presentation, we rewrite the problem by transforming the vector output space to scalar values. For that purpose, we introduce some notation to express transition matrices in vector form and rewrite (3). First, for each state vector $x_m(t) \in \mathbb{R}^d$, we create d different covariates of size $\mathbb{R}^{d^2}$. So, for $j = 1, \cdots, d$, the vector $\tilde{x}_{m,j}(t) \in \mathbb{R}^{d^2}$ contains $x_m(t)$ in the j-th block of size d and $0's$ elsewhere.

Then, we express the system matrix $A_m \in \mathbb{R}^{d\times d}$ as a vector $\tilde{A}_m \in \mathbb{R}^{d^2}$. Similarly, the concatenation of all vectors $\tilde{A}_m$ can be coalesced into the matrix $\tilde{\Theta} \in \mathbb{R}^{d^2\times M}$. Analogously, $\tilde{\eta}_m(t)$ will denote the concatenated dt dimensional vector of noise vectors for system m. Thus, the structural assumption in (2) can be written as:

$$\tilde{A}_m = W^*\beta_m^*, \tag{12}$$

where $W^* \in \mathbb{R}^{d^2\times k}$ and $\beta_m^* \in \mathbb{R}^k$. Similarly, the overall parameter set can be factorized as $\tilde{\Theta}^* = W^*B^*$, where the matrix $B^* = [\beta_1^*|\beta_2^*|\cdots\beta_M^*] \in \mathbb{R}^{k\times M}$ contains the true weight vectors β_m^*. Thus, expressing the system matrices A_m in this manner leads to a low rank structure in (12), so that the matrix $\tilde{\Theta}^*$ is of rank k. Using the vectorized parameters, the evolution for the components $j \in [d]$ of all state vectors $x_m(t)$ can be written as:

$$x_m(t+1)[j] = \tilde{A}_m\tilde{x}_{m,j}(t) + \eta_m(t+1)[j]. \tag{13}$$

For each system $m \in [M]$, we therefore have a total of dT samples, where the statistical dependence now follows a block structure: d covariates of $x_m(1)$ are all constructed using $x_m(0)$, next d using $x_m(1)$ and so forth. To estimate the parameters, we solve the following optimization problem:

$$\begin{aligned}
&\widehat{W}, \{\widehat{\beta}_m\}_{m=1}^M \\
&:= \underset{W,\{\beta_m\}_{m=1}^M}{\operatorname{argmin}} \underbrace{\sum_{m,t}\sum_{j=1}^{d}\left(x_m(t+1)[j] - \langle W\beta_m, \tilde{x}_{m,j}(t)\rangle\right)^2}_{\mathcal{L}(W,\beta)} \\
&= \underset{W,\{\beta_m\}_{m=1}^M}{\operatorname{argmin}} \sum_{m=1}^{M}\left\|y_m - \tilde{X}_mW\beta_m\right\|_2^2,
\end{aligned} \tag{14}$$

where $y_m \in \mathbb{R}^{Td}$ contains all T state vectors stacked vertically and $\tilde{X}_m \in \mathbb{R}^{Td\times d^2}$ contains the corresponding matrix input. We denote the covariance matrices for the vectorized form by $\tilde{\Sigma}_m = \sum_{t=0}^{T-1}\tilde{x}_m(t)\tilde{x}_m(t)'$. Recall, that the sample covariance matrices for all systems are denoted by $\Sigma_m = \sum_{t=0}^{T-1}x_m(t)x_m(t)'$. We further use the following notation: for any parameter set $\Theta = WB \in \mathbb{R}^{d^2\times M}$, we define $\mathcal{X}(\Theta) \in \mathbb{R}^{dT\times M}$ as $\mathcal{X}(\Theta) := [\mathcal{X}_1(\Theta)|\mathcal{X}_2(\Theta)\cdots|\mathcal{X}_M(\Theta)]$, where each column $\mathcal{X}_m(\Theta) \in \mathbb{R}^{dT}$ is the prediction of states $x_m(t+1)$ with Θ_m. That is,

$$\mathcal{X}_m(\Theta) = (x_m(0)', x_m(0)'\Theta_m', \ldots, x_m(T-1)'\Theta_m')'.$$

Thus, $\mathcal{X}(\tilde{\Theta}^*) \in \mathbb{R}^{Td\times M}$ denotes the ground truth mapping for the training data of the M systems and $\mathcal{X}(\tilde{\Theta}^* - \widehat{\Theta}) \in \mathbb{R}^{Td\times M}$ is the prediction error across all coordinates of the MT state vectors, each of dimension d.

By Assumption 3, we have $\Delta := \tilde{\Theta}^* - \widehat{\Theta} = UR$, where $U \in O^{d^2\times 2k}$ is an orthonormal matrix and $R \in \mathbb{R}^{2k\times M}$. We start by the fact that the estimates $\widehat{W}$ and $\widehat{\beta}_m$ minimize (3), and therefore, have a smaller squared prediction error than (W^*, B^*). Hence, we get the following inequality:

$$\begin{aligned}
&\frac{1}{2}\sum_{m=1}^{M}\left\|\tilde{X}_m(W^*\beta_m^* - \widehat{W}\widehat{\beta}_m)\right\|_2^2 \\
&\leq \sum_{m=1}^{M}\left\langle\tilde{\eta}_m, \tilde{X}_m\left(\widehat{W}\widehat{\beta}_m - W^*\beta_m^*\right)\right\rangle.
\end{aligned} \tag{15}$$

We can rewrite $\widehat{W}\widehat{\beta}_m - W^*\beta_m^* = Ur_m$, for all $m \in [M]$, where $r_m \in \mathbb{R}^{2k}$ is an idiosyncratic projection vector for system m. Since our joint estimator is a least squares objective with bilinear terms, we first decompose the prediction error for the estimator, similar to the linear regression setting [14,44]. In subsequent analyses, we use

different matrix concentration results and LTI estimation theory in order to account for the temporal dependence and spectral properties of the systems. Our first step is to bound the prediction error for all systems.

Lemma 3 *For any fixed orthonormal matrix $\bar{U} \in \mathbb{R}^{d^2 \times 2k}$, the total squared prediction error in (3) for $(\widehat{W}, \widehat{B})$ can be decomposed as follows:*

$$\begin{aligned}
&\frac{1}{2}\sum_{m=1}^{M}\left\|\tilde{X}_m(W^*\beta_m^* - \widehat{W}\widehat{\beta}_m)\right\|_F^2 \\
&\leq \sqrt{\sum_{m=1}^{M}\left\|\tilde{\eta}_m^\top \tilde{X}_m \bar{U}\right\|_{\bar{V}_m^{-1}}^2}\sqrt{2\sum_{m=1}^{M}\left\|\tilde{X}_m(W^*\beta_m^* - \widehat{W}\widehat{\beta}_m)\right\|_2^2} \\
&\quad + \sqrt{\sum_{m=1}^{M}\left\|\tilde{\eta}_m^\top \tilde{X}_m \bar{U}\right\|_{\bar{V}_m^{-1}}^2}\sqrt{\sum_{m=1}^{M}\left\|\tilde{X}_m(\bar{U}-U)r_m\right\|^2} \\
&\quad + \sum_{m=1}^{M}\left\langle \tilde{\eta}_m, \tilde{X}_m(U-\bar{U})r_m\right\rangle. \qquad (16)
\end{aligned}$$

The proof of Lemma 3 can be found in the extended version of this paper [34]. Our next step is to bound each term on the RHS of (16). To that end, let $\mathcal{N}_\epsilon$ be an ϵ-cover of the set of orthonormal matrices in $\mathbb{R}^{d^2\times 2k}$. In (16), we select the matrix $\bar{U}$ to be an element of $\mathcal{N}_\epsilon$ such that $\left\|\bar{U}-U\right\|_F \leq \epsilon$. Note that since $\mathcal{N}_\epsilon$ is an ϵ-cover, such matrix $\bar{U}$ exists. We can bound the size of such a cover using Lemma 5, and obtain $|\mathcal{N}_\epsilon| \leq \left(\frac{6\sqrt{d}}{\epsilon}\right)^{2d^2k}$.

We now bound each term in the following propositions using the auxiliary results in Section 9 and covariance matrix bounds in the previous section. The detailed proofs for the following results are available in the extended version [34]. Using Proposition 9, we bound the expression in the second term of (16), as follows.

Proposition 2 *Under Assumption 3, for the noise process $\{\eta_m(t)\}_{t=1}^{\infty}$ defined for each system, with probability at least $1-\delta_Z$, we have:*

$$\sum_{m=1}^{M}\left\|\tilde{X}_m(\bar{U}-U)r_m\right\|^2 \lesssim \boldsymbol{\kappa}\epsilon^2\left(MT\,\mathrm{tr}\,(C) + \sigma^2\log\frac{2}{\delta_Z}\right).$$

Based on the bound in Proposition 2, we can bound the third term in (16) as follows:

Proposition 3 *Under Assumption 2 and Assumption 3, with probability at least $1-\delta_Z$, we have:*

$$\sum_{m=1}^{M}\left\langle \tilde{\eta}_m, \tilde{X}_m(U-\bar{U})r_m\right\rangle \lesssim \sqrt{\boldsymbol{\kappa}}\epsilon\left(MT\,\mathrm{tr}\,(C) + \sigma^2\log\tfrac{1}{\delta_Z}\right). \qquad (17)$$

Next, we show a multitask concentration of martingales projected on a low-rank subspace.

Proposition 4 *For an arbitrary orthonormal matrix $\bar{U} \in \mathbb{R}^{d^2\times 2k}$ in the ϵ-cover $\mathcal{N}_\epsilon$ defined in Lemma 5, let $\Sigma \in \mathbb{R}^{d^2\times d^2}$ be a positive definite matrix, and define $S_m(\tau) = \tilde{\eta}_m(\tau)^\top \tilde{X}_m(\tau)\bar{U}$, $\bar{V}_m(\tau) = \bar{U}'\left(\tilde{\Sigma}_m(\tau)+\Sigma\right)\bar{U}$, and $V_0 = \bar{U}'\Sigma\bar{U}$. Then, letting $\mathcal{E}_1(\delta_U)$ be the event*

$$\sum_{m=1}^{M}\|S_m(T)\|^2_{\bar{V}_m^{-1}(T)} \leq 2\sigma^2\log\left(\frac{\Pi_{m=1}^{M}\frac{\det(\bar{V}_m(T))}{\det(V_0)}}{\delta_U}\right),$$

we have

$$\mathbb{P}\left[\mathcal{E}_1(\delta_U)\right] \geq 1 - \left(\frac{6\sqrt{2k}}{\epsilon}\right)^{2d^2k}\delta_U. \qquad (18)$$

7.1 Proof of Estimation Error in Theorem 2

Proof We now use the bounds we have shown for each term before and give the final steps by using the error decomposition in Lemma 3. Let $|\mathcal{N}_\epsilon|$ be the cardinality of the ϵ-cover of the set of orthonormal matrices in $\mathbb{R}^{d^2\times 2k}$ that we defined in Lemma 3. Let $\mathbb{V}$ denote the expression $\Pi_{m=1}^{M}\frac{\det(\bar{V}_m(t))}{\det(V_0)}$. So, substituting the termwise bounds from Proposition 2, Proposition 3, and Proposition 4 in Lemma 3, with probability at least $1-|\mathcal{N}_\epsilon|\,\delta_U - \delta_Z$, it holds that:

$$\begin{aligned}
&\frac{1}{2}\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F^2 \\
&\lesssim \sqrt{\sigma^2\log\left(\frac{\mathbb{V}}{\delta_U}\right)}\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F \\
&\quad + \sqrt{\sigma^2\log\left(\frac{\mathbb{V}}{\delta_U}\right)}\sqrt{\boldsymbol{\kappa}\epsilon^2\left(MT\,\mathrm{tr}\,(C) + \sigma^2\log\frac{1}{\delta_Z}\right)} \\
&\quad + \sqrt{\boldsymbol{\kappa}}\epsilon\left(MT\,\mathrm{tr}\,(C) + \sigma^2\log\frac{1}{\delta_Z}\right). \qquad (19)
\end{aligned}$$

For the matrix V_0, we now substitute $\Sigma = \underline{\lambda}I_{d^2}$, which implies that $\det(V_0)^{-1} = \det(1/\underline{\lambda}I_{2k}) = (1/\underline{\lambda})^{2k}$. Similarly, for $\bar{V}_m(T)$, we get $\det(\bar{V}_m(T)) \leq \bar{\lambda}^{2k}$. Thus, substituting $\delta_U = \delta/3\,|\mathcal{N}|_\epsilon$ and $\delta_C = \delta/3$ in Theorem 1,

with probability at least $1-2\delta/3$, the upper-bound in Proposition 4 becomes:

$$\sum_{m=1}^{M}\left\|\tilde{\eta}_m^\top \tilde{X}_m \bar{U}\right\|^2_{\bar{V}_m^{-1}} \lesssim \sigma^2 Mk\log\boldsymbol{\kappa}_\infty + \sigma^2 d^2 k\log\frac{k}{\delta\epsilon}.$$

Substituting this in (19) with $\delta_Z = \delta/3$, $c^2 = \max(\sigma^2, \lambda_1(C))$, with probability at least $1-\delta$, we have:

$$\begin{aligned}
&\frac{1}{2}\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F^2 \\
&\lesssim \sqrt{c^2 Mk\log\boldsymbol{\kappa}_\infty + c^2 d^2 k\log\frac{k}{\delta\epsilon}}\left(\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F\right. \\
&\quad \left. + \sqrt{\boldsymbol{\kappa}\epsilon^2\left(c^2 dMT + c^2\log\frac{1}{\delta}\right)}\right) \\
&\quad + \sqrt{\boldsymbol{\kappa}}\epsilon\left(c^2 dMT + c^2\log\frac{1}{\delta}\right).
\end{aligned}$$

Noting that $\log\frac{1}{\delta} \lesssim d^2 k\log\frac{k}{\delta\epsilon}$ for $\epsilon = \frac{k}{\sqrt{\boldsymbol{\kappa}}d^2 T}$, with probability at least $1-\delta$, we get:

$$\begin{aligned}
&\frac{1}{2}\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F^2 \\
&\lesssim \left(\sqrt{c^2 Mk\log\boldsymbol{\kappa}_\infty + c^2 d^2 k\log\frac{\boldsymbol{\kappa}dT}{\delta}}\right)\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F \\
&\quad + \sqrt{c^2 Mk\log\boldsymbol{\kappa}_\infty + c^2 d^2 k\log\frac{\boldsymbol{\kappa}dT}{\delta}}\sqrt{c^2\left(\frac{k^2 M}{d^3 T} + \frac{k^3}{d^2 T^2}\log\frac{\boldsymbol{\kappa}dT}{\delta}\right)} \\
&\quad + c^2\left(\frac{Mk}{d} + \frac{k^2}{T}\log\frac{\boldsymbol{\kappa}dT}{\delta}\right).
\end{aligned}$$

As $k \le d^2$, we can rewrite the above inequality as:

$$\begin{aligned}
&\frac{1}{2}\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F^2 \\
&\lesssim \sqrt{c^2\left(Mk\log\boldsymbol{\kappa}_\infty + d^2 k\log\frac{\boldsymbol{\kappa}dT}{\delta}\right)}\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F \\
&\quad + c^2\left(Mk\log\boldsymbol{\kappa}_\infty + \frac{d^2 k}{T}\log\frac{\boldsymbol{\kappa}dT}{\delta}\right).
\end{aligned}$$

The above quadratic inequality for the prediction error $\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F^2$ implies the following bound, which holds with probability at least $1-\delta$:

$$\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F^2 \lesssim c^2\left(Mk\log\boldsymbol{\kappa}_\infty + d^2 k\log\frac{\boldsymbol{\kappa}dT}{\delta}\right).$$

Since the smallest eigenvalue of the matrix $\Sigma_m = \sum_{t=0}^{T} X_m(t)X_m(t)'$ is at least $\underline{\lambda}$ (Theorem 1), we can convert the above prediction error bound to an estimation error bound and get

$$\left\|W^*B^* - \widehat{W}\widehat{B}\right\|_F^2 \lesssim \frac{c^2\left(Mk\log\boldsymbol{\kappa}_\infty + d^2 k\log\frac{\boldsymbol{\kappa}dT}{\delta}\right)}{\underline{\lambda}},$$

which implies the desired bound for the solution of (3).

■

8 Proof of Theorem 3

Here, we provide the key steps for bounding the average estimation error across the M systems for the estimator in (3) in presence of misspecifications $D_m \in \mathbb{R}^{d\times d}$:

$$A_m = \left(\sum_{i=1}^{k}\beta_m^*[i]W_i^*\right) + D_m,$$

where we use ζ_m to denote the bound on misspecification in task m and set $\bar{\zeta}^2 = \sum_{m=1}^{M}\zeta_m^2$. In the presence of misspecifications, we have $\Delta \coloneqq \widetilde{\Theta}^* - \widehat{\Theta} = VR + D$, where $V \in O^{d^2\times 2k}$ is an orthonormal matrix, $R \in \mathbb{R}^{2k\times M}$, and $D \in \mathbb{R}^{d^2\times M}$ is the misspecification error. As the analysis here shares its template with the proof of Theorem 2, we provide a sketch with the complete details delegated to the extended version [34]. Same as in Section 7, we start with the fact that $(\widehat{W}, \widehat{B})$ minimize the squared loss in (3). However, in this case, we get an additional term caused by on the misspecifications D_m:

$$\begin{aligned}
&\frac{1}{2}\sum_{m=1}^{M}\left\|\tilde{X}_m(W^*\beta_m^* - \widehat{W}\widehat{\beta}_m)\right\|_2^2 \\
&\le \sum_{m=1}^{M}\left\langle\tilde{\eta}_m, \tilde{X}_m\left(\widehat{W}\widehat{\beta}_m - W^*\beta_m^*\right)\right\rangle \\
&\quad + \sum_{m=1}^{M} 2\left\langle\tilde{X}_m\tilde{D}_m, \tilde{X}_m\left(\widehat{W}\widehat{\beta}_m - W^*\beta_m^*\right)\right\rangle. \qquad (20)
\end{aligned}$$

We follow a similar proof strategy as in Section 7 and account for the additional terms arising due to the misspecifications D_m. The error in the shared part, $\widehat{W}\widehat{\beta}_m - W^*\beta_m^*$, can still be rewritten as Ur_m where $U \in \mathbb{R}^{d^2\times 2k}$ is a matrix containing an orthonormal basis of size $2k$ in $\mathbb{R}^{d^2}$ and $r_m \in \mathbb{R}^{2k}$ is the system specific vector. We now show a decomposition similar to Lemma 3:

Lemma 4 *Under the misspecified shared linear basis structure in* (10)*, for any fixed orthonormal matrix* $\bar{U} \in \mathbb{R}^{d^2\times 2k}$*, the low rank part of the total squared error*

can be decomposed as follows:

$$\begin{aligned}
&\frac{1}{2}\sum_{m=1}^{M}\left\|\tilde{X}_m(W^*\beta_m^* - \widehat{W}\widehat{\beta}_m)\right\|_F^2 \\
&\leq \sqrt{\sum_{m=1}^{M}\left\|\tilde{\eta}_m^\top \tilde{X}_m \bar{U}\right\|_{\bar{V}_m^{-1}}^2}\sqrt{2\sum_{m=1}^{M}\left\|\tilde{X}_m(W^*\beta_m^* - \widehat{W}\widehat{\beta}_m)\right\|_2^2} \\
&\quad + \sum_{m=1}^{M}\left\langle \tilde{\eta}_m, \tilde{X}_m(U-\bar{U})r_m\right\rangle \\
&\quad + \sqrt{\sum_{m=1}^{M}\left\|\tilde{\eta}_m^\top \tilde{X}_m \bar{U}\right\|_{\bar{V}_m^{-1}}^2}\sqrt{2\sum_{m=1}^{M}\left\|\tilde{X}_m(\bar{U}-U)r_m\right\|^2} \\
&\quad + 2\sqrt{\bar{\lambda}}\bar{\zeta}\sqrt{\sum_{m=1}^{M}\left\|\tilde{X}_m\left(\widehat{W}\widehat{\beta}_m - W^*\beta_m^*\right)\right\|_2^2}. \qquad (21)
\end{aligned}$$

We bound each term on the RHS of (21) individually. Similar to Section 7, we choose the orthonormal $\mathbb{R}^{d^2\times 2k}$ matrix $\bar{U} \in \mathcal{N}_\epsilon$. Then, we use the following results, for which the proofs are provided in the longer version [34].

Proposition 5 (Bounding $\sum_{m=1}^{M}\left\|\tilde{X}_m(\bar{U}-U)r_m\right\|^2$) *For the model in* (10)*, with probability at least* $1-\delta_Z$*, it holds that*

$$\sum_{m=1}^{M}\left\|\tilde{X}_m(\bar{U}-U)r_m\right\|^2 \lesssim \boldsymbol{\kappa}\epsilon^2\left(MT\,\mathrm{tr}\,(C) + \sigma^2\log\tfrac{2}{\delta_Z} + \bar{\lambda}\bar{\zeta}^2\right). \qquad (22)$$

Proposition 6 (Bounding $\sum_{m=1}^{M}\left\langle \tilde{\eta}_m, \tilde{X}_m(U-\bar{U})r_m\right\rangle$) *Under Assumption 2 and* (10)*, with probability at least* $1-\delta_Z$ *we have:*

$$\begin{aligned}
&\sum_{m=1}^{M}\left\langle \tilde{\eta}_m, \tilde{X}_m(U-\bar{U})r_m\right\rangle \\
&\lesssim \sqrt{\boldsymbol{\kappa}}\epsilon\left(MT\,\mathrm{tr}\,(C) + \sigma^2\log\frac{1}{\delta_Z}\right) \\
&\quad + \sqrt{\boldsymbol{\kappa}\bar{\lambda}}\sqrt{MT\,\mathrm{tr}\,(C) + \sigma^2\log\frac{1}{\delta_Z}}\epsilon\bar{\zeta}. \qquad (23)
\end{aligned}$$

Finally, we are ready to put the above intermediate results together. Using the decomposition in Lemma 4 and the term-wise upper bounds above, one can derive the desired estimation error rate. Below, we show the final steps with appropriate substitution for constants. The full details are available online in [34].

As before, we substitute the termwise bounds from Propositions 5, 6 and 4 in Lemma 4 with values $\delta_U = \delta/3\,|\mathcal{N}|_\epsilon$, $\delta_C = \delta/3$ (in Theorem 1), $\delta_Z = \delta/3$. Noting that $k \leq d^2$ and $\log\frac{1}{\delta} \lesssim d^2 k\log\frac{k}{\delta\epsilon}$, by setting $\epsilon = \frac{k}{\sqrt{\boldsymbol{\kappa}}d^2T}$ we finally get the following quadratic inequality in the error term $\Xi \coloneqq \left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F$:

$$\begin{aligned}
\frac{1}{2}\Xi^2 \lesssim &\left(\sqrt{c^2\left(Mk\log\boldsymbol{\kappa}_\infty + d^2k\log\frac{\boldsymbol{\kappa}dT}{\delta}\right)} + \sqrt{\bar{\lambda}}\bar{\zeta}\right)\Xi \\
&+ c^2\left(Mk\log\boldsymbol{\kappa}_\infty + \frac{d^2k}{T}\log\frac{\boldsymbol{\kappa}dT}{\delta}\right) \\
&+ c\sqrt{\frac{\bar{\lambda}\bar{\zeta}^2}{T}\left(Mk\log\boldsymbol{\kappa}_\infty + \frac{d^2k}{T}\log\frac{\boldsymbol{\kappa}dT}{\delta}\right)}.
\end{aligned}$$

The quadratic inequality for the prediction error $\left\|\mathcal{X}(W^*B^* - \widehat{W}\widehat{B})\right\|_F^2$ implies the following bound with probability at least $1-\delta$:

$$\Xi^2 \lesssim c^2\left(Mk\log\boldsymbol{\kappa}_\infty + d^2k\log\frac{\boldsymbol{\kappa}dT}{\delta}\right) + \bar{\lambda}\bar{\zeta}^2.$$

Since $\underline{\lambda} = \min_m \underline{\lambda}_m$, an estimation error bound for the solution of (3):

$$\begin{aligned}
&\sum_{m=1}^{M}\left\|\widehat{A}_m - A_m\right\|_F^2 \\
&\lesssim \frac{c^2\left(Mk\log\boldsymbol{\kappa}_\infty + d^2k\log\frac{\boldsymbol{\kappa}dT}{\delta}\right)}{\underline{\lambda}} + (\boldsymbol{\kappa}_\infty + 1)\bar{\zeta}^2.
\end{aligned}$$

9 Auxiliary Probabilistic Inequalities

In this section, we state the general probabilistic inequalities which we used in proving the main results in the previous sections. The proofs for these results can be found in the full preprint [34].

Proposition 7 (Bounding the noise sequence) *For $T = 0, 1, \ldots,$ and $0 < \delta < 1$, let $\mathcal{E}_{\mathrm{bdd}}$ be the event*

$$\mathcal{E}_{\mathrm{bdd}}(\delta) \coloneqq \left\{\max_{1\leq t\leq T, m\in[M]}\|\eta_m(t)\|_\infty \leq \sqrt{2\sigma^2\log\tfrac{2dMT}{\delta}}\right\}. \qquad (24)$$

Then, we have $\mathbb{P}[\mathcal{E}_{\mathrm{bdd}}] \geq 1-\delta$. For simplicity, we denote the above upper-bound by $b_T(\delta)$.

Proposition 8 (Noise covariance concentration) *For T and $0 < \delta < 1$, let $\mathcal{E}_\eta$ be the event*

$$\mathcal{E}_\eta(\delta) \coloneqq \left\{\tfrac{3\lambda_{\min}(C)}{4}I \preceq \tfrac{1}{T}\sum_{t=1}^{T}\eta_m(t)\eta_m(t)' \preceq \tfrac{5\lambda_{\max}(C)}{4}I\right\}.$$

Then, if $T \geq T_\eta(\delta) := \frac{c_\eta d\sigma^2}{\lambda_{\min}(C)^2} \log 18/\delta$, *we have* $\mathbb{P}[\mathcal{E}_{\text{bdd}}(\delta) \cap \mathcal{E}_\eta(\delta)] \geq 1 - 2\delta$.

Define $Z \in \mathbb{R}^{dT \times M}$ as the pooled noise matrix as follows:

$$Z = [\tilde{\eta}_1(T) | \tilde{\eta}_2(T) \cdots | \tilde{\eta}_M(T)], \tag{25}$$

with each column vector $\eta_m(T) \in \mathbb{R}^{dT}$ as the concatenated noise vector $(\eta_m(1), \eta_m(2), \ldots, \eta_m(T))$ for the m-th system.

Proposition 9 (Bounding total magnitude of noise) *For the joint noise matrix* $Z \in \mathbb{R}^{dT \times M}$ *defined in* (25), *with probability at least* $1 - \delta$, *we have:*

$$\|Z\|_F^2 \leq MT \operatorname{tr}(C) + \log \frac{2}{\delta}.$$

We denote the above event by $\mathcal{E}_Z(\delta)$.

The following result shows a self-normalized martingale bound for vector valued noise processes.

Proposition 10 *For the system in* (1), *for any* $0 < \delta < 1$ *and system* $m \in [M]$, *with prob. at least* $1 - \delta$, *we have:*

$$\left\| \bar{V}_m^{-1/2} (T-1) \sum_{t=0}^{T-1} x_m(t) \eta_m(t+1)' \right\|_2 \leq \sigma \sqrt{8d \log \left(\frac{5 \det\left(\bar{V}_m(T-1)\right)^{1/2d} \det(V)^{-1/2d}}{\delta^{1/d}} \right)},$$

where $\bar{V}_m(s) = \sum_{t=0}^{s} x_m(t) x_m(t)' + V$ *and* V *is a deterministic positive definite matrix.*

Lemma 5 (Covering low-rank matrices [14]) *For the set of orthonormal matrices* $O^{d \times d'}$ *(with* $d > d'$*), there exists* $\mathcal{N}_\epsilon \subset O^{d \times d'}$ *that forms an* ϵ*-net of* $O^{d \times d'}$ *in Frobenius norm such that* $|\mathcal{N}_\epsilon| \leq (\frac{6\sqrt{d'}}{\epsilon})^{dd'}$, *i.e., for every* $V \in O^{d \times d'}$, *there exists* $V' \in \mathcal{N}_\epsilon$ *and* $\|V - V'\|_F \leq \epsilon$.

10 Concluding Remarks

We studied the problem of jointly learning multiple linear time-invariant dynamical systems, under the assumption that their transition matrices can be expressed based on an unknown shared basis. Our finite-time analysis for the proposed joint estimator shows that pooling data across systems can provably improve over individual estimators, even in presence of moderate misspecifications. The results highlight the critical roles of the spectral properties of the system matrices and the number of the basis matrices, in the efficiency of joint estimation. Further, we characterize fundamental differences between joint estimation of system dynamics using dependent state trajectories and learning from independent stationary observations. Considering different shared structures, extensions of the presented results to explosive systems, or those with high-dimensional transition matrices, as well as joint learning of multiple nonlinear dynamical systems, all are interesting avenues for future work that this paper paves the road towards.

Acknowledgements

The authors appreciate the helpful comments of the reviewers on the initial version of this paper. During this work, AM was supported in part by a grant from the Open Philanthropy Project to the CHAI and NSF CAREER IIS-1452099. AT and GM acknowledge the support from NSF grants IIS-2007055 and DMS-2348640 respectively.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[3] Pierre Alquier, Mai The Tien, Massimiliano Pontil, et al. Regret bounds for lifelong learning. In *Artificial Intelligence and Statistics*, pages 261–269. PMLR, 2017.

[4] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[5] Sumanta Basu, Ali Shojaie, and George Michailidis. Network granger causality with inherent grouping structure. *The Journal of Machine Learning Research*, 16(1):417–453, 2015.

[6] John T Bosworth. *Linearized aerodynamic and control law models of the X-29A airplane and comparison with flight data*, volume 4356. National Aeronautics and Space Administration, Office of Management ..., 1992.

[7] Stephen Boyd and Sosale Shankara Sastry. Necessary and sufficient conditions for parameter convergence in adaptive control. *Automatica*, 22(6):629–639, 1986.

[8] Boris Buchmann and Ngai Hang Chan. Asymptotic theory of least squares estimators for nearly unstable processes under strong dependence. *The Annals of statistics*, 35(5):2001–2017, 2007.

[9] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[10] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[11] Yanxi Chen and H Vincent Poor. Learning mixtures of linear dynamical systems. In *International Conference on Machine Learning*, pages 3507–3557. PMLR, 2022.

[12] Alexander Chudik, Kamiar Mohaddes, M Hashem Pesaran, and Mehdi Raissi. Debt, inflation and growth-robust estimation of long-run effects in dynamic panel data models. 2013.

[13] Valentina Ciccone, Augusto Ferrante, and Mattia Zorzi. Factor models with real data: A robust estimation of the number of factors. *IEEE Transactions on Automatic Control*, 64(6):2412–2425, 2018.

[14] Simon Shaolei Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2020.

[15] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite-time adaptive stabilization of linear systems. *IEEE Transactions on Automatic Control*, 64(8):3498–3505, 2018.

[16] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

[17] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive control and learning. *Automatica*, 117:108950, 2020.

[18] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear–quadratic regulators. *Automatica*, 117:108982, 2020.

[19] André Fujita, Joao R Sato, Humberto M Garay-Malpartida, Rui Yamaguchi, Satoru Miyano, Mari C Sogayar, and Carlos E Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC systems biology*, 1(1):1–11, 2007.

[20] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

[21] Michael Green and John B Moore. Persistence of excitation in linear systems. *Systems & control letters*, 7(5):351–360, 1986.

[22] Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, and Liwei Wang. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, pages 4349–4358. PMLR, 2021.

[23] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

[24] Benjamin M Jenkins, Anuradha M Annaswamy, Eugene Lavretsky, and Travis E Gibson. Convergence properties of adaptive systems and the definition of exponential stability. *SIAM journal on control and optimization*, 56(4):2463–2484, 2018.

[25] Katarina Juselius and Zorica Mladenovic. *High inflation, hyperinflation and explosive roots: the case of Yugoslavia*. Citeseer, 2002.

[26] Thomas Kailath, Ali H Sayed, and Babak Hassibi. *Linear estimation*. Prentice Hall, 2000.

[27] Wei Kang. Approximate linearization of nonlinear control systems. In *Proceedings of 32nd IEEE Conference on Decision and Control*, pages 2766–2771. IEEE, 1993.

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[29] TL Lai and CZ Wei. Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *Journal of multivariate analysis*, 13(1):1–23, 1983.

[30] Weiwei Li and Emanuel Todorov. Iterative linear quadratic regulator design for nonlinear biological movement systems. In *ICINCO (1)*, pages 222–229. Citeseer, 2004.

[31] Rui Lu, Gao Huang, and Simon S Du. On the power of multitask representation learning in linear mdp. *arXiv preprint arXiv:2106.08053*, 2021.

[32] Andreas Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7(Jan):117–139, 2006.

[33] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

[34] Aditya Modi, Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Joint learning of linear time-invariant dynamical systems. *arXiv preprint arXiv:2112.10955*, 2021.

[35] M Hashem Pesaran. *Time series and panel data econometrics*. Oxford University Press, 2015.

[36] Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618, 2019.

[37] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[38] Anil K Seth, Adam B Barrett, and Lionel Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

[39] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

[40] A Skripnikov and G Michailidis. Joint estimation of multiple network granger causal models. *Econometrics and Statistics*, 10:120–133, 2019.

[41] Andrey Skripnikov and George Michailidis. Regularized joint estimation of related vector autoregressive models. *Computational statistics & data analysis*, 139:164–177, 2019.

[42] James H Stock and Mark W Watson. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics*, volume 2, pages 415–525. Elsevier, 2016.

[43] Sagar Sudhakara, Aditya Mahajan, Ashutosh Nayyar, and Yi Ouyang. Scalable regret for learning to control network-coupled subsystems with unknown dynamics. *IEEE Transactions on Control of Network Systems*, 2022.

[44] Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

[45] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[46] Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Artificial Intelligence and Statistics*, pages 981–990. PMLR, 2017.