Learning Sparse Dynamical Systems from a Single Sample Trajectory

Salar Fattahi, Nikolai Matni, Somayeh Sojoudi *

Abstract

This paper addresses the problem of identifying sparse linear time-invariant (LTI) systems from a single sample trajectory generated by the system dynamics. We introduce a Lasso-like estimator for the parameters of the system, taking into account their sparse nature. Assuming that the system is stable, or that it is equipped with an initial stabilizing controller, we provide sharp finite-time guarantees on the accurate recovery of both the sparsity structure and the parameter values of the system. In particular, we show that the proposed estimator can correctly identify the sparsity pattern of the system matrices with high probability, provided that the length of the sample trajectory exceeds a threshold. Furthermore, we show that this threshold scales polynomially in the number of nonzero elements in the system matrices, but logarithmically in the system dimensions — this improves on existing sample complexity bounds for the sparse system identification problem. We further extend these results to obtain sharp bounds on the ℓ_∞-norm of the estimation error and show how different properties of the system—such as its stability level and *mutual incoherency*—affect this bound. Finally, an extensive case study on power systems is presented to illustrate the performance of the proposed estimation method.

1 Introduction

Modern cyber-physical systems, such as power grids, autonomous transportation systems, and distributed computing and sensing networks, are characterized by being large scale, spatially distributed, and by having complex ever changing dynamics and interconnected topologies. The distributed optimal control literature addresses set-point tracking and regulation in the distributed setting by assuming known dynamics with a sparse interconnections. Indeed, the underlying sparsity structure of a distributed system is aggressively (and necessarily) exploited, with foundational results showing that both tractability [1] and scalability [2, 3, 4, 5] in controller synthesis are only possible when the underlying dynamical system is suitably sparse. However, in this large-scale,

*Salar Fattahi is with the Department of Industrial Engineering and Operations Research, University of California, Berkeley. Nikolai Matni is with the Department of Electrical Engineering and Computer Sciences, 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, NSF Award 1808859 and AFSOR Award FA9550-19-1-0055.

dynamic, and complex setting, it is unclear how to obtain the necessary models of the dynamical systems. To address this issue, we use data-driven approaches to identify both the interconnected topology and the dynamic behavior of these systems for which first-principle modeling becomes either intractable or impractical for such large-scale dynamic systems.

This then raises a more fundamental question: how can data-driven methods be appropriately integrated into safety-critical control loops? This question has been addressed in the context of learning [6, 7], and control of a small-scale and dense unknown systems, e.g., a single autonomous vehicle or robot [8, 9, 10, 11, 12]. These works make clear that if a learned model is to be integrated into a safety-critical control loop, then it is essential that the uncertainty associated with the learned model be explicitly quantified. This way, the learned model and the uncertainty bounds can be integrated with tools from robust control to provide strong guarantees of system performance and stability. This paper takes a first step towards extending these results to the large-scale distributed setting by providing a sample efficient and computationally tractable algorithm for the identification of sparse dynamical systems, as well as providing sharp estimates on the corresponding model uncertainty.

Main contributions: We show that large-scale sparse system models can be identified with a complexity scaling quadratically with the number of nonzero elements in the underlying dynamical system—for systems composed of a large number of subsystems that only interact with a small number of local neighbors, this computational saving can be significant. We further provide sharp bounds on the corresponding model uncertainty, paving the way for the use of these models in safety-critical control loops. Finally, in contrast to previous work, we show that such models can be extracted from a single trajectory of the system. In the context of large-scale systems, the system resets needed by methods relying on independent trajectories become prohibitively more expensive and impractical—indeed contrast resetting a robotic arm and a power distribution network, and the increase in difficulty becomes apparent. Note that we defer a detailed comparison of our results to prior work to Section 3.

Paper organization: In Section 2, we formally define the sparse system-identification task that we consider, and introduce our Lasso-like estimator based on a single system trajectory. Section 3 presents our main result, and compares and contrasts it with existing results in the literature. We also show that some of the technical assumptions that we make are necessary for a well-posed problem. We then present an overview of our proof technique in Section **??**, and follow this up with an empirical study of our method on a power system in Section 4. We end with conclusions in Section 5.

Notation: For a matrix M, the symbols $|||M|||$, $|||M|||_\infty$, $\|M\|_F$, $\|M\|_1$, and $\|M\|_\infty$ are used to denote its induced spectral, induced infinity, Frobenius, element-wise ℓ_1/ℓ_1, and element-wise ℓ_∞/ℓ_∞ norms, respectively. Furthermore, $\|M\|_0$ refers to the number of nonzero elements in M. The symbols $M_{:j}$ and $M_{j:}$ indicate the j^{th} column and row of M, respectively. For a set $\mathcal{I}$, the symbol $|\mathcal{I}|$ denotes its cardinality. Given the index sets $\mathcal{U}$ and $\mathcal{V}$, define $M_{\mathcal{U}\mathcal{V}}$ as the $|\mathcal{U}| \times |\mathcal{V}|$ submatrix of M obtained by removing the rows and columns with indices not belonging to $\mathcal{U}$ and $\mathcal{V}$. The symbols c and c_i play the role of universal constants throughout the paper. $\mathbb{E}\{x\}$ denotes the expected value of a random variable x. For an event $\mathcal{E}$, the notation $\mathbb{P}(\mathcal{E})$ refers to its probability of occurrence. The notation $x_n \overset{a.s.}{\to} x$ means that a sequence of random variables x_n converges to x almost surely.

2 Problem Statement

Consider the linear time-invariant (LTI) system

$$x(t+1) = Ax(t) + Bu(t) + w(t) \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are the unknown state and input matrices, respectively. Furthermore, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $w(t) \in \mathbb{R}^n$ are the respective state, input, and disturbance vectors at time t.

The goal of this work is to estimate the underlying parameters of the dynamics, 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 and T is the length of each sample trajectory. To simplify the notations, the superscript i is dropped from the sample trajectories when $d = 1$.

This paper is concerned with the identification of high dimensional but sparse system matrices (A, B). Such high-dimensional sparse parameters arise in the context of large-scale distributed and multi-agent systems, where dynamic coupling arises due to local interactions between subsystems– it is this local interaction structure that results in correspondingly sparse system matrices. Examples of such systems include power grids, intelligent transportation systems, and distributed computation and sensing networks.

We now compare and contrast two approaches to collecting sample trajectories from a dynamical system (1):

Fixed d and variable T: In this method, the number of sample trajectories d is set to a fixed value (e.g., $d = 1$) and instead, a sufficiently long time horizon (also referred to as learning time) T is chosen to collect enough information about the dynamics. This approach is most suitable when the open-loop system is stable, or if a stabilizing controller is provided—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 [6]. From a practical perspective, system instability may also impose limits on how large the learning time can be in order to ensure system safety, thereby restricting the amount of data that can be collected.

Fixed T and variable d: In this approach, the learning time T is fixed and instead, the number of sample trajectories is chosen to be sufficiently large. Notice that this method is not dependent on the system stability. However, one needs to reset the initial state of the system at the beginning of each sample trajectory, which may not be possible in practice, especially in the case of large-scale systems.

This work focuses on *sparse* system identification using a single trajectory, where it is assumed that the system is either stable, or equipped with an initial stabilizing controller, and our goal is to both identify the supports of the sparse system matrices (A, B) and estimate their values, using a single sample trajectory. As mentioned in [8], in many applications, the existence of an initial stabilizing controller for the unknown system (1) is not restrictive. In fact, [9] introduces an offline procedure for designing such an initial stabilizing controller.

Indeed, one can cast the sparse system identification task as a *supervised learning* problem, where the goal is to fit the linear model (1)—parameterized by (A, B)—to a limited number of measurements $\{(x(\tau), u(\tau))\}_{\tau=0}^{T}$. Motivated by this observation, one can consider the following

M-estimator:

$$(\hat{A},\hat{B}) = \arg\min_{A,B} \frac{1}{2T}\sum_{t=0}^{T-1} \|x(t+1) - (Ax(t)+Bu(t))\|_2^2 + \lambda(\|A\|_1 + \|B\|_1). \tag{2}$$

where the first term corresponds to the maximum likelihood estimation of (A, B) when the disturbance noise has a zero-mean Gaussian distribution, and the second term has the role of promoting sparsity in the estimated $(\hat{A}, \hat{B})$.

Before proceeding, it is essential to note that there are fundamental limits on the performance of the introduced estimator. In particular, the above optimization problem may not have a unique solution for any length of the sample trajectory. To see this, suppose that $u(t) = K_0 x(t)$ and K_0 is equal to the identity matrix. Then, the above optimization problem reduces to

$$(\hat{A},\hat{B}) = \arg\min_{A,B} \frac{1}{2T}\sum_{t=0}^{T-1} \|x(t+1) - (A+B)\,x(t)\|_2^2 + \lambda(\|A\|_1 + \|B\|_1).$$

It is easy to see that, given any optimal solution $(\hat{A}, \hat{B})$ to the above optimization, $(\tilde{A}, \tilde{B}) = (\alpha\hat{A}, (1-\alpha)\hat{B})$ is also optimal for any $0 \leq \alpha \leq 1$. To break this symmetry and to guarantee the identifiability of the parameters, it is essential to inject an *input noise* to the system at every time t. In particular, we assume that $u(t) = K_0 x(t) + v(t)$, where $v(t)$ is a random vector with a user-defined distribution. As another example, if A is stable and $K_0 = 0$, the need to introduce noise in the input is inevitable in order to identify the matrix B.

To further analyze the properties of the above estimator, one can write (1) in a compact form. Let $\Psi^* = \begin{bmatrix} A & B \end{bmatrix}^\top$ denote the true parameters of the system. Furthermore, define

$$Y = \begin{bmatrix} x(1)^\top \\ \vdots \\ x(T)^\top \end{bmatrix}, X = \begin{bmatrix} x(0)^\top & u(0)^\top \\ \vdots & \vdots \\ x(T-1)^\top & u(T-1)^\top \end{bmatrix}, W = \begin{bmatrix} w(0)^\top \\ \vdots \\ w(T-1)^\top \end{bmatrix}. \tag{3}$$

The system identification problem is then reduced to estimating the unknown parameter Ψ^* given the *design matrix* X, and the *observation matrix* Y that is corrupted with the *noise matrix* W. We can therefore rewrite optimization problem (2) compactly as

$$\hat{\Psi} = \arg\min_{\Psi} \frac{1}{2T}\|Y - X\Psi\|_F^2 + \lambda\|\Psi\|_1 \tag{4}$$

which corresponds to the so-called *Lasso* estimator, initially popularized in statistics and machine learning to estimate the support parameter values of a sparse linear model [13]. The non-asymptotic properties of this estimator have been widely studied in the literature [14, 15, 16], all highlighting its sub-linear sample complexity under suitable technical conditions. In particular, they show that under the so-called *mutual incoherency* of the design matrix and the sparsity of the unknown parameters, the minimum number of observations for the accurate estimation of the Lasso scales logarithmically in the dimension of Ψ. Motivated by these results, one may speculate that the proposed estimator (2) benefits from a similar logarithmic sample complexity. However, the validity

of the derived non-asymptotic estimation error bounds on the Lasso is contingent upon a number of assumptions on the independence between the design matrix X and the noise matrix W [14, 17]; such assumptions do not necessarily hold in the sparse system identification problem, partly due to the dependency between the states, the inputs and the disturbance noise. The problematic nature of this dependency becomes more evident by noting that the Lasso may not be consistent when the design and noise matrices are dependent [18].

This lack of independence in the design and noise matrices of the sparse system identification problem has been the main roadblock in deriving similar sub-linear sample complexity bounds for the sparse system identification problem and it leaves the following question unanswered:

Is the estimator (2) consistent, and if so, what is its sample complexity?

3 Main Results

Despite the fact that in general, the Lasso may not be a consistent estimator when the design and noise matrices are dependent, we exploit the underlying structure of the system identification problem to control this dependency and provide an affirmative answer to the posed question. In other words, we show that not only is the proposed estimator (2) consistent, but that it also enjoys a logarithmic sample complexity in the state and input dimensions, under appropriate conditions. To this goal, we first provide a number of definitions.

Definition 1. *A zero-mean (centered) random variable x is **sub-Gaussian** with parameter b if its moment generating function satisfies*

$$\mathbb{E}\{\exp(tx)\} \leq \exp\left(\frac{b^2t^2}{2}\right)$$

for every t.

For a centered sub-Gaussian random variable x with parameter b, one can easily verify that $\mathbb{P}(|x| > t) \leq 2\exp\left(\frac{t^2}{2b^2}\right)$. The most commonly known examples of such random variables are Gaussian, Bernoulli, and any bounded random variable.

Definition 2. *Given a sub-Gaussian random variable x, its **sub-Gaussian norm**, denoted by $\|x\|_\psi$ is defined as the smallest $r > 0$ such that the inequality $\mathbb{E}\{x^2/r^2\} \leq 2$ is satisfied.*

It is well-known that the above two definitions are closely related. In particular, it can be verified that $\frac{1}{\sqrt{5}}b \leq \|x\|_\psi \leq \sqrt{\frac{8}{3}}b$ for a sub-Gaussian random variable with parameter b.[1] For a random vector x with sub-Gaussian elements, $\|x\|_\psi$ is defined as $\max_i\{\|x_i\|_\psi\}$.

As mentioned before, we assume that the dynamical system is equipped with an initial static and stabilizing state-feedback controller K_0. More specifically, we assume that at any given time t, the input $u(t)$ is equal to $K_0x(t) + v(t)$, where $v(t)$ is a user-defined input noise with independent and centered sub-Gaussian elements whose non-zero variance is upper bounded by σ_v^2 (for stable systems, K_0 can be set to zero). Similarly, we assume that the disturbance noise at

[1] This is a standard result; see [19] and [20] for a simple proof.

every time t is a random vector with independent and centered sub-Gaussian elements whose variance is upper bounded by σ_u^2. Further, let $\eta > 0$ be the smallest positive constant such that $\max\{\|w(t)\|_\psi, \|v(t)\|_\psi\} \leq \eta$; such a constant is guaranteed to exist as w and v are assumed to be centered sub-Gaussian random variables.

Remark 1. *Most of the existing results on the sample complexity of the system identification problem assume a centered Gaussian distribution for the input noise [7, 21, 9]. Despite having desirable finite-time properties, these types of Gaussian inputs may jeopardize the safety of the dynamical system due to their unbounded range. Accordingly, in many control systems, the input is constrained to have a limited power. These types of constraints can be translated into ℓ_∞ or ℓ_2 bounds on the input signal. Due to the fact that such bounded random signals are sub-Gaussian, our results are readily applied to system identification problems with input constraints.*

Notice that for LTI systems, the uniform asymptotic stability of the closed-loop system is equivalent to its exponential stability. In other words, an LTI system is uniformly asymptotically stable if and only if there exist constants $C \geq 1$ and $0 < \rho < 1$ such that $\|\!|(A + BK_0)^\tau|\!\| \leq C\rho^\tau$ for every time τ. Without loss of generality, let $C \geq 1$ and $0 \leq \rho < 1$ be the smallest constants such that $\|\!|(A + BK_0)^\tau B|\!\| \leq C\rho^\tau$, $\|\!|K_0(A + BK_0)^\tau|\!\| \leq C\rho^\tau$ and $\|\!|K_0(A + BK_0)^\tau B|\!\| \leq C\rho^\tau$ for every time τ. Note that the existence of such $C \geq 1$ and $0 < \rho < 1$ is guaranteed due to the exponential stability of the closed-loop system.

Furthermore, we assume that the initial state $x(0)$ rests at its stationary distribution or, equivalently, the following equality holds:

$$x(0) = \lim_{\tilde{T}\to\infty} \sum_{\tau=-\tilde{T}}^{-1} (A + BK_0)^{-\tau-1}(w(\tau) + Bv(\tau))$$

Note that, for exponentially stable systems, the state converges to its stationary distribution exponentially fast and therefore, the stationarity of $x(0)$ is a reasonable assumption. Furthermore, using the above equality, it is easy to see that $x(0)$ is a random vector whose elements are (dependent) centered sub-Gaussian random variables with bounded parameters. Moreover, one can verify that its covariance $\mathbb{E}\{x(0)x(0)^\top\} = Q^*$ satisfies the following Lyapunov equation:

$$(A + BK_0)Q^*(A + BK_0)^\top - Q^* + \sigma_w^2 I + \sigma_v^2 BB^\top = 0 \tag{5}$$

Accordingly, Q^* can be used to derive the covariance matrix M^* for the random vector $\begin{bmatrix} x(0)^\top & (K_0x(0) + v(0))^\top \end{bmatrix}^\top$

$$M^* = \begin{bmatrix} Q^* & Q^*K_0^T \\ K_0Q^* & K_0Q^*K_0^T + \sigma_v^2 I \end{bmatrix}$$

Define $\mathcal{A}_j = \{i : \Psi^*_{ij} \neq 0\}$ and let $\mathcal{A}_j^c$ refer to its complement. Denote k as the maximum number of nonzero elements in any column of Ψ^*.

Assumption 1. *The following inequalities are satisfied*

A1 (Mutual incoherence)

$$\max_{1\leq j\leq n} \left\{ \max_{i\in\mathcal{A}_j^c} \left\{ \left\| M^*_{i\mathcal{A}_j} (M^*_{\mathcal{A}_j\mathcal{A}_j})^{-1} \right\|_1 \right\} \right\} \leq 1 - \gamma$$

A2 (Bounded eigenvalue)

$$\min_{1\le j\le n} \lambda_{\min}(M^*_{\mathcal{A}_j\mathcal{A}_j}) \ge C_{\min}$$

A3 (Bounded infinity norm)

$$\max_{1\le j\le n} \left|\!\left|\!\left| (M^*_{\mathcal{A}_j\mathcal{A}_j})^{-1} \right|\!\right|\!\right|_\infty \le D_{\max}$$

A4 (Nonzero gap)

$$\min_{1\le j\le n} \left\{ \max_{i\in\mathcal{A}_j} \{|\Psi^*_{ij}|\} \right\} \ge \Psi_{\min}$$

for some constants $0 < \gamma < 1$, $1 \ge C_{\min} > 0$, $D_{\max} \ge 1$ *and* $1 \ge \Psi_{\min} > 0$.

Next, we present the main result of the paper.

Theorem 1. *Assume that* $k \ge 2$ *and*

$$\lambda = c_1 \cdot \frac{C}{1-\rho} \cdot \frac{\eta^2}{\gamma} \sqrt{\frac{\log((n+m)/\delta)}{T}} \tag{6}$$

$$T \ge c_2 \cdot \frac{C^4}{(1-\rho)^4} \cdot \frac{D^2_{\max}}{\gamma^2 C^2_{\min} \Psi^2_{\min}} \cdot k^2 \log((n+m)/\delta), \tag{7}$$

where c_1 *and* c_2 *are universal constants. Then, the following statements hold with probability of at least* $1-\delta$*:*

1. *(Correct sparsity recovery) (4) has a unique solution and recovers the true sparsity pattern of* Ψ^*.
2. *(ℓ_∞-norm error) We have*

$$\|\hat{\Psi} - \Psi^*\|_\infty \le c_3 \cdot \frac{C}{1-\rho} \cdot \frac{D_{\max}\eta^2}{\gamma} \sqrt{\frac{\log((n+m)/\delta)}{T}} \tag{8}$$

where c_3 *is a universal constant.*

Remark 2. *As mentioned before, the injection of a random input noise is essential to guarantee the identifiability of the parameters. This is also reflected in the above theorem: in order to guarantee a finite sample complexity for the proposed estimator, it is crucial to have* $C_{\min} > 0$*, which is only possible if* $\sigma_v > 0$.

A number of observations can be made based on Theorem 1. First, it implies that if γ, C, $D_{\max}$, $C_{\min}$, $\Psi_{\min}$, and ρ do not scale with the system dimension, then $T = \Omega(k^2 \log(n+m))$ is enough to guarantee the correct sparsity recovery and a small estimation error. Notice that for sparse systems, this quantity can be much smaller than the system dimension. Second, the sample complexity of the proposed estimator depends on $\frac{C}{1-\rho}$, which is a measure of the system

stability. In particular, for highly stable systems, $\frac{C}{1-\rho}$ is small, resulting in an improved accuracy of the proposed estimator with smaller T. In contrast, when the system is close to its stability margin, $\frac{C}{1-\rho}$ will grow which negatively affects the estimation error as well as the lower bound on T. Another intuitive interpretation of $\frac{C}{1-\rho}$ is that it measures the amount of *dependency* between the states at different times: for highly stable systems where ρ is small, $(x(t), u(t))$ is only weakly dependent on $(x(\tau), u(\tau))$ for $\tau = 0, \dots, t-1$, thereby facilitating the estimation of the unknown parameters. We finally mention that this dependency is in contrast with the recent discoveries on the sample complexity of the least-squares estimator, which support the favorable effect of a large ρ on the accuracy of the estimator [22]. We leave investigating whether this seemingly contradictory observation is an artifact of our methodology (e.g., mixing the initial state to the stationary distribution), or is fundamental to the sparse system identification problem, to future work.

Remark 3. *In order to further enhance the accuracy of the proposed estimator, one can perform a least-squares estimation restricted to the nonzero elements of the estimated parameter, after obtaining its sparsity pattern via the proposed method. Although, theoretically, this post-model-selection estimation method may not improve the estimation error rate, it will incur less bias [23]. We will show in our simulations that the effect of this post-processing step can be significant in the accuracy of the estimation.*

3.1 Comparison to prior art

As mentioned before, another line of work focuses on unstructured system identification, where either the learning time T or the number of sample trajectories d is allowed to grow. In [9], the authors consider the sample complexity of the system identification problem with multiple sample trajectories via least-squares, where it is shown that the proposed estimator incurs a small error, provided that $d = \Omega(n+m)$. Revisiting (20) reveals that the proposed method outperforms the sample complexity of ordinary least-squares when k is significantly smaller than $n+m$, i.e., exploiting prior knowledge of the system sparsity leads to a reduction in sample complexity. In [6, 22, 11, 12], the authors consider unstructured system identification from a single sample trajectory under different assumptions on system stability and/or the initial state of the system. However, similar to [9], none of these works take advantage of the underlying sparsity structures of the system matrices. As a result, they cannot correctly estimate the sparsity structure of (A, B) and suffer from poor dependencies on the system dimensions in the large-scale and structure setting.

Subsequently, a Lasso-type estimator is proposed in [21] to further exploit the underlying sparsity pattern of (A, B) with d sample trajectories, each with a zero initial state. In particular, it is shown that $d = \Omega\left(\frac{\kappa(\Sigma)^2}{\gamma^2 \Psi_{\min}^2} k \log(n+m)\right)$ is enough to ensure the correct sparsity recovery and a small estimation error with high probability, where $\kappa(\Sigma)$ is the condition number of the finite-time *controllability matrix* of the system. Comparing this quantity with (20), one can observe that the former has a better dependency on k. However, $\kappa(\Sigma)$ is highly dependent on the learning time T. In fact, it is easy to show that for unstable systems, $\kappa(\Sigma)$ may grow exponentially fast with respect to T. On the other hand, (20) is free of such dependency and instead, it is in terms of the stationary distributions of the state and input vectors.

Moreover, our work is a major extension to the results of [7], where the authors address a similar sparse system identification problem with a single sample trajectory. First, unlike the presented

results, [7] only considers autonomous systems, i.e., systems (1) with B=0. Second, [7] only ensures the correct sparsity recovery of the true parameters. In contrast, we extend these results to obtain non-asymptotic bounds on the estimation error. As demonstrated in [9, 8], having these bounds is essential for the design of near-optimal and robustly stabilizing controllers. Third, [7] requires that the closed-loop system be contractive with respect to the spectral norm, i.e., that $|||(A + BK_0)||| < 1$, whereas we only require system stability. Notice that the former condition is much stronger, as in practice, stable systems are often not contractive in spectral norm. Finally, the validity of the non-asymptotic bounds introduced in [7] heavily relies on the Gaussian nature of the disturbance and input noises. As an extension to this result, our proposed method targets a larger class of uncertainties for the disturbance and input noises, thereby allowing for norm bounded disturbance and input signals.

3.2 Mutual incoherency

In this subsection, we analyze the mutual incoherence condition on the steady-state covariance matrix M^*. In particular, we explain why this assumption is not an artifact of the proposed method, but that it rather stems from a fundamental limitation of *any* sparsity-promoting technique for the system identification problem. We show that similar mutual incoherence assumptions are indeed necessary to recover the correct sparsity of system parameters by using a class of *oracle estimators*.

We assume that the oracle estimator can measure the disturbance matrix W and that it can work with sample trajectories of an arbitrary length. With these assumptions, the oracle estimator solves the following optimization problem to estimate the parameters of the system:

$$\min_{\Psi} \ \|\Psi\|_0 \tag{9a}$$

$$\text{s.t.} \ \ X\Psi = Y - W \tag{9b}$$

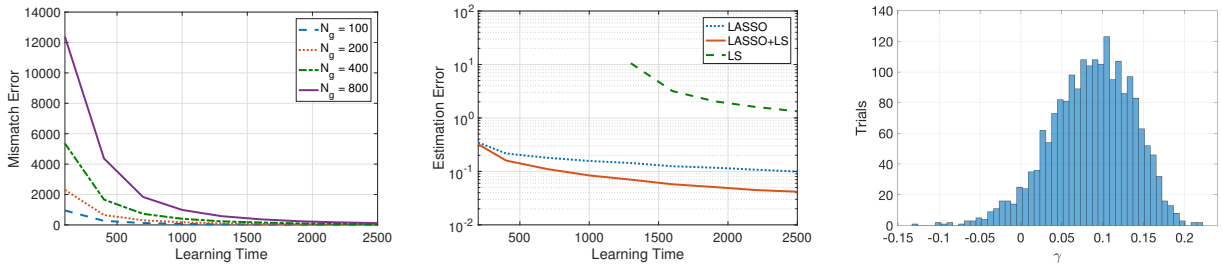
Clearly, this oracle estimator cannot be used in practice since 1) the disturbance matrix W is unknown, 2) the learning time T is finite, and 3) the corresponding optimization problem is non-convex and NP-hard in its worst case. Setting aside these restrictions for now, there are fundamental limits on the consistency of this estimator. To explain this, we introduce the mutual-coherence metric for a matrix (note the difference between this definition and Assumption A1). For a given matrix $A \in \mathbb{R}^{t_1 \times t_2}$, its mutual-coherence $\mu(A)$ is defined as

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

In other words, $\mu(A)$ measures the maximum correlation between distinct columns of A. Reminiscent of the classical results in the compressive sensing literature, it is well-known that the optimal solution Ψ^* of (9) is unique if the following *identifiability* condition

$$\|\Psi^*_{:,j}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right) \tag{10}$$

holds for $j = 1, 2, ..., n$ (see, e.g., Theorem 2.5 in [24]). Furthermore, this bound is tight, implying that there exists an instance of the problem for which the violation of $\|\Psi^*_{:,j}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right)$



(a) Relative mismatch error (b) Normalized estimation error (c) The distribution of γ

Figure 1: (a) The mismatch error with respect to the learning time for different number of generators in the system. The values are averaged over 10 independent trials. (b) The normalized estimation error for Lasso (abbreviated as `LASSO`), Lasso + least-squares (abbreviated as `LASSO+LS`), and least-squares (abbreviated as `LS`) estimators with respect to the learning time. The values are averaged over 10 independent trials. (c) The distribution of mutual incoherence parameter γ for 2000 randomly generated instances of the problem.

for some j results in the non-uniqueness of the optimal solution. On the other hand, according to Lemma 3 (to be introduced later) and the Borel-Cantelli lemma, $\frac{1}{T}X^\top X$ converges to M^* almost surely, as $T \to \infty$. This implies that

$$\mu(X) = \max_{1\le i<j\le m+n} \frac{|X_{:,i}^\top X_{:,j}|}{\|X_{:,i}\|_2 \|X_{:,j}\|_2} \overset{a.s.}{\to} \max_{1\le i<j\le m+n} \frac{|M^*_{ij}|}{\sqrt{M^*_{ii}M^*_{jj}}}$$

The above analysis reveals that the off-diagonal entries of M^* play a crucial role in the identifiability of the true parameters: as these elements become smaller relative to the diagonal entries, the oracle estimator can correctly identify the structure of Ψ for a wider range of sparsity levels. Similarly, our proposed mutual incoherence assumption is expected to be satisfied when the off-diagonals of M^* have small magnitudes, relative to the diagonal entries. This implies that Assumption A1 is a natural condition to impose in order to ensure the correct sparsity recovery of Ψ. Furthermore, in practice, M^* will be close to a diagonally dominant matrix with exponentially decaying off-diagonal entries, provided that the matrices A, B, and K_0 have sparse structures [25].

4 Numerical Experiments

As a 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. Assume that there exist N_g generators in the system. It is common to describe the per-unit swing equations using the well-known direct current (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 is defined as follows:

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

where n is the number of generators in the network, $\mathcal{N}_i$ 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

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

The goal is to identify the underlying dynamical system based on a single sample trajectory consisting of a sequence of mechanical power inputs and their effects on the angles and frequencies of different generators. To assess the performance of the proposed method, we generate several instances of the problem according to the following rules:

- the generators are connected via a randomly generated tree with a maximum degree of 10.
- the parameters B_{ij}, M_i, D_i are uniformly chosen from $[0.5, 1]$, $[1, 2]$, $[0.5, 1.5]$, respectively.

Furthermore, the sampling time dt is set to 0.1. We assume that the disturbance noise has a zero-mean Gaussian distribution with covariance $0.01 I_{2\times 2}$. Notice that the magnitude of the noise is comparable to those of the nonzero elements in A and B. Furthermore, the mechanical input is set to $u_i(t) = -0.1(\theta_i + \dot{\theta}_i) + v_i(t)$, where $v_i(t)$ is a randomly generated input noise, distributed according to a zero-mean Gaussian distribution with variance 0.05. Notice that the first term in the input signal is used to ensure the closed-loop stability.

The reported results are for a serial implementation in MATLAB R2017b, and the function `lasso` is used to solve (2). It is worthwhile to note that the running time can be further reduced via parallelization; this is trivially possible due to the decomposable nature of the problem. The *mismatch error* is defined as the total number of false positives and false negatives in the sparsity pattern of the estimated parameters $(\hat{A}, \hat{B})$. Furthermore, *relative learning time* (RLT) is defined as the learning time normalized by the dimension of the system, and *relative mismatch error* (RME) is used to denote the mismatch error normalized by the total number of elements in A and B. In all of our experiments, the regularization coefficient λ is set to $\lambda = \sqrt{\frac{0.03 \log(n+m)}{T}}$. Note that this value does not require any additional fine-tuning and is at most a constant factor away from (6).

Figure 1a illustrates the mismatch error (averaged over 10 different trials) with respect to the learning time T and for different number of generators N_g that are chosen from $\{100, 200, 400, 800\}$. These correspond to the total system dimensions of $\{300, 600, 1200, 2400\}$. Note that the largest instance has more than 3.84 million unknown parameters. Not surprisingly, the learning time needed to achieve a small mismatch error increases as the dimension of the system grows. Conversely, a smaller value for RLT is needed to achieve infinitesimal RME for larger systems. In particular, when N_g is equal to 100, 200, 400, and 800, the minimum RLT to guarantee RME $\leq 0.1\%$ is equal to 3.83, 1.42, 0.50, and 0.16, respectively.

As mentioned before, the accuracy of the proposed estimator can be improved by additionally applying the least-squares over the nonzero elements of $(\hat{A}, \hat{B})$. Figure 1b illustrates the normalized 2-norm estimation error of this approach (abbreviated as `LASSO+LS`), compared to the proposed method without any post-processing step (abbreviated as `LASSO`), and the least-squares estimator (abbreviated as `LS`) when N_g is set to 200. It can be observed that both `LASSO+LS` and `LS` significantly outperform `LS`; in fact, `LS` is not even well-defined if the learning time is strictly less than the system dimensions. Furthermore, on average, the estimation error for `LASSO+LS` is 1.91 times smaller than that of `LASSO`.

Finally, only 32 out of 360 generated instances did not satisfy the proposed mutual incoherence condition. However, this violation did not have a significant effect on the accuracy of the proposed estimator. To further investigate the frequency of the instances that satisfy this condition, we plot the histogram of the mutual incoherence parameter γ for 2000 randomly generated instances with fixed $N_g = 200$. It can be seen in Figure 1c that the mutual incoherence condition is violated only for 5.15% of the instances.

5 Conclusions

The problem of sparse system identification of linear time-invariant (LTI) systems is considered in this work, where the goal is to estimate the sparse structure of the system matrices based on a single sample trajectory of the dynamics. A Lasso-type estimator is introduced to identify the parameters of the system, while promoting their sparsity via a ℓ_1-regularization technique. By carefully examining the underlying properties of the system—such as its stability and mutual incoherency—we provide non-asymptotic bounds on the accuracy of the proposed estimator. In particular, we show that it correctly identifies the sparsity structure of the system matrices and enjoys a sharp upper bound on its estimation error, provided that the learning time exceeds a threshold. We further show that this threshold scales polynomially in the number of nonzero elements but logarithmically in the system dimensions.

References

[1] M. Rotkowitz and S. Lall, "A characterization of convex problems in decentralized control," *IEEE Transactions on Automatic Control*, vol. 50, no. 12, pp. 1984–1996, 2005.

[2] Y.-S. Wang, N. Matni, and J. C. Doyle, "A system level approach to controller synthesis," *arXiv preprint arXiv:1610.04815*, 2016.

[3] Y.-S. Wang, N. Matni, and J. C. Doyle, "Separable and localized system-level synthesis for large-scale systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4234–4249, 2018.

[4] M. Kheirandishfard, F. Zohrizadch, M. Adil, and R. Madani, "Convex relaxation of bilinear matrix inequalities part ii: Applications to optimal control synthesis," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 75–82.

[5] S. Fattahi, G. Fazelnia, J. Lavaei, and M. Arcak, "Transformation of optimal centralized controllers into near-globally optimal static distributed controllers," *IEEE Transactions on Automatic Control*, vol. 64, no. 1, pp. 63–77, 2019.

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

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

[8] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," in *Advances in Neural Information Processing Systems*, 2018, pp. 4192–4201.

[9] 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.

[10] S. Dean, S. Tu, N. Matni, and B. Recht, "Safely learning to control the constrained linear quadratic regulator," *arXiv preprint arXiv:1809.10121*, 2018.

[11] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proceedings of the 24th Annual Conference on Learning Theory*, 2011, pp. 1–26.

[12] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[14] 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.

[15] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

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

[17] S. N. Negahban, P. Ravikumar, M. J. Wainwright, B. Yu *et al.*, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[18] J. Fan and Y. Liao, "Endogeneity in high dimensions," *Annals of statistics*, vol. 42, no. 3, p. 872, 2014.

[19] O. Rivasplata, "Subgaussian random variables: An expository note," *Internet publication, PDF*, 2012.

[20] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[21] S. Fattahi and S. Sojoudi, "Sample complexity of sparse system identification problem," *arXiv preprint arXiv:1803.07753v2*, 2018.

[22] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," *arXiv preprint arXiv:1802.08334*, 2018.

[23] A. Belloni and V. Chernozhukov, "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, vol. 19, no. 2, pp. 521–547, 2013.

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

[25] V. Simoncini, "The lyapunov matrix equation. matrix analysis from a computational perspective," *arXiv preprint arXiv:1501.07564*, 2015.

[26] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-gaussian concentration," *Electronic Communications in Probability*, vol. 18, 2013.

A Proof of Theorem 1

In this section, we present the sketch of the proof for the main theorem. Define

$$L(\Psi_{:,j}) = \|Y - X\Psi_{:,j}\|_2^2$$

and

$$\hat{\Psi}_{:,j} = \arg\min \frac{1}{2T} L(\Psi_{:,j}) + \lambda \|\Psi_{:,j}\|_1 \tag{11}$$

for every $j \in \{1, 2, ..., n\}$. It is easy to verify that

$$\hat{\Psi} = \begin{bmatrix} \hat{\Psi}_{:,1} & \hat{\Psi}_{:,2} & \cdots & \hat{\Psi}_{:,n} \end{bmatrix}$$

Furthermore, the Gradient and Hessian of $L(\cdot)$ are equal to

$$G = -\nabla L(\Psi_{:,j})|_{\Psi_{:,j}=\Psi^*_{:,j}} = \frac{1}{T} X^T W_{:,j},$$
$$M = \nabla^2 L(\Psi_{:,j})|_{\Psi_{:,j}=\Psi^*_{:,j}} = \frac{1}{T} X^T X$$

Note that G can be different for every j. However, we keep this dependency implicit in the notations to streamline the presentation. The following Lemma is at the core of our subsequent analysis:

Lemma 1 (Proposition 4.1 [7]). *Suppose that the following conditions are satisfied:*

$$\begin{aligned}
&\|G\|_\infty \leq \frac{\lambda\gamma}{3},\\
&\|G_{\mathcal{A}_j}\|_\infty \leq \frac{\Psi_{\min} C_{\min}}{4k} - \lambda\\
&\left|\left|\left| M_{\mathcal{A}_j^c \mathcal{A}_j} - M^*_{\mathcal{A}_j^c \mathcal{A}_j} \right|\right|\right|_\infty \leq \frac{\gamma C_{\min}}{12\sqrt{k}},\\
&\left|\left|\left| M_{\mathcal{A}_j \mathcal{A}_j} - M^*_{\mathcal{A}_j \mathcal{A}_j} \right|\right|\right|_\infty \leq \frac{\gamma C_{\min}}{12\sqrt{k}}
\end{aligned}$$

Then, (11) *recovers the true sparsity pattern of* $\Psi^*_{:,j}$.

The first step in proving Theorem 1 is to verify that the conditions of Lemma 1 hold with high probability. To this goal, first we write $x(t)$ and $u(t)$ in terms of $x(0)$, $w(\tau)$ and $v(\tau)$ for $\tau = 0, 1, \ldots, t$:

$$\begin{aligned}
x(t) =& (A + BK_0)^t x(0) + \sum_{\tau=0}^{t-1} (A + BK_0)^{t-\tau-1}(w(\tau) + Bv(\tau))\\
u(t) =& v(t) + K_0(A + BK_0)^t x(0) + \sum_{\tau=0}^{t-1} K_0(A + BK_0)^{t-\tau-1}(w(\tau) + Bv(\tau))
\end{aligned}$$

Instead of initiating the system at $x(0)$ with the stationary distribution, we will start at the time $-T_0$, with a modified initial state $x(-T_0) = w(-T_0 - 1) + Bv(-T_0 - 1)$, where $w(-T_0 - 1)$ and $v(-T_0 - 1)$ have the same distributions as the disturbance and input noises, respectively. Since the system is stable, by taking $T_0 \to \infty$ and invoking the Continuous Mapping Theorem, the matrices

$$\begin{bmatrix} x(0) & x(1) & \ldots & x(T-1) \end{bmatrix}$$

and

$$\begin{bmatrix} K_0x(0)+v(0) & K_0x(1)+v(1) & \ldots & K_0x(T-1)+v(T-1) \end{bmatrix}$$

converge in distribution to the same matrices when the system is initialized at a state with the stationary distribution. Therefore, without loss of generality, we will focus on the former. Based on this observation, one can write

$$\begin{aligned}
x(t) &= \lim_{T_0 \to \infty} \sum_{\tau=-T_0-1}^{t-1} (A + BK_0)^{t-\tau-1}(w(\tau) + Bv(\tau))\\
u(t) &= v(t) + \lim_{T_0 \to \infty} \sum_{\tau=-T_0-1}^{t-1} K_0(A + BK_0)^{t-\tau-1}(w(\tau) + Bv(\tau))
\end{aligned}$$

This implies that the elements in G and M can be written as quadratic functions of the disturbance and input noises in the form of $G_i = z^\top R_G z$ and $M_{ij} = z^\top R_M z$, where $z \in \mathbb{R}^{(n+m)(t+T_0+1)}$ is a random vector, defined as

$$z = \begin{bmatrix} w(-T_0-1)^\top & \cdots & w(t-1)^\top & v(-T_0-1)^\top & \cdots & v(t-1)^\top \end{bmatrix}^\top$$

The following theorem will be used in our analysis to provide concentration bounds on G and M.

Theorem 2 (Hanson-Wright inequality [26]). *Let* $x = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$ *be a random vector with independent zero-mean sub-Gaussian elements. Given a square and symmetric matrix* P*, the following inequality holds*

$$\mathbb{P}\left(\left|x^\top Px - \mathbb{E}\left\{x^\top Px\right\}\right| > t\right) \leq 2\exp\left(-c \cdot \min\left\{\frac{t^2}{\|x\|_\psi^4 \|P\|_F^2}, \frac{t}{\|x\|_\psi^2 |||P|||}\right\}\right)$$

for every $t \geq 0$*, where* c *is a universal constant.*

For a symmetric matrix P, we have $\|P\|_F^2 = \sum_{k=1}^n \lambda_k^2$. Therefore, the above theorem implies that, for a sub-Gaussian random vector z with independent elements, we have

$$\mathbb{P}\left(\left|z^\top Pz - \mathbb{E}\left\{z^\top Pz\right\}\right| > t\right) \leq 2\exp\left(-c \cdot \frac{t^2}{\|z\|_\psi^4 \left(\sum_{k=1}^n \lambda_k^2\right)}\right)$$

provided that $t \leq \left(\frac{\sum_k \lambda_k^2}{\max_k |\lambda_k|}\right) \|z\|_\psi^2$. The assumptions of Lemma 1 can be seen to hold directly as a consequence of the following two lemmas:

Lemma 2. *Let* $i \in \{1, 2, ..., n+m\}$ *and suppose that* $\epsilon < \frac{3C\eta^2}{1-\rho}$*. Then, there exists a universal constant* c_4 *such that*

$$\mathbb{P}\{|G_i| > \epsilon\} \leq 2\exp\left(-c_4 \frac{(1-\rho)^2}{C^2\eta^4} T\epsilon^2\right)$$

Proof. See Appendix B.1. □

Lemma 3. *Let* $i, j \in \{1, 2, ..., n+m\}$ *and suppose that* $\epsilon \leq \frac{4C^2\eta^2}{(1-\rho)^2}$*. Then, there exists a universal constant* c_5 *such that*

$$\mathbb{P}\{|M_{ij} - M_{ij}^*| > \epsilon\} \leq 2\exp\left(-c_5 \frac{(1-\rho)^4}{C^4\eta^4} T\epsilon^2\right)$$

Proof. See Appendix B.2. □

The following proposition shows that for a fixed column j, the proposed estimator (11) correctly recovers the sparsity pattern with high probability.

Proposition 1. *Assume that* $k \geq 2$ *and the following conditions are satisfied:*

$$\lambda = c_6 \cdot \sqrt{\frac{C^2\eta^4}{\gamma^2 T(1-\rho)} \log(n+m/\delta)} \tag{12}$$

$$T \geq c_7 \cdot \frac{C^4\eta^4 k^2}{\gamma^2 C_{\min}^2 \Psi_{\min}^2 (1-\rho)^4} \log(n+m/\delta) \tag{13}$$

for universal constants $c_6, c_7 \geq 0$*. Then,* (11) *recovers the true sparsity pattern of* $\Psi_{:,j}^*$ *with probability of at least* $1-\delta$*.*

Proof. The Lemmas 2 and 3 can be used to prove statement. The details are provided in Appendix B.3. □

The next lemma provides a deterministic upper bound on the estimation error in terms of the deviations of M and G from their mean.

Lemma 4. *Assume that*

$$\left|\left|\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\right|\right|_\infty \leq \frac{\min\{1, 2\eta^2\}}{2D_{\max}} \tag{14}$$

and (11) *recovers the correct sparsity pattern of* $\Psi^*_{:,j}$*. Then, the following inequality holds for* $E = \hat{\Psi}_{:,j} - \Psi^*_{:,j}$*:*

$$\begin{aligned} &E_{\mathcal{A}_j^c} = 0 \\ &\|E_{\mathcal{A}_j}\|_\infty \leq \left(2D^2_{\max}\left|\left|\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\right|\right|_\infty + D_{\max}\right)\left(\|G_{\mathcal{A}_j}\|_\infty + \lambda\right) \end{aligned} \tag{15}$$

Proof. See Appendix B.4. □

The next lemma shows that the condition of Proposition 4 holds with high probability, provided that T is large enough.

Proposition 2. *Assume that*

$$T \geq c_8 \cdot \frac{D^2_{\max} C^4}{(1-\rho)^4} k^2 \log(k/\delta) \tag{16}$$

for some universal constant $c_5 \geq 0$*. Then, the following inequality holds with probability of at least* $1 - \delta$

$$\left|\left|\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\right|\right|_\infty \leq \frac{\min\{1, 2\eta^2\}}{2D_{\max}} \tag{17}$$

Proof. Notice that $|\mathcal{A}_j| \leq k$. One can verify that

$$\mathbb{P}\left(\left|\left|\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\right|\right|_\infty > \epsilon\right) \leq 2k^2 \exp\left(-c_5 \cdot \frac{(1-\rho)^4}{C^4\eta^4}\frac{T}{k^2}\epsilon^2\right) \tag{18}$$

provided that $\frac{\epsilon}{k} \leq \frac{4C^2\eta^2}{(1-\rho)^2}$. Setting $\epsilon = \frac{\min\{1,2\eta^2\}}{2D_{\max}}$ and recalling that $D_{\max}, C \geq 1$, one can verify that $\frac{\epsilon}{k} \leq \frac{4C^2\eta^2}{(1-\rho)^2}$ is satisfied. Furthermore, by choosing $c_8 = \frac{16}{c_5}$, one can certify that (16) is enough to ensure that the right hand side of the above inequality is upper bounded by δ, thereby completing the proof. □

Proof of Theorem 1: First note that (4) can be decomposed into n disjoint sub-problems over different columns of Ψ, each in the form of (11). Consider the following choices for λ and T:

$$\lambda = c_6 \cdot \sqrt{\frac{C^2\eta^4}{\gamma^2 T(1-\rho)^2}\log(4(n+m)/\delta)} \tag{19}$$

$$T \geq \max\left\{c_7, c_8, \frac{1}{c_4}, \frac{2}{c_5}\right\} \cdot \frac{C^4 D^2_{\max} k^2}{\gamma^2 C^2_{\min}\Psi^2_{\min}(1-\rho)^4}\log((n+m)/\delta) \tag{20}$$

where c_4, c_5, c_6, c_7, and c_6 are introduced in Lemmas 2, 3, and Propositions 1, 2. Based on the Proposition 1 and the above choices for λ and T, (11) recovers the sparsity pattern of $\Psi^*_{:,j}$ for a

given column index j with probability of at least $1-\delta$. Furthermore, based on Proposition 2, the lower bound on T guarantees that the inequality

$$\left|\!\left|\!\left| Q_{\mathcal{A}_j,\mathcal{A}_j} - Q^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty \leq \frac{\min\{1, 2\eta^2\}}{2D_{\max}} \tag{21}$$

holds with probability of at least $1-\delta$. This, together with Proposition 4 results in

$$\|E_{:,j}\|_\infty \leq \left(2D^2_{\max}\left|\!\left|\!\left| Q_{\mathcal{A}_j,\mathcal{A}_j} - Q^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty + D_{\max}\right)\left(\|G_{\mathcal{A}_j}\|_\infty + \lambda\right) \tag{22}$$

with probability of at least $1-2\delta$. Now, it suffices to obtain concentration bounds for different terms of the above inequality. Based on (18) and Lemma 2, one can write

$$\mathbb{P}\left(\|G_{\mathcal{A}_j}\|_\infty > \epsilon_1\right) \leq \exp\left(\log(2k) - c_4 \cdot \frac{(1-\rho)^2}{C^2\eta^4} T\epsilon_1^2\right) \tag{23}$$

$$\mathbb{P}\left(\left|\!\left|\!\left| Q_{\mathcal{A}_j,\mathcal{A}_j} - Q^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty > \epsilon_2\right) \leq \exp\left(2\log(2k) - c_5 \cdot \frac{(1-\rho)^4}{C^4\eta^4}\frac{T}{k^2}\epsilon_2^2\right) \tag{24}$$

This implies that, with the following choices

$$\epsilon_1(\zeta_1) = \sqrt{\zeta_1 \cdot \frac{C^2\eta^4}{c_4 T(1-\rho)^2}\log(2k)} \tag{25}$$

$$\epsilon_2(\zeta_2) = \sqrt{\zeta_2 \cdot \frac{C^4\eta^4 k^2}{c_5 T(1-\rho)^4}\log(2k)} \tag{26}$$

for any $\zeta_1 > 1, \zeta_2 > 2$ that satisfy

$$\epsilon_1(\zeta_1) \leq \frac{3C\eta^2}{1-\rho}, \quad \epsilon_2(\zeta_2) \leq \frac{4C^2\eta^2}{(1-\rho)^2}k, \tag{27}$$

we have

$$\begin{aligned}\mathbb{P}\left(\|E_{:,j}\|_\infty \leq \left(2D^2_{\max}\epsilon_2(\zeta_2) + D_{\max}\right)\left(\epsilon_1(\zeta_1) + \lambda\right)\right) \geq 1 - \exp\left(-(\zeta_2 - 2)\log(2k)\right) \\ - \exp\left(-(\zeta_1 - 1)\log(2k)\right) - 2\delta\end{aligned} \tag{28}$$

Note that the last term on the right hand side is due to a simple union bound on the events that (21) holds and (11) recovers the correct sparsity pattern of $\Psi^*_{:,j}$. Now, upon defining

$$\zeta_1 = \frac{\log(2/\delta)}{\log(2k)} + 1 \tag{29}$$

$$\zeta_2 = \frac{\log(2/\delta)}{\log(2k)} + 2 \tag{30}$$

the inequalities in (27) are satisfied, provided that $T \geq \max\{\frac{1}{c_4}, \frac{2}{c_5}\} \cdot \log(4k/\delta)$. Furthermore, combining (29) and (30) with (28) results in

$$\mathbb{P}\left(\|E_{:,j}\|_\infty \leq \left(2D^2_{\max}\epsilon_2(\zeta_2) + D_{\max}\right)\left(\epsilon_1(\zeta_1) + \lambda\right)\right) \geq 1 - 3\delta \tag{31}$$

After plugging (29) and (30) into (26) and (25), the above inequality is reduced to

$$\|E_{:,j}\|_\infty \leq \left(2D_{\max}^2\sqrt{\frac{2}{c_5}\cdot\frac{C^4\eta^4}{T(1-\rho)^4}k^2\log(4k/\delta)} + D_{\max}\right) \times \left(\sqrt{\frac{1}{c_4}\cdot\frac{C^2\eta^4}{T(1-\rho)^2}\log(4k/\delta)} + c_6\sqrt{\frac{C^2\eta^4}{\gamma^2T(1-\rho)^2}\log(4(n+m)/\delta)}\right) \tag{32}$$

with probability of at least $1-3\delta$. Due to (20), one can write

$$D_{\max}^2\sqrt{\frac{2}{c_5}\cdot\frac{C^4\eta^4}{T(1-\rho)^4}k^2\log(4k/\delta)} \leq D_{\max} \tag{33}$$

Therefore,

$$\begin{aligned}\|E_{:,j}\|_\infty \leq& 3D_{\max}\left(\frac{1}{\sqrt{c_4}}+c_6\right)\sqrt{\frac{C^2\eta^4}{\gamma^2T(1-\rho)^2}\log(4(n+m)/\delta)}\\ =&\left(\frac{3}{\sqrt{c_4}}+3c_6\right)\frac{D_{\max}C\eta^2}{\gamma(1-\rho)}\sqrt{\frac{\log(4(n+m)/\delta)}{T}}\end{aligned} \tag{34}$$

with probability of at least $1-3\delta$. Now, to conclude the proof, it suffices to perform a union bound on different columns of the solution with indices $1\leq j\leq n$. This results in

$$\|E\|_\infty \leq \left(\frac{3}{\sqrt{c_4}}+3c_6\right)\frac{D_{\max}C\eta^2}{\gamma(1-\rho)}\sqrt{\frac{\log(4(n+m)/\delta)}{T}} \tag{35}$$

with probability of at least $1-3n\delta$. Replacing δ with $\frac{\delta}{3n}$ in the above inequality concludes the proof. □

B Proof of Auxiliary Lemmas

B.1 Proof of Lemma 2

To prove this lemma, we first introduce some notations. Define the matrix

$$R_1(X(\tau)) = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0\\ X(T_0) & X(T_0-1) & \dots & X(1) & X(0) & 0 & \dots & 0 & 0\\ X(T_0+1) & X(T_0) & \dots & X(2) & X(1) & X(0) & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ X(T_0+T-1) & X(T_0+T-2) & \dots & X(T) & X(T-1) & X(T-2) & \dots & X(0) & 0 \end{bmatrix} \tag{36}$$

where $X(\tau)$ is a matrix valued time-dependent signal. Furthermore, define the symmetrized matrix $\tilde{R}_1(\cdot) = \left(R_1(\cdot)+R_1(\cdot)^T\right)/2$. Finally, for a matrix N, define $[N]_{i\to j}$ as a matrix with the same size as H and with all rows equal to zero except for the j^{th} row which is equal to the i^{th} row of N.

Lemma 5. *Let λ_k be the k^{th} eigenvalue of the matrix R_G defined as*

$$R_G = \begin{bmatrix} \tilde{R}_1\left([(A+BK)^\tau]_{i\to j}\right)\eta^2 & \frac{1}{2}R_1\left([(A+BK)^\tau B]_{i\to j}\right)\eta^2 \\ \frac{1}{2}R_1\left([(A+BK)^\tau B]_{i\to j}\right)^T\eta^2 & 0 \end{bmatrix} \tag{37}$$

Then, the following relations hold

$$\max_k |\lambda_k| \leq \frac{3}{2}\frac{C\eta^2}{1-\rho} \tag{38}$$

$$\sum_k^{(n+m)(T+T_0+1)} \lambda_k^2 \leq \frac{9}{2}\frac{C^2\eta^4 T}{(1-\rho)^2} \tag{39}$$

Proof. Notice that

$$\|R_G\| \leq \eta^2 \left\|\tilde{R}_1\left([(A+BK)^\tau]_{i\to j}\right)\right\| + \frac{1}{2}\eta^2 \left\|R_1\left([(A+BK)^\tau B]_{i\to j}\right)\right\| \tag{40}$$

Similar to the proof of Lemma A.3 in [7], one can verify that

$$\left\|\tilde{R}_1\left([(A+BK)^\tau]_{i\to j}\right)\right\| \leq \frac{C}{1-\rho} \tag{41}$$

$$\left\|R_1\left([(A+BK)^\tau B]_{i\to j}\right)\right\| \leq \frac{C}{1-\rho} \tag{42}$$

This completes the proof of the second statement. Finally, it is easy to see that the rank of R_G is upper bounded by $2T$. This, together with the bound on the maximum eigenvalue completes the proof of the third statement. □

Define the matrix $P_{ji} \in \mathbb{R}^{n(T+T_0+1)\times m(T+T_0+1)}$ as

$$P_{ji} = \begin{bmatrix} 0_{(T_0+1)\times(T_0+1)} & 0_{(T_0+1)\times T} \\ 0_{T\times(T_0+1)} & I_{T\times T} \end{bmatrix} \otimes E_{ji} \tag{43}$$

where $E_{ji} \in \mathbb{R}^{n\times m}$ is a 0-1 matrix with 1 at its $(j,i)^{th}$ entry and 0 otherwise.

Lemma 6. *Let λ_k be the k^{th} eigenvalue of the matrix $\tilde{R}_G$ defined as*

$$\tilde{R}_G = \begin{bmatrix} \tilde{R}_1\left([K(A+BK)^\tau]_{i\to j}\right)\eta^2 & \frac{1}{2}R_1\left([K(A+BK)^\tau B]_{i\to j}\right)\eta^2 + \frac{1}{2}P_{ji}\eta^2 \\ \frac{1}{2}R_1\left([K(A+BK)^\tau B]_{i\to j}\right)^T\eta^2 + \frac{1}{2}P_{ji}^T\eta^2 & 0 \end{bmatrix} \tag{44}$$

Then, the following relations hold

$$\max_k |\lambda_k| \leq \frac{2C\eta^2}{1-\rho} \tag{45}$$

$$\sum_k^{(n+m)(T+T_0+1)} \lambda_k^2 \leq \frac{16C^2\eta^4 T}{(1-\rho)^2} \tag{46}$$

Proof. The proof of the first statement follows directly from Lemma 5. Furthermore, it is easy to verify that the rank of $\tilde{R}_G$ is upper bounded by $4T$. This, together with the upper bound on the maximum eigenvalue completes the proof of the third statement. □

Proof of Lemma 2: One can easily verify that

- if $i \in \{1, 2, \ldots, n\}$, then $G_i = \frac{1}{T} X_{:,i}^T W_{:,j} = \frac{1}{T} z^T R_G z$ where $z \in \mathbb{R}^{(n+m)(T+T_0+1)}$ is a random vector with independent zero-mean sub-Gaussian elements and $\|z\|_\psi \leq 1$.
- if $i \in \{n+1, \ldots, n+m\}$, then $G_i = \frac{1}{T} X_{:,i}^T W_{:,j} = \frac{1}{T} z^T \tilde{R}_G z$ where $z \in \mathbb{R}^{(n+m)(T+T_0+1)}$ is a random vector with independent zero-mean sub-Gaussian elements and $\|z\|_\psi \leq 1$.

Furthermore, note that the diagonal entries of both R_G and $\tilde{R}_G$ are zero and hence, $\mathbb{E}\left\{\frac{1}{T} z^T R_G z\right\} = \mathbb{E}\left\{\frac{1}{T} z^T \tilde{R}_G z\right\} = 0$. This, together with Hanson-Wright inequality and Lemmas 5 and 6 completes the proof. □

B.2 Proof of Lemma 3

Define the matrix

$$R_2(X(\tau)) = \begin{bmatrix} X(T_0) & X(T_0-1) & \ldots & X(1) & X(0) & 0 & \ldots & 0 & 0 \\ X(T_0+1) & X(T_0) & \ldots & X(2) & X(1) & X(0) & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ X(T_0+T-1) & X(T_0+T-2) & \ldots & X(T) & X(T-1) & X(T-2) & \ldots & X(0) & 0 \end{bmatrix} \tag{47}$$

and

$$\begin{aligned} H_{1i} &= R_2\left([(A+BK_0)^\tau]_{i,:}\right)\eta \in \mathbb{R}^{T\times n(T+T_0+1)} \\ H_{1j} &= R_2\left([(A+BK_0)^\tau]_{j,:}\right)\eta \in \mathbb{R}^{T\times n(T+T_0+1)} \\ H_{2i} &= R_2\left([(A+BK_0)^\tau B]_{i,:}\right)\eta \in \mathbb{R}^{T\times m(T+T_0+1)} \\ H_{2j} &= R_2\left([(A+BK_0)^\tau B]_{j,:}\right)\eta \in \mathbb{R}^{T\times m(T+T_0+1)} \\ H_{3i} &= R_2\left([K_0(A+BK_0)^\tau]_{i,:}\right)\eta \in \mathbb{R}^{T\times n(T+T_0+1)} \\ H_{3j} &= R_2\left([K_0(A+BK_0)^\tau]_{j,:}\right)\eta \in \mathbb{R}^{T\times n(T+T_0+1)} \\ H_{4i} &= R_2\left([K_0(A+BK_0)^\tau B]_{i,:}\right)\eta^2 + P_i\eta \in \mathbb{R}^{T\times m(T+T_0+1)} \\ H_{4j} &= R_2\left([K_0(A+BK_0)^\tau B]_{j,:}\right)\eta^2 + P_j\eta \in \mathbb{R}^{T\times m(T+T_0+1)} \end{aligned} \tag{48}$$

where the matrix $P_j \in \mathbb{R}^{T\times m(T+T_0+1)}$ has the form

$$P_j = \begin{bmatrix} 0_{T\times(T_0+1)} & I_{T\times T} \end{bmatrix} \otimes e_j \tag{49}$$

and $e_j \in \mathbb{R}^{1\times m}$ with 1 at its j^{th} entry and 0 otherwise. These notations will be used in the subsequent lemma.

Lemma 7. *Let $\{k_1, k_2, k_3, k_4\} \in \{1, 2, 3, 4\}^4$, where $k_1 \neq k_4$ and $k_2 \neq k_3$. Furthermore, let λ_k be the k^{th} eigenvalue of the following matrix*

$$R_M(k_1, k_2, k_3, k_4) = \begin{bmatrix} \frac{1}{2}(H_{k_1 i}^\top H_{k_3 j} + H_{k_3 j}^\top H_{k_1 i}) & \frac{1}{2}(H_{k_1 i}^\top H_{k_4 j} + H_{k_3 j}^\top H_{k_2 i}) \\ \frac{1}{2}(H_{k_4 j}^\top H_{k_1 i} + H_{k_2 i}^\top H_{k_3 j}) & \frac{1}{2}(H_{k_2 i}^\top H_{k_4 j} + H_{k_4 j}^\top H_{k_2 i}) \end{bmatrix} \\ \in \mathbb{R}^{(n+m)(T+T_0+1)\times(n+m)(T+T_0+1)} \tag{50}$$

Then, the following relations hold

$$\max_k |\lambda_k| \leq \frac{6C^2\eta^2}{(1-\rho)^2} \tag{51}$$

$$\sum_{k=1}^{(n+m)(T+T_0+1)} \lambda_k^2 \leq \frac{72C^4\eta^4}{(1-\rho)^4} \tag{52}$$

Proof. To show the validity of the first statement, one can write

$$\begin{aligned} &|||R_M(k_1, k_2, k_3, k_4)||| \\ &\leq \frac{1}{2}\max\{|||H_{k_1 i}^\top H_{k_3 j} + H_{k_3 j}^\top H_{k_1 i}|||, |||H_{k_2 i}^\top H_{k_4 j} + H_{k_4 j}^\top H_{k_2 i}|||\} + \frac{1}{2}|||H_{k_1 i}^\top H_{k_4 j} + H_{k_3 j}^\top H_{k_2 i}||| \\ &\leq \frac{1}{2}\max\{|||H_{k_1 i}^\top|||\,|||H_{k_3 j}||| + |||H_{k_3 j}^\top|||\,|||H_{k_1 i}|||, |||H_{k_2 i}^\top|||\,|||H_{k_4 j}||| + |||H_{k_4 j}^\top|||\,|||H_{k_2 i}|||\} \\ &\quad + \frac{1}{2}\left(|||H_{k_1 i}^\top|||\,|||H_{k_4 j}||| + |||H_{k_3 j}^\top|||\,|||H_{k_2 i}|||\right) \end{aligned} \tag{53}$$

Furthermore, similar to the proof of Lemma A.4 in [7], one can verify that

$$|||H_{ri}|||, |||H_{rj}||| \leq \frac{C}{1-\rho} \qquad \text{if } r = 1, 2, 3$$

$$|||H_{ri}|||, |||H_{rj}||| \leq \frac{2C}{1-\rho} \qquad \text{if } r = 4$$

Combining this with the above inequality completes the proof of the first statement. Finally, note that $R_M(k_1, k_2, k_3, k_4)$ can be written as

$$R_M^{(1)} = \frac{1}{2}\begin{bmatrix} H_{k_1 i}^\top \\ H_{k_2 i}^\top \end{bmatrix}\begin{bmatrix} H_{k_3 j} & H_{k_4 j} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} H_{k_3 j}^\top \\ H_{k_4 j}^\top \end{bmatrix}\begin{bmatrix} H_{k_1 i} & H_{k_2 i} \end{bmatrix} \tag{54}$$

which implies that its rank is upper bounded by $2T$. This, together with the upper bound on the maximum eigenvalue completes the proof. □

Lemma 8. *We have $\mathbb{E}(M) = M^*$.*

Proof. Define

$$X_1 = \begin{bmatrix} x(0) & \dots & x(T-1) \end{bmatrix}$$
$$X_2 = \begin{bmatrix} Kx(0) + v(0) & \dots & Kx(T-1) + v(T-1) \end{bmatrix}$$

The theorem can be proven by showing

$$\begin{aligned}\frac{1}{T}\mathbb{E}(X_1X_1^T) &= Q^*,\\ \frac{1}{T}\mathbb{E}(X_2X_1^T) &= KQ^*,\\ \frac{1}{T}\mathbb{E}(X_2X_2^T) &= KQ^*K^T + \sigma_v^2 I,\end{aligned} \tag{55}$$

In what follows, we show the validity of the first equality. The other equalities can be proven in a similar manner. We have

$$\frac{1}{T}\mathbb{E}(X_1X_1^T) = \frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}(x(\tau)x(\tau)^T) \tag{56}$$

Furthermore, notice that $x(0)$ has a stationary distribution and hence, $\mathbb{E}(x(0)x(0)^T) = Q^*$. Furthermore,

$$\mathbb{E}(x(1)x(1)^T) = (A+BK)Q^*(A+BK)^T + \sigma_w^2 I + \sigma_v^2 BB^T = Q^* \tag{57}$$

where the second inequality is due to (**??**). Similarly, one can show that $\mathbb{E}(x(\tau)x(\tau)^T) = Q^*$ for every $\tau \in \{2,3,\ldots,T-1\}$ and hence,

$$\frac{1}{T}\mathbb{E}(X_1X_1^T) = \frac{1}{T}\sum_{\tau=0}^{T-1}Q^* = Q^* \tag{58}$$

This completes the proof. □

Proof of Lemma 3: Due to Lemma 8 and upon taking $T_0 \to \infty$, we have

$$\mathbb{P}\{|M_{ij} - M_{ij}^*| > \epsilon\} = \mathbb{P}\{|M_{ij} - \mathbb{E}(M_{ij})| > \epsilon\} \tag{59}$$

and hence, it suffices to obtain a bound for $\mathbb{P}\{|M_{ij} - \mathbb{E}(M_{ij})| > \epsilon\}$. We should consider four cases:

- If $i,j \in \{1,2,\ldots,n\}$, then $M_{ij} = \frac{1}{T}z^T R_M(1,2,1,2)z$, where $z \in \mathbb{R}^{(n+m)(T+T_0+1)}$ is a random vector with independent zero-mean sub-Gaussian elements and $\|z\|_\psi \leq 1$.
- If $i \in \{1,2,\ldots,n\}$ and $j \in \{n+1,n+2,\ldots,n+m\}$, then $M_{ij} = \frac{1}{T}z^T R_M(1,2,3,4)z$, where $z \in \mathbb{R}^{(n+m)(T+T_0+1)}$ is a random vector with independent zero-mean sub-Gaussian elements and $\|z\|_\psi \leq 1$.
- If $i \in \{n+1,n+2,\ldots,n+m\}$ and $j \in \{1,2,\ldots,n\}$, then $M_{ij} = \frac{1}{T}z^T R_M(3,4,1,2)z$, where $z \in \mathbb{R}^{(n+m)(T+T_0+1)}$ is a random vector with independent zero-mean sub-Gaussian elements and $\|z\|_\psi \leq 1$.
- If $i \in \{n+1,n+2,\ldots,n+m\}$ and $j \in \{n+1,n+2,\ldots,n+m\}$, then $M_{ij} = \frac{1}{T}z^T R_M^{(4)}(3,4,3,4)z$, where $z \in \mathbb{R}^{(n+m)(T+T_0+1)}$ is a random vector with independent zero-mean sub-Gaussian elements and $\|z\|_\psi \leq 1$.

Invoking the Hanson-Wright inequality and Lemma 7 for the aforementioned cases completes the proof. □

B.3 The proof of Proposition 1

We need the following lemma:

Lemma 9. *We have*

$$\|M^*\| \leq \frac{85C^2\eta^2}{1-\rho} \tag{60}$$

Proof. One can easily verify that

$$Q^* = \sum_{\tau=0}^{\infty} \begin{bmatrix}\sigma_w(A+BK_0)^\tau & \sigma_v(A+BK_0)^\tau B\end{bmatrix} \begin{bmatrix}\sigma_w(A+BK_0)^\tau & \sigma_v(A+BK_0)^\tau B\end{bmatrix}^T \tag{61}$$

and hence

$$\begin{aligned} M^* &= \begin{bmatrix}0 & 0\\ 0 & \sigma_v^2 I\end{bmatrix} \\ &+ \sum_{\tau=0}^{\infty} \begin{bmatrix}\sigma_w(A+BK_0)^\tau & \sigma_v(A+BK_0)^\tau B\\ \sigma_w K_0(A+BK_0)^\tau & \sigma_v K_0(A+BK_0)^\tau B\end{bmatrix} \begin{bmatrix}\sigma_w(A+BK_0)^\tau & \sigma_v(A+BK_0)^\tau B\\ \sigma_w K_0(A+BK_0)^\tau & \sigma_v K_0(A+BK_0)^\tau B\end{bmatrix}^T \end{aligned} \tag{62}$$

Therefore, with the assumption $\sigma_w, \sigma_w \leq 1$ and the fact that $\sigma_u, \sigma_v \leq \sqrt{5}\eta$ (the proof of which is simple and can be found, e.g., in [19]), one can write

$$\begin{aligned} |||M^*||| &\leq 5\eta^2 + 5\eta^2 \sum_{\tau=0}^{\infty} \left|\left|\left| \begin{bmatrix}(A+BK_0)^\tau & (A+BK_0)^\tau B\\ K_0(A+BK_0)^\tau & K_0(A+BK_0)^\tau B\end{bmatrix} \right|\right|\right|^2 \\ &\leq 5\eta^2 + 5\eta^2 \sum_{\tau=0}^{\infty} (|||(A+BK_0)^\tau||| + |||K_0(A+BK_0)^\tau B||| + |||K_0(A+BK_0)^\tau||| \\ &\qquad + |||(A+BK_0)^\tau B|||)^2 \\ &\leq 5\eta^2 + 80\eta^2 \sum_{\tau=0}^{\infty} C^2\rho^{2\tau} \\ &\leq \frac{85C^2\eta^2}{1-\rho} \end{aligned} \tag{63}$$

This completes the proof. □

Based on this lemma, we will take a similar approach to the proof of Theorem 3.1 in [7] to prove the correct sparsity recovery of the system matrices.

Proof of Proposition 1: To prove this proposition, we need to show that the conditions of Lemma 1 holds with high probability. To ensure that the first condition on G implies the second one, it suffices to have

$$\frac{\lambda\gamma}{3} \leq \frac{\Psi_{\min} C_{\min}}{4k} - \lambda \tag{64}$$

Noting that $0 < \gamma < 1$, one can verify that the following bound on λ is enough to guarantee that the above inequality holds:

$$\lambda \leq \frac{\Psi_{\min} C_{\min}}{8k} \tag{65}$$

Furthermore, to ensure the last two conditions on M, it suffices to have

$$\left|\left|\left| M_{:\mathcal{A}_j} - M^*_{:\mathcal{A}_j} \right|\right|\right|_\infty \leq \frac{\gamma C_{\min}}{12\sqrt{k}} \tag{66}$$

Based on the above analysis, it suffices to have

$$\mathbb{P}\left(\|G\|_\infty > \frac{\gamma\lambda}{3} \right) \leq \frac{\delta}{2} \tag{67a}$$

$$\mathbb{P}\left(\left|\left|\left| M_{:\mathcal{A}_j} - M^*_{:\mathcal{A}_j} \right|\right|\right|_\infty > \frac{\gamma C_{\min}}{12\sqrt{k}} \right) \leq \frac{\delta}{2} \tag{67b}$$

in order to ensure the exact recovery with probability of at least $1 - \delta$. First, we derive conditions under which (67a) holds. Based on Lemma 2, one needs to ensure the following inequalities

$$2(n+m)\exp\left(-c_4 \cdot \frac{(1-\rho)^2}{C^2\eta^4} \frac{\gamma^2\lambda^2}{9} T \right) \leq \frac{\delta}{2} \tag{68a}$$

$$\lambda \leq \frac{\Psi_{\min} C_{\min}}{8k} \tag{68b}$$

$$\frac{\gamma\lambda}{3} \leq \frac{3C\eta^2}{1-\rho} \tag{68c}$$

where (68c) is a technical condition that is required by Lemma 2. It can be easily verified that (68a) is satisfied with the choice of

$$\lambda = \sqrt{\frac{9}{c_4} \cdot \frac{C^2\eta^4}{\gamma^2 T (1-\rho)^2} \log(4(n+m)/\delta)} \tag{69}$$

Based on the chosen value for λ and in order to satisfy (68b), we should have the following lower bound on T

$$T \geq \frac{576}{c_4} \cdot \frac{C^2\eta^4 k^2}{\Psi^2_{\min} C^2_{\min} \gamma^2 (1-\rho)^2} \log(4(n+m)/\delta) \tag{70}$$

Similarly, to ensure the validity of (68c), we should have

$$T \geq \frac{1}{c_4} \cdot \log(4(n+m)/\delta) \tag{71}$$

Now, we will derive the conditions under which (67b) is satisfied using Lemma 3. To this goal, first we need to show that the following condition is satisfied:

$$0 < \epsilon < \frac{4C^2\eta^2}{(1-\rho)^2} \tag{72a}$$

which is reduced to

$$\frac{\gamma C_{\min}}{12\sqrt{k}} < \frac{4C^2\eta^2}{(1-\rho)^2}k \tag{73}$$

with the choice of $\epsilon = \frac{\gamma C_{\min}}{12\sqrt{k}}$. However, the above inequality implies that

$$k^{3/2} > \frac{1}{48}\frac{\gamma C_{\min}(1-\rho)^2}{C^2\eta^2} \tag{74}$$

A sufficient condition for the correctness of the above inequality is to have $k \geq 2$. To see this, note that

$$C_{\min} \leq \lambda_{\min}(M^*_{\mathcal{A}_j,\mathcal{A}_j}) \leq \lambda_{\max}(M^*) \leq \frac{85C^2\eta^2}{1-\rho} \tag{75}$$

where the last inequality is due to Lemma 9. Therefore,

$$\frac{1}{48}\frac{\gamma C_{\min}(1-\rho)^2}{C^2\eta^2} \leq \frac{85}{48} < 2 \tag{76}$$

which implies $k \geq 2$. Finally, to verify (67b) and according to Lemma 3, it suffices to have

$$2(n+m)k\exp\left(-c_5 \cdot \frac{(1-\rho)^4}{C^4\eta^4}\frac{\gamma^2 C^2_{\min}}{144k}T\right) \leq \frac{\delta}{2} \tag{77}$$

This implies that

$$T \geq \frac{144}{c_5} \cdot \frac{C^4\eta^4 k}{(1-\rho)^4\gamma^2 C^2_{\min}} \log(4(n+m)k/\delta) \tag{78}$$

Based on the above analysis, the inequalities (70), (71), and (78) impose lower bounds on T. Comparing these inequalities with (20), one can verify that the latter dominates all of them. This completes the proof. □

B.4 Proof of Lemma 4

To prove this lemma, first we introduce the KKT conditions for (11).

Lemma 10 (KKT conditions)**.** *$\hat{\Psi}_{:,j}$ is an optimal solution for (11) if and only if it satisfies*

$$M(\hat{\Psi}_{:,j} - \Psi^*_{:,j}) - G + \lambda S = 0 \tag{79}$$

for some $S \in \partial\|\hat{\Psi}_{:,j}\|_1$, where $\partial\|\hat{\Psi}_{:,j}\|_1$ is the sub-differential of $\|\cdot\|_1$ at $\hat{\Psi}_{:,j}$.

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

The following lemma is an immediate consequence of the KKT conditions.

Lemma 11. *Assuming that (11) recovers the correct sparsity pattern of $\Psi^*_{:,j}$, the following equalities hold for $E = \hat{\Psi}_{:,j} - \Psi^*_{:,j}$:*

$$E_{\mathcal{A}_j^c} = 0 \tag{80}$$

$$E_{\mathcal{A}_j} = (M_{\mathcal{A}_j,\mathcal{A}_j})^{-1}G_{\mathcal{A}_j} - \lambda(M_{\mathcal{A}_j,\mathcal{A}_j})^{-1}S_{\mathcal{A}_j} \tag{81}$$

Proof. Due to the correct sparsity recovery, we have $E_{\mathcal{A}_j^c} = 0$. This, together with the KKT conditions imply that

$$M_{\mathcal{A}_j\mathcal{A}_j} E_{\mathcal{A}_j} - G_{\mathcal{A}_j} + \lambda S_{\mathcal{A}_j} = 0 \tag{82}$$

Solving the above equation with respect to $E_{\mathcal{A}_j}$ will conclude the proof. □

Proof of Lemma 4: Based on Lemma 11, one can write

$$\|E_{\mathcal{A}_j}\|_\infty \leq \underbrace{\left\|(M_{\mathcal{A}_j\mathcal{A}_j})^{-1} G_{\mathcal{A}_j}\right\|_\infty}_{Z_1} + \underbrace{\lambda \left\|(M_{\mathcal{A}_j\mathcal{A}_j})^{-1} S_{\mathcal{A}_j}\right\|_\infty}_{Z_2} \tag{83}$$

In what follows, we will provide a bound for each term in the above inequality. For Z_2, one can write

$$\begin{aligned} Z_2 &\leq \lambda \left\| \left((M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right) S_{\mathcal{A}_j} \right\|_\infty + \lambda \left\| (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} S_{\mathcal{A}_j} \right\|_\infty \\ &\leq \lambda \left(\left|\!\left|\!\left| (M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right|\!\right|\!\right|_\infty + \left|\!\left|\!\left| (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right|\!\right|\!\right|_\infty \right) \\ &\leq \lambda \left(\underbrace{\left|\!\left|\!\left| (Q_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right|\!\right|\!\right|_\infty}_{\Delta} + D_{\max} \right) \end{aligned} \tag{84}$$

On the other hand, we have

$$\begin{aligned} (M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} =& (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \left(M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right) (M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \\ =& (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \\ & - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \left(M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right) \left((M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} + \left((M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right) \right) \end{aligned} \tag{85}$$

and therefore

$$\Delta \leq \left|\!\left|\!\left| (M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right|\!\right|\!\right|_\infty \left|\!\left|\!\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty \left(\left|\!\left|\!\left| (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right|\!\right|\!\right|_\infty + \Delta \right) \tag{86}$$

This leads to

$$\begin{aligned} \Delta &\leq \frac{D^2_{\max}}{1 - D_{\max} \left|\!\left|\!\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty} \left|\!\left|\!\left| Q_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty \\ &\leq \frac{D^2_{\max}}{1 - \min\{1/2, \eta^2\}} \left|\!\left|\!\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty \\ &\leq 2 D^2_{\max} \left|\!\left|\!\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty \end{aligned} \tag{87}$$

where the last inequality is due to the assumption (14). Combining the above inequality with (84) gives rise to

$$Z_2 \leq \lambda \left(2 D^2_{\max} \left|\!\left|\!\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty + D_{\max} \right) \tag{88}$$

Now we will bound Z_1. Similar to Z_2, we have

$$\begin{aligned} Z_1 &\leq \left(\left|\!\left|\!\left| (M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right|\!\right|\!\right|_\infty + \left|\!\left|\!\left| (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right|\!\right|\!\right|_\infty \right) \|G_{\mathcal{A}_j}\|_\infty \\ &\leq \left(\Delta + \left|\!\left|\!\left| (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right|\!\right|\!\right|_\infty \right) \|G_{\mathcal{A}_j}\|_\infty \\ &\leq \left(2D^2_{\max} \left|\!\left|\!\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty + D_{\max} \right) \|G_{\mathcal{A}_j}\|_\infty \end{aligned} \tag{89}$$

Putting together (89) and (88) completes the proof. □