Sample Complexity of Block-Sparse System Identification Problem

Salar Fattahi and Somayeh Sojoudi

***Abstract*—In this article, 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 elementwise 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 switching networks and power systems are offered to demonstrate the effectiveness of the proposed method.**

***Index Terms*—High-dimensional statistics, statistical learning, system identification.**

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.

In this work, our main goal is to characterize the sample complexity of learning block-sparse linear time-invariant (LTI) systems from noisy input–output trajectories. More specifically, we study the efficient learning of LTI systems in high-dimensional settings, where the system dimension is significantly larger than the number of collected samples. This type of dynamical system forms the basis of many classical control 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 block-sparse representation of the state-space equations of the system. Driven by the existing nonasymptotic results on the classical Lasso problem, the main focus of this article is on the block-regularized estimators for the system identification problem, where the goal is to characterize the number of required sample trajectories to reliably estimate the block-sparse interactions of the system. To this goal, the ℓ_∞-norms of the blocks are penalized instead of their ℓ_1-norms.

In many real-world systems, such as power networks and multiagent systems, the local state and input behavior of the physical agents/subsystems can be captured and characterized via block matrices in their dynamical models. For instance, in the system identification problem for power systems, each block of the system matrices corresponds to the local states/inputs of an individual generator, and the goal is to learn the sparse interactions among generators given a limited number of measurements from phasor measurement units and supervisory control and data acquisition systems [1], [2]. In this context, it is reasonable to assume that the unknown dynamical interactions among the generators enjoy a block-sparse structure. As another example, consider the problem of planar vertical takeoff and landing for a fleet of interconnected aircraft. In this context, the number of blocks in the state-space equation of the system corresponds to the number of aircraft that is known *a priori*, and the goal is to infer the time-varying and uncertain interactions among the aerial vehicles based on the local sensory data [3], [4]. Indeed, such local interactions can be captured via a block-sparse dynamical model.

A. Related Works

1) Asymptotic Guarantees: System identification is a well-established area of research in control theory, with related

Manuscript received November 15, 2020; revised March 21, 2021; accepted May 28, 2021. Date of publication June 14, 2021; date of current version December 3, 2021. This work was supported in part by the Air Force Office of Scientific Research, in part by the Office of Naval Research, and in part by the National Science Foundation. Recommended by Associate Editor W. X. Zheng. *(Corresponding author: Salar Fattahi.)*

Salar Fattahi is with the Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: fattahi@umich.edu).

Somayeh Sojoudi is with the Department of Electrical Engineering and Computer Sciences, the Department of Mechanical Engineering, and the Tsinghua–Berkeley Shenzhen Institute, University of California, Berkeley, CA 94720 USA (e-mail: sojoudi@berkeley.edu).

Digital Object Identifier 10.1109/TCNS.2021.3089141

preliminary results dating back to the 1960s. Standard reference textbooks on the topic include [5] and [6], both focusing on establishing *asymptotic* consistency of different types of estimators. 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 and transportation networks, the dimensionality of the system is overwhelmingly large, often surpassing the number of available input–output data [7], [8]. Under such circumstances, the classical approaches for checking the asymptotic consistency of an estimator face major breakdowns.

2) 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 [9], 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 1) marginally stable systems in [10], 2) unstable systems in [11], and 3) partially observed stable systems in [12].

3) 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 [13], identification based on compressive sensing [14]–[17], sparse estimation of polynomial system dynamics [18], Kernel-based regularization [19], low rank estimation in the frequency domain [20], and sparse system identification of time-varying systems [21]. On the other hand, with the unprecedented interest in data-driven control approaches, such as model-free reinforcement learning (RL) [22], 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 [9], [23]. 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.

4) Autoregressive Processes With Sparse Graphical Models: Another closely related line of research studies the inference of autoregressive processes, whose structures can be captured via sparse graphical models. Earlier works on the inference of sparse autoregressive graphical models were based on hypothesis testing [24], [25]. More recently, the work [26] has proposed an ℓ_1-regularized maximum likelihood estimator for estimating the precision matrices of autoregressive Gaussian processes. A similar regularized estimator is also used in [27] to infer autoregressive processes with sparse latent-variable graphical models. Alternatively, the work [28] introduced a Bayesian approach for the inference of autoregressive graphical models. While being related to our proposed method, these works rely upon a different underlying generative model for the system and, hence, are not directly applicable to the system identification of LTI systems.

B. Contributions

In this work, we introduce a regularized estimator for recovering the true block-sparsity of an LTI system. In particular, we use an ℓ_1/ℓ_∞-regularized estimator, i.e., a least-squares estimator accompanied by an ℓ_∞ regularizer on different blocks. We show that the required number of sample trajectories to recover the nonzero blocks of the system matrices and to guarantee a small estimation error scales polynomially with the maximum block sizes and the number of row- and columnwise nonzero elements, but only logarithmically with the number of blocks in the system.

Our work makes a significant improvement over the recently studied least-squares estimator, whose sample complexity scales linearly with the system dimensions. Most interconnected systems consist of many smaller subsystems (blocks) with sparse or localized interactions. Under such circumstances, it may be costly, if not impossible, to collect as many samples as the system dimension. 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, Pereira *et al.* [29] ignore the block structure of the system and assume autonomy and inherent stability, all of which will be relaxed in this work.

This work is a significant extension of our previous conference papers on Lasso-type estimators for system identification [30] and nonasymptotic analysis of block-regularized linear regression problems [31]. 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 [30] and [31] 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.

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 jth column and the ith 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 $f_1(n) \leq c_1 f_2(n)$ and $f_1(n) \geq c_2 f_2(n)$, 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.

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 independent identically distributed (i.i.d.) with distribution $N(0, \Sigma_w)$ 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 $i = 1, 2, \ldots, d$, where d is the number of available sample trajectories. The ith 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 [11].

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. Such a reset-and-run approach is possible and even crucial in many problems of practical relevance. For instance, having the ability to reset cyber-physical systems to a zero or safe state at any given time is deemed crucial to ensure the safety of the system and to protect it from malicious attacks [32], [33]. Moreover, the recent advancement in RL lends itself to the user's ability to run the system in different and independent sample trajectories (also known as *rollouts* or *episodes* in the RL literature), each with a controlled and independent initial state.

Given the sample trajectories $\{(x^{(i)}[\tau], u^{(i)}[\tau])\}_{\tau=0}^{\top}$ for $i = 1, 2, \ldots, d$, 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

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

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

for every sample trajectory $i = 1, 2, \ldots, d$. Furthermore, let Y, X, and W be defined as vertical concatenations of $Y^{(i)}$, $X^{(i)}$, and $W^{(i)}$ for $i = 1, 2, \ldots, d$, respectively. Finally, denote $\Psi = [A \;\; B]^\top$ 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, Dean *et al.* [9] simplify 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 following 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, Dean *et al.* [9] show that, with input vectors $u^{(i)}[t]$ chosen randomly from $N(0, \Sigma_u)$ 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 $\|E\|_2 = \mathcal{O}(\sqrt{(m+n)\log(1/\delta)/d})$ 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 the Appendixes.

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)th block of A with size $n_i \times n_j$. Similarly, $B^{(k,l)}$ is the (k,l)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 $\Psi = [A \;\; B]^\top$. Note that $\Psi^{(i,j)} = (A^{(j,i)})^\top$ for $i \in \{1, \ldots, \bar{n}\}$ and $\Psi^{(i,j)} = (B^{(j,i-\bar{n})})^\top$ for $i \in \{\bar{n}+1, \ldots, \bar{n}+\bar{m}\}$. 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:

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

where $\|\Psi\|_{\text{block}}$ is defined as the summation of $\|\Psi^{(i,j)}\|_\infty$ over $(i,j) \in \{1, \ldots, \bar{n}+\bar{m}\} \times \{1, \ldots, \bar{n}\}$. 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 nonasymptotic 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 $d = \Omega(k_{\max}^2(D\log(\bar{n}+\bar{m}) + D^2\log(1/\delta)))$. Here, $k_{\max}$ is the maximum number of nonzero elements in the columns of $[A \;\; B]^\top$. 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

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

Suppose that $u^{(i)}[t]$ and $w^{(i)}[t]$ are i.i.d. samples of $N(0, \Sigma_u)$ and $N(0, \Sigma_w)$, respectively. Moreover, we assume that the initial state is random with a Gaussian distribution $N(0, \Sigma_x)$. Therefore, (8) and (10) imply that

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

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

$$\tilde{\Sigma} = \begin{bmatrix} C^\top C & 0 \\ 0 & \Sigma_u \end{bmatrix}, C = \begin{bmatrix} F_T^\top \\ G_T^\top \end{bmatrix} \tag{12a}$$

$$F_T = \begin{bmatrix} A^{T-2}B\Sigma_u^{1/2} & A^{T-3}B\Sigma_u^{1/2} & \ldots & B\Sigma_u^{1/2} \end{bmatrix} \tag{12b}$$

$$G_T = \begin{bmatrix} A^{T-1}\Sigma_x^{1/2} & A^{T-2}\Sigma_w^{1/2} & A^{T-3}\Sigma_w^{1/2} & \ldots & \Sigma_w^{1/2} \end{bmatrix}. \tag{12c}$$

The matrix C is referred to as the *combined controllability matrix* in the following. 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 $\mathcal{T} \subseteq \{1, \ldots, \bar{n}+\bar{m}\}$, 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 jth 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 article.

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

A1) (Mutual incoherency property): There exists a number $\gamma \in (0, 1]$ such that

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

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

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

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

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

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

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

The mutual incoherency property in Assumption A1 is a commonly known condition for the exact recovery of unknown parameters in compressive sensing and classical Lasso problems [34], [35]. 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 $W = 0$. It is also worth noting that this condition can be further relaxed under additional conditions [36]. 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.

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 eigenvalues of Σ_u and Σ_w do not scale with the system dimension, it is easy to verify that $\min\{\lambda_{\min}(\Sigma_u), \lambda_{\min}(\Sigma_w)\} \leq \Lambda_{\min} \leq \lambda_{\min}(\Sigma_w)$. 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 $\max_{1 \leq i \leq n+m} \tilde{\Sigma}_{ii} \leq \bar{\sigma}_{\max}^2$.

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 $n_{\max} = \Omega((\bar{n} + \bar{m})^{\log n})$ and $m_{\max} = \Omega((\bar{n} + \bar{m})^{\log m})$. It is worthwhile to mention that Assumption A4 results in $k_{\max} = O((\bar{n} + \bar{m})^{\alpha_k})$ for some number $\alpha_k < \infty$; this will be used later in the derivations.

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

Theorem 1 (Blockwise 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

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

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 are 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.

Remark 1: In this article, we assume that A and B are partitioned into blocks with known sizes, each with a maximum size of D. If the blocks sizes are unknown, an alternative approach is to treat A and B as sparse matrices, where each block is of size $D = 1$. This lack of prior knowledge on the block sizes of the system matrices can be compensated with a higher number of collected sample trajectories from the system. In particular, as it is shown in [30], an elementwise regularized estimator (i.e., vanilla Lasso) can still recover the correct sparsity pattern of the true system matrices with no prior knowledge on the block sizes, albeit with potentially a higher number of sample trajectories and worse estimation error. In Section IV, we showcase the performance of these regularized estimators with and without prior knowledge on the block sizes.

It is worth noting that, based on Theorem 1, one may speculate that setting $D = 1$ (i.e., not using the prior information on the block sizes) may lead to a better statistical guarantee. However, note that the derived bound is based on a customized λ_d that is designed to obtain a logarithmic dependence on $\bar{n} + \bar{m}$. This λ_d is specifically designed to offer a small value in terms of $\bar{n} + \bar{m}$ without optimizing its dependence on D. To obtain a tighter bound with respect to D (instead of $\bar{n} + \bar{m}$), one may need to select another λ_d that 1) would depend on D in a more sophisticated way, and 2) similar to [37], would potentially depend on the level of "overlap" in the blockwise support of the unknown parameters. We consider obtaining a better dependence on D as an enticing challenge for future research.

Remark 2: Similar to the classical results on the regularized linear regression [37], [38], the particular choice of the regularization coefficient λ_d in our analysis depends on the unknown parameters of the true system, such as σ_w, $\sigma_{\max}$, and γ. As will be shown in the next section, in practice, we do not rely on these unknown parameters. In particular, the chosen value for λ_d in our simulations will merely depend on the known parameters of the system, such as d, $\bar{n} + \bar{m}$, and D when we know the block sizes, or d and $n + m$ when the block sizes are unknown.

Remark 3: Another alternative approach to promote the block sparsity in the identification of dynamical systems is the ℓ_1/ℓ_2-regularized estimator (also known as group Lasso), where the ℓ_∞ regularization on different blocks is replaced by an ℓ_2 regularization [39], [40]. In Section IV, it is empirically shown that these estimators offer a similar performance in terms of the estimation error. However, an important advantage of the ℓ_1/ℓ_∞-regularized estimator over the group Lasso is in terms of its computational complexity. As pointed out in [41], one of the main benefits of the ℓ_1/ℓ_∞-regularized estimator lies in the efficient computation of its *entire solution path* over a compact range of regularization coefficients (as opposed to a single regularization coefficient). In particular, contrary to the group Lasso, the solution path for the ℓ_1/ℓ_∞-regularized estimator is piecewise linear with easily computable breakpoints. This, in turn, can be used in sensitivity analysis and boosting methods [41], [42].

A. Comparison to Least Squares

In this subsection, we 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. The proof is omitted for brevity and can be found in [43]. □

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

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

Then, we have

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

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

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

provided that

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

Proof. The proof is omitted for brevity and can be found in [43]. □

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 multiagent 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 elementwise error of the estimated parameters depend on $\kappa(\tilde{\Sigma})$. Recall that $\min\{\lambda_{\min}(\Sigma_u), \lambda_{\min}(\Sigma_w)\} \leq \Lambda_{\min} \leq \lambda_{\min}(\Sigma_w)$, 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 implies 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 are required to accurately identify the dynamics of the system if all of its modes are easily excitable. The dependence of the estimation error on the modes of the system is also reflected in the nonasymptotic error bound of the least-squares estimator in [9]. 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.

On the other hand, for fixed σ_w, the performance of the estimator deteriorates as the expected energy of the input decreases. In the extreme case of zero input, we inevitably have $\Lambda_{\min} = 0$, which, in turn, implies that the proposed estimator provides no guarantee on the accuracy of the estimated parameters.

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 1; it is diagonalizable, and n is fixed. One can easily verify that the following chain of inequalities holds:

$$\begin{aligned}\lambda_{\max}(\tilde{\Sigma}) &\geq \lambda_{\max}(\sigma_u^2 F_T F_T^\top + \sigma_w^2 G_T G_T^\top)\\ &\geq \lambda_{\min}(\Sigma_w)\lambda_{\max}\left(A^{T-2}(A^{T-2})^\top\right)\\ &\geq \lambda_{\min}(\Sigma_w)\max_i\left\{\left(\left(A^{T-2}\left(A^{T-2}\right)^\top\right)_{ii}\right)^2\right\}\\ &\geq \frac{\lambda_{\min}(\Sigma_w)}{n}\|A^{T-2}\|_\infty \geq \frac{\lambda_{\min}(\Sigma_w)}{n}\rho(A)^{T-2}.\end{aligned} \tag{26}$$

This exponential dependence is also empirically observed in our numerical experiments.

C. Mutual Incoherency

In this subsection, we will analyze the mutual incoherency condition (14). In particular, we will show that the proposed mutual incoherency condition is tightly related to the so-called *identifiability* condition and, hence, cannot be relaxed for specific classes of problems. 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:

$$\min_{\Psi} \quad \|\Psi\|_0 \tag{27a}$$

$$\text{s.t.} \quad X\Psi = Y - W. \tag{27b}$$

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 nonconvex 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

$$\mu(A) = \max_{1\leq i<j\leq t_2}\frac{|A_{:,i}^\top A_{:,j}|}{\|A_{:,i}\|_2\|A_{:,j}\|_2}. \tag{28}$$

In other words, $\mu(A)$ measures the maximum correlation between distinct columns of A (with a slight abuse of notation, we assume that

$$\frac{1}{\mu(A)} = +\infty \text{ if } \mu(A) = 0).$$

Reminiscent of the classical results in the compressive sensing literature, it is well known that the optimal solution Ψ^* of (27) is unique if the *identifiability* condition

$$\|\Psi^*_{:,j}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right) \tag{29}$$

holds for every $j = 1, 2, \ldots, n$ (see, e.g., [44, Th. 2.5]). Furthermore, this bound cannot be tightened, since there exist instances of the problem, for which the violation of

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

for some j results in the nonuniqueness 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 (12a). This, together with the definition of $\tilde{\Sigma}$, implies that

$$\begin{aligned}\mu(X) &= \max_{1\leq i<j\leq m+n}\frac{|X_{:,i}^\top X_{:,j}|}{\|X_{:,i}\|_2\|X_{:,j}\|_2}\\ &= \max_{1\leq i<j\leq n}\frac{|C_{:,i}^\top C_{:,j}|}{\|C_{:,i}\|_2\|C_{:,j}\|_2} = \mu(C).\end{aligned} \tag{30}$$

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 1:

$$\left\|(C_{:,\mathcal{A}_j}^\top C_{:,\mathcal{A}_j})^{-1}C_{:,\mathcal{A}_j}^\top C_{:,k}\right\|_1 \leq 1-\alpha \qquad \forall k\in\mathcal{A}_j^c,\ j=1,2,\ldots,n \tag{31}$$

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 (29), 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 (29) (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.

1) 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 $\varphi > 0$.
2) T is set to 3.
3) $\Sigma_u = \sigma_u^2 I$ and $\Sigma_w = \sigma_w^2 I$, where σ_u and σ_w are less than or equal to 1. Moreover, $\Sigma_x = 0$.

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

(1) There exists an instance for which the identifiability condition fails to hold for the oracle estimator if $\varphi \geq \frac{3}{k}$.

(2) The mutual incoherency condition holds if $\varphi < \frac{\sigma_u + \sigma_w}{9^\frown k}$.

Proof. The proof is omitted for brevity and can be found in [43]. □

The tightness of the identifiability condition (29) together with the Proposition 1 implies that for some specific classes of problems, it is not possible to a have a consistent sparsity promoting technique with significantly more relaxed conditions than the ones introduced in this article. However, we point out that, in general, the mutual incoherency condition may be improved by resorting to more sophisticated (and potentially nonconvex) estimators [36]. Moreover, it will be shown in Section IV that the incoherency condition is expected to hold in many cases of practical relevance.

IV. Numerical Results

In this section, we illustrate the performance of the block-regularized estimator and compare it with its least-squares counterpart. We consider two case studies on switching networks and power systems.

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.

A. Case Study 1: Switching Networks

In this case study, we study a network of multiagent systems that are interconnected through a switching information exchange topology. Recently, a special attention has been devoted to multiagent systems with a time-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 [45]. 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:

$$\dot{x}_i(t) = \sum_{(i,j)\in\mathcal{N}_x(i)} A^{(i,j)} x_j(t) + \sum_{(i,j)\in\mathcal{N}_u(i)} B^{(i,j)} u_j(t) + w_i(t) \tag{32}$$

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)th blocks of A and B, respectively. 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 five 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 $[-0.4 \ \ -0.3] \cup [0.3 \ 0.4]$. Moreover, the regularization coefficient λ_d is set to

$$\sqrt{\frac{2(D^2 + D\log(\bar{n} + \bar{m}))}{d}}. \tag{33}$$

Note that this choice of λ_d does not rely on the unknown parameters of the system, and it does not require any additional fine-tuning. 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 s. This implies that each sample trajectory is collected within 0.6 s. The number of block mismatch and 2-norm estimation errors are depicted in Fig. 1(a) and (b) 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.

Fig. 1(a) 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, 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.

Fig. 1(b) 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 $d \geq 200$.

Finally, we compare the proposed estimator with group Lasso, where the ℓ_∞ regularization is replaced by an ℓ_2 regularization. Suppose that $D = 25$, and the regularization coefficient for the group Lasso (i.e., ℓ_1/ℓ_2-regularized estimator) is chosen as

$$\lambda = \sqrt{\frac{0.5(D^2 + D\log(\bar{n} + \bar{m}))}{d}}$$

(the constant factor is fine-tuned for this case study). According to Fig. 1(c), the proposed ℓ_1/ℓ_∞ slightly outperforms group Lasso in terms of the mismatch error. On the other hand, Fig. 1(d) illustrates that neither of the estimators is superior in terms of the estimation error. As a future research direction, we plan to conduct a more comprehensive study on the group Lasso and its statistical performance in the context of system identification.

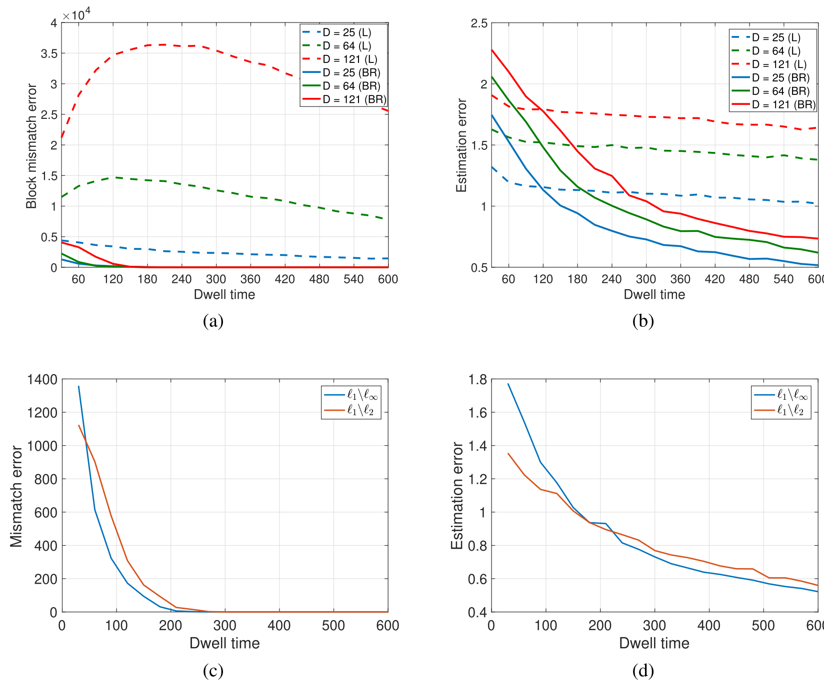


Fig. 1. (a) Block mismatch error for the block-regularized (abbreviated as BR) and Lasso (abbreviated as L) estimators. (b) Estimation error for the block-regularized and Lasso estimators. (c) Block mismatch error for the block-regularized and Lasso estimators. (d) Normalized estimation error for the block-regularized and Lasso estimators.

B. Case Study 2: Power Systems

For the second case study, we consider the frequency control problem for power systems, where the goal is to control the governing frequency of the entire network based on the so-called *swing* equations [46]. Assume that there exist N_g generators in the system. It is easy to describe the swing equations using the well-known dc approximation

$$M_i\ddot{\theta}_i + D_i\dot{\theta}_i = P_{M_i} - P_{E_i}$$

where θ_i is the voltage angle at generator i, P_{M_i} is the mechanical power input at generator i, and P_{E_i} denotes the active power injection at the bus connected to generator i. Furthermore, M_i and D_i are the inertia and damping coefficients at generator i, respectively. Under the dc approximation, the relationship between active power injection and voltage can be written as

$$P_{E_i} = \sum_{j\in\mathcal{N}_i} B_{ij}(\theta_i - \theta_j)$$

where $\mathcal{N}_i$ collects the neighbors of generator i, and B_{ij} is the susceptance of the line (i, j). After discretization with the sampling time dt, the system of swing equations is reduced to the following dynamical system:

$$x_i[t+1] = \left(A_{ii}x_i[t] + \sum_{j\in\mathcal{N}_i} A_{ij}x_j[t]\right) + B_{ii}u_i[t] + w_i[t]$$

where $x_i = \begin{bmatrix}\theta_i & \dot{\theta}_i\end{bmatrix}^\top$, $u_i(t) = P_{M_i}$, and

$$A_{ii} = \begin{bmatrix} 1 & dt \\ -\frac{\sum_{j\in\mathcal{N}_i} B_{ij}}{M_i}dt & 1 - \frac{D_i}{M_i}dt \end{bmatrix}, A_{ij} = \begin{bmatrix} 0 & 0 \\ \frac{B_{ij}}{M_i}dt & 0 \end{bmatrix}$$

$$B_{ii} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Realistic power systems are often equipped with an initial distributed controller whose sensing and actuation communication topology is limited by the underlying physical structure of the system [47]. In particular, consider a static distributed controller as follows:

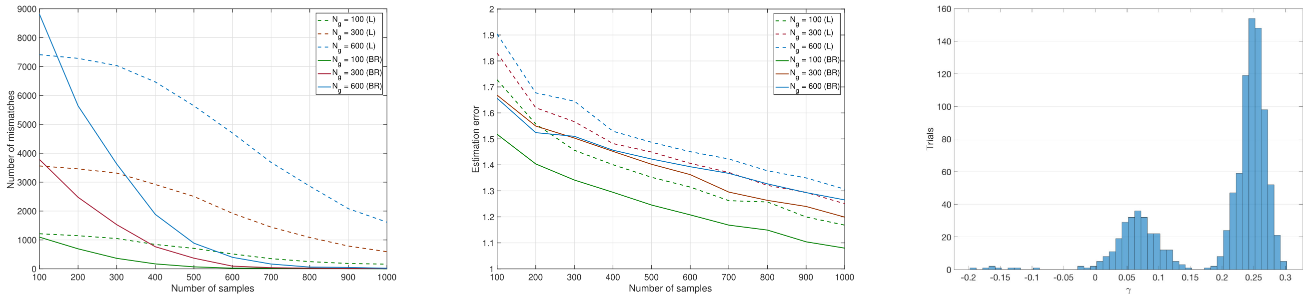
$$u_i[t] = K_{ii}x_i[t] + \sum_{j\in\mathcal{N}_i} K_{ij}x_j[t] + v_i[t] \tag{34}$$

where K is a matrix with (i, j)th block equal to zero if the generators i and j are not connected. Moreover, $v_i[t]$ is an exogenous input. Therefore, the closed-loop dynamics of the power system can be written as

$$x_i[t+1] = \left(A^c_{ii}x_i[t] + \sum_{j\in\mathcal{N}_i} A^c_{ij}x_j[t]\right) + B_{ii}v_i[t] + w_i[t]$$

where

$$A^c_{ii} = \begin{bmatrix} 1 & dt \\ -\frac{\sum_{j\in\mathcal{N}_i} B_{ij}}{M_i}dt + K^1_{ii} & 1 - \frac{D_i}{M_i}dt + K^2_{ii} \end{bmatrix}$$



(a) (b) (c)

Fig. 2. (a) Block mismatch error for the block-regularized and Lasso estimators. (b) Estimation error for the block-regularized and Lasso estimators. (c) Distribution of mutual incoherency parameter γ over 1000 instances of the problem.

$$A^c_{ij} = \begin{bmatrix} 0 & 0 \\ \frac{B_{ij}}{M_i}dt + K^1_{ij} & K^2_{ij} \end{bmatrix} \tag{35}$$

and $K_{ij} = \begin{bmatrix} K^1_{ij} & K^2_{ij} \end{bmatrix}$ for every block (i, j)th. Our goal is to identify the closed-loop dynamics of the power system and the underlying topology of the network, based on the sample trajectories collected from the system. Note that the underlying topology structure of the network can be naturally obtained from the block-sparsity structure of A^c: the block A^c_{ij} is equal to zero if and only if the generators i and j are not connected. Therefore, the topology inference problem reduces to obtaining the correct block-sparsity pattern of the system matrices A^c and B^c. To assess the performance of the proposed method, we generate different instances of the problem according to the following rules.

(1) The generators are connected via a randomly generated graph with the average degree of 6.
(2) The parameters B_{ij}, M_i, and D_i are uniformly chosen from the intervals [0.5,1], [1,2], and [0.5,1.5], respectively.
(3) The nonzero elements of K are uniformly chosen from the interval $[0.1, 0.2]$.

The sampling time dt is set to 0.1. We assume that the disturbance noise has a zero-mean Gaussian distribution with variance 0.01. The mechanical input $v_i(t)$ is randomly generated according to a zero-mean Gaussian distribution with variance 0.05. In this case study, we compare the performance of the block-regularized and Lasso estimators. The regularization coefficients for these estimators are chosen as $\sqrt{0.1(D^2 + D\log(\bar{n} + \bar{m}))/d}$ with $D = 4$ (i.e., the maximum block size), and $\sqrt{0.01(1 + \log(n + m))/d}$, respectively.

Fig. 2(a) illustrates the mismatch error of these estimators for different numbers of generators N_g chosen from $\{100, 300, 600\}$. Not surprisingly, the learning time needed to achieve a small mismatch error for both estimators increases as the dimension of the system grows. Conversely, a smaller value for RST is needed to achieve infinitesimal RME for larger systems. In particular, when N_g is equal to 100, 300, and 600, the minimum RST for the proposed block-regularized estimator to guarantee RME $\leq 0.1\%$ is equal to 0.67, 0.45, and 0.29, respectively. On the other hand, the minimum RST for the Lasso to achieve the same RME is on average 2.45 times larger than that of the block-regularized estimator.

Fig. 2(b) depicts the 2-norm of the estimation error of the block-regularized and Lasso estimators. In can be seen that the estimation error of the block-regularized estimator is strictly smaller than that of the Lasso, highlighting its superior performance in the block-sparse systems.

Finally, Fig. 2(c) illustrates the distribution of the mutual incoherency parameter γ for 1000 randomly generated instances of power systems with 300 generators. It can be seen that only 1.2% of the instances violate the mutual incoherency condition (14) due to the negative values of γ. This highlights the nonconservativeness of this condition in practice.

V. Conclusion

We consider the problem of identifying the parameters of LTI systems. In many real-world problems, the state-space equation 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. We derive sharp nonasymptotic bounds on the minimum number of input-state data samples to guarantee a small elementwise 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 elementwise error. Through different case studies on switching networks and power systems, we demonstrate the performance of the proposed estimator.

Appendix A
Proof of Main Theorem

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

$$Y^{(:,j)} = X\Psi^{(:,j)} + W^{(:,j)} \quad \forall j \in \{1, \ldots, \bar{n}\} \tag{36}$$

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

$$\hat{\Psi}^{(:,j)} = \arg\min \frac{1}{2\tilde{d}}\|Y^{(:,j)} - X\Psi^{(:,j)}\|_2^2 + \lambda_d\|\Psi^{(:,j)}\|_{\text{block}}. \tag{37}$$

Define $D_j = p_{\max}n_j$. The next two lemmas are at the core of our proof for Theorem 1. Due to space restrictions, we have deferred their proofs to the extended version of this article [43].

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

$$\lambda_d \geq \sqrt{\frac{32c_1\lambda_{\max}(\Sigma_w)^2\sigma_{\max}^2}{\gamma^2}\cdot\frac{(D_j)^2 + D_j\log(\bar{n}+\bar{m})}{d}} \tag{38a}$$

$$d \geq \frac{72c_2\sigma_{\max}^2}{\gamma^2\Lambda_{\min}}\cdot k_j(D_j^2 + D_j\log(\bar{n}+\bar{m})). \tag{38b}$$

Then, with probability of at least

$$\begin{aligned} 1 - 3\exp\left(-(c_1-1)(D_j + \log(\bar{n}+\bar{m}))\right)& \\ - 4\exp\left(-(c_2-1)(D_j+\log(\bar{n}+\bar{m}))\right)& \end{aligned} \tag{39}$$

$\hat{\Psi}^{(:,j)}$ is unique and its nonzero blocks exclude the zero blocks of $\Psi^{*(:,j)}$.

Due to Assumption A4, we have $n_{\max} = O((\bar{n}+\bar{m})^{\alpha_n})$ and $k_{\max} = O((\bar{n}+\bar{m})^{\alpha_k})$ for some $\alpha_n \geq 0$ and $\alpha_k \geq 0$.

Lemma 2 (Elementwise 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

$$\begin{aligned} 1 - 2\exp(-(k_jn_j + c_3\log(\bar{n}+\bar{m}))/2) - 2\exp\left(-d/2\right)& \\ - 2\exp\left(-2(c_4-1)(\alpha_n+\alpha_k)\log(\bar{n}+\bar{m})\right)& \end{aligned} \tag{40}$$

we have

$$\begin{aligned} &\|\hat{\Psi}^{(:,j)} - \Psi^{*(:,j)}\|_\infty \leq \\ &\sqrt{\frac{36c_4(\alpha_n+\alpha_k)\lambda_{\max}(\Sigma_w)^2\log(\bar{n}+\bar{m})}{\Lambda_{\min}d}} \\ &+ \frac{\lambda_d}{\Lambda_{\min}}\left(8\sqrt{k_j}\sqrt{\frac{k_jn_j + c_3\log(\bar{n}+\bar{m})}{d}} + 1\right) = g_j. \end{aligned} \tag{41}$$

Furthermore, the zero blocks of $\hat{\Psi}^{(:,j)}$ exclude the nonzero blocks of $\Psi^{*(:,j)}$ if $\min_{i\in\mathcal{A}_j}\|\Psi^{(i,j)}\|_\infty > g_j$.

Most of the existing block-sparsity methods in linear regression focus on the problems, where the blocks have row or column dimension of one [37], [40], [48]–[50] and, hence, are not applicable to problems with arbitrary block sizes. On the other hand, recall that many large-scale dynamical systems are composed of interacting subsystems, each with its own local states/inputs with potentially different sizes. This imposes a general block structure on different rows and columns of the matrices A and B, and hence, the existing results on block-regularized estimators cannot be readily used in these settings. Lemmas 1 and 2 are precisely aimed to address this issue and will play key roles in proving the main theorem of this article. The proofs of Lemmas 1 and 2 are based on the extended version of the so-called primal-dual witness approach, which was initially proposed in [38] for element- or rowwise sparse structures. The details of this generalization can be found in the extended version of this article [43].

Appendix B
Proof of Theorem 1

First, we present the proof in a few steps.

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

$$\hat{\Psi} = \arg\min_{\Psi}\sum_{j=1}^{n}\left(\frac{1}{2\tilde{d}}\|Y^{(:,j)} - X\Psi^{(:,j)}\|_2^2 + \lambda\|\Psi^{(:,j)}\|_{\text{block}}\right). \tag{42}$$

The above optimization problem can be decomposed into $\bar{n}$ disjoint block-regularized subproblems in the form of (37).

Step 2: Assume that (38b) and (38a) hold for every $1 \leq j \leq \bar{n}$. Upon defining $\mathcal{T}_j$ as the event that Lemmas 1 and 2 hold, one can write

$$\begin{aligned} \mathbb{P}(\mathcal{T}_j) \geq 1 &- 5\exp\left(-(c_1-1)(D_j+\log(\bar{n}+\bar{m}))\right) \\ &- 4\exp\left(-(c_2-1)(D_j+\log(\bar{n}+\bar{m}))\right) \\ &- 2\exp(-(k_jn_j + c_3\log(\bar{n}+\bar{m}))/2) \\ &- 2\exp\left(-2(c_4-1)(\alpha_n+\alpha_k)\log(\bar{n}+\bar{m})\right) \end{aligned} \tag{43}$$

for every $1 \leq j \leq \bar{n}$.

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 (43) and a simple union bound, one can write

$$\begin{aligned} \mathbb{P}(\mathcal{T}) \geq 1 &- \underbrace{K_1(\bar{n}+\bar{m})^{-(c_1-2)}}_{(a)} - \underbrace{K_2(\bar{n}+\bar{m})^{-(c_2-2)}}_{(b)} \\ &- \underbrace{K_3(\bar{n}+\bar{m})^{-(\frac{c_3}{2}-1)}}_{(c)} \\ &- \underbrace{K_4(\bar{n}+\bar{m})^{-(2(\alpha_n+\alpha_k)(c_4-1)-1)}}_{(d)} \end{aligned} \tag{44}$$

for some constants $K_1, K_2, K_3,$ and K_4. One can easily verify that the following equalities are enough to guarantee that the right-hand side of (44) is equal to $1-\delta$:

$$\begin{aligned} c_1 &= \frac{\log(4K_1/\delta)}{\log(\bar{n}+\bar{m})} + 2,\ c_2 = \frac{\log(4K_2/\delta)}{\log(\bar{n}+\bar{m})} + 2 \\ c_3 &= \frac{2\log(4K_3/\delta)}{\log(\bar{n}+\bar{m})} + 2 \\ c_4 &= \frac{\log(4K_4/\delta)}{2(\alpha_n+\alpha_k)\log(\bar{n}+\bar{m})} + \frac{1}{2(\alpha_n+\alpha_k)} + 1. \end{aligned} \tag{45}$$

Substituting (45) into Lemmas 1 and 2 leads to two observations.

(1) If λ_d and d satisfy (18a) and (18b), then they also satisfy (38a) and (38b).

(2) The parameter g defined in (19) is greater than or equal to g_j for every $j = 1, \ldots, \bar{n}$.

Therefore, (18a) and (18b) guarantee that: 1) $\hat{\Psi}$ is unique and does not have any false positive in its blocks, and 2) its elementwise 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

$$d = \Omega\left(C_\Psi \kappa(\tilde{\Sigma})^2 k_{\max}\left(D\log(\bar{n}+\bar{m}) + D^2\log(1/\delta)\right)\right). \tag{46}$$

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

$$g \leq K\left(\frac{1}{\sqrt{C_\Psi k_{\max}}} + \frac{1}{C_\Psi \kappa(\tilde{\Sigma})}\right) \tag{47}$$

for some constant K. Therefore,

$$C_\Psi = \frac{2/K}{t_{\min}\kappa(\tilde{\Sigma})} + \frac{4/K}{t_{\min}^2 k_{\max}} = O(1) \tag{48}$$

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

References

[1] M. Pai, *Energy Function Analysis for Power System Stability*. New York, NY, USA: Springer, 2012.

[2] Y. Wang, D. J. Hill, and G. Guo, "Robust decentralized control for multimachine power systems," *IEEE Trans. Circuits Syst. I: Fundam. Theory Appl.*, vol. 45, no. 3, pp. 271–279, Mar. 1998.

[3] I. Sa, M. Kamel, R. Khanna, M. Popovic, J. Nieto, and R. Siegwart, "Dynamic system identification, and control for a cost effective open-source VTOL MAV," 2017, *arXiv:1701.08623*.

[4] M. Arcak, "Synchronization and pattern formation in diffusively coupled systems," in *Proc. IEEE 51st IEEE Conf. Decis. Control*, 2012, pp. 7184–7192.

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

[6] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

[7] 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," *Proc. Nat. Acad. Sci.*, vol. 109, no. 15, pp. 5868–5873, 2012.

[8] S. Sun, R. Huang, and Y. Gao, "Network-scale traffic modeling and forecasting with graphical lasso and neural networks," *J. Transp. Eng.*, vol. 138, no. 11, pp. 1358–1367, 2012.

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

[10] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Proc. 31st Conf. Learn. Theory*, 2018, pp. 439–473.

[11] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems?" in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 5610–5618.

[12] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," in *Proc. Amer. Control Conf.*, 2019, pp. 5655–5661.

[13] V. L. Le, F. Lauer, and G. Bloch, "Selective ℓ_1 minimization for sparse recovery," *IEEE Trans. Autom. Control*, vol. 59, no. 11, pp. 3008–3013, Nov. 2014.

[14] B. M. Sanandaji, T. L. Vincent, M. B. Wakin, R. Tóth, and K. Poolla, "Compressive system identification of LTI and LTV ARX models," in *Proc. 50th IEEE Conf. Decis. Control Eur. Control Conf.*, 2011, pp. 791–798.

[15] R. Tóth, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Compressive system identification in the linear time-invariant framework," in *Proc. 50th IEEE Conf. Decis. Control Eur. Control Conf.*, 2011, pp. 783–790.

[16] X. Jiang, Y. Yao, H. Liu, and L. Guibas, "Compressive network analysis," *IEEE Trans. Autom. Control*, vol. 59, no. 11, pp. 2946–2961, Nov. 2014.

[17] V. Cerone, S. M. Fosson, and D. Regruto, "Sparse linear regression with compressed and low-precision data via concave quadratic programming," in *Proc. IEEE 58th Conf. Decis. Control*, 2019, pp. 6971–6976.

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

[19] 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 Trans. Autom. Control*, vol. 59, no. 11, pp. 2933–2945, Nov. 2014.

[20] R. S. Smith, "Frequency domain subspace identification using nuclear norm minimization and Hankel matrix realizations," *IEEE Trans. Autom. Control*, vol. 59, no. 11, pp. 2886–2896, Nov. 2014.

[21] S. M. Fosson, "Online optimization in dynamic environments: A regret analysis for sparse problems," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 7225–7230.

[22] S. Ross and J. A. Bagnell, "Agnostic system identification for model-based reinforcement learning," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1905–1912.

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

[24] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, no. 2, pp. 157–172, 2000.

[25] M. Eichler, "Testing nonparametric and semiparametric hypotheses in vector stationary processes," *J. Multivariate Anal.*, vol. 99, no. 5, pp. 968–1009, 2008.

[26] J. Songsiri and L. Vandenberghe, "Topology selection in graphical models of autoregressive processes," *J. Mach. Learn. Res.*, vol. 11, pp. 2671–2705, 2010.

[27] M. Zorzi and R. Sepulchre, "AR identification of latent-variable graphical models," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2327–2340, Sep. 2016.

[28] D. F. Ahelegbey, M. Billio, and R. Casarin, "Sparse graphical vector autoregression: A Bayesian approach," *Ann. Econ. Statist.*, GENES, nos. 123/ 124, pp. 333–361, 2016.

[29] J. Pereira, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 172–180.

[30] S. Fattahi and S. Sojoudi, "Data-driven sparse system identification," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput.*, 2018, pp. 462–469.

[31] S. Fattahi and S. Sojoudi, "Non-asymptotic analysis of block-regularized regression problem," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 27–34.

[32] F. Abdi, M. Hasan, S. Mohan, D. Agarwal, and M. Caccamo, "ReSecure: A restart-based security protocol for tightly actuated hard real-time systems," *in Proc. IEEE 1st Workshop Secur. Dependability Crit. Embedded Real-Time Syst.*, 2016, pp. 47–54.

[33] K. Andersson, B. Lennartson, and M. Fabian, "Restarting manufacturing systems; restart states and restartability," *IEEE Trans. Autom. Sci. Eng.*, vol. 7, no. 3, pp. 486–499, Jul. 2010.

[34] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, 2006.

[35] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Probl.*, vol. 23, no. 3, 2007, Art. no. 969.

[36] V. Cerone, S. Fosson, D. Regruto, and A. Salam, "Sparse learning with concave regularization: Relaxation of the irrepresentable condition," in *Proc. 59th IEEE Conf. Decis. Control*, 2020, pp. 396–401.

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

[38] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[39] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Union support recovery in high-dimensional multivariate regression," in *Proc. 46th Annu. Allerton Conf. Commun., Control, Comput.*, 2008, pp. 21–26.

[40] J. Huang and T. Zhang, "The benefit of group sparsity," *Ann. Statist.*, vol. 38, no. 4, pp. 1978–2004, 2010.

[41] P. Zhao, G. Rocha, and B. Yu, "Grouped and hierarchical model selection through composite absolute penalties," Dept. Statist., Univ. California, Berkeley, CA, USA, Tech. Rep. 703, 2006.

[42] S. Rosset, "Topics in regularization and boosting," Ph.D. dissertation, Dept. Statist., Stanford Univ., Stanford, CA, USA, 2003.

[43] S. Fattahi and S. Sojoudi, "Sample complexity of sparse system identification problem," 2018, *arXiv:1803.07753*.

[44] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York, NY, USA: Springer, 2010.
[45] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton, NJ, USA: Princeton Univ. Press, 2010.
[46] X.-F. Wang, Y. Song, and M. Irving, *Modern Power Systems Analysis*. New York, NY, USA: Springer, 2010.
[47] M. Andreasson, D. V. Dimarogonas, K. H. Johansson, and H. Sandberg, "Distributed vs. centralized power systems frequency control," in *Proc. Eur. Control Conf.*, 2013, pp. 3524–3529.
[48] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *Ann. Statist.*, vol. 39, no. 1, pp. 1–47, 2011.
[49] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group Lasso," *J. Comput. Graph. Statist.*, vol. 22, no. 2, pp. 231–245, 2013.
[50] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statist. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

Salar Fattahi received the Ph.D. degree in industrial engineering and operations research from UC Berkeley in 2020. His research interests include optimization, machine learning, and control theory. He is an Assistant Professor with the University of Michigan, Ann Arbor, MI, USA. His research interests include the intersection of optimization, statistics, and control.

Prof. Fattahi was the recipient of several awards, including the 2020 INFORMS ENRE Best Student Paper Award (as a co-author), the 2020 IEEE Power and Energy Society General Meeting Best-of-the-Best Paper Award, and the 2018 INFORMS Data Mining Best Paper Award. He was also a finalist for the 2018 American Control Conference Best Paper Award.

Somayeh Sojoudi received the Ph.D. degree in control and dynamical systems from the California Institute of Technology in 2013. Her research interests include optimization, control theory, machine learning, and power systems. She is an Assistant Professor with the Department of Electrical Engineering and Computer Sciences and the Department of Mechanical Engineering, University of California, Berkeley, CA, USA.

Prof. Sojoudi is an Associate Editor for IEEE Transactions on Smart Grid, IEEE Access, and *Systems and Control Letters*. She is also a member of the Conference Editorial Board of the IEEE Control Systems Society. She received many awards and honors, including the INFORMS Optimization Society Prize for Young Researchers, the INFORMS Energy Best Publication Award, the INFORMS Data Mining Best Paper Award, the National Science Foundation CAREER Award, and the Office of Naval Research Young Investigator Award. She has also received several Best Student Conference Paper Awards (as advisor or co-author) from the IEEE Control Systems Society.