IEEE Open Journal of Control Systems

Received 24 July 2024; revised 22 October 2024; accepted 19 November 2024. Date of publication 27 November 2024; date of current version 9 January 2025. Recommended by Senior Editor Dr. Alberto Speranzon.

Digital Object Identifier 10.1109/OJCSYS.2024.3507452

Exact Recovery for System Identification With More Corrupt Data Than Clean Data

BATURALP YALCIN [1], HAIXIANG ZHANG [2], JAVAD LAVAEI [1], AND MURAT ARCAK [3]

[1]Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720 USA
[2]Department of Mathematics, University of California, Berkeley, CA 94720 USA
[3]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA

CORRESPONDING AUTHOR: BATURALP YALCIN (e-mail: baturalp_yalcin@berkeley.edu).

This work was supported in part by the U. S. Army Research Laboratory and the U. S. Army Research Office under Grant W911NF2010219, in part by the Office of Naval Research under Grant N000142412673, in part by AFOSR, in part by NSF, and in part by the UC Noyce Initiative.

ABSTRACT This paper investigates the system identification problem for linear discrete-time systems under adversaries and analyzes two lasso-type estimators. We examine non-asymptotic properties of these estimators in two separate scenarios, corresponding to deterministic and stochastic models for the attack times. 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 emphasize that the attack vectors are allowed to be correlated with each other in this work. 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 Linear systems, robust control, statistical learning, system identification.

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, 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 [1]. Despite several advances in this field, most results in system identification focus on the asymptotic properties of the proposed estimators, i.e., their behavior as sample size approaches infinity [2], [3]. Nonetheless, the non-asymptotic analysis of the system identification problem has gained interest in recent years [4], [5], [6], [7]. Non-asymptotic analysis is crucial to understand the required sample complexity for online control problems.

Robust learning of dynamical systems is crucial for safety-critical applications, such as autonomous driving [8], unmanned aerial vehicles [9], and robotic arms [10]. 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 [11], [12], [13]. 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 [14] 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 [15]. 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 [16], [17], [18], [19]. ℓ_1-based non-smooth estimators are extensively investigated in the context of robust learning in the presence of adversaries and outliers [20], [21], [22]. Compressed sensing and sparse error detection are such problems that are closely related to the proposed estimators [23], [24], [25], [26]. 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 [27]. 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. Block Martingale Small Ball (BMSB) techniques are used to analyze the properties of the estimator [11], [28], [29]. It employs normalized martingale bounds for the estimation error when the excitation is large enough [11].

Unlike the non-asymptotic analysis of correlated data, the least-squares estimator offers a closed-form solution [30], [31], [32]. As long as the noise magnitudes are not large, the least-squares estimator performs relatively well. The estimation error asymptotically converges to zero with the optimal rate of $T^{-1/2}$, where T is the number of samples collected from the system [11]. However, it is not robust to adversarial attacks, and the literature on robust learning of dynamical systems is limited. The work [33] uses compressed sensing to learn system parameters for FIR systems. However, the impulse vectors are assumed to be Gaussian independent and identically distributed, which does not contain the correlated state vectors. The work by [34] 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 [34] and study 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 consider an LTI 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. We aim to learn these matrices from the samples $\{x_i\}_{i=0}^{T}$ and $\{u_i\}_{i=0}^{T-1}$ 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 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 $\{x_i\}_{i=0}^{T}$ and $\{u_i\}_{i=0}^{T-1}$:

$$\min_{A\in\mathbb{R}^{n\times n}, B\in\mathbb{R}^{n\times m}} \sum_{i=0}^{T-1} \|x_{i+1} - Ax_i - Bu_i\|_\circ, \quad \circ \in \{1, 2\}.$$

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 Δ. We show that both of our estimators exactly recover the true system matrices when the system is stable and the number of samples, i.e., T, is larger than $n + \Delta$.

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. Then, the estimators achieve exact recovery with probability at least $1 - \delta$ if the time horizon T satisfies the inequality $T \geq \Theta(\max\{T^1_{\text{sample}}, T^2_{\text{sample}}\})$, where T^1_{sample} and T^2_{sample} are defined as

$$n^2 R_1 \log\left(\frac{nR_1}{\delta}\right) \quad \text{and} \quad nmR_2 \log\left(\frac{nR_2}{\delta}\right),$$

with the constants R_1 and R_2 defined in Theorem 4. 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 II and III, we introduce the notations used in the paper and formulate the problem, respectively. In Section IV, we study the convergence and sample complexity properties of our estimators

in the case when the system is autonomous. In Section V, we generalize the results to non-autonomous systems. In Section VI, we demonstrate the results on synthetic simulations and a biomedical system that models blood sugar levels with the injection of bolus insulin.

II. NOTATION AND PRELIMINARIES

For a matrix Z, $\|Z\|_F$ and $\|Z\|_{op}$ denote the Frobenius norm and operator 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 $f(x) = \Theta[g(x)]$ 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. The relation $f(x) \gtrsim g(x)$ holds if $g(x) \lesssim f(x)$. Given the function f, ∂f denotes the subdifferential of the function. $|S|$ shows the cardinality of a given set S. 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.

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.

Lemma 1: (Hoeffding's Bound [35]) Suppose that the variable X has mean μ and sub-Gaussian parameter σ. Then, for all $t > 0$, we have

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

The subdifferential of the ℓ_2 norm of 0 vector is the ℓ_2 norm unit ball, whereas the subdifferential of the ℓ_1 norm of 0 vector is the ℓ_∞ norm unit ball, which is $\mathbb{B}_\infty(1) = \{x \in \mathbb{R}^n : \|x\|_\infty \leq 1\}$. 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. We also define the unit ball $\mathbb{S}_2(1)$ as $\mathbb{S}_2(1) = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$, that is the set of all the points on the sphere with radius 1.

III. PROBLEM FORMULATION

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 $\mathcal{K}$, that is $\mathcal{K} = \{i : \bar{d}_i \neq 0, i \in 0, 1, \ldots, T-1\}$. These time instances are called the attack times and $\mathcal{K}$ is the set of attack times. Similarly, the set of time instances without attack or corrupted data is shown with $\mathcal{K}^c$ and these time instances are called the 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. To exactly recover the system matrices $\bar{A}$ and $\bar{B}$, we analyze the following convex optimization problems with non-smooth objective functions:

$$\min_{A\in\mathbb{R}^{n\times n}, B\in\mathbb{R}^{n\times m}} \sum_{i=0}^{T-1} \|x_{i+1} - Ax_i - Bu_i\|_2 \qquad \text{(CO-L2)}$$

and

$$\min_{A\in\mathbb{R}^{n\times n}, B\in\mathbb{R}^{n\times m}} \sum_{i=0}^{T-1} \|x_{i+1} - Ax_i - Bu_i\|_1 \qquad \text{(CO-L1)}$$

where the states $\{x_i\}_{i=0}^{T}$ are generated according to $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i, \quad i = 0, \ldots, T-1$. The difference between problems (CO-L2) and (CO-L1) is their objective functions. 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 $\{d_i\}_{i=0}^{T-1}$ because the occurrence of large injections of disturbances is rare and not frequent. We highlight that the vectors $\{\bar{d}_i\}_{i=0}^{T-1}$ 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 $\{d_i\}_{i=0}^{T-1}$ 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. 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.

The optimization problems (CO-L2) and (CO-L1) are equivalent to an empirical risk minimization problem for which the loss function is the ℓ_2 and ℓ_1 norms. We remark that classical statistical theory on empirical risk minimization is not applicable here due to the correlated data at each time instance. By representing the data points X_i as tuples (x_{i+1}, x_i, u_i), it is impossible to claim that X_i and X_{i+1} are independent, which is a key assumption in the empirical risk minimization literature. We can also transform this problem into the compressed sensing problem. Let $Y = [x_1, \ldots, x_T]$, $X = [x_0, \ldots, x_{T-1}]$, and $U = [u_0, \ldots, u_{T-1}]$ be the matrices where the state, input and noise vectors are given as columns. Then, the problem is equivalent to

$$\min_{A\in\mathbb{R}^{n\times n}, B\in\mathbb{R}^{n\times m}} \left\| Y - [A, B]\begin{bmatrix} X \\ U \end{bmatrix} \right\|_{\circ,1},$$

where $\|\cdot\|_{\circ,1}$ is the sum of the $\circ$-norms of the columns for a given matrix. However, some entries of the feature matrix $[X^T, U^T]^T$ are not independent of each other due to the auto-correlation between the state vectors. Therefore, the classical compressed sensing results are not applicable to this system identification problem. In addition, the existing sufficient conditions, such as NSP [34], do not provide explicit conditions for exact recovery and are difficult to analyze. As the first step

of the proof, the Karush-Kuhn-Tucker (KKT) conditions will be used to analyze the properties of these estimators because (CO-L2) and (CO-L1) are convex optimization problems.

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

$$0 \in \sum_{i \notin \mathscr{K}} x_i \otimes \partial \|(\bar{A} - \hat{A})x_i + (\bar{B} - \hat{B})u_i\|_\circ + \sum_{i \in \mathscr{K}} x_i \otimes \partial \|(\bar{A} - \hat{A})x_i + (\bar{B} - \hat{B})u_i + \bar{d}_i\|_\circ, \quad (1)$$

$$0 \in \sum_{i \notin \mathscr{K}} u_i \otimes \partial \|(\bar{A} - \hat{A})x_i + (\bar{B} - \hat{B})u_i\|_\circ + \sum_{i \in \mathscr{K}} u_i \otimes \partial \|(\bar{A} - \hat{A})x_i + (\bar{B} - \hat{B})u_i + \bar{d}_i\|_\circ, \quad (2)$$

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

We emphasize that $(\bar{A}, \bar{B})$ represent the unknown ground truth matrices, (A, B) are the decision variables in the convex optimization problems, and $(\hat{A}, \hat{B})$ are the solutions to those convex optimization problems. The proof for the KKT conditions when $\circ = 2$ is provided in [36], and the proof for the case $\circ = 1$ 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:

$$0 \in \sum_{i \notin \mathscr{K}} x_i \otimes \partial \|0\|_\circ + \sum_{i \in \mathscr{K}} x_i \otimes \partial \|\bar{d}_i\|_\circ,$$

$$0 \in \sum_{i \notin \mathscr{K}} u_i \otimes \partial \|0\|_\circ + \sum_{i \in \mathscr{K}} u_i \otimes \partial \|\bar{d}_i\|_\circ.$$

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 $i = 0, \ldots, T-1$. We study noiseless systems under an adversary to obtain exact recovery results, meaning that if there is no attack at time i, $i \in \mathscr{K}^c$, then $\bar{d}_i = 0$. We are interested in recovering the system matrix $\bar{A}$ using the following convex optimization problems for autonomous systems:

$$\min_{A \in \mathbb{R}^{n \times n}} \sum_{i=0}^{T-1} \|x_{i+1} - Ax_i\|_2 \quad \text{(CO-L2-Aut)}$$

and

$$\min_{A \in \mathbb{R}^{n \times n}} \sum_{i=0}^{T-1} \|x_{i+1} - Ax_i\|_1 \quad \text{(CO-L1-Aut)}$$

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

$$0 \in \sum_{i \notin \mathscr{K}} x_i \otimes \partial \|(\bar{A} - A)x_i\|_\circ + \sum_{i \in \mathscr{K}} x_i \otimes \partial \|((\bar{A} - A)x_i + \bar{d}_i)\|_\circ.$$

Remark 1: As a remark, although the set of attack times $\mathscr{K}$ 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. Moreover, although the attacker can choose a zero attack, $\bar{d}_i = 0$, the time period i is not included in the set of attack vectors, i.e. $i \notin \mathscr{K}$, to facilitate the mathematical derivations regarding subdifferentials. However, due to the continuity of optimal solutions, one can select an infinitesimal small value for the attack vector. Thus, its limit corresponds to a zero attack.

We examine two types of attack structures: Δ-spaced and probabilistic 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 $\mathscr{K}$.

A. Δ-SPACED ATTACK STRUCTURE

The first attack structure is a deterministic attack model for which the attacks occur at every Δ time period. For instance, if $\Delta = 2$, the set $\mathscr{K}$ could be $\{1, 3, 5, \ldots, 2k+1\}$, meaning that an agent injects a disturbance vector into the system at every odd time instance. We first define the deterministic attack model, borrowed from [36].

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

Initially, we consider first-order ans stable systems where $x_i \in \mathbb{R}$ and $|\bar{A}| < 1$. When $n = 1$, the problems (CO-L1-Aut) and (CO-L2-Aut) are equivalent, and therefore, we only focus on (CO-L2-Aut). We will show that the convex formulation (CO-L2-Aut) exactly recovers $\bar{A}$ in the case of Δ-spaced disturbance sequence with $\Delta \geq 2$.

Proposition 1: Consider a first-order autonomous system with $|\bar{A}| < 1$ and Δ-spaced disturbance sequence with $\Delta \geq 2$. Then, whenever $x_0 = 0$ or $0 \notin \mathscr{K}$, the convex formulation (CO-L2-Aut) has the unique solution $\bar{A}$ as long as the sample complexity satisfies the inequality $T \geq \Delta + 1$.

This proposition implies that whenever there are more than $\Delta + 1$ data samples, the exact recovery is guaranteed to be achieved. Note that Proposition 1 does not make any assumption on the vector set $\{\bar{d}_i : i \in \mathscr{K}\}$ 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-L2-Aut). One important implication of Proposition 1 is for the case where there is a Δ-spaced disturbance sequence with $\Delta = 2$, 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 [36]) Consider the convex optimization problem (CO-L2-Aut). If $\sum_{i\notin\mathscr{K}}|x_i| > \sum_{i\in\mathscr{K}}|x_i|$, then $\bar{A}$ is the unique solution to the problem.

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 Δ-spaced disturbance sequence with $\Delta \geq n+1$.

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

$$\text{span}\{\bar{d}_i, \bar{A}\bar{d}_i, \ldots, \bar{A}^{n-1}\bar{d}_i\} \in \mathbb{R}^n, \quad \forall i \in \mathscr{K} \tag{3}$$

is satisfied. Then, whenever $x_0 = 0$ or $\{0, \ldots, \Delta-1\} \notin \mathscr{K}$, $\bar{A}$ is a solution to the convex formulation (CO-L2-Aut) if $T \geq n+\Delta$, provided that

$$\left|\sum_{k_1+\ldots+k_n=\Delta-n} \bar{\lambda}(k_1,\ldots,k_n)\right| \leq \sum_{i=0}^{\Delta-n-1}\left|\sum_{k_1+\cdots+k_n=i} \bar{\lambda}(k_1,\ldots,k_n)\right|, \tag{4}$$

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

This result is a generalization of Proposition 1. The condition (3) is necessary to ensure that the KKT condition is satisfied, which guarantees that the disturbance vector excites the system to explore the entire system space in the next n time instances. In real-life applications, this condition can be attained by injecting a random small perturbation to the system. To gain insight into (4), which involves the product of eigenvalues, consider a special case where $\bar{A}$ has the eigenvalue λ with multiplicity n with n linearly independent eigenvectors. In this case, we can simplify (4) as follows. Define $k := \Delta - n$. Then, (4) is equivalent to

$$\binom{n+k-1}{k}|\lambda|^k - \sum_{i=0}^{k-1}\binom{n+i-1}{i}|\lambda|^i < 0.$$

This condition is satisfied if $|\lambda| \leq C_{n,k}$, where $C_{n,k}$ denotes the upper bound on the eigenvalue magnitudes given the parameters n and k. Fig. 1 summarizes the values of $C_{n,k}$ for different choices of n and k. Note that $C_{n,k} \leq C_{m,k}$ if $n > m$ and $C_{n,k} \leq C_{n,l}$ if $k < l$, due to the definition of $C_{n,k}$. It can be shown that $C_{1,k} \to 2$ as $k \to \infty$. As a result, $|\lambda| \leq C_{n,k} \leq C_{1,k} \to 2$. 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 $\Delta = 2n$, $|\lambda| < 1$ is sufficient for exact recovery. 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.

$C_{n,k}$	k=1	k=2	k=3	k=5	k=7	k=10
n=1	1.0000	1.6180	1.8393	1.9659	1.9920	1.9990
n=2	0.5000	1.0000	1.2886	1.5725	1.7010	1.7951
n=3	0.3300	0.7287	1.0000	1.3181	1.4892	1.6310
n=5	0.2000	0.4740	0.6938	1.0000	1.1956	1.3087
n=7	0.1429	0.3516	0.5320	0.8069	1.0000	1.1979
n=10	0.1000	0.2535	0.3944	0.6263	0.8036	1.0000

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

Remark 2: We choose the Δ-attack structure with uniform attacks as the strictest attack structure. Here, Δ can be generalized as the smallest interval over which there is no attack. The sufficient conditions hold for every Δ-time interval. If the sufficiency conditions are satisfied for the interval of length Δ, they will also be satisfied for intervals of length greater than Δ. In the latter case, we add additional terms to the summation for $i \notin \mathscr{K}$ in the KKT condition, which enlarges the set of possible values for the subgradients. In fact, this relaxes the condition of the KKT condition for the interval of length Δ.

B. PROBABILISTIC ATTACK STRUCTURE

Because the minimum value of Δ is 2, the deterministic attack structure does not allow the size of corrupted data to exceed the size of clean data. Thus, we investigate a probabilistic attack structure for which 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 other time periods. Specifically, 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. 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 assumptions throughout this section.

Assumption 1: Given an autonomous system $x_{i+1} = \bar{A}x_i + \bar{d}_i$ for $i = 0, \ldots, T-1$ with dimension n, assume that $x_0 = 0$ and all singular values of $\bar{A}$ are less than 1, i.e. $\|\bar{A}\|_{op} < 1$.

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. In addition, we make the following stealth attack assumption.

Assumption 2: For each $k \in \mathscr{K}$, the attack vector is defined as $\bar{d}_k := \bar{\ell}_k\bar{f}_k$ where $\bar{\ell}_k \in \mathbb{R}$ and $\bar{f}_k \in \mathbb{S}_2(1)$. $\bar{f}_k$ plays the role of the direction of the attack while $\bar{\ell}_k$ plays the role of the length. Define the filtration

$$\mathscr{F}_k := \sigma\{x_1, \ldots, x_k\}, \quad \forall k \in \{0, \ldots, T-1\}.$$

For all $k \in \mathscr{K}$, conditioning on $\mathscr{F}_k$, 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 $\mathbb{S}_2(1)$;
3) $\bar{\ell}_k$ is mean-zero and sub-Gaussian with parameter σ;
4) The variance of $\bar{\ell}_k$ is $\sigma_k^2 \in [c^2\sigma^2, \sigma^2]$ 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 $k' < k$. For example, a stealthy attack vector $\bar{d}_i$ could have the distribution $\mathscr{N}(0, \min\{c, \widehat{\|x_i\|_2}\})$ where $\widehat{\|x_i\|_2} = \sum_{k=0}^{i}(i+1)^{-1}\|x_i\|_2$ is the average norm of the states between time periods 0 and i. The magnitude of the attack vector is dependent on the past. We note that the above assumption of symmetry of the disturbance vectors reflected in $\bar{f}_k$ is not restrictive and corresponds to stealth attacks. If this does not hold, the attacks may be detectable, and their effects could be nullified, or the system could be stopped to investigate the possible influence from outside agents [37], [38]. To understand the notion of stealthy attack, consider the practice in power systems where the system operator performs some hypothesis testing on sensory data to detect anomaly before making any decisions using the data. If the mean of the data is not zero, it would not pass the test and therefore the attack would be isolated or nullified. If the symmetric assumption does not hold, there is an unavoidable bias in estimation [39].

It is not possible to obtain deterministic sample complexity for exact recovery due to the randomness in attack structure. Therefore, it is essential to quantify the required number of samples for the exact recovery with high probability. Under Assumption 2, the attack vector at time i, $\bar{d}_i$, has a sub-Gaussian distribution with parameter σ given $\mathscr{F}_i$. The sub-Gaussian 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 mild.

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

$$\exists \gamma_i \in \partial\|0\|_\circ, \quad \forall i \notin \mathscr{K} \quad \text{s.t.} \sum_{i\notin\mathscr{K}} x_i \otimes \gamma_i = \sum_{i\in\mathscr{K}} x_i \otimes \partial\|\bar{d}_i\|_\circ.$$

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 ℓ_∞ and ℓ_2 norms, that is $\mathbb{B}_\infty(1)/\sqrt{n} \subseteq \mathbb{B}_2(1)$. Additionally, we examine the results for each coordinate of the subdifferentials since they are separable due to the properties of the ℓ_∞ norm. Therefore, the following propositions provide sufficient conditions to satisfy the KKT conditions.

Proposition 3: The KKT conditions for the problem (CO-L2-Aut) and (CO-L1-Aut) are satisfied if there exist scalars $\gamma_i^l \in [-1, 1]$, $i \notin \mathscr{K}$, $l = 1, \ldots, n$ such that

$$\sum_{i\notin\mathscr{K}} \gamma_i^l x_i/\sqrt{n} = \sum_{i\in\mathscr{K}} \partial\|\bar{d}_i\|_2^l x_i, \quad \forall l = 1, \ldots, n \tag{5}$$

and

$$\sum_{i\notin\mathscr{K}} \gamma_i^l x_i = \sum_{i\in\mathscr{K}} \partial\|\bar{d}_i\|_1^l x_i, \quad \forall l = 1, \ldots, n. \tag{6}$$

Here, $\partial\|\bar{d}_i\|_\circ^l$ is the l-th element of the subgradient.

Because analyzing the conditions (5) and (6) directly is cumbersome, we investigate the equivalent condition provided in the lemma below, derived using Farkas' lemma [40] and the duality of linear programs.

Lemma 3: Given a matrix $\mathbf{F} \in \mathbb{R}^{n\times m}$ 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 $\mathbf{F}w = g$.
ii) For every $z \in \mathbb{R}^n$ with $\|z\|_2 = 1$, it holds that $f(z) := z^T g + \|z^T\mathbf{F}\|_1 \geq 0$.

It is important to notice that the conditions (5) and (6) amount to finding a vector for the set of equations in the form of $\mathbf{F}w = g$ where w is restricted as $\|w\|_\infty \leq 1$. Given a coordinate l, the matrix $\mathbf{F} \in \mathbb{R}^{n\times(T-|\mathscr{K}|)}$ associated with the conditions (5) and (6) is a matrix with columns $\frac{x_i}{\sqrt{n}}$ and x_i, and the vector $g \in \mathbb{R}^n$ is $\sum_{i\in\mathscr{K}} \partial\|\bar{d}_i\|_2^l x_i$ and $\sum_{i\in\mathscr{K}} \partial\|\bar{d}_i\|_1^l x_i$, respectively. Moreover, the vector $w \in \mathbb{R}^{T-|\mathscr{K}|}$ has the elements γ_i^l, $i \notin \mathscr{K}$ for both conditions. Hence, we study the condition ii) in Lemma 3. However, there are infinitely many points on the ℓ_2 unit circle $\mathbb{S}_2(1)$. In order to show that the function $f(z) = z^T g + \|z^T\mathbf{F}\|_1$ is non-negative at every point on the ℓ_2 unit circle, we employ the discretization technique that chooses a finite set of points. These points are chosen as the ϵ-cover of the unit ball.

Definition 2 (Covering Number [35]): ϵ-cover of the compact set $\mathbb{T}$ with respect to the norm ρ is a set $\{\theta^1, \theta^2, \ldots, \theta^N\} \subset \mathbb{T}$ such that for each $\theta \in \mathbb{T}$, there exists some $i \in \{1, \ldots, N\}$ with $\rho(\theta, \theta^i) \leq \epsilon$. The ϵ-covering number $\mathscr{N}(\epsilon, \mathbb{T}, \rho)$ 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 ϵ 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 [35]): Given an n-dimensional unit ball $\mathbb{B}(1)$ with the norm $\|\cdot\|$, the metric entropy of the unit ball can be upper bounded by

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

We show that the function $f(z)$ can be lower bounded by some positive number $\theta > 0$ at every point in the ϵ-cover of the unit ball with high probability, and that the function value inside the ϵ-ball does not change more than this positive number θ 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 sample complexity for the exact recovery grows with $n^2\log(n)$ 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 $\delta \in (0, 1]$, if the time horizon satisfies $T \geq \Theta(T_{\text{sample}})$, where T_{sample} is defined as

$$nR\left[n\log(nR) + \log\left(\frac{1}{\delta}\right)\right],$$

and

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

with ρ denoting the largest magnitude of the singular values of $\bar{A}$, then $\bar{A}$ is the unique solution to the convex optimization (CO-L2-Aut) with probability at least $1-\delta$.

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 assumptions of Theorem 2, if the time horizon T satisfies $T \geq \Theta(T_{\text{sample}})$, where T_{sample} is

$$R\left[n\log(nR) + \log\left(\frac{1}{\delta}\right)\right],$$

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

The proof of Theorem 3 is highly similar to that of Theorem 2. Because the conditions (5) and (6) differ by a factor of $\sqrt{n}$, 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. In addition, if the system is at the verge of instability, the sample complexity increases significantly. Even in the case when the probability p is close to 1, significantly more corrupt data than clean data, this result guarantees asymptotic exact recovery as long as there are a sufficient number of clean samples. We make the following remarks regarding various generalizations of the above results.

Remark 3: We note that the dependence on $p^{-1}(1-p)^{-1}$ is an artifact of the high probability bound. Specifically, this dependence ensures that the number of attacks is bounded by $\Theta(pT)$ with high probability. When p is very small or even zero, learning the system becomes a classic problem in control theory, where it is known that artificial noise (referred to as an excitation signal) must be added to the system in order to enable learning. There is a rich literature explaining why an excitation signal is necessary when a system is (nearly) deterministic. For instance, consider the system $x_{i+1} = \bar{A}x_i$. If x_0 is zero, then x_i will always remain zero, preventing us from identifying $\bar{A}$. To avoid this, we must excite the system as $x_{i+1} = \bar{A}x_i + w_i$, where w_i is, for example, Gaussian noise. When p is not close to zero, the adversarial attack serves as an excitation signal, helping in this regard.

Remark 4: When the system is initialized randomly at x_0 following Assumption 2, the results of Theorems 2 and 3 continue to hold. A system with a nonzero initial state x_0 at time 0 is equivalent to the system initialized at the origin at time -1 under the attack x_0 at time -1, namely, the new system could be constructed as $x_{-1} = 0$ and $\bar{d}_{-1} = x_0$, leading to the state value x_0 at time 0. Thus, the sample complexities of the theorems only shift by a constant.

Remark 5: Furthermore, we mention that the uniform distribution assumption of $\bar{f}_k$ can be relaxed to any distribution on the sphere with zero mean and full-rank covariance matrix. In that case, the sample complexity in Theorems 2–5 will depend on the conditional number of the covariance matrix.

V. SYSTEMS WITH INPUT SEQUENCE

It is desirable to understand the role of an input sequence in exact recovery. Since the input sequence is generated by a controller, one can design it in such a way that it accelerates the learning. In this 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 $\bar{B} \in \mathbb{R}^{n\times m}$. The goal is to obtain these matrices using the state trajectories and the sequence of inputs. We will investigate the estimators (CO-L2) and (CO-L1) defined earlier.

We choose the input vectors u_i to be Gaussian given $\mathscr{F}_i$. 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 is mild, and it is satisfied when u_i is designed in the linear feedback form as $u_i = Kx_i + \omega$. Conditioning on $\mathscr{F}_i$, the input is excited with Gaussian noise ω. Note that the closed loop system could be written as $x_{i+1} = (\bar{A} + \bar{B}K)x_i + \bar{B}\omega + \bar{d}_i$. Thus, the problem is equivalent to estimating the matrices $(\bar{A} + \bar{B}K)$ and $\bar{B}$ when the linear feedback control is used. Therefore, the most common input sequence used in optimal control satisfies this assumption. 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 $\gamma_i^l, \mu_i^l \in [-1, 1]$ for all $i \notin \mathscr{K}$, $l \in \{1, \ldots, n\}$ such that

$$\sum_{i\notin\mathscr{K}} \gamma_i^l x_i/\sqrt{n} = \sum_{i\in\mathscr{K}} \partial\|\bar{d}_i\|_2^l x_i, \quad \forall l = 1, \ldots, n, \tag{7}$$

and

$$\sum_{i\notin\mathscr{K}} \mu_i^l u_i/\sqrt{n} = \sum_{i\in\mathscr{K}} \partial\|\bar{d}_i\|_2^l u_i, \quad \forall l = 1, \ldots, n, \tag{8}$$

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

The proof of Proposition 4 relies on the same technique as in Proposition 3. Similar to autonomous systems, we can omit the factor $\sqrt{n}$ from above equations to guarantee the satisfaction of the KKT conditions for problem (CO-L1). We require the following controllability assumption.

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

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

Intuitively, the controllability of a non-autonomous system denotes the ability to move a system around in its entire state space using the input sequence $\{u_i\}_{i=0}^{T-1}$. 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.

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

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

and

$$T^2_{\text{sample}} := nR_2\left[m\log(nR_2) + \log\left(\frac{1}{\delta}\right)\right],$$

where

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

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

Here, constants $c \in (0, 1]$ and $\kappa \geq (1-\rho)^{-1}$ depend on m, n, σ, ξ and $\bar{B}$. If the time horizon satisfies the inequality $T \geq \Theta[\max\{T^1_{\text{sample}}, T^2_{\text{sample}}\}]$, then $(\bar{A}, \bar{B})$ is the unique solution to (CO-L2) with probability at least $1-\delta$.

The proof of Theorem 4 is deferred to the online version [41]. We have obtained a high probability bound for the exact recovery of the system matrices $\bar{A}$ and $\bar{B}$. The T^1_{sample} in the sample complexity corresponds to the satisfaction of the KKT conditions for the state measurements, whereas the T^2_{sample} corresponds to the satisfaction of the KKT conditions for the input sequence. Just like autonomous systems, the sample complexity increases as the probability of disturbances increases. 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 $1/[(1-\rho)^3\log^2(1/\rho)]$ to $1/[(1-\rho)^2\log^2(1/\rho)]$. 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.The following theorem studies problem (CO-L1).

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

$$R_1\left[n\log(nR_1)+\log\left(\frac{1}{\delta}\right)\right] \text{ and } R_2\left[m\log(nR_2)+\log\left(\frac{1}{\delta}\right)\right],$$

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

Remark 6: When the input sequence $u_i = Kx_i$ is used to control the system, 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 $\hat{B} = 0$. Still, 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. It could be avoided if the input is excited with a small noise in the form of $u_i = Kx_i + \omega$. Moreover, if all the input vectors u_i are set to zero, it is not possible to uniquely recover the system matrix $\bar{B}$. Because the input sequence is zero, the KKT conditions are trivially satisfied; thus, we have multiple optimum solutions.

Remark 7: The results in Sections IV and V can be extended to the case where there is a small-in-magnitude dense measurement noise, e_i, in addition to potentially large-in-magnitude adversarial noise d_i. Note that the estimator can be written as a constrained optimization problem as

$$\min_{\substack{A\in\mathbb{R}^{n\times n}, B\in\mathbb{R}^{n\times m},\\ d_i\in\mathbb{R}^n,\forall i}} \sum_{i=0}^{T-1}\|d_i\|_\circ$$

$$s.t. \quad x_{i+1} - Ax_i + Bu_i + d_i = 0, \quad i = 0, \ldots, T-1$$

Adding the dense measurement vector e_i is equivalent to perturbing the constraint from $x_{i+1} - Ax_i - Bu_i - d_i = 0$ to $x_{i+1} - Ax_i - Bu_i - d_i = e_i$. This implies that the optimal solution will be perturbed as well. Different results are readily available on how to calculate the change in the optimal solution (Theorem 2 in [42]).

VI. NUMERICAL EXPERIMENT

We conduct numerical experiments with synthetically generated dynamical systems and a real-life biomedical application involving insulin injections.

A. SYNTHETIC SIMULATIONS

We generate LTI dynamical systems to verify the theoretical results. For each experiment, we generate 10 different random matrices $\bar{A}$ and $\bar{B}$ with singular values uniformly distributed between $(-1, 1)$. We generate the trajectory of the system, $\{x_i\}_{i=0}^{T}$, from a system initialized at the origin by using the

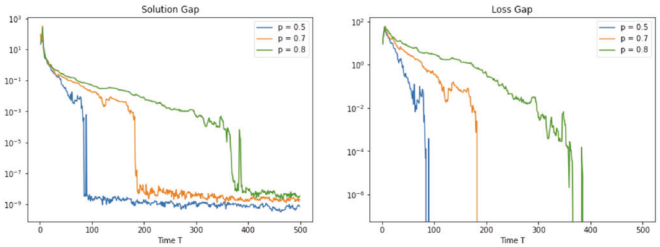

FIGURE 2. Performance of (CO-L2-Aut) with Probability of Attacks $p \in \{0.5, 0.7, 0.8\}$ with $n = 5$.

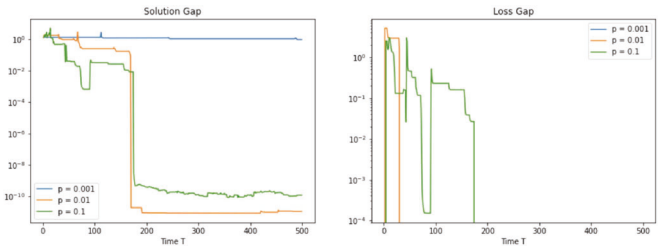

FIGURE 3. Performance of (CO-L2-Aut) with Probability of Attacks $p \in \{0.001, 0.01, 0.1\}$ with $n = 5$.

disturbance vector $\bar{d}_i = \ell_i \hat{f}_i$ where

$$\ell_i \sim \mathcal{N}(0, \min\{100/n, 100\widehat{\|x_i\|_2}\}), \quad \hat{f}_i \sim \text{Uniform}(\mathbb{S}^{n-1})$$

whenever $i \in \mathscr{K}$ and using i.i.d. zero mean and Gaussian input vectors u_i with the covariance matrix I_m/m. We solve the following optimization problem for every time period t between $[1, T]$ using the CVX solver

$$(\hat{A}_t, \hat{B}_t) = \arg \min_{A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}} f_t(A, B) = \sum_{i=0}^{t-1} \|x_{i+1} - Ax_i - Bu_i\|_\circ$$

For any time period $t = 1, \ldots, T$, we obtain the solution gap, $\|(\hat{A}_t, \hat{B}_t) - (\bar{A}, \bar{B})\|_F$, and the loss gap, $f_t(\hat{A}_t, \hat{B}_t) - f_t(\bar{A}, \bar{B})$. We plot the trajectory of the average solution gap and the loss gap of the 10 independent simulation runs. First, we analyze the performance of the estimator with respect to the probability of having an attack for the autonomous systems. We use three different values of $p \in \{0.5, 0.7, 0.8\}$. In Fig. 2, we report the results for the estimator (CO-L2-Aut). In this case, as the probability of attack p increases, the number of required samples grows. This aligns with the theoretical results since the sample complexity scales with $(1 - p)^{-2}$.

Moreover, we conduct an experiment with a very small probability of having an attack, namely the probability of attack p being equal to $\{0.001, 0.01, 0.1\}$. Unsurprisingly, an extremely small excitation or probability of attack, such as $p = 0.001$, leads to the failure of exact recovery in $T = 500$ time periods in Fig. 3 because of the lack of excitation.

In addition, we test the impact of the dimension of the system using the estimator (CO-L2). Setting $p = 0.5$, we create systems with dimensions $(n, m) \in \{(5, 5), (10, 10), (15, 15)\}$

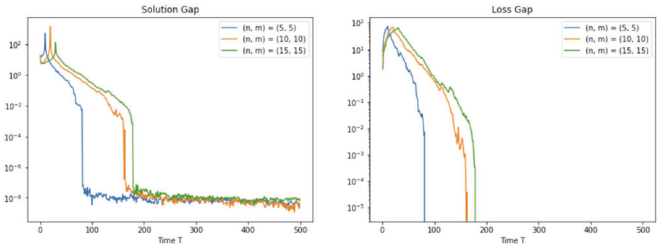

FIGURE 4. Performance of (CO-L2) with Dimensions $(n, m) \in \{(5, 5), (10, 10), (15, 15)\}$ with $p = 0.5$.

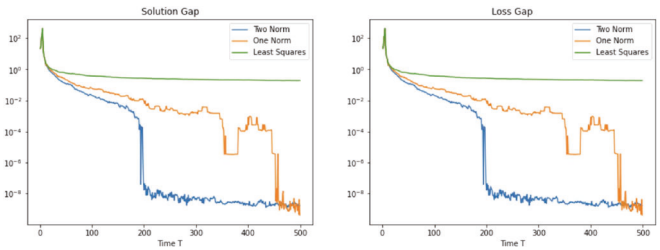

FIGURE 5. Performance of (CO-L2-Aut), (CO-L1-Aut), and Least-Squares with $n = 5$ and $p = 0.7$.

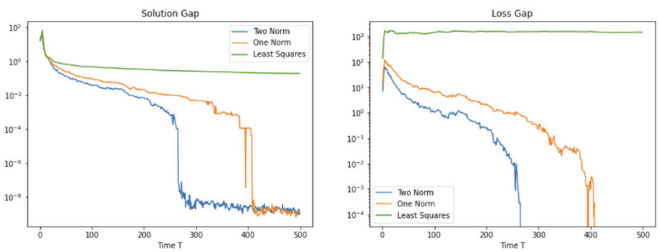

FIGURE 6. Performance of (CO-L2), (CO-L1) and Least-Squares with $(n, m) = 5$ and $p = 0.7$.

for (CO-L2). In Fig. 4, it is observed that the sample complexity for exact recovery grows with the dimension of the system. Theoretically, T grows with roughly $n \log(n)$ when the term R in the sample complexity bound is dominated by the terms that scales with n^{-1}. The smaller empirical sample complexity hints stricter lower bounds since the theoretical results are only sufficient conditions and are derived for the worst-case scenario.

Moreover, we test the relationship between the sample complexities of the estimators with ℓ_2, ℓ_1 norms, and the least-squares method. Figs. 5 and 6 show the performance for autonomous systems with $n = 5$ and $p = 0.7$, and non-autonomous systems with $(n, m) = (5, 5)$ and $p = 0.7$, respectively. The solution gap and loss gap plateau for the least-squares estimator, whereas the ℓ_2 and ℓ_1 norm estimators successfully learn the ground truth system matrices. Since the attack vectors themselves are not sparse, the ℓ_2 estimator requires fewer number of samples to achieve exact recovery than the ℓ_1 estimator.

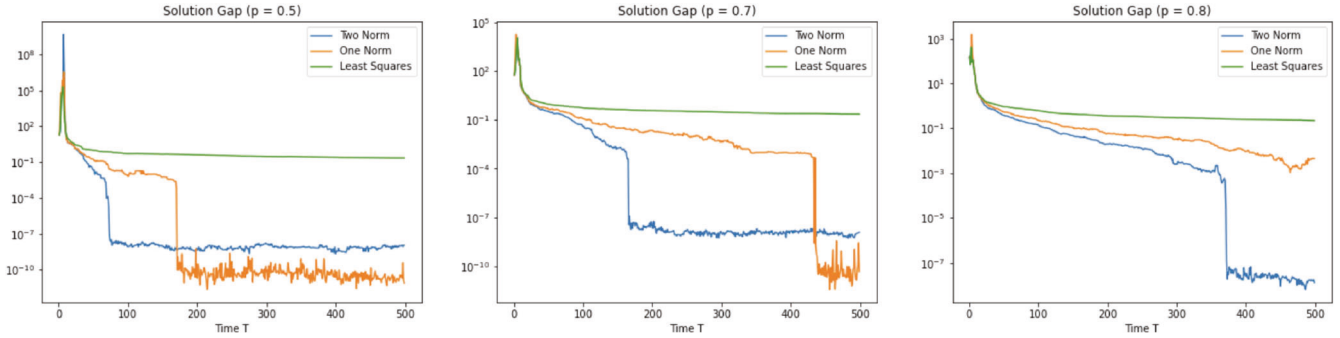

FIGURE 7. Solution Gap for (CO-L2), (CO-L1) and Least-Squares with $p \in \{0.5, 0.7, 0.8\}$ for the Insulin Application.

B. BIOMEDICAL APPLICATION

We conduct a numerical experiment inspired by biomedical applications to demonstrate results for a real-life biomedical application. We consider a compartmental model of blood sugar and insulin dynamics in the human body, as described in [43]. 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 [44]:

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

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 glucose distribution/transport, glucose disposal, and endogenous production, respectively. S_1 and S_2 represent the absorption rate of insulin. 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 represent other effects not captured by the model. 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.

We discretize the continuous-time system to obtain an LTI system using $\Delta_i = 0.5$. The resulting matrix $\bar{A}$ is stable. Our objective is to estimate the parameters $(k_{ai}, k_{bi}, t_{\max,I}, V_I, k_e)$. The attack vectors are modeled using the same distribution as in the synthetic simulations. We run our model with the probability of an attack being $p = 0.5$, $p = 0.7$, and $p = 0.8$. We report the solution gap for the least-squares estimator, problem (CO-L2), and problem (CO-L1).

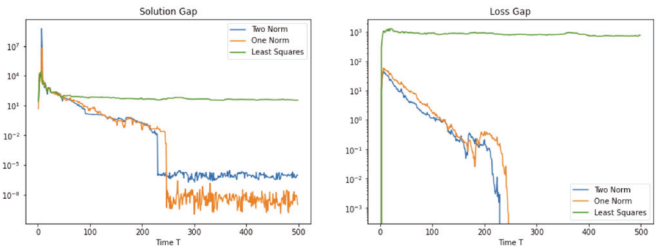

FIGURE 8. Performance of (CO-L2), (CO-L1) and Least-Squares with $p = 0.7$ for the Insulin Application with Sparse Vector Injections.

Fig. 7 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. Note that there are more corrupted data than clean data in the case of $p = 0.7$ and $p = 0.8$. 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 disturbances. When the disturbances are sparse, d_1, d_2, d_3, d_5 are set to zero. Fig. 8 shows that the ℓ_2 and ℓ_1 estimators 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 conditions for the exact recovery of the system dynamics. When the attacks occur deterministically every Δ period, exact recovery is possible after $n + \Delta$ 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 $\Theta((1-p)^{-2})$ and a polynomial in the dimension of the problem. These findings were supported by

a numerical experiments. This work provides the first set of mathematical guarantees for the robust non-asymptotic analysis of dynamic systems.

APPENDIX

A. PROOF OF PROPOSITION 1

Let $i_1, i_2, \dots$ be the set of attack times over time horizon, i.e. $\mathcal{K} = \{i_1, i_2, \dots\}$. We show that sufficient conditions satisfied for every attack interval $[i_k + 1, i_{k+1}], \forall k \geq 1$. Due to Δ-spaced attack model, we have $i_1 \leq \Delta$. We can utilize Lemma 2 to show that $\bar{A}$ is the unique solution.

Case 1: $x_0 = 0$

We have $x_i = 0$ for $i = 0, 1, \dots, i_1$. As a result, we will show that the condition of Lemma 2 holds for every time period in the time intervals $[i_k + 1, i_{k+1}], \forall k \geq 1$, where $i_{k+1} = i_k + \Delta$. For any such interval with $k \geq 1$, the following holds

$$\sum_{i=i_k+1}^{i_k+\Delta-1} |x_i| - |x_{i_{k+1}}| = \sum_{i=i_k+1}^{i_k+\Delta-1} |x_i| - |\bar{A}||x_{i_k+\Delta-1}|$$
$$= \sum_{i=i_k+1}^{i_k+\Delta-2} |x_i| + (1 - |\bar{A}|)|x_{i_k+\Delta-1}| > 0$$

The last statement is positive because $|\bar{A}| < 1$ and $\Delta \geq 2$. Note that $\sum_i^j |x_i| = 0$ whenever $j < i$. The condition $\sum_{i\notin\mathcal{K}} |x_i| - \sum_{i\in\mathcal{K}} |x_i| > 0$ holds after the first attack time period. Thus, whenever $T \geq i_1 + 1 \geq \Delta + 1$, exact recovery is achieved.

Case 2: $0 \notin \mathcal{K}$ and $x_0 \neq 0$

Because $0 \notin \mathcal{K}$, we have $i_1 \geq 1$. From the above statements, we know that the sufficient condition holds for every time interval $[i_k + 1, i_{k+1}], \forall k \geq 1$. Hence, we only need to show that the condition in Lemma 2 is satisfied for the time interval $[0, i_1]$ as well. It is apparent that

$$\sum_{i=0}^{i_1-1} |x_i| - |x_{i_1}| = \sum_{i=0}^{i_1-1} |x_i| - |\bar{A}||x_{i_1-1}|$$
$$= \sum_{i=0}^{i_1-2} |x_i| + (1 - |\bar{A}|)|x_{i_1-1}| > 0.$$

The last statement is positive because $|\bar{A}| < 1$ and $i_1 \geq 1$.

B. PROOF OF PROPOSITION 2

By using KKT conditions, $\bar{A}$ is a solution to the problem if and only if

$$0 \in \sum_{i\notin\mathcal{K}} x_i \otimes \partial\|0\|_2 + \sum_{i\in\mathcal{K}} x_i \otimes \partial\|\bar{d}_i\|_2. \tag{9}$$

Let i_1 be the time stamp of the first attack time. Then, we have $i_1 \leq \Delta$ due to Δ-attack structure and the assumptions in the theorem. The set of attack times is $\mathcal{K} = \{i_1, i_1 + \Delta, i_1 + 2\Delta, i_1 + 3\Delta, \dots\}$. Since $x_0 = 0$, we have $x_i = 0$ whenever $i = 0, 1, \dots, i_1$ and $x_{i_1+1} = \bar{d}_{i_1}$. Let $T = \Delta + i_1$, i.e., the time step at which a cycle of disturbance is completed. In this case, the sufficient condition using KKT condition can be written as

$$0 \in \sum_{t=1}^{\Delta-1} x_{i_1+t} \otimes \partial\|0\|_2 + x_{i_1+\Delta} \otimes \partial\|\bar{d}_{i_1+\Delta}\|_2$$
$$\in \sum_{t=0}^{\Delta-2} \bar{A}^t \bar{d}_{i_1} \otimes \partial\|0\|_2 + \bar{A}^{\Delta-1}\bar{d}_{i_1} \otimes \frac{\bar{d}_{i_1+\Delta}}{\|\bar{d}_{i_1+\Delta}\|}.$$

The matrix 0 may belong to the right-hand side term for arbitrary $\bar{d}_{i_1+\Delta}$ if $\bar{d}_{i_1+\Delta} \in span\{\bar{d}_{i_1}, \bar{A}\bar{d}_{i_1}, \dots, \bar{A}^{\Delta-2}\bar{d}_{i_1}\}$. This is satisfied by the assumption in the proposition statement. However, this is not sufficient because the vectors chosen for $\partial\|0\|_2$ have a bounded norm. Therefore, we need to bound the norm of the columns of $\bar{A}^{\Delta-1}\bar{d}_{i_1} \otimes \frac{\bar{d}_{i_1+\Delta}}{\|\bar{d}_{i_1+\Delta}\|_2}$, so it can be expressed as a linear combination of the vectors $\{\bar{d}_{i_1}, \bar{A}\bar{d}_{i_1}, \dots, \bar{A}^{\Delta-2}\bar{d}_{i_1}\}$. Let (λ_j, v_j) be eigenvalue-eigenvector pairs for the matrix $\bar{A}^T$. Let $e_1, \dots, e_{\Delta-1} \in \partial\|0\|_2$. Then, the KKT condition can be written as follows:

$$0 \in e_1\bar{d}^T + e_2\bar{d}^T\bar{A}^T + \dots + e_{\Delta-1}\bar{d}^T(\bar{A}^T)^{\Delta-2} + f\bar{d}^T(\bar{A}^T)^{\Delta-1},$$

where $f = \frac{\bar{d}_{i_1+\Delta}}{\|\bar{d}_{i_1+\Delta}\|_2}$ and $\|f\|_2 = 1$. If we multiply the equation above by the eigenvector v_j of $\bar{A}^T$, we obtain

$$0 \in e_1\bar{d}^T v_j + \dots + e_{\Delta-1}\bar{d}^T(\bar{A}^T)^{\Delta-2}v_j + f\bar{d}^T(\bar{A}^T)^{\Delta-1}v_j$$
$$\in (e_1 + \lambda_j e_2 + \dots + \lambda_j^{\Delta-2}e_{\Delta-1} + \lambda_j^{\Delta-1}f)\bar{d}^T v_j.$$

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

$$0 = e_1 + \lambda_j e_2 + \dots + \lambda_j^{\Delta-2}e_{\Delta-1} + \lambda_j^{\Delta-1}f, \quad \forall j = 1, \dots, n.$$

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

$$\sum_{k_1+\dots+k_n=\Delta-n} \lambda(k_1, \dots, k_n)f$$
$$= \sum_{i=0}^{\Delta-n-2} \sum_{k_1+\dots+k_n=i} \lambda(k_1, \dots, k_n)e_{i+n+1}.$$

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

$$\left|\sum_{k_1+\dots+k_n=\Delta-n} \lambda(k_1, \dots, k_n)\right| \|f\|_2$$
$$\leq \sum_{i=0}^{\Delta-n-1} \left|\sum_{k_1+\dots+k_n=i} \lambda(k_1, \dots, k_n)\right| \|e_{i+n+1}\|_2.$$

Using the fact that $\|e_j\|_2 = 1$ for all j and $\|f\|_2 = 1$, we obtain

$$\left|\sum_{k_1+\dots+k_n=\Delta-n} \lambda(k_1, \dots, k_n)\right| \leq \sum_{i=0}^{\Delta-n-1} \left|\sum_{k_1+\dots+k_n=i} \lambda(k_1, \dots, k_n)\right|.$$

Moreover, if x_0 is not 0, then we can show that KKT condition is satisfied for the interval $\{0, 1, \ldots, \Delta\}$. This completes the proof for the proposition.

C. PROOF OF PROPOSITION 3

The KKT condition for the exact recovery is

$$\exists \gamma_i \in \partial\|0\|_\circ, i \notin \mathscr{K} \text{ s.t. } \sum_{i\notin\mathscr{K}} x_i \otimes \gamma_i = \sum_{i\in\mathscr{K}} x_i \otimes \partial\|\bar{d}_i\|_\circ. \tag{10}$$

For (CO-L2-Aut) with $\circ = 2$. the condition (10) becomes

$$\exists \gamma_i \in \partial\|0\|_2, i \notin \mathscr{K} \text{ s.t. } \sum_{i\notin\mathscr{K}} x_i \otimes \gamma_i = \sum_{i\in\mathscr{K}} x_i \otimes \partial\|\bar{d}_i\|_2.$$

Since $\partial\|0\|_1/\sqrt{n} = \mathbb{B}_\infty/\sqrt{n}(1) \subseteq \mathbb{B}_2(1) = \partial\|0\|_2$, we can rewrite it as

$$\exists \gamma_i \in \partial\|0\|_1, i \notin \mathscr{K} \text{ s.t. } \sum_{i\notin\mathscr{K}} \frac{x_i}{\sqrt{n}} \otimes \gamma_i = \sum_{i\in\mathscr{K}} x_i \otimes \partial\|\bar{d}_i\|_2.$$

We can check the condition at each coordinate because the set $\mathbb{B}_\infty(1)$ is coordinate wise separable. Thus, KKT condition holds for (CO-L2-Aut) if there exist scalars $\gamma_i^l \in [-1, 1]$, $i \notin \mathscr{K}, l = 1, \ldots, n$ such that

$$\sum_{i\notin\mathscr{K}} \gamma_i^l x_i/\sqrt{n} = \sum_{i\in\mathscr{K}} \partial\|\bar{d}_i\|_2^l x_i, \quad \forall l = 1, \ldots, n,$$

where $\partial\|\bar{d}_i\|_\circ^l$ is the l-th element of the subgradient. Similar algebraic manipulation can be done for (CO-L1-Aut).

D. PROOF OF LEMMA 3

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

$$\max_{w\in\mathbb{R}^m} \quad 0 \quad \text{s.t.} \quad \mathbf{F}w = g, \quad \|w\|_\infty \le 1.$$

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

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

or equivalently,

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

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

E. PROOF OF THEOREM 2

Since $x_0 = 0$, x_i can be expressed as

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

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 $l \in \{1, \ldots, n\}$, the optimality condition for the recovery of $\bar{A}$ is equivalent to

$$f(z) := z^T g + \|z^T\mathbf{F}\|_1 \ge 0, \quad \forall z \in \mathbb{S}_2(1), \tag{11}$$

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

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

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

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

We do the proof in multiple steps.

1) SHOWING $f(z) > 0$ FOR ANY GIVEN $z \in \mathbb{S}_2(1)$

We first prove that condition (11) holds with high probability for a fixed $z \in \mathbb{S}_2(1)$.

a) Analysis of the term $\|z^T\mathbf{F}\|_1$:

$$\mathbb{E}\|z^T\mathbf{F}\|_1 = \frac{1}{\sqrt{n}} \sum_{i\notin\mathscr{K}} \mathbb{E}\left|\sum_{k\in\mathscr{K}} z^T\bar{A}^{(i-k-1)_+}\bar{d}_k\right|. \tag{12}$$

We construct the index set

$$\mathscr{I}_1 := \{i \mid i \notin \mathscr{K},\ i-1 \in \mathscr{K}\}.$$

Let

$$S := \left\lceil \log_\rho \Theta\left[\frac{c^5}{\log(|\mathscr{I}_1|/\delta)}\right]\right\rceil = \Theta\left[\frac{\log\log(|\mathscr{I}_1|/\delta) + \log(1/c)}{\log(1/\rho)}\right],$$

where $\lceil x\rceil$ is the minimal integer that is not smaller than x and $\delta \in (0, 1)$ is the specified probability. We construct a subset of $\mathscr{I}_1$ in the following way:

$$\mathscr{I} := \{i_1, \ldots, i_I \mid i_j \in \mathscr{I}_1,\ i_j - i_{j-1} \ge S,\ \forall j\}.$$

It is straightforward to construct $\mathscr{I}$ such that

$$I = |\mathscr{I}| \ge \frac{1}{S}|\mathscr{I}_1|.$$

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

$$|\mathscr{I}_1| \ge \frac{p(1-p)T}{2}.$$

Therefore, we have an estimate on the size of $\mathscr{I}$:

$$\mathbb{P}\left(I \ge \frac{p(1-p)T}{2S}\right) \ge 1 - \exp[-\Theta[p(1-p)T]]. \tag{13}$$

For each $j \in \{1, \dots, I\}$, we define $\mathscr{K}_j := \{k \in \mathscr{K} \,|\, i_{j-1} < k < i_j\}$, where we denote $i_0 := -1$. Moreover, we define

$$X_{j,\ell} := \sum_{k \in \mathscr{K}_j} z^T \bar{A}^{i_\ell - k - 1} \bar{d}_k, \quad \forall j, \ell \in \{1, \dots, I\} \quad \text{s.t. } j \le \ell.$$

Using (12), we can calculate that

$$\begin{aligned} \|z^T \mathbf{F}\|_1 &\ge \frac{1}{\sqrt{n}} \sum_{\ell=1}^{I} \left| \sum_{j=1}^{\ell} X_{j,\ell} \right| \\ &\ge \frac{1}{\sqrt{n}} \sum_{j=1}^{I} \left(|X_{j,j}| - \sum_{\ell=j+1}^{I} |X_{j,\ell}| \right). \end{aligned} \tag{14}$$

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

$$\mathbb{P}\left(|X| \ge \tilde{\sigma}_X\right) \ge \frac{\tilde{\sigma}_X^4}{64 \sigma_X^4}.$$

For all $j \in \{1, \dots, I\}$, Assumption 2 implies that the standard deviation and the sub-Gaussian parameter of $X_{j,\ell}$ are

$$\tilde{\sigma}_{j,\ell} := \sqrt{\frac{1}{n} \sum_{k \in \mathscr{K}_j} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 \sigma_k^2},$$

$$\sigma_{j,\ell} := \sqrt{\frac{1}{n} \sum_{k \in \mathscr{K}_j} \|z^T \bar{A}^{i_\ell - k - 1}\|_2^2 \sigma^2},$$

respectively. It follows from Lemma 5 that

$$\mathbb{P}\left(|X_{j,j}| \ge \tilde{\sigma}_{j,j}\right) \ge \frac{\tilde{\sigma}_{j,j}^4}{64 \sigma_{j,j}^4},$$

which further leads to

$$\mathbb{P}\left(|X_{j,j}| \ge c \sigma_{j,j}\right) \ge \frac{c^4}{64}. \tag{15}$$

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

$$\sum_{\ell=j+1}^{I} \sigma_{j,\ell} \le \sum_{\ell=j+1}^{I} \rho^{(\ell-j)S} \sigma_{j,j} \le \frac{\rho^S}{1-\rho^S} \sigma_{j,j}.$$

Therefore, it holds with probability at least $1 - \delta/(4I)$ that

$$\begin{aligned} -\sum_{\ell=j+1}^{I} |X_{j,\ell}| &\ge -\frac{\rho^S}{1-\rho^S} \sigma_{j,j} \cdot \sqrt{2 \log(4I/\delta)} \\ &\ge -\frac{\rho^S}{1-\rho^S} \sigma_{j,j} \cdot \sqrt{2 \log(4|\mathscr{I}_1|/\delta)} \\ &\ge -\frac{c^4}{512} \cdot c \sigma_{j,j}, \end{aligned} \tag{16}$$

where the last step is by the choice of S. Using the bound in (13), if we choose

$$T \ge \Theta\left(\frac{\log\log(1/\delta) + \log(1/c)}{p(1-p)c^4 \log(1/\rho)} \right),$$

it holds with high probability that

$$\frac{c^4}{64} - \frac{\delta}{4I} \ge \frac{c^4}{128}.$$

Note that we have dropped the $|\mathscr{I}_1|$ term in the definition of S since $\log\log(|\mathscr{I}_1|)$ 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 $|X_{j,j}| - \sum_{\ell=j+1}^{I} |X_{j,\ell}|$.

We define the indicator function

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

Since the value of the Bernoulli random variable $\mathbf{1}_j$ only depends on attacks in $\mathscr{K}_j$, which are disjoint from each other, the random variables

$$\mathbf{1}_1 - c^4/128, \ \dots, \ \mathbf{1}_I - c^4/128$$

form a martingale sequence with respect to filtration $\mathscr{F}_{i_1}, \dots, \mathscr{F}_{i_I}$. For all $j \in \{1, \dots, I\}$, we can calculate that

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

By the tower property of expectation, we have

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

For conditional probabilities and expectations of a random variable X given another random variable Y, we use the notation $\mathbb{E}_Y[X] := \mathbb{E}[X|Y]$ and $\mathbb{P}_Y(X) := \mathbb{P}(X|Y)$ for the remainder of the proof. Therefore, $\mathbb{E}_I[\cdot]$ denotes the conditional expectation of the term given the value of the random variable I. Therefore, by applying Chernoff's bound and choosing $s := -\log(2)$, it follows that

$$\begin{aligned} \mathbb{P}_I \left(\sum_{j=1}^{I} \mathbf{1}_j \le \frac{c^4}{256} \cdot I \right) &\le \exp\left[-\frac{c^4 I}{256} \cdot s + \frac{c^4 I}{128} \left(e^s - 1\right) \right] \\ &= \exp\left[-\Theta\left(\frac{c^4}{128} \cdot I \right) \right]. \end{aligned}$$

Equivalently, we know

$$\mathbb{P}_I \left(\sum_{j=1}^{I} \mathbf{1}_j \ge \frac{c^4}{256} \cdot I \right) \ge 1 - \exp\left[-\Theta\left(\frac{c^4}{128} \cdot I \right) \right]. \tag{17}$$

Furthermore, since $i_j - 1 \in \mathscr{K}_j$, we can estimate that

$$\sigma_{j,j} \ge \sqrt{\frac{1}{n} \|z\|_2^2 \sigma^2} = \frac{1}{\sqrt{n}} \sigma.$$

By the definition of q_j and $\mathbf{1}_j$, when the event in inequality (17) happens, inequalities (15) and (16) imply that

$$\begin{aligned}\|z^T\mathbf{F}\|_1 &\geq \frac{1}{\sqrt{n}}\sum_{j=1}^{I}\left(|X_{j,j}| - \sum_{\ell=j+1}^{I}|X_{j,\ell}|\right)\\ &\geq \frac{1}{\sqrt{n}}\sum_{j=1}^{I}\left[\frac{c^4}{256}\cdot c\sigma_{j,j} - \frac{c^4}{512}\cdot c\sigma_{j,j}\right] \geq \frac{c^5\sigma}{512n}\cdot I\end{aligned}$$

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

$$\mathbb{P}_I\left[\|z^T\mathbf{F}\|_1 \geq \frac{c^5\sigma}{512n}\cdot I\right] \geq 1-\exp\left[-\Theta\left(c^4 I\right)\right] - \frac{\delta}{4}. \quad (18)$$

b) Upper bounding of the term $z^T g$:

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

Define the filtration $\mathscr{F}^f := \sigma\{\bar{f}_i, i\in\mathscr{K}\}$. By the stealth assumption, for each $k\in\mathscr{K}$, conditional on $\mathscr{F}_k$ and $\mathscr{F}^f$, we have

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

Let T' be the second last time instance in $\mathscr{K}$. We have

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda\sum_{i\in\mathscr{K}}\sum_{k\in\mathscr{K}} z^T\bar{A}^{(i-k-1)_+}\bar{d}_k\cdot\bar{f}_i^l\right)\right]\\ &= \mathbb{E}\left[\exp\left(\lambda\sum_{k\in\mathscr{K},k<T'}\sum_{i\in\mathscr{K}} z^T\bar{A}^{(i-k-1)_+}\bar{d}_k\cdot\bar{f}_i^l\right)\right.\\ &\left.\times\ \mathbb{E}\left[\exp\left(\lambda\sum_{i\in\mathscr{K}} z^T\bar{A}^{(i-1-T')_+}\bar{d}_T'\cdot\bar{f}_i^l\right)\middle|\mathscr{F}_{T'},\mathscr{F}^f\right]\right].\end{aligned} \quad (19)$$

Using the decomposition in Assumption 2, we have

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda\sum_{i\in\mathscr{K}} z^T\bar{A}^{(i-1-T')_+}\bar{d}_{T'}\cdot\bar{f}_i^l\right)\middle|\mathscr{F}_{T'},\mathscr{F}^f\right]\\ &\leq \exp\left[\frac{\lambda^2\sigma^2}{2}\left(\sum_{i\in\mathscr{K}} z^T\bar{A}^{(i-1-T')_+}\bar{f}_{T'}\bar{f}_i^l\right)^2\right].\end{aligned}$$

Substituting back into (19) and continuing the process for all $k\in\mathscr{K}$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda\sum_{i\in\mathscr{K}}\sum_{k\in\mathscr{K}} z^T\bar{A}^{(i-k-1)_+}\bar{d}_k\cdot\bar{f}_i^l\right)\right]\\ &\leq \mathbb{E}\left[\exp\left[\frac{\lambda^2\sigma^2}{2}\sum_{k\in\mathscr{K}}\left(\sum_{i\in\mathscr{K}}\left|z^T\bar{A}^{(i-1-k)_+}\bar{f}_k\right|\right)^2\right]\right], \quad (20)\end{aligned}$$

where the last inequality holds because $\bar{f}_i^l$ is bounded in $[-1,1]$. For each $i,k\in\mathscr{K}$, the value of $(z^T\bar{A}^{(i-1-k)_+}\bar{f}_k)^2$ concentrates around its expectation $\|z^T\bar{A}^{(i-1-k)_+}\|_2^2/n$. Therefore, inequality (20) leads to

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda\sum_{i\in\mathscr{K}}\sum_{k\in\mathscr{K}} z^T\bar{A}^{(i-k-1)_+}\bar{d}_k\cdot\bar{f}_i^l\right)\right]\\ &\leq \exp\left[\Theta\left[\frac{\lambda^2\sigma^2}{2n}\sum_{k\in\mathscr{K}}\left(\sum_{i\in\mathscr{K}}\rho^{(i-k-1)_+}\right)^2\right]\right]. \quad (21)\end{aligned}$$

Suppose the elements in $\mathscr{K}$ are

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

Define

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

We can calculate that

$$\sum_{i\in\mathscr{K}}\rho^{(i-1-j_k)_+} \leq \frac{\rho^{\Delta_k}}{1-\rho}.$$

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

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

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

$$\sum_{k=2}^{|\mathscr{K}|}\rho^{2\Delta_k} \lesssim \frac{|\mathscr{K}|p}{1-(1-p)\rho^2} \leq \frac{|\mathscr{K}|p}{1-\rho}.$$

Hence, inequality (21) implies that with the same probability, $z^T g$ is sub-Gaussian with parameter

$$\Theta\left[\sqrt{\frac{\sigma^2}{n}\sum_{k=2}^{|\mathscr{K}|}\frac{\rho^{2\Delta_k}}{(1-\rho)^2}}\right] \leq \Theta\left[\sqrt{\frac{|\mathscr{K}|p\sigma^2}{n(1-\rho)^3}}\right].$$

Therefore, Hoeffding's inequality leads to

$$\mathbb{P}_{|\mathscr{K}|}\left[z^T g \leq -\Theta\left(\sqrt{\frac{|\mathscr{K}|p\sigma^2}{n(1-\rho)^3}\log\left(\frac{4}{\delta}\right)}\right)\right] \leq \frac{\delta}{4}. \quad (22)$$

By combining inequalities (18) and (22), it holds with probability at least

$$1-\exp\left[-\Theta(c^4 I)\right] - \frac{\delta}{2}$$

that

$$f(z) \geq \Theta\left[\frac{c^5\sigma I}{n} - \sqrt{\frac{|\mathscr{K}|p\sigma^2}{n(1-\rho)^3}\log\left(\frac{1}{\delta}\right)}\right].$$

Similar to the bound in (13), it holds with probability at least $1-\exp[-\Theta(pT)]$ that

$$|\mathscr{K}| \leq 2pT.$$

As a result, if we choose

$$T \geq \Theta\Bigg[\max\Bigg\{\frac{\log\log(1/\delta)+\log(1/c)}{c^4 p(1-p)\log(1/\rho)}\log\left(\frac{1}{\delta}\right), \frac{1}{p(1-p)}\log\left(\frac{1}{\delta}\right), \frac{n\log(1/c)^2}{c^{10}(1-p)^2(1-\rho)^3\log^2(1/\rho)}\log\left(\frac{1}{\delta}\right)\Bigg\}\Bigg] = \Theta\left[nR\log\left(\frac{1}{\delta}\right)\right], \tag{23}$$

where

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

we have

$$\mathbb{P}_I\left[f(z) \geq \Theta\left(\frac{c^5\sigma I}{n}\right)\right] \geq 1-\delta. \tag{24}$$

F. USING DISCRETIZATION BOUND OVER THE UNIT SPHERE

In the second step, we apply discretization techniques to prove that condition (11) holds for all $z \in \mathbb{S}_2(1)$ with high probability. Suppose that $\epsilon > 0$ is a small constant. We construct an ϵ-cover of the unit sphere $\mathbb{S}_2(1)$, denoted as

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

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

$$\log(N) \leq \log[\mathcal{N}(\epsilon, \mathbb{S}_2(1), \|\cdot\|_2)] \leq n\log\left(1+\frac{2}{\epsilon}\right).$$

Define a to be the lower bound of $f(z)$ in inequality (24). Then, we have

$$a = \Theta\left(\frac{c^5\sigma I}{n}\right).$$

Our goal is to prove that

$$f(z) - f(z') \geq -a, \quad \forall z, z' \in \mathbb{S}_2(1) \text{ s.t. } \|z - z'\|_2 \leq \epsilon$$

holds with high probability. Notice that

$$\begin{aligned} f(z) - f(z') &= (z-z')^T g + (\|z^T\mathbf{F}\|_1 - \|(z')^T\mathbf{F}\|_1) \\ &\geq (z-z')^T g - \|(z-z')^T\mathbf{F}\|_1 \\ &\geq -\|z-z'\|_2\|g\|_2 - \|z-z'\|_2\sum_{i\notin\mathscr{K}}\|\mathbf{F}^i\|_2 \\ &\geq -\epsilon\Bigg(\left\|\sum_{i\in\mathscr{K}}\sum_{k\in\mathscr{K}}\bar{A}^{(i-k-1)_+}\bar{d}_k\right\|_2 + \frac{1}{\sqrt{n}}\sum_{i\notin\mathscr{K}}\left\|\sum_{k\in\mathscr{K}}\bar{A}^{(i-k-1)_+}\bar{d}_k\right\|_2\Bigg) \\ &\geq -\epsilon\sum_{k\in\mathscr{K}}\sum_{i>k}\rho^{(i-k-1)}|\bar{\ell}_k|. \end{aligned}$$

Using the property of exponential sequences, we have

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

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

$$\mathbb{P}_{|\mathscr{K}|}\left(\frac{1}{1-\rho}\sum_{k\in\mathscr{K}}|\bar{\ell}_k| > \frac{a}{\epsilon}\right) \leq 2\exp\left[-\frac{(1-\rho)^2a^2}{2\epsilon^2|\mathscr{K}|^2\sigma^2}\right].$$

Letting

$$\epsilon := \frac{(1-\rho)a}{|\mathscr{K}|\sigma\sqrt{2\log(4/\delta)}},$$

it holds that

$$\begin{aligned} &\mathbb{P}\left[f(z) - f(z') \geq -a, \quad \forall z, z' \in \mathbb{S}_2(1) \text{ s.t.} \|z-z'\|_2 \leq \epsilon\right] \\ &\geq \mathbb{P}_{|\mathscr{K}|}\left(\frac{1}{1-\rho}\sum_{k\in\mathscr{K}}|\bar{\ell}_k| \leq \frac{a}{\epsilon}\right) \geq 1 - \frac{\delta}{2}. \end{aligned}$$

Now, after we replace δ in (23) with $\delta/(2N)$, it holds with probability at least $1-\delta/2$ that

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

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

$$\mathbb{P}\left[f(z) \geq 0, \quad \forall z \in \mathbb{S}_2(1)\right] \geq 1-\delta.$$

The corresponding sample complexity is

$$T \geq \Theta\left[nR\log\left(\frac{2N}{\delta}\right)\right].$$

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

$$|\mathscr{I}_1| = \Theta[p(1-p)T], \quad |\mathscr{K}| = \Theta(pT),$$

we get the estimate

$$\begin{aligned} \log(N) &\leq n\log\left(1+\frac{2}{\epsilon}\right) \\ &= n\log\left[1+\Theta\left(\frac{n\sqrt{\log(1/\delta)}\log(1/c)}{(1-p)c^5(1-\rho)\log(1/\rho)}\right)\right] \\ &= \Theta\left[n\log(nR)\right]. \end{aligned}$$

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

$$T \geq \Theta\left[nR\left[n\log(nR) + \log\left(\frac{1}{\delta}\right)\right]\right].$$

Finally, we replace δ with δ/n and apply the union bound to all coordinates $\ell \in \{1, \ldots, n\}$.

G. SHOWING UNIQUENESS OF SOLUTION

$\bar{A}$ is the unique solution to the (CO-L2-Aut) if and only if the objective function value evaluated at $\bar{A}$ is strictly less than the objective function value evaluated at $\bar{A}+\Omega$, where Ω is any small perturbation to the matrix $\bar{A}$. Let $f_T(A)=\sum_{i=0}^{T-1}\|(\bar{A}-A)x_i+\bar{d}_t\|$ be the objective function of the (CO-L2-Aut). Then, $\bar{A}$ is the unique solution if

$$f_T(\bar{A}) < f_T(\bar{A}+\Omega) \quad \Rightarrow$$

$$\sum_{i=0}^{T-1}\|\bar{d}_i\|_2 < \sum_{i=0}^{T-1}\|\Omega x_i+\bar{d}_i\|_2 \quad \Rightarrow$$

$$\sum_{i\in\mathcal{K}}\|\bar{d}_i\|_2 < \sum_{i\notin\mathcal{K}}\|\Omega x_i\|_2+\sum_{i\in\mathcal{K}}\|\bar{d}_i\|_2+\sum_{i\in\mathcal{K}}\left\langle \Omega x_i, \frac{\bar{d}_i}{\|\bar{d}_i\|_2}\right\rangle$$

$$\Rightarrow \quad 0 < \sum_{i\notin\mathcal{K}}\|\Omega x_i\|_2+\sum_{i\in\mathcal{K}}\left\langle \Omega x_i, \frac{\bar{d}_i}{\|\bar{d}_i\|_2}\right\rangle+\mathcal{O}(\|\Omega\|_F^2)$$

In the second-to-last step, we used Taylor's expansion for $\|x\|_2$ whenever $x\neq 0$. Taking Ω sufficiently small ensures that $\|\Omega x_i+\bar{d}_i\|_2$ is not zero. The terms $\mathcal{O}(|\Omega|_F^2)$ is non-negative. As a result, $\bar{A}$ is the unique solution whenever

$$\sum_{i\notin\mathcal{K}}\|\Omega x_i\|_2+\sum_{i\in\mathcal{K}}\left\langle \Omega x_i, \frac{\bar{d}_i}{\|\bar{d}_i\|_2}\right\rangle > 0, \quad \forall\Omega : \|\Omega\|_F \leq \epsilon.$$

We can bound the norm of Ω with 1 instead of ϵ thanks to homogeneity. Let Ω^l denote the l-th row of the matrix Ω. Using $\|x\|_2 \geq \|x\|_1/\sqrt{n}$ and, we have the following sufficient condition for the exact recovery:

$$\sum_{l=1}^{n}\left(\sum_{i\notin\mathcal{K}}\frac{1}{\sqrt{n}}\|\Omega^l x_i\|_1+\Omega^l x_i\cdot\frac{\bar{d}_i^l}{\|\bar{d}_i\|_2}\right) > 0,$$

for all Ω such that $\|\Omega^l\|_2\leq 1, l=1,\ldots,n$. This condition can be simplified as

$$\sum_{l=1}^{n}\Omega^l g+\|\Omega^l\mathbf{F}\|_1 > 0, \quad \forall\Omega \text{ s.t. } \|\Omega^l\|_2\leq 1, l=1,\ldots,n$$

The matrix $\mathbf{F}\in\mathbb{R}^{n\times(T-|\mathcal{K}|)}$ and the vector $g\in\mathbb{R}^n$ are defined at the beginning of the proof. The above condition is the same as (11), except that we require strict positivity of $f(z)$ rather than non-negativity. The number of samples required to satisfy both inequalities will remain the same, due to the continuous distribution of the attack vectors.

ACKNOWLEDGMENT

We extend our sincere gratitude to Jihun Kim for his valuable contributions to the revisions of the paper and the refinement of the mathematical details.

REFERENCES

[1] H.-F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*. Berlin, Germany: Springer Science & Business Media, 2012.

[2] L. Ljung, "Convergence analysis of parametric identification methods," *IEEE Trans. Autom. Control*, vol. TAC-23, no. 5, pp. 770–783, Oct. 1978.

[3] L. Lennart et al., "Theory for the User," in *System Identification*, vol. 28. Upper Saddle River, NJ, USA: PTR Prentice Hall, 1987.

[4] E. Hazan, H. Lee, K. Singh, C. Zhang, and Y. Zhang, "Spectral filtering for general linear dynamical systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 4639–4648.

[5] H. Mania, S. Tu, and B. Recht, "Certainty equivalence is efficient for linear quadratic control," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32,pp. 10154–10164.

[6] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite time LTI system identification," *J. Mach. Learn. Res.*, vol. 22, no. 26, 2021.

[7] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite sample perspective," *IEEE Control Syst. Mag.*, vol. 43, no. 6, pp. 67–97, 2023.

[8] 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 Trans. Control Syst. Technol.*, vol. 31, no. 6, pp. 2744–2759, Nov. 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:249461776

[9] L. Wang, E. A. Theodorou, and M. Egerstedt, "Safe learning of quadrotor dynamics using barrier certificates," in *Proc. 2018 IEEE Int. Conf. Robot. Automat.*, 2017, pp. 2460–2465. [Online]. Available: https://api.semanticscholar.org/CorpusID:35948052

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

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

[12] M. Simchowitz and D. Foster, "Naive exploration is optimal for online LQR," in *Int. Conf. Mach. Learn.*, 2020, pp. 8937–8948.

[13] R. Zhang, Y. Li, and N. Li, "On the regret analysis of online LQR control with predictions," in *2021 Amer. Control Conf.*, 2021, pp. 697–703.

[14] I. Ziemann, A. Tsiamis, B. Lee, Y. Jedra, N. Matni, and G. J. Pappas, "A tutorial on the non-asymptotic theory of system identification," in *Proc. 62nd IEEE Conf. Decis. Control*, 2023, pp. 8921–8939.

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

[16] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *J. Mach. Learn. Res.*, vol. 10, no. 7, pp. 1485–1510, 2009.

[17] L. Bako, "On a class of optimization-based robust estimators," *IEEE Trans. Autom. Control*, vol. 62, no. 11, pp. 5990–5997, Nov. 2017.

[18] D. Bertsimas and M. S. Copenhaver, "Characterization of the equivalence of robustification and regularization in linear and matrix regression," *Eur. J. Oper. Res.*, vol. 270, no. 3, pp. 931–942, Nov. 2018.

[19] S. Pesme and N. Flammarion, "Online robust regression via SGD on the l-1 loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 2540–2552.

[20] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[21] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, *Introduction to Compressed Sensing*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[22] S. Foucart and H. Rauhut, *A Math. Introduction to Compressive Sens.*. Basel, Switzerland: Birkhäuser, 2013.

[23] D. L. Donoho and J. Tanner, "Neighborliness of randomly projected simplices in high dimensions," *Proc. Nat. Acad. Sci.*, vol. 102, no. 27, pp. 9452–9457, 2005.

[24] D. L. Donoho, "High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension," *Discrete Comput. Geometry*, vol. 35, pp. 617–652, 2006.

[25] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[26] E. J. Candes and P. A. Randall, "Highly robust error correction byconvex programming," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2829–2840, Jul. 2008.

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

[28] S. Mendelson, "Learning without concentration," *J. ACM*, vol. 62, no. 3, pp. 1–25, 2015.
[29] Y. Li, S. Das, J. Shamma, and N. Li, "Safe adaptive learning-based control for constrained linear quadratic regulators with regret guarantees," 2021, *arXiv:2111.00411*.
[30] S. Fattahi, N. Matni, and S. Sojoudi, "Learning sparse dynamical systems from a single sample trajectory," in *2019 IEEE 58th Conf. Decis. Control*, 2019, pp. 2682–2689.
[31] Y. Jedra and A. Proutiere, "Finite-time identification of stable linear systems optimality of the least-squares estimator," in *2020 59th IEEE Conf. Decis. Control*, 2020, pp. 996–1001.
[32] A. Wagenmaker and K. Jamieson, "Active learning for identification of linear dynamical systems," in *Conf. Learn. Theory*, 2020, pp. 3487–3582.
[33] W. Xu, E.-W. Bai, and M. Cho, "System identification in the presence of outliers and random noises: A compressed sensing approach," *Automatica*, vol. 50, no. 11, pp. 2905–2911, 2014.
[34] H. Feng, B. Yalcin, and J. Lavaei, "Learning of dynamical systems under adversarial attacks - null space property perspective," in *2023 Amer. Control Conf.*, 2023, pp. 4179–4184.
[35] M. J. Wainwright, "High-Dimensional Statistics: A Non-Asymptotic Viewpoint," in *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[36] H. Feng and J. Lavaei, "Learning of dynamical systems under adversarial attacks," in *2021 60th IEEE Conf. Decis. Control*, 2021, pp. 3010–3017.
[37] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *2012 50th Annu. Allerton Conf. Commun., Control, Comput.*, 2012, pp. 1806–1813.
[38] P. Pradhan and P. Venkitasubramaniam, "Stealthy attacks in dynamical systems: Tradeoffs between utility and detectability with application in anonymous systems," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 4, pp. 779–792, Apr. 2017.
[39] Y. Chen, J. Fan, C. Ma, and Y. Yan, "Bridging convex and nonconvex optimization in robust PCA: Noise, outliers, and missing data," *Ann. Statist.*, vol. 49, no. 5, 2021, Art. no. 2948.
[40] J. Farkas, "Theorie der einfachen ungleichungen," *J. für die reine und angewandte Mathematik*, vol. 124, pp. 1–27, 1902. [Online]. Available: http://eudml.org/doc/149129
[41] B. Yalcin, H. Zhang, J. Lavaei, and M. Arcak, "Exact recovery for system identification with more corrupt data than clean data," 2024. [Online]. Available: https://arxiv.org/abs/2305.10506
[42] Y. Zhang, R. Madani, and J. Lavaei, "Conic relaxations for power system state estimation with line measurements," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1193–1205, Sep. 2018.
[43] R. Hovorka et al., "Partitioning glucose distribution/transport, disposal, and endogenous production during IVGTT," *Amer. J. Physiol. Endocrinol. Metab.*, vol. 282, no. 5, pp. E992–E1007, 2002.
[44] I. Hajizadeh, M. Rashid, and A. Cinar, "Integrating compartment models with recursive system identification," in *2018 Annu. Amer. Control Conf.*, 2018, pp. 3583–3588.

BATURALP YALCIN received the B.S. degree in industrial engineering from Bogazici University, Istanbul, Türkiye, and the M.S. degree in industrial engineering and operations research from the University of California, Berkeley, CA, USA, where he is currently working toward the Ph.D. degree. His research interests include the landscape of optimization and adversarial learning problems.

HAIXIANG ZHANG received the B.S. degree in computer science and technology and computational mathematics from Peking University, Beijing, China. He is currently working toward the Ph.D. degree in applied mathematics with the University of California, Berkeley, CA, USA. His research interests include non-convex optimization, especially low-rank matrix optimization and optimization via simulation. He was the recipient of Two Sigma Ph.D. Fellowship.

JAVAD LAVAEI received the Ph.D. degree in control and dynamical systems from the California Institute of Technology, Pasadena, CA, USA, in 2011. He is currently an Associate Professor with the Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA. Dr. Lavaei was the recipient of multiple awards, including the NSF CAREER Award, Office of Naval Research Young Investigator Award, and Donald P. Eckman Award.

MURAT ARCAK is currently a Professor with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA. His research interests include dynamical systems and control theory with applications to synthetic biology, multiagent systems, and transportation. Prof. Arcak was the recipient of the NSF CAREER Award and Donald P. Eckman Award.