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

Baturalp Yalcin[1], **Javad Lavaei**[1], **Murat Arcak**[2]
[1]UC Berkeley, Industrial Engineering and Operations Research
[2]UC Berkeley, Electrical Engineering and Computer Sciences
{byalcin, lavaei, arcak}@berkeley.edu

Abstract

In this paper, we study the system identification problem for linear discrete-time systems under adversaries and analyze two lasso-type estimators. We study both asymptotic and non-asymptotic properties of these estimators in two separate scenarios, corresponding to deterministic and stochastic models for the attack times. Since the samples collected from the system are correlated, the existing results on lasso are not applicable. We show that when the system is stable and the attacks are injected periodically, the sample complexity for the exact recovery of the system dynamics is $\mathcal{O}(n)$, where n is the dimension of the states. When the adversarial attacks occur at each time instance with probability p, the required sample complexity for the exact recovery scales as $\mathcal{O}(\log(n)p/(1-p)^2)$. This result implies the almost sure convergence to the true system dynamics under the asymptotic regime. As a by-product, even when more than half of the data is compromised, our estimators still learn the system correctly. 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.

1 Introduction

Dynamic systems are the building block of 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 is essential in robotics, control, time-series, and reinforcement learning applications. Despite being ubiquitously studied, most results in system identification were focused on the asymptotic properties of the proposed estimators until recently. Nonetheless, the non-asymptotic analysis of the system identification problem gained popularity in recent years [Hazan et al., 2018, Mania et al., 2019, Sarkar et al., 2019, Tsiamis et al., 2022]. Although the non-asymptotic analysis is harder because of the correlation between the sample points, it is crucial to understand the required sample complexity for online control problems.

The robust learning of dynamical systems is also crucial for safety-critical applications, such as autonomous driving, unmanned aerial vehicles, and biomedical applications. Although there are several recent papers on online non-asymptotic control of linear time-invariant (LTI) systems, their methods are often designed to only handle small noise in the measurements and do not consider large values of noise corresponding to adversarial attacks or other types of data corruption. [Simchowitz et al., 2018, Simchowitz and Foster, 2020, Zhang et al., 2021]. 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 estimators inspired by the lasso problem and robust regression literature [Bako and Ohlsson, 2016]. 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.

Preprint. Under review.

The robust regression and learning problems under adversaries are ubiquitously studied in the literature [Xu et al., 2009, Bako, 2017, Bertsimas and Copenhaver, 2018, Pesme and Flammarion, 2020]. However, existing methods for analyzing the estimators cannot be directly generalized to control problems due to the correlation between the samples. Therefore, different strategies were 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 Dean et al. [2020]. However, obtaining multiple trajectories is not viable for most safety-critical applications. One of the methods with a single trajectory relies on the persistent excitation of the states so that the dynamics could be explored thoroughly. This is achieved by injecting Gaussian noise input into the system. The small ball techniques are used to analyze the properties of the estimator [Mendelson, 2015, Simchowitz et al., 2018, Li et al., 2021]. This technique uses normalized martingale bounds for the estimation error when the excitation is large enough [Simchowitz et al., 2018].

Unlike the cumbersome non-asymptotic analysis of correlated data, the least-squares estimator offers a closed-form solution when the system is subjected to small white noise [Fattahi et al., 2019, Jedra and Proutiere, 2020, Wagenmaker and Jamieson, 2020]. When systems are not injected with large and sparse noise vectors, the least-squares estimator performs relatively well. However, the least-squares estimator is not robust to adversarial attacks and the literature on robust learning of dynamical systems is limited. The work by Feng et al. [2022] defines 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 Feng et al. [2022] and study robust estimators from a non-asymptotic point of view under generic assumptions such as the system being stable and the attacks being sub-Gaussian.

Contributions: We study discrete-time linear time-invariant systems of the form $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i, i = 0, 1, \ldots, T-1$, where $\bar{A} \in \mathbb{R}^{n\times n}$ and $\bar{B} \in \mathbb{R}^{m\times n}$ are unknown matrices of the model. We aim to learn these matrices from the samples $\{x_i, u_i\}_{i=0}^{T-1}$ when the disturbance vectors $\bar{d}_i$ are adversarial. If $\bar{d}_i$ is zero, there is no attack at time i. If $\bar{d}_i$ is nonzero, then an attack has occurred at time i and we have no information on the value of $\bar{d}_i$, which is designed by an attacker to adversely affect the states of the system.

i) We study two convex estimators based on the minimization of the ℓ_2 and ℓ_1 norm of the estimated disturbance vectors, $\sum_{i=0}^{T-1} \|d_i\|_2$ and $\sum_{i=0}^{T-1} \|d_i\|_1$, with the variables A and B subject to $x_{i+1} = Ax_i + Bu_i + d_i$, given the samples $\{x_i, u_i\}_{i=0}^{T-1}$. The ℓ_2 norm estimator is most effective when the set of attack times is sparse while the vector $\bar{d}_i$ at each attack time i could have several nonzero entries (meaning that most entries of the state x_{i+1} are affected by the attack). In contrast, the ℓ_1 norm estimator is preferable when the vector $\bar{d}_i$ at each attack time is sparse.

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

iii) We then consider a probabilistic model for the attack occurrences in which there is an attack at each time t with probability p, independently of previous time periods. We show that our estimators find the true system matrices almost surely when the attack vectors are stealthy.

iv) We study the required sample complexity of our estimators for exact recovery. Let $\bar{\lambda}_l$'s be the eigenvalues of $\bar{A}$. 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

$$T \gtrsim \max\left\{ \frac{p}{(1-p)^2} \max_l \left\{ \frac{1}{|\bar{\lambda}_l|^2(1-|\bar{\lambda}_l|)^2} \right\} \log(n^2/\delta), \frac{1}{(1-p)^2} \log(mn/\delta) \right\}.$$

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

2 Notation

For a matrix Z, $\|Z\|_F$ denotes the Frobenius norm of a matrix. For a vector z, $\|z\|_1$ and $\|z\|_2$ denote its ℓ_1 and ℓ_2 norms, respectively. Moreover, for a given function f, $\partial f(z)$ shows the subdifferential for the function f at the point z. Given two functions f and g, the relations $f(x) \lesssim g(x)$ and $f(x) \gtrsim g(x)$ mean that there exist universal constants c_1 and c_2 such that $f(x) \leq c_1 g(x)$ and $f(x) \geq c_2 g(x)$, respectively. For two vectors v and w, $\langle v, w \rangle$ is the inner product between those vectors in their respective vector space. Furthermore, we use the notation $v \otimes w = vw^T$ to denote the outer product. $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ denote the probability of an event and the expectation of a random variable. A Gaussian random variable X with mean μ and covariance matrix Σ is written as $X \sim N(\mu, \Sigma)$. A random variable $X \sim sG(\sigma)$ is sub-Gaussian with parameter σ and $X \sim sE(\nu, \alpha)$ is sub-Exponential with parameters ν and α. Given a time-dependent variable $z(t)$, $\dot{z}(t)$ represents its derivative with respect to time t. $|S|$ shows the cardinality of a given set S.

3 Problem Formulation

We consider a linear time-invariant dynamical system over the time horizon $[0, T]$, $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i, i = 0, 1, \ldots, T-1$, where $\bar{A} \in \mathbb{R}^{n \times n}$ and $\bar{B} \in \mathbb{R}^{n \times m}$ are system matrices, and $\bar{d}_i \in \mathbb{R}^n$ are unknown system disturbances. Given the set of state measurements $\{x_i\}_{i=0}^{T}$ and the set of inputs $\{u_i\}_{i=0}^{T-1}$, the goal is to estimate the unknown system matrices $\bar{A}$ and $\bar{B}$. In ordinary systems without attacks, the disturbance vectors $\{\bar{d}_i\}_{i=0}^{T-1}$ represent small disturbances on the input and small modeling errors. However, the disturbance vectors $\{\bar{d}_i\}_{i=0}^{T-1}$ can be engineered to be large if there is an outside attack on the system from an agent or there is a sensor/actuation fault. Define $D := [d_0, \ldots, d_{T-1}]$, as well as $\|D\|_{1,col} := \sum_i \|d_i\|_1$, and $\|D\|_{2,col} := \sum_i \|d_i\|_2$. To recover the system matrices $\bar{A}$ and $\bar{B}$, we analyze the following convex optimization problems:

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

and

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

where the states $x_i, i \in \{0, \ldots, T\}$, 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. Note that these two problems are equivalent if the system has a single state. In problem (CO-L2), the ℓ_1 norm is applied at the group level to $\{d_i\}_{i=0}^{T-1}$. 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). If the disturbance vectors are assumed to be sparse, (CO-L1) is more suitable than (CO-L2). Furthermore, the states x_t are correlated to each other due to the system's dynamics, which makes the non-asymptotic analysis of the problem more challenging than robust regression literature where the samples are assumed to be independently generated.

Since (CO-L2) and (CO-L1) are convex optimization problems with linear equalities, the Karush-Kuhn-Tucker (KKT) conditions are sufficient to guarantee optimality, as stated below.

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

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

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

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

The proof for the KKT conditions when $\circ = 2$ is given in Feng and Lavaei [2021] and the proof for the case $\circ = 1$ can be done similarly. We will utilize the conditions above to show when the exact recovery is possible. As a simple corollary to Theorem 1, we can state that $(\bar{A}, \bar{B})$ is a solution to our estimator if the following conditions holds:

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

4 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 assume that the system is stable and it is initialized at the origin 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 eigenvalues of $\bar{A}$ are inside the unit circle.*

The stability assumption is generic in system identification problems to avoid unbounded growth of the states. Without loss of generality, we initialize the trajectories at the origin, and initialization at other points affects the results only with a constant factor. We consider noiseless systems to obtain exact recovery results, meaning that if there is no attack at time period i, then $\bar{d}_i = 0$. The noisy case when $\bar{d}_i$ is small can be addressed using our framework via perturbation analysis which allows us to bound how far the recovered solution is away from the true solution in terms of the values of small noise vectors. We denote the set of time instances at which the attack occurs with $\mathscr{K}$. Mathematically, it is represented as $\mathscr{K} = \{i | \bar{d}_i \neq 0, i \in \{0, 1, \ldots, T-1\}\}$. As a result, we are interested in recovering the system matrices $\bar{A}$ and $\bar{B}$ using the following convex optimization problems for autonomous systems:

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

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

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:

$$0 \in \sum_{i \notin \mathscr{K}} x_i \otimes \partial \|(\bar{A} - A)\|_\circ + \sum_{i \in \mathscr{K}} x_i \otimes \partial \|((\bar{A} - A)x_i + \bar{d}_i)\|_\circ \tag{3}$$

4.1 Single State Case

We first assume that the problem is one-dimensional. When $n = 1$, the problems (CO-L1-Aut) and (CO-L2-Aut) are equivalent and, therefore, we only focus on (CO-L2-Aut). After establishing the optimality conditions for these problems, we will examine two types of attack structures. The first one is a deterministic attack model for which attacks occur at every Δ time period. Later, we investigate a probabilistic attack structure where each attack occurs with the same probability at each time period. We first define the deterministic attack model, which is borrowed from Feng and Lavaei [2021].

Definition 1. *Given a non-negative integer Δ, 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 $j \in \{i+1, \ldots, i+\Delta-1\}$.*

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 one-dimensional autonomous system with Δ-spaced disturbance sequence with $\Delta \geq 2$. Then, the convex formulation* (CO-L2-Aut) *(or equivalently* (CO-L1-Aut)*) has the unique solution $\bar{A}$ as long as $T \geq \Delta + 1$.*

Note that Proposition 1 does not make any assumption on the vector set $\{\bar{d}_i : i \in \mathscr{K}\}$ and it could be arbitrarily large or correlated. As a result, regardless of the severity of the attack, an exact recovery is guaranteed for (CO-L1-Aut) and (CO-L2-Aut). One important implication of Proposition 1 is that whenever we have Δ-spaced disturbance sequence with $\Delta = 2$, it implies that half of the observations are corrupted. In the robust regression estimation literature, the 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 the exact recovery for the convex formulation even if half of the data is wrong.

Next, it is natural to ask whether it is possible to learn the system when there is more corrupt data than clean data. We cannot use a Δ-spaced disturbance sequence model because the minimum value of Δ is 2, which does not allow the size of corrupt data to exceed the size of clean data. To address this, we consider a probabilistic attack model for which there is a parameter p specifying the probability of having an attack at each time. Specifically, given a time instance i, $\bar{d}_i$ is nonzero with probability p and this is independent of all previous and future time instances. As a result, the event of having an attack at each time is identically and independently distributed with Bernoulli distribution with parameter p. Our goal is to discover the properties of (CO-L1-Aut) and (CO-L2-Aut) when $p > 0.5$. The next theorem states that as long as the attacks have the same probability of being negative or positive, the estimators recover the true system matrix $\bar{A}$.

Theorem 2. *Consider a one-dimensional autonomous system, and suppose that there is an attack at each time i, i.e., $i \in \mathscr{K}$, with probability p and this is independent of the other time periods. Assume that the attack vector $\bar{d}_i$ comes from an arbitrary probability distribution such that $\mathbb{P}(\bar{d}_i < 0) = \mathbb{P}(\bar{d}_i > 0)$ for all $i \in \mathscr{K}$. Then, the solution A^* is almost surely a solution to the convex optimization* (CO-L2-Aut)*, or equivalently* (CO-L1-Aut)*.*

Even in the case when the probability p is close to 1, leading to much more corrupt data than clean data, Theorem 2 guarantees the asymptotic exact recovery when there are a sufficient number of samples. Note that the symmetricity assumption on the disturbance vectors is non-restrictive and corresponds to stealth attacks. For an attack to be stealthy, its value should be zero on expectation and our assumption has a similar flavor.

Although almost sure convergence is a desirable property, the number of required samples for exact recovery may be large. Therefore, it is desirable to quantify the sample complexity for the exact recovery. We require a non-asymptotic analysis of the system with a slightly stronger assumption on the attack values. It is assumed that the attack values have a sub-Gaussian distribution. [1]

Theorem 3. *Consider a one-dimensional autonomous system, and suppose that there is an attack at each time i, i.e., $i \in \mathscr{K}$, with probability p and this is independent of the other time periods. Assume that the attack vectors $\bar{d}_i$ for all $i \in \mathscr{K}$ are zero mean sub-Gaussian with parameter σ. Given a positive number δ, if the time horizon T is*

$$T \gtrsim \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{A}|)^2|\bar{A}|^2} \log(1/\delta),$$

the $\bar{A}$ is a solution of the convex formulation (CO-L2-Aut)*, or equivalently* (CO-L1-Aut)*, with probability at least $1-\delta$.*

If we have sub-Gaussian attack values, the sample complexity can be obtained if $|\bar{A}| < 1$ meaning that the system is stable. Due to the logarithmic probability bound, Theorem 3 implies asymptotic convergence as well. The required amount of data increases with the value $p/(1-p)^2$. Hence, as p increases, the number of samples for an exact recovery with high probability blows up. Moreover, sample complexity scales with the inverse of $|\bar{A}|^2(1-|\bar{A}|)^2$, which is maximized at $|\bar{A}| = 0.5$.

4.2 General Case with State Size n

We first generalize the result obtained in Proposition 1 to generic autonomous dynamical systems with an arbitrary n under a Δ-spaced disturbance sequence with $\Delta \geq n+1$.

Proposition 2. *Consider an autonomous system with dimension n under a Δ-spaced disturbance sequence with $\Delta \geq n+1$. Suppose that $\bar{A}$ has n eigenvalues, $\bar{\lambda}_j, j = 1,2,\ldots,n$, with linearly independent eigenvectors such that*

$$\bar{A}\bar{d}_i \neq \bar{\lambda}_j \bar{d}_i, \forall i \in \mathscr{K}, j = 1,2,\ldots,n. \tag{4}$$

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

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

[1] A random variable X with mean μ is sub-Gaussian with parameter σ if $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2/2}, \forall \lambda \in \mathbb{R}$

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

This result is the generalization of Proposition 1 in the case of $n=1$. The equation (4) implies that the disturbance vectors are not aligned with the eigenvectors of the matrix $\bar{A}$. To gain insight into the condition (5) that involves the product of eigenvalues, consider a special case where $\bar{A}$ has the eigenvalue λ with multiplicity n with n distinct eigenvectors. In this case, we can simplify (5) as follows. Define $k := \Delta - n$. Then, (5) 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. Table 1 in the appendix 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, $|\lambda| \leq C_{n,k} \leq C_{1,k} \to 2$. In addition, whenever $k = n$ or $\Delta = 2n$, $|\lambda| < 1$ is sufficient for this condition to hold, which in turn is sufficient for exact recovery given that the other conditions in Proposition 2 are satisfied. This conclusion is analogous to the stability of the system.

In higher-dimensional cases, Propositions 1 and 2 still apply for problem (CO-L1-Aut). However, the KKT conditions will differ by the subdifferential of the ℓ_2 and ℓ_1 norms. In fact, they both have a similar shape. Therefore, one can show that these propositions hold with the same conditions when we use the convex formulation (CO-L1-Aut) using the ℓ_1 norms of the disturbance vectors.

Theorem 4. *Consider an autonomous system with dimension n. Suppose that there is an attack at each time i, i.e.,* $i \in \mathscr{K}$*, with probability p and this is independent of the other time periods. Let* $\bar{d}_i^j$ *denote the j-th entry of the vector* $\bar{d}_i$*. Assume that* $\mathbb{P}(\bar{d}_i^j < 0) = \mathbb{P}(\bar{d}_i^j > 0)$ *for all* $i \in \mathscr{K}$ *and* $j = 1,\ldots n$*. Then,* $\bar{A}$ *is almost surely a solution of convex formulation* (CO-L2-Aut)*.*

We obtained almost sure convergence to the true solution when each entry of the disturbance vector is symmetric around the origin. One could argue that using the objective $\|D\|_{1,col} = \sum_{i=0}^{T-1}\|d_i\|_1$ could be a better alternative to $\|D\|_{2,col} = \sum_{i=0}^{T-1}\|d_i\|_2$. The results above hold in this case as well since the expectation arguments are still the same despite the change in subdifferentials. Therefore, asymptotic results continue to hold despite this change in the objective function.

The next theorem shows that the sample complexity for the exact recovery grows with $\log(n)$ and $p/(1-p)^2$ similar to the one-dimensional case. We again assume that the attack vectors $\bar{d}_i$ are sub-Gaussian random vectors with parameter σ. As a result, $\partial\|\bar{d}_i\|_2 = \frac{\bar{d}_i}{\|d_i\|_2}$ will have bounded entries between $[-1,1]$ and the bounded random variables are known to be sub-Gaussian as well. In addition, we will utilize some concentration inequalities for sub-Exponential random variables.[2]

Theorem 5. *Consider an autonomous system with dimension n and suppose that* $\bar{A}$ *has linearly independent eigenvectors with eigenvalues* $|\bar{\lambda}_l| < 1$ *for* $l = 1,\ldots,n$*. Assume also that there is an attack at time i, i.e.,* $i \in \mathscr{K}$*, with probability p and this is independent of the other time periods. Consider the attack vectors* $\bar{d}_i$ *to be zero mean sub-Gaussian vectors with parameter* σ*. Given a positive* δ*, if the time horizon T satisfies*

$$T \gtrsim \max\left\{\frac{p}{(1-p)^2}\max_l\left\{\frac{1}{|\bar{\lambda}_l|^2(1-|\bar{\lambda}_l|)^2}\right\}\log(n^2/\delta), \frac{1}{(1-p)}\max_l\left\{\frac{1}{|\bar{\lambda}_l|}\right\}\log(n^2/\delta)\right\},$$

then $\bar{A}$ *is a solution to the convex optimization* (CO-L2-Aut) *with probability at least* $1-\delta$*.*

The two different terms for the sample complexity stem from the concentration inequality for the sub-Exponential random variables. We require sub-Exponential results because the KKT condition involves the multiplication of two sub-Gaussian random variables, which are known to be sub-Exponential. We can obtain a similar result if one prefers to use the problem (CO-L1-Aut) to recover the system matrix $\bar{A}$.

[2] A random variable X with mean μ is sub-Exponential with parameters (ν,α) if $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\nu^2\lambda^2/2}, \forall|\lambda| \in 1/\alpha$

Theorem 6. *Under the assumption of Theorem 5, if the time horizon T satisfies*

$$T \gtrsim \frac{p}{(1-p)^2} \max_l \left\{ \frac{1}{|\bar{\lambda}_l|^2(1-|\bar{\lambda}_l|)^2} \right\} \log(n^2/\delta),$$

then $\bar{A}$ is a solution to the convex optimization (CO-L1-Aut) *with probability at least $1-\delta$.*

The results on sample complexity are intuitive. As the probability of having an attack increases, we require a larger time horizon for exact recovery. In addition, if the system is barely stable with eigenvalues close to the unit circle, the sample complexity blows up. Furthermore, since the sample complexity scales with the logarithm of the dimension, the proposed estimators are scalable with respect to the dimension.

5 Systems with Input Sequence

It is desirable to understand the role of an input sequence in exact recovery because the majority of dynamic systems are controlled by an external input. Since the input sequence is generated by a controller, one can design it in such a way that it accelerates the exact recovery.

In the non-autonomous case, the system dynamics is given as $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i, i = 0, \ldots, T-1$, where $\bar{A} \in \mathbb{R}^{n\times n}$ and $\bar{B} \in \mathbb{R}^{n\times m}$, similar to the autonomous case, the true system matrices $\bar{A}$ and $\bar{B}$ are not known and the goal is to obtain these matrices using the state trajectory. We will investigate estimators (CO-L2) and (CO-L1) defined earlier. The inputs $u_i, i \in \{0, \ldots, T-1\}$ are known unlike the disturbances.

5.1 Single State with Single Input Case

We study the non-asymptotic properties of the problems (CO-L2) and (CO-L1) when there is a single state and a single input. These problems are equivalent when the state space is one-dimensional. Thus, we only consider (CO-L2). We assume that an independent and identically distributed sub-Gaussian sequence of inputs is injected into the system at each period. This allows us to obtain a high probability bound for the exact recovery of the matrices $\bar{A}$ and $\bar{B}$. A random input sequence is commonly used in system identification and online learning because it enables the exploration of the system to learn the system dynamics faster. Moreover, the sub-Gaussian input assumption is satisfied when u_i is designed in the feedback form as $u_i = Kx_i + \omega$ if the states x_i are sub-Gaussian and the input is excited with sub-Gaussian noise ω. The following theorem follows from these observations.

Theorem 7. *Consider a stable single-input single-state system with the dynamics $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i$ for $i = 0, \ldots, T-1$. Assume also that there is an attack at time i, i.e., $i \in \mathscr{K}$, with probability p and this is independent of the other time periods. Assume that the attack vectors $\bar{d}_i, \forall i \in \mathscr{K}$ are zero-mean sub-Gaussian with parameter σ and the inputs are sub-Gaussian with parameter ε. Given a positive number δ, if the time horizon T satisfies*

$$T \gtrsim \max \left\{ \frac{1}{(1-p)^2} \log(1/\delta), \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{A}|)^2|\bar{A}|^2}] \log(1/\delta) \right\},$$

then then $(\bar{A}, \bar{B})$ is a solution to (CO-L2) *with probability at least $1-\delta$.*

We obtained a high probability bound for the exact recovery of the system matrices $\bar{A}$ and $\bar{B}$. The first term in the sample complexity corresponds to the satisfaction of the KKT condition for the input sequence $\{u_i\}_{i=0}^{T-1}$, whereas the second term corresponds to the satisfaction of the KKT condition for the state measurements $\{x_i\}_{i=0}^{T}$. Similar to autonomous systems, the sample complexity increases as the probability of a large disturbance increases. Because we have a logarithmic dependence on the satisfaction of the probability bound, Theorem 7 implies almost sure asymptotic convergence to the correct matrices $\bar{A}$ and $\bar{B}$.

5.2 General Case with State Size n and Input Size m

In this section, we present our most general results when the state size is n and input size m. Our assumptions for the exact recovery are mild: system stability and sub-Gaussian inputs.

Theorem 8. *Consider a stable system with n states and m inputs with system dynamics* $x_{i+1} = \bar{A}x_i + \bar{B}u_i + \bar{d}_i$ *for* $i = 0,\dots,T-1$ *and suppose that* $\bar{A}$ *has linearly independent eigenvectors with eigenvalues* $|\bar{\lambda}_l| < 1$ *for* $l = 1,\dots,n$*. Assume also that there is an attack at time i, i.e.,* $i \in \mathscr{K}$*, with probability p and this is independent of the other time periods. Assume that the attack vectors* $\bar{d}_i, \forall i \in \mathscr{K}$ *are zero-mean sub-Gaussian with parameter* σ *and the inputs are sub-Gaussian with parameter* ε*. Given a positive number* δ*, if the time horizon T is chosen as* $\max\{T_1, T_2\}$ *where*

$$T_1 \gtrsim \max\left\{ \frac{p}{(1-p)^2} \max_l \left\{ \frac{1}{|\bar{\lambda}_l|^2(1-|\bar{\lambda}_l|)^2} \right\} \log(n^2/\delta), \frac{1}{(1-p)\min_l |\bar{\lambda}_l|} \log(n^2/\delta) \right\},$$

and

$$T_2 \gtrsim \frac{1}{(1-p)^2} \log(mn/\delta),$$

then $(\bar{A}, \bar{B})$ *is a solution to* (CO-L2) *with probability at least* $1-\delta$.

Similar to previous theorems for the autonomous case, we require a sample complexity that scales with $p/(1-p)^2$ and terms depending on the eigenvalues of $\bar{A}$. The sample complexity T_2 is needed to satisfy the KKT condition that depends on the input sequence. The number of required samples increases with the logarithm of the dimension of the unknown matrices n^2 and mn.

Theorem 9. *Under assumptions of Theorem 8, if the time horizon T satisfies*

$$T \gtrsim \max\left\{ \frac{1}{(1-p)^2} \log(mn/\delta), \frac{p}{(1-p)^2} \max_l \left\{ \frac{1}{|\bar{\lambda}_l|^2(1-|\bar{\lambda}_l|)^2} \right\} \log(n^2/\delta) \right\},$$

then $(\bar{A}, \bar{B})$ *is a solution to* (CO-L1) *with probability at least* $1-\delta$.

We note that when the input sequence $u_i = Kx_i$ is used to control the system, this input sequence satisfies the assumptions in the above theorems if x_i are sub-Gaussian. The closed-loop system with the matrix $(\bar{A} + \bar{B}K)$ results in the second solution $A^* = \bar{A} + \bar{B}K$ and $B^* = 0$. Nevertheless, the ground-truth system matrix pair $(\bar{A}, \bar{B})$ is also a solution to our estimators. This phenomenon occurs due to the existence of multiple optimal solutions and it could be avoided if the input is excited with a small sub-Gaussian noise in the form of $u_i = Kx_i + \omega$.

6 Numerical Experiment

We provide a numerical experiment inspired by biomedical applications to demonstrate the results of this paper. We consider a compartmental model of blood sugar and insulin dynamics in the human body; see Hovorka et al. [2002]. It is crucial to accurately estimate the parameters of the dynamics when the blood sugar level is regulated through the injection of a bolus of insulin into the system. Due to the complex structure of the human body, the dynamics are not the same for different individuals. We consider a linear system based on Hovarka's model given below [Hajizadeh et al., 2018]:

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

where the states x_1, x_2, x_3 represent the influence of insulin on the system of the body. S_1 and S_2 represent the absorption rate of insulin in the, directly and indirectly, accessible compartment models, respectively. Lastly, the state I stands for the blood-sugar level in the body. The disturbance d_4 shows the bolus injection to the body and the remaining disturbance vectors model sudden changes in the body due to diseases such as diabetes. Although the injected insulin amount could be known, the amount of insulin and when it reaches the effective body parts are not known exactly. Hence, d_i values are treated as unknown. Even though the disturbance in this application is not a malicious attack, it has similar features for identification purposes: the arrival time of the bolus is unknown and, once it arrives, it has a large magnitude.

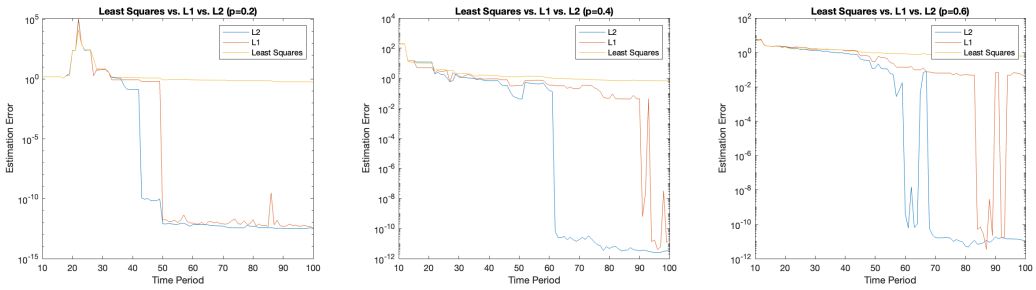

Figure 1: Estimation errors for Least-Squares, CO-L2, and CO-L1 with attack probability of $p = 0.2, 0.4, 0.6$ (left-to-right)

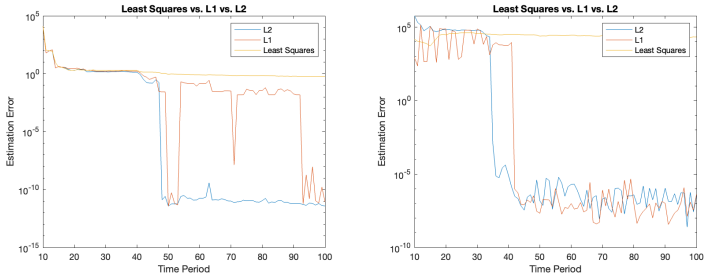

Figure 2: Estimation errors for Least-Squares, CO-L2, and CO-L1 with attack probability $p = 0.6$ (left) not Sparse d (right) Sparse d

We discretize the continuous time system to obtain an LTI system using $\Delta t = 0.5$. The obtained matrix $\bar{A}$ is stable and our goal is to estimate the parameters $(k_{ai}, k_{bi}, t_{max,I}, V_I, k_e)$ where the true values are obtained from Table 1 in Hovorka et al. [2004]. We model the attacks as zero-mean Gaussian random vectors with identity covariance matrix with variance 10 and we run our model with the probability of attack being $p = 0.2$, $p = 0.4$, and $p = 0.6$. We report the estimation error $\|A^* - \bar{A}\|_F$ for the least-squares estimator, problem CO-L2 and problem CO-L1. Figure 1 suggests that our proposed estimators attain the exact recovery while the least-squares estimator fails to do such. As the probability of having an attack p increases, the number of required time periods for exact recovery grows proportional to $p/(1-p)^2$. Note that there are more corrupted data than clean data in the case of $p = 0.6$. In addition, 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 while d_4 and d_6 have the same Gaussian distribution as before. Figure 2 shows that the two methods perform similarly when the attack vectors are also sparse.

7 Conclusion and Future Work

We studied the problem of learning LTI systems under adversarial attacks by studying two lasso-type estimators. We considered both deterministic and probabilistic attack models in terms of the time occurrence of the attack and developed strong conditions for the exact recovery of the system dynamics. When the attack occurs deterministically at every Δ period, the exact recovery is possible after $n + \Delta$ time steps. Moreover, if the system is attacked at each time instance with probability p, then the system matrices are recovered with high probability when T is on the order of $\mathcal{O}(p/(1-p)^2)$ and the logarithm of the dimension of the problem. We obtained similar results when the system is controlled by an input sequence. The results are corroborated by a numerical experiment in biology that supports the nonasymptotic analytic results. This work provides the first set of mathematical guarantees for the robust non-asymptotic analysis of dynamic systems. We leave the study of noisy systems and online control of dynamic systems under adversaries as future work.

References

Laurent Bako. On a Class of Optimization-Based Robust Estimators. *IEEE Transactions on Automatic Control*, 62(11):5990–5997, November 2017. ISSN 0018-9286, 1558-2523. doi: 10.1109/TAC.2017.2703308.

Laurent Bako and Henrik Ohlsson. Analysis of a nonsmooth optimization approach to robust estimation. *Automatica*, 66:132–145, April 2016. ISSN 00051098. doi: 10.1016/j.automatica.2015.12.024.

Dimitris Bertsimas and Martin S. Copenhaver. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270 (3):931–942, November 2018. ISSN 0377-2217. doi: 10.1016/j.ejor.2017.03.051.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020. doi: 10.1007/s10208-019-09426-y.

Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. Learning sparse dynamical systems from a single sample trajectory. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2682–2689. IEEE, 2019.

Han Feng and Javad Lavaei. Learning of dynamical systems under adversarial attacks. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3010–3017, 2021. doi: 10.1109/CDC45484.2021.9683149.

Han Feng, Baturalp Yalcin, and Javad Lavaei. Learning of dynamical systems under adversarial attacks–null space property perspective. *arXiv preprint arXiv:2210.01421*, 2022.

Iman Hajizadeh, Mudassir Rashid, and Ali Cinar. Integrating compartment models with recursive system identification. In *2018 Annual American Control Conference (ACC)*, pages 3583–3588, 2018. doi: 10.23919/ACC.2018.8431822.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. *Advances in Neural Information Processing Systems*, 31, 2018.

Roman Hovorka, Fariba Shojaee-Moradie, Paul V Carroll, Ludovic J Chassin, Ian J Gowrie, Nicola C Jackson, Romulus S Tudor, A Margot Umpleby, and Richard H Jones. Partitioning glucose distribution/transport, disposal, and endogenous production during ivgtt. *American Journal of Physiology-Endocrinology and Metabolism*, 282(5):E992–E1007, 2002.

Roman Hovorka, Valentina Canonico, Ludovic J Chassin, Ulrich Haueter, Massimo Massi-Benedetti, Marco Orsini Federici, Thomas R Pieber, Helga C Schaller, Lukas Schaupp, Thomas Vering, et al. Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological measurement*, 25(4):905, 2004.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems optimality of the least-squares estimator. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 996–1001. IEEE, 2020.

Yingying Li, Subhro Das, Jeff Shamma, and Na Li. Safe adaptive learning-based control for constrained linear quadratic regulators with regret guarantees. *arXiv preprint arXiv:2111.00411*, 2021.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

Shahar Mendelson. Learning without concentration. *Journal of the ACM (JACM)*, 62(3):1–25, 2015.

Scott Pesme and Nicolas Flammarion. Online robust regression via SGD on the l-1 loss. *Advances in Neural Information Processing Systems*, 33:2540–2552, 2020.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Nonparametric finite time LTI system identification. *arXiv preprint arXiv:1902.01848*, 2019.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online LQR. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite sample perspective. *arXiv preprint arXiv:2209.05423*, 2022.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and Regularization of Support Vector Machines. *Journal of machine learning research*, 10(7), 2009.

Runyu Zhang, Yingying Li, and Na Li. On the regret analysis of online LQR control with predictions. In *2021 American Control Conference (ACC)*, pages 697–703. IEEE, 2021.

A Table in Section 4

$C_{n,k}$	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

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

B Proofs of the Results

B.1 Proof of Proposition 1

Proof. The proof of Proposition 1 is established based on Lemma 1 below.

Lemma 1. *(Theorem 1 in Feng and Lavaei [2021]) Consider the convex optimization problem* (CO-L2-Aut) *and assume that $n = 1$. 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.*

Let $i_1, i_2, \dots$ be the set of attack times over time horizon T. Therefore, $\mathscr{K} = \{i_1, i_2, \dots\}$. Due to Δ-spaced attack model, the first attack time must be smaller than Δ, i.e., $i_1 \leq \Delta$. Since $x_0 = 0$, we have $x_t = 0$ for $t = 0, 1, \dots, i_1$. Define $\mathbb{N}$ as the set of natural numbers. We can utilize Lemma 1 to show that $\bar{A}$ is the unique solution. Using these facts, we can decompose the sum of the magnitudes of the states at non-attack times as

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

where $\mathscr{K}^c = (\mathbb{N}\backslash\mathscr{K})\backslash\{0, 1, \dots, i_1 - 1\}$, $\mathscr{K}' = \mathscr{K}^c\backslash\mathscr{K}$", and $\mathscr{K}$" $= \{i_2 - 1, i_3 - 1, \dots\}$. The second term is the sum of magnitudes at the time step just before the attack while the first term covers the rest of the magnitudes of the states. In addition, the magnitudes of the states at attack times can be written as

$$\sum_{i\in\mathscr{K}} |x_i| = \sum_{i\in\mathscr{K},\, i\geq i_2} |x_i| = \sum_{i\in\mathscr{K}"} |\bar{A}x_i| = \sum_{i\in\mathscr{K}"} |\bar{A}||x_i|.$$

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

$$\sum_{i\notin\mathscr{K}} |x_i| - \sum_{i\in\mathscr{K}} |x_i| = \sum_{i\in\mathscr{K}'} |x_i| + \sum_{i\in\mathscr{K}''} |x_i| - \sum_{i\in\mathscr{K}"} |\bar{A}||x_i| = \sum_{i\in\mathscr{K}'} |x_i| + (1 - |\bar{A}|) \sum_{i\in\mathscr{K}"} |x_i| > 0. \tag{6}$$

Note that the term $\sum_{i\notin\mathscr{K}} |x_i|$ becomes positive at time period $i_1 + 1$ while $\sum_{i\in\mathscr{K}} |x_i|$ is positive first time at time step i_2. Consequently, the strict inequality for (6) holds for every time step after i_1 because $(1 - |\bar{A}|) > 0$ by assumption. As a result, we have a unique and exact recovery for every time period $T \geq \Delta + 1$. □

B.2 Proof of Theorem 2

Proof. One can write the KKT condition for the solution $\hat{A}$ as follows:

$$0 \in \sum_{i\notin\mathscr{K}} x_i \partial \|(\bar{A} - \hat{A})\|_2 + \sum_{i\in\mathscr{K}} x_i \partial \|((\bar{A} - \hat{A})x_i + \bar{d}_i)\|_2.$$

Based on the KKT condition and the convexity of the problem, $\bar{A}$ is the solution to the problem if and only if

$$0 \in \sum_{i\notin\mathscr{K}} x_i \partial \|0\|_2 + \sum_{i\in\mathscr{K}} x_i \partial \|\bar{d}_i\|_2.$$

Our goal is to show that 0 is included in the expectation of the terms on the right-hand side. The randomness of the system stems from the set of attack vectors $\bar{d}_i, i \in \mathscr{K}$. Define $S := |\mathscr{K}|$ and $\mathscr{K} := \{i_1, i_2, \ldots, i_S\}$. Let the set of attack vectors be $\bar{d}_{\mathscr{K}} = \{\bar{d}_{i_1}, \bar{d}_{i_2}, \ldots, \bar{d}_{i_S}\}$. Taking expectations with respect to the random vectors $\bar{d}_i, i \in \mathscr{K}$, gives the following:

$$\begin{aligned}\mathbb{E}_{\bar{d}_{\mathscr{K}}}\left[\sum_{i\notin\mathscr{K}}|x_i|\partial\|0\|_2+\sum_{i\in\mathscr{K}}x_i\partial\|\bar{d}_i\|_2\right] &= \sum_{i\notin\mathscr{K}}\mathbb{E}_{\bar{d}_{\mathscr{K}}}\left[|x_i|\right]\partial\|0\|_2+\mathbb{E}_{\bar{d}_{\mathscr{K}\setminus\{i_S\}}}\left[\mathbb{E}_{\bar{d}_{i_S}}\left[\sum_{i\in\mathscr{K}}x_i\partial\|\bar{d}_i\|_2\middle|\bar{d}_{\mathscr{K}\setminus\{i_S\}}\right]\right]\\ &= \sum_{i\notin\mathscr{K}}\mathbb{E}_{\bar{d}_{\mathscr{K}}}\left[|x_i|\right]\partial\|0\|_2+\\ &\quad \mathbb{E}_{\bar{d}_{\mathscr{K}\setminus\{i_S\}}}\left[\sum_{i\in\mathscr{K}\setminus\{i_S\}}x_i\partial\|\bar{d}_i\|_2+\mathbb{P}(\bar{d}_{i_S}=0)x_{i_S}\partial\|0\|_2\right].\end{aligned}$$

The conditional expectation is taken in the first line. Given all the attack vectors until the last attack vector, every state x_i is deterministic until x_{i_S}. Since $\mathbb{P}(\bar{d}_{i_S}<0)=\mathbb{P}(\bar{d}_{i_S}>0)$, we have $\mathbb{E}[x_{i_S}\partial\|\bar{d}_{i_S}\|_2]=\mathbb{P}(\bar{d}_{i_S}=0)x_{i_S}\partial\|0\|_2$. If the conditional expectation is taken iteratively, we obtain that

$$\mathbb{E}_{\bar{d}_K}\left[\sum_{i\notin\mathscr{K}}|x_i|\partial\|0\|_2+\sum_{i\in\mathscr{K}}x_i\partial\|\bar{d}_i\|_2\right]=\sum_{i\notin\mathscr{K}}\mathbb{E}_{\bar{d}_K}\left[|x_i|\right]\partial\|0\|_2+\sum_{i\in\mathscr{K}}\mathbb{P}(\bar{d}_i=0)x_i\partial\|0\|_2.$$

For the subdifferential of the $\|\cdot\|_2$, it is known that $0\in\partial\|0\|_2$. Consequently, we have

$$0\in\mathbb{E}\left[\sum_{i\notin\mathscr{K}}|x_i|\partial\|0\|_2+\sum_{i\in\mathscr{K}}x_i\partial\|\bar{d}_i\|_2\right].$$

Therefore, $\bar{A}$ is a solution almost surely due to the KKT condition and the strong law of large numbers. □

B.3 Proof of Theorem 3

Proof. $\bar{A}$ is a solution to the problem if the KKT condition holds:

$$0\in\sum_{i\notin\mathscr{K}}x_i\partial\|0\|_2+\sum_{i\in\mathscr{K}}x_i\partial\|\bar{d}_i\|_2.$$

Due to the system dynamics and given $x_0=0$, x_i can be expressed as

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

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

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

where I is the identity matrix. Therefore, substituting this into the KKT condition yields that

$$0\in\sum_{i\notin\mathscr{K}}\sum_{k\in\mathscr{K}}\bar{A}^{(i-1-k)_+}\partial\|0\|_2\bar{d}_k+\sum_{i\in\mathscr{K}}\sum_{k\in\mathscr{K}}\bar{A}^{(i-1-k)_+}\bar{d}_k\partial\|\bar{d}_i\|_2. \tag{7}$$

Due to the definition of $A^{(i)_+}$, this can be written as

$$0\in\sum_{i\notin\mathscr{K}}\sum_{\substack{k\in\mathscr{K}\\k<i}}\bar{A}^{(i-1-k)_+}\partial\|0\|_2\bar{d}_k+\sum_{i\in\mathscr{K}}\sum_{\substack{k\in\mathscr{K}\\k<i}}\bar{A}^{(i-1-k)_+}\bar{d}_k\partial\|\bar{d}_i\|_2.$$

By changing the order of summation, the following is obtained:

$$0\in\sum_{k\in\mathscr{K}}\left(\sum_{\substack{i\notin\mathscr{K}\\i>k}}\bar{A}^{(i-1-k)_+}\partial\|0\|_2+\sum_{\substack{i\in\mathscr{K}\\i>k}}\bar{A}^{(i-1-k)_+}\partial\|\bar{d}_i\|_2\right)\bar{d}_k.$$

The right-hand side term has the minimum value

$$L := \sum_{k\in\mathscr{K}} \left(-\sum_{\substack{i\notin\mathscr{K}\\ i>k}} |\bar{A}^{(i-1-k)_+}| \partial \|\bar{d}_k\|_2 + \sum_{\substack{i\in\mathscr{K}\\ i>k}} \bar{A}^{(i-1-k)_+} \partial \|\bar{d}_i\|_2 \right) \bar{d}_k,$$

and the maximum value

$$U := \sum_{k\in\mathscr{K}} \left(\sum_{\substack{i\notin\mathscr{K}\\ i>k}} |\bar{A}^{(i-1-k)_+}| \partial \|\bar{d}_k\|_2 + \sum_{\substack{i\in\mathscr{K}\\ i>k}} \bar{A}^{(i-1-k)_+} \partial \|\bar{d}_i\|_2 \right) \bar{d}_k.$$

The goal is to show that $L < 0 < U$ with high probability. Because $\bar{d}_i$ is a zero-mean, symmetric sub-Gaussian random variable, $Y_i = \partial\|\bar{d}_i\|_2$ is a Rademacher random variable. In other words, $\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = -1) = 0.5$. It can be seen that $\bar{d}_k \partial \|\bar{d}_i\|_2$ is sub-Gaussian with parameter σ:

$$\mathbb{E}\left[e^{s\bar{d}_k \partial \|\bar{d}_i\|_2}\right] = \mathbb{E}\left[\frac{1}{2}e^{s\bar{d}_k} + \frac{1}{2}e^{-s\bar{d}_k}\right] \leq \frac{1}{2}e^{s^2\sigma^2/2} + \frac{1}{2}e^{s^2\sigma^2/2} = e^{s^2\sigma^2/2}$$

In addition, the random variables $\bar{d}_k \partial \|\bar{d}_i\|_2$ and $\bar{d}_k \partial \|\bar{d}_j\|_2$ with different i and j are independent of each other. Similarly, $\bar{d}_i \partial \|\bar{d}_i\|_2 = |\bar{d}_i|$ is a sub-Gaussian random variable with parameter σ. (See Proposition 2.5.2 in Vershynin [2018].) Consequently, L and U are sub-Gaussian random variables because each of them is a sum of independent sub-Gaussian random variables. We have

$$L \sim sG\left(\sqrt{ \left(\sum_{k\in\mathscr{K}} \left(-\sum_{\substack{i\notin\mathscr{K}\\ i>k}} |\bar{A}^{(i-1-k)_+}| \right)^2 + \left(\sum_{\substack{i\in\mathscr{K}\\ i>k}} \bar{A}^{(i-1-k)_+} \right)^2 \right) \sigma^2 } \right) = sG(\sigma_1(\bar{A})),$$

and

$$U \sim sG\left(\sqrt{ \left(\sum_{k\in\mathscr{K}} \left(\sum_{\substack{i\notin\mathscr{K}\\ i>k}} |\bar{A}^{(i-1-k)_+}| \right)^2 + \left(\sum_{\substack{i\in\mathscr{K}\\ i>k}} \bar{A}^{(i-1-k)_+} \right)^2 \right) \sigma^2 } \right) = sG(\sigma_2(\bar{A})).$$

In addition, $\mathbb{E}[|\bar{d}_i|] = c\sigma$ with some positive constant c since $\bar{d}_i$ is sub-Gaussian with parameter σ. Thus, the expectations of L and U can be written as

$$\mathbb{E}[U] = -\mathbb{E}[L] = c\sigma \left(\sum_{k\in\mathscr{K}} \sum_{i\notin\mathscr{K}} |\bar{A}^{(i-1-k)_+}| \right).$$

Since L and U are sub-Gaussian, we have the following concentration bounds:

$$\mathbb{P}\left(L > \mathbb{E}[L] + t\right) \leq e^{-\frac{t^2}{2\sigma_1^2(\bar{A})}}, \qquad \mathbb{P}\left(U < \mathbb{E}[U] - t\right) \leq e^{-\frac{t^2}{2\sigma_2^2(\bar{A})}}.$$

Therefore, $1 - \mathbb{P}(L < 0 < U)$ can be upper-bounded using the union bound by substituting $t = -\mathbb{E}[L]$ and $t = \mathbb{E}[U]$:

$$1 - \mathbb{P}(L < 0 < U) \leq e^{-\frac{\mathbb{E}[L]^2}{2\sigma_1^2(\bar{A})}} + e^{-\frac{\mathbb{E}[U]^2}{2\sigma_2^2(\bar{A})}} \leq 2e^{-\frac{\mathbb{E}[U]^2}{2\sigma_3^2(\bar{A})}}.$$

where $\sigma_3^2(\bar{A}) = \max\{\sigma_1^2(\bar{A}), \sigma_2^2(\bar{A})\}$. Moreover, we have

$$\mathbb{E}[U] = c\sigma \left(\sum_{k\in\mathscr{K}} \sum_{i\notin\mathscr{K}} |\bar{A}^{(i-1-k)_+}| \right) = c\sigma \left(\sum_{i\notin\mathscr{K}} \sum_{k\in\mathscr{K}} |\bar{A}^{(i-1-k)_+}| \right) \gtrsim \sigma(1-p)T|\bar{A}|,$$

and

$$\begin{aligned}\sigma_3^2(\bar{A}) &\leq \left(\sum_{k\in\mathscr{K}} \left(\sum_{\substack{i\notin\mathscr{K}\\ i>k}} |\bar{A}^{(i-1-k)_+}| \right)^2 + \left(\sum_{\substack{i\in\mathscr{K}\\ i>k}} |\bar{A}^{(i-1-k)_+}| \right)^2 \right) \sigma^2 \\ &\lesssim |K| \frac{\sigma^2}{(1-|\bar{A}|)^2} \lesssim \frac{pT\sigma^2}{(1-|\bar{A}|)^2}.\end{aligned}$$

As a result,

$$1-\mathbb{P}(L<0<U) \lesssim 2\exp\left\{-\frac{(1-p)^2}{p}(1-|\bar{A}|)^2|\bar{A}|^2 T\right\} = \delta.$$

If $1-\mathbb{P}(L<0<U) \leq \delta$, solving for T implies that if $T \gtrsim \frac{p}{(1-p)^2}\frac{1}{(1-|\bar{A}|)^2|\bar{A}|^2}\log(2/\delta)$, $\bar{A}$ is a solution to the problem with probability at least $1-\delta$. □

B.4 Proof of Proposition 2

Proof. By using (3), the necessary and sufficient condition for this problem is

$$0 \in \sum_{i\notin\mathscr{K}} x_i \otimes \partial\|(\bar{A}-A)x_i\|_2 + \sum_{i\in\mathscr{K}} x_i \otimes \partial\|(\bar{A}-A)x_i + \bar{d}_i\|_2.$$

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

$$0 \in \sum_{i\notin\mathscr{K}} x_i \otimes \partial\|0\|_2 + \sum_{i\in\mathscr{K}} x_i \otimes \partial\|\bar{d}_i\|_2. \tag{8}$$

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

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

The matrix 0 belongs to the right-hand side term for arbitrary $\bar{d}_{i_1+\Delta}$ if $span\{\bar{d}_{i_1}, \bar{A}\bar{d}_{i_1},\dots,\bar{A}^{\Delta-2}\bar{d}_{i_1}\} \in \mathbb{R}^n$. Hence, we require those vectors to span $\mathbb{R}^n$. Because $\bar{A}$ has distinct eigenvalues, it has n distinct eigenvectors as well. As a result, as long as $\bar{d}_{i_1}$ is not a multiple of the eigenvectors of $\bar{A}$ as stated in the theorem, we have

$$span\{\bar{d}_{i_1}, \bar{A}\bar{d}_{i_1},\dots,\bar{A}^{\Delta-2}\bar{d}_{i_1}\} \in \mathbb{R}^n.$$

However, this is not sufficient to ensure that KKT condition (3) holds. The reason is that $\partial\|0\|_2 = \{x\in\mathbb{R}^n : \|x\|_2 \leq 1\}$. The vectors chosen for $\partial\|0\|_2$ have a bounded norm. Therefore, we need a condition that bounds 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 after dropping the subindex i_1:

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

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

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

Because the disturbance vectors are not aligned with the eigenvectors of $\bar{A}^T$ by assumption, we have $\bar{d}^T v_j \neq 0, j = 1,\dots,n$. Therefore, the KKT condition holds if

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

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

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

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

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

Using the fact that $\|e_j\|_2 = 1$ for $j = 1,\ldots,\Delta-n$ and $\|f\|_2 = 1$, we obtain

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

This completes the proof for the proposition. □

B.5 Proof of Theorem 4

Proof. From Proposition 2, $\bar{A}$ is a solution if and only if

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

Our goal is to show that 0 is included in the expectation of the right-hand side term. The randomness of the system stems from the set of attack vectors $\bar{d}_i, \forall i \in \mathscr{K}$. Define $S := |\mathscr{K}|$ and $\mathscr{K} := \{i_1, i_2, \ldots, i_S\}$. Let the set of attack vectors be $\bar{d}_{\mathscr{K}} = \{\bar{d}_{i_1}, \bar{d}_{i_2}, \ldots, \bar{d}_{i_S}\}$. Taking the expectation with respect to the random vectors $\bar{d}_i, \forall i \in \mathscr{K}$, give rise to

$$\begin{aligned}
\mathbb{E}_{\bar{d}_{\mathscr{K}}}\left[\sum_{i\notin\mathscr{K}} x_i \otimes \partial\|0\|_2 + \sum_{i\in\mathscr{K}} x_i \otimes \partial\|\bar{d}_i\|_2\right] &= \sum_{i\notin\mathscr{K}} \mathbb{E}_{\bar{d}_{\mathscr{K}}}[x_i] \otimes \partial\|0\|_2 \\
&\quad + \mathbb{E}_{\bar{d}_{\mathscr{K}\setminus\{i_S\}}}\left[\mathbb{E}_{\bar{d}_{i_S}}\left[\sum_{i\in\mathscr{K}} x_i \otimes \partial\|\bar{d}_i\|_2 \middle| \bar{d}_{\mathscr{K}\setminus\{i_S\}}\right]\right] \\
&= \sum_{i\notin\mathscr{K}} \mathbb{E}_{\bar{d}_{\mathscr{K}}}[x_i] \otimes \partial\|0\|_2 + \\
&\quad \mathbb{E}_{\bar{d}_{\mathscr{K}\setminus\{i_S\}}}\left[\sum_{i\in\mathscr{K}\setminus\{i_S\}} x_i \otimes \partial\|\bar{d}_i\|_2 + x_{i_S} \otimes (\partial\|0\|_2 \odot \mathbb{P}(\bar{d}_{i_S} = 0))\right].
\end{aligned}$$

Here, $\mathbb{P}(\bar{d}_{i_S} = 0)$ is a vector of dimension n and the j-th entry of this vector is equal to $\mathbb{P}(\bar{d}^j_{i_S} = 0)$. Since $\mathbb{P}(\bar{d}^j_{i_S} < 0) = \mathbb{P}(\bar{d}^j_{i_S} > 0)$, we have $\mathbb{E}[x_{i_S} \otimes \partial\|\bar{d}_{i_S}\|_2 | \bar{d}_{\mathscr{K}\setminus\{i_S\}}] = x_{i_S} \otimes (\partial\|0\|_2 \odot \mathbb{P}(\bar{d}_{i_S} = 0))$. If the conditional expectation is taken iteratively, we obtain that

$$\mathbb{E}_{\bar{d}_K}\left[\sum_{i\notin\mathscr{K}} x_i \otimes \partial\|0\|_2 + \sum_{i\in\mathscr{K}} x_i \otimes \partial\|\bar{d}_i\|_2\right] = \sum_{i\notin\mathscr{K}} \mathbb{E}_{\bar{d}_{\mathscr{K}}}[|x_i|] \otimes \partial\|0\|_2 + \sum_{i\in\mathscr{K}} x_i \otimes (\partial\|0\|_2 \odot \mathbb{P}(\bar{d}_i = 0)).$$

For the subdifferential of $\|\cdot\|_2$, it is known that $0 \in \partial\|0\|_2$. Consequently,

$$0 \in \mathbb{E}_{\bar{d}_{\mathscr{K}}}\left[\sum_{i\notin\mathscr{K}} x_i \otimes \partial\|0\|_2 + \sum_{i\in\mathscr{K}} x_i \otimes \partial\|\bar{d}_i\|_2\right].$$

Therefore, $\bar{A}$ is a solution almost surely due to the KKT condition and the strong law of large numbers. □

B.6 Proof of Theorem 5

Proof. Since the matrix $\bar{A}$ is diagonalizable due to the existence of n linearly independent eigenvectors, it can be written as $\bar{A} = O\bar{\Lambda}O^T$, where $O \in \mathbb{R}^n$ is an orthonormal matrix and $\bar{\Lambda}$ is the diagonal matrix with diagonal entries $\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_n$. From Proposition 2, we know that $\bar{A}$ is a solution if and only if

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

Moreover,

$$x_i = \sum_{k\in\mathscr{K}} \bar{A}^{(i-1-k)_+}\bar{d}_k = O\left(\sum_{k\in\mathscr{K}} \bar{\Lambda}^{(i-1-k)_+}\right) O^T\bar{d}_k$$

due to the system dynamics, diagonalization of $\bar{A}$, and $x_0 = 0$. We can substitute this into the KKT condition to arrive at

$$0 \in O\sum_{i\notin\mathscr{K}}\left(\sum_{k\in\mathscr{K}} \bar{\Lambda}^{(i-1-k)_+}\right) O^T\bar{d}_k \otimes \partial\|0\|_2 + O\sum_{i\in\mathscr{K}}\left(\sum_{k\in\mathscr{K}} \bar{\Lambda}^{(i-1-k)_+}\right) O^T\bar{d}_k \otimes \frac{\bar{d}_i}{\|\bar{d}_i\|_2}.$$

Since O is an orthonormal matrix, $\tilde{d}_k := O^T\bar{d}_k$ is still a sub-Gaussian vector with the parameter σ. In addition, the random variable $\frac{\bar{d}_i}{\|\bar{d}_i\|_2}$ is distributed over the unit sphere $\mathbb{S}_{n-1}$. As a result, the KKT condition can be rewritten as

$$0 \in \sum_{i\notin\mathscr{K}}\sum_{k\in\mathscr{K}} \bar{\Lambda}^{(i-1-k)_+}\tilde{d}_k \otimes \partial\|0\|_2 + \sum_{i\in\mathscr{K}}\sum_{k\in\mathscr{K}} \bar{\Lambda}^{(i-1-k)_+}\tilde{d}_k \otimes \frac{\bar{d}_i}{\|\bar{d}_i\|_2}.$$

Note that the KKT condition is in the matrix form due to the outer product. It is desirable to find the sample complexity for each entry of the matrix, for which we utilize the union bound to guarantee the satisfaction of the KKT condition. The (l,j)-th entry of the matrix in the KKT condition can be written as

$$0_{l,j} \in \sum_{i\notin\mathscr{K}}\sum_{k\in\mathscr{K}} \bar{\lambda}_l^{(i-1-k)_+}\tilde{d}_k^l \otimes \partial\|0\|_2^j + \sum_{i\in\mathscr{K}}\sum_{k\in\mathscr{K}} \bar{\lambda}_l^{(i-1-k)_+}\tilde{d}_k^l \otimes \frac{\bar{d}_i^j}{\|\bar{d}_i\|_2},$$

where $\bar{\lambda}_l$ represents the l-th eigenvalue of the matrix $\bar{A}$ and $\bar{d}_k^l$ denotes the l-th entry of the vector $\bar{d}_k$. The right-hand side of the above condition has the minimum value

$$L := \sum_{k\in\mathscr{K}}\left(-\sum_{\substack{i\notin\mathscr{K}\\ i>k}} |\bar{\lambda}_l^{(i-1-k)_+}|\partial\|\bar{d}_k^l\|_2 + \sum_{\substack{i\in\mathscr{K}\\ i>k}} \bar{\lambda}_l^{(i-1-k)_+}\frac{\bar{d}_i^j}{\|\bar{d}_i\|_2}\right)\bar{d}_k^l,$$

and the maximum value

$$U := \sum_{k\in\mathscr{K}}\left(\sum_{\substack{i\notin\mathscr{K}\\ i>k}} |\bar{\lambda}_l^{(i-1-k)_+}|\partial\|\bar{d}_k^l\|_2 + \sum_{\substack{i\in\mathscr{K}\\ i>k}} \bar{\lambda}_l^{(i-1-k)_+}\frac{\bar{d}_i^j}{\|\bar{d}_i\|_2}\right)\bar{d}_k^l.$$

Since each vector $\bar{d}_k$ is sub-Gaussian vector with parameter σ, its entries $\bar{d}_k^l$ for $l = 1,\ldots,n$ are sub-Gaussian with parameter σ. In addition, we know that $\frac{\bar{d}_i^j}{\|\bar{d}_i\|_2}$ is a bounded random variable between $[-1,1]$ since it is on the unit sphere. Therefore, $\frac{\bar{d}_i^j}{\|\bar{d}_i\|_2}$ is sub-Gaussian with parameter 1. We know that $\tilde{d}_k^l\frac{\bar{d}_i^j}{\|\bar{d}_i\|_2}$ are independent. Consequently, $\tilde{d}_k^l\frac{\bar{d}_i^j}{\|\bar{d}_i\|_2}$ is sub-Exponential with $(v_1\sigma, v_2\sigma)$ for some universal constants v_1 and v_2 [Vershynin, 2018]. Using similar arguments as in the proof of Theorem 3, one can conclude that the expectation of L and U are

$$\mathbb{E}[U] = -\mathbb{E}[L] = c\sigma\left(\sum_{k\in\mathscr{K}}\sum_{i\notin\mathscr{K}} |\bar{\lambda}_l^{(i-1-k)_+}|\right).$$

Since the sum of independent sub-Exponential random variables is sub-Exponential as well, it holds that

$$\begin{aligned} L \sim U &\sim sE\left(v_1\sqrt{\left(\sum_{k\in\mathscr{K}}\left(-\sum_{\substack{i\notin\mathscr{K}\\ i>k}} |\bar{\lambda}_l^{(i-1-k)_+}|\right)^2 + \left(\sum_{\substack{i\in\mathscr{K}\\ i>k}} \bar{\lambda}_l^{(i-1-k)_+}\right)^2\right)\sigma^2}, v_2\sigma\right) \\ &= sE(v_1\sigma_1(\bar{\lambda}_l), v_2\sigma). \end{aligned}$$

Since L and U are sub-Exponential, we have the following concentration bounds:

$$\mathbb{P}(L > \mathbb{E}[L] + t) \leq e^{-\min\left\{\frac{t^2}{2v_1^2\sigma_1^2(\bar{\lambda}_l)}, \frac{t}{2v_2\sigma}\right\}}, \qquad \mathbb{P}(U < \mathbb{E}[U] - t) \leq e^{-\min\left\{\frac{t^2}{2v_1^2\sigma_1^2(\bar{\lambda}_l)}, \frac{t}{2v_2\sigma}\right\}}.$$

Therefore, $1 - \mathbb{P}(L < 0 < U)$ can be upper-bounded using the union bound by substituting $t = -\mathbb{E}[L]$ and $t = \mathbb{E}[U]$:

$$1 - \mathbb{P}(L < 0 < U) \leq 2e^{-\min\left\{\frac{\mathbb{E}[U]^2}{2v_1^2\sigma_1^2(\bar{\lambda}_l)}, \frac{\mathbb{E}[U]}{2v_2\sigma}\right\}}.$$

The first part of the minimum in the exponential term results in the sample complexity of

$$T \gtrsim \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{\lambda}_l|)^2|\bar{\lambda}_l|^2} \log(2/\delta)$$

in light of Theorem 3, and the second part of the minimum in the exponential term results in the sample complexity of

$$T \gtrsim \frac{1}{(1-p)} \frac{1}{|\bar{\lambda}_l|} \log(2/\delta).$$

Therefore, the (l, j)-th entry of the matrix in the KKT condition is satisfied with probability at least $1 - \delta$ if

$$T \gtrsim \max\left\{ \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{\lambda}_l|)^2|\bar{\lambda}_l|^2} \log(2/\delta), \frac{1}{(1-p)} \frac{1}{|\bar{\lambda}_l|} \log(2/\delta) \right\}.$$

Using the union-bound over the n^2 entries of the matrix in the KKT condition, it can be concluded that $\bar{A}$ is a solution with probability at least $1 - \delta$ if

$$T \gtrsim \max\left\{ \frac{p}{(1-p)^2} \max_l \left\{ \frac{1}{|\bar{\lambda}_l|^2(1-|\bar{\lambda}_l|)^2} \right\} \log(n/\delta), \frac{1}{(1-p)} \max_l \left\{ \frac{1}{|\bar{\lambda}_l|} \right\} \log(n/\delta) \right\}.$$

□

B.7 Proof of Theorem 6

Proof. We have the same KKT condition as in the previous theorem. However, the subdifferentials of the ℓ_1 and ℓ_2 norms differ. For a vector $\bar{d}$, the i-th entry of subdifferential of its ℓ_1-norm is given as

$$[\partial \|\bar{d}\|_1]_i = \begin{cases} \text{sign}(\bar{d}^i)1, & \text{if } \bar{d}^i \neq 0 \\ [-1, 1], & \text{if } \bar{d}^i = 0 \end{cases}.$$

Consequently, the (l, j)-th entry of the matrix in the KKT condition can be written as follows using similar arguments as in the proof of Theorem 5:

$$0_{l,j} \in \sum_{i \notin \mathscr{K}} \sum_{k \in \mathscr{K}} \bar{\lambda}_l^{(i-1-k)_+} \tilde{d}_k^l \text{sign}(\bar{d}_k^l) + \sum_{i \in \mathscr{K}} \sum_{k \in \mathscr{K}} \bar{\lambda}_l^{(i-1-k)_+} \tilde{d}_k^l \text{sign}(\bar{d}_i^j).$$

The right-hand side of the above condition has the minimum value

$$L := \sum_{k \in \mathscr{K}} \left(- \sum_{\substack{i \notin \mathscr{K} \\ i > k}} |\bar{\lambda}_l^{(i-1-k)_+}| \text{sign}(\bar{d}_k^l) + \sum_{\substack{i \in \mathscr{K} \\ i > k}} \bar{\lambda}_l^{(i-1-k)_+} \text{sign}(\bar{d}_i^j) \right) \bar{d}_k^l,$$

and the maximum value

$$U := \sum_{k \in \mathscr{K}} \left(\sum_{\substack{i \notin \mathscr{K} \\ i > k}} |\bar{\lambda}_l^{(i-1-k)_+}| \text{sign}(\bar{d}_k^l) + \sum_{\substack{i \in \mathscr{K} \\ i > k}} \bar{\lambda}_l^{(i-1-k)_+} \text{sign}(\bar{d}_i^j) \right) \bar{d}_k^l.$$

This is similar to L and U in Theorem 3, where we have $\bar{\lambda}_l$ instead of $\bar{A}$. Therefore, the (l, j)-th entry of the matrix in the KKT condition is satisfied with probability at least $1 - \delta$ if

$$T \gtrsim \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{\lambda}_l|)^2|\bar{\lambda}_l|^2} \log(2/\delta).$$

Using the union-bound over the n^2 entries of the matrix in the KKT condition, it can be concluded that $\bar{A}$ is a solution with probability at least $1 - \delta$ if

$$T \gtrsim \frac{p}{(1-p)^2} \log(2/\delta).$$

□

B.8 Proof of Theorem 7

Proof. Due to the system dynamics and given $x_0 = 0$, x_i can be expressed as

$$x_i = \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{B} u_k + \sum_{k \in \mathscr{K}} \bar{A}^{(i-1-k)_+} \bar{d}_k.$$

Since $n = m = 1$, it is desirable to show that

$$0 \in \sum_{i \notin \mathscr{K}} x_i \partial \|0\|_2 + \sum_{i \in \mathscr{K}} x_i \partial \|\bar{d}_i\|_2, \tag{9a}$$

$$0 \in \sum_{i \notin \mathscr{K}} u_i \partial \|0\|_2 + \sum_{i \in \mathscr{K}} u_i \partial \|\bar{d}_i\|_2, \tag{9b}$$

hold simultaneously. We first investigate (9a). Substituting the expression for x_i into (9a) results in the relation.

$$\begin{aligned} 0 \in & \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{B} u_k \partial \|0\|_2 + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{B} u_k \partial \|\bar{d}_i\|_2 \\ & + \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{d}_k \partial \|0\|_2 + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{d}_k \partial \|\bar{d}_i\|_2. \end{aligned}$$

It is known from Theorem 3 that if $T \gtrsim \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{A}|)^2 |\bar{A}|^2} \log(8/\delta)$, then

$$\mathbb{P}\left(0 \notin \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{d}_k \partial \|0\|_2 + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{d}_k \partial \|\bar{d}_i\|_2 \right) \leq \frac{\delta}{4}.$$

Similarly, if $T \gtrsim \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{A}|)^2 |\bar{A}|^2} \log(8/\delta)$, then

$$\mathbb{P}\left(0 \notin \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{B} u_k \partial \|0\|_2 + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{B} u_k \partial \|\bar{d}_i\|_2 \right) \leq \frac{\delta}{4}.$$

Therefore, if $T \gtrsim \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{A}|)^2 |\bar{A}|^2} \log(8/\delta)$, then the first KKT condition (9a) holds with probability at least $1 - \delta/2$ by using the union bound. The right-hand side of the term in the probability term can be lower and upper-bounded as follows:

$$L := \sum_{k=0}^{T-1} \left(- \sum_{\substack{i \notin \mathscr{K} \\ i > k}} |\bar{A}^{(i-1-k)_+} \bar{B}| \partial \|u_k\|_2 + \sum_{\substack{i \in \mathscr{K} \\ i > k}} \bar{A}^{(i-1-k)_+} \bar{B} \partial \|\bar{d}_i\|_2 \right) u_k,$$

$$U := \sum_{k=0}^{T-1} \left(\sum_{\substack{i \notin \mathscr{K} \\ i > k}} |\bar{A}^{(i-1-k)_+} \bar{B}| \partial \|u_k\|_2 + \sum_{\substack{i \in \mathscr{K} \\ i > k}} \bar{A}^{(i-1-k)_+} \bar{B} \partial \|\bar{d}_i\|_2 \right) u_k.$$

We aim to show that $L < 0 < U$ with high probability. Similar to the calculations in the proof of Theorem 3, we have that $u_k \partial \|\bar{d}_i\|_2$ and $u_k \partial \|u_k\|_2 = |u_k|$ are sub-Gaussian with parameter ε. Therefore,

$$L \sim U \sim sG\left(|\bar{B}| \varepsilon \sqrt{\sum_{k=0}^{T-1} \left(\left(\sum_{\substack{i \notin \mathscr{K} \\ i > k}} |\bar{A}^{(i-1-k)_+}| \right)^2 + \left(\sum_{\substack{i \in \mathscr{K} \\ i > k}} \bar{A}^{(i-1-k)_+} \right)^2 \right)} \right) = sG(\varepsilon(\bar{A}, \bar{B})).$$

In addition, $\mathbb{E}[|u_k|] = c\varepsilon$ since u_k is sub-Gaussian with parameter ε. Thus, the expectations of L and U can be written as

$$\mathbb{E}[U] = -\mathbb{E}[L] = c\varepsilon \left(\sum_{k=0}^{T-1} \sum_{i \notin \mathscr{K}} |\bar{A}^{(i-1-k)_+}| \right) |\bar{B}|.$$

Since L and U are sub-Gaussian, the following concentration bounds hold:

$$\mathbb{P}(L > \mathbb{E}[L] + t) \leq e^{-\frac{t^2}{2\varepsilon^2(\bar{A},\bar{B})}}, \qquad \mathbb{P}(U < \mathbb{E}[U] - t) \leq e^{-\frac{t^2}{2\varepsilon^2(\bar{A},\bar{B})}}.$$

Therefore, $1 - \mathbb{P}(L < 0 < U)$ can be upper-bounded using the union bound by substituting $t = -\mathbb{E}[L]$ and $t = \mathbb{E}[U]$:

$$1 - \mathbb{P}(L < 0 < U) \leq 2e^{-\frac{\mathbb{E}[U]^2}{2\varepsilon^2(\bar{A},\bar{B})}}.$$

Moreover,

$$\mathbb{E}[U] = c\varepsilon \left(\sum_{k=0}^{T-1} \sum_{i \notin \mathscr{K}} |\bar{A}^{(i-1-k)_+}| \right) |\bar{B}| \gtrsim \varepsilon(1-p)T|\bar{A}||\bar{B}|,$$

and

$$\varepsilon^2(\bar{A},\bar{B}) \lesssim pT \frac{|\bar{B}|^2}{(1-|\bar{A}|)^2} \varepsilon^2.$$

As a result,

$$1 - \mathbb{P}(L < 0 < U) \lesssim 2\exp\left\{ -\frac{(1-p)^2}{p} |\bar{A}|^2 (1-|\bar{A}|)^2 T \right\} = \delta/4.$$

Therefore, if $T \gtrsim \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{A}|)^2|\bar{A}|^2} \log(8/\delta)$, then the first KKT condition (9a) is satisfied with probability at least $1 - \delta/2$.

Checking the second KKT condition (9b) is more straightforward than the first one. Note that lower and upper bounds for (9b) can be expressed as

$$L := -\sum_{i \notin \mathscr{K}} |u_i| + \sum_{i \in \mathscr{K}} u_i \partial \|\bar{d}_i\|_2, \tag{10a}$$

$$U := \sum_{i \notin \mathscr{K}} |u_i| + \sum_{i \in \mathscr{K}} u_i \partial \|\bar{d}_i\|_2. \tag{10b}$$

Hence,

$$L \sim U \sim sG(\varepsilon\sqrt{T}),$$

and

$$\mathbb{E}[U] = -\mathbb{E}[L] = c\varepsilon(1-p)T.$$

Then, similar calculations yield the probability bound below:

$$1 - \mathbb{P}(L < 0 < U) \lesssim 2\exp\left\{ -(1-p)^2 T \right\} = \delta/2.$$

Therefore, if $T \gtrsim \frac{1}{(1-p)^2} \log(4/\delta)$, then the second KKT condition (9b) is satisfied with probability at least $1 - \delta/2$. As a result, using the union bound, if

$$T \gtrsim \max\left\{ \frac{1}{(1-p)^2} \log(4/\delta), \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{A}|)^2|\bar{A}|^2} \log(8/\delta) \right\},$$

$(\bar{A},\bar{B})$ is a solution to the estimator with probability at least $1 - \delta$. □

B.9 Proof of Theorem 8

Proof. If the matrices $\bar{A}$ and $\bar{B}$ satisfy the KKT conditions provided in Theorem 1, we conclude that $(\bar{A},\bar{B})$ is the solution pair to the (CO-L2). The KKT conditions can be written as

$$0 \in \sum_{i \notin \mathscr{K}} x_i \otimes \partial \|0\|_2 + \sum_{i \in \mathscr{K}} x_i \otimes \partial \|\bar{d}_i\|_2, \tag{11a}$$

$$0 \in \sum_{i \notin \mathscr{K}} u_i \otimes \partial \|0\|_2 + \sum_{i \in \mathscr{K}} u_i \otimes \partial \|\bar{d}_i\|_2. \tag{11b}$$

It is desirable to show that (11a) and (11b) are satisfied with probability at least $1 - \delta/2$ when the theorem statements are satisfied separately, which will imply that exact recovery is guaranteed using

the union bound with probability at least $1-\delta$. Due to the system dynamics and given $x_0 = 0$, x_i can be expressed as

$$x_i = \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{B} u_k + \sum_{k \in \mathscr{K}} \bar{A}^{(i-1-k)_+} \bar{d}_k.$$

Substituting the expression for x_i into (11a) results in

$$\begin{aligned} 0 \in & \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{B} u_k \otimes \partial \|0\|_2 + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{B} u_k \otimes \partial \|\bar{d}_i\|_2 \\ & + \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{d}_k \otimes \partial \|0\|_2 + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{A}^{(i-1-k)_+} \bar{d}_k \otimes \partial \|\bar{d}_i\|_2. \end{aligned}$$

Since the matrix $\bar{A}$ is diagonalizable due to the existence of n linearly independent eigenvectors, it can be written as $\bar{A} = O \bar{\Lambda} O^T$, where $O \in \mathbb{R}^n$ is an orthonormal matrix and $\bar{\Lambda}$ is the diagonal matrix with diagonal entries $\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_n$. Similar to the proof of Theorem 5, the (l, j)-th entry of the matrix in the KKT condition above can be transformed to the following condition

$$\begin{aligned} 0_{l,j} \in & \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} (\tilde{B} u_k)^l \partial \|0\|_2^j + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} (\tilde{B} u_k)^l \frac{\bar{d}_i^j}{\|\bar{d}_i\|_2} \\ & + \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} \tilde{d}_k^l \partial \|0\|_2^j + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} \tilde{d}_k^l \frac{\bar{d}_i^j}{\|\bar{d}_i\|_2}, \end{aligned}$$

where $\tilde{B} = O^T \bar{B}$, and $\tilde{d}_k = O^T \bar{d}_k$. It is known from Theorem 5 that if

$$T \gtrsim \max \left\{ \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{\lambda}_l|)^2 |\bar{\lambda}_l|^2} \log(8/\delta), \frac{1}{(1-p)} \frac{1}{|\bar{\lambda}_l|} \log(8/\delta) \right\},$$

then

$$\mathbb{P}\left(0_{l,j} \notin \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} \tilde{d}_k^l \partial \|0\|_2^j + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} \tilde{d}_k^l \frac{\bar{d}_i^j}{\|\bar{d}_i\|_2} \right) \leq \frac{\delta}{4}.$$

Similarly, when we have the same order of the samples, the following probability bound holds:

$$\mathbb{P}\left(0_{l,j} \notin \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} (\tilde{B} u_k)^l \partial \|0\|_2^j + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} (\tilde{B} u_k)^l \frac{\bar{d}_i^j}{\|\bar{d}_i\|_2} \right) \leq \frac{\delta}{4}. \tag{12}$$

Note that $(\tilde{B} u_k)^l$ is a sub-Gaussian random variable with parameter $\varepsilon \|\tilde{B}_l\|_2$, where $\tilde{B}_l$ denotes the l-th row of the $\tilde{B}$. As a result, we can lower and upper bound the term in the probability bound above. The lower bound is

$$L := \sum_{k=0}^{T-1} \left(- \sum_{\substack{i \notin \mathscr{K} \\ i > k}} |\bar{\lambda}_l^{(i-1-k)_+}| \partial \|(\tilde{B} u_k)^l\|_2 + \sum_{\substack{i \in \mathscr{K} \\ i > k}} \bar{A}^{(i-1-k)_+} \bar{B} \frac{\bar{d}_i^j}{\|\bar{d}_i\|_2} \right) (\tilde{B} u_k)^l,$$

and the upper bound is

$$U := \sum_{k=0}^{T-1} \left(\sum_{\substack{i \notin \mathscr{K} \\ i > k}} |\bar{\lambda}_l^{(i-1-k)_+}| \partial \|(\tilde{B} u_k)^l\|_2 + \sum_{\substack{i \in \mathscr{K} \\ i > k}} \bar{A}^{(i-1-k)_+} \bar{B} \frac{\bar{d}_i^j}{\|\bar{d}_i\|_2} \right) (\tilde{B} u_k)^l.$$

By performing similar calculations as in the earlier proofs, one can obtain the sub-Exponential parameters for L and U, besides the expectation of L and U. More precisely,

$$\mathbb{E}[U] = -\mathbb{E}[L] = c\varepsilon \|\tilde{B}_l\|_2 \left(\sum_{k \in \mathscr{K}} \sum_{i \notin \mathscr{K}} |\bar{\lambda}_l^{(i-1-k)_+}| \right),$$

and

$$L \sim U \sim \left(v_1 \sqrt{\left(\sum_{k \in \mathscr{K}} \left(- \sum_{\substack{i \notin \mathscr{K} \\ i>k}} |\bar{\lambda}_l^{(i-1-k)_+}| \right)^2 + \left(\sum_{\substack{i \in \mathscr{K} \\ i>k}} \bar{\lambda}_l^{(i-1-k)_+} \right)^2 \right) \varepsilon^2 \|\tilde{B}_l\|_2^2}, v_2 \varepsilon \|\tilde{B}_l\|_2 \right)$$
$$= sE(v_1 \|\tilde{B}_l\|_2 \varepsilon_1(\bar{\lambda}_l), v_2 \varepsilon \|\tilde{B}_l\|_2).$$

Since L and U are sub-Exponential, we have the following concentration bounds:

$$\mathbb{P}(L > \mathbb{E}[L] + t) \leq e^{-\min\left\{ \frac{t^2}{2v_1^2 \|\tilde{B}_l\|_2^2 \varepsilon_1^2(\bar{\lambda}_l)}, \frac{t}{2v_2 \varepsilon \|\tilde{B}_l\|_2} \right\}}, \qquad \mathbb{P}(U < \mathbb{E}[U] - t) \leq e^{-\min\left\{ \frac{t^2}{2v_1^2 \|\tilde{B}_l\|_2^2 \varepsilon_1^2(\bar{\lambda}_l)}, \frac{t}{2v_2 \varepsilon \|\tilde{B}_l\|_2} \right\}}.$$

Therefore, $1 - \mathbb{P}(L < 0 < U)$ can be upper-bounded using the union bound by substituting $t = -\mathbb{E}[L]$ and $t = \mathbb{E}[U]$:

$$1 - \mathbb{P}(L < 0 < U) \leq 2e^{-\min\left\{ \frac{\mathbb{E}[U]^2}{2v_1^2 \|\tilde{B}_l\|_2^2 \varepsilon_1^2(\bar{\lambda}_l)}, \frac{\mathbb{E}[U]}{2v_2 \varepsilon \|\tilde{B}_l\|_2} \right\}}.$$

Solving the probability bound above in terms of the sample complexity T gives that if

$$T \gtrsim \max\left\{ \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{\lambda}_l|)^2 |\bar{\lambda}_l|^2} \log(8/\delta), \frac{1}{(1-p)} \frac{1}{|\bar{\lambda}_l|} \log(8/\delta) \right\},$$

then (12) holds. Hence, by taking the union bound over the n^2 entries of the matrix in the first KKT condition, if the sample complexity is

$$T \gtrsim \max\left\{ \frac{p}{(1-p)^2} \max_l \left\{ \frac{1}{|\bar{\lambda}_l|^2 (1-|\bar{\lambda}_l|)^2} \right\} \log(8n^2/\delta), \frac{1}{(1-p) \min_l |\bar{\lambda}_l|} \log(8n^2/\delta) \right\},$$

then the first KKT condition is satisfied with probability at least $1 - \delta/2$. As a next step, we can show that the second KKT condition (11b) holds with probability at least $1 - \delta/2$ as well. The (l, j)-th entry of the second KKT condition above can be written as

$$0_{l,j} \in \sum_{i \notin \mathscr{K}} u_i^l \partial \|0\|_2^j + \sum_{i \in \mathscr{K}} u_i^l \frac{\bar{d}_i^j}{\|\bar{d}_i\|}.$$

We again define lower and upper bounds for the above expression using the flexibility of the subdifferential at the origin, given as

$$L := -\sum_{i \notin \mathscr{K}} |u_i^l| + \sum_{i \in \mathscr{K}} u_i^l \frac{\bar{d}_i^j}{\|\bar{d}_i\|},$$

and

$$U := \sum_{i \notin \mathscr{K}} |u_i^l| + \sum_{i \in \mathscr{K}} u_i^l \frac{\bar{d}_i^j}{\|\bar{d}_i\|},$$

respectively. Note that L and U are sub-Exponential random variables:

$$L \sim U \sim sE(v_1 \varepsilon \sqrt{T}, v_2 \varepsilon).$$

Moreover, $\mathbb{E}[U] = -\mathbb{E}[L] = c\varepsilon(1-p)T$. As before, we obtain that

$$1 - \mathbb{P}(L < 0 < U) \lesssim e^{-\min\left\{ \frac{\mathbb{E}[U]^2}{2v_1^2 \varepsilon^2 T}, \frac{\mathbb{E}[U]}{2v_2^2 \varepsilon} \right\}}.$$

Therefore, if $T \gtrsim \max\left\{ \frac{1}{(1-p)^2} \log(4/\delta), \frac{1}{(1-p)} \log(4/\delta) \right\} = \frac{1}{(1-p)^2} \log(4/\delta)$, then each element of the second KKT condition is satisfied with probability at least $1 - \delta/2$. If we take the union bound over the mn entries of the second KKT condition, we conclude that the condition (11b) is satisfied with probability at least $1 - \delta/2$ if

$$T \gtrsim \frac{1}{(1-p)^2} \log(4mn/\delta).$$

Hence, the union bound over the two KKT conditions completes the proof.

□

B.10 Proof of Theorem 9

Proof. Similar to the proof of Theorem 8, the KKT conditions for problem (CO-L1) can be written as

$$0 \in \sum_{i \notin \mathscr{K}} x_i \otimes \partial \|0\|_1 + \sum_{i \in \mathscr{K}} x_i \otimes \partial \|\bar{d}_i\|_1, \tag{13a}$$

$$0 \in \sum_{i \notin \mathscr{K}} u_i \otimes \partial \|0\|_1 + \sum_{i \in \mathscr{K}} u_i \otimes \partial \|\bar{d}_i\|_1. \tag{13b}$$

Using similar transformations and arguments, one can show that if the sample complexity is

$$T \gtrsim \frac{p}{(1-p)^2} \frac{1}{(1-|\bar{\lambda}_l|)^2 |\bar{\lambda}_l|^2} \log(8/\delta),$$

then the following probability bounds hold for each entry of the first KKT condition (13a):

$$\mathbb{P}\left(0_{l,j} \notin \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} \bar{d}_k^l \partial \|0\|_1^j + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} \bar{d}_k^l \partial \|\bar{d}_i^j\|_1 \right) \leq \frac{\delta}{4},$$

and

$$\mathbb{P}\left(0_{l,j} \notin \sum_{i \notin \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} (\tilde{B} u_k)^l \partial \|0\|_1^j + \sum_{i \in \mathscr{K}} \sum_{k=0}^{T-1} \bar{\lambda}_l^{(i-1-k)_+} (\tilde{B} u_k)^l \partial \|\bar{d}_i^j\|_1 \right) \leq \frac{\delta}{4}.$$

Thus, whenever

$$T \gtrsim \frac{p}{(1-p)^2} \max_l \left\{ \frac{1}{|\bar{\lambda}_l|^2 (1-|\bar{\lambda}_l|)^2} \right\} \log(8n^2/\delta),$$

the condition (13a) is satisfied with probability at least $1-\delta/2$. By adopting the same arguments as in the proof of Theorem 8, we can show that each entry of the second KKT condition (13b) is satisfied with probability at least $1-\delta/2$ if

$$T \gtrsim \frac{1}{(1-p)^2} \log(4/\delta).$$

Therefore, the condition (13b) is satisfied with probability at least $1-\delta/2$ when

$$T \gtrsim \frac{1}{(1-p)^2} \log(4mn/\delta).$$

Hence, the union bound over the two KKT conditions completes the proof. □

C Background on High Dimensional Statistics

Definition 2 (sub-Gaussian Random Variable [Wainwright, 2019])**.** *A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian with parameter σ if*

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

It is denoted as $X \sim sG(\sigma)$.

Based on the definition above, it is trivial to show that the sum of sub-Gaussian random variables is also sub-Gaussian. Specifically, if the independent random variables X_i are sub-Gaussian with mean μ_i and parameter σ_i, then $\sum_{i=1}^n (X_i - \mu_i)$ is sub-Gaussian with mean 0 and parameter $\sqrt{\sum_{i=1}^n \sigma_i^2}$. Therefore, we can have the following concentration bound for the sum of sub-Gaussian random variables.

Lemma 2. *(Hoeffding Bound [Wainwright, 2019]) Suppose that the variables $X_i, i = 1, \ldots, n$, are independent, and that X_i has mean μ_i and sub-Gaussian parameter σ_i. Then, for all $t \geq 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \geq t \right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2} \right).$$

Since the subgradients involve the absolute values of the sub-Gaussian random variables, it is useful to understand the properties of the random variable $|X|$ when X is sub-Gaussian with parameter σ. The following lemma characterizes the properties of $|X|$.

Lemma 3. *[Vershynin, 2018] Given a zero mean sub-Gaussian random variable X with mean 0 and parameter σ, the random variable $|X|$ is sub-Gaussian with parameter σ. Moreover, there exists a constant c such that $\mathbb{E}[|X|] = c\sigma$.*

Because subgradients are bounded random variables, we utilize the following Lemma in our proofs.

Lemma 4. *Suppose that the random variable X is bounded and belongs to the interval $[a,b]$. Then, X is sub-Gaussian with parameter $(b-a)/2$.*

For high-dimensional problems, we define sub-Gaussian vectors. Each element of a sub-Gaussian vector with parameter σ is also sub-Gaussian with parameter σ.

Definition 3. *(Sub-Gaussian Vector [Wainwright, 2019]) A random vector $X \in \mathbb{R}^d$ with zero mean is sub-Gaussian with parameter σ if*

$$\mathbb{E}[e^{\lambda\langle v,X\rangle}] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in \mathbb{R},$$

for every vector $v \in \mathbb{R}^d$ with unit norm, i.e., $\|v\|_2 = 1$.

Definition 4 (sub-Exponential Random Variable [Wainwright, 2019])**.** *A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Exponential with parameters (ν, α) if*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\nu^2\lambda^2/2}, \quad \forall |\lambda| \in 1/\alpha.$$

It is denoted as $X \sim sE(\nu, \alpha)$

We have the following tail bound for sub-Exponential random variables, which is analogous to the Hoefding bound for sub-Gaussian random variables.

Lemma 5. *(Sub-Exponential Tail Bound [Wainwright, 2019]) Suppose that X is sub-Exponential with parameters (v, α). Then,*

$$\mathbb{P}(X-\mu \geq t) \leq \begin{cases} e^{-t^2/2v^2}, & \text{if } 0 \leq t \leq v^2/\alpha \\ e^{-t/2\alpha}, & \text{if } t \geq v^2/\alpha \end{cases}.$$

The sum of independent sub-Exponential random variables with mean μ_i and parameters (ν_i, α_i) is also sub-Exponential with parameters $\left(\sqrt{\sum_i \nu_i^2}, \max_i \alpha_i\right)$. We also provide a lemma for the multiplication of two sub-Gaussian random variables.

Lemma 6. *[Vershynin, 2018] Given two independent zero-mean sub-Gaussian random variables X_1 and X_2 with paramters σ_1 and σ_2, there exist scalar constants c_1 and c_2 such that X_1X_2 is sub-Exponential with parameters $(c_1\sigma_1\sigma_2, c_2\sigma_1\sigma_2)$.*