Exact Recovery for System Identification with More Corrupt Data than Clean Data

Baturalp Yalcin, Haixiang Zhang, Javad Lavaei, and Murat Arcak

***Abstract*— This paper investigates the system identification problem for linear discrete-time systems under adversaries and analyzes two lasso-type estimators. We examine both asymptotic and non-asymptotic properties of these estimators in two separate scenarios, corresponding to deterministic and stochastic models for the attack times. Since the samples collected from the system are correlated, the existing results on lasso are not applicable. We prove that when the system is stable and attacks are injected periodically, the sample complexity for exact recovery of the system dynamics is linear in terms of the dimension of the states. When adversarial attacks occur at each time instance with probability p, the required sample complexity for exact recovery scales polynomially in the dimension of the states and the probability p. This result implies almost sure convergence to the true system dynamics under the asymptotic regime. As a by-product, our estimators still learn the system correctly even when more than half of the data is compromised. We highlight that the attack vectors are allowed to be correlated with each other in this work, whereas we make some assumptions about the times at which the attacks happen. This paper provides the first mathematical guarantee in the literature on learning from correlated data for dynamical systems in the case when there is less clean data than corrupt data.**

***Index Terms*— System Identification, Robust Control, Statistical Learning, Linear Systems, Uncertain Systems**

I. Introduction

Dynamical systems serve as the fundamental components in reinforcement learning and control systems. The system dynamics may not be known exactly when the system is complex. Therefore, learning the underlying system dynamics, named the system identification problem, and using the data collected from the system are essential in robotics, control theory, time-series, and reinforcement learning applications. The system identification problem with small disturbances using the least square estimator has been ubiquitously studied, and the literature for this problem is overly rich [1]. Despite several advances in this field, most results in system identification focus on the asymptotic properties, i.e., properties of the estimators at infinity, of the proposed estimators only. Nonetheless, the non-asymptotic analysis of the system identification problem has gained interest in recent years [2]–[5]. Although non-asymptotic analysis is harder, it is crucial to understand the required sample complexity for online control problems.

The robust learning of dynamical systems is crucial for safety-critical applications, such as autonomous driving [6], unmanned aerial vehicles [7], and robotic arms [8]. While recent papers have addressed online non-asymptotic control of linear time-invariant (LTI) systems, their applicability often hinges on the assumption of small noise in measurements, neglecting scenarios involving large magnitudes of noise indicative of adversarial attacks or data corruption [9]–[11]. These papers utilize recent advances in high-dimensional statistics and learning theory to analyze the properties of the solution even when the data samples are correlated. The work [12] provides a tutorial on proof techniques. Least-square estimators are the main tool in those works, which are susceptible to outliers and large noise in the system. Consequently, we propose two new non-smooth estimators inspired by the lasso problem and robust regression literature [13]. We study the required sample complexity for the exact recovery of LTI systems using these estimators when there are sporadic large disturbance injections to the system.

The robust regression and learning problems under adversaries are ubiquitously studied in the literature [14]–[17]. However, existing methods for analyzing the estimators cannot be directly generalized to control problems due to the correlation between the samples. Therefore, different strategies have been developed recently to tackle this challenge. Firstly, the system is initiated multiple times, and the data point at the end of each run is used to obtain uncorrelated data points, as in [18]. However, obtaining multiple trajectories is not viable and cost-efficient for most safety-critical applications. One method with a single trajectory relies on the persistent excitation of the states so that the dynamics can be explored thoroughly. This is achieved by injecting a Gaussian noise input into the system. Small ball techniques are used to analyze the properties of the estimator [9], [19], [20]. This technique employs normalized martingale bounds for the estimation error when the excitation is large enough [9].

Unlike the non-asymptotic analysis of correlated data, the least-squares estimator offers a closed-form solution when the system is subjected to small white noise [21]–[23]. As long as the noise magnitudes are not large, the least-squares estimator performs relatively well. The estimation error asymptotically

This work was supported by grants from AFOSR, ARO, ONR, and NSF.

B. Yalcin and J. Lavaei are with the Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA, USA, 94720 (e-mail: baturalp_yalcin@berkeley.edu; lavaei@berkeley.edu).

H. Zhang is with the Department of Mathematics, University of California, Berkeley, CA, USA, 94720 (e-mail: haixiang_zhang@berkeley.edu).

M. Arcak is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA, 94720 (e-mail: arcak@berkeley.edu).

converges to zero with the optimal rate of $T^{-1/2}$, where T is the number of samples collected from the system [9]. However, it is not robust to adversarial attacks, and the literature on robust learning of dynamical systems is limited. The work by [24] defines the null space property (NSP) to analyze a lasso-type estimator for the system. It provides necessary and sufficient conditions for exact recovery when NSP is satisfied, which is NP-hard to check. To circumvent the computational complexity, we build upon [24] and study robust estimators from a non-asymptotic point of view under standard assumptions, such as the system being stable and the attacks being sub-Gaussian.

Contributions: We study discrete-time linear time-invariant systems of the form $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i$, where $\bar{A} \in \mathbb{R}^{n \times n}$ and $\bar{B} \in \mathbb{R}^{n \times m}$ are unknown matrices of the model. We aim to learn these matrices from the samples $\{x_i; u_i\}_{i=0}^{T+1}$ of a single initialization of the system when the disturbance vectors $\bar{d}_i$ are adversarial. Here, the adversarial noise refers to a vector that is designed to deteriorate the performance of the estimator. Thus, the adversarial vectors $\{\bar{d}_i\}_{i=0}^{T-1}$ can take arbitrarily large finite values, be dependent over time, and can have any undesirable structures. We say that an adversarial attack occurs whenever $\bar{d}_i$ is non-zero, and we have no information on the value of $\bar{d}_i$. If $\bar{d}_i$ is zero, there is no attack or adversary at time i. In our setting, we study systems that are not subject to ordinary minor measurement or modeling errors, and instead the non-zero noise or disturbance stems from an adversarial event.

We study two convex estimators based on the minimization of the ℓ_2 and ℓ_1 norms of the estimated disturbance vectors, $\sum_{i=0}^{T-1} \|d_i\|_2$ and $\sum_{i=0}^{T-1} \|d_i\|_1$, with the decision variables A, B, and $\{d_i\}_{i=0}^{T-1}$ subject to $x_{i+1} = Ax_i + Bu_i + d_i$, given the samples $\{x_i; u_i\}_{i=0}^{T-1}$:

$$\min_{A \in \mathbb{R}^{n \times n}; B \in \mathbb{R}^{n \times m}} \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t - Bu_t\|_\ell ; \qquad \ell \in \{1; 2\}.$$

This is equivalent to an empirical risk minimization problem for which the loss function is the ℓ_1 and ℓ_2 norms depending on the choice of ℓ. We employ a non-smooth objective function to obtain a robust estimator. The arbitrary injection of adversaries may happen infrequently in time. In that case, the attacks occur sparsely in time. Conversely, the vector $\bar{d}_i$ at each attack time i could be dense, and there is no limitation on how sparse the vector is. The ℓ_2 norm estimator is the most effective in this case. In contrast, the ℓ_1 norm estimator is preferable if the vector $\bar{d}_i$ at each attack time is structured and known to be sparse. We summarize our contributions below.

i) We first consider the case when the adversarial noise injections, i.e., adversarial attacks, happen periodically over time with the period D. We show that both of our estimators exactly recover the true system matrices $\bar{A}$ and $\bar{B}$ when the system is stable and the number of samples, i.e., T, is larger than $n + D$.

ii) We then consider a probabilistic model for the occurrence of attacks, in which there is an arbitrary noise injection at each time instance i with probability p, independent of previous time periods. Nevertheless, we allow these noise injections, or attack vectors, to be dependent. We study the required sample complexity of our estimators for exact recovery when the attack vectors are stealthy. Suppose that the adversarial noise and the input sequence are sub-Gaussian random vectors and possibly dependent. Then, the estimators achieve exact recovery with probability at least $1 - \delta$ if the time horizon T satisfies the inequality $T \geq Q(\max\{T^1_{\text{sample}}; T^2_{\text{sample}}\})$, where T^1_{sample} and T^2_{sample} are defined as

$$n^2 R_1 \log\left(\frac{nR_1}{\delta}\right);$$

and

$$nmR_2 \log\left(\frac{nR_2}{\delta}\right);$$

with the constants R_1 and R_2 defined in Theorem 4.

iii) As a corollary to the previous result, we show that the estimators converge to true system matrices almost surely when the attack vectors are stealthy. Otherwise, if the attack vectors are not stealthy, the system operator could detect the abnormalities and stop the system, which is not a desired outcome for the adversarial agent or attacker. This is the first paper that studies the adversarial attack structure for the system identification problem to obtain sample complexity using non-asymptotic analysis techniques.

This paper is organized as follows. In Sections 2 and 3, we introduce the notations used in the paper and formulate the problem, respectively. In Section 4, we study the convergence and sample complexity properties of our estimators in the case when the system is autonomous. In Section 5, we generalize the results to non-autonomous systems. In Section 6, we demonstrate the results on a biomedical system that models blood sugar levels with the injection of bolus insulin. This work provides the first bound in the literature on sample complexity for dynamical systems under adversaries, and its techniques can be adopted to study other robust online learning problems.

II. Notation and Preliminaries

For a matrix Z, $\|Z\|_F$ denotes the Frobenius norm of a matrix. For a vector z, $\|z\|_1$, $\|z\|_2$, and $\|z\|_\infty$ denote its ℓ_1, ℓ_2, and ℓ_∞ norms, respectively. Given two functions f and g, the notation $f(x) = Q(g(x))$ means that there exist universal positive constants c_1 and c_2 such that $c_1 g(x) \leq f(x) \leq c_2 g(x)$. The relation $f(x) \lesssim g(x)$ holds if there exists a universal positive constant c_3 such that $f(x) \leq c_3 g(x)$ holds with high probability when T is large. The relation $f(x) \asymp g(x)$ holds if $g(x) \lesssim f(x)$. $|S|$ shows the cardinality of a given set S. For two vectors v and w, $\langle v; w\rangle$ is the inner product between those vectors in their respective vector space. Furthermore, we use the notation $v \otimes w = vw^T$ to denote the outer product. $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ denote the probability of an event and the expectation of a random variable. A Gaussian random variable X with mean m and covariance matrix Σ is written as $X \sim N(m; \Sigma)$. Since we restrict the disturbance vectors to be sub-Gaussian, we formally define them below.

Definition 1 (Sub-Gaussian Random Variable [25]): A random variable $X \in \mathbb{R}$ with mean $m = \mathbb{E}[X]$ is sub-Gaussian

with parameter s if

$$\mathbb{E}[e^{\lambda(X-m)}] \leq e^{\lambda^2 s^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

Moreover, a random vector $X \in \mathbb{R}^n$ with mean $m = \mathbb{E}[X]$ is sub-Gaussian with parameter s if

$$\mathbb{E}[e^{\lambda\langle u, X-m\rangle}] \leq e^{\lambda^2 s^2/2}, \quad \forall \lambda \in \mathbb{R}, u \in \mathbb{R}^n, \|u\|_2 = 1.$$

Informally, a sub-Gaussian random variable with parameter s has the property that its tails are less dense than those of a Gaussian random variable with variance s^2. We will utilize concentration bounds for sub-Gaussian random variables to verify that the optimality conditions for our proposed estimators are satisfied with high probability. The main concentration inequality for sub-Gaussian random variables is Hoeffding's bound.

Lemma 1: (Hoeffding's Bound [25]) Suppose that the variable X has mean m and sub-Gaussian parameter s. Then, for all $t > 0$, we have

$$\mathbb{P}(|X-m| > t) \leq 2\exp\left(\frac{-t^2}{2s^2}\right).$$

We use the union bound over the set of coordinates and other sets with finite cardinality. Let S be a set with finite cardinality, $|S| < \infty$, and E_i be the event related to element i in the set S. Then, we can write the union bound as

$$\mathbb{P}\left(\bigcup_{i\in S} E_i\right) \leq \sum_{i\in S} \mathbb{P}(E_i).$$

Since we use non-smooth objective functions with ℓ_1 and ℓ_2 norms, we introduce the subdifferentials of the ℓ_1 and ℓ_2 norms.

Definition 2 (Subdifferential of ℓ_2 Norm): Given a vector $z \in \mathbb{R}^n$, the subdifferential of $\|z\|_2$ is denoted as $\partial\|z\|_2$ and is given as

$$\partial\|z\|_2 = \begin{cases} \frac{z}{\|z\|_2}, & \text{if } z \neq 0, \\ B_2(1), & \text{otherwise.} \end{cases}$$

where $B_2(1) = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ is the ℓ_2 norm unit ball.

Definition 3 (Subdifferential of ℓ_1 Norm): Given a vector $z \in \mathbb{R}^n$ with entries $z_i, i = 1,\ldots,n$, the subdifferential of the $\|z\|_1$ is denoted as $\partial\|z\|_1$ and is given as

$$\partial\|z\|_1^i = \begin{cases} 1, & \text{if } z_i > 0, \\ -1, & \text{if } z_i < 0, \\ [-1, 1], & \text{otherwise,} \end{cases}$$

where $\partial\|z\|_1^i$ is the i-th coordinate of the subdifferential of $\|z\|_1$.

Note that while the subdifferential of the ℓ_1 norm is coordinate-wise separable, the subdifferential of the ℓ_2 norm is not coordinate-wise separable. Whenever the vector z is equal to 0, the subdifferential of the ℓ_2 norm is the ℓ_2 norm unit ball, whereas the subdifferential of the ℓ_1 norm is the ℓ_∞ norm unit ball, which is

$$B_\infty(1) = \{x \in \mathbb{R}^n : \|x\|_\infty \leq 1\}.$$

We also define the unit ball $S_2(1)$ as

$$S_2(1) = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$$

that is the set of all the points on the sphere with radius 1.

The asymptotic analysis of the system identification problem concerns the convergence rate to the true parameter at an infinite-time horizon. However, historically, asymptotic analysis has not provided the required sample complexity to obtain a solution within a given error tolerance. In contrast, non-asymptotic analysis deals with the finite-time behavior of the estimators using learning theory and high-dimensional statistics. It provides the required sample complexity to bound the estimation error within the specified tolerance with high probability. Consequently, non-asymptotic analysis is more challenging than asymptotic analysis. Our goal is to provide the minimum required number of samples to recover the true parameters of the system with a high probability using techniques designed for non-asymptotic analysis.

III. Problem Formulation

We consider a linear time-invariant dynamical system over the time horizon $[0, T]$, $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i, i = 0, 1, \ldots, T-1$, where $\bar{A} \in \mathbb{R}^{n\times n}$ and $\bar{B} \in \mathbb{R}^{n\times m}$ are unknown system matrices, and $\bar{d}_i \in \mathbb{R}^n$ are unknown system disturbances. Given the set of state measurements $\{x_i\}_{i=0}^{T}$ and the set of inputs $\{u_i\}_{i=0}^{T-1}$, the goal is to estimate the unknown system matrices $\bar{A}$ and $\bar{B}$. In this paper, the disturbance vectors $\{\bar{d}_i\}_{i=0}^{T-1}$ can be engineered to be large if there is an outside attack on the system from an agent or there is a sensor/actuation fault that leads to major corruption in the system dynamics. Throughout the paper, the disturbance vectors $\{\bar{d}_i\}_{i=0}^{T-1}$ are also called (adversarial) attack vectors. Moreover, the agent who engineers the disturbance vectors is called an attacker. As opposed the majority of the literature, we assume that the disturbance vectors $\{\bar{d}_i\}_{i=0}^{T-1}$ can be dependent on the disturbance vectors from the previous time instances and there is no specific distribution assumption for these vectors except the sub-Gaussian assumption. We represent the time indices of the attacks or large disturbance vectors with the set K, that is $K = \{i : \bar{d}_i \neq 0, i \in 0, 1, \ldots, T-1\}$. These time instances are called the attack times and K is the set of attack times. Similarly, the set of time instances without attack or corrupted data is shown with $K^c = \{i : \bar{d}_i = 0, i \in 0, 1, \ldots, T-1\}$. These time instances are called the no-attack times, and K^c is the set of no-attack times. The data corresponding to attack times are corrupted, whereas the data corresponding to no-attack times are uncorrupted.

We establish the exact recovery of the proposed estimators when there are large disturbances in the system. In such cases, the least-squares method cannot achieve exact recovery, a fact that can be easily verified from its closed-form solution. Define the matrices $X := [x_0, \ldots, x_{T-1}]$ and $\bar{D} := [\bar{d}_0, \ldots, \bar{d}_{T-1}]$. The solution for the least-squares problem is $\hat{A} = (\bar{A}X + \bar{D})X^T(XX^T)^{-1}$ in the absence of the input sequence $\{u_i\}_{i=0}^{T-1}$. Thus, the estimation error is $\|\bar{D}X^T(XX^T)^{-1}\|$, which is non-zero and arbitrarily large in the presence of arbitrarily large disturbance vectors. A similar calculation can be made in the presence of an input sequence. Consequently, the least-squares estimator cannot achieve a zero estimation error, leading to a plateau in the estimation error of the least-squares estimator in our numerical experiments in Section

6. We define the matrix $D := [d_0; \cdots; d_{T-1}]$ with its columns being estimated disturbances, as well as the norms of matrices $\|D\|_{1,1} := \sum_i \|d_i\|_1$, and $\|D\|_{2,1} := \sum_i \|d_i\|_2$. To exactly recover the system matrices $\bar{A}$ and $\bar{B}$, we analyze the following convex optimization problems with non-smooth objective functions:

$$\min_{\substack{A \in \mathbb{R}^{n \times n}; B \in \mathbb{R}^{n \times m}; \\ D \in \mathbb{R}^{n \times T}}} \|D\|_{2,1} \qquad \text{(CO-L2)}$$
$$\text{s.t.} \quad x_{i+1} = Ax_i + Bu_i + d_i; \quad i = 0; \cdots; T-1;$$

and

$$\min_{\substack{A \in \mathbb{R}^{n \times n}; B \in \mathbb{R}^{n \times m}; \\ D \in \mathbb{R}^{n \times T}}} \|D\|_{1,1} \qquad \text{(CO-L1)}$$
$$\text{s.t.} \quad x_{i+1} = Ax_i + Bu_i + d_i; \quad i = 0; \cdots; T-1;$$

where the states $\{x_i\}_{i=0}^{T}$ are generated according to $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i; \quad i = 0; \cdots; T-1$. The difference between problems (CO-L2) and (CO-L1) is their objective functions. Note that these two problems are equivalent when we have a first-order system with $x_i \in \mathbb{R}; i \in 0; \cdots; T-1$. In problem (CO-L2), the sum of the ℓ_2 norm columns is analogous to the ℓ_1 norm minimization in the lasso problem. In other words, the ℓ_1 norm is applied at the group level to $\{\|d_i\|_2\}_{i=0}^{T-1}$ because the occurrence of large injections of disturbances is rare and not frequent. We highlight that the vectors $\{\|\bar{d}_i\|_2\}_{i=0}^{T-1}$ are not necessarily sparse. On the other hand, the ℓ_1 norm is applied both at the group level and the in-group levels to $\{\|d_i\|_2\}_{i=0}^{T-1}$ for problem (CO-L1). For those applications that the disturbance vectors can be assumed to be sparse, (CO-L1) is more suitable than (CO-L2). Furthermore, the states x_i are correlated to each other due to the system dynamics, which makes the non-asymptotic analysis of the problem more challenging than the robust regression literature for which the samples are assumed to be independently generated. One can write the optimization problems (CO-L2) and (CO-L1) as follows using the ℓ_2 and ℓ_1 norms, respectively:

$$\min_{A \in \mathbb{R}^{n \times n}; B \in \mathbb{R}^{n \times m}} \sum_{t=0}^{T-1} \|x_{t+1} - Ax_t - Bu_t\|_{\ell} :$$

This is equivalent to an empirical risk minimization problem for which the loss function is the ℓ_1 or ℓ_2 norm, depending on the choice of ℓ. Although these types of sum-of-norm minimization non-smooth loss functions are utilized in other applications, this paper marks the first non-asymptotic analysis of these loss functions in the context of control and system identification with serially correlated data.

We remark that classical statistical theory on empirical risk minimization is not applicable to the problem under study in this paper due to the correlated data at each time instance. By representing the data points X_t as tuples $(x_{t+1}; x_t; u_t)$, it is impossible to claim that X_t and X_{t+1} are independent, which is a key assumption in the empirical risk minimization literature. As the first step of our proof technique, the Karush-Kuhn-Tucker (KKT) conditions will be used to analyze the properties of these estimators. Since (CO-L2) and (CO-L1) are convex optimization problems with linear equalities, the KKT conditions are necessary and sufficient to guarantee optimality, as stated below.

Theorem 1: Consider the convex optimization problems (CO-L2) and (CO-L1) and let $\ell \in \{1; 2\}$. Given a pair of matrices $(\hat{A}; \hat{B})$, if the following conditions hold simultaneously

$$0 \in \sum_{i \in K} x_i \left\| (\bar{A} - \hat{A})x_i + (\bar{B} - \hat{B})u_i \right\|_{\ell} + \sum_{i \in K^c} x_i \left\| (\bar{A} - \hat{A})x_i + (\bar{B} - \hat{B})u_i + \bar{d}_i \right\|_{\ell}; \quad (1)$$

$$0 \in \sum_{i \in K} u_i \left\| (\bar{A} - \hat{A})x_i + (\bar{B} - \hat{B})u_i \right\|_{\ell} + \sum_{i \in K^c} u_i \left\| (\bar{A} - \hat{A})x_i + (\bar{B} - \hat{B})u_i + \bar{d}_i \right\|_{\ell}; \quad (2)$$

then $(\hat{A}; \hat{B})$ is a solution to (CO-L1) when $\ell = 1$ and a solution to (CO-L2) when $\ell = 2$.

The proof for the KKT conditions when $\ell = 2$ is provided in [26], and the proof for the case $\ell = 1$ can be done similarly. We will utilize the conditions above to study in what scenarios the exact recovery is achievable. As a simple corollary to Theorem 1, we can state that $(\bar{A}; \bar{B})$ is a solution to our estimator(s) if the following conditions hold:

$$0 \in \sum_{i \in K} x_i \|0\|_{\ell} + \sum_{i \in K^c} x_i \|\bar{d}_i\|_{\ell};$$
$$0 \in \sum_{i \in K} u_i \|0\|_{\ell} + \sum_{i \in K^c} u_i \|\bar{d}_i\|_{\ell}:$$

IV. Autonomous Systems

In this section, we consider autonomous systems, meaning that $u_0 = \cdots = u_{T-1} = 0$. Therefore, the system dynamics could be written as $x_{i+1} = \bar{A}x_i + \bar{d}_i$ for $i = 0; \cdots; T-1$. Throughout this section, we assume that the system is stable and that it is initialized at the origin.

Assumption 1: Given an autonomous system $x_{i+1} = \bar{A}x_i + \bar{d}_i$ for $i = 0; \cdots; T-1$ with dimension n, assume that $x_0 = 0$ and all eigenvalues of $\bar{A}$ are inside the unit circle.

The stability assumption is standard in system identification problems to avoid an unbounded growth of the states during the learning process. Without loss of generality, we initialize the trajectories at the origin since an initialization at other points affects the results only with a constant factor. We study noiseless systems under an adversary to obtain exact recovery results, meaning that if there is no attack at time i, $i \in K^c$, then $\bar{d}_i = 0$.

In the noisy case, one can consider the following setup. If there is no attack at time i, $i \in K^c$, then $\bar{d}_i$ is likely non-zero with a small variance and its value is independent of those for other time periods. If there is an attack at time $i \in K$, then $\bar{d}_i$ is a combination of two terms: a small noise vector that is Gaussian and independent of past time periods, and a large noise vector that could have an arbitrary distribution and possibly be dependent on past time instances. The noisy case, where the system is subjected to small independent and identically distributed Gaussian errors due to measurements and modeling errors, in addition to the adversarial vectors, can be easily addressed using our framework. The perturbation analysis allows us to bound how far the recovered solution is

from the true solution in terms of the values of small noise vectors.

Therefore, we only study the noiseless case as described above. Thus, we are interested in recovering the system matrix $\bar{A}$ using the following convex optimization problems for autonomous systems:

$$\min_{A\in\mathbb{R}^{n\times n},\, D\in\mathbb{R}^{n\times T}} \sum_{i=0}^{T-1} \|d_i\|_2 \quad \text{s.t.: } x_{i+1} = Ax_i + d_i; \tag{CO-L2-Aut}$$

and

$$\min_{A\in\mathbb{R}^{n\times n},\, D\in\mathbb{R}^{n\times T}} \sum_{i=0}^{T-1} \|d_i\|_1 \quad \text{s.t.: } x_{i+1} = Ax_i + d_i. \tag{CO-L1-Aut}$$

The optimality conditions for problem (CO-L2-Aut) with $p = 2$ and problem (CO-L1-Aut) with $p = 1$ can be written as follows using Theorem 1:

$$0 \in \sum_{i\notin K} x_i \, \P\|(\bar{A} - A)x_i\| + \sum_{i\in K} x_i \, \P\|((\bar{A} - A)x_i + \bar{d}_i)\|. \tag{3}$$

As a remark, although the set of attack times K appears in the optimality conditions, this set is not known a priori to the system operator. The set is only used during the analysis of the proposed estimators to derive sufficient conditions for exact recovery.

We first consider first-order systems where $x_i, \bar{d}_i \in \mathbb{R}, i = 0, 1, \ldots, T-1$ and $\bar{A} \in \mathbb{R}$. We examine the first-order case to gain some insight into the ideas behind the proof techniques for general systems. When $n = 1$, the problems (CO-L1-Aut) and (CO-L2-Aut) are equivalent, and therefore, we only focus on (CO-L2-Aut). After establishing the optimality conditions for these problems, we will examine two types of attack structures. An attack structure refers to the pattern of attack occurrences. In other words, it involves the distribution of each time instance at which a large disturbance vector is injected into the system. Namely, we inspect the structure of the set K.

The first attack structure is a deterministic attack model for which the attacks occur at every D time period. For instance, if $D = 2$, the set K could be $\{1, 3, 5, \ldots, 2k+1\}$, meaning that an agent injects a disturbance vector into the system at every odd time instance. Later, we investigate a probabilistic attack structure where each attack may occur with probability p at each time instance, independent of the past periods. We first define the deterministic attack model, borrowed from [26].

Definition 4 (D-spaced Attack Structure): Given a positive integer $D > 2$, the disturbance sequence $\{\bar{d}_i\}_{i=0}^{T-1}$ is said to be D-spaced if for every $j \in \{0, 1, \ldots, T-D-1\}$ such that $\bar{d}_j \neq 0$, we have $\bar{d}_j = 0$, for all $j \in \{j+1, \ldots, j+D-1\}$ and $\bar{d}_{j+D} \neq 0$. In addition, for $i \in \{0, 1, \ldots, D-1\}$, we must have at least one non-zero disturbance vector, i.e., $\bar{d}_i \neq 0$.

We will show that the convex formulation (CO-L2-Aut) exactly recovers $\bar{A}$ in the case of D-spaced disturbance sequence with $D \geq 2$.

Proposition 1: Consider a first-order autonomous system with D-spaced disturbance sequence with $D \geq 2$. Then, the convex formulation (CO-L2-Aut) (or equivalently (CO-L1-Aut)) has the unique solution $\bar{A}$ as long as the sample complexity satisfies the inequality $T \geq D+1$.

This proposition implies that whenever there are more than $D+1$ data samples, the exact recovery is guaranteed to be achieved. Note that Proposition 1 does not make any assumption on the vector set $\{\bar{d}_i : i \in K\}$ and each element of the set could be arbitrarily large and correlated as long as they are finite. As a result, regardless of the severity of the attack, an exact recovery is guaranteed for (CO-L1-Aut) and (CO-L2-Aut). One important implication of Proposition 1 is for the case where there is D-spaced disturbance sequence with $D = 2$, meaning that half of the observations are corrupted. In the robust regression estimation literature, exact recovery is possible only if the number of attacked observations is less than half of the total observations. The main difference between robust regression and system identification problems is that the observations are correlated with each other in the latter. This enables exact recovery for the convex formulation even if half of the data is corrupted via an adversarial agent. The proof of Proposition 1 is based on the following lemma.

Lemma 2: (Theorem 1 in [26]) Consider the convex optimization problem (CO-L2-Aut). If $\sum_{i\notin K} |x_i| > \sum_{i\in K} |x_i|$, then $\bar{A}$ is the unique solution to the problem.

The proof of Lemma 2 is based on the KKT conditions of the problem provided earlier. A natural question arises as to whether one can generalize the above result to higher-order systems. The next proposition extends Proposition 1 to autonomous dynamical systems with an arbitrary order n under a D-spaced disturbance sequence with $D \geq n+1$.

Proposition 2: Consider an autonomous system of order n under a D-spaced disturbance sequence with $D \geq n+1$. Suppose that $\bar{A}$ is diagonalizable with eigenvalues $\bar{\lambda}_i, i = 1, 2, \ldots, n$, and that the condition

$$\bar{d}_{i+D} \in \operatorname{span}\{\bar{d}_i, \bar{A}\bar{d}_i, \ldots, \bar{A}^{D-2}\bar{d}_i\}, \quad \forall i = 0, 1, \ldots, T-1 \tag{4}$$

is satisfied. Then, $\bar{A}$ is a solution to the convex formulation (CO-L2-Aut) if $T \geq n+D$, provided that

$$\sum_{k_1+\cdots+k_n=D} |\bar{\lambda}(k_1, \ldots, k_n)| > \sum_{t=0}^{D-n-1} \sum_{k_1+\cdots+k_n=t} |\bar{\lambda}(k_1, \ldots, k_n)|, \tag{5}$$

where the notation $\bar{\lambda}(k_1, \ldots, k_n)$ denotes $\bar{\lambda}_1^{k_1}\bar{\lambda}_2^{k_2}\cdots\bar{\lambda}_n^{k_n}$.

This result is a generalization of Proposition 1, and we do not require all the eigenvalues of $\bar{A}$ to lie inside the unit circle (i.e., it allows the violation of Assumption 1). The condition (4) is necessary to ensure that the KKT condition is satisfied, which eliminates the alignment of the attack vectors with eigenspaces of the matrix $\bar{A}$. In real-life applications, this circumstance can be avoided by injecting a small perturbation to the system. To gain insight into equation (5), which involves the product of eigenvalues, consider a special case where $\bar{A}$ has the eigenvalue $\bar{\lambda}$ with multiplicity n and n distinct

Fig. 1. Upper-Bound Value $C_{n;k}$ for Different Values of n and k.

eigenvectors. In this case, we can simplify (5) as follows. De ne $k := D \quad n$. Then, (5) is equivalent to

$$\binom{n+k \quad 1}{k} |l|^k \quad \sum_{i=0}^{k \quad 1} \binom{n+i \quad 1}{i} |l|^i < 0:$$

This condition is satis ed if $|l| \quad C_{n;k}$, where $C_{n;k}$ denotes the upper bound on the eigenvalue magnitudes given the parameters n and k. Figure 1 summarizes the values of $C_{n;k}$ for different choices of n and k. Note that $C_{n;k} \quad C_{m;k}$ if $n > m$ and $C_{n;k} \quad C_{n;l}$ if $k < l$, due to the de nition of $C_{n;k}$. It can be shown that $C_{1;k} \quad 2$ as $k \quad ¥$. As a result, $|l| \quad C_{n;k} \quad C_{1;k} \quad 2$. This shows that the stability of the system is not necessary for exact recovery when the attack vectors are injected less frequently. In addition, whenever $k = n$ or $D = 2n$, $|l| < 1$ is suf cient for exact recovery as suggested by Proposition 2. This conclusion is analogous to the stability of the system. Proposition 2 can still be applied to problem (CO-L1-Aut). However, the KKT conditions will differ due to the subdifferential of the ℓ_2 and ℓ_1 norms. In fact, they both have a similar shape. Therefore, one can show that this proposition still holds with the same condition even if convex formulation (CO-L1-Aut) with the ℓ_1 norm of the disturbance vectors is used.

It is natural to ask whether it is possible to learn the system when there is more corrupted data than clean data. We cannot use a D-spaced disturbance sequence model because the minimum value of D is 2, which does not allow the size of corrupted data to exceed the size of clean data. Thus, we investigate a probabilistic attack structure. In this structure, a non-zero disturbance vector $\bar{d}_i$ is injected into the system at time instance i with probability $p > 0$, which is independent of the past and future time periods. To address this, we consider a probabilistic attack model where there is a parameter p specifying the probability of an attack at each time instance. Speci cally, given a time instance i, $\bar{d}_i$ is non-zero with probability p, and this is independent of all previous and future time instances. As a result, the event of having an attack at each time instance is identically and independently distributed with a Bernoulli distribution with parameter p. Nevertheless, the attack vectors are still allowed to be correlated with each other. Our goal is to discover the properties of (CO-L1-Aut) and (CO-L2-Aut) for an arbitrary value of p, especially $p > 0.5$. We make the following stealth attack assumption.

Assumption 2: For each $k \quad 2 \quad K$, the attack vector is de ned by

$$\bar{d}_k := \bar{\ell}_k \bar{f}_k; \quad \text{where } \bar{\ell}_k \quad 2 \quad R \text{ and } \bar{f}_k \quad 2 \quad S_2(1):$$

where $\bar{f}_k$ plays the role of the direction of the attack while $\bar{\ell}_k$ plays the role of the length (that is allowed to take negative values too). De ne the ltration

$$F_k := s f x_1; : : : ; x_k g; \quad 8k \quad 2 f 0; : : : ; T \quad 1g:$$

For all $k \quad 2 \quad K$, conditioning on F_k, the following statements hold:

1) $\bar{\ell}_k$ is independent from the direction $\bar{f}_k$;
2) The direction $\bar{f}_k$ obeys the uniform distribution on $S_2(1)$;
3) $\bar{\ell}_k$ is mean-zero and sub-Gaussian with parameter s;
4) The variance of $\bar{\ell}_k$ is $s_k^2 \quad 2 [c^2 s^2; s^2]$ for some constant $c > 0$.

Under the stealth assumption, the length $\bar{\ell}_k$ can depend on the previous attacks $\bar{d}_{k'}$, and in particular $\bar{\ell}_{k'}$ and $\bar{f}_{k'}$ for $k' < k$. In addition, we note that the above assumption of symmetry of the disturbance vectors re ected in $\bar{f}_k$ is not restrictive and corresponds to stealth attacks. If this assumption does not hold, the attacks may be detectable, and their effects could be nulli ed, or the system could be stopped to investigate the possible in uence from outside agents. For an attack to be stealthy, its value should be zero on expectation, and our assumption has a similar avor. If the symmetric assumption does not hold, it has been shown that there is a bias in estimation, and there is no way to avoid this bias [27]. In the special cases when the length distribution is Gaussian or bounded, the constant c is equal to 1. Furthermore, we mention that the uniform distribution assumption of $\bar{f}_k$ can be relaxed to an arbitrary distribution on the sphere with zero mean and full-rank covariance matrix. In that more general case, the sample complexity in Theorems 2-5 will depend on the conditional number of the covariance matrix, which is equal to 1 under Assumption 2.

Since the KKT conditions include random variables and random sets due to the randomness in the attack structure, it is not possible to obtain deterministic sample complexity for exact recovery as in Proposition 2. Therefore, it is essential to quantify the required number of samples for exact recovery with high probability using non-asymptotic analysis. Under Assumption 2, the attack vector at time i, $\bar{d}_i$, has a sub-Gaussian distribution with parameter s given F_i, as described in Assumption 2. The sub-Gaussianity assumption does not specify the distribution of the disturbance vector but assures that the disturbance vectors have light tails. For instance, any distribution over a bounded space is sub-Gaussian, making this assumption extremely mild. As a result, the sub-Gaussian assumption is not restrictive.

The KKT conditions for exact recovery, which are necessary and suf cient, can be restated as

$$9g_i \quad 2 \quad ¶k0k ; \quad 8i \quad 62K \quad s:t: \sum_{i62K} x_i \quad g_i = \sum_{i2K} x_i \quad ¶k\bar{d}_i k :$$

because of the properties of the subdifferentials at the origin. In order to simplify the analysis, we use the relationship between the unit balls of the $\ell_¥$ and ℓ_2 norms, that is $\frac{1}{\sqrt{n}} B_¥(1) \quad B_2(1)$. Additionally, we examine the results for each coordinate of the subdifferentials since they are separable due to the properties of the $\ell_¥$ norm. Therefore, the following propositions provide suf cient conditions to satisfy the KKT conditions.

Proposition 3: The KKT conditions for the problem (CO-L2-Aut) and (CO-L1-Aut) are satis ed if there exist

scalars $g_i^l \in [-1,1]; i \in \mathcal{K}, l = 1,\dots,n$ such that

$$\sum_{i \in \mathcal{K}} g_i^l x_i = \sqrt{\bar{n}} = \sum_{i \in \mathcal{K}} \P\|d_i - \bar{k}\|_2^l x_i, \quad \forall l = 1,\dots,n \tag{6}$$

and

$$\sum_{i \in \mathcal{K}} g_i^l x_i = \sum_{i \in \mathcal{K}} \P\|d_i - \bar{k}\|_1^l x_i, \quad \forall l = 1,\dots,n, \tag{7}$$

respectively. Here $\P\|d_i - \bar{k}\|^l$ is the l-th element of the subgradient.

Because analyzing the conditions (6) and (7) directly is cumbersome, we investigate the equivalent condition provided in the lemma below, derived using Farkas' lemma [28] and the duality of linear programs.

Lemma 3: Given a matrix $F \in \mathbb{R}^{n \times m}$ and the vector $g \in \mathbb{R}^n$, the following statements are equivalent:

i) There exists a vector $w \in \mathbb{R}^m$ with $\|w\|_\infty \leq 1$ satisfying $Fw = g$.
ii) For every $z \in \mathbb{R}^n$ with $\|z\|_2 = 1$, it holds that $f(z) := z^T g + \|z^T F\|_1 \geq 0$.

It is important to notice that the conditions (6) and (7) amount to finding a vector for the set of equations in the form of $Fw = g$ where w is restricted as $\|w\|_\infty \leq 1$. Given a coordinate l, the matrix $F \in \mathbb{R}^{n \times (T-K)}$ associated with the conditions (6) and (7) is a matrix with columns $\frac{x_i}{\sqrt{\bar{n}}}$ and x_i, and the vector $g \in \mathbb{R}^n$ is $\sum_{i \in \mathcal{K}} \P\|d_i - \bar{k}\|_2^l x_i$ and $\sum_{i \in \mathcal{K}} \P\|d_i - \bar{k}\|_1^l x_i$, respectively. Moreover, the vector $w \in \mathbb{R}^{T-K}$ has the elements $g_i^l; i \in \mathcal{K}$ for both conditions. Hence, we study the second statement in Lemma 3. We use the union bound to study the satisfaction of this condition. However, there are infinitely many points inside the ℓ_2 unit ball $B_2(1)$. In order to show that the function $f(z) = z^T g + \|z^T F\|_1$ is non-negative at every point inside the ℓ_2 unit ball, we employ the discretization technique that uses a finite set of points. The set of such points is called the cover of the unit ball.

Definition 5 (Covering Number [25]): Let (T, r) be a compact metric space with a set T and a norm operator r. An ϵ-cover of the set T with respect to the norm r is a set $\{q^1, q^2, \dots, q^N\} \subseteq T$ such that for each $q \in T$, there exists some $i \in \{1,\dots,N\}$ such that $r(q, q^i) \leq \epsilon$. The ϵ-covering number $N(\epsilon, T, r)$ is the cardinality of the smallest ϵ-cover.

Given a $\epsilon > 0$, the logarithm of the covering number of the unit ball or the metric entropy of the unit ball can be upper bounded using the volumetric arguments of the balls. Indeed, the number of balls exceeding $\exp\{n \log(1 + 2/\epsilon)\}$ is sufficient to cover the unit ball with balls of radius ϵ.

Lemma 4 (Covering Number of the Unit Ball [25]): Given an n-dimensional unit ball $B(1)$ with the norm $\|\cdot\|$,

$$B(1) = \{x \in \mathbb{R}^n : \|x\| \leq 1\},$$

the logarithm of the ϵ-covering number, i.e., the metric entropy of the unit ball, can be upper bounded by

$$\log N(\epsilon, B(1), \|\cdot\|) \leq n \log\left(1 + \frac{2}{\epsilon}\right).$$

We show that the function $f(z)$ can be lower bounded by some positive number $q > 0$ at every point in the ϵ-cover of the unit circle with high probability, and that the function value inside the ϵ-ball does not change more than this positive number q with high probability. Thus, $f(z)$ must be non-negative at every point of the unit circle with high probability. Utilizing this idea, the next theorem shows that the required number of samples for the exact recovery grows with n^2 and $(1-p)^{-2}$ for the general systems of order n.

Theorem 2: Consider an autonomous system of order n under a probabilistic attack model with frequency p. Suppose that Assumptions 1 and 2 hold. Then, for all $d \in (0,1]$, if the time horizon satisfies $T \geq Q(T_{\text{sample}})$, where T_{sample} is defined as

$$n^2 R \log\left(\frac{nR}{d}\right),$$

and

$$R := \max\left(\frac{\log(1/c)}{nc^4 p(1-p)\log(1/r)}, \frac{\log^2(1/c)}{c^{10}(1-p)^2(1-r)^3\log^2(1/r)}, \frac{1}{np(1-p)}\right),$$

with $\bar{r}$ denoting the largest magnitude of the eigenvalues of $\bar{A}$, then $\bar{A}$ is a solution to the convex optimization (CO-L2-Aut) with probability at least $1-d$.

An implication of the above theorem is that even when p is large (e.g., $p > 0.5$) corresponding to the system being under attack frequently, exact recovery of the system dynamics is still possible as long as the time horizon is above the threshold. Similar results can be obtained if one prefers to use problem (CO-L1-Aut) to recover the system matrix $\bar{A}$.

Theorem 3: Under the same assumptions as in Theorem 2, if the time horizon T satisfies $T \geq Q(T_{\text{sample}})$, where T_{sample} is defined as

$$nR \log\left(\frac{nR}{d}\right),$$

and R is defined in Theorem 2, then $\bar{A}$ is a solution to the convex optimization (CO-L1-Aut) with probability at least $1-d$.

The proof of Theorem 3 is highly similar to that of Theorem 2 and therefore, it is omitted. Because the conditions (6) and (7) differ by a factor of $\sqrt{\bar{n}}$, the sample complexity results in those theorems differ by a factor of n.

The required amount of data increases with the value $(1-p)^{-2}$ and the order of the system n. Hence, as p and n increase, the number of samples for exact recovery with high probability grows. The results on sample complexity are intuitive: as the probability of having an attack increases, a larger time horizon is required for exact recovery. We note that the dependence on p $(1-p)^{-1}$ is an artifact of the high probability bound. More specifically, this dependence guarantees that the number of attacks is bounded by $Q(pT)$ with high probability. In addition, if the system is at the verge of instability with eigenvalues close to the unit circle, the sample complexity increases significantly. Even in the case when the probability p is close to 1, resulting in significantly more corrupt data than clean data, this result guarantees asymptotic exact recovery as long as there are a sufficient number of clean samples.

Last but not at least, due to the logarithmic probability bound and the Borel-Cantelli lemma, Theorems 2 and 3 imply almost sure asymptotic convergence as a corollary. Almost sure convergence of random variables implies the convergence in probability and convergence in distribution for a sequence of random variables. Almost sure convergence of random variables is defined as below for completeness.

Definition 6 (Almost Sure Convergence): A sequence of random variables $X_1, X_2, X_3, \ldots$ converges to X almost surely if

$$P\left[\lim_{n\to\infty} X_n = X\right] = 1.$$

The following corollary states that the sequence of estimators over time converges to the true system matrices almost surely.

Corollary 1: Under the same assumptions as in Theorem (2), $\bar{A}$ is almost surely a solution of convex formulations (CO-L2-Aut) and (CO-L1-Aut) when T goes to infinity.

V. Systems with Input Sequence

It is desirable to understand the role of an input sequence in exact recovery because the majority of dynamical systems are controlled by an external input. Since the input sequence is generated by a controller, one can design it in such a way that it accelerates the exact recovery. In the non-autonomous case, the system dynamics is given as $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i; i = 0, \ldots, T-1$, where $\bar{A} \in \mathbb{R}^{n\times n}$ and $\bar{B} \in \mathbb{R}^{n\times m}$. Similar to the autonomous case, the true system matrices $\bar{A}$ and $\bar{B}$ are not known and the goal is to obtain these matrices using the state trajectories and the sequence of inputs. Unlike the disturbance vectors $\{\bar{d}_i; i \in \{0, \ldots, T-1\}\}$, the sequence of system states $\{x_i; i \in \{0, \ldots, T\}\}$, and the sequence of the inputs $\{u_i; i \in \{0, \ldots, T-1\}\}$ are known. We will investigate the estimators (CO-L2) and (CO-L1) defined earlier.

We choose the input vectors u_i to be Gaussian given F_i. This allows us to obtain a high-probability bound for the exact recovery of the matrices $\bar{A}$ and $\bar{B}$. A random input sequence is commonly used in system identification and online learning because it enables the exploration of the system to learn the system dynamics faster. The Gaussian input assumption may seem restrictive. Nevertheless, it is satisfied when u_i is designed in the linear feedback form as $u_i = Kx_i + w$. Conditioning on F_i, if the input is excited with Gaussian noise w, the input vector u_i is also Gaussian. Therefore, the most common input sequence used in optimal control satisfies this assumption. Note that the closed loop system could be written as $x_{i+1} = (\bar{A} + \bar{B}K)x_i + \bar{B}w + \bar{d}_i$. Thus, the problem is equivalent to estimating the matrices $(\bar{A} + \bar{B}K)$ and $\bar{B}$ when the linear feedback control is used.

The KKT conditions for the exact recovery that are both necessary and sufficient can be restated as

$$\exists g_i \in \|\cdot\|_0 \quad \forall\, i \in \mathcal{K} \quad \text{s.t.} \sum_{i\in\mathcal{K}} x_i \quad g_i = \sum_{i\in\mathcal{K}} x_i \quad \|\bar{d}_i\|$$

and

$$\exists m_i \in \|\cdot\|_0 \quad \forall\, i \in \mathcal{K} \quad \text{s.t.} \sum_{i\in\mathcal{K}} u_i \quad m_i = \sum_{i\in\mathcal{K}} u_i \quad \|\bar{d}_i\| :$$

The first set of conditions corresponds to the KKT conditions for the system states while the second set is for the KKT conditions for the input sequence. Similar to Proposition 3, the sufficient conditions can be tightened so that the equations become coordinate-wise separable.

Proposition 4: The KKT conditions for problem (CO-L2) are satisfied if there exist scalars $g_i^l, m_i^l \in [-1, 1]$ for all $i \notin \mathcal{K}, l \in \{1, \ldots, n\}$ such that

$$\sum_{i\notin\mathcal{K}} g_i^l x_i = \sqrt{n} = \sum_{i\in\mathcal{K}} \|\bar{d}_i\|_2 x_i, \quad \forall l = 1, \ldots, n, \tag{8}$$

and

$$\sum_{i\notin\mathcal{K}} m_i^l u_i = \sqrt{n} = \sum_{i\in\mathcal{K}} \|\bar{d}_i\|_2 u_i, \quad \forall l = 1, \ldots, n, \tag{9}$$

where $\|\bar{d}_i\|_2^l$ denotes the l-th element of the subgradient.

The proof of Proposition 4 is omitted because it relies on the same technique as in Proposition 3. As in the case of autonomous systems, two sets of equations that guarantee the satisfaction of the KKT conditions can be written for problem (CO-L1) by omitting the factor $\sqrt{n}$. To establish the exact recovery guarantees, we require the following controllability assumption.

Assumption 3: The ground truth $(\bar{A}, \bar{B})$ satisfies

$$\operatorname{rank}\begin{bmatrix} \bar{B} & \bar{A}\bar{B} & \cdots & \bar{A}^{n-1}\bar{B} \end{bmatrix} = n.$$

Intuitively, the controllability of a non-autonomous system denotes the ability to move a system around in its entire state space using the admissible manipulations, namely, the input sequence $\{u_t\}_{t=0}^{T-1}$. Controllability is an important property of a control system and plays a crucial role in many control problems, such as stabilization of unstable systems by feedback. Under the above assumption, we implement the non-asymptotic analysis of the general non-autonomous system in a similar fashion to Theorem 2 using the covering arguments and Farkas' lemma.

Theorem 4: Consider an autonomous system of order n under a probabilistic attack model with frequency p. Suppose that Assumptions 1, 3 and the first three conditions in Assumption 2 hold. Assume also that the input vectors $u_i | F_i$ are selected to be independent from the attack vectors and obey the Gaussian distribution $\mathcal{N}(0, \frac{x^2}{m} I_m)$. For all $\delta \in (0, 1]$, let

$$T^1_{\text{sample}} := n^2 R_1 \log\left(\frac{nR_1}{\delta}\right)$$

and

$$T^2_{\text{sample}} := nmR_2 \log\left(\frac{nR_2}{\delta}\right);$$

where

$$R_1 := \max\left(\frac{\log(k/c)}{nc^4\log(1/r)}, \frac{pk^2}{c^{10}(1-p)^2(1-r)^2}, \frac{pk^2\log^2(k/c)}{c^{10}(1-r)^2\log^2(1/r)}, \frac{1}{np}\right)$$

$$R_2 := \max\left(\frac{1}{np}, \frac{p}{(1-p)^2}, \frac{m}{n}\right) :$$

Here, constants $s \in (0,1]$ and $k = (1-r)^{-1}$ depend on m, n, s, x and $\bar{B}$. If the time horizon T satisfies the inequality $T \geq Q[\max\{T^1_{\text{sample}}, T^2_{\text{sample}}\}]$, then $(\bar{A}, \bar{B})$ is a solution to (CO-L2) with probability at least $1-d$.

We have obtained a high probability bound for the exact recovery of the system matrices $\bar{A}$ and $\bar{B}$. The first term in the sample complexity corresponds to the satisfaction of the KKT conditions for the state measurements $\{x_i\}_{i=0}^{T}$, whereas the second term corresponds to the satisfaction of the KKT conditions for the input sequence $\{u_i\}_{i=0}^{T-1}$. Similar to the case of autonomous systems, the sample complexity increases as the probability of disturbances increases. Because there is a logarithmic dependence on the satisfaction of the probability bound, Theorem 4 and the application of the Borel-Cantelli lemma imply almost sure asymptotic convergence to the correct matrices $\bar{A}$ and $\bar{B}$. The sample complexity T^2_{sample} is needed to satisfy the KKT conditions associated with on the input sequence. Compared with the previous theorems for the autonomous case, we require a sample complexity that scales with $p/(1-p)^2$ and terms depending on the spectral norm of $\bar{A}$. The introduction of the input sequence removes the requirement on the variance of the attack vectors. In addition, the dependence of the sample complexity on p is improved from $1/(1-p)^2$ to $p/(1-p)^2$. Moreover, the dependence on the spectrum of $\bar{A}$ is reduced from $1/[(1-r)^3 \log^2(1/r)]$ to $1/[(1-r)^2 \log^2(1/r)]$. Finally, we mention that the dependence on $1/(np)$ is also to guarantee that the number of attacks is bounded by $Q(pT)$ with high probability.

The following theorem studies problem (CO-L1).

Theorem 5: Under the assumptions of Theorem 4, for all $d \in (0,1]$, let T^1_{sample} and T^2_{sample} be defined as

$$nR_1 \log\left(\frac{nR_1}{d}\right) \quad \text{and} \quad mR_2 \log\left(\frac{nR_2}{d}\right);$$

where R_1 and R_2 are given in Theorem 4. If the time horizon T satisfies the inequality $T \geq Q[\max\{T^1_{\text{sample}}, T^2_{\text{sample}}\}]$, then $(\bar{A}, \bar{B})$ is a solution to (CO-L1) with probability at least $1-d$.

As expected, even if more than half of the data are corrupted, that is $p > 1/2$, the exact recovery is still attainable with high probability. We note that when the input sequence $u_i = Kx_i$ is used to control the system, this input sequence satisfies the assumptions in the above theorems if x_i are sub-Gaussian. The closed-loop system with the matrix $(\bar{A} + \bar{B}K)$ results in a second solution $\hat{A} = \bar{A} + \bar{B}K$ and $\hat{B} = 0$. Nevertheless, the ground-truth system matrix pair $(\bar{A}, \bar{B})$ is also a solution to our estimators. This phenomenon occurs due to the existence of multiple optimal solutions and it could be avoided if the input is excited with a small noise in the form of $u_i = Kx_i + w$. Moreover, if all the input vectors u_i are set to zero, it is not possible to uniquely recover the system matrix $\bar{B}$. Nevertheless, because the input sequence $\{u_i\}_{i=0}^{T-1}$ is zero, the KKT conditions are trivially satisfied. Therefore, the estimators have multiple optimum solutions where $\bar{A}$ and 0 matrices are possible solutions among all optimum solutions.

VI. Numerical Experiment

We conduct a numerical experiment inspired by biomedical applications to demonstrate the results of this paper. We consider a compartmental model of blood sugar and insulin dynamics in the human body, as described in [29]. Accurately estimating the parameters of the dynamics is crucial when regulating the blood sugar level through the injection of a bolus of insulin into the system. Due to the complex structure of the human body, the dynamics vary among individuals. We consider a linear system based on Hovarka's model as follows [30]:

$$\begin{aligned}
\dot{x}_1 &= -k_{a1}x_1 - k_{b1}I + d_1;\\
\dot{x}_2 &= -k_{a2}x_1 - k_{b2}I + d_2;\\
\dot{x}_3 &= -k_{a3}x_1 - k_{b3}I + d_3;\\
\dot{S}_1 &= -S_1/t_{\max,I} + d_4;\\
\dot{S}_2 &= S_1/t_{\max,I} - S_2/t_{\max,I} + d_5;\\
\dot{I} &= S_2/(t_{\max,I}V_I) - k_e I + d_6;
\end{aligned}$$

where given a time-dependent variable $z(t)$, $\dot{z}(t)$ represents its derivative with respect to time t. The states x_1, x_2, x_3 represent the influence of insulin on the system of the body. S_1 and S_2 represent the absorption rate of insulin in the, directly and indirectly, accessible compartment models, respectively. Lastly, the state I represents the blood sugar level in the body. The disturbance d_4 corresponds to the bolus injection into the body, while the remaining disturbance vectors model sudden changes in the body due to diseases such as diabetes. Although the injected insulin amount could be known, the exact amount of insulin and its timing reaching the effective body parts are unknown. Hence, the d_i values are treated as unknown. Even though the disturbance in this application is not a malicious attack, it exhibits similar characteristics for identification purposes: the arrival time of the bolus is unknown, and once it arrives, it has a large magnitude.

In this experiment, we discretize the continuous-time system to obtain an LTI system using $D_t = 0.5$. The resulting matrix $\bar{A}$ is stable. Our objective is to estimate the parameters $(k_{ai}, k_{bi}, t_{\max,I}, V_I, k_e)$, where the true values are obtained from Table 1 in [31]. We model the attack vectors given the historical data as zero-mean Gaussian random vectors with an identity covariance matrix with variance 10. Thus, the attack vectors are conditionally independent, although they are dependent. We run our model with the probability of an attack being $p = 0.2$, $p = 0.4$, and $p = 0.6$. We report the estimation error $\|\hat{A} - \bar{A}\|_F$ for the least-squares estimator, problem (CO-L2), and problem (CO-L1).

Figure 2 suggests that our proposed estimators attain exact recovery while the least-squares estimator fails to do so. As the probability of having an attack p increases, the number of required time periods for exact recovery grows proportionally to $p/(1-p)^2$. Note that there are more corrupted data than clean data in the case of $p = 0.6$. Additionally, because there is no sparsity assumption on the attack vectors, (CO-L2) performs slightly better than (CO-L1).

We compare the performance of (CO-L2) and (CO-L1) by running a similar experiment with and without sparse

Fig. 2. Estimation errors for Least-Squares, (CO-L2), and (CO-L1) with attack probability of $p = 0.2, 0.4, 0.6$ (left-to-right).

Fig. 3. Estimation errors for Least-Squares, (CO-L2), and (CO-L1) with attack probability $p = 0.6$ not Sparse d (top) Sparse d (bottom).

disturbances. When the disturbances are sparse, d_1, d_2, d_3, d_5 are set to zero while d_4 and d_6 have the same Gaussian distribution as before. Figure 3 shows that the two methods perform similarly when the attack vectors are also sparse.

VII. Discussion and Conclusion

We investigated the problem of learning LTI systems under adversarial attacks by studying two lasso-type estimators. We considered both deterministic and probabilistic attack models regarding the time occurrence of the attack and developed strong conditions for the exact recovery of the system dynamics. When the attacks occur deterministically every D period, exact recovery is possible after $n+D$ time steps. Moreover, if the system is attacked at each time instance with probability p, the system matrices are recovered with high probability when T is on the order of $O((1-p)^2)$ and a polynomial in the dimension of the problem. Similar results were obtained when the system is controlled by an input sequence. These findings were supported by a numerical experiment in biology that to validate the non-asymptotic analytic results. This work provides the first set of mathematical guarantees for the robust non-asymptotic analysis of dynamic systems.

Since our estimators have non-smooth objective functions, closed-form solutions to the optimization problem are not obtainable. We did not provide any specific numerical algorithm to solve the provided estimation problems. However, both (CO-L2) and (CO-L1) are convex optimization problems, allowing the use of the subgradient descent algorithm to obtain these estimators. It is a well-established result that the subgradient algorithm has a convergence rate on the order of $\frac{1}{\sqrt{k}}$, where k is the iteration update number. Although the algorithm is considered fast, one possible extension of this work would be to design an algorithm to predict and update $(\hat{A}_{t+1}, \hat{B}_{t+1})$ using the latest estimation $(\hat{A}_t, \hat{B}_t)$ and the new data (x_{t+1}, u_t), instead of solving the problem from scratch at each time period. Initial experiments hinted that a single subgradient update at each iteration using the new information, (x_{t+1}, u_t), asymptotically converges to the true system matrices. We leave the analysis of this algorithm and online control of dynamic systems under adversaries as future work.

Appendix

A. Proofs for Results in Main Part

1) Proof of Proposition 1: The proof of Proposition 1 is established based on Lemma 2 defined in the paper. Let $i_1, i_2, \ldots$ be the set of attack times over time horizon T. Therefore, $K = \{i_1, i_2, \ldots\}$. Due to D-spaced attack model, the first attack time must be smaller than D, i.e., $i_1 \leq D$. Since $x_0 = 0$, we have $x_t = 0$ for $t = 0, 1, \ldots, i_1$. Define $\mathcal{N}$ as the set of natural numbers. We can utilize Lemma 2 to show that the unique solution. Using these facts, we can decompose the sum of the magnitudes of the states at non-attack times as

$$\sum_{i \notin K} |x_i| = \sum_{i \notin K,\, i > i_1} |x_i| = \sum_{i \in K'} |x_i| + \sum_{i \in K''} |x_i|;$$

where $K^{c+} = Nn(K \setminus [f\ 0;1;\dots;i_1 \quad 1g)$, $K^{0} = K^{c+} \setminus K^{00}$, and $K^{00} = f i_2 \quad 1;i_3 \quad 1;\dots g$. The second term on the right-hand side is the sum of magnitudes at the time step just before the attack while the first term covers the rest of the magnitudes of the states. In addition, the magnitudes of the states at attack times can be written as

$$\sum_{i2K} |x_i| = \sum_{i2K;\, i \neq i_2} |x_i| = \sum_{i2K^{00}} |\bar{A}x_i| = \sum_{i2K^{00}} |\bar{A}||x_i|:$$

The second equality follows from the fact that $x_{i_k} = \bar{A}x_{i_k 1}$ due to lack of attack. We compare the sum of the magnitudes of the states at attack times for the non-attack times to check if the condition in Lemma 2 holds:

$$\begin{aligned}\sum_{i2K} |x_i| \quad \sum_{i2K} |x_i| &= \sum_{i2K^{0}} |x_i| + \sum_{i2K^{00}} |x_i| \quad \sum_{i2K^{00}} |\bar{A}||x_i| \\ &= \sum_{i2K^{0}} |x_i| + (1 \ j\bar{A}j) \sum_{i2K^{00}} |x_i| > 0: \quad (10)\end{aligned}$$

Note that the term $\sum_{i2K} |x_i|$ becomes positive at time period $i_1 + 1$ while $\sum_{i2K} |x_i|$ is positive first time at time step i_2. Consequently, the strict inequality for (10) holds for every time step after i_1 because $(1 \ j\bar{A}j) > 0$ by assumption. As a result, we have a unique and exact recovery for every time period $T \quad D+1 \quad i_1 + 1$.

2) Proof of Proposition 2: By using (3), the necessary and sufficient condition for this problem is

$$0 2 \sum_{i2K} x_i \quad \P k(\bar{A} \quad A)x_i k_2 + \sum_{i2K} x_i \quad \P k(\bar{A} \quad A)x_i + \bar{d}_i k_2:$$

Then, $\bar{A}$ is a solution to the problem if and only if

$$0 2 \sum_{i2K} x_i \quad \P k0k_2 + \sum_{i2K} x_i \quad \P k\bar{d}_i k_2: \quad (11)$$

Let i_1 be the time stamp of the first attack time. Then, we have $i_1 2 f 1;\dots;Dg$. The set of attack times is $K = f i_1; i_1 + D; i_1 + 2D; i_1 + 3D;\dots g$. Since $x_0 = 0$, we have $x_t = 0$ whenever $t = 0;1;\dots;i_1$ and $x_{i_1+1} = \bar{d}_{i_1}$. Let $T = D+ i_1$, i.e., the time step at which a cycle of disturbance is completed. In this case, the sufficient condition (3) can be written as

$$\begin{aligned}0 2 \sum_{t=1}^{D \ 1} x_{i_1+t} \quad \P k0k_2 + x_{i_1+D} \quad \P k\bar{d}_{i_1+D}k_2 \\ = \sum_{t=0}^{D \ 2} \bar{A}^t\bar{d}_{i_1} \quad \P k0k_2 + \bar{A}^{D \ 1}\bar{d}_{i_1} \quad \frac{\bar{d}_{i_1+D}}{k\bar{d}_{i_1+D}k}:\end{aligned}$$

The matrix 0 may belong to the right-hand side term for arbitrary $\bar{d}_{i_1+D}$ if $\bar{d}_{i_1+D} 2 \text{span} f\bar{d}_{i_1}; \bar{A}\bar{d}_{i_1};\dots;\bar{A}^{D \ 2}\bar{d}_{i_1} g$. This is satisfied by the assumption in the proposition statement.

However, this is not sufficient to ensure that KKT condition (3) holds. The reason is that $\P k0k_2 = f x 2 R^n : kxk_2 \quad 1g$. The vectors chosen for $\P k0k_2$ have a bounded norm. Therefore, we need a condition that bounds the norm of the columns of $\bar{A}^{D \ 1}\bar{d}_{i_1} \quad \frac{\bar{d}_{i_1+D}}{k\bar{d}_{i_1+D}k_2}$, so it can be expressed as a linear combination of the vectors $f\bar{d}_{i_1}; \bar{A}\bar{d}_{i_1};\dots;\bar{A}^{D \ 2}\bar{d}_{i_1} g$. Let $(l_j; v_j)$ be eigenvalue-eigenvector pairs for the matrix $\bar{A}$. Let $e_1;\dots;e_{D \ 1} 2 \P k0k_2$. Then, the KKT condition can be written as follows after dropping the sub-index i_2

$$0 2 e_1\bar{d}^T + e_2\bar{d}^T\bar{A}^T + \quad + e_{D \ 1}\bar{d}^T(\bar{A}^T)^{D \ 2} + f\bar{d}^T(\bar{A}^T)^{D \ 1};$$

where $f = \frac{\bar{d}_{i_1+D}}{k\bar{d}_{i_1+D}k_2}$ and $kfk_2 = 1$. If we multiply the equation above by the eigenvector v_j of $\bar{A}^T$, we obtain

$$\begin{aligned}0 2 e_1\bar{d}^T v_j + \quad + e_{D \ 1}\bar{d}^T(\bar{A}^T)^{D \ 2}v_j + f\bar{d}^T(\bar{A}^T)^{D \ 1}v_j \\ 2 (e_1 + l_j e_2 + \quad + l_j^{D \ 2}e_{D \ 1} + l_j^{D \ 1}f)\bar{d}^T v_j:\end{aligned}$$

Note that because $\bar{A}$ is diagonalizable, we only need to satisfy this condition along the direction of each eigenvector, since all eigenvectors span the whole space. Therefore, the KKT condition holds if

$$0 2 e_1 + l_j e_2 + \quad + l_j^{D \ 2}e_{D \ 1} + l_j^{D \ 1}f; \quad 8j = 1;\dots;n:$$

There are $(D \quad 1)n$ free variables and n^2 equations. One can use the substitution to eliminate n^2 variables, which leads to

$$\begin{aligned}\sum_{k_1+ \ +k_n=D \ n} l(k_1;\dots;k_n) f = \\ \sum_{t=0}^{D \ n \ 2} \sum_{k_1+ \ +k_n=t} l(k_1;\dots;k_n)e_{t+n+1}:\end{aligned}$$

Taking the norm of both sides and using the triangle inequality yields that

$$\begin{aligned}\sum_{k_1+ \ +k_n=D \ n} l(k_1;\dots;k_n) \ kfk_2 \\ \sum_{t=0}^{D \ n \ 1} \sum_{k_1+ \ +k_n=t} l(k_1;\dots;k_n) \ ke_{t+n+1}k_2:\end{aligned}$$

Using the fact that $ke_j k_2 = 1$ for all j and $kfk_2 = 1$, we obtain

$$\sum_{k_1+ \ +k_n=D \ n} l(k_1;\dots;k_n) \quad \sum_{t=0}^{D \ n \ 1} \sum_{k_1+ \ +k_n=t} l(k_1;\dots;k_n) :$$

This completes the proof for the proposition.

3) Proof of Proposition 3: The KKT condition for the exact recovery that is the necessary and sufficient condition can be restated as

$$9g 2 \P k0k_{\ ;i 62K} \text{ s.t. } \sum_{i2K} x_i \quad g_i = \sum_{i2K} x_i \quad \P k\bar{d}_i k: \quad (12)$$

For the problem (CO-L2-Aut) with $= 2$. the condition (12) becomes

$$9g 2 \P k0k_2; i 62K \text{ s.t. } \sum_{i2K} x_i \quad g_i = \sum_{i2K} x_i \quad \P k\bar{d}_i k_2:$$

Since $\frac{1}{\sqrt{n}}\P k0k_1 = \frac{1}{\sqrt{n}}B_¥(1) \quad B_2(1) = \P k0k_2$, we can rewrite it as

$$9g 2 \P k0k_1; i 62K \text{ s.t. } \sum_{i2K} \frac{x_i}{\sqrt{n}} \quad g_i = \sum_{i2K} x_i \quad \P k\bar{d}_i k_2:$$

We can check the condition at each coordinate because the set $B_¥(1)$ is coordinate wise separable. Thus, the condition becomes that KKT condition holds for (CO-L2-Aut) if there exist scalars $g_i^l 2 [\ 1;1]; i 62K \ ; l = 1;\dots;n$ such that

$$\sum_{i2K} g_i^l x_i = \sqrt{n} = \sum_{i2K} \P k\bar{d}_i k_2 x_i; \quad 8l = 1;\dots;n;$$

where $\P k\bar{d}_i k^l$ is the l-th element of the subgradient. Similar algebraic manipulation can be done for (CO-L1-Aut) except for the transforming subdifferential of the ℓ_2 norm to subdifferential of the ℓ_1 norm to obtain the second part of the result.

4) Proof of Lemma 3: The condition "Given a matrix $F \in \mathbb{R}^{n \times m}$ and the vector $g \in \mathbb{R}^n$, there exists a vector $w \in \mathbb{R}^m$ with $\|w\|_\infty \leq 1$ satisfying $Fw = g$." is equivalent to the feasibility of the linear programming (LP) below with objective function equal to 0:

$$\begin{aligned} \max_{w \in \mathbb{R}^m} \quad & 0 \\ \text{s.t.} \quad & Fw = g, \\ & \|w\|_\infty \leq 1. \end{aligned}$$

Due to the strong duality, the dual problem of the LP above must have the optimum objective value equal to 0. The dual problem can be formulated as

$$\begin{aligned} \min_{y \in \mathbb{R}^m, z \in \mathbb{R}^n} \quad & z^T g + \|y^T\|_1 \\ \text{s.t.} \quad & z^T F + y^T = 0, \end{aligned}$$

or equivalently,

$$\min_{z \in \mathbb{R}^n} \quad f(z) := z^T g + \|z^T F\|_1.$$

Thus, for any $z \in \mathbb{R}^n$, $f(z)$ must be nonnegative, i.e., $f(z) \geq 0$. Because $f(cz) = |c| f(z)$ for all $c > 0$, the condition $f(z) \geq 0$ for all $z \in \mathbb{R}^n$ is satisfied if $f(z) \geq 0$ for all $z \in \mathbb{R}^n$ such that $\|z\|_2 = 1$. This completes the proof.

5) Proof of Theorem 2: Due to the system dynamics and given $x_0 = 0$, x_i can be expressed as

$$x_i = \sum_{k \in K} \bar{A}^{(i-k-1)^+} \bar{d}_k,$$

where $A^{(i)^+}$ is defined as

$$A^{(i)^+} := \begin{cases} 0, & \text{if } i < 0 \\ I, & \text{if } i = 0 \\ A^i, & \text{if } i > 0 \end{cases}$$

By Lemma 3, given a coordinate $\bar{i} \in \{1, \ldots, n\}$, the optimality condition for the recovery of $\bar{A}$ is equivalent to

$$f(z) := z^T g + \|z^T F\|_1 \geq 0, \quad \forall z \in S_2(1), \tag{13}$$

where the unit sphere $S_2(1)$ is $\{z \in \mathbb{R}^n : \|z\|_2 = 1\}$, the matrix $F \in \mathbb{R}^{n \times (T|K|)}$ has the columns

$$F^i := \sum_{k \in K} \frac{\bar{A}^{(i-k-1)^+} \bar{d}_k}{\sqrt{n}}, \quad \forall i \in \bar{K},$$

and the vector $g \in \mathbb{R}^n$ is

$$g := \sum_{i \in K} \sum_{k \in K} \bar{A}^{(i-k-1)^+} \bar{d}_k \, \bar{f}_i^T.$$

We prove that condition (13) holds with high probability in two steps.

Step 1: We first prove that condition (13) holds with high probability for a fixed $z \in S_2(1)$.

a) Step 1-1: We first analyze the term $\|z^T F\|_1$, namely,

$$\mathbb{E}\|z^T F\|_1 = \frac{1}{\sqrt{n}} \sum_{i \in \bar{K}} \mathbb{E} \sum_{k \in K} z^T \bar{A}^{(i-k-1)^+} \bar{d}_k. \tag{14}$$

We construct the index set

$$I_1 := \{i \mid i \in \bar{K}, \; i - 1 \in K\}.$$

Let

$$\begin{aligned} S &:= \log_r Q \frac{c^5}{\log(|I_1|/d)} \\ &= Q \frac{\log\log(|I_1|/d) + \log(1/c)}{\log(1/r)}, \end{aligned}$$

where $\lceil x \rceil$ is the minimal integer that is not smaller than x and $d \in (0,1)$ is the specified probability. We construct a subset of I_1 in the following way:

$$I := \{i_1, \ldots, i_l \mid i_j \in I_1, \; i_j - i_{j-1} \leq S, \; \forall j\}.$$

It is straightforward to construct I such that

$$|I| \geq |I_1| - \frac{1}{S}|I_1|.$$

In addition, due to the probabilistic attack model, it holds with probability at least $1 - \exp[-Q[p(1-p)T]]$ that

$$|I_1| \geq \frac{p(1-p)T}{2}.$$

Therefore, we have an estimate on the size of I

$$P\left(|I| \geq \frac{p(1-p)T}{2S}\right) \geq 1 - \exp[-Q[p(1-p)T]]. \tag{15}$$

For each $j \in \{1, \ldots, |I|\}$, we define

$$K_j := \{k \in K \mid i_{j-1} < k < i_j\},$$

where we denote $i_0 := -1$. Moreover, we define

$$X_{j,\ell} := \sum_{k \in K_j} z^T \bar{A}^{\ell-k-1} \bar{d}_k, \quad \forall j, \ell \in \{1, \ldots, |I|\} \text{ s.t. } j \leq \ell.$$

Using equation (14), we can calculate that

$$\begin{aligned} \|z^T F\|_1 &\geq \frac{1}{\sqrt{n}} \sum_{\ell=1}^{|I|} \sum_{j=1}^{\ell} X_{j,\ell} \\ &\geq \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{|I|} X_{j,j} - \sum_{\ell=j+1}^{|I|} X_{j,\ell} \right). \end{aligned} \tag{16}$$

We utilize the following lemma to bound $|X_{j,\ell}|$.

Lemma 5: Suppose that a random variable X is sub-Gaussian with parameter σ_X, where the mean and the variance of X are 0 and $\tilde{\sigma}_X^2$, respectively. Then, we have

$$P(|X| \geq \tilde{\sigma}_X) \geq \frac{\tilde{\sigma}_X^4}{64 \sigma_X^4}.$$

For all $j \in \{1, \ldots, |I|\}$, the stealthy assumption (Assumption 2) implies that the standard deviation and the sub-Gaussian parameter of $X_{j,\ell}$ are

$$\begin{aligned} \tilde{\sigma}_{j,\ell} &:= \sqrt{\frac{1}{n} \sum_{k \in K_j} \|z^T \bar{A}^{\ell-k-1}\|_2^2 \sigma_k^2}, \\ \sigma_{j,\ell} &:= \sqrt{\frac{1}{n} \sum_{k \in K_j} \|z^T \bar{A}^{\ell-k-1}\|_2^2 \sigma^2}, \end{aligned}$$

respectively. It follows from Lemma 5 that

$$P(|X_{j,j}| \geq \tilde{s}_{j,j}) \leq \frac{\tilde{s}_{j,j}^4}{64 s_{j,j}^4},$$

which further leads to

$$P(|X_{j,j}| \geq c s_{j,j}) \leq \frac{c^4}{64}. \tag{17}$$

On the other hand, the sub-Gaussian parameter of $\sum_{\ell=j+1}^{l} X_{j,\ell}$ is at most

$$\sum_{\ell=j+1}^{l} s_{j,\ell} \leq \sum_{\ell=j+1}^{l} r^{(\ell-j)} s_{j,j} \leq \frac{r}{1-r} s_{j,j}.$$

Therefore, it holds with probability at least $1-d/(4l)$ that

$$\begin{aligned} \sum_{\ell=j+1}^{l} X_{j,\ell} &\leq \frac{r}{1-r} s_{j,j} \sqrt{2\log(4l/d)} \\ &\leq \frac{r}{1-r} s_{j,j} \sqrt{2\log(4|l-j|/d)} \\ &\leq \frac{c^4}{512} c s_{j,j}, \end{aligned} \tag{18}$$

where the last step is by the choice of S. Using the bound in (15), if we choose

$$T \geq \frac{\log\log(1/d) + \log(1/c)}{p(1-p)c^4 \log(1/r)},$$

it holds with high probability that

$$\frac{c^4}{64} \geq \frac{d}{4l} \geq \frac{c^4}{128}.$$

Note that we have dropped the $|l-j|$ term in the definition of S since $\log\log(|l-j|)$ is bounded by $\log\log(T)$ and will not change the order of the above bound. Let q_j be the $(1-c^4/128)$-quantile of $X_{j,j} - \sum_{\ell=j+1}^{l} X_{j,\ell}$. We define the indicator function

$$1_j := \begin{cases} 1, & \text{if } X_{j,j} - \sum_{\ell=j+1}^{l} X_{j,\ell} \geq q_j, \\ 0, & \text{otherwise,} \end{cases} \qquad \forall j \in \{1,\dots,l\}.$$

Since the value of the Bernoulli random variable 1_j only depends on attacks in K_j, which are disjoint from each other, the random variables

$$1_1 - c^4/128, \dots, 1_l - c^4/128$$

form a martingale sequence with respect to filtration $F_{i_1}, \dots, F_{i_l}$. For all $j \in \{1,\dots,l\}$, we can calculate that

$$E[\exp(s 1_j)] \leq \exp\left(\frac{c^4}{128}(e^s - 1)\right), \quad \forall s \in R.$$

By the tower property of expectation, we have

$$E\left[\exp\left(s \sum_{j=1}^{l} 1_j\right)\right] \leq \exp\left(\frac{c^4 l}{128}(e^s - 1)\right), \quad \forall s \in R.$$

Therefore, by applying Chernoff's bound and choosing $s = \log(2)$, it follows that

$$\begin{aligned} P\left(\sum_{j=1}^{l} 1_j \geq \frac{c^4}{256} l\right) &\leq \exp\left(-\frac{c^4 l}{256} s + \frac{c^4 l}{128}(e^s - 1)\right) \\ &\leq \exp\left(-\frac{c^4 l}{256}\log\frac{1}{2} - \frac{c^4 l}{128}\frac{1}{2}\right) \\ &= \exp\left(-Q\frac{c^4}{128} l\right). \end{aligned}$$

Equivalently, we know

$$P\left(\sum_{j=1}^{l} 1_j \leq \frac{c^4}{256} l\right) \geq 1 - \exp\left(-Q\frac{c^4}{128} l\right). \tag{19}$$

Furthermore, since $e_j \in K_j$, we can estimate that

$$s_{j,j} \leq \sqrt{\frac{1}{n}\|z\|_2^2 s^2} = \frac{1}{\sqrt{n}} s.$$

By the definition of q_j and 1_j, when the event in inequality (19) happens, inequalities (17) and (18) imply that

$$\begin{aligned} \|z^T F\|_1 &\leq \frac{1}{\sqrt{n}} \sum_{j=1,\dots} \left(X_{j,j} - \sum_{\ell=j+1}^{l} X_{j,\ell} \right) \\ &\leq \frac{1}{\sqrt{n}} \sum_{j=1}^{l} \left(\frac{c^4}{256} c s_{j,j} - \frac{c^4}{512} c s_{j,j} \right) \leq \frac{c^5 s}{512\sqrt{n}} l \end{aligned}$$

holds with probability at least $1-d/4$. Hence, we obtain

$$P\left(\|z^T F\|_1 \leq \frac{c^5 s}{512\sqrt{n}} l\right) \geq 1 - \exp\left(-Q c^4 l\right) - \frac{d}{4}. \tag{20}$$

b) Step 1-2: For the term $z^T g$, we can establish an upper bound on

$$\begin{aligned} &E\left[\exp\left(l z^T g\right)\right] \\ &= E\left[\exp\left(l \sum_{k \in K} \sum_{i \in K} z^T \bar{A}^{(i-k-1)+} \bar{d}_k f_i^T\right)\right]. \end{aligned}$$

Define the filtration

$$F^f := \{\sigma(\bar{f}_t); t \in K^c\}.$$

By the stealth assumption, for each $k \in K$, conditional on F_k and F^f, we have

$\bar{\ell}_k$ is sub-Gaussian with parameter s.

Let T' be the second last time instance in K. We have

$$\begin{aligned} &E\left[\exp\left(l \sum_{i \in K} \sum_{k \in K} z^T \bar{A}^{(i-k-1)+} \bar{d}_k f_i^T\right)\right] \\ &= E\left[\exp\left(l \sum_{k \in K, k < T'} \sum_{i \in K} z^T \bar{A}^{(i-k-1)+} \bar{d}_k f_i^T\right)\right] \\ &\quad E\left[\exp\left(l \sum_{i \in K} z^T \bar{A}^{(i-T'-1)+} \bar{d}_{T'}^0 f_i^T \,\middle|\, F_{T'}, F^f\right)\right]. \end{aligned} \tag{21}$$

Using the decomposition in Assumption 2, we have

$$E\left[\exp\left(\lambda \sum_{i\in K} z^T \bar{A}^{(i-1,T_0)+} \bar{d}_{T_0} \bar{f}_i^T \,\middle|\, F_{T_0}, F^f\right)\right] = E\left[\exp\left(\lambda \sum_{i\in K} z^T \bar{A}^{(i-1,T_0)+} \bar{f}_{T_0} \bar{f}_i^T \bar{\epsilon}_{T_0} \,\middle|\, F_{T_0}, F^f\right)\right]$$

$$\le \exp\left[\frac{\lambda^2\sigma^2}{2} \sum_{i\in K} z^T \bar{A}^{(i-1,T_0)+} \bar{f}_{T_0} \bar{f}_i^T\right].$$

Substituting back into (21), it follows that

$$E\left[\exp\left(\lambda \sum_{i\in K}\sum_{k\in K} z^T \bar{A}^{(i-k-1)+} \bar{d}_k \bar{f}_i^T\right)\right] \le E\left[\exp\left(\lambda \sum_{k\in K,\, k<T_0}\sum_{i\in K} z^T \bar{A}^{(i-k-1)+} \bar{d}_k \bar{f}_i^T\right)\right] \exp\left[\frac{\lambda^2\sigma^2}{2} \sum_{i\in K} z^T \bar{A}^{(i-1,T_0)+} \bar{f}_{T_0} \bar{f}_i^T\right].$$

Continuing the process for all $k \in K$, we obtain

$$E\left[\exp\left(\lambda \sum_{i\in K}\sum_{k\in K} z^T \bar{A}^{(i-k-1)+} \bar{d}_k \bar{f}_i^T\right)\right] \le \exp\left[\frac{\lambda^2\sigma^2}{2} \sum_{k\in K}\sum_{i\in K} z^T \bar{A}^{(i-k-1)+} \bar{f}_k \bar{f}_i^T\right] \le \exp\left[\frac{\lambda^2\sigma^2}{2} \sum_{k\in K}\sum_{i\in K} z^T \bar{A}^{(i-k-1)+} \bar{f}_k\right], \tag{22}$$

where the last inequality holds because $\bar{f}_i^T \bar{f}_k$ is bounded in $[-1,1]$. For each $i,k \in K$, the value of $z^T \bar{A}^{(i-k-1)+} \bar{f}_k$ concentrates around its expectation $z^T \bar{A}^{(i-k-1)+} \bar{r}_k$ with probability at least $1-\exp[-Q(pT)]$. Therefore, inequality (22) leads to

$$E\left[\exp\left(\lambda \sum_{i\in K}\sum_{k\in K} z^T \bar{A}^{(i-k-1)+} \bar{d}_k \bar{f}_i^T\right)\right] \le \exp\left[Q\left(\frac{\lambda^2\sigma^2}{2n} \sum_{k\in K}\sum_{i\in K} z^T \bar{A}^{(i-k-1)+}\right)^2\right] \exp\left[Q\left(\frac{\lambda^2\sigma^2}{2n} \sum_{k\in K}\sum_{i\in K} r^{(i-k-1)+}\right)^2\right]. \tag{23}$$

Suppose the elements of K are

$$j_1 < j_2 < \cdots < j_{|K|}.$$

Define

$$D_k := j_k - j_{k-1} - 1, \quad \forall k \in \{2,\ldots,|K|\}.$$

We can calculate that

$$\sum_{i\in K} r^{(i-j_k)+} \le \frac{r^{D_k}}{1-r}.$$

Since $r^{D_k} \in [0,1]$ are bounded random variables, they are sub-Gaussian and concentrate around the mean with high probability. The expectation of r^{2D_k} is

$$\sum_{D=0}^{\infty} p(1-p)^D r^{2D} = \frac{p}{1-(1-p)r^2}.$$

Therefore, with probability at least $1-\exp[-Q(pT)]$, we have

$$\sum_{k=2}^{|K|} r^{2D_k} \le \frac{|K|\,p}{1-(1-p)r^2} \le \frac{|K|\,p}{1-r}.$$

Hence, inequality (23) implies that with the same probability, $z^T g$ is sub-Gaussian with parameter

$$Q\left[\sigma\sqrt{\frac{\sigma^2}{n}\sum_{k=2}^{|K|}\frac{r^{2D_k}}{(1-r)^2}}\right] \le Q\left[\sigma\sqrt{\frac{|K|\,p\sigma^2}{n(1-r)^3}}\right].$$

Therefore, Hoeffding's inequality leads to

$$P\left[z^T g \ge Q\sqrt{\frac{|K|\,p\sigma^2}{n(1-r)^3}\log\frac{4}{d}}\right] \le \frac{d}{4}. \tag{24}$$

By combining inequalities (20) and (24), it holds with probability at least

$$1-\exp[-Q(c^4\lambda)] - \frac{d}{2}$$

that

$$f(z) \le Q\left[\frac{c^5\sigma\lambda}{n} + \sqrt{\frac{|K|\,p\sigma^2}{n(1-r)^3}\log\frac{1}{d}}\right].$$

Similar to the bound in (15), it holds with probability at least $1-\exp[-Q(pT)]$ that

$$|K| \le 2pT.$$

As a result, if we choose

$$T = Q\left(\max\left\{\frac{\log\log(1/d)+\log(1/c)}{c^4 p(1-p)\log(1/r)}\log\frac{1}{d};\ \frac{1}{p(1-p)}\log\frac{1}{d};\ \frac{n\log(1/c)^2}{c^{10}(1-p)^2(1-r)^3\log^2(1/r)}\log\frac{1}{d}\right\}\right)$$

$$= Q\left(nR\log\frac{1}{d}\right),$$

where

$$R := \max\left\{\frac{\log(1/c)}{c^4 p(1-p)\log(1/r)}\log\frac{1}{d};\ \frac{\log^2(1/c)}{c^{10}(1-p)^2(1-r)^3\log^2(1/r)};\ \frac{1}{np(1-p)}\right\},$$

we have

$$P\left[f(z) \le Q\left(\frac{c^5\sigma\lambda}{n} + 1\right) \le d\right]. \tag{26}$$

Step 2: In the second step, we apply discretization techniques to prove that condition (13) holds for all $z \in S_2(1)$ with high probability. Suppose that $\varepsilon > 0$ is a small constant. We construct an ε-cover of the unit sphere $S_2(1)$, denoted as

$$\{z^1, \ldots, z^N\},$$

Namely, for all $z \in S_2(1)$, we can find $r \in \{1, 2, \ldots, N\}$ such that $\|z - z^r\|_2 \le \varepsilon$. The number of points N can be bounded by

$$\log(N) \le \log[N(\varepsilon, S_2(1), \|\cdot\|_2)] \le n\log\left(1 + \frac{2}{\varepsilon}\right).$$

Define a to be the lower bound of $f(z)$ in inequality (26). Then, we have

$$a = Q\left(\frac{c^5 s_l}{n}\right).$$

Our goal is to prove that

$$f(z) - f(z^0) \ge a, \quad \forall z, z^0 \in S_2(1) \text{ s.t. } \|z - z^0\|_2 \le \varepsilon$$

holds with high probability. Notice that

$$\begin{aligned} f(z) - f(z^0) &= (z - z^0)^T g + (\|Fz\|_1 - \|Fz^0\|_1) \\ &\le (z - z^0)^T g + \|z - z^0\|_2 \|g\|_2 + \|z - z^0\|_2 \sum_{i \in K} \|F^i\|_2 \\ &\le \left\| \sum_{i \in K} \sum_{k \in K} \bar{A}^{(i-k-1)_+} \bar{d}_k \right\|_2 \\ &\quad + p\frac{1}{\bar{n}} \left\| \sum_{i \in K} \sum_{k \in K} \bar{A}^{(i-k-1)_+} \bar{d}_k \right\|_2 \\ &\le \sum_{k \in K} \sum_{i > k} r^{(i-k-1)} |\bar{\ell}_k|. \end{aligned}$$

Using the property of exponential sequences, we have

$$\sum_{k \in K} \sum_{i > k} r^{(i-k-1)} |\bar{\ell}_k| \le \frac{1}{1-r} \sum_{k \in K} |\bar{\ell}_k|.$$

Using a similar proof, we can show that $\sum_{k \in K} |\bar{\ell}_k|$ is sub-Gaussian with parameter $\sqrt{|K|}s$. Therefore, Hoeffding's inequality implies that

$$P\left\{\frac{1}{1-r} \sum_{k \in K} |\bar{\ell}_k| > \frac{a}{e}\right\} \le 2\exp\left(\frac{(1-r)^2 a^2}{2e^2 |K| s^2}\right).$$

Letting

$$e := \frac{(1-r)a}{\sqrt{|K|}s\sqrt{2\log(4/d)}},$$

it holds that

$$\begin{aligned} &P\{f(z) - f(z^0) \ge a, \ \forall z, z^0 \in S_2(1) \text{ s.t. } \|z - z^0\|_2 \le \varepsilon\} \\ &\ge P\left\{\frac{1}{1-r} \sum_{k \in K} |\bar{\ell}_k| \le \frac{a}{e}\right\} \ge 1 - \frac{d}{2}. \end{aligned}$$

Now, after we replaced d in (25) with $d/(2N)$, it holds with probability at least $1 - d/2$ that

$$f(z^r) \ge a, \quad \forall r \in \{1, \ldots, N\}.$$

After combining the above two inequalities, we apply the union bound to obtain

$$P[f(z) \ge 0, \ \forall z \in S_2(1)] \ge 1 - d.$$

The corresponding sample complexity is

$$T \ge Q\left(nR\log\left(\frac{2N}{d}\right)\right).$$

Since it holds with probability $1 - \exp(-Q[p(1-p)T])$ that

$$|I_{-1}| = Q[p(1-p)T], \quad |K| = Q(pT),$$

we get the estimate

$$\begin{aligned} \log(N) &\le n\log\left(1 + \frac{2}{\varepsilon}\right) \\ &= n\log\left(1 + Q\left(\frac{n\sqrt{\log(1/d)\log(1/c)}}{(1-p)c^5(1-r)\log(1/r)}\right)\right) \\ &= Q[n\log(nR)]. \end{aligned}$$

By omitting the constants in the expression, the final sample complexity can be written as

$$T \ge Q\left(n^2 R\log\left(\frac{nR}{d}\right)\right).$$

Finally, we replaced d with d/n and apply the union bound to all coordinates $\ell \in \{1, \ldots, n\}$. The sample complexity remains on the same order as the above expression.

6) Proof of Theorem 4: Due to the system dynamics and given $x_0 = 0$, x_i can be expressed as

$$x_i = \sum_{k \in K} \bar{A}^{(i-k-1)_+} \bar{B} u_k + \sum_{k \in K} \bar{A}^{(i-k-1)_+} (\bar{B} u_k + \bar{d}_k).$$

From Proposition 4, we want to show that there exist scalars $g_i^\ell, m_i^\ell \in [-1, 1]$ for all $i \in K$, $\ell \in \{1, \ldots, n\}$ such that

$$\sum_{i \in K} g_i^\ell x_i / \sqrt{\bar{n}} = \sum_{i \in K} \|\bar{d}_i^\ell\|_2 x_i, \quad \forall \ell = 1, \ldots, n; \tag{27}$$

and

$$\sum_{i \in K} m_i^\ell u_i / \sqrt{\bar{n}} = \sum_{i \in K} \|\bar{d}_i^\ell\|_2 u_i, \quad \forall \ell = 1, \ldots, n. \tag{28}$$

We finish the proof in two steps.

Step 1: We first analyze condition (27) with a given coordinate $\ell \in \{1, \ldots, n\}$. From Lemma 3, condition (27) is equivalent to

$$f(z) := z^T g + \|Fz\|_1 \ge 0, \quad \forall z \in S_2(1),$$

where the matrix $F \in R^{n(T-|K|)}$ has the columns

$$F^i := \sum_{k \in K} \frac{\bar{A}^{(i-k-1)_+} \bar{B} u_k}{\sqrt{\bar{n}}} + \sum_{k \in K} \frac{\bar{A}^{(i-k-1)_+} (\bar{B} u_k + \bar{d}_k)}{\sqrt{\bar{n}}}, \ \forall i \in K,$$

and the vector $g \in R^n$ is

$$g := \sum_{i \in K} \left(\sum_{k \in K} \bar{A}^{(i-k-1)_+} \bar{B} u_k + \sum_{k \in K} \bar{A}^{(i-k-1)_+} (\bar{B} u_k + \bar{d}_k) \right) f_i^T.$$

Similar to the proof of Theorem 2, we first prove that $f(z) \ge a$ holds with high probability for a fixed $z \in S_2(1)$ and some

positive constant α. For each $k \in K_-$, the standard deviation and sub-Gaussian parameter of $z^T \bar{A}^{(i_k-1)+} \bar{B} u_k$ are both

$$\frac{1}{\sqrt{m}} \|z^T \bar{A}^{(i_k-1)+} \bar{B}\|_2 \chi.$$

For each $k \in K_-$, the standard deviation and sub-Gaussian parameter of $z^T \bar{A}^{(i_k-1)+} (\bar{B} u_k + \bar{d}_k)$ are, respectively,

$$\sqrt{\frac{1}{m} \|z^T \bar{A}^{(i_k-1)+} \bar{B}\|_2^2 \chi^2 + \frac{1}{n} \|z^T \bar{A}^{(i_k-1)+}\|_2^2 s_k^2},$$
$$\sqrt{\frac{1}{m} \|z^T \bar{A}^{(i_k-1)+} \bar{B}\|_2^2 \chi^2 + \frac{1}{n} \|z^T \bar{A}^{(i_k-1)+}\|_2^2 s^2}.$$

Note that we have utilized the independence between u_k and $\bar{d}_k$ in the above calculation. Let

$$S := \left\lceil \log_r \left(4 \sqrt{\frac{\frac{1}{m} h_B^2 \chi^2 + \frac{p}{n} s^2}{\frac{1}{m} r_B^2 \chi^2 + \frac{p}{n} s^2}} \, \frac{c^5}{\sqrt{\log(1/\delta)}} \right) \right\rceil,$$

where r_B is the maximal singular value of $\bar{B}$ and h_B is the minimal singular value of the matrix

$$\frac{1}{(1-r)^2} \left[\bar{B} \;\; \bar{A}\bar{B} \;\; \cdots \;\; \bar{A}^{n-1} \bar{B} \right].$$

By the controllability assumption, the above matrix is rank n and thus, the parameter h_B is strictly positive. We define $i_0 := -1$ and construct the index set

$$I := \{ i_1, \ldots, i_l \mid i_j \in K_-, \; i_j - i_{j-1} \le S, \; \forall j \}.$$

It is straightforward to construct I such that $l = |I|$ is on the order of

$$\min \left\{ (1-p)T, \frac{T}{S} \right\}.$$

For each $j \in \{1, \ldots, l\}$, we define

$$K_j := \{ k \in K_- \mid i_{j-1} < k < i_j \}, \quad K_j^c := \{ k \in K_- \mid i_{j-1} < k < i_j \}.$$

Moreover, we define

$$X_{j,\ell} := \sum_{k \in K_j} z^T \bar{A}^{i_\ell - k - 1} \bar{B} u_k + \bar{d}_k + \sum_{k \in K_j^c} z^T \bar{A}^{i_\ell - k - 1} \bar{B} u_k, \quad \forall j, \ell \in \{1, \ldots, l\} \text{ s.t. } j < \ell.$$

For all $j \in \{1, \ldots, l\}$, the stealthy assumption (Assumption 2) implies that the standard deviation and the sub-Gaussian parameter of $X_{j,\ell}$ is

$$\tilde{s}_{j,\ell} := \sqrt{\frac{1}{m} \sum_{k \in K_j^c} \|z^T \bar{A}^{i_\ell - k - 1} \bar{B}\|_2^2 \chi^2 + \frac{1}{n} \sum_{k \in K_j} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 s_k^2},$$

$$s_{j,\ell} := \sqrt{\frac{1}{m} \sum_{k \in K_j^c} \|z^T \bar{A}^{i_\ell - k - 1} \bar{B}\|_2^2 \chi^2 + \frac{1}{n} \sum_{k \in K_j} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 s^2},$$

respectively. Define

$$c_{j,\ell} := \frac{\tilde{s}_{j,\ell}}{s_{j,\ell}}, \quad \forall j, \ell \in \{1, \ldots, l\} \text{ s.t. } j < \ell.$$

Similar to the proof of Theorem 2, we have the bound

$$\|z^T F\|_1 \le \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{l} |X_{j,j}| - \sum_{\ell=j+1}^{l} |X_{j,\ell}| \right).$$

By Lemma 5, we have

$$P\left(|X_{j,j}| \le c_{j,j} s_{j,j} \right) \ge \frac{c_{j,j}^4}{64}. \tag{29}$$

For all vector $y \in \mathbb{R}^n$, the controllability assumption leads to

$$\sum_{k=0}^{n-1} \|y^T \bar{A}^k \bar{B}\|_2^2 \ge \frac{h_B^2}{(1-r)^2} \|y\|_2^2 \ge \frac{h_B^2}{(1-r)^2} (1-r)^2 \sum_{k=0}^{n} r^{2k} \|y\|_2^2 \ge h_B^2 \sum_{k=0}^{n} \|y^T \bar{A}^k\|_2^2. \tag{30}$$

Therefore, we can divide the set $[1, T] \setminus K_-$ into segments with n consecutive time instances and apply inequality (30) to each segment. When T is large enough such that $T \ge Q(n)$, we obtain the estimation

$$\sum_{k \in K_j \setminus K_j^c} \|z^T \bar{A}^{i_\ell - k - 1} \bar{B}\|_2^2 \ge \sum_{k=i_{j-1}+1}^{i_j - 1} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 h_B^2.$$

Applying concentration inequalities to set K_j, the distribution of its elements will surround their expected values. Therefore, for the simplicity of presentation, we use the following approximation:

$$s_{j,j}^2 \approx \frac{1}{m} h_B^2 \chi^2 + \frac{p}{n} s^2 := \bar{s}^2.$$

In addition, the parameter $c_{j,j}$ can be bounded by

$$c_{j,j}^2 \le \frac{\frac{1}{m} \sum_{k \in K_j \setminus K_j^c} \|z^T \bar{A}^{i_\ell - k - 1} \bar{B}\|_2^2 \chi^2}{\frac{1}{m} \sum_{k \in K_j \setminus K_j^c} \|z^T \bar{A}^{i_\ell - k - 1} \bar{B}\|_2^2 \chi^2 + \frac{1}{n} \sum_{k \in K_j} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 s^2}$$
$$= 1 + \left(\frac{1}{m} \sum_{k \in K_j \setminus K_j^c} \|z^T \bar{A}^{i_\ell - k - 1} \bar{B}\|_2^2 \chi^2 \right)^{-1} \left(\frac{1}{n} \sum_{k \in K_j} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 s^2 \right)^{-1}.$$

For the numerator, we can estimate that

$$\sum_{k \in K_j \setminus K_j^c} \|z^T \bar{A}^{i_\ell - k - 1} \bar{B}\|_2^2 \ge \sum_{k=i_{j-1}+1}^{i_j - 1} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 h_B^2.$$

On the other hand, since

$$\sum_{k \in K_j \setminus K_j^c} \|z^T \bar{A}^{i_\ell - k - 1} \bar{B}\|_2^2 \chi^2 \le \sum_{k \in K_j \setminus K_j^c} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 r_B^2 \chi^2,$$
$$\sum_{k \in K_j} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 s^2 \le \sqrt{p} \sum_{k \in K_j \setminus K_j^c} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 s^2,$$

we get

$$c_{j,j}^2 \le \frac{\frac{1}{m} h_B^2 \chi^2}{\frac{1}{m} r_B^2 \chi^2 + \frac{p}{n} s^2} := c.$$

Therefore, inequality (29) implies

$$P(|X_{j,j}| \geq c\bar{s}) \leq \frac{c^4}{64}. \quad (31)$$

Since the sub-Gaussian parameter of $\sum_{\ell=j+1}^{l} |X_{j,\ell}|$ is $\sum_{\ell=j+1}^{l} s_{j,\ell}$, Hoeffding's inequality implies that

$$P\left(\sum_{\ell=j+1}^{l} |X_{j,\ell}| \geq \sum_{\ell=j+1}^{l} s_{j,\ell} + 2\log\frac{4l}{d}\right) \leq 1 - \frac{d}{4l}. \quad (32)$$

We can bound the sub-Gaussian parameter by

$$\begin{aligned}
&\sum_{\ell=j+1}^{l} s_{j,\ell} \\
&\sqrt{\frac{1}{m}\sum_{k\in K_j\cup K_j^c} r^{2(i_\ell-k-1)} r_B^2 x^2 + \frac{1}{n}\sum_{k\in K_j} r^{2(i_\ell-k-1)} s^2} \\
&\frac{r^S}{1-r^S}\sqrt{\frac{1}{m}\sum_{k\in K_j\cup K_j^c} r^{2(i_j-k-1)} r_B^2 x^2 + \frac{1}{n}\sum_{k\in K_j} r^{2(i_j-k-1)} s^2} \\
&\frac{r^S}{1-r^S}\sqrt{\frac{1}{m(1-r)} r_B^2 x^2 + \frac{1}{n}\sum_{k\in K_j} r^{2(i_j-k-1)} s^2}.
\end{aligned}$$

In the same way, we have the following bound with high probability:

$$\sum_{k\in K_j} r^{2(i_j-k-1)} \cdot p \leq \sum_{k\in K_j\cup K_j^c} r^{2(i_j-k-1)} \leq \frac{p}{1-r},$$

which holds with high probability when T is large. Therefore, we have the bound

$$\begin{aligned}
&\sum_{\ell=j+1}^{l} s_{j,\ell} \cdot \frac{r^S}{1-r^S}\sqrt{\frac{1}{m(1-r)} r_B^2 x^2 + \frac{p}{n(1-r)} s^2} \\
&:= \frac{r^S}{1-r^S}\tilde{s}.
\end{aligned}$$

By the choice of S, we get

$$\sum_{\ell=j+1}^{l} s_{j,\ell} \cdot \frac{c^4}{256} c\bar{s} \leq 2\log\frac{4l}{d}\Bigg)^{-1};$$

Therefore, inequality (32) leads to

$$P\left(\sum_{\ell=j+1}^{l} |X_{j,\ell}| \geq \frac{c^4}{256} c\bar{s}\right) \leq 1 - \frac{d}{4l}. \quad (33)$$

Choosing

$$T \geq \frac{\log\log(1/d)}{c^4 \min\{1-p; 1/S\}};$$

we have

$$\frac{c^4}{64} \leq \frac{d}{4l} \leq \frac{c^4}{128}.$$

By the same construction of the martingale sequence and the application of Azuma-Hoeffding's inequality, inequalities (29) and (32) imply that

$$\begin{aligned}
\|z^T F\|_1 &\leq p\frac{1}{n}\sum_{j=1}^{l} |X_{j,j}| \geq \sum_{\ell=j+1}^{l} |X_{j,\ell}|\Bigg) \\
&\geq p\frac{1}{n}\left(\frac{c^4 l}{256} c\bar{s} - \frac{c^4 l}{512} c\bar{s}\right) = \frac{c^5\bar{s}}{512\sqrt{pn}} l
\end{aligned} \quad (34)$$

holds with probability at least

$$1 - \exp[-Q(c^4 l)] - d/4.$$

On the other hand, for the term $z^T g$, we can bound its sub-Gaussian parameter by

$$\sqrt{\frac{|K|r_B^2 x^2}{m(1-r)^3} + \frac{|K|ps^2}{n(1-r)^3}} = \sqrt{\frac{|K|}{(1-r)^2}}\tilde{s}.$$

Then, Hoeffding's inequality leads to

$$P\left(z^T g \geq Q\sqrt{\frac{|K|\tilde{s}^2}{(1-r)^2}\log\frac{4}{d}}\right) \leq \frac{d}{4}. \quad (35)$$

Combining inequalities (34) and (35), it holds with probability at least

$$1 - \exp[-Q(c^4 l)] - \frac{d}{2}$$

that

$$f(z) \geq Q\left(\frac{c^5\bar{s}l}{\sqrt{pn}} - \sqrt{\frac{\tilde{s}^2|K|}{(1-r)^2}\log\frac{1}{d}}\right).$$

Similar to the bound in (15), it holds with probability at least $1 - \exp[-Q(pT)]$ that

$$|K| \leq 2pT.$$

As a result, if we choose

$$\begin{aligned}
T \geq Q\max\Bigg(&\frac{1}{c^4\min\{1-p; 1/S\}}\log\frac{1}{d};\\
&\frac{1}{p}\log\frac{1}{d};\\
&\frac{npk^2}{c^{10}(1-r)^2\min\{(1-p)^2; 1/S^2\}}\log\frac{1}{d}\Bigg)\\
\geq Q\Bigg(&nR_1\log\frac{1}{d}\Bigg);
\end{aligned}$$

where $k := \tilde{s}/\bar{s} \leq (1-r)^{-1}$ and

$$\begin{aligned}
R_1 := \max\Bigg(&\frac{1}{c^4(1-p)}; \frac{\log(k/c)}{c^4\log(1/r)};\\
&\frac{pk^2}{c^{10}(1-p)^2(1-r)^2}; \frac{pk^2\log^2(k/c)}{c^{10}(1-r)^2\log^2(1/r)}; \frac{1}{np}\Bigg)
\end{aligned}$$

we have

$$P\left(f(z) \geq Q\frac{c^5\bar{s}l}{\sqrt{pn}}\right) \leq 1 - d. \quad (36)$$

Next, we apply the discretization techniques and estimate the size of ϵ-net, which we denote as N. Similar to the proof of Theorem 2, it is sufficient to choose

$$\log(N) \quad n\log\left(1+\frac{2}{\epsilon}\right);$$

and

$$\epsilon := Q\left(\frac{a}{\|g\|_2+\sum_{i\in K}\|F^i\|_2}\right);$$

where $a>0$ is the lower bound of $f(z)$ in (36). We can estimate that

$$\|g\|_2 \quad \sum_{i\in K}\sum_{k\in\bar{K}}\bar{A}^{(i\ k\ 1)+}\bar{B}u_k+\sum_{k\in K}\bar{A}^{(i\ k\ 1)+}(\bar{B}u_k+\bar{d}_k)\Big\|_2$$
$$\frac{r_B}{1\ r}\sum_{k=0}^{T\ 1}\|u_k\|_2+\frac{1}{1\ r}\sum_{j\in K}\|\bar{d}_k\|_2:$$

Therefore, the sub-Gaussian parameter of $\|g\|_2$ is bounded by

$$\frac{1}{1\ r}\sqrt{r_B^2Tx^2+jK\ js^2}.\ \frac{1}{1\ r}\sqrt{r_B^2Tx^2+pTs^2}:=s\sigma\sqrt{T}:$$

Similarly, the sub-Gaussian parameter of $\sum_{i\in K}\|F^i\|_2$ is bounded by

$$\frac{1}{(1\ r)\sqrt{n}}\sqrt{r_B^2Tx^2+pTs^2}=s\sigma\sqrt{\frac{pT}{n}}$$

with high probability. Hoeffding's inequality implies

$$\|g\|_2+\sum_{i\in K}\|F^i\|_2 \quad Q\left(s\sigma\sqrt{T}\ \sqrt{\log\frac{1}{d}}\right)$$

with probability at least $1\ d/2$. With the same probability, we have

$$\log\left(1+\frac{2}{\epsilon}\right) \quad \log\left(1+Q\left(\frac{\sqrt{n}s\sigma\sqrt{\log(1/d)}}{c^5\bar{s}\min\{1\ p;1/s\}\sqrt{T}}\right)\right)$$
$$Q[\log(nR_1)];$$

which further leads to

$$\log(N)\ .\ nQ[\log(nR_1)]:$$

Replacing d with d/N in (36), the final sample complexity bound is

$$T \quad Q\left(n^2R_1\log\frac{nR_1}{d}\right):$$

Step 2: In the second step, we consider condition (28). From Lemma 3, given a coordinate $i\in\{1;:::;n\}$, (28) is equivalent to

$$f(z):=z^Tg+\|z^TF\|_1 \quad 0;\quad 8z\in S_2(1);$$

where the matrix $F\in R^{m\ (T\ j\ K\ j)}$ has the columns

$$F^i:=\frac{u_i}{\sqrt{n}};\quad 8i\ 62K\ ;$$

and the vector $g\in R^m$ is

$$g:=\sum_{i\in K}u_if_i^T:$$

For a given $z\in S_2(1)$, we have

$$E[f(z)]=E\|z^TF\|_1=\sum_{i\in\bar{K}}E|z^TF^i|$$
$$=Q\left(\frac{(T\ j\ K\ j)x}{\sqrt{mn}}\right)\ \&\ Q\left(\frac{(1\ p)Tx}{\sqrt{mn}}\right):$$

The sub-Gaussian parameter of $z^TF\|_1+z^Tg$ is

$$\sqrt{\frac{(T\ j\ K\ j)x^2}{mn}+\frac{jK\ jx^2}{m}}.\ \sqrt{\frac{1\ p}{mn}+\frac{p}{m}}\ Tx^2:$$

Therefore, Hoeffding's inequality implies that

$$f(z) \quad Q\left(\frac{(1\ p)Tx}{\sqrt{mn}}\right)$$

holds with probability at least

$$1\ \exp\left(Q\left(\frac{(1\ p)^2T}{1\ p+np}\right)\right):$$

Choosing

$$T \quad Q\left(\max\left\{\frac{1}{p}\log\frac{1}{d};\frac{1\ p+np}{(1\ p)^2}\log\frac{1}{d}\right\}\right)$$
$$=Q\left(\max\left\{\frac{1}{p}\log\frac{1}{d};\frac{np}{(1\ p)^2}\log\frac{1}{d}\right\}\right);$$

we have

$$P\left\{f(z) \quad Q\left(\frac{(1\ p)Tx}{\sqrt{mn}}\right)\right\} \quad 1\ \frac{d}{2}:$$

Similarly, applying the discretization techniques, it is sufficient to choose N points, where

$$\log(N)$$
$$=m\log\left(1+Q\left(\frac{\sum_{i\in K}\|u_i\|_2/\sqrt{n}+\sum_{i\in K}\|u_i\|_2}{(1\ p)Tx/\sqrt{mn}}\right)\right)$$
$$.\ m\log\left(1+Q\left(\frac{\sqrt{(1\ p)T/n}+\sqrt{pTx}\sqrt{\log(1/d)}}{(1\ p)Tx/\sqrt{mn}}\right)\right)$$
$$=m\log\left(1+Q\left(\frac{\sqrt{1\ p+np}\sqrt{\log(1/d)}}{(1\ p)\sqrt{T/m}}\right)\right)$$
$$m\log\ m\log\frac{1}{d}\quad:$$

Hence, the overall sample complexity is

$$T \quad Q\left(mnR_2\log\frac{nR_2}{d}\right);$$

where we define

$$R_2:=\max\left\{\frac{1}{np};\frac{p}{(1\ p)^2};\frac{m}{n}\right\}:$$

Combining the two steps, we get the conclusion of this theorem.

B. Proofs for Results in Appendix

1) Proof of Lemma 5: For the notational simplicity, we omit the X in subscripts. Let

$$h := \frac{\tilde{s}}{s}; \quad d := 1 - \frac{\tilde{s}^4}{64s^4}.$$

Assume conversely that

$$P(|X| \leq hs) < 1 - d.$$

Then, we can calculate that

$$\begin{aligned}
E(X^2) &= \int_0^\infty q^2 \, d[-P(|X| \geq q)] = \int_0^\infty 2q P(|X| \geq q) \, dq \\
&= (hs)^2 + \int_{hs}^\infty 2q \min\left\{1 - d; 2\exp\left(-\frac{q^2}{2s^2}\right)\right\} dq \\
&= (hs)^2 + (1-d)(s')^2 - (hs)^2 \\
&\quad + \int_{s'}^\infty 2q \cdot 2\exp\left(-\frac{q^2}{2s^2}\right) dq \\
&= (hs)^2 + (1-d)(s')^2 - (hs)^2 \\
&\quad + 4s^2 \exp\left(-\frac{(s')^2}{2s^2}\right),
\end{aligned}$$

where we define

$$s' := \sqrt{2s^2 \log\left(\frac{2}{1-d}\right)}.$$

Rearranging the above inequality, we get

$$h^2 \leq d + 2(1-d)\log\left(\frac{2}{1-d}\right) + 2(1-d)\frac{\tilde{s}^2}{s^2}.$$

Hence, it holds that

$$\begin{aligned}
h^2 &\leq \frac{1}{d}\frac{\tilde{s}^2}{s^2} \cdot 2(1-d)\log\left(\frac{2}{1-d}\right) \cdot 2(1-d) \\
&> 2\frac{\tilde{s}^2}{s^2} + 4(1-d)\log\left(\frac{1-d}{2}\right),
\end{aligned}$$

where the second inequality holds because $d > 1/2$ and $\log[2/(1-d)] > 1$. Using the fact that

$$(1-d)\log\left(\frac{1-d}{2}\right) \geq -\sqrt{1-d}\,\frac{\tilde{s}^2}{8s^2},$$

we get

$$h^2 > \frac{\tilde{s}^2}{s^2},$$

which contradicts with the definition of h. Therefore, we have proved that

$$P(|X| \leq \tilde{s}) \leq \frac{\tilde{s}^4}{64s^4}.$$

REFERENCES

[1] H.-F. Chen and L. Guo, *Identification and stochastic adaptive control*. Springer Science & Business Media, 2012.

[2] E. Hazan, H. Lee, K. Singh, C. Zhang, and Y. Zhang, "Spectral filtering for general linear dynamical systems," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[3] H. Mania, S. Tu, and B. Recht, "Certainty equivalence is efficient for linear quadratic control," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[4] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Nonparametric finite time LTI system identification," *arXiv preprint arXiv:1902.01848*, 2019.

[5] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite sample perspective," *arXiv preprint arXiv:2209.05423*, 2022.

[6] A. Alan, A. J. Taylor, C. R. He, A. Ames, and G. Orosz, "Control barrier functions and input-to-state safety with application to automated vehicles," *IEEE Transactions on Control Systems Technology*, vol. 31, pp. 2744–2759, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249461776

[7] L. Wang, E. A. Theodorou, and M. Egerstedt, "Safe learning of quadrotor dynamics using barrier certificates," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2460–2465, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:35948052

[8] S. M. Khansari-Zadeh and A. Billard, "Learning control lyapunov function to ensure stability of dynamical system-based robot reaching motions," *Robotics Auton. Syst.*, vol. 62, pp. 752–765, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14374268

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

[10] M. Simchowitz and D. Foster, "Naive exploration is optimal for online LQR," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8937–8948.

[11] R. Zhang, Y. Li, and N. Li, "On the regret analysis of online LQR control with predictions," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 697–703.

[12] I. Ziemann, A. Tsiamis, B. Lee, Y. Jedra, N. Matni, and G. J. Pappas, "A tutorial on the non-asymptotic theory of system identification," 2023.

[13] L. Bako and H. Ohlsson, "Analysis of a nonsmooth optimization approach to robust estimation," *Automatica*, vol. 66, pp. 132–145, Apr. 2016.

[14] H. Xu, C. Caramanis, and S. Mannor, "Robustness and Regularization of Support Vector Machines," *Journal of machine learning research*, vol. 10, no. 7, 2009.

[15] L. Bako, "On a Class of Optimization-Based Robust Estimators," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5990–5997, Nov. 2017.

[16] D. Bertsimas and M. S. Copenhaver, "Characterization of the equivalence of robustification and regularization in linear and matrix regression," *European Journal of Operational Research*, vol. 270, no. 3, pp. 931–942, Nov. 2018.

[17] S. Pesme and N. Flammarion, "Online robust regression via SGD on the l1 loss," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2540–2552, 2020.

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

[19] S. Mendelson, "Learning without concentration," *Journal of the ACM (JACM)*, vol. 62, no. 3, pp. 1–25, 2015.

[20] Y. Li, S. Das, J. Shamma, and N. Li, "Safe adaptive learning-based control for constrained linear quadratic regulators with regret guarantees," *arXiv preprint arXiv:2111.00411*, 2021.

[21] S. Fattahi, N. Matni, and S. Sojoudi, "Learning sparse dynamical systems from a single sample trajectory," *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 2682–2689.

[22] Y. Jedra and A. Proutiere, "Finite-time identification of stable linear systems optimality of the least-squares estimator," *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 996–1001.

[23] A. Wagenmaker and K. Jamieson, "Active learning for identification of linear dynamical systems," in *Conference on Learning Theory*. PMLR, 2020, pp. 3487–3582.

[24] H. Feng, B. Yalcin, and J. Lavaei, "Learning of dynamical systems under adversarial attacks - null space property perspective," *2023 American Control Conference (ACC)*, pp. 4179–4184, 2023.

[25] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[26] H. Feng and J. Lavaei, "Learning of dynamical systems under adversarial attacks," in *2021 60th IEEE Conference on Decision and Control (CDC)*. 2021, pp. 3010–3017.

[27] Y. Chen, J. Fan, C. Ma, and Y. Yan, "Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data," *Annals of statistics*, vol. 49, no. 5, p. 2948, 2021.