Sample Complexity of Sparse System Identification Problem

Salar Fattahi and Somayeh Sojoudi

Abstract

In this paper, we study the system identification problem for sparse linear time-invariant systems. We propose a sparsity promoting block-regularized estimator to identify the dynamics of the system with only a limited number of input-state data samples. We characterize the properties of this estimator under high-dimensional scaling, where the growth rate of the system dimension is comparable to or even faster than that of the number of available sample trajectories. In particular, using contemporary results on high-dimensional statistics, we show that the proposed estimator results in a small element-wise error, provided that the number of sample trajectories is above a threshold. This threshold depends polynomially on the size of each block and the number of nonzero elements at different rows of input and state matrices, but only logarithmically on the system dimension. A by-product of this result is that the number of sample trajectories required for sparse system identification is significantly smaller than the dimension of the system. Furthermore, we show that, unlike the recently celebrated least-squares estimators for system identification problems, the method developed in this work is capable of *exact recovery* of the underlying sparsity structure of the system with the aforementioned number of data samples. Extensive case studies on synthetically generated systems and multi-agent systems are offered to demonstrate the effectiveness of the proposed method.

I. Introduction

With their ever-growing size and complexity, real-world dynamical systems are hard to model. Today's systems are complex and large, often with a massive number of unknown parameters, which render them doomed to the so-called *curse of dimensionality*. Therefore, system operators should rely on simple and tractable estimation methods to identify the dynamics of the system via a limited number of recorded input-output interactions, and then design control policies to ensure the desired behavior of the entire system. The area of *system identification* is created to address this problem [1].

In this work, the objective is to employ modern results on high-dimensional statistics to reduce the sample complexity for one of the most fundamental classes of systems in control theory, namely linear time-invariant (LTI) systems with perfect state measurements. This type of dynamical system forms the basis of many classical control

Email: fattahi@berkeley.edu and sojoudi@berkeley.edu.

Salar Fattahi is with the Department of Industrial Engineering and Operations Research, University of California, Berkeley. Somayeh Sojoudi is with the Departments of Electrical Engineering and Computer Sciences and Mechanical Engineering as well as the Tsinghua-Berkeley Shenzhen Institute, University of California, Berkeley. This work was supported by the ONR Award N00014-18-1-2526 and an NSF EPCN Grant.

problems, such as Linear Quadratic Regulator and Linear Quadratic Gaussian problems. Our results are built upon the fact that, in many practical large-scale systems, the states and inputs exhibit sparse interactions with one another, which in turn translates into a sparse representation of the state-space equations of the system. Driven by the existing non-asymptotic results on the classical Lasso problem, the main focus of this paper is on the block-regularized estimators for the system identification problem, where the goal is to promote sparsity on different blocks of input and state matrices. To this goal, the ℓ_∞-norms of the blocks are penalized instead of their ℓ_1-norms. One motivation behind employing this type of estimator stems from topology extraction in consensus networks, especially in the multi-agent setting [2], [3]. In this problem, given a number of subsystems (agents) whose interactions are defined via an unknown sparse topology network, the objective is to estimate the state-space model governing the entire system based on a limited number of input-output sample trajectories. Since the subsystems have their own local state and input vectors with potentially different sizes, the parameters of the state-space model admit a block-sparse structure.

A. Related Works

Asymptotic Guarantees: System identification is a well-established area of research in control theory, with related preliminary results dating back to 1960s. Standard reference textbooks on the topic include [4]–[7], all focusing on establishing *asymptotic* consistency of different types of estimators (e.g. least-squares, prediction error, and maximum likelihood). Although these results shed light on the theoretical consistency of the existing methodologies, they are not applicable in the finite time/sample settings. In many applications, including neuroscience, transportation networks, and gene regulatory networks, the dimensionality of the system is overwhelmingly large, often surpassing the number of available input-output data [8]–[10]. Under such circumstances, the dynamics of the system should be estimated under the *large dimension-small sample size* regime and classical approaches for checking the asymptotic consistency of an estimator face major breakdowns. Simple examples of such failures are widespread in high-dimensional statistics. For instance, it is well-known that the least-squares estimators, which are widely used in system identification problems, cease to exist uniquely when the sample size is smaller than the dimension of the system [11].

Finite-Time Guarantees: Contemporary results in statistical learning as applied to system identification seek to characterize *finite time and finite data* rates, relying heavily on tools from sample complexity analysis and concentration of measure. Such finite-time guarantees provide estimates of both system parameters and their uncertainty, which allows for a natural bridge to robust/optimal control. In [12], it was shown that under full state observation, if the system is driven by Gaussian noise, the ordinary least squares estimate of the system matrices constructed from independent data points achieves order optimal rates that are linear in the system dimension. This result was later generalized to the single trajectory setting for (i) marginally stable systems in [13], (ii) unstable systems in [14], and (iii) partially observed stable systems in [15]–[18].

Sparse System Identification: Recently, special attention has been devoted to the *sparse* system identification problem, where the states and inputs are assumed to possess localized or low-order interactions. These methods

include, but are not restricted to, selective ℓ_1-regularized estimator [19], identification based on compressive sensing [20], sparse estimation of polynomial system dynamics [21], kernel-based regularization [22], and low rank estimation in frequency domain [23]. On the other hand, with the unprecedented interest in data-driven control approaches, such as model-free reinforcement learning, robust control, and adaptive control [24]–[26], a question arises as to what the minimum number of input-output data samples should be to guarantee a small error in the estimated model. Answering this question has been the subject of many recent studies on the sample complexity of the system identification problem [12], [27]–[30]. Most of these results are tailored to a specific type of dynamics, depend on the stability of the open-loop system, or do not exploit the *a priori* information about the structure of the system.

B. Contributions:

In this work, we introduce a regularized estimator for recovering the true block-sparsity of an LTI system. We show that the proposed estimator is guaranteed to achieve infinitesimal estimation error with a small number of samples. In particular, we use an ℓ_1/ℓ_∞-regularized least-squares estimator, i.e., a least-squares estimator accompanied by a ℓ_∞ regularizer on different blocks, and show that, with an appropriate scaling of the regularization coefficient, $\Omega(k_{\max}(D\log(\bar{n}+\bar{m})+D^2))$ sample trajectories are enough to guarantee a small estimation error with a high probability, where $k_{\max}$ is the maximum number of nonzero elements in the rows of input and state matrices, D is the size of the largest block in these matrices, and $\bar{n}$ and $\bar{m}$ are the number of row blocks in the state and input matrices, respectively. This is a significant improvement over the recently derived sample complexity of $\Omega(n+m)$ for the least-squares estimator (n and m are the state and input dimensions, respectively), in the case where the system is sparse and the sizes of all blocks are small relative to the system dimension. While the traditional Lasso is heavily studied in the literature [31], [32], the high-dimensional behavior of the block-regularized estimator is less known when the dimensions of blocks are arbitrary. The paper [33] analyzes the high-dimensional consistency of this estimator when each block of the regression parameter is a row vector. Furthermore, it assumes that the regression parameter consists of only one column of blocks. In an effort to make these results applicable to the block-sparse system identification problem, we significantly generalize the existing non-asymptotic properties to problems with an arbitrary number of blocks, each with general sizes.

Moreover, we derive upper bounds on the element-wise error of the proposed estimator. In particular, we prove that $\Omega(k_{\max}^2(D\log(\bar{n}+\bar{m})+D^2))$ sample trajectories is enough to ensure that the estimation error decreases at the rate $O(\sqrt{(D\log(\bar{n}+\bar{m})+D^2)/d})$, where d is the number of available sample trajectories. We show that if the number of nonzero elements in the columns (in addition to the rows) of input and state matrices are upper bounded by $k_{\max}$, the operator norm of the estimation error of the proposed estimator is *arbitrarily smaller* than that of its un-regularized least-squares counterpart introduced in [12]. Another advantage of the proposed estimator over its least-squares analog is its *exact recovery* property. More specifically, we show that while the least-squares estimator is unable to identify the sparsity pattern of the input and state matrices for *any* finite number of samples, the proposed estimator recovers the true sparsity pattern of these matrices with a sublinear number of sample

trajectories. It is worthwhile to mention that this work generalizes the results in [29], where the authors use a similar regularized estimator to learn the dynamics of a particular type of systems. However, [29] ignores the block structure of the system and assumes autonomy and inherent stability, all of which will be relaxed in this work. To demonstrate the efficacy of the developed regularized estimator, two case studies are offered on synthetically generated systems and multi-agent systems.

This work is a significant extension of our previous conference papers on Lasso-type estimators for system identification [34] and non-asymptotic analysis of block-regularized linear regression problems [35]. In particular, by combining the properties of the block-regularized regression and the characteristics of LTI systems, we provide a unified sparsity-promoting framework for estimating the parameters of the system with arbitrary block structures. To this goal, we have generalized our theoretical results in [34] and [35] to account for partially-sparse structures. We explain the effect of different parameters of the problem—such as input energy and the length of the time horizon—on the sample complexity of the proposed estimator. Furthermore, it is shown that the required conditions for the validity of the proposed results are not an artifact of the proposed estimator, but are rather inherent to the problem. Based on these results, we introduce a class of k-sparse systems where the conditions of our theorem translate into a set of sufficient and (almost) necessary conditions for the correct recovery of the system dynamics. Furthermore, we relax certain assumptions on the structure of the true system that were initially required in [34], and provide comprehensive discussions and more relevant simulations on the performance of the proposed method.

Notations: For a matrix M, the symbols $\|M\|_F$, $\|M\|_2$, $\|M\|_0$, $\|M\|_1$, and $\|M\|_\infty$ denote its Frobenius, operator, number of nonzero elements, ℓ_1/ℓ_1, and ℓ_∞/ℓ_∞ norms, respectively. Furthermore, $\kappa(M)$ refers to its 2-norm condition number, i.e., the ratio between its maximum and minimum singular values. Given integer sets I and J, the notation M_{IJ} refers to the submatrix of M whose rows and columns are indexed by I and J, respectively. The symbols $M_{:,j}$ and $M_{i,:}$ refer to the j^{th} column and i^{th} row of M, respectively. Given the sequences $f_1(n)$ and $f_2(n)$, the notations $f_1(n) = O(f_2(n))$ and $f_1(n) = \Omega(f_2(n))$ imply that there exist $c_1 < \infty$ and $c_2 > 0$ such that $f_1(n) \le c_1 f_2(n)$ and $f_1(n) \ge c_2 f_2(n)$, respectively. Furthermore, $f_1(n) = \Theta(f_2(n))$ is used to imply that $f_1(n) = O(f_2(n))$ and $f_1(n) = \Omega(f_2(n))$. Finally, $f_1(n) = o(f_2(n))$ is used to show that $f_1(n)/f_2(n) \to 0$ as $n \to \infty$. A zero-mean Gaussian distribution with covariance Σ is shown as $N(0, \Sigma)$. Given a function $f(x)$, the expression $\arg\min f(x)$ refers to its minimizer. For a set $\mathcal{I}$, the symbol $|\mathcal{I}|$ denotes its cardinality.

II. Problem Formulation

Consider the LTI system

$$x[t+1] = Ax[t] + Bu[t] + w[t] \tag{1a}$$

where t is the time step, $A \in \mathbb{R}^{n\times n}$ is the state matrix, and $B \in \mathbb{R}^{n\times m}$ is the input matrix. Furthermore, $x[t] \in \mathbb{R}^n$, $u[t] \in \mathbb{R}^m$, and $w[t] \in \mathbb{R}^n$ are the state, input, and disturbance vectors at time t, respectively. The dimension of the system is defined as $m + n$. It is assumed that the input disturbance vectors are identically distributed and independent (i.i.d.) with distribution $N(0, \sigma_w^2 I)$ across different times. In this work, we assume that the matrices A

and B are sparse and the goal is to estimate them based on a limited number of *sample trajectories*, i.e. a sequence $\{(x^{(i)}[\tau], u^{(i)}[\tau])\}_{\tau=0}^{T}$ with $i = 1, 2, ..., d$, where d is the number of available sample trajectories. The i^{th} sample trajectory $\{(x^{(i)}[\tau], u^{(i)}[\tau])\}_{\tau=0}^{T}$ is obtained by running the system from $t = 0$ to $t = T$ and collecting the input and state vectors. Note that in general, one may consider two general approaches to obtain the sample input-output trajectories for the system identification problem:

Fixed d, and variable T: In this approach, one sets the number of sample trajectories d to a fixed value (e.g., $d = 1$) and instead, chooses a sufficiently long time horizon T to obtain enough information about the dynamics of the system. Notice that *this is only viable when the system is stable*. In other words, one needs to assume that either the system is inherently stable, or there exists an initial stabilizing controller in place to be able to use this approach. Note that this assumption of stability is necessary, as even a simple least-squares estimator may not be consistent if the system has unstable modes [14].

Fixed T, and variable d: In this approach, the length of the time horizon T is fixed and instead, the number of sample trajectories is chosen to be sufficiently large to collect enough information about the dynamics of the system. Notice that in this method, one needs to reset the initial state of the system at the beginning of each sample trajectory. However, unlike the previous method, *its applicability is not contingent upon the stability of the true system*.

Due to the aforementioned theoretical and practical limitations, one can only use the second approach for unstable systems.

Given the sample trajectories $\{(x^{(i)}[\tau], u^{(i)}[\tau])\}_{\tau=0}^{\intercal}$ for $i = 1, 2, ..., d$, one can obtain an estimate of (A, B) by solving the following least-squares optimization problem:

$$\min_{A,B} \sum_{i=1}^{d} \sum_{t=0}^{T-1} \left\| x^{(i)}[t+1] - \left(Ax^{(i)}[t] + Bu^{(i)}[t] \right) \right\|_2^2 \tag{2}$$

In order to describe the behavior of the least-squares estimator, define

$$Y^{(i)} = \begin{bmatrix} x^{(i)}[1]^\intercal \\ \vdots \\ x^{(i)}[T]^\intercal \end{bmatrix}, \quad X^{(i)} = \begin{bmatrix} x^{(i)}[0]^\intercal & u^{(i)}[0]^\intercal \\ \vdots & \vdots \\ x^{(i)}[T-1]^\intercal & u^{(i)}[T-1]^\intercal \end{bmatrix},$$

$$W^{(i)} = \begin{bmatrix} w^{(i)}[0]^\intercal \\ \vdots \\ w^{(i)}[T-1]^\intercal \end{bmatrix}. \tag{3}$$

for every sample trajectory $i = 1, 2, ..., d$. Furthermore, let Y, X, and W be defined as vertical concatenations of $Y^{(i)}$, $X^{(i)}$, and $W^{(i)}$ for $i = 1, 2, ..., d$, respectively. Finally, denote $\Psi = [A \;\; B]^\intercal$ as the unknown system parameter and Ψ^* as its true value. Based on these definitions, it follows from (1) that

$$Y = X \cdot \Psi + W \tag{4}$$

The system identification problem is then reduced to estimating Ψ based on the *observation matrix* Y and the

design matrix X. Consider the following least-squares estimator:

$$\Psi_{\text{ls}} = \arg\min_{\Psi} \|Y - X\Psi\|_F^2 \tag{5}$$

One can easily verify the equivalence of (2) and (5). The optimal solution of (5) can be written as

$$\Psi_{\text{ls}} = (X^\top X)^{-1} X^\top Y = \Psi^* + (X^\top X)^{-1} X^\top W \tag{6}$$

Notice that Ψ_{ls} is well-defined and unique if and only if $X^\top X$ is invertible, which necessitates $d \geq n + m$. The estimation error is then defined as

$$E = \Psi_{\text{ls}} - \Psi^* = (X^\top X)^{-1} X^\top W \tag{7}$$

Thus, one needs to study the behavior of $(X^\top X)^{-1} X^\top W$ in order to control the estimation error of the least-squares estimator. However, since the state of the system at time t is affected by random input disturbances at times $0, 1, \ldots t-1$, the matrices X and W are correlated, which renders (7) hard to analyze. In order to circumvent this issue, [12] simplifies the estimator and considers only the state of the system at time T in $Y^{(i)}$. By ignoring the first $T-1$ rows in $Y^{(i)}$, $X^{(i)}$, and $W^{(i)}$, one can ensure that the random matrix $(X^\top X)^{-1} X^\top$ is independent of W. Therefore, it is assumed in the sequel that

$$Y = \begin{bmatrix} x^{(1)}[T]^\top \\ \vdots \\ x^{(d)}[T]^\top \end{bmatrix}, \qquad X = \begin{bmatrix} x^{(1)}[T-1]^\top & u^{(1)}[T-1]^\top \\ \vdots & \vdots \\ x^{(d)}[T-1]^\top & u^{(d)}[T-1]^\top \end{bmatrix},$$
$$W = \begin{bmatrix} w^{(1)}[T-1]^\top \\ \vdots \\ w^{(d)}[T-1]^\top \end{bmatrix} \tag{8}$$

With this simplification, [12] shows that, with input vectors $u^{(i)}[t]$ chosen randomly from $N(0, \sigma_u^2 I)$ for every $t = 1, 2, \ldots, T-1$ and $i = 1, 2, \ldots, d$, the least-squares estimator requires at least $d = \Omega(m + n + \log(1/\delta))$ sample trajectories to guarantee $\|E\|_2 = \mathcal{O}\left(\sqrt{(m+n)\log(1/\delta)/d}\right)$ with probability of at least $1 - \delta$. In what follows, a block-regularized estimator will be introduced that exploits the underlying sparsity structure of the system dynamics to significantly reduce the number of sample trajectories for an accurate estimation of the parameters. To streamline the presentation, the main technical proofs are deferred to Section IV.

Remark 1. *We assume that the covariance matrices for the input and the disturbance noise have diagonal structure (shown as $\sigma_u^2 I$ and $\sigma_w^2 I$, respectively), which implies that there is no dependency between different elements of the input and disturbance vectors. This assumption is made without loss of generality to simplify the presentation of the technical results of this paper. Indeed, these covariance matrices can be replaced by Σ_u and Σ_w without substantially affecting the findings of this work.*

III. Main Results

Suppose that A and B can be partitioned as $A = [A^{(i,j)}]$ and $B = [B^{(k,l)}]$ where $(i,j) \in \{1, \ldots, \bar{n}\} \times \{1, \ldots, \bar{n}\}$ and $(k,l) \in \{1, \ldots, \bar{n}\} \times \{1, \ldots, \bar{m}\}$. $A^{(i,j)}$ is the $(i,j)^{\text{th}}$ block of A with size $n_i \times n_j$. Similarly, $B^{(k,l)}$ is the $(k,l)^{\text{th}}$ block

of B with size $n_k \times m_l$. Note that $\sum_{i=1}^{\bar{n}} n_i = n$ and $\sum_{i=1}^{\bar{m}} m_i = m$. Suppose that it is known *a priori* that all elements in each block $A^{(i,j)}$ or $B^{(k,l)}$ are simultaneously zero or nonzero. This implies that, as long as one element in $A^{(i,j)}$ or $B^{(k,l)}$ is nonzero, there is no reason to promote sparsity in the remaining elements of the corresponding block. Clearly, this kind of block-sparsity constraint is not correctly reflected in (2). To simplify the presentation, we use the notation $\Psi = [A \quad B]^\top$. Note that $\Psi^{(i,j)} = (A^{(j,i)})^\top$ for $i \in \{1, ..., \bar{n}\}$ and $\Psi^{(i,j)} = (B^{(j,i-\bar{n})})^\top$ for $i \in \{\bar{n}+1, ..., \bar{n}+\bar{m}\}$. In order to recover the true block-sparsity of A and B, one can resort to an ℓ_1/ℓ_∞ variant of the Lasso problem—known as the block-regularized least-squares (or simply block-regularized) problem:

$$\hat{\Psi} = \arg\min_{\Psi} \frac{1}{2d}\|Y - X\Psi\|_F^2 + \lambda_d \|\Psi\|_{\text{block}} \tag{9}$$

where $\|\Psi\|_{\text{block}}$ is defined as the summation of $\|\Psi^{(i,j)}\|_\infty$ over $(i,j) \in \{1, ..., \bar{n}+\bar{m}\} \times \{1, ..., \bar{n}\}$. D is used to denote the maximum size of the blocks of Ψ. Under the sparsity assumption on (A, B), we will show that the non-asymptotic statistical properties of $\hat{\Psi}$ significantly outperform those of Ψ_{ls}. In particular, the primary objective is to prove that $\|\hat{\Psi} - \Psi^*\|_\infty$ decreases at the rate $\mathcal{O}(\sqrt{D\log(n+m) + D^2\log(1/\delta)/d})$ with probability of at least $1-\delta$ and with an appropriate scaling of the regularization coefficient, provided that $d = \Omega\left(k_{\max}^2\left(D\log(\bar{n}+\bar{m}) + D^2\log(1/\delta)\right)\right)$. Here, $k_{\max}$ is the maximum number of nonzero elements in the columns of $[A \quad B]^\top$. Comparing this number with the required lower bound $\Omega(n+m+\log(1/\delta))$ on the number of sample trajectories for the least-squares estimator, we conclude that the proposed method needs significantly fewer samples when A and B are sparse. The third objective is to prove that this method is able to find the correct block-sparsity structure of A and B with high probability. In contrast, it will be shown that the solution of the least-squares estimator is fully dense for any finite number of sample trajectories, and hence, it cannot correctly extract the sparsity structures of A and B. We will showcase the superior performance of the block-regularized estimator both in sparsity identification and estimation accuracy in simulations.

To present the main results of this work, first note that

$$\begin{aligned} & x^{(i)}[T-1] \\ & = A^{T-2}Bu^{(i)}[0] + A^{T-3}Bu^{(i)}[1] + \cdots + Bu^{(i)}[T-2] \\ & + A^{T-2}w^{(i)}[0] + A^{T-3}w^{(i)}[1] + \cdots + w^{(i)}[T-2] \end{aligned} \tag{10}$$

where, without loss of generality, the initial state is assumed to be zero for every sample trajectory. The results can be readily extended to the case where the initial state is an unknown random vector with Gaussian distribution. Suppose that $u^{(i)}[t]$ and $w^{(i)}[t]$ are i.i.d samples of $N(0, \sigma_u^2 I)$ and $N(0, \sigma_w^2 I)$, respectively. Therefore, (10) and (8) imply that

$$X_{i,:}^\top \sim N\left(0, \tilde{\Sigma}\right) \tag{11}$$

where $X_{i,:}$ is the i^{th} row of X and

$$\tilde{\Sigma} = \begin{bmatrix} C^\top C & 0 \\ 0 & \sigma_u^2 I \end{bmatrix} \tag{12a}$$

$$C = \begin{bmatrix} \sigma_u F_T^\top \\ \sigma_w G_T^\top \end{bmatrix} \tag{12b}$$

$$F_T = \begin{bmatrix} A^{T-2}B & A^{T-3}B & \dots & B \end{bmatrix} \tag{12c}$$

$$G_T = \begin{bmatrix} A^{T-2} & A^{T-3} & \dots & I \end{bmatrix} \tag{12d}$$

The matrix C is referred to as the *combined controllability matrix* in the sequel. Define $\mathcal{A}_j(\Psi) = \{i : \Psi^{(i,j)} \neq 0\}$. Unless stated otherwise, $\mathcal{A}_j$ is used to refer to $\mathcal{A}_j(\Psi^*)$. Define $\mathcal{A}_j^c$ as the complement of $\mathcal{A}_j$. For $\mathcal{T} \subseteq \{1, ..., \bar{n}+\bar{m}\}$, denote $I(\mathcal{T})$ as the index set of rows in Ψ^* corresponding to the blocks $\{\Psi^{*(i,:)} : i \in \mathcal{T}\}$. For an index set $\mathcal{U}$, define $X_{\mathcal{U}}$ as a $d \times |\mathcal{U}|$ submatrix of X after removing the columns with indices not belonging to $\mathcal{U}$. With a slight abuse of notation, $X_{(i)}$, $X_{\mathcal{A}_j}$, and $X_{\mathcal{A}_j^c}$ are used to denote $X_{I(\{i\})}$, $X_{I(\mathcal{A}_j)}$, and $X_{I(\mathcal{A}_j^c)}$ when there is no ambiguity. Similarly, $\tilde{\Sigma}_{(i),\mathcal{A}_j}$ and $\tilde{\Sigma}_{\mathcal{A}_j,\mathcal{A}_j}$ are used in lieu of $\tilde{\Sigma}_{I(\{i\}),I(\mathcal{A}_j)}$ and $\tilde{\Sigma}_{I(\mathcal{A}_j),I(\mathcal{A}_j)}$, respectively. Denote k_j as the maximum number of nonzero elements in any column of $\Psi^{*(:,j)}$ which is the j^{th} block column of Ψ^*. Finally, define

$$\begin{aligned} n_{\max} &= \max_{1\le i\le \bar{n}} n_i, & m_{\max} &= \max_{1\le i\le \bar{m}} m_i, \\ p_{\max} &= \max\{n_{\max}, m_{\max}\}, & k_{\max} &= \max_{1\le j\le \bar{n}} k_j, \\ \sigma_{\max}^2 &= \max_{1\le i\le n+m} \tilde{\Sigma}_{ii} \end{aligned} \tag{13}$$

The following set of assumptions plays a key role in deriving the main result of this paper:

Assumption 1. By fixing the time horizon T, we assume that the following conditions hold for all finite system dimensions:

A1. (Mutual Incoherence Property): There exists a number $\gamma \in (0, 1]$ such that

$$\max_{j=1,\dots,\bar{n}} \left\{ \max_{i\in\mathcal{A}_j^c} \left\| \tilde{\Sigma}_{(i),\mathcal{A}_j} (\tilde{\Sigma}_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right\|_1 \right\} \le 1 - \gamma \tag{14}$$

A2. (Bounded eigenvalue): There exist numbers $0 < \Lambda_{\min} < \infty$ and $0 < \Lambda_{\max} < \infty$ such that

$$\Lambda_{\min} \le \lambda_{\min}(\tilde{\Sigma}) \le \lambda_{\max}(\tilde{\Sigma}) \le \Lambda_{\max} \tag{15}$$

A3. (Bounded minimum value): There exists a number $t_{\min} > 0$ such that

$$t_{\min} \le \min_{1\le j\le \bar{n}} \min_{i\in\mathcal{A}_j} \left\| \Psi^{*(i,j)} \right\|_\infty \tag{16}$$

A4. (Block sizes): There exist numbers $\alpha_n, \alpha_m < \infty$ such that

$$n_{\max} = O\left((\bar{n} + \bar{m})^{\alpha_n}\right) \tag{17a}$$

$$m_{\max} = O\left((\bar{n} + \bar{m})^{\alpha_m}\right) \tag{17b}$$

The mutual incoherence property in Assumption A1 is a commonly known condition for the exact recovery of unknown parameters in compressive sensing and classical Lasso problems [32], [36]–[38]. This assumption entails that the effect of those submatrices of $\tilde{\Sigma}$ corresponding to zero (unimportant) elements of Ψ on the remaining entries of $\tilde{\Sigma}$ should not be large. Roughly speaking, this condition guarantees that the unknown parameters are *recoverable* in the noiseless scenario, i.e. when $W = 0$. If the recovery cannot be guaranteed in the noise-free setting, then there is little hope for the block-regularized estimator to recover the true structure of A and B when the system is subject to noise. This assumption is satisfied in all of our simulations.

The bounded eigenvalue condition in Assumption A2 entails that the condition number of $\tilde{\Sigma}$ is bounded away from 0 and ∞ for all finite system dimensions. Assuming that the values σ_u and σ_w do not scale with the system dimension, it is easy to verify that $\min\{\sigma_u^2, \sigma_w^2\} \leq \Lambda_{\min} \leq \sigma_u^2$. However, as will be shown later, the value of $\Lambda_{\max}$ can change with respect to the time horizon T. In particular, it will be later shown that for highly unstable systems, $\tilde{\Sigma}$ becomes severely ill-conditioned as the time horizon increases, which in turn makes the system identification problem difficult to solve. Furthermore, this assumption implies that there exists a constant $\bar{\sigma}_{\max}^2 < \infty$ such that $\max_{1\leq i\leq n+m} \Sigma_{ii} \leq \bar{\sigma}_{\max}^2$ for every finite system dimension.

Assumption A3 implies that, independent of the system dimensions, there always exists a strictly positive gap between the zero and nonzero elements of A and B. This assumption holds in almost all practical settings and will facilitate the exact sparsity recovery of the parameters of the system.

Finally, Assumption A4 requires that the maximum size of the blocks in Ψ^* be polynomially bounded by the number of its block columns. For instance, $\bar{n} = O(1)$ and $\bar{m} = O(1)$ violate this assumption since it implies that $n_{\max} = \Omega\big((\bar{n}+\bar{m})^{\log n}\big)$ and $m_{\max} = \Omega\big((\bar{n}+\bar{m})^{\log m}\big)$. It is worthwhile to mention that Assumption A4 results in $k_{\max} = O\big((\bar{n}+\bar{m})^{\alpha_k}\big)$ for some number $\alpha_k < \infty$; this will be used later in the paper.

Remark 2. *Note that, due to Assumption A2, $\kappa(\tilde{\Sigma}) = O(1)$. However, this quantity will not be removed from the big-O analysis of our subsequent theorems and corollaries to demonstrate its effect on the high-dimensional properties of the developed estimator.*

Define $D = p_{\max} n_{\max}$, which is the maximum size of the blocks in Ψ.

Theorem 1 (block-wise regularization)**.** *Upon choosing*

$$\lambda_d = \Theta\left(\sigma_{\max}\sqrt{\frac{D\log(\bar{n}+\bar{m}) + D^2\log(1/\delta)}{d}}\right) \tag{18a}$$

$$d = \Omega\left(\kappa(\tilde{\Sigma})^2 k_{\max}\left(D\log(\bar{n}+\bar{m}) + D^2\log(1/\delta)\right)\right) \tag{18b}$$

the following statements hold with probability of at least $1-\delta$:

1. *$\hat{\Psi}$ is unique and has the same nonzero blocks as Ψ^*.*

2. *We have*

$$\begin{aligned} g &= \|\hat{\Psi}-\Psi^*\|_\infty \\ &= O\Bigg(\kappa(\tilde{\Sigma})\left(1+\sqrt{\frac{k_{\max}(k_{\max}n_{\max}+\log(\bar{n}+\bar{m})+\log(1/\delta))}{d}}\right) \\ &\qquad \times\sqrt{\frac{D\log(\bar{n}+\bar{m})+D^2\log(1/\delta)}{d}}\Bigg) \end{aligned} \tag{19}$$

Theorem 1 shows that the minimum number of required sample trajectories is a quadratic function of the maximum block size. Therefore, only a small number of samples is enough to guarantee the uniqueness, exact block-sparsity recovery, and small estimation error for sparse systems, assuming that the sizes of the blocks are significantly smaller than the system dimensions.

Corollary 1. *Assume that* $n_{\max} = O(n^{\beta_n})$ *and* $m_{\max} = O(m^{\beta_m})$ *for some* $\beta_n > 0$ *and* $\beta_m > 0$*. Then,*

$$\lambda_d = \Theta\left(\sigma_{\max}(n+m)^{(\beta_n+\beta_m)}\sqrt{\frac{\log(1/\delta)}{d}}\right) \tag{20a}$$

$$d = \Omega(\kappa(\tilde{\Sigma})^2 k_{\max}^2 (n+m)^{2(\beta_n+\beta_m)}\log(1/\delta)) \tag{20b}$$

is enough to guarantee the exact sparsity recovery of Ψ^* *and*

$$\|\hat{\Psi}-\Psi^*\|_\infty = O\left(\kappa(\tilde{\Sigma})(n+m)^{(\beta_n+\beta_m)}\sqrt{\frac{\log(1/\delta)}{d}}\right) \tag{21}$$

with probability of at least $1-\delta$*.*

Proof. The proof follows from Theorem 1. The details are omitted for brevity. □

Corollary 1 analyzes the behavior of the proposed estimator for the *polynomial scaling* of the block size. It can be seen that the size of the required sample trajectories heavily depends on the growth rate of the maximum block size of Ψ. Although the sampling rate is still sublinear when $\beta_n + \beta_m < 1/2$, it may surpass the system dimension if $\beta_n + \beta_m > 1/2$. A question arises as to whether one can resort to the ordinary least-squares estimator in lieu of the proposed block-regularized estimator for the cases where $\beta_n + \beta_m > 1/2$ since the proposed estimator requires $d = \Omega((n+m)^{1+\epsilon}\log(1/\delta))$ for some $\epsilon > 0$ whereas $d = \Theta(n+m+\log(1/\delta))$ is enough to guarantee the uniqueness of the least-squares estimator. This will be addressed in the next subsection.

A. Comparison to Least-Squares

In this subsection, we will prove that the least-squares estimator does not extract the correct sparsity structure of Ψ for *any finite number* of sample trajectories.

Theorem 2. *If* A *and* B *are not fully dense matrices,* Ψ_{ls} *does not recover the support of* Ψ^* *for any finite number of sample trajectories with probability 1.*

Proof. Define $R = ((X^\top X)^{-1}X^\top)^\top$, and note that R and W are independent random variables due to the construction of X. Now, suppose that $\Psi^*_{ij} = 0$. We show that, with probability zero, $E_{ij} = |(\Psi_{\text{ls}})_{ij} - \Psi^*_{ij}| = 0$

holds. Note that $E_{ij} = R_{:,i}^\top W_{:,j}$. If $R_{:,i} \neq 0$, then E_{ij} is a linear combination (with at least one nonzero coefficient) of identically distributed normal random variables with mean zero and variance $(\Sigma_w)_{jj}$. Since $R_{:,i}$ and $W_{:,j}$ are independent, we have $E_{ij} = 0$ with probability zero. Now, assume that $R_{:,i} = 0$. This means that the i^{th} row of $R^\top$ is a zero vector. This, in turn, implies that the i^{th} row of $R^\top X$ is zero. However, $R^\top X = (X^\top X)^{-1} X^\top X = I$, which is a contradiction. This completes the proof. □

Define $h(n,m) = \sqrt{(n+m)\log(1/\delta)/d}$ and recall that $\|\Psi_{\text{ls}} - \Psi^*\|_2 = O(h(n,m))$. In the next corollary, we show that, under additional sparsity conditions, the operator norm of the estimation error for $\hat{\Psi}$ becomes arbitrarily smaller than $h(n,m)$ as the system dimension grows.

Corollary 2. *Assume that the number of nonzero elements at different rows and columns of Ψ^* is upper bounded by $k_{\max}$. Furthermore, suppose that λ_d satisfies* (18a) *and*

$$d = \Omega\left(\kappa(\tilde{\Sigma})^2 k_{\max}^2 \left(D\log(\bar{n}+\bar{m}) + D^2\log(1/\delta)\right)\right) \tag{22}$$

Then, we have

$$\|\hat{\Psi} - \Psi^*\|_2 = O\Bigg(\underbrace{\kappa(\tilde{\Sigma}) k_{\max}\sqrt{\frac{D\log(\bar{n}+\bar{m}) + D^2\log(1/\delta)}{d}}}_{v(n,m)}\Bigg) \tag{23}$$

with probability of at least $1-\delta$. Furthermore, we have

$$\frac{v(n,m)}{h(n,m)} \to 0 \quad \textit{as} \quad (n,m) \to \infty \tag{24}$$

provided that

$$k_{\max} D = o\left(\sqrt{\frac{n+m}{\log(n+m)}}\right) \tag{25}$$

Proof. One can use Holder's inequality to write

$$\|\hat{\Psi} - \Psi^*\|_2 \leq \sqrt{\|\hat{\Psi} - \Psi^*\|_1 \|\hat{\Psi} - \Psi^*\|_\infty} \leq k_{\max}\|\hat{\Psi} - \Psi^*\|_\infty \tag{26}$$

On the other hand, under (22), it can be verified that

$$\sqrt{\frac{k_{\max}(k_{\max} n_{\max} + \log(\bar{n}+\bar{m}) + \log(1/\delta))}{d}} = O(1) \tag{27}$$

Combined with (26) and Theorem 1, this certifies the validity of (23). It remains to prove the correctness of (24). Note that under (25), we have

$$k_{\max}^2 D\log(\bar{n}+\bar{m}) = o(n+m) \tag{28a}$$

$$k_{\max}^2 D^2 = o(n+m) \tag{28b}$$

Combined with the definitions of $h(n,m)$ and $v(n,m)$, this completes the proof. □

Corollary 2 describes the settings under which our proposed method significantly outperforms the least-squares estimator in terms of the operator norm of the errors. This improvement is more evident for those systems where the states and inputs have sparse interactions and the block sizes in A and B are smaller than the system dimensions.

A class of such systems is multi-agent networks where the agents interact only locally and their total number dominates the dimension of each individual agent.

B. Controllability and the Effect of T

Notice that the minimum number of required sample trajectories and the element-wise error of the estimated parameters depend on $\kappa(\tilde{\Sigma})$. Recall that $\min\{\sigma_w^2, \sigma_u^2\} \leq \Lambda_{\min} \leq \sigma_w^2$, independent of T. Therefore, the value of $\kappa(\tilde{\Sigma})$ is governed by the maximum eigenvalue of $C^\top C$. Roughly speaking, $\lambda_{\max}(C^\top C)$ quantifies the easiest-to-identify mode of the dynamical system. Therefore, Theorem 1 imply that the sample complexity of the proposed block-regularized estimator depends on the modes of the system, as well as the *expected energy of the input and disturbance noise*. In particular, by fixing σ_u and σ_w, only a small number of samples is required to accurately identify the dynamics of the system if all of its modes are easily excitable. The dependency of the estimation error on the modes of the system is also reflected in the non-asymptotic error bound of the least-squares estimator in [12]. This is completely in line with the conventional results on the identifiability of dynamical systems: independent of the method in use, it is significantly harder to identify the parameters of the system accurately if it possesses nearly-hidden modes.

Furthermore, notice that F_T, G_T, and, hence, $\lambda_{\max}(C^\top C)$ depend directly on the length of the time horizon T for each sample trajectory. In what follows, we will show that for highly unstable systems, $\lambda_{\max}(C^\top C)$ can grow *exponentially fast* in terms of T and, hence, short sample trajectories are more desirable in estimating the parameters of such unstable systems. To better understand this, assume that the spectral radius of A—shown as $\rho(A)$—is greater than one, it is diagonalizable, and n is fixed. One can easily verify that the following chain of inequalities holds:

$$\begin{aligned}
\lambda_{\max}(\tilde{\Sigma}) &\geq \lambda_{\max}(\sigma_u^2 F_T F_T^\top + \sigma_w^2 G_T G_T^\top) \\
&\geq \max_i \left\{ \left(\left(A^{T-2} \left(A^{T-2} \right)^\top \right)_{ii} \right)^2 \right\} \\
&\geq \frac{1}{n} \| A^{T-2} \|_\infty \\
&\geq \frac{1}{n} \rho(A)^{T-2}
\end{aligned} \tag{29}$$

This exponential dependency is also empirically observed in our numerical experiments. Furthermore, the connection between the identifiability of the system and the number of required sample trajectories to guarantee a small estimation error will be elaborated through different case studies in Section V.

C. Mutual Incoherency

In what follows, we will analyze the Assumption A1 about the mutual incoherency of the covariance matrix $\tilde{\Sigma}$. In particular, we will show that the possible limitations that arise from Assumption A1 are not artifacts of the proposed method, but rather stem from a fundamental limitation of *any* sparsity recovery technique for the system identification problem that is based on sparsity promoting techniques. For simplicity of the subsequent arguments,

assume that the size of each block is equal to 1, and that the oracle estimator can measure the disturbance matrix W. Furthermore, suppose that the estimator can collect and work with an infinite number of sample trajectories. Under these assumptions, the oracle estimator should solve the following optimization problem to estimate the parameters of the system:

$$\min_{\Psi} \quad \|\Psi\|_0 \tag{30a}$$

$$\text{s.t.} \quad X\Psi = Y - W \tag{30b}$$

Notice that the oracle estimator cannot be obtained in practice since: 1) the exact value of the disturbance noise is not available, 2) only a finite number of sample trajectories can be collected, and 3) the corresponding optimization is non-convex and NP-hard in its worst case.

As mentioned before, there are fundamental limits on the performance of the introduced oracle estimator. To explain this, we introduce the mutual-coherence metric for a matrix. For a given matrix $A \in \mathbb{R}^{t_1 \times t_2}$, its mutual-coherence $\mu(A)$ is defined as

$$\mu(A) = \max_{1 \le i < j \le t_2} \frac{|A_{:,i}^\top A_{:,j}|}{\|A_{:,i}\|_2 \|A_{:,j}\|_2} \tag{31}$$

In other words, $\mu(A)$ measures the maximum correlation between distinct columns of A (with a slight abuse of notation, we assume that $\frac{1}{\mu(A)} = +\infty$ if $\mu(A) = 0$). Reminiscent of the classical results in the compressive sensing literature, it is well-known that the optimal solution Ψ^* of (30) is unique if the *identifiability* condition

$$\|\Psi^*_{:,j}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right) \tag{32}$$

holds for every $j = 1, 2, ..., n$ (see, e.g., Theorem 2.5 in [39]). Furthermore, this bound is tight, implying that there exists an instance of the problem for which the violation of $\|\Psi^*_{:,j}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right)$ for some j results in the non-uniqueness of the optimal solution. On the other hand, one can invoke the Central Limit Theorem to show that $\frac{1}{d}X^\top X = \tilde{\Sigma}$ almost surely as $d \to \infty$. Furthermore, recall the definition of the combined controllability matrix C in 12b. This, together with the definition of $\tilde{\Sigma}$, implies that

$$\begin{aligned} \mu(X) &= \max_{1 \le i < j \le m+n} \frac{|X_{:,i}^\top X_{:,j}|}{\|X_{:,i}\|_2 \|X_{:,j}\|_2} \\ &= \max_{1 \le i < j \le n} \frac{|C_{:,i}^\top C_{:,j}|}{\|C_{:,i}\|_2 \|C_{:,j}\|_2} = \mu(C) \end{aligned} \tag{33}$$

According to the above equality, the correlation between different columns of C plays a crucial role in the identifiability of the true parameters: as $\mu(C)$ becomes smaller, the oracle estimator can correctly identify the structure of Ψ for a wider range of sparsity levels.

Revisiting Assumption A1, one can verify that the mutual incoherency condition is reduced to the following inequality when the size of each block is equal to one:

$$\left\| (C_{:,\mathcal{A}_j}^\top C_{:,\mathcal{A}_j})^{-1} C_{:,\mathcal{A}_j}^\top C_{:,k} \right\|_1 \le 1 - \alpha, \qquad \forall k \in \mathcal{A}_j^c,\ j = 1, 2, \ldots, n \tag{34}$$

where, with a slight abuse of notation, we use $\mathcal{A}_j$ to denote the set $\{i : A_{ij} \neq 0\}$. Notice that, similar to (32), the above condition is expected to be satisfied when different columns of C are nearly orthogonal, i.e., when the elements in $C_{:,\mathcal{A}_j}^\top C_{:,k}$ have small magnitudes. In particular, we introduce a class of k-sparse dynamical systems for which the above condition is equivalent to (32) (modulo a constant factor).

k-sparse systems: Consider a class of problems where each row or column of A has at most k nonzero entries and B is diagonal. Without loss of generality and to simplify the subsequent derivations, suppose that the following assumptions hold:

- B is equal to identity matrix and diagonal entries of A are equal to 1. Moreover, the magnitude of each off-diagonal entry of A is upper bounded by $\varphi > 0$.
- T is set to 3.
- σ_u and σ_w are less than or equal to 1.

Proposition 1. *For k-sparse systems with $k \geq 3$, the following statements hold:*

- *There exists an instance for which the identifiability condition fails to hold for the oracle estimator if $\varphi \geq \frac{3}{k}$.*
- *The mutual incoherency condition holds if $\varphi < \frac{\sigma_u+\sigma_w}{9k}$.*

Proof. The first statement can be easily verified. To prove the second statement, it suffices to provide separate upper bounds for $\left\|(C_{:,\mathcal{A}_j}^\top C_{:,\mathcal{A}_j})^{-1}\right\|_1$ and $\left\|C_{:,\mathcal{A}_j}^\top C_{:,k}\right\|_1$. In particular, one can verify that the $\left\|(C_{:,\mathcal{A}_j}^\top C_{:,\mathcal{A}_j})^{-1}\right\|_1$ is upper bounded by $\frac{1}{(\sigma_u+\sigma_w)-3(k-1)\varphi}$ after controlling different terms of its Taylor expansion. Similarly, $\left\|C_{:,\mathcal{A}_j}^\top C_{:,k}\right\|_1$ is upper bounded by $3(\sigma_u+\sigma_w)k\varphi$. Combining these bounds implies that (34) holds for a strictly positive α, provided that $\varphi < \frac{\sigma_u+\sigma_w}{9k}$. The details are omitted for brevity. □

The above proposition shows that, for this class of dynamical systems, the mutual incoherency is at most a constant factor away from the aforementioned identifiability condition for the oracle estimator, confirming the non-conservativeness of the proposed condition.

IV. Proofs

A number of preliminary definitions and lemmas are required to present the proof of Theorem 1.

Definition 1 (sub-Gaussian random variable)**.** A zero-mean random variable x is *sub-Gaussian* with parameter σ^2 if there exists a constant number $c < \infty$ such that

$$\mathbb{P}(|x| > t) \leq c \cdot \exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{35}$$

Lemma 1. *Given a set of zero-mean sub-Gaussian variables x_i with parameters σ_i for $i = 1, 2, ..., m$, the inequality*

$$\mathbb{P}\left(\max_i |x_i| > t\right) \leq c \cdot \exp\left(-\frac{t^2}{2\max_i \sigma_i^2} + \log m\right) \tag{36}$$

holds for some constant $c < \infty$.

Define I_d as the $d \times d$ identity matrix. The next two lemmas are borrowed from [33] and [31], respectively.

Lemma 2. *Given a set of random vectors $X_i \sim N(0, \sigma_i^2 I_d)$ for $i = 1, 2, ..., m$ and $d > 2 \log m$, the inequality*

$$\mathbb{P}\left(\max_i \|X_i\|_2^2 \geq 4\sigma^2 d\right) \leq \exp\left(-\frac{d}{2} + \log m\right) \tag{37}$$

holds, where $\sigma = \max_i \sigma_i$.

Lemma 3. *Consider a matrix $X \in \mathbb{R}^{m \times n}$ whose rows are drawn from $N(0, \Sigma)$. Assuming that $n \leq m$, we have*

$$\mathbb{P}\left(\left\|\left(\frac{1}{d}X^\top X\right)^{-1} - \Sigma^{-1}\right\|_2 \geq \frac{8}{\Lambda_{\min}}\sqrt{\frac{t}{m}}\right) \leq 2\exp\left(-\frac{t}{2}\right) \tag{38}$$

for every $n \leq t \leq m$.

The basic inequalities given below will be used frequently in our subsequent arguments.

Lemma 4. *The following statements hold true:*

- *Given a number of (not necessarily independent) events $\mathcal{T}_i$ for $i = 1, 2, ..., n$, the following inequality is satisfied:*

$$\sum_{i=1}^{n} \mathbb{P}(\mathcal{T}_i) - (n-1) \leq \mathbb{P}(\mathcal{T}_1 \cap \mathcal{T}_2 \cap ... \cap \mathcal{T}_n) \tag{39}$$

- *Given events $\mathcal{B}$ and $\mathcal{C}$ together with the complement of $\mathcal{C}$, denoted as $\mathcal{C}^c$, the following inequality holds:*

$$\mathbb{P}(\mathcal{B}) \leq \mathbb{P}(\mathcal{B}|\mathcal{C}) + \mathbb{P}(\mathcal{C}^c) \tag{40}$$

The next lemma characterizes the first-order optimality conditions for (9).

Lemma 5 (KKT conditions)**.** *$\hat{\Psi}$ is an optimal solution for* (9) *if and only if it satisfies*

$$\frac{1}{d}X^\top X(\hat{\Psi} - \Psi^*) - \frac{1}{d}X^\top W + \lambda_d \hat{S} = 0 \tag{41}$$

for some $\hat{S} \in \mathbb{R}^{(n+m)\times n} \in \partial\|\hat{\Psi}\|_{\text{block}}$, where $\partial\|\hat{\Psi}\|_{\text{block}}$ denotes the sub-differential of $\|\cdot\|_{\text{block}}$ at $\hat{\Psi}$.

Proof. The proof is straightforward and omitted for brevity. □

$\hat{S}_{\mathcal{A}}$ and $\hat{S}_{\mathcal{A}^c}$ are obtained by removing those blocks of $\hat{S}$ with indices not belonging to $\mathcal{A}$ and $\mathcal{A}^c$, respectively. The equation (4) can be reformulated as the set of linear equations

$$Y^{(:,j)} = X\Psi^{(:,j)} + W^{(:,j)} \quad \forall j \in \{1, ..., \bar{n}\} \tag{42}$$

where $Y^{(:,j)}$, $\Psi^{(:,j)}$, and $W^{(:,j)}$ are the j^{th} block column of Y, Ψ, and W, respectively. Based on this definition, consider the following set of block-regularized subproblems:

$$\hat{\Psi}^{(:,j)} = \arg\min \frac{1}{2d}\|Y^{(:,j)} - X\Psi^{(:,j)}\|_2^2 + \lambda_d \|\Psi^{(:,j)}\|_{\text{block}} \tag{43}$$

Define $D_j = p_{\max} n_j$. The next two lemmas are at the core of our proof for Theorem 1.

Lemma 6 (No false positives)**.** *Given arbitrary constants $c_1, c_2 > 1$, suppose that λ_d and d are chosen such that*

$$\lambda_d \geq \sqrt{\frac{32 c_1 \sigma_w^2 \sigma_{\max}^2}{\gamma^2} \cdot \frac{(D_j)^2 + D_j \log(\bar{n} + \bar{m})}{d}} \tag{44a}$$

$$d \geq \frac{72 c_2 \sigma_{\max}^2}{\gamma^2 \Lambda_{\min}} \cdot k_j (D_j^2 + D_j \log(\bar{n} + \bar{m})) \tag{44b}$$

Then, with probability of at least

$$\begin{aligned} 1 - 3\exp\big(-(c_1-1)(D_j+\log(\bar{n}+\bar{m}))\big) \\ -4\exp\big(-(c_2-1)(D_j+\log(\bar{n}+\bar{m}))\big) \end{aligned} \tag{45}$$

$\hat{\Psi}^{(:,j)}$ *is unique and its nonzero blocks exclude the zero blocks of* $\Psi^{*(:,j)}$*. In other words,* $\hat{\Psi}^{(:,j)}$ *does not have any false positives.*

Recall that due to Assumption A4, one can write $n_{\max} = O\left((\bar{n}+\bar{m})^{\alpha_n}\right)$ and $k_{\max} = O\left((\bar{n}+\bar{m})^{\alpha_k}\right)$ for some $\alpha_n \geq 0$ and $\alpha_k \geq 0$.

Lemma 7 (Element-wise error)**.** *Given arbitrary constants* $c_3 > 0$ *and* $c_4 > 1$*, suppose that* $\hat{\Psi}$ *is unique and the set of its nonzero blocks excludes the zero blocks of* Ψ^**. Then, with probability of at least*

$$\begin{aligned} 1 - 2\exp(-(k_j n_j + c_3\log(\bar{n}+\bar{m}))/2) - 2\exp(-d/2) \\ -2\exp\big(-2(c_4-1)(\alpha_n+\alpha_k)\log(\bar{n}+\bar{m})\big)\big) \end{aligned} \tag{46}$$

we have

$$\begin{aligned} \|\hat{\Psi}^{(:,j)} - \Psi^{*(:,j)}\|_\infty \leq \sqrt{\frac{36c_4(\alpha_n+\alpha_k)\sigma_w^2\log(\bar{n}+\bar{m})}{\Lambda_{\min} d}} \\ + \frac{\lambda_d}{\Lambda_{\min}}\left(8\sqrt{k_j}\sqrt{\frac{k_j n_j + c_3\log(\bar{n}+\bar{m})}{d}} + 1\right) = g_j \end{aligned} \tag{47}$$

Furthermore, the zero blocks of $\hat{\Psi}^{(:,j)}$ *exclude the nonzero blocks of* $\Psi^{*(:,j)}$ *if* $\min_{i\in\mathcal{A}_j}\|\Psi^{(i,j)}\|_\infty > g_j$*. In other words,* $\hat{\Psi}^{(:,j)}$ *does not have any false negatives if* $\min_{i\in\mathcal{A}_j}\|\Psi^{(i,j)}\|_\infty > g_j$*.*

In what follows, we will present some preliminaries that are essential in proving Lemmas 6 and 7. Notice that $\hat{S}$ and W have the same dimensions as $\hat{\Psi}$, and hence, can be similarly partitioned into different blocks. Since Lemmas 6 and 7 hold for any given column block index j, $\Psi^{(i,j)}$ and $\mathcal{A}_j$ will be referred to as $\Psi^{(i)}$ and $\mathcal{A}$ in order to streamline the presentation.

Lemma 8. $Q \in \partial\|\tilde{\Psi}\|_{\text{block}}$ *if and only if the following conditions are satisfied for every* $i \in \{1,2,...,\bar{n}+\bar{m}\}$*:*

- *If* $\|\tilde{\Psi}^{(i)}\|_\infty \neq 0$*, define* $M^{(i)} = \{(k,l) : \tilde{\Psi}^{(i)}_{kl} = \|\tilde{\Psi}^{(i)}\|_\infty\}$*. Then,* $Q^{(i)}_{kl} = \eta_{kl}\cdot\text{sign}(\tilde{\Psi}^{(i)}_{kl})$*, where* $\sum_{(k,l)\in M^{(i)}}\eta_{kl} = 1$ *and* $\eta_{kl} = 0$ *if* $(k,l) \notin M^{(i)}$*.*
- *If* $\|\tilde{\Psi}^{(i)}\|_\infty = 0$*, then* $\|Q^{(i)}\|_1 \leq 1$*.*

The proofs of Lemmas 6 and 7 are based on the well-known primal-dual witness approach introduced in [31], [33], which is defined as follows:

Primal-dual witness approach ([31], [33]):

- *Step 1:* Define the restricted regularized problem as

$$\tilde{\Psi} = \arg \min_{\Psi \in \mathbb{R}^{p \times r}} \quad \frac{1}{2d}\|Y - X\Psi\|_F^2 + \lambda_d \|\Psi\|_{\text{block}} \tag{48a}$$

$$\text{s.t.} \quad \Psi^{(i)} = 0 \quad \forall i \in \mathcal{A}^c \tag{48b}$$

  whose solution is unique if $X_{\mathcal{A}}^\top X_{\mathcal{A}}$ is invertible.
- *Step 2:* With a slight abuse of notation, $\tilde{\Psi}$ can be written as $(\tilde{\Psi}_{\mathcal{A}}, 0)$. Choose $\tilde{S}_{\mathcal{A}}$ as an element of the sub-differential $\partial\|\tilde{\Psi}_{\mathcal{A}}\|_{\text{block}}$.
- *Step 3:* Find $\tilde{S}_{\mathcal{A}^c}$ by solving the KKT equations (41), given $\tilde{\Psi}$ and $\tilde{S}_{\mathcal{A}}$. Then, verify

$$\|\tilde{S}^{(i)}\|_1 < 1 \quad \forall i \in \mathcal{A}^c \tag{49}$$

If (49) can be verified in the last step, it is said that the primal-dual witness (PDW) approach *succeeds*. The next lemma unveils a close relationship between the block-regularized estimator, the PDW approach, and the true regression parameter Ψ^*.

Lemma 9. *The following statements hold:*

- *If the PDW approach succeeds, then $\tilde{\Psi}$ is the unique optimal solution of* (9)*, i.e. $\hat{\Psi} = \tilde{\Psi}$.*
- *Conversely, suppose that $\hat{\Psi}$ is the optimal solution of* (9) *such that $\hat{\Psi}^{(i)} = 0$ for every $i \in \mathcal{A}^c$. Then, the PDW approach succeeds.*

Proof. The proof is a simple generalization of Lemma 2 in [33]. The details are omitted for brevity.□ Lemma 9 is the building block of our proofs for Lemmas 6 and 7. In particular, Lemma 9 indicates that in order to show that the solution of (42) is unique and excludes false positive errors, it is enough to verify that the PDW approach succeeds with high probability. Then, conditioned on the success of the PDW approach, our focus can be devoted to the optimal solution of the restricted problem (48) and bounding its difference from the true parameters.

Lemma 10. *Define $\tilde{\Psi} - \Psi = E$. The following equalities hold:*

$$E_{\mathcal{A}^c} = 0 \tag{50a}$$

$$E_{\mathcal{A}} = (\frac{1}{d}X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}\frac{1}{d}X_{\mathcal{A}}^\top W - (\frac{1}{d}X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}\lambda_d \tilde{S}_{\mathcal{A}} \tag{50b}$$

$$\tilde{S}_{\mathcal{A}^c} = \frac{1}{d\lambda_d}\left(X_{\mathcal{A}^c}^\top - (X_{\mathcal{A}^c}^\top X_{\mathcal{A}})(X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}X_{\mathcal{A}}^\top\right)W + \frac{1}{d}X_{\mathcal{A}^c}^\top X_{\mathcal{A}}(\frac{1}{d}X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}} \tag{50c}$$

Proof. To verify (50b) and (50c), note that the KKT condition in Lemma 5 reduces to

$$\frac{1}{d}(X_{\mathcal{A}}^\top X_{\mathcal{A}})E_{\mathcal{A}} - \frac{1}{d}X_{\mathcal{A}}^\top W + \lambda_d \tilde{S}_{\mathcal{A}} = 0 \tag{51a}$$

$$\frac{1}{d}(X_{\mathcal{A}^c}^\top X_{\mathcal{A}})E_{\mathcal{A}} - \frac{1}{d}X_{\mathcal{A}^c}^\top W + \lambda_d \tilde{S}_{\mathcal{A}^c} = 0 \tag{51b}$$

Solving (51a) with respect to $E_{\mathcal{A}}$ and substituting the solution in (51b) completes the proof. □

A. Proof of Lemma 6:

As shown in Lemma 9, it is enough to prove that the PDW succeeds with high probability. To this goal, we show that $\max_{i\in\mathcal{A}^c}\|\tilde{S}^{(i)}\|_1<1$ with high probability, which results in the success of the PDW approach. Lemma 10 yields that

$$\|\tilde{S}^{(i)}\|_1 \le \underbrace{\left\|\frac{1}{d\lambda_d}\left({X_{(i)}}^\top-({X_{(i)}}^\top X_{\mathcal{A}})(X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}X_{\mathcal{A}}^\top\right)W\right\|_1}_{Z_1^{(i)}} + \underbrace{\left\|\frac{1}{d}{X_{(i)}}^\top X_{\mathcal{A}}(\frac{1}{d}X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}}\right\|_1}_{Z_2^{(i)}} \tag{52}$$

Similar to [33], we will show that $\max_{i\in\mathcal{A}^c}Z_1^{(i)}<\gamma/2$ and $\max_{i\in\mathcal{A}^c}Z_2^{(i)}<1-\gamma/2$ with high probability. First, consider $\max_{i\in\mathcal{A}^c}Z_1^{(i)}$. We have

$$Z_1^{(i)}=\sum_{(k,l)\in\Psi^{(i)}}\Big|\underbrace{\frac{1}{d\lambda_d}({X_{(i)}})_{:,k}{}^\top(I-X_{\mathcal{A}}(X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}X_{\mathcal{A}}^\top)W_{:,l}}_{R_{kl}^{(i)}}\Big|$$

Given X, note that $R_{kl}^{(i)}$ is Gaussian with variance

$$\frac{\sigma_w^2}{d^2\lambda_d^2}\left(({X_{(i)}})_{:,k}{}^\top(I-X_{\mathcal{A}}(X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}X_{\mathcal{A}}^\top)^2({X_{(i)}})_{:,k}\right) \tag{53}$$

Moreover, $X_{\mathcal{A}}(X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}X_{\mathcal{A}}^\top$ is an orthogonal projection onto the range of $X_{\mathcal{A}}$. Therefore,

$$\begin{aligned}&\frac{\sigma_w^2}{d^2\lambda_d^2}\left(({X_{(i)}})_{:,k}{}^\top(I-X_{\mathcal{A}}(X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}X_{\mathcal{A}}^\top)^2({X_{(i)}})_{:,k}\right)\\&=\frac{\sigma_w^2}{d^2\lambda_d^2}\left(({X_{(i)}})_{:,k}{}^\top(I-X_{\mathcal{A}}(X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1}X_{\mathcal{A}}^\top)({X_{(i)}})_{:,k}\right)\\&\le\frac{\sigma_w^2}{d^2\lambda_d^2}\|({X_{(i)}})_{:,k}\|_2^2\end{aligned} \tag{54}$$

Define $p_i=n_i$ if $1\le i\le\bar{n}$ and $p_i=m_i$ if $\bar{n}+1\le i\le\bar{n}+\bar{m}$. Due to Lemma 2, the last inequality is upper bounded by $4\sigma_w^2\sigma_{\max}^2/d\lambda_d^2$ for every $k\in\{1,...,p_i\}$ with probability of at least $1-\exp(-d/2+\log p_i)$ for $d>2\log p_i$. Conditioned on this event, one can write:

$$Z_1^{(i)}=\max_{\epsilon\in\{-1,+1\}^{p_i\times n_j}}\sum_{(k,l)\in\Psi^{(i)}}\epsilon_{kl}R_{kl}^{(i)} \tag{55}$$

which means that $\sum_{(k,l)\in\Psi^{(i)}}\epsilon_{kl}R_{kl}^{(i)}$ is sub-Gaussian with the parameter $4D_j\sigma_w^2\sigma_{\max}^2/d\lambda_d^2$. This implies that

$$\begin{aligned}\mathbb{P}(\max_{i\in\mathcal{A}^c}Z_1^{(i)}\ge\zeta)&=\mathbb{P}\left(\max_{i\in\mathcal{A}^c}\max_{\epsilon\in\{-1,+1\}^{p_i\times n_j}}\sum_{(k,l)\in\Psi^{(i)}}\epsilon_{kl}R_{kl}^{(i)}\ge\zeta\right)\\&\le 2\exp\left(-\frac{d\lambda_d^2\zeta^2}{8D_j\sigma_w^2\sigma_{\max}^2}+D_j+\log(\bar{n}+\bar{m})\right)\\&\quad+\exp\left(-d/2+\log p_{\max}+\log(\bar{n}+\bar{m})\right)\end{aligned}$$

where we have used Lemma 1, the second statement of Lemma 4 and the facts that $p_i \le p_{\max}$ and $|\mathcal{A}^c| \le \bar{n} + \bar{m}$ in the last inequality. Now, setting $\zeta = \gamma/2$ and

$$\lambda_d \ge \sqrt{\frac{32 c_1 \sigma_w^2 \sigma_{\max}^2}{\gamma^2} \cdot \frac{(D_j)^2 + D_j \log(\bar{n}+\bar{m})}{d}} \tag{56}$$

for some arbitrary constant $c_1 > 1$ yields that

$$\begin{aligned}\mathbb{P}(\max_{i \in \mathcal{A}^c} Z_1^{(i)} < \gamma/2) \ge & 1 - 2\exp(-(c_1-1)(D_j + \log(\bar{n}+\bar{m}))) \\ & - \exp(-d/2 + \log p_{\max} + \log(\bar{n}+\bar{m})) \\ \ge & 1 - 3\exp(-(c_1-1)(D_j + \log(\bar{n}+\bar{m})))\end{aligned} \tag{57}$$

where the last inequality is due to the lower bound (44b) on d. Next, an upper bound on $\max_{i \in \mathcal{A}^c} Z_2^{(i)}$ will be derived. Since each row of X is drawn from $N(0, \tilde{\Sigma})$, one can write the distribution of $X_{\mathcal{A}^c}^\top$, conditioned on $X_{\mathcal{A}}$ as

$$N\Big(\tilde{\Sigma}_{\mathcal{A}^c,\mathcal{A}}(\tilde{\Sigma}_{\mathcal{A},\mathcal{A}})^{-1} X_{\mathcal{A}}^\top, \underbrace{\tilde{\Sigma}_{\mathcal{A}^c,\mathcal{A}^c} - \tilde{\Sigma}_{\mathcal{A}^c,\mathcal{A}}(\tilde{\Sigma}_{\mathcal{A},\mathcal{A}})^{-1}\tilde{\Sigma}_{\mathcal{A},\mathcal{A}^c}}_{\tilde{\Sigma}_{\mathcal{A}^c|\mathcal{A}}}\Big) \tag{58}$$

Based on (58), one can verify that $\frac{1}{d} X_{\mathcal{A}^c}^\top X_{\mathcal{A}} (\frac{1}{d} X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} \tilde{S}_{\mathcal{A}}$ has the same distribution as

$$\tilde{\Sigma}_{\mathcal{A}^c,\mathcal{A}}(\Sigma_{\mathcal{A},\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}} + \frac{1}{d} V^\top X_{\mathcal{A}} (\frac{1}{d} X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} \tilde{S}_{\mathcal{A}} \tag{59}$$

where V is a random matrix with zero mean, covariance $\tilde{\Sigma}_{\mathcal{A}^c|\mathcal{A}}$, and independent of X. In light of the definition of $\tilde{\Sigma}_{\mathcal{A}^c|\mathcal{A}}$, it can be easily seen that the elements of V are sub-Gaussian with parameters of at most $\sigma_{\max}^2$. This implies that

$$\begin{aligned}\max_{i \in \mathcal{A}^c} Z_2^{(i)} \le & \max_{i \in \mathcal{A}^c} \big\| \Sigma_{i,\mathcal{A}} (\Sigma_{\mathcal{A},\mathcal{A}})^{-1} \tilde{S}_{\mathcal{A}} \big\|_1 \\ & + \max_{i \in \mathcal{A}^c} \Big\| \frac{1}{d} {V_{(i)}}^\top X_{\mathcal{A}} (\frac{1}{d} X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} \tilde{S}_{\mathcal{A}} \Big\|_1 \\ \le & 1 - \gamma + \max_{i \in \mathcal{A}^c} \underbrace{\Big\| \frac{1}{d} {V_{(i)}}^\top X_{\mathcal{A}} (\frac{1}{d} X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} \tilde{S}_{\mathcal{A}} \Big\|_1}_{Z_3^{(i)}}\end{aligned} \tag{60}$$

where we have used the mutual incoherence property and the fact that $\|\tilde{S}^{(i)}\|_1 = 1$ for every $i \in \mathcal{A}$. Now, it remains to show that $\max_{i \in \mathcal{A}^c} Z_3^{(i)} < \gamma/2$ with high probability. Similar to $Z_1^{(i)}$, one can write:

$$Z_3^{(i)} = \sum_{(k,l) \in \Psi^{(i)}} \underbrace{\Big| \frac{1}{d} (V_{(i)})_{:,k}^\top X_{\mathcal{A}} (\frac{1}{d} X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} (\tilde{S}_{\mathcal{A}})_{:,l} \Big|}_{T_{kl}^{(i)}} \tag{61}$$

Given X, note that $T_{kl}^{(i)}$ is Gaussian with variance

$$\sigma_{\max}^2 (\tilde{S}_{\mathcal{A}})_{:,l}^\top \Big(\frac{1}{d} X_{\mathcal{A}}^\top X_{\mathcal{A}} \Big)^{-1} (\tilde{S}_{\mathcal{A}})_{:,l} \tag{62}$$

Also, $\|(\tilde{S}_{\mathcal{A}})_{:,l}\|_2^2 \leq k_j$. Therefore, Lemma 3 can be used to bound (62) as follows:

$$\begin{aligned}(\tilde{S}_{\mathcal{A}})_{:,l}^{\top}\left(\frac{1}{d}X_{\mathcal{A}}^{\top}X_{\mathcal{A}}\right)^{-1}(\tilde{S}_{\mathcal{A}})_{:,l} &\leq \sigma_{\max}^2 k_j \left\|\frac{1}{d}(\frac{1}{d}X_{\mathcal{A}}^{\top}X_{\mathcal{A}})^{-1}\right\|_2 \\ &\leq \sigma_{\max}^2 k_j \left(\frac{1}{d}\frac{8}{\Lambda_{\min}} + \frac{1}{d}\left\|\Sigma_{\mathcal{A},\mathcal{A}}^{-1}\right\|_2\right) \\ &\leq \sigma_{\max}^2 k_j \left(\frac{1}{d}\cdot\frac{8}{\Lambda_{\min}} + \frac{1}{d}\cdot\frac{1}{\Lambda_{\min}}\right) \\ &\leq \frac{9\sigma_{\max}^2 k_j}{\Lambda_{\min} d} \end{aligned} \tag{63}$$

with probability of at least $1-2\exp(-d/2)$. Similar to the arguments made for bounding $\max_{i\in\mathcal{A}^c} Z_1^{(i)}$, one can verify that

$$\begin{aligned}\mathbb{P}\left(\max_{i\in\mathcal{A}^c} Z_3^{(i)} < \gamma/2\right) \geq 1 - 2\exp\Big(-\frac{\Lambda_{\min} d\gamma^2}{72\sigma_{\max}^2 k_j D_j} + D_j \\ + \log(\bar{n}+\bar{m})\Big) - 2\exp\Big(-\frac{d}{2}\Big)\end{aligned} \tag{64}$$

Now, choosing

$$d \geq \frac{72 c_2 \sigma_{\max}^2 k_j D_j}{\Lambda_{\min}\gamma^2}\cdot(D_j + \log(\bar{n}+\bar{m})) \tag{65}$$

for some arbitrary constant $c_2 > 1$ results in

$$\mathbb{P}\left(\max_{i\in\mathcal{A}^c} Z_3^{(i)} < \gamma/2\right) \geq 1 - 4\exp(-(c_2-1)(D_j + \log(\bar{n}+\bar{m})))$$

Therefore, $\max_{i\in\mathcal{A}^c}\|\tilde{S}^{(i)}\|_1 < 1$ and, hence, PDW succeeds with a probability that is lower bounded by (45). □

B. Proof of Lemma 7:

In order to bound the estimation error, an upper bound on $\|E\|_\infty$ will be derived, conditioning on the success of the PDW approach. Note that $E_{\mathcal{A}^c} = 0$ according to Lemma 10 and, hence, it suffices to bound $\|E_{\mathcal{A}}\|_\infty$. Again, due to Lemma 10, one can write:

$$\begin{aligned}\max_{k=1,\ldots,n_j}\|(E_{\mathcal{A}})_{:,k}\|_\infty \leq \max_{k=1,\ldots,n_j} \underbrace{\left\|(\frac{1}{d}X_{\mathcal{A}}^{\top}X_{\mathcal{A}})^{-1}\frac{1}{d}X_{\mathcal{A}}^{\top}W_{:,k}\right\|_\infty}_{Z_4^k} \\ + \max_{k=1,\ldots,n_j} \underbrace{\left\|(\frac{1}{d}X_{\mathcal{A}}^{\top}X_{\mathcal{A}})^{-1}\lambda_d(\tilde{S}_{\mathcal{A}})_{:,k}\right\|_\infty}_{Z_5^k}\end{aligned}$$

for $k = 1, 2, ..., n_j$. For bounding Z_5^k, it can be argued similarly to (63) that

$$\begin{aligned}\max_{k=1,\ldots,n_j} Z_5^k &\leq \max_{k=1,\ldots,n_j}\left\|\left((\frac{1}{d}X_{\mathcal{A}}^{\top}X_{\mathcal{A}})^{-1} - \Sigma_{\mathcal{A},\mathcal{A}}^{-1}\right)\lambda_d(\tilde{S}_{\mathcal{A}})_{:,k}\right\|_\infty \\ &\quad + \max_{k=1,\ldots,n_j}\left\|\Sigma_{\mathcal{A},\mathcal{A}}^{-1}\lambda_d(\tilde{S}_{\mathcal{A}})_{:,k}\right\|_\infty \\ &\leq \left\|(\frac{1}{d}X_{\mathcal{A}}^{\top}X_{\mathcal{A}})^{-1} - \Sigma_{\mathcal{A},\mathcal{A}}^{-1}\right\|_2 \lambda_d\sqrt{k_j} + \frac{\lambda_d}{\Lambda_{\min}} \\ &\leq \frac{\lambda_d}{\Lambda_{\min}}\left(8\sqrt{k_j}\sqrt{\frac{k_j n_j + c_3\log(\bar{n}+\bar{m})}{d}} + 1\right)\end{aligned} \tag{66}$$

for some $c_3 > 0$ with probability of at least $1 - 2\exp(-(k_j n_j + c_3 \log(\bar{n}+\bar{m}))/2)$, where we have used the matrix norm properties and Lemma 3 with $t = k_j n_j + c_3 \log(\bar{n}+\bar{m})$ (note that $|I(\mathcal{A})| \leq k_j n_j$). Now, it remains to bound $\max_{k=1,\ldots,n_j} Z_4^k$. This can be carried out similar to the previous arguments, i.e., by making use of (63) and obtaining a sub-Gaussian parameter for $(\frac{1}{d} X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} \frac{1}{d} X_{\mathcal{A}}^\top W_{:,k}$. For brevity, only the final key relation is stated below:

$$\begin{aligned}\mathbb{P}(\max_{k=1,\ldots,n_j} Z_4^k \geq \zeta) \leq & 2\exp\Big(-\frac{d\Lambda_{\min}\zeta^2}{18\sigma_w^2} + \log n_j + \log(k_j n_j)\Big) \\ & + 2\exp(-d/2) \\ \leq & 2\exp\Big(-\frac{d\Lambda_{\min}\zeta^2}{18\sigma_w^2} + 2(\alpha_n+\alpha_k)\log(\bar{n}+\bar{m})\Big) \\ & + 2\exp(-d/2) \end{aligned} \tag{67}$$

where the last inequality is due to the assumption that $n_j \leq n_{\max} = O\left((\bar{n}+\bar{m})^{\alpha_n}\right)$ and $k_j \leq k_{\max} = O\left((\bar{n}+\bar{m})^{\alpha_k}\right)$. Now, setting

$$\zeta = \sqrt{\frac{36 c_4 (\alpha_n + \alpha_k)\sigma_w^2 \log(\bar{n}+\bar{m})}{d\Lambda_{\min}}} \tag{68}$$

for an arbitrary constant $c_4 > 1$, together with the inequality $\log r_1 \leq \log(k_j D_j)$, leads to

$$\max_{k=1,\ldots,n_j} Z_4^k \leq \sqrt{\frac{36 c_4 (\alpha_n + \alpha_k)\sigma_w^2 \log(\bar{n}+\bar{m})}{d\Lambda_{\min}}} \tag{69}$$

with probability of at least

$$1 - 2\exp\left(-2(c_4-1)(\alpha_n+\alpha_k)\log(\bar{n}+\bar{m})\right) - 2\exp(-d/2) \tag{70}$$

Combining this inequality with (66) results in the elementwise error bound (47) with probability of at least (46). This concludes the proof. □

C. Proof of Theorem 1:

First, we present the sketch of the proof in a few steps:

1. We decompose the block-regularized problem (9) into $\bar{n}$ disjoint block-regularized subproblems defined in (43).
2. For each of these subproblems, we consider the event that Lemmas 6 and 7 hold.
3. We consider the intersection of these $\bar{n}$ events and show that, together with (18a) and (18b), they lead to the element-wise error (19) with probability of at least $1 - \delta$.

Step 1: (9) can be rewritten as follows:

$$\hat{\Psi} = \arg\min_{\Psi} \sum_{j=1}^{n} \left(\frac{1}{2d} \|Y^{(:,j)} - X\Psi^{(:,j)}\|_2^2 + \lambda \|\Psi^{(:,j)}\|_{\text{block}} \right) \tag{71}$$

The above optimization problem can be naturally decomposed into $\bar{n}$ disjoint block-regularized subproblems in the form of (43).

Step 2: Assume that (44b) and (44a) hold for every $1 \le j \le \bar{n}$. Upon defining $\mathcal{T}_j$ as the event that Lemmas 6 and 7 hold, one can write:

$$\begin{aligned}\mathbb{P}(\mathcal{T}_j) \ge 1 &- 5\exp\big(-(c_1-1)(D_j+\log(\bar{n}+\bar{m}))\big)\\ &- 4\exp\big(-(c_2-1)(D_j+\log(\bar{n}+\bar{m}))\big)\\ &- 2\exp(-(k_j n_j + c_3\log(\bar{n}+\bar{m}))/2)\\ &- 2\exp\big(-2(c_4-1)(\alpha_n+\alpha_k)\log(\bar{n}+\bar{m}))\big)\end{aligned} \tag{72}$$

For every $1 \le j \le \bar{n}$.

Step 3: Assume that $c_1, c_2, c_4 > 2$ and $c_3 > 1$. Consider the event $\mathcal{T} = \mathcal{T}_1 \cap \mathcal{T}_2 \cap \cdots \cap \mathcal{T}_n$. Based on (72) and Lemma 4, one can write:

$$\begin{aligned}\mathbb{P}(\mathcal{T}) \ge 1 &- \underbrace{K_1(\bar{n}+\bar{m})^{-(c_1-2)}}_{(a)} - \underbrace{K_2(\bar{n}+\bar{m})^{-(c_2-2)}}_{(b)}\\ &- \underbrace{K_3(\bar{n}+\bar{m})^{-(\frac{c_3}{2}-1)}}_{(c)} - \underbrace{K_4(\bar{n}+\bar{m})^{-(2(\alpha_n+\alpha_k)(c_4-1)-1)}}_{(d)}\end{aligned} \tag{73}$$

for some constants K_1, K_2, K_3, K_4. One can easily verify that the following equalities are enough to guarantee that the right hand side of (73) is equal to $1-\delta$:

$$\begin{aligned}c_1 &= \frac{\log(4K_1/\delta)}{\log(\bar{n}+\bar{m})} + 2,\\ c_2 &= \frac{\log(4K_2/\delta)}{\log(\bar{n}+\bar{m})} + 2,\\ c_3 = c_1 &= \frac{2\log(4K_3/\delta)}{\log(\bar{n}+\bar{m})} + 2,\\ c_4 &= \frac{\log(4K_4/\delta)}{2(\alpha_n+\alpha_k)\log(\bar{n}+\bar{m})} + \frac{1}{2(\alpha_n+\alpha_k)} + 1.\end{aligned} \tag{74}$$

Substituting (74) in Lemmas 6 and 7 leads to two observations:

- If λ_d and d satisfy (18a) and (18b), then they also satisfy (44a) and (44b).
- The parameter g defined in (19) is greater than or equal to g_j for every $j = 1, ..., \bar{n}$.

Therefore, (18a) and (18b) guarantee that: 1) $\hat{\Psi}$ is unique and does not have any false positive in its blocks, and 2) its element-wise error is upper bounded by (19). Now, it only remains to show that $\hat{\Psi}$ excludes false negatives (the blocks that are mistakenly estimated to have nonzero values). To this goal, it suffices to show that (18b) guarantees $g < t_{\min}$. Suppose that

$$d = \Omega\left(C_\Psi \kappa(\tilde{\Sigma})^2 k_{\max}\left(D\log(\bar{n}+\bar{m}) + D^2\log(1/\delta)\right)\right) \tag{75}$$

In what follows, we will show that $C_\Psi = O(1)$ is enough to have $g < t_{\min}$. The lower bound on d in (18b) yields that

$$g \le K\left(\frac{1}{\sqrt{C_\Psi k_{\max}}} + \frac{1}{C_\Psi \kappa(\tilde{\Sigma})}\right) \tag{76}$$

for some constant K. Therefore,

$$C_\Psi = \frac{2/K}{t_{\min}\kappa(\tilde{\Sigma})} + \frac{4/K}{t_{\min}^2 k_{\max}} = O(1) \tag{77}$$

is enough to ensure $g < t_{\min}$. This completes the proof. □

V. Numerical Results

In this section, we illustrate the performance of the block-regularized estimator and compare it with its least-squares counterpart. We consider three case studies on synthetically generated systems and multi-agent systems. The built-in `lasso` function in MATLAB and the `PQN` package from [40] are used to obtain the Lasso and block-regularized estimators, respectively. These solvers are relatively fast in practice; they can solve the largest instance of the problem (with approximately 9.7 million unknown parameters) in approximately 9.28 and 7.69 minutes, respectively.

Define the (block) mismatch error as the total number of false positives and false negatives in the (block) sparsity pattern of the estimator. Moreover, define *relative number of sample trajectories* (RST) as the number of sample trajectories normalized by the dimension of the system, and *relative (block) mismatch error* (RME) as the mismatch error normalized by total number of elements (blocks) in Ψ. To verify the developed theoretical results, λ_d is set to

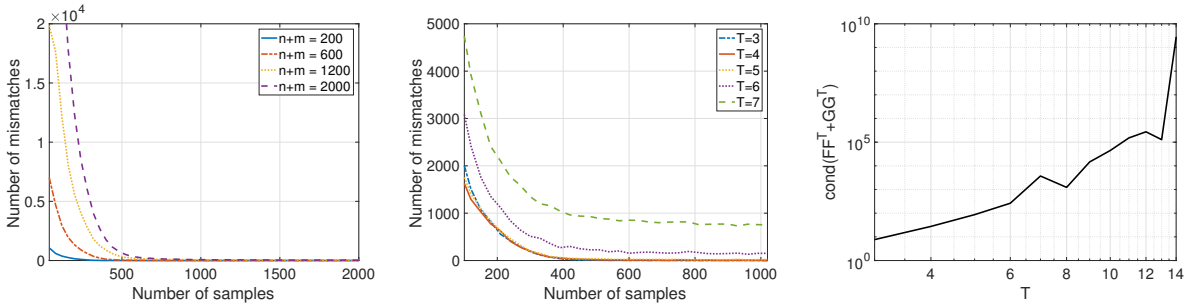
$$\sqrt{\frac{2(D^2 + D\log(\bar{n}+\bar{m}))}{d}} \tag{78}$$

in all of the experiments. Note that this choice of λ_d does not require any additional fine-tuning.

A. Case Study 1: Synthetically Generated Systems

Given the numbers n and w, and for each instance of the problem, the state and input matrices are constructed as follows: The size of each block in A and B is set to 1. The diagonal elements of $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ are set to 1 (the dimensions of the inputs and states are chosen to be equal). The elements of the first w upper and lower diagonals of A and B are set to 0.3 or -0.3 with equal probability. Furthermore, at each row of A, another w elements are randomly chosen from the elements not belonging to the first w upper and lower diagonals and set to 0.3 or -0.3 with equal probability. We set $\Sigma_u = I$ and $\Sigma_w = 0.5I$. The mutual incoherence property is satisfied for most of the constructed instances.

In the first set of experiments, we consider the mismatch error of $\hat{\Psi}$ with respect to the number of sample trajectories and for different system dimensions. The length of the time horizon T is set to 3. The results are illustrated in Figure 1a for $n+m$ equal to 200, 600, 1200, and 2000. In all of these test cases, w is chosen in such a way that the number of nonzero elements in each column of Ψ is between $(n+m)^{0.3}$ and $(n+m)^{0.4}$. It can be observed that as the dimension of the system increases, a higher number of sample trajectories is required to have a small mismatch error in the block-regularized estimator. Conversely, the required value of RST to achieve a small RME reduces as the dimension of the system grows. More precisely, RST should be at least 1.80, 1.13, 0.37, and 0.20 to guarantee RME $\leq 0.1\%$, when $m+n$ is equal to 200, 600, 1200, and 2000, respectively.



(a) Mismatch error for different d (b) Mismatch error for different T (c) Condition number for different T

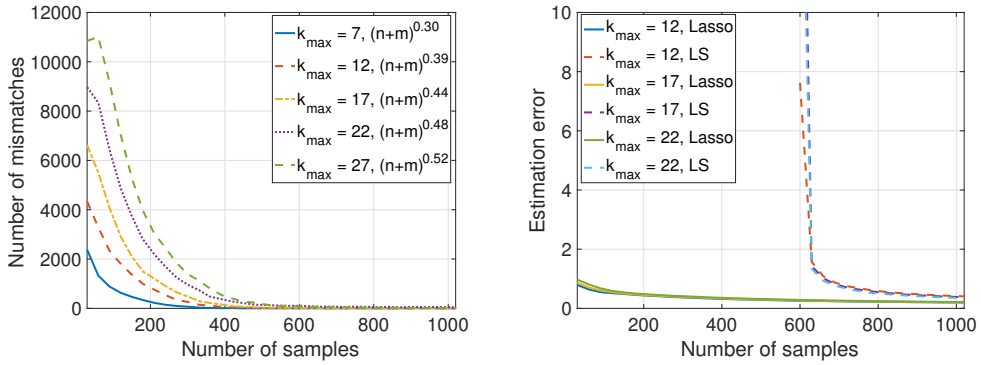
Fig. 1: (a) The mismatch error with respect to the number of sample trajectories for different system dimensions, (b) the mismatch error with respect to the number of sample trajectories for different time horizons, (c) the condition number of $FF^{\mathsf{T}} + GG^{\mathsf{T}}$ with respect to the time horizon.

In the next set of experiments, we consider the mismatch error for different time horizons $T = 3, 4, ..., 7$, by fixing $m + n = 600$ and $w = 2$. As mentioned before, large values of T tend to inflate the easily identifiable modes of the system and suppress the nearly hidden ones, thereby making it hard to obtain an accurate estimation of the parameters. It is pointed out that $\kappa(F_T F_T^{\mathsf{T}} + G_T G_T^{\mathsf{T}})$ is a good indicator of the gap between these modes. This relationship is clearly reflected in Figures 1b and 1c. As can be observed in Figure 1b, 330 sample trajectories are enough to guarantee RME $\leq 0.1\%$ for $T = 3$. However, for $T = 7$, RME cannot be reduced below 0.42% even with 1000 sample trajectories. To further elaborate on this dependency, Figure 1c is used to illustrate the value of $\kappa(F_T F_T^{\mathsf{T}} + G_T G_T^{\mathsf{T}})$ with respect to T in a log-log scale. One can easily verify that $\kappa(F_T F_T^{\mathsf{T}} + G_T G_T^{\mathsf{T}})$ associated with $T = 7$ is 485 times greater than this parameter for $T = 3$.

Finally, we study the block-regularized estimator for different per-column numbers of nonzero elements in Ψ and compare its accuracy to the least-squares estimator. Fixing $T = 3$ and $m + n = 600$, Figure 2a depicts the mismatch error of the block-regularized estimator when the maximum number of nonzero elements at each column of Ψ ranges from 7 (corresponding to $w = 1$) to 27 (corresponding to $w = 5$). Not surprisingly, the required number of samples to achieve a small mismatch error increases as the number of nonzero elements in each column of Ψ grows. On the other hand, the least-squares estimator is fully dense in all of these experiments, regardless of the number of sample trajectories. To have a better comparison between the two estimators, we consider the 2-norm of the estimation errors normalized by the 2-norm of Ψ^*, for different numbers of nonzero elements in each column of Ψ^*. As it is evident in Figure 2b, the block-regularized estimator significantly outperforms the least-squares one for any number of sample trajectories. Furthermore, the least-squares estimator is not defined for $d < 600$.

B. Case Study 2: Switching Networks

In this case study, we study a network of multi-agent systems that are interconnected through a switching information exchange topology. Recently, a special attention has been devoted to multi-agent systems with a time-



(a) Mismatch error for different $k_{\max}$

(b) Estimation error for different $k_{\max}$

Fig. 2: (a) The mismatch error with respect to the number of sample trajectories for different per-column number of nonzero elements in Ψ^*, (b) the normalized estimation error for Lasso and least-squares (abbreviated as LS) estimators with respect to the number of sample trajectories.

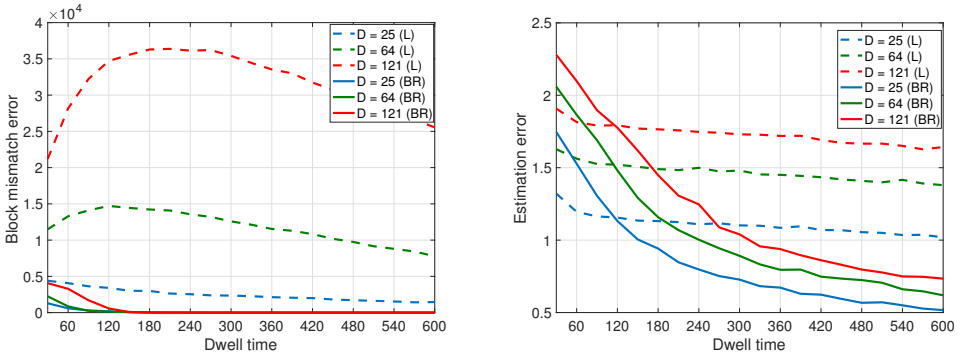
varying network topology; in many communication networks, each sensor has access only to the information of its neighbors. Therefore, when the location of these sensors changes over time, so does the topology of the interconnecting links [41]. The *dwell time* is defined as the time interval in which the network topology is unchanged. The goal is to identify the structure of the network within the dwell time. The state-space equation of agent i admits the following general form:

$$\dot{x}_i(t) = \sum_{(i,j)\in\mathcal{N}_x(i)} A^{(i,j)} x_j(t) + \sum_{(i,j)\in\mathcal{N}_u(i)} B^{(i,j)} u_j(t) + w_i(t) \tag{79}$$

where, as before, $A^{(i,j)} \in \mathbb{R}^{n_i\times n_i}$ and $B^{(i,j)} \in \mathbb{R}^{n_i\times m_i}$ are the $(i,j)^{\text{th}}$ blocks of A and B. Furthermore, $\mathcal{N}_x(i)$ and $\mathcal{N}_u(i)$ are the sets of neighbors of agent i whose respective state and input actions affect the state of agent i.

We consider 200 agents connected through a randomly generated sparse network. In particular, we assume that each agent is connected to 5 other agents. If $j \in \mathcal{N}_x(i)$ or $j \in \mathcal{N}_u(i)$, then each element of $A^{(i,j)}$ or $B^{(i,j)}$ is randomly selected from $[-0.4\ \ -0.3]\cup[0.3\ \ 0.4]$. The behavior of the proposed block-regularized estimator will be examined for different dimensions of the agents. In particular, we investigate the performance of this estimator in comparison with the Lasso for which the sparsity of the system matrices is promoted on different elements independent of the block structures. In these experiments, (n_i, m_i) is chosen from $\{(5,5),(8,8),(11,11)\}$. This entails that $D \in \{25, 64, 121\}$ and $(n, m) \in \{(1000, 1000), (1600, 1600), (2200, 2200)\}$. Furthermore, T is set to 3 and the system is discretized using the forward Euler method with the sampling time of 0.2 seconds. This implies that each sample trajectory is collected within 0.6 seconds. The number of block mismatch and 2-norm estimation errors are depicted in Figures 3a and 3b with respect to the dwell time. As can be seen in these figures, the incorporation of the block sizes in the estimation procedure can significantly improve the accuracy.

Figure 3a shows the number of block mismatch error for the block-regularized and Lasso estimators. Evidently, the former substantially outperforms the latter in terms of the correct sparsity recovery. In particular, 252, 260,



(a) Mismatch error

(b) Normalized estimation error

Fig. 3: (a) The block mismatch error for the block-regularized (abbreviated as `BR`) and Lasso (abbreviated as `L`) estimators with respect to the dwell time and for different block sizes in Ψ^*, (b) the normalized estimation error for the block-regularized (abbreviated as `BR`) and Lasso (abbreviated as `L`) estimators with respect to the dwell time for different block sizes in Ψ^*.

and 302 sample trajectories are enough to achieve RME $\leq 0.1\%$ when D is equal to 25, 64, and 121, respectively (notice that the largest instance has more than 9 million parameters to be estimated). However, the Lasso estimator cannot achieve this accuracy with even 2000 sample trajectories.

Figure 3b demonstrates the 2-norm of the estimation error for these estimators. Although the Lasso has a smaller estimation error for $d < 200$, it is strictly dominated by that of the block-regularized estimator when $d \geq 200$.

VI. Conclusion

We consider the problem of identifying the parameters of linear time-invariant (LTI) systems. In many real-world problems, the state-space equation describing the evolution of the system admits a block-sparse representation due to localized or internally limited interactions of its states and inputs. In this work, we leverage this property and introduce a block-regularized estimator to identify the sparse representation of the system. Using modern high-dimensional statistics, we derive sharp non-asymptotic bounds on the minimum number of input-state data samples to guarantee a small element-wise estimation error. In particular, we show that the number of available sample trajectories can be significantly smaller than the system dimension and yet, the proposed block-regularized estimator can correctly recover the block-sparsity of the state and input matrices and result in a small element-wise error. Through different case studies on synthetically generated systems and multi-agent systems, we demonstrate substantial improvements in the accuracy of the proposed estimator, compared to its well-known least-squares counterpart.

References

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

[2] K. Chernyshov, "Towards the knowledge-based multi-agent system identification," in *IEEE 10th Conference on Industrial Electronics and Applications*, 2015, pp. 399–404.

[3] S. Hassan-Moghaddam, N. K. Dhingra, and M. R. Jovanović, "Topology identification of undirected consensus networks via sparse inverse covariance estimation," in *IEEE 55th Conference on Decision and Control*, 2016, pp. 4624–4629.

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

[5] L. Ljung, "System identification," *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 1–19, 1999.

[6] H.-F. Chen and L. Guo, *Identification and stochastic adaptive control*. Springer Science & Business Media, 2012, original work published 1991.

[7] G. C. Goodwin and R. L. Payne, *Dynamic system identification: experiment design and data analysis*. Academic press, 1977.

[8] P. E. Vértes, A. F. Alexander-Bloch, N. Gogtay, J. N. Giedd, J. L. Rapoport, and E. T. Bullmore, "Simple models of human brain functional networks," *Proceedings of the National Academy of Sciences*, vol. 109, no. 15, pp. 5868–5873, 2012.

[9] S. Sun, R. Huang, and Y. Gao, "Network-scale traffic modeling and forecasting with graphical lasso and neural networks," *Journal of Transportation Engineering*, vol. 138, no. 11, pp. 1358–1367, 2012.

[10] N. Omranian, J. M. Eloundou-Mbebi, B. Mueller-Roeber, and Z. Nikoloski, "Gene regulatory network inference using fused lasso on multiple data sets," *Scientific reports*, vol. 6, p. 20533, 2016.

[11] D. R. Cox and D. V. Hinkley, *Theoretical statistics*. CRC Press, 1979.

[12] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *arXiv preprint arXiv:1710.01688*, 2017.

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

[14] T. Sarkar and A. Rakhlin, "How fast can linear dynamical systems be learned?" *arXiv preprint arXiv:1812.01251*, 2018.

[15] S. Oymak and N. Ozay, "Non-asymptotic identification of lti systems from a single trajectory," *arXiv preprint arXiv:1806.05722*, 2018.

[16] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite-time system identification for partially observed lti systems of unknown order," *arXiv preprint arXiv:1902.01848*, 2019.

[17] A. Tsiamis and G. J. Pappas, "Finite sample analysis of stochastic system identification," *arXiv preprint arXiv:1903.09122*, 2019.

[18] M. Simchowitz, R. Boczar, and B. Recht, "Learning linear dynamical systems with semi-parametric least squares," *arXiv preprint arXiv:1902.00768*, 2019.

[19] V. L. Le, F. Lauer, and G. Bloch, "Selective ℓ_1 minimization for sparse recovery," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 3008–3013, 2014.

[20] X. Jiang, Y. Yao, H. Liu, and L. Guibas, "Compressive network analysis," *IEEE transactions on automatic control*, vol. 59, no. 11, pp. 2946–2961, 2014.

[21] C. R. Rojas, R. Tóth, and H. Hjalmarsson, "Sparse estimation of polynomial and rational dynamical models." *IEEE Trans. Automat. Contr.*, vol. 59, no. 11, pp. 2962–2977, 2014.

[22] T. Chen, M. S. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto, "System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 2933–2945, 2014.

[23] R. S. Smith, "Frequency domain subspace identification using nuclear norm minimization and hankel matrix realizations," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 2886–2896, 2014.

[24] S. Ross and J. A. Bagnell, "Agnostic system identification for model-based reinforcement learning," *arXiv preprint arXiv:1203.1007*, 2012.

[25] S. Sadraddini and C. Belta, "Formal guarantees in data-driven model identification and control synthesis," in *21st ACM International Conference on Hybrid Systems: Computation and Control*. ACM, 2018.

[26] Z. Hou and S. Jin, "Data-driven model-free adaptive control for a class of mimo nonlinear discrete-time systems," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2173–2188, 2011.

[27] E. Weyer, R. C. Williamson, and I. M. Mareels, "Finite sample properties of linear model identification," *IEEE Transactions on Automatic Control*, vol. 44, no. 7, pp. 1370–1383, 1999.

[28] E. Weyer, "Finite sample properties of system identification of arx models under mixing conditions," *Automatica*, vol. 36, no. 9, pp. 1291–1299, 2000.

[29] J. Pereira, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," in *Advances in Neural Information Processing Systems*, 2010, pp. 172–180.

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

[31] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE transactions on information theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[32] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine learning research*, vol. 7, no. Nov, pp. 2541–2563, 2006.

[33] S. N. Negahban and M. J. Wainwright, "Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3841–3863, 2011.

[34] S. Fattahi and S. Sojoudi, "Data-driven sparse system identification," *to appear in IEEE 57th Conference on Decision and Control*, 2018.

[35] S. Fattahi and S. Sojoudi, "Non-asymptotic analysis of block-regularized regression problem," *to appear in 56th Annual Allerton Conference on Communication, Control, and Computing*, 2018.

[36] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, pp. 1436–1462, 2006.

[37] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[38] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, no. 3, p. 969, 2007.

[39] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

[40] M. Schmidt, E. van den Berg, M. P. Friedlander, and K. Murphy, "Pqn: Optimizing costly functions with simple constraints," 2009. [Online]. Available: https://www.cs.ubc.ca/~schmidtm/Software/PQN.html

[41] M. Mesbahi and M. Egerstedt, *Graph theoretic methods in multiagent networks*. Princeton University Press, 2010.