**Title**

Optimization Problems with Matrix Variables in Machine Learning and Control Applications



**ISBN**

9798288861833

**Author**

Yalcin, Baturalp

**Publication Date**

2025-05-15

Peer reviewed|Thesis/dissertation

# Optimization Problems with Matrix Variables in Machine Learning and Control Applications

by

Baturalp Yalcin

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Javad Lavaei, Chair
Associate Professor Paul Grigas
Associate Professor Somayeh Sojoudi

Spring 2025

Optimization Problems with Matrix Variables in Machine Learning and Control Applications

Abstract

Optimization Problems with Matrix Variables in Machine Learning and Control Applications

by

Baturalp Yalcin

Doctor of Philosophy in Industrial Engineering and Operations Research

University of California, Berkeley

Associate Professor Javad Lavaei, Chair

This dissertation aims to advance the understanding and solution methods for optimization problems involving matrix variables. These problems are inherently complex and require sophisticated solution approaches. Despite their difficulty, they have critical applications in machine learning and control systems. We integrate high-dimensional statistics, machine learning, system theory, and mathematical optimization techniques to address key challenges in analyzing their solution landscapes and computational feasibility. The dissertation is divided into two parts.

In the first part, we study low-rank matrix sensing and completion problems, which are fundamental in recommendation systems and collaborative filtering. The restricted isometry property (RIP) is sufficient for solving matrix sensing problems using the Burer-Monteiro (B-M) factorization approach. However, when the RIP condition is not satisfied, these problems are considered hard. A widely studied case of matrix recovery without RIP is the matrix completion problem. First, we identify a class of matrix completion problems that can be solved using simple graph traversal algorithms in polynomial time. Next, we demonstrate that for this class of polynomial-time solvable problems, the B-M factorization approach gives rise to exponentially many spurious local minima, causing gradient descent algorithms to fail in recovering ground-truth matrices almost surely. This highlights a gap between the information-theoretic and optimization complexity of the problem. On the other hand, another widely used convex semidefinite programming (SDP) formulation approach achieves the exact recovery for the same class of problems. However, we also identify another class of matrix completion problems where the SDP formulation fails, whereas the B-M factorization produces an optimization landscape where all local minima are global. Consequently, when the RIP conditions are not met, neither approach strictly outperforms the other for matrix completion.

In the second part, we focus on the system identification problem for dynamical systems in the presence of adversarial disturbances. System identification involves learning a system matrix that governs the relationship between system states and inputs across time. In safety-critical systems, such as power grids, autonomous vehicles, and unmanned aerial vehicles, adversarial agents may

attempt to manipulate system dynamics. The least-squares estimator, commonly used for learning system dynamics, is highly susceptible to these outliers and adversarial perturbations. To address this vulnerability, we analyze non-smooth estimators and establish their finite-time recovery guarantees. We show that for both linear time-invariant and parametrized nonlinear systems, non-smooth estimators can recover the exact system matrices in finite time with high probability. In contrast, the least-squares estimator fails to do so. We propose a first-order subgradient method to iteratively update estimates at each time step, leveraging information from previous estimates. This approach eliminates the need to solve a new optimization problem from scratch at each time step, significantly reducing computational complexity.

*To my family, for ensuring every educational opportunity available to me.*

# Contents

# List of Figures

# Acknowledgments

Completing this dissertation and my Ph.D. has been challenging and deeply rewarding. This milestone would not have been possible without the unwavering support, guidance, and encouragement of many colleagues, friends, and family members. I am profoundly grateful to each of them and extend my sincerest appreciation for their contributions to this journey.

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Javad Lavaei. His continuous support, invaluable insights, and mentorship throughout my Ph.D. have shaped this research and my growth as a scholar and thinker. His ability to balance rigor with encouragement pushed me to refine my ideas and strive for excellence. I am incredibly grateful for his patience, generosity with his time, and unwavering belief in my abilities, even during moments of self-doubt. His mentorship has been instrumental in shaping my academic and professional trajectory, and I will carry the lessons I have learned from him throughout my career.

I am also deeply thankful to my dissertation committee members, Prof. Somayeh Sojoudi and Prof. Paul Grigas, for their time, expertise, and thoughtful feedback, which have significantly strengthened this work. Their guidance has helped me become a better researcher and refine the quality of this dissertation. I sincerely appreciate their insightful questions and constructive critiques, which challenged me to think more critically and rigorously. Their support and encouragement throughout this process have been invaluable, and I feel incredibly fortunate to have had them as part of my committee.

This journey would not have been possible without my esteemed co-authors and colleagues. I am incredibly grateful for the opportunity to collaborate with Prof. Eduardo Sontag from Northeastern University. His detailed insights and precision significantly contributed to refining aspects of this research. Working with Prof. Murat Arcak was also a privilege; his system theory expertise and engaging classes were highlights of my time at Berkeley. I am also thankful to Haixiang Zhang, Ziye (Gaven) Ma, and Han Feng for their invaluable insights, mentorship, and collaboration.

I owe a great deal to the UC Berkeley IEOR department, which has been both an academic home and a source of support throughout my Ph.D. Among my cohort, Dilara Aykanat and Anna Deza deserve special recognition for their constant mental and emotional support. They were the first two friends I made in Berkeley, and our friendship has only grown stronger over the years. During my most challenging times, the IEOR department staff provided immense support. I am deeply indebted to Heather Iwata for her kindness, regular check-ins, and reassuring presence during challenging moments. I am also grateful to my research group peers: Sang Woo Park, Igor Molybog, Yuhao Ding, Julie Mulvaney-Kemp, Donghao Ying, Ying Chen, Hyunin Lee, Jihun Kim, and Eli Brock for the engaging discussions and Friday meetings that made this journey more enjoyable.

Moving to a different country to pursue my dream of earning a Ph.D. came with its challenges, but I am incredibly fortunate to have found a family abroad. Mariana Tejada, Sid Sun, Duncan Aronson, and Anushka Drescher, your wisdom, cheerful conversations, and unwavering support made Berkeley feel like home. I am also indebted to Efe Erginbas, Kadircan Godeneli, Tanya Deniz Ipek, Oguzhan Akcin, Ekin Karasan, and Mert Cemri for their incredible companionship

and for creating a sense of belonging. Most importantly, I am deeply grateful to Noah Snable for his unwavering support, steady presence, and the light he brought to even my hardest days.

I must also acknowledge my friends in Turkey, who have been more than just friends; they have been family. I am incredibly fortunate to have them in my life. Special thanks to Seda Dilek Yetut for our mental health phone calls, Gülce Yavuz, Fevzi Berk Colakoglu, Sercan Demirkilic, and İrem Bilgun for our biweekly Zoom calls that provided much needed connection and support. I am also grateful to İrem Işık, Doğuş Baştürk, Gökhan Özdemir, Utku Emre Ergin, and İbrahim Erol for their continuous encouragement throughout this journey.

Finally, my deepest gratitude goes to my family, who have provided the foundation for all my achievements. Their belief in the power of education made this journey possible. As a first-generation student, I owe everything to their dedication and support. To my mother, Münevver Yalcin, my father, Mevlüt Yalcin, and my brother, Erdem Yalcin—thank you for believing in me, even when I doubted myself.

# Chapter 1

# Introduction

Matrices capture intricate relationships between entities. With the surge of machine learning applications and the growth of big data, optimization problems involving matrix variables have become fundamental to data-driven applications. These problems aim to accurately learn relationships between entities using observations from unknown matrices or systems. Although optimization problems with matrix variables conceptually resemble those with vector variables, they are more challenging to analyze, and their solution algorithms are generally less scalable.

Recommendation systems have gained significant popularity in recent years with the rise of streaming services and online retail platforms. These services aim to learn user preferences for products by identifying users with similar tastes. As platforms grow, the user-product matrix can reach millions of rows and columns, yet the observations from this matrix remain sparse since users cannot interact with all products. Completing the missing preferences in this matrix from limited observations is generally NP-hard. However, convex relaxations of NP-hard formulations and first-order algorithms, such as stochastic gradient descent, can efficiently recover the matrix. Despite this empirical success, a gap remains in the theoretical understanding of these problems, particularly regarding (i) how solution algorithms behave when sufficient conditions for recovery are not met and (ii) which solution method is most practical given the structure of the available observations.

Another key application of optimization problems with matrix variables is system identification for safety-critical dynamical systems, including autonomous vehicles, power grids, and robotics. In such systems, the current state determines the next state through a complex set of differential equations. Here, matrices represent system dynamics, and the objective is to learn the unknown dynamics, i.e., the unknown system matrices. Although the system identification problem can be formulated as a convex optimization problem with convex loss functions, there is growing interest in understanding how estimation error scales with the number of observations. As more observations become available, the optimization landscape evolves, causing the global solution of the convex loss function to converge to the true system dynamics. However, smooth convex loss functions are not robust to large outliers or adversarial perturbations. Yet, accurate estimation of safety-critical systems under adversaries is essential for effective control. This motivates the need (i) to utilize non-smooth estimators for robust system identification and (ii) to develop tractable

algorithms to solve convex problems with non-smooth objectives in control applications.

This chapter first provides a general overview of optimization problems involving matrix variables and explores how their solution landscapes evolve. It then summarizes the contributions of each chapter and the related publications that form the basis of this dissertation. Finally, it introduces the notations used throughout the dissertation.

## 1.1 Landscape of Optimization Problems with Matrix Variables

This dissertation studies the optimization problem with the matrix variable in the following form:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \quad f(\mathbf{X};\mathcal{A},\mathbf{X}^*) \quad \text{s.t.} \quad \mathbf{X}\in\mathcal{X}(\mathcal{A},\mathbf{X}^*),$$

where;

- The term $\mathcal{A}$ represents the set of observations and measurements from the ground-truth matrix $\mathbf{X}^*$. The set of observations determines how well-specified the function $f$ and the set $\mathcal{X}$ are. When the number of observations increases, we collect more information from the ground-truth matrix. Hence, the problem becomes well-specified and easier to solve. In the recommendation system application, $\mathcal{A}$ may be the set of all the rated movies by the users on the streaming platform. Alternatively, it also denotes the duration of the data collection from the dynamical system when the unknown system dynamics are learned using the system identification problem.

- *The decision variable* $\mathbf{X}\in\mathbb{R}^{m\times n}$ denotes $m\times n$ dimensional matrix. The goal is to find a matrix $\mathbf{X}$ as close to the ground-truth matrix $\mathbf{X}^*$. For example, in the context of recommendation systems for the streaming platform, $\mathbf{X}$ will be the rating matrix of the $n$ movies by $m$ users. Based on the solution $\mathbf{X}$, the relevant movies with highly estimated ratings by the user can be recommended to the users. In another application, $\mathbf{X}\in\mathbb{R}^{n\times n}$ may stand for the system dynamics equations for a system with $n$ states.

- The function $f(\mathbf{X};\mathcal{A})$ is *the loss function* that penalizes the mismatch between the decision variable $\mathbf{X}$ and the observations from the ground-truth matrix $\mathbf{X}^*$. Throughout this dissertation, the function $f$ is a convex loss function with respect to $\mathbf{X}$. The most common loss function used in the context of this dissertation is the squared-loss function. Alternatively, it might be a non-smooth loss function for robust learning applications. We note that the objective function and its landscape change over time with the observation set growing.

- The set $\mathcal{X}(\mathcal{A},\mathbf{X}^*)$ is *the feasible set* for the optimization problem. It consists of all the feasible matrices with respect to the observations from the ground-truth matrix and the system. In the context of the low-rank matrix optimization problems, the feasible set contains all

the matrices with a rank lower than $r$ and conforms to all the observations we have from the ground truth matrix. Moreover, the system identification problem must have a feasible set that the matrix must satisfy all the system update equations over time. We note that the feasible set evolves as more observations become available.

Most optimization problems with matrix variables, including those discussed in this dissertation, can be formulated within the framework outlined above. When the loss function $f$ is smooth and convex, and the feasible set $\mathcal{X}$ is convex with respect to the observations, the optimization problem is tractable. It can be solved using convex optimization techniques. However, specific challenges related to the objective function and feasible set can disrupt these favorable properties, necessitating further study.

- The **number of observations and measurements** plays a crucial role in determining the feasibility of matrix recovery. In the trivial case of no observations, recovering the ground-truth matrix is impossible. The number of available observations directly influences the amount of information about the underlying matrix: the more observations we have, the more well-specified the problem becomes. Ideally, the ground-truth matrix should be the unique solution that satisfies all given measurements. However, learning the matrix may become infeasible if the measurements are concentrated in specific regions of the matrix or follow a particular pattern. For example, an observation structure that only captures zero entries of a sparse matrix fails to provide useful information. Therefore, it is essential to ensure that the measurement operator $\mathcal{A}$ collects a sufficient number of measurements with adequate randomness to guarantee a well-posed optimization problem. The study of the number of required measurements, known as sample complexity analysis, is of fundamental importance in this context.

- Another challenge arises from **the non-convexity of the feasible set** in low-rank matrix optimization problems. Typically, the rank of the decision variable is constrained by a pre-specified upper bound, which is enforced using the rank function that is a non-smooth and non-convex operator. A common approach to addressing this issue is to replace the rank function with its convex surrogate, the nuclear norm, in its convex relaxation problem. However, it is crucial to understand the properties of the solution obtained through this relaxation. Specifically, we seek to determine the conditions under which the solution to the relaxed problem satisfies the original rank constraint despite the non-convex nature of the feasible set.

- An alternative approach to handling rank constraints is through factorization of the decision variable. If the matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is required to have rank at most $r$, it can be parameterized as $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{m \times r}$ and $\mathbf{V} \in \mathbb{R}^{n \times r}$. This factorization enforces the rank constraint automatically but introduces **non-convexity into the objective function** $f(\mathbf{U}\mathbf{V}^\top; \mathcal{A})$ with respect to the decision variables $\mathbf{U}$ and $\mathbf{V}$. The resulting optimization landscape may contain multiple local minima that are not global, making the problem intractable in general. However, in the context of low-rank matrix optimization, it is possible to show that under

certain conditions, the landscape demonstrates benign properties. When these conditions hold, simple first-order algorithms, such as gradient descent, can efficiently find the global minimum. Therefore, a key objective is identifying the necessary and sufficient conditions that characterize the optimization landscape of factorized formulations.

- Another complication arises from **the non-smoothness of the loss function.** While smooth loss functions are commonly used, they often lack robustness in learning the ground-truth matrix when measurements are contaminated by noise or adversarial perturbations. This issue is particularly critical in safety-critical systems, where malicious attacks can significantly corrupt the measurements. Analyzing non-smooth loss functions presents several challenges: (i) first-order optimality conditions involve subdifferentials, which consist of sets of subgradients rather than single gradients; (ii) unlike squared loss functions, non-smooth loss functions often lack closed-form solutions; and (iii), optimization algorithms for non-smooth functions typically exhibit slower convergence rates compared to their smooth counterparts.

Throughout this dissertation, we aim to address these challenges in optimization problems involving matrix variables, with a particular focus on applications in machine learning and control.

## 1.2 Summary of Contributions

Due to the challenges outlined above, this dissertation plays a fundamental role in advancing the understanding of the optimization landscape for problems involving matrix variables. Effectively addressing these complexities will lead to better-informed choices of scalable solution methods and algorithms. In this section, we present a summary of the key contributions of this dissertation.

### Low-Rank Matrix Optimization

- **Chapter 2**

  It is well-known that the Burer-Monteiro (B-M) factorization approach can efficiently solve low-rank matrix optimization problems under the RIP condition. It is natural to ask whether B-M factorization-based methods can succeed on any low-rank matrix optimization problems with low information-theoretic complexity, i.e., polynomial-time solvable problems that have a unique solution. In this chapter, we provide a negative answer to the above question. We investigate the landscape of B-M factorized polynomial-time solvable matrix completion (MC) problems, which are the most popular subclass of low-rank matrix optimization problems without the RIP condition. We construct an instance of polynomial-time solvable MC problems with exponentially many spurious local minima, leading to the failure of most gradient-based methods. Based on those results, we define a new complexity metric that potentially measures the solvability of low-rank matrix optimization problems based on

the B-M factorization approach. In addition, we show that more measurements of the ground truth matrix can deteriorate the landscape, which further reveals the unfavorable behavior of the B-M factorization on general low-rank matrix optimization problems.

- **Chapter 3**

  Many fundamental low-rank optimization problems, such as matrix completion, phase synchronization/retrieval, power system state estimation, and robust PCA, can be formulated as matrix sensing problems. Two main approaches for solving matrix sensing are based on semidefinite programming (SDP) and Burer-Monteiro (B-M) factorization. The SDP method suffers from high computational and space complexities, whereas the B-M method may return a spurious solution due to the non-convexity of the problem. The existing theoretical guarantees for the success of these methods have led to similar conservative conditions, which may wrongly imply that these methods have comparable performances. In this chapter, we shed light on the major differences between these two methods. First, we present a class of *structured* matrix completion problems for which the B-M methods fail with an overwhelming probability while the SDP method works correctly. Second, we identify a class of *highly sparse* matrix completion problems for which the B-M method works and the SDP method fails. Third, we prove that although the B-M method exhibits the same performance independent of the rank of the unknown solution, the success of the SDP method is correlated to the rank of the solution and improves as the rank increases.

## System Identification under Adversarial Disturbances

- **Chapter 4**

  This chapter investigates the system identification problem for linear discrete-time systems under adversaries and analyzes two lasso-type estimators. We examine the 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 the 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. This chapter provides the guarantee of learning from correlated data for dynamical systems when there is less clean data than corrupt data.

- **Chapter 5**

  We study the problem of learning a non-linear dynamical system by parameterizing its dynamics using basis functions. We assume that disturbances occur at each time step with an

arbitrary probability $p$, which models the sparsity level of the disturbance vectors over time. These disturbances are drawn from an arbitrary, unknown probability distribution, which may depend on past disturbances, provided that it satisfies a zero-mean assumption. The primary objective of this chapter is to learn the system's dynamics within a finite time and analyze the sample complexity as a function of $p$. To achieve this, we examine a LASSO-type non-smooth estimator and establish necessary and sufficient conditions for its well-specifiedness and the uniqueness of the global solution to the underlying optimization problem. We then provide exact recovery guarantees for the estimator under two distinct conditions: boundedness and Lipschitz continuity of the basis functions. We show that finite-time exact recovery is achieved with high probability, even when $p$ approaches 1. Unlike prior literature, which primarily focuses on independent and identically distributed (i.i.d.) disturbances and provides only asymptotic guarantees for system learning, this chapter presents the first finite-time analysis of non-linear dynamical systems under a highly general disturbance model. Our framework allows for possible temporal correlations in the disturbances and accommodates semi-oblivious adversarial attacks, significantly broadening the scope of existing theoretical results.

- **Chapter 6** This chapter investigates a subgradient-based algorithm to solve the system identification problem for linear time-invariant systems with non-smooth objectives. This is essential for robust system identification in safety-critical applications. While existing literature provides theoretical exact recovery guarantees using optimization solvers, the fast learning algorithms with convergence guarantees for practical use remain unexplored. We analyze the subgradient method in this setting and establish linear convergence results for both the best and Polyak step sizes after a burn-in period. Additionally, we characterize the asymptotic convergence of the best average sub-optimality gap under diminishing and constant step sizes. Finally, we compare the time complexity of standard solvers with the subgradient algorithm and support our findings with experimental results.

## 1.3 Related Publications

- **Chapter 2**

  *Main Paper*

  - Baturalp Yalçın, Haixiang Zhang, Javad Lavaei, and Somayeh Sojoudi. "Factorization Approach for Low-Complexity Matrix Completion Problems: Exponential Number of Spurious Solutions and Failure of Gradient Methods". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 319–341

  *Related Paper*

  - Haixiang Zhang, Baturalp Yalcin, Javad Lavaei, and Somayeh Sojoudi. "A New Complexity Metric for Nonconvex Rank-One Generalized Matrix Completion". In: *Mathematical Programming* 207.1 (2024), pp. 227–268

- **Chapter 3**

  *Main Paper*

  - Baturalp Yalçın, Ziye Ma, Javad Lavaei, and Somayeh Sojoudi. "Semidefinite Programming versus Burer-Monteiro Factorization for Matrix Sensing". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 9. 2023, pp. 10702–10710

- **Chapter 4**

  *Main Paper*

  - Baturalp Yalcin, Haixiang Zhang, Javad Lavaei, and Murat Arcak. "Exact Recovery for System Identification with More Corrupt Data than Clean Data". In: *IEEE Open Journal of Control Systems* (2024)

  *Related Paper*

  - Han Feng, Baturalp Yalcin, and Javad Lavaei. "Learning of Dynamical Systems under Adversarial Attacks-Null Space Property Perspective". In: *2023 American Control Conference (ACC)*. IEEE. 2023, pp. 4179–4184

- **Chapter 5**

  *Main Paper*

  - Haixiang Zhang, Baturalp Yalcin, Javad Lavaei, and Eduardo D. Sontag. "Exact Recovery Guarantees for Parameterized Non-linear System Identification Problem under Adversarial Attacks". 2024. arXiv: `2409.00276`

- **Chapter 6**

  *Main Paper*

  - Baturalp Yalcin and Javad Lavaei. "Subgradient Method for System Identification with Non-Smooth Objectives". 2025. arXiv: `2503.16673`

## 1.4 Notations

**Scalars, vectors, matrices, and sets:** We use lower case bold letters $\mathbf{x}$ to represent vectors and capital bold letters $\mathbf{X}$ to represent matrices. For a vector $\mathbf{x}$, $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, and $\|\mathbf{x}\|_\infty$ denote its $\ell_1$, $\ell_2$, and $\ell_\infty$ norms, respectively. Moreover, $\mathbb{S}^{n-1}$ is the unit ball $\{\mathbf{x} \in \mathbb{R}^n | \|\mathbf{x}\|_2 = 1\}$. For every vector $\mathbf{x}$, $[\mathbf{x}]_i$ denotes the $i$-th entry and $[\mathbf{x}]_{i:j}$ denotes the subvector of entries from index $i$ to index $j$ for $i < j$. Similarly, for every matrix $\mathbf{X}$, $[\mathbf{X}]_{i:j,k:l}$ denotes the submatrix with rows between $i$ and $j$ and columns between $k$ and $l$ with $i < j$ and $k < l$. For a vector $\mathbf{x}$ and a matrix $\mathbf{X}$, $\mathbf{x}^\top$ and $\mathbf{X}^\top$ represent their transpose, respectively. For a matrix $\mathbf{X}$, $\|\mathbf{X}\|_F$ and $\|\mathbf{X}\|$ denote the Frobenius norm and operator norm of a matrix. For two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, we use $\langle \mathbf{X}_1, \mathbf{X}_2 \rangle = \mathrm{Tr}({\mathbf{X}_1}^\top \mathbf{X}_2)$ and $\mathbf{X}_1 \otimes \mathbf{X}_2$ to denote the inner product and Kronecker product, respectively. The notations $\mathbf{X} \succeq 0$ and $\mathbf{X} \succ 0$ mean that the matrix $\mathbf{X}$ is positive semidefinite and positive definite, respectively.

The set of $n \times n$ positive semidefinite matrices is denoted as $\mathbb{S}_+^n$. For a matrix $\mathbf{X}$, $\mathrm{vec}(\mathbf{X})$ is the usual vectorization operation by stacking the columns of the matrix $\mathbf{X}$ into a vector. For a vector $\mathbf{x} \in \mathbb{R}^{n^2}$, $\mathrm{mat}(\mathbf{x})$ converts $\mathbf{x}$ to a square matrix and $\mathrm{mat}_S(\mathbf{x})$ converts $\mathbf{x}$ to a symmetric matrix, i.e., $\mathrm{mat}(\mathbf{x}) = \mathbf{x}$ and $\mathrm{mat}_S(\mathbf{x}) = (\mathbf{X} + \mathbf{X}^\top)/2$, where $\mathbf{X} \in \mathbb{R}^{n \times n}$ is the unique matrix satisfying $\mathbf{x} = \mathrm{vec}(\mathbf{X})$. The set $[n]$ represents the set of integers from 1 to $n$. $|S|$ shows the cardinality of a given set $S$. $\mathbf{I}_n$ refers to the identity matrix of size $n \times n$ and $\mathbf{0}_{n \times n}$ refers to the $n \times n$ dimensional matrix with zero entries. We use $\mathbf{0}_n$ to denote the $n$-dimensional vector with all entries 0.

**Functions:** Given two functions $f$ and $g$, the notation $f(\mathbf{x}) = \Theta[g(\mathbf{x})]$ means that there exist universal positive constants $c_1$ and $c_2$ such that $c_1 g(\mathbf{x}) \leq f(\mathbf{x}) \leq c_2 g(\mathbf{x})$. The relation $f(\mathbf{x}) \lesssim g(\mathbf{x})$ holds if there exists a universal positive constant $c_3$ such that $f(\mathbf{x}) \leq c_3 g(\mathbf{x})$ holds with high probability. Equivalently, we use $f(\mathbf{x}) \leq \mathcal{O}(g(\mathbf{x}))$ as well. The relation $f(\mathbf{x}) \gtrsim g(\mathbf{x})$ holds if $g(\mathbf{x}) \lesssim f(\mathbf{x})$. Given the function $f$, $\partial f$ denotes the subdifferential of the function. For a function $f : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$, we denote the gradient and the Hessian as $\nabla f(\cdot)$ and $\nabla^2 f(\cdot)$, respectively. The Hessian is a four-dimensional tensor with $[\nabla^2 f(\mathbf{X})]_{i,j,k,l} = \frac{\partial^2 f(\mathbf{X})}{\partial \mathbf{X}_{ij} \partial \mathbf{X}_{k,l}}$ for all $i, j \in [m]$ and $k, l \in [n]$. The quadratic form of the Hessian in the direction $\Delta \in \mathbb{R}^{m \times n}$ is defined as $\Delta : \nabla^2 f(\mathbf{X}) : \Delta = \sum_{i,j,k,l} [\nabla^2 f(\mathbf{X})]_{i,j,k,l} \Delta_{ij} \Delta_{kl}$. We use $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ to denote the ceiling and floor operators, respectively.

**Random Events:** $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ denote the probability of an event and the expectation of a random variable. A Gaussian random vector $X$ with mean $\mu$ and covariance matrix $\Sigma$ is written as $X \sim \mathcal{N}(\mu, \Sigma)$.

# Part I

# Low-Rank Matrix Optimization

# Chapter 2

# Factorization Approach for Matrix Completion Problems

The low-rank matrix optimization problem aims to recover a low-rank ground truth matrix $\mathbf{M}^*$ through some measurements modeled as $\mathcal{A}(\mathbf{M}^*)$, where the measurement operator $\mathcal{A}$ is a function from $\mathbb{R}^{n\times n}$ to $\mathbb{R}^d$. The operator $\mathcal{A}$ can be either linear, as in the linear matrix sensing problem and the matrix completion problem [23, 91], or non-linear, as in the one-bit matrix sensing problem [31] and the phase retrieval problem [99]. The problem has two variants, known as symmetric and asymmetric problems. The first one assumes that $\mathbf{M}^*$ is a positive semidefinite (PSD) matrix, whereas the second one makes no such assumption and allows $\mathbf{M}^*$ to be non-symmetric or sign indefinite. Since the asymmetric problem can be equivalently transformed into a symmetric problem [125], we focus on the latter.

There are, in general, two different approaches to overcome the non-convex low-rank constraint. The first approach is to design a convex penalty function that prefers low-rank matrices and then optimize the penalty function under the measurement constraint [23, 91, 24]. However, this approach works in the matrix space $\mathbb{R}^{n\times n}$ and has a high computational complexity. The other widely accepted technique is the Burer-Monteiro (B-M) factorization approach [17], which converts the original problem into an unconstrained one by replacing the original PSD matrix variable $\mathbf{M} \in \mathbb{R}^{n\times n}$ with the product of a low-dimensional variable $\mathbf{X} \in \mathbb{R}^{n\times r}$ and its transpose. The optimization problem based on the B-M factorization approach can be written as

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times r}} \; g\left[\mathcal{A}(\mathbf{X}\mathbf{X}^\top) - \mathcal{A}(\mathbf{M}^*)\right],$$

where $g(\cdot)$ is a loss function that penalizes the mismatch between $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{M}^*$. The objective function is generally non-convex using the B-M factorization, even if the loss function $g(\cdot)$ is convex. Nonetheless, it has been proved that under certain strong conditions, such as the Restricted Isometry Property (RIP) condition, saddle-escaping methods can converge to the ground truth solution with a random initialization [125, 14] and first-order methods with spectral initialization converge locally [108, 11]; see [29] for an overview.

Then, it is natural to ask whether optimization methods based on the B-M factorization approach can succeed on general low-rank matrix optimization problems with low information-

theoretic complexity (i.e., problems that have a unique global solution and can be solved in polynomial time), especially when the RIP condition does not hold. This chapter focuses on a common class of problems without the RIP condition, namely the matrix completion (MC) problem. For the MC problem, the measurement operator $\mathcal{A}_\Omega : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ is given by

$$\mathcal{A}_\Omega(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise,} \end{cases}$$

where $\Omega$ is the set of indices of observed entries. We denote the measurement operator as $\mathbf{M}_\Omega := \mathcal{A}_\Omega(\mathbf{M})$ for simplicity. An instance of the MC problem, denoted as $\mathcal{P}_{\mathbf{M}^*,\Omega,n,r}$, can be formulated as

$$\begin{aligned} \text{find} \quad & \mathbf{M} \in \mathbb{R}^{n\times n} \\ \text{s.t.} \quad & \operatorname{rank}(\mathbf{M}) \le r, \quad \mathbf{M} \succeq 0, \quad \mathbf{M}_\Omega = \mathbf{M}^*_\Omega. \end{aligned} \tag{$\mathcal{P}_{\mathbf{M}^*,\Omega,n,r}$}$$

If $\mathbf{M}^*$ is the only solution to this problem, we will say that $\mathcal{P}_{\mathbf{M}^*,\Omega,n,r}$ has a unique solution. Using the B-M factorization approach, the MC problem can be solved via the optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times r}} f(\mathbf{X}), \tag{2.1}$$

where $f(\mathbf{X}) := g[(\mathbf{X}\mathbf{X}^\top - \mathbf{M}^*)_\Omega]$. For example, if the $\ell_2$-loss function is used, the problem becomes

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times r}} \left\|(\mathbf{X}\mathbf{X}^\top - \mathbf{M}^*)_\Omega\right\|_F^2. \tag{2.2}$$

**Contributions.** We provide a negative answer to the preceding question by constructing MC problem instances for which the *optimization complexity* of local search methods using the B-M factorization does not align with the *information-theoretic complexity* of the underlying MC problem instance. The information-theoretic complexity refers to the minimum number of operations that the best possible algorithm takes to find the ground truth matrix. In contrast, the optimization complexity refers to the minimum number of operations that a given optimization method takes to find the ground truth matrix. In general, the optimization complexity of local search methods depends on the properties of spurious solutions to the optimization problem, e.g., the number, the sharpness, and the regions of attraction of spurious solutions. The optimization complexity predicts the performance of an algorithm and provides a hint on which algorithm to use for a given problem. Therefore, the results in this chapter imply that the popular B-M factorization approach cannot capture the benign properties of the low-rank problem when the RIP condition does not hold. We summarize our contributions as follows:

**i)** Given natural numbers $n$ and $r$ with $n \ge 2r$, we construct a class of MC problem instances $\mathcal{L}(\mathcal{G}, n, r)$, whose ground matrix $\mathbf{M}^* \in \mathbb{S}^n_+$ has rank $r$. A unique global solution exists for every instance in this class, and the solution can be found in polynomial time via graph-theoretic algorithms.

**ii)** Next, we show the existence of an instance in $\mathcal{L}(\mathcal{G}, n, r)$ whose B-M factorization formulation (2.2) has at least $\mathcal{O}(2^{n-2r})$ equivalent classes of spurious local solutions. A solution is called spurious if it is a local minimum but has a larger objective value than the optimal objective value. Note that this claim holds for general loss functions under a weak assumption.

**iii)** Moreover, for the rank-1 case, we prove that most gradient-based methods with a random initialization converge to a spurious local minimum with probability at least $1 - \mathcal{O}(2^{-n/2})$. Numerical studies verify that the failure of the gradient-based methods also happens for general rank cases.

**iv)** We present an instance with no spurious solution under the B-M factorization formulation (2.2), but introducing additional observations of the ground truth matrix leads to at least exponentially many spurious solutions. This example further reveals the unfavorable behavior of the B-M factorization approach on general low-rank matrix optimization problems.

Based on these results, we define a new complexity metric that potentially captures the optimization complexity of optimization methods based on the B-M factorization.

**Related Work.** The low-rank optimization problem has been well studied under the RIP condition [91]. Several recent works [129, 13, 125] showed that the non-convex formulation has no spurious local minima with a small RIP parameter. To understand how conservative the RIP condition is, we consider a class of polynomial-time solvable problems without the RIP condition and study the behavior of optimization methods in this class. Specifically, we consider the polynomial-time solvable MC problems. Most existing literature on the MC problem is based on the assumption that the measurement set is randomly constructed and the global solution is coherent [23, 24, 48, 47, 74, 27]. In comparison, there is a small range of works that have focused on the deterministic MC problem [10, 63, 88, 70]. Furthermore, efficient graph-theoretic algorithms utilizing the special structures of a deterministic measurement set can be designed [75]. Existing works on a deterministic measurement set case have focused on the completability problem and the convex relaxation approach, while the B-M factorization approach has not been analyzed. Moreover, several existing works also provided negative results on the low-rank matrix optimization problem. The counterexamples in [24, 10] have non-unique global solutions, which make the recovery of the ground truth matrix impossible. The counterexamples in [112] have a unique global solution, but the objective function must be linear. We refer the reader to [29] for a review of the low-rank matrix optimization problem. This chapter is the first one in the literature to study optimization complexity in cases when information-theoretic complexity is low.

## 2.1 Exponential Number of Spurious Local Minima

This section shows that MC problem instances with a low information-theoretic complexity may have exponentially many spurious local minima if the B-M factorization is employed. We first construct a class of such instances and then identify the problematic instances.

## Low-complexity Class of MC Problems

Suppose that $r \geq 1$ and $n \geq 2r$ are two given integers. We construct a class of MC problem instances whose ground truth matrix $\mathbf{M}^* \in \mathbb{S}_+^n$ is rank-$r$. For every instance in this class, the global solution is unique and can be found in polynomial time in terms of $n$ and $r$. Let $m := \lfloor n/r \rfloor \geq 2$. We divide the first $mr$ rows and the first $mr$ columns of the matrix $\mathbf{M}^*$ into $m \times m$ block matrices, where each block has dimension $r \times r$. For every $i, j \in [m]$, we denote the block matrix at position $(i, j)$ as $\mathbf{M}_{i,j}^*$. We now define the block measurement patterns induced by a given graph.

**Definition 1** (Induced measurement set)**.** *Let $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2) = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2)$ be a pair of undirected graphs with the node set $\mathcal{V} = [m]$ and the disjoint edge sets $\mathcal{E}_1, \mathcal{E}_2 \subset [m] \times [m]$, respectively. The induced measurement set $\Omega(\mathcal{G})$ is defined as follows: if $(i, j) \in \mathcal{E}_1$, then the entire block $\mathbf{M}_{i,j}^*$ are observed; if $(i, j) \in \mathcal{E}_2$, then all non-diagonal entries of the block $\mathbf{M}_{i,j}^*$ are observed; otherwise, none of the entries of the block is observed. In addition, the last $n - mr$ rows and the last $n - mr$ columns of the matrix $\mathbf{M}^*$ are fully observed. We refer to the graph $\mathcal{G}$ as the block sparsity graph.*

The following definition introduces a low-complexity class of MC problem instances.

**Definition 2** (Low-complexity class of MC problems)**.** *Define $\mathcal{L}(\mathcal{G}, n, r)$ to be the class of low-complexity MC problems $\mathcal{P}_{\mathbf{M}^*,\Omega,n,r}$ with the following properties:*

- ***i)*** *The ground truth $\mathbf{M}^* \in \mathbb{S}_+^n$ is rank-r.*
- ***ii)*** *The matrix $\mathbf{M}_{i,j}^* \in \mathbb{R}^{r \times r}$ is rank-$r$ for all $i, j \in [m]$.*
- ***iii)*** *The measurement set $\Omega = \Omega(\mathcal{G})$ is induced by $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2)$, where $\mathcal{G}_1$ is connected and non-bipartite.*

The following proposition states that every MC problem instance in $\mathcal{L}(\mathcal{G}, n, r)$ is polynomial-time solvable.

**Proposition 1.** *For an arbitrary instance $\mathcal{P}_{\mathbf{M}^*,\Omega,n,r}$ in $\mathcal{L}(\mathcal{G}, n, r)$, the ground truth $\mathbf{M}^*$ is the unique solution of this problem and can be found in $\mathcal{O}(n^2/r^2 + nr^2)$ time.*

## Intuition for Rank-1 Case with $\ell_2$-loss Function

We start with the case when the rank $r$ equals 1 and the loss function $g(\cdot)$ is the $\ell_2$-loss. We study two instances in the class $\mathcal{L}(\mathcal{G}, n, 1)$ with $\mathcal{O}(n)$ and $\mathcal{O}(n^2)$ observations, respectively. The B-M formulation (2.2) of both instances contains exponentially many spurious local minima. Since the decomposition variable $\mathbf{X}$ is a column vector in the rank-1 case, we write it as $\mathbf{x}$.

**Example 1.** *We first provide an instance with $\mathcal{O}(n)$ observations. Note that the number of blocks, namely $m$, equals $n$ in the rank-1 case. Let the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2)$ be chosen as $\mathcal{V} := [n]$ and*

$$\mathcal{E}_1 := \{(i, j) \mid i, j \in [n],\ |i - j| \leq 1\}, \quad \mathcal{E}_2 := \emptyset.$$

*The measurement set is the induced set $\Omega := \Omega(\mathcal{G})$. Namely, we observe the ground truth matrix's diagonal, sub-diagonal, and super-diagonal entries. One can verify that the subgraph $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}_1)$*

*is connected and non-bipartite. Now, we construct a specific ground truth matrix. We define the vector* $\mathbf{x}^* \in \mathbb{R}^n$ *by*

$$\mathbf{x}^*_{2k+1} := 1,\ \forall k \in [\lceil n/2 \rceil], \quad \mathbf{x}^*_{2k} := 0,\ \forall k \in [\lfloor n/2 \rfloor],$$

*and let* $\mathbf{M}^* := \mathbf{x}^*(\mathbf{x}^*)^\top$. *For the B-M factorization formulation* (2.2)*, the set of global minima is given by*

$$\mathcal{X}^* := \big\{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^2_{2k+1} = 1, \quad \forall k \in [\lceil n/2 \rceil], \quad \mathbf{x}_{2k} = 0, \quad \forall k \in [\lfloor n/2 \rfloor] \big\},$$

*which has cardinality* $2^{\lceil \frac{n}{2} \rceil}$. *For every global solution* $\hat{\mathbf{x}} \in \mathcal{X}^*$ *and every* $\Delta \in \mathbb{R}^n \backslash \{0\}$*, the Hessian satisfies*

$$\Delta : \nabla^2 f(\hat{\mathbf{x}}) : \Delta = 2 \left\| \left( \hat{\mathbf{x}} \Delta^\top + \Delta \hat{\mathbf{x}}^\top \right)_\Omega \right\|_F^2 = 8\|\Delta\|^2 - \mathbb{1}[n \text{ is even}] 4\Delta_n^2 > 0,$$

*where* $\mathbb{1}[\cdot]$ *is the indicator function. Therefore, the Hessian is positive definite at every global minimum. Then, we perturb the ground truth solution* $\mathbf{M}^*$ *to*

$$\mathbf{M}^*(\epsilon) := \mathbf{x}^*(\epsilon) \left[ \mathbf{x}^*(\epsilon) \right]^\top = (\mathbf{x}^* + \epsilon)(\mathbf{x}^* + \epsilon)^\top,$$

*where* $\mathbf{x}^*(\epsilon) := \mathbf{x}^* + \epsilon$ *and* $\epsilon \in \mathbb{R}^n$ *is a small perturbation. We denote the associated problem* (2.2) *as*

$$\min_{\mathbf{x} \in \mathbb{R}^n} \tilde{f}(\mathbf{x}; \epsilon), \tag{2.3}$$

*where* $\tilde{f}(\mathbf{x}; \epsilon) := \|(\mathbf{x}\mathbf{x}^\top - \mathbf{M}(\epsilon)^*)_\Omega\|_F^2$. *For a generic perturbation* $\epsilon$*, all components of* $\epsilon$ *are non-zero and problem* $\mathcal{P}_{\mathbf{M}^*(\epsilon),\Omega,n,1}$ *belongs to the class* $\mathcal{L}(\mathcal{G}, n, 1)$*. This implies that the global solution of problem* (2.3) *is unique up to a sign flip.*

*We analyze the relation between the local minima of the original problem and those of the perturbed problem. Consider the equation* $\nabla_{\mathbf{x}} \tilde{f}(\mathbf{x}; \epsilon) = 0$ *near an unperturbed global minimum* $\hat{\mathbf{x}} \in \mathcal{X}^*$*. Since* $(\hat{\mathbf{x}}; 0)$ *is a solution to the gradient equation and the Jacobian matrix with respect to* $\mathbf{x}$ *is equal to the positive definite Hessian* $\nabla^2 f(\hat{\mathbf{x}})$*, the Implicit Function Theorem (IFT) states that there exists a unique solution* $\hat{\mathbf{x}}(\epsilon)$ *in a neighborhood of* $\hat{\mathbf{x}}$ *for all values of* $\epsilon$ *with a small norm. In addition, the continuity of Hessian implies that* $\nabla_{\mathbf{x}\mathbf{x}} \tilde{f}(\hat{\mathbf{x}}^*(\epsilon); \epsilon) \succ 0$*.*

*Thus,* $\hat{\mathbf{x}}(\epsilon)$ *is a local minimum of the perturbed problem* (2.3)*. As a result, we have proved the existence of a local minimum for the perturbed problem corresponding to each of the* $2^{\lceil n/2 \rceil}$ *global minima of the unperturbed problem. Hence, the problem* (2.3) *has at least* $2^{\lceil n/2 \rceil}$ *local minima, while only two of them are global minima. In summary, we have constructed an instance in* $\mathcal{L}(\mathcal{G}, n, 1)$ *that has exponentially many spurious local solutions.*

**Example 2.** *Next, we construct an MC problem instance with exponentially many spurious local minima and* $\mathcal{O}(n^2)$ *observations. We choose the same ground truth matrix* $\mathbf{M}^*(\epsilon)$ *as in the last example, but assume that the measurement set* $\Omega$ *is induced by the graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2)$ *with* $\mathcal{V} := [n]$*,* $\mathcal{E}_2 = \emptyset$ *and*

$$\mathcal{E}_1 := \{(i,i), (i,2k), (2k,i) \mid \forall i \in [n],\ k \in [\lfloor n/2 \rfloor]\}.$$

*Since the subgraph $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}_1)$ is connected and non-bipartite, the problem $\mathcal{P}_{\mathbf{M}^*(\epsilon),\Omega,n,1}$ belongs to the class $\mathcal{L}(\mathcal{G}, n, 1)$. Moreover, one can verify that the set of global minima of this problem is still $\mathcal{X}^*$, and the Hessian at every global solution is positive definite. By the same argument, IFT implies that problem* (2.3) *has at least $2^{\lceil n/2 \rceil} - 2$ spurious local minima for a generic and small perturbation $\epsilon$.*

Note that the instances analyzed in this section, as well as those in the remainder of this chapter, satisfy the incoherence condition [23] with the parameter $\mu = \mathcal{O}(1)$. The results in this subsection will be formalized next with a unified framework.

## Rank-1 Case with General Measurement Sets

In this subsection, we estimate the largest lower bound on the number of spurious local minima for the given parameters $n$ and $r$. We address the problem by first finding a lower bound on the number of spurious local minima given a general measurement set $\Omega$ and then maximizing the lower bound over $\Omega$. The following theorem utilizes the topology of $\mathcal{G}$ to quantify a lower bound on the number of spurious solutions for the measurement set $\Omega(\mathcal{G})$. It uses the independent sets for the graph $\mathcal{G}$. For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the set $\mathcal{S} \subset \mathcal{V}$ is called an independent set if no two nodes in $\mathcal{S}$ are adjacent. The set $\mathcal{S}$ is called a maximal independent set if it is an independent set and is not a strict subset of another independent set.

**Theorem 1.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \emptyset)$ such that $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}_1)$ is connected and non-bipartite with $n$ vertices. Assume that there exists a maximal independent set $\mathcal{S}(\mathcal{G}_1)$ of $\mathcal{G}_1$ such that every vertex in the set has a self-loop. There exists an instance in $\mathcal{L}(\mathcal{G}, n, 1)$ for which the problem* (2.2) *has at least $2^{|\mathcal{S}(\mathcal{G}_1)|} - 2$ spurious local minima.*

In both Examples 1 and 2, a maximal independent set is $\mathcal{S} = \{2k+1 \mid \forall k \in [\lceil n/2 \rceil]\}$ . Hence, Theorem 1 implies that there are $2^{\lceil n/2 \rceil} - 2$ spurious local minima, which is consistent with our analysis. Since a maximal independent set of a connected and non-bipartite graph can have up to $n-1$ vertices, the number of spurious local minima can be as large as $2^{n-1} - 2$.

**Corollary 1.** *There exist a graph $\mathcal{G}$ and an instance in $\mathcal{L}(\mathcal{G}, n, 1)$ such that problem* (2.2) *has $2^{n-1} - 2$ spurious solutions. In addition, there exist a graph $\mathcal{G}$ and an instance in $\mathcal{L}(\mathcal{G}, n, 1)$ with $|\Omega| = n^2 - 2$ such that the problem* (2.2) *has spurious solutions.*

Corollary 1 implies that the B-M factorization may not be an efficient approach to the MC problem since it has a spurious solution even in the highly ideal case when almost all matrix entries are measured. Generally, the proof of Theorem 1 implies that, as a necessary condition for not having a spurious solution in formulation (2.2), the elements of $\mathbf{x}^*$ associated with the nodes outside of the maximal independent set $\mathcal{S}$ should not be much smaller than those associated with the nodes in $\mathcal{S}$.

**Corollary 2.** *Under the setting of Theorem 1, there exists a function $h_{\mathcal{S}}(\cdot) : (0,\infty) \mapsto (0,\infty)$ such that $\mathcal{P}_{\mathbf{x}^*(\mathbf{x}^*)^\top,\Omega(\mathcal{G}),n,1}$ has at least $2^{|\mathcal{S}|}-2$ spurious local minima in formulation* (2.2) *for every generic $\mathbf{x}^* \in \mathbb{R}^n$ satisfying*

$$\|\mathbf{x}^*_{\mathcal{S}^c}\| \leq h_{\mathcal{S}}(\min_{i\in\mathcal{S}}|x^*_i|) \cdot \|\mathbf{x}^*_{\mathcal{S}}\|,$$

*where $\mathcal{S}^c := [n]\backslash\mathcal{S}$ and $\mathbf{x}_{\mathcal{S}^c} := (x_i : i \notin \mathcal{S})$.*

Because the maximal independent set of a graph $\mathcal{G}$ is not necessarily unique, the set of functions $h_{\mathcal{S},M}(\cdot)$ over all maximal independent sets $\mathcal{S}$ designates a necessary condition for the non-existence of spurious local minima given a measurement set $\Omega(\mathcal{G})$.

## Extension to General Rank-$r$ Case

We generalize the results to the case when the ground truth matrix has a general rank. [37] showed that the rank-$r$ MC problem is $\mathcal{NP}$-hard in the worst case for every $r \geq 2$. However, we focus on instances in the low-complexity class $\mathcal{L}(\mathcal{G}, n, r)$ and show that there are instances in this class whose B-M factorization formulation (2.2) has a highly undesirable landscape. We cannot simply extend the proof of the rank-1 case to the rank-$r$ case since there exists an infinite number of matrices $\mathbf{X}^*$ such that $\mathbf{M}^* = \mathbf{X}^*(\mathbf{X}^*)^\top$ when $r > 1$. The global optimality of a solution $\mathbf{X}^*$ is not lost under any orthogonal transformation. This implies that the Hessian at the global solutions of problem (2.2) cannot be positive definite, which fails the applicability of IFT. Therefore, we consider the quotient manifold $\mathbb{R}^{n\times r}/O_r$, where $O_r$ is the lie group of $r \times r$ orthogonal matrices. To simplify the analysis, we instead consider the following lower-diagonal subspace:

$$\mathcal{W}^{n\times r} := \{\mathbf{X} \in \mathbb{R}^{n\times r} \mid \mathbf{X}_{ij} = 0,\ \forall i \in [n],\ j \in [r] \quad \text{s.t.} \quad i < j\}.$$

We define an embedding of the manifold $\mathbb{R}^{n\times r}/O_r$ into $\mathcal{W}^{n\times r}$ and composite it with the quotient map.

**Definition 3** (Restriction map)**.** *Given a matrix $\mathbf{X} \in \mathbb{R}^{n\times r}$, we define the embedding $\phi^{emb}([\mathbf{X}]) := \mathbf{R} \in \mathcal{W}^{n\times r}$, where $\mathbf{X} = \mathbf{RQ}$ is the RQ decomposition with $\mathbf{Q}$ being an orthogonal matrix and $\mathbf{R}$ having non-negative diagonal elements. The restriction map is defined as $\phi(\mathbf{X}) := \phi^{emb}([\mathbf{X}])$.*

When the RQ decomposition is not unique, we choose an arbitrary decomposition for the embedding $\phi^{emb}(\cdot)$. However, the properties of the RQ decomposition ensure that $\phi^{emb}(\cdot)$ is a bijection in a small neighborhood of each matrix $\mathbf{X}$ whose first $r$ rows are linearly independent. Consider the restricted version of problem (2.2):

$$\min_{\mathbf{X}\in\mathcal{W}^{n\times r}} f(\mathbf{X}), \tag{2.4}$$

Results in Section 2.1 can be extended to the problem (2.4) and then be translated back to the problem (2.2).

**Lemma 1.** *Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2)$ with $m = \lfloor n/r \rfloor$ vertices for which the subgraph $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}_1)$ is connected and non-bipartite. Assume that there exists a maximal independent set $\mathcal{S}(\mathcal{G}_1)$ of $\mathcal{G}_1$ whose vertices each have a self-loop. If the induced subgraph (see [52] for the definition.) $\mathcal{G}_2[\mathcal{S}]$ is connected, then there exists an instance in $\mathcal{L}(\mathcal{G}, n, r)$ for which the problem* (2.4) *has at least $2^{r|\mathcal{S}(\mathcal{G}_1)|} - 2^r$ spurious local minima. In addition, the first $r$ rows of each local minimum are linearly independent.*

The Hessian at each local minimum of the unperturbed problem is positive definite along the tangent space of $\mathcal{W}^{n\times r}$, where the off-diagonal observations of $\Omega(\mathcal{G})$ play a key role. If the first $r$ rows of a local minimum $\hat{\mathbf{X}}$ are linearly independent, the diagonal elements of $\hat{\mathbf{X}}$ are non-zero. Therefore, by flipping the signs of columns, we can find an equivalent local minimum $\tilde{\mathbf{X}}$ with positive diagonal elements, i.e., $\tilde{\mathbf{X}}$ lies in the range of $\phi^{emb}(\cdot)$. By symmetry, the total number of such local minima is $2^{r(|\mathcal{S}(\mathcal{G}_1)|-1)} - 1$. Since the restriction map $\phi^{emb}(\cdot)$ is a bijection in a neighborhood of $\tilde{\mathbf{X}}$, the equivalent class $[\tilde{\mathbf{X}}] \in \mathbb{R}^{n\times r}/O_r$ is a local minimum of problem (2.2) on the quotient manifold and thus $\tilde{\mathbf{X}}$ is a local minimum of problem (2.2). The above argument leads to Theorem 2.

**Theorem 2.** *Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2)$ satisfying the conditions of Lemma 1. There exists an instance in $\mathcal{L}(\mathcal{G}, n, r)$ for which problem* (2.2) *has at least $2^{r(|\mathcal{S}(\mathcal{G}_1)|-1)} - 1$ equivalent classes of spurious solutions.*

Finally, we estimate the largest lower bound for the number of spurious local minima.

**Corollary 3.** *There exists an instance in $\mathcal{L}(\mathcal{G}, n, r)$ for which the problem* (2.2) *has at least $2^{n-2r} - 1$ equivalent classes of spurious solutions. In addition, there exists an instance in $\mathcal{L}(\mathcal{G}, n, r)$ with $|\Omega| = n^2 - 2r$ for which the problem* (2.2) *has spurious solutions.*

## General Loss Functions

In this part, we generalize the preceding results to the problem (2.1). To extend the constructions to a general loss function $g(\cdot)$, we require a few weak assumptions on the loss function $g(\cdot)$.

**Assumption 1.** *The following conditions hold for the function $g(\cdot)$:*

- ***i)*** *$g(\cdot)$ is twice continuously differentiable;*
- ***ii)*** *the matrix $\mathbf{0}_{n\times r}$ is the unique minimizer of $g(\cdot)$;*
- ***iii)*** *the Hessian of $g(\cdot)$ at $\mathbf{0}_{n\times r}$ is positive definite.*

Now, we can extend the results in Section 2.1 to the general loss function case under the above assumption.

**Theorem 3.** *Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2)$ satisfying the conditions of Lemma 1 and suppose that Assumption 1 holds. There exists an instance in $\mathcal{L}(\mathcal{G}, n, r)$ for which the problem* (2.1) *has at least $2^{r(|\mathcal{S}(\mathcal{G}_1)|-1)} - 1$ equivalent classes of spurious local minima.*

$$\mathbf{M}_{\Omega_k} = \left[\begin{array}{c|c|c} \mathbf{0} & \begin{matrix}\mathbf{M}_{1,k}\\ \vdots\end{matrix} & \mathbf{0} \\ \hline \mathbf{M}_{k,1} \quad \cdots & \mathbf{M}_{k,k} & \cdots \quad \mathbf{M}_{k,m} \\ \hline \mathbf{0} & \begin{matrix}\vdots\\ \mathbf{M}_{m,k}\end{matrix} & \mathbf{0} \end{array}\right]$$

Figure 2.1: Measurement Operator $\mathcal{A}_{\Omega_k}$ on Matrix $\mathbf{M}$.

The proof is similar to those of Theorems 1 and 2. We can prove that the Hessian is positive definite at all global solutions for each loss function $g(\cdot)$ satisfying Assumption 1. We note that the $\ell_2$-loss function $g(\cdot) = \|\cdot\|_F^2$ satisfies the conditions in Assumption 1. As another example, regularizers are ubiquitously used in the low-rank matrix optimization literature [48, 47, 42]. As a corollary to Theorem 3, the regularized version of the problem (2.2) also suffers from the same issue. In this case, the loss function is equal to

$$g(\mathbf{X}) := \left\|(\mathbf{X}\mathbf{X}^\top - \mathbf{M}^*)_\Omega\right\|_F^2 + Q(\mathbf{X}),$$

where $Q(\mathbf{X}) := \lambda \left(\|\mathbf{X}_i\| - \alpha\right)_+^4$ is from [48], $(x)_+ := \max\{x, 0\}$ and $\alpha, \lambda > 0$ are constants. Since the regularizer does not change the landscape around global solutions with a large $\alpha$, Assumption 1 is satisfied and Theorem 3 is applicable to the regularized problem.

## 2.2 More Observations Lead to Spurious Local Minima

In Section 2.1, we showed that the B-M factorization formulation (2.2) has an exponential number of spurious local minima on low-complexity MC problem instances. In this section, we exhibit another unfavorable behavior of the B-M factorization. We identify an MC problem instance in $\mathcal{L}(\mathcal{G}, n, r)$ with some pattern $\Omega$ that has no spurious solution, while adding observations to $\Omega$ leads to spurious solutions. Let $m$ and $r$ be natural numbers with $m \geq 2r$. We define $n := mr$ and let the graph be $\mathcal{G}_k := (\mathcal{V}, \mathcal{E}_k)$ where $k \in [m]$ is an arbitrary index and

$$\mathcal{V} := [m], \quad \mathcal{E}_k := \{(k, j),\ (j, k) \mid \forall j \in [m]\}.$$

In the measurement set $\Omega_k$, we observe the blocks $\mathbf{M}_{i,j}$ for all $(i, j) \in \mathcal{E}_k$; see Figure 2.1. $\Omega_k$ contains only full block observations induced by $\mathcal{G}_k$. Second-order critical points are defined as those points that satisfy the first-order and the second-order necessary optimality conditions. The next proposition states that if the ground truth matrix is generic, every SOCP of the problem (2.2) is a global minimum.

**Proposition 2.** *Given an index $k \in [m]$, let the measurement set $\Omega$ be equal to $\Omega_k$. Assume that the block $\mathbf{M}^*_{i,j}$ of the ground truth matrix $\mathbf{M}^*$ has rank $r$ for all $i, j \in [m]$. Then, every SOCP of problem* (2.2) *is a global minimum.*

Next, we construct a graph $\tilde{\mathcal{G}}_k := (\mathcal{V}, \tilde{\mathcal{E}}_k, \mathcal{E}_2)$, where

$$\tilde{\mathcal{E}}_k := \mathcal{E}_k \cup \{(i,i) \mid \forall i \in [m]\}, \quad \mathcal{E}_2 := \{(i,j) \mid \forall i, j \in [m],\ i \neq j\}.$$

Namely, we have included each block's self-loops and non-diagonal observations in the new graph. A maximal independent set for the subgraph $\tilde{\mathcal{G}}_{k,1} := (\mathcal{V}, \tilde{\mathcal{E}}_k)$ is $\mathcal{S} := [m] \backslash \{k\}$. We define a new measurement set $\tilde{\Omega}_k := \Omega(\tilde{\mathcal{G}}_k)$. Since $\tilde{\mathcal{E}}_k$ is a super set of $\mathcal{E}_k$, the measurement set $\tilde{\Omega}_k$ is larger than $\Omega_k$. We obtain the following result using Theorem 2.

**Corollary 4.** *Every instance of the MC problem with the measurement set $\tilde{\Omega}_k$ and full rank blocks $\mathbf{M}^*_{i,j}$ for all $i, j \in [m]$ belongs to $\mathcal{L}(\mathcal{G}, n, r)$. The formulation* (2.2) *of an instance of the problem has at least $2^{r(m-2)} - 1$ equivalent classes of spurious local minima. In contrast, all spurious solutions disappear when using the smaller set $\Omega_k$.*

Results of Proposition 2 and Corollary 4 conclude that the landscape of the problem (2.2) deteriorates when the number of observations is increased. This phenomenon further reveals the unfavorable behavior of the B-M factorization on low-rank matrix optimization problems, even if the information-theoretic complexity is low.

## 2.3 Measure of Complexity for Factorization Approach

In Section 2.1, we showed that if there is an MC problem with a non-unique completion, a slightly perturbed problem will have exponentially many spurious local minima in the B-M factorization formulation (2.1). Hence, bifurcation behaviors appear around measurement matrices $\mathbf{M}^*_\Omega$ that are associated with multiple global solutions. For a given measurement operator $\mathcal{A}$, a measurement matrix $\mathcal{A}(\mathbf{M})$ that allows multiple global solutions designates an unacceptable region in the space of ground truth solutions. Based on this intuition, we define a metric to capture the extent of the bifurcation behavior. We define the set of measurements that allow non-unique solutions:

$$\mathcal{T}_\mathcal{A} := \Big\{\mathcal{A}(\mathbf{M}) :\ \exists \mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{n\times r} \text{s.t. } \mathbf{X}_1\mathbf{X}_1^\top \neq \mathbf{X}_2\mathbf{X}_2^\top,\ \mathcal{A}(\mathbf{M}) = \mathcal{A}(\mathbf{X_1}\mathbf{X_1}^\top) = \mathcal{A}(\mathbf{X_2}\mathbf{X_2}^\top)\Big\}.$$

Then, we define a complexity metric below.

**Definition 4** (Complexity metric)**.** *The complexity metric for operator $\mathcal{A}$ and ground truth $\mathbf{M}^*$ is defined as*

$$\operatorname{dist}(\mathcal{A}(\mathbf{M}^*), \mathcal{T}_\mathcal{A}) := \min_{\mathcal{A}(\mathbf{M}) \in \mathcal{T}_\mathcal{A}} \|\mathcal{A}(\mathbf{M}^*) - \mathcal{A}(\mathbf{M})\|_F.$$

It is expected that for instances with a large complexity metric, the optimization complexity of algorithms based on the B-M factorization approach will be aligned with the corresponding

information-theoretic complexity. For example, if the RIP condition is satisfied, the set $\mathcal{T}$ is empty and therefore the complexity metric is always $\infty$. Oppositely, the instances studied in Section 2.1 with spurious solutions all have small complexity metrics. Consequently, the complexity metric is a possible measure of the optimization complexity for the MC problem with the B-M factorization: the optimization problem should be more difficult if the complexity metric is lower.

## 2.4 Gradient-Based Methods Fail With High Probability

We show that the exponential number of spurious local minima in preceding instances will make most randomly initialized gradient-based methods fail with a high probability. The existence of spurious local minima does not necessarily imply the failure of gradient-based methods; see [75, 27]. The analysis in this section is based on the gradient flow

$$\dot{\mathbf{X}}(t) = -\nabla_{\mathbf{X}} f(\mathbf{X}(t)), \quad \mathbf{X}(0) = \mathbf{X}_0. \tag{2.5}$$

It is known that the trajectories of gradient-based methods with a small enough step size are close to those of the gradient flow. We can view various gradient-based methods as ordinary differential equation (ODE) solvers applied to the gradient flow (2.5). Then, the convergence of the discrete trajectories when the step size goes to $0$ can be guaranteed by the consistency and stability of the ODE solver. [97] proved that the consistency and stability conditions are satisfied by several commonly used gradient-based methods, such as the gradient descent, the proximal point, and the accelerated gradient descent methods. Although [97] considered minimizing a strongly convex function, the consistency and stability conditions only depend on the ODE solver and the Lipschitz continuity of the underlying gradient flow. We need the following assumption on the loss function to characterize the global landscape.

**Assumption 2.** *The loss function $g(\cdot)$ satisfies the sparse $(\delta, r)$-RIP condition in the $\Omega$-norm for some constant $\delta \in [0, 1)$ and integer $r \geq 1$. Namely, the inequality*

$$(1-\delta)\|\mathbf{N}_\Omega\|_F^2 \leq \mathbf{N} : \nabla^2 g(\mathbf{M}_\Omega) : \mathbf{N} \leq (1+\delta)\|\mathbf{N}_\Omega\|_F^2$$

*holds for all matrices $\mathbf{M}$ and $\mathbf{N}$ with rank at most $2r$.*

The sparse RIP condition is remarkably different from the conservative RIP condition. For example, the $\ell_2$-loss function satisfies the sparse $(0, r)$-RIP condition for every $r$, while the RIP condition does not hold if $\Omega$ is not a complete graph. Under the above assumption, we show that for the unperturbed example constructed in Section 2.1, the gradient flow will converge to each global minimum with equal probability in the rank-1 case. The main difficulty is to show that all saddle points of the objective function $f(\mathbf{X})$ are strict, and therefore, their region of attractions (ROAs) have measure zero [67].

**Lemma 2.** *Suppose that Assumption 2 holds for $r = 1$ and $\delta = \mathcal{O}(1/n)$, where $n \geq 3$ is the size of the ground truth matrix. There exists an MC problem instance such that the following statements hold for the problem* (2.1)*:*

- *there are $2^{\lceil n/2 \rceil}$ equivalent global minima;*
- *if the gradient flow* (2.5) *is initialized with an absolutely continuous radial probability distribution, it converges to each global minimum with the equal probability $2^{-\lceil n/2 \rceil}$,*

*where a probability distribution is called radial if its density function at point $x$ only depends on $\|x\|$ and the absolute continuity is with respect to the Lebesgue measure.*

Examples of absolutely continuous radial probability distributions include zero-mean Gaussian distributions and uniform distributions over a ball centered at the origin. Note that the $\ell_2$-loss function satisfies the assumption of Lemma 2. Next, we show that with a sufficiently small perturbation to the previous instance of the problem, the ROA of each local minimum will not shrink significantly. Therefore, the gradient flow will converge to each global minimum or spurious solution with approximately the same probability.

**Theorem 4.** *Consider an absolutely continuous radial probability distribution under the setting of Lemma 2. There exists an instance in $\mathcal{L}(\mathcal{G}, n, 1)$ for which the problem* (2.1) *satisfies the following properties:*

- *the global minima are unique up to a sign flip;*
- *if the gradient flow* (2.5) *is initialized with the given distribution, it converges to a global minimum associated with the ground truth solution with probability at most $\mathcal{O}(2^{-\lceil n/2 \rceil})$.*

The results of Theorem 4 imply that, in the rank-1 case, most gradient-based methods with a small enough step size and a suitable random initialization will converge to a spurious solution with an overwhelming probability. The proof works for the general rank case if it can be shown that there are no degenerate saddle points for the above-mentioned unperturbed instances of the problem.

**Remark.** *We remark that the trajectories of stochastic gradient descent (SGD) methods cannot be approximated by those of the gradient flow. Hence, our analysis cannot automatically imply the failure of SGD. However, the proof of Lemma 2 can be adopted to conclude that SGD methods with a random initialization will converge to each global minimum with equal probability in the unperturbed case. Since the trajectories of SGD methods will not vary dramatically with a sufficiently small perturbation, they still converge to each solution with approximately the same probability. Therefore, we also expect the SGD methods to fail with high probability.*

## 2.5 Numerical Experiments

Numerical results are presented to support the failure of the gradient descent algorithm. Each MC problem with the B-M factorization formulation (2.2) is solved by the gradient descent algorithm with a constant step size, where the step size is chosen to be small enough to guarantee that the algorithm converges to a stationary point. Regarding the measurement set, the graph $\mathcal{G}_1 := (\mathcal{V}, \mathcal{E}_1)$ is generated randomly by the Erdös–Rényi model $G(m, p)$, where $\mathcal{V} := [m]$ and each edge of the graph is included independently with probability $p$. If $\mathcal{G}_1$ is not connected or a node in the maximal independent set $\mathcal{S}$ does not have a self-loop, the missed edges are added

to satisfy these conditions. In addition, a connected subtree $\mathcal{G}_2 = (\mathcal{S}, \mathcal{E}_2(\mathcal{S}))$ is generated for non-diagonal observations. We define $\mathcal{G} := (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2)$ and subsequently the measurement set $\Omega(\mathcal{G})$. In addition, the unperturbed ground truth matrix $\mathbf{M}^* = \mathbf{X}^*(\mathbf{X}^*)^\top$ is defined as $\mathbf{M}^*_{i,j} := \mathbf{I}_r$ for all $i, j \in \mathcal{S}$, and $\mathbf{M}^*_{i,j} := \mathbf{0}_{r\times r}$ otherwise. Lastly, a Gaussian random perturbation matrix $\epsilon \in \mathbb{R}^{n\times r}$ is generated and normalized, e.g. $\|\epsilon\|_F = 1$. Then, the perturbed ground truth matrix $\mathbf{M}^*(\epsilon) := (\mathbf{X}^* + \gamma\epsilon)(\mathbf{X}^* + \gamma\epsilon)^\top$ is generated, where $\gamma > 0$ is the perturbation size. We evaluate the algorithm's success rate at $100$ equally distributed values of $\gamma \in (0, 0.5)$ with $300$ random initializations of the gradient algorithm for each instance and each $\gamma$.

The top figure in Figure 2.2 illustrates the success rate of the gradient descent algorithm with the rank $r = 1$, dimension $n = 20$, and various maximum independent set sizes $|\mathcal{S}|$. The results conform to Theorem 4 implying that the success rate is less than $1/(2^{|\mathcal{S}|-1})$ when the perturbation size $\gamma$ is small. The bottom figure in Figure 2.2 makes similar observations with different ranks and maximum independent set sizes $|\mathcal{S}|$ when $m$ equals $10$. These observations imply that Theorem 4 can be extended to the general rank case, since the success rate is less than $1/(2^{r(|\mathcal{S}|-1)})$ when $\gamma$ is small. We note that the algorithm's behavior may change when $\gamma$ is large. Specifically, significant improvements in success rate are observed when $\gamma > 0.2$ for most problem instances. This is in accordance with our notion of complexity metric.

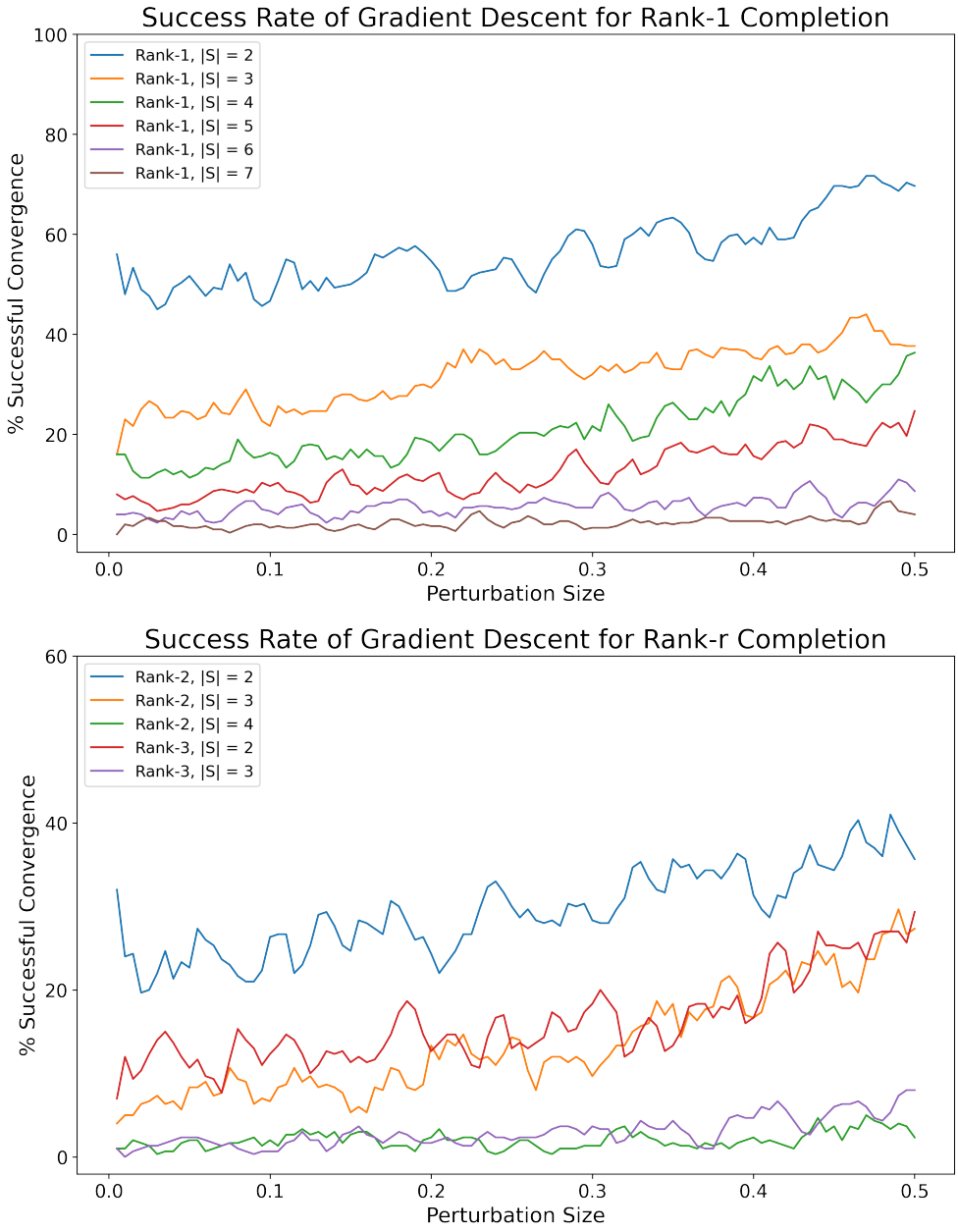


Figure 2.2: Success Rate of Gradient Descent Method for *(top)* Rank-1 and *(bottom)* Rank-r MC Problem Instances with Randomly Generated Measurement Sets.

# Appendices

## 2.A Proof of Proposition 1

*Proof.* The condition $\mathcal{P}_{\mathbf{M}^*,\Omega,n,r} \in \mathcal{L}(\mathcal{G}, n, r)$ implies that $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}_1)$ is connected and non-bipartite. Since the graph is non-bipartite, there exists a cycle with an odd number of vertices $\mathcal{C}_{odd} = (\mathcal{V}_{odd}, \mathcal{E}_{odd})$ in $\mathcal{G}_1$ in which $\mathcal{E}_{odd} \subset \mathcal{E}_1$. To numerically find an odd cycle, the breadth-first search method requires $\mathcal{O}(|\mathcal{V}| + |\mathcal{E}_1|) = \mathcal{O}(m^2)$ operations. Without loss of generality, we assume that the set of vertices of the cycle is $\mathcal{V}_{odd} = \{1, 2, \ldots, 2k+1\}$ and the set of edges is $\mathcal{E}_{odd} = \{(1,2), (2,3), \ldots, (2k+1, 1)\}$, where $k$ is a non-negative integer. Suppose that the matrix $\mathbf{X}^* \in \mathbb{R}^{n\times r}$ satisfies $\mathbf{M}^* = \mathbf{X}^*(\mathbf{X}^*)^\top$. We denote the $i$-th $r \times r$ block of $\mathbf{X}^*$ as $\mathbf{X}_i^*$ for all $i \in [m]$, i.e.,

$$\mathbf{M}^* = \mathbf{X}^*(\mathbf{X}^*)^\top = \begin{bmatrix} \mathbf{X}_1^* \\ \vdots \\ \mathbf{X}_m^* \end{bmatrix} \begin{bmatrix} (\mathbf{X}_1^*)^\top & \ldots & (\mathbf{X}_m^*)^\top \end{bmatrix}.$$

Since $\mathcal{P}_{\mathbf{M}^*,\Omega,n,r} \in \mathcal{L}(\mathcal{G}, n, r)$, the block $\mathbf{X}_i^*$ is non-singular for every $i \in [m]$, which further implies that the block $\mathbf{M}_{i,j}^*$ is non-singular for all $i, j \in [m]$. Using the relation that $\mathbf{M}_{i,j}^* = \mathbf{X}_i^*(\mathbf{X}_j^*)^\top$, we can calculate that

$$\left[\prod_{i=1}^{k} \left(\mathbf{M}_{2i-1,2i}^* \left(\mathbf{M}_{2i,2i+1}^*\right)^{-T}\right)\right] \mathbf{M}_{2k+1,1}^* = \mathbf{X}_1^*(\mathbf{X}_1^*)^\top,$$

Since the left-hand side only contains observed blocks, the matrix $\mathbf{X}_1^*(\mathbf{X}_1^*)^\top$ can be computed via observed blocks. Since computing the inverse of an $r \times r$ matrix and computing the product of two $r \times r$ matrices both require $\mathcal{O}(r^3)$ operations, the total number of operations required for computing $\mathbf{X}_1^*(\mathbf{X}_1^*)^\top$ is $\mathcal{O}[(2k+1)r^3] = \mathcal{O}(mr^3)$. In addition, computing the Cholesky decomposition of $\mathbf{X}_1^*(\mathbf{X}_1^*)^\top$ requires $\mathcal{O}(r^3)$ operations, which produces a matrix $\mathbf{X}_1^*\mathbf{R}$ for some orthogonal matrix $\mathbf{R} \in \mathbb{R}^{r\times r}$.

With the knowledge of $\mathbf{X}_1^*\mathbf{R}$, we can recursively compute the block $\mathbf{X}_i^*$ using the connectivity of $\mathcal{G}_1$. More specifically, we use $\mathcal{P} \subset [m]$ to denote the set of vertices $i$ or which we have computed $\mathbf{X}_i^*\mathbf{R}$. We start with $\mathcal{P} = \{1\}$. At each iteration, we choose indices $i \in \mathcal{P}$ and $j \notin \mathcal{P}$ such that $(i, j) \in \mathcal{E}_1$. Such a pair of indices always exist unless $\mathcal{P} = [m]$, since the graph $\mathcal{G}_1$ is connected. Then, using the observation

$$\mathbf{M}_{j,i}^* = \mathbf{X}_j^*(\mathbf{X}_i^*)^\top = (\mathbf{X}_j^*\mathbf{R})(\mathbf{X}_i^*\mathbf{R})^\top,$$

we first compute the matrix $\mathbf{X}_j^*\mathbf{R}$ with $\mathcal{O}(r^3)$ operations and then add $j$ to the set $\mathcal{P}$. We stop the iteration when $\mathcal{P} = [m]$. After this process, we can concatenate $\mathbf{X}_i^*\mathbf{R}$ for all $i \in [m]$ to obtain the matrix $\mathbf{X}^*\mathbf{R}$. The number of iterations is $m - 1$; thus, the total number of operations is $\mathcal{O}(mr^3)$.

Summarizing the two parts, the total number of operations to compute the matrix $\mathbf{X}^*\mathbf{U}$ is $\mathcal{O}(m^2 + mr^3) = \mathcal{O}[n^2/r^2 + nr^2]$.

□

### Gradient and Hessian of the Problem (2.2)

Before proceeding with the analysis for the proofs in the remainder of this chapter, we first derive the gradient and the Hessian of the objective function of the problem (2.2). We omit the proof since the calculation can be done using basic calculus. The gradient of the objective function can be written as

$$\nabla f(\mathbf{X}) = 2(\mathbf{X}\mathbf{X}^\top - \mathbf{M}^*)_\Omega \mathbf{X}. \tag{2.6}$$

Similarly, the quadratic variant of the Hessian can be written as

$$\Delta : \nabla^2 f(\mathbf{X}) : \Delta = 4\langle (\mathbf{X}\mathbf{X}^\top - \mathbf{M}^*)_\Omega, \Delta\Delta^\top \rangle + 2 \left\| \left(\mathbf{X}\Delta^\top + \Delta\mathbf{X}^\top\right)_\Omega \right\|_F^2. \tag{2.7}$$

## 2.B Proof of Theorem 1

*Proof.* Let $\mathcal{S}(\mathcal{G}_1)$ be a maximal independent set of $\mathcal{G}_1$ such that every vertex in the set has a self-loop. We define the global solution as $\mathbf{M}^* := \mathbf{x}^*(\mathbf{x}^*)^\top$, where

$$x_i^* := c_i, \quad \forall i \in \mathcal{S}, \quad x_i^* := 0, \quad \forall i \notin \mathcal{S}$$

and $\{c_i \mid \forall i \in \mathcal{S}\}$ is a set of non-zero constants. We note that in the case when $r = 1$, the factor $\mathbf{X} \in \mathbb{R}^n$ is a vector. Therefore, we represent it using the notation for vectors, i.e., $\mathbf{x}$. Considering the problem instance $\mathcal{P}_{\mathbf{M}^*,\Omega(\mathcal{G}),n,1}$, the set of global solutions of the problem (2.2) is given by

$$\mathcal{X}^* := \left\{ \mathbf{x} \in \mathbb{R}^n \mid x_i^2 = c_i^2,\ \forall i \in \mathcal{S},\ x_i = 0,\ \forall i \notin \mathcal{S} \right\},$$

which has the cardinality $|\mathcal{X}^*| = 2^{|\mathcal{S}|}$.

For every global solution $\hat{\mathbf{x}} \in \mathcal{X}^*$, we have $\hat{\mathbf{x}}\hat{\mathbf{x}}^\top = \mathbf{M}^*$. Thus, we know that $\hat{\mathbf{x}}$ is a first-order critical point of the problem (2.2). For every $\Delta \in \mathbb{R}^n$, the quadratic form of the Hessian (2.7) can

be written as

$$\begin{aligned}\Delta : \nabla^2 f(\hat{\mathbf{x}}) : \Delta &= 2\left\|\left(\hat{\mathbf{x}}\Delta^\top + \Delta\hat{\mathbf{x}}^\top\right)_\Omega\right\|_F^2 \\ &= \sum_{i\in\mathcal{S}} 2(\Delta_i\hat{x}_i + \hat{x}_i\Delta_i)^2 + \sum_{j\notin\mathcal{S}}\sum_{\substack{i\in\mathcal{S}\\(i,j)\in\mathcal{E}_1}} \left[2(\Delta_j\hat{x}_i)^2 + 2(\hat{x}_i\Delta_j)^2\right] \\ &= \sum_{i\in\mathcal{S}} 4(\Delta_i\hat{x}_i)^2 + \sum_{j\notin\mathcal{S}}\sum_{\substack{i\in\mathcal{S}\\(i,j)\in\mathcal{E}_1}} 4(\Delta_j\hat{x}_i)^2 \\ &= \sum_{i\in\mathcal{S}} 4\hat{x}_i^2\cdot\Delta_i^2 + \sum_{j\notin\mathcal{S}}\left(\sum\nolimits_{\substack{i\in\mathcal{S}\\(i,j)\in\mathcal{E}_1}} 4\hat{x}_i^2\right)\cdot\Delta_j^2.\end{aligned}$$

The first term in the above expression corresponds to self-loops in $\mathcal{S}$, while the second term corresponds to the edges between $\mathcal{S}$ and $\mathcal{S}^c := \mathcal{V}\backslash\mathcal{S}$. We note that the edges whose endpoints are both in $\mathcal{S}^c$ do not contribute to the quadratic form. Since $\mathcal{S}$ is a maximal independent, we know that

$$\{i\in\mathcal{S} \mid (i,j)\in\mathcal{E}_1\} \neq \emptyset, \quad \forall j\notin\mathcal{S}.$$

As a result, it holds that

$$\Delta : \nabla^2 f(\hat{\mathbf{x}}) : \Delta > 0, \quad \forall\Delta\in\mathbb{R}^n\backslash\{0\},$$

which implies that the Hessian at the global solution $\hat{\mathbf{x}}$ is positive definite. Then, we perturb the global solution of the above problem to be

$$\mathbf{M}^*(\epsilon) := \mathbf{x}^*(\epsilon)\left[\mathbf{x}^*(\epsilon)\right]^\top = (\mathbf{x}^*+\epsilon)(\mathbf{x}^*+\epsilon)^\top,$$

where $\mathbf{x}^*(\epsilon) := \mathbf{x}^* + \epsilon$ and $\epsilon\in\mathbb{R}^n$ is a small perturbation. We denote the problem (2.2) after perturbation as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \tilde{f}(\mathbf{x};\epsilon),$$

where $\tilde{f}(\mathbf{x};\epsilon) := \|(\mathbf{x}\mathbf{x}^\top - \mathbf{M}(\epsilon)^*)_\Omega\|_F^2$. For a generic perturbation $\epsilon$, all components of $\epsilon$ are non-zero and the problem $\mathcal{P}_{\mathbf{M}^*(\epsilon),\Omega,n,1}$ belongs to the class $\mathcal{L}(\mathcal{G},n,1)$. This implies that the global solution of the problem (2.3) is unique up to a sign flip.

The earlier argument implies that $\nabla_x\tilde{f}(\hat{\mathbf{x}};0) = 0$ and $\nabla_{xx}\tilde{f}(\hat{\mathbf{x}};0)\succ 0$. Now, we analyze the relationship between the local minima of the original problem and those of the perturbed problem. We consider the equation $\nabla_{\mathbf{x}}\tilde{f}(\mathbf{x};\epsilon) = 0$ near an unperturbed global minimum $\hat{\mathbf{x}}\in\mathcal{X}^*$. Since $(\hat{\mathbf{x}};0)$ is a solution to the gradient equation and the Jacobian matrix with respect to $\mathbf{x}$ is equal to the positive definite Hessian $\nabla^2 f(\hat{\mathbf{x}})$, the Implicit Function Theorem (IFT) states that there exists a unique solution $\hat{\mathbf{x}}(\epsilon)$ in a neighborhood of $\hat{\mathbf{x}}$ for every $\epsilon$ with a small norm. In addition, the continuity of the Hessian implies that $\nabla_{\mathbf{xx}}\tilde{f}(\hat{\mathbf{x}}(\epsilon);\epsilon)\succ 0$. Otherwise, we can find a sequence of $\epsilon^k\to 0$ and another sequence $\{y^k\}$ with $\|y_k\| = 1$ such that $(y^k)^\top\nabla_{\mathbf{xx}}\tilde{f}(\hat{\mathbf{x}}(\epsilon);\epsilon)y^k < 0$. We arrive at a contradiction by taking the limit along a convergent subsequence of $\{y^k\}$. Therefore, every point in $\hat{\mathbf{x}}^*(\epsilon)$ still satisfies the second-order sufficient conditions. Since there are $2^{|\mathcal{S}|}$ global minima for

the unperturbed problem, an analysis through the IFT implies that there are $2^{|\mathcal{S}|}$ strict local minima for the problem (2.3). Hence, there are $2^{|\mathcal{S}|} - 2$ spurious local minima for the perturbed problem (2.3).

□

## 2.C Proof of Corollary 1

*Proof.* We first prove the claim about the largest lower bound. Theorem 1 implies that there are at least $2^{|\mathcal{S}(\mathcal{G}_1)|} - 2$ spurious local minima for a problem instance in $\mathcal{L}(\mathcal{G}, n, 1)$. For an arbitrary connected and non-bipartite graph $\mathcal{G}_1$ with $n$ vertices, the maximal possible size of a maximal independent set with self-loops is $n - 1$. More specifically, the graph $\mathcal{G}_1^*$ that attains this maximal value is the star graph $K_{1,n-1}$ complemented with self-loops for the $n - 1$ independent vertices. Hence, Theorem 1 implies that there exists a problem instance in $\mathcal{L}(\mathcal{G}^*, n, 1)$ with at least $2^{n-1} - 2$ local minima, where we define $\mathcal{G}^* := (\mathcal{G}_1^*, \emptyset)$.

We then consider the second claim of this corollary. Consider the measurement set $\Omega$ that observes all entries of the ground truth $\mathbf{M}^*$ except $\mathbf{M}_{12}^*$ and $\mathbf{M}_{21}^*$. We have $|\Omega| = n^2 - 2$ in this case. Choose the set of vertices to be $\mathcal{V} := [n]$ and the set of edges to be $\mathcal{E}_1 := [n] \times [n] \backslash \{(1, 2), (2, 1)\}$. Then, the graph $\mathcal{G} := (\mathcal{V}, \mathcal{E}_1, \emptyset)$ satisfies that $\Omega = \Omega(\mathcal{G})$ and the maximal independent set is $\mathcal{S} := \{i, j\}$. Therefore, Theorem 1 implies that there exists a problem instance in $\mathcal{L}(\mathcal{G}, n, 1)$ that has at least $2^{|\mathcal{S}|} - 2 = 2$ spurious local minima. □

## 2.D Proof of Corollary 2

*Proof.* For a generic vector $\mathbf{x}^* \in \mathbb{R}^n$, all elements of $\mathbf{x}^*$ are non-zero and we can decompose $\mathbf{x}^*$ into $\mathbf{x}^0 + \mathbf{x}^1$, where

$$\begin{aligned} x_i^0 &:= x_i^*, \quad \forall i \in \mathcal{S}, \qquad x_i^0 := 0, \qquad \forall i \notin \mathcal{S}, \\ x_i^1 &:= 0, \quad \forall i \in \mathcal{S}, \qquad x_i^1 := x_i^*, \qquad \forall i \notin \mathcal{S}. \end{aligned}$$

We first consider the problem $\mathcal{P}_{\mathbf{x}^0(\mathbf{x}^0)^\top, \Omega(\mathcal{G}), n, 1}$. Using a similar proof as Theorem 1, we can prove that this problem has $2^{|\mathcal{S}|}$ equivalent global solutions in formulation (2.2), which are described by the set

$$\mathcal{X}^* := \left\{ \mathbf{x} \in \mathbb{R}^n \mid x_i^2 = (x_i^*)^2, \ \forall i \in \mathcal{S}, \ x_i = 0, \ \forall i \notin \mathcal{S} \right\}.$$

In addition, the Hessian is positive definite at every global solution $\hat{\mathbf{x}} \in \mathcal{X}^*$. Using a similar argument as the proof of Theorem 1, we conclude that there exists a small constant $r_{\mathbf{x}^0} > 0$ such that the conditions

$$\|\epsilon\| \leq r_{\mathbf{x}^0}, \quad \epsilon_i \neq 0, \quad \forall i \in [m] \tag{2.8}$$

imply that the problem $\mathcal{P}_{\mathbf{x}^0(\epsilon)[\mathbf{x}^0(\epsilon)]^\top, \Omega(\mathcal{G}), n, 1}$ has at least $2^{|\mathcal{S}|} - 2$ spurious local minima in formulation (2.2), where $\mathbf{x}^0(\epsilon) := \mathbf{x}^0 + \epsilon$. Moreover, the MC problem is "scale-free" in the sense that the formulation (2.2) of the problem $\mathcal{P}_{\mathbf{x}'(\mathbf{x}')^\top, \Omega(\mathcal{G}), n, 1}$ has spurious local minima if and only if that of

the problem $\mathcal{P}_{(c\mathbf{x}')(c\mathbf{x}')^\top,\Omega(\mathcal{G}),n,1}$ has spurious local minima, where $\mathbf{x}' \in \mathbb{R}^n$ is an arbitrary vector and $c \neq 0$ is a constant. Therefore, we have the relation

$$r_{c\mathbf{x}^0} = c \cdot r_{\mathbf{x}^0}, \quad \forall c \neq 0. \tag{2.9}$$

Hence, it suffices to consider vectors $\mathbf{x}^0 \in \mathcal{X}_0$, where

$$\mathcal{X}_0 := \{\mathbf{x} \in \mathbb{R}^n \mid x_i \neq 0,\ \forall i \in \mathcal{S},\ x_i = 0,\ \forall i \notin \mathcal{S},\ \|\mathbf{x}\| = 1\}.$$

Now, we consider a vector $\hat{\mathbf{x}}^0 \in \mathbb{R}^n$ satisfying

$$\|\hat{\mathbf{x}}^0 - \mathbf{x}^0\| < r_{\mathbf{x}^0}/2.$$

Then, as long as the generic perturbation $\epsilon$ satisfies $\|\epsilon\| \leq r_{\mathbf{x}^0}/2$, the condition (2.8) implies that the problem $\mathcal{P}_{\hat{\mathbf{x}}^*(\epsilon)[\hat{\mathbf{x}}^*(\epsilon)]^\top,\Omega(\mathcal{G}),n,1}$ has at least $2^{|\mathcal{S}|} - 2$ spurious local minima in formulation (2.2), where $\hat{\mathbf{x}}^*(\epsilon) := \hat{\mathbf{x}}^0 + \epsilon$. This verifies the existence of a function $h_{\mathcal{S}}$ on the open set

$$\mathcal{N}(\mathbf{x}^0) := \left\{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}^0\| < r_{\mathbf{x}^0}/2\right\}, \quad \forall \mathbf{x}^0 \in \mathcal{X}_0.$$

Now, we consider the compact set

$$\mathcal{X}_k := \left\{\mathbf{x} \in \mathbb{R}^n \mid \min_{i\in\mathcal{S}} |x_i| \geq 1/k,\ x_i = 0,\ \forall i \notin \mathcal{S},\ \|\mathbf{x}\| = 1\right\}, \quad k = 1, 2, \ldots$$

The open set family $\{\mathcal{N}(\mathbf{x}^0) \mid \mathbf{x}^0 \in \mathcal{X}_0\}$ comprises an open covering of the compact set $\mathcal{X}_k$. Hence, there exists a finite covering of the compact set $\mathcal{X}_k$, and thus there exists a small constant $\epsilon_k$ such that

$$h_{\mathcal{S}}(\mathbf{x}^0) \geq \epsilon_k, \quad \forall \mathbf{x}^0 \in \mathcal{X}_k.$$

Moreover, using the relation

$$\mathcal{X}_{k-1} \subset \mathcal{X}_k, \quad \forall k \geq 2, \quad \mathcal{X}_0 = \bigcup_{k=1}^{\infty} \mathcal{X}_k,$$

we can decrease the value of the function $h_{\mathcal{S}}$ to be

$$h(\mathbf{x}^0) := \epsilon_k, \quad \forall \mathbf{x}^0 \in \mathcal{X}_k \backslash \mathcal{X}_{k-1},\ k \geq 2.$$

Using this definition, $h_{\mathcal{S}}$ reduces to $\min_{i\in\mathcal{S}} |x_i|$ and, with a bit of abuse of notations, we still write the new function as $h_{\mathcal{S}}(\min_{i\in\mathcal{S}} |x_i|)$.

Now, we can view the problem $\mathcal{P}_{\mathbf{x}^*(\mathbf{x}^*)^\top,\Omega(\mathcal{G}),n,1}$ as the perturbed problem, where the generic perturbation is given by $\mathbf{x}^1$. The above analysis implies that the following condition holds

$$\|\mathbf{x}_{\mathcal{S}^{c*}}\| = \|\mathbf{x}^1\| \leq h_{\mathcal{S}}\left(\min_{i\in\mathcal{S}} |x_i|\right) \cdot \|\mathbf{x}^0\| = h_{\mathcal{S}}\left(\min_{i\in\mathcal{S}} |x_i|\right) \cdot \|\mathbf{x}^*_{\mathcal{S}}\|$$

which demonstrates the existence of at least $2^{|\mathcal{S}|} - 2$ spurious local minima in formulation (2.2).

□

## 2.E Proof of Lemma 1

*Proof.* We follow a similar proof construction as in Theorem 1. Let $\mathcal{D}$ be the set of full-rank diagonal matrices and $\mathcal{D}_{\{1,-1\}}$ be the set of diagonal matrices with the diagonal entries being $+1$ or $-1$. Suppose that $\mathcal{S}$ is a maximal independent set of $\mathcal{G}_1$ in which every node has a self-loop. Then, we define the ground truth matrix $\mathbf{M}^* := \mathbf{X}^*(\mathbf{X}^*)^\top$, where $\mathbf{X}^* \in \mathbb{R}^{n\times r}$ satisfies

$$\mathbf{X}_i^* = 0, \quad \forall i \notin \mathcal{S}, \quad \mathbf{X}_i^* \in \mathcal{D}, \quad \forall i \in \mathcal{S},$$

where $\mathbf{X}_i$ is the $i$-th block of $\mathbf{X} \in \mathbb{R}^{n\times r}$; see the definition in the proof of Proposition 1. Then, the set of global solutions is given by

$$\mathcal{X}^* := \Big\{\mathbf{X} \in \mathbb{R}^{n\times r} \mid \mathbf{X}_i = \mathbf{X}_i^*\mathbf{D},\ \forall i \in \mathcal{S},\ \mathbf{D} \in \mathcal{D}_{\{1,-1\}}, \quad \mathbf{X}_i = 0,\ \forall i \notin \mathcal{S}\Big\},$$

For every global solution $\hat{\mathbf{X}} \in \mathcal{X}^*$, we have $\hat{\mathbf{X}}\hat{\mathbf{X}}^\top = \mathbf{M}^*$. Thus, every global solution is a first-order critical point of the problem (2.2). In addition, let $\Delta \in \mathbb{R}^{n\times r}$ be an arbitrary direction matrix with its $r\times r$ block matrices denoted as $\Delta_1, \Delta_2, \ldots, \Delta_m$. Then, the quadratic variant of the Hessian (2.7) in the direction $\Delta$ can be written as

$$\Delta : \nabla^2 f(\hat{\mathbf{X}}) : \Delta = 2\left\|\left(\hat{\mathbf{X}}\Delta^\top + \Delta\hat{\mathbf{X}}^\top\right)_\Omega\right\|_F^2 \tag{2.10}$$

$$= \sum_{i\in\mathcal{S}} \|\Delta_i\hat{\mathbf{X}}_i^\top + \hat{\mathbf{X}}_i\Delta_i^\top\|_F^2 \tag{2.11}$$

$$+ 2\sum_{j\notin\mathcal{S}}\left(\sum_{\substack{i\in\mathcal{S}\\(i,j)\in\mathcal{E}_1}} \|\Delta_j\hat{\mathbf{X}}_i^\top\|_F^2 + \sum_{\substack{i\in\mathcal{S}\\(i,j)\in\mathcal{E}_2}} \|(\Delta_j\hat{\mathbf{X}}_i^\top)_{nd}\|_F^2\right)$$

$$+ 2\sum_{j\in\mathcal{S}}\sum_{\substack{i\in\mathcal{S}\\(i,j)\in\mathcal{E}_2}} \|(\Delta_j\hat{\mathbf{X}}_i^\top + \hat{\mathbf{X}}_j\Delta_i^\top)_{nd}\|_F^2$$

where $(\cdot)_{nd}$ is the projection onto the matrix space with a zero diagonal. We note that the first term in (2.10) corresponds to self-loops in $\mathcal{G}_1$, while the second term corresponds to edges between $\mathcal{S}$ and $\mathcal{S}^c$. The edges whose endpoints are both in $\mathcal{S}^c$ do not contribute to the quadratic form. Moreover, the last term corresponds to partial observations with non-diagonal entries within the independent set $\mathcal{S}$.

Now, we aim to prove that the Hessian at $\hat{\mathbf{X}}$ is positive definite in the tangent space of $\mathcal{W}^{n\times r}$, namely,

$$\Delta : \nabla^2 f(\hat{\mathbf{X}}) : \Delta > 0, \quad \forall \Delta \in \mathbb{R}^{n\times r}\backslash\{0\}, \quad \Delta_1 \text{ is lower triangular.}$$

We assume that $\Delta : \nabla^2 f(\hat{\mathbf{X}}) : \Delta = 0$ for some $\Delta \in \mathbb{R}^{n\times r}$ such that $\Delta_1$ is lower triangular. Under this assumption, all three terms in (2.10) are equal to zero. Considering the second term, since $\hat{\mathbf{X}}_i$ is full-rank, we have

$$\Delta_j = 0, \quad \forall j \notin \mathcal{S}.$$

For the first term, $\|\Delta_i\hat{\mathbf{X}}_i^\top + \hat{\mathbf{X}}_i\Delta_i^\top\|_F^2$ is zero only if $\Delta_i\hat{\mathbf{X}}_i^\top = -\hat{\mathbf{X}}_i\Delta_i^\top$, i.e., $\Delta_i\hat{\mathbf{X}}_i^\top$ is skew-symmetric. Since $\hat{\mathbf{X}}_i$ is a diagonal matrix with non-zero diagonal entries, the diagonal entries of $\Delta_i$ must be zero for all $i \in \mathcal{S}$. Without loss of generality, we assumed that vertex $1 \in \mathcal{S}$. This is because we can equivalently fix the block $\mathbf{X}_i$ to be lower-diagonal for any $i \in \mathcal{S}$ and consider a similarly constrained optimization problem. Then, since $\Delta_1$ must be lower triangular, we have

$$\Delta_1 = 0,$$

We define the set

$$\mathcal{S}_0 := \{i \in \mathcal{S} \mid \Delta_i = 0\}.$$

We have shown that $i \in \mathcal{S}_0$ and aim to prove that $\mathcal{S}_0 = \mathcal{S}$. Since the induced subgraph $\mathcal{G}_2[\mathcal{S}]$ is connected, there exists a vertex $j \in \mathcal{S}$ such that $(1, j) \in \mathcal{E}_2$. Considering the third term in (2.10), we have

$$(\Delta_j\hat{\mathbf{X}}_1^\top + \hat{\mathbf{X}}_j\Delta_1^\top)_{nd} = (\Delta_j\hat{\mathbf{X}}_1^\top)_{nd} = 0,$$

which implies that $\Delta_j = 0$ because $\hat{\mathbf{X}}_j$ is a diagonal matrix with non-zero diagonal entries. Hence, we have proved that $j \in \mathcal{S}_0$.

By the connectivity of $\mathcal{G}_2[\mathcal{S}]$, we can inductively prove that all elements in $\mathcal{S}$ belong to $\mathcal{S}_0$. Therefore, it holds that $\Delta = 0$ and the quadratic form of Hessian is zero only when $\Delta = 0$. As a result, the Hessian is positive definite at every global solution $\hat{\mathbf{X}} \in \mathcal{X}^*$ of the problem (2.4).

Then, we perturb the ground truth of the above problem instance to be

$$\mathbf{M}^*(\epsilon) := \mathbf{X}^*(\epsilon)\left[\mathbf{X}^*(\epsilon)\right]^\top = (\mathbf{X}^* + \epsilon)(\mathbf{X}^* + \epsilon)^\top,$$

where $\mathbf{X}^*(\epsilon) := \mathbf{X}^* + \epsilon$ and $\epsilon \in \mathbb{R}^{n\times r}$ is a small perturbation. Similar to Theorem 1, for a generic perturbation $\epsilon$, all block components of $\epsilon$ are non-zero and full-rank. Therefore, the problem $\mathcal{P}_{\mathbf{M}^*(\epsilon),\Omega,n,r}$ belongs to the class $\mathcal{L}(\mathcal{G}, n, r)$. This implies that the global solution of the problem (2.4) is unique up to a right-multiplication with $\mathbf{D} \in \mathcal{D}_{\{1,-1\}}$. Since there are $2^{r|\mathcal{S}|}$ global minima for the unperturbed problem, IFT implies that there are $2^{r|\mathcal{S}|}$ strict local minima for the perturbed problem. Hence, there are $2^{r|\mathcal{S}|} - 2^r$ spurious local minima for the perturbed problem. □

# 2.F Proof of Corollary 3

*Proof.* We first consider the largest lower bound on the number of spurious local minima. Similar to Corollary 1, since an instance in $\mathcal{L}(\mathcal{G}, n, r)$ can have a maximum independent set of size at most $|\mathcal{S}(\mathcal{G}_1)| = m - 1 = n/r - 1$, Theorem 2 implies that the largest lower bound on the number of spurious local solution classes is $2^{r(n/r-2)} - 1 = 2^{n-2r} - 1$.

We prove the second part of this corollary next. We choose $\mathcal{G}_1$ to be a graph with $m$ vertices and $\binom{m}{2} - 1$ edges, where $(i, j)$ is the only missing edge. Then, the maximal independent set is $\mathcal{S} := \{i, j\}$. Since $\mathcal{G}_2[\mathcal{S}]$ must be connected, the non-diagonal entries of the block $\mathbf{M}^*_{i,j}$ are observed. Thus, only $2r$ entries are not observed and $|\Omega| = n^2 - 2r$. Furthermore, Theorem 2 implies that there exists a problem instance in $\mathcal{L}(\mathcal{G}, n, r)$ with at least $2^{r(|\mathcal{S}|-1)} - 1 = 2^r - 1$ equivalent classes of spurious local minima. □

## 2.G Proof of Proposition 2

*Proof.* For every matrix $\mathbf{X} \in \mathbb{R}^{n\times r}$, we denote $\mathbf{X}_i$ as the $i$-th $r \times r$ block of $\mathbf{X}$ for all $i \in [m]$. Because each block $\mathbf{M}^*_{i,j}$ is assumed to be full rank, the block $\mathbf{X}^*_i$ is also full rank for all $i \in [m]$, where $\mathbf{X}^*$ satisfies $\mathbf{M}^* = \mathbf{X}^*(\mathbf{X}^*)^\top$. It is desirable to show that every first-order critical point is either a global solution or a saddle point with a strict descent direction. For every $i \in [m]$, the gradient of the problem (2.2) with respect to the $i$-th block $\mathbf{X}_i$ is

$$\nabla_{\mathbf{X}_i} f(\mathbf{X}) = \begin{cases} 2(\mathbf{X}_i\mathbf{X}_k^\top - \mathbf{X}^*_i(\mathbf{X}^*_k)^\top)\mathbf{X}_k, & \text{if } i \neq k \\ \sum_{j=1}^m 2(\mathbf{X}_k\mathbf{X}_j^\top - \mathbf{X}^*_k(\mathbf{X}^*_j)^\top)\mathbf{X}_j, & \text{if } i = k. \end{cases}$$

Let $\hat{\mathbf{X}} \in \mathbb{R}^{n\times r}$ be a first-order critical point of problem (2.2).

We first consider the case when $\hat{\mathbf{X}}_k$ is non-singular. For every $i \in [m]\backslash\{k\}$, the condition $\nabla_{\mathbf{X}_i} f(\mathbf{X}) = 0$ implies that

$$\hat{\mathbf{X}}_i\hat{\mathbf{X}}_k^\top - \mathbf{X}^*_i(\mathbf{X}^*_k)^\top = 0.$$

Substituting the above equations into the equation $\nabla_{\mathbf{X}_k} f(\mathbf{X}) = 0$, we obtain

$$(\mathbf{X}_k\mathbf{X}_k^\top - \mathbf{X}^*_k(\mathbf{X}^*_k)^\top)\mathbf{X}_k = 0,$$

which implies that $\hat{\mathbf{X}}_k\hat{\mathbf{X}}_k^\top = \mathbf{X}^*_k(\mathbf{X}^*_k)^\top$. Therefore, the matrix $\hat{\mathbf{X}}$ is a global solution of the problem (2.2) in this case.

Now, we consider the case when $\hat{\mathbf{X}}_k$ is singular. We choose a vector $\mathbf{y}_k \in \mathbb{R}^n$ such that

$$\hat{\mathbf{X}}_k\mathbf{y}_k = 0, \quad \|\mathbf{y}_k\| = 1.$$

Given a small constant $\epsilon > 0$, the $i$-th block direction $\Delta \in \mathbb{R}^{n\times r}$ is defined as

$$\Delta_i := \begin{cases} \mathbf{z}_i\mathbf{y}_k^\top, & \text{if } i \neq k \\ \epsilon\mathbf{y}_k\mathbf{y}_k^\top, & \text{if } i = k, \end{cases}$$

where $\mathbf{z}_i \in \mathbb{R}^n$ is arbitrary. The above definition directly implies that $\hat{\mathbf{X}}_k\Delta_i^\top = 0$ for all $i \in [m]$. Then, we obtain

$$\begin{aligned} 4\left\langle(\hat{\mathbf{X}}\hat{\mathbf{X}}^\top - \mathbf{X}^*(\mathbf{X}^*)^\top)_\Omega, \Delta\Delta^\top\right\rangle &= -4\mathbf{tr}\left[\mathbf{X}^*_k(\mathbf{X}^*_k)^\top\Delta_k\Delta_k^\top\right] - \sum_{j=1,j\neq i}^m 8\mathbf{tr}\left(\mathbf{X}^*_j(\mathbf{X}^*_k)^\top\Delta_k\Delta_j^\top\right) \\ &= \sum_{j=1,j\neq i}^m -8\mathbf{tr}\left[\mathbf{X}^*_j(\mathbf{X}^*_k)^\top\mathbf{y}_k\mathbf{z}_j^\top\right]\cdot\epsilon + \mathcal{O}(\epsilon^2), \end{aligned}$$

and

$$2\|(\mathbf{X}\Delta^\top + \Delta\mathbf{X}^\top)_\Omega\|_F^2 = 8\|\mathbf{X}_k\Delta_k^\top\|_F^2 + 4\sum_{j=1,j\neq k}^m \|\mathbf{X}_j\Delta_k^\top\|_F^2 = \mathcal{O}(\epsilon^2)$$

Combining the two estimates above, the quadratic form of the Hessian (2.7) can be written as

$$\Delta : \nabla^2 f(\hat{\mathbf{X}}) : \Delta = \sum_{j=1, j\neq i}^{m} -8\mathbf{tr}\left[\mathbf{X}_j^*(\mathbf{X}_k^*)^\top \mathbf{y}_k \mathbf{z}_j^\top\right] \cdot \epsilon + \mathcal{O}(\epsilon^2).$$

Since $\mathbf{X}_i^*$ is non-singular for all $i \in [m]$, it holds that $\mathbf{X}_j^*(\mathbf{X}_k^*)^\top \mathbf{y}_k \neq 0$. Choosing

$$\mathbf{z}_j := \mathbf{X}_j^*(\mathbf{X}_k^*)^\top \mathbf{y}_k,$$

we obtain

$$\Delta : \nabla^2 f(\hat{\mathbf{X}}) : \Delta = \sum_{j=1, j\neq i}^{m} -8\|\mathbf{X}_j^*(\mathbf{X}_k^*)^\top \mathbf{y}_k\|^2 \cdot \epsilon + \mathcal{O}(\epsilon^2).$$

Hence, the quadratic form of the Hessian is negative with a sufficiently small $\epsilon$ and $\hat{\mathbf{X}}$ is a strict saddle point.

Combining the two cases, we conclude that every second-order critical point of the problem (2.2) is a global minimum.

□

## 2.H Proof of Lemma 2

In this proof and the following proofs for Section 2.4, we consider the instance of the MC problem constructed in Section 2.1. For completeness, we repeat the instance here. In the unperturbed case, the ground truth matrix is defined as $\mathbf{M}^* := \mathbf{x}^*(\mathbf{x}^*)^\top \in \mathbb{R}^{n\times n}$, where vector $\mathbf{x}^* \in \mathbb{R}^n$ satisfies

$$\mathbf{x}_{2k-1}^* = 1, \quad \forall k = 1, \ldots, \lceil n/2 \rceil, \quad \mathbf{x}_{2k}^* = 0, \quad \forall k = 1, \ldots, \lfloor n/2 \rfloor.$$

The measurement set $\Omega$ is given by

$$\Omega := \{(j,j), (2k,j), (j,2k) \mid j = 1, \ldots, n,\ k = 1, \ldots, \lfloor n/2 \rfloor\}.$$

It has been proved in Section 2.1 that the problem (2.1) has $2^{\lceil n/2 \rceil}$ global solutions, which are given by the set

$$\mathcal{X}^* := \left\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}_{2k} = 0,\ k = 1, \ldots, \lfloor n/2 \rfloor,\ \mathbf{x}_{2k+1}^2 = 1,\ k = 1, \ldots, \lceil n/2 \rceil\right\}.$$

For each vector $\mathbf{x} \in \mathbb{R}^n$, we denote

$$\mathbf{x}^o := (x_1, x_3, \ldots, x_{\lceil n/2 \rceil}), \quad \mathbf{x}^e := (x_2, x_4, \ldots, x_{\lfloor n/2 \rfloor}).$$

Before presenting the proof of Lemma 2, we first state three technical lemmas.

**Lemma 3.** *Suppose that Assumption 1 holds with* $(\delta, 1)$ *and that* $\hat{\mathbf{x}}$ *is a first-order critical point of the problem* (2.1)*. Then, it holds that*

$$\|\hat{\mathbf{x}}^e\|\|\hat{\mathbf{x}}\| \leq 2\sqrt{2}\delta\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F,$$

*where we define* $\mathbf{M} := \hat{\mathbf{x}}\hat{\mathbf{x}}^\top$.

*Proof.* Utilizing the first-order optimality condition and the gradient in (2.6), it holds that

$$\left\langle \nabla g\left[\left(\hat{\mathbf{x}}\hat{\mathbf{x}}^\top - \mathbf{x}^*\mathbf{x}^*\right)_\Omega\right], \hat{\mathbf{x}}\Delta^\top \right\rangle = \int_0^1 \left(\mathbf{M} - \mathbf{M}^*\right) : \nabla^2 g\left[(\mathbf{M}^*)_\Omega + t\left(\mathbf{M} - \mathbf{M}^*\right)_\Omega\right] : \hat{\mathbf{x}}\Delta^\top \, dt = 0,$$

$\forall \Delta \in \mathbb{R}^n$, where the first equality is from Taylor's expansion. For every fixed number $t \in [0, 1]$, the proof of Theorem 1 in [13] implies that

$$\begin{aligned}\left(\mathbf{M} - \mathbf{M}^*\right) : \nabla^2 g\left[\mathbf{M}^*{}_\Omega + t\left(\mathbf{M} - \mathbf{M}^*\right)_\Omega\right] : \hat{\mathbf{x}}\Delta^\top \\ \geq \left\langle \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega, \left(\hat{\mathbf{x}}\Delta^\top\right)_\Omega \right\rangle - 2\sqrt{2}\delta\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F \| \left(\hat{\mathbf{x}}\Delta^\top\right)_\Omega \|_F.\end{aligned}$$

Integrating over $t$, it follows that

$$\left\langle \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega, \left(\hat{\mathbf{x}}\Delta^\top\right)_\Omega \right\rangle \leq 2\sqrt{2}\delta\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F \| \left(\hat{\mathbf{x}}\Delta^\top\right)_\Omega \|_F. \tag{2.12}$$

By choosing

$$\Delta_{2k+1} = 0, \quad k = 1, \ldots, \lceil n/2 \rceil, \quad \Delta_{2k} = \hat{x}_{2k}, \quad k = 1, \ldots, \lfloor n/2 \rfloor,$$

we obtain

$$\left\langle \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega, \left(\hat{\mathbf{x}}\Delta^\top\right)_\Omega \right\rangle = \|\hat{\mathbf{x}}^e\|^2\|\hat{\mathbf{x}}\|^2, \quad \| \left(\hat{\mathbf{x}}\Delta^\top\right)_\Omega \|_F = \|\hat{\mathbf{x}}^e\|\|\hat{\mathbf{x}}\|.$$

Substituting the above two equalities into (2.12), we have

$$\|\hat{\mathbf{x}}^e\|^2\|\hat{\mathbf{x}}\|^2 \leq 2\sqrt{2}\delta\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F \|\hat{\mathbf{x}}^e\|\|\hat{\mathbf{x}}\|.$$

The above inequality implies that

$$\|\hat{\mathbf{x}}^e\|\|\hat{\mathbf{x}}\| \leq 2\sqrt{2}\delta\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F \quad \text{or} \quad \hat{\mathbf{x}}^e = 0,$$

since $\|\hat{\mathbf{x}}^e\|\|\hat{\mathbf{x}}\| = 0$ if and only if $\hat{\mathbf{x}}^e = 0$. In both cases, the claim of this lemma holds. □

**Lemma 4.** *Let* $\mathcal{D}$ *be the set of* $r \times r$ *diagonal matrices and* $\mathcal{D}_{\{1,-1\}}$ *be set of* $r \times r$ *diagonal matrices with the diagonal entries* $+1$ *or* $-1$*. Under the same setting as Lemma 1, consider the* $n \times n$ *ground truth matrix* $\mathbf{M}^* := \mathbf{X}^*(\mathbf{X}^*)^\top$ *such that* $n = mr$*, where*

$$\mathbf{X}_i^* \in \mathcal{D}, \quad \forall i \in \mathcal{S}(\mathcal{G}_1), \quad \mathbf{X}_i^* = \mathbf{0}, \quad \forall i \notin \mathcal{S}(\mathcal{G}_1).$$

*Let* $\mathbf{D} \in \mathbb{R}^{n \times n}$ *be an arbitrary matrix with its diagonal blocks denoted as* $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_m$ *such that* $\mathbf{D}_i \in \mathcal{D}_{\{1,-1\}}$ *for all* $i \in [m]$*. Then, the problem instance* $\mathcal{P}_{\mathbf{M}^*,\Omega(\mathcal{G}),n,r}$ *in formulation* (2.2) *satisfies that*

$$\nabla f(\mathbf{DX}) = \mathbf{D}\nabla f(\mathbf{X}), \quad \forall \mathbf{X} \in \mathbb{R}^{n \times r}.$$

*Proof.* Since the graph $\mathcal{G}$ satisfies the conditions in Lemma 1, the gradient $\nabla f(\mathbf{X})$ for problem (2.2) can be written as

$$\nabla_{\mathbf{X}_k} f(\mathbf{DX}) = (\mathbf{X}_k\mathbf{X}_k^\top - \mathbf{X}_k^*(\mathbf{X}_k^*)^\top)\mathbf{X}_k + \sum_{(k,j)\in\mathcal{E}_1} \mathbf{X}_k\mathbf{X}_j^\top\mathbf{X}_j + \sum_{(k,j)\in\mathcal{E}_2} (\mathbf{X}_k\mathbf{X}_j^\top - \mathbf{X}_k^*(\mathbf{X}_j^*)^\top)_{nd}\mathbf{X}_j, \quad \text{if } k \in \mathcal{S},$$

and

$$\nabla_{\mathbf{X}_k} f(\mathbf{DX}) = \sum_{(k,j)\in\mathcal{E}_1} \mathbf{X}_k\mathbf{X}_j^\top\mathbf{X}_j + \sum_{(k,j)\in\mathcal{E}_2} (\mathbf{X}_k\mathbf{X}_j^\top)_{nd}\mathbf{X}_j, \quad \text{if } k \notin \mathcal{S},$$

where $(\cdot)_{nd}$ is the projection onto the matrix space with the zero diagonals. Note that blocks of the transformed variable $\mathbf{DX}$ are $\mathbf{D}_i\mathbf{X}_i$ for all $i \in [m]$. First, we consider the change in the $i$-th block of the gradient function for $i \in \mathcal{S}$:

$$\nabla_{\mathbf{X}_i} f(\mathbf{DX}) = 2\Big((\mathbf{D}_i\mathbf{X_i}\mathbf{X_i}^\top\mathbf{D}_i^\top - \mathbf{X}_i^*(\mathbf{X}_i^*)^\top)\mathbf{D}_i\mathbf{X}_i + \sum_{(i,j)\in\mathcal{E}_1} \mathbf{D}_i\mathbf{X}_i\mathbf{X}_j^\top\mathbf{D}_j^\top\mathbf{D}_j\mathbf{X}_j + \sum_{(i,j)\in\mathcal{E}_2} (\mathbf{D}_i\mathbf{X}_i\mathbf{X_j}^\top\mathbf{D_j}^\top - \mathbf{X}_i^*(\mathbf{X}_j^*)^\top)_{nd}\mathbf{D}_j\mathbf{X}_j\Big). \tag{2.13}$$

Using $\mathbf{D}_i^\top\mathbf{D}_i = \mathbf{I}$ and $\mathbf{X}_i^* \in \mathcal{D}$, the first term in (2.13) can be written as

$$\begin{aligned}(\mathbf{D}_i\mathbf{X_i}\mathbf{X}_i^\top\mathbf{D}_i^\top - \mathbf{X}_i^*(\mathbf{X}_i^*)^\top)\mathbf{D}_i\mathbf{X}_i &= \mathbf{D}_i\mathbf{X_i}\mathbf{X}_i^\top\mathbf{X}_i - \mathbf{X}_i^*(\mathbf{X}_i^*)^\top\mathbf{D}_i\mathbf{X}_i \\ &= \mathbf{D}_i(\mathbf{X_i}\mathbf{X}_i^\top - \mathbf{X}_i^*(\mathbf{X}_i^*)^\top)\mathbf{X}_i,\end{aligned}$$

where the second equality is justified by the commutative property of diagonal matrix multiplication. Similarly, the second term in (2.13) can be written as

$$\sum_{(i,j)\in\mathcal{E}_1} \mathbf{D}_i\mathbf{X}_i\mathbf{X}_j^\top\mathbf{D}_j^\top\mathbf{D}_j\mathbf{X}_j = \mathbf{D}_i\sum_{(i,j)\in\mathcal{E}_1} \mathbf{X}_i\mathbf{X}_j^\top\mathbf{X}_j.$$

For the last term in (2.13), we use the relation $\mathbf{X}_i^* \in \mathcal{D}$ and the fact that $(\cdot)_{nd}$ is non-zero only at positions associated with the non-diagonal entries to obtain

$$\begin{aligned}\sum_{(i,j)\in\mathcal{E}_2} (\mathbf{D}_i\mathbf{X}_i\mathbf{X_j}^\top\mathbf{D_j}^\top - \mathbf{X}_i^*(\mathbf{X}_j^*)^\top)_{nd}\mathbf{D}_j\mathbf{X}_j &= \sum_{(i,j)\in\mathcal{E}_2} \mathbf{D}_i(\mathbf{X}_i\mathbf{X}_j^\top - \mathbf{X}_i^*(\mathbf{X}_j^*)^\top)_{nd}\mathbf{D}_j^\top\mathbf{D}_j\mathbf{X}_j \\ &= \mathbf{D}_i\sum_{(i,j)\in\mathcal{E}_2} (\mathbf{X}_i\mathbf{X}_j^\top - \mathbf{X}_i^*(\mathbf{X}_j^*)^\top)_{nd}\mathbf{X}_j.\end{aligned}$$

Thus, we have

$$\nabla_{\mathbf{X}_i} f(\mathbf{DX}) = \mathbf{D}_i\nabla_{\mathbf{X}_i} f(\mathbf{X}), \quad \forall i \in \mathcal{S}.$$

Now, we consider the change in the $i$-th block of the gradient function for $i \notin \mathcal{S}$:

$$\begin{aligned}\nabla_{\mathbf{X}_i} f(\mathbf{DX}) &= 2\left(\sum_{(i,j)\in\mathcal{E}_1} \mathbf{D}_i\mathbf{X}_i\mathbf{X}_j^\top\mathbf{D}_j^\top\mathbf{D}_j\mathbf{X}_j + \sum_{(i,j)\in\mathcal{E}_2} (\mathbf{D}_i\mathbf{X}_i\mathbf{X}_j^\top\mathbf{D}_j^\top)_{nd}\mathbf{D}_j\mathbf{X}_j\right) \\ &= 2\mathbf{D}_i\left(\sum_{(i,j)\in\mathcal{E}_1} \mathbf{X}_i\mathbf{X}_j^\top\mathbf{X}_j + \sum_{(i,j)\in\mathcal{E}_2} (\mathbf{X}_i\mathbf{X}_j^\top)_{nd}\mathbf{X}_j\right) = \mathbf{D}_i\nabla_{\mathbf{X}_i} f(\mathbf{X}),\end{aligned}$$

where the second equality holds by a similar argument as in the case of $i \in \mathcal{S}$. Consequently, we have

$$\nabla_{\mathbf{X}_i} f(\mathbf{DX}) = \mathbf{D}_i\nabla_{\mathbf{X}_i} f(\mathbf{X}), \quad \forall i \notin \mathcal{S}.$$

Combining the two cases, it follows that

$$\nabla f(\mathbf{DX}) = \mathbf{D}\nabla f(\mathbf{X}).$$

□

**Lemma 5.** *Consider the case $r = 1$. Given an arbitrary point $x_0 \in \mathbb{R}^n$, let $\hat{x} \in \mathbb{R}^n$ denote a point with the property that the gradient flow* (2.5) *initialized at $\mathbf{x}_0$ converges to $\hat{\mathbf{x}}$. For every diagonal matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ that satisfies*

$$\mathbf{D}_{ii}^2 = 1, \quad \forall i \in [n],$$

*the gradient flow initialized at $\mathbf{Dx}_0$ will converge to $\mathbf{D}\hat{\mathbf{x}}$.*

*Proof.* By the results of Lemma 4, we obtain

$$\nabla f(\mathbf{Dx}) = \mathbf{D}\nabla f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^n. \tag{2.14}$$

Hence, we know that the gradient flow initialized with $\mathbf{Dx}_0$ is equal to $\mathbf{Dx}(t)$ at time $t$, for all $t \geq 0$. This leads to the conclusion that the new gradient flow will converge to $\mathbf{D}\hat{\mathbf{x}}$. □

*Proof of Lemma 2.* Since it is already known that the problem (2.1) has exponentially many global solutions, it remains to prove that the gradient flow with a radial random initialization will converge to one of the above global solutions with equal probability, i.e., with probability $2^{-\lceil n/2\rceil}$. It has been proved in [67] that the gradient flow will only converge to local minima if the objective function does not have degenerate saddle points, i.e., the Hessian of every saddle point has a negative curvature. Since the global solutions of the problem (2.1) are symmetric with respect to a radial probability distribution, it follows from Lemma 5 that we only need to prove that the objective function of this problem does not have degenerate saddle points. Equivalently, we prove that all second-order critical points are global minima.

Suppose that $\hat{\mathbf{x}}$ is a second-order critical point of the problem (2.1) that is not a global minimizer. Let $\mathbf{M} := \mathbf{x}\mathbf{x}^\top$. Due to the symmetry of the landscape, we can assume, without loss of generality, that

$$\hat{x}_k \geq 0, \quad k = 1, \ldots, n.$$

We define the direction $\Delta \in \mathbb{R}^n$ as

$$\Delta_{2k+1} = \hat{x}_{2k+1} - 1, \quad k = 1, \ldots, \lceil n/2 \rceil, \quad \Delta_{2k} = \hat{x}_{2k}, \quad k = 1, \ldots, \lfloor n/2 \rfloor.$$

Then, Lemma 7 in [47] implies that

$$\begin{aligned}\Delta : \nabla^2 f\left[\hat{\mathbf{x}}\right] : \Delta =& \Delta\Delta^\top : \nabla^2 g\left[\left(\hat{\mathbf{x}}\hat{\mathbf{x}}^\top - \mathbf{x}^*\mathbf{x}^*\right)_\Omega\right] : \Delta\Delta^\top \\ &- 3(\mathbf{M} - \mathbf{M}^*) : \nabla^2 g\left[\left(\hat{\mathbf{x}}\hat{\mathbf{x}}^\top - \mathbf{x}^*\mathbf{x}^*\right)_\Omega\right] : (\mathbf{M} - \mathbf{M}^*) \\ \leq& (1+\delta)\| \left(\Delta\Delta^\top\right)_\Omega \|_F^2 - 3(1-\delta)\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F^2,\end{aligned}$$

where the last inequality is from the assumption that $g(\cdot)$ satisfies the sparse RIP condition with $(\delta, 1)$. Combining with the second-order necessary optimality condition, we obtain

$$(1+\delta)\| \left(\Delta\Delta^\top\right)_\Omega \|_F^2 \geq 3(1-\delta)\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F^2. \tag{2.15}$$

Using the expression

$$\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F^2 - \| \left(\Delta\Delta^\top\right)_\Omega \|_F^2 = \sum_{k=1}^{\lceil n/2 \rceil} \left[4\hat{x}_{2k+1} \cdot (\hat{x}_{2k+1} - 1)^2 + \|\hat{\mathbf{x}}^e\|^2(2\hat{x}_{2k+1} - 1)\right],$$

the inequality (2.15) can be written as

$$-\sum_{k=1}^{\lceil n/2 \rceil} \left[4\hat{x}_{2k+1} \cdot (\hat{x}_{2k+1} - 1)^2 + \|\hat{\mathbf{x}}^e\|^2(2\hat{x}_{2k+1} - 1)\right] \geq \frac{2-4\delta}{1+\delta}\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F^2. \tag{2.16}$$

The above inequality shows that

$$\begin{aligned}\frac{2-4\delta}{1+\delta}\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F^2 &\leq -\sum_{k=1}^{\lceil n/2 \rceil} \left[4\hat{x}_{2k+1} \cdot (\hat{x}_{2k+1} - 1)^2 + \|\hat{\mathbf{x}}^e\|^2(2\hat{x}_{2k+1} - 1)\right] \\ &\leq -\sum_{k=1}^{\lceil n/2 \rceil} \left[0 - \|\hat{\mathbf{x}}^e\|^2\right] = \lceil n/2 \rceil \cdot \|\hat{\mathbf{x}}^e\|^2 \leq (n+1)/2 \cdot \|\hat{\mathbf{x}}^e\|^2 \\ &\leq (n+1)/2 \cdot \|\hat{\mathbf{x}}^e\|\|\hat{\mathbf{x}}\| \leq \sqrt{2}(n+1)\delta\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F,\end{aligned}$$

where the second inequality is from the assumption that $\hat{x}_{2k+1} \geq 0$ and the second last inequality is from Lemma 3. The above inequality implies that

$$\| \left(\mathbf{M} - \mathbf{M}^*\right)_\Omega \|_F \leq \frac{\sqrt{2}(n+1)\delta(1+\delta)}{2-4\delta}.$$

Recalling the condition

$$n \geq 3, \quad \delta \leq \frac{1}{2n},$$

we obtain

$$\| (\mathbf{M} - \mathbf{M}^*)_\Omega \|_F \leq \frac{1}{2}.$$

Checking the diagonal entries of $\mathbf{M} - \mathbf{M}^* = \hat{\mathbf{x}}\hat{\mathbf{x}}^\top - \mathbf{x}^*(\mathbf{x}^*)^\top$, we have

$$|\hat{x}_{2k+1} - 1| \leq \frac{1}{2}, \quad k = 1, \ldots, \lceil n/2 \rceil,$$

which gives

$$\hat{x}_{2k+1} \geq \frac{1}{2}, \quad k = 1, \ldots, \lceil n/2 \rceil.$$

Applying this condition to inequality (2.16), the left-hand side of the inequality is non-positive while the right-hand side is non-negative, which implies that

$$\| (\mathbf{M} - \mathbf{M}^*)_\Omega \|_F = 0.$$

This contradicts the assumption that $\hat{\mathbf{x}}$ is not a global solution. Hence, we have completed the proof that all second-order critical points of the problem (2.1) are global minima.

Furthermore, using Lemma 5, we know that each global minimum's region of attraction (ROA) is symmetrical. Since the randomly initialized gradient flow converges to a second-order critical point with probability 1 and all second-order critical points are global minima, the gradient flow with a radial random initialization will converge to each global minimum with equal probability. □

# 2.I Proof of Theorem 4

*Proof.* In the unperturbed case, the curvature of the Hessian at each global minimum is given by

$$\Delta : \nabla^2 f(\mathbf{x}) : \Delta \geq (1 - \delta) \| \left( \mathbf{x}\Delta^\top + \Delta\mathbf{x}^\top \right)_\Omega \|_F^2, \quad \forall \Delta \in \mathbb{R}^n, \, \mathbf{x} \in \mathcal{X}^*,$$

which has been proved to be positive in Section 2.1. Therefore, we know that the Hessian at each global solution is positive definite, and global minima are asymptotically stable for the gradient flow. We choose $R > 0$ to be a large enough constant such that

$$\mathbb{P} \left[ \|\mathbf{x}_0\| \leq R \right] \geq 1 - 2^{-\lceil n/2 \rceil},$$

where the probability is chosen with respect to the initialization distribution. We consider the level set

$$\mathcal{L}_R := \{ \mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \leq c_R \},$$

where $c_R := \max\{ f(\mathbf{x}) \mid \|\mathbf{x}\| \leq R \}$. Since the function $f(\mathbf{x})$ is continuous and coercive, the level set $\mathcal{L}_R$ is compact, and the gradient flow will not leave $\mathcal{L}_R$ if it is initialized inside it. In addition, it holds that

$$\mathbb{P} \left[ \mathbf{x}_0 \in \mathcal{L}_R \right] \geq 1 - 2^{-\lceil n/2 \rceil}.$$

Conditioning on the event that $\mathbf{x}_0 \in \mathcal{L}_R$, Lemma 2 implies that the gradient flow will converge to each global minimum with the same probability. Let $\hat{\mathbf{x}} \in \mathcal{X}^*$ be an arbitrary global minimum and $\mathcal{R}_{\hat{\mathbf{x}}}$ be its ROA of the gradient flow on $\mathcal{L}_R$. Therefore, Lemma 2 implies that

$$\mathbb{P}\left[\mathbf{x}_0 \in \mathcal{R}_{\hat{\mathbf{x}}} \mid \mathbf{x}_0 \in \mathcal{L}_R\right] = 2^{-\lceil n/2 \rceil}.$$

By Theorem 4.17 in [60], there exist a smooth positive definite function $V(\mathbf{x})$ and a continuous positive definite function $W(\mathbf{x})$ such that every level set of $V(\mathbf{x})$ is compact and

$$\begin{aligned} V(\mathbf{x}) \to +\infty, \quad \forall \mathbf{x} \to \partial\mathcal{R}_{\hat{\mathbf{x}}}, \\ \left\langle \frac{dV(\mathbf{x})}{d\mathbf{x}}, -\nabla_{\mathbf{x}} f(\mathbf{x}) \right\rangle \leq -W(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{R}_{\hat{\mathbf{x}}}, \end{aligned}$$

where $\mathbf{x} \to \partial\mathcal{R}_{\hat{\mathbf{x}}}$ means that the distance between $\mathbf{x}$ and $\partial\mathcal{R}_{\hat{\mathbf{x}}}$ goes to zero, and $\partial\mathcal{R}_{\hat{\mathbf{x}}}$ denotes the boundary of the region of attraction of the solution $\hat{\mathbf{x}}$. We choose a large enough constant $M$ such that

$$\mathbb{P}\left[\mathbf{x}_0 \in \mathcal{V}_M \mid \mathbf{x}_0 \in \mathcal{R}_{\hat{\mathbf{x}}}\right] \geq 1 - 2^{-\lceil n/2 \rceil},$$

where we define the level set $\mathcal{V}_M := \{\mathbf{x} \in \mathbb{R}^n \mid V(\mathbf{x}) \leq M\}$. Since the level set $\mathcal{V}_M$ is compact, there exists a small enough constant $\epsilon_0$ such that

$$W(\mathbf{x}) \geq \epsilon_0, \quad \forall \mathbf{x} \in \mathcal{V}_M.$$

Now, we consider the perturbed case. We denote the new objective function as $\tilde{f}(\mathbf{x}; \eta)$, where $\eta \in \mathbb{R}$ is the perturbation to the global solution. More explicitly, the objective function is defined as

$$\tilde{f}(\mathbf{x}; \eta) := \left\| \left(\mathbf{x}\mathbf{x}^\top - (\mathbf{x}^* + \eta)(\mathbf{x}^* + \eta)^\top\right)_\Omega \right\|_F^2.$$

It has been proved in Section 2.1 that the global minimum of the perturbed problem is unique up to a sign flip if the perturbation is sufficiently small and generic and that there exist $2^{\lceil n/2 \rceil} - 2$ spurious local minima. Since the gradient of $\tilde{f}(\mathbf{x}; \eta)$ is a uniformly continuous function of $\eta$ on the compact set $\mathcal{V}_M$, there exists a small enough $r > 0$ such that for any generic $\eta_0$ satisfying $\|\eta_0\| \leq r$, it holds that

$$\left\langle \frac{dV(\mathbf{x})}{d\mathbf{x}}, -\nabla \tilde{f}(\mathbf{x}; \eta_0) \right\rangle \leq -\epsilon_0/2 < 0, \quad \forall \mathbf{x} \in \mathcal{V}_M.$$

This implies that the gradient flow on the perturbed problem will not leave $\mathcal{V}_M$ and will converge to a local minimum inside $\mathcal{V}_M$ if $\mathbf{x}_0 \in \mathcal{V}_M$. Therefore, if the initial point $\mathbf{x}_0$ is initialized with the given distribution, we have

$$\mathbb{P}\left[\lim_{t \to +\infty} \mathbf{x}(t) \in \mathcal{V}_M\right] \geq 2^{-\lceil n/2 \rceil} \left(1 - 2^{-\lceil n/2 \rceil}\right)^2 \geq 2^{-\lceil n/2 \rceil} \left(1 - 2^{-\lceil n/2 \rceil + 1}\right).$$

By choosing $r$ to be the minimum over all the points $\hat{\mathbf{x}} \in \mathcal{X}^*$, the gradient flow on the perturbed problem will converge to a spurious minimum with probability at least

$$\left(2^{\lceil n/2 \rceil} - 2\right) \cdot 2^{-\lceil n/2 \rceil} \left(1 - 2^{-\lceil n/2 \rceil + 1}\right) = 1 - \mathcal{O}\left(2^{-\lceil n/2 \rceil}\right).$$

Thus, we can conclude that the gradient flow on the perturbed problem will fail with probability at least $1 - \mathcal{O}(2^{-\lceil n/2 \rceil})$.

□

# Chapter 3

# Comparison of Solution Methods for Matrix Sensing

Low-rank matrix recovery problems have ubiquitous applications in machine learning and data analytics, including collaborative filtering [64], phase retrieval [19, 103, 15, 99], motion detection [42], and power system state estimation [57, 133, 58]. This problem is formally defined as follows: Given a measurement operator $\mathcal{A}(\cdot) : \mathbb{R}^{m\times n} \mapsto \mathbb{R}^d$ returning a $d$-dimensional measurement vector $\mathcal{A}(\mathbf{M}^*)$ from a low-rank ground truth matrix $\mathbf{M}^* \in \mathbb{R}^{m\times n}$ with rank $r$, the goal is to obtain a matrix with rank less than equal to $r$ that conforms with the measurements, preferably the ground truth matrix $\mathbf{M}^*$. This problem can be stated as a feasibility problem.

$$\begin{aligned} \text{find} \quad & \mathbf{M} \in \mathbb{R}^{m\times n} \\ \text{s.t.} \quad & \mathcal{A}(\mathbf{M}) = \mathcal{A}(\mathbf{M}^*) \\ & \text{rank}(\mathbf{M}) \leq r. \end{aligned} \tag{3.1}$$

While the measurement operator $\mathcal{A}$ can be non-linear as in the case of one-bit matrix sensing [31] and phase retrieval [99], matrix sensing and matrix completion that are widely studied have linear measurement operators [23, 91]. We focus on the matrix sensing and matrix completion problems throughout this chapter. Despite the linearity of $\mathcal{A}$, there are two types of problems depending on the structure of the ground truth matrix $\mathbf{M}^*$. The first type, symmetric problem, consists of a low-rank positive semidefinite ground truth matrix $\mathbf{M}^* \in \mathbb{R}^{n\times n}$, whereas the second type, asymmetric problem, consists of a ground truth matrix $\mathbf{M}^* \in \mathbb{R}^{m\times n}$ that is possibly sign indefinite and non-square. Since each asymmetric problem can be converted to an equivalent symmetric problem [125], we study only the symmetric problem in this chapter.

The matrix sensing and completion problems have linear measurements; hence, the first constraint in problem (3.1) is linear. Therefore, the only non-convexity of the problem arises from the non-convex rank constraint. Earlier works on these problems focused on their convex relaxations by penalizing high-rank solutions [23, 91, 24]. They utilized the nuclear norm of a matrix as the convex surrogate of the rank function. This led to semidefinite programming (SDP) relaxations, which solve the original non-convex problems exactly with high probability based on some assumptions on the linear measurement operator and the ground truth matrix, such as the Restricted

Isometry Property (RIP) and incoherence conditions. High computational time and storage requirements of the SDP algorithms incentivized the implementation of the B-M factorization approach [17]. This approach factorizes the symmetric matrix variable $\mathbf{M} \in \mathbb{R}^{n\times n}$ as $\mathbf{M} = \mathbf{X}\mathbf{X}^\top$ for some matrix $\mathbf{X} \in \mathbb{R}^{n\times r}$, which obviates imposing the positive semidefiniteness and rank constraints. Although the dimension of the decision variable reduces dramatically when $r$ is small, the problem is still non-convex since its objective function is non-convex in terms of the factorized $\mathbf{X}$.

## 3.1 Problem Formulation

Formally, the SDP formulation of the matrix sensing problem uses the nuclear norm of the variable, $\|\mathbf{M}\|_*$, to serve as a surrogate of the rank and replaces the rank constraint in (3.1) with an objective to minimize $\|\mathbf{M}\|_*$. Due to the symmetricity and positive semidefiniteness of the variable, the nuclear norm is equivalent to the trace of the matrix variable $\mathbf{M}$. Hence, the SDP formulation can be written as

$$\min_{\mathbf{M}\in\mathbb{R}^{n\times n}} \quad \operatorname{tr}(\mathbf{M}) \qquad \text{s.t. } \mathcal{A}(\mathbf{M}) = b,\ \mathbf{M} \succeq 0, \tag{3.2}$$

where $b = \mathcal{A}(\mathbf{M}^*) = [\langle \mathbf{A}_1, \mathbf{M}^*\rangle, \ldots, \langle \mathbf{A}_d, \mathbf{M}^*\rangle]^\top$ is given and $\{\mathbf{A}_i\}_{i=1}^d \in \mathbb{R}^{n\times n}$ are called sensing matrices. Moreover, the matrix completion problem is a special case of the matrix sensing problem, with each sensing matrix measuring only one entry of $\mathbf{M}^*$. We can represent the measurement operator $\mathcal{A}$ as $\mathcal{A}_\Omega : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ for this special case, which is defined as follows:

$$\mathcal{A}_\Omega(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise,} \end{cases}$$

where $\Omega$ is the set of indices of observed entries. We denote the measurement operator as $\mathbf{M}_\Omega := \mathcal{A}_\Omega(\mathbf{M})$ for simplicity. Besides the SDP formulation, the B-M factorization formulation of the matrix sensing (MS) and matrix completion (MC) problems can be stated as

$$\text{(MS)} \quad \min_{\mathbf{X}\in\mathbb{R}^{n\times r}} \ g\left[\mathcal{A}(\mathbf{X}\mathbf{X}^\top) - b\right], \tag{3.3a}$$

$$\text{(MC)} \quad \min_{\mathbf{X}\in\mathbb{R}^{n\times r}} \ g\left[(\mathbf{X}\mathbf{X}^\top - \mathbf{M}^*)_\Omega\right], \tag{3.3b}$$

where $g(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ is some twice continuously differentiable function such that $\mathbf{0}_{n\times n}$ is its unique minimizer and the Hessian of $g(\cdot)$ is positive definite at $\mathbf{0}_{n\times n}$. These assumptions are satisfied by the common loss functions considered in the literature. The main objective of this chapter is to compare the SDP and B-M methods for the MC and MS problems.

## 3.2 Background and Related Work

It is widely known that the SDP formulation (3.2) can be used to solve the matrix sensing problem if the sensing matrices are sampled independently from a sub-Gaussian distribution and

the number of measurements $d$ is large enough [91, 92]. This is also a sufficient condition for the sensing matrices to satisfy the RIP condition with high probability, which is defined below:

**Definition 5** (RIP)**.** *[23] The linear map $\mathcal{A} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ is said to satisfy $\delta_p$-RIP if there is a constant $\delta_p \in [0, 1)$ such that*

$$(1 - \delta_p)\|\mathbf{M}\|_F^2 \leq \|\mathcal{A}(\mathbf{M})\|^2 \leq (1 + \delta_p)\|\mathbf{M}\|_F^2$$

*holds for all matrices $\mathbf{M} \in \mathbb{R}^{n\times n}$ satisfying $\mathrm{rank}(\mathbf{M}) \leq p$.*

The RIP constant $\delta_p$ represents how similar the linear operator $\mathcal{A}$ is to an isometry, and various upper bounds on $\delta_p$ have been proposed to serve as sufficient conditions for the exact recovery (meaning that one can recover the ground truth $\mathbf{M}^*$ by solving the SDP problem). A few notable ones include $\delta_{4r} < \sqrt{2} - 1$ in [25], $\delta_{5r} < 0.607, \delta_{3r} < 0.472$ in [81], and $\delta_{2r} < 1/2, \delta_{3r} < 1/3$ in [18]. On the other hand, when the sensing matrices are not sampled independently from a sub-Gaussian distribution or when the RIP condition is not met, the SDP formulation may still recover the ground truth matrix with a high probability. This is the case for MC problems for which RIP fails to hold while SDP works as long as entries of observation follow an independent Bernoulli model [23, 24].

However, recent works have shown that if we use the B-M method instead of the SDP approach, we can still recover the ground truth matrix via first-order methods under similar RIP or coherence assumptions in both the matrix sensing and matrix completion cases [47, 12, 86, 132, 135, 129, 124, 13, 50, 134, 130, 125, 76, 77]. A local minimum is called spurious if it is not a global minimum. Namely, the state-of-the-art result states that as long as $\delta_{\tilde{r}+r} < 1/2$ for the matrix sensing problem, there exists no spurious local minima for an over-parametrized B-M formulation and the gradient descent algorithm can recover $\mathbf{M}^*$ exactly [130]. Here, $\tilde{r} \geq r$ is the search rank that we choose manually in the B-M formulation. If we know the value of $r$, we can set $\tilde{r}$ to $r$, making the B-M approach enjoy the same RIP guarantee as the SDP approach. Since the B-M approach enjoys far better scalability, it has become an increasingly popular tool for solving the matrix sensing problem.

Nevertheless, the B-M approach cannot be routinely used without careful consideration since it could fail on easy (from an information-theoretic perspective) instances of the problem as demonstrated in [121], especially in cases when the RIP condition is not satisfied.

Thus, it is essential to compare and contrast the SDP and B-M approaches to discover which is superior to the other. This comparison is timely since specialized sparse SDP algorithms have become more efficient in recent years, making the SDP method more practical than before [128, 123, 122]. In this chapter, we show that the SDP approach is more powerful than the B-M method regarding the RIP measure. We also discover that the B-M method is able to solve certain instances for which the SDP approach fails. This means that none of these techniques is universally better than the other one, and the best method should be chosen based on the nature of the problem.

### Contributions

We provide a comparative analysis between the SDP approach and the B-M method. We first present the advantages of the SDP approach over the B-M method:

1. First, we focus on an important class of MC problems recently studied in [121]. That chapter has shown that even though this class has low information-theoretic complexity, the B-M method would utterly fail, and the probability of success via first-order methods is almost zero. We prove that the SDP method successfully solves this class and, therefore, SDP may not suffer from the unusual behavior of B-M with regard to easy instances of MC. This also implies that the information-theoretic and optimization complexities are expected to be more aligned for SDP than for B-M.

2. We then investigate a class of MS problems found in the recent chapter [126]. Each MS instance belonging to this class satisfies $\delta_2$-RIP with $r = 1$ for some $\delta > 1/2$ such that the B-M formulation leads to $\mathcal{O}((1-\delta)^{-1})$ spurious solutions and this number goes to infinity as $\delta$ approaches $1$. We show that although each instance is extremely non-convex based on the number of spurious solutions, the SDP method successfully solves all of the problems in this class. This implies that, unlike the B-M method, the success of the SDP approach is not directly correlated to the presence of many spurious solutions.

3. The recent paper [125] has shown that the sharpest RIP bound for the success of the B-M method on the MS problem is $1/2$, and this is independent of the rank $r$. This is undesirable since high-rank problems have lower information-theoretic complexity than low-rank problems. We derive a sufficient RIP bound for the SDP method and show that it can increase from $1/2$ to $1$ as the rank $r$ becomes larger. This implies that the SDP approach does not suffer from a significant shortcoming of the B-M method.

Despite the above advantages, we show that the SDP approach is not universally better than the B-M method. To prove this, we identify a class of MC problems with $\mathcal{O}(n)$ observations in the rank-1 case for which B-M works while SDP fails. It is clear from these comparisons that although the B-M approach is known to be more powerful due to its scalability property, the SDP approach enjoys some unique merits and deserves to be revisited, especially in light of the advancements of fast SDP solvers [128, 123, 122]

## 3.3 Advantages of the SDP Approach

### B-M Fails While SDP Succeeds

In this section, we focus on a class of MC instances that was first proposed in [121] for which the B-M factorization fails. We focus on the matrix completion problem since it is the most common special case of the matrix sensing problem that does not satisfy the RIP condition. We will prove that while the B-M approach fails to recover $\mathbf{M}^*$, the SDP approach can provably find $\mathbf{M}^*$.

We will first give an introduction to this class of MC instances. Consider a rank-$r$ ground truth matrix $\mathbf{M}^* \in \mathbb{S}^n_+$ with $r \geq 1$ and $n \geq 2r$. Let $m := n/r$ and assume without the loss of generality that $n$ is divisible by $r$. We decompose the ground truth matrix into blocks of dimension $r \times r$; thus, $\mathbf{M}^*$ is an $m \times m$ block matrix whose block element at the position $(i, j)$ is denoted as $\mathbf{M}^*_{i,j}$ for $i, j \in [m]$. We require some graph-theoretic notions before introducing the underlying class of MC instances.

**Definition 6** (Induced Measurement Set)**.** *Let $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2) = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2)$ be a pair of undirected graphs with the node set $\mathcal{V} = [m]$ and the disjoint edge sets $\mathcal{E}_1, \mathcal{E}_2 \subset [m] \times [m]$, respectively. The induced measurement set $\Omega(\mathcal{G})$ is defined as follows: if $(i, j) \in \mathcal{E}_1$, then the entire block $\mathbf{M}^*_{i,j}$ is observed; if $(i, j) \in \mathcal{E}_2$, then all non-diagonal entries of the block $\mathbf{M}^*_{i,j}$ are observed; otherwise, none of the entries of the block is observed. The graph $\mathcal{G}$ is referred to as the block sparsity graph.*

We represent the general problem (3.1) with the linear measurement operator $\mathcal{A}$ and rank-$r$ ground truth matrix $\mathbf{M}^* \in \mathbb{R}^{n \times n}$ as $\mathcal{P}_{\mathbf{M}^*, \mathcal{A}, n, r}$. If this is a matrix completion problem with the measurement set $\Omega$, then this special case of the same problem is denoted as $\mathcal{P}_{\mathbf{M}^*, \Omega, n, r}$. Based on Definition 6 and this notation, a low-complexity class of MC instances will be introduced below. These instances have a low complexity because graph-theoretical algorithms can solve them in polynomial time in terms of $n$ and $r$.

**Definition 7** (Low-complexity class of MC instances)**.** *Define $\mathcal{L}(\mathcal{G}, n, r)$ to be the class of low-complexity MC instances $\mathcal{P}_{\mathbf{M}^*, \Omega, n, r}$ with the following properties:*

- ***i)*** *The ground truth $\mathbf{M}^* \in \mathbb{S}^n_+$ is rank-r.*
- ***ii)*** *The matrix $\mathbf{M}^*_{i,j} \in \mathbb{R}^{r \times r}$ is rank-r for all $i, j \in [m]$.*
- ***iii)*** *The measurement set $\Omega = \Omega(\mathcal{G})$ is induced by $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2)$, where $\mathcal{G}_1$ is connected, non-bipartite, and its vertices have self-loops.*

The following theorem borrowed from [121] illustrates the failure of the B-M factorization method.

**Theorem 5.** *Consider a maximal independent set $\mathcal{S}(\mathcal{G}_1)$ of $\mathcal{G}_1$ such that the induced subgraph by vertices in $\mathcal{S}$, $\mathcal{G}_2[\mathcal{S}]$, is connected. There exists an instance in $\mathcal{L}(\mathcal{G}, n, r)$ for which the problem* (3.3b) *has at least $2^{r|\mathcal{S}(\mathcal{G}_1)|} - 2^r$ spurious local minima. In addition, the randomly initialized gradient descent algorithm converges to a global minimum with probability at most $\mathcal{O}(2^{-r|\mathcal{S}(\mathcal{G}_1)|})$, while there is a graph-theoretical algorithm that can solve the problem in $\mathcal{O}(n^2/r^2 + nr^2)$ time.*

The proof of Theorem 5 utilizes the Implicit Function Theorem (IFT). Specifically, the work [121] has generated ground truth matrices $\mathbf{M}^*$ for which the B-M method has $2^{r|\mathcal{S}(\mathcal{G}_1)|}$ global solutions and only $2^r$ of them correspond to the correct completion of the $\mathbf{M}^*$. A generic small perturbation of the problem results in a new instance of an MC problem that belongs to the low-complexity class of MC instances. The conditions on $\mathcal{G}_1$ guarantee that the perturbed problem belongs to the low-complexity class, while the conditions on $\mathcal{G}_2$ guarantee that the Hessian of the objective function of the unperturbed problem is positive definite at the global solutions. Since the instances in the low-complexity class are well defined, the new perturbed problem has a unique

completion with $2^r$ possible global solutions for the B-M method. On the other hand, the other stationary points that correspond to global solutions of the unperturbed problem must be spurious local minima of the new instance. This is concluded by using the IFT. The perturbation that yields a new instance in the low-complexity class of the MC problem is achieved by perturbing the ground truth matrix $\mathbf{M}^* = \mathbf{X}^*(\mathbf{X}^*)^\top$ by a small and generic perturbation $\epsilon \in \mathbb{R}^{n\times r}$. The new ground truth matrix is $\mathbf{M}^*(\epsilon) = \mathbf{X}^*(\epsilon)(\mathbf{X}^*(\epsilon))^\top$, where $\mathbf{X}^*(\epsilon)_i = \mathbf{X}_i^* + \epsilon_i$ if $i \in \mathcal{S}(\mathcal{G}_1)$ and $\mathbf{X}^*(\epsilon)_i = \epsilon_i$ otherwise and $\operatorname{rank}(\mathbf{X}_i^*) = \operatorname{rank}(\mathbf{X}_i^* + \epsilon_i) = r, \forall i \in [m]$. A generic perturbation $\epsilon$ does not belong to a measure zero set in $\mathbb{R}^{n\times r}$.

It is desirable to study how the SDP method performs on this low-complexity class of MC instances. We will present the result for a larger class of problems that contains all instances discussed in Theorem 5.

**Theorem 6.** *Given $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2) = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2)$, consider any maximal independent set $\mathcal{S}(\mathcal{G}_1)$. Consider also $\mathbf{M}^*(\epsilon) = \mathbf{X}^*(\epsilon)(\mathbf{X}^*(\epsilon))^\top$ for any arbitrary $\epsilon \in \mathbb{R}^{n\times r}$, where $\mathbf{X}^*(\epsilon)_i = \mathbf{X}_i^* + \epsilon_i$ if $i \in \mathcal{S}(\mathcal{G}_1)$ and $\mathbf{X}^*(\epsilon)_i = \epsilon_i$ otherwise and $\operatorname{rank}(\mathbf{X}_i^*) = \operatorname{rank}(\mathbf{X}_i^* + \epsilon_i) = r, \forall i \in [m]$. The SDP formulation (3.2) with the observation set $\Omega$ induced by $\mathcal{G}_1$ uniquely recovers the ground truth matrix $\mathbf{M}^*(\epsilon)$.*

Note that we do not require $\epsilon$ to be small or have access to partial observations of blocks induced by edges in $\mathcal{G}_2$. Hence, Theorem 6 shows that SDP solves all MC instances introduced in Theorem 1 and beyond. As a result of Theorem 6, the SDP approach is a viable choice for those MC instances for which the preferable and faster B-M factorization method fails to recover the ground truth matrix. Similar to perturbing the ground truth matrix, one can perturb the linear measurement operator of the matrix completion problem $\mathcal{A}_\Omega$ as

$$\mathcal{A}_{\Omega(\epsilon)}(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij}, & \text{if } (i,j) \in \Omega \\ \epsilon \mathbf{M}_{ij}, & \text{otherwise} \end{cases}, \tag{3.4}$$

where $\epsilon > 0$ is a sufficiently small real number [126]. Note that $\mathcal{A}_{\Omega(\epsilon)}$ satisfies the RIP condition with $\delta = (1-\epsilon)/(1+\epsilon)$.

**Theorem 7.** *Suppose that $g$ is the squared loss function, i.e $g(\mathbf{x}) = \|\mathbf{x}\|^2$. Consider the measurement set $\Omega$ defined in Theorem 5. For every sufficiently small $\epsilon > 0$, there exists a low-complexity instance of the MS problem $\mathcal{P}_{\mathbf{M}^*, \mathcal{A}_{\Omega(\epsilon)}, n, r}$ with $\mathcal{O}(2^{r|\mathcal{S}(\mathcal{G}_1)|})$ spurious local minima.*

The proof of the above theorem is similar to the proof of Theorem 5 because the conditions for unperturbed problems are the same, and a different small perturbation to the problem yields a similar number of spurious solutions. Hence, the proof is omitted. The above theorem states that not only MC instances but also MS instances suffer from this undesirable behavior of the B-M factorization approach. The ground truth matrix $\mathbf{M}^*$ is generated as in Theorem 5 to have $2^{r|\mathcal{S}(\mathcal{G}_1)|}$ global solutions for the unperturbed problem. Furthermore, the number of spurious solutions for this scheme can be quantified as $\mathcal{O}((1-\delta)^{-1})$ for $\delta \geq 1/2$ in the rank-1 case [126]. Nevertheless, the SDP formulation approach trivially solves all these undesirable MS instances. This is due to

the fact the perturbed measurement operator $\mathcal{A}_{\Omega(\epsilon)}$ corresponds to observing all the entries. Hence, the feasible set only contains the ground truth matrix $\mathbf{M}^*$.

**Proposition 3.** *Given a measurement set $\Omega$, the SDP formulation* (3.2) *uniquely recovers the rank-$r$ ground truth matrix $\mathbf{M}^*$ for the MS instance $\mathcal{P}_{\mathbf{M}^*,\mathcal{A}_\Omega(\epsilon),n,r}$, where $\mathcal{A}_\Omega(\epsilon)$ is defined in* (3.4) *and $\epsilon$ is an arbitrary non-zero number.*

Hence, unlike the B-M method, the SDP approach successfully solves all the instances in Theorem 7 for which the RIP constant exists (while greater than $1/2$). Consequently, SDP could be the preferred method when sufficient conditions on RIP for exact recovery by the B-M factorization are not met. In the next part, we will provide sharper sufficiency bounds for the SDP approach, which further corroborates its strength.

## Sharper RIP bound for SDP

Since the SDP method is more powerful than the B-M factorization for certain classes of MC and MS problems, as shown in the previous section, and since specialized SDP algorithms can solve large-scale MC and MS problems, it is useful to further study the SDP method through the lens of the well-known RIP notion. We will derive a strong lower bound $\delta_{lb}$ on the RIP constant $\delta$ to guarantee convergence to the ground truth solution by using a proof technique called the inexistence of incorrect solution [129]. We aim to find a linear measurement operator $\mathcal{A}$ with the smallest RIP constant such that the SDP formulation converges to a wrong solution. To do so, we need to solve the optimization problem

$$\begin{aligned} \min_{\delta,\mathcal{A}} \quad & \delta \\ \text{s.t.} \quad & \mathcal{A}(\mathbf{M}) = \mathcal{A}(\mathbf{M}^*) \\ & \operatorname{tr}(\mathbf{M}) \leq \operatorname{tr}(\mathbf{M}^*) \\ & \mathcal{A} \text{ satisfies the } \delta_{2r}\text{-RIP property,} \end{aligned} \tag{3.5}$$

where $\mathbf{M} \neq \mathbf{M}^*$. The condition $tr(\mathbf{M}) \leq tr(\mathbf{M}^*)$ guarantees that SDP cannot uniquely recover $\mathbf{M}^*$. Checking the RIP constant for a linear measurement operator is proven to be NP-hard [105]. Therefore, it is difficult to solve the problem (3.5) analytically. To simplify the problem, we will introduce some notations. We use a matrix representation of the measurement operator $\mathcal{A}$ as follows:

$$\mathbf{A} = [\operatorname{vec}(\mathbf{A}_1), \operatorname{vec}(\mathbf{A}_2), \ldots, \operatorname{vec}(\mathbf{A}_d)]^\top \in \mathbb{R}^{d \times n^2}.$$

Then, $\mathbf{A}\operatorname{vec}(\mathbf{M}) = \mathcal{A}(\mathbf{M})$ for every matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$. We define $\mathbf{H} = \mathbf{A}^\top\mathbf{A}$, which is the matrix representation of the kernel operator $\mathcal{H} = \mathcal{A}^\top\mathcal{A}$ to simplify the last constraint of the problem (3.5).

To derive a RIP bound, we consider the following optimization problem given $\mathbf{M}$ and $\mathbf{M}^*$, where $\mathbf{M}$ is the global solution of (3.2) and $\mathbf{M}^*$ is the ground truth solution:

$$\begin{aligned} \min_{\delta,\mathbf{H}} \quad & \delta \\ \text{s.t.} \quad & \mathbf{e}^\top\mathbf{H}\mathbf{e} = 0 \\ & \mathbf{H} \text{ is symmetric and satisfies the } \delta_{2r}\text{-RIP,} \end{aligned} \tag{3.6}$$

where

$$\mathbf{e} = \mathrm{vec}(\mathbf{M}^* - \mathbf{M}).$$

For this fixed $\mathbf{M}$ and $\mathbf{M}^*$, we assume that $\mathbf{M} \neq \mathbf{M}^*$ and that $\mathrm{rank}(\mathbf{M}^* - \mathbf{M}) > 2r$, since if $\mathrm{rank}(\mathbf{M}^* - \mathbf{M}) \leq 2r$, the relation $\mathbf{M} = \mathbf{M}^*$ holds automatically by definition of $\delta_{2r}$-RIP for any $\delta$ since it implies strong convexity. Denote the optimal value to (3.6) as $\delta(\mathbf{e})$, which is a function of $\mathbf{e}$. It is desirable to find

$$\delta^* := \min_{\mathbf{e}:\mathrm{tr}(\mathbf{M})\leq\mathrm{tr}(\mathbf{M}^*)} \delta(\mathbf{e}).$$

By the logic of in-existence of counterexample, we know that if a problem $\mathbf{H} = \mathbf{A}^\top\mathbf{A}$ has $\delta_{2r}$-RIP with $\delta < \delta^*$, then the solution to (3.2) will be $\mathbf{M}^*$, which is the ground truth solution. However, since the last constraint of (3.6) is non-convex, it is useful to replace it with a surrogate condition that allows solving the problem analytically. The following problem helps to achieve this goal:

$$\begin{aligned} \min_{\delta,\mathbf{H}} \quad & \delta \\ \text{s.t.} \quad & \mathbf{e}^\top\mathbf{H}\mathbf{e} \leq 2\|\mathbf{e}_c\|^2 + 2(l-3)\delta\|\mathbf{e}_c\|^2 \\ & (1-\delta)\mathbf{I}_{n^2} \preceq \mathbf{H} \preceq (1+\delta)\mathbf{I}_{n^2}. \end{aligned} \tag{3.7}$$

Here, $l = \lceil n/r \rceil$ and we define $\{\mathbf{e}_i\}_{i=1}^l$ and $\mathbf{e}_c$ in the following fashion. First, consider the eigen-decomposition of $\mathbf{M}^* - \mathbf{M}$ and assume that the eigenvalues are ordered in terms of their absolute values, namely, $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Let $\mathbf{u}_k$'s denote the corresponding orthonormal eigenvectors:

$$\mathrm{mat}_S(\mathbf{e}) = \mathbf{M}^* - \mathbf{M} = \sum_{k=1}^{n} \lambda_k \mathbf{u}_k \mathbf{u}_k^\top.$$

Then, we define:

$$\mathbf{e}_i = \mathrm{vec}\left( \sum_{k=(i-1)*r+1}^{\min\{i*r,n\}} \lambda_k \mathbf{u}_k \mathbf{u}_k^\top \right),$$

$\mathbf{e}_{2r} = \mathbf{e}_1 + \mathbf{e}_2$, and $\mathbf{e}_c = \sum_{i=3}^{l} \mathbf{e}_i$. The next proposition allows us to replace (3.6) with (3.7) because the optimal value of the (3.7), $\delta_{lb}(\mathbf{e})$, gives a lower bound on $\delta(\mathbf{e})$.

**Proposition 4.** *The optimal objective value of the problem* (3.7)*, $\delta_{lb}(\mathbf{e})$, is always less than or equal to the optimal objective value of the problem* (3.6)*, i.e., $\delta_{lb}(\mathbf{e}) \leq \delta(\mathbf{e})$.*

The proof of this proposition is central to the construction of the sufficiency bound, which is based on using a convex program to serve as an estimate of the non-convex problem. After we extend the $\mathrm{RIP}_{2r}$ constraint in (3.7) to be $\mathrm{RIP}_n$(thus making it convex), it is necessary to somehow preserve the information that the near isometric property of $\mathbf{H}$ should only apply to low-rank matrices. This is achieved by changing the first constraint so that $\mathbf{e}$ does not need to be completely in the null space of $\mathbf{H}$. (3.7) approximately requires that $\mathbf{H}$ only maps a certain low-rank sub-manifold to 0. The full proof can be found in the Appendix. As a result of Proposition 4, it immediately follows that

$$\delta_{\mathrm{lb}} = \min_{\mathbf{e}:\mathrm{tr}(\mathbf{M})\leq\mathrm{tr}(\mathbf{M}^*)} \delta_{\mathrm{lb}}(\mathbf{e}) \leq \delta^*.$$

In fact, we can obtain a lower bound on the value $\delta_{lb}$ by solving the problem (3.7) analytically. The following lemma quantifies a lower bound on $\delta_{lb}$.

**Lemma 6.** *It holds that*

$$\delta_{lb} \geq \frac{2r}{n + (n-2r)(2l-5)}.$$

The best-known sufficiency bound presented in [18] is independent of $n$ and $r$. This sufficiency lower bound on the RIP constant presented in Lemma 6 can be tighter than $1/2$ depending on the size of the problem $n$ and the rank of the ground truth matrix $r$. For instance, the SDP formulation converges to ground truth solution whenever RIP constant $\delta$ is close to 1 as $r \to n/2$. On the other hand, whenever $r/n$ is ratio is small, e.g., rank-1 matrix sensing problem with large $n$, $\delta < 1/2$ is a stronger guarantee for the ground truth matrix recovery. Combined with the $1/2$ sufficiency bound that works for both the symmetric and asymmetric cases [18], we obtain the following result:

**Theorem 8.** *The global solution of the SDP formulation* (3.2) *will be the ground truth matrix* $\mathbf{M}^*$ *if the sensing matrix* $\mathcal{A}$ *satisfies the RIP condition with the RIP constant* $\delta_{2r}$ *satisfying the inequality:*

$$\delta_{2r} < \max\left\{1/2, \frac{2r}{n + (n-2r)(2l-5)}\right\},$$

*where* $l = \lceil n/r \rceil$.

Compared with the existing sufficiency RIP bounds, this new result has a striking advantage. The bound $\delta_{2r} < 1/2$ has already been proven to be the sharpest for the B-M formulation, which is independent of the search rank. In contrast, Theorem 8 shows that the RIP bound for SDP exceeds this bound and approaches 1 as the rank $r$ increases.

In this section, we have shown that as opposed to the popular belief that B-M enjoys very similar RIP guarantees as the SDP approach, there are real benefits to switching to the SDP formulation, making it a more competitive option since specialized SDP solvers are becoming more efficient in recent years. However, we will next provide some problem instances for which the SDP method fails to solve the problem while the B-M method contains no spurious solutions, which balances the desirable properties of the SDP method.

## 3.4 Advantages of the B-M Method

In this section, we give two classes of rank-1 matrix completion problems for which the B-M factorization does not contain any spurious solution while SDP fails to recover its ground truth matrix. Throughout this section, the rank-1 positive semidefinite ground truth matrix $\mathbf{M}^* = \mathbf{x}^*(\mathbf{x}^*)^\top$ is assumed not to contain any zero entries, meaning that $x_i^* \neq 0$ for all $i \in [n]$. Before proceeding with the results, we provide two small examples to highlight the underlying ideas behind the main results.

**Example 3.** *Consider a block sparsity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = 3$ nodes and the edge set $\mathcal{E} = \{(1,1),(1,2),(2,3)\}$. Namely, it is a chain graph with three nodes and a self-loop at the first node. First, we aim to show that only second-order critical points are the global solutions of the B-M factorization method. The B-M factorization formulation* (3.3b) *with the squared loss function can be explicitly written as $\min_{\mathbf{x}\in\mathbb{R}^3} f(\mathbf{x})$, where*

$$f(\mathbf{x}) = \frac{1}{4}\sum_{\substack{(i,j)\in\mathcal{E}\\ i=j}} (x_i^2 - (x_i^*)^2)^2 + \frac{1}{2}\sum_{\substack{(i,j)\in\mathcal{E}\\ i\neq j}} (x_i x_j - x_i^* x_j^*)^2.$$

*The corresponding gradient and Hessian are:*

$$\begin{aligned}
\frac{\partial f(\mathbf{x})}{\partial x_i} &= \sum_{\substack{i,j\in\mathcal{E}\\ i=j}} (x_i^2 - (x_i^*)^2)x_i + \sum_{\substack{i,j\in\mathcal{E}\\ i\neq j}} (x_i x_j - x_i^* x_j^*)x_j,\\
\frac{\partial^2 f(\mathbf{x})}{\partial x_i^2} &= \mathbb{1}[(i,i)\in\mathcal{E}](3x_i^2 - (x_i^*)^2) + \sum_{i,j\in\mathcal{E}} x_j^2,\\
\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} &= \begin{cases} 2x_i x_j - x_i^* x_j^*, & \text{if } i\neq j \text{ and } (i,j)\in\mathcal{E}\\ 0, & \text{otherwise}\end{cases}.
\end{aligned}$$

*Each second-order critical point $\hat{\mathbf{x}}$ must satisfy the conditions $\nabla f(\hat{\mathbf{x}}) = 0$ and $\nabla^2 f(\hat{\mathbf{x}}) \succeq 0$. The third entry of the gradient implies either $\hat{x}_2 = 0$ or $\hat{x}_2\hat{x}_3 = x_2^* x_3^*$. Whenever $\hat{x}_2 = 0$, the Hessian is not positive semidefinite since $[\nabla^2 f(\hat{x})]_{2:3,2:3} \not\succeq 0$. Thus, $\hat{x}_2\hat{x}_3 = x_2^* x_3^*$ must hold. Following this, $\partial f(\hat{\mathbf{x}})/\partial x_2$ implies either $\hat{x}_1 = 0$ or $\hat{x}_1\hat{x}_2 = x_1^* x_2^*$. However, if $\hat{x}_1 = 0$, then $\partial f(\hat{\mathbf{x}})/\partial x_1$ gives $-x_1^* x_2^* \hat{x}_2 = 0$, which implies $\hat{x}_2 = 0$. This contradicts the earlier result. Thus, each second-order critical point must have the following properties:*

$$\hat{x}_1^2 = (x_1^*)^2, \quad \hat{x}_1\hat{x}_2 = x_1^* x_2^*, \quad \hat{x}_2\hat{x}_3 = x_2^* x_3^*.$$

*The solution to this system of equations proves the exact recovery of the ground truth matrix $\mathbf{M}^*$. Hence, the only second-order critical points are the valid factors of the ground truth solution, i.e., $\pm\mathbf{x}^*$.*

*The next step is to demonstrate the failure of the SDP formulation* (3.2) *for some instances of the MC problem with this given block sparsity matrix $\mathcal{G}$. The problem* (3.2) *is equivalent to the optimization*

$$\begin{aligned}
&\min_{\mathbf{M}\in\mathbb{R}^{3\times 3}} \mathbf{M}_{2,2} + \mathbf{M}_{3,3}\\
&\quad s.t \begin{bmatrix} (x_1^*)^2 & x_1^* x_2^* & \mathbf{M}_{1,3}\\ x_1^* x_2^* & \mathbf{M}_{2,2} & x_2^* x_3^*\\ \mathbf{M}_{3,1} & x_2^* x_3^* & \mathbf{M}_{3,3}\end{bmatrix} \succeq 0.
\end{aligned}$$

*Consider a feasible solution* $\hat{\mathbf{M}}$ *with* $\hat{\mathbf{M}}_{2,2} = \hat{\mathbf{M}}_{3,3} = |x_2^* x_3^*|$ *and* $\hat{\mathbf{M}}_{1,3} = \hat{\mathbf{M}}_{3,1} = |x_1^* x_2^*|$. *Note that* $\hat{\mathbf{M}}$ *is feasible whenever* $|x_3^*| \geq |x_2^*|$. *While the objective value of the ground truth matrix* $\mathbf{M}^*$ *is* $(x_2^*)^2 + (x_3^*)^2$, *the objective value of the feasible solution* $\hat{\mathbf{M}}$ *is* $2|x_2^* x_3^*|$. *Under the assumption* $|x_3^*| > |x_2^*|$, *the feasible solution* $\hat{\mathbf{M}}$ *is strictly better than the ground truth solution. Thus, SDP fails to recover the ground truth matrix.*

This example demonstrates the MC instances for which the B-M method successfully converges to the ground truth solution while the SDP fails to find the solution. One reason is that the number of measurements is $\mathcal{O}(n)$ in this example, which is the minimum threshold for exact completion. However, the statistical guarantees on SDP often need more observations. We can generalize Example 1 to any chain graph with $n$ nodes and a single self-loop at one of the ends.

**Theorem 9.** *Consider the MC problem with a rank-1 positive definite ground truth matrix* $\mathbf{M}^* \in \mathbb{R}^{n \times n}$ *that can be factorized as* $\mathbf{M}^* = \mathbf{x}^* (\mathbf{x}^*)^\top$ *with* $x_i^* \neq 0, \forall i \in [n]$. *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *be a block sparsity graph with* $|\mathcal{V}| = n$ *and* $\mathcal{E} = \{(1,1), (1,2), (2,3), \cdots, (n-1, n)\}$. *Then, the B-M method* (3.3b) *does not contain any spurious solutions.*

The proof of Theorem 10 needs a careful treatment of the second-order optimality conditions and is deferred to the appendix. In addition to the success of the B-M factorization, the following result establishes the failure of the SDP for the instances described in the above theorem.

**Theorem 10.** *Consider the ground truth matrix* $\mathbf{M}^* \in \mathbb{R}^{n \times n}$ *satisfying the conditions in Theorem 10. Suppose that there exist two indices* $j, k$ *such that* $x_k^* > x_j^*$ *and* $j, k > 2$. *Then, the SDP problem* (3.2) *fails to recover the ground truth matrix.*

As mentioned before, SDP fails due to a lack of observations on the diagonal entries of the ground truth matrix. Note that the RIP condition is not satisfied since these are MC problems. As a result, whenever we do not have sufficient guarantees on linear measurement operator, none of the methods are superior to the other one in terms of exact recovery. The following example identifies another class of problem instances that corroborates these findings.

**Example 4.** *Consider a block sparsity graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *with* $|\mathcal{V}| = 3$ *nodes and the edge set* $\mathcal{E} = \{(1,2), (2,3), (3,1)\}$. *Namely, it is a simple cycle with three nodes. The B-M factorization formulation* (3.3b) *with the squared loss function can be written the same as in Example 1. Firstly, we can show that each second-order critical point* $\hat{\mathbf{x}}$ *only has non-zero entries, i.e.,* $\hat{x}_i \neq 0$ *for all* $i \in [n]$. *Without loss of generality, suppose by contradiction that* $\hat{x}_1 = 0$. *In order for the stationarity condition to hold, either* $\hat{x}_2 \hat{x}_3 = x_2^* x_3^*$ *or* $\hat{x}_2 = \hat{x}_3 = 0$ *should be satisfied. The latter implies* $\hat{\mathbf{x}} = 0$ *and* $\nabla^2 f(\hat{\mathbf{x}}) \not\succeq 0$ *in that case. Thus,* $\hat{x}_2 \hat{x}_3 = x_2^* x_3^*$ *must hold. Following this,* $\partial f(\hat{\mathbf{x}}) / \partial x_1$ *yields* $-x_1^* x_2^* \hat{x}_2 - x_1^* x_3^* \hat{x}_3 = 0$. *Combining these two equations yields* $(\hat{x}_3)^2 = -(x_2^*)^2$, *which does not have any real solution. As a result, each second-order critical point* $\hat{\mathbf{x}}$ *must have only non-zero entries.*

*By the condition* $\nabla f(\hat{\mathbf{x}}) = 0$, *whenever* $\hat{x}_i \hat{x}_j = x_i^* x_j^*$ *for some* $(i,j) \in \mathcal{E}$, *then* $\hat{x}_i \hat{x}_j = x_i^* x_j^*$ *holds for every* $(i,j) \in \mathcal{E}$. *This system of equations yields the ground truth solution. Accordingly,*

*a spurious solution* $\hat{\mathbf{x}}$ *must have the following characteristics:* $\hat{x}_i\hat{x}_j \neq x_i^*x_j^*, \forall (i,j) \in \mathcal{E}$ *and* $\hat{x}_i \neq 0, \forall i \in \{1,2,3\}$. *Define* $a_{i,j} = x_ix_j - x_i^*x_j^*$. *Then, the stationarity condition becomes*

$$\nabla f(\hat{\mathbf{x}}) = \begin{bmatrix} \hat{a}_{1,2}\hat{x}_2 + \hat{a}_{1,3}\hat{x}_3 \\ \hat{a}_{1,2}\hat{x}_1 + \hat{a}_{2,3}\hat{x}_3 \\ \hat{a}_{1,3}\hat{x}_1 + \hat{a}_{2,3}\hat{x}_2 \end{bmatrix} = 0.$$

*Multiplying the first entry of the gradient by* $\hat{x}_1/\hat{x}_2$ *and substituting with the second entry gives* $\hat{a}_{1,2}\hat{x}_1 = -\hat{a}_{2,3}\hat{x}_3$. *Successively, substituting this to the third entry of the gradient results in*

$$\hat{x}_3(\hat{a}_{1,3}\hat{x}_1 - \hat{a}_{2,3}\hat{x}_2) = 0.$$

*Because we search for a solution with* $\hat{x}_i \neq 0$, *we must have* $\hat{a}_{1,3}\hat{x}_1 - \hat{a}_{2,3}\hat{x}_2 = 0$. *This condition, combined with the last entry of the gradient, results in the condition* $\hat{a}_{1,3}\hat{x}_1 = 0$, *which is a contradiction. As a result, all the second-order critical points are global solutions that yield the ground truth matrix completion.*

*Our next goal is to show that the SDP formulation* (3.2) *fails for this class of instances of MC instances. Note that the SDP formulation of the matrix completion problem considered in this example is equivalent to the formulation:*

$$\begin{aligned} \min_{\mathbf{M}\in\mathbb{R}^{3\times 3}} \; & \mathbf{M}_{1,1} + \mathbf{M}_{2,2} + \mathbf{M}_{3,3} \\ \textit{s.t} \; & \begin{bmatrix} \mathbf{M}_{1,1} & x_1^*x_2^* & x_1^*x_3^* \\ x_1^*x_2^* & \mathbf{M}_{2,2} & x_2^*x_3^* \\ x_1^*x_3^* & x_2^*x_3^* & \mathbf{M}_{3,3} \end{bmatrix} \succeq 0. \end{aligned}$$

*Without loss of generality, assume that* $x_1^* \leq x_2^* \leq x_3^*$ *by the symmetry of the problem. Consider a feasible rank-2 solution* $\hat{\mathbf{M}}$ *given as* $\hat{\mathbf{M}}_{1,1} = x_1^*(x_3^* - x_2^*), \hat{\mathbf{M}}_{2,2} = x_2^*(x_3^* - x_1^*)$ *and* $\mathbf{M}_{3,3} = x_3^*(x_1^* + x_2^*)$. *Note that* $\hat{\mathbf{M}}$ *is feasible whenever* $x_3^* \geq x_1^* + x_2^*$. *The objective value of the feasible solution* $\hat{\mathbf{M}}$ *is* $-2x_1^*x_2^* + 2x_2^*x_3^* + 2x_1^*x_3^*$, *whereas the objective of the ground truth solution* $\mathbf{M}^*$ *is* $(x_1^*)^2 + (x_2^*)^2 + (x_3^*)^2$. *The feasible solution* $\hat{\mathbf{M}}$ *is strictly better than the ground truth solution if* $x_3^* > x_1^* + x_2^*$. *Hence, the SDP cannot recover the ground truth solution for all the instances with a simple cycle block sparsity graph.*

Similar to Example 1, SDP fails in this example due to a lack of diagonal observations. Next, we can generalize this instance to any simple cycle block sparsity graph with an odd number of vertices.

**Theorem 11.** *Consider the matrix completion problem with a rank-1 positive definite ground truth matrix* $\mathbf{M}^* \in \mathbb{R}^{n\times n}$ *that can be factorized as* $\mathbf{M}^* = \mathbf{x}^*(\mathbf{x}^*)^\top$ *with* $x_i^* \neq 0, \forall i \in [n]$. *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *be a block sparsity graph with* $|\mathcal{V}| = |\mathcal{E}| = n = 2k+1$ *and* $\mathcal{E} = \{(0,1),(1,2),\dots(2k-1,2k),(2k,0)\}$. *Then, the B-M factorization problem* (3.3b) *does not contain any spurious solutions.*

**Theorem 12.** *Consider the ground truth matrix* $\mathbf{M}^* \in \mathbb{R}^{n\times n}$ *satisfying the conditions in Theorem 9. Suppose that the condition* $\sum_{t=1}^{k}(x^*_{2t-1})^2 > \sum_{t=0}^{k}(x^*_{2t})^2$ *holds for some chosen node* $0$*. Then, the SDP problem* (3.2) *fails to recover the ground truth matrix.*

Note that we can choose any node as node $0$ due to the symmetry of the problem. Therefore, the condition stated in Theorem 12 is not restrictive because this condition suffices to hold for a chosen node $0$ among $2k+1$ ones. As a result of the above theorems, the B-M factorization can outperform the convex relaxation approach. A critical extension of the work presented in this section would be finding subgraphs for which the B-M method is successful while the SDP is unsuccessful and then attaching these subgraphs to generate larger block sparsity graphs. It is known that SDP will fail in these instances, and it is intriguing to investigate the behavior of the B-M method in those instances.

# Appendices

## 3.A Proof of Theorem 6

*Proof.* Let $\mathbf{M}^* \in \mathbb{R}^{n\times n}$ be the rank-r unperturbed ground truth $m\times m$ block matrix with each block having dimension $r\times r$. Hence, the ground truth matrix can be factorized as $\mathbf{M}^* = \mathbf{X}^*(\mathbf{X}^*)^\top, \mathbf{X}^* \in \mathbb{R}^{n\times r}$. Each square factor $\mathbf{X}_i^* \in \mathbb{R}^{r\times r}$ is rank-r if $i \in \mathcal{S}_1$ and is $\mathbf{0}_{r\times r}$ otherwise. We perturb the ground truth matrix by $\epsilon \in \mathbb{R}^{n\times r}$, where $\mathbf{M}^*(\epsilon) = (\mathbf{X}^* + \epsilon)(\mathbf{X}^* + \epsilon)^\top$ such that $\mathbf{X}^*(\epsilon)_i = \mathbf{X}_i^* + \epsilon_i$ if $i \in \mathcal{S}_1$ and $\mathbf{X}^*(\epsilon)_i = \epsilon_i$ otherwise. Here, we assume that $\mathrm{rank}(\mathbf{X}_i^* + \epsilon_i) = r, \forall i \in [m]$.

If all diagonal blocks are observed, then it will reduce to a feasibility problem, and we can skip the following procedure. Since $\mathcal{S}_1$ is maximal independent set, there exists two indices $i \in \mathcal{S}_1$ and $j \notin \mathcal{S}_1$ with $(j,j) \notin \mathcal{E}$ such that $(i,j) \in \mathcal{E}_1$. Consider the $2\times 2$ block sub-matrix with $i$-th and $j$-th block columns and rows:

$$\begin{bmatrix} \mathbf{M}_{i,i} & \mathbf{M}_{i,j} \\ \mathbf{M}_{j,i} & \mathbf{M}_{j,j} \end{bmatrix} = \begin{bmatrix} (\mathbf{X}_i^* + \epsilon_i)(\mathbf{X}_i^* + \epsilon_i)^\top & (\mathbf{X}_i^* + \epsilon_i)\epsilon_j^\top \\ \epsilon_j(\mathbf{X}_i^* + \epsilon_i)^\top & \mathbf{M}_{j,j} \end{bmatrix} \succeq 0.$$

The equality holds because the blocks $(i,i)$ and $(i,j)$ are in $\mathcal{E}_1$. By the Schur complement argument and since $\mathbf{X}_i^* + \epsilon_i$ is full-rank, the above constraint is equivalent to

$$\mathbf{M}_{j,j} \succeq \epsilon_j\epsilon_j^\top.$$

The unique trace minimizer for the diagonal blocks is $\mathbf{M}_{j,j} = \epsilon_j\epsilon_j^\top$. By the same argument, this must hold for every $j$ that is not in the independent set without a self-loop. Thus, the objective value cannot be less than $\sum_{i=1}^n \mathrm{tr}(\mathbf{X}^*(\epsilon)_i(\mathbf{X}^*(\epsilon)_i)^\top)$. Therefore, the minimum value is achieved whenever $\mathbf{M}_{i,i}^* = \mathbf{X}^*(\epsilon)_i(\mathbf{X}^*(\epsilon))^\top$. This makes $\mathbf{M}(\epsilon)^*$ an optimal solution.

We now prove the uniqueness of the solution. Since the graph is connected, a node $k$ exists adjacent to the node $j$, such that the edges $(i,j)$ and $(j,k)$ exist in the graph. Consider the $3\times 3$ block submatrix

$$\begin{bmatrix} \mathbf{M}_{i,i} & \mathbf{M}_{i,j} & \mathbf{M}_{i,k} \\ \mathbf{M}_{j,i} & \mathbf{M}_{j,j} & \mathbf{M}_{j,k} \\ \mathbf{M}_{k,i} & \mathbf{M}_{k,j} & \mathbf{M}_{k,k} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^*(\epsilon)_i\mathbf{X}^*(\epsilon)_i^\top & \mathbf{X}^*(\epsilon)_i\mathbf{X}^*(\epsilon)_j^\top & \mathbf{M}_{i,k} \\ \mathbf{X}^*(\epsilon)_j\mathbf{X}^*(\epsilon)_i^\top & \mathbf{X}^*(\epsilon)_j\mathbf{X}^*(\epsilon)_j^\top & \mathbf{X}^*(\epsilon)_j\mathbf{X}^*(\epsilon)_k^\top \\ \mathbf{M}_{k,i} & \mathbf{X}^*(\epsilon)_k\mathbf{X}^*(\epsilon)_j^\top & \mathbf{X}^*(\epsilon)_k\mathbf{X}^*(\epsilon)_k^\top \end{bmatrix} \succeq 0.$$

This is equivalent to following constraints by Schur's complement

$$\begin{bmatrix} \mathbf{X}^*(\epsilon)_i\mathbf{X}^*(\epsilon)_i^\top & \mathbf{X}^*(\epsilon)_i\mathbf{X}^*(\epsilon)_j^\top \\ \mathbf{X}^*(\epsilon)_j\mathbf{X}^*(\epsilon)_i^\top & \mathbf{X}^*(\epsilon)_j\mathbf{X}^*(\epsilon)_j^\top \end{bmatrix} - \begin{bmatrix} \mathbf{M}_{i,k}\mathbf{X}^*(\epsilon)_k^{-T}\mathbf{X}^*(\epsilon)_k^{-1}\mathbf{M}_{k,i} & \mathbf{M}_{i,k}\mathbf{X}^*(\epsilon)_k^{-T}\mathbf{X}^*(\epsilon)_j^T \\ \mathbf{X}^*(\epsilon)_j\mathbf{X}^*(\epsilon)_k^{-1}\mathbf{M}_{k,i} & \mathbf{X}^*(\epsilon)_j\mathbf{X}^*(\epsilon)_j^T \end{bmatrix} \succeq 0,$$

$$\begin{bmatrix} \mathbf{X}^*(\epsilon)_i\mathbf{X}^*(\epsilon)_i^\top - \mathbf{M}_{i,k}\mathbf{X}^*(\epsilon)_k^{-T}\mathbf{X}^*(\epsilon)_k^{-1}\mathbf{M}_{k,i} & (\mathbf{X}^*(\epsilon)_i - \mathbf{M}_{i,k}\mathbf{X}^*(\epsilon)_k^{-T})\mathbf{X}^*(\epsilon)_j^\top \\ \mathbf{X}^*(\epsilon)_j(\mathbf{X}^*(\epsilon)_i - \mathbf{M}_{i,k}\mathbf{X}^*(\epsilon)_k^{-T})^\top & 0 \end{bmatrix} \succeq 0.$$

Another Schur complement argument gives

$$(\mathbf{X}^*(\epsilon)_i - \mathbf{M}_{i,k}\mathbf{X}^*(\epsilon)_k^{-T})\mathbf{X}^*(\epsilon)_j^\top = 0.$$

Since $\mathbf{X}^*(\epsilon)_j^\top$ is full rank, we have $\mathbf{X}^*(\epsilon)_i - \mathbf{M}_{i,k}\mathbf{X}^*(\epsilon)_k^{-T} = 0$. Thus, we obtain $\mathbf{M}_{i,k} = \mathbf{X}^*(\epsilon)_i\mathbf{X}^*(\epsilon)_k^\top$. Note that filling out the unobserved non-diagonal blocks is equivalent to adding the edge $(i,k)$ to the graph $\mathcal{G}_1$. Hence, we can always find such triple $(i,j,k)$ defined above until all missing entries are filled out. As a result, we obtain the unique solution $\mathbf{M}^*(\epsilon)$ by continuing iteratively. □

## 3.B Proof of Proposition 4

*Proof.* To prove Proposition 4, we study the intermediary problem.

$$\begin{aligned} \min_{\delta,\hat{\mathbf{H}}} \quad & \delta \\ \text{s.t.} \quad & \hat{\mathbf{e}}^\top\hat{\mathbf{H}}\hat{\mathbf{e}} \le (1+\delta)\|\mathbf{e}_c\|^2 + 2(l-3)\delta\|\mathbf{e}_c\|^2 \\ & (1-\delta)\mathbf{I}_{4r^2} \preceq \hat{\mathbf{H}} \preceq (1+\delta)\mathbf{I}_{4r^2}, \end{aligned} \tag{3.8}$$

where

$$\hat{\mathbf{e}} = \mathbf{P}^\top\mathbf{e}, \qquad \mathbf{P} \in \mathbb{R}^{n^2\times 4r^2} = P \otimes P,$$

and $P \in \mathbb{R}^{n\times 2r}$ is defined to be

$$P = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_{2r}, \end{bmatrix}$$

where $\mathbf{u}_i$'s are orthonormal eigenvectors of $\mathbf{M}^* - \mathbf{M}$ so that $P^\top P = \mathbf{I}$. Denote the optimal solution to (3.8) as $\delta_P(\mathbf{e})$. Then, the following two lemmas will suffice to prove Proposition 4.

**Lemma 7.** *Given a fixed vector $\mathbf{e} \in \mathbb{R}^{n^2}$, we have*

$$\delta_P(\mathbf{e}) \le \delta(\mathbf{e}). \tag{3.9}$$

**Lemma 8.** *Given a fixed vector $\mathbf{e} \in \mathbb{R}^{n^2}$, we have*

$$\delta_{lb}(\mathbf{e}) \le \delta_P(\mathbf{e}). \tag{3.10}$$

*Proof of Lemma 7.* It suffices to show that for any feasible pair $(\delta, \bar{\mathbf{H}})$ of (3.6), we can construct a feasible solution $(\delta, \hat{\mathbf{H}})$ to (3.8) characterized as below

$$\delta = \delta, \qquad \hat{\mathbf{H}} = \mathbf{P}^\top \bar{\mathbf{H}} \mathbf{P},$$

which directly proves the lemma. We can verify the feasibility of $(\delta, \hat{\mathbf{H}})$ as follows. The feasibility of the first constraint is certified by the following argument:

$$\hat{\mathbf{e}}^\top \hat{\mathbf{H}} \hat{\mathbf{e}} = \mathbf{e}^\top \mathbf{P}\mathbf{P}^\top \bar{\mathbf{H}} \mathbf{P}\mathbf{P}^\top \mathbf{e},$$

By the definition of $\mathbf{P}$, one can write

$$\mathbf{P}\mathbf{P}^\top \mathbf{e} = (PP^\top \otimes PP^\top)\mathbf{e} = \mathrm{vec}(PP^\top(\mathbf{M}^* - \mathbf{M})PP^\top) = \mathbf{e}_1 + \mathbf{e}_2 = \mathbf{e}_{2r},$$

Since $\mathbf{e}^\top \bar{\mathbf{H}} \mathbf{e} = 0$ and $\bar{\mathbf{H}}$ is symmetric, $\bar{\mathbf{H}}$ admits a factorization $\bar{\mathbf{H}} = \bar{\mathbf{A}}^\top \bar{\mathbf{A}}$, making $\bar{\mathbf{A}}\mathbf{e} = 0$. Also, we know that $\mathbf{e} = \mathbf{e}_{2r} + \mathbf{e}_c$, meaning that

$$\bar{\mathbf{A}}\mathbf{e}_{2r} = -\bar{\mathbf{A}}\mathbf{e}_c.$$

Therefore,

$$\hat{\mathbf{e}}^\top \hat{\mathbf{H}} \hat{\mathbf{e}} = \mathbf{e}_{2r}^\top \bar{\mathbf{H}} \mathbf{e}_{2r} = \mathbf{e}_c^\top \bar{\mathbf{H}} \mathbf{e}_c = (\sum_{i=3}^{l} \mathbf{e}_i)^\top \bar{\mathbf{H}} (\sum_{i=3}^{l} \mathbf{e}_i).$$

Since $\bar{\mathbf{H}}$ satisfies $\delta_{2r}$-RIP, for every $(i, j)$ such that $i \neq j$, we have:

$$(\mathbf{e}_i + \mathbf{e}_j)^\top \bar{\mathbf{H}} (\mathbf{e}_i + \mathbf{e}_j) \leq (1+\delta)\|\mathbf{e}_i + \mathbf{e}_j\|^2 = (1+\delta)(\|\mathbf{e}_i\|^2 + \|\mathbf{e}_j\|^2), \tag{3.11}$$

where the last equality follows from the facts that $\mathbf{e}_i^\top \mathbf{e}_j = 0$ and

$$(\mathbf{e}_i + \mathbf{e}_j)^\top \bar{\mathbf{H}} (\mathbf{e}_i + \mathbf{e}_j) = \mathbf{e}_i^\top \bar{\mathbf{H}} \mathbf{e}_i + 2\mathbf{e}_i^\top \bar{\mathbf{H}} \mathbf{e}_j + \mathbf{e}_j^\top \bar{\mathbf{H}} \mathbf{e}_j \geq 2\mathbf{e}_i^\top \bar{\mathbf{H}} \mathbf{e}_j + (1-\delta)(\|\mathbf{e}_i\|^2 + \|\mathbf{e}_j\|^2). \tag{3.12}$$

Combining (3.11) and (3.12) yields that

$$\mathbf{e}_i^\top \bar{\mathbf{H}} \mathbf{e}_j \leq \delta(\|\mathbf{e}_i\|^2 + \|\mathbf{e}_j\|^2) \qquad \forall i \neq j.$$

Therefore,

$$\begin{aligned}\hat{\mathbf{e}}^\top \hat{\mathbf{H}} \hat{\mathbf{e}} = (\sum_{i=3}^{l} \mathbf{e}_i)^\top \bar{\mathbf{H}} (\sum_{i=3}^{l} \mathbf{e}_i) &\leq (1+\delta)(\sum_{i=3}^{l} \|\mathbf{e}_i\|^2) + 2\delta(l-3)(\sum_{i=3}^{l} \|\mathbf{e}_i\|^2) \\ &= (1+\delta)\|\mathbf{e}_c\|^2 + 2\delta(l-3)\|\mathbf{e}_c\|^2.\end{aligned}$$

The above inequality directly verifies the satisfaction of the first constraint. For the second constraint, consider an arbitrary vector $\tilde{\mathbf{e}} \in \mathbb{R}^{4r^2}$. Then,

$$\begin{aligned}\tilde{\mathbf{e}}^\top \hat{\mathbf{H}} \tilde{\mathbf{e}} = \tilde{\mathbf{e}}^\top \mathbf{P}^\top \bar{\mathbf{H}} \mathbf{P} \tilde{\mathbf{e}} &= \tilde{\mathbf{e}}^\top (P^\top \otimes P^\top) \bar{\mathbf{H}} (P \otimes P) \tilde{\mathbf{e}} \\ &= \mathrm{vec}(P\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top)^\top \bar{\mathbf{H}}\, \mathrm{vec}(P\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top).\end{aligned}$$

By orthogonal projection, we know that $P\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top \in \mathbb{R}^{n\times n}$ has rank $2r$. Therefore, the following holds by the $\delta_{2r}$-RIP property of $\bar{\mathbf{H}}$:

$$(1-\delta)\|P\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top\|_F^2 \le \mathrm{vec}(P\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top)^\top \bar{\mathbf{H}}\,\mathrm{vec}(P\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top) \le (1+\delta)\|P\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top\|_F^2 \tag{3.13}$$

and since

$$\begin{aligned}\|P\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top\|_F^2 &= \mathrm{tr}(P\,\mathrm{mat}(\tilde{\mathbf{e}})^\top P^\top P\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top)\\ &= \mathrm{tr}(P\,\mathrm{mat}(\tilde{\mathbf{e}})^\top\,\mathrm{mat}(\tilde{\mathbf{e}})P^\top)\\ &= \mathrm{tr}(P^\top P\,\mathrm{mat}(\tilde{\mathbf{e}})^\top\,\mathrm{mat}(\tilde{\mathbf{e}}))\\ &= \mathrm{tr}(\mathrm{mat}(\tilde{\mathbf{e}})^\top\,\mathrm{mat}(\tilde{\mathbf{e}}))\\ &= \|\tilde{\mathbf{e}}\|_2^2,\end{aligned}$$

(3.13) automatically implies the satisfaction of the second constraint. □

*Proof of Lemma 8.* It suffices to show that for any feasible pair $(\delta, \hat{\mathbf{H}})$ of (3.8), we can construct a feasible solution $(\delta, \mathbf{H})$ to (3.7) characterized as

$$\delta = \delta, \qquad \mathbf{H} = \mathbf{P}\hat{\mathbf{H}}\mathbf{P}^\top + (1-\delta)(\mathbf{I}_{n^2} - \mathbf{P}\mathbf{P}^\top).$$

To prove the lemma, it is enough to verify that the above pair $(\delta, \mathbf{H})$ is feasible to (3.7). We first verify the second constraint. Given an arbitrary vector $\mathbf{e} \in \mathbb{R}^{n^2}$, we have that

$$\mathbf{e}^\top\mathbf{H}\mathbf{e} = \mathbf{e}^\top\mathbf{P}\hat{\mathbf{H}}\mathbf{P}^\top\mathbf{e} + (1-\delta)\left[\mathbf{e}^\top\mathbf{e} - \mathbf{e}^\top\mathbf{P}\mathbf{P}^\top\mathbf{e}\right]$$

and defining $\tilde{\mathbf{e}} \coloneqq \mathbf{P}^\top\mathbf{e} \in \mathbb{R}^{4r^2}$, we obtain:

$$\mathbf{e}^\top\mathbf{P}\hat{\mathbf{H}}\mathbf{P}^\top\mathbf{e} + (1-\delta)\left[\mathbf{e}^\top\mathbf{e} - \mathbf{e}^\top\mathbf{P}\mathbf{P}^\top\mathbf{e}\right] \ge (1-\delta)\|\tilde{\mathbf{e}}\|_2^2 + (1-\delta)[\|\mathbf{e}\|_2^2 - \|\tilde{\mathbf{e}}\|_2^2] = (1-\delta)\|\mathbf{e}\|_2^2.$$

Also, since $\|\tilde{\mathbf{e}}\|_2^2 \le \|\mathbf{e}\|_2^2$ and $P$ is a projection matrix, one can write:

$$(1+\delta)[\|\mathbf{e}\|_2^2 - \|\tilde{\mathbf{e}}\|_2^2] \ge (1-\delta)[\|\mathbf{e}\|_2^2 - \|\tilde{\mathbf{e}}\|_2^2],$$

which further implies that

$$\mathbf{e}^\top\mathbf{P}\hat{\mathbf{H}}\mathbf{P}^\top\mathbf{e} + (1-\delta)\left[\mathbf{e}^\top\mathbf{e} - \mathbf{e}^\top\mathbf{P}\mathbf{P}^\top\mathbf{e}\right] \le (1+\delta)\|\tilde{\mathbf{e}}\|_2^2 + (1+\delta)[\|\mathbf{e}\|_2^2 - \|\tilde{\mathbf{e}}\|_2^2] = (1+\delta)\|\mathbf{e}\|_2^2.$$

Combining the above equations, we recover the second constraint of (3.7):

$$(1-\delta)\|\mathbf{e}\|_2^2 \le e^\top\mathbf{H}\mathbf{e} \le (1+\delta)\|\mathbf{e}\|_2^2.$$

To study the first constraint, we have that

$$\begin{aligned}e^\top\mathbf{H}e &= \hat{\mathbf{e}}^\top\hat{\mathbf{H}}\hat{\mathbf{e}} + (1-\delta)\left[\|\mathbf{e}\|_2^2 - \|\hat{\mathbf{e}}\|_2^2\right]\\ &\le (1+\delta)\|\mathbf{e}_c\|_2^2 + 2(l-3)\delta\|\mathbf{e}_c\|_2^2 + (1-\delta)\|\mathbf{e}_c\|_2^2\\ &= 2\|\mathbf{e}_c\|_2^2 + 2(l-3)\delta\|\mathbf{e}_c\|_2^2.\end{aligned}$$

Note that $\|\mathbf{e}\|_2^2 - \|\hat{\mathbf{e}}\|_2^2 = \|\mathbf{e}_c\|_2^2$ due to

$$\|\mathbf{e}\|_2^2 = \sum_{i=1}^{n} \lambda_i^2, \qquad \|\hat{\mathbf{e}}\|_2^2 = \sum_{i=1}^{2r} \lambda_i^2.$$

The proof of Proposition 4 follows directly from combining Lemma 7 and 8.

□

## 3.C Proof of Lemma 6

*Proof.* We aim to solve (3.7) analytically to obtain a sufficient RIP bound for problem (3.2). This amounts to deriving a closed-form expression for $\delta_{\text{lb}}(\mathbf{e})$. We consider a simpler problem to solve (3.7):

$$\begin{aligned} \max_{\eta, \tilde{\mathbf{H}}} \quad & \eta \\ \text{s.t.} \quad & \mathbf{e}^\top \tilde{\mathbf{H}} \mathbf{e} \leq \frac{1-\eta}{2} c^2 + \frac{1+\eta}{2} d^2 \\ & \eta I_{n^2} \preceq \tilde{\mathbf{H}} \preceq I_{n^2} \end{aligned} \tag{3.14}$$

with $c^2 = 2(l-3)\|\mathbf{e}_c\|_2^2$ and $d^2 = 2\|\mathbf{e}_c\|_2^2$. Given any feasible solution $(\delta, \mathbf{H})$ to (3.7), the tuple

$$\left( \frac{1-\delta}{1+\delta}, \frac{1}{1+\delta} \mathbf{H} \right)$$

is a feasible solution to problem (3.14). Therefore, if we denote the optimal value of (3.14) as $\eta(\mathbf{e})$, then it holds that

$$\eta(\mathbf{e}) \geq \frac{1-\delta_{\text{lb}}(\mathbf{e})}{1+\delta_{\text{lb}}(\mathbf{e})} \implies \delta_{\text{lb}}(\mathbf{e}) \geq \frac{1-\eta(\mathbf{e})}{1+\eta(\mathbf{e})}. \tag{3.15}$$

We use the dual problem to solve for $\eta(\mathbf{e})$:

$$\begin{aligned} \min_{\mathbf{U}_1, \mathbf{U}_2, \gamma} \quad & \operatorname{tr}(\mathbf{U}_2) + \frac{\gamma}{2}(c^2 + d^2) \\ \text{s.t.} \quad & \operatorname{tr}(\mathbf{U}_1) + \frac{\gamma}{2}(c^2 - d^2) = 1 \\ & \gamma \mathbf{e}\mathbf{e}^\top = \mathbf{U}_1 - \mathbf{U}_2, \quad \mathbf{U}_1, \mathbf{U}_2 \succeq 0, \ \gamma \geq 0. \end{aligned} \tag{3.16}$$

Since Slater's condition holds for the convex program (3.14), the optimal solution to (3.16) is equivalent to that of (3.14), which is $\eta(\mathbf{e})$. Using a Lagrangian argument, $\eta(\mathbf{e})$ can be solved as

follows:

$$\begin{aligned}
\eta(\mathbf{e}) &= \max_{\beta\in\mathbb{R}} \min_{\gamma\geq 0} \left\{ \beta(1-\frac{\gamma}{2}(c^2-d^2)) + \gamma\frac{c^2+d^2}{2} + \min_{\substack{\mathbf{U}_1\succeq 0 \\ \mathbf{U}_1-\gamma\mathbf{e}\mathbf{e}^\top\succeq 0}} \left[\operatorname{tr}(\mathbf{U}_1-\gamma\mathbf{e}\mathbf{e}^\top) - \beta\operatorname{tr}(\mathbf{U}_1)\right] \right\} \\
&= \max_{\beta\leq 1} \min_{\gamma\geq 0} \left\{ \beta(1-\frac{\gamma}{2}(c^2-d^2)) + \gamma\frac{c^2+d^2}{2} + \min_{\substack{\mathbf{U}_1\succeq 0 \\ \mathbf{U}_1-\gamma\mathbf{e}\mathbf{e}^\top\succeq 0}} \left[(1-\beta)\operatorname{tr}(\mathbf{U}_1) - \gamma\|\mathbf{e}\|_2^2\right] \right\} \\
&= \max_{\beta\leq 1} \min_{\gamma\geq 0} \left\{ \beta(1-\frac{\gamma}{2}(c^2-d^2)) + \gamma\frac{c^2+d^2}{2} + \gamma(1-\beta)\|\mathbf{e}\|_2^2 - \gamma\|\mathbf{e}\|_2^2 \right\} \\
&= \max_{\beta\leq 1} \left\{ \beta + \min_{\gamma\geq 0} \left[ \gamma(\frac{c^2+d^2}{2} - \beta(\frac{c^2-d^2}{2} + \|\mathbf{e}\|_2^2)) \right] \right\} \\
&= \max_{\beta\leq 1} \left\{ \beta : \frac{c^2+d^2}{2} - \beta(\frac{c^2-d^2}{2} + \|\mathbf{e}\|_2^2) \geq 0 \right\} \\
&= \min\left\{ 1, \frac{c^2+d^2}{2\|\mathbf{e}\|_2^2 + c^2 - d^2} \right\},
\end{aligned}$$

where the first equality uses the Lagrangian argument by introducing the Lagrange multiplier $\beta$, and the second equality constraints $\beta \leq 1$ since $(1-\beta)\operatorname{tr}(\mathbf{U}_1)$ will be unbounded otherwise. The third equality results from the obvious choice of $\mathbf{U}_1 = \gamma\mathbf{e}\mathbf{e}^\top$ given that $(1-\beta)$ is non-negative. The fifth equality results from the choice of $\gamma = 0$ constrained to the requirement that its coefficient must be non-negative.

Substituting $\eta(\mathbf{e}) = 1$ into (3.15) results in the trivial lower bound $\delta_{\text{lb}}(\mathbf{e})$ of 0, which means that the lower bound indeed will not be negative. Hence, we will focus on $\frac{c^2+d^2}{2\|\mathbf{e}\|_2^2+c^2-d^2}$ from now on. We know from (3.15) that in order to obtain a lower bound on $\delta_{\text{lb}}$, we need to derive an upper bound on $\eta(\mathbf{e})$. Note that

$$\frac{c^2+d^2}{2\|\mathbf{e}\|_2^2 + c^2 - d^2} = \frac{2(l-2)\|\mathbf{e}_c\|_2^2}{2\|\mathbf{e}\|_2^2 + 2(l-4)\|\mathbf{e}_c\|_2^2} = \frac{(l-2)\|\mathbf{e}_c\|_2^2}{\|\mathbf{e}_{2r}\|_2^2 + (l-3)\|\mathbf{e}_c\|_2^2}. \tag{3.17}$$

The last equality follows from the relations

$$\|\mathbf{e}\|_2^2 = \|\mathbf{e}_c + \mathbf{e}_{2r}\|_2^2 = \|\mathbf{e}_{2r}\|_2^2 + \|\mathbf{e}_c\|_2^2.$$

If we fix $\|\mathbf{e}_{2r}\|_2^2$, then we can maximize (3.17) with respect to $\|\mathbf{e}_c\|_2^2$ first. In this case, taking the derivative of (3.17) yields that

$$\frac{\partial}{\partial\|\mathbf{e}_c\|_2}\left(\frac{(l-2)\|\mathbf{e}_c\|_2^2}{\|\mathbf{e}_{2r}\|_2^2 + (l-3)\|\mathbf{e}_c\|_2^2}\right) = 2\frac{(l-2)\|\mathbf{e}_c\|_2\|\mathbf{e}_{2r}\|_2^2}{(\|\mathbf{e}_{2r}\|_2^2 + (l-3)\|\mathbf{e}_c\|_2^2)^2} \geq 0.$$

Therefore, (3.17) is maximized when $\|\mathbf{e}_c\|_2^2$ is set to be as large as possible. Before we derive the maximum value of $\|\mathbf{e}_c\|_2^2$ in terms of $\|\mathbf{e}_1\|_2^2$ and $\|\mathbf{e}_2\|_2^2$, we introduce one key lemma.

**Lemma 9.** *Consider two PSD matrices* $\mathbf{M}$ *and* $\mathbf{M}^*$ *such that* $\mathrm{tr}(\mathbf{M}) \leq \mathrm{tr}(\mathbf{M}^*)$ *and* $\mathrm{rank}(\mathbf{M}^*) = r$. *Then,*

$$\sigma_{(1)}(\mathbf{M}^* - \mathbf{M}) + \cdots + \sigma_{(r)}(\mathbf{M}^* - \mathbf{M}) \geq \sigma_{(r+1)}(\mathbf{M}^* - \mathbf{M}) + \cdots \sigma_{(n)}(\mathbf{M}^* - \mathbf{M}), \tag{3.18}$$

*where* $\sigma_i$ *denote the* $i$*-th largest singular value of the matrix* $\mathbf{M}^* - \mathbf{M}$.

*Proof.* For each matrix $\mathbf{A}$, we denote the $i^{\text{th}}$ eigenvalue as $\lambda_{(i)}(\cdot)$, meaning that

$$\lambda_{(1)}(\mathbf{A}) \geq \lambda_{(2)}(\mathbf{A}) \geq \cdots \geq \lambda_{(n)}(\mathbf{A}).$$

By Weyl's inequality, we know that

$$\lambda_{(i+j-1)}(\mathbf{M}^* - \mathbf{M}) \leq \lambda_{(i)}(\mathbf{M}^*) + \lambda_{(j)}(-\mathbf{M}).$$

Hence,

$$\lambda_{(r+1)}(\mathbf{M}^* - \mathbf{M}) \leq \lambda_{(r+1)}(\mathbf{M}^*) + \lambda_{(1)}(-\mathbf{M}) \leq 0$$

since $\mathbf{M}^*$ is of rank-$r$ and $\mathbf{M} \succeq 0$. Therefore, we know that $\mathbf{M}^* - \mathbf{M}$ has at most $r$ positive eigenvalues. Also, since $\mathrm{tr}(\mathbf{M}^* - \mathbf{M}) \geq 0$, it holds that

$$\lambda_{(1)}(\mathbf{M}^* - \mathbf{M}) + \cdots + \lambda_{(r)}(\mathbf{M}^* - \mathbf{M}) \geq -\lambda_{(r+1)}(\mathbf{M}^* - \mathbf{M}) - \cdots - \lambda_{(n)}(\mathbf{M}^* - \mathbf{M}),$$

which implies that

$$|\lambda_{(1)}(\mathbf{M}^* - \mathbf{M})| + \cdots + |\lambda_{(r)}(\mathbf{M}^* - \mathbf{M})| \geq |\lambda_{(r+1)}(\mathbf{M}^* - \mathbf{M})| + \cdots + |\lambda_{(n)}(\mathbf{M}^* - \mathbf{M})| \tag{3.19}$$

since $\lambda_{(k)}(\mathbf{M}^* - \mathbf{M}) \leq 0$ for all $k > r$. According to the definition, we have

$$\begin{aligned} &|\lambda_1(\mathbf{M}^* - \mathbf{M})| + \cdots + |\lambda_r(\mathbf{M}^* - \mathbf{M})| \geq |\lambda_{(1)}(\mathbf{M}^* - \mathbf{M})| + \cdots + |\lambda_{(r)}(\mathbf{M}^* - \mathbf{M})|, \\ &|\lambda_{(r+1)}(\mathbf{M}^* - \mathbf{M})| + \cdots + |\lambda_{(n)}(\mathbf{M}^* - \mathbf{M})| \geq |\lambda_{r+1}(\mathbf{M}^* - \mathbf{M})| + \ldots |\lambda_n(\mathbf{M}^* - \mathbf{M})| \end{aligned} \tag{3.20}$$

since $\lambda_1(\mathbf{M}^* - \mathbf{M}), \ldots, \lambda_n(\mathbf{M}^* - \mathbf{M})$ are ordered with respect to their absolute values. As per the main text, we abbreviate $\lambda_i(\mathbf{M}^* - \mathbf{M})$ as $\lambda_i$ for the sake of brevity for any $i \in [n]$. Combining (3.19) with (3.20) proves the original lemma. □

Denote $S_1 \coloneqq \sum_{i=1}^{r} |\lambda_i|$ and $S_2 \coloneqq \sum_{i=r+1}^{n} |\lambda_i|$. Given $S_2$, since $|\lambda_i| \leq |\lambda_r|$ as long as $i > r$, we know that $\sum_{i=r+1}^{n} \lambda_i^2$ is maximized when every absolute value is chosen to be as large as possible, namely

$$|\lambda_{r+1}|, \ldots, |\lambda_{r+\lfloor S_2/|\lambda_r| \rfloor}| = |\lambda_r|, |\lambda_{r+\lceil S_2/|\lambda_r| \rceil}| = S_2 - \lfloor S_2/|\lambda_r| \rfloor |\lambda_r| \coloneqq \tilde{\lambda} \leq |\lambda_r|.$$

Therefore,

$$\sum_{i=r+1}^{n} \lambda_i^2 \leq \lfloor S_2/|\lambda_r| \rfloor \lambda_r^2 + \tilde{\lambda}^2 \leq \lfloor S_2/|\lambda_r| \rfloor \lambda_r^2 + \frac{\tilde{\lambda}}{|\lambda_r|} \lambda_r^2 = \frac{S_2}{|\lambda_r|} \lambda_r^2.$$

As a result,

$$\frac{S_2}{|\lambda_r|}\lambda_r^2 = S_2|\lambda_r| \leq S_1|\lambda_r| \leq S_1\frac{S_1}{r} \leq \frac{S_1^2}{r} \leq \frac{r\|\mathbf{e}_1\|_2^2}{r} = \|\mathbf{e}_1\|_2^2,$$

where the last inequality follows from Cauchy-Schwartz. Combining the above two inequalities, we obtain

$$\sum_{i=r+1}^{n} \lambda_i^2 \leq \|\mathbf{e}_1\|_2^2. \tag{3.21}$$

Furthermore, since $|\lambda_{r+1}| \geq \cdots \geq |\lambda_n|$, one can write:

$$\|\mathbf{e}_c\|_2^2 = \sum_{i=2r+1}^{n} \lambda_i^2 \leq \frac{n-2r}{n-r}\sum_{i=r+1}^{n} \lambda_i^2$$

with equality holding if and only if $|\lambda_{r+1}| = \cdots = |\lambda_n|$. Combined with (3.21), we obtain

$$\|\mathbf{e}_c\|_2^2 \leq \frac{n-2r}{n-r}\|\mathbf{e}_1\|_2^2. \tag{3.22}$$

Consequently,

$$\|\mathbf{e}_c\|_2^2 \leq \frac{n-2r}{n-r}(\|\mathbf{e}_2\|_2^2 + \|\mathbf{e}_c\|_2^2) \implies \|\mathbf{e}_2\|_2^2 \geq \frac{r}{n-2r}\|\mathbf{e}_c\|_2^2. \tag{3.23}$$

It results from (3.22) and (3.23) that

$$\begin{aligned}
\max_{\mathbf{e}:\mathrm{tr}(M)\leq\mathrm{tr}(M^*)} \eta(\mathbf{e}) &= \max_{\mathbf{e}:\mathrm{tr}(M)\leq\mathrm{tr}(M^*)} \frac{(l-2)\|\mathbf{e}_c\|_2^2}{\|\mathbf{e}_{2r}\|_2^2 + (l-3)\|\mathbf{e}_c\|_2^2} \\
&\leq \frac{(l-2)\|\mathbf{e}_c\|_2^2}{\|\mathbf{e}_1\|_2^2 + \frac{r}{n-2r}\|\mathbf{e}_c\|_2^2 + (l-3)\|\mathbf{e}_c\|_2^2} \\
&\leq \frac{(l-2)\frac{n-2r}{n-r}}{1 + (\frac{r}{n-2r} + l - 3)\frac{n-2r}{n-r}} \\
&= \frac{(l-2)(n-2r)}{n + (n-2r)(l-3)}.
\end{aligned}$$

Thus,

$$\delta_{\text{lb}} \geq \max_{\mathbf{e}:\mathrm{tr}(M)\leq\mathrm{tr}(M^*)} \frac{1-\delta(\mathbf{e})}{1+\delta(\mathbf{e})} = \frac{2r}{n + (n-2r)(2l-5)}.$$

□

## 3.D Proof of Theorem 9

*Proof.* The matrix completion problem with the least-squares objective function can be written as

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = (x_1^2 - (x_1^*)^2)^2 + 2\sum_{i=1}^{n-1}(x_i x_{i+1} - x_i^* x_{i+1}^*)^2. \tag{3.24}$$

We assume that every element of the vector $\mathbf{x}^*$ is non-zero, i.e. $x_i^* \neq 0, \forall i \in [n]$. In this case, the problem (3.24) does not have any spurious solutions. In order to prove the inexistence of a non-global second-order critical solution, we need to investigate the second-order necessary optimality conditions. The gradient and Hessian of the above objective can be written as

$$[\nabla f(x)]_i = \begin{cases} (x_1^2 - (x_1^*)^2)x_1 + (x_1x_2 - x_1^*x_2^*)x_2, & \text{if } i = 1 \\ (x_{i-1}x_i - x_{i-1}^*x_i^*)x_{i-1} + (x_ix_{i+1} - x_i^*x_{i+1}^*)x_{i+1}, & \text{if } i \notin \{1, n\} \\ (x_{n-1}x_n - x_{n-1}^*x_n^*)x_{n-1}, & \text{if } i = n \end{cases},$$

$$[\nabla^2 f(x)]_{i,j} = \begin{cases} 3x_1^2 - (x_1^*)^2 + x_2^2, & \text{if } i,j = 1 \\ x_{i-1}^2 + x_{i+1}^2, & \text{if } i = j, i, j \notin \{1, n\} \\ x_{n-1}^2, & \text{if } i,j = n \\ 2x_ix_j - x_i^*x_j^*, & \text{if } |i-j| = 1 \\ 0, & \text{otherwise} \end{cases}$$

Note that for the point $\hat{\mathbf{x}}$ to be a second-order stationary point, we require $\nabla f(\hat{\mathbf{x}}) = 0$ and $\nabla^2 f(\hat{\mathbf{x}}) \succeq 0$. $[\nabla f(\hat{\mathbf{x}})]_n = 0$ implies either $\hat{x}_{n-1}\hat{x}_n = x_{n-1}^*x_n^*$ or $\hat{x}_{n-1} = 0$. The latter results in $\nabla^2 f(\hat{\mathbf{x}}) \not\succeq 0$ since the following principal minor is not positive semidefinite:

$$\begin{aligned}[\nabla^2 f(\hat{\mathbf{x}})]_{n-1:n,n-1:n} &= \begin{bmatrix} (\hat{x}_{n-2})^2 + (\hat{x}_n)^2 & 2\hat{x}_{n-1}\hat{x}_n - x_{n-1}^*x_n^* \\ 2\hat{x}_{n-1}\hat{x}_n - x_{n-1}^*x_n^* & (\hat{x}_{n-1})^2 \end{bmatrix} \\ &= \begin{bmatrix} (\hat{x}_{n-2})^2 + (\hat{x}_n)^2 & -x_{n-1}^*x_n^* \\ -x_{n-1}^*x_n^* & 0 \end{bmatrix} \not\succeq 0.\end{aligned}$$

Hence, $\hat{x}_{n-1}\hat{x}_n^* = x_{n-1}^*x_n^*$ is the only feasible option for $\hat{\mathbf{x}}$ to be a second-order critical point. Next, we show that no second-order critical point $\hat{\mathbf{x}}$ can have a zero entry, i.e. $\hat{x}_i \neq 0$ for all $i \in [n]$.

The first goal is to identify the structure of the stationary points with at least one zero entry. If $\hat{x}_1 = 0$, $[\nabla f(\hat{\mathbf{x}})]_1$ implies $\hat{x}_2 = 0$. Then, $[\nabla f(\hat{\mathbf{x}})]_2$ gives $\hat{x}_3 = 0$. Continuing iteratively gives that the only possible stationary point with $\hat{x}_1 = 0$ is $\hat{\mathbf{x}} = 0$, which cannot be a second-order critical point. Similarly, if $\hat{x}_n = 0$, then $[\nabla f(\hat{\mathbf{x}})]_n$ implies $\hat{x}_{n-1} = 0$. Then, $[\nabla f(\hat{\mathbf{x}})]_{n-1}$ gives $\hat{x}_{n-2} = 0$. Hence, the only stationary point with $\hat{x}_n = 0$ is $\hat{\mathbf{x}} = 0$. Consequently, $\hat{x}_1 \neq 0$ and $\hat{x}_n \neq 0$ for a second-order stationary point.

Let $k$ be defined as an index with the property that $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{k-1} \neq 0$ and $\hat{x}_k = 0$. Then, $[\nabla f(\hat{\mathbf{x}})]_{k-2}$ yields $\hat{x}_{k-2}\hat{x}_{k-1} = x_{k-2}^*x_{k-1}^*$. Continuing iteratively backward on $[\nabla f(\hat{\mathbf{x}}))]_i$ for $1 \leq$

$i \leq k-2$ gives the following conditions on second-order critical points, which are correct values for the corresponding edges:

$$\begin{aligned}
\hat{x}_{k-2}\hat{x}_{k-1} &= x^*_{k-2}x^*_{k-1},\\
\hat{x}_{k-3}\hat{x}_{k-2} &= x^*_{k-3}x^*_{k-2},\\
&\vdots\\
\hat{x}_1\hat{x}_2 &= x^*_1x^*_2,\\
(\hat{x}_1)^2 &= (x^*_1)^2.
\end{aligned}$$

We can focus on entries of the stationary point $\hat{\mathbf{x}}$ corresponding to $\hat{x}_i$ for $i > k$. We show that $\hat{x}_{k+1} \neq 0$ and $\hat{x}_{k+2} \neq 0$ for every second-order critical point. If $\hat{x}_{k+1} = 0$, then $[\nabla f(\hat{\mathbf{x}})]_k = 0$ gives $x^*_{k-1}x^*_k = 0$, which contradicts the assumption on the ground truth matrix. Since $\hat{x}_{n-1}, \hat{x}_n \neq 0$, we have $k \leq n-2$ for each second-order critical point. If $\hat{x}_{k+2} = 0$, then the $2 \times 2$ submatrix of the Hessian will have the form

$$\begin{aligned}
[\nabla^2 f(\hat{\mathbf{x}})]_{k:k+1,k:k+1} &= \begin{bmatrix} (\hat{x}_{k-1})^2 + (\hat{x}_{k+1})^2 & 2\hat{x}_k\hat{x}_{k+1} - x^*_kx^*_{k+1} \\ 2\hat{x}_k\hat{x}_{k+1} - x^*_kx^*_{k+1} & (\hat{x}_k)^2 + (\hat{x}_{k+2})^2 \end{bmatrix}\\
&= \begin{bmatrix} (\hat{x}_{k-1})^2 + (\hat{x}_{k+1})^2 & -x^*_kx^*_{k+1} \\ -x^*_kx^*_{k+1} & 0 \end{bmatrix} \not\succeq 0.
\end{aligned}$$

Let $m$ be defined as an index with the property that $\hat{x}_{k+1}, \hat{x}_{k+2}, \ldots, \hat{x}_m \neq 0$ and either $\hat{x}_{m+1} = 0$ or $m = n$. By the previous arguments and the definition of $m$, we have the condition $k+3 \leq m \leq n$. We are not interested in the entries after the $m$-th entry because we can show that a stationary point with this structure cannot be a second-order critical point because $[\nabla^2 f(\hat{\mathbf{x}})]_{1:m,1:m} \not\succeq 0$. By using the first-order partial derivatives and algebra, we obtain the following equations for the stationary point $\hat{\mathbf{x}}$:

$$\begin{aligned}
&[\nabla f(\hat{\mathbf{x}})]_k = 0 \rightarrow \hat{x}_{k+1} = -\left(\frac{x^*_{k-1}}{x^*_{k+1}}\right)^2 x^*_{k+1} = -\alpha x^*_{k+1},\\
&[\nabla f(\hat{\mathbf{x}})]_{k+1} = 0 \rightarrow \hat{x}_{k+2} = -\left(\frac{x^*_{k-1}}{x^*_{k+1}}\right)^{-2} x^*_{k+2} = -\alpha^{-1} x^*_{k+2},\\
&\vdots\\
&[\nabla f(\hat{\mathbf{x}})]_{m-1} = 0 \rightarrow \hat{x}_m = -\left(\frac{x^*_{k-1}}{x^*_{k+1}}\right)^{2(-1)^{(m-k+1)}} x^*_m = -\alpha^{((-1)^{m-k+1})} x^*_m,
\end{aligned}$$

where $\alpha = \left(\frac{x^*_{k-1}}{x^*_{k+1}}\right)^2$. As a result, the possible candidates for second-order critical points with at least one zero entry have the form:

$$\hat{\mathbf{x}} = [x^*_1, x^*_2, \ldots, x^*_{k-1}, 0, -\alpha x^*_{k+1}, -\alpha^{-1}x^*_{k-2}, \ldots, -\alpha^{((-1)^{m-k+1})}x^*_m, \ldots]$$

or the form

$$\hat{\mathbf{x}} = [-x_1^*, -x_2^*, \dots, -x_{k-1}^*, 0, \alpha x_{k+1}^*, \alpha^{-1} x_{k-2}^*, \dots, \alpha^{((-1)^{m-k+1})} x_m^*, \dots].$$

These points correspond to the same matrix completion, and they lead to the same Hessian matrix. As mentioned before, we focus on the $m \times m$ Hessian submatrix $[\nabla^2 f(\hat{\mathbf{x}})]_{1:m,1:m}$:

$$[\nabla^2 f(\hat{\mathbf{x}})]_{1:m,1:m} = \left(\begin{array}{c|cc} \mathbf{E} & \multicolumn{2}{c}{\mathbf{D}} \\ \hline \mathbf{D}^\top & \mathbf{A} & \mathbf{B} \\ & \mathbf{B}^\top & \mathbf{C} \end{array}\right),$$

where $\mathbf{A} \in \mathbb{R}^{3\times 3}, \mathbf{B} \in \mathbb{R}^{3\times(m-k-1)}, \mathbf{C} \in \mathbb{R}^{(m-k-1)\times(m-k-1)}, \mathbf{D} \in \mathbb{R}^{(k-2)\times(m-k+2)}$ and $\mathbf{E} \in \mathbb{R}^{(k-2)\times(k-2)}$. The submatrices can be written as

$$\mathbf{A} = \begin{bmatrix} (x_{k-2}^*)^2 & -x_{k-1}^* x_k^* & 0 \\ -x_{k-1}^* x_k^* & (x_{k-1}^*)^2 + \frac{(x_{k-1}^*)^4}{(x_{k+1}^*)^2} & -x_k^* x_{k+1}^* \\ 0 & -x_k^* x_{k+1}^* & \frac{(x_{k+1}^*)^4 (x_{k+2}^*)^2}{(x_{k-1}^*)^4} \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ x_{k+1}^* x_{k+2}^* & 0 & \dots & 0 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} \alpha^2 (x_{k+1}^*)^2 + \alpha^2 (x_{k+3}^*)^2 & x_{k+2}^* x_{k+3}^* & \dots & 0 \\ x_{k+2}^* x_{k+3}^* & \alpha^{-2}(x_{k+2}^*)^2 + \alpha^{-2}(x_{k+4}^*)^2 & \dots & 0 \\ 0 & x_{k+3}^* x_{k+4}^* & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_{m-1}^* x_m^* \\ 0 & 0 & \dots & (\alpha^2)^{((-1)^{m-k})} (x_{m-1}^*)^2 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_{k-2}^* x_{k-1}^* & 0 & \dots & 0 \end{bmatrix},$$

$$\mathbf{E} = [\nabla^2 f(\mathbf{x}^*)]_{1:k-2,1:k-2}.$$

We investigate three different cases: $\mathbf{A} \not\succeq 0$, $\mathbf{A} \succeq 0$ with at least one of the eigenvalues being equal to $0$, and $\mathbf{A} \succ 0$. If $\mathbf{A} \not\succeq 0$, then $[\nabla^2 f(\hat{\mathbf{x}})]_{1:m,1:m} \not\succeq 0$ because $\mathbf{A}$ is a principal minor of the Hessian. Therefore, $\hat{\mathbf{x}}$ cannot be a second-order critical point. If $\mathbf{A} \succeq 0$ with an eigenvalue equal to $0$, we consider the following $4 \times 4$ principal minors of the Hessian:

$$\left(\begin{array}{c|ccc} (x_{k-3}^*)^2 + (x_{k-1}^*)^2 & x_{k-2}^* x_{k-1}^* & 0 & 0 \\ \hline x_{k-2}^* x_{k-1}^* & \multicolumn{3}{c}{\multirow{2}{*}{$\mathbf{A}$}} \\ 0 & & & \\ 0 & & & \end{array}\right), \qquad \left(\begin{array}{ccc|c} & & & 0 \\ & \mathbf{A} & & 0 \\ & & & x_{k+1}^* x_{k+2}^* \\ \hline 0 & 0 & x_{k+1}^* x_{k+2}^* & \alpha^2 (x_{k+1}^*)^2 + \alpha^2 (x_{k+3}^*)^2 \end{array}\right).$$

The submatrices are not positive semidefinite by Schur complement if

$$\mathbf{A} - \frac{1}{(x^*_{k-3})^2 + (x^*_{k-1})^2} \begin{bmatrix} (x^*_{k-2})^2(x^*_{k-1})^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \not\succeq 0$$

and

$$\mathbf{A} - \frac{1}{\alpha^2(x^*_{k+1})^2 + \alpha^2(x^*_{k+3})^2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & (x^*_{k+1})^2(x^*_{k+2})^2 \end{bmatrix} \not\succeq 0,$$

respectively. If $\mathbf{A}\mathbf{v} = 0$, then $v_1$ and $v_3$ cannot be equal to 0 at the same time because $(x^*_{k-1})^2 + \frac{(x^*_{k-1})^4}{(x^*_{k+1})^2} > 0$. If $v_1 \neq 0$, then

$$\mathbf{v}^\top \mathbf{A}\mathbf{v} - \mathbf{v}^\top \frac{1}{(x^*_{k-3})^2 + (x^*_{k-1})^2} \begin{bmatrix} (x^*_{k-2})^2(x^*_{k-1})^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v} = 0 - \frac{v_1^2(x^*_{k-2})^2(x^*_{k-1})^2}{(x^*_{k-3})^2 + (x^*_{k-1})^2} < 0.$$

Hence, the quadratic form of the Hessian has a descent direction. If $v_1 = 0$, then $v_3 \neq 0$, and we can use the second submatrix to show that the submatrix is not positive semidefinite. Thus, $\hat{\mathbf{x}}$ cannot be a second-order critical point. As a result, if $\hat{\mathbf{x}}$ is a second-order critical point, then $\mathbf{A} \succ 0$. In that case, we can show that the submatrix

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix}$$

cannot be positive semidefinite. This requires proving that $\mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1} \mathbf{B} \not\succeq 0$. Note that

$$\mathbf{B}^\top \mathbf{A}^{-1} \mathbf{B} = \begin{bmatrix} \mathbf{A}^{-1}_{3,3}(x^*_{k+1})^2(x^*_{k+2})^2 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

We can calculate $\mathbf{A}^{-1}_{3,3}$ by using the cofactors of $\mathbf{A}$ as

$$\mathbf{A}^{-1}_{3,3} = \frac{1}{\det(\mathbf{A})} \Big( (x^*_{k-2})^2(x^*_{k-1})^2 + (x^*_{k-2})^2 \frac{(x^*_{k-1})^4}{(x^*_{k+1})^2} - (x^*_{k-1})^2(x^*_k)^2 \Big),$$

where

$$\det(\mathbf{A}) = (x^*_{k-2})^2 \Big( \frac{(x^*_{k+1})^4}{(x^*_{k-1})^2}(x^*_{k+2})^2 + (x^*_{k+1})^2(x^*_{k+2})^2 - (x^*_k)^2(x^*_{k+1})^2 \Big) - \frac{(x^*_k)^2(x^*_{k+1})^4(x^*_{k+2})^2}{(x^*_{k-1})^2}.$$

An algebraic manipulation shows that $\frac{d\mathbf{A}_{3,3}^{-1}}{d(x_k^*)^2} > 0$. Thus, the value of $\mathbf{A}_{3,3}^{-1}$ is minimized when $(x_k^*)^2 \to 0$. Whenever $(x_k^*)^2 = 0$, $\mathbf{A}_{3,3}^{-1}(x_{k+1}^*)^2(x_{k+2}^*)^2$ is equal to $\alpha^2(x_{k+1}^*)^2$. As a result, $\mathbf{A}_{3,3}^{-1}(x_{k+1}^*)^2(x_{k+2}^*)^2 > \alpha^2(x_{k+1}^*)^2$ by the non-zero assumption on $x_k^*$. We can write the matrix $\mathbf{C} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}$ as the summation of $m-k-1$ matrices as

$$\mathbf{C} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B} = \begin{bmatrix} \alpha^2(x_{k+1}^*)^2 - \mathbf{A}_{3,3}^{-1}(x_{k+1}^*)^2(x_{k+2}^*)^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} + \begin{bmatrix} \alpha^2(x_{k+3}^*)^2 & x_{k+2}^*x_{k+3}^* & \dots & 0 & 0 \\ x_{k+2}^*x_{k+3}^* & \alpha^{-2}(x_{k+2}^*)^2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & (\alpha^2)^{(-1)^{m-k+1}}(x_m^*)^2 & x_{m-1}^*x_m^* \\ 0 & 0 & \dots & x_{m-1}^*x_m^* & (\alpha^2)^{(-1)^{m-k}}(x_{m-1}^*)^2 \end{bmatrix}.$$

Consider the vector $\mathbf{v}$ defined as $v_{2t+1} = \frac{x_{k+2t+2}^*}{x_{k+2}^*}$ and $v_{2t} = \alpha^2\frac{x_{k+2t+1}^*}{x_{k+2}^*}$ for $t = 0, \dots, \lfloor (m-k-1)/2 \rfloor$. Then,

$$\mathbf{v}^\top(\mathbf{C} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B})\mathbf{v} = \alpha^2(x_{k+1}^*)^2 - \mathbf{A}_{3,3}^{-1}(x_{k+1}^*)^2(x_{k+2}^*)^2 + 0 + \cdots + 0 < 0.$$

As a result, $\mathbf{C} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}$ is not positive definite. Hence, the Hessian is not positive definite either. Consequently, the problem cannot have a second-order critical point $\hat{x}$ with $\hat{x}_i = 0$ for some $i \in [n]$. Then, the only possible second-order-critical points must satisfy $(\hat{x}_1)^2 = (x_1^*)^2$ and $\hat{x}_i\hat{x}_{i+1} = x_i^*x_{i+1}^*, i = 1, \dots, n-1$, which correspond to the valid factors of the ground truth completion of the matrix $\mathbf{M}^*$.

□

## 3.E Proof of Theorem 10

*Proof.* To show that the SDP formulated as (3.2) fails to solve the problem, consider two indices $j, k$ such that $x_k^* > x_j^*$ and $j, k > 2$. Then, we construct a feasible solution $\hat{\mathbf{M}}$ that is strictly better than the ground truth solution, which shows the failure of SDP. Let $\hat{\mathbf{M}} = \mathbf{y}\mathbf{y}^\top + \mathbf{z}\mathbf{z}^\top$, sum of two rank-1 matrices, where

$$y_i = x_i^*, \forall i \in [n]\backslash\{k\}, \qquad y_k = x_j^*$$

and

$$z_i = 0, \forall i \in [n] \backslash \{j, k\}, \qquad z_j = z_k = (|x_j^* x_k^*| - (x_j^*)^2)^{1/2}.$$

Since the sum of PSD matrices is PSD and $\hat{\mathbf{M}}$ satisfies the observed entries, $\hat{\mathbf{M}}$ is a feasible solution. Moreover, the objective value corresponding to $\hat{\mathbf{M}}$ is $\sum_{i \neq j, i \neq k} (x_i^*)^2 + 2|x_j^* x_k^*|$. Hence, by the assumption, the feasible solution $\hat{\mathbf{M}}$ is strictly better than the ground truth solution $\mathbf{M}^*$. Thus, SDP fails to recover the true solution. □

## 3.F Proof of Theorem 11

*Proof.* Note that the nodes are numbered from 0 to $2k$ as opposed to earlier examples. Hence, the matrix completion problem with the least-squares objective function can be written as

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = 2 \sum_{i=0}^{2k} (x_i x_{i+1} - x_i^* x_{i+1}^*)^2 + 2(x_0 x_{2k} - x_0^* x_{2k}^*)^2. \tag{3.25}$$

The gradient and Hessian of the above objective can be written as

$$[\nabla f(\mathbf{x})]_i = \begin{cases} (x_0 x_{2k} - x_0^* x_{2k}^*) x_{2k} + (x_0 x_1 - x_0^* x_1^*) x_1, & \text{if } i = 0 \\ (x_{i-1} x_i - x_{i-1}^* x_i^*) x_{i-1} + (x_i x_{i+1} - x_i^* x_{i+1}^*) x_{i+1}, & \text{if } i \notin \{0, 2k\} \\ (x_{2k-1} x_{2k} - x_{2k-1}^* x_{2k}^*) x_{2k-1} + (x_0 x_{2k} - x_0^* x_{2k}^*) x_0, & \text{if } i = 2k \end{cases},$$

$$[\nabla^2 f(x)]_{i,j} = \begin{cases} x_1^2 + x_{2k}^2, & \text{if } i, j = 0 \\ x_{i-1}^2 + x_{i+1}^2, & \text{if } i = j,\ i, j \notin \{0, 2k\} \\ x_0^2 + x_{2k-1}^2, & \text{if } i, j = 2k \\ 2x_i x_j - x_i^* x_j^*, & \text{if } |i - j| = 1 \\ 0, & \text{otherwise} \end{cases}.$$

The optimization problem (3.25) does not have any spurious solution $\hat{\mathbf{x}}$ such that $\hat{x}_i = 0$ for some $i = 0, \dots 2k$. To prove this, assume without loss of generality that $\hat{x}_0 = 0$ for a stationary point $\hat{\mathbf{x}}$. By the proof of Theorem 9, we know that $\hat{x}_{2k-1}, \hat{x}_{2k}, \hat{x}_1, \hat{x}_2 \neq 0$. Let $m$ be defined as an index such that $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_m \neq 0$ and either $\hat{x}_{m+1} = 0$ or $m = n$. Thus, $2 < m < 2k - 2$. One can characterize the stationary points as.

$$\begin{aligned} &[\nabla f(\hat{\mathbf{x}})]_{2k} = 0 \to \hat{x}_{2k-1} \hat{x}_{2k} = x_{2k-1}^* x_{2k}^*, \\ &[\nabla f(\hat{\mathbf{x}})]_0 = 0 \to -x_0^* x_{2k}^* \hat{x}_{2k} = x_0^* x_1^* \hat{x}_1, \\ &[\nabla f(\hat{\mathbf{x}})]_1 = 0 \to \hat{x}_1 \hat{x}_2 = x_1^* x_2^*, \\ &[\nabla f(\hat{\mathbf{x}})]_2 = 0 \to \hat{x}_2 \hat{x}_3 = x_2^* x_3^*, \\ &\vdots \\ &[\nabla f(\hat{\mathbf{x}})]_m = 0 \to \hat{x}_{m-1} \hat{x}_m = x_{m-1}^* x_m^*. \end{aligned}$$

Setting $\hat{x}_{2k}$ as a free variable gives the following characterization of the stationary points:

$$\begin{aligned}
\hat{x}_{2k-2} &= \frac{x^*_{2k-2}}{x^*_{2k}}\hat{x}_{2k},\\
\hat{x}_{2k-1} &= \frac{x^*_{2k-1}x^*_{2k}}{\hat{x}_{2k}},\\
\hat{x}_{2k} &= \hat{x}_{2k},\\
\hat{x}_0 &= 0,\\
\hat{x}_1 &= -\frac{x^*_{2k}\hat{x}_{2k}}{(x^*_1)^2}x^*_1 = -\alpha x^*_1,\\
\hat{x}_2 &= -\frac{(x^*_1)^2}{x^*_{2k}\hat{x}_{2k}}x^*_2 = -\alpha^{-1}x^*_2,\\
\hat{x}_3 &= -\alpha x^*_3,\\
&\vdots\\
\hat{x}_m &= -\alpha^{(-1)^{m-1}}x^*_m.
\end{aligned}$$

Similar to proof of Theorem 9, we focus on the following $(m+2)\times(m+2)$ submatrix of the Hessian that is $[\nabla^2 f(\hat{\mathbf{x}})]_{(2k-1):m,(2k-1):m}$ where $(2k-1):m$ denotes the rows/columns corresponding to $(2k-1), 2k, 0, 1, \ldots, m$ in that respective order:

$$[\nabla^2 f(\hat{\mathbf{x}})]_{(2k-1):m,(2k-1):m} = \left(\begin{array}{c|cc} \mathbf{E} & \multicolumn{2}{c}{\mathbf{D}} \\ \hline \mathbf{D}^\top & \mathbf{A} & \mathbf{B} \\ & \mathbf{B}^\top & \mathbf{C} \end{array}\right),$$

where $\mathbf{A} \in \mathbb{R}^{3\times 3}, \mathbf{B} \in \mathbb{R}^{3\times(m-1)}, \mathbf{C} \in \mathbb{R}^{(m-1)\times(m-1)}, \mathbf{D} \in \mathbb{R}^{1\times(m+2)}$ and $\mathbf{E} \in \mathbb{R}$. The submatrices can be written as

$$\mathbf{A} = \begin{bmatrix} \frac{(x^*_{2k-1})^2(x^*_{2k})^2}{(\hat{x}_{2k})^2} & -x^*_{2k}x^*_0 & 0 \\ -x^*_{2k}x^*_0 & (\hat{x}_{2k})^2 + \frac{(x^*_{2k})^2(\hat{x}_{2k})^2}{(x^*_1)^2} & -x^*_0x^*_1 \\ 0 & -x^*_0x^*_1 & \frac{(x^*_1)^4(x^*_2)^2}{(x^*_{2k})^2(\hat{x}_{2k})^2} \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ x^*_1x^*_2 & 0 & \dots & 0 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} \alpha^2(x^*_1)^2 + \alpha^2(x^*_3)^2 & x^*_2x^*_3 & \dots & 0 \\ x^*_2x^*_3 & \alpha^{-2}(x^*_2)^2 + \alpha^{-2}(x^*_4)^2 & \dots & 0 \\ 0 & x^*_3x^*_4 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x^*_{m-1}x^*_m \\ 0 & 0 & \dots & (\alpha^2)^{((-1)^{m-2})}(x^*_{m-1})^2 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} x^*_{2k-1}x^*_{2k} & 0 & \dots & 0 \end{bmatrix},$$

$$\mathbf{E} = (\hat{x}_{2k})^2 + \frac{(x^*_{2k-2})^2}{(x^*_{2k})^2}(\hat{x}_{2k})^2.$$

Similar to the proof of Theorem 9, we can investigate three different cases to demonstrate that $\hat{\mathbf{x}}$ cannot be a second-order critical point, which are $\mathbf{A} \not\succeq 0$, $\mathbf{A} \succeq 0$ with at least one zero eigenvalue, and $\mathbf{A} \succ 0$. Firstly, if $\mathbf{A} \not\succeq 0$, $\nabla^2 f(\hat{\mathbf{x}})$ cannot be positive semidefinite because $\mathbf{A}$ is a principal minor of the Hessian. The second case is when $\mathbf{A}$ is positive semidefinite but not positive definite. In that case, for a vector $\mathbf{v}$ that satisfies $\mathbf{Av} = 0$, we know that $v_1$ and $v_3$ cannot be zero simultaneously. Consider following principal minors that correspond to $[\nabla^2 f(\hat{\mathbf{x}})]_{(2k-1):1,(2k-1):1}$, which includes the rows/columns corresponding to $(2k-1), 2k, 0, 1$ in that order, and $[\nabla^2 f(\hat{\mathbf{x}})]_{(2k):2,(2k):2}$, which includes the rows/columns corresponding to $2k, 0, 1, 2$ in that order, respectively:

$$\left(\begin{array}{c|cccc} (\hat{x}_{2k})^2 + \frac{(x^*_{2k-2})^2}{(x^*_{2k})^2}(\hat{x}_{2k})^2 & x^*_{2k-1}x^*_{2k} & 0 & 0 \\ \hline x^*_{2k-1}x^*_{2k} & & & \\ 0 & & \mathbf{A} & \\ 0 & & & \end{array}\right), \qquad \left(\begin{array}{ccc|c} & & & 0 \\ & \mathbf{A} & & 0 \\ & & & x^*_1x^*_2 \\ \hline 0 & 0 & x^*_1x^*_2 & \alpha^2(x^*_1)^2 + \alpha^2(x^*_3)^2 \end{array}\right).$$

The submatrices are not positive semidefinite by Schur complement if

$$\mathbf{A} - \frac{1}{(\hat{x}_{2k})^2 + \frac{(x^*_{2k-2})^2}{(x^*_{2k})^2}(\hat{x}_{2k})^2} \begin{bmatrix} (x^*_{2k-1})^2(x^*_{2k})^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \not\succeq 0$$

and

$$\mathbf{A} - \frac{1}{\alpha^2(x_1^*)^2 + \alpha^2(x_3^*)^2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & (x_1^*)^2(x_2^*)^2 \end{bmatrix} \not\succeq 0.$$

If $v_1 \neq 0$, then

$$\mathbf{v}^\top \mathbf{A}\mathbf{v} - \mathbf{v}^\top \frac{1}{(\hat{x}_{2k})^2 + \frac{(x_{2k-2}^*)^2}{(x_{2k}^*)^2}(\hat{x}_{2k})^2} \begin{bmatrix} (x_{2k-1}^*)^2(x_{2k}^*)^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{v} = 0 - \frac{v_1^2(x_{2k-1}^*)^2(x_{2k}^*)^2}{(\hat{x}_{2k})^2 + \frac{(x_{2k-2}^*)^2}{(x_{2k}^*)^2}(\hat{x}_{2k})^2}$$

This term is negative. Hence, the quadratic form of the Hessian has a descent direction. If $v_1 = 0$, then $v_3 \neq 0$, and we can use the second submatrix to show that the submatrix is not positive semidefinite. Thus, $\hat{\mathbf{x}}$ is not a second-order critical point. As a result, if $\hat{\mathbf{x}}$ is a second-order critical point, then $\mathbf{A}$ must be positive definite. In that case, however, we can show that the submatrix

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix}$$

cannot be positive semidefinite, which implies that $\hat{\mathbf{x}}$ cannot be a second-order critical point. We prove the last proposition by showing that $\mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1}\mathbf{B} \not\succeq 0$ by using Schur complement idea. Note that

$$\mathbf{B}^\top \mathbf{A}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{A}_{3,3}^{-1}(x_1^*)^2(x_2^*)^2 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

We calculate $\mathbf{A}_{3,3}^{-1}$ by using the cofactors of $\mathbf{A}$ as follows:

$$\mathbf{A}_{3,3}^{-1} = \frac{1}{\det(\mathbf{A})}\Big((x_{2k-1}^*)^2(x_{2k}^*)^2 + \frac{(x_{2k-1}^*)^2(x_{2k}^*)^4}{(x_1^*)^2} - (x_{2k}^*)^2(x_0^*)^2\Big),$$

where

$$\det(\mathbf{A}) = \frac{(x_{2k-1}^*)^2(x_{2k}^*)^2}{(\hat{x}_{2k})^2}\Big(\frac{(x_1^*)^4(x_2^*)^2}{(x_{2k}^*)^2} + (x_1^*)^2(x_2^*)^2 - (x_0^*)^2(x_1^*)^2\Big) - \frac{(x_0^*)^2(x_1^*)^4(x_2^*)^2}{(\hat{x}_{2k})^2}.$$

An algebraic manipulation shows that $\frac{d\mathbf{A}_{3,3}^{-1}}{d(x_0^*)^2} > 0$. Thus, the value of $\mathbf{A}_{3,3}^{-1}$ is minimized when $(x_0^*)^2 \to 0$. Whenever $x_0^* = 0$, $\mathbf{A}_{3,3}^{-1}(x_1^*)^2(x_2^*)^2$ is equal to $\alpha^2(x_1^*)^2$. As a result, $\mathbf{A}_{3,3}^{-1}(x_1^*)^2(x_2^*)^2 >$

$\alpha^2(x_1^*)^2$ by the non-zero assumption on $u_0$. We can write the matrix $\mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1}\mathbf{B}$ as the summation of $m-1$ matrices as

$$\mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1}\mathbf{B} = \begin{bmatrix} \alpha^2(x_1^*)^2 - \mathbf{A}_{3,3}^{-1}(x_1^*)^2(x_2^*)^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} + \begin{bmatrix} \alpha^2(x_3^*)^2 & x_2^*x_3^* & \dots & 0 & 0 \\ x_2^*x_3^* & \alpha^{-2}(x_2^*)^2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & (\alpha^2)^{(-1)^{m-1}}(x_m^*)^2 & x_{m-1}^*x_m^* \\ 0 & 0 & \dots & x_{m-1}^*x_m^* & (\alpha^2)^{(-1)^{m-2}}(x_{m-1}^*)^2 \end{bmatrix}.$$

Consider the vector $\mathbf{v}$ defined as $v_{2t+1} = \frac{x_{2t+2}^*}{x_2^*}$ and $v_{2t} = \alpha^2 \frac{x_{2t+1}^*}{x_2^*}$ for $t = 0, \dots, \lfloor (m-1)/2 \rfloor$ so that $\mathbf{v}^\top(\mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1}\mathbf{B})\mathbf{v}$ results in zero for all the matrices above except for the first one. Then,

$$\mathbf{v}^\top(\mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1}\mathbf{B})\mathbf{v} = \alpha^2(x_1^*)^2 - \mathbf{A}_{3,3}^{-1}(x_1^*)^2(x_2^*)^2 + 0 + \cdots + 0 < 0.$$

As a result, $\mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1}\mathbf{B}$ is not positive definite. Hence, the Hessian is not positive definite either. Consequently, the problem cannot have a second-order critical point $\hat{\mathbf{x}}$ with $\hat{x}_i^* = 0$ for some $i \in [n]$.

Any solution that meets the requirements of second-order necessary optimality conditions cannot contain any zero entries. In addition, if a stationary point $\hat{\mathbf{x}}$ satisfies $\hat{x}_i\hat{x}_j = x_i^*x_j^*$ for some $(i,j) \in \mathcal{E}$, then $\hat{x}_i\hat{x}_j = x_i^*x_j^*$ for all $(i,j) \in \mathcal{E}$ due to the stationarity condition. Therefore, a spurious solution has the following properties: $\hat{x}_i\hat{x}_j \neq x_i^*x_j^*$ for all $(i,j) \in \mathcal{E}$ and $\hat{x}_i \neq 0$ for all $i = 0, \dots 2k$.

Define $a_{i,j} = x_ix_j - x_i^*x_j^*$. Then, the stationarity condition for the B-M factorized problem (3.25) can be written as

$$[\nabla f(\hat{\mathbf{x}})]_i = \begin{cases} \hat{a}_{0,2k}\hat{x}_{2k} + \hat{a}_{0,1}\hat{x}_1 = 0, & \text{if } i = 0 \\ \hat{a}_{i-1,i}\hat{x}_{i-1} + \hat{a}_{i,i+1}\hat{x}_{i+1} = 0, & \text{if } i \notin \{0, 2k\} \\ \hat{a}_{2k-1,2k}\hat{x}_{2k-1} + \hat{a}_{0,2k}\hat{x}_0 = 0, & \text{if } i = 2k \end{cases}.$$

The following calculation yields a contradiction because none of the terms can be zero:

$$\sum_{i=0}^{2k}(-1)^i[\nabla f(\hat{\mathbf{x}})]_i = 2\hat{a}_{0,2k}\hat{x}_0\hat{x}_{2k} = 0.$$

As a result, all the second-order critical points are global solutions, and we obtain the valid factors for the ground truth matrix. □

## 3.G Proof of Theorem 12

*Proof.* We aim to find an instance of the matrix completion problem with an odd-numbered cycle graph for which the SDP problem (3.2) fails. Consider the following rank-1 feasible solution $\hat{\mathbf{M}} = \mathbf{y}\mathbf{y}^\top$ with

$$y_0 = (\hat{\mathbf{M}}_{0,0})^{1/2} \geq x_0^*, \quad y_{2t-1} = \left(\frac{(x_{2t-1}^*)^2 (x_0^*)^2}{\hat{\mathbf{M}}_{0,0}}\right)^{1/2}, \quad y_{2t} = \left(\frac{(x_{2t}^*)^2}{(x_0^*)^2}\hat{\mathbf{M}}_{0,0}\right)^{1/2},$$

where $t = 1, \ldots, k$. Note that $\hat{\mathbf{M}}_{0,0}$ is free to choose, and we can minimize the trace to find the best rank-1 solution. This is equivalent to the following optimization problem:

$$\min_{\hat{\mathbf{M}}_{0,0} \geq (x_0^*)^2} \frac{\hat{\mathbf{M}}_{0,0}}{(x_0^*)^2} \sum_{t=0}^{k} (x_{2t}^*)^2 + \frac{(x_0^*)^2}{\hat{\mathbf{M}}_{0,0}} \sum_{t=1}^{k} (x_{2t-1}^*)^2.$$

Note that whenever $\hat{\mathbf{M}}_{0,0} = (x_0^*)^2$, the solution $\hat{\mathbf{M}}$ is the same as ground truth solution. A basic first-order optimality condition implies that whenever $\sum_{t=1}^{k} (x_{2t-1}^*)^2 > \sum_{t=0}^{k} (x_{2t}^*)^2$, the optimal $\hat{\mathbf{M}}_{0,0}$ is strictly larger than $(x_0^*)^2$. However, the objective value is strictly better than the trace of the ground truth matrix. Due to the symmetry of the problem, any node can be chosen as node 0. Therefore, if the condition $\sum_{t=1}^{k} (x_{2t-1}^*)^2 > \sum_{t=0}^{k} (x_{2t}^*)^2$ holds for some chosen node $0$, the SDP fails to solve the matrix completion problem. □

# Part II

# System Identification under Adversarial Disturbances

# Chapter 4

# System Identification for Linear Systems

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 is 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 [26]. 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 [72, 68]. Nonetheless, the non-asymptotic analysis of the system identification problem has gained interest in recent years [53, 78, 94, 107]. Non-asymptotic analysis is crucial to understanding the required sample complexity for online control problems.

Robust learning of dynamical systems is crucial for safety-critical applications, such as autonomous driving [2], unmanned aerial vehicles [113], and robotic arms [61]. 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 [102, 101, 131]. 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 [136] 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 [6]. We study the required sample complexity for the exact recovery of LTI systems using these estimators when there are sporadic large disturbance injections.

The robust regression and learning problems under adversaries are ubiquitously studied in the literature [116, 5, 9, 87]. $\ell_1$-based non-smooth estimators are extensively investigated in the context of robust learning in the presence of adversaries and outliers [38, 30, 46]. Compressed sensing and sparse error detection are such problems that are closely related to the proposed estimators [36, 35, 21, 20]. However, existing methods for analyzing the estimators cannot be directly generalized to control problems due to the sample correlation. 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 [32]. 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 [79, 102, 69]. It employs normalized martingale bounds for the estimation error when the excitation is large enough [102].

Unlike the non-asymptotic analysis of correlated data, the least-squares estimator offers a closed-form solution [41, 56, 110]. The least-squares estimator performs relatively well as long as the noise magnitudes are not large. 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 [102]. However, it is not robust to adversarial attacks, and the literature on robust learning of dynamical systems is limited. The work [117] 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 [44] 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 [44] and study estimators from a non-asymptotic point of view under standard assumptions, such as the stable system and sub-Gaussian attacks.

**Contributions:** We consider an LTI dynamical system over the time horizon $[0, T]$, $\mathbf{x}_{i+1} = \bar{\mathbf{A}}\mathbf{x}_i + \bar{\mathbf{B}}\mathbf{u}_i + \bar{\mathbf{d}}_i, i = 0, 1, \ldots, T-1$, where $\bar{\mathbf{A}} \in \mathbb{R}^{n\times n}$ and $\bar{\mathbf{B}} \in \mathbb{R}^{n\times m}$ are unknown system matrices, and $\bar{\mathbf{d}}_i \in \mathbb{R}^n$ are unknown system disturbances. We aim to learn these matrices from the samples $\{\mathbf{x}_i\}_{i=0}^{T}$ and $\{\mathbf{u}_i\}_{i=0}^{T-1}$ of a single initialization of the system when the disturbance vectors $\bar{\mathbf{d}}_i$ are adversarial. Here, the adversarial noise refers to a vector designed to deteriorate the estimator's performance. Thus, the adversarial vectors $\{\bar{\mathbf{d}}_i\}_{i=0}^{T-1}$ can take arbitrarily large finite values, be dependent over time, and have undesirable structures. We say that an adversarial attack occurs whenever $\bar{\mathbf{d}}_i$ is non-zero, and we have no information on the value of $\bar{\mathbf{d}}_i$. If $\bar{\mathbf{d}}_i$ 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 $\ell_2$ and $\ell_1$ norms of the estimated disturbance vectors, $\sum_{i=0}^{T-1} \|\mathbf{d}_i\|_2$ and $\sum_{i=0}^{T-1} \|\mathbf{d}_i\|_1$, with the decision variables $\mathbf{A}$, $\mathbf{B}$, and $\{\mathbf{d}_i\}_{i=0}^{T-1}$ subject to $\mathbf{x}_{i+1} = \mathbf{A}\mathbf{x}_i + \mathbf{B}\mathbf{u}_i + \mathbf{d}_i$, given the samples $\{\mathbf{x}_i\}_{i=0}^{T}$ and $\{\mathbf{u}_i\}_{i=0}^{T-1}$:

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

We employ a non-smooth objective function to obtain a robust estimator. The arbitrary injection of adversaries may happen infrequently. In that case, the attacks occur sparsely in time. Conversely, the vector $\bar{\mathbf{d}}_i$ at each attack time $i$ could be dense, and there is no limitation on how sparse the vector

is. The $\ell_2$ norm estimator is the most effective in this case. In contrast, the $\ell_1$ norm estimator is preferable if the vector $\bar{\mathbf{d}}_i$ 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 $\Delta$. 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 $T^1_{\text{sample}}$ and $T^2_{\text{sample}}$ 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 16. 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.

## 4.1 Preliminaries

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 10.** *(Hoeffding's Bound [111]) Suppose that the variable $X$ has mean $\mu$ and sub-Gaussian parameter $\sigma$. 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 $\ell_2$ norm of 0 vector is the $\ell_2$ norm unit ball, whereas the subdifferential of the $\ell_1$ norm of 0 vector is the $\ell_\infty$ norm unit ball, which is $\mathbb{B}_\infty(1) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_\infty \leq 1\}$. Note that while the subdifferential of the $\ell_1$ norm is coordinate-wise separable, the subdifferential of the $\ell_2$ norm is not coordinate-wise separable. We also define the unit sphere $\mathbb{S}_2(1)$ as $\mathbb{S}_2(1) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$, that is the set of all the points on the sphere with radius 1.

## 4.2 Problem Formulation

We assume that the disturbance vectors $\{\bar{\mathbf{d}}_i\}_{i=0}^{T-1}$ 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 $\mathcal{K} = \{i : \bar{\mathbf{d}}_i \neq \mathbf{0}, i \in 0, 1, \ldots, T-1\}$. 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 $\mathcal{K}^c$, which are called the no-attack times. The data corresponding to attack times are corrupted, whereas those 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 $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$, we analyze the following convex optimization problems with non-smooth objective functions:

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times n},\mathbf{B}\in\mathbb{R}^{n\times m}} \sum_{i=0}^{T-1} \|\mathbf{x}_{i+1} - \mathbf{A}\mathbf{x}_i - \mathbf{B}\mathbf{u}_i\|_2 \tag{CO-L2}$$

and

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times n},\mathbf{B}\in\mathbb{R}^{n\times m}} \sum_{i=0}^{T-1} \|\mathbf{x}_{i+1} - \mathbf{A}\mathbf{x}_i - \mathbf{B}\mathbf{u}_i\|_1 \tag{CO-L1}$$

where the states $\{\mathbf{x}_i\}_{i=0}^{T}$ are generated according to $\mathbf{x}_{i+1} = \bar{\mathbf{A}}\mathbf{x}_i + \bar{\mathbf{B}}\mathbf{u}_i + \bar{\mathbf{d}}_i, \quad i = 0, \ldots, T-1$. The difference between problems (CO-L2) and (CO-L1) is their objective functions. In problem (CO-L2), the sum of the $\ell_2$ norm columns is analogous to the $\ell_1$ norm minimization in the lasso problem. In other words, the $\ell_1$ norm is applied at the group level to $\{\bar{\mathbf{d}}_i\}_{i=0}^{T-1}$ because the occurrence of large injections of disturbances is rare and not frequent. We highlight that the vectors $\{\bar{\mathbf{d}}_i\}_{i=0}^{T-1}$ are not necessarily sparse. On the other hand, the $\ell_1$ norm is applied both at the group level and the in-group levels to $\{\bar{\mathbf{d}}_i\}_{i=0}^{T-1}$ 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 $\mathbf{x}_i$ 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 chapter 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 $\ell_2$ and $\ell_1$ 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 $\mathbf{x}_i$ as tuples $(\mathbf{x}_{i+1}, \mathbf{x}_i, \mathbf{u}_i)$, it is impossible to claim that $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$ are independent, which is a key assumption in the empirical risk minimization literature. We can also transform this problem into a compressed sensing problem. Let $\mathbf{Y} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$, $\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}]$, and $\mathbf{U} = [\mathbf{u}_0, \ldots, \mathbf{u}_{T-1}]$ be the matrices where the state, input, and noise vectors are given as columns. Then, the problem is equivalent to

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

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 $[\mathbf{X}^\top, \mathbf{U}^\top]^\top$ are not independent of each other due to the auto-correlation between the state vectors. Therefore, the classical compressed sensing results do not apply to this system identification problem. In addition, the existing sufficient conditions, such as NSP [44], 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 13.** *Consider the convex optimization problems* (CO-L2) *and* (CO-L1) *and let* $\circ \in \{1, 2\}$. *Given a pair of matrices* $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$*, if the following conditions hold simultaneously*

$$\mathbf{0} \in \sum_{i \notin \mathcal{K}} \mathbf{x}_i \otimes \partial \|(\bar{\mathbf{A}} - \hat{\mathbf{A}})\mathbf{x}_i + (\bar{\mathbf{B}} - \hat{\mathbf{B}})\mathbf{u}_i\|_\circ + \sum_{i \in \mathcal{K}} \mathbf{x}_i \otimes \partial \|(\bar{\mathbf{A}} - \hat{\mathbf{A}})\mathbf{x}_i + (\bar{\mathbf{B}} - \hat{\mathbf{B}})\mathbf{u}_i + \bar{\mathbf{d}}_i\|_\circ, \tag{4.1}$$

$$\mathbf{0} \in \sum_{i \notin \mathcal{K}} \mathbf{u}_i \otimes \partial \|(\bar{\mathbf{A}} - \hat{\mathbf{A}})\mathbf{x}_i + (\bar{\mathbf{B}} - \hat{\mathbf{B}})\mathbf{u}_i\|_\circ + \sum_{i \in \mathcal{K}} \mathbf{u}_i \otimes \partial \|(\bar{\mathbf{A}} - \hat{\mathbf{A}})\mathbf{x}_i + (\bar{\mathbf{B}} - \hat{\mathbf{B}})\mathbf{u}_i + \bar{\mathbf{d}}_i\|_\circ, \tag{4.2}$$

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

We emphasize that $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ represent the unknown ground truth matrices, $(\mathbf{A}, \mathbf{B})$ are the decision variables in the convex optimization problems, and $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ are the solutions to those convex optimization problems. The proof for the KKT conditions when $\circ = 2$ is provided in [43], and the proof for the case $\circ = 1$ can be done similarly. We will utilize the conditions above to study the scenarios where the exact recovery is achievable. As a simple corollary to Theorem 13, we can state that $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is a solution to our estimator(s) if the following conditions hold:

$$\mathbf{0} \in \sum_{i \notin \mathcal{K}} \mathbf{x}_i \otimes \partial \|\mathbf{0}\|_\circ + \sum_{i \in \mathcal{K}} \mathbf{x}_i \otimes \partial \|\bar{\mathbf{d}}_i\|_\circ,$$
$$\mathbf{0} \in \sum_{i \notin \mathcal{K}} \mathbf{u}_i \otimes \partial \|\mathbf{0}\|_\circ + \sum_{i \in \mathcal{K}} \mathbf{u}_i \otimes \partial \|\bar{\mathbf{d}}_i\|_\circ.$$

## 4.3 Autonomous Systems

In this section, we consider autonomous systems, meaning that $\mathbf{u}_0 = \cdots = \mathbf{u}_{T-1} = \mathbf{0}$. Therefore, the system dynamics could be written as $\mathbf{x}_{i+1} = \bar{\mathbf{A}}\mathbf{x}_i + \bar{\mathbf{d}}_i$ 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 $i$, $i \in \mathcal{K}^c$, then $\bar{\mathbf{d}}_i = \mathbf{0}$. We are interested in recovering the system matrix $\bar{\mathbf{A}}$ using the following convex optimization problems for autonomous systems:

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times n}} \sum_{i=0}^{T-1} \|\mathbf{x}_{i+1} - \mathbf{A}\mathbf{x}_i\|_2 \tag{CO-L2-Aut}$$

and

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times n}} \sum_{i=0}^{T-1} \|\mathbf{x}_{i+1} - \mathbf{A}\mathbf{x}_i\|_1 \tag{CO-L1-Aut}$$

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 13:

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

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

We examine two types of attack structures: $\Delta$-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 $\mathcal{K}$.

## $\Delta$-spaced Attack Structure

The first attack structure is a deterministic attack model for which the attacks occur at every $\Delta$ time period. For instance, if $\Delta = 2$, the set $\mathcal{K}$ 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 [43].

**Definition 8** ($\Delta$-spaced Attack Structure)**.** *Given a positive integer $\Delta > 2$, the disturbance sequence $\{\bar{\mathbf{d}}_i\}_{i=0}^{T-1}$ is said to be $\Delta$-spaced if for every $i \in \{0, 1, \ldots, T-\Delta-1\}$ such that $\bar{\mathbf{d}}_i \neq \mathbf{0}$, we have $\bar{\mathbf{d}}_j = \mathbf{0}$, for all $j \in \{i+1, \ldots, i+\Delta-1\}$ and $\bar{\mathbf{d}}_{i+\Delta} \neq \mathbf{0}$. In addition, for $i \in \{0, 1, \ldots, \Delta-1\}$, we must have at least one non-zero disturbance vector, i.e. $\bar{\mathbf{d}}_i \neq \mathbf{0}$.*

Initially, we consider first-order and 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 $\Delta$-spaced disturbance sequence with $\Delta \geq 2$.

**Proposition 5.** *Consider a first-order autonomous system with $|\bar{A}| < 1$ and $\Delta$-spaced disturbance sequence with $\Delta \geq 2$. Then, whenever $x_0 = 0$ or $0 \notin \mathcal{K}$, the convex formulation* (CO-L2-Aut) *has the unique solution $\bar{\mathbf{A}}$ as long as the sample complexity satisfies the inequality $T \geq \Delta + 1$.*

This proposition implies that the exact recovery is guaranteed to be achieved whenever there are more than $\Delta + 1$ data samples. Note that Proposition 5 does not make any assumption on the vector set $\{\bar{\mathbf{d}}_i : i \in \mathcal{K}\}$ 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 5 is when there is a $\Delta$-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 5 is based on the following lemma.

**Lemma 11.** *(Theorem 1 in [43]) Consider the convex optimization problem* (CO-L2-Aut). *If $\sum_{i \notin \mathcal{K}} |x_i| > \sum_{i \in \mathcal{K}} |x_i|$, then $\bar{\mathbf{A}}$ is the unique solution to the problem.*

A natural question arises about whether one can generalize the above result to higher-order systems. The next proposition extends Proposition 5 to autonomous dynamical systems with an arbitrary order $n$ under a $\Delta$-spaced disturbance sequence with $\Delta \geq n + 1$.

**Proposition 6.** *Consider an autonomous system of order $n$ under a $\Delta$-spaced disturbance sequence with $\Delta \geq n + 1$. Suppose that $\bar{\mathbf{A}}$ is diagonalizable with eigenvalues, $\bar{\lambda}_l, l = 1, 2, \ldots, n$, and that the condition*

$$\operatorname{span}\{\bar{\mathbf{d}}_i, \bar{\mathbf{A}}\bar{\mathbf{d}}_i, \ldots, \bar{\mathbf{A}}^{n-1}\bar{\mathbf{d}}_i\} \in \mathbb{R}^n, \quad \forall i \in \mathcal{K} \tag{4.3}$$

*is satisfied. Then, whenever $\mathbf{x}_0 = \mathbf{0}$ or $\{0, \ldots, \Delta - 1\} \notin \mathcal{K}$, $\bar{\mathbf{A}}$ is a solution to the convex formulation* (CO-L2-Aut) *if $T \geq n + \Delta$, provided that*

$$\left| \sum_{k_1 + \cdots + k_n = \Delta - n} \bar{\lambda}(k_1, \ldots, k_n) \right| \leq \sum_{i=0}^{\Delta - n - 1} \left| \sum_{k_1 + \cdots + k_n = i} \bar{\lambda}(k_1, \ldots, k_n) \right|, \tag{4.4}$$

*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 generalizes Proposition 5. The condition (4.3) 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 into the system. To gain insight into equation

| $C_{n,k}$ | 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 4.1: Upper-Bound Value $C_{n,k}$ for Different Values of $n$ and $k$.

(4.4), which involves the product of eigenvalues, consider a special case where $\bar{\mathbf{A}}$ has the eigenvalue $\lambda$ with multiplicity $n$ with $n$ linearly independent eigenvectors. In this case, we can simplify (4.4) as follows. Define $k := \Delta - n$. Then, (4.4) 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 $k$. Figure 4.1 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 system's stability 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 6 can still be applied to problem (CO-L1-Aut). However, the KKT conditions will differ due to the subdifferential of the $\ell_2$ and $\ell_1$ norms. 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 $\ell_1$ norm of the disturbance vectors is used.

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

## Probabilistic Attack Structure

Because the minimum value of $\Delta$ 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 $\bar{\mathbf{d}}_i$ is injected into the system at time instance $i$ with probability $p > 0$, independent of the other time periods. Specifically, given a time instance $i$, $\bar{\mathbf{d}}_i$ is

non-zero with probability $p$, 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 3.** *Given an autonomous system $\mathbf{x}_{i+1} = \bar{\mathbf{A}}\mathbf{x}_i + \bar{\mathbf{d}}_i$ for $i = 0, \ldots, T-1$ with dimension $n$, assume that $\mathbf{x}_0 = \mathbf{0}$ and all singular values of $\bar{\mathbf{A}}$ are less than 1, i.e. $\|\bar{\mathbf{A}}\| < 1$.*

The stability assumption is standard in system identification problems to avoid 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 4.** *For each $k \in \mathcal{K}$, the attack vector is defined as $\bar{\mathbf{d}}_k := \bar{\ell}_k \bar{\mathbf{f}}_k$ where $\bar{\ell}_k \in \mathbb{R}$ and $\bar{\mathbf{f}}_k \in \mathbb{S}_2(1)$. $\bar{\mathbf{f}}_k$ plays the role of the direction of the attack while $\bar{\ell}_k$ plays the role of the length. Define the filtration*

$$\mathcal{F}_k := \sigma\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}, \quad \forall k \in \{0, \ldots, T-1\}.$$

*For all $k \in \mathcal{K}$, conditioning on $\mathcal{F}_k$, the following statements hold:*

1. *$\bar{\ell}_k$ is independent from the direction $\bar{\mathbf{f}}_k$;*
2. *The direction $\bar{\mathbf{f}}_k$ obeys the uniform distribution on $\mathbb{S}_2(1)$;*
3. *$\bar{\ell}_k$ is mean-zero and sub-Gaussian with parameter $\sigma$;*
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 $\bar{\mathbf{d}}_{k'}$, and in particular $\bar{\ell}_{k'}$ and $\bar{\mathbf{f}}_{k'}$ for $k' < k$. For example, a stealthy attack vector $\bar{\mathbf{d}}_i$ could have the distribution $\mathcal{N}(0, \min\{c, \widehat{\|\mathbf{x}_i\|_2}\})$ where $\widehat{\|\mathbf{x}_i\|_2} = \sum_{k=0}^{i} (i+1)^{-1} \|\mathbf{x}_i\|_2$ is the average norm of the states between time periods 0 and $i$. The magnitude of the attack vector is dependent on the past. The above symmetry assumption of the disturbance vectors reflected in $\bar{\mathbf{f}}_k$ 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 [104, 90]. 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 anomalies 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 [28].

It is impossible to obtain deterministic sample complexity for exact recovery due to the randomness in the attack structure. Therefore, it is essential to quantify the required number of samples for the exact recovery with high probability. Under Assumption 4, the attack vector at time $i$, $\bar{\mathbf{d}}_i$, has a sub-Gaussian distribution with parameter $\sigma$ given $\mathcal{F}_i$. 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

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

because of the properties of the subdifferentials at the origin. To simplify the analysis, we use the relationship between the unit balls of the $\ell_\infty$ and $\ell_2$ 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 $\ell_\infty$ norm. Therefore, the following propositions provide sufficient conditions to satisfy the KKT conditions.

**Proposition 7.** *The KKT conditions for the problem* (CO-L2-Aut) *and* (CO-L1-Aut) *are satisfied if there exist scalars* $\gamma_i^l \in [-1, 1], i \notin \mathcal{K}, l = 1, \ldots, n$ *such that*

$$\sum_{i \notin \mathcal{K}} \gamma_i^l \mathbf{x}_i / \sqrt{n} = \sum_{i \in \mathcal{K}} \partial \|\bar{\mathbf{d}}_i\|_2^l \mathbf{x}_i, \quad \forall l = 1, \ldots, n \tag{4.5}$$

*and*

$$\sum_{i \notin \mathcal{K}} \gamma_i^l \mathbf{x}_i = \sum_{i \in \mathcal{K}} \partial \|\bar{\mathbf{d}}_i\|_1^l \mathbf{x}_i, \quad \forall l = 1, \ldots, n. \tag{4.6}$$

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

Because analyzing the conditions (4.5) and (4.6) 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 12.** *Given a matrix* $\mathbf{F} \in \mathbb{R}^{n \times m}$ *and the vector* $\mathbf{g} \in \mathbb{R}^n$*, the following statements are equivalent:*

*i) There exists a vector* $\mathbf{w} \in \mathbb{R}^m$ *with* $\|\mathbf{w}\|_\infty \leq 1$ *satisfying* $\mathbf{F}\mathbf{w} = \mathbf{g}$.

*ii) For every* $\mathbf{z} \in \mathbb{R}^n$ *with* $\|\mathbf{z}\|_2 = 1$*, it holds that* $f(\mathbf{z}) := \mathbf{z}^\top \mathbf{g} + \|\mathbf{z}^\top \mathbf{F}\|_1 \geq 0$.

It is important to notice that the conditions (4.5) and (4.6) amount to finding a vector for the set of equations in the form of $\mathbf{F}\mathbf{w} = \mathbf{g}$ where $\mathbf{w}$ is restricted as $\|\mathbf{w}\|_\infty \leq 1$. Given a coordinate $l$, the matrix $\mathbf{F} \in \mathbb{R}^{n \times (T - |\mathcal{K}|)}$ associated with the conditions (4.5) and (4.6) is a matrix with columns $\frac{\mathbf{x}_i}{\sqrt{n}}$ and $\mathbf{x}_i$, and the vector $\mathbf{g} \in \mathbb{R}^n$ is $\sum_{i \in \mathcal{K}} \partial \|\bar{\mathbf{d}}_i\|_2^l \mathbf{x}_i$ and $\sum_{i \in \mathcal{K}} \partial \|\bar{\mathbf{d}}_i\|_1^l \mathbf{x}_i$, respectively. Moreover, the vector $\mathbf{w} \in \mathbb{R}^{T - |\mathcal{K}|}$ has the elements $\gamma_i^l, i \notin \mathcal{K}$ for both conditions. Hence, we study the condition ii) in Lemma 12. However, infinitely many points are on the $\ell_2$ unit circle $\mathbb{S}_2(1)$. To show that the function $f(\mathbf{z}) = \mathbf{z}^\top \mathbf{g} + \|\mathbf{z}^\top \mathbf{F}\|_1$ is non-negative at every point on the $\ell_2$ unit circle, we employ the discretization technique that chooses a finite set of points. These points are chosen as the $\epsilon$-cover of the unit ball.

**Definition 9** (Covering Number [111]). *$\epsilon$-cover of the compact set $\mathbb{T}$ with respect to the norm $\rho$ 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 $\epsilon$-covering number $\mathcal{N}(\epsilon, \mathbb{T}, \rho)$ is the cardinality of the smallest $\epsilon$-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 $\epsilon$ balls exceeding $\exp\{n \log(1 + 2/\epsilon)\}$ is sufficient to cover the unit ball with balls of radius $\epsilon$.

**Lemma 13** (Covering Number of the Unit Ball [111]). *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*

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

We show that the function $f(\mathbf{z})$ can be lower bounded by some positive number $\theta > 0$ at every point in the $\epsilon$-cover of the unit ball with high probability and that the function value inside the $\epsilon$-ball does not change more than this positive number $\theta$ with high probability. Thus, $f(\mathbf{z})$ 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 14.** *Consider an autonomous system of order $n$ under a probabilistic attack model with frequency $p$. Suppose that Assumptions 3 and 4 hold. Then, for all $\delta \in (0, 1]$, if the time horizon satisfies $T \geq \Theta(T_{sample})$, where $T_{sample}$ is defined as*

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

*and*

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

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

The above theorem implies 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 uses problem (CO-L1-Aut) to recover the system matrix $\bar{\mathbf{A}}$.

**Theorem 15.** *Under the assumptions of Theorem 14, if the time horizon $T$ satisfies $T \geq \Theta(T_{sample})$, where $T_{sample}$ is*

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

*and $R$ is defined in Theorem 14, then $\bar{\mathbf{A}}$ is the unique solution to the convex optimization problem* (CO-L1-Aut) *with probability at least $1-\delta$.*

The proof of Theorem 15 is highly similar to that of Theorem 14. Because the conditions (4.5) and (4.6) 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 on the verge of instability, the sample complexity increases significantly. Even 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 sufficient clean samples. We make the following remarks regarding various generalizations of the above results.

**Remark.** *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 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 $\mathbf{x}_{i+1} = \bar{\mathbf{A}}\mathbf{x}_i$. If $\mathbf{x}_0$ is zero, then $\mathbf{x}_i$ will always remain zero, preventing us from identifying $\bar{\mathbf{A}}$. To avoid this, we must excite the system as $\mathbf{x}_{i+1} = \bar{\mathbf{A}}\mathbf{x}_i + \mathbf{w}_i$, where $\mathbf{w}_i$ 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.** *When the system is initialized randomly at $\mathbf{x}_0$ following Assumption 4, the results of Theorems 14 and 15 continue to hold. A system with a non-zero initial state $\mathbf{x}_0$ at time $0$ is equivalent to the system initialized at the origin at time $-1$ under the attack $\mathbf{x}_0$ at time $-1$, namely, the new system could be constructed as $\mathbf{x}_{-1} = \mathbf{0}$ and $\bar{\mathbf{d}}_{-1} = \mathbf{x}_0$, leading to the state value $\mathbf{x}_0$ at time $0$. Thus, the sample complexities of the theorems only shift by a constant.*

**Remark.** *Furthermore, we mention that the uniform distribution assumption of $\bar{\mathbf{f}}_k$ 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 14-17 will depend on the conditional number of the covariance matrix.*

## 4.4 Systems with Input Sequence

It is desirable to understand the role of an input sequence in exact recovery. Since a controller generates the input sequence, one can design it in such a way that it accelerates the learning. In this case, the system dynamics is given as $\mathbf{x}_{i+1} = \bar{\mathbf{A}}\mathbf{x}_i + \bar{\mathbf{B}}\mathbf{u}_i + \bar{\mathbf{d}}_i, i = 0, \ldots, T-1$, where $\bar{\mathbf{A}} \in \mathbb{R}^{n\times n}$ and $\bar{\mathbf{B}} \in \mathbb{R}^{n\times m}$. 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 $\mathbf{u}_i$ to be Gaussian given $\mathcal{F}_i$. 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 $\mathbf{u}_i$ is designed in the linear feedback form as $\mathbf{u}_i = \mathbf{K}\mathbf{x}_i + \mathbf{\Omega}$. Conditioning on $\mathcal{F}_i$, the input is excited with Gaussian noise $\mathbf{\Omega}$. Note that the closed-loop system could be written as $\mathbf{x}_{i+1} = (\bar{\mathbf{A}} + \bar{\mathbf{B}}\mathbf{K})\mathbf{x}_i + \bar{\mathbf{B}}\mathbf{\Omega} + \bar{\mathbf{d}}_i$. Thus, the problem is equivalent to estimating the matrices $(\bar{\mathbf{A}} + \bar{\mathbf{B}}\mathbf{K})$ and $\bar{\mathbf{B}}$ when the linear feedback control is used. Therefore, the most common input sequence used in optimal control satisfies this assumption. Similar to Proposition 7, sufficient conditions can be tightened so that the equations become coordinate-wise separable.

**Proposition 8.** *The KKT conditions for problem* (CO-L2) *are satisfied if there exist scalars* $\gamma_i^l, \mu_i^l \in [-1, 1]$ *for all* $i \notin \mathcal{K}, l \in \{1, \ldots, n\}$ *such that*

$$\sum_{i \notin \mathcal{K}} \gamma_i^l \mathbf{x}_i / \sqrt{n} = \sum_{i \in \mathcal{K}} \partial \|\bar{\mathbf{d}}_i\|_2^l \mathbf{x}_i, \quad \forall l = 1, \ldots, n, \tag{4.7}$$

*and*

$$\sum_{i \notin \mathcal{K}} \mu_i^l \mathbf{u}_i / \sqrt{n} = \sum_{i \in \mathcal{K}} \partial \|\bar{\mathbf{d}}_i\|_2^l \mathbf{u}_i, \quad \forall l = 1, \ldots, n, \tag{4.8}$$

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

The proof of Proposition 8 relies on the same technique as in Proposition 7. Similar to autonomous systems, we can omit the factor $\sqrt{n}$ from the above equations to guarantee the satisfaction of the KKT conditions for problem (CO-L1). We require the following controllability assumption.

**Assumption 5.** *The ground truth* $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ *satisfies*

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

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 $\{\mathbf{u}_i\}_{i=0}^{T-1}$. Controllability is an important property of a control system and plays a crucial role in many control problems, such as the 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 14.

**Theorem 16.** *Consider an autonomous system of order* $n$ *under a probabilistic attack model with frequency* $p$*. Suppose that Assumptions 3, 4, and 5 hold. Assume also that the input vectors* $\mathbf{u}_i | \mathcal{F}_i$ *are selected to be independent of the attack vectors and obey the Gaussian distribution* $\mathcal{N}(0, \frac{\xi^2}{m}\mathbf{I}_m)$*. For all* $\delta \in (0, 1]$*, let*

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

*and*

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

*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 $m$, $n$, $\sigma$, $\xi$ and $\bar{\mathbf{B}}$. If the time horizon satisfies the inequality $T \geq \Theta[\max\{T^1_{\textit{sample}}, T^2_{\textit{sample}}\}]$, then $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is the unique solution to* (CO-L2) *with probability at least $1-\delta$.*

We have obtained a high probability bound for the exact recovery of the system matrices $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$. The $T^1_{\text{sample}}$ in the sample complexity corresponds to the satisfaction of the KKT conditions for the state measurements, whereas the $T^2_{\text{sample}}$ corresponds to the satisfaction of the KKT conditions for the input sequence. 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 $\bar{\mathbf{A}}$. Introducing 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 $\bar{\mathbf{A}}$ 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 17.** *Under the assumptions of Theorem 16, for all $\delta \in (0,1]$, let $T^1_{\textit{sample}}$ and $T^2_{\textit{sample}}$ 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 16. If $T$ satisfies the inequality*

$$T \geq \Theta[\max\{T^1_{\textit{sample}}, T^2_{\textit{sample}}\}]$$

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

**Remark.** *When the input sequence $\mathbf{u}_i = \mathbf{K}\mathbf{x}_i$ is used to control the system, the closed-loop system with the matrix $(\bar{\mathbf{A}} + \bar{\mathbf{B}}\mathbf{K})$ results in a second solution $\hat{\mathbf{A}} = \bar{\mathbf{A}} + \bar{\mathbf{B}}\mathbf{K}$ and $\hat{\mathbf{B}} = \mathbf{0}$. Still, the ground-truth system matrix pair $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ 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* $\mathbf{u}_i = \mathbf{K}\mathbf{x}_i + \mathbf{\Omega}$. *Moreover, if all the input vectors* $\mathbf{u}_i$ *are set to zero, it is not possible to uniquely recover the system matrix* $\bar{\mathbf{B}}$. *Because the input sequence is zero, the KKT conditions are trivially satisfied; thus, we have multiple optimum solutions.*

**Remark.** *The results in Sections IV and V can be extended to the case where there is a small-in-magnitude dense measurement noise,* $\mathbf{e}_i$*, in addition to potentially large-in-magnitude adversarial noise* $\mathbf{d}_i$. *Note that the estimator can be written as a constrained optimization problem as*

$$\begin{aligned} \min_{\substack{\mathbf{A}\in\mathbb{R}^{n\times n},\mathbf{B}\in\mathbb{R}^{n\times m},\\ \mathbf{d}_i\in\mathbb{R}^n,\forall i}} \quad & \sum_{i=0}^{T-1} \|\mathbf{d}_i\|_\circ \\ s.t. \quad & \mathbf{x}_{i+1} - \mathbf{A}\mathbf{x}_i + \mathbf{B}\mathbf{u}_i + \mathbf{d}_i = \mathbf{0}, \quad i = 0,\dots,T-1 \end{aligned}$$

*Adding the dense measurement vector* $\mathbf{e}_i$ *is equivalent to perturbing the constraint from* $\mathbf{x}_{i+1} - \mathbf{A}\mathbf{x}_i - \mathbf{B}\mathbf{u}_i - \mathbf{d}_i = \mathbf{0}$ *to* $\mathbf{x}_{i+1} - \mathbf{A}\mathbf{x}_i - \mathbf{B}\mathbf{u}_i - \mathbf{d}_i = \mathbf{e}_i$. *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 [133]).*

## 4.5 Numerical Experiments

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

### Synthetic Simulations

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

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

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

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

For any time period $t = 1, \dots, T$, we obtain the solution gap, $\|(\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t) - (\bar{\mathbf{A}}, \bar{\mathbf{B}})\|_F$, and the loss gap, $f_t(\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t) - f_t(\bar{\mathbf{A}}, \bar{\mathbf{B}})$. We plot the trajectory of the average solution gap and the loss gap

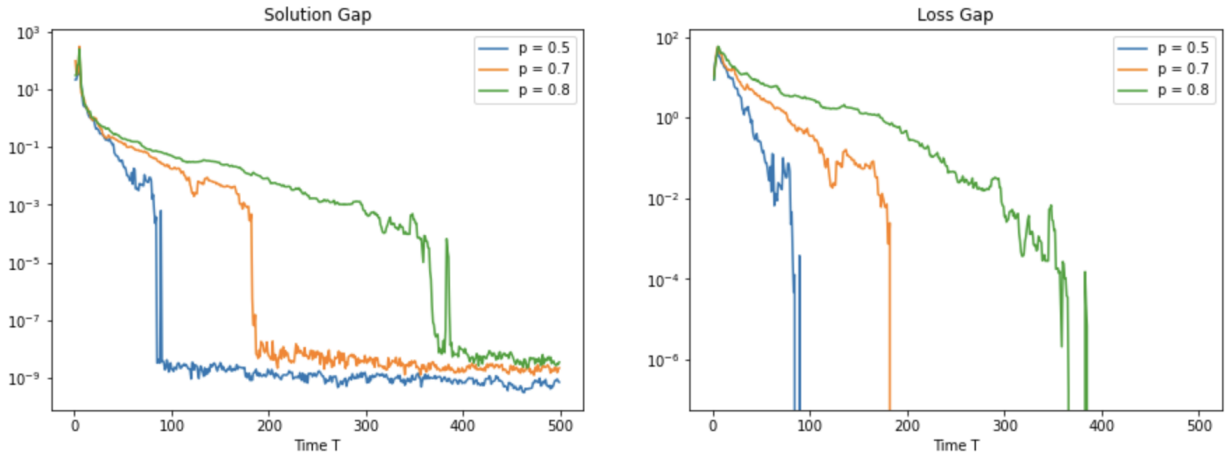


Figure 4.2: Performance of (CO-L2-Aut) with Probability of Attacks $p \in \{0.5, 0.7, 0.8\}$ with $n = 5$

of the 10 independent simulation runs. First, we analyze the performance of the estimator with respect to the probability of having an attack on the autonomous systems. We use three different values of $p \in \{0.5, 0.7, 0.8\}$. In Figure 4.2, 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 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 Figure 4.3 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 $(n, m) \in \{(5, 5), (10, 10), (15, 15)\}$ for (CO-L2). In Figure 4.4, 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 scale with $n^{-1}$. The smaller empirical sample complexity hints at 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 $\ell_2$, $\ell_1$ norms, and the least-squares method. Figures 4.5 and 4.6 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 $\ell_2$ and $\ell_1$ norm estimators successfully learn the ground truth system matrices. Since the attack vectors themselves are not sparse, the $\ell_2$ estimator requires fewer samples to achieve exact recovery than the $\ell_1$ estimator.

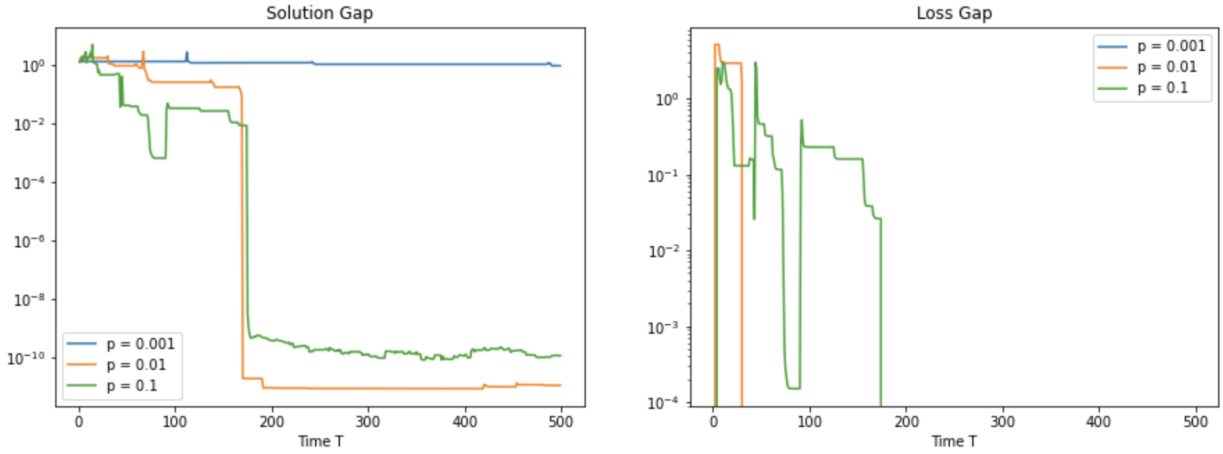


Figure 4.3: Performance of (CO-L2-Aut) with Probability of Attacks $p \in \{0.001, 0.01, 0.1\}$ with $n = 5$

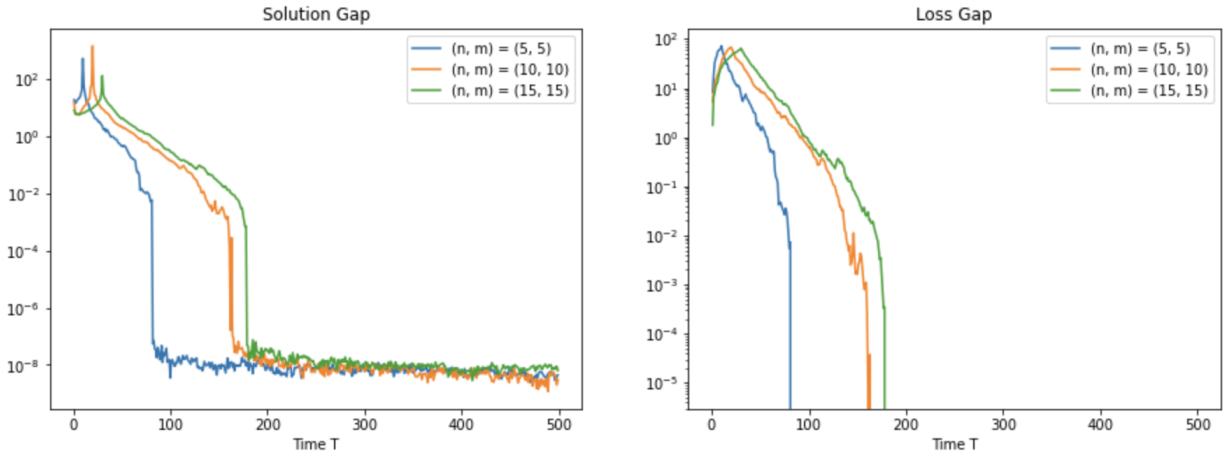


Figure 4.4: Performance of (CO-L2) with Dimensions $(n, m) \in \{(5, 5), (10, 10), (15, 15)\}$ with $p = 0.5$

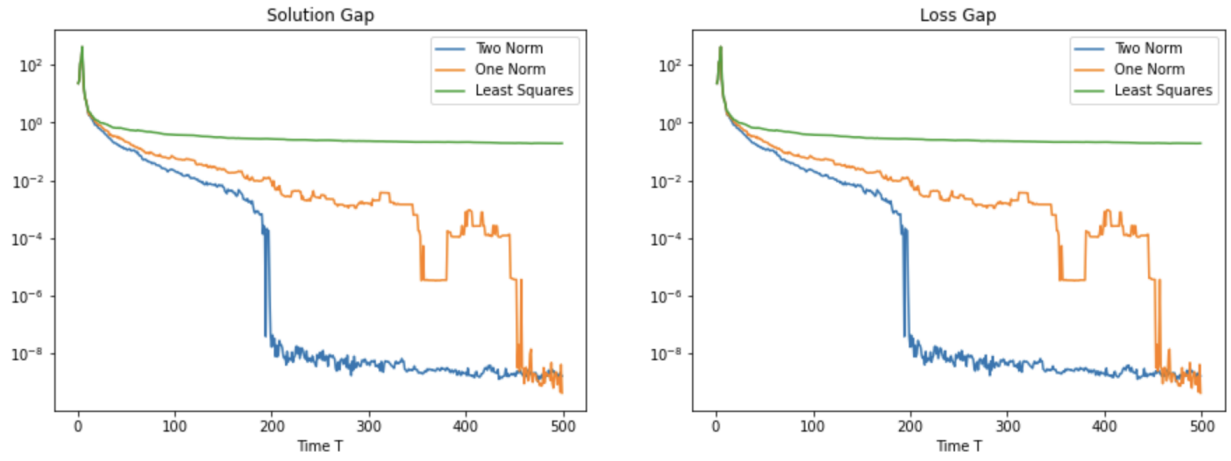


Figure 4.5: Performance of (CO-L2-Aut), (CO-L1-Aut), and Least-Squares with $n = 5$ and $p = 0.7$

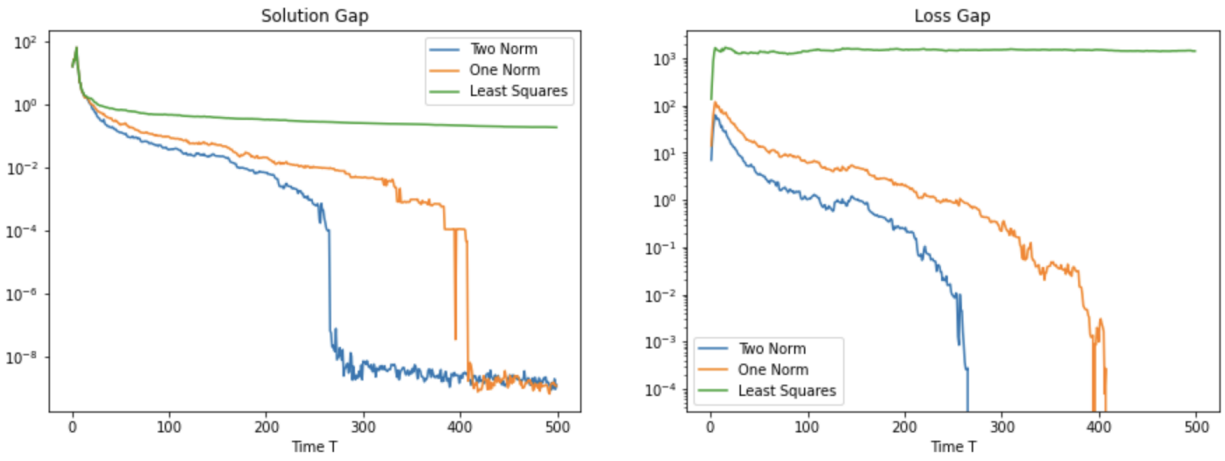


Figure 4.6: Performance of (CO-L2), (CO-L1) and Least-Squares with $(n, m) = 5$ and $p = 0.7$

## 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 [54]. Accurately estimating the parameters of the dynamics is crucial when regulating the blood sugar level by injecting 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 [51]:

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

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 $\bar{\mathbf{A}}$ is stable. Our objective is to estimate the parameters $(k_{ai}, k_{bi}, t_{\max,I}, \mathbf{v}_I, k_e)$. 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 4.7 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. Figure 4.8 shows that the $\ell_2$ and $\ell_1$ estimators perform similarly when the attack vectors are also sparse.

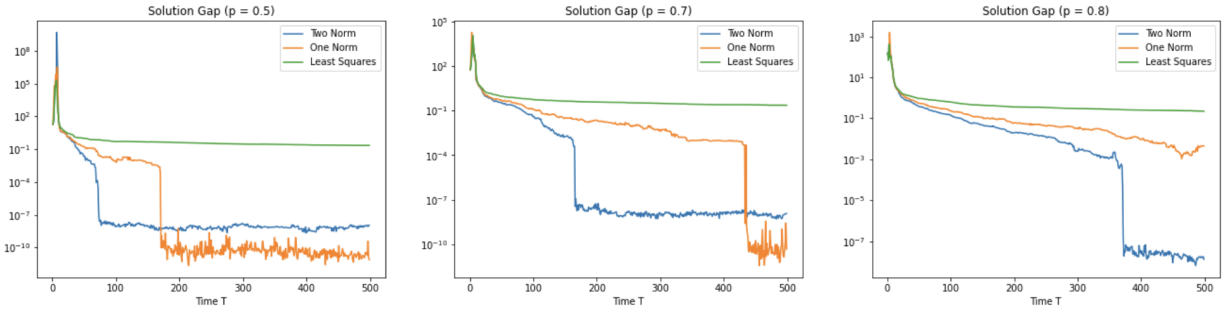


Figure 4.7: Solution Gap for (CO-L2), (CO-L1) and Least-Squares with $p \in \{0.5, 0.7, 0.8\}$ for the Insulin Application

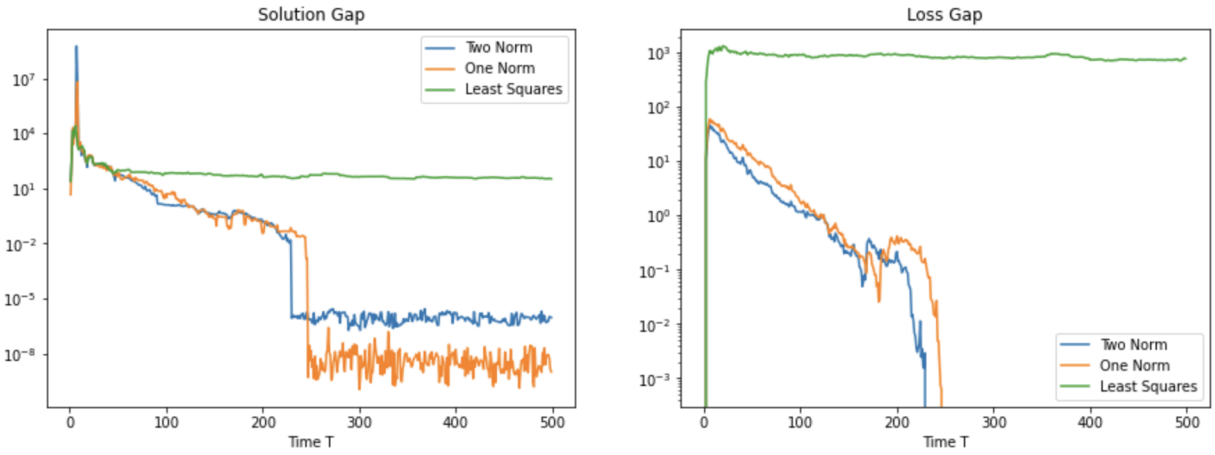


Figure 4.8: Performance of (CO-L2), (CO-L1) and Least-Squares with $p = 0.7$ for the Insulin Application with Sparse Vector Injections

# Appendices

## 4.A Proof of Proposition 5

Let $i_1, i_2, \dots$ be the set of attack times over the time horizon, i.e., $\mathcal{K} = \{i_1, i_2, \dots\}$. We show that sufficient conditions are satisfied for every attack interval $[i_k + 1, i_{k+1}], \forall k \geq 1$. Due to $\Delta$-spaced attack model, we have $i_1 \leq \Delta$. We can utilize Lemma 11 to show that $\bar{\mathbf{A}}$ 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 11 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 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 11 is also satisfied for the time interval $[0, i_1]$. 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 $i_1 \geq 1$.

## 4.B Proof of Proposition 6

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

$$\mathbf{0} \in \sum_{i \notin \mathcal{K}} \mathbf{x}_i \otimes \partial \|\mathbf{0}\|_2 + \sum_{i \in \mathcal{K}} \mathbf{x}_i \otimes \partial \|\bar{\mathbf{d}}_i\|_2. \tag{4.9}$$

Let $i_1$ be the time stamp of the first attack time. Then, we have $i_1 \leq \Delta$ due to the $\Delta$-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 $\mathbf{x}_0 = \mathbf{0}$, we have $\mathbf{x}_i = \mathbf{0}$ whenever $i = 0, 1, \dots, i_1$ and $\mathbf{x}_{i_1+1} = \bar{\mathbf{d}}_{i_1}$. Let $T = \Delta + i_1$, i.e., the time step at which a disturbance cycle is completed. In this case, the sufficient condition using the KKT condition can be written as

$$\begin{aligned}\mathbf{0} &\in \sum_{t=1}^{\Delta-1} \mathbf{x}_{i_1+t} \otimes \partial\|\mathbf{0}\|_2 + \mathbf{x}_{i_1+\Delta} \otimes \partial\|\bar{\mathbf{d}}_{i_1+\Delta}\|_2 \\ &\in \sum_{t=0}^{\Delta-2} \bar{\mathbf{A}}^t \bar{\mathbf{d}}_{i_1} \otimes \partial\|\mathbf{0}\|_2 + \bar{\mathbf{A}}^{\Delta-1}\bar{\mathbf{d}}_{i_1} \otimes \frac{\bar{\mathbf{d}}_{i_1+\Delta}}{\|\bar{\mathbf{d}}_{i_1+\Delta}\|}.\end{aligned}$$

The matrix $0$ may belong to the right-hand side term for arbitrary $\bar{\mathbf{d}}_{i_1+\Delta}$ if

$$\bar{\mathbf{d}}_{i_1+\Delta} \in span\{\bar{\mathbf{d}}_{i_1}, \bar{\mathbf{A}}\bar{\mathbf{d}}_{i_1}, \dots, \bar{\mathbf{A}}^{\Delta-2}\bar{\mathbf{d}}_{i_1}\}.$$

This is satisfied by the assumption in the proposition statement. However, this is not sufficient because the vectors chosen for $\partial\|\mathbf{0}\|_2$ have a bounded norm. Therefore, we need to bound the norm of the columns of $\bar{\mathbf{A}}^{\Delta-1}\bar{\mathbf{d}}_{i_1} \otimes \frac{\bar{\mathbf{d}}_{i_1+\Delta}}{\|\bar{\mathbf{d}}_{i_1+\Delta}\|_2}$, so it can be expressed as a linear combination of the vectors $\{\bar{\mathbf{d}}_{i_1}, \bar{\mathbf{A}}\bar{\mathbf{d}}_{i_1}, \dots, \bar{\mathbf{A}}^{\Delta-2}\bar{\mathbf{d}}_{i_1}\}$. Let $(\lambda_j, \mathbf{v}_j)$ be eigenvalue-eigenvector pairs for the matrix $\bar{\mathbf{A}}^T$. Let $\mathbf{e}_1, \dots, \mathbf{e}_{\Delta-1} \in \partial\|\mathbf{0}\|_2$. Then, the KKT condition can be written as follows:

$$\mathbf{0} \in \mathbf{e}_1\bar{\mathbf{d}}^\top + \mathbf{e}_2\bar{\mathbf{d}}^\top\bar{\mathbf{A}}^\top + \cdots + \mathbf{e}_{\Delta-1}\bar{\mathbf{d}}^\top(\bar{\mathbf{A}}^\top)^{\Delta-2} + \mathbf{f}\bar{\mathbf{d}}^\top(\bar{\mathbf{A}}^\top)^{\Delta-1},$$

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

$$\begin{aligned}0 &\in \mathbf{e}_1\bar{\mathbf{d}}^\top\mathbf{v}_j + \cdots + \mathbf{e}_{\Delta-1}\bar{\mathbf{d}}^\top(\bar{\mathbf{A}}^\top)^{\Delta-2}\mathbf{v}_j + \mathbf{f}\bar{\mathbf{d}}^\top(\bar{\mathbf{A}}^\top)^{\Delta-1}\mathbf{v}_j \\ &\in (\mathbf{e}_1 + \lambda_j\mathbf{e}_2 + \cdots + \lambda_j^{\Delta-2}\mathbf{e}_{\Delta-1} + \lambda_j^{\Delta-1}\mathbf{f})\bar{\mathbf{d}}^\top\mathbf{v}_j.\end{aligned}$$

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

$$\mathbf{0} = \mathbf{e}_1 + \lambda_j\mathbf{e}_2 + \cdots + \lambda_j^{\Delta-2}\mathbf{e}_{\Delta-1} + \lambda_j^{\Delta-1}\mathbf{f}, \quad \forall j = 1, \dots, n.$$

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

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

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

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

Using the fact that $\|\mathbf{e}_j\|_2 = 1$ for all $j$ and $\|\mathbf{f}\|_2 = 1$, we obtain

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

Moreover, if $\mathbf{x}_0$ is not $0$, then we can show that the KKT condition is satisfied for the interval $\{0,1,\ldots,\Delta\}$. This completes the proof for the proposition.

## 4.C Proof of Proposition 7

The KKT condition for the exact recovery is

$$\exists\gamma_i \in \partial\|\mathbf{0}\|_\circ, i \notin \mathcal{K} \text{ s.t. } \sum_{i\notin\mathcal{K}}\mathbf{x}_i\otimes\gamma_i = \sum_{i\in\mathcal{K}}\mathbf{x}_i\otimes\partial\|\bar{\mathbf{d}}_i\|_\circ. \tag{4.10}$$

For (CO-L2-Aut) with $\circ = 2$. the condition (4.10) becomes

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

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

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

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 $\gamma_i^l \in [-1,1], i \notin \mathcal{K}, l = 1,\ldots,n$ such that

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

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

## 4.D Proof of Lemma 12

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

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

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

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

or equivalently,

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

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

## 4.E Proof of Theorem 14

Since $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{x}_i$ can be expressed as

$$\mathbf{x}_i = \sum_{k\in\mathcal{K}} \bar{\mathbf{A}}^{(i-k-1)_+}\bar{\mathbf{d}}_k,$$

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

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

By Lemma 12, given a coordinate $l \in \{1, \ldots, n\}$, the optimality condition for the recovery of $\bar{\mathbf{A}}$ is equivalent to

$$f(\mathbf{z}) := \mathbf{z}^\top\mathbf{g} + \|\mathbf{z}^\top\mathbf{F}\|_1 \geq 0, \quad \forall \mathbf{z} \in \mathbb{S}_2(1), \tag{4.11}$$

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

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

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

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

We do the proof in multiple steps.

## Showing $f(\mathbf{z}) > 0$ for any given $\mathbf{z} \in \mathbb{S}_2(1)$

We first prove that condition (4.11) holds with high probability for a fixed $\mathbf{z} \in \mathbb{S}_2(1)$.

**Analysis of the term $\|\mathbf{z}^\top \mathbf{F}\|_1$**

$$\mathbb{E}\|\mathbf{z}^\top \mathbf{F}\|_1 = \frac{1}{\sqrt{n}} \sum_{i \notin \mathcal{K}} \mathbb{E} \left| \sum_{k \in \mathcal{K}} \mathbf{z}^\top \bar{\mathbf{A}}^{(i-k-1)_+} \bar{\mathbf{d}}_k \right|. \tag{4.12}$$

We construct the index set

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

Let

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

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 $\mathcal{I}_1$ in the following way:

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

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

$$I = |\mathcal{I}| \geq \frac{1}{S} |\mathcal{I}_1|.$$

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

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

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

$$\mathbb{P}\left( I \geq \frac{p(1-p)T}{2S} \right) \geq 1 - \exp[-\Theta[p(1-p)T]]. \tag{4.13}$$

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

$$X_{j,\ell} := \sum_{k \in \mathcal{K}_j} \mathbf{z}^\top \bar{\mathbf{A}}^{i_\ell - k - 1} \bar{\mathbf{d}}_k, \quad \forall j, \ell \in \{1, \ldots, I\} \quad \text{s.t. } j \leq \ell.$$

Using equation (4.12), we can calculate that

$$\|\mathbf{z}^\top \mathbf{F}\|_1 \geq \frac{1}{\sqrt{n}} \sum_{\ell=1}^{I} \left| \sum_{j=1}^{\ell} X_{j,\ell} \right| \geq \frac{1}{\sqrt{n}} \sum_{j=1}^{I} \left( |X_{j,j}| - \sum_{\ell=j+1}^{I} |X_{j,\ell}| \right). \tag{4.14}$$

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

**Lemma 14.** *Suppose that a random variable $X$ is sub-Gaussian with parameter $\sigma_X$, where the mean and the variance of $X$ are $0$ and $\tilde{\sigma}_X^2$, respectively. Then, we have*

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

For all $j \in \{1, \ldots, I\}$, Assumption 4 implies that the standard deviation and the sub-Gaussian parameter of $X_{j,\ell}$ are

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

respectively. It follows from Lemma 14 that

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

which further leads to

$$\mathbb{P}\left(|X_{j,j}| \geq c\sigma_{j,j}\right) \geq \frac{c^4}{64}. \tag{4.15}$$

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

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

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

$$\begin{aligned} -\sum_{\ell=j+1}^{I} |X_{j,\ell}| &\geq -\frac{\rho^S}{1-\rho^S}\sigma_{j,j} \cdot \sqrt{2\log(4I/\delta)} \\ &\geq -\frac{\rho^S}{1-\rho^S}\sigma_{j,j} \cdot \sqrt{2\log(4|\mathcal{I}_1|/\delta)} \\ &\geq -\frac{c^4}{512} \cdot c\sigma_{j,j}, \end{aligned} \tag{4.16}$$

where the last step is by the choice of $S$. Using the bound in (4.13), if we choose

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

it holds with high probability that

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

Note that we have dropped the $|\mathcal{I}_1|$ term in the definition of $S$ since $\log\log(|\mathcal{I}_1|)$ 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

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

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

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

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

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

By the tower property of expectation, we have

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

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

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

Equivalently, we know

$$\mathbb{P}_I\left(\sum_{j=1}^{I}\mathbf{1}_j \geq \frac{c^4}{256}\cdot I\right) \geq 1 - \exp\left[-\Theta\left(\frac{c^4}{128}\cdot I\right)\right]. \tag{4.17}$$

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

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

By the definition of $q_j$ and $\mathbf{1}_j$, when the event in inequality (4.17) happens, inequalities (4.15) and (4.16) imply that

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

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

$$\mathbb{P}_I \left[ \|\mathbf{z}^\top \mathbf{F}\|_1 \geq \frac{c^5\sigma}{512n} \cdot I \right] \geq 1 - \exp\left[ -\Theta\left(c^4 I\right) \right] - \frac{\delta}{4}. \tag{4.18}$$

**Upper bounding of the term $\mathbf{z}^\top \mathbf{g}$**

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

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

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

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

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda \sum_{i\in\mathcal{K}} \sum_{k\in\mathcal{K}} \mathbf{z}^\top \bar{\mathbf{A}}^{(i-k-1)_+} \bar{\mathbf{d}}_k \cdot \bar{\mathbf{f}}_i^l\right)\right] \\ =&\mathbb{E}\left[\exp\left(\lambda \sum_{k\in\mathcal{K}, k<T'} \sum_{i\in\mathcal{K}} \mathbf{z}^\top \bar{\mathbf{A}}^{(i-k-1)_+} \bar{\mathbf{d}}_k \cdot \bar{\mathbf{f}}_i^l\right)\right. \\ &\left.\times \mathbb{E}\left[\exp\left(\lambda \sum_{i\in\mathcal{K}} \mathbf{z}^\top \bar{\mathbf{A}}^{(i-1-T')_+} \bar{\mathbf{d}}'_T \cdot \bar{\mathbf{f}}_i^l\right) \middle| \mathcal{F}_{T'}, \mathcal{F}^f\right]\right].\end{aligned} \tag{4.19}$$

Using the decomposition in Assumption 4, we have

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

Substituting back into (4.19) and continuing the process for all $k\in\mathcal{K}$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda\sum_{i\in\mathcal{K}}\sum_{k\in\mathcal{K}}\mathbf{z}^\top\bar{\mathbf{A}}^{(i-k-1)_+}\bar{\mathbf{d}}_k\cdot\bar{\mathbf{f}}_i^l\right)\right] \qquad (4.20)\\
\leq&\mathbb{E}\left[\exp\left[\frac{\lambda^2\sigma^2}{2}\sum_{k\in\mathcal{K}}\left(\sum_{i\in\mathcal{K}}\mathbf{z}^\top\bar{\mathbf{A}}^{(i-1-k)_+}\bar{\mathbf{f}}_k\bar{\mathbf{f}}_i^l\right)^2\right]\right]\\
\leq&\mathbb{E}\left[\exp\left[\frac{\lambda^2\sigma^2}{2}\sum_{k\in\mathcal{K}}\left(\sum_{i\in\mathcal{K}}\left|\mathbf{z}^\top\bar{\mathbf{A}}^{(i-1-k)_+}\bar{\mathbf{f}}_k\right|\right)^2\right]\right]\\
\leq&\exp\left[\frac{\lambda^2\sigma^2}{2}\sum_{k\in\mathcal{K}}\left(\sum_{i\in\mathcal{K}}\rho^{(i-1-k)_+}\right)^2\right],
\end{aligned}$$

where the last inequality holds because $\bar{\mathbf{f}}_i^l$ is bounded in $[-1,1]$. For each $i,k\in\mathcal{K}$, the value of $\left(\mathbf{z}^\top\bar{\mathbf{A}}^{(i-1-k)_+}\bar{\mathbf{f}}_k\right)^2$ concentrates around its expectation $\|\mathbf{z}^\top\bar{\mathbf{A}}^{(i-1-k)_+}\|_2^2/n$. Therefore, inequality (4.20) leads to

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\sum_{i\in\mathcal{K}}\sum_{k\in\mathcal{K}}\mathbf{z}^\top\bar{\mathbf{A}}^{(i-k-1)_+}\bar{\mathbf{d}}_k\cdot\bar{\mathbf{f}}_i^l\right)\right]&\leq\exp\left[\Theta\left[\frac{\lambda^2\sigma^2}{2n}\sum_{k\in\mathcal{K}}\left(\sum_{i\in\mathcal{K}}\left\|\mathbf{z}^\top\bar{\mathbf{A}}^{(i-1-k)_+}\right\|_2\right)^2\right]\right] \qquad (4.21)\\
&\leq\exp\left[\Theta\left[\frac{\lambda^2\sigma^2}{2n}\sum_{k\in\mathcal{K}}\left(\sum_{i\in\mathcal{K}}\rho^{(i-k-1)_+}\right)^2\right]\right].
\end{aligned}$$

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

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

Define

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

We can calculate that

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

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

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

Hence, inequality (4.21) implies that with the same probability, $\mathbf{z}^\top \mathbf{g}$ is sub-Gaussian with parameter

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

Therefore, Hoeffding's inequality leads to

$$\mathbb{P}_{|\mathcal{K}|}\left[\mathbf{z}^\top \mathbf{g} \leq -\Theta\left(\sqrt{\frac{|\mathcal{K}|p\sigma^2}{n(1-\rho)^3} \log\left(\frac{4}{\delta}\right)}\right)\right] \leq \frac{\delta}{4}. \tag{4.22}$$

By combining inequalities (4.18) and (4.22), it holds with probability at least

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

that

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

Similar to the bound in (4.13), it holds with probability at least $1 - \exp[-\Theta(pT)]$ that

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

As a result, if we choose

$$T \geq \Theta\Bigg[\max\Bigg\{\frac{\log\log(1/\delta)+\log(1/c)}{c^4 p(1-p)\log(1/\rho)}\log\left(\frac{1}{\delta}\right), \frac{1}{p(1-p)}\log\left(\frac{1}{\delta}\right), \frac{n\log(1/c)^2}{c^{10}(1-p)^2(1-\rho)^3\log^2(1/\rho)}\log\left(\frac{1}{\delta}\right)\Bigg\}\Bigg] = \Theta\left[nR\log\left(\frac{1}{\delta}\right)\right], \tag{4.23}$$

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

$$\mathbb{P}_I\left[f(\mathbf{z}) \geq \Theta\left(\frac{c^5\sigma I}{n}\right)\right] \geq 1-\delta. \tag{4.24}$$

**Using Discretization Bound over the Unit Sphere**

In the second step, we apply discretization techniques to prove that condition (4.11) holds for all $\mathbf{z} \in \mathbb{S}_2(1)$ with high probability. Suppose that $\epsilon > 0$ is a small constant. We construct an $\epsilon$-cover of the unit sphere $\mathbb{S}_2(1)$, denoted as

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

Namely, for all $\mathbf{z} \in \mathbb{S}_2(1)$, we can find $r \in \{1, 2, \ldots, N\}$ such that $\|\mathbf{z} - \mathbf{z}^r\|_2 \leq \epsilon$. 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 $f(\mathbf{z})$ in inequality (4.24). Then, we have

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

Our goal is to prove that

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

holds with high probability. Notice that

$$\begin{aligned}
f(\mathbf{z}) - f(\mathbf{z}') &= (\mathbf{z}-\mathbf{z}')^\top \mathbf{g} + (\|\mathbf{z}^\top \mathbf{F}\|_1 - \|(\mathbf{z}')^\top \mathbf{F}\|_1) \\
&\geq (\mathbf{z}-\mathbf{z}')^\top \mathbf{g} - \|(\mathbf{z}-\mathbf{z}')^\top \mathbf{F}\|_1 \\
&\geq -\|\mathbf{z}-\mathbf{z}'\|_2 \|\mathbf{g}\|_2 - \|\mathbf{z}-\mathbf{z}'\|_2 \sum_{i \notin \mathcal{K}} \|\mathbf{F}^i\|_2 \\
&\geq -\epsilon \left( \left\| \sum_{i \in \mathcal{K}} \sum_{k \in \mathcal{K}} \bar{\mathbf{A}}^{(i-k-1)_+} \bar{\mathbf{d}}_k \right\|_2 + \frac{1}{\sqrt{n}} \sum_{i \notin \mathcal{K}} \left\| \sum_{k \in \mathcal{K}} \bar{\mathbf{A}}^{(i-k-1)_+} \bar{\mathbf{d}}_k \right\|_2 \right) \\
&\geq -\epsilon \sum_{k \in \mathcal{K}} \sum_{i > k} \rho^{(i-k-1)} |\bar{\ell}_k|.
\end{aligned}$$

Using the property of exponential sequences, we have

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

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

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

Letting

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

it holds that

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

Now, after we replace $\delta$ in (4.23) with $\delta/(2N)$, it holds with probability at least $1 - \delta/2$ that

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

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

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

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

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

we get the estimate

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

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

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

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

**Showing Uniqueness of Solution**

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

$$\begin{aligned} f_T(\bar{\mathbf{A}}) &< f_T(\bar{\mathbf{A}} + \mathbf{\Omega}) \quad \Rightarrow \\ \sum_{i=0}^{T-1} \|\bar{\mathbf{d}}_i\|_2 &< \sum_{i=0}^{T-1} \|\mathbf{\Omega}\mathbf{x}_i + \bar{\mathbf{d}}_i\|_2 \quad \Rightarrow \\ \sum_{i\in\mathcal{K}} \|\bar{\mathbf{d}}_i\|_2 &< \sum_{i\notin\mathcal{K}} \|\mathbf{\Omega}\mathbf{x}_i\|_2 + \sum_{i\in\mathcal{K}} \|\bar{\mathbf{d}}_i\|_2 + \sum_{i\in\mathcal{K}} \left\langle \mathbf{\Omega}\mathbf{x}_i, \frac{\bar{\mathbf{d}}_i}{\|\bar{\mathbf{d}}_i\|_2} \right\rangle \\ \Rightarrow \quad 0 &< \sum_{i\notin\mathcal{K}} \|\mathbf{\Omega}\mathbf{x}_i\|_2 + \sum_{i\in\mathcal{K}} \left\langle \mathbf{\Omega}\mathbf{x}_i, \frac{\bar{\mathbf{d}}_i}{\|\bar{\mathbf{d}}_i\|_2} \right\rangle + \mathcal{O}(\|\mathbf{\Omega}\|_F^2) \end{aligned}$$

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

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

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

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

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

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

The matrix $\mathbf{F} \in \mathbb{R}^{n \times (T - |\mathcal{K}|)}$ and the vector $\mathbf{g} \in \mathbb{R}^n$ are defined at the beginning of the proof. The above condition is the same as (4.11), except that we require strict positivity of $f(\mathbf{z})$ 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.

# Chapter 5

# System Identification for Non-Linear Systems

Dynamical systems serve as the foundation for several fields, including sequential decision-making, reinforcement learning, control theory, and recurrent neural networks. They are essential for analyzing and controlling the behavior of real-world physical systems. However, accurately modeling dynamical systems is challenging due to their complexity and the large-scale nature of modern systems. The problem of estimating or learning a system's dynamics from past observations is known as the *system identification* problem. This problem is extensively studied in the control theory literature, typically under the restrictive assumption that disturbances are small in magnitude and follow an independent and identically distributed (i.i.d.) probability distribution, accounting for modeling errors, measurement noise, and sensor inaccuracies. In contrast to the conventional small-value, dense-over-time i.i.d. disturbance model, this chapter considers a sparse-over-time, large-value, and temporally correlated disturbance model. Specifically, we analyze settings where the disturbance is often zero, but when non-zero, it can take large values and exhibit correlations with past disturbances.

The primary motivation for this chapter arises from emerging safety-critical applications, such as smart grids, autonomous vehicles, and unmanned aerial vehicles, which require robust estimation of system dynamics in the presence of sparse but large and potentially adversarial disturbances. While machine learning techniques have demonstrated significant success in various domains, such as computer vision and natural language processing, their application to safety-critical systems remains limited due to a lack of theoretical guarantees. This chapter addresses this gap by providing strong theoretical results for learning dynamical systems using machine learning techniques.

As a motivating example, we consider the dynamical system associated with a power grid, such as the U.S. electrical grid or a regional interconnection. The system states capture various physical parameters, including voltage magnitudes and frequencies across different parts of the network. To enhance the sustainability, resiliency, and efficiency of energy systems, modern power grids integrate large volumes of renewable energy sources, such as wind turbines, solar panels, and electric vehicles.

The operation of power systems has become increasingly complex due to the active participa-

tion of consumers, who strategically respond to electricity prices by adjusting their consumption based on price signals. Simultaneously, the widespread deployment of sensors across the grid has enabled data-driven grid operation by continuously collecting and analyzing system data. However, this advancement introduces a significant vulnerability: even a small, strategically executed data manipulation could mislead power suppliers, causing them to overestimate or underestimate electricity demand. Such miscalculations could result in severe consequences, including system-wide blackouts. This scenario can be modeled as a non-linear dynamical system in which the system input is subject to *semi-oblivious attacks* at various locations, resulting in the injection of incorrect electricity values into the grid. Given the integration of a large number of new devices into the system, coupled with strategic human behavior, power operators lack a complete model of the underlying dynamical system. Consequently, they must simultaneously learn the system dynamics and detect potential adversarial attacks to effectively mitigate disruptions and restore normal operation. If an input attack remains unaddressed, it can destabilize the system's transient behavior, leading to signal instability and potentially triggering a cascading failure across the grid.

Prior research on robust system learning has primarily focused on unreliable measurements, where the objective is to extract knowledge from noisy and corrupted observations—such as in the matrix sensing problem in machine learning. However, this chapter addresses an emerging and largely overlooked aspect of robust learning in safety-critical systems, where the system input itself is manipulated (potentially by an adversary), thereby affecting the system states and inducing instability. Existing results in system identification have largely concentrated on the asymptotic properties of the least squares estimator (LSE) [26, 71, 73, 7]. With the advent of statistical learning theory, research in this area has evolved to study the required number of samples necessary to achieve a given error threshold [107]. While early non-asymptotic analyses focused on linear time-invariant (LTI) systems with i.i.d. disturbances using mixing arguments [66, 93], more recent studies employ martingale and small-ball techniques to derive sample complexity guarantees for LTI systems [102, 39, 106]. For non-linear systems, parameterized models have been explored in recent studies [84, 85, 45, 95, 137], demonstrating the convergence of recursive and gradient-based algorithms to the true parameters with a convergence rate of $T^{-1/2}$ using martingale techniques and mixing time arguments. Additionally, there has been some progress toward developing non-smooth estimators for both linear and non-linear systems [43, 44, 119], particularly in handling large-but-sparse noise vectors with dependencies. However, robust regression techniques incorporating regularization [116, 9, 55] remain relatively unexplored in the context of dynamical systems, particularly with respect to non-asymptotic sample complexity analysis. This gap is largely due to the inherent autocorrelation in system samples, making traditional statistical tools less applicable. A more detailed literature review is provided in Section 5.1.

This chapter lays the foundation for advancing online optimal control in the presence of large but sparse disturbances, such as certain adversarial attacks. A crucial first step in achieving this goal is accurately learning the system dynamics. To this end, we focus on the system identification problem for parameterized non-linear systems, making necessary assumptions about the disturbance model, which will be discussed in later sections. We model the unknown non-linear functions describing the system via a linear combination of some given basis functions by taking advantage of their representation properties. Our objective is to estimate the parameters of these

basis functions, which dictate the updates of the dynamical system. Formally, we consider the following autonomous dynamical system:

$$\mathbf{x}_0 = \mathbf{0}_n, \quad \mathbf{x}_{t+1} = \bar{\mathbf{A}} f(\mathbf{x}_t) + \bar{\mathbf{d}}_t, \quad \forall t \in \{0, \dots, T-1\}, \tag{5.1}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is a combination of $m$ known basis functions and $\bar{\mathbf{A}} \in \mathbb{R}^{n \times m}$ is the unknown matrix of system parameters. The system trajectory is also influenced by the disturbance term $\bar{\mathbf{d}}_t \in \mathbb{R}^n$. We consider a *sparse-but-large* disturbance scenario, where the disturbance is sometimes zero but, when non-zero, takes large values. In the context of an adversarial attack, a zero disturbance value for $\bar{\mathbf{d}}_t$ indicates the absence of attack at time $t$, while a non-zero disturbance is typically large, designed to maximize the impact of the attack. The goal of the system identification problem is to recover the ground truth matrix $\bar{\mathbf{A}}$ using observations of the system states, i.e., $\{\mathbf{x}_0, \dots, \mathbf{x}_T\}$. The model generating $\bar{\mathbf{d}}_t$'s is unknown to the user; however, to ensure unique recoverability of the system from measurements, we will later introduce necessary assumptions on the disturbance model, referred to as *semi-oblivious attacks*.

Unlike empirical risk minimization problems, where samples are typically assumed to be i.i.d., the system states $\{\mathbf{x}_0, \dots, \mathbf{x}_T\}$ exhibit auto-correlation. As a result, the standard i.i.d. assumption on the data-generating distribution is violated. This auto-correlation introduces significant theoretical challenges, which we address in this chapter by proposing a novel and non-trivial extension of exact recovery guarantees to the system identification problem. Since the disturbance $\bar{\mathbf{d}}_t$ is unknown to the system operator and can take a large value, it is necessary to utilize estimators to the ground truth $\bar{\mathbf{A}}$ that is robust to variations in $\bar{\mathbf{d}}_t$ and converge to $\bar{\mathbf{A}}$ within a finite time horizon $T$. This chapter is inspired by [119] that studied the above problem for linear systems. The linear case is noticeably simpler than the non-linear system identification problem since each observation $\mathbf{x}_t$ becomes a linear function of previous disturbances. In contrast, for non-linear systems, the relationship between measurements and disturbances is considerably more complex, requiring substantial technical advancements beyond the methods used in [119].

The existing literature has primarily focused on the smooth least-squares estimator:

$$\hat{\mathbf{A}} \in \arg\min_{\mathbf{A} \in \mathbb{R}^{n \times m}} \sum_{t=0}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{A} f(\mathbf{x}_t)\|_2^2. \tag{5.2}$$

In contrast, motivated by the exact recovery properties of non-smooth loss functions (e.g., the $\ell_1$-norm and the nuclear norm), we consider the following alternative estimator:

$$\hat{\mathbf{A}} \in \arg\min_{\mathbf{A} \in \mathbb{R}^{n \times m}} \sum_{t=0}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{A} f(\mathbf{x}_t)\|_2. \tag{5.3}$$

We note that the optimization problem (5.3) remains convex in $\mathbf{A}$ despite its non-smooth objective function; therefore, it can be solved efficiently by existing optimization solvers. The estimator (5.3) is closely related to the LASSO estimator, as its loss function can be interpreted as a generalization

of the $\ell_1$-loss function. More specifically, when $n = 1$, the estimator (5.3) simplifies to

$$\hat{\mathbf{A}} \in \arg\min_{A \in \mathbb{R}^{1\times m}} \sum_{t=0}^{T-1} |\mathbf{x}_{t+1} - \mathbf{A} f(\mathbf{x}_t)|,$$

which corresponds to the auto-correlated general linear regression estimator with an $\ell_1$-loss function. In this chapter, the goal is to prove the efficacy of the above estimator by obtaining mild conditions under which the ground truth $\bar{\mathbf{A}}$ can be *exactly recovered* by the estimator (5.3). More specifically, we seek to address the following key questions:

i) What are the *necessary and sufficient* conditions under which $\bar{\mathbf{A}}$ is an optimal solution to the optimization problem (5.3) or the unique optimal solution?

ii) What is the minimum number of samples required to ensure that the necessary and sufficient conditions for exact recovery hold with high probability under certain assumptions?

In this chapter, we provide answers to the above questions. Section 5.1 presents a comprehensive literature review on system identification and robust estimation problems. In Section 5.2, we analyze the necessary and sufficient conditions for the global optimality of $\bar{\mathbf{A}}$ for the problem (5.3). Furthermore, we establish conditions under which $\bar{\mathbf{A}}$ is the unique solution, thereby addressing Question (i). Next, in Sections 5.4 and 5.5, we derive lower bounds on the number of samples $T$ required to guarantee that $\bar{\mathbf{A}}$ is the unique solution with high probability. We consider two distinct cases: when the basis function $f$ is bounded when it is Lipschitz continuous. These results provide an answer to Question (ii). In Section 5.6, we introduce a faster learning strategy when non-zero disturbances fail to sufficiently explore the system state space. Finally, in Section 5.7, we present numerical experiments that support our theoretical findings. This chapter constitutes the first non-asymptotic sample complexity analysis for exact recovery in the non-linear system identification problem.

## 5.1 Literature Overview

Until recently, research on the system identification problem primarily focused on the asymptotic properties of the least squares estimator (LSE) [26, 71, 73, 7]. However, with the increasing prominence of statistical learning theory [109, 111], understanding the number of samples required to achieve a given error threshold in system identification has become a topic of significant interest. For a comprehensive overview of existing results and proof techniques, we refer the reader to the survey by [107]. The non-asymptotic analysis of system identification has primarily focused on linear time-invariant (LTI) systems under the assumption of i.i.d. noise. Early research in this area relied on mixing arguments, which heavily depended on system stability [66, 93]. More recent studies have employed martingale and small-ball techniques to establish sample complexity guarantees for least squares estimators applied to LTI systems [102, 39, 106]. These works demonstrate that the LSE converges to the true system parameters at a rate of $T^{-1/2}$, where $T$ denotes

the number of samples. Furthermore, these results have been applied to the linear-quadratic regulator problem, where adaptive control techniques leverage system identification results to achieve optimal regret bounds [33, 1, 34].

The non-linear system identification problem has been extensively studied [84, 85]. However, research on the non-asymptotic analysis of non-linear system identification remains in its early stages and has primarily focused on parameterized non-linear systems. Recursive and gradient-based algorithms designed for the least squares loss function have been shown to asymptotically converge to the true system parameters at a rate of $T^{-1/2}$ for non-linear systems with a known link function $\phi$ of the form $\phi(\bar{\mathbf{A}}\mathbf{x}_t)$, using martingale techniques [45] and mixing time arguments [95]. More recently, [137] established sample complexity guarantees for non-parametric learning of non-linear system dynamics, demonstrating a convergence rate that scales as $T^{-1/(2+q)}$, where $q$ depends on the complexity of the function class in which the true dynamics reside. Notably, existing studies on both linear and non-linear system identification have largely assumed i.i.d. (sub)-Gaussian noise structures, limiting their applicability to more general disturbance models.

Despite the growing interest in non-asymptotic system identification, research on system identification using non-smooth estimators capable of handling dependent and adversarial noise vectors remains limited to linear systems. [43] and [44] investigated a non-smooth convex estimator in the form of the least absolute deviation estimator, analyzing the conditions required for exact recovery of system dynamics using Karush-Kuhn-Tucker (KKT) conditions and the Null Space Property from the LASSO literature. Subsequently, [119] demonstrated that exact recovery of system parameters is achievable with high probability, even when more than half of the data is corrupted, opening new avenues for adversarially robust system identification. Compared to [119], the presence of non-linear basis functions in our setting makes it impossible to analyze the optimization problem by explicitly expressing $\mathbf{x}_t$; see the proof of Theorem 2 in [119]. Note that when the system is in the form of $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t$, then $\mathbf{x}_t$ can be written directly as $\mathbf{A}^t\mathbf{x}_0$ and we only need to analyze the eigenvalues of $\mathbf{A}$. For a non-linear system in the form of $\mathbf{x}_{t+1} = f(\mathbf{x}_t)$, writing $\mathbf{x}_t$ in terms of $\mathbf{x}_0$ needs the composition of $t$ functions, and this cannot be done analytically. Unlike linear systems, there is no direct counterpart to eigenvalue analysis for non-linear systems. This fundamental challenge is widely acknowledged in non-linear systems literature within control theory, and as a result, many results known for linear systems do not extend to the non-linear setting. Consequently, our proof for the bounded case is novel and fundamentally different from the approach in [119]. Finally, by leveraging the generalized Farkas' lemma, we establish necessary and sufficient conditions in Section 5.2 that are both novel and stronger than the sufficient conditions provided in [119].

On the other hand, robust regression techniques have been developed by incorporating regularizers into the objective function [116, 9, 55]. Additionally, the robust estimation literature has introduced several non-smooth estimators, including M-estimators, least absolute deviation estimators, convex estimators, least median squares, and least trimmed squares [98]. The convex estimator in (5.3) was proposed in [6, 5] in the context of robust regression, demonstrating that exact recovery is achievable given an infinite number of samples. However, these studies lack a non-asymptotic analysis of sample complexity. Furthermore, their analytical techniques are not directly applicable to dynamical systems due to the presence of autocorrelation among samples.

Recent works [115, 65] have focused on reinforcement learning (RL), where the primary objective is to maximize a reward function. In contrast, system identification aims to recover the underlying system dynamics, and in many applications, a naturally defined reward function may not exist. Moreover, both of these RL studies assume that perturbations are bounded, a restrictive assumption that may not hold in real-world scenarios. More importantly, attempting to control a system without first learning its dynamics (e.g., using model-free RL techniques) poses significant risks. During exploration, an inadequate policy could shift the system state beyond safe operational limits, potentially leading to instability; see the survey by [80]. For safety-critical systems, it is typically necessary to first learn the system dynamics before applying a control strategy, whether it be a classical optimal control method or an RL-based approach. This chapter focuses on learning the system model in the presence of adversaries on its dynamics. Existing RL approaches, including those in [115, 65], address a fundamentally different problem. Additionally, while the field of robust model-based RL is well-developed, our setting involves unknown system dynamics, necessitating the use of model-free RL techniques.

## 5.2 Global Optimality of Ground Truth and Uniqueness of Global Solutions

In this section, we derive conditions under which the ground truth $\bar{\mathbf{A}}$ is a global minimizer to the optimization problem (5.3). Given the system dynamics, this optimization problem can be reformulated as

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times m}} \sum_{t=0}^{T-1} \|(\bar{\mathbf{A}} - \mathbf{A})f(\mathbf{x}_t) + \bar{\mathbf{d}}_t\|_2, \tag{5.4}$$

where $\mathbf{x}_0, \ldots, \mathbf{x}_T$ are generated according to the unknown system under disturbances. To facilitate the analysis, we define the set of time instances where disturbances are non-zero as $\mathcal{K} := \{t \mid \bar{\mathbf{d}}_t \neq 0\}$ and introduce the normalized disturbances as

$$\hat{\mathbf{d}}_t := \bar{\mathbf{d}}_t / \|\bar{\mathbf{d}}_t\|_2, \quad \forall t \in \mathcal{K}.$$

The following theorem provides a necessary and sufficient condition for $\bar{\mathbf{A}}$ to be a global minimizer of the problem in (5.4).

**Theorem 18** (Necessary and sufficient condition for optimality)**.** *The ground truth matrix $\bar{\mathbf{A}}$ is a global solution to problem* (5.4) *if and only if*

$$\sum_{t\in\mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z} f(\mathbf{x}_t) \leq \sum_{t\in\mathcal{K}^c} \|\mathbf{Z} f(\mathbf{x}_t)\|_2, \quad \forall \mathbf{Z} \in \mathbb{R}^{n\times m}, \tag{5.5}$$

*where $\mathcal{K}^c := \{0, \ldots, T-1\} \backslash \mathcal{K}$ denotes the set of time indices where disturbances are zero.*

Theorem 18 provides a necessary and sufficient condition for the well-specifiedness of the optimization problem in (5.4). Intuitively, the left-hand side of (5.5) represents the impact of non-zero disturbances, while the right-hand side corresponds to the normal system dynamics. If the disturbances do not dominate the correct system dynamics, then the estimator can successfully recover the ground truth dynamics. The condition in (5.5) is derived using the generalized Farkas' lemma, which eliminates the need for inner approximation of the $\ell_2$-ball by an $\ell_\infty$-ball, as in [119]. Consequently, the sample complexity bounds obtained in this chapter are stronger than those in [119] when applied to the special case of linear systems. Further details on these bounds are provided in Sections 5.4 and 5.5.

Using the condition established in Theorem 18, we derive both sufficient and necessary conditions for the optimality of $\bar{\mathbf{A}}$ in Corollaries 5 and 6.

**Corollary 5** (Sufficient condition for optimality)**.** *If it holds that*

$$\sum_{t\in\mathcal{K}} \|\mathbf{Z}f(\mathbf{x}_t)\|_2 \leq \sum_{t\in\mathcal{K}^c} \|\mathbf{Z}f(\mathbf{x}_t)\|_2, \quad \forall \mathbf{Z}\in\mathbb{R}^{n\times m}, \tag{5.6}$$

*then the ground truth matrix $\bar{\mathbf{A}}$ is a global solution to problem* (5.4)*.*

**Corollary 6** (Necessary condition for optimality)**.** *If the ground truth matrix $\bar{\mathbf{A}}$ is a global solution to problem* (5.4)*, then it holds that*

$$\left\| \sum_{t\in\mathcal{K}} f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top \right\|_F \leq \sum_{t\in\mathcal{K}^c} \|f(\mathbf{x}_t)\|_2. \tag{5.7}$$

*In the case when $m = 1$, condition* (5.7) *is necessary and sufficient.*

The proofs of Corollaries 5 and 6 are provided in the Appendix. The conditions established above are more general than many existing results in the literature; see Examples 1 and 2 in Appendix 5.A for further discussion.

Next, we derive conditions under which the ground truth matrix $\bar{\mathbf{A}}$ is the unique solution to the optimization problem in (5.4). We present the following necessary and sufficient conditions for the uniqueness of global solutions, which extends the result of Theorem 18.

**Theorem 19** (Necessary and sufficient condition for uniqueness)**.** *Suppose that condition* (5.5) *holds. The ground truth $\bar{\mathbf{A}}$ is the unique global solution to problem* (5.4) *if and only if the following logical condition holds for every non-zero $\mathbf{Z}\in\mathbb{R}^{n\times m}$:*

$$\sum_{t\in\mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z}f(\mathbf{x}_t) = \sum_{t\in\mathcal{K}^c} \|\mathbf{Z}f(\mathbf{x}_t)\|_2 \implies \sum_{t\in\mathcal{K}} \left|\hat{\mathbf{d}}_t^\top \mathbf{Z}f(\mathbf{x}_t)\right| < \sum_{t\in\mathcal{K}} \|\mathbf{Z}f(\mathbf{x}_t)\|_2, \tag{5.8}$$

*which means that whenever the left-hand side equality is satisfied for some non-zero $\mathbf{Z}$, the right-hand side inequality must also hold.*

Based on Theorem 19, the following corollary provides a sufficient condition for the uniqueness of $\bar{\mathbf{A}}$, which is more practical to verify compared to condition (5.8). Notably, this corollary also generalizes the sufficiency part of Corollary 6 to the multi-dimensional setting.

**Corollary 7** (Sufficient condition for uniqueness)**.** *Suppose that condition* (5.5) *holds. If*

$$\sum_{t\in\mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z} f(\mathbf{x}_t) < \sum_{t\in\mathcal{K}^c} \|\mathbf{Z} f(\mathbf{x}_t)\|_2, \quad \forall \mathbf{Z} \in \mathbb{R}^{n\times m} \quad \text{s.t. } \mathbf{Z} \neq 0, \tag{5.9}$$

*then the ground truth matrix* $\bar{\mathbf{A}}$ *is the unique global solution to problem* (5.4)*.*

*Proof.* The logical condition in (5.8) states that whenever the left-hand side equality is satisfied for some non-zero $\mathbf{Z}$, the right-hand side inequality must also hold. Under the assumption in (5.9), there is no non-zero $\mathbf{Z}$ satisfying the left-hand-side equality. This implies that the logical condition in (5.8) automatically holds and, thus, Theorem 19 implies the uniqueness of $\bar{\mathbf{A}}$. □

Theorem 19 strengthens and generalizes existing results for first-order systems, specifically Theorem 1 in [43]. For further discussion, see Example 3 in Appendix 5.A.

## 5.3 Disturbance Model and Semi-Oblivious Attacks

We consider the frequency at which the system experiences a non-zero disturbance to model sparse disturbances. This is captured by the sparsity probability $p$, defined as follows:

**Definition 10** (Probabilistic sparsity model)**.** *For each time instance* $t$*, the disturbance vector* $\bar{\mathbf{d}}_t$ *is non-zero with probability* $p \in (0, 1)$*. Furthermore, the occurrences of disturbances are independent across time instances.*

Existing results have predominantly focused on the case where $p = 1$, under which the system dynamics cannot be learned in finite time. To illustrate the impact of sparsity, consider the scenario where $\bar{\mathbf{d}}_t$ follows a Gaussian distribution for all $t \in \mathcal{K}$ and independent of previous disturbances $\bar{\mathbf{d}}_1, ..., \bar{\mathbf{d}}_{t-1}$. When $p = 1$, it follows that there exists no finite time $T$ for which the correct dynamics matrix $\bar{\mathbf{A}}$ is a solution to the least-square estimator (5.2) almost surely.

Moreover, when $p < 1$, the LSE fails to achieve exact recovery even under an i.i.d zero-mean Gaussian disturbance structure. To illustrate the failure of LSE, we define the following matrices:

$$\mathbf{F}_T = \begin{bmatrix} f(\mathbf{x}_0) & f(\mathbf{x}_1) & \cdots & f(\mathbf{x}_{T-1)} \end{bmatrix} \text{ and } \mathbf{X}_T = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_T \end{bmatrix},$$

where $\mathbf{F}_T, \mathbf{X}_T \in \mathbb{R}^{n\times T}$. In addition, we define the disturbance matrix $\bar{\mathbf{D}}_T = \begin{bmatrix} \bar{\mathbf{d}}_0 & \bar{\mathbf{d}}_1 & \cdots & \bar{\mathbf{d}}_{T-1} \end{bmatrix}$, where $\bar{\mathbf{D}}_T \in \mathbb{R}^{n\times T}$. Under these definitions, the system updates can be expressed as $\mathbf{X}_T = \bar{\mathbf{A}}\mathbf{F}_T + \bar{\mathbf{D}}_T$. Thus, the closed-form solution of the LSE is given by $\hat{\mathbf{A}} = (\mathbf{F}_T\mathbf{F}_T^\top)^{-1}\mathbf{F}_T\mathbf{X}_T^\top$, and the estimation error can be written as $\hat{\mathbf{A}} - \bar{\mathbf{A}} = (\mathbf{F}_T\mathbf{F}_T^\top)^{-1}\mathbf{F}_T\bar{\mathbf{D}}_T^\top$. When $T$ is sufficiently large, the matrices $\mathbf{F}_T$ and $\bar{\mathbf{D}}_T$ become full-rank; consequently, the LSE estimation error, $\hat{\mathbf{A}} - \bar{\mathbf{A}}$ never attains zero error. In contrast, the results in Sections 5.4 and 5.5 conclude that for every $p \in$

$(0, 1)$, the ground truth matrix $\bar{\mathbf{A}}$ becomes the unique solution of the non-smooth estimator (5.3), provided that $T$ exceeds a certain threshold. This result represents a highly specialized case of the broader findings presented in this chapter and highlights the advantage of incorporating zero disturbances. Specifically, having some instances of zero disturbances enables a transition from asymptotic learning to finite-time learning, demonstrating the efficacy of the proposed approach.

It is important to note that in this chapter, the disturbance vectors $\bar{\mathbf{d}}_t$'s are allowed to be correlated over time. Definition 10 is only about the times at which disturbances are non-zero. Additionally, we do not assume that the sparsity probability $p$ or the model generating the disturbances is known. Recalling the definition $\mathcal{K} := \{t \mid \bar{\mathbf{d}}_t \neq 0\}$, we observe that with probability at least $1 - \exp[-\Theta(pT)]$, the cardinality of $\mathcal{K}$ satisfies $|\mathcal{K}| = \Theta(pT)$. When $p$ is close to 0, the system is rarely affected by disturbances. The absence of disturbances, however, may lessen the rate of exact recovery due to insufficient exploration of the system space. A potential approach to mitigate this issue is outlined in Section 5.6. Nevertheless, this chapter primarily focuses on the case where $p$ is close to 1, meaning that the system experiences disturbances at nearly all time steps. In the following theorem, we demonstrate that in the absence of assumptions on the disturbance vectors, no estimator can reliably learn the system dynamics.

**Theorem 20.** *Consider the linear system $\mathbf{x}_{t+1} = \bar{\mathbf{A}}\mathbf{x}_t + \bar{\mathbf{d}}_t, \mathbf{x}_0 = \mathbf{0}_n$, together with a linear subspace $D \subset \mathbb{R}^n$ whose dimension is less than rank of $\bar{\mathbf{A}}$. Assume that the disturbance $\bar{\mathbf{d}}_t$ is always chosen from this not-full-dimensional subspace. Then, there does not exist any estimator of the form*

$$\hat{\mathbf{A}}_T \in \arg\min_{\mathbf{A}\in\mathbb{R}^{n\times n}} g(\mathbf{A}; \{\mathbf{x}_t\}_{t=0}^T)$$

*that uniquely recovers the matrix $\bar{\mathbf{A}}$ all the time from the states $\mathbf{x}_0, ..., \mathbf{x}_T$ no matter how large $T$ is. Here, $g(\mathbf{A}; \{\mathbf{x}_t\}_{t=0}^T)$ is the function of $\mathbf{A}$ that depends on the system states over time.*

Consider the scenario where the disturbances affecting the system are adversarially designed. In this case, we refer to these disturbances as attacks, with $p$ representing the frequency at which the system is under attack. Two key problems arise in the context of attack analysis: *(1) attack problem* where the adversary aims to design the attack vectors $\bar{\mathbf{d}}_t$ to maximize the disruption to the system, and *(2) defense problem* where the system operator seeks to detect any suspicious attack and nullify it. Two common strategies often used in combination to design an effective defense mechanism are (i) inspecting the system inputs to detect anomalies indicative of attacks and (ii) analyzing collected state values to infer whether an attack has occurred. Based on Theorem 20, if the attacker has complete freedom in selecting attack values, the most effective strategy is to constrain them to a specific subspace where no estimator can function reliably. However, such biased attack strategies are easier to detect using defense strategy (i), as the operator can analyze the statistical behavior of the inputs and identify deviations from natural disturbances, flagging them as attacks. Thus, the risk for an attacker employing extreme attacks is that they increase the likelihood of detection and nullification by a well-designed defense mechanism. Consequently, a strategic adversary should avoid highly conspicuous attack patterns and instead focus on attacks that have a lower probability of detection.

As an example, consider the first-order system $\mathbf{x}_{t+1} = \mathbf{x}_t + \bar{\mathbf{d}}_t$ where $\bar{\mathbf{d}}_t \in \mathbb{R}$. Suppose this represents a physical system that cannot accept inputs exceeding a given magnitude limit $\gamma$. The most severe attack in this case would be to set $\bar{\mathbf{d}}_t$ to be equal to $\gamma$ at all times, driving the state to grow as rapidly as possible. However, a well-designed defense mechanism could quickly detect this attack by recognizing that the injected input is not a natural disturbance but rather a deliberately crafted adversarial input. To evade detection, the attacker must adopt a more subtle approach. Instead of consistently applying the maximum allowable input, the attacker may alternate between positive and negative disturbances, choosing $\bar{\mathbf{d}}_t$ to be $+\bar{\gamma}$ and $-\bar{\gamma}$ for some $\bar{\gamma} < \gamma$. This strategy achieves two key objectives: (i) avoiding hitting the maximum limit, and (ii) making the attack look like a disturbance by having a zero mean. Although this alternating attack pattern is less detectable, it still has a significant impact on the system. Specifically, it increases the variance of $\mathbf{x}_t$, causing oscillatory behavior that can degrade system performance. This example illustrates a broader principle: an effective attack strategy should involve alternating positive and negative perturbations rather than consistently applying unidirectional inputs. To formally capture this concept, we introduce the notion of attack/disturbance direction and define a semi-oblivious condition in the attack/disturbance process. This condition aligns with existing robustness criteria in the literature [22, 28]. To do so, we define the filtration $\mathcal{F}_t := \sigma\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t\}$.

**Assumption 6** (Semi-oblivious condition)**.** *Conditional on the past information $\mathcal{F}_t$ and the event that $\bar{\mathbf{d}}_t \neq \mathbf{0}_n$, the disturbance direction $\hat{\mathbf{d}}_t = \bar{\mathbf{d}}_t / \|\bar{\mathbf{d}}_t\|_2$ is zero-mean.*

The semi-oblivious condition has been widely studied in real-world systems, particularly in the context of stealthy attacks. To illustrate this concept, consider another example from an energy system comprising two nodes: Node 1 represents a neighborhood of homes, and Node 2 is a power supplier owned by a utility company. Every five minutes, Node 1 reports to Node 2 the amount of electrical power required for the next five-minute period. Suppose the neighborhood requires a constant power demand of 1 unit for the next five hours. However, an attacker has compromised the communication channel between Node 1 and Node 2, altering the reported demand from 1 unit to either $1 - e$ or $1 + e$ every half hour, where $e$ is a large perturbation relative to 1 unit. Since the average of $1 - e$ and $1 + e$ is still 1, this could serve as a semi-oblivious attack. However, this manipulation disrupts the power balance: when Node 1 actually requires 1 unit of power, but Node 2 instead generates either $1 - e$ or $1 + e$, the resulting mismatch violates the fundamental physical constraints of the grid. Such discrepancies can trigger grid instability, potentially leading to large-scale blackouts. Semi-oblivious attacks of this kind have been observed in various parts of the world, resulting in severe power outages. The cyberattack problem in power systems aligns closely with our mathematical models. Notably, power system operators employ hypothesis testing techniques to detect anomalies in reported data. If the injected disturbances exhibit a non-zero mean, they would be flagged as suspicious. However, by maintaining a zero mean, semi-oblivious evade detection while still destabilizing the system.

Due to the poor performance of LSE under the disturbance structure characterized by the semi-oblivious condition and the probabilistic sparsity model, this chapter focuses on analyzing the LASSO-type estimator and understanding the conditions under which exact recovery is attainable

in such settings. We emphasize that this disturbance structure encompasses not only some sparse disturbances but also certain types of attacks. For instance, the disturbance structure may consist of a combination of sparse, zero-mean Gaussian input vectors, denoted as $\bar{\mathbf{e}}_t$, and semi-oblivious attack vectors, $\bar{\mathbf{f}}_t$, both of which satisfy the required assumptions. If the $\bar{\mathbf{e}}_t$ and $\bar{\mathbf{f}}_t$ follow the probabilistic sparsity model with probabilities $p'$ and $p$, respectively, then the overall disturbance, given by $\bar{\mathbf{d}}_T = \bar{\mathbf{e}}_t + \bar{\mathbf{f}}_t$, is also semi-oblivious and follows the probabilistic sparsity model with probability at most $p' + p < 1$.

In the next two sections, we provide lower bounds on the sample complexity $T$ such that the ground truth $\bar{\mathbf{A}}$ is the unique solution to problem (5.4). The next section derives results under Assumption 6. Following that, we extend our analysis to unbounded basis functions, which require Assumption 11. While Assumption 11 is somewhat stronger than Assumption 6, it remains a practical assumption. Note that the latter assumption requires the disturbance directions to have a zero mean meanwhile the former assumption requires the disturbance directions to follow a uniform distribution. In both cases, the magnitudes of the non-zero disturbances/attacks remain unrestricted, meaning that adversaries can inject disturbances of arbitrary values to maximize their impact on the system.

## 5.4 Bounded Basis Function

In this section, we analyze the case where the basis function $f$ is bounded.

**Assumption 7** (Bounded basis function)**.** *The basis function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ satisfies*

$$\|f(\mathbf{x})\|_\infty \leq B, \quad \forall \mathbf{x} \in \mathbb{R}^n,$$

*where $B > 0$ is a constant.*

Finally, to avoid the degenerate case, we assume that the norm of basis function is lower bounded under conditional expectation after a non-zero disturbance.

**Assumption 8** (Non-degenerate condition)**.** *Conditional on the past information $\mathcal{F}_t$ and the event that $\bar{\mathbf{d}}_t \neq \mathbf{0}_n$, the disturbance vector and the basis function satisfy*

$$\lambda_{min}\left[\mathbb{E}\left[f(\mathbf{x}+\bar{\mathbf{d}}_t)f(\mathbf{x}+\bar{\mathbf{d}}_t)^\top \mid \mathcal{F}_t, \bar{\mathbf{d}}_t \neq \mathbf{0}_n\right]\right] \geq \lambda^2, \quad \forall \mathbf{x} \in \mathbb{R}^n,$$

*where $\lambda_{min}(F)$ is the minimal eigenvalue of matrix $\mathbf{F}$ and $\lambda > 0$ is a constant.*

Intuitively, the non-degenerate assumption ensures sufficient exploration of the trajectory in the state space. More specifically, for the condition in (5.9) to hold, it is necessary that the matrix

$$[f(\mathbf{x}_t),\ t \in \mathcal{K}^c] \in \mathbb{R}^{m \times (T-|\mathcal{K}|)} \tag{5.10}$$

has rank-$m$; see the proof of Theorem 22 for further details. The non-degenerate assumption guarantees that the basis function evaluations $f(\mathbf{x}+\bar{\mathbf{d}}_t)$ spans the entire state space in expectation.

As a result, the matrix in (5.10) is full-rank with high probability when the number of samples $T$ is sufficiently large.

The following theorem establishes that when the sample complexity is sufficiently high, the estimator in (5.3) achieves exact recovery of the ground truth matrix $\bar{\mathbf{A}}$ with high probability.

**Theorem 21** (Exact recovery for bounded basis function)**.** *Suppose that Assumptions 6-8 hold and define $\kappa := B/\lambda \geq 1$. For all $\delta \in (0, 1]$, if the sample complexity $T$ satisfies*

$$T \geq \Theta \left[ \frac{m^2 \kappa^4}{p(1-p)^2} \left[ mn \log \left( \frac{m\kappa}{p(1-p)} \right) + \log \left( \frac{1}{\delta} \right) \right] \right], \tag{5.11}$$

*then $\bar{\mathbf{A}}$ is the unique global solution to problem* (5.4) *with probability at least $1 - \delta$.*

The above theorem provides a non-asymptotic bound on the sample complexity required for exact recovery with a specified probability of at least $1 - \delta$. The lower bound scales as $m^3 n$, indicating that the required number of samples increases as the number of states $n$ and the number of basis functions $m$ grow. In addition, the sample complexity increases when $B$ is larger or $\lambda$ is smaller. This aligns with the intuition that $B$ reflects the size of the space spanned by the basis function and $\lambda$ measures the *speed* of exploring the spanned space.

Regarding the dependence on the sparsity probability $p$, the next theorem demonstrates that the scaling factor $1/[p(1-p)]$ is unavoidable under the probabilistic sparsity model. Furthermore, the theorem also establishes a lower bound on the sample complexity that depends on $m$ and $\log(1/\delta)$.

**Theorem 22.** *Suppose that the sample complexity satisfies*

$$T < \frac{m}{2p(1-p)}.$$

*Then, there exists a basis function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ and a disturbance model such that Assumptions 6-8 hold and the global solutions to problem* (5.4) *are not unique with probability at least $\max\left\{1 - 2\exp\left(-m/3\right), 2[p(1-p)]^{T/2}\right\}$. Furthermore, given a constant $\delta \in (0, 1]$, if*

$$T < \max\left\{ \frac{m}{2p(1-p)}, \frac{2}{-\log[p(1-p)]} \log\left(\frac{2}{\delta}\right) \right\},$$

*then the global solutions to problem* (5.4) *are not unique with probability at least*

$$\max\left\{1 - 2\exp\left(-m/3\right), \delta\right\}$$

.

**Remark.** *A key objective of this chapter is to show that the exact recovery in finite time is achievable when $p$ is close to 1, indicating that the system operates under semi-oblivious attacks most of the time. In Theorem 21, we establish an upper bound on the required time horizon for exact recovery, a result with significant implications for real-world systems. Conversely, the lower bound*

*derived in Theorem 22 serves primarily as a theoretical result. Unlike large-scale machine learning problems, where the problem size can reach tens of millions, real-world dynamical systems typically have far fewer states, often fewer than several thousand. As a result, our upper bound already provides a practical estimate of the required sample complexity. While further tightening the lower bound remains a relevant and theoretically interesting problem, its practical impact is likely to be marginal, given that the current upper bound is already within a feasible range for real-world applications.*

## 5.5 Lipschitz Basis Function

In this section, we consider the case when the basis function $f(\mathbf{x})$ is Lipschitz continuous in $\mathbf{x}$. Specifically, we make the following assumption.

**Assumption 9** (Lipschitz basis function)**.** *The basis function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ satisfies*

$$f(\mathbf{0}_n) = \mathbf{0}_m \quad \textit{and} \quad \|f(\mathbf{x}) - f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

*where $L > 0$ is the Lipschitz constant.*

As a special case of Assumption 9, the basis function of a linear system is $f(\mathbf{x}) = \mathbf{x}$, which satisfies the Lipschitz condition with $L = 1$.

**Remark.** *It is important to note that the assumptions of boundedness or Lipschitz continuity are always satisfied in dynamical systems, as the user has the flexibility to select appropriate basis functions to ensure these properties. Specifically, the user can choose an arbitrary set of basis functions to approximate the unknown function as a linear combination of these bases. This setting differs from classical machine learning problems, where the model is trained to learn an unknown function, and there is no direct control over its Lipschitz continuity. Conversely, in dynamical systems, the user can impose constraints on the basis functions to ensure desirable mathematical properties. However, if unbounded basis functions or functions with high Lipschitz constants are restricted, a larger number of basis functions may be required to achieve an accurate approximation of the unknown function. Moreover, many real-world dynamical systems, ranging from robotics to energy systems, inherently exhibit smooth and well-behaved dynamics due to their foundation in physical laws, such as Newtonian mechanics and Kirchhoff's circuit laws. This stands in contrast to various machine learning problems, where the optimal policy may be inherently non-smooth and highly complex, making function approximation significantly more challenging.*

Additionally, we assume that the spectral norm of $\bar{\mathbf{A}}$ is bounded to ensure system stability.

**Assumption 10** (System stability)**.** *The ground truth $\bar{\mathbf{A}}$ satisfies*

$$\rho := \left\|\bar{\mathbf{A}}\right\| < \frac{1}{L}.$$

Assumption 10 is related to the asymptotic stability of the dynamical system and is sufficient to prevent the finite-time divergence of the system trajectories. In Theorem 24, we demonstrate that this stability condition may also be necessary for exact recovery. Finally, we make the additional assumption that each disturbance follows a sub-Gaussian distribution.

**Assumption 11** (Sub-Gaussian disturbances)**.** *Conditional on the filtration $\mathcal{F}_t$ and the event that $\bar{\mathbf{d}}_t \neq \mathbf{0}_n$, the disturbance vector $\bar{\mathbf{d}}_t$ is expressed as the product $\ell_t \hat{\mathbf{d}}_t$, where*

1. *$\hat{\mathbf{d}}_t \in \mathbb{R}^n$ and $\ell_t \in \mathbb{R}$ are independent conditional on $\mathcal{F}_t$ and $\bar{\mathbf{d}}_t \neq \mathbf{0}_n$;*
2. *$\hat{\mathbf{d}}_t$ is a zero-mean unit vector, namely, $\mathbb{E}(\hat{\mathbf{d}}_t \mid \mathcal{F}_t, \bar{\mathbf{d}}_t \neq \mathbf{0}_n) = \mathbf{0}_n$ and $\|\hat{\mathbf{d}}_t\|_2 = 1$;*
3. *$\ell_t$ is zero-mean and sub-Gaussian with parameter $\sigma$.*

As a special case, the sub-Gaussian assumption is guaranteed to hold if there is an upper bound on the magnitude of each disturbance. The bounded disturbance case is common in practical applications since real-world systems do not accept inputs that are arbitrarily large. For example, physical devices have a clear limitation on the input size and the disturbances/attacks cannot exceed that limit. In Assumption 11, the components $\hat{\mathbf{d}}_t$ and $\ell_t$ respectively represent the direction and intensity (such as magnitude) of each attack/disturbance. While the intensity parameters $\ell_t$'s could be correlated over time, both $\hat{\mathbf{d}}_t$ and $\ell_t$ are assumed to be zero-mean, aligning with the semi-oblivious condition stated before.

Under these assumptions, we can establish that high-probability exact recovery is achievable, provided that the sample size $T$ is sufficiently large.

**Theorem 23** (Exact recovery for Lipschitz basis function)**.** *Suppose that Assumptions 8-11 hold and define $\kappa := \sigma L/\lambda \geq 1$. If the sample complexity $T$ satisfies*

$$T \geq \Theta\Bigg[ \max\left\{ \frac{\kappa^{10}}{(1-\rho L)^3(1-p)^2}, \frac{\kappa^4}{p(1-p)} \right\} \times \left[ mn \log\left( \frac{1}{(1-\rho L)\kappa p(1-p)} \right) + \log\left(\frac{1}{\delta}\right) \right] \Bigg], \tag{5.12}$$

*then $\bar{\mathbf{A}}$ is the unique global solution to problem* (5.4) *with probability at least $1-\delta$.*

Theorem 23 provides a non-asymptotic sample complexity bound for the case where the basis function is Lipschitz continuous. As a special case, when the basis function is linear, i.e., $f(\mathbf{x}) = \mathbf{x}$, and the attack/disturbance vector $\bar{\mathbf{d}}_t$ follows the Gaussian distribution $\mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$ conditional on $\mathcal{F}_t$, we have $\kappa = 1$. Compared to Theorem 21, the dependence on the non-zero disturbance probability $p$ is improved from $1/[p(1-p)^2]$ to $1/[p(1-p)]$, a consequence of the stability condition (Assumption 10). In addition, the dependence on the dimension $m$ is improved from $m^3$ to $m$. Intuitively, the improvement is achieved by improving the upper bound on the norm $\|f(\mathbf{x}_t)\|_2$. In the bounded basis function case, the norm is bounded by $\sqrt{m}B$; while in the Lipschitz basis

function case, the norm is bounded by $\sigma L$ with high probability, which is independent from the dimension $m$. Finally, the sample complexity bound grows with the parameter $\kappa = \sigma L/\lambda$ and the gap $1 - \rho L$, which is also consistent with the intuition.

Conversely, we can construct counterexamples demonstrating that when the stability condition (Assumption 10) is violated, exact recovery fails with probability at least $p$.

**Theorem 24** (Failure of exact recovery for unstable systems)**.** *There exists a system such that Assumptions 8, 9 and 11 are satisfied but for all $T \geq 1$, the ground truth $\bar{\mathbf{A}}$ is not a global solution to problem* (5.4) *with probability at least $p[1 - (1 - p)^{T-1}]$.*

## 5.6 Absence of Semi-Oblivious Attacks

In Sections 5.4 and 5.5, we derived the required number of samples for exact recovery in the presence of sparse and semi-oblivious disturbances using bounded and Lipschitz basis functions. For bounded basis functions, the sample complexity scales as $p^{-1}(1-p)^{-2}$ in terms of the non-zero disturbance probability, up to a logarithm factor, as stated in Theorem 21. Similarly, Theorem 23 established that, for Lipschitz basis functions, the sample complexity scales as $\max\{(1-p)^{-2}, p^{-1}(1-p)^{-1}\}$ with respect to the non-zero disturbance probability. Although the theoretical results in these sections are quite promising whenever non-zero disturbance probability $p > 0.5$, it may seem counterintuitive that the sample complexity increases as $p$ decreases and approaches zero, given that a smaller $p$ implies fewer non-zero disturbances and attack injections into the system. This phenomenon can be explained through the concept of exploration. When $p$ is small, the lack of disturbances reduces the rate at which the system state space is explored, thereby increasing the number of samples required for exact recovery. This theoretical observation is further supported by numerical experiments, the results of which are presented in Appendix 5.L.

It is well-known that random excitation signals into dynamical systems can accelerate convergence in system identification problems. To leverage the sample complexity bounds derived in earlier sections and enhance the learning rate, we define the disturbance vector $\bar{\mathbf{d}}_t$ as a combination of the random input vectors injected by the controller, $\bar{\mathbf{e}}_t$, and the semi-oblivious attacks designed by an adversarial agent, $\bar{\mathbf{f}}_t$, such that $\bar{\mathbf{d}}_t = \bar{\mathbf{e}}_t + \bar{\mathbf{f}}_t$. In the absence of random excitation input injections from the controller, we have $\bar{\mathbf{d}}_t = \bar{\mathbf{f}}_t$. In this case, as the non-zero attack disturbance probability $p$ approaches to zero, attaining exact recovery requires more samples. The controller can counteract this trend by injecting random zero-mean i.i.d. Gaussian inputs into the system following the probabilistic sparsity model with probability $p'$. Under this formulation, $\bar{\mathbf{d}}_t$ as the combination of $\bar{\mathbf{e}}_t$ and $\bar{\mathbf{f}}_t$, continues to satisfy the assumptions required for Theorem 21 and Theorem 23. The non-zero disturbance probability for $\bar{\mathbf{d}}_t$ becomes at most $p' + p$, and the exact recovery remains achievable as long as there exist sufficient time periods when both vectors $\bar{\mathbf{e}}_t$ and $\bar{\mathbf{f}}_t$ are zero, i.e., $p' + p < 1$.

For the bounded basis functions, the sample complexity is minimized when the non-zero disturbance probability $p$ is set to $p^* = 1/3$, as the bound in Theorem 21 scales with $p^{-1}(1-p)^{-2}$. As a result, it is optimal to inject zero-mean i.i.d Gaussian inputs with probability $p' = p^* = 1/3$

as this value minimizes the sample complexity bound. Furthermore, for Lipschitz basis functions, the sample complexity scales as $\max\{(1-p)^{-2}, p^{-1}(1-p)^{-1}\}$ according to Theorem 23. In this case, the optimal $p^*$ value is $p^* = 1/2$, meaning that it is preferable to inject random excitations approximately half of the time when the semi-oblivious attacks are nearly absent, i.e., $p \approx 0$. Finally, although $p$ is unknown to the controller, when the probability of attacks is lower than $p^*$ values specified above for bounded and Lipschitz basis functions, injecting excitation signals with zero-mean i.i.d Gaussian inputs at a probability of $p' = p^* - p$ will accelerate the exact recovery rate of the system dynamics.

## 5.7 Numerical Experiments

We conduct numerical experiments for both the bounded and Lipschitz continuous basis function cases to validate the exact recovery guarantees presented in Sections 5.4 and 5.5. Specifically, we examine the convergence behavior of the estimator in (5.3) under varying values of the sparsity probability $p$ and problem dimensions $(n, m)$. Additionally, we numerically verify the necessary and sufficient conditions established in Section 5.2. Further numerical experiments exploring additional problem parameters are provided in Appendix 5.L.

**Evaluation metrics.** Given a trajectory $\{\mathbf{x}_0, \ldots, \mathbf{x}_T\}$, we compute the estimators

$$\hat{\mathbf{A}}^{T'} \in \arg\min_{\mathbf{A}\in\mathbb{R}^{n\times m}} g_{T'}(\mathbf{A}), \quad \forall T' \in \{1, \ldots, T\},$$

where the loss function is defined as $g_{T'}(\mathbf{A}) := \sum_{t=0}^{T'-1} \|\mathbf{x}_{t+1} - \mathbf{A}f(\mathbf{x}_t)\|_2$. In our experiments, we solve the convex optimization problem using the CVX solver [49]. To evaluate the recovery quality of the estimator for each $T'$, we consider the following three metrics:

- The **Loss Gap** is defined as $g_{T'}(\bar{\mathbf{A}}) - g_{T'}(\hat{\mathbf{A}}_{T'})$. The ground truth $\bar{\mathbf{A}}$ is a global solution if and only if the loss gap is $0$.
- The **Solution Gap** is defined as $\|\bar{\mathbf{A}} - \hat{\mathbf{A}}_{T'}\|_F$. The ground truth $\bar{\mathbf{A}}$ is the unique solution only if the solution gap is $0$.
- The **Optimality Certificate** is defined as

  $$\min_{\mathbf{Z}\in\mathbb{R}^{n\times m}} \sum_{t\in\mathcal{K}^c} \|\mathbf{Z}f(\mathbf{x}_t)\|_2 - \sum_{t\in\mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z}f(\mathbf{x}_t) \quad \text{s.t. } \|\mathbf{Z}\|_F \leq 1,$$

  which is a convex optimization problem that can be solved using the CVX solver. The ground truth is a global solution if and only if the optimality certificate is $0$.

We evaluate these metrics to assess the performance of the estimator in (5.3) and validate the proposed optimality conditions. For each parameter setting, we independently generate 10 trajectories using the system dynamics in (5.1) and compute the average of the three metrics.

Since we need to solve estimator (5.3) many times (for different trajectories and steps $T'$), we consider relatively small-scale problems. In practice, the estimator (5.3) is only required for

$T' = T$ and we only need to solve a single optimization problem. As a result, estimator (5.3) can be solved for large-scale real-world systems since it is convex and should be solved only once. We provided experiment results for large-scale systems in Appendix 5.L. Next, we explain the experiment setup and the numerical results for the bounded basis functions.

**Bounded basis function.** Given a state space dimension $n$, we set $m = 5n$ and define the basis function as

$$f(\mathbf{x}) := \begin{bmatrix} \tilde{f}(\mathbf{x}_1) \\ \vdots \\ \tilde{f}(\mathbf{x}_n) \end{bmatrix}, \quad \text{where } \tilde{f}(y) := \begin{bmatrix} \sin(y) \\ \vdots \\ \sin(5y) \end{bmatrix}, \quad \forall \mathbf{x} \in \mathbb{R}^n,\ y \in \mathbb{R}.$$

The basis function satisfies Assumption 7 with $B = 1$. For each time instance $t \in \mathcal{K}$ and for each $i \in \{1, \ldots, n\}$, the disturbance $\bar{\mathbf{d}}_{t,i}$ is independently generated by

$$\bar{\mathbf{d}}_{t,i} \sim \text{Uniform}\left(-c_{t,i}\pi, c_{t,i}\pi\right), \quad \text{where } c_{i,t} := \min\{\max\{|\mathbf{x}_{t,i}|, 0.1\}, 0.5\}.$$

Here, $\bar{\mathbf{d}}_{t,i}$ and $\mathbf{x}_{t,i}$ denote $i$-th components of $\bar{\mathbf{d}}_t$ and $\mathbf{x}_t$, respectively. Since the disturbance distribution is symmetric about the origin, it satisfies Assumption 6. Additionally, as the sine functions $\sin(y), \ldots, \sin(5y)$ are linearly independent, the non-degeneracy condition (Assumption 8) is satisfied. Finally, the ground truth matrix $\bar{\mathbf{A}}$ is constructed such that

$$\bar{\mathbf{A}} f(\mathbf{x}) = \begin{bmatrix} \sum_{k=1}^{5} \bar{\mathbf{A}}_{1,k} \sin(k\mathbf{x}_1) \\ \vdots \\ \sum_{k=1}^{5} \bar{\mathbf{A}}_{n,k} \sin(k\mathbf{x}_n) \end{bmatrix},$$

where the coefficients are sampled independently as

$$\bar{\mathbf{A}}_{i,k} \overset{\text{i.i.d.}}{\sim} \text{Uniform}(-100, 100), \quad \forall i \in \{1, \ldots, n\},\ k \in \{1, \ldots, 5\}.$$

We choose a large upper bound for the coefficients $\bar{\mathbf{A}}_{i,k}$ (i.e., greater than 1) to illustrate that the stability condition (Assumption 10) is not required in the case of bounded basis functions.

We first compare the performance of estimator (5.3) under different non-zero disturbance probabilities $p$. Specifically, we set $T = 500$, $n = 1$ and $p \in \{0.7, 0.8, 0.85\}$. The results, presented in Figure 5.1, align with the Theorem 21. In particular, the optimality certificate accurately reflects the exact recovery of the estimator in (5.3), and the required sample complexity increases as the attack/disturbance probability $p$ grows.

Next, we analyze the estimator's performance across different problem dimensions $(n, m)$. We set $T = 500$, $p = 0.7$, and evaluate the cases where $n \in \{1, 2, 4\}$. The results are plotted in Figure 5.2 demonstrate that exact recovery requires more samples as $(n, m)$ increases, further validating the theoretical results established in Theorem 21. After verifying the behavior of the estimator for bounded basis functions, we proceed to simulate its performance using Lipschitz continuous basis functions.

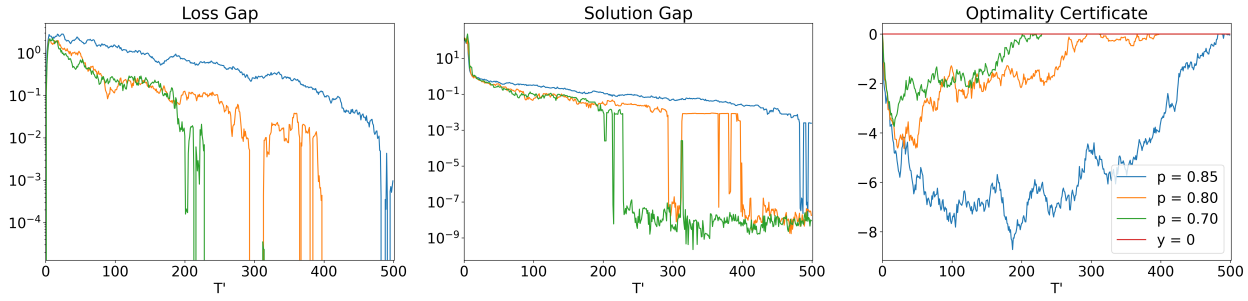


Figure 5.1: Loss Gap, Solution Gap, and Optimality Certificate of the Bounded Basis Function Case with Attack/Disturbance Probability $p = 0.7, 0.8$ and $0.85$.

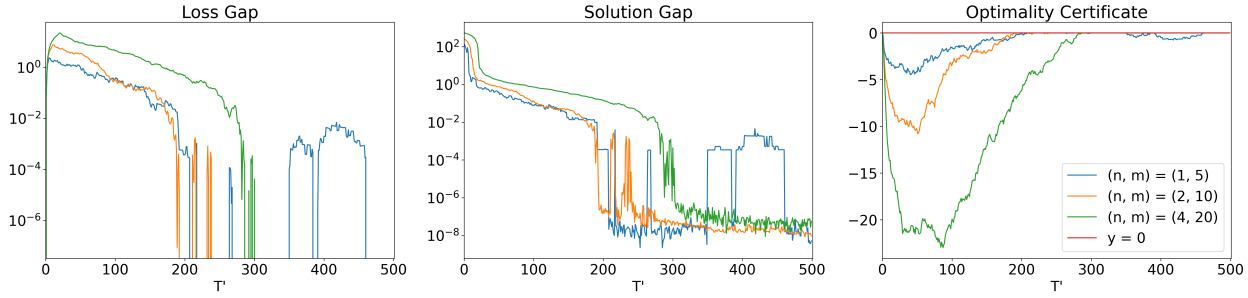


Figure 5.2: Loss Gap, Solution Gap, and Optimality Certificate of the Bounded Basis Function Case with Dimension $(n, m) = (1, 5), (2, 10)$ and $(4, 20)$.

**Lipschitz basis function.** Given a state space dimension $n$, we set $m = n$ and define the basis function as

$$f(\mathbf{x}) := \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\|\mathbf{x} - \mathbf{x}_1\|_2^2 + 1} - \sqrt{\|\mathbf{x}_1\|_2^2 + 1} \\ \vdots \\ \sqrt{\|\mathbf{x} - \mathbf{x}_n\|_2^2 + 1} - \sqrt{\|\mathbf{x}_n\|_2^2 + 1} \end{bmatrix}, \quad \forall \mathbf{x} \in \mathbb{R}^n,$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^n$ are i.i.d. standard Gaussian random vectors. We can verify that the basis function is Lipschitz continuous with a Lipschitz constant of $L = 1$, thus satisfying Assumption 9. For each time instance $t \in \mathcal{K}$, the disturbance $\bar{\mathbf{d}}_t$ is generated by

$$\bar{\mathbf{d}}_t := \ell_t \hat{\mathbf{d}}_t, \quad \text{where } \ell_t \sim \mathcal{N}(0, \sigma_t^2),\ \hat{\mathbf{d}}_t \sim \text{uniform}(\mathbb{S}^{n-1}),\ \ell_t \text{ and } \hat{\mathbf{d}}_t \text{ are independent.}$$

Here, we define $\sigma_t^2 := \min\{\|\mathbf{x}_t\|_2^2, 1/n\}$. We verify that the random variable $\ell_t$ is zero-mean and sub-Gaussian with parameter $\sigma = 1$. Additionally, since the random vector $\hat{\mathbf{d}}_t$ follows a uniform distribution on the unit sphere $\mathbb{S}^{n-1}$, Assumption 11 is satisfied. Note that $\bar{\mathbf{d}}_0, \ldots, \bar{\mathbf{d}}_{T-1}$ are correlated, violating the i.i.d. assumption commonly made in the literature. Our disturbance

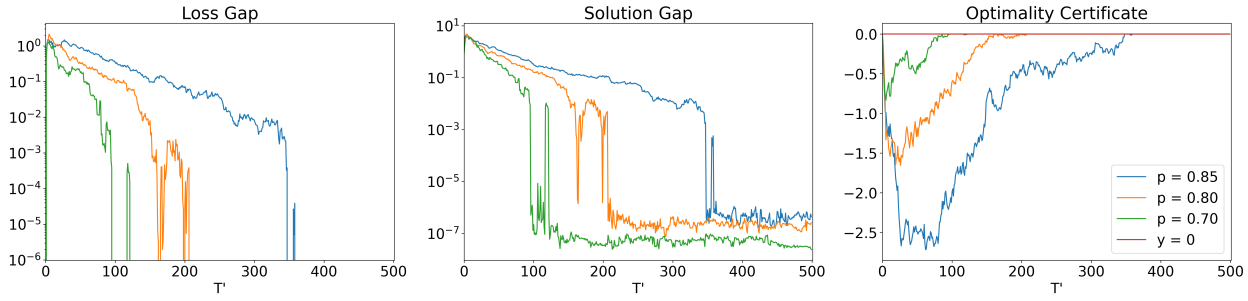


Figure 5.3: Loss Gap, Solution Gap, and Optimality Certificate of the Lipschitz Basis Function Case with Attack/Disturbance Probability $p = 0.7, 0.8$ and $0.85$.

model further implies that the intensity of a disturbance vector (represented by $\ell_t$) depends on the current state, which itself is influenced by previous disturbances Since the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are randomly generated, the multiquadric radial basis functions are linearly independent with probability 1. Functions $g_1(y), \ldots, g_k(y)$ are said to be linearly independent if there do not exist constants $c_1, \ldots, c_k$ such that $\sum_{i=1}^{k} c_i g_i(y) = 0$ for all $y$. Thus, the non-degeneracy condition (Assumption 8) is satisfied. Finally, the ground truth matrix $\bar{\mathbf{A}}$ is constructed as $U\Sigma V^\top$, where $U, V \in \mathbb{R}^{n\times n}$ are random orthogonal matrices and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ is a diagonal matrix. The singular values $\sigma_i$ are independently drawn from a uniform distribution:

$$\sigma_i \overset{\text{i.i.d.}}{\sim} \mathrm{uniform}(0, \rho), \quad \forall i \in \{1, \ldots, n\},$$

where $\rho > 0$ is the upper bound on the spectral norm of $\bar{\mathbf{A}}$.

We first evaluate the performance of the estimator in (5.3) under varying values of the sparsity probability $p$. We set $T = 500$, $n = 3$, and consider $p \in \{0.7, 0.8, 0.85\}$. Additionally, we set the upper bound $\rho$ to be 1, which guarantees the stability condition (Assumption 10). The results are plotted in Figure 5.3. The results demonstrate that both the loss gap and solution gap converge to zero as the number of samples $T$ increases. This implies that the estimator in (5.3) achieves exact recovery of the ground truth matrix $\bar{\mathbf{A}}$ when a sufficient number of samples is available. Furthermore, the optimality certificate also converges to zero simultaneously with the solution gap, thereby confirming the validity of the necessary and sufficient conditions established in Section 5.2. Additionally, we observe that the required number of samples increases with the non-zero disturbance probability $p$, which is consistent with the upper bound derived in Theorem 23.

Next, we evaluate the performance of the estimator in (5.3) across different problem dimensions $(n, m)$. Specifically, we set $T = 500$, $p = 0.75$, and $\rho = 1$, while varying $n \in \{3, 5, 7\}$. The results are presented in Figure 5.4. We observe that as the problem dimension $(n, m)$ increases, a larger number of samples is required to ensure exact recovery. This finding aligns with the theoretical bound established in Theorem 21.

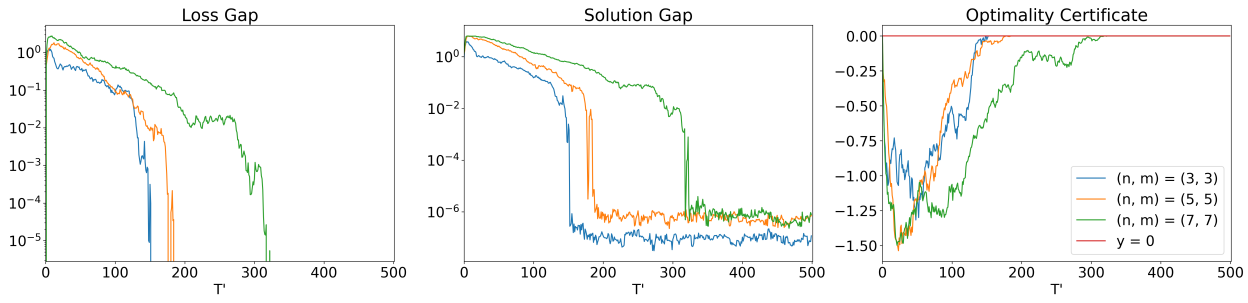


Figure 5.4: Loss Gap, Solution Gap, and Optimality Certificate of the Lipschitz Basis Function Case with Dimension $(n, m) = (3, 3), (5, 5)$ and $(7, 7)$.

# Appendices

## 5.A Comparing Results to Existing Work

**Example 5** (First-order systems)**.** *In the special case when $n = m = 1$ and the basis function is $f(x) = x$, condition* (5.7) *reduces to*

$$\left| \sum_{t \in \mathcal{K}} \hat{d}_t x_t \right| \leq \sum_{t \in \mathcal{K}^c} |x_t|,$$

*which is the same as Theorem 1 in [43].*

**Example 6** (Linear systems)**.** *We consider the case when $m = n$ and the basis function is $f(\mathbf{x}) = \mathbf{x}$. We also assume the $\Delta$-spaced disturbance model; see the definition in [119]. By considering the disturbance period starting at the time step $t_1$, a sufficient condition to guarantee condition* (5.5) *is given by*

$$\hat{\mathbf{d}}^\top \mathbf{Z} \bar{\mathbf{A}}^{\Delta-1} \bar{\mathbf{d}}_{t_1} \leq \sum_{t=0}^{\Delta-2} \left\| \mathbf{Z} \bar{\mathbf{A}}^t \bar{\mathbf{d}}_{t_1} \right\|_2, \quad \forall \mathbf{Z} \in \mathbb{R}^{n \times n}, \tag{5.13}$$

*where we denote $\hat{\mathbf{d}} := \hat{\mathbf{d}}_{t_1}$ for simplicity. Let $\hat{\mathbf{d}} \in \mathbb{R}^{n \times (n-1)}$ be the matrix of orthonormal bases of the orthogonal complementary space of $f$, namely, $\hat{\mathbf{d}}^\top \hat{\mathbf{d}} = \mathbf{0}$, $\hat{\mathbf{d}}^\top \hat{\mathbf{d}} = \mathbf{I}_{n-1}$, and $\hat{\mathbf{d}} \hat{\mathbf{d}}^\top = \mathbf{I}_n - \hat{\mathbf{d}} \hat{\mathbf{d}}^\top$. Then, we can calculate that*

$$\left\| \mathbf{Z} \bar{\mathbf{A}}^t \bar{\mathbf{d}}_{t_1} \right\|_2^2 \geq \left( \mathbf{Z} \bar{\mathbf{A}}^t \bar{\mathbf{d}}_{t_1} \right)^\top \hat{\mathbf{d}} \hat{\mathbf{d}}^\top \left( \mathbf{Z} \bar{\mathbf{A}}^t \bar{\mathbf{d}}_{t_1} \right),$$

*where the equality holds when $\hat{\mathbf{d}}^\top \mathbf{Z} \bar{\mathbf{A}}^t \bar{\mathbf{d}}_{t_1} = \mathbf{0}$, i.e., $\mathbf{Z} \bar{\mathbf{A}}^t \bar{\mathbf{d}}_{t_1}$ is parallel with $\hat{\mathbf{d}}$. Therefore, for condition* (5.13) *to hold, it is equivalent to consider $\mathbf{Z}$ with the form $\mathbf{Z} = \hat{\mathbf{d}} \mathbf{z}^\top$ for some vector $\mathbf{Z} \in \mathbb{R}^n$. In this case, condition* (5.13) *reduces to*

$$\mathbf{z}^\top \bar{\mathbf{A}}^{\Delta-1} \bar{\mathbf{d}}_{t_1} \leq \sum_{t=0}^{\Delta-2} \left| \mathbf{z}^\top \bar{\mathbf{A}}^t \bar{\mathbf{d}}_{t_1} \right|, \quad \forall z \in \mathbb{R}^n. \tag{5.14}$$

*Condition* (5.14) *leads to a better sufficient condition than that in [119]. To illustrate the improvement, we consider the special case when the ground truth matrix is $\bar{\mathbf{A}} = \lambda \mathbf{I}_n$ for some $\lambda \in \mathbb{R}$.*

*Then, condition* (5.14) *becomes*

$$|\lambda|^{\Delta-1} \leq \sum_{t=0}^{\Delta-2} |\lambda|^t = \frac{1-|\lambda|^{\Delta-1}}{1-|\lambda|} \quad \text{, which is further equivalent to } |\lambda| + |\lambda|^{1-\Delta} \leq 2,$$

*which is a stronger condition than that in [119]. When the disturbance period $\Delta$ is large, we approximately have $|\lambda| \leq 2 - 2^{1-\Delta}$, which is a better condition than that in Figure 1 of [119].*

**Example 7** (First-order linear systems)**.** *In the case when $m = n = 1$ and $f(x) = x$, our results state that the uniqueness of global solutions is equivalent to*

$$\left| \sum_{t \in \mathcal{K}} \hat{d}_t x_t \right| < \sum_{t \in \mathcal{K}^c} |x_t|. \tag{5.15}$$

*As a comparison, the sufficient condition in Theorem 1 in [43] is*

$$\sum_{t \in \mathcal{K}} |x_t| < \sum_{t \in \mathcal{K}^c} |x_t|.$$

*Since $|\hat{d}_t| = 1$ for all $t \in \mathcal{K}$, our results* (5.15)*, as well as Theorem 19, are more general and stronger than that in [43].*

## 5.B Proof of Theorem 18

*Proof of Theorem 18.* Since problem (5.4) is convex in $\mathbf{A}$, the ground truth matrix $\bar{\mathbf{A}}$ is a global optimum if and only if

$$\mathbf{0} \in \sum_{t \in \mathcal{K}^c} f(\mathbf{x}_t) \otimes \partial \|\mathbf{0}_n\|_2 + \sum_{t \in \mathcal{K}} f(\mathbf{x}_t) \otimes \hat{\mathbf{d}}_t. \tag{5.16}$$

Using the form of the subgradient of the $\ell_2$-norm, condition (5.16) holds if and only if there exist vectors

$$\mathbf{g}_t \in \mathbb{R}^n, \quad \forall t \in \mathcal{K}^c$$

such that

$$\sum_{t \in \mathcal{K}^c} f(\mathbf{x}_t)\mathbf{g}_t^\top + \sum_{t \in \mathcal{K}} f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top = \mathbf{0}_{n \times n}, \quad \|\mathbf{g}_t\|_2 \leq 1, \quad \forall t \in \mathcal{K}^c. \tag{5.17}$$

Define the matrices

$$\mathbf{B} := \begin{bmatrix} f(\mathbf{x}_t) & \forall t \in \mathcal{K}^c \end{bmatrix} \in \mathbb{R}^{m \times (T - |\mathcal{K}|)}, \quad \mathbf{V} := \begin{bmatrix} f(\mathbf{x}_t) & \forall t \in \mathcal{K} \end{bmatrix} \in \mathbb{R}^{m \times |\mathcal{K}|},$$
$$\mathbf{G} := \begin{bmatrix} \mathbf{g}_t & \forall t \in \mathcal{K}^c \end{bmatrix} \in \mathbb{R}^{n \times (T - |\mathcal{K}|)}, \quad \mathbf{F} := \begin{bmatrix} \hat{\mathbf{d}}_t & \forall t \in \mathcal{K} \end{bmatrix} \in \mathbb{R}^{n \times |\mathcal{K}|}.$$

Condition (5.17) can be written as a combination of second-order cone constraints and linear constraints:

$$\exists \mathbf{G} \in \mathbb{R}^{n\times(T-|\mathcal{K}|)}, s, r \in \mathbb{R} \quad \text{s.t. } \mathbf{B}\mathbf{G}^\top + \mathbf{V}\mathbf{F}^\top = \mathbf{0}_{m\times n}, \quad \|\mathbf{G}_{:,t}\|_2 \le s,\ \forall t, \\ s + r = 1, \quad s, r \ge 0, \tag{5.18}$$

where $\mathbf{G}_{:,t}$ is the $t$-th column of $G$ for all $t \in \{1, \dots, T - |\mathcal{K}|\}$. We define the closed convex cone

$$\mathcal{S} := \left\{ \mathbf{z} \in \mathbb{R}^{(T-|\mathcal{K}|)n+2} \middle| \sqrt{\sum_{i=1}^{n} \mathbf{z}(T-|\mathcal{K}|)i + t^2} \le \mathbf{z}(T-|\mathcal{K}|)n + 1,\ \forall t \in \{0, \dots, T-|\mathcal{K}|-1\}, \right. \\ \left. \mathbf{z}(T-|\mathcal{K}|)n + 1, \mathbf{z}(T-|\mathcal{K}|)n + 2 \ge 0 \right\},$$

and we define the matrix and vector

$$\mathcal{A} := \begin{bmatrix} \mathbf{I}_n \otimes \mathbf{B} & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \in \mathbb{R}^{(mn+1)\times[(T-|\mathcal{K}|)n+2]}, \quad \mathbf{b} := \begin{bmatrix} -(\mathbf{V}\mathbf{F}^\top)_{:,1} \\ -(\mathbf{V}\mathbf{F}^\top)_{:,2} \\ \vdots \\ -(\mathbf{V}\mathbf{F}^\top)_{:,n} \\ 1 \end{bmatrix} \in \mathbb{R}^{mn+1},$$

where $(\mathbf{V}\mathbf{F}^\top)_{:,i}$ is the $i$-th column of $\mathbf{V}\mathbf{F}^\top$. Then, condition (5.18) can be equivalently written as

$$\exists \mathbf{z} \in \mathbb{R}^{(T-|\mathcal{K}|)n+2} \quad \text{s.t. } \mathcal{A}\mathbf{z} = \mathbf{b}, \quad \mathbf{z} \in \mathcal{S}. \tag{5.19}$$

Since the cone $\mathcal{S}$ is closed and convex, we can apply the *generalized Farka's lemma* to conclude that condition (5.19) is equivalent to

$$\forall \mathbf{y} \in \mathbb{R}^{mn+1}, \quad \left(\mathcal{A}^\top \mathbf{y} \in \mathcal{S}^* \implies \mathbf{b}^\top \mathbf{y} \ge 0\right), \tag{5.20}$$

where $\mathcal{S}^*$ is the dual cone of $\mathcal{S}$. It can be verified that the dual cone is

$$\mathcal{S}^* = \left\{ \mathbf{z} \in \mathbb{R}^{(T-|\mathcal{K}|)n+2} \middle| \sum_{t=0}^{T-|\mathcal{K}|-1} \sqrt{\sum_{i=1}^{n} \mathbf{z}(T-|\mathcal{K}|)i + t^2} \le \mathbf{z}(T-|\mathcal{K}|)n + 1, \right. \\ \left. \mathbf{z}(T-|\mathcal{K}|)n + 1, \mathbf{z}(T-|\mathcal{K}|)n + 2 \ge 0 \right\}.$$

We can equivalently write condition (5.20) as

$$\forall \mathbf{Z} \in \mathbb{R}^{n\times m},\ p \in \mathbb{R}, \quad \left(\|\mathbf{Z}\mathbf{B}\|_{2,1} \le p, \quad p \ge 0 \implies \langle \mathbf{V}\mathbf{F}^\top, \mathbf{Z}^\top \rangle \le p\right),$$

By eliminating variable $p$, we get

$$\langle \mathbf{V}\mathbf{F}^\top, \mathbf{Z}^\top \rangle \leq \|\mathbf{Z}\mathbf{B}\|_{2,1}, \quad \forall \mathbf{Z} \in \mathbb{R}^{n \times m},$$

where the $\ell_{2,1}$-norm is defined as

$$\|\mathbf{M}\|_{2,1} := \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} M_{ij}^2}, \quad \forall \mathbf{M} \in \mathbb{R}^{n \times m}.$$

The above condition is equivalent to condition (5.5), and this completes the proof. □

## 5.C Proof of Corollary 5

*Proof of Corollary 5.* The sufficient condition follows from the fact that $\|\hat{\mathbf{d}}_t\|_2 = 1$ and

$$\hat{\mathbf{d}}_t^\top \mathbf{Z} f(\mathbf{x}_t) \leq \|\mathbf{Z} f(\mathbf{x}_t)\|_2, \quad \forall t \in \mathcal{K}.$$

This completes the proof. □

## 5.D Proof of Corollary 6

*Proof of Corollary 6.* We choose

$$\mathbf{Z} := \frac{\sum_{t \in \mathcal{K}} \hat{\mathbf{d}}_t f(\mathbf{x}_t)^\top}{\left\| \sum_{t \in \mathcal{K}} \hat{\mathbf{d}}_t f(\mathbf{x}_t)^\top \right\|_F}.$$

Then, condition (5.5) implies

$$\left\| \sum_{t \in \mathcal{K}} f(\mathbf{x}_t) \hat{\mathbf{d}}_t^\top \right\|_F = \sum_{t \in \mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z} f(\mathbf{x}_t) \leq \sum_{t \in \mathcal{K}^c} \|\mathbf{Z} f(\mathbf{x}_t)\|_2 \leq \sum_{t \in \mathcal{K}^c} \|f(\mathbf{x}_t)\|_2,$$

where the last step is because $\|\mathbf{Z}\| \leq \|\mathbf{Z}\|_F = 1$. Now, suppose that the basis dimension is $m = 1$. In this case, we have

$$\sum_{t \in \mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z} f(\mathbf{x}_t) = \left( \sum_{t \in \mathcal{K}} f(\mathbf{x}_t) \hat{\mathbf{d}}_t \right)^\top \mathbf{Z}^\top \leq \left\| \sum_{t \in \mathcal{K}} f(\mathbf{x}_t) \hat{\mathbf{d}}_t \right\|_F \|\mathbf{Z}\|,$$
$$\sum_{t \in \mathcal{K}^c} \|\mathbf{Z} f(\mathbf{x}_t)\|_2 = \sum_{t \in \mathcal{K}^c} |f(\mathbf{x}_t)| \|\mathbf{Z}\| = \sum_{t \in \mathcal{K}^c} \|f(\mathbf{x}_t)\|_2 \|\mathbf{Z}\|.$$

Combining the above two inequalities shows that condition (5.7) is also a sufficient condition. □

## 5.E Proof of Theorem 19

We establish the sufficient and the necessary parts of Theorem 19 by the following two lemmas.

**Lemma 15** (Sufficient condition for uniqueness)**.** *Suppose that condition* (5.5) *holds. If for every non-zero* $\mathbf{Z} \in \mathbb{R}^{n\times m}$ *such that*

$$\sum_{t\in\mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z} f(\mathbf{x}_t) = \sum_{t\in\mathcal{K}^c} \|\mathbf{Z} f(\mathbf{x}_t)\|_2,$$

*it holds that*

$$\sum_{t\in\mathcal{K}} \left|\hat{\mathbf{d}}_t^\top \mathbf{Z} f(\mathbf{x}_t)\right| < \sum_{t\in\mathcal{K}} \|\mathbf{Z} f(\mathbf{x}_t)\|_2.$$

*Then, the ground truth matrix* $\bar{\mathbf{A}}$ *is the unique global solution to problem* (5.4)*.*

*Proof.* The ground truth $\bar{\mathbf{A}}$ is the unique solution if and only if for every matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ such that $\mathbf{A} \neq \bar{\mathbf{A}}$, the loss function of $\mathbf{A}$ is larger than that of $\bar{\mathbf{A}}$, namely,

$$\sum_{t\in\mathcal{K}} \|\bar{\mathbf{d}}_t\|_2 < \sum_{t\in\mathcal{K}^c} \|(\bar{\mathbf{A}} - \mathbf{A}) f(\mathbf{x}_t)\|_2 + \sum_{t\in\mathcal{K}} \|(\bar{\mathbf{A}} - \mathbf{A}) f(\mathbf{x}_t) + \bar{\mathbf{d}}_t\|_2. \tag{5.21}$$

Denote

$$\mathbf{Z} := \mathbf{A} - \bar{\mathbf{A}} \in \mathbb{R}^{n\times m}.$$

The inequality (5.21) becomes

$$\sum_{t\in\mathcal{K}^c} \| - \mathbf{Z} f(\mathbf{x}_t)\|_2 + \sum_{t\in\mathcal{K}} \big(\| - \mathbf{Z} f(\mathbf{x}_t) + \bar{\mathbf{d}}_t\|_2 - \|\bar{\mathbf{d}}_t\|_2\big) > 0. \tag{5.22}$$

Since problem (5.4) is convex in $\mathbf{A}$, it is sufficient to guarantee that $\bar{\mathbf{A}}$ is a strict local minimum. Therefore, the uniqueness of global solutions can be formulated as

$$\text{condition (5.22) holds}, \quad \forall \mathbf{Z} \in \mathbb{R}^{n\times m} \quad \text{s.t. } 0 < \|\mathbf{Z}\|_F \leq \epsilon, \tag{5.23}$$

where $\epsilon > 0$ is a sufficiently small constant. In the following, we fix the direction $Z$ and discuss two different cases.

**Case I.** We first consider the case when condition (5.5) holds strictly, namely,

$$\sum_{t\in\mathcal{K}^c} \|\mathbf{Z} f(\mathbf{x}_t)\|_2 - \sum_{t\in\mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z} f(\mathbf{x}_t) > 0.$$

Since the $\ell_2$-norm is a convex function, it holds that

$$\| - \mathbf{Z} f(\mathbf{x}_t) + \bar{\mathbf{d}}_t\|_2 - \|\bar{\mathbf{d}}_t\|_2 \geq \big\langle \partial \|\bar{\mathbf{d}}_t\|_2, -\mathbf{Z} f(\mathbf{x}_t)\big\rangle = -\hat{\mathbf{d}}_t^\top \mathbf{Z} f(\mathbf{x}_t).$$

Therefore, we get

$$\begin{aligned}&\sum_{t\in\mathcal{K}^c}\|-\mathbf{Z}f(\mathbf{x}_t)\|_2+\sum_{t\in\mathcal{K}}\left(\|-\mathbf{Z}f(\mathbf{x}_t)+\bar{\mathbf{d}}_t\|_2-\|\bar{\mathbf{d}}_t\|_2\right)\\&\geq\sum_{t\in\mathcal{K}^c}\|-\mathbf{Z}f(\mathbf{x}_t)\|_2+\sum_{t\in\mathcal{K}}-\hat{\mathbf{d}}_t^\top\mathbf{Z}f(\mathbf{x}_t)>0,\end{aligned}$$

which exactly leads to inequality (5.22).

**Case II.** Next, we consider the case when

$$\sum_{t\in\mathcal{K}}\hat{\mathbf{d}}_t^\top\mathbf{Z}f(\mathbf{x}_t)=\sum_{t\in\mathcal{K}^c}\|\mathbf{Z}f(\mathbf{x}_t)\|_2,\quad\sum_{t\in\mathcal{K}}\left|\hat{\mathbf{d}}_t^\top\mathbf{Z}f(\mathbf{x}_t)\right|<\sum_{t\in\mathcal{K}}\|\mathbf{Z}f(\mathbf{x}_t)\|_2.\tag{5.24}$$

Since $\epsilon$ is a sufficiently small constant, we know

$$\bar{\mathbf{d}}_t^\alpha:=-\alpha\mathbf{Z}f(\mathbf{x}_t)+\bar{\mathbf{d}}_t\neq0,\quad\forall\alpha\in[0,1],$$

and the $\ell_2$-norm is second-order continuously differentiable in an open set that contains the line. Therefore, the *mean value theorem* implies that there exists $\alpha\in[0,1]$ such that for each $t\in\mathcal{K}$, it holds

$$\begin{aligned}\|-\mathbf{Z}f(\mathbf{x}_t)+\bar{\mathbf{d}}_t\|_2-\|\bar{\mathbf{d}}_t\|_2=&\left\langle\hat{\mathbf{d}}_t,-\mathbf{Z}f(\mathbf{x}_t)\right\rangle\\&+\frac{1}{2}\left[-\mathbf{Z}f(\mathbf{x}_t)\right]^\top\left(\frac{\mathbf{I}_n}{\|\bar{\mathbf{d}}_t^\alpha\|_2}-\frac{\bar{\mathbf{d}}_t^\alpha\left(\bar{\mathbf{d}}_t^\alpha\right)^\top}{\|\bar{\mathbf{d}}_t^\alpha\|_2^3}\right)\left[-\mathbf{Z}f(\mathbf{x}_t)\right].\end{aligned}\tag{5.25}$$

We can calculate that

$$\left[-\mathbf{Z}f(\mathbf{x}_t)\right]^\top\left(\frac{\mathbf{I}}{\|\bar{\mathbf{d}}_t^\alpha\|_2}-\frac{\bar{\mathbf{d}}_t^\alpha\left(\bar{\mathbf{d}}_t^\alpha\right)^\top}{\|\bar{\mathbf{d}}_t^\alpha\|_2^3}\right)\left[-\mathbf{Z}f(\mathbf{x}_t)\right]=\frac{\|\mathbf{Z}f(\mathbf{x}_t)\|_2^2}{\|\bar{\mathbf{d}}_t^\alpha\|_2}-\frac{\left\langle\bar{\mathbf{d}}_t^\alpha,\mathbf{Z}f(\mathbf{x}_t)\right\rangle^2}{\|\bar{\mathbf{d}}_t^\alpha\|_2^3}\geq0,\tag{5.26}$$

where the equality holds if and only if $\mathbf{Z}f(\mathbf{x}_t)$ is parallel with $\bar{\mathbf{d}}_t^\alpha$. By the definition of $\bar{\mathbf{d}}_t^\alpha$, the equality holds if and only if $\mathbf{Z}f(\mathbf{x}_t)$ is parallel with $\bar{\mathbf{d}}_t$, which is further equivalent to

$$\left|\left\langle\hat{\mathbf{d}}_t,\mathbf{Z}f(\mathbf{x}_t)\right\rangle\right|=\left\|\mathbf{Z}f(\mathbf{x}_t)\right\|_2.$$

Substituting (5.25) and (5.26) into (5.22), we have

$$\begin{aligned}\sum_{t\in\mathcal{K}^c}\|-\mathbf{Z}f(\mathbf{x}_t)\|_2+\sum_{t\in\mathcal{K}}\left(\|-\mathbf{Z}f(\mathbf{x}_t)+\bar{\mathbf{d}}_t\|_2-\|\bar{\mathbf{d}}_t\|_2\right)&\geq\sum_{t\in\mathcal{K}^c}\|\mathbf{Z}f(\mathbf{x}_t)\|_2-\sum_{t\in\mathcal{K}}\left\langle\hat{\mathbf{d}}_t,\mathbf{Z}f(\mathbf{x}_t)\right\rangle\\&=0,\end{aligned}$$

where the equality holds if and only if

$$\left|\left\langle \hat{\mathbf{d}}_t, \mathbf{Z}f(\mathbf{x}_t)\right\rangle\right| = \|\mathbf{Z}f(\mathbf{x}_t)\|_2, \quad \forall t \in \mathcal{K}.$$

Considering the second condition in (5.24), the above equality condition is violated by some $t \in \mathcal{K}$. Therefore, we have proven that condition (5.22) holds strictly.

Combining the two cases, we complete the proof. □

Next, we prove that the condition in Lemma 15 is also necessary for the uniqueness.

**Lemma 16** (Necessary condition for uniqueness)**.** *Suppose that condition* (5.5) *holds. If the ground truth matrix* $\bar{\mathbf{A}}$ *is the unique global solution to problem* (5.4)*, then for every non-zero* $\mathbf{Z} \in \mathbb{R}^{n\times m}$*, we have*

$$\sum_{t\in\mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z}f(\mathbf{x}_t) < \sum_{t\in\mathcal{K}^c} \|\mathbf{Z}f(\mathbf{x}_t)\|_2 \quad \textit{or} \quad \sum_{t\in\mathcal{K}} \left|\hat{\mathbf{d}}_t^\top \mathbf{Z}f(\mathbf{x}_t)\right| < \sum_{t\in\mathcal{K}} \|\mathbf{Z}f(\mathbf{x}_t)\|_2. \tag{5.27}$$

*Proof.* Assume conversely that there exists a non-zero $\mathbf{Z} \in \mathbb{R}^{n\times m}$ such that

$$\sum_{t\in\mathcal{K}} \hat{\mathbf{d}}_t^\top \mathbf{Z}f(\mathbf{x}_t) = \sum_{t\in\mathcal{K}^c} \|\mathbf{Z}f(\mathbf{x}_t)\|_2, \quad \sum_{t\in\mathcal{K}} \left|\hat{\mathbf{d}}_t^\top \mathbf{Z}f(\mathbf{x}_t)\right| = \sum_{t\in\mathcal{K}} \|\mathbf{Z}f(\mathbf{x}_t)\|_2. \tag{5.28}$$

Without loss of generality, we assume that

$$0 < \|\mathbf{Z}\| \leq \epsilon$$

for a sufficiently small $\epsilon$. In this case, the second condition in (5.28) implies that

$$\left|\hat{\mathbf{d}}_t^\top \mathbf{Z}f(\mathbf{x}_t)\right| = \|\mathbf{Z}f(\mathbf{x}_t)\|_2, \text{ and } \quad \mathbf{Z}f(\mathbf{x}_t) \text{ is parallel with } \bar{\mathbf{d}}_t, \quad \forall t \in \mathcal{K}.$$

Therefore, when $\epsilon$ is sufficiently small, equations (5.26) and (5.24) lead to

$$\| - \mathbf{Z}f(\mathbf{x}_t) + \bar{\mathbf{d}}_t\|_2 - \|\bar{\mathbf{d}}_t\|_2 = -\left\langle \hat{\mathbf{d}}_t, \mathbf{Z}f(\mathbf{x}_t)\right\rangle, \quad \forall t \in \mathcal{K}.$$

We now show that condition (5.22) fails:

$$\begin{aligned}\sum_{t\in\mathcal{K}^c} \| - \mathbf{Z}f(\mathbf{x}_t)\|_2 + \sum_{t\in\mathcal{K}} \left(\| - \mathbf{Z}f(\mathbf{x}_t) + \bar{\mathbf{d}}_t\|_2 - \|\bar{\mathbf{d}}_t\|_2\right) &= \sum_{t\in\mathcal{K}} \left\langle \hat{\mathbf{d}}_t, \mathbf{Z}f(\mathbf{x}_t)\right\rangle - \sum_{t\in\mathcal{K}} \left\langle \hat{\mathbf{d}}_t, \mathbf{Z}f(\mathbf{x}_t)\right\rangle \\ &= 0.\end{aligned}$$

This contradicts the assumption that $\bar{\mathbf{A}}$ is the unique solution to the problem (5.4). □

Combining Lemmas 15 and 16, we have the following necessary and sufficient condition for the uniqueness of the ground truth solution $\bar{\mathbf{A}}$.

## 5.F Proof of Theorem 20

*Proof.* Without loss of generality, we can assume that $\bar{\mathbf{A}}$ is full-rank. Otherwise, the system space can be reduced to the subspace spanned by $\bar{\mathbf{A}}$. Similarly, we assume that $\bar{\mathbf{A}}$ is diagonalizable, such that $\bar{\mathbf{A}} = \mathbf{P}\bar{\mathbf{\Lambda}}\mathbf{P}^{-1}$. If this is not the case, define $\tilde{\mathbf{A}} = \bar{\mathbf{A}} + \epsilon\mathbf{I}_n$ for a sufficiently small $\epsilon$ to ensure diagonalizability. Taking the limit as $\epsilon \to 0$ preserves the validity of the argument.

We redefine the linear system updates $\mathbf{x}_{t+1} = \bar{\mathbf{A}}\mathbf{x}_t + \bar{\mathbf{d}}_t$ as follows:

$$\mathbf{x}_{t+1} = \bar{\mathbf{A}}\mathbf{x}_t + \bar{\mathbf{d}}_T \Rightarrow \tilde{\mathbf{x}}_{t+1} = \bar{\mathbf{\Lambda}}\tilde{\mathbf{x}}_t + \tilde{\mathbf{d}}_t,$$

where $\tilde{\mathbf{x}}_t = \mathbf{P}^{-1}\mathbf{x}_t$ and $\tilde{\mathbf{d}}_t = \mathbf{P}^{-1}\bar{\mathbf{d}}_t$. These transformations can be thought of as coordinate transformations in $n$ dimensional space. Thus, we can establish the result for the LTI systems with a diagonal ground truth matrix $\bar{\mathbf{\Lambda}}$, whose diagonal entries are $\bar{\lambda}_1, \ldots, \bar{\lambda}_n$. In this case, the system update equations decompose into $n$ separate equations of the form: $\tilde{x}^i_{t+1} = \bar{\lambda}_i\tilde{x}^i_t + \tilde{d}^i_t, \forall i = 1, \ldots, n$ were $\tilde{x}^i_t \in \mathbb{R}$ is the i-th element of the system state at time $t$ in the new coordinate system.

Now, suppose that the non-zero disturbance vectors are chosen from the subspace $\tilde{D} := \{\tilde{\mathbf{d}} : \mathbf{P}\tilde{\mathbf{d}} \in D\}$ where $D = \{\bar{\mathbf{d}} : \tilde{\mathbf{d}} = \mathbf{P}^{-1}\bar{\mathbf{d}}, \tilde{\mathbf{d}}^1 = 0\}$. Due to the constraint on the first entry of the $\tilde{\mathbf{d}}$, $\tilde{\mathbf{d}}^1$, the subspace $D$ has dimension less than $n$. Consequently, since $\mathbf{P}^{-1}\mathbf{x}^1_0 = \tilde{\mathbf{x}}^1_0 = \mathbf{0}$, it follows that $\tilde{\mathbf{x}}^1_t = \mathbf{0}, \forall t \geq 0$. Note that setting the first entry to zero is done without loss of generality, as this argument applies to any coordinate or any subset of coordinates.

Next, define $\hat{\mathbf{\Lambda}}$ as the diagonal matrix with diagonal entries $(\hat{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_n)$ where $\hat{\lambda}_1 \neq \bar{\lambda}_1$, where all the diagonal entries are identical to those of $\bar{\mathbf{\Lambda}}$ except for the first one. Suppose that we start two dynamical systems with ground truth matrices $\bar{\mathbf{A}} = \mathbf{P}\bar{\mathbf{\Lambda}}\mathbf{P}^{-1}$ and $\hat{\mathbf{A}} = \mathbf{P}\hat{\mathbf{\Lambda}}\mathbf{P}^{-1}$, where the non-zero disturbances $\{\bar{\mathbf{d}}_t\}_{t=0}^{T-1}$ are chosen from $D$. Since $\tilde{\mathbf{x}}^1_t = \mathbf{0}, \forall t \geq 0$ in both systems, their trajectories over time must be identical, i.e., $\{\tilde{\mathbf{x}}_t\}_{t=0}^T$ is the same for both systems.

Suppose, for contradiction, that there exists an estimator that uniquely recovers $\bar{\mathbf{\Lambda}}$ in finite time $T$, given by:

$$\bar{\mathbf{\Lambda}} \in \arg\min_{\mathbf{\Lambda} \in \text{Diag(n)}} g(\mathbf{\Lambda}; \{\tilde{\mathbf{x}}_t\}_{t=0}^T),$$

where the decision variable $\mathbf{\Lambda}$ is restricted to be a diagonal matrix. Since $\bar{\mathbf{\Lambda}}$ is the unique solution, we must have

$$g(\bar{\mathbf{\Lambda}}; \{\tilde{\mathbf{x}}_t\}_{t=0}^T) < g(\hat{\mathbf{\Lambda}}; \{\tilde{\mathbf{x}}_t\}_{t=0}^T)$$

However, since the objective function $g(\mathbf{\Lambda}; \{\tilde{\mathbf{x}}_t\}_{t=0}^T)$ depends solely on system states over time, and these states are identical for both systems, it follows that $g(\hat{\mathbf{\Lambda}}; \{\tilde{\mathbf{x}}_t\}_{t=0}^T) = g(\bar{\mathbf{\Lambda}}; \{\tilde{\mathbf{x}}_t\}_{t=0}^T)$. This contradicts the uniqueness assumption of the solution, which concludes the proof. □

## 5.G Proof of Theorem 21

*Proof of Theorem 21.* Since both sides of inequality (5.9) are affine in $Z$, it suffices to prove that

$$\mathbb{P}\left[\hat{d}_1(\mathbf{Z}) - \hat{d}_2(\mathbf{Z}) < 0, \ \forall \mathbf{Z} \in \mathbb{S}_F\right] \geq 1 - \delta, \tag{5.29}$$

where $\mathbb{S}_F$ is the Frobenius-norm unit sphere in $\mathbb{R}^{n\times m}$ and

$$\hat{d}_1(\mathbf{Z}) := \sum_{t\in\mathcal{K}} \langle \mathbf{Z}^\top, f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top \rangle, \quad \hat{d}_2(\mathbf{Z}) := \sum_{t\in\mathcal{K}^c} \|\mathbf{Z} f(\mathbf{x}_t)\|_2.$$

The proof is divided into two steps.

**Step 1.** First, we fix the vector $\mathbf{Z} \in \mathbb{S}_F$ and prove that

$$\mathbb{P}\left[\hat{d}_1(\mathbf{Z}) - \hat{d}_2(\mathbf{Z}) < -\theta\right] \geq 1 - \delta,$$

holds for some constant $\theta > 0$. Using Markov's inequality, it is sufficient to prove that for some $\nu > 0$, it holds that

$$\mathbb{E}\left[\exp\left(\nu\left[\hat{d}_1(\mathbf{Z}) - \hat{d}_2(\mathbf{Z})\right]\right)\right] \leq \exp(-\nu\theta)\delta. \tag{5.30}$$

We focus on the case when $\mathcal{K}$ is not empty, which happens with high probability. The proof of this step is also divided into two sub-steps.

**Step 1-1.** We first analyze the term $\hat{d}_1(\mathbf{Z})$. Let $T'$ be the last non-zero disturbance time instance, i.e.,

$$T' := \max\{t \mid t \in \mathcal{K}\}.$$

Then, we have $\mathbb{E}\left[\exp\left[\nu\hat{d}_1(\mathbf{Z})\right]\right]$ equaling to the term

$$\mathbb{E}\left[\exp\left(\nu \sum_{t\in\mathcal{K}\setminus\{T'\}} \left\langle \mathbf{Z}^\top, f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top \right\rangle\right) \times \mathbb{E}\left[\exp\left[\nu\left\langle \mathbf{Z}^\top, f(\mathbf{x}_{T'})\hat{\mathbf{d}}_{T'}^\top \right\rangle\right] \mid \mathcal{F}_{T'}\right]\right]. \tag{5.31}$$

According to Assumption 6, the direction $\hat{\mathbf{d}}_{T'}$ is a unit vector. Since

$$\left|\left[\mathbf{Z}f(\mathbf{x}_{T'})\right]^\top \hat{\mathbf{d}}_{T'}\right| \leq \|\mathbf{Z}f(\mathbf{x}_{T'})\|_2 \leq \|\mathbf{Z}\|\|f(\mathbf{x}_{T'})\|_2 \leq \|\mathbf{Z}\|_F\sqrt{m}\|f(\mathbf{x}_{T'})\|_\infty \leq \sqrt{m}B,$$

the random variable $\left[\mathbf{Z}f(\mathbf{x}_{T'})\right]^\top \hat{\mathbf{d}}_{T'}$ is sub-Gaussian with parameter $mB^2$. Therefore, the property of sub-Gaussian random variables implies that

$$\mathbb{E}\left[\exp\left[\nu\left\langle \mathbf{Z}^\top, f(\mathbf{x}_{T'})\hat{\mathbf{d}}_{T'}^\top \right\rangle\right] \mid \mathcal{F}_{T'}\right] \leq \exp\left(\frac{\nu^2 \cdot mB^2}{2}\right).$$

Substituting into (5.31), we get

$$\mathbb{E}\left[\exp\left[\nu\hat{d}_1(\mathbf{Z})\right]\right] \leq \mathbb{E}\left[\exp\left(\nu \sum_{t\in\mathcal{K}\setminus\{T'\}} \left\langle \mathbf{Z}^\top, f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top \right\rangle\right)\right] \cdot \exp\left(\frac{\nu^2 \cdot mB^2}{2}\right).$$

Continuing this process for all $t \in \mathcal{K}$, it follows that

$$\mathbb{E}\left[\exp\left[\nu\hat{d}_1(\mathbf{Z})\right]\right] \leq \exp\left(\frac{\nu^2 \cdot mB^2|\mathcal{K}|}{2}\right). \tag{5.32}$$

**Step 1-2.** Now, we consider the second term in (5.30), namely, $-\hat{d}_2(\mathbf{Z})$. Define

$$\mathcal{K}' := \{t \mid 1 \le t \le T,\ t \in \mathcal{K}^c,\ t-1 \in \mathcal{K}\}.$$

With probability at least $1 - \exp[-\Theta[p(1-p)T]]$, we have

$$|\mathcal{K}'| = \Theta[p(1-p)T].$$

Therefore, $\mathcal{K}'$ is non-empty with high probability. Since $\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \ge 0$ for all $t \in \mathcal{K}^c$, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left[-\nu\hat{d}_2(\mathbf{Z})\right]\right] \le \mathbb{E}\left[\exp\left(-\nu\sum_{t\in\mathcal{K}'}\|\mathbf{Z}f(\mathbf{x}_t)\|_2\right)\right] \qquad (5.33)\\ =\mathbb{E}\left[\exp\left(-\nu\sum_{t\in\mathcal{K}'\setminus\{T'\}}\|\mathbf{Z}f(\mathbf{x}_t)\|_2\right)\times\mathbb{E}\left[\exp\left(-\nu\|\mathbf{Z}f(\mathbf{x}_{T'})\|_2\right)\mid\mathcal{F}_{T'}\right]\right],\end{aligned}$$

where $T'$ is the last time instance in $\mathcal{K}'$, namely,

$$T' := \max\{t \mid t \in \mathcal{K}'\}.$$

By Bernstein's inequality [111], we can estimate that

$$\begin{aligned}&\mathbb{E}\left[\exp\left(-\nu\|\mathbf{Z}f(\mathbf{x}_{T'})\|_2\right)\mid\mathcal{F}_{T'}\right]\\ &\quad\le\exp\left[-\nu\mathbb{E}\left(\|\mathbf{Z}f(\mathbf{x}_{T'})\|_2\mid\mathcal{F}_{T'}\right)+\frac{\nu^2}{2}\mathbb{E}\left(\|\mathbf{Z}f(\mathbf{x}_{T'})\|_2^2\mid\mathcal{F}_{T'}\right)\right]\\ &\quad\le\exp\left[-\frac{\nu}{\sqrt{m}B}\mathbb{E}\left(\|\mathbf{Z}f(\mathbf{x}_{T'})\|_2^2\mid\mathcal{F}_{T'}\right)+\frac{\nu^2}{2}\mathbb{E}\left(\|\mathbf{Z}f(\mathbf{x}_{T'})\|_2^2\mid\mathcal{F}_{T'}\right)\right],\end{aligned}$$

where the last inequality is from

$$\|\mathbf{Z}f(\mathbf{x}_{T'})\|_2 \le \sqrt{m}B.$$

Assumption 8 implies that

$$\mathbb{E}\left(\|\mathbf{Z}f(\mathbf{x}_{T'})\|_2^2\mid\mathcal{F}_{T'}\right) = \left\langle\mathbf{Z}\mathbf{Z}^\top,\mathbb{E}\left[f(\mathbf{x}_{T'})f(\mathbf{x}_{T'})^\top\mid\mathcal{F}_{T'}\right]\right\rangle \ge \lambda^2\|\mathbf{Z}\|_F^2 = \lambda^2.$$

If we choose $\nu$ such that

$$0 < \nu < \frac{2}{\sqrt{m}B}, \qquad (5.34)$$

we have

$$\mathbb{E}\left[\exp\left(-\nu\|\mathbf{Z}f(\mathbf{x}_{T'})\|_2\right)\mid\mathcal{F}_{T'}\right] \le \exp\left[\left(\frac{\nu^2}{2}-\frac{\nu}{\sqrt{m}B}\right)\lambda^2\right].$$

Substituting into inequality (5.33), it follows that

$$\mathbb{E}\left[\exp\left[-\nu\hat{d}_2(\mathbf{Z})\right]\right] \leq \mathbb{E}\left[\exp\left(-\nu\sum_{t\in\mathcal{K}'\setminus\{T'\}}\|\mathbf{Z}f(\mathbf{x}_t)\|_2\right)\times\exp\left[\left(\frac{\nu^2}{2}-\frac{\nu}{\sqrt{m}B}\right)\lambda^2\right]\right].$$

Continuing this process for all $t \in \mathcal{K}'$, we have

$$\mathbb{E}\left[\exp\left[-\nu\hat{d}_2(\mathbf{Z})\right]\right] \leq \exp\left[\left(\frac{\nu^2}{2}-\frac{\nu}{\sqrt{m}B}\right)\lambda^2|\mathcal{K}'|\right]. \tag{5.35}$$

Combining the inequalities (5.32) and (5.35), we have

$$\mathbb{E}\left[\exp\left(\nu\left[\hat{d}_1(\mathbf{Z})-\hat{d}_2(\mathbf{Z})\right]\right)\right] \leq \exp\left[\frac{m\nu^2B^2}{2}|\mathcal{K}|+\left(\frac{\nu^2}{2}-\frac{\nu}{\sqrt{m}B}\right)\lambda^2|\mathcal{K}'|\right].$$

We choose

$$\theta := \frac{\lambda^2 p(1-p)T}{4\sqrt{m}B}.$$

In order to satisfy condition (5.30), it is equivalent to have

$$\frac{m\nu^2B^2}{2}|\mathcal{K}|+\left(\frac{\nu^2}{2}-\frac{\nu}{\sqrt{m}B}\right)\lambda^2|\mathcal{K}'|+\frac{\lambda^2\nu p(1-p)T}{4\sqrt{m}B} \leq \log(\delta). \tag{5.36}$$

Now, we consider the fact that $\mathcal{K}$ is generated by the probabilistic sparsity model. Using the Bernoulli bound, it holds with probability at least $1-\exp[-\Theta[p(1-p)T]]$ that

$$|\mathcal{K}| \leq 2pT, \quad |\mathcal{K}'| \geq \frac{p(1-p)T}{2}. \tag{5.37}$$

Thus, with the same probability, we have the estimation

$$\begin{aligned}\frac{m\nu^2B^2}{2}|\mathcal{K}|+\left(\frac{\nu^2}{2}-\frac{\nu}{\sqrt{m}B}\right)\lambda^2|\mathcal{K}'|+\frac{\lambda^2\nu p(1-p)T}{4\sqrt{m}B}\\ \leq \frac{m\nu^2B^2}{2}\cdot 2pT+\left(\frac{\nu^2}{2}-\frac{\nu}{2\sqrt{m}B}\right)\lambda^2\cdot\frac{p(1-p)T}{2}.\end{aligned}$$

Choosing

$$\nu := \frac{\lambda^2(1-p)}{2\sqrt{m}B[4mB^2+\lambda^2(1-p)]},$$

we get

$$\frac{m\nu^2B^2}{2}|\mathcal{K}|+\left(\frac{\nu^2}{2}-\frac{\nu}{\sqrt{m}B}\right)\lambda^2|\mathcal{K}'|+\frac{\lambda^2\nu p(1-p)T}{4\sqrt{m}B} \leq -\frac{p(1-p)^2}{16m\kappa^2(4m\kappa^2+1-p)}\cdot T,$$

where we define $\kappa := B/\lambda \geq 1$. Note that our choice of $\nu$ satisfies the condition (5.34). Therefore, in order for inequality (5.36) to hold, the sample complexity should satisfy

$$T \geq \frac{16m\kappa^2(4m\kappa^2 + 1 - p)}{p(1-p)^2} \log\left(\frac{1}{\delta}\right).$$

By considering the Bernoulli bound (5.37), the sample complexity bound becomes

$$\begin{aligned} T &\geq \Theta\left[\max\left\{\frac{m\kappa^2(m\kappa^2 + 1 - p)}{p(1-p)^2}, \frac{1}{p(1-p)}\right\} \log\left(\frac{1}{\delta}\right)\right] \\ &= \Theta\left[\frac{m^2\kappa^4}{p(1-p)^2} \log\left(\frac{1}{\delta}\right)\right]. \end{aligned} \tag{5.38}$$

**Step 2.** Next, we establish the bound (5.29) by discretization techniques. More specifically, suppose that $\epsilon > 0$ is a constant and $\{\mathbf{Z}^1, \ldots, \mathbf{Z}^N\} \subset \mathbb{S}_F$ is an $\epsilon$-net of the sphere $\mathbb{S}_F$ under the Frobenius norm, where we can bound

$$\log(N) \leq mn \cdot \log\left(1 + \frac{2}{\epsilon}\right).$$

Then, for every $\mathbf{Z} \in \mathbb{S}_F$, we can find a point in the $\epsilon$-net, denoted as $\mathbf{Z}'$, such that

$$\|\mathbf{Z} - \mathbf{Z}'\|_F \leq \epsilon.$$

Now, we upper bound the difference $f(\mathbf{Z}) - f(\mathbf{Z}')$, where we define the function

$$f(\mathbf{Z}) := \hat{d}_1(\mathbf{Z}) - \hat{d}_2(\mathbf{Z}), \quad \forall \mathbf{Z} \in \mathbb{R}^{n \times m}.$$

We can calculate that

$$\begin{aligned} f(\mathbf{Z}) - f(\mathbf{Z}') &= \sum_{t \in \mathcal{K}} \hat{\mathbf{d}}_t(\mathbf{Z} - \mathbf{Z}')f(\mathbf{x}_t) - \sum_{t \in \mathcal{K}^c} (\|\mathbf{Z}f(\mathbf{x}_t)\|_2 - \|\mathbf{Z}'f(\mathbf{x}_t)\|_2) \\ &\leq \sum_{t \in \mathcal{K}} \hat{\mathbf{d}}_t(\mathbf{Z} - \mathbf{Z}')f(\mathbf{x}_t) + \sum_{t \in \mathcal{K}^c} \|(\mathbf{Z} - \mathbf{Z}')f(\mathbf{x}_t)\|_2 \\ &\leq \sum_{t \in \mathcal{K}} \|\mathbf{Z} - \mathbf{Z}'\|_F \|f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top\|_F + \sum_{t \in \mathcal{K}^c} \|\mathbf{Z} - \mathbf{Z}'\|_2 \|f(\mathbf{x}_t)\|_2 \\ &\leq \sum_{t \in \mathcal{K}} \|\mathbf{Z} - \mathbf{Z}'\|_F \|f(\mathbf{x}_t)\|_2 + \sum_{t \in \mathcal{K}^c} \|\mathbf{Z} - \mathbf{Z}'\|_F \|f(\mathbf{x}_t)\|_2 \\ &\leq T \cdot \epsilon\sqrt{m}B = \sqrt{m}TB \cdot \epsilon. \end{aligned}$$

We choose

$$\epsilon := \frac{\theta}{\sqrt{m}TB} = \Theta\left[\frac{p(1-p)}{m\kappa^2}\right].$$

Therefore, under the event that

$$f(\mathbf{Z}^i) < -\theta, \quad \forall i = 1, \dots, N, \tag{5.39}$$

we have

$$f(\mathbf{Z}) < -\theta + \sqrt{m}TB \cdot \epsilon = 0, \quad \forall \mathbf{Z} \in \mathbb{S}_F.$$

Hence, it suffices to estimate the probability that event (5.39) happens. To bound the failing probability, we replace $\delta$ with $\delta/N$ in (5.38) and it follows that

$$\mathbb{P}\left[f(\mathbf{Z}^i) < -\theta\right] \geq 1 - \frac{\delta}{N}, \quad \forall i = 1, \dots, N.$$

Applying the union bound over all $i \in \{1, \dots, N\}$, the event (5.39) happens with probability at least $1 - \delta$, namely,

$$\mathbb{P}\left[f(\mathbf{Z}^i) < -\theta, \ \forall i = 1, \dots, N\right] \geq 1 - \delta.$$

With this choice of $\delta$, the sample complexity should be at least

$$T \geq \Theta\left[\frac{m^2\kappa^4}{p(1-p)^2} \log\left(\frac{N}{\delta}\right)\right] = \Theta\left[\frac{m^2\kappa^4}{p(1-p)^2}\left[mn \log\left(\frac{m\kappa}{p(1-p)}\right) + \log\left(\frac{1}{\delta}\right)\right]\right].$$

This completes the proof. □

## 5.H Proof of Theorem 22

*Proof of Theorem 22.* We only need to show that condition (5.8) fails with probability at least $1 - \exp(-m/3)$. We choose the matrix

$$\bar{\mathbf{A}} := \begin{bmatrix} 1 & 0_{1\times(m-1)} \\ 0_{n-1} & 0_{(n-1)\times(m-1)} \end{bmatrix} \in \mathbb{R}^{n\times m}.$$

As a result, the last $n-1$ elements of $\bar{\mathbf{A}}f(\mathbf{x})$ are zero for every state $\mathbf{x} \in \mathbb{R}^n$. Moreover, we will choose the basis function $f$ such that its values will only depend on the first element of state $\mathbf{x} \in \mathbb{R}^n$. With these definitions, the dynamics of $\mathbf{x}_t$ reduces to the dynamics of its first element $(\mathbf{x}_t)_1$. Hence, we can assume without loss of generality that $n = 1$ in the remainder of the proof.

We define the basis function $f : \mathbb{R} \mapsto \mathbb{R}^m$ as

$$\tilde{f}(x) := \left[\frac{x}{\max\{|x|,1\}} \quad \sin(x) \quad \sin(2x) \quad \cdots \quad \sin[(m-1)x]\right], \quad \forall x \in \mathbb{R}.$$

Under the above definitions, it is straightforward to show that the following properties hold and we omit the proof:

$$f(\mathbf{0}) = \mathbf{0}_m, \quad f\left[\bar{\mathbf{A}}f(\mathbf{x})\right] = f(\mathbf{x}), \quad \forall x \in \mathbb{R}. \tag{5.40}$$

Finally, the attack/disturbance vector is defined as

$$\bar{\mathbf{d}}_t | \mathcal{F}_t \sim \text{Uniform}\left\{[-(|\mathbf{x}_t| + 2\pi), -(|\mathbf{x}_t| + \pi)] \cup [|\mathbf{x}_t| + \pi, |\mathbf{x}_t| + 2\pi]\right\}, \quad \forall t \in \mathcal{K}.$$

The remainder of the proof is divided into three steps.

**Step 1.** In the first step, we prove that Assumptions 6-8 hold. By the definition of $f(\mathbf{x})$, we have

$$\|f(\mathbf{x})\|_\infty = \max\left\{\frac{|x|}{\max\{|x|,1\}}, |\sin(x)|, \dots, |\sin[(m-1)x]|\right\} \le 1, \quad \forall x \in \mathbb{R},$$

which implies that Assumption 7 holds with $B = 1$. Moreover, the semi-oblivious condition (Assumption 6) is a result of the symmetric distribution of $\bar{\mathbf{d}}_t|\mathcal{F}_t$.

Finally, we prove that Assumption 8 holds. For the notational simplicity, in this step, we omit the subscript $t$, the conditioning on the filtration $\mathcal{F}_t$, and the event $t \in \mathcal{K}$. The model of disturbance vector $d$ implies that

$$|x+d| \ge |d| - |x| \ge \pi > 1.$$

Therefore, we have

$$f(x+d) = \begin{bmatrix} \frac{x+d}{|x+d|} & \sin[(x+d)] & \cdots & \sin[(m-1)(x+d)] \end{bmatrix}.$$

For any vector $\nu \in \mathbb{R}^m$, we want to estimate

$$\nu^\top \mathbb{E}\left[f(x+d)f(x+d)^\top\right]\nu = \mathbb{E}\left[\nu_1 \frac{x+d}{|x+d|} + \sum_{i=1}^{m-1} \nu_{i+1}\sin[i(x+d)]\right]^2.$$

First, we can calculate that

$$\mathbb{E}\left(\nu_1 \frac{x+d}{|x+d|}\right)^2 = \nu_1^2,\ \mathbb{E}\left[\nu_{i+1}\sin[i(x+d)]\right]^2 = \nu_{i+1}^2 \cdot \frac{1}{2}, \quad \forall i \in \{1,\dots,m-1\}. \tag{5.41}$$

Then, for every $i \in \{1,\dots,m-1\}$, we have

$$\begin{aligned} &\mathbb{E}\left[\nu_1 \frac{x+d}{|x+d|} \cdot \nu_{i+1}\sin[i(x+d)]\right] \\ =&\nu_1\nu_{i+1}\left[\int_{-|x|-2\pi}^{-|x|-\pi} \frac{x+d}{|x+d|}\sin[i(x+d)]\,\mathrm{d}d + \int_{|x|+\pi}^{|x|+2\pi} \frac{x+d}{|x+d|}\sin[i(x+d)]\,\mathrm{d}d\right] \\ =&\nu_1\nu_{i+1}\left[\int_{-|x|-2\pi}^{-|x|-\pi} -\sin[i(x+d)]\,\mathrm{d}d + \int_{|x|+\pi}^{|x|+2\pi} \sin[i(x+d)]\,\mathrm{d}d\right] = 0. \end{aligned} \tag{5.42}$$

For every $i, j \in \{1,\dots,m-1\}$ such that $i \ne j$, it holds that

$$\begin{aligned} \mathbb{E}\left[\nu_{i+1}\sin[i(x+d)]\cdot\nu_{j+1}\sin[j(x+d)]\right] =&\nu_{i+1}\nu_{j+1}\left[\int_{-|x|-2\pi}^{-|x|-\pi} \sin[i(x+d)]\sin[j(x+d)]\,\mathrm{d}d\right. \\ &\left. + \int_{|x|+\pi}^{|x|+2\pi} \sin[i(x+d)]\sin[j(x+d)]\,\mathrm{d}d\right] = 0. \end{aligned} \tag{5.43}$$

Combining equations (5.41)-(5.43), it follows that

$$\nu^\top \mathbb{E}\left[f(x+d)f(x+d)^\top\right]\nu = \nu_1^2 + \frac{1}{2}\sum_{i=1}^{m-1}\nu_{i+1}^2 \geq \frac{1}{2}\|\nu\|_2^2,$$

which implies that Assumption 8 holds with $\lambda^2 = 1/2$.

**Step 2.** In this step, we prove that the linear space spanned by the set of vectors

$$\mathcal{F}^c := \{f(\mathbf{x}_t) \mid t \in \mathcal{K}^c\}$$

has dimension at most $m-1$ with probability at least $1-\delta$. By the second property in (5.40), the subspace spanned by $\mathcal{F}^c$ is equivalent to that spanned by

$$\mathcal{F}' := \{f(\mathbf{x}_t) \mid t \in \mathcal{K}'\},$$

where we define

$$\mathcal{K}' := \{t \mid t-1 \in \mathcal{K},\ t \in \mathcal{K}^c\}.$$

Therefore, the dimension of the subspace is at most $|\mathcal{K}'|$.

To estimate the cardinality of $\mathcal{K}'$, we divide $\mathcal{K}'$ into the following two disjoint sets:

$$\mathcal{K}_1' := \{2t+1 \mid 2t \in \mathcal{K},\ 2t+1 \in \mathcal{K}^c\}, \quad \mathcal{K}_2' := \{2t \mid 2t-1 \in \mathcal{K},\ 2t \in \mathcal{K}^c\}.$$

The size of $\mathcal{K}_1'$ is the summation of $\lceil T/2 \rceil$ independent Bernoulli random variables with parameter $p(1-p)$. Therefore, the Chernoff bound implies

$$\mathbb{P}\left[|\mathcal{K}_1'| \leq 2p(1-p)\cdot\left\lceil\frac{T}{2}\right\rceil\right] \geq 1-\exp\left[-\frac{p(1-p)}{3}\cdot\left\lceil\frac{T}{2}\right\rceil\right]. \tag{5.44}$$

Similarly, the size of $\mathcal{K}_2'$ is the summation of $\lfloor T/2 \rfloor$ independent Bernoulli random variables with parameter $p(1-p)$. Therefore, the Chernoff bound implies

$$\mathbb{P}\left[|\mathcal{K}_2'| \leq 2p(1-p)\cdot\left\lfloor\frac{T}{2}\right\rfloor\right] \geq 1-\exp\left[-\frac{p(1-p)}{3}\cdot\left\lfloor\frac{T}{2}\right\rfloor\right]. \tag{5.45}$$

Combining the bounds (5.44) and (5.45) and applying the union bound, it holds that

$$\begin{aligned}\mathbb{P}\left[|\mathcal{K}'| \leq 2p(1-p)T\right] &\geq 1-\exp\left[-\frac{p(1-p)}{3}\cdot\left\lceil\frac{T}{2}\right\rceil\right]-\exp\left[-\frac{p(1-p)}{3}\cdot\left\lfloor\frac{T}{2}\right\rfloor\right]\\ &\geq 1-2\exp\left[-\frac{p(1-p)T}{3}\right],\end{aligned}$$

where the last inequality is because $\lfloor T/2 \rfloor \leq \lceil T/2 \rceil \leq T$. Since

$$T < \frac{m}{2p(1-p)},$$

we know

$$\mathbb{P}\left[|\mathcal{K}'| < m\right] \geq 1 - 2\exp\left(-m/3\right). \tag{5.46}$$

In addition, when $\mathcal{K}$ is the empty set $\emptyset$ or the full set $\{0, \ldots, T-1\}$, the set $\mathcal{K}'$ is an empty set, which implies that $|\mathcal{K}'|$ is smaller than $m$. This event happens with probability

$$p^{\top} + (1-p)^{\top} \geq 2[p(1-p)]^{T/2}.$$

Combining with inequality (5.46), we get

$$\mathbb{P}\left[|\mathcal{K}'| < m\right] \geq \max\left\{1 - 2\exp\left(-m/3\right), 2[p(1-p)]^{T/2}\right\}.$$

**Step 3.** Finally, we prove that if the dimension of the subspace spanned by $\mathcal{F}^c$ is smaller than $m$, the condition (5.8) cannot hold. Since the dimension of the subspace is at most $m-1$, there exists $\mathbf{Z} \in \mathbb{R}^m$ such that

$$\mathbf{Z} f(\mathbf{x}_t) = 0, \quad \forall t \in \mathcal{K}^c.$$

With this choice of $Z$, the condition on the left-hand-side of (5.8) holds while the strict inequality on the right-hand-side fails. Therefore, we know that $\bar{\mathbf{A}}$ is not the unique global solution to (5.4). □

# 5.I Proof of Theorem 23

*Proof of Theorem 23.* The proof is similar to that of Theorem 21. Since both sides of inequality (5.9) are affine in $Z$, it suffices to prove that

$$\mathbb{P}\left[\hat{d}_1(\mathbf{Z}) - \hat{d}_2(\mathbf{Z}) < 0, \ \forall \mathbf{Z} \in \mathbb{S}_F\right] \geq 1 - \delta,$$

where $\mathbb{S}_F$ is the Frobenius-norm unit sphere in $\mathbb{R}^{n \times m}$ and

$$\hat{d}_1(\mathbf{Z}) := \sum_{t \in \mathcal{K}} \left\langle \mathbf{Z}^{\top}, f(\mathbf{x}_t)\hat{\mathbf{d}}_t^{\top} \right\rangle, \quad \hat{d}_2(\mathbf{Z}) := \sum_{t \in \mathcal{K}^c} \left\| \mathbf{Z} f(\mathbf{x}_t) \right\|_2.$$

The proof is divided into two steps.

**Step 1.** First, we fix the vector $\mathbf{Z} \in \mathbb{S}_F$ and prove that

$$\mathbb{P}\left[\hat{d}_1(\mathbf{Z}) - \hat{d}_2(\mathbf{Z}) < -\theta\right] \geq 1 - \delta,$$

holds for some constant $\theta > 0$. The proof of this step is divided into two steps.

**Step 1-1.** We first analyze the term $\hat{d}_1(\mathbf{Z})$. For each $k \in \mathcal{K}$, we define the following disturbance vectors:

$$\bar{\mathbf{d}}_t^k := \begin{cases} \bar{\mathbf{d}}_t & \text{if } t \leq k, \\ \mathbf{0}_n & \text{otherwise,} \end{cases} \quad \forall t \in \{0, \ldots, T-1\}.$$

Then, we define the trajectory generated by the above disturbance vectors:

$$\mathbf{x}_0^k = \mathbf{0}_m, \quad \mathbf{x}_{t+1}^k = \bar{\mathbf{A}} f(\mathbf{x}_t^k) + \bar{\mathbf{d}}_t^k, \quad \forall t \in \{0, \ldots, T-1\}.$$

Let

$$\mathcal{K} = \{k_1, \ldots, k_{|\mathcal{K}|}\},$$

where the elements are sorted as $k_1 < k_2 < \cdots < k_{|\mathcal{K}|}$. Under the above definition, we know $\mathbf{x}_t^{k_{|\mathcal{K}|}} = \mathbf{x}_t$ for all $t$. We define

$$\mathbf{g}_t^{k_j} := \begin{cases} f(\mathbf{x}_t^{k_j}) - f(\mathbf{x}_t^{k_{j-1}}) & \text{if } j > 1, \\ f(\mathbf{x}_t^{k_1}) & \text{if } j = 1, \end{cases} \quad \forall j \in \{1, \ldots, |\mathcal{K}|\}.$$

We note that $\mathbf{g}_t^{k_j}$ is measurable on $\mathcal{F}_{k_j}$. Using these introduced notations, we can write $\hat{d}_1(\mathbf{Z})$ as

$$\hat{d}_1(\mathbf{Z}) = \sum_{j=1}^{|\mathcal{K}|} \left\langle \mathbf{Z}^\top, f(\mathbf{x}_{k_j}) \hat{\mathbf{d}}_{k_j}^\top \right\rangle = \sum_{j=1}^{|\mathcal{K}|} \left\langle \mathbf{Z}^\top, \sum_{\ell=1}^{j-1} \mathbf{g}_{k_j}^{k_\ell} \hat{\mathbf{d}}_{k_j}^\top \right\rangle = \sum_{\ell=1}^{|\mathcal{K}|} \sum_{j=\ell+1}^{|\mathcal{K}|} \hat{\mathbf{d}}_{k_j}^\top \mathbf{Z} \mathbf{g}_{k_j}^{k_\ell}.$$

Then, Assumption 11 implies that $\bar{\mathbf{d}}_t$ is sub-Gaussian with parameter $\sigma$ conditional on $\mathcal{F}_t$. Now, we estimate the expectation

$$\mathbb{E}\left[\exp\left[\nu \hat{d}_1(\mathbf{Z})\right]\right],$$

where $\nu \in \mathbb{R}$ is an arbitrary constant. First, for each $\ell \in \{1, \ldots, |\mathcal{K}|-1\}$, we estimate the following probability:

$$\mathbb{P}\left( \left| \sum_{j=\ell+1}^{|\mathcal{K}|} \hat{\mathbf{d}}_{k_j}^\top \mathbf{Z} \mathbf{g}_{k_j}^{k_\ell} \right| \geq \epsilon \;\middle|\; \mathcal{F}_{k_\ell} \right).$$

Since $\hat{\mathbf{d}}_{k_j}$ is a unit vector and $\|\mathbf{Z}\|_F = 1$, we know

$$\left\| \hat{\mathbf{d}}_{k_j}^\top \mathbf{Z} \right\|_2 \leq \|\hat{\mathbf{d}}_{k_j}^\top\|_2 \|\mathbf{Z}\| \leq \|\hat{\mathbf{d}}_{k_j}^\top\|_2 \|\mathbf{Z}\|_F = 1. \tag{5.47}$$

Moreover, we can estimate that

$$\begin{aligned}\left\|g_{k_j}^{k_\ell}\right\|_2 &= \left\|f(\mathbf{x}_{k_j}^{k_\ell}) - f(\mathbf{x}_{k_j}^{k_{\ell-1}})\right\|_2 \leq L\left\|\mathbf{x}_{k_j}^{k_\ell} - \mathbf{x}_{k_j}^{k_{\ell-1}}\right\|_2 \\ &= L\left\|\bar{\mathbf{A}}\left[f\left(\mathbf{x}_{k_j-1}^{k_\ell}\right) - f\left(\mathbf{x}_{k_j-1}^{k_{\ell-1}}\right)\right]\right\|_2 \leq \rho L\left\|f\left(\mathbf{x}_{k_j-1}^{k_\ell}\right) - f\left(\mathbf{x}_{k_j-1}^{k_{\ell-1}}\right)\right\|_2 \\ &\leq L(\rho L)\left\|\mathbf{x}_{k_j-1}^{k_\ell} - \mathbf{x}_{k_j-1}^{k_{\ell-1}}\right\|_2 \leq \cdots \leq L(\rho L)^{k_j-k_\ell-1}\left\|\mathbf{x}_{k_\ell+1}^{k_\ell} - \mathbf{x}_{k_\ell+1}^{k_{\ell-1}}\right\|_2 \\ &= L(\rho L)^{k_j-k_\ell-1}\|\bar{\mathbf{d}}_{k_\ell}\|_2,\end{aligned} \tag{5.48}$$

where the first inequality holds because $f$ has Lipschitz constant $L$, the second inequality is from $\|\bar{\mathbf{A}}\| \leq \rho$ and the last equality holds because

$$\mathbf{x}_{k_\ell+1}^{k_\ell} = \bar{\mathbf{A}}f\left(\mathbf{x}_{k_\ell}^{k_\ell}\right) + \bar{\mathbf{d}}_{k_\ell}, \quad \mathbf{x}_{k_\ell+1}^{k_{\ell-1}} = \bar{\mathbf{A}}f\left(\mathbf{x}_{k_\ell}^{k_{\ell-1}}\right) = \bar{\mathbf{A}}f\left(\mathbf{x}_{k_\ell}^{k_\ell}\right).$$

By the sub-Gaussian assumption (Assumption 11), it holds that

$$\mathbb{P}\left(\|\bar{\mathbf{d}}_{k_\ell}\|_2 \geq \eta \,\middle|\, \mathcal{F}_{k_\ell}\right) \leq 2\exp\left(-\frac{\eta^2}{2\sigma^2}\right), \quad \forall \eta \geq 0. \tag{5.49}$$

Combining inequalities (5.47)-(5.49), we get

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{j=\ell+1}^{|\mathcal{K}|}\hat{\mathbf{d}}_{k_j}^\top\mathbf{Z}^\top\mathbf{g}_{k_j}^{k_\ell}\right| \geq \epsilon \,\middle|\, \mathcal{F}_{k_\ell}\right) &\leq \mathbb{P}\left(\sum_{j=\ell+1}^{|\mathcal{K}|}\left\|g_{k_j}^{k_\ell}\right\|_2 \geq \epsilon \,\middle|\, \mathcal{F}_{k_\ell}\right) \\ &\leq \mathbb{P}\left(\sum_{j=\ell+1}^{|\mathcal{K}|}L(\rho L)^{k_j-k_\ell-1}\|\bar{\mathbf{d}}_{k_\ell}\|_2 \geq \epsilon \,\middle|\, \mathcal{F}_{k_\ell}\right) \\ &\leq \mathbb{P}\left(\frac{L(\rho L)^{\Delta_j}}{1-\rho L}\|\bar{\mathbf{d}}_{k_\ell}\|_2 \geq \epsilon \,\middle|\, \mathcal{F}_{k_\ell}\right) \\ &\leq 2\exp\left[-\frac{(1-\rho L)^2\epsilon^2}{2\sigma^2L^2(\rho L)^{2\Delta_j}}\right],\end{aligned} \tag{5.50}$$

where $\Delta_j := k_j - k_{j-1} - 1$ and the second last inequality is from

$$\sum_{j=\ell+1}^{|\mathcal{K}|}L(\rho L)^{k_j-k_\ell-1} < \sum_{i=\Delta_j}^{\infty}L(\rho L)^i = \frac{L(\rho L)^{\Delta_j}}{1-\rho L}.$$

Since

$$\mathbb{E}\left(\sum_{j=\ell+1}^{|\mathcal{K}|}\hat{\mathbf{d}}_{k_j}^\top\mathbf{Z}\mathbf{g}_{k_j}^{k_\ell} \,\middle|\, \mathcal{F}_{k_\ell}\right) = 0,$$

inequality (5.50) implies the random variable $\sum_{j=\ell+1}^{|\mathcal{K}|} \hat{\mathbf{d}}_{k_j}^\top \mathbf{Z}^\top \mathbf{g}_{k_j}^{k_\ell}$ is zero-mean and sub-Gaussian with parameter $\sigma L/(1-\rho L)$ conditional on $\mathcal{F}_{k_\ell}$. By the property of sub-Gaussian random variables, we have

$$\mathbb{E}\left[\exp\left(\nu \sum_{j=\ell+1}^{|\mathcal{K}|} \hat{\mathbf{d}}_{k_j}^\top \mathbf{Z}\mathbf{g}_{k_j}^{k_\ell}\right) \,\middle|\, \mathcal{F}_{k_\ell}\right] \le \exp\left[\frac{\nu^2\sigma^2 L^2(\rho L)^{2\Delta_j}}{2(1-\rho L)^2}\right], \quad \forall \nu \ge 0.$$

Finally, utilizing the tower property of conditional expectation, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left[\nu \hat{d}_1(\mathbf{Z})\right]\right] &= \mathbb{E}\left[\exp\left(\nu \sum_{\ell=1}^{|\mathcal{K}|-2}\sum_{j=\ell+1}^{|\mathcal{K}|} \hat{\mathbf{d}}_{k_j}^\top \mathbf{Z}\mathbf{g}_{k_j}^{k_\ell}\right) \times \mathbb{E}\left[\exp\left(\nu \sum_{j=|\mathcal{K}|}^{|\mathcal{K}|} \hat{\mathbf{d}}_{k_j}^\top \mathbf{Z}\mathbf{g}_{k_j}^{k_\ell}\right) \,\middle|\, \mathcal{F}_{k_{|\mathcal{K}|-1}}\right]\right] \\
&\le \mathbb{E}\left[\exp\left(\nu \sum_{\ell=1}^{|\mathcal{K}|-2}\sum_{j=\ell+1}^{|\mathcal{K}|} \hat{\mathbf{d}}_{k_j}^\top \mathbf{Z}\mathbf{g}_{k_j}^{k_\ell}\right) \times \exp\left[\frac{\nu^2\sigma^2 L^2(\rho L)^{2\Delta_j}}{2(1-\rho L)^2}\right]\right] \\
&\le \cdots \le \exp\left[\frac{\nu^2\sigma^2 L^2}{2(1-\rho L)^2}\sum_{j\in\mathcal{K}}(\rho L)^{2\Delta_j}\right], \quad \forall \nu \ge 0.
\end{aligned} \tag{5.51}$$

Since the random variable $(\rho L)^{\Delta_j}$ is bounded in $[0,1]$ and thus, it is sub-Gaussian with parameter $1/2$. Therefore, with a constant number of samples, the mean of $(\rho L)^{2\Delta_j}$ will concentrate around its expectation, which is approximately

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

Then, the bound in (5.51) becomes

$$\mathbb{E}\left[\exp\left[\nu \hat{d}_1(\mathbf{Z})\right]\right] \lesssim \exp\left[\frac{\nu^2\sigma^2 L^2 p|\mathcal{K}|}{2(1-\rho L)^3}\right], \quad \forall \nu \ge 0. \tag{5.52}$$

Applying Chernoff's bound to (5.52), we get

$$\mathbb{P}\left[\hat{d}_1(\mathbf{Z}) \le \epsilon\right] \ge 1 - \exp\left[-\frac{(1-\rho L)^3}{2\sigma^2 L^2 p|\mathcal{K}|}\cdot \epsilon^2\right], \quad \forall \epsilon \ge 0. \tag{5.53}$$

**Step 1-2.** Next, we analyze the term $\hat{d}_2(\mathbf{Z})$. Define the set

$$\mathcal{K}' := \{t \mid 1 \le t \le T,\ t \in \mathcal{K}^c,\ t-1 \in \mathcal{K}\}.$$

With probability at least $1 - \exp[-\Theta[p(1-p)T]]$, we have

$$|\mathcal{K}'| = \Theta[p(1-p)T].$$

Therefore, $\mathcal{K}'$ is non-empty with high probability. Since $\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq 0$ for all $t \in \mathcal{K}^c$, we know

$$\hat{d}_2(\mathbf{Z}) \geq \sum_{k \in \mathcal{K}'} \|\mathbf{Z}f(\mathbf{x}_t)\|_2.$$

To establish a high-probability lower bound of $\|\mathbf{Z}f(\mathbf{x}_t)\|_2$, we prove the following lemma.

**Lemma 17.** *For each $t \in \mathcal{K}'$, it holds that*

$$\mathbb{P}\left[\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \frac{\lambda}{2} \,\middle|\, \mathcal{F}_t\right] \geq \frac{c\lambda^4}{\sigma^4 L^4},$$

*where $c := 1/1058$ is an absolute constant.*

For each $t \in \mathcal{K}'$, let $\mathbf{1}_t$ be the indicator of the event that $\|\mathbf{Z}f(\mathbf{x}_t)\|_2$ is larger than the $\frac{c\lambda^4}{\sigma^4 L^4}$-quantile conditional on $\mathcal{F}_t$. Then, it holds that

$$\mathbb{P}(\mathbf{1}_t = 1 \mid \mathcal{F}_t) = 1 - \mathbb{P}(\mathbf{1}_t = 0 \mid \mathcal{F}_t) = \frac{c\lambda^4}{\sigma^4 L^4}.$$

Therefore, we know

$$\left\{\mathbf{1}_t - \frac{c\lambda^4}{\sigma^4 L^4},\ t \in \mathcal{K}'\right\}$$

is a martingale with respect to filtration set $\{\mathcal{F}_t,\ t \in \mathcal{K}'\}$. Applying Azuma's inequality, it holds with probability at least $1 - \exp[-\Theta(\frac{\lambda^4|\mathcal{K}'|}{\sigma^4 L^4})]$ that

$$\sum_{t \in \mathcal{K}'} \mathbf{1}_t \geq \frac{c\lambda^4|\mathcal{K}'|}{2\sigma^4 L^4},$$

which means that for at least $\frac{c\lambda^4|\mathcal{K}'|}{2\sigma^4 L^4}$ elements in $\mathcal{K}'$, the event that $\|\mathbf{Z}f(\mathbf{x}_t)\|_2$ is larger than the $\frac{c\lambda^4}{\sigma^4 L^4}$-quantile conditional on $\mathcal{F}_t$ happens. Using the lower bound on the quantile in Lemma 17, we know

$$\sum_{t \in \mathcal{K}'} \|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \frac{c\lambda^4|\mathcal{K}'|}{2\sigma^4 L^4} \cdot \frac{\lambda}{2} + \left(|\mathcal{K}'| - \frac{c\lambda^4|\mathcal{K}'|}{2\sigma^4 L^4}\right) \cdot 0 = \frac{c\lambda^5|\mathcal{K}'|}{4\sigma^4 L^4} \tag{5.54}$$

holds with the same probability.

Combining inequalities (5.53) and (5.54), we get

$$\mathbb{P}\left[f(\mathbf{Z}) \leq \epsilon - \frac{c\lambda^5|\mathcal{K}'|}{4\sigma^4 L^4}\right] \geq 1 - \exp\left[-\frac{(1-\rho L)^3}{2\sigma^2 L^2 p|\mathcal{K}|} \cdot \epsilon^2\right] - \exp\left[-\Theta\left(\frac{\lambda^4|\mathcal{K}'|}{\sigma^4 L^4}\right)\right],$$

where we define $f(\mathbf{Z}) := \hat{d}_1(\mathbf{Z}) - \hat{d}_2(\mathbf{Z})$. Choosing

$$\epsilon := \frac{c\lambda^5|\mathcal{K}'|}{8\sigma^4 L^4},$$

it follows that

$$\mathbb{P}\left[f(\mathbf{Z}) \leq -\frac{c\lambda^5|\mathcal{K}'|}{8\sigma^4L^4}\right] \geq 1 - \exp\left[-\Theta\left(\frac{(1-\rho L)^3\lambda^{10}|\mathcal{K}'|^2}{\sigma^{10}L^{10}p|\mathcal{K}|}\right)\right] - \exp\left[-\Theta\left(\frac{\lambda^4|\mathcal{K}'|}{\sigma^4L^4}\right)\right]. \tag{5.55}$$

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

$$|\mathcal{K}| \leq 2pT, \quad |\mathcal{K}'| \geq \frac{p(1-p)T}{2}. \tag{5.56}$$

Therefore, the probability bound in (5.55) becomes

$$\begin{aligned}\mathbb{P}\left[f(\mathbf{Z}) \leq -\frac{c\lambda^5p(1-p)T}{16\sigma^4L^4}\right] \geq & 1 - \exp\left[-\Theta\left(\frac{(1-\rho L)^3\lambda^{10}(1-p)^2T}{\sigma^{10}L^{10}}\right)\right] \\ & - \exp\left[-\Theta\left(\frac{\lambda^4p(1-p)T}{\sigma^4L^4}\right)\right] - \exp[-\Theta[p(1-p)T]].\end{aligned}$$

Now, if the sample complexity satisfies

$$T \geq \Theta\left[\max\left\{\frac{\kappa^{10}}{(1-\rho L)^3(1-p)^2}, \frac{\kappa^4}{p(1-p)}\right\}\log\left(\frac{1}{\delta}\right)\right], \tag{5.57}$$

we know

$$\mathbb{P}\left[f(\mathbf{Z}) \leq -\theta\right] \geq 1 - \delta, \tag{5.58}$$

where we define

$$\kappa := \frac{\sigma L}{\lambda}, \quad \theta := \frac{c\lambda^5p(1-p)T}{16\sigma^4L^4}.$$

**Step 2.** In the second step, we apply discretization techniques to prove that condition (5.58) holds for all $\mathbf{Z} \in \mathbb{S}_F$. For a sufficiently small constant $\epsilon > 0$, let

$$\{\mathbf{Z}^1, \ldots, \mathbf{Z}^N\}$$

be an $\epsilon$-cover of the unit ball $\mathbb{S}_F$. Namely, for all $\mathbf{Z} \in \mathbb{S}_F$, we can find $r \in \{1, 2, \ldots, N\}$ such that $\|\mathbf{Z} - \mathbf{Z}^r\|_F \leq \epsilon$. It is proved in [111] that the number of points $N$ can be bounded by

$$\log(N) \leq mn\log\left(1 + \frac{2}{\epsilon}\right).$$

Now, we estimate the Lipschitz constant of $f(\mathbf{Z})$ and construct a high-probability upper bound for the Lipschitz constant. For all $\mathbf{Z}, \mathbf{Z}' \in \mathbb{R}^{n\times m}$, we can calculate that

$$\begin{aligned} f(\mathbf{Z}) - f(\mathbf{Z}') &= \sum_{t\in\mathcal{K}} \left\langle (\mathbf{Z}-\mathbf{Z}')^\top, f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top \right\rangle - \sum_{t\in\mathcal{K}^c} \left(\|\mathbf{Z}f(\mathbf{x}_t)\|_2 - \|\mathbf{Z}'f(\mathbf{x}_t)\|_2\right) \\ &\le \|\mathbf{Z}-\mathbf{Z}'\|_F \sum_{t\in\mathcal{K}} \left\| f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top \right\|_F + \|\mathbf{Z}-\mathbf{Z}'\|_2 \sum_{t\in\mathcal{K}^c} \|f(\mathbf{x}_t)\|_2 \\ &\le \|\mathbf{Z}-\mathbf{Z}'\|_F \sum_{t=0}^{T-1} \|f(\mathbf{x}_t)\|_2 . \end{aligned} \tag{5.59}$$

Using the decomposition in **Step 1-1**, we have

$$f(\mathbf{x}_t) = \sum_{\ell=1}^{j} g_t^{k_\ell},$$

where $k_j$ is the maximal element in $\mathcal{K}$ such that $k_j < t$. Therefore, we can calculate that

$$\sum_{t=0}^{T-1} \|f(\mathbf{x}_t)\|_2 \le \sum_{j=1}^{|\mathcal{K}|} \sum_{t=k_j+1}^{T-1} \left\| \mathbf{g}_t^{k_j} \right\|_2 . \tag{5.60}$$

For each $j \in \{1, \dots, |\mathcal{K}|\}$, we can prove in the same way as (5.48) that

$$\left\| \mathbf{g}_t^{k_j} \right\|_2 \le L(\rho L)^{k_j - t - 1} \|\bar{\mathbf{d}}_{k_j}\|_2, \quad \forall t > k_j.$$

Substituting into inequality (5.60), it follows that

$$\sum_{t=0}^{T-1} \|f(\mathbf{x}_t)\|_2 \le \sum_{j=1}^{|\mathcal{K}|} \sum_{t=k_j+1}^{T-1} L(\rho L)^{k_j-t-1} \|\bar{\mathbf{d}}_{k_j}\|_2 \le \frac{L}{1-\rho L} \sum_{j=1}^{|\mathcal{K}|} \|\bar{\mathbf{d}}_{k_j}\|_2.$$

Using Assumption 11 and the same technique as in (5.51), we know

$$\mathbb{P}\left( \sum_{j=1}^{|\mathcal{K}|} \|\bar{\mathbf{d}}_{k_j}\|_2 \le \eta \right) \ge 1 - 2\exp\left( -\frac{\eta^2}{2\sigma^2 |\mathcal{K}|} \right) \ge 1 - 2\exp\left( -\frac{\eta^2}{4\sigma^2 pT} \right),$$

where the second inequality is from the high probability bound in (5.56). Hence, it holds that

$$\mathbb{P}\left( \sum_{t=0}^{T-1} \|f(\mathbf{x}_t)\|_2 \le \eta \right) \ge 1 - 2\exp\left( -\frac{\eta^2 (1-\rho L)^2}{4\sigma^2 L^2 pT} \right), \tag{5.61}$$

Choosing

$$\eta := \frac{\theta}{2\epsilon},$$

the bound in (5.61) becomes

$$\begin{aligned}\mathbb{P}\left(\sum_{t=0}^{T-1}\|f(\mathbf{x}_t)\|_2 \leq \frac{\theta}{2\epsilon}\right) &\geq 1-2\exp\left(-\frac{(1-\rho L)^2}{4\sigma^2L^2pT\epsilon^2}\cdot\theta^2\right) \\ &= 1-2\exp\left[-\Theta\left[\frac{(1-\rho L)^2}{4\sigma^2L^2pT\epsilon^2}\cdot\left(\frac{\lambda^5p(1-p)T}{\sigma^4L^4}\right)^2\right]\right] \\ &= 1-2\exp\left[-\Theta\left[\frac{(1-\rho L)^2\kappa^{10}p(1-p)^2T}{\epsilon^2}\right]\right].\end{aligned} \tag{5.62}$$

We set

$$\epsilon := \Theta\left[\sqrt{(1-\rho L)^2\kappa^{10}p(1-p)^2}\right].$$

Then, it follows that

$$\exp\left[-\Theta\left[\frac{(1-\rho L)^2\kappa^{10}p(1-p)^2T}{\epsilon^2}\right]\right] = \exp\left[-\Theta(T)\right] \leq \frac{\delta}{4},$$

where the last inequality is from the choice of $T$ in (5.57). Substituting back into (5.62), we get

$$\mathbb{P}\left(\sum_{t=0}^{T-1}\|f(\mathbf{x}_t)\|_2 \leq \frac{\theta}{2\epsilon}\right) \geq 1-\frac{\delta}{2}. \tag{5.63}$$

Under the event in (5.63), for all $\mathbf{Z}\in\mathbb{S}_F$, there exists an element $\mathbf{Z}^r$ in the $\epsilon$-net such that

$$f(\mathbf{Z}) \leq f(\mathbf{Z}^r) + \epsilon\cdot\sum_{t=0}^{T-1}\|f(\mathbf{x}_t)\|_2 \leq f(\mathbf{Z}^r)+\frac{\theta}{2}.$$

If we replace $\delta$ with $\delta/(2N)$ in (5.58) and choose $\mathbf{Z}=\mathbf{Z}^r$ for all $r\in\{1,\ldots,N\}$, the union bound implies that

$$\mathbb{P}\left[f(\mathbf{Z}^r)\leq -\theta,\ r=1,\ldots,N\right] \geq 1-\frac{\delta}{2}. \tag{5.64}$$

Under the above conditions, we have

$$f(\mathbf{Z}) \leq f(\mathbf{Z}^r)+\frac{\theta}{2} \leq -\frac{\theta}{2} < 0.$$

To satisfy condition (5.64), the sample complexity bound (5.57) becomes

$$\begin{aligned}T &\geq \Theta\left[\max\left\{\frac{\kappa^{10}}{(1-\rho L)^3(1-p)^2},\frac{\kappa^4}{p(1-p)}\right\}\log\left(\frac{2N}{\delta}\right)\right] \\ &= \Theta\Bigg[\max\left\{\frac{\kappa^{10}}{(1-\rho L)^3(1-p)^2},\frac{\kappa^4}{p(1-p)}\right\} \\ &\qquad\times\left[mn\log\left(\frac{1}{(1-\rho L)\kappa p(1-p)}\right)+\log\left(\frac{1}{\delta}\right)\right]\Bigg],\end{aligned}$$

which is the desired sample complexity bound in the theorem.

**Lower bound of $\kappa$.** Before we close the proof, we provide a lower bound of $\kappa = \sigma L/\lambda$. Equivalently, we provide an upper bound on $\lambda^2$, which is at most the minimal eigenvalue of

$$\mathbb{E}\left[f(\mathbf{x}+\bar{\mathbf{d}}_t)f(\mathbf{x}+\bar{\mathbf{d}}_t)^\top \mid \mathcal{F}_t, \bar{\mathbf{d}}_t \neq \mathbf{0}_n\right].$$

Let $\nu \in \mathbb{R}^m$ be a vector satisfying

$$\|\nu\|_2 = 1, \quad \nu^\top f(\mathbf{x}) = 0.$$

Then, we know

$$\begin{aligned}\nu^\top f(\mathbf{x}+\bar{\mathbf{d}}_t)f(\mathbf{x}+\bar{\mathbf{d}}_t)^\top \nu &= \nu^\top \left[f(\mathbf{x}+\bar{\mathbf{d}}_t)-f(\mathbf{x})\right]\left[f(\mathbf{x}+\bar{\mathbf{d}}_t)-f(\mathbf{x})\right]^\top \nu \\ &= \left[\left[f(\mathbf{x}+\bar{\mathbf{d}}_t)-f(\mathbf{x})\right]^\top \nu\right]^2 \leq \left\|f(\mathbf{x}+\bar{\mathbf{d}}_t)-f(\mathbf{x})\right\|_2^2 \\ &\leq L^2\|\bar{\mathbf{d}}_t\|_2^2,\end{aligned} \tag{5.65}$$

where the last inequality is from the Lipschitz continuity of $f$. Using the sub-Gaussian assumption, it follows that

$$\mathbb{E}\left[\|\bar{\mathbf{d}}_t\|_2^2 \mid \mathcal{F}_t,\ \bar{\mathbf{d}}_t \neq \mathbf{0}_n\right] \leq \sigma^2, \tag{5.66}$$

where we utilize the fact that the standard deviation of sub-Gaussian random variables with parameter $\sigma$ is at most $\sigma$. Combining inequalities (5.65) and (5.66), it follows that

$$\nu^\top \mathbb{E}\left[f(\mathbf{x}+\bar{\mathbf{d}}_t)f(\mathbf{x}+\bar{\mathbf{d}}_t)^\top \mid \mathcal{F}_t, \bar{\mathbf{d}}_t \neq \mathbf{0}_n\right]\nu \leq \sigma^2 L^2.$$

Therefore, it holds that

$$\lambda^2 \leq \lambda_{min}\left[\mathbb{E}\left[f(\mathbf{x}+\bar{\mathbf{d}}_t)f(\mathbf{x}+\bar{\mathbf{d}}_t)^\top \mid \mathcal{F}_t, \bar{\mathbf{d}}_t \neq \mathbf{0}_n\right]\right] \leq \sigma^2 L^2, \quad \forall \mathbf{x} \in \mathbb{R}^n,$$

which further leads to

$$\kappa = \frac{\sigma L}{\lambda} \geq 1.$$

This completes the proof. □

# 5.J Proof of Lemma 17

*Proof of Lemma 17.* Let

$$\delta := \frac{c\lambda^4}{\sigma^4 L^4}, \quad \theta_t := \left\|\mathbf{Z}^\top f\left[\bar{\mathbf{A}} f(\mathbf{x}_{t-1})\right]\right\|_2.$$

We finish the proof by discussing two cases.

**Case 1.** We first consider the case when

$$\theta_t \geq \frac{\lambda}{2} + \sqrt{2\sigma^2 L^2 \log\left(\frac{2}{1-\delta}\right)}.$$

Using the Lipschitz continuity of $f$, we have

$$\begin{aligned} \|\mathbf{Z}f(\mathbf{x}_t)\|_2 &= \left\| \left[\mathbf{Z}f(\mathbf{x}_t) - \mathbf{Z}^\top f\left[\bar{\mathbf{A}}f(\mathbf{x}_{t-1})\right]\right] + \mathbf{Z}f\left[\bar{\mathbf{A}}f(\mathbf{x}_{t-1})\right] \right\|_2 \\ &\geq \left\|\mathbf{Z}f\left[\bar{\mathbf{A}}f(\mathbf{x}_{t-1})\right]\right\|_2 - \left\|\mathbf{Z}f(\mathbf{x}_t) - \mathbf{Z}f\left[\bar{\mathbf{A}}f(\mathbf{x}_{t-1})\right]\right\|_2 \\ &\geq \theta_t - \|\mathbf{Z}\| \left\|f(\mathbf{x}_t) - f\left[\bar{\mathbf{A}}f(\mathbf{x}_{t-1})\right]\right\|_2 \\ &\geq \theta_t - \|\mathbf{Z}\|_F \cdot L \left\|\bar{\mathbf{d}}_t\right\|_2 \geq \theta_t - L\left\|\bar{\mathbf{d}}_t\right\|_2. \end{aligned} \tag{5.67}$$

By Assumption 11, we know $\left\|\bar{\mathbf{d}}_t\right\|_2 = |\ell_t|$ and it follows that

$$\mathbb{P}\left(\left\|\bar{\mathbf{d}}_t\right\|_2 \geq \epsilon \mid \mathcal{F}_t\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right), \quad \forall \epsilon \geq 0.$$

Therefore, we get the estimation

$$\begin{aligned} \mathbb{P}\left(\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \leq \frac{\lambda}{2} \,\middle|\, \mathcal{F}_t\right) &\leq \mathbb{P}\left(\theta_t - L\left\|\bar{\mathbf{d}}_t\right\|_2 \leq \frac{\lambda}{2} \,\middle|\, \mathcal{F}_t\right) \\ &= \mathbb{P}\left(\left\|\bar{\mathbf{d}}_t\right\|_2 \geq \frac{\theta_t - \lambda/2}{L} \,\middle|\, \mathcal{F}_t\right) \\ &\leq \mathbb{P}\left(\left\|\bar{\mathbf{d}}_t\right\|_2 \geq \sqrt{2\sigma^2 \log\left(\frac{2}{1-\delta}\right)} \,\middle|\, \mathcal{F}_t\right) \leq 1-\delta. \end{aligned}$$

Therefore, we have proved that

$$\mathbb{P}\left(\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \frac{\lambda}{2} \,\middle|\, \mathcal{F}_t\right) \geq \delta.$$

**Case 2.** Then, we focus on the case when

$$\theta_t \leq \frac{\lambda}{2} + \sqrt{2\sigma^2 L^2 \log\left(\frac{2}{1-\delta}\right)}. \tag{5.68}$$

Assume conversely that

$$\mathbb{P}\left(\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \frac{\lambda}{2} \,\middle|\, \mathcal{F}_t\right) < \delta. \tag{5.69}$$

Similar to inequality (5.67), the Lipschitz continuity of $f$ implies

$$\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \leq \theta_t + L\left\|\bar{\mathbf{d}}_t\right\|_2.$$

Therefore, by applying Assumption 11, we get the tail bound

$$\begin{aligned}&\mathbb{P}\left(\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \theta \mid \mathcal{F}_t\right) \leq \mathbb{P}\left(\theta_t + L\left\|\bar{\mathbf{d}}_t\right\|_2 \geq \theta \mid \mathcal{F}_t\right)\\ =&\mathbb{P}\left(\left\|\bar{\mathbf{d}}_t\right\|_2 \geq \frac{\theta-\theta_t}{L} \;\middle|\; \mathcal{F}_t\right) \leq 2\exp\left[-\frac{(\theta-\theta_t)^2}{2\sigma^2L^2}\right], \quad \forall\theta \geq \theta_t.\end{aligned}$$

Define $(x)_+ := \max\{x, 0\}$. The bound above leads to

$$\mathbb{P}\left(\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \theta \mid \mathcal{F}_t\right) \leq 2\exp\left[-\frac{(\theta-\theta_t)_+^2}{2\sigma^2L^2}\right], \quad \forall\theta \in \mathbb{R}. \tag{5.70}$$

Using the definition of expectation, we can calculate that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{Z}f(\mathbf{x}_t)\|_2^2 \mid \mathcal{F}_t\right] &= \int_0^\infty 2\theta\cdot\mathbb{P}\left[\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \theta \mid \mathcal{F}_t\right]\, d\theta\\ &\leq \frac{\lambda^2}{4} + \int_{\lambda/2}^\infty 2\theta\cdot\mathbb{P}\left[\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \theta \mid \mathcal{F}_t\right]\, d\theta.\end{aligned}$$

By condition (5.69), we get

$$\mathbb{P}\left[\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \theta \mid \mathcal{F}_t\right] \leq \mathbb{P}\left[\|\mathbf{Z}f(\mathbf{x}_t)\|_2 \geq \frac{\lambda}{2} \;\middle|\; \mathcal{F}_t\right] \leq \delta, \quad \forall\theta \geq \frac{\lambda}{2}.$$

Combining with inequality (5.70), it follows that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{Z}f(\mathbf{x}_t)\|_2^2 \mid \mathcal{F}_t\right] &\leq \frac{\lambda^2}{4} + \int_{\lambda/2}^\infty 2\theta\cdot\min\left\{\delta, 2\exp\left[-\frac{(\theta-\theta_t)_+^2}{2\sigma^2L^2}\right]\right\}\, d\theta \qquad (5.71)\\ &= \frac{\lambda^2}{4} + \delta\left(\theta_1^2 - \frac{\lambda^2}{4}\right) + \int_{\theta_1}^\infty 4\theta\exp\left[-\frac{(\theta-\theta_t)^2}{2\sigma^2L^2}\right]\, d\theta,\end{aligned}$$

where we define

$$\theta_1 := \max\left\{\frac{\lambda}{2}, \theta_t + \sqrt{2\sigma^2L^2\log\left(\frac{2}{\delta}\right)}\right\} \geq \theta_t.$$

Using condition (5.68), we know

$$\begin{aligned}\theta_1^2 &\leq \left(\frac{\lambda}{2} + \sqrt{2\sigma^2L^2\log\left(\frac{2}{1-\delta}\right)} + \sqrt{2\sigma^2L^2\log\left(\frac{2}{\delta}\right)}\right)^2 \qquad (5.72)\\ &\leq \left(\frac{\lambda}{2} + 2\sqrt{2\sigma^2L^2\log\left(\frac{2}{\delta}\right)}\right)^2 \leq \frac{\lambda^2}{2} + 16\sigma^2L^2\log\left(\frac{2}{\delta}\right),\end{aligned}$$

where the last inequality is from Cauchy's inequality. Moreover, we can estimate that

$$\begin{aligned}&\int_{\theta_1}^{\infty} 4\theta \exp\left[-\frac{(\theta-\theta_t)^2}{2\sigma^2 L^2}\right]\, d\theta \le \int_{\theta_2}^{\infty} 4\theta \exp\left[-\frac{(\theta-\theta_t)^2}{2\sigma^2 L^2}\right]\, d\theta \\ =&\int_{\theta_2}^{\infty} 4\theta_t \exp\left[-\frac{(\theta-\theta_t)^2}{2\sigma^2 L^2}\right]\, d\theta + \int_{\theta_2}^{\infty} 4(\theta-\theta_t) \exp\left[-\frac{(\theta-\theta_t)^2}{2\sigma^2 L^2}\right]\, d\theta \\ =&\int_{\theta_2}^{\infty} 4\theta_t \exp\left[-\frac{(\theta-\theta_t)^2}{2\sigma^2 L^2}\right]\, d\theta + 2\delta\sigma^2 L^2,\end{aligned} \tag{5.73}$$

where we denote $\theta_2 := \theta_t + \sqrt{2\sigma^2 L^2 \log\left(\frac{2}{\delta}\right)} \le \theta_1$. Utilizing the following bound on the cumulative density function of the standard Gaussian distribution:

$$\int_{\eta}^{\infty} e^{-\frac{x^2}{2}}\, dx \le \eta^{-1} e^{-\frac{\eta^2}{2}}, \quad \forall \eta > 0,$$

we have

$$\int_{\theta_2}^{\infty} 4\theta_t \exp\left[-\frac{(\theta-\theta_t)^2}{2\sigma^2 L^2}\right]\, d\theta \le 4\theta_t \sigma L \cdot \frac{1}{\sqrt{2\log\left(\frac{2}{\delta}\right)}} \cdot \frac{\delta}{2} \le \sqrt{2}\theta_t \cdot \delta\sigma L.$$

Combining with (5.73), it follows that

$$\int_{\theta_1}^{\infty} 4\theta \exp\left[-\frac{(\theta-\theta_t)^2}{2\sigma^2 L^2}\right]\, d\theta \le \sqrt{2}\theta_t \cdot \delta\sigma L + 2\delta\sigma^2 L^2 \le 4\delta\theta_t^2 + 4\delta\sigma^2 L^2, \tag{5.74}$$

where the last inequality is from Cauchy's inequality. Substituting inequalities (5.72) and (5.74) back into (5.71), we get

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{Z}f(\mathbf{x}_t)\|_2^2 \mid \mathcal{F}_t\right] &\le \frac{\lambda^2}{4} + \delta\left[\frac{\lambda^2}{4} + 16\sigma^2 L^2 \log\left(\frac{2}{\delta}\right)\right] + 4\delta\theta_t^2 + 4\delta\sigma^2 L^2 \\ &\le \frac{(1+\delta)\lambda^2}{4} + 16\sigma^2 L^2 \cdot \delta \log\left(\frac{2}{\delta}\right) \\ &\qquad + \delta\left[\frac{\lambda}{2} + \sqrt{2\sigma^2 L^2 \log\left(\frac{2}{1-\delta}\right)}\right]^2 + 4\delta\sigma^2 L^2 \\ &\le \frac{(1+\delta)\lambda^2}{4} + 16\sigma^2 L^2 \cdot \delta \log\left(\frac{2}{\delta}\right) + \frac{\delta\lambda^2}{2} + 4\sigma^2 L^2 \cdot \delta \log\left(\frac{2}{\delta}\right) + 4\delta\sigma^2 L^2 \\ &\le \frac{(1+3\delta)\lambda^2}{4} + 24\sigma^2 L^2 \cdot \delta \log\left(\frac{2}{\delta}\right).\end{aligned}$$

where the second inequality is from (5.68) and the last inequality is from Cauchy's inequality and $\delta < 1/2$. On the other hand, Assumption 8 implies that

$$\mathbb{E}\left(\|\mathbf{Z}f(\mathbf{x}_t)\|_2^2 \mid \mathcal{F}_t\right) = \left\langle \mathbf{Z}\mathbf{Z}^\top, \mathbb{E}\left[f(\mathbf{x}_t)f(\mathbf{x}_t)^\top \mid \mathcal{F}_t\right]\right\rangle \ge \lambda^2 \|\mathbf{Z}\|_F^2 = \lambda^2.$$

Combining the last two inequalities, we get

$$\lambda^2 \leq \frac{(1+3\delta)\lambda^2}{4} + 24\sigma^2 L^2 \cdot \delta \log\left(\frac{2}{\delta}\right),$$

which is equivalent to

$$\delta \log\left(\frac{2}{\delta}\right) \geq \frac{(3-3\delta)\lambda^2}{96\sigma^2 L^2} \geq \frac{\lambda^2}{23\sigma^2 L^2}.$$

For all $x \in (0,1)$, it holds that $x\log(2/x) < \sqrt{2x}$. Hence, we have

$$\sqrt{2\delta} > \frac{\lambda^2}{23\sigma^2 L^2},$$

which contradicts with our assumption (5.69).

□

## 5.K Proof of Theorem 24

*Proof of Theorem 24.* In this proof, we focus on the case when $m = n$, and the counterexample can be easily extended into more general cases. We construct the following system dynamics:

$$\bar{\mathbf{A}} := \rho \mathbf{I}_n, \quad f(\mathbf{x}) := \mathbf{x}, \quad \forall x \in \mathbb{R}^n,$$

where $\rho \geq 2 + \sqrt{6}$ is a constant. One can verify Assumption 9 holds with Lipschitz constant $L = 1$. Therefore, the stability condition (Assumption 10) is violated since $\rho > 1/L$. The system dynamics can be written as

$$\mathbf{x}_t = \sum_{k \in \mathcal{K}, k<t} \rho^{t-k-1} \mathbf{d}_k, \quad \forall t \in \{0, \ldots, T\}. \tag{5.75}$$

Conditional on $\mathcal{F}_t$ and $t \in \mathcal{K}$, the disturbance vector is generated as

$$\mathbf{d}_t \sim \mathrm{Uniform}(\mathbb{S}^{n-1}),$$

where $\mathbb{S}^{n-1}$ is the unit ball $\{\mathbf{d} \in \mathbb{R}^n \mid \|\mathbf{d}\|_2 = 1\}$. The attack/disturbance model satisfies Assumption 8 with $\lambda = 1/\sqrt{n}$ and Assumption 11 with $\sigma = 1/\sqrt{n}$. Define the event

$$\mathcal{E} := \{T - 1 \in \mathcal{K}, |\mathcal{K}| > 1\}.$$

By the definition of the probabilistic sparsity model, we can calculate that

$$\mathbb{P}(\mathcal{E}) = p\left[1 - (1-p)^{T-1}\right].$$

Our goal is to prove that

$$\mathbb{P}\left[\hat{d}_1(\mathbf{Z}) - \hat{d}_2(\mathbf{Z}) > 0 \mid \mathcal{E}\right] = 1,$$

where we define

$$\hat{d}_1(\mathbf{Z}) := \sum_{t\in\mathcal{K}} \left\langle \mathbf{Z}^\top, f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top \right\rangle, \quad \hat{d}_2(\mathbf{Z}) := \sum_{t\in\mathcal{K}^c} \|\mathbf{Z}f(\mathbf{x}_t)\|_2 .$$

Then, by Theorem 18, we know that $\bar{\mathbf{A}}$ is not a global solution to problem (5.4) with probability at least

$$p\left[1-(1-p)^{T-1}\right].$$

Let $t_1$ be the smallest element in $\mathcal{K}$, namely, the first time instance when there is a non-zero disturbance. Under event $\mathcal{E}$, it holds that $t_1 < T-1$. We first prove that

$$\mathbf{x}_t \neq \mathbf{0}_n, \quad \forall t \in \{t_1+1, \ldots, T-1\}.$$

By the system dynamics (5.75) and the triangle inequality, we have

$$\begin{aligned}\|\mathbf{x}_t\|_2 &\geq \rho^{t-t_1-1}\|\mathbf{d}_{t_1}\|_2 - \sum_{k\in\mathcal{K},t_1<k<t} \rho^{t-k-1}\|\mathbf{d}_k\|_2 = \rho^{t-t_1-1} - \sum_{k\in\mathcal{K},t_1<k<t} \rho^{t-k-1} \\ &\geq \rho^{t-t_1-1} - \sum_{i=0}^{t-t_1-2} \rho^i = \frac{\rho^{t-t_1} - 2\rho^{t-t_1-1}+1}{\rho-1} > 0,\end{aligned}$$

where the last inequality holds because $\rho \geq 2$. Then, we choose

$$\mathbf{Z} := \mathbf{x}_{T-1}\hat{\mathbf{d}}_{T-1}^\top \neq 0.$$

It follows that

$$\begin{aligned}\hat{d}_1(\mathbf{Z}) &= \sum_{t\in\mathcal{K}} \left\langle \mathbf{Z}^\top, f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top \right\rangle = \left\| \mathbf{x}_{T-1}\hat{\mathbf{d}}_{T-1}^\top \right\|_F^2 + \sum_{t\in\mathcal{K},t<T-1} \left\langle \mathbf{x}_{T-1}\hat{\mathbf{d}}_{T-1}^\top, f(\mathbf{x}_t)\hat{\mathbf{d}}_t^\top \right\rangle \\ &\geq \|\mathbf{x}_{T-1}\|_2^2 - \sum_{t\in\mathcal{K},t<T-1} \|\mathbf{x}_{T-1}\|_2 \|\mathbf{x}_t\|_2 , \\ \hat{d}_2(\mathbf{Z}) &= \sum_{t\in\mathcal{K}^c} \|\mathbf{Z}f(\mathbf{x}_t)\|_2 = \sum_{t\in\mathcal{K}^c} \left\| \mathbf{x}_{T-1}\hat{\mathbf{d}}_{T-1}^\top \mathbf{x}_t \right\|_2 \leq \sum_{t\in\mathcal{K}^c} \|\mathbf{x}_{T-1}\|_2 \|\mathbf{x}_t\|_2 .\end{aligned}$$

Combining the above two inequalities, we get

$$\hat{d}_1(\mathbf{Z}) - \hat{d}_2(\mathbf{Z}) \leq \|\mathbf{x}_{T-1}\|_2 \left( \|\mathbf{x}_{T-1}\|_2 - \sum_{t=0}^{T-2} \|\mathbf{x}_t\|_2 \right) = \|\mathbf{x}_{T-1}\|_2 \left( \|\mathbf{x}_{T-1}\|_2 - \sum_{t=t_1+1}^{T-2} \|\mathbf{x}_t\|_2 \right),$$

where the last equality holds because $\mathbf{x}_t = \mathbf{0}_n$ for all $t \leq t_1$. Since $\|\mathbf{x}_{T-1}\|_2 > 0$, it is sufficient to prove that

$$\|\mathbf{x}_{T-1}\|_2 > \sum_{t=t_1+1}^{T-2} \|\mathbf{x}_t\|_2 . \tag{5.76}$$

Considering the system dynamics (5.75) and the fact that $\|\mathbf{d}_k\|_2 = 1$ for all $k \in \mathcal{K}$ , we have the estimation

$$\rho^{t-t_1-1} - \sum_{k\in\mathcal{K}, t_1<k<t} \rho^{t-k-1} \leq \|\mathbf{x}_t\|_2 \leq \sum_{k\in\mathcal{K}, k<t} \rho^{t-k-1}.$$

The desired inequality (5.76) holds if we can show

$$\rho^{T-1-t_1-1} - \sum_{k\in\mathcal{K}, t_1<k<T-1} \rho^{T-1-k-1} > \sum_{t=t_1+1}^{T-2} \sum_{k\in\mathcal{K}, k<t} \rho^{t-k-1},$$

which is further equivalent to

$$\begin{aligned}
&2\rho^{T-t_1-2} > \sum_{t=t_1+1}^{T-1} \sum_{k\in\mathcal{K}, k<t} \rho^{t-k-1} \\
\Longleftarrow\ &2\rho^{T-t_1-2} > \sum_{t=t_1+1}^{T-1} \sum_{k=t_1}^{t-1} \rho^{t-k-1} = \sum_{t=t_1+1}^{T-1} \frac{\rho^{t-t_1}-1}{\rho-1} = \frac{\rho^{T-t_1} - \rho - (T-t_1-1)(\rho-1)}{(\rho-1)^2} \\
\Longleftarrow\ &2\rho^{T-t_1-2} \geq \frac{\rho^{T-t_1}}{(\rho-1)^2} \iff \rho^2 - 4\rho - 2 \geq 0 \impliedby \rho \geq 2+\sqrt{6}.
\end{aligned}$$

By our choice of $\rho$, we know condition (5.76) holds and this completes our proof.

□

# 5.L Extensions of Numerical Experiments for Different Problem Parameters

## Numerical Experiments on Spectral Norm of $\bar{\mathbf{A}}$

In this section, we use the same experimental setup as in Section 5.7 for Lipschitz continuous basis functions. We examine the relationship between sample complexity and the spectral norm $\rho$. Specifically, we set $T = 100$, $p = 0.75$, and $n = 3$. To eliminate randomness in the spectral norm $\|\bar{\mathbf{A}}\|$, we assign the singular values of $\bar{\mathbf{A}}$ as $\sigma_1 = \cdots = \sigma_n = \rho$, where $\rho \in \{0.5, 0.95, 1.5\}$. In the case where $\rho = 1.5$, we terminate the simulation when $\|\mathbf{x}_t\|_2 \geq 10^{14}$, as this indicates that the trajectory diverges to infinity, causing numerical issues for the CVX solver.

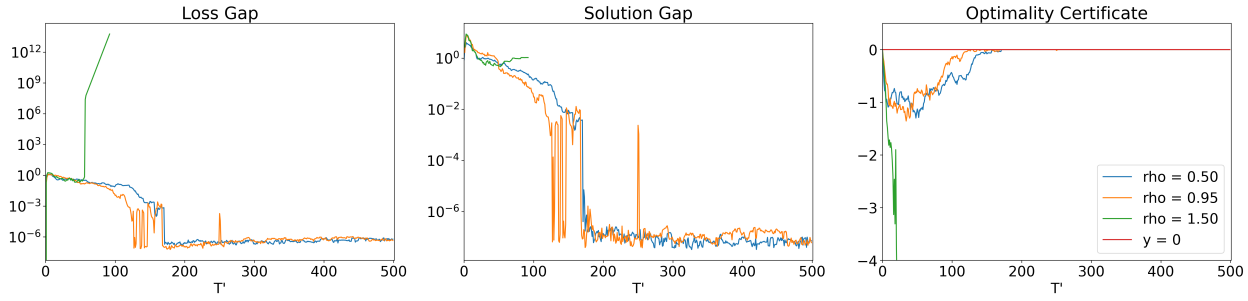


Figure 5.5: Loss Gap, Solution Gap, and Optimality Certificate of the Lipschitz Basis Function Case with Spectral Norm $\rho = 0.5, 0.95$ and $1.5$.

The results, presented in Figure 5.5, reveal that the required sample complexity increases slightly as $\rho$ increases from $0.5$ to $0.95$, which is consistent with Theorem 23. In addition, when $\rho = 1.5$, the system is not asymptotically stable, violating Assumption 10. The resulting divergence of the system state ($\|\mathbf{x}_t\|_2 \to \infty$) leads to numerical instabilities in computing the estimator (5.3). However, it is possible that the estimator in (5.3) could still achieve exact recovery for large values of $\rho$, if a numerically stable method is employed for optimization. This does not contradict our theoretical findings, as Theorem 23 provides only a sufficient condition for exact recovery, rather than a necessary one.

## Numerical Experiments with Small Probabilistic Sparsity Model

In this section, we repeat the experiments in Figure 5.3 with $p \in \{0.001, 0.1, 0.3\}$ and $n = 5$. The results are presented in Figure 5.6. We observe that the predictor fails to recover the ground truth within $500$ steps when $p = 0.001$, whereas it successfully converges when $p = 0.1$ and $0.3$. Notably, both the loss gap and the optimality certificate are equal to zero in the case of $p = 0.001$. This occurs because multiple global solutions exist, causing the estimator to fail in recovering the unique ground truth solution within $500$ iterations. However, given a larger number of samples, the algorithm will eventually converge to the correct solution. That said, the primary focus of this chapter is the regime where $p > 0.5$. When $p$ is very small or zero, learning the system falls within the classical control theory framework, where it is well established that an artificial excitation signal must be introduced to facilitate learning. The necessity of excitation signals in nearly deterministic systems is well documented in the control literature. For example, consider the linear system $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t$, where the objective is to learn $\mathbf{A}$ from observations of $\mathbf{x}_t$. If the initial condition $\mathbf{x}_0$ is zero, then $\mathbf{x}_t$ remains zero for all $t$, making it impossible to infer $\mathbf{A}$. To circumvent this issue, an artificial excitation signal is typically introduced, yielding a system of the form $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{w}_t$ where $\mathbf{w}_t$ is, for instance, Gaussian noise. Interestingly, when $p$ is sufficiently large, the adversarial attack itself serves as an effective excitation signal, aiding in the learning process by introducing necessary perturbations into the system.

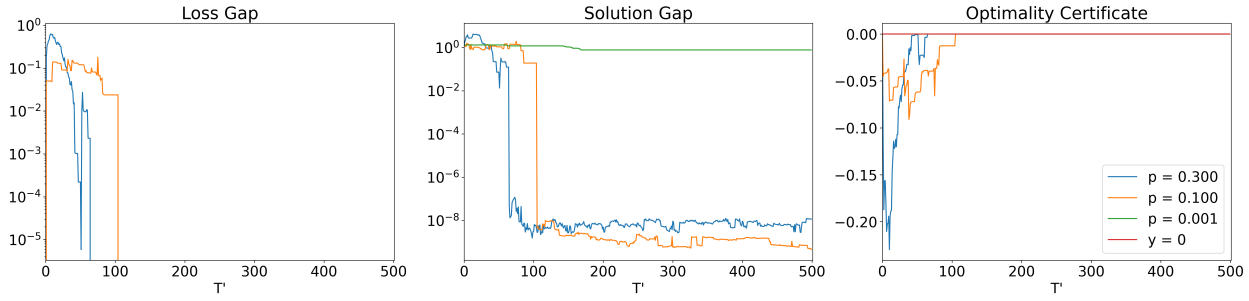


Figure 5.6: Loss Gap, Solution Gap, and Optimality Certificate of the Lipschitz Basis Function Case with Attack/Disturbance Probability $p = 0.001, 0.1$ and $0.3$.

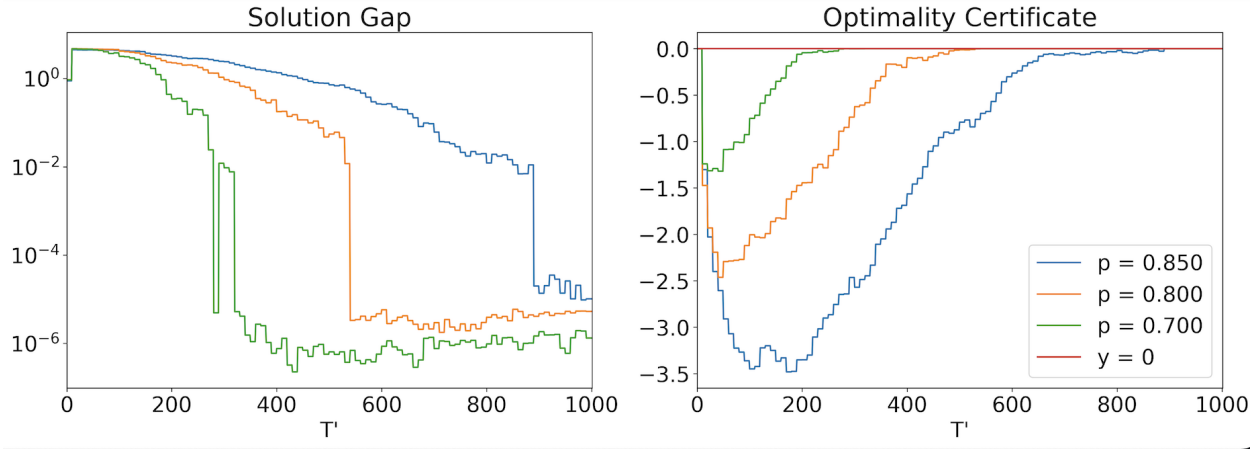


Figure 5.7: Loss Gap, Solution Gap, and Optimality Certificate of the Lipschitz Basis Function Case with $p \in \{0.7, 0.8, 0.85\}$ and $n = m = 10$.

## Numerical Experiments with Sparse $\bar{\mathbf{A}}$

In this section, we replicate the experiments presented in Figure 5.3, using a sparse ground truth matrix $\bar{\mathbf{A}}$. Specifically, we generate a sparse matrix $\bar{\mathbf{A}}$ as a tridiagonal matrix, where each entry $\bar{\mathbf{A}}_{i,j}$ is set to zero whenever $|i - j| > 1$. We conduct the experiments for Lipschitz basis functions with non-zero disturbance probabilities $p \in 0.7, 0.8, 0.85$ and system dimension $n = 10$. Additionally, we extend the simulation period to $T = 1000$, compared to $T = 500$ in the previous experiments. To improve computational efficiency, we solve the optimization problem (5.3) every $10$ time steps. Consequently, the plots exhibit discrete jumps corresponding to time periods that are multiples of ten. The loss gap is omitted from the figures since the estimator is computed only at a subset of time points. Figure 5.7 suggests that exact recovery is achieved despite the sparse structure of the ground truth matrix $\bar{\mathbf{A}}$. This result is expected, as our theoretical analysis does

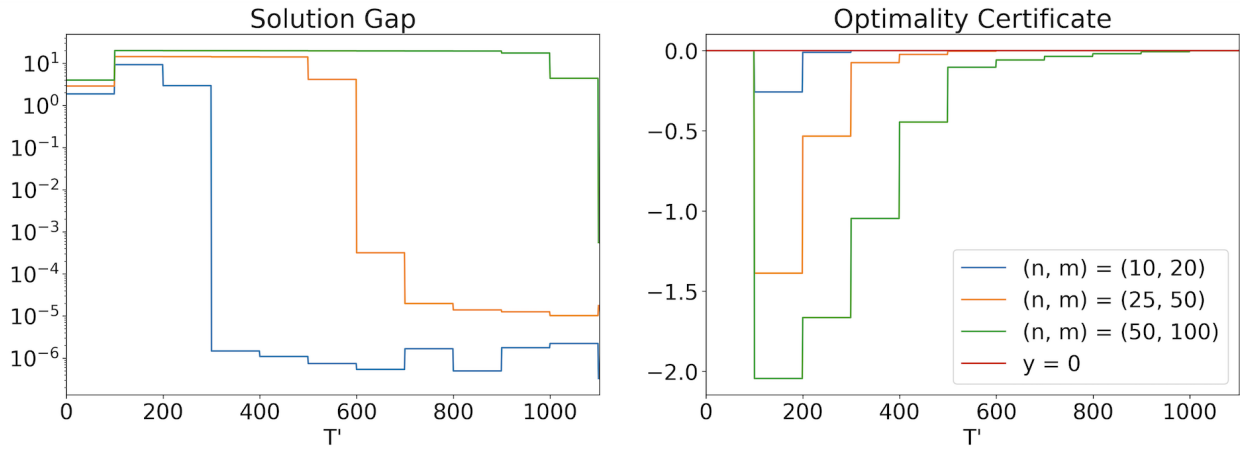


Figure 5.8: Loss Gap, Solution Gap, and Optimality Certificate of the Lipschitz Basis Function Case with Dimension $(n, m) = (10, 20), (25, 50)$ and $(50, 100)$.

not impose any dependency on the sparsity structure of $\bar{\mathbf{A}}$. Beyond demonstrating robustness, the non-smooth objective function in (5.3) also acts as an implicit regularization mechanism that aligns with the specific matrix structure.

## Numerical Experiments with Larger Order Systems

In this section, we extend the experiments presented in Figure 5.4 to significantly higher-order dynamical systems and a larger number of basis functions. Specifically, we consider system dimensions and basis function counts of $(n, m) \in (10, 20), (25, 50), (50, 100)$. The probability of a non-zero disturbance occurring is set to $p = 0.6$. Additionally, we increase the simulation period to $T = 1100$, compared to $T = 500$ in the previous experiments. To optimize computational efficiency, we solve the optimization problem (5.3) every 100 time periods. As a result, the plots exhibit discrete jumps corresponding to time periods that are multiples of 100. Due to this sampling strategy, the loss gap is omitted from the figures, as the estimator is computed only at a subset of time points.

In Figure 5.8, we observe that exact recovery is achieved even for large-scale system identification problems, where both the system dimension and the number of basis functions are significantly high. However, achieving exact recovery requires the system to evolve over a sufficiently long time horizon, as indicated by our theoretical results.

# Chapter 6

# Algorithm for Non-Smooth System Identification

Dynamical systems form the foundation of modern problems, such as reinforcement learning, auto-regressive models, and control systems. These systems evolve based on their current state, with the system dynamics equations determining the next state. Due to the complexity of real-world systems, these dynamics are often unknown or difficult to model precisely. Learning the unknown dynamics of a dynamical system is known as the system identification problem.

The system identification problem has been extensively studied in the literature [26, 68]. The least-squares estimator (LSE) is the most widely studied approach for this problem as it provides a closed-form solution. Early research focused on the asymptotic (infinite-time) convergence properties of LSE [72]. Recently, finite-time (non-asymptotic) analyses gained popularity, leveraging high-dimensional statistical techniques for both linear [102, 32, 106, 39] and non-linear dynamical systems [45, 95, 137]. See [107] for a survey. Traditionally, asymptotic and non-asymptotic analyses of LSE focused on systems with independent and identically distributed (sub)-Gaussian disturbances. However, this assumption is often unrealistic for safety-critical systems such as autonomous vehicles, unmanned aerial vehicles, and power grids. In these settings, disturbances may be designed by an adversarial agent, leading to dependencies across time and arbitrary variations.

Since LSE is susceptible to outliers and temporal dependencies, recent research focused on robust non-smooth estimators for safety-critical systems, particularly lasso-type estimators. [43] and [44] analyzed the properties of these estimators through the lens of the Null Space Property, a key concept in lasso analysis. Later works introduced a disturbance structure where disturbance vectors are zero with probability $1 - p$ and non-zero with probability $p$. For any $p$ between $0$ and $1$, [119] and [127] proved that lasso-type estimators with non-smooth objectives can exactly recover system dynamics for linear and parameterized non-linear systems, respectively, with probability approaching one. Additionally, [62] showed that asymmetric disturbances with non-zero probability $p$ could be transformed into zero-mean disturbances with probability $2p$, enabling exact recovery for any disturbance structure where $p < 1/2$. These non-asymptotic analyses leverage techniques from modern high-dimensional probability and statistics [111, 109]. Despite their strong learning properties, these approaches rely on convex optimization solvers, which become computationally

prohibitive as the system state dimension and time horizon grow. Safety-critical systems require rapid estimation to ensure real-time control without delays. Thus, it is imperative to design fast algorithms that efficiently solve lasso-type non-smooth estimators within a reasonable time frame.

Gradient- and Hessian-based first-order and second-order algorithms are not viable options in this context due to the non-smoothness of the objective. Instead, we employ the subgradient method, a well-established technique in the literature, as the lasso-type estimator is a convex optimization problem. Unlike gradient-based methods, subgradient methods do not guarantee descent at each iteration. However, existing results show that the best sub-optimality gap asymptotically converges to zero over time with constant and diminishing step sizes. For a comprehensive overview of subgradient methods for non-smooth objectives, see [100, 89, 8, 83, 82]. Closed-form solutions and gradient-based methods are widely used in system identification problems, as the literature has primarily focused on LSE or other smooth estimators for linear systems [59]. Nonlinear system identification is inherently more complex than its linear counterpart [96], yet existing estimation techniques still rely on smooth penalty functions. In contrast, subgradient methods have not been explored in the context of system identification. It is important to note that we do not apply the subgradient method to a single optimization problem but instead apply it to a sequence of time-varying problems because as new measurements are taken, the optimization problem to be solved changes as well.

## 6.1 Non-smooth Estimator and Disturbance Model

We consider a linear time-invariant dynamical system of order $n$ with the system update equation

$$\mathbf{x}_{t+1} = \bar{\mathbf{A}}\mathbf{x}_t + \bar{\mathbf{d}}_t, \quad t = 0, 1, \ldots, T-1,$$

where $\bar{\mathbf{A}} \in \mathbb{R}^{n\times n}$ is the unknown system matrix and $\bar{\mathbf{d}}_t \in \mathbb{R}^n$ are unknown system disturbances. Our goal is to estimate $\bar{\mathbf{A}}$ using the state samples $\{\mathbf{x}_t\}_{t=0}^{T}$ obtained from a single system initialization. Note that we only study the autonomous case and the generalization of the case with inputs is straightforward. Traditionally, the following LSE has been widely studied in the literature:

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times n}} \sum_{t=0}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t\|_2^2 . \tag{LSE}$$

However, the LSE is vulnerable to outliers and adversarial disturbances. In safety-critical system identification, a new popular alternative is the following non-smooth estimator:

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times n}} \sum_{t=0}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t\|_2 . \tag{NSE}$$

The primary objective of this non-smooth estimator is to achieve exact recovery in finite time, an essential feature for safety-critical applications such as power-grid monitoring, unmanned aerial vehicles, and autonomous vehicles. In these scenarios, the system should be learned from a single

trajectory (since multiple initialization is often infeasible) and can be exposed to adversarial agents. Even under i.i.d. disturbance vectors, if $\bar{\mathbf{d}}_t$ is non-zero at every time step $t$, the exact recovery of the $\bar{\mathbf{A}}$ is not attainable in finite time. Motivated by [127], we introduce a disturbance model in which the disturbance vector $\bar{\mathbf{d}}_t$ is zero at certain time instances.

**Definition 11** (Probabilistic sparsity model [127])**.** *For each time instance $t$, the disturbance vector $\bar{\mathbf{d}}_t$ is zero with probability $1-p$, where $p \in (0,1)$. Furthermore, the time instances at which the disturbances are non-zero are independent of each other while the non-zero disturbance values could be correlated.*

[127] showed that the estimator (LSE) fails to achieve exact recovery in finite time even when $p < 1$ and in the simplest case when the disturbance vectors are i.i.d. Gaussian. Consequently, employing non-smooth estimators to enable exact recovery within a finite time horizon becomes essential. Furthermore, without specific assumptions on the system dynamics and disturbance vectors, exact recovery is unattainable in finite time. See Theorem 3 in [127]. First, we impose a sufficient condition for the system stability of the LTI system in Assumption 12 to guarantee learning under possibly large correlated disturbances.

**Assumption 12.** *The ground-truth $\bar{\mathbf{A}}$ satisfies $\rho := \left\|\bar{\mathbf{A}}\right\| < 1$.*

If the disturbance vectors take arbitrary values and are chosen from a low-dimensional subspace, it is shown in [127] that learning the whole system dynamics becomes infeasible. To address this, we define the filtration $\mathcal{F}_t := \sigma\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t\}$, and we add the semi-oblivious and sub-Gaussian disturbances assumption below based on [127].

**Assumption 13** (Semi-Oblivious and sub-Gaussian disturbances)**.** *Conditional on the filtration $\mathcal{F}_t$ and the event that $\bar{\mathbf{d}}_t \neq \mathbf{0}_n$, the attack vector $\bar{\mathbf{d}}_t$ is defined by the product $\ell_t \hat{\mathbf{d}}_t$, where*

1. *$\hat{\mathbf{d}}_t \in \mathbb{R}^n$ and $\ell_t \in \mathbb{R}$ are independent conditional on $\mathcal{F}_t$ and $\bar{\mathbf{d}}_t \neq \mathbf{0}_n$;*
2. *$\hat{\mathbf{d}}_t$ obeys the uniform distribution on the unit sphere whenever $\bar{\mathbf{d}}_t \neq \mathbf{0}_n$;*
3. *$\ell_t$ is zero-mean and sub-Gaussian with parameter $\sigma$.*

Assumption 13 implies that the disturbance vectors $\bar{\mathbf{d}}_t$ are sub-Gaussian and temporally correlated since $\ell_t$ may depend on previous disturbances. This correlation better reflects real-world semi-oblivious attacks, where an adversarial agent can inject disturbances that depend on prior injections. Moreover, Assumption 13 ensures exploration of the entire state space, as the direction vector $\hat{\mathbf{d}}_t$ is uniformly distributed on the unit sphere.

In recent years, non-smooth estimators for system identification gained traction [43, 44, 119, 127, 62]. These studies analyzed non-smooth estimators in a non-asymptotic framework under various disturbance models. Specifically, under Assumptions 12 and 13, [119] and [127] derived sample complexity bounds for exact recovery in finite time with high probability. However, they do not propose numerical algorithms and rely on existing convex optimization solvers, which suffer from two drawbacks. The first issue is that in practice, (NSE) should be repeatedly solved for

$T = 1, 2, ...$ as new measurements are taken until a satisfactory model is obtained. This means the methods in [15] and [16] require solving a convex optimization problem at each time step. In addition, the problem (NSE) at time $T$ is translated into a second-order conic program with $T + n^2$ scalar variables and $T$ conic inequalities. Hence, the solver performance degrades significantly as the order of the system and the time horizon grow. Since safety-critical applications demand fast and efficient identification, convex optimization solvers are inefficient for real-time computation. We investigate the subgradient method for learning system parameters to address these limitations.

## 6.2 Subgradient Method

Let $f_T(\mathbf{A})$ denote the objective function of (NSE) and $\partial f_T(\mathbf{A})$ denote its subdifferential at a point $\mathbf{A}$ and time period $T$. We denote a subgradient in the subdifferential as $\mathbf{G}_{\mathbf{A},T} \in \partial f_T(\mathbf{A})$. We first define the subdifferential of the $\ell_2$ norm.

**Definition 12** (Subdifferential of $\ell_2$ Norm)**.** *Let $\partial\|\mathbf{x}\|_2$ denote the subdifferential of the $\ell_2$ norm, i.e.,*

$$\partial\|\mathbf{x}\|_2 = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, & \text{if } \mathbf{x} \neq 0, \\ \mathbb{B}_2(1), & \text{otherwise .} \end{cases}$$

The subdifferential is a unique point when $\mathbf{x}$ is non-zero, but it consists of all points within the unit ball when $\mathbf{x} = 0$. Using the previous definition, we obtain the subdifferential of the objective function, $\partial f_T(\mathbf{A})$.

**Lemma 18.** *The subdifferential of $f_T(\cdot)$ at a point $\mathbf{A}$ is*

$$\partial f_T(\mathbf{A}) = \left\{ -\sum_{t=0}^{T-1} \partial\|\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t\|_2 \otimes \mathbf{x}_t \right\},$$

*and the subdifferential of the $f_T(\cdot)$ at the point $\bar{\mathbf{A}}$ is*

$$\partial f_T(\bar{\mathbf{A}}) = \left\{ -\sum_{t\in\mathcal{K}} \hat{\mathbf{d}}_t \otimes \mathbf{x}_t - \sum_{t\notin\mathcal{K}} \mathbf{e}_t \otimes \mathbf{x}_t : \|\mathbf{e}_t\|_2 \leq 1, \forall t \notin \mathcal{K} \right\},$$

*where $\mathcal{K} := \{t \mid \bar{\mathbf{d}}_t \neq 0\}$ is the set of time indices of non-zero disturbances and*

$$\hat{\mathbf{d}}_t := \bar{\mathbf{d}}_t / \|\bar{\mathbf{d}}_t\|_2, \forall t \in \mathcal{K},$$

*are the normalized disturbance directions.*

The proof of Lemma 18 is straightforward and follows from a property of the Minkowski sum over convex sets. Furthermore, due to the convexity of $f_T(\mathbf{A})$, the following lemma provides a useful bound on the difference between the objective function values at two distinct points.

**Lemma 19** ([16]). *Given two arbitrary points $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ and any subgradient $\mathbf{G}_{\mathbf{A},T} \in \partial f_T(\mathbf{A})$, it holds that*

$$f_T(\mathbf{B}) \geq f_T(\mathbf{A}) + \langle \mathbf{G}_{\mathbf{A},T}, \mathbf{B} - \mathbf{A} \rangle.$$

To find the matrix $\bar{\mathbf{A}}$, we propose Algorithm 1 that relies on the subgradient method. In Step (i), the subdifferential of the objective function at the current estimate is calculated. Since this subdifferential possibly contains multiple values, we choose a subgradient from the subdifferential in Step (ii). After that, a step size is selected in Step (iii). In Step (iv), a subgradient update is performed. Lastly, we update the objective function with the new data point in Step (v). The main feature of this algorithm is that we do not apply the subgradient algorithm to solve the current optimization problem (NSE), and instead, we take one iteration to update our estimate of the matrix value at time $t$. Then, after a new measurement is taken, the optimization problem at time $t+1$ changes since an additional term is added to its objective function. Thus, as we update our estimates, the optimization problems also change over time. We analyze the convergence of Algorithm 1 under different step size selection rules and show that although the objective function changes at each iteration, our algorithm convergences to the $\bar{\mathbf{A}}$.

**Algorithm 1** Subgradient Method for (NSE)

Initialize $\hat{\mathbf{A}}^{(0)}$ randomly
**for** $t = 0, \ldots, T-1$ **do**
(i) Find $\partial f_t(\hat{\mathbf{A}}^{(t)}) = -\sum_{\ell=0}^{t} \partial \left\| (\bar{\mathbf{A}} - \hat{\mathbf{A}}^{(t)})\mathbf{x}_\ell + \bar{\mathbf{d}}_\ell \right\|_2 \otimes \mathbf{x}_\ell$
(ii) Choose $\mathbf{G}_{\hat{\mathbf{A}}^{(t)},t}$ from $\partial f_t(\hat{\mathbf{A}}^{(t)})$
(iii) Choose $\beta^{(t)}$ using an appropriate step size rule
(iv) Update $\hat{\mathbf{A}}^{(t+1)} = \hat{\mathbf{A}}^{(t)} - \beta^{(t)}\mathbf{G}_{\hat{\mathbf{A}}^{(t)},t}$
(v) Update $f_{t+1}(\mathbf{A}) = f_t(\mathbf{A}) + \|\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t\|_2$
**end for**
Return $\hat{\mathbf{A}}^{(T)}$

## Subgradient Method with Best Step Size

We analyze the subgradient method using the best step size, which is defined as the step size that maximizes improvement at each iteration. However, computing this step size requires knowledge of the ground-truth matrix $\bar{\mathbf{A}}$, making it impractical in real-world applications. Nonetheless, studying this case offers valuable insights into the behavior of subgradient updates. Let the step size in Algorithm 1 be chosen as

$$\beta^{(t)} = \frac{\langle \mathbf{G}_{\hat{\mathbf{A}}^{(t)},t}, \hat{\mathbf{A}}^{(t)} - \bar{\mathbf{A}} \rangle}{\|\mathbf{G}_{\hat{\mathbf{A}}^{(t)},t}\|_F^2},$$

It is shown in Appendix 6.A that this is the best step size possible. According to Theorem 25, the distance to the ground-truth matrix decreases by a factor of $\sqrt{1-\cos^2(\hat{\theta}_t)}$ at each update with best step size. As long as this angle is not $\pm 90^\circ$ or $\pm 270^\circ$, Algorithm 1 progresses toward $\bar{\mathbf{A}}$.

**Theorem 25.** *Under the best step size for Algorithm 1, it holds that*

$$\|\hat{\mathbf{A}}^{(t+1)} - \bar{\mathbf{A}}\|_F \leq \sqrt{1-\cos^2(\hat{\theta}_t)}\|\hat{\mathbf{A}}^{(t)} - \bar{\mathbf{A}}\|_F,$$

*where $\hat{\theta}_T$ denotes the angle between $\hat{\mathbf{A}}^{(t)} - \bar{\mathbf{A}}$ and $\mathbf{G}_{\hat{\mathbf{A}}^{(t)},t}$.*

If $\hat{\theta}_t$ is obtuse, the step size will be negative, which must be avoided. To this end, we establish a positive lower bound on $\cos(\hat{\theta}_t)$ in Theorem 26 and show that this angle remains acute. To guarantee that, we define the burn-in period as the number of samples required for a theoretical exact recovery. After the burn-in period, $\bar{\mathbf{A}}$ is the unique optimal solution of $\min f_T(\mathbf{A})$ with high probability.

**Definition 13.** *Consider an arbitrary $\delta \in (0,1)$. Let $T^{(burn)}(n,p,\rho,\sigma,\delta)$ be defined as*

$$T^{(burn)} := \Theta\Bigg( \max\left\{ \frac{\sigma^{10}}{(1-\rho)^3(1-p)^2}, \frac{\sigma^4}{p(1-p)} \right\} \times \left[ n^2 \log\left( \frac{1}{(1-\rho)\sigma p(1-p)} \right) + \log\left(\frac{1}{\delta}\right) \right] \Bigg).$$

*For every $t \geq T^{(burn)}(n,p,\rho,\sigma,\delta)$, $\bar{\mathbf{A}}$ is the unique optimal solution to* (NSE) *with probability at least $1-\delta$.*

For convenience, we omit the argument $(n,p,\rho,\sigma,\delta)$ and denote the burn-in period as $T^{(burn)}$. We analyze the algorithm's behavior after the burn-in period because it happens that the iterates of Algorithm 1 are chaotic when $t$ is not large enough or equivalently when $\bar{\mathbf{A}}$ is not the unique solution of (NSE). The following theorem establishes that if the number of samples exceeds $T^{(burn)}$, the cosine of the angle can be lower-bounded by a positive term.

**Theorem 26.** *Define $\hat{\theta}_T$ to be the angle between $\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}$ and $\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}$, and let $\widetilde{\|\mathbf{x}_T\|_2} = \sum_{t=0}^{T-1}\|\mathbf{x}_t\|_2/T$ be the average norm of the states until time $T$. Suppose that Assumptions 12 and 13 hold. Then, as long as the sample complexity $T \geq T^{(burn)}(n,p,\rho,\sigma,\delta)$ and $\hat{\mathbf{A}}^{(T)} \neq \bar{\mathbf{A}}$, with probability at least $1-\delta$, the term $\cos(\hat{\theta}_T)$ can be lower-bounded as*

$$\cos(\hat{\theta}_T) \geq \Theta\left( \min\left\{ 1, \frac{p(1-p)}{\sigma^4 \widetilde{\|\mathbf{x}_T\|_2}} \right\} \right) > 0.$$

Theorem 26 provides a probabilistic lower bound on the cosine of the angle of interest when the sample complexity $T$ is sufficiently large. The algorithm iterates move toward the ground truth $\bar{\mathbf{A}}$ when the $\bar{\mathbf{A}}$ is the global solution of (NSE). In this lower bound, the average norm of the system states over time, $\widetilde{\|\mathbf{x}_T\|_2}$, appears as a random variable. This term is known to have an upper bound on the order of $p^{1/2}/\sigma^9(1-\rho)$ with high probability.

**Lemma 20** (Equation (63) in [127])**.** *Under the assumptions of Theorem 26, when $T$ is greater than $T^{(burn)}(n,p,\rho,\sigma,\delta)$, the inequality*

$$\widetilde{\|\mathbf{x}_T\|_2} \leq \Theta\left(\frac{p^{1/2}}{\sigma^9(1-\rho)}\right)$$

*holds with probability at least $1-\delta$*

In order to establish the result in Lemma 20, we select $\theta := \Theta(p(1-p)T/\sigma^4)$ and $\epsilon := \Theta(\sqrt{p(1-p)^2(1-\rho)^2\sigma^{10}})$ in Equation (63) of [127]. Combining this result with Theorem 26 leads to the main result of this subsection.

**Corollary 8.** *Let $D = \|\hat{\mathbf{A}}^{(T^{(burn)}(n,p,\rho,\sigma,\delta/2))} - \bar{\mathbf{A}}\|_F$ denote the distance of the solution of the last iterate to the ground-truth $\bar{\mathbf{A}}$ at the end of the burn-in time. Define $\gamma(p,\rho,\sigma)$ to be*

$$\gamma(p,\rho,\sigma) := \sqrt{1-\min\{1,\Theta(p(1-p)^2\sigma^{10}(1-\rho)^2)\}}.$$

*Then, given an arbitrary $\epsilon > 0$, Algorithm 1 under the best step size achieves $\|\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\|_F \leq \epsilon$ with probability $1-\delta$, when the number of iterations $T$ is sufficiently large such that $T > T^{(\epsilon)} := T^{(burn)}(n,p,\rho,\sigma,\delta/2) + T^{(conv)}(p,\rho,\sigma,D,\epsilon)$, where*

$$T^{(conv)}(p,\rho,\sigma,D,\epsilon) := \log\left(\frac{D}{\epsilon}\right)\log\left(\frac{1}{\gamma(p,\rho,\sigma)}\right).$$

Algorithm 1 achieves linear convergence with high probability after the burn-in period due to the $\log(D/\epsilon)$ term. Corollary 8 implies that to achieve an $\epsilon$-level estimation error, the total number of required samples should be the sum of those needed for the burn-in phase and the linear convergence phase. With high probability, the burn-in period scales as $\max\{(1-p)^{-2}, p^{-1}(1-p)^{-1}\}$ with respect to $p$ and as $n^2$ with respect to the system dimension. After the burn-in phase, the algorithm converges to the ground truth at a linear rate because $\cos(\hat{\theta}_T)$ is lower-bounded by a positive factor given by $\min\{1, p^{1/2}(1-p)\sigma^5(1-\rho)\}$. Consequently, the linear convergence is upper-bounded by $\gamma(p,\rho,\sigma)$, and the sample complexity scales logarithmically with the problem parameters. Therefore, the sample complexity of the burn-in period dominates the sample complexity of the linear convergence phase. Next, we generalize the results to more practical step size rules.

## Subgradient Method with Polyak Step Size

We analyze the Polyak step size. Unlike the best step size selection, the Polyak step size does not require knowledge of $\bar{\mathbf{A}}$. However, it requires information about the optimal objective value at

$\bar{\mathbf{A}}$ at time $T$, that is, $f_T(\bar{\mathbf{A}})$. Under this step size rule, we can establish a result similar to Corollary 8.

**Theorem 27.** *Consider Algorithm 1 under the Polyak step size defined as*

$$\beta^{(T)} = \frac{(f_T(\hat{\mathbf{A}}^{(T)}) - f_T(\bar{\mathbf{A}}))}{\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2}.$$

*Suppose that Assumptions 12 and 13 hold. When the number of samples $T$ is greater than $T^{(\epsilon)}$, Algorithm 1 achieves $\|\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\|_F \leq \epsilon$, where $T^{(\epsilon)}$ defined in Corollary 1.*

Although the Polyak step size relies on less information about the ground-truth system dynamics than the best step size (i.e., it needs to know an optimal scalar rather than an optimal matrix), it requires the same number of iterations as the best step size selection. These step sizes are implementable in practice (subject to some approximations) if some prior knowledge about the system dynamics is available.

## Subgradient Method with Constant and Diminishing Step Sizes

The step sizes discussed in previous sections require information about the ground-truth matrix $\bar{\mathbf{A}}$ and the optimal objective value $f_T(\bar{\mathbf{A}})$. While theoretically sound, these step sizes are impractical in real-world applications. Therefore, we explore the subgradient method with constant and diminishing step sizes. Unlike the best and Polyak step sizes, constant and diminishing step sizes do not guarantee descent at every iteration. Instead, they lead to convergence within a neighborhood of the true solution rather than exact recovery. Consequently, results concerning the solution gap, $\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}$, are less relevant in this setting. Nevertheless, we establish the asymptotic convergence of the best average sub-optimality gap over time, defined as:

$$\min_{t \in [T^{(burn)}(n,p,\rho,\sigma,\delta),T]} \left\{ \frac{1}{t}(f_t(\hat{\mathbf{A}}^{(t)}) - f_t(\bar{\mathbf{A}})) \right\}.$$

The best average sub-optimality gap represents the algorithm's performance over its entire horizon, from the burn-in period to the final iteration. We scale the sub-optimality gap by $1/t$ to ensure comparability across time steps, as the objective function $f_t(\mathbf{A})$ in (NSE) evolves with $t$ and consists of a sum of $t$ terms. An $\epsilon$-level estimation error at time $t$ , i.e. $\|\hat{\mathbf{A}}^{(t)} - \bar{\mathbf{A}}\|_F \leq \epsilon$, corresponds to a sub-optimality gap of $t\epsilon$. Furthermore, we focus on the best average sub-optimality gap rather than the gap at the final iteration, as iterates fluctuate around the neighborhood of $\bar{\mathbf{A}}$ due to the nature of these step sizes. These step sizes may potentially remain large, preventing exact convergence, or diminish too quickly or too slowly to achieve a precise limit. Additionally, we analyze the best average sub-optimality gap only after the burn-in period since $\bar{\mathbf{A}}$ may not be the global minimizer during this phase, leading to potential negative sub-optimality terms.

Next, we demonstrate that under a constant step size, the best average sub-optimality gap of the subgradient algorithm asymptotically converges at a rate of $T^{-1/2}$. When $T$ is significantly larger than $T^{(burn)}$, the upper bound for the best average sub-optimality gap can be simplified as $\Theta(Dp^{1/2}/\sigma^9(1-\rho)T^{1/2})$. Hence, it asymptotically converges to zero with the rate $T^{-1/2}$.

**Theorem 28.** *When the step size $\beta^{(t)}$ is set to the constant,*

$$\beta^{(t)} = \beta = \Theta\left(\frac{D\sigma^9(1-\rho)}{p^{1/2}(\sum_{t=T^{(burn)}}^{T} t^2)^{1/2}}\right),$$

*the best average sub-optimality gap for Algorithm 1 iterates can be bounded with probability at least $1-\delta$ as*

$$\min_{t\in[T^{(burn)},T]}\left\{\frac{1}{t}(f_t(\hat{\mathbf{A}}^{(t)}) - f_t(\bar{\mathbf{A}}))\right\} \leq \Theta\left(\frac{Dp^{1/2}}{\sigma^9(1-\rho)} \times \frac{(\sum_{t=T^{(burn)}}^{T} t^2)^{1/2}}{\sum_{t=T^{(burn)}}^{T} t}\right).$$

We establish a similar result for the diminishing step size.

**Theorem 29.** *Suppose that the step size $\beta^{(t)}$ takes the diminishing form $\beta^{(t)} = \beta/t$, where $\beta$ is defined as*

$$\beta := \Theta\left(\frac{D\sigma^9(1-\rho)}{p^{1/2}(T-T^{(burn)})^{1/2}}\right).$$

*Then, with probability at least $1-\delta$, the best average sub-optimality gap for Algorithm 1 iterates can be bounded as*

$$\min_{t\in[T^{(burn)},T]}\left\{\frac{1}{t}(f_t(\hat{\mathbf{A}}^{(t)}) - f_t(\bar{\mathbf{A}}))\right\} \leq \Theta\left(\frac{Dp^{1/2}}{\sigma^9(1-\rho)} \times \frac{1}{(T-T^{(burn)})^{1/2}}\right).$$

The proof of Theorem 29 is omitted as it follows the same reasoning as Theorem 28. The diminishing step size is chosen to satisfy $\sum \beta^{(t)} \to \infty$ and $\sum(\beta^{(t)})^2 < \infty$, ensuring sufficient exploration and stable convergence without excessive oscillations. Theorem 29 establishes that, under a diminishing step size, the best average sub-optimality gap asymptotically converges at a rate of $T^{-1/2}$.

## 6.3 Time Complexity Analysis

We compare the time complexities of the subgradient method and the exact solution of (NSE) using CVX. For the subgradient algorithm, each time period involves computing the subgradient, which has a per-iteration cost of $\mathcal{O}(Tn^2)$. Running the algorithm for $T$ iterations results in an overall complexity of $\mathcal{O}(T^2n^2)$. On the other hand, solving (NSE) using CVX reduces the problem to a second-order conic program (SOCP) with $T+n^2$ variables with $T$ second-order cone constraints (minimization of the sum of norms can be written as a SOCP [3]). This problem is formulated as

$$\begin{aligned}\min_{\substack{vec(\mathbf{A})\in\mathbb{R}^{n^2}\\ f_t\in\mathbb{R},\, t\in[T]}} \quad & \sum_{t=0}^{T} f_t \\ \text{s.t.} \quad & f_t \geq \|\mathbf{x}_{t+1} - (\mathbf{x}_t \otimes \mathbf{I}_n)vec(\mathbf{A})\|_2,\ t \in \{0,1,\ldots,T-1\}.\end{aligned} \tag{SOCP}$$

The CVX solver employs primal-dual interior point methods to solve (SOCP). The number of iterations required to reach an $\epsilon$-accurate solution is $\mathcal{O}(T^{1/2}\log(1/\epsilon))$ [114]. Since solving conic constraints has a worst-case complexity of $\mathcal{O}(n^6)$, the per time period complexity is $\mathcal{O}(T^{1/2}n^6)$, leading to an overall complexity of $\mathcal{O}(T^{3/2}n^6)$. We know that $T$ and $T^{(burn)}$ scale as $\Theta(n^2)$. Consequently, the per-time period complexity of the subgradient method and the exact SOCP solution are $\mathcal{O}(n^4)$ and $\mathcal{O}(n^7)$, respectively. This highlights that solving (SOCP) requires significantly more computational power than the subgradient approach. Next, we present numerical experiments to validate the convergence results.

## 6.4 Numerical Experiments

We generate a system with the ground-truth matrix $\bar{\mathbf{A}}$ whose singular values are uniform in $(0, 1)$. Given the probability $p$, disturbance vectors are set to zero with probability $1 - p$. With probability $p$, the disturbance $\bar{\mathbf{d}}_t$ is defined as $\bar{\mathbf{d}}_t := \ell_t \hat{\mathbf{d}}_t$. Then, we sample $\ell_t \sim \mathcal{N}(0, \sigma_t^2)$, where $\sigma_t^2 := \min\{\|\mathbf{x}_t\|_2^2, 1/n\}$, and we sample $\hat{\mathbf{d}}_t \sim \text{uniform}(\mathbb{S}^{n-1})$. This disturbance vector conforms to Assumption 13, and the values of the disturbance vectors are correlated over time. We generate $10$ independent trajectories of the subgradient algorithm using $10$ different systems. Then, we report the average solution gap and the average loss gap of Algorithm 1 at each period. The solution gap and the loss gap for Algorithm 1 at time $T$ are $\|\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\|_F$ and $f_T(\hat{\mathbf{A}}^{(T)}) - f_T(\bar{\mathbf{A}})$, respectively.

Algorithm 1 is experimented with different step size rules. The system has order $n = 5$, and $p = 0.7$. In addition to the best, Polyak, diminishing, and constant step sizes, we also introduce the backtracking step size. We initialize the step size at a constant value, and it is shrunk by constant factor $\gamma \in (0, 1)$ until we achieve a sufficiently large descent. See Armijo step size in [4]. Figure 6.1 depicts the performance of Algorithm 1 under these different step sizes. Best and Polyak step sizes achieve fast and linear convergence as proven theoretically. Moreover, the backtracking step size performs similarly to the best and Polyak step sizes without using the knowledge of the ground truth $\bar{\mathbf{A}}$ or the optimal objective value $f_T(\bar{\mathbf{A}})$. Thus, the theoretical analysis of the backtracking step size is an interesting future direction. Furthermore, despite the slower rates, the diminishing and constant step sizes converged to the neighborhood of the solution.

We test the performance of Algorithm 1 with different system dimensions $n$ and probabilistic sparsity model parameters $p$. We choose backtracking and best step sizes to test the impact of different dimensions of the system states, $n \in \{5, 10, 15\}$ on the solution gap and loss gap when $p$ is set to $0.7$. Figure 6.2 shows that as $n$ grows, the number of samples needed for convergence grows as $n^2$. This aligns with the theoretical results from earlier sections. Moreover, the backtracking step size performs similarly to the best step size. Likewise, Figure 6.3 depicts the performance of the subgradient algorithm for the different probabilities of non-zero disturbance $p \in \{0.5, 0.7, 0.8\}$ with $n = 5$ using best and backtracking step sizes. As $p$ increases, the convergence for the estimator needs more samples, as supported by the theoretical results.

We experimented with the subgradient selection in Step (ii) of Algorithm 1. Whenever the $\ell_2$ norm is less than $10^{-5}$, we consider its subdifferential as the unit ball. We select the subgradient randomly on the unit sphere for random subgradient selection. For the zero subgradients, we set

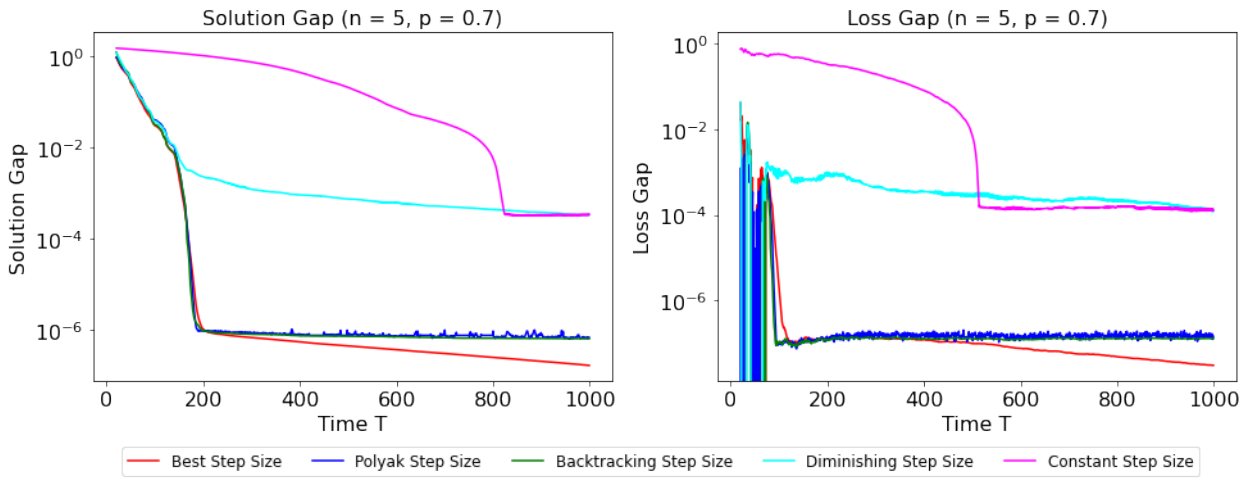


Figure 6.1: Solution Gap and Loss Gap of Algorithm 1 with Different Step Sizes for $n = 5$, and $p = 0.7$

the subgradient to zero. Lastly, we chose the minimum norm subgradient among all the available ones in the subdifferential. Using the best step size, we test these selection criteria for a system with order $n = 10$ and $p = 0.7$. Figure 6.4 demonstrates that there is not a significant difference in terms of solution gap. Last, we provide simulation results using the best step size for large systems of order $n \in \{25, 50, 75\}$. We run these experiments for $T = 10000$. This is not possible with the CVX solver due to the high computational complexity of (SOCP). Nevertheless, the subgradient algorithm provides an accurate estimation for large systems within a reasonable time frame, as seen in Figure 6.5.

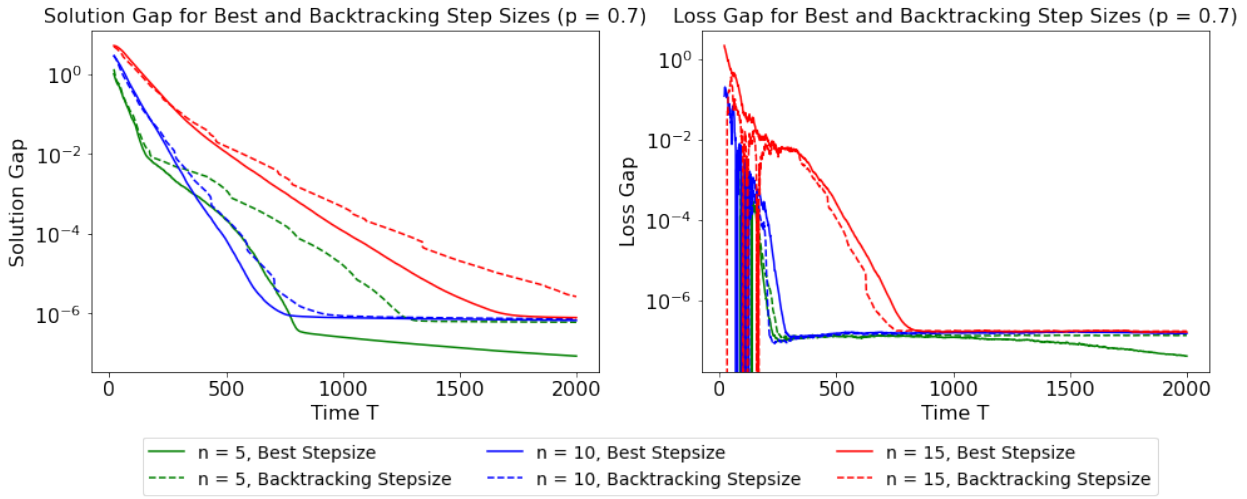


Figure 6.2: Solution Gap and Loss Gap of Algorithm 1 with $p = 0.7$, Best and Backtracking Step Sizes, and $n \in \{5, 10, 15\}$

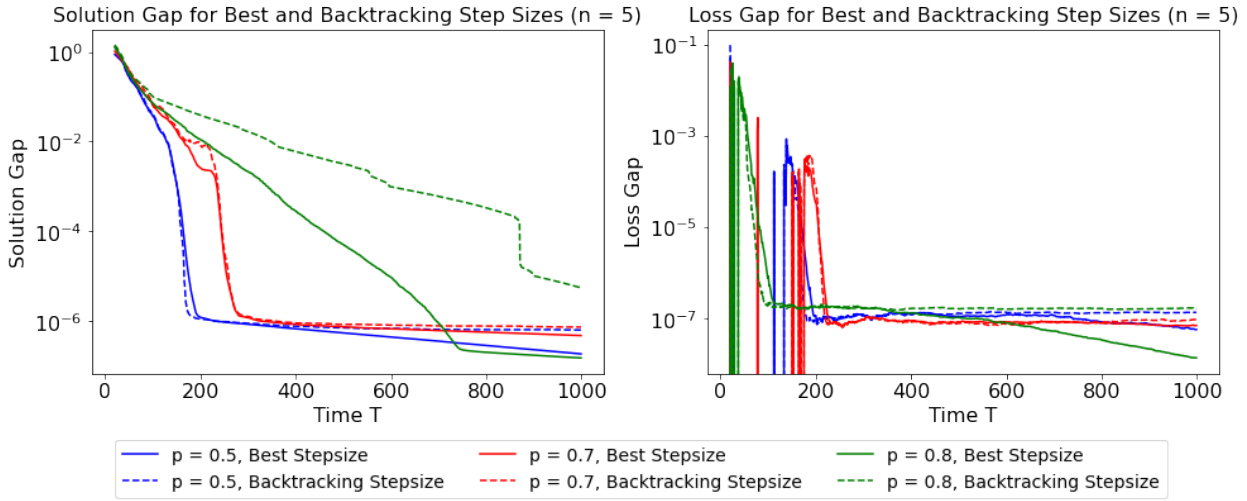


Figure 6.3: Solution Gap and Loss Gap of Algorithm 1 with $n = 5$, Best and Backtracking Step Sizes, and $p \in \{0.5, 0.7, 0.8\}$

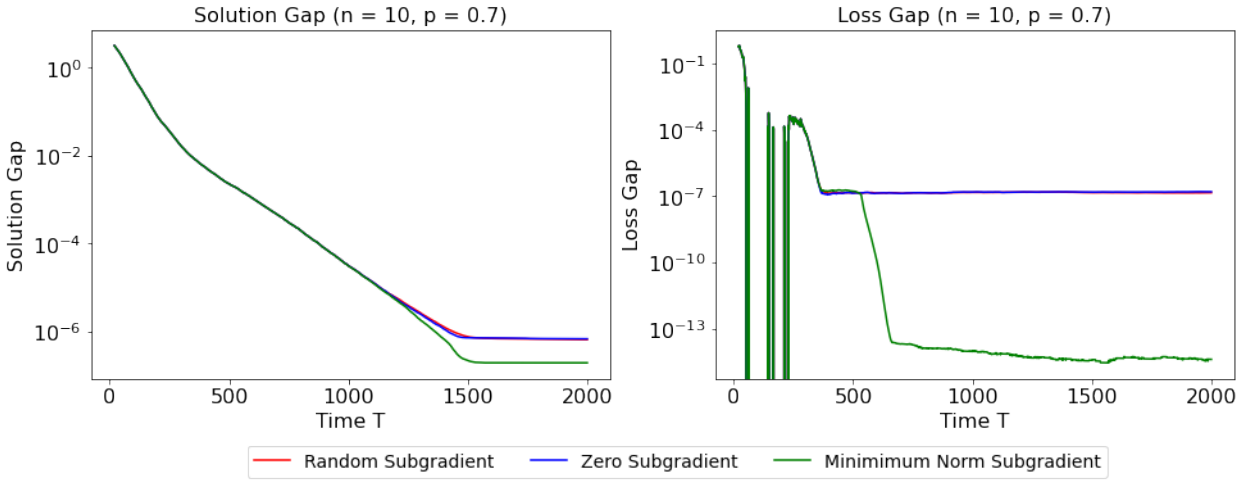


Figure 6.4: Solution Gap and Loss Gap of Algorithm 1 with Different Subgradient Selection Methods for $n = 10$ and $p = 0.7$

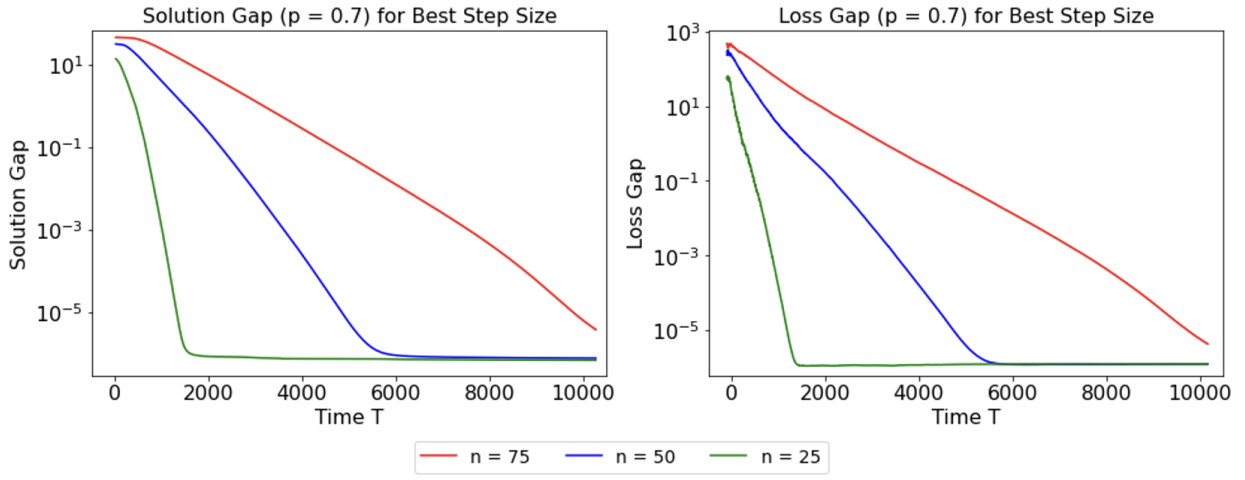


Figure 6.5: Solution Gap and Loss Gap of Algorithm 1 with $p = 0.7$, Best Step Size, and $n \in \{50, 75, 100\}$

# Appendices

## 6.A Proof of Theorem 25

*Proof.* Using the subgradient update iteration at time $T$, $\hat{\mathbf{A}}^{(T+1)} = \hat{\mathbf{A}}^{(T)} - \beta^{(T)}\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}$, we have

$$\begin{aligned}\|\hat{\mathbf{A}}^{(T+1)} - \bar{\mathbf{A}}\|_F^2 &= \|(\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}) - \beta^{(T)}\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2 \\ &= \|\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\|_F^2 + (\beta^{(T)})^2\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2 - 2\beta^{(T)}\langle \mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}, \hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\rangle. \quad (6.1)\end{aligned}$$

Substituting the step size stated in the theorem yields

$$\begin{aligned}\|\hat{\mathbf{A}}^{(T+1)} - \bar{\mathbf{A}}\|_F^2 &= \|\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\|_F^2 - \frac{(\langle \mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}, \hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\rangle)^2}{\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2} \\ &\leq \sqrt{1 - \cos^2(\hat{\theta}_T)}\|\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\|_F.\end{aligned}$$

The last inequality uses the fact that $\langle \mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}, \hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\rangle = \|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F\|\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\|_F \cos(\hat{\theta}_T)$. □

## 6.B Proof of Theorem 26

*Proof.* Using Lemma 19, we can show that

$$\cos(\hat{\theta}_T) = \frac{\langle \hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}, \mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\rangle}{\|\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\|_F\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F} \geq \frac{f_T(\hat{\mathbf{A}}^{(T)}) - f_T(\bar{\mathbf{A}})}{\|\bar{\mathbf{A}} - \hat{\mathbf{A}}^{(T)}\|_F\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F}.$$

We lower bound the difference between the objectives as

$$\begin{aligned}f_T(\hat{\mathbf{A}}^{(T)}) - f_T(\bar{\mathbf{A}}) &\geq \sup_{\mathbf{G}_{\bar{\mathbf{A}},T} \in \partial f_T(\bar{\mathbf{A}})} \left\{ \langle \hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}, \mathbf{G}_{\bar{\mathbf{A}},T}\rangle \right\} \\ &= \sup_{\|e_t\| \leq 1, t \notin \mathcal{K}} \left\{ \left\langle \hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}, -\sum_{t \in \mathcal{K}} \hat{\mathbf{d}}_t \otimes \mathbf{x}_t - \sum_{t \notin \mathcal{K}} e_t \otimes \mathbf{x}_t \right\rangle \right\} \\ &= \sum_{t \in \mathcal{K}} \left\langle (\bar{\mathbf{A}} - \hat{\mathbf{A}}^{(T)})\mathbf{x}_t, \frac{\bar{\mathbf{d}}_t}{\|\bar{\mathbf{d}}_t\|_2} \right\rangle + \sum_{t \notin K} \|(\bar{\mathbf{A}} - \hat{\mathbf{A}}^{(T)})\mathbf{x}_t\|_2.\end{aligned}$$

The term on the right-hand side is equivalent to the condition in Corollary 3 in [127] with $f(\mathbf{x}_t) = \mathbf{x}_t$ and $Z = \bar{\mathbf{A}} - \hat{\mathbf{A}}^{(T)}$.

**Lemma 21** (Theorem 6 in [127])**.** *Suppose the Assumptions 12 and 13 are satisfied. Then, for any given* $\mathbf{B} \in \mathbb{R}^{n\times n}$ *, the following holds with probability at least* $1-\delta$*:*

$$\sum_{t\in\mathcal{K}} \left\langle \mathbf{B} - \bar{\mathbf{A}}, \frac{\bar{\mathbf{d}}_t}{\|\bar{\mathbf{d}}_t\|_2} \right\rangle + \sum_{t\notin K} \|(\mathbf{B} - \bar{\mathbf{A}})\mathbf{x}_t\|_2 \geq \Theta\left[\frac{cp(1-p)\|\mathbf{B} - \bar{\mathbf{A}}\|_F T}{\sigma^4}\right]$$

*when the number of samples* $T \geq T^{(burn)}(n, p, \rho, \sigma, \delta)$*.*

Using Lemma 21 and $\mathbf{B} = \hat{\mathbf{A}}^{(T)}$, if $T \geq T^{(1)}(\delta)$, we have

$$f_T(\hat{\mathbf{A}}^{(T)}) - f_T(\bar{\mathbf{A}}) \geq \frac{cp(1-p)\|\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\|_F T}{\sigma^4} \tag{6.2}$$

with probability at least $1-\delta$. Next, we upper-bound the norm of the subgradient as below:

$$\begin{aligned}
\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2 = \Big\langle & -\sum_{t=0}^{T} \partial \left\| (\bar{\mathbf{A}} - \hat{\mathbf{A}}^{(T)})\mathbf{x}_t + \bar{\mathbf{d}}_t \right\|_2 \otimes \mathbf{x}_t, \\
& -\sum_{t=0}^{T} \partial \left\| (\bar{\mathbf{A}} - \hat{\mathbf{A}}^{(T)})\mathbf{x}_t + \bar{\mathbf{d}}_t \right\|_2 \otimes \mathbf{x}_t \Big\rangle \\
& \leq \sum_{t=0}^{T}\sum_{m=0}^{T} \|\mathbf{x}_t\|_2 \|\mathbf{x}_m\|_2 = \left(\sum_{t=0}^{T-1} \|\mathbf{x}_t\|_2\right)^2 .
\end{aligned} \tag{6.3}$$

The second-to-last inequality holds due to the norm of subgradients being at most one and the Cauchy-Schwarz inequality. We simplify the expression as $\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F \leq T\widetilde{\|\mathbf{x}_T\|_2}$. We combine (6.2) and (6.3) with the lower bound on $\cos(\hat{\theta}_T)$. With probability at least $1-\delta$, when the number of samples is greater than $T \geq T^{(burn)}(n, p, \rho, \sigma, \delta)$, we have

$$\cos(\hat{\theta}_T) \geq \Theta\left(\min\left\{1, \frac{p(1-p)}{\sigma^4 \widetilde{\|\mathbf{x}_T\|_2}}\right\}\right) > 0.$$

□

# 6.C Proof of Theorem 27

*Proof.* Using the subgradient update iteration at time $T$, $\hat{\mathbf{A}}^{(T+1)} = \hat{\mathbf{A}}^{(T)} - \beta^{(T)}\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}$, we have

$$\begin{aligned}
\|\hat{\mathbf{A}}^{(T+1)} - \bar{\mathbf{A}}\|_F^2 &= \|(\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}) - \beta^{(T)}\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2 \\
&= \|(\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}})\|_F^2 + (\beta^{(T)})^2\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2 - 2\beta^{(T)}\langle \mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}, \hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}}\rangle \\
&\leq \|(\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}})\|_F^2 + (\beta^{(T)})^2\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2 - 2\beta^{(T)}(f_T(\hat{\mathbf{A}}^{(T)}) - f_T(\bar{\mathbf{A}})).
\end{aligned} \tag{6.4}$$

The last inequality uses the fact that $(f_T(\hat{\mathbf{A}}^{(T)}) - f_T(\bar{\mathbf{A}})) \geq \langle \mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}, \hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}} \rangle$ due to the properties of subgradient. Substituting the step size stated in the theorem yields

$$\|\hat{\mathbf{A}}^{(T+1)} - \bar{\mathbf{A}}\|_F^2 \leq \|(\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}})\|_F^2 - \frac{(f_T(\hat{\mathbf{A}}^{(T)}) - f_T(\bar{\mathbf{A}}))^2}{\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2}.$$

In Theorem 26 and Corollary 8, it is proved that whenever the number of samples is greater than $T^{(burn)}(n, p, \rho, \sigma, \delta)$, we have

$$\frac{(f_T(\hat{\mathbf{A}}^{(T)}) - f_T(\bar{\mathbf{A}}))^2}{\|\mathbf{G}_{\hat{\mathbf{A}}^{(T)},T}\|_F^2} \leq (1-\gamma^2)\|(\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}})\|_F^2$$

with probability at least $1-\delta$ where $\gamma$ is a positive constant that is less than 1. Thus, we have convergence to the ground truth with the linear rate $\gamma$ because

$$\|\hat{\mathbf{A}}^{(T+1)} - \bar{\mathbf{A}}\|_F \leq \gamma\|(\hat{\mathbf{A}}^{(T)} - \bar{\mathbf{A}})\|_F.$$

The rest of the arguments are similar to those in Corollary 8. □

## 6.D Proof of Theorem 28

*Proof.* We first define the following Lemma.

**Lemma 22.** *Let $D = \|\hat{\mathbf{A}}^{(T^{(burn)})} - \bar{\mathbf{A}}\|_F$ be the distance of the iterate to the ground-truth $\bar{\mathbf{A}}$ at the end of the burn-in time. Then, the following holds:*

$$\sum_{t=T^{(burn)}}^{T} \beta^{(t)}(f_t(\hat{\mathbf{A}}^{(t)}) - f_t(\bar{\mathbf{A}})) \leq D^2 + \sum_{t=T^{(burn)}}^{T} (\beta^{(t)})^2 \left(\sum_{\ell=0}^{t} \|\mathbf{x}_\ell\|_2\right)^2.$$

Using Lemma 22 and taking the minimum of the left-hand side over periods $[T^{(burn)}, T]$ yields

$$\min_{t\in[T^{(burn)},T]} \left\{\frac{1}{t}(f_t(\hat{\mathbf{A}}^{(t)}) - f_t(\bar{\mathbf{A}}))\right\} \leq \frac{D^2 + \sum_{t=T^{(burn)}}^{T}(\beta^{(t)})^2 \left(\sum_{\ell=0}^{t} \|\mathbf{x}_\ell\|_2\right)^2}{2\sum_{t=T^{(burn)}}^{T}(\beta^{(t)}t)}.$$

Using Lemma 20, we know that $\left(\sum_{\ell=0}^{t} \|\mathbf{x}_\ell\|_2\right)$ can be upper-bounded by $\Theta(tp^{1/2}/\sigma^9(1-\rho))$ with probability at least $1-\delta$ whenever $t$ is larger than the burn-in time. We have

$$\min_{t\in[T^{(burn)},T]} \left\{\frac{1}{t}(f_t(\hat{\mathbf{A}}^{(t)}) - f_t(\bar{\mathbf{A}}))\right\} \leq \Theta\left(\frac{D^2 + \sum_{t=T^{(burn)}}^{T}(\beta^{(t)})^2 t^2 p/\sigma^{18}(1-\rho)^2}{\sum_{t=T^{(burn)}}^{T}(\beta^{(t)}t)}\right)$$

with probability $1-\delta$. Setting the step size to be $\beta^{(t)} = \beta$ implies the term on the right-hand side to be

$$\Theta\left(\frac{D^2 + \frac{\beta^2 p}{\sigma^{18}(1-\rho)^2}\sum_{t=T^{(burn)}}^{T} t^2}{\beta\sum_{t=T^{(burn)}}^{T} t}\right).$$

Minimizing the right-hand side with respect to $\beta$ results in the following constant step size:

$$\beta^{(t)} = \beta = \Theta\left(\frac{D\sigma^9(1-\rho)}{p^{1/2}(\sum_{t=T^{(burn)}}^{T} t^2)^{1/2}}\right).$$

Substituting this constant step size to the expression on the sub-optimality gap inequality provides the simplified upper bound that holds with probability at least $1-\delta$:

$$\min_{t\in[T^{(burn)},T]}\left\{\frac{1}{t}(f_t(\hat{\mathbf{A}}^{(t)}) - f_t(\bar{\mathbf{A}}))\right\} \leq \Theta\left(\frac{Dp^{1/2}}{\sigma^9(1-\rho)} \times \frac{(\sum_{t=T^{(burn)}}^{T} t^2)^{1/2}}{\sum_{t=T^{(burn)}}^{T} t}\right).$$

□

# 6.E Proof of Lemma 22

*Proof.* Using (6.4) and the iterative substitution between the time periods from $T^{(burn)}(n,p,\rho,\sigma,\delta)$ to $T$, we obtain

$$\begin{aligned}\|\hat{\mathbf{A}}^{(T+1)} - \bar{\mathbf{A}}\|_F^2 &\leq \|(\hat{\mathbf{A}}^{(T^{(burn)})} - \bar{\mathbf{A}})\|_F^2 + \sum_{t=T^{(burn)}}^{T} (\beta^{(t)})^2 \|\mathbf{G}_{\hat{\mathbf{A}}^{(t)},t}\|_F^2 \\ &\qquad - 2\sum_{t=T^{(burn)}}^{T} \beta^{(t)}(f_t(\hat{\mathbf{A}}^{(t)}) - f_t(\bar{\mathbf{A}})) \\ &\leq D^2 + \sum_{t=T^{(burn)}}^{T} (\beta^{(t)})^2 \left(\sum_{\ell=0}^{t} \|\mathbf{x}_\ell\|_2\right)^2 - 2\sum_{t=T^{(burn)}}^{T} \beta^{(t)}(f_t(\hat{\mathbf{A}}^{(t)}) - f_t(\bar{\mathbf{A}})).\end{aligned}$$

The last inequality follows from the fact that $D = \|\hat{\mathbf{A}}^{(T^{(burn)})} - \bar{\mathbf{A}}\|_F$ and $\|\mathbf{G}_{\hat{\mathbf{A}}^{(t)},t}\|_F$ can be upper-bounded by $\left(\sum_{\ell=0}^{t} \|\mathbf{x}_\ell\|_2\right)$. Since $\|\hat{\mathbf{A}}^{(T+1)} - \bar{\mathbf{A}}\|_F^2$ is non-negative, the last term on the right-hand side can be lower-bounded by 0. Switching the term with the negative coefficient from the right-hand side to the left-hand side completes the proof. □

# Chapter 7

# Conclusion and Future Direction

This dissertation explored the landscape of optimization problems involving matrix variables and develops scalable solution approaches. These problems present significant challenges due to factors such as limited observations, the non-convexity of the feasible set and objective function, and the non-smoothness of the objective function. Addressing these complexities through optimization and statistical learning techniques is essential for achieving tractable and accurate solutions in machine learning and control applications. In this dissertation, we tackled these challenges in the context of low-rank matrix optimization and system identification under adversarial disturbances. The following sections summarize the key contributions of each chapter and outline promising directions for future research.

## 7.1 Low-Rank Matrix Optimization

In Chapter 2, we provided a negative answer to the question of whether the B-M factorization approach can capture the benign properties of low-complexity MC problem instances. More specifically, we defined a class of MC problem instances that could be solved in polynomial time. We showed that there exist MC problem instances in this class that have exponentially many spurious local minima in the B-M factorization formulation. The results hold for a general class of loss functions, including the commonly used regularized formulation. In addition, for the rank-$1$ case, we proved that gradient-based methods fail with high probability for such instances. Numerical results verify that similar behaviors also hold for higher rank cases. These results imply that the optimization complexity of methods based on the factorized problem is not aligned with the information-theoretic complexity of the MC problem. Furthermore, we derived a complexity metric that potentially captures the complexity of the B-M factorization formulation.

In Chapter 3, we conducted a comparison between two main approaches to matrix completion and matrix sensing problems: a convex relaxation that gives an SDP formulation and the B-M factorization method. It is well-known that both of these methods enjoy mathematical guarantees for the recovery of the ground truth matrix whenever the RIP assumption is satisfied with a sufficiently small $\delta$. We offered the first result in the literature that compares these two methods whenever the

RIP condition is not satisfied or only satisfied with a large constant. We discovered classes of problems for which B-M factorization fails while the SDP recovers the ground truth matrix. The fact that specialized SDP algorithms have improved in recent years and can compete with simple first-order descent algorithms inspired us to investigate sharper bounds on sufficient conditions for the SDP formulation. We provided RIP bounds for the SDP formulation that depend on the rank of the solution and are automatically satisfied for high-rank problems, unlike the B-M method. On the other hand, when the number of measurements from the ground truth matrix is not high, we showed that SDP fails drastically while the B-M method does not contain any spurious solutions on its optimization landscape. As a result, we concluded that none of the methods outperforms the other one whenever the sufficiency guarantees are not met. The parameters of the problem, such as dimension, rank, and linear measurement operator, determine which solution method performs better. Consequently, it is prudent to apply both solution methods in case the RIP and incoherence are not satisfied.

Although sufficient conditions for learning low-rank matrices have been studied, the literature still lacks effective metrics for assessing the complexity of solution methods based on the structure of the measurements. A promising direction for future research is the development of a new complexity metric that quantifies the difficulty of solving matrix sensing problems using measurement properties. Such a metric could provide deeper insights by capturing not only sufficient but also necessary conditions for exact recovery. Additionally, further investigation into SDP formulations could refine the bounds established in Chapter 3. With recent advancements in SDP solvers, it is crucial to explore problem structures that can mitigate the limitations of the B-M factorization approach. Understanding these structures could lead to more efficient recovery guarantees and improved computational performance in practical applications.

## 7.2 System Identification under Adversarial Disturbances

In Chapter 4, 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 $\Delta$ 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 numerical experiments. We provided the first set of mathematical guarantees for the robust non-asymptotic analysis of dynamic systems.

In Chapter 5, we addressed the problem of parameterized non-linear system identification in the presence of *sparse-but-large* disturbances or *semi-oblivious attacks*. The non-smooth estimator is utilized to achieve the exact recovery of the underlying parameter $\bar{\mathbf{A}}$. First, we established necessary and sufficient conditions for the well-specifiedness of the estimator and the uniqueness of optimal solutions to the embedded optimization problem. Subsequently, we derived sample complexity bounds for the exact recovery of the system dynamics under two different conditions:

bounded basis functions and Lipschitz basis functions. For bounded basis functions, the sample complexity scales as $m^3n$ in terms of the problem dimension and as $p^{-1}(1-p)^{-2}$ in terms of the non-zero disturbance probability, up to a logarithm factor. For Lipschitz basis functions, the sample complexity scales as $mn$ in terms of the problem dimension and as $\max\{(1-p)^{-2}, p^{-1}(1-p)^{-1}\}$ in terms of the non-zero disturbance probability, up to a logarithm factor. Furthermore, we established that if the sample complexity scales at an order smaller than $p^{-1}(1-p)^{-1}$, high-probability exact recovery is unattainable. This result implies that the term $p^{-1}(1-p)^{-1}$ in our bounds is fundamental and unavoidable. Finally, we validated our theoretical findings through numerical experiments.

In Chapter 6, we proposed and analyzed a subgradient algorithm for non-smooth system identification of linear-time invariant systems with disturbances following a probabilistic sparsity model. While the non-asymptotic theory for this setting is well-established in the literature, existing methods rely on solvers that use primal-dual interior-point methods for second-order conic programs, leading to high computational costs. We proposed a subgradient method that enables fast exact recovery to address this. We demonstrated that the subgradient method with the best and Polyak step sizes converges linearly to the ground-truth system dynamics after a burn-in period. Additionally, we showed that constant and diminishing step sizes lead to convergence within a neighborhood of the true solution, with the best average sub-optimality gap asymptotically approaching zero. Numerical simulations support these theoretical results.

Given the novelty of non-smooth estimators in system identification problems, there are numerous promising directions for future research. One key avenue is the analysis of sample complexity for non-smooth estimators in parametrized non-linear systems with input sequences, as well as in non-parametric non-linear systems. Understanding these complexities could provide deeper insights into estimator performance and its practical implications. Additionally, accurate estimation of system dynamics from a single initialization is crucial for model predictive control (MPC), where control inputs are designed based on the best available estimates. A particularly compelling future direction is the regret analysis of non-smooth estimators when used in MPC for safety-critical applications. Comparing the performance of non-smooth estimators against least-squares estimators in the context of regret analysis could offer valuable insights into their relative advantages and trade-offs. Moreover, our experiments highlight the practical effectiveness of backtracking step size in subgradient-based algorithm design, suggesting it as a promising alternative. However, its theoretical properties remain an open question, warranting further investigation. Another area for improvement lies in the computational efficiency of our current implementation, which computes the full subgradient at each iteration. Exploring a partial subgradient approach akin to stochastic gradient descent could significantly enhance time complexity while maintaining robust performance. These future directions not only build upon the foundational work in this dissertation but also open new pathways for advancing the theory and application of non-smooth estimators in system identification and control.

# Bibliography


[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. "Regret Bounds for the Adaptive Control of Linear Quadratic Systems". In: *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings. 2011, pp. 1–26.

[2] Anil Alan, Andrew J. Taylor, Chaozhe R. He, A. Ames, and Gábor Orosz. "Control Barrier Functions and Input-to-State Safety With Application to Automated Vehicles". In: *IEEE Transactions on Control Systems Technology* 31 (2022), pp. 2744–2759. URL: https://api.semanticscholar.org/CorpusID:249461776.

[3] Farid Alizadeh and Donald Goldfarb. "Second-Order Cone Programming". In: *Mathematical Programming* 95.1 (2003), pp. 3–51.

[4] Larry Armijo. "Minimization of Functions Having Lipschitz Continuous First Partial Derivatives". In: *Pacific Journal of Mathematics* 16.1 (1966), pp. 1–3.

[5] Laurent Bako. "On a Class of Optimization-Based Robust Estimators". In: *IEEE Transactions on Automatic Control* 62.11 (Nov. 2017), pp. 5990–5997. ISSN: 0018-9286, 1558-2523. DOI: 10.1109/TAC.2017.2703308.

[6] Laurent Bako and Henrik Ohlsson. "Analysis of a Nonsmooth Optimization Approach to Robust Estimation". In: *Automatica* 66 (Apr. 2016), pp. 132–145. ISSN: 00051098. DOI: 10.1016/j.automatica.2015.12.024.

[7] Dietmar Bauer, Manfred Deistler, and Wolfgang Scherrer. "Consistency and Asymptotic Normality of Some Subspace Algorithms for Systems without Observed Inputs". In: *Automatica* 35.7 (1999), pp. 1243–1254.

[8] Dimitri P Bertsekas. "Nonlinear Programming". In: *Journal of the Operational Research Society* 48.3 (1997), pp. 334–334.

[9] Dimitris Bertsimas and Martin S. Copenhaver. "Characterization of the Equivalence of Robustification and Regularization in Linear and Matrix Regression". In: *European Journal of Operational Research* 270.3 (Nov. 2018), pp. 931–942. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2017.03.051.

[10] Srinadh Bhojanapalli and Prateek Jain. "Universal Matrix Completion". In: *International Conference on Machine Learning*. PMLR. 2014, pp. 1881–1889.

[11] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. "Dropping Convexity for Faster Semi-Definite Optimization". In: *Conference on Learning Theory*. PMLR. 2016, pp. 530–582.

[12] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. "Global Optimality of Local Search for Low Rank Matrix Recovery". In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016.

[13] Yingjie Bi and Javad Lavaei. "On the Absence of Spurious Local Minima in Nonlinear Low-Rank Matrix Recovery Problems". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 379–387.

[14] Yingjie Bi, Haixiang Zhang, and Javad Lavaei. "Local and global linear convergence of general low-rank matrix recovery problems". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 9. 2022, pp. 10129–10137.

[15] Nicolas Boumal. "Nonconvex Phase Synchronization". In: *SIAM Journal on Optimization* 26.4 (2016), pp. 2355–2377.

[16] Stephen P Boyd and Lieven Vandenberghe. "Convex Optimization". Cambridge University Press, 2004.

[17] Samuel Burer and Renato DC Monteiro. "A Nonlinear Programming Algorithm for Solving Semidefinite Programs via Low-Rank Factorization". In: *Mathematical Programming* 95.2 (2003), pp. 329–357.

[18] T Tony Cai and Anru Zhang. "Sharp RIP Bound for Sparse Signal and Low-Rank Matrix Recovery". In: *Applied and Computational Harmonic Analysis* 35.1 (2013), pp. 74–93.

[19] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. "Phase Retrieval via Matrix Completion". In: *SIAM Review* 57.2 (2015), pp. 225–251.

[20] Emmanuel J Candes and Paige A Randall. "Highly Robust Error Correction by Convex Programming". In: *IEEE Transactions on Information Theory* 54.7 (2008), pp. 2829–2840.

[21] Emmanuel J Candes and Terence Tao. "Decoding by Linear Programming". In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215.

[22] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. "Robust Principal Component Analysis?" In: *Journal of the ACM (JACM)* 58.3 (2011), pp. 1–37.

[23] Emmanuel J Candès and Benjamin Recht. "Exact Matrix Completion via Convex Optimization". In: *Foundations of Computational Mathematics* 9.6 (2009), pp. 717–772.

[24] Emmanuel J Candès and Terence Tao. "The Power of Convex Relaxation: Near-optimal Matrix Completion". In: *IEEE Transactions on Information Theory* 56.5 (2010), pp. 2053–2080.

[25] Emmanuel J. Candès and Yaniv Plan. "Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements". In: *IEEE Transactions on Information Theory* 57.4 (Apr. 2011), pp. 2342–2359.

[26] Han-Fu Chen and Lei Guo. "Identification and Stochastic Adaptive Control". Springer Science & Business Media, 2012.

[27] Ji Chen, Dekai Liu, and Xiaodong Li. "Nonconvex Rectangular Matrix Completion via Gradient Descent without $\ell_{2,\infty}$ Regularization". In: *IEEE Transactions on Information Theory* 66.9 (2020), pp. 5806–5841.

[28] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. "Bridging Convex and Nonconvex Optimization in Robust PCA: Noise, Outliers, and Missing Data". In: *Annals of Statistics* 49.5 (2021), p. 2948.

[29] Yuejie Chi, Yue M Lu, and Yuxin Chen. "Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview". In: *IEEE Transactions on Signal Processing* 67.20 (2019), pp. 5239–5269.

[30] Mark A Davenport, Marco F Duarte, Yonina C Eldar, and Gitta Kutyniok. "Introduction to Compressed Sensing." 2012.

[31] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. "1-Bit Matrix Completion". In: *Information and Inference: A Journal of the IMA* 3.3 (2014), pp. 189–223.

[32] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. "On the Sample Complexity of the Linear Quadratic Regulator". In: *Foundations of Computational Mathematics* 20.4 (2020), pp. 633–679. DOI: 10.1007/s10208-019-09426-y.

[33] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. "On the Sample Complexity of the Linear Quadratic Regulator". In: *Foundations of Computational Mathematics* 20.4 (2020), pp. 633–679.

[34] Sarah Dean, Stephen Tu, Nikolai Matni, and Benjamin Recht. "Safely Learning to Control the Constrained Linear Quadratic Regulator". In: *2019 American Control Conference (ACC)*. IEEE. 2019, pp. 5582–5588.

[35] David L Donoho. "High-Dimensional Centrally Symmetric Polytopes with Neighborliness Proportional to Dimension". In: *Discrete & Computational Geometry* 35 (2006), pp. 617–652.

[36] David L Donoho and Jared Tanner. "Neighborliness of Randomly Projected Simplices in High Dimensions". In: *Proceedings of the National Academy of Sciences* 102.27 (2005), pp. 9452–9457.

[37] M Eisenberg-Nagy, M Laurent, and A Varvitsiotis. "Complexity of the Positive Semidefinite Matrix Completion Problem with a Rank Constraint." In: *Springer, Fields Institute Communications, vol. 69* (2013).

[38] Yonina C Eldar and Gitta Kutyniok. "Compressed Sensing: Theory and Applications". Cambridge University Press, 2012.

[39] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. "Finite Time Identification in Unstable Linear Systems". In: *Automatica* 96 (2018), pp. 342–353.

[40] Julius Farkas. “Theorie der Einfachen Ungleichungen.” In: *Journal für die Reine und Angewandte Mathematik* 124 (1902), pp. 1–27. URL: http://eudml.org/doc/149129.

[41] 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)*. IEEE. 2019, pp. 2682–2689.

[42] Salar Fattahi and Somayeh Sojoudi. “Exact Guarantees on the Absence of Spurious Local Minima for Non-Negative Rank-1 Robust Principal Component Analysis”. In: *Journal of Machine Learning Research* (2020).

[43] Han Feng and Javad Lavaei. “Learning of Dynamical Systems under Adversarial Attacks”. In: *2021 60th IEEE Conference on Decision and Control (CDC)*. 2021, pp. 3010–3017. DOI: 10.1109/CDC45484.2021.9683149.

[44] Han Feng, Baturalp Yalcin, and Javad Lavaei. “Learning of Dynamical Systems under Adversarial Attacks-Null Space Property Perspective”. In: *2023 American Control Conference (ACC)*. IEEE. 2023, pp. 4179–4184.

[45] Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. “Learning Nonlinear Dynamical Systems from a Single Trajectory”. In: *Learning for Dynamics and Control*. PMLR. 2020, pp. 851–861.

[46] Simon Foucart and Holger Rauhut. “A Mathematical Introduction to Compressive Sensing”. Birkhäuser Basel, 2013. ISBN: 0817649476.

[47] Rong Ge, Chi Jin, and Yi Zheng. “No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1233–1242.

[48] Rong Ge, Jason D Lee, and Tengyu Ma. “Matrix Completion Has no Spurious Local Minimum”. In: *Advances in Neural Information Processing Systems* 29 (2016).

[49] Michael Grant and Stephen Boyd. “CVX: Matlab Software for Disciplined Convex Programming, version 2.1”. https://cvxr.com/cvx. Mar. 2014.

[50] Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. “An Equivalence Between Critical Points for Rank Constraints Versus Low-Rank Factorizations”. In: *SIAM Journal on Optimization* 30.4 (2020), pp. 2927–2955.

[51] Iman Hajizadeh, Mudassir Rashid, and Ali Cinar. “Integrating Compartment Models with Recursive System Identification”. In: *2018 Annual American Control Conference (ACC)*. 2018, pp. 3583–3588. DOI: 10.23919/ACC.2018.8431822.

[52] Frank Harary. “Graph Theory”. In: *CRC Press* (2018).

[53] Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. “Spectral Filtering for General Linear Dynamical Systems”. In: *Advances in Neural Information Processing Systems* 31 (2018).

[54] 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". In: *American Journal of Physiology-Endocrinology and Metabolism* 282.5 (2002), E992–E1007.

[55] Dong Huang, Ricardo Cabral, and Fernando De la Torre. "Robust Regression". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.2 (2016), pp. 363–375. DOI: 10.1109/TPAMI.2015.2448091.

[56] 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)*. IEEE. 2020, pp. 996–1001.

[57] Ming Jin, Javad Lavaei, Somayeh Sojoudi, and Ross Baldick. "Boundary Defense Against Cyber Threat for Power System State Estimation". In: *IEEE Transactions on Information Forensics and Security* 16 (2020), pp. 1752–1767.

[58] Ming Jin, Igor Molybog, Reza Mohammadi-Ghazi, and Javad Lavaei. "Towards Robust and Scalable Power System State Estimation". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE. 2019, pp. 3245–3252.

[59] Karel J Keesman. "System Identification: An Introduction". Springer Science & Business Media, 2011.

[60] Hassan K Khalil. "Nonlinear Systems; 3rd ed." Upper Saddle River, NJ: Prentice-Hall, 2002.

[61] Seyed Mohammad Khansari-Zadeh and Aude Billard. "Learning Control Lyapunov Function to Ensure Stability of Dynamical System-based Robot Reaching Motions". In: *Robotics Auton. Syst.* 62 (2014), pp. 752–765. URL: https://api.semanticscholar.org/CorpusID:14374268.

[62] Jihun Kim and Javad Lavaei. "Prevailing against Adversarial Noncentral Disturbances: Exact Recovery of Linear Systems with the $\ell_1$-norm Estimator". In: *arXiv preprint arXiv: 2410.03218* (2024).

[63] Franz J Király, Louis Theran, and Ryota Tomioka. "The Algebraic Combinatorial Approach for Low-Rank Matrix Completion." In: *Journal of Machine Learning Research* 16.1 (2015), pp. 1391–1436.

[64] Yehuda Koren, Robert Bell, and Chris Volinsky. "Matrix Factorization Techniques for Recommender Systems". In: *Computer* 42.8 (2009), pp. 30–37.

[65] A Kumar, A Levine, and S Feizi. "Policy Smoothing for Provably Robust Reinforcement Learning". In: *International Conference on Learning Representations (ICLR)*. 2022.

[66] Vitaly Kuznetsov and Mehryar Mohri. "Generalization Bounds for Non-Stationary Mixing Processes". In: *Machine Learning* 106.1 (2017), pp. 93–117.

[67] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. "Gradient Descent Only Converges to Minimizers". In: *Conference on Learning Theory*. PMLR. 2016, pp. 1246–1257.

[68] Ljung Lennart. "System Identification: Theory for the User". In: *PTR Prentice Hall, Upper Saddle River, NJ* 28 (1999), p. 540.

[69] Yingying Li, Subhro Das, Jeff Shamma, and Na Li. "Safe Adaptive Learning-based Control for Constrained Linear Quadratic Regulators with Regret Guarantees". In: *arXiv preprint arXiv:2111.00411* (2021).

[70] Yuanzhi Li, Yingyu Liang, and Andrej Risteski. "Recovery Guarantee of Weighted Low-Rank Approximation via Alternating Minimization". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 2358–2367.

[71] L Ljung, R Pintelon, and J Schoukens. "System Identification: Wiley Encyclopedia of Electrical and Electronics Engineering". In: *System Identification: A Frequency Domain Approach*. John Wiley & Sons Cham, 1999, pp. 1–19.

[72] Lennart Ljung. "Convergence Analysis of Parametric Identification Methods". In: *IEEE Transactions on Automatic Control* 23.5 (1978), pp. 770–783.

[73] Lennart Ljung and Bo Wahlberg. "Asymptotic Properties of the Least-Squares Method for Estimating Transfer Functions and Disturbance Spectra". In: *Advances in Applied Probability* 24.2 (1992), pp. 412–440.

[74] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. "Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval, Matrix Completion, and Blind Deconvolution". In: *Foundations of Computational Mathematics* (2019).

[75] Yao Ma, Alexander Olshevsky, Csaba Szepesvari, and Venkatesh Saligrama. "Gradient Descent for Sparse Rank-One Matrix Completion for Crowd-Sourced Aggregation of Sparsely Interacting Workers". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3335–3344.

[76] Ziye Ma, Yingjie Bi, Javad Lavaei, and Somayeh Sojoudi. "Sharp Restricted Isometry Property Bounds for Low-rank Matrix Recovery Problems with Corrupted Measurements". In: *AAAI-22* (2022).

[77] Ziye Ma and Somayeh Sojoudi. "Noisy Low-Rank Matrix Optimization: Geometry of Local Minima and Convergence Rate". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 3125–3150.

[78] Horia Mania, Stephen Tu, and Benjamin Recht. "Certainty Equivalence is Efficient for Linear Quadratic Control". In: *Advances in Neural Information Processing Systems* 32 (2019).

[79] Shahar Mendelson. "Learning without Concentration". In: *Journal of the ACM (JACM)* 62.3 (2015), pp. 1–25.

[80] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. "Model-Based Reinforcement Learning: A Survey". In: *Foundations and Trends in Machine Learning* 16.1 (2023), pp. 1–118.

[81] Karthik Mohan and Maryam Fazel. "New Restricted Isometry Results for Noisy Low-Rank Recovery". In: *2010 IEEE International Symposium on Information Theory*. IEEE. 2010, pp. 1573–1577.

[82] Yurii Nesterov. "Introductory Lectures on Convex Optimization: A Basic Course". Vol. 87. Springer Science & Business Media, 2013.

[83] Yurii Nesterov. "Primal-Dual Subgradient Methods for Convex Problems". In: *Mathematical Programming* 120.1 (2009), pp. 221–259.

[84] Jean-Philippe Noël and Gaëtan Kerschen. "Nonlinear System Identification in Structural Dynamics: 10 More Years of Progress". In: *Mechanical Systems and Signal Processing* 83 (2017), pp. 2–35.

[85] Robert D Nowak. "Nonlinear System Identification". In: *Circuits, Systems and Signal Processing* 21 (2002), pp. 109–122.

[86] Dohuyng Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. "Non-square Matrix Sensing Without Spurious Local Minima via the Burer–Monteiro Approach". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Vol. 54. Proceedings of Machine Learning Research. 2017, pp. 65–74.

[87] Scott Pesme and Nicolas Flammarion. "Online Robust Regression via SGD on the l-1 Loss". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 2540–2552.

[88] Daniel L Pimentel-Alarcón, Nigel Boston, and Robert D Nowak. "A Characterization of Deterministic Sampling Patterns for Low-Rank Matrix Completion". In: *IEEE Journal of Selected Topics in Signal Processing* 10.4 (2016), pp. 623–636.

[89] Boris T Polyak. "Introduction to Optimization". In: (1987).

[90] Parth Pradhan and Parv Venkitasubramaniam. "Stealthy Attacks in Dynamical Systems: Tradeoffs Between Utility and Detectability With Application in Anonymous Systems". In: *IEEE Transactions on Information Forensics and Security* 12.4 (2017), pp. 779–792. DOI: 10.1109/TIFS.2016.2607695.

[91] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization". In: *SIAM Review* 52.3 (2010), pp. 471–501.

[92] Benjamin Recht, Weiyu Xu, and Babak Hassibi. "Necessary and Sufficient Conditions for Success of the Nuclear Norm Heuristic for Rank Minimization". In: *2008 47th IEEE Conference on Decision and Control*. IEEE. 2008, pp. 3065–3070.

[93] Afshin Rostamizadeh and Mehryar Mohri. "Stability Bounds for Non-IID Processes". In: *Neural Information Processing Systems (NIPS)*. 2007.

[94] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. "Finite Time LTI System Identification". In: *Journal of Machine Learning Research* 22.26 (2021), pp. 1–61.

[95] Yahya Sattar and Samet Oymak. "Non-Asymptotic and Accurate Learning of Nonlinear Dynamical Systems". In: *Journal of Machine Learning Research* 23.140 (2022), pp. 1–49.

[96] Johan Schoukens and Lennart Ljung. "Nonlinear System Identification: A User-Oriented Road Map". In: *IEEE Control Systems Magazine* 39.6 (2019), pp. 28–99.

[97] Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d'Aspremont. "Integration Methods and Optimization Algorithms". In: *Advances in Neural Information Processing Systems* 30 (2017).

[98] George AF Seber and Alan J Lee. "Linear Regression Analysis". John Wiley & Sons, 2012.

[99] Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. "Phase Retrieval with Application to Optical Imaging: A Contemporary Overview". In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 87–109.

[100] Naum Zuselevich Shor. "Minimization Methods for Non-Differentiable Functions". Vol. 3. Springer Science & Business Media, 2012.

[101] Max Simchowitz and Dylan Foster. "Naive Exploration is Pptimal for Online LQR". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8937–8948.

[102] 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*. PMLR. 2018, pp. 439–473.

[103] Amit Singer. "Angular Synchronization by Eigenvectors and Semidefinite Programming". In: *Applied and Computational Harmonic Analysis* 30.1 (2011), pp. 20–36.

[104] André Teixeira, Iman Shames, Henrik Sandberg, and Karl H. Johansson. "Revealing Stealthy Attacks in Control Systems". In: *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2012, pp. 1806–1813. DOI: 10.1109/Allerton.2012.6483441.

[105] Andreas M Tillmann and Marc E Pfetsch. "The Computational Complexity of the Restricted Isometry Property, the Nullspace Property, and Related Concepts in Compressed Sensing". In: *IEEE Transactions on Information Theory* 60.2 (2013), pp. 1248–1259.

[106] Anastasios Tsiamis and George J Pappas. "Finite Sample Analysis of Stochastic System Identification". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE. 2019, pp. 3648–3654.

[107] Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J. Pappas. "Statistical Learning Theory for Control: A Finite-Sample Perspective". In: *IEEE Control Systems Magazine* 43.6 (2023), pp. 67–97. DOI: 10.1109/MCS.2023.3310345.

[108] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. "Low-Rank Solutions of Linear Matrix Equations via Procrustes Flow". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 964–973.

[109] 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.

[110] Andrew Wagenmaker and Kevin Jamieson. "Active Learning for Identification of Linear Dynamical Systems". In: *Conference on Learning Theory*. PMLR. 2020, pp. 3487–3582.

[111] 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.

[112] Irene Waldspurger and Alden Waters. "Rank Optimality for the Burer–Monteiro Factorization". In: *SIAM Journal on Optimization* 30.3 (2020), pp. 2577–2602.

[113] Li Wang, Evangelos A. Theodorou, and Magnus Egerstedt. "Safe Learning of Quadrotor Dynamics Using Barrier Certificates". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)* (2017), pp. 2460–2465.

[114] Stephen J Wright. "Primal-Dual Interior-Point Methods". SIAM, 1997.

[115] Fan Wu, Linyi Li, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. "CROP: Certifying Robust Policies for Reinforcement Learning through Functional Smoothing". In: *International Conference on Learning Representations*. 2022.

[116] Huan Xu, Constantine Caramanis, and Shie Mannor. "Robustness and Regularization of Support Vector Machines." In: *Journal of Machine Learning Research* 10.7 (2009).

[117] Weiyu Xu, Er-Wei Bai, and Myung Cho. "System Identification in the Presence of Outliers and Random Noises: A Compressed Sensing Approach". In: *Automatica* 50.11 (2014), pp. 2905–2911.

[118] Baturalp Yalcin and Javad Lavaei. "Subgradient Method for System Identification with Non-Smooth Objectives". 2025. arXiv: 2503.16673.

[119] Baturalp Yalcin, Haixiang Zhang, Javad Lavaei, and Murat Arcak. "Exact Recovery for System Identification with More Corrupt Data than Clean Data". In: *IEEE Open Journal of Control Systems* (2024).

[120] Baturalp Yalçın, Ziye Ma, Javad Lavaei, and Somayeh Sojoudi. "Semidefinite Programming versus Burer-Monteiro Factorization for Matrix Sensing". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 9. 2023, pp. 10702–10710.

[121] Baturalp Yalçın, Haixiang Zhang, Javad Lavaei, and Somayeh Sojoudi. "Factorization Approach for Low-Complexity Matrix Completion Problems: Exponential Number of Spurious Solutions and Failure of Gradient Methods". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 319–341.

[122] Alp Yurtsever, Joel A. Tropp, Olivier Fercoq, Madeleine Udell, and Volkan Cevher. "Scalable Semidefinite Programming". In: *SIAM Journal on Mathematics of Data Science* 3.1 (2021), pp. 171–200. DOI: 10.1137/19M1305045. eprint: https://doi.org/10.1137/19M1305045.

[123] Alp Yurtsever, Madeleine Udell, Joel Tropp, and Volkan Cevher. "Sketchy Decisions: Convex Low-Rank Matrix Optimization with Optimal Storage". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, Apr. 2017, pp. 1188–1196.

[124] Gavin Zhang and Richard Y. Zhang. "How Many Samples Is a Good Initial Point Worth in Low-Rank Matrix Recovery?" In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 12583–12592.

[125] Haixiang Zhang, Yingjie Bi, and Javad Lavaei. "General Low-Rank Matrix Optimization: Geometric Analysis and Sharper Bounds". In: *Advances in Neural Information Processing Systems*. 2021.

[126] Haixiang Zhang, Baturalp Yalcin, Javad Lavaei, and Somayeh Sojoudi. "A New Complexity Metric for Nonconvex Rank-One Generalized Matrix Completion". In: *Mathematical Programming* 207.1 (2024), pp. 227–268.

[127] Haixiang Zhang, Baturalp Yalcin, Javad Lavaei, and Eduardo D. Sontag. "Exact Recovery Guarantees for Parameterized Non-linear System Identification Problem under Adversarial Attacks". 2024. arXiv: 2409.00276.

[128] Richard Y Zhang and Javad Lavaei. "Sparse Semidefinite Programs with Guaranteed Near-Linear Time Complexity via Dualized Clique Tree Conversion". In: *Mathematical Programming* 188.1 (2021), pp. 351–393.

[129] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. "Sharp Restricted Isometry Bounds for the Inexistence of Spurious Local Minima in Nonconvex Matrix Recovery". In: *Journal of Machine Learning Research* 20.114 (2019), pp. 1–34.

[130] Richard Y. Zhang. "Sharp Global Guarantees for Nonconvex Low-Rank Matrix Recovery in the Overparameterized Regime". arXiv:2104.10790. 2021.

[131] Runyu Zhang, Yingying Li, and Na Li. "On the Regret Analysis of Online LQR Control with Predictions". In: *2021 American Control Conference (ACC)*. IEEE. 2021, pp. 697–703.

[132] Xiao Zhang, Lingxiao Wang, Yaodong Yu, and Quanquan Gu. "A Primal-Dual Analysis of Global Optimality in Nonconvex Low-Rank Matrix Recovery". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. 2018, pp. 5862–5871.

[133] Yu Zhang, Ramtin Madani, and Javad Lavaei. "Conic Relaxations for Power System State Estimation with Line Measurements". In: *IEEE Transactions on Control of Network Systems* 5.3 (2017), pp. 1193–1205.

[134] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. "The Global Optimization Geometry of Low-Rank Matrix Optimization". In: *IEEE Transactions on Information Theory* 67.2 (2021), pp. 1308–1331.

[135] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B. Wakin. "Global Optimality in Low-Rank Matrix Optimization". In: *IEEE Transactions on Signal Processing* 66.13 (2018), pp. 3614–3628.

[136] Ingvar Ziemann, Anastasios Tsiamis, Bruce Lee, Yassir Jedra, Nikolai Matni, and George J. Pappas. "A Tutorial on the Non-Asymptotic Theory of System Identification". 2023. arXiv: 2309.03873 [eess.SY].

[137] Ingvar M Ziemann, Henrik Sandberg, and Nikolai Matni. "Single Trajectory Nonparametric Learning of Nonlinear Dynamics". In: *Conference on Learning Theory*. PMLR. 2022, pp. 3333–3364.