Title
Structure-Aware Methods in Large-Scale Computational Problems: Machine Learning, Optimization, and Control

Author
Fattahi, Salar

Publication Date
2020

Peer reviewed|Thesis/dissertation

Structure-Aware Methods in Large-Scale Computational Problems: Machine Learning, Optimization, and Control

by

Salar Fattahi

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering- Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Javad Lavaei, Co-chair
Assistant Professor Somayeh Sojoudi, Co-chair
Professor Alper Atamturk
Professor Shmuel Oren
Professor Murat Arcak

Summer 2020

Structure-Aware Methods in Large-Scale Computational Problems: Machine Learning, Optimization, and Control

Abstract

Structure-Aware Methods in Large-Scale Computational Problems: Machine Learning, Optimization, and Control

by

Salar Fattahi

Doctor of Philosophy in Engineering- Industrial Engineering and Operations Research

University of California, Berkeley

Associate Professor Javad Lavaei, Co-chair

Assistant Professor Somayeh Sojoudi, Co-chair

Within the realm of computational methods, there has been a long-standing trade-off between the scalability of different techniques and their optimality guarantees. However, most of today's systems—such as transportation, power, and brain networks—are large-scale and safety-critical, thereby requiring both scalability and optimality guarantees. To address these challenges, in this dissertation we develop structure-aware, scalable, and guaranteed computational methods for the *learning*, *optimization*, and *control* of safety-critical systems.

In the first part of the dissertation, we consider two classes of machine learning problems, namely graphical model inference and robust matrix recovery. First, we provide a massively-scalable algorithm for the graphical model inference, where the goal is to reveal hidden correlation structures of high-dimensional datasets. We introduce a graph-based method that is capable of solving instances with billions of variables in less than an hour, significantly outperforming other state-of-the-art methods. Next, we consider a class of nonconvex and nonsmooth optimization problems in safe machine learning. We show that, despite their nonconvexity, a large class of problems in robust matrix recovery is devoid of spurious and sub-optimal solutions, thereby leading to the guaranteed success of fast local-search algorithms.

The second part of the dissertation is devoted to different classes of network optimization problems. In particular, we consider a class of generalized network flow problems that are at the backbone of many modern interconnected systems, such as power, water, and gas networks. Unlike many of its classical counterparts, the generalized network flow problem is highly nonconvex due to the incorporation of nonlinear losses in its formulation. To address this issue, we propose an efficient convex relaxation of the problem, and provide conditions under which the proposed relaxation is exact. Next, we focus on a specialized network opti-

mization problem in power systems, namely optimal transmission switching, where the goal is to find the optimal topology of a power grid to minimize its cost of operation, while satisfying operational and security constraints in the network. The optimal transmission switching is a NP-hard optimization problem with mixed-integer variables. However, by exploiting the *tree-like* structure of realistic power grids, we introduce an strengthened formulation of the problem that can be solved efficiently in practice.

The third part of the dissertation is concerned with the design of robust and distributed control policies for dynamical systems with uncertain models. To this end, first we propose a sparsity-exploiting technique for the efficient learning of a structured dynamical system, based on a limited number of collected input-output sample trajectories from the system. In particular, we quantify the sample complexity of the sparse system identification problem in a high-dimensional setting, where the dimension of the system is significantly greater than the number of available data samples. Given the estimated dynamics, our next goal is to design a robust and distributed control policy for the system by taking into account the uncertainty of its estimated model. We show that near-optimal distributed controllers can be learned with logarithmic sample complexity and computed with near-linear time complexity.

To my parents, for their unconditional love and support

Contents

List of Figures

List of Tables

Acknowledgments

This thesis would not have been possible without the help and support from my advisors, collaborators, friends, and family in the past five years.

First and foremost, I would like to thank my advisor, Professor Javad Lavaei. His intelligence, vision, and versatile expertise helped me greatly in my research and inspired my future career directions. I am greatly indebted to Javad for his endless help and support throughout my PhD, for helping me become an independent researcher, and for giving me the courage to tackle hard problems. He was always available whenever I ran into a trouble or had a problem in my research. I will always be grateful to him for having faith in me, and for providing me with many research opportunities.

I am thankful to my co-advisor, Somayeh Sojoudi, for introducing me to the world of machine learning. Under Somayeh's supervision, I had the opportunity to add machine learning as a new dimension to my research agenda. Aside from her keen eye towards interesting research topics, her friendly and supportive character was like a breath of fresh air throughout my PhD life. I am thankful to Somayeh for always being supportive and making sure that I have a well-rounded academic and personal life.

I could not have asked for a better thesis committee member, mentor, and collaborator than Alper Atamturk. Throughout my PhD, I was fortunate to receive significant advice and assistance from Alper, whom I consider as a true role model in my academic life. I also thank my other thesis committee members, Shmuel Oren and Murat Arcak, who provided me with helpful feedbacks during my PhD. I am truly grateful to Shmuel for his support and guidance during my academic job applications. I also feel fortunate to have had the opportunity to collaborate with Murat on the problem of distributed control, which lead to a series of publications.

I had the pleasure of working with two superb researchers, Richard Y. Zhang and Cedric Josz, who are now faculty members at UIUC and Columbia University. I have also learned a great deal from Ramtin Madani and Andres Gomez, who are now faculty members at UT Arlington and USC. I would welcome the opportunity to work with and alongside these stellar researchers on joint projects in the future. My gratitude extends to my other co-authors: John Lygeros, Nikolai Matni, Reza Mohammadi, Julie Mulvaney-Kemp, Morteza Ashraphijou, Ghazal Fazelnia, and Georgios Darivianakis.

I consider myself incredibly lucky to have been in such a friendly environment at UC Berkeley. My sincere gratitude goes to Mahbod Olfat, Georgios Patsakis, Pedro Hespanhol, Dean Grosbard, Yonatan Mintz, Quico Spaen, Alfonso Lobos, Han Feng, SangWoo Park, Igor Molybog, Ming Jin, Yuhao Ding, Armin Askari, Mahan Tajrobehkar, and Arman Jabbari. A special shoutout goes to Mahbod Olfat, my roommate and friend who became like a brother to me. I will forever remember our late night discussions, from philosophy and politics to different twists of Game of Thrones. I thank Mahbod, George, and Pedro for co-founding the "Farkas Group" (aka the most exclusive group at UC Berkeley!). My personal well-being during my PhD was hugely indebted to my amazing friends outside UC Berkeley IEOR,

including Rafegh Aghamohammadi, Sina Akhbari, Pouria Kourehpaz, Sajjad Moazeni, and Ahmad Zareei.

Words cannot describe my gratitude for my parents, Mohammadreza and Ghamarrokh, and my sister, Sarvin, who have always filled my life with their selfless love and support. I am wholeheartedly grateful to them for making so many sacrifices in their lives, and for tolerating thousands of miles between us. Finally, I want to thank my best friend, Behnaz, who has continuously helped me throughout my academic and personal life. No words can express my gratitude for her unconditional support.

Chapter 1

Introduction

This dissertation focuses on developing data-driven and large-scale computational methods for modern interconnected and safety-critical problems. Today's systems are complex and large, often with a massive number of unknown parameters which render them doomed to the so-called *curse of dimensionality*. The ever-growing and dynamic interconnections between smart systems (such as smart grids and cities) have been a major impediment to their safe and resilient operation. The goal of this dissertation is to identify, study, and exploit the underlying hidden-but-useful structures of these large-scale and real-world problems with the goal of designing certifiable computational methods that, at the same time, can be easily implemented and used in practice.

Our main goal is to strike a balance between two major paradigms, namely *theory vs. application* of the computational methods, and their *efficiency vs. accuracy*. In particular, we will make use of cutting-edge techniques in learning, optimization, and control to solve massive-scale problems that stem from real-life applications, with a special focus on interconnected and safety-critical systems, such as power, transportation, and brain networks. Indeed, modern computational problems are complex and, consequently, most of the available algorithms lean towards enhancing their *efficiency* or *accuracy*, at the expense of sacrificing the other. We strive to develop structure-promoting algorithms that can provide the best of both worlds. In particular, by taking advantage of application-specific structure of the problem (such as sparsity, locality, low-rankness), our goal is to guarantee their efficient solvability by developing practical algorithms, while ensuring the near-global optimality of the obtained solutions.

In the following sections of this chapter, we first provide a general introductory overview of the problems that are considered in this dissertation, as well as the challenges we may face towards solving them. Next, we provide a brief summary of our contributions, together with the relevant publications. We conclude this chapter by presenting the basic notations that are used throughout the dissertation.

1.1 Optimization as an Overarching Framework

A major part of this dissertation is devoted to solving optimization problems in the form of

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x};\theta) \tag{1.1a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}(\theta) \tag{1.1b}$$

where:

- $\mathbf{x} \in \mathbb{R}^n$ is the targeted multivariate *decision variable*. For instance, it may capture the amount of generations for different generators in a power system; it can correspond to the unknown interactions between different brain regions in response to various physical or mental activities; or it may indicate an optimal control policy for a dynamical system.

- $\theta \in \mathbb{R}^m$ is the exogenous vector that (directly or indirectly) captures the parameters of the problem. For instance, it may include the generation capacities of different generators in a power system; it may correspond to the functional MRI scans that are collected from a brain network; or it can encapsulate specific parameters of a dynamical system.

- $f(\mathbf{x};\theta)$ is the objective function in terms of $\mathbf{x}$ and parameterized by θ. For example, it may correspond to the operational cost of a power system; it can capture the estimation error of an inferred brain connectivity network; or it may be equivalent to some notion of robustness in a dynamical system.

- $\mathcal{X}(\theta)$ is the feasible set of the optimization problem (parameterized by θ), i.e., the set of all feasible values that can be attributed to the decision variable $\mathbf{x}$. The feasible set $\mathcal{X}(\theta)$ can be either explicitly characterized by a set of inequality or equality constraints, or it may be given implicitly via a set of (noisy) observations from the problem. For instance, it may correspond to different security and operational constraints in a power system; it can correspond to various structural constraints on a brain connectivity network; or it may capture certain communication constraints on the set of feasible control policies for an interconnected dynamical system.

Evidently, our ultimate goal is to obtain a *globally-optimal* solution $\mathbf{x}^*$ to (1.1) that universally minimizes the objective function $f(\mathbf{x},\theta)$ over all possible feasible points in $\mathcal{X}(\theta)$. However, as will be delineated later, depending on the complexity of the optimization problem, one may only hope to obtain a *locally optimal* solution[1], or merely a feasible solution without any guarantee on its local or global optimality.

[1] A solution $\bar{\mathbf{x}}$ is locally optimal if there exists $\epsilon > 0$ such that $f(\bar{\mathbf{x}},\theta) \leq f(\mathbf{x},\theta)$ for every $\mathbf{x} \in \mathcal{X}(\theta) \cap \mathcal{B}(\bar{\mathbf{x}},\epsilon)$, where $\mathcal{B}(\bar{\mathbf{x}},\epsilon)$ is a Euclidean ball centered at $\bar{\mathbf{x}}$ with radius ϵ.

As will be shown later in the dissertation, many real-world problems can be cast as instances of (1.1). Before delving into the details of such problems, first we will present a number of universal challenges in solving the aforementioned optimization problem.

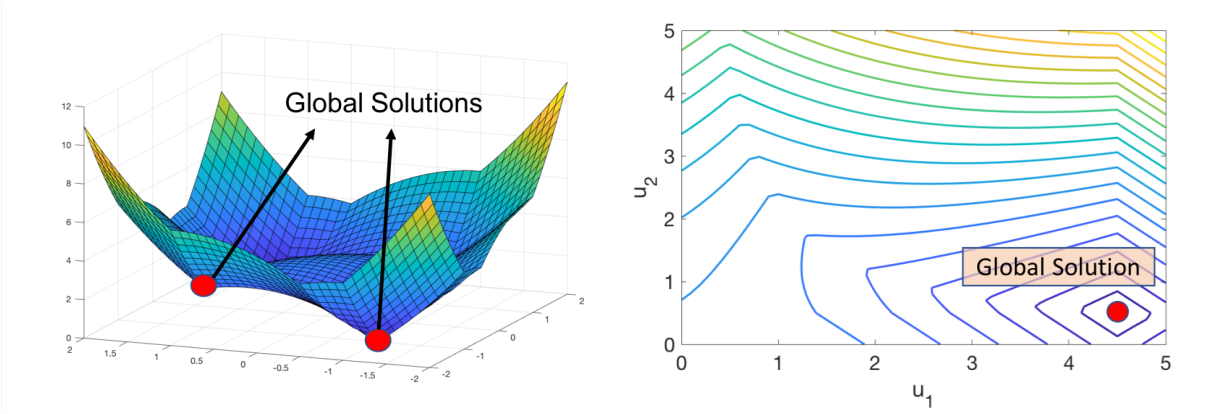
1. *Convexity vs. nonconvexity:* It is a conventional wisdom that the complexity of solving an optimization problem is closely tied to its convexity. Roughly speaking, an optimization problem is convex if it satisfies two conditions: 1) the objective function $f(\mathbf{x}, \theta)$ is convex, i.e., the segment between any two points on the function lies above the function[2], and 2) the feasible set $\mathcal{X}(\theta)$ is convex, i.e., any point on the segment between any pair of feasible points is also feasible[3]. It is well-known that convex optimization problems are theoretically easy to solve due to an equivalence between their local and global optimality conditions: any locally-optimal solution is also globally-optimal. This important property enables different local-search algorithms to solve (1.1) to global optimality. On the other hand, a nonconvex optimization may possess multiple local/global solutions, any of which may be recovered and returned as a candidate solution using our numerical algorithms.

   However, a recent line of research reveals that a criterion solely based on "convexity vs. nonconvexity" is not enough to characterize the difficulty of solving an optimization problem. A case study that best exemplifies such phenomenon is the famous *low-rank matrix recovery* problem, where the goal is to recover a low-rank matrix given a limited number of (possibly noisy) observations. Due to the inherent nonconvexity of the low-rank matrix recovery problem, the most commonly-used methods for solving this problem are based on *convex relaxation* techniques, where the problem is relaxed into a convex optimization problem (typically a semidefinite programming) in a lifted space, where the number of variables is often significantly greater than that of the original formulation. However, it has been recently observed that such convex relaxations are not necessary to guarantee the recovery of a globally-optimal solution. In fact, it is shown that, for different classes of low-rank matrix recovery problems, globally-optimal solutions may be obtained via their nonconvex formulations much faster than those obtained using convex relaxation techniques. This counter-intuitive observation gives rise to two important points:

   - The equivalence between local and global optimality by itself is not enough to guarantee the efficient solvability of an optimization problem. In other words, even if an algorithm is guaranteed to converge to a globally-optimal solution, it may still have overwhelmingly high per-iteration complexity, thereby making it prohibitive to use in practice.
   - The convexity of an optimization problem is only a sufficient condition for the absence of bad local minima. In other words, an optimization problem can be

[2]More formally, for any $\mathbf{x}, \bar{\mathbf{x}} \in \mathbb{R}^n$ and $\alpha \in [0, 1]$, the inequality $f(\alpha\mathbf{x} + (1-\alpha)\bar{\mathbf{x}}, \theta) \leq \alpha f(\mathbf{x}, \theta) + (1-\alpha) f(\bar{\mathbf{x}}, \theta)$ holds.

[3]More formally, for any $\mathbf{x}, \bar{\mathbf{x}} \in \mathcal{X}(\theta)$ and $\alpha \in [0, 1]$, we have $\alpha\mathbf{x} + (1-\alpha)\bar{\mathbf{x}} \in \mathcal{X}(\theta)$.



(a) 3D Landscape (b) Level-sets

Figure 1.1: An instance of low-rank matrix recovery problem that is devoid of spurious local solutions (see Chapter 3 for more details). a) 3D landscape of the function shows that it has two globally-optimal minima without any spurious local minima, b) Level-sets of the function reveal that the it is neither convex nor quasiconvex.

nonconvex, and yet, it may be devoid of spurious and undesirable local minima; see Figure 1.1 for an example of such functions.

A key takeaway of the aforementioned observations is: *Understanding the true complexity of modern optimization problems requires rethinking convexity as a measure of their difficulty*; a subject that is at the core of Chapter 3 of this dissertation.

2. *Stochasticity:* Thus far, our discussion was based on the assumption that we have a full knowledge of the parameter vector θ of the optimization problem (1.1). Indeed, such assumption is rarely valid in practice. For instance, the functional MRI scans can only provide limited observations of the brain activity, and are often subject to random noise. Similarly, the true model of a dynamical system is rarely known in practice and, instead, it is estimated *indirectly* by analyzing its behavior in response to different inputs. This indeed adds a new dimension to the complexity of solving (1.1): Not only do we need to design efficient algorithms for solving (1.1), but we also need to infer an accurate estimate $\hat{\theta}$ of θ based on a limited number of noisy observations/samples $\{w_i\}_{i=1}^N$. More formally, our goal is to design an estimator $\hat{\theta} = \phi(w_1, \ldots, w_N)$ and solve the following *surrogate* optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}; \hat{\theta}) \tag{1.2a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}(\hat{\theta}) \tag{1.2b}$$

Furthermore, our aim is to characterize the *sample complexity* of the above optimization problem, i.e., the number of samples N that is required to guarantee the closeness of the surrogate optimization problem (1.2) and its solutions to its true counterpart (1.1). We will elaborate more on this connection in Chapters 3, 6, and 7 of the dissertation.

3. *Robustness:* As mentioned before, most of the problems that are considered in this dissertation are motivated by safety-critical applications. In many cases, the reliability requirements for a safety-critical system can be translated into some measure of *robustness* of its corresponding optimization problem. For instance, suppose that, instead of directly observing the true parameter θ, we know *a priori* that it belongs to a pre-defined set of parameters Θ, i.e., $\theta \in \Theta$. Then, instead of estimating the true parameter, one may take a more conservative approach of solving the following optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\tilde{\theta}\in\mathbb{R}^m} \quad f(\mathbf{x};\tilde{\theta}) \tag{1.3a}$$

$$\text{subject to} \quad \mathbf{x} \in \mathcal{X}(\tilde{\theta}) \tag{1.3b}$$

$$\tilde{\theta} \in \Theta \tag{1.3c}$$

which can be described as follows: Due to the unknown nature of the true parameter θ, our goal is to pessimistically obtain a solution that is governed by a *worst-case* parameter $\tilde{\theta} \in \Theta$. Despite its favorable worst-case guarantees, it is not surprising that the robust variants of optimization problems are often significantly harder to solve. In Chapter 7, we will consider a special class of robust optimization problems that arise in the distributed control of dynamical systems.

1.2 Summary of Contributions

The intelligent, efficient, and resilient operation of safety-critical systems is contingent upon developments at different fronts of data analytics and computational methods, including the scalability of the optimization techniques, their robustness against uncertainties, and the efficiency in the learning methods. The dimensionality and complexity of modern safety-critical problems is overwhelmingly high, often surpassing what existing methods can solve in a reasonable amount of time. Throughout the dissertation, we show that exploiting the underlying structure of real-world systems, such as their sparsity, locality, or low-rankness, is a key game-changer in the pursuit of better computational methods. In this section, we will briefly summarize the contributions of the dissertation.

Machine Learning

Chapters 2 and 3 of the dissertation are devoted to two classes of problems in safe machine learning, namely sparse inverse covariance estimation and robust matrix recovery. These

problems are extensively used in brain and transportation networks, safe recommendation systems, and self-driving cars.

- **Chapter 2:** Graphical models are fundamental methods for obtaining interpretable descriptions of large-scale datasets. For instance, in Neuroscience, it is known that brain connectivity can be studied by inferring an associated graphical model based on functional MRI measurements. As another example, graphical models can be used in short- and long-term traffic flow prediction and control of intelligent transportation systems (ITSs). At the heart of graphical model inference is sparse inverse covariance estimation. The best known algorithms for sparse inverse covariance estimation have time complexities on the order of $\mathcal{O}(n^4)$, making them prohibitive to solve massive-scale instances of the problem. This is despite the fact that in high-dimensional settings, the sample covariance matrix can be efficiently constructed in $\mathcal{O}(n^2)$. The prohibitive computational cost of the current solvers for sparse inverse covariance estimation motivated us to investigate the following question: *Is it possible to design low-complexity algorithms for sparse inverse covariance estimation?*

  In Chapter 2, we provide an affirmative answer to the above question. In particular, we show that, under mild assumptions, a simple thresholding operation on the sample covariance matrix reveals the sparsity pattern of the inverse covariance matrix. By building upon this result, we prove that sparse inverse covariance estimation can be solved to near-optimality in $\mathcal{O}(n^2)$ time and $\mathcal{O}(n)$ memory complexities. Furthermore, we show the graceful scalability of the proposed method on real-life functional MRI data and traffic flows for transportation networks. In practice, our method obtains accurate estimates of the inverse covariance matrix for instances with more than 3.2 billion variables in less than 30 minutes on a laptop computer, while other methods do not converge within 4 hours.

- **Chapter 3:** A recent line of work has shown that a surprisingly large class of smooth-but-nonconvex low-rank optimization problems—including matrix completion/sensing, phase retrieval, and dictionary learning—has a *benign landscape*, i.e., every local solution is also global. Despite the nonconvexity of these problems, their benign landscape implies that simple local-search algorithms are guaranteed to converge to a globally-optimal solution, thus leading to significant computational savings and zero optimality gap. In general, the validity of these results relies heavily on the smoothness of the objective function. However, such smooth objective functions are not *robust* against outliers, i.e., they cannot correctly identify and reject large-and-sparse noise values. Inspired by this deficiency in existing methods, we studied the following open problem: *Does robust low-rank optimization in the presence of a nonsmooth objective function still have a benign landscape?*

  In Chapter 3, we consider an important class of such problems, namely *non-negative robust principal component analysis* (NRPCA), in which the goal is to *exactly* recover the non-negative and low-rank component of a measurement matrix, despite a subset

of the measurements being grossly corrupted with large noise values. We prove that NRPCA has no spurious local minima under a set of necessary and sufficient conditions, such as strict positivity of the true components, as well as the absence of bipartite components in its sparsity graph. This implies that, despite the highly nonsmooth and nonconvex nature of NRPCA (see Figure 1.1 for an illustrative example), simple local search algorithms can efficiently recover its globally-optimal solution. By building upon this result and leveraging contemporary techniques in random graph theory, we provide probabilistic guarantees on the absence of spurious local minima under random sampling and noise regime. In particular, we show that up to a constant factor of the measurements could be corrupted by large amounts of noise without creating any spurious local solution.

Network Optimization

Chapters 4 and 5 of the dissertation consider two classes of network optimization problems, namely generalized network flow and optimal transmission switching, with primary applications in power systems.

- **Chapter 4:** Network flow problems play a crucial role in operations research with a myriad of applications in assignment, electrical power, and production networks, to name a few. Most of the classical results on the network flow problem are contingent upon the lossless nature of the network. However, physical systems are lossy, where the loss is often a nonconvex function of the flows. An example is power networks where the loss over each line is given by a parabolic function due to Kirchhoff's circuit laws. Indeed, the accurate incorporation of these nonlinearities in the optimization of such realistic network flow problems can ensure their cost-efficient and safe operation, thereby leading to tremendous economic and environmental benefits.

  In Chapter 4, we investigate optimization over lossy networks in the context of the generalized network flow (GNF) problem. GNF aims to minimize the operational cost of a lossy network by optimizing over the nodal injections subject to flow constraints. Solving GNF to optimality is a daunting task due to the incorporation of nonlinear losses in its formulation. However, we introduce an efficient convex relaxation of the problem that incurs zero optimality gap. In particular, we prove that, under practical conditions, the globally optimal cost and nodal injections can be efficiently obtained by simply relaxing the nonconvex equality constraints to convex inequalities. Unlike the computationally-expensive convexification techniques—such as sum-of-squares (SOS)—that are based on lifting the problem to higher dimensions, our proposed convex relaxation is defined over the original space of variables, making it suitable for the real-time operation of lossy networks in realistic scales.

- **Chapter 5:** Optimal transmission switching (OTS) problem is a recently-developed control paradigm to optimize the topology of the power networks with the goal of

improving the dispatch of electricity, while satisfying physical and operational constraints. The nonlinear and mixed-integer nature of this problem has been the major impediment to the scalability and reliability of its existing solvers.

In Chapter 5, we introduce an efficient bound strengthening method for solving the OTS by leveraging the graph structure of the power systems. Our proposed method leads to a 10-fold speedup in the solution time of the mixed-integer solvers for large-scale power systems, including Polish networks.

System Identification and Control

Chapters 6 and 7 are devoted to the system identification and distributed control of interconnected systems with unknown dynamics, with applications in the control of multi-agent systems, such as self-driving cars.

- **Chapter 6:** With their ever-growing size and complexity, real-world dynamical systems are hard to model. Therefore, system operators should rely on efficient estimation methods to identify the dynamics of the system via a limited number of recorded input-output interactions. The area of system identification is created to address this problem.

  In chapter 6, our objective is to employ modern results on high-dimensional statistics to reduce the sample complexity of a fundamental class of system identification problems in control theory, namely linear time-invariant (LTI) systems with perfect state measurements. Our results are built upon the fact that, in many practical large-scale systems, the states and inputs exhibit sparse interactions with one another, which in turn translates into a sparse representation of the state-space equations of the system. In particular, we propose a sparsity-promoting estimator that can correctly identify the underlying structure of the system matrices with high probability, provided that the length of the sample trajectory exceeds a threshold. Furthermore, we show that this threshold scales polynomially in the number of nonzero elements in the system matrices, but only logarithmically in the system dimensions. Finally, we present an extensive case study on power systems to illustrate the performance of the proposed estimation method.

- **Chapter 7:** The efficient operation of intelligent and dynamical infrastructures—such as smart cities and grids—demands a shift from classical centralized control policies toward efficient edge computing methods with *distributed* control schemes. The main objective in distributed control problem is to design a hierarchy of interacting sub-controllers with a prescribed structure—as opposed to the traditional unstructured and centralized control architectures—for an interconnected system consisting of local sub-systems.

  Another challenge in the control of dynamical systems is uncertainty in their models. The unknown nature of a dynamical system implies that any viable control policy

should actively interact with the system to learn the model, and then make robust decisions by taking into account the uncertainty of the learned model. Indeed, a practical data-driven control framework should not have a "long interaction" with an unknown system in the learning phase to avoid jeopardizing its safety, and it should be efficient to design and implement. We address these challenges in Chapter 7, where we introduce a robust and learning-based distributed control scheme for linear systems that benefits from efficient sample and computational complexities. Our scheme only makes a logarithmic number of interactions with the unknown system to learn the model, and then designs a controller in near-linear time complexity.

Related Publications

- **Chapter 2.**
  Main paper:

  1. Salar Fattahi and Somayeh Sojoudi, "Graphical Lasso and Thresholding: Equivalence and Closed-form Solutions", *Journal of Machine Learning Research (JMLR)*, 2019

     - INFORMS Data Mining Best Paper Award - Applied Track, 2018,
     - Katta G. Murty Best Paper Award, 2018.

  Related papers:

  2. Richard Y. Zhang, Salar Fattahi and Somayeh Sojoudi, "Large-Scale Sparse Inverse Covariance Estimation via Thresholding and Max-Det Matrix Completion", *International Conference on Machine Learning (ICML)*, 2018
  3. Salar Fattahi and Somayeh Sojoudi, "Closed-Form Solution and Sparsity Path for Inverse Covariance Estimation Problem", *American Control Conference (ACC)*, 2018

     - ACC Best Student Paper Award - Finalist, 2018,
  4. Salar Fattahi, Richard Y. Zhang, and Somayeh Sojoudi, "Sparse Inverse Covariance Estimation for Chordal Structures", *European Control Conference (ECC)*, 2018
  5. Salar Fattahi, Richard Y. Zhang, and Somayeh Sojoudi, "Linear-Time Algorithm for Learning Large-Scale Sparse Graphical Models", *IEEE Access*, 2019

- **Chapter 3.**
  Main paper:

  6. Salar Fattahi and Somayeh Sojoudi, "Exact Guarantees on the Absence of Spurious Local Minima for Rank-1 Non-negative Robust Principal Component Analysis", *Journal of Machine Learning Research (JMLR)*, 2020

Related papers:

7. Salar Fattahi, Cedric Josz, Reza Mohammadi, Javad Lavaei, and Somayeh Sojoudi, "Absence of Spurious Local Trajectories in Time-varying Optimization", *submitted for journal publication*, 2019
8. Julie Mulvaney-Kemp, Salar Fattahi, and Javad Lavaei, "Smoothing Property of Load Variation Promotes Finding Global Solutions of Time-Varying Optimal Power Flow", *submitted for journal publication*, 2020
9. Julie Mulvaney-Kemp, Salar Fattahi, and Javad Lavaei, "Load Variation Enables Escaping Poor Solutions of Time-Varying Optimal Power Flow", *IEEE Power & Energy Society General Meeting*, 2020

- **Chapter 4.**
Main paper:

10. Somayeh Sojoudi, Salar Fattahi, and Javad Lavaei, "Convexification of Generalized Network Flow Problem", *Mathematical Programming*, 2019

Related paper:

12. Salar Fattahi and Javad Lavaei, "Convex Analysis of Generalized Flow Networks", *IEEE Conference on Decision and Control (CDC)*, 2015

- **Chapter 5.**
Main paper:

13. Salar Fattahi, Javad Lavaei, and Alper Atamturk, "A Bound Strengthening Method for Optimal Transmission Switching in Power Systems with Fixed Connected Subgraph", *IEEE Transactions on Power Systems*, 2019

Related papers:

14. Salar Fattahi, Javad Lavaei, and Alper Atamturk, "Promises of Conic Relaxations in Optimal Transmission Switching of Power Systems", *IEEE Transactions on Power Systems*, 2019
15. Salar Fattahi, Morteza Ashraphijou, Javad Lavaei, and Alper Atamturk, "Conic Relaxation of the Unit Commitment Problem", *Energy*, 2017
16. Morteza Ashraphijou, Salar Fattahi, Javad Lavaei, and Alper Atamturk, "A Strong Semidefinite Programming Relaxation of the Unit Commitment Problem", *IEEE Conference on Decision and Control (CDC)*, 2016

- **Chapter 6.**
Main paper:

17. Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi, "Learning Sparse Dynamical Systems from a Single Sample Trajectory", *IEEE Conference on Decision and Control (CDC)*, 2019

Related papers:

18. Salar Fattahi and Somayeh Sojoudi, "Sample Complexity of Sparse System Identification Problem for Linear Time-Invariant Systems", *submitted for journal publication*, 2019
19. Salar Fattahi and Somayeh Sojoudi, "Data-Driven Sparse System Identification", *Annual Allerton Conference on Communication, Control, and Computing*, 2018
20. Salar Fattahi and Somayeh Sojoudi, "Non-Asymptotic Analysis of Block-Regularized Regression Problem", *IEEE Conference on Decision and Control (CDC)*, 2018

- **Chapter 7.**
Main paper:

21. Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi, "Efficient Learning of Distributed Linear-Quadratic Regulators", *submitted for journal publication*, 2019

Related papers:

22. Salar Fattahi, Ghazal Fazelnia, Javad Lavaei, and Murat Arcak, "Transformation of Optimal Centralized Controllers Into Near-Global Static Distributed Controllers", *IEEE Transactions on Automatic Control*, 2019
23. Georgios Darivianakis, Salar Fattahi, Javad Lavaei, and John Lygeros, "High-Performance Cooperative Distributed Model Predictive Control for Linear Systems", *American Control Conference (ACC)*, 2018
24. Salar Fattahi, Javad Lavaei, and Murat Arcak, "A Scalable Method for Designing Distributed Controllers for Systems with Unknown Initial State", *IEEE Conference on Decision and Control (CDC)*, 2017
25. Salar Fattahi and Javad Lavaei, "Theoretical Guarantees for the Design of Near Globally Optimal Static", *Annual Allerton Conference on Communication, Control, and Computing*, 2016

1.3 Notations

Scalars, vectors, matrices, and sets: Lowercase, bold lowercase and uppercase letters are used for scalars, vectors and matrices, respectively (say $x, \mathbf{x}, X$). The symbols $\mathbb{R}^d$, $\mathbb{S}^d$, $\mathbb{S}^d_+$, and $\mathbb{S}^d_{++}$ are used to denote the sets of $d \times 1$ real vectors, $d \times d$ symmetric matrices, $d \times d$ symmetric positive-semidefinite matrices, and $d \times d$ symmetric positive-definite matrices,

respectively. The notations $\text{trace}(M)$ and $\log\det(M)$ refer to the trace and the logarithm of the determinant of a matrix M, respectively. The notation $M \bullet N$ or $\langle M, N\rangle$ denote the inner product between the matrices M and N of the same size. The $(i,j)^{\text{th}}$ entry of the matrix M is denoted by M_{ij}. The symbols $M_{:j}$ and $M_{j:}$ indicate the j^{th} column and row of M, respectively. Given the index sets $\mathcal{U}$ and $\mathcal{V}$, define $M_{\mathcal{U}\mathcal{V}}$ as the $|\mathcal{U}| \times |\mathcal{V}|$ submatrix of M after removing the rows and columns with indices not belonging to $\mathcal{U}$ and $\mathcal{V}$. Moreover, I_n denotes the $n \times n$ identity matrix. The sign of a scalar x is shown as $\text{sign}(x)$. The notations $|x|$, $\|M\|_1$ and $\|M\|_F$ denote the absolute value of the scalar x, the element-wise ℓ_1, and Frobenius norm of the matrix M, respectively. The symbols $|||M|||$, $|||M|||_\infty$, and $|||M|||_1$ are used to denote its induced spectral, infinity, and ℓ_1/ℓ_1 norms, respectively. We will frequently write $M \succeq 0$ to mean $M \in \mathbb{S}^n_+$ and write $M \succ 0$ to mean $M \in \mathbb{S}^n_{++}$. Given a sparsity pattern $G \in \{1,\ldots,n\}^2$, we define $\mathbb{S}^n_G \subseteq \mathbb{S}^n$ as the set of $n \times n$ real symmetric matrices with this sparsity pattern. Let $P_H(M)$ denote the projection operator from $\mathbb{S}^n$ onto $\mathbb{S}^n_H$, i.e. by setting all $S_{ij} = 0$ if $(i,j) \notin H$. The ceiling function is denoted as $\lceil\cdot\rceil$. The cardinality of a discrete set $\mathcal{D}$ is denoted as $|\mathcal{D}|_0$. Given a matrix $M \in \mathbb{S}^d$, define

$$\begin{aligned}
\|M\|_{1,\text{off}} &= \sum_{i=1}^{n}\sum_{j=1}^{n} |M_{ij}| - \sum_{i=1}^{n} |M_{ii}|,\\
\|M\|_{\max} &= \max_{i\neq j} |M_{ij}|.\\
\|M\|_{\infty} &= \max_{i,j} |M_{ij}|
\end{aligned}$$

Probability: For an event $\mathcal{E}$, the notation $\mathbb{P}(\mathcal{E})$ is used to show the probability of its occurrence. For a random variable x, the symbol $\mathbb{E}\{x\}$ shows its expected value. The notation $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ implies that $\mathbf{x}$ is a random vector drawn from a Gaussian distribution with mean vector μ and covariance matrix Σ. The notation $x_n \overset{a.s.}{\to} x$ is used to show that a sequence of random variables x_n converges to x almost surely.

Functions: Given the sequences $f_1(n)$ and $f_2(n)$, the notation $f_1(n) \lesssim f_2(n)$ or equivalently $f_1(n) = O(f_2(n))$ means that there exists a number $c_1 \in [0,\infty)$ such that $f_1(n) \leq c_1 f_2(n)$ for all $n \geq 1$. Similarly, the notation $f_1(n) \gtrsim f_2(n)$ or $f_1(n) = \Omega(f_2(n))$ means that there exists a number $c_2 > 0$ such that $f_1(n) \geq c_2 f_2(n)$ for all $n \geq 1$. The indicator function $\mathbb{I}_{x\geq\alpha}$ takes the value 1 if $x \geq \alpha$ and 0 otherwise. To streamline the presentation and whenever the equivalence is clear by the context, we abuse notation and use boldface upper- and lower-case letters to denote transfer matrices and vector-valued signals, respectively. The symbols $\mathcal{H}_2$ and $\mathcal{H}_\infty$ are endowed with the standard definitions of the Hardy spaces, i.e., the class of holomorphic transfer functions on the open unit disk with bounded mean square and maximum norms, respectively. Accordingly, let $\mathcal{RH}_2$ and $\mathcal{RH}_\infty$ correspond to the restriction of these spaces to the set of real, rational, and proper functions. For a transfer matrix $\mathbf{M} \in \mathcal{RH}_\infty$, one can write $\mathbf{M} = \sum_{\tau=0}^{\infty} M(\tau) z^{-\tau}$, where $M(\tau)$ is the τ^{th} spectral component of M.

Part I

Machine Learning

Chapter 2

Closed-form Solutions for Sparse Inverse Covariance Estimation

Sparse inverse covariance estimation is a popular method for learning the structure of undirected Gaussian graphical models, which is commonly solved using an l_1-regularized Gaussian maximum likelihood estimator known as "Graphical Lasso" (GL). Despite the convexity of the problem, its computational cost becomes prohibitive for large-scale instances.

The first objective of this chapter is to compare the computationally-heavy GL technique with a numerically-cheap heuristic method that is based on simply thresholding the sample covariance matrix. To this end, two notions of sign-consistent and inverse-consistent matrices are developed, and then it is shown that the thresholding and GL methods are equivalent if: (i) the thresholded sample covariance matrix is both sign-consistent and inverse-consistent, and (ii) the gap between the largest thresholded and the smallest un-thresholded entries of the sample covariance matrix is not too small.

By building upon this result, we prove that the GL—as a conic optimization problem—has an explicit closed-form solution if the thresholded sample covariance matrix has an acyclic structure. This result is then generalized to arbitrary sparse support graphs, where a formula is found to obtain an approximate solution of GL. Furthermore, it is shown that the approximation error of the derived explicit formula decreases exponentially fast with respect to the length of the minimum-length cycle of the sparsity graph.

The developed results are demonstrated on synthetic data, as well as on massive real-world datasets, such as functional MRI data, traffic flows for transportation systems, and chemical networks. We show that the proposed method can obtain an accurate approximation of the GL for instances with the sizes as large as $80,000 \times 80,000$ (more than 2 billion variables) in less than 30 minutes on a standard laptop computer running MATLAB, while other state-of-the-art methods do not converge within 4 hours.

2.1 Introduction

There has been a pressing need in developing new and efficient computational methods to analyze and learn the characteristics of high-dimensional data with a structured or randomized nature. Real-world data sets are often overwhelmingly complex, and therefore it is important to obtain a simple description of the data that can be processed efficiently. In an effort to address this problem, there has been a great deal of interest in sparsity-promoting techniques for large-scale optimization problems [57, 15, 26]. These techniques have become essential to the tractability of big-data analyses in many applications, including data mining [98, 192, 269], pattern recognition [267, 210], human brain functional connectivity [239], distributed controller design [78, 82], and compressive sensing [47, 95]. Similar approaches have been used to arrive at a parsimonious estimation of high-dimensional data. However, most of the existing statistical learning techniques in data analytics are contingent upon the availability of a sufficient number of samples (compared to the number of parameters), which is difficult to satisfy for many applications [40, 76]. To remedy the aforementioned issues, a special attention has been paid to the augmentation of these problems with sparsity-inducing penalty functions to obtain sparse and easy-to-analyze solutions.

Graphical lasso (GL) is one of the most commonly used techniques for estimating the inverse covariance matrix from a limited number of data samples [96, 21, 274].

2.2 Problem Formulation

Consider a random vector $\mathbf{x} = (x_1, x_2, ..., x_n)$ with a multivariate normal distribution. Let $\Sigma_* \in \mathbb{S}^n_+$ denote the covariance matrix associated with the vector $\mathbf{x}$. The inverse of the covariance matrix can be used to determine the conditional independence between the random variables $x_1, x_2, ..., x_n$. In particular, if the $(i,j)^{\text{th}}$ entry of Σ_*^{-1} is zero for two disparate indices i and j, then x_i and x_j are conditionally independent given the rest of the variables.

Definition 1. *Given a symmetric matrix $S \in \mathbb{S}^n$, the **support graph or sparsity graph** of S is defined as a graph with the vertex set $\mathcal{V} := \{1, 2, ..., n\}$ and the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ such that $(i,j) \in \mathcal{V}$ if and only if $S_{ij} \neq 0$, for every two different vertices $i, j \in \mathcal{V}$. The support graph of S captures the sparsity pattern of the matrix S and is denoted as $\mathcal{G}(S)$.*

Definition 2. *Given a graph $\mathcal{G}$, define $\mathcal{G}^{(c)}$ as the complement of $\mathcal{G}$, which is obtained by removing the existing edges of $\mathcal{G}$ and drawing an edge between every two vertices of $\mathcal{G}$ that were not originally connected.*

Definition 3. *Given two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ with the same vertex set, $\mathcal{G}_1$ is called a subgraph of $\mathcal{G}_2$ if the edge set of $\mathcal{G}_1$ is a subset of the edge set of $\mathcal{G}_2$. The notation $\mathcal{G}_1 \subseteq \mathcal{G}_2$ is used to denote this inclusion.*

The graph $\mathcal{G}\left(\Sigma_*^{-1}\right)$ (i.e., the sparsity graph of Σ_*^{-1}) represents a graphical model capturing the conditional independence between the elements of $\mathbf{x}$. Assume that Σ_* is nonsingular and that $\mathcal{G}\left(\Sigma_*^{-1}\right)$ is a sparse graph. Finding this graph is cumbersome in practice because the exact covariance matrix Σ_* is rarely known. More precisely, $\mathcal{G}\left(\Sigma_*^{-1}\right)$ should be constructed from a given sample covariance matrix (constructed from N samples), as opposed to Σ_*. Let Σ denote an arbitrary $n \times n$ positive-semidefinite matrix, which is provided as an estimate of Σ_*. Consider the convex optimization problem

$$\min_{S \in \mathbb{S}_+^n} -\log\det(S) + \operatorname{trace}(\Sigma S). \tag{2.1}$$

It is easy to verify that the optimal solution of the above problem is equal to $S^{\text{opt}} = \Sigma^{-1}$. However, there are two issues with this solution. First, since the number of samples available in many applications is small or modest compared to the dimension of Σ, the matrix Σ is ill-conditioned or even singular. Second, although Σ_*^{-1} is assumed to be sparse, a small random difference between Σ_* and Σ would make S^{opt} highly dense. In order to address the aforementioned issues, consider the problem

$$\min_{S \in \mathbb{S}_+^n} -\log\det(S) + \operatorname{trace}(\Sigma S) + \lambda \|S\|_{1,\text{off}}, \tag{2.2}$$

where $\lambda \in \mathbb{R}_+$ is a regularization parameter. This problem is referred to as *Graphical Lasso* (GL). Intuitively, the term $\|S\|_{1,\text{off}}$ in the objective function serves as a surrogate for promoting sparsity among the off-diagonal entries of S, while ensuring that the problem is well-defined even with a singular input Σ. Henceforth, the notation S^{opt} will be used to denote a solution of the GL instead of the unregularized optimization problem (2.1).

While the ℓ_1-regularized problem (2.2) is technically convex, it is commonly considered intractable for large-scale datasets. The decision variable is an $n \times n$ matrix, so simply fitting all $O(n^2)$ variables into memory is already a significant issue. General-purpose algorithms have either prohibitively high complexity or slow convergence. In practice, (2.2) is solved using problem-specific algorithms. The state-of-the-art include GLASSO [96], QUIC [126], and its "big-data" extension BIG-QUIC [125]. These algorithms use between $O(n)$ and $O(n^3)$ time and between $O(n^2)$ and $O(n)$ memory per iteration, but the number of iterations needed to converge to an accurate solution can be very large.

2.3 Related Work

Algorithms for GL. Algorithms for GL are usually based on some mixture of Newton [205], proximal Newton [125, 126], iterative thresholding [215], and (block) coordinate descent [96, 248]. All of these methods suffer fundamentally from the need to keep track and act on all $O(n^2)$ elements in the matrix variable S. Even if the final solution matrix were sparse with $O(n)$ nonzeros, it is still possible for the algorithm to traverse through a "dense region" in which the iterate S must be fully dense. Thresholding heuristics have

been proposed to address issue, but these may adversely affect the outer algorithm and prevent convergence. It is generally impossible to guarantee a figure lower than $O(n^2)$ time per-iteration, even if the solution contains only $O(n)$ nonzeros. Most of the algorithms mentioned above actually have worst-case per-iteration costs of $O(n^3)$.

GL and Thresholding. Recently, it has been empirically verified that a simple thresholding of the sample covariance matrix reveals the true sparsity pattern of the optimal solution to GL [238]. Despite its practical significance, the theoretical justification of this equivalence was unclear. Another line of work has been devoted to studying the connectivity structure of the optimal solution of the GL. In particular, [179] and [264] have shown that the connected components induced by thresholding the covariance matrix and those in the support graph of the optimal solution of the GL lead to the same vertex partitioning. Although this result does not require any particular condition, it cannot provide any information about the edge structure of the support graph and one needs to solve (2.2) for each connected component using an iterative algorithm, which may take up to $\mathcal{O}(n^3)$ per iteration [96, 21, 179].

GL with prior information. A number of approaches are available in the literature to introduce prior information to GL (also known as restricted GL, or RGL). The paper [72] introduced a class of RGL in which the true graphical model is assumed to have Laplacian structure. This structure commonly appears in signal and image processing [182]. For the *a priori* graph-based correlation structure described above, [112] introduced a *pathway* graphical lasso method similar to RGL.

2.4 GL and Thresholding

Suppose that it is known *a priori* that the true graph $\mathcal{G}\left(\Sigma_*^{-1}\right)$ has k edges, for some given number k. With no loss of generality, assume that all nonzero off-diagonal entries of Σ have different magnitudes. Two methods for finding an estimate of $\mathcal{G}\left(\Sigma_*^{-1}\right)$ are as follows:

- **Graphical Lasso:** We solve the optimization problem (2.2) repeatedly for different values of λ until a solution S^{opt} with exactly $2k$ nonzero off-diagonal entries are found.
- **Thresholding:** Without solving any optimization problem, we simply identify those $2k$ entries of Σ that have the largest magnitudes among all off-diagonal entries of Σ. We then replace the remaining $n^2 - n - 2k$ off-diagonal entries of Σ with zero and denote the thresholded sample covariance matrix as Σ_k. Note that Σ and Σ_k have the same diagonal entries. Finally, we consider the sparsity graph of Σ_k, namely $\mathcal{G}(\Sigma_k)$, as an estimate for $\mathcal{G}\left(\Sigma_*^{-1}\right)$.

Definition 4. *It is said that the sparsity structures of Graphical Lasso and thresholding are equivalent if there exists a regularization coefficient λ such that $\mathcal{G}(S^{opt}) = \mathcal{G}(\Sigma_k)$.*

In this section, it is aimed to understand under what conditions the easy-to-find graph $\mathcal{G}(\Sigma_k)$ is equal to the hard-to-obtain graph $\mathcal{G}(S^{\text{opt}})$, without having to solve the GL.

In particular, we derive sufficient conditions to guarantee that the GL and thresholding methods result in the same sparsity graph. These conditions are only dependent on λ and Σ, and are expected to hold whenever λ is large enough or a sparse graph is sought.

Definition 5. *A matrix $M \in \mathbb{S}^n$ is called **inverse-consistent** if there exists a matrix $N \in \mathbb{S}^n$ with zero diagonal elements such that*

$$\begin{aligned} &M + N \succ 0, \\ &\mathcal{G}(N) \subseteq (\mathcal{G}(M))^{(c)}, \\ &\mathcal{G}\left((M+N)^{-1}\right) \subseteq \mathcal{G}(M). \end{aligned}$$

*The matrix N is called **inverse-consistent complement** of M and is denoted as $M^{(c)}$.*

The next Lemma will shed light on the definition of inverse-consistency by introducing an important class of such matrices that satisfy this property, namely the *set of matrices with positive-definite completions*.

Lemma 1. *Any arbitrary matrix with positive-definite completion is inverse-consistent and has a unique inverse-consistent complement.*

Proof: Consider the optimization problem

$$\begin{aligned} \min_{S \in \mathbb{S}^n} \quad & \operatorname{trace}(MS) - \operatorname{logdet}(S) && \text{(2.4a)} \\ \text{subject to} \quad & S_{ij} = 0, \qquad \forall (i,j) \in (\mathcal{G}(M))^{(c)} && \text{(2.4b)} \\ & S \succeq 0, && \text{(2.4c)} \end{aligned}$$

and its dual

$$\begin{aligned} \max_{\Pi \in \mathbb{S}^n} \quad & \det(M + \Pi) && \text{(2.5a)} \\ \text{subject to} \quad & M + \Pi \succeq 0 && \text{(2.5b)} \\ & \mathcal{G}(\Pi) \subseteq (\mathcal{G}(M))^{(c)} && \text{(2.5c)} \\ & \Pi_{ii} = 0, \qquad i = 1, ..., n. && \text{(2.5d)} \end{aligned}$$

Note that Π_{ij} is equal to the Lagrange multiplier for (2.4b) and every $(i,j) \in (\mathcal{G}(M))^{(c)}$, and is zero otherwise. Since the matrix M has a positive-definite completion, the dual problem is strictly feasible. Moreover, $S = I_n$ is a feasible solution of (2.4). Therefore, strong duality holds and the primal solution is attainable. On the other hand, the objective function (2.4a) is strictly convex, which makes the solution of the primal problem unique. Let S^{opt} denote the globally optimal solution of (2.4). It follows from the first-order optimality conditions that

$$S^{\text{opt}} = (M + \Pi^{\text{opt}})^{-1}.$$

This implies that

$$\begin{aligned}&\mathcal{G}(\Pi^{\text{opt}}) \subseteq (\mathcal{G}(M))^{(c)}\\&\mathcal{G}((M+\Pi^{\text{opt}})^{-1}) \subseteq \mathcal{G}(M)\\&M+\Pi^{\text{opt}} \succ 0.\end{aligned}$$

As a result, $M \in \mathbb{S}^n$ is inverse-consistent and Π^{opt} is its complement. To prove the uniqueness of the inverse-consistent complement of M, let Π denote an arbitrary complement of M. It follows from Definition 5 and the first-order optimality conditions that $(M+\Pi)^{-1}$ is a solution of (2.4). Since S^{opt} is the unique solution of (2.4), it can be concluded that $\Pi = \Pi^{\text{opt}}$. This implies that M has a unique inverse-consistent complement. □

Remark 1. *Two observations can be made based on Lemma 1. First, the positive-definiteness of a matrix is sufficient to guarantee that it belongs to the cone of matrices with positive-definite completion. Therefore, positive-definite matrices are inverse-consistent. Second, upon existence, the inverse-consistent complement of a matrix with positive-definite completion is equal to the difference between the matrix and its unique* ***maximum determinant completion****.*

Definition 6. *An inverse-consistent matrix M is called* ***sign-consistent*** *if the (i,j) entries of M and $(M+M^{(c)})^{-1}$ are nonzero and have opposite signs for every $(i,j) \in \mathcal{G}(M)$.*

Example 1 (**An inverse- and sign-consistent matrix).** *To illustrate Definitions 5 and 6, consider the matrix*

$$M = \begin{bmatrix} 1 & 0.3 & 0 & 0 \\ 0.3 & 1 & -0.4 & 0 \\ 0 & -0.4 & 1 & 0.2 \\ 0 & 0 & 0.2 & 1 \end{bmatrix}.$$

The graph $\mathcal{G}(M)$ is a path graph with the vertex set $\{1,2,3,4\}$ and the edge set $\{(1,2),(2,3),(3,4)\}$. To show that M is inverse-consistent, let the matrix $M^{(c)}$ be chosen as

$$M^{(c)} = \begin{bmatrix} 0 & 0 & -0.120 & -0.024 \\ 0 & 0 & 0 & -0.080 \\ -0.120 & 0 & 0 & 0 \\ -0.024 & -0.080 & 0 & 0 \end{bmatrix}.$$

The inverse matrix $(M+M^{(c)})^{-1}$ is equal to

$$\begin{bmatrix} \frac{1}{0.91} & \frac{-0.3}{0.91} & 0 & 0 \\ \frac{-0.3}{0.91} & 1+\frac{0.09}{0.91}+\frac{0.16}{0.84} & \frac{0.4}{0.84} & 0 \\ 0 & \frac{0.4}{0.84} & 1+\frac{0.16}{0.84}+\frac{0.04}{0.96} & \frac{-0.2}{0.96} \\ 0 & 0 & \frac{-0.2}{0.96} & \frac{1}{0.96} \end{bmatrix}.$$

Observe that:

- *M and $M + M^{(c)}$ are both positive-definite.*
- *The sparsity graphs of M and $M^{(c)}$ are complements of each other.*
- *The sparsity graphs of M and $(M + M^{(c)})^{-1}$ are identical.*
- *The nonzero off-diagonal entries of M and $(M + M^{(c)})^{-1}$ have opposite signs.*

The above properties imply that M is both inverse-consistent and sign-consistent, and $M^{(c)}$ is its complement.

Definition 7. *Given a graph $\mathcal{G}$ and a scalar α, define $\beta(\mathcal{G}, \alpha)$ as the maximum of $\|M^{(c)}\|_{\max}$ over all matrices M with positive-definite completions and with the diagonal entries all equal to 1 such that $\mathcal{G}(M) = \mathcal{G}$ and $\|M\|_{\max} \leq \alpha$.*

Consider the dual solution Π^{opt} introduced in the proof of Lemma 1 and note that it is a function of M. Roughly speaking, the function $\beta(\mathcal{G}, \alpha)$ in the above definition provides an upper bound on $\|\Pi^{\text{opt}}\|_{\max}$ over all matrices M with positive-definite completions and with the diagonal entries equal to 1 such that $\mathcal{G}(M) = \mathcal{G}$ and $\|M\|_{\max} \leq \alpha$. As will be shown later, this function will be used as a *certificate* to verify the optimality conditions for the GL.

Since Σ_* is non-singular and we have a finite number of samples, the elements of the upper triangular part of Σ (excluding its diagonal elements) are all nonzero and distinct with probability one. Let $\sigma_1, \sigma_2, ..., \sigma_{n(n-1)/2}$ denote the absolute values of those upper-triangular entries such that

$$\sigma_1 > \sigma_2 > ... > \sigma_{n(n-1)/2} > 0.$$

Definition 8. *Consider an arbitrary positive regularization parameter λ that does not belong to the discrete set $\{\sigma_1, \sigma_2, ..., \sigma_{n(n-1)/2}\}$. Define the index k associated with λ as an integer number satisfying the relation $\lambda \in (\sigma_{k+1}, \sigma_k)$. If λ is greater than σ_1, then k is set to 0.*

Throughout this chapter, the index k refers to the number introduced in Definition 8, which depends on λ.

Definition 9. *Define the **residue of** Σ **relative to** λ as a matrix $\Sigma^{\text{res}}(\lambda) \in \mathbb{S}^n$ such that the (i, j) entry of $\Sigma^{\text{res}}(\lambda)$ is equal to $\Sigma_{ij} - \lambda \times \text{sign}(\Sigma_{ij})$ if $i \neq j$ and $|\Sigma_{ij}| > \lambda$, and it is equal to 0 otherwise. Furthermore, define **normalized residue of** Σ **relative to** λ as*

$$\tilde{\Sigma}^{\text{res}}(\lambda) = D^{-1/2} \times \Sigma^{\text{res}}(\lambda) \times D^{-1/2},$$

where D is diagonal matrix with $D_{ii} = \Sigma_{ii}$ for every $i \in \{1, ..., d\}$.

Notice that $\Sigma^{\text{res}}(\lambda)$ is in fact the soft-thresholded sample covariance matrix with the threshold λ. For notational simplicity, we will use Σ^{res} or $\tilde{\Sigma}^{\text{res}}$ instead of $\Sigma^{\text{res}}(\lambda)$ or $\tilde{\Sigma}^{\text{res}}(\lambda)$ whenever the equivalence is implied by the context. One of the main theorems of this chapter is presented below.

Theorem 1. *The sparsity structures of Σ^{res} and S^{opt} are equivalent if the following conditions are satisfied:*

- ***Condition 1-i:*** *$I_n + \tilde{\Sigma}^{\text{res}}$ has a positive-definite completion.*
- ***Condition 1-ii:*** *$I_n + \tilde{\Sigma}^{\text{res}}$ is sign-consistent.*
- ***Condition 1-iii:*** *The relation*

$$\beta\left(\mathcal{G}(\Sigma^{\text{res}}), \|\tilde{\Sigma}^{\text{res}}\|_{\max}\right) \leq \min_{\substack{i\neq j \\ |\Sigma_{ij}|\leq\lambda}} \frac{\lambda - |\Sigma_{ij}|}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$$

holds.

A number of observations can be made based on Theorem 1. First note that, due to Lemma 1, Condition (1-i) guarantees that $I_n + \tilde{\Sigma}^{\text{res}}$ is inverse-consistent; in fact it holds when $I_n + \tilde{\Sigma}^{\text{res}}$ itself is positive-definite. Note that the positive-definiteness of $I_n + \tilde{\Sigma}^{\text{res}}$ is guaranteed to hold if the eigenvalues of the normalized residue of the matrix Σ relative to λ are greater than -1. Recall that $\lambda \in (\sigma_{k+1}, \sigma_k)$ for some integer k and the off-diagonal entries of $I_n + \tilde{\Sigma}^{\text{res}}$ are in the range $[-1, 1]$. In the case where the number k is significantly smaller than n^2, the residue matrix has many zero entries. Hence, the satisfaction of Condition (1-i) is expected for a large class of residue matrices; this will be verified extensively in our case studies on the real-world and synthetically generated data sets. Specifically, this condition is automatically satisfied if $I_n + \tilde{\Sigma}^{\text{res}}$ is diagonally dominant. Conditions (1-ii) and (1-iii) of Theorem 1 are harder to check. These conditions depend on the support graph of the residue matrix $\tilde{\Sigma}^{\text{res}}$ and/or how small the nonzero entries of $\tilde{\Sigma}^{\text{res}}$ are. The next two lemmas further analyze these conditions to show that they are expected to be satisfied for large λ.

Lemma 2. *Given an arbitrary graph $\mathcal{G}$, there is a strictly positive constant number $\zeta(\mathcal{G})$ such that*

$$\beta(\mathcal{G}, \alpha) \leq \zeta(\mathcal{G})\alpha^2, \qquad \forall\ \alpha \in (0, 1) \tag{2.7}$$

and therefore, Condition (1-iii) is reduced to

$$\zeta(\mathcal{G}(\Sigma^{\text{res}})) \times \max_{\substack{k\neq l \\ |\Sigma_{kl}|>\lambda}} \left(\frac{|\Sigma_{kl}| - \lambda}{\sqrt{\Sigma_{kk}\Sigma_{ll}}}\right)^2 \leq \min_{\substack{i\neq j \\ |\Sigma_{ij}|\leq\lambda}} \frac{\lambda - |\Sigma_{ij}|}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}.$$

Lemma 3. *Consider a matrix M with a positive-definite completion and with unit diagonal entries. Define $\alpha = \|M\|_{\max}$ and $\mathcal{G} = \mathcal{G}(M)$. There exist strictly positive constant numbers $\alpha_0(\mathcal{G})$ and $\gamma(\mathcal{G})$ such that M is sign-consistent if $\alpha \leq \alpha_0(\mathcal{G})$ and the absolute value of the off-diagonal nonzero entries of M is lower bounded by $\gamma(\mathcal{G})\alpha^2$. This implies that Condition (i-ii) is satisfied if $\|\tilde{\Sigma}^{\text{res}}\|_{\max} \leq \alpha_0(\mathcal{G}(\Sigma^{\text{res}}))$ and*

$$\gamma(\mathcal{G}(\Sigma^{\text{res}})) \times \max_{\substack{k\neq l \\ |\Sigma_{kl}|>\lambda}} \left(\frac{|\Sigma_{kl}|-\lambda}{\sqrt{\Sigma_{kk}\Sigma_{ll}}}\right)^2 \leq \min_{\substack{i\neq j \\ |\Sigma_{ij}|>\lambda}} \frac{|\Sigma_{ij}|-\lambda}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}. \tag{2.8}$$

For simplicity of notation, define $r = \frac{\max_i \Sigma_{ii}}{\min_j \Sigma_{jj}}$ and $\Sigma_{\max} = \max_i \Sigma_{ii}$. Assuming that $\|\tilde{\Sigma}^{\text{res}}\|_{\max} \leq \alpha_0(\mathcal{G}(\Sigma^{\text{res}}))$, Conditions (1-ii) and (1-iii) of Theorem 1 are guaranteed to be satisfied if

$$\zeta(\mathcal{G}(\Sigma^{\text{res}})) \leq \frac{1}{r^2} \cdot \frac{\frac{\lambda-\sigma_{k+1}}{\Sigma_{\max}}}{\left(\frac{\sigma_1-\lambda}{\Sigma_{\max}}\right)^2}, \qquad \gamma(\mathcal{G}(\Sigma^{\text{res}})) \leq \frac{1}{r^2} \cdot \frac{\frac{\sigma_k-\lambda}{\Sigma_{\max}}}{\left(\frac{\sigma_1-\lambda}{\Sigma_{\max}}\right)^2}, \tag{2.9}$$

which is equivalent to

$$\max\left\{\gamma(\mathcal{G}(\Sigma^{\text{res}})), \zeta(\mathcal{G}(\Sigma^{\text{res}}))\right\} \leq \frac{2}{r^2} \cdot \frac{\frac{\sigma_k-\sigma_{k+1}}{\Sigma_{\max}}}{\left(\frac{2\sigma_1-\sigma_k-\sigma_{k+1}}{\Sigma_{\max}}\right)^2}.$$

for the choice $\lambda = \frac{\sigma_k+\sigma_{k+1}}{2}$. Consider the set

$$\mathcal{T} = \left\{|\Sigma_{ij}| \;\middle|\; i = 1, 2, ..., n-1,\ j = i+1, ..., n\right\}.$$

This set has $\frac{n(n-1)}{2}$ elements. The cardinality of $\{\sigma_1, ..., \sigma_{n-1}\}$, as a subset of $\mathcal{T}$, is smaller than the cardinality of $\mathcal{T}$ by a factor of $\frac{n}{2}$. Combined with the fact that $|\sigma_i| < \Sigma_{\max}$ for every $i = 1, ..., \frac{n(n-1)}{2}$, this implies that the term $\frac{2\sigma_1-\sigma_{n-1}-\sigma_n}{\Sigma_{\max}}$ is expected to be small and its square is likely to be much smaller than 1, provided that the elements of $\mathcal{T}$ are sufficiently spread. If the number $(2\sigma_1 - \sigma_{n-1} - \sigma_n)$ is relatively smaller than the gap $\sigma_{n-1} - \sigma_n$ and $k = O(n)$, then (2.7) and as a result Conditions (1-ii) and (1-iii) would be satisfied. The satisfaction of this condition will be studied for acyclic graphs in the next section.

2.5 Closed-form Solution: Acyclic Sparsity Graphs

In the previous subsection, we provided a set of sufficient conditions for the equivalence of the GL and thresholding methods. Although these conditions are merely based on the known parameters of the problem, i.e., the regularization coefficient and sample covariance matrix, their verification is contingent upon knowing the value of $\beta(\mathcal{G}(\Sigma^{\text{res}}), \|\tilde{\Sigma}^{\text{res}}\|_{\max})$ and whether $I_n + \tilde{\Sigma}^{\text{res}}$ is sign-consistent and has a positive-definite completion. The objective of this part is to greatly simplify the conditions in the case where the thresholded sample covariance

matrix has an acyclic support graph. First, notice that if $I_n + \tilde{\Sigma}^{\text{res}}$ is positive-definite, it has a trivial positive-definite completion. Furthermore, we will prove that $\zeta(\text{supp}(\Sigma^{\text{res}}))$ in Lemma 2 is equal to 1 when $\text{supp}(\Sigma^{\text{res}})$ is acyclic. This reduces Condition (1-iii) to the simple inequality

$$\|\tilde{\Sigma}^{\text{res}}\|_{\max}^2 \leq \min_{\substack{i\neq j \\ |\Sigma_{ij}|\leq\lambda}} \frac{\lambda - |\Sigma_{ij}|}{\sqrt{\Sigma_{ii}\Sigma_{jj}}},$$

which can be verified efficiently and is expected to hold in practice (see Section **??**). Then, we will show that the sign-consistency of $I_n + \tilde{\Sigma}^{\text{res}}$ is automatically implied by the fact that it has a positive-definite completion if $\text{supp}(\Sigma^{\text{res}})$ is acyclic.

Lemma 4. *Given an arbitrary acyclic graph $\mathcal{G}$, the relation*

$$\beta(\mathcal{G}, \alpha) \leq \alpha^2 \tag{2.10}$$

holds for every $0 \leq \alpha < 1$. Furthermore, strict equality holds for (2.10) *if $\mathcal{G}$ includes a path of length at least 2.*

Lemma 4 is at the core of our subsequent arguments. It shows that the function $\beta(\mathcal{G}, \alpha)$ has a simple and explicit formula since its inverse-consistent complement can be easily obtained. Furthermore, it will be used to derive *approximate* inverse-consistent complement of the matrices with sparse, but not necessarily acyclic support graphs.

Lemma 5. *Condition (1-ii) of Theorem 1 is implied by its Condition (1-i) if the graph $\mathcal{G}(\Sigma^{\text{res}})$ is acyclic.*

Proof: Consider an arbitrary matrix $M \in \mathbb{S}^n$ with a positive-definite completion. It suffices to show that if $\mathcal{G}(M)$ is acyclic, then M is sign-consistent. To this end, consider the matrix Π^{opt} introduced in the proof of Lemma 1, which is indeed the unique inverse-consistent complement of M. For an arbitrary pair $(i, j) \in \mathcal{G}(M)$, define a diagonal matrix $\Phi \in \mathbb{S}^n$ as follows:

- Consider the graph $\mathcal{G}(M)\backslash\{(i, j)\}$, which is obtained from the acyclic graph $\mathcal{G}(M)$ by removing its edge (i, j). The resulting graph is disconnected because there is no path between nodes i and j.
- Divide the disconnected graph $\mathcal{G}(M)\backslash\{(i, j)\}$ into two groups 1 and 2 such that group 1 contains node i and group 2 includes node 2.
- For every $l \in \{1, ..., n\}$, define Φ_{ll} as 1 if l is in group 1, and as -1 otherwise.

In light of Lemma 1, $(M+\Pi)^{-1}$ is the unique solution of (2.4). Similarly, $\Phi(M+\Pi)^{-1}\Phi$ is a feasible point for (2.4). As a result, the following inequality must hold

$$\Big\{\text{trace}(M(M+\Pi^{\text{opt}})^{-1}) - \text{logdet}((M+\Pi^{\text{opt}})^{-1})\Big\}$$
$$-\Big\{\text{trace}(M\Phi(M+\Pi^{\text{opt}})^{-1}\Phi) - \text{logdet}(\Phi(M+\Pi^{\text{opt}})^{-1}\Phi)\Big\} < 0.$$

It is easy to verify that the left side of the above inequality is equal to twice the product of the (i,j) entries of M and $(M+\Pi)^{-1}$. This implies that the (i,j) entries of M and $(M+\Pi)^{-1}$ have opposite signs. As a result, M is sign-consistent. □

Definition 10. *Define $T(\lambda)$ as a $n \times n$ symmetric matrix whose $(i,j)^{th}$ entry is equal to $\Sigma_{ij} + \lambda \times sign(S^{opt}_{ij})$ for every $(i,j) \in supp(S^{opt})$, and it is equal to zero otherwise.*

The next result of this chapter is a consequence of Lemmas 4 and 5 and Theorem 1.

Theorem 2. *Assume that the graph $supp(S^{opt})$ is acyclic and the matrix $D+T(\lambda)$ is positive-definite. Then, the relation $\mathcal{E}^{opt} \subseteq \mathcal{E}^{\text{res}}$ holds and the optimal solution S^{opt} of the GL can be computed via the explicit formula*

$$S^{opt}_{ij} = \begin{cases} \frac{1}{\Sigma_{ii}}\left(1 + \sum\limits_{(i,m)\in\mathcal{E}^{opt}} \frac{(\Sigma^{\text{res}}_{im})^2}{\Sigma_{ii}\Sigma_{mm}-(\Sigma^{\text{res}}_{im})^2}\right) & \text{if} \quad i=j, \\ \frac{-\Sigma^{\text{res}}_{ij}}{\Sigma_{ii}\Sigma_{jj}-(\Sigma^{\text{res}}_{ij})^2} & \text{if} \quad (i,j)\in\mathcal{E}^{opt}, \\ 0 & \text{otherwise}, \end{cases} \tag{2.11}$$

where $\mathcal{E}^{opt}$ and $\mathcal{E}^{\text{res}}$ denote the edge sets of $\mathcal{G}(S^{opt})$ and $\mathcal{G}(\Sigma^{\text{res}})$, respectively.

When the regularization parameter λ is large, the graph $\text{supp}(S^{\text{opt}})$ is expected to be sparse and possibly acyclic. In this case, the matrix $T(\lambda)$ is sparse with small nonzero entries. If $D+T(\lambda)$ is positive-definite and $\text{supp}(S^{\text{opt}})$ is acyclic, Theorem 2 reveals two important properties of the solution of the GL: 1) its support graph is contained in the sparsity graph of the thresholded sample covariance matrix, and 2) the entries of this matrix can be found using the explicit formula (2.11). However, this formula requires to know the locations of the nonzero elements of S^{opt}. In what follows, we will replace the assumptions of the above theorem with easily verifiable rules that are independent from the optimal solution S^{opt} or the locations of its nonzero entries. Furthermore, it will be shown that these conditions are expected to hold when λ is large enough, i.e., if a sparse matrix S^{opt} is sought.

Theorem 3. *Assume that the following conditions are satisfied:*

- ***Condition 2-i.*** *The graph $supp(\Sigma^{\text{res}})$ is acyclic.*
- ***Condition 2-ii.*** *$I_n + \tilde{\Sigma}^{\text{res}}$ is positive-definite.*

- ***Condition 2-iii.*** $\|\tilde{\Sigma}^{\text{res}}\|_{\max}^2 \leq \min_{\substack{i\neq j \\ |\Sigma_{ij}|\leq\lambda}} \frac{\lambda-|\Sigma_{ij}|}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$.

Then, the sparsity pattern of the optimal solution S^{opt} corresponds to the sparsity pattern of Σ^{res} and, in addition, S^{opt} can be obtained via the explicit formula (2.11).

The above theorem states that if a sparse graph is sought, then as long as some easy-to-verify conditions are met, there is an explicit formula for the optimal solution. It will later be shown that Condition (2-i) is exactly or approximately satisfied if the regularization coefficient is sufficiently large. Condition (2-ii) implies that the eigenvalues of the normalized residue of Σ with respect to λ should be greater than -1. This condition is expected to be automatically satisfied since most of the elements of $\tilde{\Sigma}^{\text{res}}$ are equal to zero and the nonzero elements have small magnitude. In particular, this condition is satisfied if $I_n + \tilde{\Sigma}^{\text{res}}$ is diagonally dominant. Finally, using (4.58), it can be verified that Condition (2-iii) is satisfied if

$$\frac{\left(\frac{2\sigma_1-\sigma_k-\sigma_{k+1}}{\Sigma_{\max}}\right)^2}{\frac{\sigma_k-\sigma_{k+1}}{\Sigma_{\max}}} \leq \frac{2}{r^2}. \tag{2.12}$$

Similar to the arguments made in the previous subsection, (4.40) shows that Condition (2-iii) is satisfied if $\frac{2\sigma_1-\sigma_k-\sigma_{k+1}}{\Sigma_{\max}}$ is small. This is expected to hold in practice since the choice of λ entails that $2\sigma_1 - \sigma_k - \sigma_{k+1}$ is much smaller than $\Sigma_{\max}$. Under such circumstances, one can use Theorem 3 to obtain the solution of the GL without having to solve (2.2) numerically.

Having computed the sample covariance matrix, we will next show that checking the conditions in Theorem 3 and finding S^{opt} using (2.11) can all be carried out efficiently.

Corollary 1. *Given Σ and λ, the total time complexity of checking the conditions in Theorem 3 and finding S^{opt} using* (2.11) *is $\mathcal{O}(n^2)$.*

Another line of work has been devoted to studying the connectivity structure of the optimal solution of the GL. In particular, [179] and [264] have shown that the connected components induced by thresholding the covariance matrix and those in the support graph of the optimal solution of the GL lead to the same vertex partitioning. Although this result does not require any particular condition, it cannot provide any information about the edge structure of the support graph and one needs to solve (2.2) for each connected component using an iterative algorithm, which may take up to $\mathcal{O}(n^3)$ per iteration [96, 21, 179]. Corollary 1 states that this complexity could be reduced significantly for sparse graphs.

Remark 2. *The results introduced in Theorem 1 can indeed be categorized as a set of "safe rules" that correctly determine sparsity pattern of the optimal solution of the GL. These rules are subsequently reduced to a set of easily verifiable conditions in Theorem 3 to safely obtain the correct sparsity pattern of the acyclic components in the optimal solution. On the other hand, there is a large body of literature on simple and cheap safe rules to pre-screen and simplify the sparse learning and estimation problems, including Lasso, logistic*

regression, support vector machine, group Lasso, etc [103, 246, 90, 195]. Roughly speaking, these methods are based on constructing a sequence of safe regions that encompass the optimal solution for the dual of the problem at hand. These safe regions, together with the Karush—Kuhn—Tucker (KKT) conditions, give rise to a set of rules that facilitate inferring the sparsity pattern of the optimal solution. Our results are similar to these methods since we also analyze the special structure of the KKT conditions and resort to the dual of the GL to obtain the correct sparsity structure of the optimal solution. However, according to the seminal work [195], most of the developed results on safe screening rules rely on strong Lipschitz assumptions on the objective function; an assumption that is violated in the GL. This calls for a new machinery to derive theoretically correct rules for this problem; a goal that is at the core of Theorems 1 and 3.

2.6 Approximate Closed-form Solution: Sparse Graphs

In the preceding subsection, it was shown that, under some mild assumptions, the GL has an explicit closed-form solution if the support graph of the thresholded sample covariance matrix is acyclic. In this part, a similar approach will be taken to find approximate solutions of the GL with an arbitrary underlying sparsity graph. In particular, by closely examining the hard-to-check conditions of Theorem 1, a set of simple and easy-to-verify surrogates will be introduced which give rise to an approximate closed-form solution for the general sparse GL. Furthermore, we will derive a strong upper bound on the approximation error and show that it decreases exponentially fast with respect to the length of the minimum-length cycle in the support graph of the thresholded sample covariance matrix. Indeed, the formula obtained earlier for acyclic graphs could be regarded as a by-product of this generalization since the length of the minimum-length cycle can be considered as infinity for such graphs. The significance of this result is twofold:

- Recall that the support graph corresponding to the optimal solution of the GL is sparse (but not necessarily acyclic) for a large regularization coefficient. In this case, the approximate error is provably small and the derived closed-form solution can serve as a good approximation for the exact solution of the GL. This will later be demonstrated in different simulations.

- The performance and runtime of numerical (iterative) algorithms for solving the GL heavily depend on their initializations. It is known that if the initial point is chosen close enough to the optimal solution, these algorithms converge to the optimal solution in just a few iterations [96, 126, 277]. The approximate closed-form solution designed in this chapter can be used as an initial point for the existing numerical algorithms to significantly improve their runtime.

The proposed approximate solution for the GL with an arbitrary support graph has the following form:

$$A_{ij} = \begin{cases} \frac{1}{\Sigma_{ii}} \left(1 + \sum\limits_{(i,m)\in\mathcal{E}^{\text{opt}}} \frac{(\Sigma^{\text{res}}_{im})^2}{\Sigma_{ii}\Sigma_{mm} - (\Sigma^{\text{res}}_{im})^2} \right) & \text{if} \quad i = j, \\ \frac{-\Sigma^{\text{res}}_{ij}}{\Sigma_{ii}\Sigma_{jj} - (\Sigma^{\text{res}}_{ij})^2} & \text{if} \quad (i,j) \in \mathcal{E}^{\text{res}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.13}$$

The definition of this matrix does not make any assumption on the structure of the graph $\mathcal{E}^{\text{res}}$. Recall that Σ^{res} in the above formula is the shorthand notation for $\Sigma^{\text{res}}(\lambda)$. As a result, the matrix A is a function of λ. To prove that the above matrix is an approximate solution of the GL, a few steps need to be taken. First, recall that—according to the proof of Lemma 4—it is possible to explicitly build the inverse-consistent complement of the thresholded sample covariance matrix if its sparsity graph is acyclic. This matrix serves as a *certificate* to confirm that the explicit solution (2.13) indeed satisfies the KKT conditions for the GL. By adopting a similar approach, it will then be proved that if the support graph of the thresholded sample covariance matrix is sparse, but not necessarily acyclic, one can find an approximate inverse-consistent complement of the proposed closed-form solution to approximately satisfy the KKT conditions.

Definition 11. *Given a number $\epsilon \geq 0$, a $n \times n$ matrix B is called an ϵ-**relaxed inverse** of matrix A if $A \times B = I_n + E$ such that $|E_{ij}| \leq \epsilon$ for every $(i,j) \in \{1,2,...,n\}^2$.*

The next lemma offers optimality (KKT) conditions for the unique solution of the GL.

Lemma 6 ([238])**.** *A matrix S^{opt} is the optimal solution of the GL if and only if it satisfies the following conditions for every $i,j \in \{1,2,...,n\}$*

$$(S^{opt})^{-1}_{ij} = \Sigma_{ij} \qquad \text{if} \quad i = j, \tag{2.14a}$$

$$(S^{opt})^{-1}_{ij} = \Sigma_{ij} + \lambda \times sign(S^{opt}_{ij}) \qquad \text{if} \quad S^{opt}_{ij} \neq 0, \tag{2.14b}$$

$$\Sigma_{ij} - \lambda \leq (S^{opt})^{-1}_{ij} \leq \Sigma_{ij} + \lambda \qquad \text{if} \quad S^{opt}_{ij} = 0, \tag{2.14c}$$

where $(S^{opt})^{-1}_{ij}$ denotes the $(i,j)^{th}$ entry of $(S^{opt})^{-1}$.

The following definition introduces a relaxed version of the first-order optimality conditions given in (2.14).

Definition 12. *Given a number $\epsilon \geq 0$, it is said that the $n \times n$ matrix A satisfies the ϵ-**relaxed KKT conditions** for the GL problem if there exists a $n \times n$ matrix B such that*

- *B is an ϵ-relaxed inverse of the matrix A.*

- *The pair* (A, B) *satisfies the conditions*

$$B_{ij} = \Sigma_{ij} \qquad \text{if} \quad i = j, \tag{2.15a}$$
$$|B_{ij} - (\Sigma_{ij} + \lambda \times \text{sign}(A_{ij}))\,| \leq \epsilon \qquad \text{if} \quad A_{ij} \neq 0, \tag{2.15b}$$
$$|B_{ij} - \Sigma_{ij}| \leq \lambda + \epsilon \qquad \text{if} \quad A_{ij} = 0. \tag{2.15c}$$

By leveraging the above definition, the objective is to prove that the explicit solution introduced in (2.13) satisfies the ϵ-relaxed KKT conditions for some number ϵ to be defined later.

Definition 13. *Given a graph* $\mathcal{G}$*, define the function* $c(\mathcal{G})$ *as the length of the minimum-length cycle of* $\mathcal{G}$ *(the number* $c(\mathcal{G})$ *is set to* $+\infty$ *if* $\mathcal{G}$ *is acyclic). Let* $\deg(\mathcal{G})$ *refer to the maximum degree of* $\mathcal{G}$*. Furthermore, define* $\mathcal{P}_{ij}(\mathcal{G})$ *as the set of all simple paths between nodes* i *and* j *in* $\mathcal{G}$*, and denote the maximum of* $|\mathcal{P}_{ij}(\mathcal{G})|_0$ *over all pairs* (i, j) *as* $P_{\max}(\mathcal{G})$*.*

Define $\Sigma_{\max}$ and $\Sigma_{\min}$ as the maximum and minimum diagonal elements of Σ, respectively.

Theorem 4. *Under the assumption* $\lambda < \sigma_1$*, the explicit solution* (2.13) *satisfies the* ϵ*-relaxed KKT conditions for the GL with* ϵ *chosen as*

$$\epsilon = \max\left\{\Sigma_{\max}, \sqrt{\frac{\Sigma_{\max}}{\Sigma_{\min}}}\right\} \cdot \delta \cdot (P_{\max}(\mathcal{G}(\Sigma^{\text{res}})) - 1) \cdot \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\left\lceil \frac{c(\mathcal{G}(\Sigma^{\text{res}}))}{2} \right\rceil}, \tag{2.16}$$

where

$$\delta = 1 + \frac{\deg(\mathcal{G}(\Sigma^{\text{res}})) \cdot \|\tilde{\Sigma}^{\text{res}}\|_{\max}^2}{1 - \|\tilde{\Sigma}^{\text{res}}\|_{\max}^2} + \frac{(\deg(\mathcal{G}(\Sigma^{res})) - 1)}{1 - \|\tilde{\Sigma}^{\text{res}}\|_{\max}^2}, \tag{2.17}$$

if the following conditions are satisfied:

- ***Condition 3-i.*** $I_n + \tilde{\Sigma}^{\text{res}}$ *is positive-definite.*
- ***Condition 3-ii.*** $\|\tilde{\Sigma}^{\text{res}}\|_{\max}^2 \leq \min\limits_{\substack{i \neq j \\ (i,j) \notin \mathcal{G}(\Sigma^{\text{res}})}} \frac{\lambda - |\Sigma_{ij}|}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$.

The number ϵ given in Theorem 4 is comprised of different parts:

- $\|\tilde{\Sigma}^{\text{res}}\|_{\max}$: Notice that $\|\tilde{\Sigma}^{\text{res}}\|_{\max}$ is strictly less than 1 and λ is large when a sparse graph is sought. Therefore, $\|\tilde{\Sigma}^{\text{res}}\|_{\max}$ is expected to be small for sparse graphs. Under this assumption, we have $0 \leq \|\tilde{\Sigma}^{\text{res}}\|_{\max} \ll 1$.

- $c(\mathcal{G}(\Sigma^{\text{res}}))$: It is straightforward to verify that $c(\mathcal{G}(\Sigma^{\text{res}}))$ is a non-decreasing function of λ. This is due to the fact that as λ increases, $\Sigma^{\text{res}}(\lambda)$ becomes sparser and this results in a support graph with fewer edges. In particular, if $n \geq 3$, then $c(\mathcal{G}(\Sigma^{\text{res}})) = 3$ for $\lambda = 0$ and $c(\mathcal{G}(\Sigma^{\text{res}})) = +\infty$ for $\lambda = \sigma_1$ almost surely.

- $P_{\max}(\mathcal{G}(\Sigma^{\text{res}}))$ and $\deg(\mathcal{G}(\Sigma^{\text{res}}))$: These two parameters are also non-decreasing functions of λ and likely to be small for large λ. For a small λ, the numbers $P_{\max}(\mathcal{G}(\Sigma^{\text{res}}))$ and $\deg(\mathcal{G}(\Sigma^{\text{res}}))$ could be on the order of $\mathcal{O}(n!)$ and $\mathcal{O}(n)$, respectively. However, these values are expected to be small for sparse graphs. In particular, it is easy to verify that for nonempty and acyclic graphs, $P_{\max}(\mathcal{G}(\Sigma^{\text{res}})) = 1$.

The above observations imply that if λ is large enough and the support graph of Σ^{res} is sparse, (2.13) serves as a good approximation of the optimal solution of the GL. In other words, it results from (2.16) that if $\text{supp}(\Sigma^{\text{res}})$ has a structure that is close to an acyclic graph, i.e., it has only a few cycles with moderate lengths, we have $\epsilon \approx 0$. In Section **??**, we will present illustrative examples to show the accuracy of the closed-form approximate solution with respect to the size of the cycles in the sparsity graph.

Consider the matrix A given in (2.13), and let $\mu_{\min}(A)$ and $\mu_{\max}(A)$ denote its minimum and maximum eigenvalues, respectively. If $\lambda = \sigma_1$, then $A = D^{-1}$ (recall that D collects the diagonal entries of Σ) and subsequently $\mu_{\min}(A) > 0$. Since $\mu_{\min}(\cdot)$ is a continuous function of λ, there exists a number λ_0 in the interval $(0, 1)$ such that the matrix A (implicitly defined based on λ) is positive-definite for every $\lambda \geq \lambda_0$. The following theorem further elaborates on the connection between the closed-form formula and the optimal solution of the GL.

Theorem 5. *There exists an strictly positive number λ_0 such that, for every $\lambda \geq \lambda_0$, the matrix A given in* (2.13) *is the optimal solution of the GL problem after replacing Σ with some perturbed matrix $\hat{\Sigma}$ that satisfies the inequality*

$$\left|\left|\left|\Sigma - \hat{\Sigma}\right|\right|\right|_2 \leq d_{\max}(A) \left(\frac{1}{\mu_{\min}(A)} + 1\right) \epsilon, \tag{2.18}$$

where $d_{\max}(A)$ is the maximum vertex cardinality of the connected components in the graph $\mathcal{G}(A)$ and ϵ is given in (2.16)*. Furthermore, 2.18 implies that*

$$f(A) - f^* \leq \left(\mu_{\max}(A) + \mu_{\max}(S^{\text{opt}})\right) d_{\max}(A) \left(\frac{1}{\mu_{\min}(A)} + 1\right) \epsilon, \tag{2.19}$$

where $f(A)$ and f^* are the objective functions of the GL evaluated at A and the optimal solution, respectively.

As mentioned before, if a sparse solution is sought for the GL, the regularization coefficient would be large and this helps with the satisfaction of the inequality $\lambda \geq \lambda_0$. In fact, it will be shown through different simulations that λ_0 is small in practice and hence, this condition is not restrictive. Under this circumstance, Theorem 5 states that the easy-to-construct matrix A is 1) the exact optimal solution of the GL problem with a perturbed sample covariance matrix, and 2) it is the approximate solution of the GL with the original sample covariance matrix. The magnitudes of this perturbation and approximation error are a function of $d_{\max}(A)$, $\mu_{\min}(A)$, $\mu_{\max}(A)$, $\mu_{\max}(S^{\text{opt}})$, and ϵ. Furthermore, it should be clear that A and ϵ are functions of λ and Σ (we dropped this dependency for simplicity of

notation). Recall that the disjoint components (or the vertex partitions) of supp(A) satisfy a nested property: given $1 \geq \lambda_1 > \lambda_2 \geq 0$, the components of supp(A) for $\lambda = \lambda_1$ are nested within the components of supp(A) for $\lambda = \lambda_2$ (see [179] for a simple proof of this statement). This implies that $d_{\max}(A)$ is a decreasing function of λ. In particular, it can be observed that $d_{\max}(A) = d$ if $\lambda = 0$ and $d_{\max}(A) = 1$ if $\lambda = \sigma_1$. Now, consider $\mu_{\min}(A)$, $\mu_{\max}(A)$, and $\mu_{\max}(S^{\text{opt}})$. First, note that if $\lambda = \sigma_1$, then $A = S^{\text{opt}} = D^{-1}$. Furthermore, it is easy to verify that both A and S^{opt} are continuous functions of λ. Therefore, for large values of λ, $\mu_{\min}(A)$, $\mu_{\max}(A)$, and $\mu_{\max}(S^{\text{opt}})$ are expected to be close to $1/\Sigma_{\max}$, $1/\Sigma_{\min}$, and $1/\Sigma_{\min}$, respectively. In addition, as discussed earlier, ϵ is a decreasing function of λ and vanishes when λ is large enough. Based on these observations, it can be concluded that the upper bound presented in (2.18) is small if λ is chosen to be large.

Notice that although the aforementioned value of ϵ in (2.16) and the upper bound in (2.18) were essential in the study of the effect of the sparsity of the support graph on the accuracy of the presented closed-form solution, they are conservative in practice. These numbers may be tightened significantly for specific sample covariance matrices. We will further discuss the approximation error of the closed-form solution in Section 2.7.

Warm-start algorithm As delineated before, one of the main strengths of the proposed closed-form solution is that it can be used as an initial point (warm-start) for the numerical algorithms specialized for solving the GL. To this goal, the following warm-start procedure is proposed.

Algorithm 1 Warm-start algorithm

input: data samples ($\mathbf{x}$), and regularization coefficient (λ)
output: Solution of the GL (S^{opt})
Obtain the residue matrix Σ^{res} based on Definition 9 and the closed-form solution A from (2.13)
for each component i in $\mathcal{G}(\Sigma^{\text{res}})$ **do**
  if Conditions 2-i, 2-ii, 2-iii are satisfied **then**
    $S^{\text{opt}}[i] \leftarrow A[i]$
  else
    Find $S^{\text{opt}}[i]$ by numerically solving the GL for component i with initial point $A[i]$
  end if
end for

In the above algorithm, $S^{\text{opt}}[i]$ and $A[i]$ are the submatrices of S^{opt} and A corresponding to the i^{th} component of $\mathcal{G}(\Sigma_{\text{res}})$. The warm-start algorithm is based on the key fact that the GL decomposes over the disjoint components of $\mathcal{G}(\Sigma_{\text{res}})$ [179, 264]. In particular, in the first step, the warm-start algorithm obtains the residue matrix according to Definition 9. Next, for every disjoint component of the residue matrix, if its support graph is acyclic and the conditions of Theorem 3 are satisfied, then the corresponding component in S^{opt} is found

using the closed-form solution (2.11). Otherwise, this closed-form solution is provided as an initial point to a numerical algorithm, such as GLASSO and QUIC [96, 126], in order to boost the runtime of solving the GL for the considered component. The results of the warm-start algorithm will be evaluated in the next section.

Remark 3. *The statistical analysis of the GL entails that λ should converge to zero as the number of samples grows to infinity. It is worthwhile to mention that our results may not be applicable in the high sampling regime, where λ is close to zero and consequently the thresholded sample covariance matrix is dense. However, notice that the main strength of the GL lies in the high dimensional-low sampling regime where n is much smaller than n and is in the order of $\log d$. Under such circumstances, the proposed explicit formula results in highly accurate solutions for the GL. In fact, it will be shown through massive-scale simulations that in practice, the required conditions on λ—such as the positive-definiteness of $I_n + \tilde{\Sigma}^{\text{res}}$—for the validity of the presented results are much more relaxed than the known conditions on λ to guarantee the statistical consistency of the GL.*

2.7 Numerical Results

In this section, we will demonstrate the effectiveness of the proposed methods on synthetically generated data, as well as on real data collected from the brain networks and transportation systems.

Case Study on Synthetic Data

Given a nonnegative number ω, consider an arbitrary sample covariance matrix Σ with the following properties:

- Its diagonal elements are normalized to 1.
- The entries corresponding to an arbitrary spanning tree of $\text{supp}(\Sigma)$ belong to the union of the intervals $[-0.85, -0.95]$ and $[0.85, 0.95]$.
- The off-diagonal entries that do not belong to the spanning tree are in the interval $[-0.85 + \omega, 0.85 - \omega]$.

The goal is to find conditions on λ, ω and the size of the covariance matrix such that Theorem 3 can be used to obtain a closed-form solution for the GL problem. One can choose the value of λ to be greater than σ_n to ensure that the graph $\text{supp}(\Sigma^{\text{res}})$ is acyclic. In particular, if we pick λ in the interval (σ_n, σ_{n-1}), the graph $\text{supp}(\Sigma^{\text{res}})$ becomes a spanning tree.

Select λ as $0.85 - \epsilon$ for a sufficiently small number ϵ and consider Condition (2-ii) in Theorem 3. One can easily verify that $I_n + \Sigma^{\text{res}}$ is positive-definite if the inequality $\frac{1}{\deg(v)} > (\sigma_1 - \lambda)^2$ holds for every node v in $\text{supp}(\Sigma^{\text{res}})$, where $\deg(v)$ is the degree of node v. This condition is guaranteed to be satisfied for all possible acyclic graphs if $(\deg(v))(0.95-0.85)^2 < 1$ or equivalently $\deg(v) \leq 100$ for every node v. Regarding Condition (2-iii), it can be

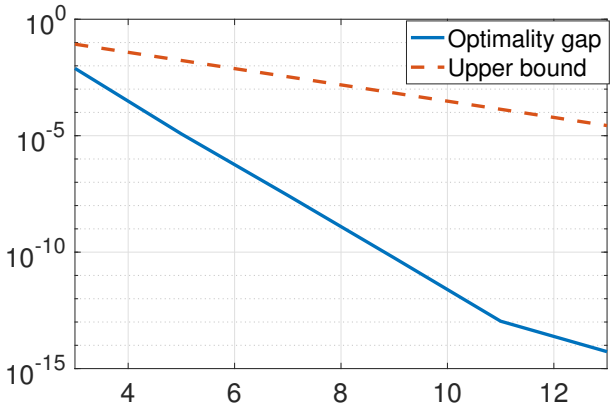


Figure 2.1: The optimality gap between the closed-form and optimal solutions for the GL

observed that the relation $(\sigma_1 - \lambda)^2 \leq \lambda - \sigma_{k+1}$ holds if $(0.95 - 0.85)^2 < 0.85 - (0.85 - \omega)$. This implies that the inequality $\omega > 0.01$ guarantees the satisfaction of Condition (2-iii) for every acyclic graph $\text{supp}(\Sigma^{\text{res}})$. In other words, one can find the optimal solution of the GL problem using the explicit formula in Theorem 3 as long as: 1) a spanning tree structure for the optimal solution of the GL problem is sought, 2) the degree of each node in the spanning tree is not greater than 100, and (3) the difference between σ_{n-1} and σ_n is greater than 0.01. Note that Condition (2) is conservative and can be dropped for certain types of graphs (e.g., path graphs). In practice, the positive-definiteness of $I_n + \Sigma^{\text{res}}$ is not restrictive; we have verified that this matrix is positive-definite for randomly generated instances with the sizes up to $n = 200,000$ even when $\deg(v) > 100$.

Now, consider the following modifications in the experiment:

- The elements of Σ corresponding to a cycle of length n are randomly set to -0.8 or 0.8 with equal probability.
- The off-diagonal entries that do not correspond to the above cycle are in the interval $[-0.7, 0.7]$.

If λ is chosen as 0.75, then the graph $\text{supp}(\Sigma^{\text{res}})$ coincides with a cycle of length n. Furthermore, $I_n + \Sigma^{\text{res}}$ is diagonally dominant and hence positive-definite for every n. Figure 2.1 shows the optimality gap of the proposed closed-form solution and its derived theoretical upper bound (i.e. the left and right hand sides of (2.19), respectively) with respect to the length of the cycle n in log-linear scale. (note that $\deg(\mathcal{G}(\Sigma^{\text{res}}))$ and $P_{\max}(\mathcal{G}(\Sigma^{\text{res}}))$ in (2.19) are both equal to 2). Two important observations can be made based on this figure.

- In practice, the performance of the derived closed-form solution is significantly better than its theoretical upper bounds. In fact, this error is less than 10^{-6} when the length of the minimum-length cycle is at least 6. The high accuracy of the closed-form solution will become more evident in the subsequent case studies on large-scale problems.

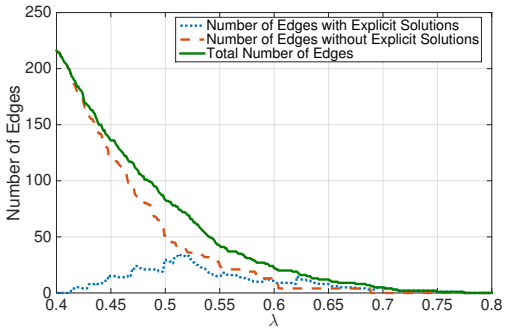
- It can be seen that the logarithm of the optimality gap is approximately a linear function of the cycle length. This matches the behavior of the theoretical bounds introduced in Theorems 4 and 5: the approximation error is exponentially decreasing with respect to the length of the minimum-length cycle.

Case Study on Brain Networks

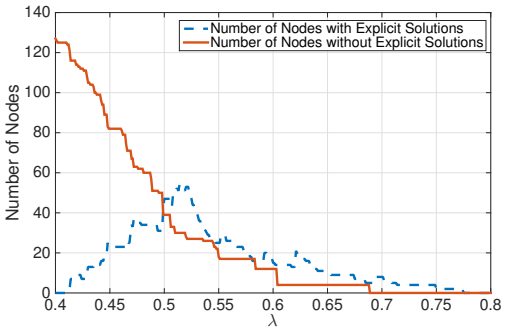
Consider the problem of estimating the brain functional connectivity network based on a set of resting state functional MRI (fMRI) data collected from 20 individual subjects [254]. The data for each subject correspond to disjoint brain activities and are correlated due to the underlying functional connectivity structure of the brain. In order to represent these dependencies, each disjoint region of the brain can be considered as a node and the correlation between two different regions can be resembled by an edge between the nodes. The data set for each subject consists of 134 samples of low frequency oscillations taken from 140 different cortical brain regions. We construct a normalized sample covariance matrix by combining the data sets of all 20 subjects (note that the data for each individual is limited and not informative enough, but the combined data provides rich information about the brain network). The goal is to use the GL to estimate the underlying functional connectivity network of different regions of the brain based on the obtained 140×140 sample covariance matrix. We study the thresholded sample covariance matrix and the derived closed-form solution for different values of the regularization coefficient in order to analyze their accuracy.

Figure 2.2a shows the number of edges in the sparsity graph of the thresholded sample covariance matrix that belong to those connected components satisfying the conditions in Theorem 3. The formula derived in this chapter is able to find the optimal values of the entries of the solution corresponding to these edges. It can be observed that if λ is greater than 0.51, then almost half of the edges in the sparsity graph of the optimal solution can be found using the proposed explicit formula. This is due to the fact that the corresponding entries in the residue matrix belong to the acyclic components of its sparsity graph and satisfy the conditions of Theorem 3. Figure 2.2b depicts the number of nodes that belong to the components (with sizes greater than 1) for which the corresponding submatrices of the solution of the GL have an explicit formula. Note that those entries in the optimal solution that correspond to isolated nodes are trivially equal to 0. Therefore, in order to better reflect the significance of the derived solution, we have only considered the components with at least two nodes. It can be observed that if λ is greater than 0.5, then the number of nodes belonging to the components with explicit formula is greater than the number of those nodes associated with inexact closed-form solutions. Figure 2.2c demonstrates the number of edges in the sparsity graph of the optimal solution, together with the number of mismatches in the edge sets of the sparsity graphs of the optimal and thresholded solutions. Notice that the number of mismatches is less than 10% when λ is greater than 0.35 and is almost 0 when λ is greater than 0.5.

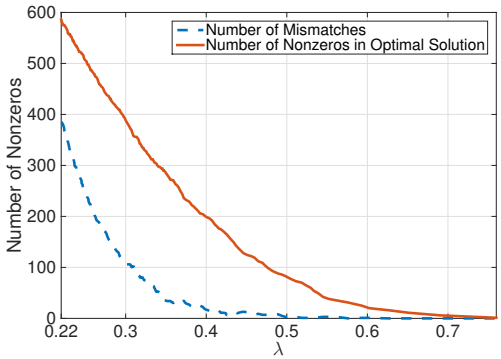
Figure 2.2d shows the minimum eigenvalues of the optimal and closed-form approximate solutions for different values of λ. The approximate solution is positive-definite when λ is



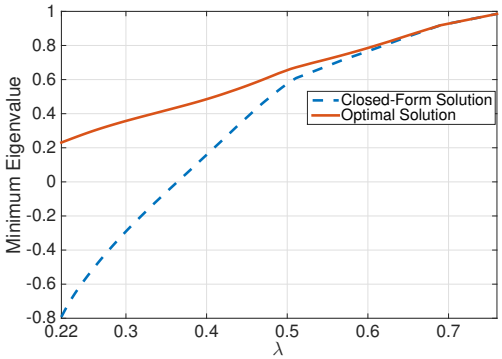
(a) Number of nodes with an exact closed-form solution



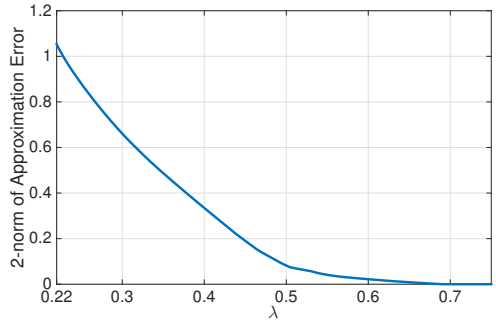
(b) Number of nodes with an exact closed-form solution



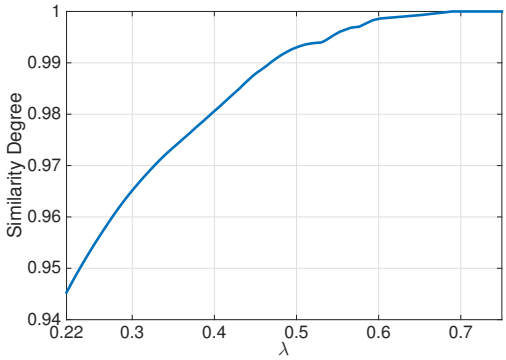
(c) Number of mismatches



(d) Minimum eigenvalue



(e) 2-norm of the approximation error



(f) Similarity degree

Figure 2.2: The performance of the proposed closed-form solution for the brain network.

greater than 0.37. This implies that λ_0 in Corollary 5 is equal to 0.37. Figures 2.2e and 2.2f depict the 2-norm of the approximation error (the difference between the optimal and closed-form approximate solutions) and the similarity degree between these two solutions, which is defined as

$$\text{similarity degree} = \frac{\text{trace}(\tilde{S}^{\text{opt}} \times \tilde{A})}{\|\tilde{S}^{\text{opt}}\|_F \times \|\tilde{A}\|_F},$$

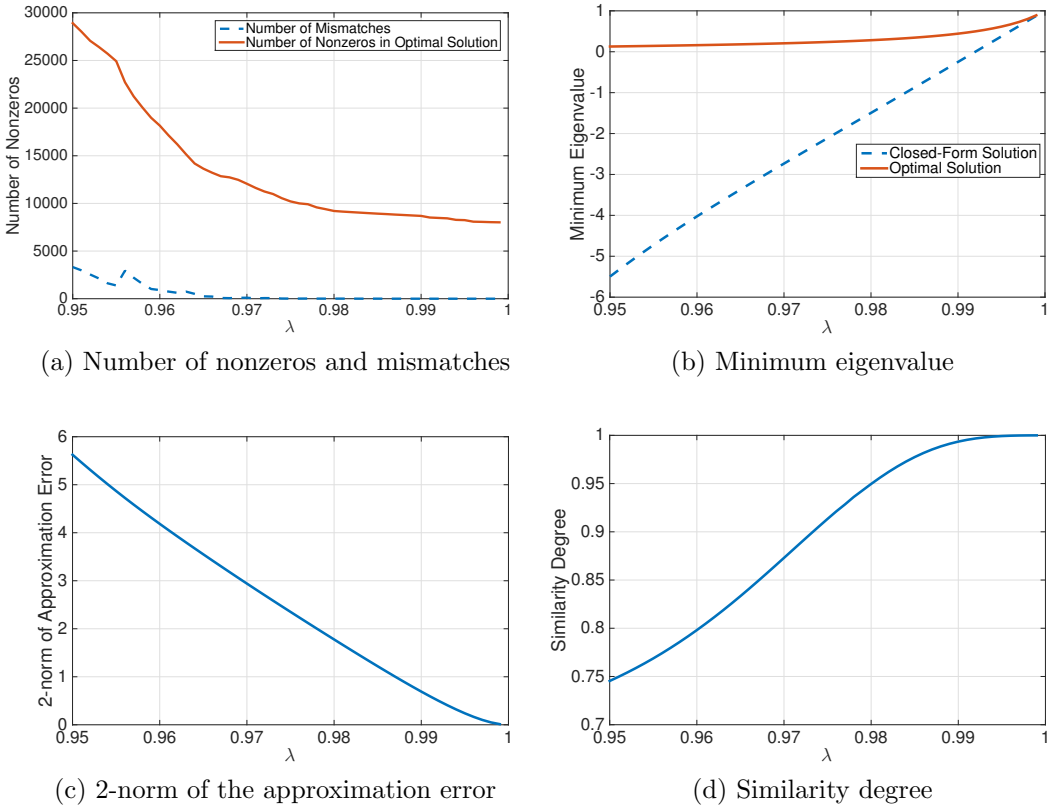
where $\tilde{S}^{\text{opt}} = S^{\text{opt}} - I_n$ and $\tilde{A} = A - I_n$. Subtracting the identity matrix from A and S^{opt} is due to the observation that both matrices have diagonal entries close to 1 when the support graph is sparse. This leads to an artificially inflated similarity degree between A and S^{opt}. Therefore, in order to have a better assessment of the similarity between the closed-form and optimal solutions, we measure the similarity between A and S^{opt} after softening the effect of their diagonal entries. The similarity degree of 1 means that the optimal and approximate solutions are exactly equal.

It can be observed that the approximation error is small and the similarity degree is high for a wide range of values of λ. For instance, if λ is greater than 0.4, then the 2-norm of the approximation error is less than 0.37 and the similarity degree is greater than 0.98. For these values of λ, the number of edges in the sparsity graph of the optimal solution ranges from 200 to 0. In all of these cases, the structure and values of the optimal solution can be estimated efficiently, without solving the optimization problem numerically.

Case Study on Transportation Networks

In recent years, the problem of short- and long-term traffic flow prediction and control has attracted much attention in Intelligent Transportation Systems (ITSs) [91]. Estimating the correlation between the traffic flows on different links of a transportation network is one of the crucial steps toward the traffic congestion control in the network; it can also serve as an initial block in different traffic forecasting methods. Substantial research has been devoted to extracting these dependencies and performing predictions based on the measured data (see [273, 193] and the references therein). In this case study, the objective is to construct a sparse matrix representing the conditional covariance between the traffic flows of different links in the network. The data is collected from the Caltrans Performance Measurement System (PeMS) database, which consists of traffic information of freeways on the a statewide scale across California [43]. We consider the data measured by the stations deployed in District 3 of California, which is collected and aggregated every 5 minutes from 1277 stations during March 6^{th} to March 12^{th} of the year 2017 (one-week interval). Due to the malfunctioning of some of the detectors, a non-negligible portion of the traffic flows was missing from the raw data set. Therefore, the following steps were taken before solving the GL problem in order to obtain a useful representation of the raw data:

- Since 228 stations did not have sufficient number of measurements during the one-week period, they were removed from the sampled data.



(a) Number of nonzeros and mismatches

(b) Minimum eigenvalue

(c) 2-norm of the approximation error

(d) Similarity degree

Figure 2.3: The performance of the proposed closed-form solution for the transportation network.

- In a few stations, the detectors did not measure the traffic flow for some periods of time. For these data samples, we used a linear interpolation method to estimate the missing values.

After performing the aforementioned data-cleaning steps, a 1049×1049 normalized sample covariance matrix was constructed from the combined 2016 data samples (288 samples for each day of the week). In Figure 2.3, the accuracy of the thresholding technique and its corresponding closed-form approximate solution is compared to the optimal solution of the GL problem for different values of the regularization coefficient.

Since the number of entries in the upper triangular part of the sample covariance matrix is large (roughly 550,000 entries), we have only considered large values of λ in order to obtain a sparse solution for the GL. Figure 2.3a shows the number of edges in the sparsity graph of the optimal solution, compared to the number of mismatches between the edge sets of the sparsity graphs of the optimal and closed-form solutions. It can be observed that as λ increases, the support graph of the optimal solution becomes sparser and the number of

mismatches decreases. In particular, the number of mismatches is almost zero if λ is chosen to be greater than 0.97. Figure 2.3b depicts the minimum eigenvalues of the optimal and closed-form approximate solutions of the GL with respect to λ. The approximate solution becomes positive-definite if λ is greater than 0.991. Furthermore, Figures 2.3c and 2.3d show that, for those values of λ between 0.991 and 0.999, the 2-norm of the approximation error is between 0.5 and 0.01, and that the similarity degree is greater than 0.99. For this range of λ, the number of edges in the sparsity graph of the optimal solution is 7.82 to 7.40 times higher the number of nodes.

Case Study on Large-Scale Data

In this case study, we evaluate the performance of the proposed closed-form solution on massive randomly generated data sets. Given n (the dimension of each sample) and similar to [126] and [275], a sparse inverse covariance matrix is generated for each test case according to the following procedure: first, a sparse matrix $U \in \mathbb{R}^{d\times d}$ is generated whose nonzero elements are randomly set to $+1$ or -1, with equal probability. Then, the inverse covariance matrix is set to $UU^\top + 2I$. Depending on the test case, the number of nonzero elements in U is controlled so that the resulted inverse covariance matrix has approximately $5d$ or $10d$ nonzero elements. $n = d/2$ number of i.i.d. samples are drawn from the corresponding multivariate Gaussian distribution in all experiments, except for the largest test case with $d = 80000$. This instance has more than 3.2 billion variables and only $n = 20000$ samples are collected to solve the GL due to the memory limitations. Furthermore, the regularization coefficient is chosen such that the estimated solution has approximately the same number of nonzero elements as the ground truth.

Table 2.1 reports the runtime of the closed-form solution, compared to two state-of-the-art methods for solving the GL, namely QUIC [126] and GLASSO [96] algorithms, as well as elementary estimator [271]. The GLASSO is the most widely-used algorithm for the GL, while the QUIC algorithm is commonly regarded as the fastest available solver for this problem. The elementary estimator is recently proposed in lieu of the GL to remove its computational burden, while preserving its desired high-dimensional properties. We use the source codes for latest versions of QUIC and GLASSO in our simulations. In particular, we use the QUIC 1.1 (available in `http://bigdata.ices.utexas.edu/software/1035/`) which is implemented in C++ with MATLAB interface. The GLASSO is downloaded from `http://statweb.stanford.edu/~tibs/glasso/` and is implemented in FORTRAN with MATLAB interface. We implemented the elementary estimator and the proposed closed-form solution in MATLAB using its sparse package. A time limit of 4 hours is considered in all experiments. Table 2.1 has the following columns:

- n: The dimension of the samples.
- m: The number of nonzero elements in the true inverse covariance matrix.
- Closed-form: The runtime of the proposed method.

n	m	Closed-Form	QUIC-C	QUIC-W	GLASSO-C	GLASSO-W	Elem.
2000	9894	0.1	2.0	1.4	42.8	13.5	0.2
2000	20022	0.1	3.0	2.1	43.8	15.3	0.2
4000	20094	0.5	13.9	7.5	460.8	135.1	2.1
4000	40382	0.5	21.5	12.0	467.6	156.2	2.9
8000	40218	2.5	78.7	49.3	3675.1	1011.2	11.3
8000	79890	2.5	111.7	88.4	3784.3	1278.8	22.2
12000	60192	7.8	243.8	153.1	$\star$	3233.0	31.8
12000	119676	7.4	333.6	251.0	$\star$	3437.2	70.2
16000	80064	17.1	570.0	322.8	$\star$	6545.0	67.2
16000	160094	18.5	787.4	616.4	$\star$	9960.8	174.8
20000	99954	39.4	1266.5	539.4	$\star$	$\star$	107.8
20000	200018	37.4	1683.8	1392.5	$\star$	$\star$	211.5
40000	200290	495.4	$\star$	$\star$	$\star$	$\star$	$\star$
80000	401798	1450.4	$\star$	$\star$	$\star$	$\star$	$\star$

Table 2.1: The runtime of different methods for solving the GL.

- QUIC-C and GLASSO-C: The runtime of the QUIC and GLASSO without initialization.
- QUIC-W and GLASSO-W: The runtime of the QUIC and GLASSO using the warm-start Algorithm 1.
- Elem.: The runtime of the elementary estimator.

In all of the test cases, the resulted closed-form solution is positive-definite and hence, feasible. It can be seen that the proposed method significantly outperforms QUIC, GLASSO and elementary estimator in terms of its runtime. In particular, the presented method is on average 6, 36, and 951 times faster than elementary, QUIC, and GLASSO methods, respectively, provided that they can obtain the solution within the predefined time limit. Furthermore, for the cases where the GL can be solved to optimality using QUIC, the relative optimality gap of the closed-form solution, i.e., $(f(A) - f^*)/f^*$, is 2.1×10^{-3} on average. For the cases with $d = 40000$ and $d = 80000$, none of these methods converge to a meaningful solution, while the proposed method can obtain an accurate solution in less than 30 minutes. On the other hand, the warm-start Algorithm 1 accompanied by QUIC and GLASSO yields up to 2.35 and 4.45 times speedups in their runtime, respectively. Moreover, the warm-start algorithm doubles the size of the instances that are solvable using the GLASSO.

Table 2.2 compares the accuracy of the estimated inverse covariance matrix using different methods. This table includes the following columns:

		Closed-Form			Graphical Lasso			Elementary		
n	m	ℓ_F	TPR	FPR	ℓ_F	TPR	FPR	ℓ_F	TPR	FPR
2000	9894	0.41	0.71	0.00	0.41	0.71	0.00	0.40	0.63	0.00
2000	20022	0.50	0.59	0.00	0.65	0.59	0.00	0.49	0.34	0.01
4000	20094	0.39	0.83	0.00	0.38	0.84	0.00	0.37	0.76	0.00
4000	40382	0.48	0.74	0.00	0.48	0.75	0.00	0.48	0.54	0.00
8000	40218	0.36	0.92	0.00	0.35	0.93	0.00	0.33	0.87	0.00
8000	79890	0.45	0.87	0.00	0.44	0.88	0.00	0.44	0.71	0.00
12000	60192	0.33	0.96	0.00	0.32	0.97	0.00	0.30	0.93	0.00
12000	119676	0.43	0.93	0.00	0.41	0.94	0.00	0.42	0.81	0.00
16000	80064	0.32	0.97	0.00	0.30	0.98	0.00	0.28	0.96	0.00
16000	160094	0.42	0.95	0.00	0.40	0.96	0.00	0.40	0.86	0.00
20000	99954	0.31	0.99	0.00	0.30	0.99	0.00	0.28	0.96	0.00
20000	200018	0.41	0.96	0.00	0.39	0.97	0.00	0.39	0.89	0.00
40000	200290	0.28	1.00	0.00	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$
80000	401798	0.27	1.00	0.00	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$

Table 2.2: The accuracy of different methods for solving the GL.

- ℓ_F: The Frobenius norm of the difference between the true and estimated inverse covariance matrices, normalized by the Frobenius norm of the true inverse covariance matrix.
- TPR and FPR: The true positive rate (TPR) and false positive rate (FPR) defined as

$$\text{TPR} = \frac{\left|(i,j) : i \neq j, S_{ij} \neq 0, (\Sigma_*^{-1})_{ij} \neq 0\right|_0}{\left|(i,j) : i \neq j, (\Sigma_*^{-1})_{ij} \neq 0\right|_0},$$
$$\text{FPR} = \frac{\left|(i,j) : i \neq j, S_{ij} \neq 0, (\Sigma_*^{-1})_{ij} = 0\right|_0}{\left|(i,j) : i \neq j, (\Sigma_*^{-1})_{ij} = 0\right|_0},$$

  where S corresponds to the explicit formula, the optimal solution of the GL, or the elementary estimator.

It can be seen that, while the elementary estimator has slightly better estimation error, its TPR is significantly outperformed by the those of the GL and closed-form solutions. Furthermore, it can be seen that the closed-form estimator has almost the same accuracy as the optimal solution of the GL. The superiority of the proposed closed-form solution over the other methods becomes more evident in the larger instances, where it (almost) exactly recovers the true sparsity pattern of the inverse covariance matrix and results in small estimation error, while becoming the only viable method for estimating the inverse covariance matrix.

Appendix

2.A Omitted Proofs of Section 2.4

Proof of Theorem 1

Before presenting the proof of Theorem 1, consider the *normalized GL*, defined as

$$\min_{S\in\mathbb{S}^n_+} -\log\det(S) + \operatorname{trace}(\tilde{\Sigma}S) + \sum_{i\neq j}\tilde{\lambda}_{ij}|S_{ij}|, \tag{2.20}$$

where $\tilde{\Sigma}$ is the normalized sample covariance, i.e., $\tilde{\Sigma}_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$ for every $(i,j) \in \{1,2,...,n\}^2$ (also known as sample correlation matrix). Similarly, $\tilde{\lambda}_{ij}$ is defined as $\frac{\lambda}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$. Upon denoting the optimal solution of the normalized GL as $\tilde{S}$, we consider the relationship between $\tilde{S}$ and S^{opt}. Recall that D is defined as a matrix collecting the diagonal elements of Σ.

Lemma 7. *We have* $S^{\text{opt}} = D^{-1/2}\tilde{S}D^{-1/2}$.

Proof. Notice that the GL (2.2) can be re-written as follows

$$\min_{S\in\mathbb{S}^n_+} -\log\det(S) + \operatorname{trace}(\tilde{\Sigma}D^{1/2}SD^{1/2}) + \sum_{i\neq j}\lambda|S_{ij}|, \tag{2.21}$$

where we have used the equality

$$\operatorname{trace}(\Sigma S) = \operatorname{trace}(D^{1/2}\tilde{\Sigma}D^{1/2}S) = \operatorname{trace}(\tilde{\Sigma}D^{1/2}SD^{1/2}).$$

Upon defining

$$\tilde{S} = D^{1/2}SD^{1/2} \tag{2.22}$$

and following some algebra, one can verify that (4.33) is equivalent to

$$\min_{\tilde{S}\in\mathbb{S}^n_+} -\log\det(\tilde{S}) + \operatorname{trace}(\tilde{\Sigma}\tilde{S}) + \sum_{i\neq j}\tilde{\lambda}_{ij}|\tilde{S}_{ij}| + \log\det(D). \tag{2.23}$$

Dropping the constant term in (2.23) gives rise to the normalized GL (2.20). Therefore, $S^{\text{opt}} = D^{-1/2}\tilde{S}D^{-1/2}$ holds in light of 2.22. This completes the proof. □

Proof of Theorem 1. Note that, due to the Definition 9 and Lemma 7, $\tilde{\Sigma}^{\text{res}}$ and $\tilde{S}$ have the same sparsity pattern as Σ^{res} and S^{opt}, respectively. Therefore, it suffices to show that the sparsity structures of $\tilde{\Sigma}^{\text{res}}$ and $\tilde{S}$ are the same.

To verify this, we focus on the optimality conditions for optimization (2.20). Define M as $I_n + \tilde{\Sigma}^{\text{res}}$. Due to Condition (1-i) and Lemma 1, M is inverse-consistent and has a unique inverse-consistent complement, which is denoted by N. First, will show that $(M+N)^{-1}$ is the optimal solution of (2.20). For an arbitrary pair $(i,j) \in \{1,...,d\}^2$, the KKT conditions, introduced in Lemma 6, imply that one of the following cases holds:

1) $i = j$: We have $(M+N)_{ij} = M_{ii} = \tilde{\Sigma}_{ii}$.

2) $(i,j) \in \mathcal{G}(\tilde{\Sigma}^{\text{res}})$: In this case, we have

$$(M+N)_{ij} = M_{ij} = \tilde{\Sigma}_{ij} - \tilde{\lambda}_{ij} \times \text{sign}(\tilde{\Sigma}_{ij}).$$

Note that since $|\Sigma_{ij}| > \lambda$, we have that $\text{sign}(M_{ij}) = \text{sign}(\tilde{\Sigma}_{ij})$. On the other hand, due to the sign-consistency of M, we have $\text{sign}(M_{ij}) = -\text{sign}\Big(((M+N)^{-1})_{ij}\Big)$. This implies that

$$(M+N)_{ij} = M_{ij} = \tilde{\Sigma}_{ij} + \tilde{\lambda}_{ij} \times \text{sign}((M+N)^{-1}).$$

3) $(i,j) \notin \mathcal{G}(\tilde{\Sigma}^{\text{res}})$: One can verify that $(M+N)_{ij} = N_{ij}$. Therefore, due to Condition (1-iii), we have

$$\begin{aligned} |(M+N)_{ij}| &\leq \beta\left(\mathcal{G}(\tilde{\Sigma}^{\text{res}}), \|\tilde{\Sigma}^{\text{res}}\|_{\max}\right) \\ &\leq \min_{\substack{k\neq l \\ (k,l)\notin\mathcal{G}(\Sigma^{\text{res}})}} \frac{\lambda - |\Sigma_{kl}|}{\sqrt{\Sigma_{kk}\Sigma_{ll}}} \\ &= \min_{\substack{k\neq l \\ (k,l)\notin\mathcal{G}(\Sigma^{\text{res}})}} \tilde{\lambda}_{kl} - |\tilde{\Sigma}_{kl}|. \end{aligned} \tag{2.24}$$

This leads to

$$|(M+N)_{ij} - \tilde{\Sigma}_{ij}| \leq |(M+N)_{ij}| + |\tilde{\Sigma}_{ij}| \leq \min_{\substack{k\neq l \\ (k,l)\notin\mathcal{G}(\Sigma^{\text{res}})}} \Big(\tilde{\lambda}_{kl} - |\tilde{\Sigma}_{kl}|\Big) + |\tilde{\Sigma}_{ij}| \leq \tilde{\lambda}_{ij}. \tag{2.25}$$

Therefore, it can be concluded that $(M+N)^{-1}$ satisfies the KKT conditions for (2.20)[1]. On the other hand, note that $\mathcal{G}((M+N)^{-1}) = \mathcal{G}(\tilde{\Sigma}^{\text{res}})$. This concludes the proof. □

Proof of Lemma 2

To proceed with the proof of Lemma 2, we need the following lemma.

[1]The KKT conditions for the normalized GL are equivalent to (2.14) after replacing λ with $\tilde{\lambda}_{ij}$

Lemma 8. *Consider a matrix $M \in \mathbb{S}^n$ with positive-definite completion. Assume that $\left|\left|\left|M^{(c)}\right|\right|\right|_1 \leq \eta |||M - I_n|||_1$ and $|||M - I_n|||_1 < \frac{1}{\eta+1}$, for some number η. The relation*

$$\left|\left|\left|M^{(c)}\right|\right|\right|_1 \leq (1+\eta)^2 \frac{|||M - I_n|||_1^2}{1 - (\eta+1)|||M - I_n|||_1}$$

holds.

Proof. Note that $M \in \mathbb{S}^n$ has a positive-definite completion and hence, is inverse-consistent due to Lemma 1. One can write

$$\left|\left|\left|(M - I_n) + M^{(c)}\right|\right|\right|_1 \leq |||M - I_n|||_1 + \left|\left|\left|M^{(c)}\right|\right|\right|_1 \leq (\eta+1)|||M - I_n|||_1 < 1.$$

Therefore,

$$\begin{aligned}(M + M^{(c)})^{-1} = (I_n + (M - I_n + M^{(c)}))^{-1} + I_n - (M - I_n + M^{(c)}) \\ + (M - I_n + M^{(c)})^2 \times \sum_{i=0}^{\infty} (-M + I_n - M^{(c)})^i.\end{aligned}$$

Since $\mathcal{G}((M + M^{(c)})^{-1}) \subseteq \mathcal{G}(M)$, it can be concluded that the (i,j) entries of $M^{(c)}$ and

$$(M - I_n + M^{(c)})^2 \times \sum_{i=0}^{\infty} (-M + I_n - M^{(c)})^i$$

are equal for every $(i,j) \in \mathcal{G}(M^{(c)})$. Since the (i,j) entry of $M^{(c)}$ is zero if $(i,j) \notin \mathcal{G}(M^{(c)})$, we have

$$\left|\left|\left|M^{(c)}\right|\right|\right|_1 \leq \left|\left|\left|(M - I_n + M^{(c)})^2 \sum_{i=0}^{\infty} (M - I_n + M^{(c)})^i\right|\right|\right|_1.$$

Since 1-norm is sub-multiplicative, the above inequality can be simplified as

$$\begin{aligned}\left|\left|\left|M^{(c)}\right|\right|\right|_1 &\leq (|||M - I_n|||_1 + \left|\left|\left|M^{(c)}\right|\right|\right|_1)^2 \times \sum_{i=0}^{\infty} (|||M - I_n|||_1 + \left|\left|\left|M^{(c)}\right|\right|\right|_1)^i \\ &= \frac{(|||M - I_n|||_1 + \left|\left|\left|M^{(c)}\right|\right|\right|_1)^2}{1 - |||M - I_n|||_1 - |||M^{(c)}|||_1} \\ &\leq \frac{(|||M - I_n|||_1 + \eta|||M - I_n|||_1)^2}{1 - |||M - I_n|||_1 - \eta|||M - I_n|||_1} \\ &= (1+\eta)^2 \frac{|||M - I_n|||_1^2}{1 - (\eta+1)|||M - I_n|||_1}.\end{aligned}$$

This completes the proof. □

Proof of Lemma 2. Given an arbitrary graph $\mathcal{G}$, consider a matrix variable M with 1's on the diagonal such that $\mathcal{G}(M) \subseteq \mathcal{G}$. The first objective is to find a matrix in terms of M, denoted by the matrix function $N(M)$, satisfying the following properties

$$\begin{aligned}&\mathcal{G}\left((M+N(M))^{-1}\right) \subseteq \mathcal{G},\\ &\mathcal{G}(N(M)) \subseteq \mathcal{G}^{(c)}.\end{aligned}$$

To this end, define the matrix function $A(M)$ as

$$A(M) = (M + N(M))^{-1}.$$

Observe that

- As long as $A(M)$ exists and $\mathcal{G}(A(M)) \subseteq \mathcal{G}$, there is a continuously differentiable mapping from $A(M)$ to M because M can be found by setting those entries of $A(M)^{-1}$ corresponding to the edges of $\mathcal{G}^{(c)}$ to zero. Moreover, the Jacobian of this mapping has full rank at $M = I_n$. Due to the inverse function theorem, the mapping from M to $A(M)$ exists and is continuously differentiable.
- Similarly, as long as $A(M)$ exists and $\mathcal{G}(A(M)) \subseteq \mathcal{G}$, there is a continuously differentiable mapping from $A(M)$ to $N(M)$.
- If $M = I_n$, then $N(M) = 0$.

It follows from the above properties that if M is sufficiently small, the function $N(M)$ exists and satisfies the following properties: (i) $0 = N(I_n)$, and (ii) $N(\cdot)$ is differentiable at $M = I_n$. This implies that there are sufficiently small nonzero numbers η and α_0 such that $|||N(M)|||_1 \leq \eta |||M - I_n|||_1$ whenever $\|M\|_{\max} \leq \alpha_0$. Now, it follows from Lemma 8 that

$$|||N(M)|||_1 \leq (1+\eta)^2 \frac{|||M - I_n|||_1^2}{1 - (\eta+1)|||M - I_n|||_1},$$

or

$$\|N(M)\|_{\max} \leq \frac{(1+\eta)^2 \times (\deg(\mathcal{G}))^2}{1 - (\eta+1)\alpha_0 \times \deg(\mathcal{G})} \|M\|_{\max}^2,$$

if $\|M\|_{\max} \leq \alpha_0$. The inequality (2.7) is satisfied for the number ζ defined as the maximum of

$$\frac{(1+\eta)^2 \times (\deg(\mathcal{G}))^2}{1 - (\eta+1)\alpha_0 \times \deg(\mathcal{G})}$$

and the finite number

$$\max\left\{ \frac{\beta(\mathcal{G},\alpha)}{\alpha^2} \middle| \alpha \in (\alpha_0, 1) \right\}.$$

This completes the proof. □

Proof of Lemma 3

It can be easily verified that

$$(M + M^{(c)})^{-1} = I - (M + M^{(c)} - I) + (M + M^{(c)})^{-1}(M + M^{(c)} - I)^2.$$

This implies that, for a given pair $(i, j) \in \mathcal{G}$, one can write

$$\left((M + M^{(c)})^{-1}\right)_{ij} = -M_{ij} + \left((M + M^{(c)})^{-1}\right)_{i:} \left((M + M^{(c)} - I)^2\right)_{:j}, \tag{2.27}$$

where $\left((M + M^{(c)})^{-1}\right)_{i:}$ and $\left((M + M^{(c)} - I)^2\right)_{:j}$ are the i^{th} row and j^{th} column of $(M + M^{(c)})^{-1}$ and $(M + M^{(c)} - I)^2$, respectively. Based on (2.27), the (i, j) entries of M and $(M + M^{(c)})^{-1}$ have opposite signs if

$$|M_{ij}| > \left|\left((M + M^{(c)})^{-1}\right)_{i:} \left((M + M^{(c)} - I)^2\right)_{:j}\right|. \tag{2.28}$$

To streamline the presentation, $\|M\|_{\max}$ is redefined as $\max_{i,j} |M_{ij}|$ in the rest of the proof. One can write

$$\begin{aligned}
\left\|(M+M^{(c)}-I)^2\right\|_{\max} &\leq \left\|(M-I)^2\right\|_{\max} + \left\|\left(M^{(c)}\right)^2\right\|_{\max} + \left\|M^{(c)}(M-I)\right\|_{\max} + \left\|(M-I)M^{(c)}\right\|_{\max} \\
&\leq \deg(\mathcal{G})\alpha^2 + (d - \deg(\mathcal{G}))\zeta(\mathcal{G})^2\alpha^4 + 2\deg(\mathcal{G})\zeta(\mathcal{G})\alpha^3 \\
&\leq 3\deg(\mathcal{G}) \max\{\alpha^2, \zeta(\mathcal{G})\alpha^3\} + (d - \deg(\mathcal{G}))\zeta(\mathcal{G})^2\alpha^4 \\
&\leq K\alpha^2,
\end{aligned} \tag{2.29}$$

for some K that only depends on $\deg(\mathcal{G}), \zeta(\mathcal{G})$, and d. Furthermore, assume that

$$\alpha \leq \frac{1}{2\deg(\mathcal{G})\sqrt{\zeta(\mathcal{G})}} = \alpha_0(\mathcal{G}). \tag{2.30}$$

Note that

$$(M + M^{(c)})^{-1} = I - (M + M^{(c)} - I)(M + M^{(c)})^{-1},$$

which implies that

$$\left\|(M + M^{(c)})^{-1}\right\|_{\max} = 1 + \deg(\mathcal{G}) \max\{\alpha, \zeta(\mathcal{G})\alpha^2\} \left\|(M + M^{(c)})^{-1}\right\|_{\max}, \tag{2.31}$$

where we have used the fact that $\mathcal{G}((M + M^{(c)})^{-1}) \subseteq \mathcal{G}$ and hence, its maximum degree is upper bounded by $\deg(\mathcal{G})$. (2.31), together with the assumption (2.30) implies that

$$\left\|(M + M^{(c)})^{-1}\right\|_{\max} \leq \frac{1}{1 - \deg(\mathcal{G}) \max\{\alpha, \zeta(\mathcal{G})\alpha^2\}} \leq 2. \tag{2.32}$$

Combining (2.29) and (2.32) with (2.28) completes the proof. □

2.B Omitted Proofs of Section 2.5

Proof of Lemma 4

Without loss of generality, assume that $\mathcal{G}$ is a tree. Note that if there are disjoint components, the argument made in the sequel can be applied to each connected component of $\mathcal{G}$ separately. Let d_{ij} denote the unique path between every two disparate nodes i and j in $\mathcal{G}$. Furthermore, define $\mathcal{N}(i)$ as the set of all neighbors of node i in $\mathcal{G}$. Consider a matrix M with positive-definite completion and with diagonal elements equal to 1 such that $\|M\|_{\max} \leq \alpha$ and $\text{supp}(M) = \mathcal{G}$. Let N be a matrix with the following entries

$$N_{ij} = \begin{cases} \prod_{(m,t)\in d_{ij}} M_{mt} & \text{if} \quad (i,j) \in (\mathcal{G}(M))^{(c)}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.33}$$

Moreover, define

$$A_{ij} = \begin{cases} 1 + \sum_{m\in\mathcal{N}(i)} \frac{M_{mi}^2}{1-M_{mi}^2} & \text{if} \quad i = j, \\ \frac{-M_{ij}}{1-M_{ij}^2} & \text{if} \quad (i,j) \in \mathcal{G}(M), \\ 0 & \text{otherwise.} \end{cases} \tag{2.34}$$

The goal is to show that the matrix N is the unique inverse-consistent complement of M. First, note that $\text{supp}(N) = (\text{supp}(M))^{(c)}$ and $\text{supp}(M) = \text{supp}(A)$. Next, it is desirable to prove that $(M+N)^{-1} = A$ or equivalently $(M+N)A = I$. Upon defining $T = (M+N)A$, one can write

$$T_{ii} = \sum_{m=1}^{n} (M_{im} + \mathcal{N}_{im}) A_{mi} = 1 + \sum_{m\in\mathcal{N}(i)} \frac{M_{mi}^2}{1-M_{mi}^2} - \sum_{m\in\mathcal{N}(i)} \frac{M_{mi}^2}{1-M_{mi}^2} = 1.$$

Moreover, for every pair of nodes i and j, define D_{ij} as $\prod_{(k,t)\in d_{ij}} M_{kt}$ if $i \neq j$ and as 1 if $i = j$.

Consider a pair of distinct nodes i and j. Let t denote the node adjacent to j in d_{ij} (note that we may have $t = i$). It can be verified that

$$\begin{aligned} T_{ij} = \sum_{m=1}^{n} (M_{im} + N_{im}) A_{mj} = D_{ij} \left(1 + \sum_{m\in\mathcal{N}(j)} \frac{M_{mj}^2}{1-M_{mj}^2} \right) - D_{it} \left(\frac{M_{tj}}{1-M_{tj}^2} \right) \\ - \sum_{\substack{m\in\mathcal{N}(j) \\ m\neq t}} D_{im} \frac{M_{mj}}{1-M_{mj}^2}. \end{aligned} \tag{2.35}$$

Furthermore,

$$\begin{aligned} &D_{ij} = D_{it} M_{tj}, \\ &D_{im} = D_{it} M_{tj} M_{jm}, \qquad \forall\ m \in \mathcal{N}(j),\ m \neq t. \end{aligned} \tag{2.36}$$

Plugging (2.36) into (2.35) yields that

$$T_{ij} = D_{it}M_{tj}\left(\frac{1}{1-M_{tj}^2} + \sum_{\substack{m\in\mathcal{N}(j)\\ m\neq t}} \frac{M_{mj}^2}{1-M_{mj}^2}\right) - D_{it}\left(\frac{M_{tj}}{1-M_{tj}^2}\right) - D_{it}M_{tj}\sum_{\substack{m\in\mathcal{N}(j)\\ m\neq t}} \frac{M_{mj}^2}{1-M_{mj}^2} = 0.$$

Hence, $T = I$. Finally, we need to show that $M + N \succ 0$. To this end, it suffices to prove that $A \succ 0$. Note that A can be written as $I + \sum_{(i,j)\in\mathcal{G}} L^{(i,j)}$, where $L^{(i,j)}$ is defined as

$$L_{rl}^{(i,j)} = \begin{cases} \frac{M_{ij}^2}{1-M_{ij}^2} & \text{if} \quad r = l = i \text{ or } j, \\ \frac{-M_{ij}}{1-M_{ij}^2} & \text{if} \quad (r,l) = (i,j), \\ 0 & \text{otherwise.} \end{cases}$$

Consider the term x^TAx for an arbitrary vector $x \in \mathbb{R}^n$. One can verify that

$$\begin{aligned} x^TAx =& \sum_{i=1}^n x_i^2 + \sum_{(i,j)\in\mathcal{G}} x^TL^{(i,j)}x \\ =& \sum_{i=1}^n x_i^2 + \sum_{(i,j)\in\mathcal{G}} \left(\frac{M_{ij}^2}{1-M_{ij}^2}\right)x_i^2 + \left(\frac{M_{ij}^2}{1-M_{ij}^2}\right)x_j^2 - \left(\frac{2M_{ij}}{1-M_{ij}^2}\right)x_ix_j. \end{aligned} \tag{2.37}$$

Without loss of generality, assume that the graph is a rooted tree with the root at node n. Assume that each edge (i,j) defines a direction that is toward the root. Then, it follows from (2.37) that

$$\begin{aligned} x^TAx =& x_n^2 + \sum_{(i,j)\in\mathcal{G}} x_i^2 + \left(\frac{M_{ij}^2}{1-M_{ij}^2}\right)x_i^2 + \left(\frac{M_{ij}^2}{1-M_{ij}^2}\right)x_j^2 - \left(\frac{2M_{ij}}{1-M_{ij}^2}\right)x_ix_j \\ =& x_n^2 + \sum_{(i,j)\in\mathcal{G}} \left(\frac{1}{1-M_{ij}^2}\right)x_i^2 + \left(\frac{M_{ij}^2}{1-M_{ij}^2}\right)x_j^2 - \left(\frac{2M_{ij}}{1-M_{ij}^2}\right)x_ix_j \\ =& x_n^2 + \sum_{(i,j)\in\mathcal{G}} \frac{(x_i - M_{ij}x_j)^2}{1-M_{ij}^2} \geq 0. \end{aligned}$$

Therefore, $M + N \succeq 0$ and subsequently $M + N \succ 0$ (because it is invertible). Hence, according to Definition 5 and Lemma 1, the matrix N is the unique inverse-consistent compliment of M. On the other hand, it follows from the definition of N that $\|N\|_{\max} \leq \alpha^2$ and consequently $\beta(\mathcal{G},\alpha) \leq \alpha^2$. Now, suppose that $\mathcal{G}$ includes a path of length at least 2, e.g., the edges $(1,2)$ and $(2,3)$ belong to $\mathcal{G}$. By setting $M_{12} = M_{23} = \alpha$ and choosing sufficiently small values for those entries of M corresponding to the remaining edges in $\mathcal{G}$, the matrix M becomes positive-definite with a trivial positive-definite completion and we obtain $\|N\|_{\max} = \alpha^2$. This completes the proof. □

Proof of Theorem 2

To prove this theorem, first consider the following matrix

$$\hat{S}_{ij} = \begin{cases} 1 + \sum_{(i,m)\in\mathcal{E}^{\text{opt}}} \frac{(\tilde{\Sigma}^{\text{res}}_{im})^2}{1-(\tilde{\Sigma}^{\text{res}}_{im})^2} & \text{if} \quad i = j, \\ \frac{-\tilde{\Sigma}^{\text{res}}_{ij}}{1-(\tilde{\Sigma}^{\text{res}}_{ij})^2} & \text{if} \quad (i,j) \in \mathcal{E}^{\text{opt}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.38}$$

In what follows, we will show that $\hat{S} = \tilde{S}$, where $\tilde{S}$ is the optimal solution for the normalized GL. This, together with Lemma 7 implies that (2.11) is indeed optimal for the GL.

First, note that there exists a matrix N such that $\tilde{S}^{-1} = M + N$, where M is defined as

$$M_{ij} = \begin{cases} \tilde{\Sigma}_{ij} + \tilde{\lambda}_{ij} \times \text{sign}(\tilde{S}_{ij}) & \text{if} \quad (i,j) \in \text{supp}(\tilde{S}), \\ 1 & \text{if} \quad i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.39}$$

Clearly, $\text{supp}(\tilde{S}) = \text{supp}(M)$. Furthermore, $M = I_n + \tilde{T}(\lambda)$, where $(i,j)^{\text{th}}$ entry of $\tilde{T}(\lambda)$ is equal to $\tilde{\Sigma}_{ij} + \tilde{\lambda}_{ij}\text{sign}(S^{\text{opt}}_{ij})$ for every $(i,j) \in \text{supp}(S^{\text{opt}})$ and it is equal to zero otherwise. Subsequently, $M = D^{-1/2}(D + T(\lambda))D^{-1/2}$ and hence, $D + T(\lambda) \succ 0$ implies $M \succ 0$. By combining $N = (\tilde{S})^{-1} - M$ with (2.39) and exploiting the optimality conditions in (2.14), one can verify that $\text{supp}(N) \subseteq (\text{supp}(M))^{(c)}$ and $\text{supp}(\tilde{S}) = \text{supp}\left((M+N)^{-1}\right) \subseteq \text{supp}(M)$. Therefore, according to Lemma 1, the matrix N is the unique inverse-complement of M. Moreover, since M is sign-consistent, the equation $\text{sign}(M_{ij}) = -\text{sign}(\tilde{S}_{ij})$ holds for every $(i,j) \in \text{supp}(\tilde{S})$. This leads to the relations $\text{sign}(\Sigma_{ij}) = -\text{sign}(\tilde{S}_{ij})$ and

$$M_{ij} = \tilde{\Sigma}^{\text{res}}_{ij}, \tag{2.40a}$$

$$|\tilde{\Sigma}_{ij}| > \tilde{\lambda}_{ij}, \tag{2.40b}$$

for every $(i,j) \in \text{supp}(\tilde{S})$. Part 1 of the theorem is an immediate consequence of (2.40b). On the other hand, based on the argument made in the proof of Lemma 4, the matrix N can be obtained as

$$N_{ij} = \begin{cases} \prod_{(m,t)\in d_{ij}} M_{mt} & \text{if} \quad d_{ij} \neq \emptyset \text{ and } (i,j) \in (\text{supp}(M))^{(c)}, \\ 0 & \text{otherwise,} \end{cases} \tag{2.41}$$

where d_{ij} denotes the unique path between nodes i and j in $\text{supp}(\tilde{S})$ if they belong to the same connected component in $\text{supp}(\tilde{S})$, and d_{ij} is empty if there is no path between nodes i and j. Similar to the proof of Lemma 4, one can show that (2.11) is equal to $(M+N)^{-1}$. This completes the proof of the second part of the theorem. □

Proof of Theorem 3

Based on Lemmas 4 and 5, the conditions introduced in Theorem 1 can be reduced to conditions (2-ii) and (2-iii) in Theorem 3 if $\text{supp}(\Sigma^{\text{res}})$ is acyclic and therefore, $\mathcal{E}^{\text{opt}} = \mathcal{E}^{\text{res}}$. Moreover, suppose that M is set to $I_n + \tilde{\Sigma}^{\text{res}}$, and that the matrices N and A are defined as (2.33) and (2.34), respectively. Similar to the proof of Theorem 1, it can be verified that (2.38) satisfies all the KKT conditions for the normalized GL (2.20). Therefore, due to Lemma 7, (2.11) is the unique solution of the GL. The details are omitted for brevity. □

Proof of Corollary 1

Given Σ and λ, the matrix Σ^{res} can be computed in $\mathcal{O}(n^2)$. Moreover, Condition (2-i) in Theorem 3 can be checked using the Depth-First-Search algorithm, which has the time complexity of $\mathcal{O}(n^2)$ in the worst case [6]. If the graph is cyclic, Theorem 3 cannot be used. Otherwise, we consider Condition (2-ii). For matrices with acyclic support graphs, the Cholesky Decomposition can be computed in $\mathcal{O}(n)$, from which the positive-definiteness of the matrix can be checked [252]. The complexity of checking Condition (2-iii) is equivalent to that of finding its left and right hand sides, which can be done in $\mathcal{O}(n)$ and $\mathcal{O}(n^2)$, respectively. Finally, since (2.11) can be used only if the support graph of Σ^{res} is acyclic, one can easily verify that the complexity of obtaining S^{opt} using (2.11) is at most $\mathcal{O}(n)$. This completes the proof of Corollary 1. □

2.C Omitted Proofs of Section 2.6

This section is devoted to proving approximation bounds for the derived closed-form solution when the acyclic assumption on the support graph of the thresholded sample covariance matrix is not necessarily acyclic. The shorthand notations c, deg, $\mathcal{P}_{ij}$ and $P_{\max}$ will be used instead of $c(\mathcal{G}(\Sigma^{\text{res}}))$, $\deg(\mathcal{G}(\Sigma^{\text{res}}))$, $\mathcal{P}_{ij}(\mathcal{G}(\Sigma^{\text{res}}))$ and $P_{\max}(\mathcal{G}(\Sigma^{\text{res}}))$, respectively. First, the approximation error of the closed-form solution for the normalized GL will be analyzed. Then, the result will be generalized to the GL via the key equality in Lemma 7.

Proof of Theorem 4

To prove Theorem 4, the first step is to generalize the definition of the matrix N in (2.41) and show that this generalized matrix is an approximate inverse-consistent complement of $I_n + \tilde{\Sigma}^{\text{res}}$. Without loss of generality, assume that $\text{supp}(\Sigma^{\text{res}})$ is connected. If there are disjoint components in $\text{supp}(\Sigma^{\text{res}})$, the argument made in the sequel can be used for every connected component due to the decomposition rule for the GL (see [179]). Let M be equal

to $I_n + \tilde{\Sigma}^{\text{res}}$. Consider the matrix N as

$$N_{ij} = \begin{cases} \sum_{d_{ij} \in \mathcal{P}_{ij}} \prod_{(m,t) \in d_{ij}} M_{mt} & \text{if} \quad (i,j) \in (\text{supp}(M))^{(c)}, \\ \sum_{d_{ij} \in \mathcal{P}_{ij} \backslash \{(i,j)\}} \prod_{(m,t) \in d_{ij}} M_{mt} & \text{if} \quad (i,j) \in (\text{supp}(M)), \\ 0 & \text{otherwise.} \end{cases} \tag{2.42}$$

It can be verified that $M + N = R$, where

$$R_{ij} = \begin{cases} \sum_{d_{ij} \in \mathcal{P}_{ij}} \prod_{(m,t) \in d_{ij}} M_{mt} & \text{if} \quad i \neq j, \\ 1 & \text{if} \quad i = j. \end{cases} \tag{2.43}$$

For each simple path between the pair of nodes i and j, define its length as the multiplication of the entries of M corresponding to the edges of the path. Based on this definition, R_{ij} is equal to the sum of the lengths of all nonidentical simple paths between nodes i and j in $\text{supp}(M)$. Denote d_{ij}^s as any shortest path between nodes i and j in $\text{supp}(M)$ (recall that $\text{supp}(M)$ is unweighted), and let R^s be given by

$$R_{ij}^s = \begin{cases} \prod_{(m,t) \in d_{ij}^s} M_{mt} & \text{if} \quad i \neq j, \\ 1 & \text{if} \quad i = j. \end{cases}$$

Note that R^s collects the length of the shortest path between every two nodes in $\text{supp}(M)$. The following lemmas are crucial to prove Theorem 4.

Lemma 9. *Given two nodes i and j in $\mathcal{G}(\Sigma^{\text{res}})$, suppose that $\mathcal{P}_{ij} \backslash d_{ij}^s$ is non-empty. Then, the length of every path d_{ij} in $\mathcal{P}_{ij} \backslash d_{ij}^s$ is at least $\lceil c/2 \rceil$.*

Proof. Consider a path d_{ij} in $\mathcal{P}_{ij} \backslash d_{ij}^s$. The subgraph $d_{ij} \cup d_{ij}^s$ has a cycle. Since the length of this cycle is at least c, the segment of this cycle that resides in d_{ij} should have the length of at least $\lceil c/2 \rceil$; otherwise d_{ij}^s is not the shortest path between the nodes i and j. This implies that the length of d_{ij} is at least $\lceil c/2 \rceil$. □

Lemma 10. *Let M be equal to $I_n + \tilde{\Sigma}^{\text{res}}$. The inequalities*

$$\left| R_{ij} - R_{k'j}^s M_{ik'} \right| \leq (|\mathcal{P}_{ij}|_0 - 1) \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max} \right)^{\lceil \frac{c}{2} \rceil}, \tag{2.44a}$$

$$\left| R_{kj} - R_{k'j}^s M_{ik'} M_{ik} \right| \leq (|\mathcal{P}_{kj}|_0 - 1) \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max} \right)^{\lceil \frac{c}{2} \rceil - 1} \tag{2.44b}$$

hold if $i \neq j$, where k' is the node adjacent to i in d_{ij}^s and $k \in \mathcal{N}(i) \backslash k'$.

Proof. First, we show the validity of (2.44a). Due to (2.43), one can write

$$R_{ij} = R_{ij}^s + \sum_{d_{ij} \in \mathcal{P}_{ij} \backslash d_{ij}^s} \prod_{(m,t) \in d_{ij}} M_{mt}. \tag{2.45}$$

If $\mathcal{P}_{ij}\backslash d_{ij}^s$ is empty, then the equation $R_{ij} = R_{k'j}^s M_{ik'}$ and therefore (2.44a) hold. Now, assume that $\mathcal{P}_{ij}\backslash d_{ij}^s$ is not empty. Due to Lemma 9, we have

$$-\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\lceil \frac{c}{2} \rceil} \leq \prod_{(m,t)\in d_{ij}} M_{mt} \leq \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\lceil \frac{c}{2} \rceil},$$

for every $d_{ij} \in \mathcal{P}_{ij}\backslash d_{ij}^s$. The above inequalities, together with (2.45) and the equation $R_{ij}^s = R_{k'j}^s M_{ik'}$, result in (2.44a). To prove (2.44b), define $\hat{d}_{kj}$ as $d_{ij}^s \cup \{(i,k)\}$ (note that $\hat{d}_{kj}$ is not necessarily equal to d_{kj}^s). It yields that

$$R_{kj} = R_{ij}^s M_{ik} + \sum_{d_{kj}\in\mathcal{P}_{kj}\backslash\hat{d}_{kj}} \prod_{(m,t)\in d_{kj}} M_{mt}. \tag{2.46}$$

In light of Lemma 9, the length of every path $d_{kj} \in \mathcal{P}_{kj}\backslash\hat{d}_{kj}$ is lower bounded by $\lceil c/2 \rceil - 1$. This implies that

$$-\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\lceil \frac{c}{2} \rceil - 1} \leq \prod_{(m,t)\in d_{ij}} M_{mt} \leq \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\lceil \frac{c}{2} \rceil - 1}, \tag{2.47}$$

for every $d_{kj} \in \mathcal{P}_{kj}\backslash\hat{d}_{kj}$. Combining $R_{ij}^s M_{ik} = R_{k'j}^s M_{ik'} M_{ik}$ with (3.39) and (2.47) leads to the inequality (2.44b). □

Lemma 11. *The following inequality holds*

$$\frac{\deg}{1 - \|\tilde{\Sigma}^{\text{res}}\|_{\max}^2} \leq \delta,$$

where δ defined as (2.17).

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

Proof of Theorem 4 Consider the normalized GL and define the following explicit formula for $\tilde{A}$

$$\tilde{A}_{ij} = \begin{cases} 1 + \sum\limits_{(i,m)\in\mathcal{E}^{\text{opt}}} \frac{(\tilde{\Sigma}_{im}^{\text{res}})^2}{1-(\tilde{\Sigma}_{im}^{\text{res}})^2} & \text{if} \quad i = j, \\ \frac{-\tilde{\Sigma}_{ij}^{\text{res}}}{1-(\tilde{\Sigma}_{ij}^{\text{res}})^2} & \text{if} \quad (i,j) \in \mathcal{E}^{\text{res}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.48}$$

Let M be equal to $I_n + \tilde{\Sigma}^{\text{res}}$. Furthermore, define

$$\tilde{\epsilon} = \delta \cdot (P_{\max}(\mathcal{G}(\Sigma^{\text{res}})) - 1) \cdot \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\left\lceil \frac{c(\mathcal{G}(\Sigma^{\text{res}}))}{2} \right\rceil}.$$

In order to prove the theorem, we use the matrix N defined in (2.42), and first show that $M+N$ is an $\tilde{\epsilon}$-relaxed inverse of $\tilde{A}$ and that the pair $(\tilde{A}, M+N)$ satisfies the $\tilde{\epsilon}$-relaxed KKT conditions.

Consider the matrix T defined as $T = \tilde{A}(M+N)$ and recall that $M+N = R$. One can write

$$T_{ii} = \sum_{m=1}^{n} \tilde{A}_{im} R_{mi} = \left(1 + \sum_{m \in \mathcal{N}(i)} \frac{{M_{im}}^2}{1-{M_{im}}^2}\right) - \sum_{m \in \mathcal{N}(i)} \frac{M_{im}}{1-{M_{im}}^2} R_{mi}. \tag{2.49}$$

Note that since $\{(m,i)\} \in \mathcal{P}_{mi}$ for every $m \in \mathcal{N}(i)$, we have

$$R_{mi} = M_{mi} + \sum_{d_{mi} \in \mathcal{P}_{mi} \backslash \{(m,i)\}} \prod_{(r,t) \in d_{mi}} M_{rt}.$$

If $\mathcal{P}_{mi}\backslash\{(m,i)\}$ is empty, then $R_{mi} = M_{mi}$ and $T_{ii} = 1$. Otherwise, since the length of the minimum-length cycle is c, the length of every path $d_{mi} \in \mathcal{P}_{mi}\backslash\{(m,i)\}$ is at least $c-1$. This yields that

$$M_{mi} - (|\mathcal{P}_{mi}|_0 - 1)\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{c-1} \leq R_{mi} \leq M_{mi} + (|\mathcal{P}_{mi}|_0 - 1)\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{c-1}. \tag{2.50}$$

Combining (2.50) and (2.49) leads to

$$|T_{ii} - 1| \leq (|\mathcal{P}_{mi}|_0 - 1)\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{c-1}\left(\sum_{m \in \mathcal{N}(i)} \frac{M_{im}}{1-{M_{im}}^2}\right) \leq \deg(P_{\max} - 1)\frac{\|\tilde{\Sigma}^{\text{res}}\|_{\max}^{c}}{1-\|\tilde{\Sigma}^{\text{res}}\|_{\max}^{2}} \leq \tilde{\epsilon}, \tag{2.51}$$

where the last inequality is due to Lemma 11 and the fact that $\lceil \frac{c}{2} \rceil \leq c$ for $c \geq 3$. Now, consider T_{ij} for a pair (i,j) such that $i \neq j$. We have

$$T_{ij} = \sum_{m=1}^{n} \tilde{A}_{im} R_{mj} = \left(1 + \sum_{m \in \mathcal{N}(i)} \frac{{M_{im}}^2}{1-{M_{im}}^2}\right) R_{ij} - \sum_{m \in \mathcal{N}(i)} \frac{M_{im}}{1-{M_{im}}^2} R_{mj}. \tag{2.52}$$

According to Lemma 9, one can write

$$R^s_{m'j} M_{im'} - (|\mathcal{P}_{ij}|_0 - 1)\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\lceil \frac{c}{2} \rceil} \leq R_{ij} \leq R^s_{m'j} M_{im'} + (|\mathcal{P}_{ij}|_0 - 1)\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\lceil \frac{c}{2} \rceil}, \tag{2.53a}$$

$$\begin{aligned} R^s_{m'j} M_{im'} M_{im} - (|\mathcal{P}_{mj}|_0 - 1)\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\lceil \frac{c}{2} \rceil - 1} &\leq R_{mj} \\ &\leq R^s_{m'j} M_{im'} M_{im} + (|\mathcal{P}_{mj}|_0 - 1)\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\lceil \frac{c}{2} \rceil - 1}, \end{aligned} \tag{2.53b}$$

where m' is the node adjacent to i in d_{ij}^s and $m \in \mathcal{N}(i)\backslash m'$. Note that if $\mathcal{N}(i)\backslash m'$ is empty, then $R_{ij} = R_{m'j}^s M_{im'}$ and $R_{mj} = R_{m'j}^s M_{im'} \tilde{\Sigma}_{im}^{\text{res}}$. In this case, an argument similar to the proof of Lemma 4 can be made to show that $T_{ij} = 0$. Now, assume that $\mathcal{N}(i)\backslash m'$ is not empty. One can write

$$|T_{ij} - F_{ij}| \overset{(a)}{=} |T_{ij}| \overset{(b)}{\leq} \tilde{\epsilon}, \tag{2.54}$$

where

$$F_{ij} = \left(\frac{1}{1 - {M_{im'}}^2} + \sum_{m \in \mathcal{N}(i)\backslash m'} \frac{{M_{im}}^2}{1 - {M_{im}}^2} \right) R_{m'j}^s M_{im'} - \frac{M_{im'}}{1 - {M_{im'}}^2} R_{m'j}^s$$
$$- \sum_{m \in \mathcal{N}(i)\backslash m'} \frac{{M_{im}}^2}{1 - {M_{im}}^2} R_{m'j}^s M_{im'} M_{im}.$$

Note that the relation (a) can be verified by the fact that $F_{ij} = 0$ and the inequality (b) is obtained by combining (2.52) with (2.53a) and (2.53b). The inequalities (2.51) and (2.54) imply that $M + N$ is an $\tilde{\epsilon}$-relaxed inverse of $\tilde{A}$.

Now, it will be shown that the pair $(\tilde{A}, M + N)$ satisfies the $\tilde{\epsilon}$-relaxed KKT conditions. Note that $M_{ii} + N_{ii} = M_{ii} = \tilde{\Sigma}_{ii}$ and, hence, (2.15a) is satisfied. To prove (2.15b), since $\text{sign}(\tilde{A}_{ij}) = -\text{sign}(M_{ij}) = -\text{sign}(\tilde{\Sigma}_{ij})$, it can be concluded that

$$M_{ij} + N_{ij} = (\tilde{\Sigma}_{ij} - \tilde{\lambda}_{ij} \times \text{sign}(\Sigma_{ij})) + N_{ij} = (\tilde{\Sigma}_{ij} + \tilde{\lambda}_{ij} \times \text{sign}(\tilde{A}_{ij})) + N_{ij},$$

for every (i, j) such that $i \neq j$ and $\tilde{A}_{ij} \neq 0$. Due to the definition of N and the fact that $(i, j) \in \text{supp}(M)$, we have $|N_{ij}| \leq (P_{\max} - 1) \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max} \right)^{c-1}$. Hence,

$$|M_{ij} + N_{ij} - (\tilde{\Sigma}_{ij} + \tilde{\lambda}_{ij} \times \text{sign}(\tilde{A}_{ij}))| \leq \epsilon,$$

for every (i, j) such that $i \neq j$ and $\tilde{A}_{ij} \neq 0$. Therefore, the pair $(\tilde{A}, M + N)$ satisfies (2.15b). Finally, consider a pair (i, j) such that $i \neq j$ and $\tilde{A}_{ij} = 0$. One can write

$$M_{ij} + N_{ij} = R_{ij}^s + \sum_{d_{ij} \in \mathcal{P}_{ij}\backslash d_{ij}^s} \prod_{(m,t) \in d_{ij}} \tilde{\Sigma}_{mt}^{\text{res}}.$$

If $\mathcal{P}_{ij}\backslash d_{ij}^s$ is empty, a set of inequalities similar to (2.24) and (2.25) can be obtained to prove (2.15c). Now, assume that $\mathcal{P}_{ij}\backslash d_{ij}^s$ is not empty. The length of d_{ij}^s is at least 2 since there is no direct edge between nodes i and j. Hence, $|R_{ij}^s| \leq \|\tilde{\Sigma}^{\text{res}}\|_{\max}^2$. Furthermore, due to Lemma (9), the length of every path $d_{ij} \in \mathcal{P}_{ij}\backslash d_{ij}^s$ is at least $\lceil c/2 \rceil$. This leads to

$$|M_{ij} + N_{ij}| \leq \|\tilde{\Sigma}^{\text{res}}\|_{\max}^2 + (P_{\max} - 1) \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max} \right)^{\lceil \frac{c}{2} \rceil}$$
$$\leq \min_{\substack{k \neq l \\ (k,l) \notin \mathcal{G}(\Sigma^{\text{res}})}} (\tilde{\lambda}_{kl} - |\tilde{\Sigma}_{kl}^{\text{res}}|) + (P_{\max} - 1) \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max} \right)^{\lceil \frac{c}{2} \rceil}$$
$$\leq \tilde{\lambda}_{ij} - |\tilde{\Sigma}_{ij}^{\text{res}}| + (P_{\max} - 1) \left(\|\tilde{\Sigma}^{\text{res}}\|_{\max} \right)^{\lceil \frac{c}{2} \rceil},$$

where the last inequality follows from Condition (2-ii) in the theorem. Therefore,

$$|M_{ij}+N_{ij}-\tilde{\Sigma}_{ij}| \leq |M_{ij}+N_{ij}|+|\tilde{\Sigma}_{ij}| \leq \tilde{\lambda}_{ij} - |\tilde{\Sigma}^{\text{res}}_{ij}| + |\tilde{\Sigma}^{\text{res}}_{ij}| + (P_{\max}-1)\left(\|\tilde{\Sigma}^{\text{res}}\|_{\max}\right)^{\lceil \frac{c}{2} \rceil} \leq \tilde{\lambda}_{ij} + \tilde{\epsilon}.$$

This shows that $(\tilde{A}, M+N)$ indeed satisfies the $\tilde{\epsilon}$-relaxed KKT conditions for the normalized GL. Finally, we consider the explicit solution A defined as (2.13). The following statements hold:

1. the matrix $D^{1/2}(M+N)D^{1/2}$ is ϵ-relaxed inverse of A. To see this, note that

$$\begin{aligned} A\left(D^{1/2}(M+N)D^{1/2}\right) &= D^{-1/2}\tilde{A}D^{-1/2}D^{1/2}(M+N)D^{1/2} \\ &= D^{-1/2}TD^{1/2} \\ &= I_n + E, \end{aligned}$$

   where $\|E\|_{\max} \leq \sqrt{\frac{\Sigma_{\max}}{\Sigma_{\min}}}\tilde{\epsilon} \leq \epsilon$.

2. The pair $(A, D^{1/2}(M+N)D^{1/2})$ satisfies the ϵ-relaxed KKT conditions. Note that it is already shown that $(\tilde{A}, M+N)$ satisfies the following inequalities

$$(M+N)_{ij} = \tilde{\Sigma}_{ij} \qquad \text{if} \quad i=j, \tag{2.55a}$$

$$\left|(M+N)_{ij} - \left(\tilde{\Sigma}_{ij} + \tilde{\lambda}_{ij} \times \text{sign}(\tilde{A}_{ij})\right)\right| \leq \tilde{\epsilon} \qquad \text{if} \quad \tilde{A}_{ij} \neq 0, \tag{2.55b}$$

$$\left|(M+N)_{ij} - \tilde{\Sigma}_{ij}\right| \leq \tilde{\lambda}_{ij} + \tilde{\epsilon} \qquad \text{if} \quad \tilde{A}_{ij} = 0. \tag{2.55c}$$

   Replacing $M+N$ with $D^{1/2}(M+N)D^{1/2}$ and modifying (2.55) accordingly, one can verify that $(A, D^{1/2}(M+N)D^{1/2})$ satisfies ϵ-relaxed KKT conditions for the GL, where

$$\epsilon = \max\left\{\Sigma_{\max}, \sqrt{\frac{\Sigma_{\max}}{\Sigma_{\min}}}\right\}\tilde{\epsilon}.$$

This completes the proof. □

Proof of Theorem 5

Due to Theorem 4, the equation

$$D^{1/2}(M+N)D^{1/2} = A^{-1} + A^{-1}E \tag{2.56}$$

holds for every λ greater than or equal to λ_0, where $\|E\|_{\max} \leq \epsilon$. Since the pair $(A, D^{1/2}(M+N)D^{1/2})$ satisfies the ϵ-relaxed KKT conditions, it follows from (2.56) that

$$(A)_{ij}^{-1} = \Sigma_{ij} - (A^{-1}E)_{ij} = \hat{\Sigma}_{ij} \qquad \text{if} \quad i = j, \tag{2.57a}$$

$$(A)_{ij}^{-1} = \underbrace{\Sigma_{ij} + t_{ij}\epsilon - (A^{-1}E)_{ij}}_{\hat{\Sigma}_{ij}} + \lambda \times \operatorname{sign}(A_{ij}) \qquad \text{if} \quad A_{ij} \neq 0, \tag{2.57b}$$

$$\underbrace{\Sigma_{ij} + s_{ij}\epsilon - (A^{-1}E)_{ij}}_{\hat{\Sigma}_{ij}} - \lambda \leq (A)_{ij}^{-1} \leq \underbrace{\Sigma_{ij} + s_{ij}\epsilon - (A^{-1}E)_{ij}}_{\hat{\Sigma}_{ij}} + \lambda \qquad \text{if} \quad A_{ij} = 0, \tag{2.57c}$$

for some numbers t_{ij} and s_{ij} in the interval $[-1, 1]$. To complete the proof, it suffices to show that the matrix F defined as

$$\Sigma_{ij} - \hat{\Sigma}_{ij} = F_{ij} = \begin{cases} -(A^{-1}E)_{ij} & \text{if} \quad i = j, \\ t_{ij}\epsilon - (A^{-1}E)_{ij} & \text{if} \quad A_{ij} \neq 0, \\ s_{ij}\epsilon - (A^{-1}E)_{ij} & \text{if} \quad A_{ij} = 0 \end{cases} \tag{2.58}$$

satisfies the inequality $|||F|||_2 \leq d_{\max}\left(1/\mu_{\min}(A) + 1\right)\epsilon$. To this end, it is enough to prove that $|||A^{-1}E|||_2 \leq (d_{\max}/\mu_{\min}(A))\epsilon$, since $|||F - A^{-1}E|||_2 \leq d_{\max}(A)\epsilon$. One can write

$$\left|\left|\left|A^{-1}E\right|\right|\right|_2 \leq \left|\left|\left|A^{-1}\right|\right|\right|_2 |||E|||_2 \leq d_{\max}(A)\left|\left|\left|A^{-1}\right|\right|\right|_2 \|E\|_{\max} = \left(\frac{d_{\max}(A)}{\mu_{\min}(A)}\right)\epsilon,$$

which shows the validity of (2.18).

Next, we prove the inequality (2.19). The following chain of inequalities hold

$$\begin{aligned} -\log\det(A) + \operatorname{trace}(\hat{\Sigma}A) + \lambda\|A\|_{1,\text{off}} &= \underbrace{-\log\det(A) + \operatorname{trace}(\Sigma A) + \lambda\|A\|_{1,\text{off}}}_{f(A)} \\ &\quad + \operatorname{trace}((\hat{\Sigma} - \Sigma)A) \\ &\overset{(a)}{\leq} -\log\det(S^{\text{opt}}) + \operatorname{trace}(\hat{\Sigma}S^{\text{opt}}) + \lambda\|S^{\text{opt}}\|_{1,\text{off}} \\ &= \underbrace{-\log\det(S^{\text{opt}}) + \operatorname{trace}(\Sigma S^{\text{opt}}) + \lambda\|S^{\text{opt}}\|_{1,\text{off}}}_{f^*} \\ &\quad + \operatorname{trace}((\hat{\Sigma} - \Sigma)S^{\text{opt}}), \end{aligned}$$

where (a) is due to the fact that A is optimal for the GL with the perturbed sample covariances. This implies that

$$\begin{aligned} f(A) - f^* &\leq \operatorname{trace}((\tilde{\Sigma} - \Sigma)(S^{\text{opt}} - A)) \\ &\leq \left|\left|\left|\tilde{\Sigma} - \Sigma\right|\right|\right|_2 (|||S^{\text{opt}}|||_2 + |||A|||_2) \\ &\leq \left(\mu_{\max}(A) + \mu_{\max}(S^{\text{opt}})\right) d_{\max}(A)\left(\frac{1}{\mu_{\min}(A)} + 1\right)\epsilon. \end{aligned}$$

□

Chapter 3

Global Guarantees on Robust Matrix Recovery

This chapter is concerned with the non-negative rank-1 robust principal component analysis (RPCA), where the goal is to recover the dominant non-negative principal components of a data matrix *precisely*, where a number of measurements could be grossly corrupted with sparse and arbitrary large noise. Most of the known techniques for solving the RPCA rely on convex relaxation methods by lifting the problem to a higher dimension, which significantly increase the number of variables. As an alternative, the well-known Burer-Monteiro approach can be used to cast the RPCA as a non-convex and non-smooth ℓ_1 optimization problem with a significantly smaller number of variables. In this work, we show that the low-dimensional formulation of the symmetric and asymmetric positive rank-1 RPCA based on the Burer-Monteiro approach has benign landscape, i.e., 1) it does not have any spurious local solution, 2) has a unique global solution, and 3) its unique global solution coincides with the *true* components. An implication of this result is that simple local search algorithms are guaranteed to achieve a zero global optimality gap when directly applied to the low-dimensional formulation. Furthermore, we provide strong deterministic and probabilistic guarantees for the exact recovery of the true principal components. In particular, it is shown that a constant fraction of the measurements could be grossly corrupted and yet they would not create any spurious local solution.

3.1 Introduction

The principal component analysis (PCA) is perhaps the most widely-used dimension-reduction method that reveals the components with maximum variability in high-dimensional datasets. In particular, given the data matrix $X \in \mathbb{R}^{m\times n}$, where each row corresponds to a data sample with size n, the goal is to recover its most dominant component under the

rank-1 spiked model[1]

$$X = \beta \mathbf{u}\mathbf{v}^\top + S \tag{3.1}$$

where β determines the signal-to-noise ratio, S is the additive noise matrix, and $\mathbf{u}$ and $\mathbf{v}$ are two unknown unit norm vectors. If the data matrix X is symmetric (for instance, it corresponds to a sample covariance matrix), then (3.1) can be modified as

$$X = \beta \mathbf{v}\mathbf{v}^\top + S \tag{3.2}$$

Depending on the nature of the noise matrix, different methods have been proposed in the literature to recover the principal components from (partial) observations of X. The problem of recovering β, $\mathbf{u}$, and $\mathbf{v}$ under a Gaussian and sparse noise is conventionally referred to as PCA and robust PCA (or RPCA), respectively.

The properties of both PCA and its robust analog have been heavily studied in the literature and their applications span from quantitative finance to health care and neuroscience ([128, 48, 39]). Recently, a special focus has been devoted to further exploiting the prior knowledge on the principal components, such as sparsity ([286]) and nonlinearity ([110]). Accordingly, one such knowledge appearing in different applications is the non-negativity of the principal components ([189]). In this scenario, one needs to solve the PCA or the RPCA under the additional constraints $\mathbf{u}, \mathbf{v} \geq 0$. While the non-negative PCA has been recently studied in [189], the main focus of our work is on its robust variant, where the noise matrix is assumed to be sparse and the goal is the *exact* recovery of the non-negative vectors $\mathbf{u}$ and $\mathbf{v}$. Note that the non-negativity of principal components naturally arises in many real-world problems. In what follows, we will present two classes of real-world applications for which the non-negative RPCA is useful.

1. Non-negative matrix factorization: Extracting the dominant principal component of a symmetric or asymmetric data matrix appears in many applications and the examples are ubiquitous. For instance, an important problem in astronomy is the recovery of non-negative astronomical signals from the covariance matrix of photometric observations ([213]). The measured data samples are prone to sparse and random outliers. Similarly, one can extract moving objects from video frames via non-negative matrix factorization by treating the background as the dominant low-rank component in the video frames and the moving object as sparse noise (the non-negativity of the data is due to the non-negative values of the pixels) ([156, 46]). We will conduct a case study on this application later in this chapter.

2. Gene networks: Gene activities can be captured by the samples collected from different organs, and are described by multi-spiked models ([155]):

$$X = X_0 + \sum_{i=1}^{k} \mathbf{u}_{(i)} \mathbf{v}_{(i)}^\top \tag{3.3}$$

[1]There are more general models under which the PCA is shown to be useful (see [131] for more details). We use the rank-1 spiked model since it fits into our framework and is often used as a baseline to evaluate the performance of the PCA.

where $(i,j)^{\text{th}}$ entry of X measures the strength of the participation of gene i in sample j and X_0 is an offset. Furthermore, k is the number of the gene-block, and $\mathbf{u}_{(i)}$ and $\mathbf{v}_{(i)}$ measure the participation of different genes and samples in the i^{th} gene-block. The participation vectors are non-negative and the measurements can be subject to malfunctioning of the measurement tools. Therefore, the problem of obtaining $\mathbf{u}_{(i)}$ and $\mathbf{v}_{(i)}$ can be cast as a non-negative RPCA with multiple principal components.

The seminal work by [46] proposes a sparsity promoting convex relaxation for the RPCA that is capable of the exact recovery of $\mathbf{u}$ and $\mathbf{v}$. Upon defining $W = \mathbf{u}\mathbf{v}^\top$, the convex relaxation of the RPCA is defined as

$$\min_{W \in \mathbb{R}^{m\times n}} \quad \|W\|_* + \lambda \|\mathcal{P}_\Omega(X - W)\|_1 \tag{3.4}$$

where $\|W\|_*$ is the nuclear norm of W, serving as a penalty on the rank of the recovered matrix W, and $\|\cdot\|_1$ is used to denote the element-wise ℓ_1 norm. Furthermore, $\mathcal{P}_\Omega(\cdot)$ is the projection onto the set of matrices with the same support as the measurement set Ω. Therefore, upon defining $S = X - W$ as the corruption or noise matrix, $\|\mathcal{P}_\Omega(X-W)\|_1$ plays the role of promoting sparsity in the estimated noise matrix. After finding an optimal value of W, the matrix can then be decomposed into the desired vectors $\mathbf{u}$ and $\mathbf{v}$, provided that the relaxation is exact. Notice that the problem is convexified via lifting from $n+m$ variables on $(\mathbf{u},\mathbf{v})$ to nm variables on W. Despite the convexity of the lifted problem, its dimension makes it prohibitive to solve in high-dimensional settings. To circumvent this issue, one popular approach is to resort to an alternative formulation, inspired by [41] (commonly known as the Burer-Monteiro technique):

$$\min_{\mathbf{u}\in\mathbb{R}^m_+, \mathbf{v}\in\mathbb{R}^n_+} \quad \|\mathcal{P}_\Omega(X - \mathbf{u}\mathbf{v}^\top)\|_1 \tag{3.5}$$

Despite the non-convexity of (3.5), its smooth counterpart (with or without non-negativity constraints) defined as

$$\min_{\mathbf{u}\in\mathbb{R}^m, \mathbf{v}\in\mathbb{R}^n} \quad \underbrace{\|\mathcal{P}_\Omega(X - \mathbf{u}\mathbf{v}^\top)\|_F^2}_{g(\mathbf{u},\mathbf{v})} \tag{3.6}$$

has been widely used in matrix completion/sensing and is known to possess *benign global landscape*, i.e., every local solution is also global and every saddle point has a direction with a strictly negative curvature ([31, 101, 100]). This will be stated below.

Theorem 6 (Informal, Benign Landscape ([100]))**.** *Under some technical conditions, a regularized version of* (3.6) *has benign landscape: every local minimum is global and every saddle point has a direction with a strictly negative curvature.*

In particular, both symmetric and asymmetric matrix completion (or matrix sensing) under dense Gaussian noise can be cast as (3.6) and in light of the above theorem, they

have benign landscape. However, it is well-known that such smooth norms are incapable of correctly identifying and rejecting sparse-but-large noise/outliers in the measurements.

Despite the generality of Theorem 6 within the realm of smooth norms, it does not address the following important question: *Does the non-smooth and non-negative rank-1 RPCA* (3.5) *have benign landscape?*

The Issue with the Known Proof Techniques

To understand the inherent difficulty of examining the landscape of (3.5), it is essential to explain why the existing proof techniques for the absence of spurious local minima in matrix sensing/completion cannot naturally be extended to their robust counterparts. In general, the main idea in the literature behind proving the benign landscape of matrix sensing/completion is based on analyzing the gradient and the Hessian of the objective function. More precisely, for every point that satisfies $\nabla g(\mathbf{u}, \mathbf{v}) = 0$ and does not correspond to a globally optimal minimum, it suffices to find a *global* direction of descent $\mathbf{d}$ such that $\text{vec}(\mathbf{d})^\top \nabla^2 g(\mathbf{u}, \mathbf{v}) \text{vec}(\mathbf{d}) < 0$, where $\text{vec}(\mathbf{d})$ is the vectorized version of $\mathbf{d}$ and $\nabla^2 g(\mathbf{u}, \mathbf{v})$ is the Hessian of $g(\mathbf{u}, \mathbf{v})$. Such a direction certifies that every stationary point that is not globally optimal must be either a local maximum or a saddle point with a strictly negative direction. However, this approach cannot be used to prove similar results for (3.5) mainly because the objective function of (3.5) is non-differentiable and, hence, the Hessian is not well-defined. This difficulty calls for a new methodology for analyzing the landscape of the robust and non-smooth PCA; a goal that is at the core of this work.

3.2 Overview of Contributions

In this work, we characterize the landscape of both the symmetric non-negative rank-1 RPCA defined as

$$\min_{\mathbf{u} \in \mathbb{R}^n_+} \quad \underbrace{\|\mathcal{P}_\Omega(X - \mathbf{u}\mathbf{u}^\top)\|_1 + R_\beta(\mathbf{u})}_{f_{\text{reg}}(\mathbf{u})} \tag{SN-RPCA}$$

and its asymmetric counterpart defined as

$$\min_{\mathbf{u} \in \mathbb{R}^m_+, \mathbf{v} \in \mathbb{R}^n_+} \quad \underbrace{\|\mathcal{P}_\Omega(X - \mathbf{u}\mathbf{v}^\top)\|_1 + R_\beta(\mathbf{u}, \mathbf{v})}_{f_{\text{reg}}(\mathbf{u}, \mathbf{v})} \tag{AN-RPCA}$$

In particular, we fully characterize the stationary points of these optimization problems, under both deterministic and probabilistic models for the measurement index Ω and the noise matrix S. The functions $R(\mathbf{u})$ and $R(\mathbf{u}, \mathbf{v})$ are regularization functions that prevent the solutions from *blowing up*; roughly speaking, they penalize the points whose norm is greater than β, but do not change the landscape otherwise. The exact definitions of these regularization functions will be presented later in Section 3.5.

Remark 4. *The focus of this chapter is on the symmetric and non-symmetric RPCA under the rank-1 spiked model. A natural extension to this model is its rank-r variant:*

$$X = UV^\top + S \tag{3.7}$$

where $U := \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_r\end{bmatrix} \in \mathbb{R}_+^{m\times r}$ *and* $V := \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_r\end{bmatrix} \in \mathbb{R}_+^{n\times r}$ *are non-negative matrices encompassing the r principal components of the model (the symmetric version can be defined in a similar manner). Furthermore, similar to the rank-1 case, S is a sparse noise matrix. Under this rank-r spiked model, the aim of the non-negative* ***rank-r*** *RPCA is to recover the non-negative matrices U and V given a subset of the elements of the noisy measurement matrix X. In Section 3.8, we will elaborate on the technical difficulties behind this extension. In addition, we will provide some empirical evidence to support that the developed results may hold for the general non-negative rank-r RPCA with* $r \geq 2$.

Definition 14. *Given the set* Ω*, two graphs are defined below:*

- *The sparsity graph* $\mathcal{G}(\Omega)$ *induced by* Ω *for an instance of* (SN-RPCA) *is defined as a graph with the vertex set* $V := \{1, 2, ..., n\}$ *that includes an edge* (i, j) *if* $(i, j) \in \Omega$.
- *The bipartite sparsity graph* $\mathcal{G}_{m,n}(\Omega)$ *induced by* Ω *for an instance of* (AN-RPCA) *is defined as a graph with the vertex partitions* $V_u := \{1, 2, ..., m\}$ *and* $V_v := \{m + 1, m + 2, ..., m + n\}$ *that includes an edge* (i, j) *if* $(i, j - m) \in \Omega$.

Furthermore, define $\Delta(\mathcal{G}(\Omega))$ *and* $\delta(\mathcal{G}(\Omega))$ *as the maximum and minimum degrees of the nodes in* $\mathcal{G}(\Omega)$*, respectively. Similarly,* $\Delta(\mathcal{G}_{m,n}(\Omega))$ *and* $\delta(\mathcal{G}_{m,n}(\Omega))$ *are used to refer to the maximum and minimum degrees of the nodes in* $\mathcal{G}_{m,n}(\Omega)$*, respectively.*

Definition 15. *The sets of* ***bad/corrupted*** *and* ***good/correct*** *measurements are defined as* $B = \{(i, j)|(i, j) \in \Omega, S_{ij} \neq 0\}$ *and* $G = \{(i, j)|(i, j) \in \Omega, S_{ij} = 0\}$*, respectively.*

Based on the above definitions, the sparsity graph is allowed to include self-loops. For a positive vector $\mathbf{x}$, we denote its maximum and minimum values with $x_{\max}$ and $x_{\min}$, respectively. Furthermore, define $\kappa(\mathbf{x}) = \frac{x_{\max}}{x_{\min}}$ as the condition number of the vector $\mathbf{x}$. The first result of this chapter develops deterministic conditions on the measurement set Ω and the sparsity pattern of the noise matrix S to guarantee that the positive rank-1 RPCA has benign landscape. Let $\mathbf{u}^*$ and $(\mathbf{u}^*, \mathbf{v}^*)$ denote the true principal components of (SN-RPCA) and (AN-RPCA), respectively.

Theorem 7 (Informal, Deterministic Guarantee)**.** *Assuming that* $\mathbf{u}^*, \mathbf{v}^* > 0$*, there exist regularization functions* $R(\mathbf{u})$ *and* $R(\mathbf{u}, \mathbf{v})$ *such that the following statements hold with overwhelming probability:*

1. (SN-RPCA) *has no spurious local minimum and has a unique global minimum that coincides with the true component, provided that* $\mathcal{G}(G)$ *has* ***no bipartite component*** *and*

$$\kappa(\mathbf{u}^*)^4 \Delta(\mathcal{G}(B)) \lesssim \delta(\mathcal{G}(G)) \tag{3.8}$$

2. (AN-RPCA) *has no spurious local minimum and has a unique global minimum that coincides with the true components, provided that* $\mathcal{G}_{m,n}(G)$ *is* ***connected*** *and*

$$\max\left\{\kappa(\mathbf{u}^*)^4, \kappa(\mathbf{v}^*)^4\right\}\Delta(\mathcal{G}_{m,n}(B)) \lesssim \delta(\mathcal{G}_{m,n}(G)) \tag{3.9}$$

Theorem 7 puts forward a set of deterministic conditions for the absence of spurious local solutions in (SN-RPCA) and (AN-RPCA) as well as the uniqueness of the global solution. Notice that no upper bound is assumed on the values of the nonzero entries in the noise matrix. The reasoning behind the conditions imposed on the minimum and maximum degrees of the nodes in the sparsity graph of the measurement set is to ensure the identifiability of the problem. We will elaborate more on this subtle point later in Section 3.5. Furthermore, we will show later in this chapter that some of the conditions delineated in Theorem 7—such as the strict positivity of $\mathbf{u}^*$ and $\mathbf{v}^*$, as well as the absence of bipartite components in $\mathcal{G}(G)$ for (SN-RPCA)—are also necessary for the exact recovery.

The second main result of this chapter investigates (SN-RPCA) and (AN-RPCA) under random sampling and noise structures. In particular, suppose that each element (in the symmetric case, each element of the upper triangular part) of S is nonzero with probability d. Then, for every (i, j), we have

$$X_{ij} = \begin{cases} u_i^* v_j^* & \text{with probability } 1-d \\ \text{arbitrary} & \text{with probability } d \end{cases} \tag{3.10}$$

Furthermore, suppose that every element of X is measured with probability p. In other words, every (i, j) belongs to Ω with probability p. Finally, we assume that the noise and sampling events are independent.

Theorem 8 (Informal, Probabilistic Guarantee)**.** *Assuming that* $\mathbf{u}^*, \mathbf{v}^* > 0$*, there exist regularization functions* $R(\mathbf{u})$ *and* $R(\mathbf{u}, \mathbf{v})$ *such that the following statements hold with overwhelming probability:*

1. (SN-RPCA) *has no spurious local minimum and has a unique global minimum that coincides with the true component, provided that*

$$p \gtrsim \frac{\kappa(\mathbf{u}^*)^4 \log n}{n}, \qquad d \lesssim \frac{1}{\kappa(\mathbf{u}^*)^4} \tag{3.11}$$

2. (AN-RPCA) *has no spurious local minimum and has a unique global minimum that coincides with the true components, provided that*

$$p \gtrsim \frac{\kappa(\mathbf{w}^*)^4 n \log n}{m^2}, \qquad d \lesssim \frac{r}{\kappa(\mathbf{w}^*)^4} \tag{3.12}$$

where $\mathbf{w}^* = \begin{bmatrix}\mathbf{u}^{*\top} & \mathbf{v}^{*\top}\end{bmatrix}^\top$*,* $r = m/n$*, and* $n \geq m$*.*

A number of interesting corollaries can be obtained based on Theorem 8. For instance, it can be inferred that the exact recovery is guaranteed even if the number of grossly corrupted measurements is on the same order as the total number of measurements, provided that $\frac{u^*_{\max}}{u^*_{\min}}$ is uniformly bounded from above.

In addition to the absence of spurious local minima and the uniqueness of the global minimum, the next proposition states that the true solution can be recovered via local search algorithms for non-smooth optimization.

Proposition 1 (Informal, Global Convergence)**.** *Under the assumptions of Theorem 7 and 8, local search algorithms converge to the true solutions of* (SN-RPCA) *and* (AN-RPCA) *with overwhelming probability.*

Starting from Section 3.2, we will delve into the detailed analysis of the symmetric and asymmetric non-negative RPCA. In particular, we will analyze (SN-RPCA) and (AN-RPCA) under different deterministic and probabilistic settings and provide formal versions of Theorems 7 and 8.

Preliminaries

A **directional derivative** of a locally Lipschitz and possibly non-smooth function $h(\mathbf{x})$ at $\mathbf{x}$ in the direction $\mathbf{d}$ is defined as

$$h'(\mathbf{x}, \mathbf{d}) := \lim_{t \downarrow 0} \frac{h(\mathbf{x} + t\mathbf{d}) - h(\mathbf{x})}{t} \tag{3.13}$$

upon existence. Based on this definition, $\bar{\mathbf{u}}$ is **directional-minimum-stationary** (or D-min-stationary) for (SN-RPCA) if $f'(\bar{\mathbf{u}}, \mathbf{d}) \geq 0$ for every *feasible* direction $\mathbf{d}$, i.e., a direction that satisfies $d_i \geq 0$ when $u_i = 0$ for every index i. Similarly, $\bar{\mathbf{u}}$ is **directional-maximum-stationary** (or D-max-stationary) for (SN-RPCA) if $f'(\bar{\mathbf{u}}, \mathbf{d}) \leq 0$ for every feasible $\mathbf{d}$. Finally, $\bar{\mathbf{u}}$ is **directional-stationary** (or D-stationary) for (SN-RPCA) if it is either D-min- or D-max-stationary[1].

Every local minimum (maximum) $\bar{\mathbf{u}}$ should be D-min (max)-stationary for $f(\mathbf{u})$. On the other hand, $\bar{\mathbf{u}}$ cannot be a D-stationary point if $f(\mathbf{u})$ has strictly positive and negative directional derivatives at that point. In that case, $\bar{\mathbf{u}}$ is neither local maximum nor minimum. A solution to a minimization problem is referred to as **spurious local** (or simply local) if there exists another feasible point with a strictly smaller objective value; a solution is **globally optimal** (or simply global) if no such point exists.

[1]Note that the notion of D-stationary points is often used in lieu of D-min-stationary in the literature. However, we use a slightly more general definition in this chapter to account for the local maxima of (SN-RPCA).

Finally, a **vertex partitioning** of a non-empty bipartite graph is the partition of its vertices into two groups such that there exist no adjacent vertices within each group.

3.3 Related Work

Non-convex and Low-rank Optimization

A considerable amount of work has been carried out to understand the inherent difficulty of solving low-rank optimization problems both locally and globally.

Convexification: Recently, there has been a pressing need to develop efficient methods for solving large-scale nonconvex optimization problems that naturally arise in data analytics and machine learning ([69, 227, 36, 279, 199]). One promising approach for making these large-scale problems more tractable is to resort to their convex surrogates; these methods started to receive a great deal of attention after the seminal works by [65] and [44] on the *compressive sensing* and have been extended to emerging problems in machine learning, such as fairness ([199]), robust polynomial regression ([184, 173]), and neural networks ([14]), to name a few. Nonetheless, the size of today's problems has been a major impediment to the tractability of these methods. In practice, the dimension of the real-world problems is overwhelmingly large, often surpassing the ability of these seemingly efficient convex methods to solve the problem in a reasonable amount of time. Due to this so-called *curse of dimensionality*, the common practice is to deploy fast local search algorithms directly applied to the original nonconvex problem with the hope of converging to acceptable solutions. Roughly speaking, these methods can only guarantee the local optimality, thus exposing themselves to potentially large optimality gaps. However, a recent line of work has shown that a surprisingly large class of nonconvex problems, including matrix completion/sensing ([31, 101, 100, 284]), phase retrieval ([242]), and dictionary recovery ([243]) have *benign global landscape*, i.e., every local solution is also global and every saddle point has a direction with a strictly negative curvature (see [53] for a comprehensive survey on the related problems). More recently, the work by [280] has introduced a unified framework that shows the benign landscape of nonconvex low-rank optimization problems with general loss functions, provided that they satisfy certain restricted convexity and smoothness properties. This enables most of the saddle-escaping local search algorithms to converge to a global solution, thereby resulting in a zero optimality gap ([102]).

Benign landscape: As mentioned before, it has been recently shown that many low-rank optimization problems can be cast as smooth-but-nonconvex optimization problems that are free of spurious local minima. These methods heavily rely on the notion of *restricted isometry property* (RIP)—a property that was initially introduced by [45] and has been used ever since as a metric to measure a norm-preserving property of the objective function. In general, these methods have two major drawbacks: 1) they can only target a narrow set of nearly-isotropic instances ([278]), and 2) their proof technique depends on the differentiability

of the objective function; a condition that is not satisfied for non-smooth norms, such as ℓ_1. To the best of our knowledge, the work by [134] is the only one that studies the landscape of the ℓ_1 minimization problem, where the authors consider the tensor decomposition problem under the full and perfect measurements. Our work is somewhat related to [168] that derives similar conditions for the absence of spurious local solution of the non-negative rank-1 matrix completion but for the smooth Frobenius norm minimization problem.

PCA with prior information: With an exponential growth in the size and dimensionality of the real-world datasets, it is often required to exploit the additional prior information in the PCA. In many real-world applications, prior knowledge from the underlying physics of the problem—such as non-negativity ([189]), sparsity ([286]), robustness ([46]), and nonlinearity ([110])—can be taken into account to perform more efficient, consistent, and accurate PCA.

Numerical algorithms for non-smooth optimization: Numerical algorithms for non-smooth optimization problems can be dated back to the work by Clarke on the extended definitions of gradients and directional derivatives, commonly known as generalized derivatives ([54]). Intuitively, for non-smooth functions, the gradient in the classical sense seize to exist at a subset of the points in the domain. The Clarke generalized derivative is introduced to circumvent this issue by associating a convex differential to these points, even if the original problem is non-convex. In the domain of unconstrained non-smooth optimization, earlier works have introduced simple algorithms that converge to approximate Clarke-stationary points ([107, 51]). More recent methods take advantage of the fact that many non-smooth optimization problems are smooth in every open dense subset of their domains. This implies that the objective function is smooth with probability one at a randomly drawn point. This observation lays the groundwork for several gradient-sampling-based algorithms for both unconstrained and constrained non-smooth optimization problems ([42, 61]). As mentioned before, a sub-gradient method has been recently proposed by [163] for solving the RPCA, where the authors prove linear convergence of the algorithm to the true components, provided that the initial point is chosen sufficiently close to the globally optimal solution.

Comparison to the Existing Results on RPCA

Similar to the non-convex matrix sensing and completion, most of the existing results on the RPCA work on a *lifted* space of the variables via different convex relaxations and they do not incorporate the positivity constraints in the problem. In what follows, we will explain the advantages of our proposed method compared to these results.

Positivity constraints: In the present work, we show that the positivity of the true components is both sufficient and (almost) necessary for the absence of spurious local solutions. We use this prior knowledge to obtain sharp deterministic and probabilistic guarantees on the absence of spurious local minima for the RPCA based on the Burer-Monteiro formulation. For instance, we show that up to a constant factor of the measurements can be grossly corrupted and yet they do not introduce any spurious local solution. Considering the

fact that these results heavily rely on the positivity of the true components, it is unclear if similar "no spurious local minima" results hold for the general case without the positivity assumption. The statistical properties of these types of constraints have also been shown to be useful in the classical PCA by [189], where the authors show that by imposing positivity constraints on the principal components, one can guarantee its consistent recovery with smaller signal-to-noise ratio. It is also worthwhile to mention that the incorporation of the non-negativity/positivity constraints in the low-rank matrix recovery can be traced back to some earlier works on the non-negative matrix factorization problem ([156, 124]).

Computational savings: Similar to the convexification techniques in nonconvex optimization, most of the classical results on the RPCA relax the inherent non-convexity of the problem by lifting it to higher dimensions ([46, 50, 283, 127]). In particular, by moving from vector to matrix variables, they guarantee the convexity of the problem at the expense of significantly increasing the number of variables. In this work, we show that such lifting is not necessary for the positive rank-1 RPCA since—despite the non-convexity of the problem—it is free of spurious local solutions and, hence, simple local search algorithms converge to the true components when directly applied to its original formulation.

Sharp guarantees with mild conditions: In general, most of the existing results on RPCA for guaranteeing the recovery of the true components fall into two categories. First, a large class of methods rely on some deterministic conditions on the spectra of the dominant components and/or the structure of the sparse noise ([127, 50, 272]). For instance, the works by [127, 50] require the regularization coefficient to be within a specific interval that is defined in terms of the true principal components. Furthermore, the algorithm proposed by [272] requires prior knowledge on the density of the sparse noise matrix. Although being theoretically significant, these types of conditions cannot be easily verified and met in practice. With the goal of bypassing such stringent conditions, the second category of research has studied the RPCA under probabilistic models. These types of guarantees were popularized by [46, 266] and they do not rely on any prior knowledge on the true components or the density of the noise matrix. However, their success is contingent upon specific random models on the sparse noise or the spectra of the true components, neither of which may be satisfied in practice.

In contrast, the method proposed here does not rely on any prior knowledge on the true solution, other than the availability of an upper bound on the maximum absolute value of the elements in the principal components[2]. Furthermore, unlike the previous works, our results encompass *both deterministic and probabilistic* models under random sampling.

[2]Note that in most cases, these types of upper bounds can be immediately inferred by the domain knowledge; see e.g. our discussion on the moving object detection problem.

3.4 Base Case: Noiseless Non-negative RPCA

In this section, we consider the noiseless version of both symmetric and asymmetric non-negative RPCA. While not entirely obvious, the subsequent arguments are at the core of our proofs for the general noisy case. In the noiseless scenario, (SN-RPCA) is reduced to

$$\min_{\mathbf{u}\geq 0} \quad \underbrace{\sum_{(i,j)\in\Omega} |u_i u_j - u_i^* u_j^*|}_{f(\mathbf{u})} \tag{P1-Sym}$$

For the asymmetric problem (AN-RPCA), the solution is invariant to scaling. In other words, if $(\mathbf{u}, \mathbf{v})$ is a solution to (AN-RPCA), then $(\frac{1}{q}\mathbf{u}, q\mathbf{v})$ is also a valid solution with the same objective value, for every scalar $q > 0$. To circumvent the issue of invariance to scaling, it is common to balance the norms of $\mathbf{u}$ and $\mathbf{v}$ by penalizing their difference. Therefore, similar to the works by [100, 282, 272], we consider the following regularized variant of (AN-RPCA):

$$\min_{\mathbf{u}\geq 0, \mathbf{v}\geq 0} \quad \underbrace{\|\mathcal{P}_\Omega(X - \mathbf{u}\mathbf{v}^\top)\|_1 + \alpha|\mathbf{u}^\top\mathbf{u} - \mathbf{v}^\top\mathbf{v}|}_{f_{\text{asym}}(\mathbf{u},\mathbf{v})} \tag{3.14}$$

for an arbitrary constant $\alpha > 0$ (note that the positivity of α is the only condition required in this work). To deal with the asymmetric case, we first convert it to a symmetric problem after a simple concatenation of variables. Define $\mathbf{w} = [\mathbf{u}^\top \quad \mathbf{v}^\top]^\top$, $\mathbf{w}^* = [\mathbf{u}^{*\top} \quad \mathbf{v}^{*\top}]^\top$, and $\bar{\Omega} = \{(i,j)|(i,j-m) \in \Omega\}$. Based on these definitions, one can symmetrize (3.14) as follows:

$$\min_{\mathbf{w}\geq 0} \quad \underbrace{\sum_{(i,j)\in\bar{\Omega}} |w_i w_j - w_i^* w_j^*| + \alpha \left|\sum_{i=1}^{m} w_i^2 - \sum_{j=m+1}^{m+n} w_j^2\right|}_{f_{\text{sym}}(\mathbf{w})} \tag{P1-Asym}$$

To simplify the notation, we drop the subscript from $f_{\text{sym}}(\mathbf{w})$ whenever there is no ambiguity in the context.

Deterministic Guarantees

Symmetric case: First, we introduce deterministic conditions to guarantee a benign landscape for (P1-Sym).

Theorem 9. *Suppose that* $\mathbf{u}^* > 0$ *and* $\mathcal{G}(\Omega)$ *has no bipartite component. Then, the following statements hold for* (P1-Sym)*:*

1. *It does not have any spurious local minimum;*
2. *The point* $\mathbf{u} = \mathbf{u}^*$ *is the unique global minimum;*

3. In the positive orthant, the point $\mathbf{u} = \mathbf{u}^*$ *is the only D-stationary point.*

Additionally, if $\mathcal{G}(\Omega)$ *is connected, the following statements hold for* (P1-Sym)*:*

4. The points $\mathbf{u} = \mathbf{u}^*$ *and* $\mathbf{u} = 0$ *are the only D-min-stationary points;*

5. The point $\mathbf{u} = 0$ *is a local maximum.*

The above theorem has a number of important implications for (P1-Sym): 1) it has no spurious local solution, 2) $\mathbf{u} = \mathbf{u}^*$ is its unique global solution, and 3) every feasible point $\mathbf{u} > 0$ such that $\mathbf{u} \neq \mathbf{u}^*$ has at least a strictly negative directional derivative. Additionally, if $\mathcal{G}(\Omega)$ is connected, the feasible points of (P1-Sym) with zero entries either have a strictly negative directional derivative or correspond to the origin that is a local maximum with a strictly negative curvature. Therefore, these points are not local/global minima and can be easily avoided using local search algorithms.

To prove Theorem 9, we first need the following important lemma.

Lemma 12. *Suppose that* $\mathcal{G}(\Omega)$ *has no bipartite component and* $\mathbf{u}^* > 0$*. Then, for every D-min-stationary point* $\mathbf{u}$ *of* (P1-Sym)*, we have* $\mathbf{u}[c] > 0$ *or* $\mathbf{u}[c] = 0$*, where* $\mathbf{u}[c]$ *is a sub-vector of* $\mathbf{u}$ *induced by the* c^{th} *component of* $\mathcal{G}(\Omega)$*.*

Now, we are ready to present the proof of Theorem 9.

Proof of Theorem 9: We prove the first three statements. Note that Statement 5 can be easily verified and Statement 4 is implied by Lemma 12 and Statement 3.

Suppose that $\mathbf{u} \neq \mathbf{u}^*$ is a local minimum. Note that if $u_i = 0$ for some i, Lemma 12 implies that $\mathbf{u}[c] = 0$ for the c^{th} component that includes node i. However, a strictly positive perturbation of $\mathbf{u}[c]$ decreases the objective function and, therefore, $\mathbf{u}$ cannot be a local minimum. Hence, it is enough to consider the case $\mathbf{u} > 0$. We show that $\mathbf{u}$ cannot be D-stationary. This immediately certifies the validity of the first three statements. First, we prove that

$$\min_{k\in\Omega_i} \frac{u_k^*}{u_k} \leq \frac{u_i}{u_i^*} \leq \max_{k\in\Omega_i} \frac{u_k^*}{u_k} \tag{3.15}$$

for every $i \in \{1, \cdots, n\}$, where $\Omega_i = \{j|(i,j) \in \Omega\}$. By contradiction and without loss of generality, suppose that $u_i/u_i^* > \max_{k\in\Omega_i} u_k^*/u_k$ for some i. This implies that $u_i u_j > u_i^* u_j^*$ for every $j \in \Omega_i$. Therefore, a negative or positive perturbation of u_i results in respective negative or positive directional derivatives, contradicting the D-stationarity of $\mathbf{u}$. With no loss of generality, assume that the sparsity graph $\mathcal{G}(\Omega)$ is connected (since the arguments made in the sequel can be readily applied to every disjoint component of $\mathcal{G}(\Omega)$) and that the following ordering holds:

$$0 < \frac{u_1^*}{u_1} \leq \frac{u_2^*}{u_2} \leq \cdots \leq \frac{u_n^*}{u_n} \tag{3.16}$$

Therefore, due to (4.13), we have

$$0 < \frac{u_1^*}{u_1} \leq \min_{k\in\Omega_i} \frac{u_k^*}{u_k} \leq \frac{u_i}{u_i^*} \leq \max_{k\in\Omega_i} \frac{u_k^*}{u_k} \leq \frac{u_n^*}{u_n} \tag{3.17}$$

for every $i \in \{1, \cdots, n\}$.

Since $\mathbf{u} \neq \mathbf{u}^*$, there exists some index t such that $u_t \neq u_t^*$. This implies that $u_n^*/u_n > 1$; otherwise, we should have $u_n^*/u_n \leq 1$. This together with (4.54), implies that $u_t^*/u_t < 1$ and $u_t/u_t^* > 1$, which contradicts (4.20). Now, define the sets

$$T_1 = \left\{ i | \frac{u_i^*}{u_i} = \frac{u_n^*}{u_n}, 1 \leq i \leq n \right\} \tag{3.18}$$

$$T_2 = \left\{ j | \frac{u_j}{u_j^*} = \frac{u_n^*}{u_n}, 1 \leq j \leq n \right\} \tag{3.19}$$

Moreover, define the set $N = V \backslash (T_1 \cup T_2)$ and let $\mathbf{d}$ be

$$d_i = \begin{cases} \frac{u_i}{u_n} & \text{if } i \in T_1 \\ -\frac{u_i}{u_n} & \text{if } i \in T_2 \\ 0 & \text{if } i \in N \end{cases} \tag{3.20}$$

Define a perturbation of $\mathbf{u}$ as $\hat{\mathbf{u}} = \mathbf{u} + \mathbf{d}\epsilon$ where $\epsilon > 0$ is chosen to be sufficiently small. Next, the effect of the above perturbation on different terms of (P1-Sym) will be analyzed. To this goal, we divide Ω into four sets

1. $(i,j) \in \Omega$ and $i, j \in T_1$: In this case, since $u_i < u_i^*$ and $u_j < u_j^*$, one can write

$$\begin{aligned} |\hat{u}_i \hat{u}_j - u_i^* u_j^*| = u_i^* u_j^* - \hat{u}_i \hat{u}_j &= u_i^* u_j^* - \left(u_i + \frac{u_i}{u_n}\epsilon \right) \left(u_j + \frac{u_j}{u_n}\epsilon \right) \\ &= |u_i u_j - u_i^* u_j^*| - \left(\frac{2u_i u_j}{u_n} \right) \epsilon - \left(\frac{u_i u_j}{u_n^2} \right) \epsilon^2 \end{aligned} \tag{3.21}$$

   where we have used the assumption $\mathbf{u}^*, \mathbf{u} > 0$.

2. $(i,j) \in \Omega$ and $i, j \in T_2$: In this case, since $u_i > u_i^*$ and $u_j > u_j^*$, one can write

$$\begin{aligned} |\hat{u}_i \hat{u}_j - u_i^* u_j^*| = \hat{u}_i \hat{u}_j - u_i^* u_j^* &= \left(u_i - \frac{u_i}{u_n}\epsilon \right) \left(u_j - \frac{u_j}{u_n}\epsilon \right) - u_i^* u_j^* \\ &= |u_i u_j - u_i^* u_j^*| - \left(\frac{2u_i u_j}{u_n} \right) \epsilon + \left(\frac{u_i u_j}{u_n^2} \right) \epsilon^2 \end{aligned} \tag{3.22}$$

   where we have used the assumption $\mathbf{u}^*, \mathbf{u} > 0$.

3. $(i,j) \in \Omega$, $i \in N$, and $j \in T_1 \cup T_2$: According to the definitions of T_1 and T_2, we have

$$\frac{u_i}{u_i^*} < \frac{u_n^*}{u_n}, \qquad \frac{u_i^*}{u_i} < \frac{u_n^*}{u_n} \tag{3.23}$$

   Now, if $j \in T_1$, one can write

$$\frac{u_i}{u_i^*} < \frac{u_j^*}{u_j} \implies u_i u_j < u_i^* u_j^* \tag{3.24}$$

which implies that

$$|\hat{u}_i\hat{u}_j - u_i^*u_j^*| = u_i^*u_j^* - \hat{u}_i\hat{u}_j = u_i^*u_j^* - u_i\left(u_j + \frac{u_j}{u_n}\epsilon\right) = |u_iu_j - u_i^*u_j^*| - \left(\frac{u_iu_j}{u_n}\right)\epsilon \quad (3.25)$$

Similarly, if $j \in T_2$, one can verify that

$$|\hat{u}_i\hat{u}_j - u_i^*u_j^*| = |u_iu_j - u_i^*u_j^*| - \left(\frac{u_iu_j}{u_n}\right)\epsilon \quad (3.26)$$

4. $(i,j) \in \Omega$, $i \in T_1$, and $j \in T_2$: In this case, note that

$$|\hat{u}_i\hat{u}_j - u_i^*u_j^*| = \left|\left(u_i + \frac{u_i}{u_n}\epsilon\right)\left(u_j - \frac{u_j}{u_n}\epsilon\right) - u_i^*u_j^*\right| \leq |u_iu_j - u_i^*u_j^*| + \left(\frac{u_iu_j}{u_n^2}\right)\epsilon^2 \quad (3.27)$$

The above analysis entails that—unless N and the subgraphs of $\mathcal{G}(\Omega)$ induced by the nodes in T_1 or T_2 are empty—$f'(\mathbf{u},\mathbf{d}) > 0$ and $f'(\mathbf{u},-\mathbf{d}) < 0$, implying that $\mathbf{u}$ cannot be D-stationary. On the other hand, these conditions enforce $\mathcal{G}(\Omega)$ to be bipartite, which is a contradiction. This completes the proof. □

Next, we show that $\mathbf{u}^* > 0$ is *almost* necessary to guarantee the absence of spurious local minima for (P1-Sym).

Proposition 2. *Assume that* $\mathbf{u}^* \geq 0$ *and that* $\mathbf{u}^* \neq 0$ *with* $u_i^* = 0$ *for some* i*. Then, upon choosing* $\Omega = \{1,\ldots,n\}^2\backslash\{(i,i)\}$, (P1-Sym) *has a spurious local minimum.*

The above corollary shows that if $\mathbf{u}^*$ is non-negative with at least one zero element, even in the almost perfect scenario where the set Ω includes all of the measurements except for one, it may not be free of spurious local minima. The next corollary shows that the assumption on the absence of bipartite components in $\mathcal{G}(\Omega)$ is also necessary for the uniqueness of the global solution.

Proposition 3. *Given any vector* $\mathbf{u}^* > 0$ *and set* Ω*, suppose that* $\mathcal{G}(\Omega)$ *has a bipartite component. Then, the global solution of* (P1-Sym) *is not unique.*

Proof. Without loss of generality, suppose that $\mathcal{G}(\Omega)$ is a connected bipartite graph. For any vector $\mathbf{u}^* > 0$, the solution $\mathbf{u} = \mathbf{u}^*$ is globally optimal for (P1-Sym). Suppose that the bipartite graph $\mathcal{G}(\Omega)$ partitions the entries of $\mathbf{u}$ into two sets V_1 and V_2 such that $u_n \in V_1$. Based on some simple algebra, one can easily verify that, for a sufficiently small $\epsilon > 0$, the solution

$$\hat{u}_i \leftarrow \begin{cases} u_i + \frac{u_i}{u_n}\epsilon & \text{if } i \in V_1 \\ u_i - \frac{u_i}{u_n+\epsilon}\epsilon & \text{if } i \in V_2 \end{cases} \quad (3.28)$$

is also globally optimal for (P1-Sym). □

Remark 5. *Suppose that* $\mathbf{u}^*$ *is a globally optimal solution of* (P1-Sym) *and that* $\mathcal{G}(G)$ *includes a bipartite component. Then, according to Proposition 3, the part of* $\mathbf{u}^*$ *whose elements correspond to the nodes in this bipartite component can be perturbed to attain another globally optimal solution, thereby resulting in the* ***non-uniqueness of the global solution****. On the other hand, the connectedness assumption is required to eliminate the undesirable stationary points on the boundary of the feasible region. Roughly speaking, the elements of the vector variable* $\mathbf{u}$ *corresponding to different disconnected components can behave independently from each other, giving rise to spurious D-stationary points in the problem. To elaborate, recall that* $\mathbf{u}[c]$ *is a sub-vector of* $\mathbf{u}$ *induced by the* c^{th} *component of* $\mathcal{G}(G)$*. Based on Lemma 12, the D-stationary points restricted to each disjoint component of* $\mathcal{G}(G)$ *are either strictly positive or equal to zero. Therefore, upon having two disconnected components* c_1 *and* c_2*, the points* $\mathbf{u}' = \begin{bmatrix}\mathbf{u}^*[c_1]^\top & 0\end{bmatrix}^\top$ *and* $\mathbf{u}'' = \begin{bmatrix}0 & \mathbf{u}^*[c_2]^\top\end{bmatrix}^\top$ *are indeed D-stationary points of* (SN-RPCA)*, thereby resulting in* ***spurious stationary points****.*

Asymmetric case: Next, we consider (3.14) in the noiseless scenario by analyzing its symmetrized counterpart (P1-Asym). Based on the construction of $\bar{\Omega}$, the corresponding sparsity graph $\mathcal{G}(\bar{\Omega})$ is bipartite. On the other hand, according to Proposition 3, the existence of a bipartite component in $\mathcal{G}(\bar{\Omega})$ makes a part of the solution *invariant to scaling*, which subsequently results in the non-uniqueness of the global minimum. The additional regularization term in (P1-Asym) is introduced to circumvent this issue by penalizing the difference in the norms of $\mathbf{u}$ and $\mathbf{v}$.

Theorem 10. *Suppose that* $\mathbf{w}^* > 0$ *and* $\mathcal{G}(\bar{\Omega})$ *is connected. Then, the following statements hold for* (P1-Asym)*:*

1. *The points* $\mathbf{w} = 0$ *and* $\mathbf{w}$ *with the properties* $\mathbf{w}\mathbf{w}^\top = \mathbf{w}^*\mathbf{w}^{*\top}$ *and* $\sum_{i=1}^{m} w_i^2 = \sum_{j=m+1}^{m+n} w_j^2$ *are the only D-min-stationary points;*

2. *The point* $\mathbf{w} = 0$ *is a local maximum;*

3. *In the positive orthant, the point* $\mathbf{w}$ *with the properties* $\mathbf{w}\mathbf{w}^\top = \mathbf{w}^*\mathbf{w}^{*\top}$ *and* $\sum_{i=1}^{m} w_i^2 = \sum_{j=m+1}^{m+n} w_j^2$ *is the only D-stationary point.*

Remark 6. *Notice that, unlike the symmetric case, Theorem 10 requires the connectedness of* $\mathcal{G}(\bar{\Omega})$*. This is due to the additional regularization term in* (AN-RPCA)*. In particular, similar arguments do not necessarily hold for the disjoint components of* $\mathcal{G}(\bar{\Omega})$ *because of the coupling nature of the regularization term.*

Probabilistic Guarantees

Next, we consider the random sampling regime. Similar to the previous subsection, we first focus on the symmetric case.

Symmetric case: Suppose that every element of the upper triangular part of the matrix $\mathbf{u}^*\mathbf{u}^{*\top}$ is measured independently with probability p. In other words, for every $(i,j) \in \{1,2,...,n\}^2$ and $i \leq j$, the probability of (i,j) belonging to Ω is equal to p.

Theorem 11. *Suppose that $n \geq 2$, $\mathbf{u}^* > 0$, and $p \geq \min\left\{1, \frac{(2\eta+2)\log n+2}{n-1}\right\}$ for some constant $\eta \geq 1$. Then, the following statements hold for* (SN-RPCA) *with probability of at least $1 - \frac{3}{2}n^{-\eta}$:*

1. *The points $\mathbf{u} = \mathbf{u}^*$ and $\mathbf{u} = 0$ are the only D-min-stationary points;*
2. *The point $\mathbf{u} = 0$ is a local maximum;*
3. *In the positive orthant, the point $\mathbf{u} = \mathbf{u}^*$ is the only D-stationary point.*

Before presenting the proof of Theorem 11, we note that the required lower bound on p is to guarantee that the random graph $\mathcal{G}(\Omega)$ is connected with high probability. This implies that Theorem 9 can be invoked to verify the statements of Theorem 11. It is worthwhile to mention that the classical results on *Erdös-Rényi* graphs characterize the *asymptotic* properties of $\mathcal{G}(\Omega)$ as n approaches infinity. In particular, it is shown by [74] that with the choice of $p = \frac{\log n+c}{n}$ for some $c > 0$, $\mathcal{G}(\Omega)$ becomes connected with probability of at least $\Omega(e^{-e^{-c}})$ as $n \to \infty$. In contrast, we introduce the following non-asymptotic result characterizing the probability that $\mathcal{G}(\Omega)$ is connected and non-bipartite for any finite $n \geq 2$, and subsequently use it to prove Theorem 11.

Lemma 13. *Given a constant $\eta \geq 1$, suppose that $p \geq \min\left\{1, \frac{(2\eta+2)\log n+2}{n-1}\right\}$ and $n \geq 2$. Then, $\mathcal{G}(\Omega)$ is connected and non-bipartite with probability of at least $1 - \frac{3}{2}n^{-\eta}$.*

Proof of Theorem 11: The proof immediately follows from Theorem 9 and Lemma 13. □

Similar to the deterministic case, we will show that both assumptions $\mathbf{u}^* > 0$ and $p \gtrsim \log n/n$ are *almost* necessary for the successful recovery of the global solution of (P1-Sym). In particular, it will be proven that relaxing $\mathbf{u}^* > 0$ to $\mathbf{u}^* \geq 0$ will result in an instance that possesses a spurious local solution with non-negligible probability. Furthermore, it will be shown that the choice $p \approx \log n/n$ is optimal—modulo $\log n$-factor—for the unique recovery of the global solution.

Proposition 4. *Assuming that $\mathbf{u}^* \geq 0$ with $u_i^* = 0$ for some $i \in \{1,\ldots,n\}$ and that $p < 1$,* (P1-Sym) *has a spurious local minimum with probability of at least $1 - p > 0$.*

Proof. Suppose that $\mathbf{u}^* \geq 0$ and there exists an index i such that $u_i^* = 0$. The proof of Proposition 2 can be used to show that excluding the measurement (i,i) gives rise to a spurious local minimum. This occurs with probability $1 - p$. The details are omitted due to their similarities to the proof of Proposition 2. □

Proposition 5. *Given any* $\mathbf{u}^* > 0$*, suppose that* $np \to 0$ *as* $n \to \infty$*. Then, the global solution of* (P1-Sym) *is not unique with probability approaching to one.*

Asymmetric case: Consider (3.14) under a random sampling regime, where each element of $\mathbf{u}^*\mathbf{v}^{*\top}$ is independently observed with probability p. Next, the analog of Theorem 11 for the asymmetric case is provided.

Theorem 12. *Suppose that* $n, m \geq 2$*,* $\mathbf{w}^* > 0$*, and* $p \geq \min\left\{1, \frac{(m+n)((1+\eta)\log(mn)+1)}{(m-1)(n-1)}\right\}$ *for some constant* $\eta \geq 1$*. Then, the following statements hold for* (P1-Asym) *with probability of at least* $1 - 2(mn)^{-\eta} - 4(mn)^{-2\eta}$*:*

1. *The points* $\mathbf{w} = 0$ *and* $\mathbf{w}$ *with the properties* $\mathbf{w}\mathbf{w}^\top = \mathbf{w}^*\mathbf{w}^{*\top}$ *and* $\sum_{i=1}^{m} w_i^2 = \sum_{j=m+1}^{m+n} w_j^2$ *are the only D-min-stationary points;*

2. *The point* $\mathbf{w} = 0$ *is a local maximum;*

3. *In the positive orthant, the point* $\mathbf{w}$ *with the properties* $\mathbf{w}\mathbf{w}^\top = \mathbf{w}^*\mathbf{w}^{*\top}$ *and* $\sum_{i=1}^{m} w_i^2 = \sum_{j=m+1}^{m+n} w_j^2$ *is the only D-stationary point.*

Before presenting the proof of Theorem 12, we note that $\mathcal{G}(\bar{\Omega})$ no longer corresponds to an Erdös-Rényi random graph due to its bipartite structure. Therefore, we present the analog of Lemma 13 for random bipartite graphs.

Lemma 14. *Given a constant* $\eta \geq 1$*, suppose that* $p \geq \min\left\{1, \frac{(m+n)((1+\eta)\log(mn)+1)}{(m-1)(n-1)}\right\}$ *and* $m, n \geq 2$*. Then,* $\mathcal{G}(\bar{\Omega})$ *is connected with probability of at least* $1 - 2(mn)^{-\eta} - 4(mn)^{-2\eta}$*.*

Proof of Theorem 12: The proof immediately follows from Theorem 10 and Lemma 14. □

Before proceeding, we note that, similar to the classical results on the *Erdös-Rényi* graphs, there are asymptotic results guaranteeing the connectedness of a random bipartite graph as a function of p. In particular, [220] shows that $\mathcal{G}(\bar{\Omega})$ is connected with probability approaching to 1 as $m + n \to \infty$, provided that $p \geq 3\left(1 + \frac{m}{n}\right)^{-1} \frac{(n+m)\log(n+m)}{nm}$. Lemma 14 offers another lower bound on p that matches this threshold (modulo a constant factor), while being non-asymptotic in nature. In particular, it characterizes the probability that the random bipartite graph is connected for *all* $m, n \geq 2$.

3.5 Extension to Noisy Positive RPCA

In this section, we will show that an additive sparse noise with arbitrary values does not drastically change the landscape of the RPCA. In other words, a limited number of grossly wrong measurements will not introduce any spurious local solution to the positive RPCA. The key idea is to prove that the direction of descent that was introduced in the previous

section is also valid when the measurements are not perfect, i.e., when they are subject to sparse noise. To this goal, consider the following problem in the symmetric case:

$$\min_{\mathbf{u}\geq 0} \quad \underbrace{\sum_{(i,j)\in\Omega} |u_i u_j - X_{ij}|}_{f(\mathbf{u})} \tag{3.29}$$

where

$$X = \mathbf{u}^* \mathbf{u}^{*\top} + S \tag{3.30}$$

is the matrix of true measurements perturbed with sparse noise. Similarly, consider the following problem for the asymmetric case:

$$\min_{\mathbf{u}\geq 0, \mathbf{v}\geq 0} \quad \sum_{(i,j)\in\Omega} |u_i v_j - X_{ij}| + \alpha \left| \sum_{i=1}^{m} u_i^2 - \sum_{j=1}^{n} v_j^2 \right| \tag{3.31}$$

where α is an arbitrary positive number. After symmetrization, (3.31) can be re-written as

$$\min_{\mathbf{w}\geq 0} \quad \underbrace{\sum_{(i,j)\in\bar{\Omega}} |w_i w_j - \bar{X}_{ij}| + \alpha \left| \sum_{i=1}^{m} w_i^2 - \sum_{j=m+1}^{m+n} w_j^2 \right|}_{f(\mathbf{w})} \tag{3.32}$$

where

$$\bar{X} = \mathbf{w}\mathbf{w}^\top + \bar{S} \tag{3.33}$$

for $\bar{X} \in \mathbb{R}^{(n+m)\times(n+m)}$ and

$$\bar{S} = \begin{bmatrix} 0 & S \\ S^\top & 0 \end{bmatrix} \tag{3.34}$$

Furthermore, define $\bar{B} = \{(i,j) : (i,j) \in \bar{\Omega}, \bar{S}_{ij} \neq 0\}$ and $\bar{G} = \{(i,j) : (i,j) \in \bar{\Omega}, \bar{S}_{ij} = 0\}$ as the sets of bad and good measurements for the symmetrized problem, respectively. In this work, we do not impose any assumption on the maximum value of the nonzero elements of S. However, without loss of generality, one may assume that $\mathbf{u}^*\mathbf{u}^{*\top} + S > 0$ and $\mathbf{w}^*\mathbf{w}^{*\top} + \bar{S} > 0$; otherwise, the non-positive elements can be discarded due to the assumptions $\mathbf{u}^* > 0$ and $(\mathbf{u}^*, \mathbf{v}^*) > 0$. In fact, we impose a slightly more stronger condition in this work.

Assumption 1. *There exists a constant $c \in (0, 1]$ such that $S_{ij} + u_i^* u_j^* > c u_{\min}^{*2}$ and $\bar{S}_{ij} + w_i^* w_j^* > c w_{\min}^{*2}$ for (3.29) and (3.32), respectively.*

Identifiability

Intuitively, the non-negative RPCA under the unknown-but-sparse noise is more challenging to solve than its noiseless counterpart. In particular, one may consider (3.29) as a

variant of (P1-Sym) discussed in the previous section, where the locations of the bad measurements are unknown; if these locations were known, they could have been discarded to reduce the problem to (P1-Sym). If the measurements are subject to unknown noise, one of the main issues arises from the identifiability of the solution. To further elaborate, we will offer an example below.

Example 2. *Suppose that* $X(\epsilon) = (e_1 + \mathbf{1}\epsilon)(e_1 + \mathbf{1}\epsilon)^\top$*, where* e_1 *is the first unit vector and* $\mathbf{1}$ *is a vector of ones. Assuming that* $\Omega = \{1, ..., n\}^2$*, one can decompose* $X(\epsilon)$ *in two forms*

$$X(\epsilon) = \underbrace{(e_1 + \mathbf{1}\epsilon)(e_1 + \mathbf{1}\epsilon)^\top}_{\mathbf{u}_1^* \mathbf{u}_1^{*\top}} + \underbrace{0}_{S_1} \tag{3.35a}$$

$$X(\epsilon) = \underbrace{\mathbf{1}\mathbf{1}^\top \epsilon^2}_{\mathbf{u}_2^* \mathbf{u}_2^{*\top}} + \underbrace{e_1 e_1^\top + \mathbf{1} e_1^\top \epsilon + e_1 \mathbf{1}^\top \epsilon}_{S_2} \tag{3.35b}$$

For every $\epsilon > 0$*, both* S_1 *and* S_2 *can be considered as sparse matrices since the number of nonzero elements in each of these matrices is at most on the order of* $O(n)$*. However, unless more restrictions on the number of nonzero elements at each row or column of* S *are imposed, it is impossible to distinguish between these two cases. This implies that the solution is not identifiable.*

In order to ensure that the solution is identifiable in the symmetric case, we assume that $\Delta(\mathcal{G}(B)) \leq \eta \cdot \delta(\mathcal{G}(G))$ for some constant $\eta \leq 1$ to be defined later. Roughly speaking, this implies that at each row of the measurement matrix, the number of good measurements should be at least as large as the number of bad ones. Similar to the work by [101, 100], we consider the regularized version of the problem, as in

$$\min_{\mathbf{u} \geq 0} \quad \underbrace{\sum_{(i,j) \in \Omega} |u_i u_j - X_{ij}| + R(\mathbf{u})}_{f_{\text{reg}}(\mathbf{u})} \tag{P2-Sym}$$

where $R(\mathbf{u})$ is a regularizer defined as

$$R(\mathbf{u}) = \lambda \sum_{i=1}^{n} (u_i - \beta)^4 \, \mathbb{I}_{u_i \geq \beta} \tag{3.36}$$

for some fixed parameters λ and β to be specified later. Similarly, one can define an analogous regularization for (3.32) as

$$\min_{\mathbf{w} \geq 0} \quad \underbrace{\sum_{(i,j) \in \bar{\Omega}} |w_i w_j - \bar{X}_{ij}| + \alpha \left| \sum_{i=1}^{m} w_i^2 - \sum_{j=m+1}^{m+n} w_j^2 \right| + R(\mathbf{w})}_{f_{\text{reg}}(\mathbf{w})} \tag{P2-Asym}$$

with

$$R(\mathbf{u}) = \lambda \sum_{i=1}^{m+n} (w_i - \beta)^4 \, \mathbb{I}_{w_i \geq \beta} \tag{3.37}$$

for some fixed parameters λ and β to be specified later. Note that the defined regularization function is convex in its domain. In particular, it eliminates the candidate solutions that are far from the true solution. Without loss of generality and to streamline the presentation, it is assumed that $u^*_{\max} = w^*_{\max} = 1$ in the sequel.

Lemma 15. *Consider the parameter c defined in Assumption 1. The following statements hold:*

- *By choosing $\beta = 1$ and $\lambda = n/2$, any D-stationary point $\mathbf{u} > 0$ of* (P2-Sym) *satisfies the inequalities $(c/2)u^{*2}_{\min} \leq u_{\min} \leq u_{\max} \leq 2$.*
- *By choosing $\beta = 1$ and $\lambda = (m+n)/2$, any D-stationary point $\mathbf{w} > 0$ of* (P2-Asym) *satisfies the inequalities $(c/2)w^{*2}_{\min} \leq w_{\min} \leq w_{\max} \leq 2$.*

Deterministic Guarantees

In what follows, the deterministic conditions under which (P2-Sym) and (P2-Asym) have benign landscape will be investigated. The results of this subsection will be the building blocks for the derivation of the main theorems for both symmetric and asymmetric positive RPCA under the random sampling and noise regime. Note that the analysis of the landscape will be more involved in this case since the effect of the regularizer should be taken into account.

Symmetric case: Recall that, for the sparsity graph $\mathcal{G}(\Omega)$, $\Delta(\mathcal{G}(\Omega))$ and $\delta(\mathcal{G}(\Omega))$ correspond to its maximum and minimum degrees, respectively.

Theorem 13. *Suppose that*

i. $\mathbf{u}^* > 0$*;*

ii. $\delta(\mathcal{G}(G)) > (48/c^2)\kappa(\mathbf{u}^*)^4\Delta(\mathcal{G}(B))$*;*

iii. $\mathcal{G}(\Omega)$ *has no bipartite component.*

Then, with the choice of $\beta = 1$ and $\lambda = n/2$ for the parameters of the regularization function $R(\mathbf{u})$, the following statements hold for (P2-Sym)*:*

1. *It does not have any spurious local minimum;*
2. *The point $\mathbf{u} = \mathbf{u}^*$ is the unique global minimum;*
3. *In the positive orthant, the point $\mathbf{u} = \mathbf{u}^*$ is the only D-stationary point.*

Additionally, if $\mathcal{G}(\Omega)$ is connected, the following statements hold for (P2-Sym)*:*

4. *The points $\mathbf{u} = \mathbf{u}^*$ and $\mathbf{u} = 0$ are the only D-min-stationary points;*
5. *The point $\mathbf{u} = 0$ is a local maximum.*

Asymmetric case: Theorem 13 has the following natural extension to asymmetric problems.

Theorem 14. *Suppose that*

i. $\mathbf{w}^* > 0$*;*

ii. $\delta(\mathcal{G}(\bar{G})) > (48/c^2)\kappa(\mathbf{w}^*)^4\Delta(\mathcal{G}(\bar{B}))$*;*

iii. $\mathcal{G}(\bar{G})$ *is connected.*

Then, with the choice of $\beta = 1$ and $\lambda = (m+n)/2$ for the parameters of the regularization function $R(\mathbf{w})$, the following statements hold for (P2-Asym)*:*

1. *The points $\mathbf{w} = 0$ and $\mathbf{w}$ with the properties $\mathbf{w}\mathbf{w}^\top = \mathbf{w}^*\mathbf{w}^{*\top}$ and $\sum_{i=1}^{m} w_i^2 = \sum_{j=m+1}^{m+n} w_j^2$ are the only D-min-stationary points;*
2. *The point $\mathbf{w} = 0$ is a local maximum;*
3. *In the positive orthant, the point $\mathbf{w}$ with the properties $\mathbf{w}\mathbf{w}^\top = \mathbf{w}^*\mathbf{w}^{*\top}$ and $\sum_{i=1}^{m} w_i^2 = \sum_{j=m+1}^{m+n} w_j^2$ is the only D-stationary point.*

Proof. The proof is omitted due to its similarity to that of Theorem 13. □

Probabilistic Guarantees

As an extension to our previous results, we analyze the landscape of the noisy nonnegative RPCA with randomness both in the location of the samples and in the structure of the noise matrix. Suppose that for the symmetric case, with probability d, each element of the upper triangular part of X is independently corrupted with an arbitrary noise value. In other words, for every (i, j) with $i \leq j$, one can write

$$X_{ij} = \begin{cases} u_i^* u_j^* & \text{with probability } 1-d \\ \text{arbitrary} & \text{with probability } d \end{cases} \tag{3.38}$$

Furthermore, similar to the preceding section, suppose that every element of the upper triangular part of $X = \mathbf{u}^*\mathbf{u}^{*\top} + S$ is independently measured with probability p. The randomness in the location of the measurements and noise is naturally extended to the asymmetric case by considering the symmetrized $\bar{X}$ and $\bar{S}$ defined in (3.33) and (3.34), respectively.

Symmetric case: First, the main result in the symmetric case is presented below.

Theorem 15. *Suppose that*

i. $n \geq 2$,

ii. $\mathbf{u}^* > 0$,

iii. $d < \frac{1}{(144/c^2)k(\mathbf{u}^*)^4+1}$,

iv. $p > \frac{(1740/c^2)\kappa(\mathbf{u}^*)^4(1+\eta)\log n}{n}$,

for some $\eta > 0$. *Then, with the choice of* $\beta = 1$ *and* $\lambda = n/2$ *for the parameters of the regularization function* $R(\mathbf{u})$, *the following statements hold for* (P2-Sym) *with probability of at least* $1 - 3n^{-\eta}$*:*

1. *The points* $\mathbf{u} = \mathbf{u}^*$ *and* $\mathbf{u} = 0$ *are the only D-min-stationary points;*

2. *The point* $\mathbf{u} = 0$ *is a local maximum;*

3. *In the positive orthant, the point* $\mathbf{u} = \mathbf{u}^*$ *is the only D-stationary point.*

To prove Theorem 15, first we present the following lemma on the concentration of the minimum and maximum degrees of random graphs.

Lemma 16. *Consider a random graph* $\mathcal{G}(n,p)$. *Given a constant* $\eta > 0$, *the inequality:*

$$\mathbb{P}\left(\Delta(\mathcal{G}(n,p)) \geq \max\left\{\frac{3np}{2}, 18(1+\eta)\log n\right\}\right) \leq n^{-\eta} \tag{3.39}$$

holds for every $0 < p \leq 1$. *Furthermore, we have*

$$\mathbb{P}\left(\delta(\mathcal{G}(n,p)) \leq \frac{np}{2}\right) \leq n^{-\eta} \tag{3.40}$$

provided that $p \geq \frac{12(1+\eta)\log n}{n}$.

Remark 7. *Note that since the degree of each node in* $\mathcal{G}(n,p)$ *is concentrated around* np *with high probability, one may speculate that* $\Delta(\mathcal{G}(n,p))$ *and* $\delta(\mathcal{G}(n,p))$ *should also concentrate around* np *for all values of* p *and hence the inclusion of* $18(1+\eta)\log n$ *in* (3.39) *may seem redundant. Surprisingly, this is not the case in general. In fact, it can be shown that if* $p = 1/n$ *(and hence* $np = 1$*), there exists a node whose degree is lower bounded by* $\log n/\log\log n$ *with high probability. This explains the reasoning behind the inclusion of* $18(1+\eta)\log n$ *in the lemma.*

Proof of Theorem 15: In light of Lemma 13, the bounds on p and d guarantee that $\mathcal{G}(G)$ is connected and non-bipartite with probability of at least $1 - \frac{3}{2}n^{-430(1+\eta)}$. Therefore, the proof is completed by invoking Theorem 13, provided that the second condition of Theorem 13 holds. Define the events $\mathcal{E}_1 = \left\{\Delta(\mathcal{G}(B)) \leq \max\left\{\frac{3npd}{2}, 18(1+\eta)\log n\right\}\right\}$ and

$\mathcal{E}_2 = \left\{\delta(\mathcal{G}(G)) \geq \frac{np(1-d)}{2}\right\}$. Observe that Lemma 16 together with the bounds on p and d results in the inequalities

$$\mathbb{P}(\mathcal{E}_1) \geq 1 - n^{-\eta} \tag{3.41a}$$

$$\mathbb{P}(\mathcal{E}_2) \geq 1 - n^{-144\eta} \tag{3.41b}$$

This in turn implies that the events $\mathcal{E}_1$ and $\mathcal{E}_2$ occur with probability of at least $1 - n^{-\eta} - n^{-144\eta}$. Conditioned on these events, it suffices to show that

$$\frac{np(1-d)}{2} > \frac{48}{c^2}\kappa(\mathbf{u}^*)^4 \max\left\{\frac{3npd}{2}, 18(1+\eta)\log n\right\} \tag{3.42}$$

in order to certify the validity of the second condition of Theorem 13. It can be easily verified that the assumed upper and lower bounds on p and d guarantee the validity of (3.42). Therefore, a simple union bound and the fact that $n^{-\eta} > \frac{3}{2}n^{-430(1+\eta)}$ imply that the conditions of Theorem 13 are satisfied with probability of at least $1 - 3n^{-\eta}$. □

A number of interesting corollaries can be derived based on Theorem 15.

Corollary 2. *Suppose that p is a positive number independent of n and $d \lesssim \log n/n$. Then, under an appropriate choice of parameters for the regularization function, the statements of Theorem 15 hold with overwhelming probability, provided that $\kappa(\mathbf{u}^*) \lesssim (n/\log n)^{1/4}$.*

Corollary 2 implies that, roughly speaking, if the total number of measurements is sufficiently large (i.e., on the order of n^2), then up to factor of $n \log n$ bad measurements with arbitrary magnitudes will not introduce any spurious local solution to the problem. Under such circumstances, the required upper bound on the ratio between the maximum and the minimum entries of $\mathbf{u}^*$ will be more relaxed as the dimension of the problem grows.

Corollary 3. *Suppose that p is a positive number independent of n and that $d \lesssim n^{\epsilon-1}$ for some $\epsilon \in [0, 1)$. Then, under an appropriate choice of parameters for the regularization function, the statements of Theorem 15 hold with overwhelming probability, provided that $\kappa(\mathbf{u}^*) \lesssim n^{(1-\epsilon)/4}$.*

Corollary 3 describes an interesting trade-off between the sparsity level of the noise and the maximum allowable variation in the entries of $\mathbf{u}^*$; roughly speaking, as $\kappa(\mathbf{u}^*)$ decreases, a larger number of noisy elements can be added to the problem without creating any spurious local minimum. The next corollary shows that a constant fraction of the measurements can be grossly corrupted without affecting the landscape of the problem, provided that $\kappa(\mathbf{u}^*)$ is uniformly bounded from above.

Corollary 4. *Suppose that p and d are positive numbers independent of n and that $d < \frac{1}{(144/c^2)+1}$. Then, under an appropriate choice of parameters for the regularization function, the statements of Theorem 15 hold with overwhelming probability, provided that $\kappa(\mathbf{u}^*) \leq \left(\frac{1-d}{(144/c^2)d}\right)^{1/4}$.*

Asymmetric case: The aforementioned results on the symmetric positive RPCA under random sampling and noise will be generalized to the asymmetric case below.

Theorem 16. *Define $r = m/n$ and suppose that*

i. $n \geq m \geq 2$,

ii. $\mathbf{w}^* > 0$,

iii. $d < \frac{r}{(144/c^2)\kappa(\mathbf{w}^*)^4 + r}$,

iv. $p > \frac{(1740/c^2)\kappa(\mathbf{w}^*)^4(1+\eta)n\log n}{m^2}$,

for some $\eta > 0$. Then, with the choice of $\beta = 1$ and $\lambda = (m+n)/2$ for the parameters of the regularization function $R(\mathbf{u})$, the following statements hold for (P2-Sym) *with probability of at least $1 - 10n^{-\eta}$:*

1. *The points $\mathbf{w} = 0$ and $\mathbf{w}$ with the properties $\mathbf{w}\mathbf{w}^\top = \mathbf{w}^*\mathbf{w}^{*\top}$ and $\sum_{i=1}^{m} w_i^2 = \sum_{j=m+1}^{m+n} w_j^2$ are the only D-min-stationary points;*

2. *The point $\mathbf{w} = 0$ is a local maximum;*

3. *In the positive orthant, the point $\mathbf{w}$ with the properties $\mathbf{w}\mathbf{w}^\top = \mathbf{w}^*\mathbf{w}^{*\top}$ and $\sum_{i=1}^{m} w_i^2 = \sum_{j=m+1}^{m+n} w_j^2$ is the only D-stationary point.*

To prove Theorem 16, we derive a concentration bound on the minimum and maximum degree of the random bipartite graphs. Define $\mathcal{G}(m,n,p)$ as a bipartite graph with the vertex partitions $V_u = \{1,\cdots,m\}$ and $V_v = \{m+1,\cdots,m+n\}$ where each edge is independently included in the graph with probability p.

Lemma 17. *Consider a random bipartite graph $\mathcal{G}(m,n,p)$. Given a constant $\eta > 0$, the inequality*

$$\mathbb{P}\left(\Delta(\mathcal{G}(m,n,p)) \geq \max\left\{\frac{3np}{2}, \frac{18(1+\eta)n\log n}{m}\right\}\right) \leq 2n^{-\eta} \tag{3.43}$$

holds for every $0 < p \leq 1$. Furthermore, we have

$$\mathbb{P}\left(\delta(\mathcal{G}(m,n,p)) \leq \frac{mp}{2}\right) \leq 2n^{-\eta} \tag{3.44}$$

provided that $p \geq 12(1+\eta)\log n/m$.

Proof of Theorem 16: The bounds on p and d indeed guarantee that $\mathcal{G}(\bar{G})$ is connected with overwhelming probability. Based on this fact, the result of Lemma 17 and the proof of Theorem 15 can be combined to arrive at this theorem. The details are omitted for brevity. □

Remark 8. *The presented probability guarantees for RPCA share some similarities with those derived for noisy matrix completion in [100, 101]. In particular, according to Theorems 15 and 16 and similar to the results of [100, 101], the probability of having a spurious local solution decreases polynomially with respect to the dimension of the problem. Furthermore, similar to our work, the required lower bound on the sampling probability p in [100, 101] scales polynomially with respect to the condition number of the true solution. Finally, for non-symmetric noisy matrix completion problem, [100] shows that the required lower bound on p scales as* $\frac{\log n}{m}$*. Comparing this dependency with the one introduced in Theorem 16, it can be inferred that our proposed lower bound is higher by a factor of* $\frac{n}{m}$*; this is not surprising considering the fundamentally different natures of these problems.*

3.6 Global Convergence of Local Search Algorithms

So far, it has been shown that the positive RPCA is free of spurious local minima. Furthermore, it has been proven that the global solution is the only D-stationary point in the positive orthant. The question of interest in this section is: How could this unique D-stationary point be obtained? Before answering this question, we will take a detour and revisit the notion of stationarity for smooth optimization problems. Recall that $\bar{\mathbf{x}}$ is a stationary point of a differentiable function $f(\mathbf{x})$ if and only if $\nabla f(\mathbf{x}) = 0$ and, under some mild conditions, basic local search algorithms will converge to a stationary point. Therefore, the uniqueness of the stationary point for a smooth optimization problem immediately implies the convergence to global solution. Extra caution should be taken when dealing with non-smooth optimization. In particular, the convergence of classical local search algorithms may fail to hold since the gradient and/or Hessian of the function may not exist at every iteration. To deal with this issue, different local search algorithms have been introduced to guarantee convergence to generalized notions of stationary points for non-smooth optimization, such as directional-stationary (which is used in this chapter) or Clarke-stationary (to be defined next).

For a non-smooth and locally Lipschitz function $h(\mathbf{x})$ over the convex set $\mathcal{X}$, define the Clarke generalized directional derivative at the point $\bar{\mathbf{x}}$ in the feasible direction $\mathbf{d}$ as

$$h^{\circ}(\mathbf{x}, \mathbf{d}) := \limsup_{\substack{\mathbf{y} \to \mathbf{x} \\ t \downarrow 0}} \frac{h(\mathbf{y} + t\mathbf{d}) - h(\mathbf{y})}{t} \tag{3.45}$$

Note the difference between the ordinary directional derivative $h'(\mathbf{x}, \mathbf{d})$ and its Clarke generalized counterpart: in the latter, the limit is taken with respect to a *variable* vector $\mathbf{y}$ that approaches $\bar{\mathbf{x}}$, rather than taking the limit exactly at $\bar{\mathbf{x}}$. The Clarke differential of $h(\mathbf{x})$ at $\bar{\mathbf{x}}$ is defined as the following set ([54]):

$$\partial_C h(\bar{\mathbf{x}}) := \{\psi | h^{\circ}(\mathbf{x}, \mathbf{d}) \geq \langle \psi, \mathbf{d} \rangle, \forall \mathbf{d} \in \mathbb{R}^n \text{ such that } \mathbf{x} + \mathbf{d} \in \mathcal{X}\} \tag{3.46}$$

where $\mathcal{X}$ is the feasible set of the problem. A point $\bar{\mathbf{x}}$ is Clarke-stationary (or C-stationary) if $0 \in \partial_C(\bar{\mathbf{x}})$, or equivalently, $h^\circ(\bar{\mathbf{x}}, \mathbf{d}) \geq 0$ for every feasible direction $\mathbf{d}$. It is well known that C-stationary is a weaker condition than the D-min-stationarity. In particular, every D-min-stationary point is C-stationary but not all C-stationary points are D-min-stationary.

On the other hand, although some local search algorithms converge to D-min-stationary points for problems with special structures ([60]), the most well-known numerical algorithms for non-smooth optimization—such as gradient sampling, sequential quadratic programming, and exact penalty algorithms—can only guarantee the C-stationarity of the obtained solutions ([42, 61, 80]). Therefore, it remains to study whether the global solution of the positive RPCA is the only C-stationary point. To answer this question, we need the following two lemmas.

Lemma 18. *The following statements hold:*

- *If $h : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ are continuously differentiable at $\bar{\mathbf{x}} \in \mathcal{X}$, then $(h + g)^\circ(\bar{\mathbf{x}}, \mathbf{d}) = h^\circ(\bar{\mathbf{x}}, \mathbf{d}) + g^\circ(\bar{\mathbf{x}}, \mathbf{d})$ for every feasible direction $\mathbf{d}$.*
- *If $h : \mathcal{X} \to \mathbb{R}$ is continuously differentiable at $\bar{\mathbf{x}} \in \mathcal{X}$, then $h^\circ(\bar{\mathbf{x}}, \mathbf{d}) = h'(\bar{\mathbf{x}}, \mathbf{d})$ for every feasible direction $\mathbf{d}$.*

Proof. Refer to the textbook by [54]. □

Lemma 19. *Let $h_1(\mathbf{x}), h_1(\mathbf{x}), ..., h_m(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$ be continuous and locally Lipschitz functions at $\bar{\mathbf{x}} \in \mathcal{X}$. Define*

$$h(\mathbf{x}) = \max_{1 \leq i \leq m} h_i(\mathbf{x}) \tag{3.47}$$

and let $I(\bar{\mathbf{x}})$ be the set of indices i such that $h(\bar{\mathbf{x}}) = h_i(\bar{\mathbf{x}})$. Then,

$$h^\circ(\bar{\mathbf{x}}, \mathbf{d}) \leq \max_{i \in I(\bar{\mathbf{x}})} h_i^\circ(\bar{\mathbf{x}}, \mathbf{d}) \tag{3.48}$$

for every feasible direction $\mathbf{d}$.

Proof. Consider a feasible point $\mathbf{y} \in \mathcal{B}(\bar{\mathbf{x}}, \epsilon) \cap \mathcal{X}$, where $\mathcal{B}(\bar{\mathbf{x}}, \epsilon)$ is the Euclidean ball with the center $\bar{\mathbf{x}}$ and radius ϵ. First, we prove that $I(\mathbf{y}) \subseteq I(\bar{\mathbf{x}})$ for sufficiently small $\epsilon > 0$. Notice that $h_i(\bar{\mathbf{x}}) < h_j(\bar{\mathbf{x}})$ for every $i \in I(\bar{\mathbf{x}})$ and $j \in \{1, ..., m\} \backslash I(\bar{\mathbf{x}})$. Therefore, due to the continuity of $h_i(\cdot)$ for every $i \in \{1, ..., m\}$, it follows that there exists $\bar{\epsilon} > 0$ such that $h_i(\mathbf{y}) < h_j(\mathbf{y})$ for every $\mathbf{y} \in \mathcal{B}(\bar{\mathbf{x}}, \epsilon) \cap \mathcal{X}$ with $0 < \epsilon < \bar{\epsilon}$. This implies that $I(\mathbf{y} + t\mathbf{d}) \subseteq I(\mathbf{y}) \subseteq I(\bar{\mathbf{x}})$ for every $\mathbf{y} \in \mathcal{B}(\bar{\mathbf{x}}, \epsilon) \cap \mathcal{X}$ and every feasible direction $\mathbf{d}$ with sufficiently small $\epsilon > 0$ and $t > 0$. Now, note that

$$h(\mathbf{y} + t\mathbf{d}) - h(\mathbf{y}) = \max_{i \in I(\mathbf{y} + t\mathbf{d})} h_i(\mathbf{y} + t\mathbf{d}) - h_i(\mathbf{y}) \leq \max_{i \in I(\bar{\mathbf{x}})} h_i(\mathbf{y} + t\mathbf{d}) - h_i(\mathbf{y}) \tag{3.49}$$

This implies that

$$h^{\circ}(\bar{\mathbf{x}},\mathbf{d})=\limsup_{\substack{\mathbf{y}\to\mathbf{x}\\ t\downarrow 0}}\frac{h(\mathbf{y}+t\mathbf{d})-h(\mathbf{y})}{t}\leq\max_{i\in I(\bar{\mathbf{x}})}\left\{\limsup_{\substack{\mathbf{y}\to\mathbf{x}\\ t\downarrow 0}}\frac{h_i(\mathbf{y}+t\mathbf{d})-h_i(\mathbf{y})}{t}\right\}=\max_{i\in I(\bar{\mathbf{x}})}h_i^{\circ}(\bar{\mathbf{x}},\mathbf{d}) \tag{3.50}$$

This completes the proof. □

Based on the above lemmas, we develop the following theorem.

Theorem 17. *Under the conditions of Theorem 13 and assuming that $\mathcal{G}(\Omega)$ is connected, the global solution and the origin are the only C-stationary points of the symmetric positive RPCA. A similar result holds for the asymmetric positive RPCA.*

Proof. Without loss of generality, we only consider the symmetric case. At a given point $\mathbf{u}$, the function $f(\mathbf{u})$ is locally Lipschitz and can be written as

$$f(\mathbf{u})=\sum_{(i,j)\in\Omega}\max\{u_iu_j-X_{ij},-u_iu_j+X_{ij}\}=\max_{\sigma\in\mathcal{M}}f_{\sigma}(\mathbf{u}) \tag{3.51}$$

where $\mathcal{M}$ is the class of functions from Ω to $\{-1,+1\}$ and $f_{\sigma}(\mathbf{u})$ is defined as

$$f_{\sigma}(\mathbf{u})=\sum_{(i,j)\in\Omega}\sigma(i,j)(u_iu_j-X_{ij}). \tag{3.52}$$

Hence,

$$f_{\text{reg}}(\mathbf{u})=R(\mathbf{u})+\max_{\sigma\in\mathcal{M}}f_{\sigma}(\mathbf{u}) \tag{3.53}$$

Notice that each function $f_{\sigma}(\mathbf{u})$ is differentiable and locally Lipschitz for every $\sigma\in\mathcal{M}$. By contradiction, suppose that there exists $\mathbf{u}\geq 0$ such that $\mathbf{u}\notin\{\mathbf{u}^*,0\}$ and $0\in\partial_C f_{\text{reg}}(\mathbf{u})$. Furthermore, define $I(\mathbf{u})$ as the set of all functions $\sigma\in\mathcal{M}$ for which $f_{\sigma}(\mathbf{u})=f(\mathbf{u})$. Using the proof technique developed in Theorem 13, one can easily verify that there exists a feasible direction $\mathbf{d}$ such that $f'_{\sigma}(\mathbf{u},\mathbf{d})+R'(\mathbf{u},\mathbf{d})<0$ for every $\sigma\in I(\mathbf{u})$. By invoking Lemma 18 for every $\sigma\in I(\mathbf{u})$, it can be concluded that $f^{\circ}_{\sigma}(\mathbf{u},\mathbf{d})+R^{\circ}(\mathbf{u},\mathbf{d})<0$. This, together with Lemma 19, certifies that $f^{\circ}_{\text{reg}}(\mathbf{u},\mathbf{d})<0$, hence contradicting the assumption $0\in\partial_C f_{\text{reg}}(\mathbf{u})$. □

3.7 Numerical Results

In this section, we demonstrate the efficacy of the developed results in different experiments. To this goal, first we briefly introduce the recently developed sub-gradient method [163] that is specifically tailored to non-smooth and non-convex problems, such

as those considered in this chapter. The main advantage of the sub-gradient algorithm compared to other state-of-the-art methods is its extremely simple implementation; we present a sketch of the algorithm for solving the non-symmetric positive RPCA in Algorithm 2[3] (the symmetric version can be solved using a similar algorithm with slight modifications).

Algorithm 2 Sub-gradient algorithm

1: **Initialization:** Strictly positive initial point $\mathbf{w}_0^\top = \begin{bmatrix}\mathbf{u}_0^\top & \mathbf{v}_0^\top\end{bmatrix}^\top$ and step size μ_0
2: **for** $k = 0, 1, \dots$ **do**
3: set $\mathbf{d}_k$ as a sub-gradient of $f_{\text{reg}}(\mathbf{u}_0, \mathbf{v}_0)$ defined in (AN-RPCA)
4: set μ_k according to a geometrically diminishing rule such that $\mathbf{w}_k - \mu_k \mathbf{d}_k$ is strictly positive
5: set $\mathbf{w}_{k+1} = \mathbf{w}_k - \mu_k \mathbf{d}_k$
6: **end for**

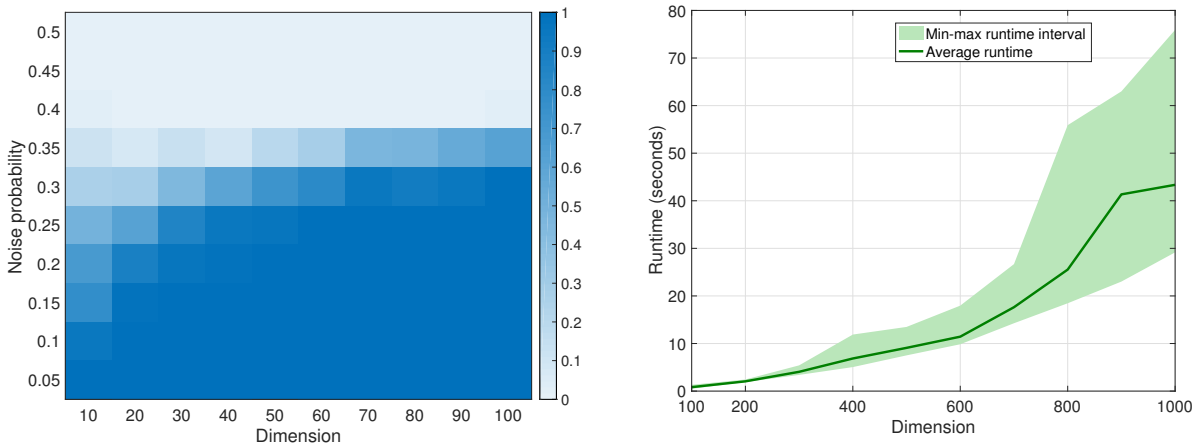
It has been shown in [163] that, under certain conditions on the initial point $\mathbf{w}_0$, the initial step size μ_0, and the update rule for μ_k, the iterates $\mathbf{w}_0, \mathbf{w}_1, \dots$ converge to the globally optimal solution at *linear* rate, provided that $\mathbf{w}_0$ is sufficiently close to the optimal solution. The closeness of $\mathbf{w}_0$ to $\mathbf{w}^*$ is required partly to avoid becoming stuck at a spurious local minima. This requirement can be relaxed for the positive RPCA due to the absence of undesired spurious local solutions, as proven in this chapter. It is also worthwhile to mention that, even though we use the sub-gradient algorithm to solve the positive RPCA, it will be shown in Section 3.6 that the results of this chapter guarantee that a large class of local-search algorithms converge to the globally optimal solution of (SN-RPCA) or (AN-RPCA).

All of the following simulations are run on a laptop computer with an Intel Core i7 quad-core 2.50 GHz CPU and 16GB RAM. The reported results are for a serial implementation in MATLAB R2017b.

Exact Recovery:

To demonstrate the strength of our results, we consider thousands of randomly generated instances of the positive rank-1 RPCA with different sizes and noise levels. In particular, the dimension of the instances ranges from 10 to 100. For each instance, the elements of $\mathbf{u}^*$ are uniformly chosen from the interval $[0, 2]$. Note that $\mathbf{u}^*$ will be strictly positive with probability one. Furthermore, each element of the upper triangular part of the symmetric noise matrix S is set to 2 with probability d and 0 with probability $1 - d$. Figure 3.7.1a shows the performance of randomly initialized sub-gradient method for the symmetric positive rank-1 RPCA. We declare that a solution is recovered exactly if $\|\mathbf{u}\mathbf{u}^\top - \mathbf{u}^*\mathbf{u}^{*\top}\|_F / \|\mathbf{u}^*\mathbf{u}^{*\top}\|_F \leq 10^{-4}$. For each dimension and noise probability, we consider 100 randomly generated instances of

[3] We present is a slightly modified version of the sub-gradient algorithm in [163] to ensure the positivity of the iterates.



(a) The performance of the randomly initialized sub-gradient method for (SN-RPCA). The intensity of the color is proportional to the exact recovery rate of the true solution (darker blue implies higher recovery rate).

(b) The runtime of the sub-gradient method for (SN-RPCA). For each dimension, it shows the average runtime and its min-max interval over 100 independent trials.

Figure 3.7.1: The performance of the sub-gradient method for RPCA.

the problem and demonstrate its exact recovery rate. The heatmap shows the exact recovery rate of the sub-gradient method, when directly applied to (SN-RPCA). It can be observed that the algorithm has recovered the globally optimal solution even when 35% of the entries in the data matrix were severely corrupted with the noise. In contrast, even a highly sparse additive noise in the data matrix prevents the sub-gradient method from recovering the true solution, when applied to the smooth problem (3.6). Figure 3.7.1b shows the graceful scalability of the sub-gradient algorithm when applied to (SN-RPCA). It can be seen that the algorithm is highly efficient. In particular, its average runtime varies from 0.88 seconds for $n = 100$ to 43.20 seconds for $n = 1000$.

The Emergence of Local Solutions

Recall that $\mathbf{u}^*$ and $\mathbf{v}^*$ are both assumed to be strictly positive. In what follows, we will illustrate that relaxing these conditions to non-negativity gives rise to spurious local solutions. Consider an instance of the symmetric non-negative rank-1 RPCA with the parameters

$$\mathbf{u}^* = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^\top, \qquad S = 0, \qquad \Omega = \{1,2,3\}^2 \backslash \{(3,3)\} \tag{3.54}$$

Notice that $\mathbf{u}^*$ consists of two strictly positive and one zero entries. Furthermore, this is a noiseless scenario where Ω consists of all possible measurements except for one. To examine

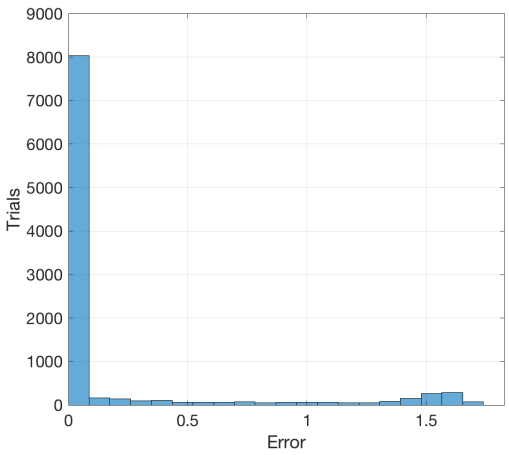


Figure 3.7.2: The distance between the recovered and true solutions for RPCA.

the existence of spurious local solutions in this example, 10000 randomly initialized trials of the sub-gradient method is ran and the normalized distances between the obtained and true solutions are displayed in Figure 3.7.2. Based on this histogram, about 20% of the trials converge to spurious local solutions, implying that they are ubiquitous in this instance. This experiment shows why the positivity of the true solution is crucial and cannot be relaxed. We formalized and proved this statement in Section 3.4.

Moving Object Detection

In video processing, one of the most important problems is to detect anomaly or moving objects in different frames of a video. In particular, given a video sequence, the goal is to separate the nearly-static or slowly-changing background from the dynamic foreground objects ([59]). Based on this observation, [46] has proposed to model the background as a low-rank component, and the dynamic foreground as the sparse noise. In particular, suppose that the video sequence consists of d_f gray-scale frames, each with the resolution of $d_m \times d_n$ pixels. The data matrix X is defined as an asymmetric $d_m d_n \times d_f$ matrix whose i^{th} column is the vectorized version of the i^{th} frame. Therefore, the moving object detection problem can be cast as the recovery of the non-negative vectors $\mathbf{u} \in \mathbb{R}_+^{d_m d_n}$ and $\mathbf{v} \in \mathbb{R}_+^{d_f}$, as well as the sparse matrix $S \in \mathbb{R}^{d_m d_n \times d_f}$, such that

$$X \approx \mathbf{u}\mathbf{v}^\top + S \tag{3.55}$$

Note that the background may not always have a rank-1 representation. However, we will show that (3.55) is sufficiently accurate if the background is relatively static. Furthermore, notice that when the background is completely static, the elements of $\mathbf{v}$ should be equal

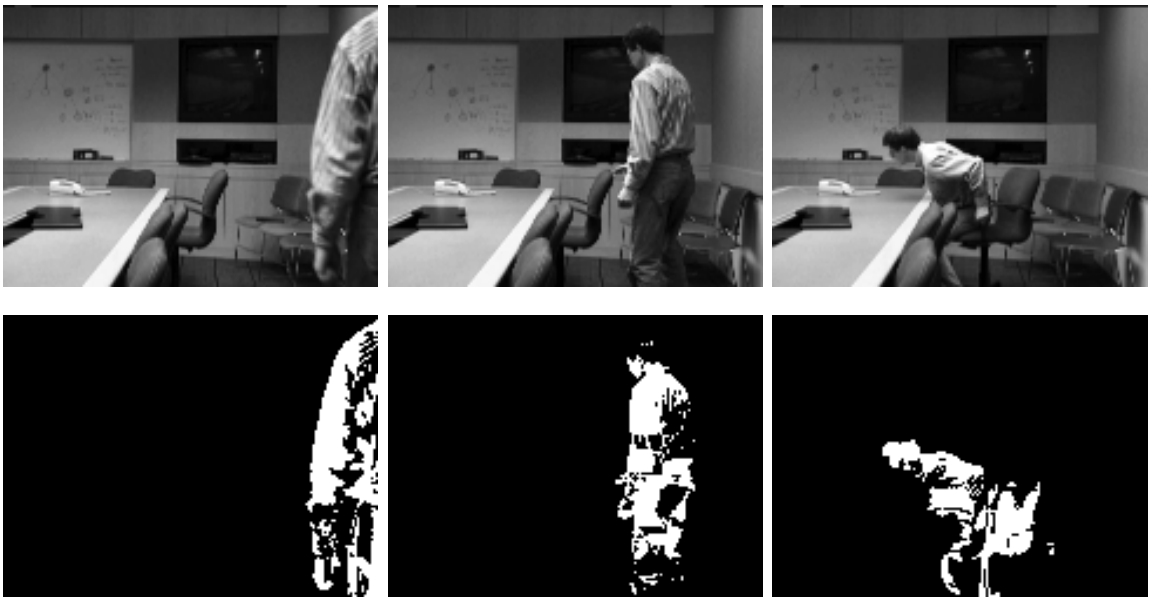


Figure 3.7.3: The performance of the sub-gradient method in the moving object detection problem.

to one. However, this is not desirable in practice since the background may change due to varying illuminations, which can be captured by the variable vector $\mathbf{v}$. Each entry of X is an integer between 0 (darkest) and 255 (brightest). To ensure the positivity of the true components, we increase each element of X by 1 without affecting the performance of the method.

The considered test case is borrowed from the work by [247][4] and is a sequence of video frames taken from a room, where a person walks in, sits on a chair, and uses a phone. We consider 100 gray-scale frames of the sequence, each with the resolution of 120×160 pixels. Therefore, X, $\mathbf{u}$, and $\mathbf{v}$ belong to $\mathbb{R}_+^{19,200\times 100}$, $\mathbb{R}_+^{19,200}$, and $\mathbb{R}_+^{100}$, respectively. Figure 3.7.3 shows that the sub-gradient method with a random initialization can recover the moving object, which is in accordance with the theoretical results of this chapter.

3.8 Discussions on Extension to Rank-r

So far, we have characterized the conditions under which the non-negative rank-1 RPCA has no spurious local solution. However, the following question has been left unanswered: *Can these results be extended to the general non-negative* ***rank*-r** *RPCA?*

[4]The video frames are publicly available at `https://www.microsoft.com/en-us/research/project/test-images-for-wallflower-paper/`.

As a first step toward answering this question and similar to our analysis in the rank-1 case, we consider the noiseless symmetric non-negative rank-r RPCA defined as

$$\min_{U \in \mathbb{R}_+^{n \times r}} \quad f(U) = \|\mathcal{P}_\Omega(U^* U^{*\top} - UU^\top)\|_1 \qquad \text{(P1-Sym-}r\text{)}$$

Indeed, a fundamental roadblock in extending the results of Section 3.4 to (P1-Sym-r) is the implicit *rotational symmetry* in the solution: given a rotation matrix R and a solution $\tilde{U}$ to (P1-Sym-r), $\tilde{U}R$ is another feasible solution with $f(\tilde{U}R) = f(\tilde{U})$, provided that $\tilde{U}R$ is a non-negative matrix. In the rank-1 case, this does not pose any problem since $R = 1$ is the only possible value. However, for the general rank-r case with $r \geq 2$, this rotational symmetry undermines the strict positivity assumption of the true components. In particular, even if the true solution U^* is strictly positive, there exists a rotation matrix R such that U^*R is non-negative with at least one zero entry. This in turn implies that Lemma 12 and, as a consequence, the technique used in Theorem 9 may not be readily extended to the rank-r cases.

Despite the theoretical difficulties in extending the presented results to the general rank-r instances, we have indeed observed—through thousands of simulations—that in general, the sub-gradient method introduced in Section 3.7 successfully converges to a solution U that satisfies $UU^\top = U^* U^{*\top}$, even if the measurement matrix is corrupted with a surprisingly dense noise matrix. To illustrate this, we consider randomly generated instances of the problem with the dimension $n = 100$ and the rank $r \in \{2, 3, 4, 5\}$. For each instance, the elements of U^* are uniformly chosen from the interval $[0.5, 2.5]$. Furthermore, each element in the upper triangular part of the noise matrix S is set to 2 and 0 with probabilities d and $1 - d$, respectively. For each rank r and the noise probability d, we consider 500 independent instances of the problem and solve them using the randomly initialized sub-gradient method. Similar to Subsection 3.7, we assume that a solution is recovered exactly if $\|UU^\top - U^* U^{*\top}\|_F / \|U^* U^{*\top}\|_F \leq 10^{-4}$. Figure 3.8.1 demonstrates the ratio of the instances for which the sub-gradient method successfully recovers the true solution. As illustrated in this figure, d can be as large as 0.30, 0.28, 0.26, and 0.25 to guarantee a success rate of at least 90% when r is equal to 2, 3, 4, and 5, respectively.

This empirical study suggests that one of the following statements may hold for the positive rank-r RPCA: (1) it is devoid of spurious local minima, or (2) its spurious local minima can be escaped efficiently using the sub-gradient method. Further investigation of this direction is left as an enticing challenge for future research.

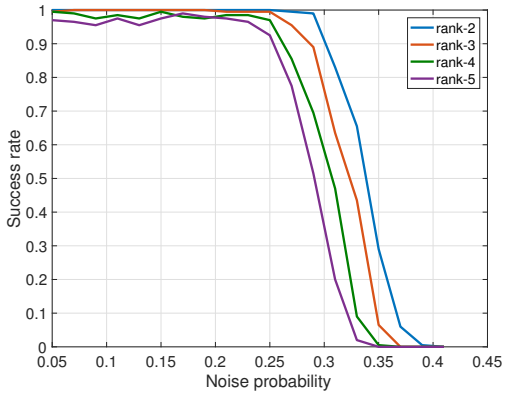


Figure 3.8.1: The success rate of the sub-gradient method for the positive rank-r RPCA.

Appendix

3.A Omitted Proofs of Section 3.4

Proof of Lemma 12

Without loss of generality and for simplicity, we will assume that $\mathcal{G}(\Omega)$ is connected since the proof can be readily applied to each disjoint component of $\mathcal{G}(\Omega)$. Consider a point $\mathbf{u} \geq 0$ with $u_k = 0$ for some k. Consider $\Omega_k = \{j|(k,j) \in \Omega\}$ and note that it is non-empty due to the assumption that $\mathcal{G}(\Omega)$ is connected and non-bipartite. Furthermore, if there exists $r \in \Omega_k$ such that $u_r > 0$, a positive perturbation of u_k will result in a feasible and negative directional derivative. Therefore, suppose that $u_r = 0$ for every $r \in \Omega_k$. Similarly, one can show that if $u_t > 0$ for some $t \in \Omega_r$ and $r \in \Omega_k$, then $\mathbf{u}$ has a feasible and strictly negative directional derivative. Invoking the same argument for the neighbors of the nodes with the zero value, one can infer that $\mathbf{u} = 0$. This completes the proof. □

Proof of Proposition 2

Suppose that $\mathbf{u}^* \geq 0$ and there exists an index i such that $u_i^* = 0$. Without loss of generality, assume that $i = 1$ and $u_j^* > 0$ for every $j \geq 2$. Next, we will show that $\mathbf{u}$ defined as $u_1 = \beta > 0$ and $u_j = 0$ for $j \geq 2$ is a local minimum of (P1-Sym). Consider the perturbed version of $\mathbf{u}$ as

$$\hat{u}_1 \leftarrow \beta + \epsilon_1 \tag{3.56}$$

$$\hat{u}_j \leftarrow \epsilon_j \qquad \forall j \in \{2, ..., n\} \tag{3.57}$$

for sufficiently small $|\epsilon_1|$ and $\epsilon_2, ..., \epsilon_n \geq 0$. Upon defining $\Omega = \{1, ..., n\}^2 \backslash \{(1,1)\}$, one can write

$$f(\mathbf{u}) = \sum_{j=2}^{n} u_j^{*2} + \sum_{j,k=2, j\neq k}^{n} u_j^* u_k^* \tag{3.58}$$

$$f(\hat{\mathbf{u}}) = \sum_{j=2}^{n} u_j^{*2} - \epsilon_j^2 + \sum_{j=2}^{n} (\beta + \epsilon_1)\epsilon_j + \sum_{j,k=2, i\neq j}^{n} |u_j^* u_k^* - \epsilon_j \epsilon_k| \geq f(\mathbf{u}) + \beta \sum_{j=2}^{n} \epsilon_j - \left(\sum_{j=1}^{n} \epsilon_j \right)^2 + \epsilon_1^2 \tag{3.59}$$

It is easy to verify that there exist constants $\bar{\epsilon}_1 > 0$ and $\bar{\epsilon} > 0$ such that for every $-\bar{\epsilon}_1 \leq \epsilon_1 \leq \bar{\epsilon}_1$ and $0 \leq \sum_{j=2}^{n} \epsilon_i \leq \bar{\epsilon}$, we have

$$\beta \sum_{j=2}^{n} \epsilon_i - \left(\sum_{i=1}^{n} \epsilon_i \right)^2 + \epsilon_1^2 \geq 0 \tag{3.60}$$

and hence $f(\hat{\mathbf{u}}) \geq f(\mathbf{u})$. This implies that $\mathbf{u}$ is a local minimum for $f(\mathbf{u})$. □

Proof of Theorem 10

First, we present a number of lemmas that are crucial to the proof of this theorem.

Lemma 20. *Suppose that $\mathcal{G}(\bar{\Omega})$ is connected and $\mathbf{w}^* > 0$. Then, for every D-min-stationary point $\mathbf{w}$, we have $\mathbf{w} > 0$ or $\mathbf{w} = 0$.*

Proof. The proof is omitted due to its similarity to that of Lemma 12. □

Lemma 21. *Suppose that $\mathcal{G}(\bar{\Omega})$ is connected and $\mathbf{w}^* > 0$. Then, $\sum_{i=1}^{m} w_i^2 = \sum_{j=m+1}^{m+n} w_j^2$ holds for every D-stationary point $\mathbf{w} > 0$ of* (P1-Asym).

Proof. By contradiction, suppose that $\sum_{i=1}^{m} w_i^2 \neq \sum_{j=m+1}^{m+n} w_j^2$ for a D-stationary point $\mathbf{w} > 0$. Without loss of generality, suppose that $\sum_{i=1}^{m} w_i^2 > \sum_{j=m+1}^{m+n} w_j^2$ and consider the following perturbation of $\mathbf{w}$

$$\hat{w}_i \leftarrow \begin{cases} w_i - w_i\epsilon & \text{if } 1 \leq i \leq n \\ w_i + w_i\epsilon & \text{if } n+1 \leq i \leq n+m \end{cases} \tag{3.61}$$

For $(i,j) \in \bar{\Omega}$, one can write

$$|\hat{w}_i\hat{w}_j - \hat{w}_i^*\hat{w}_j^*| = |(w_i - w_i\epsilon)(w_j + w_j\epsilon) - \hat{w}_i^*\hat{w}_j^*| = |w_iw_j - \hat{w}_i^*\hat{w}_j^*| + w_iw_j\epsilon^2 \tag{3.62}$$

Therefore, we have

$$f(\hat{\mathbf{w}}) - f(\mathbf{w}) \leq -2\alpha \left(\sum_{i=1}^{m} w_i^2 - \sum_{j=m+1}^{m+n} w_j^2 \right) \epsilon + O(\epsilon^2) \tag{3.63}$$

This implies the existence of strictly positive and negative directional derivatives, thus resulting in a contradiction. This completes the proof. □

Lemma 22. *$\mathcal{G}(\bar{\Omega})$ has a unique vertex partitioning.*

Proof. By contradiction, suppose that there exist two different vertex partitions (S,T) and $(\bar{S},\bar{T})$ for $\mathcal{G}(\bar{\Omega})$. Since $\mathcal{G}(\bar{\Omega})$ is a connected bipartite graph, $\bar{S}$ is not equal to S or T, and therefore, $S \cap \bar{S}$ and $T \cap \bar{T}$ are not empty. Now, it is easy to observe that the nodes in $S \cap \bar{S}$ can only be connected to those in $T \cap \bar{T}$ and, similarly, the nodes in $T \cap \bar{T}$ can only be

connected to those in $S \cap \bar{S}$. Therefore, unless $(S \cap \bar{S}) \cup (T \cap \bar{T})$ includes all the nodes, the graph will be disconnected, contradicting our assumption. On the other hand, this implies that $S \cap \bar{S} = S$ and $T \cap \bar{T} = T$, contradicting the assumption that (S, T) and $(\bar{S}, \bar{T})$ are different. □

Proof of Theorem 10. For a D-min-stationary point $\mathbf{w}$, note that if $w_i = 0$ for some index i, then Lemma 20 implies that $\mathbf{w} = 0$, which can be easily verified to be a local maximum. We assume that $\mathbf{w}^*$ satisfies $\sum_{i=1}^{m} {w_i^*}^2 = \sum_{j=m+1}^{m+n} {w_j^*}^2$, which can be ensured by an appropriate scaling of $\mathbf{u}^*$ and $\mathbf{v}^*$ while keeping $\mathbf{u}^*{\mathbf{v}^*}^\top$ intact. Now, it suffices to show that for a D-stationary point $\mathbf{w} > 0$, we have $\mathbf{w} = \mathbf{w}^*$. This proves the validity of the statements of the theorem.

By contradiction, suppose that $\mathbf{w} > 0$ with $\mathbf{w} \neq \mathbf{w}^*$ is a D-stationary point. In what follows, we will construct directions with strictly positive and negative directional derivatives at this point. Similar to the proof of Theorem 9, one can show that

$$0 < \frac{w_1^*}{w_1} \leq \min_{k \in \bar{\Omega}_i} \frac{w_k^*}{w_k} \leq \frac{w_i}{w_i^*} \leq \max_{k \in \bar{\Omega}_i} \frac{w_k^*}{w_k} \leq \frac{w_{m+n}^*}{w_{m+n}} \tag{3.64}$$

for every $1 \leq i \leq m+n$. By contradiction, suppose that $w_i \neq w_i^*$ for some index i. First, note that $w_{m+n}^*/w_{m+n} > 1$; otherwise, it holds that $w_{m+n}^*/w_{m+n} \leq 1$ and $w_i/w_i^* > 1$, which contradict with (3.64). Define

$$\begin{aligned} T_1^u &= \left\{ i | \frac{w_i^*}{w_i} = \frac{w_{m+n}^*}{w_{m+n}}, 1 \leq i \leq m \right\}, \qquad T_2^u = \left\{ i | \frac{w_i}{w_i^*} = \frac{w_{m+n}^*}{w_{m+n}}, 1 \leq i \leq m \right\} \\ T_1^v &= \left\{ i | \frac{w_i^*}{w_i} = \frac{w_{m+n}^*}{w_{m+n}}, m+1 \leq i \leq m+n \right\}, \; T_2^v = \left\{ i | \frac{w_i}{w_i^*} = \frac{w_{m+n}^*}{w_{m+n}}, m+1 \leq i \leq m+n \right\} \end{aligned} \tag{3.65}$$

and

$$N^u = \{1, \ldots, m\} \backslash (T_1^u \cup T_2^u) \tag{3.66a}$$

$$N^v = \{m+1, \ldots, m+n\} \backslash (T_1^u \cup T_2^u) \tag{3.66b}$$

Furthermore, define $\bar{\mathbf{d}}$ as

$$\bar{d}_i = \begin{cases} \frac{w_i}{w_{m+n}} - w_i \gamma & \text{if } i \in T_1^u \\ -w_i \gamma & \text{if } i \in N^u \\ -\frac{w_i}{w_{m+n}} - w_i \gamma & \text{if } i \in T_2^u \\ \frac{w_i}{w_{m+n}} + w_i \gamma & \text{if } i \in T_1^v \\ w_i \gamma & \text{if } i \in N^v \\ -\frac{w_i}{w_{m+n}} + w_i \gamma & \text{if } i \in T_2^v \end{cases} \tag{3.67}$$

where

$$\gamma = \frac{\sum_{i \in T_1^u} w_i - \sum_{i \in T_2^u} w_i - \sum_{i \in T_1^v} w_i + \sum_{i \in T_2^v} w_i}{w_n \sum_{i=1}^{m+n} w_i} \tag{3.68}$$

Similar to the symmetric case, we show that if $T_1^u \cup T_1^v$ is non-empty, then $f'(\mathbf{w}, \bar{\mathbf{d}}) < 0$ and $f'(\mathbf{w}, -\bar{\mathbf{d}}) > 0$, which contradicts the D-stationarity of $\mathbf{w}$. We will only show $f'(\mathbf{w}, \bar{\mathbf{d}}) < 0$ since $f'(\mathbf{w}, -\bar{\mathbf{d}}) > 0$ can be proven in a similar way. Define a perturbation of $\mathbf{w}$ as $\hat{\mathbf{w}} = \mathbf{w} + \mathbf{d}\epsilon$ where $\epsilon > 0$ is chosen to be sufficiently small.

First, we analyze the regularization term in (P1-Asym). One can write

$$\begin{aligned}\left|\sum_{i=1}^{m}\hat{w}_i^2 - \sum_{j=m+1}^{m+n}\hat{w}_j^2\right| \leq \Bigg|\sum_{i=1}^{m} w_i^2 - \sum_{j=m+1}^{m+n} w_j^2 \\ + 2\left(\sum_{i\in T_1^u}\frac{w_i}{w_{m+n}} - \sum_{i\in T_2^u}\frac{w_i}{w_{m+n}} - \sum_{i\in T_1^v}\frac{w_i}{w_{m+n}} + \sum_{i\in T_2^v}\frac{w_i}{w_{m+n}}\right)\epsilon \\ - 2\gamma\left(\sum_{i=1}^{m} w_i + \sum_{i=m+1}^{m+n} w_i\right)\epsilon\Bigg| + \left(\frac{1}{w_n} + \gamma\right)^2\left(\sum_{i=1}^{m+n} w_i\right)\epsilon^2 \end{aligned} \tag{3.69}$$

Now, according to the definition of γ, one can easily verify that

$$2\left(\sum_{i\in T_1^u}\frac{w_i}{w_{m+n}} - \sum_{i\in T_2^u}\frac{w_i}{w_{m+n}} - \sum_{i\in T_1^v}\frac{w_i}{w_{m+n}} + \sum_{i\in T_2^v}\frac{w_i}{w_{m+n}}\right)\epsilon - 2\gamma\left(\sum_{i=1}^{m} w_i + \sum_{i=m+1}^{m+n} w_i\right)\epsilon = 0 \tag{3.70}$$

This together with Lemma 21, reduces (3.69) to

$$\left|\sum_{i=1}^{m}\hat{w}_i^2 - \sum_{j=m+1}^{m+n}\hat{w}_j^2\right| \leq \left(\frac{1}{w_n} + \gamma\right)^2\left(\sum_{i=1}^{m+n} w_i\right)\epsilon^2 \tag{3.71}$$

To analyze the first term of (P1-Asym), similar to our previous proofs, we will divide the set $\bar{\Omega}$ into different cases (4 cases to be precise) and analyze the effect of the defined perturbation in each case. For the sake of simplicity and to streamline the presentation, we only report the final inequalities for these cases:

1. If $(i,j) \in \bar{\Omega}$ and $(i,j) \in (T_1^u \times T_1^v) \cup (T_2^u \times T_2^v)$, then

$$|\hat{w}_i\hat{w}_j - w_i^* w_j^*| \leq |w_i w_j - w_i^* w_j^*| - \frac{2w_i w_j}{w_{m+n}}\epsilon + w_i w_j\left(\frac{1}{w_{m+n}^2} - \gamma^2\right)\epsilon^2 \tag{3.72}$$

2. If $(i,j) \in \bar{\Omega}$ and $(i,j) \in (N^u \times (T_1^v \cup T_2^v)) \cup ((T_1^u \cup T_2^u) \times N^v)$, then

$$|\hat{w}_i\hat{w}_j - w_i^* w_j^*| \leq |w_i w_j - w_i^* w_j^*| - \frac{w_i w_j}{w_{m+n}}\epsilon + w_i w_j\left(\frac{\gamma}{w_{m+n}^2} - \gamma^2\right)\epsilon^2 \tag{3.73}$$

3. If $(i,j) \in \bar{\Omega}$ and $(i,j) \in (T_1^u \times T_2^v) \cup (T_2^u \times T_1^v)$, then

$$|\hat{w}_i\hat{w}_j - w_i^* w_j^*| \leq |w_i w_j - w_i^* w_j^*| + w_i w_j\left(\frac{\gamma}{w_{m+n}} - \gamma\right)^2\epsilon^2 \tag{3.74}$$

4. If $(i,j) \in \bar{\Omega}$ and $(i,j) \in N^u \times N^v$, then

$$|\hat{w}_i \hat{w}_j - w_i^* w_j^*| \leq |w_i w_j - w_i^* w_j^*| + w_i w_j \gamma^2 \epsilon^2 \tag{3.75}$$

Based on the above inequalities and due to the fact that $\mathcal{G}(\bar{\Omega})$ is connected, one can easily verify that $N^u \cup N^v$ should be empty; otherwise, $\mathbf{w}$ has a strictly negative (and positive) directional derivative. Based on the same reasoning, the graph induced by $T_1^u \cup T_1^v$ or $T_2^u \cup T_2^v$ should be empty. Therefore, $\mathcal{G}$ is bipartite with the components $T_1^u \cup T_1^v$ and $T_2^u \cup T_2^v$. Now, based on Lemma 22, $(T_1^u \cup T_1^v, T_2^u \cup T_2^v)$ induces the same vertex partitioning as (V_u, V_v) (without loss of generality, assume that $T_1^u \cup T_1^v = V_u$ and $T_2^u \cup T_2^v = V_v$). This implies that

$$\frac{w_1}{w_1^*} = \cdots = \frac{w_m}{w_m^*} = \frac{w_{m+1}^*}{w_{m+1}} = \cdots = \frac{w_{m+n}^*}{w_{m+n}} > 1 \tag{3.76}$$

Therefore,

$$\sum_{i=1}^{m} w_i > \sum_{i=1}^{m} w_i^*, \quad \sum_{i=m+1}^{m+n} w_i^* > \sum_{i=m+1}^{m+n} w_i \tag{3.77}$$

Together with the assumption $\sum_{i=1}^{m} w_i^* = \sum_{i=m+1}^{m+n} w_i^*$, this implies that

$$\sum_{i=1}^{m} w_i > \sum_{i=m+1}^{m+n} w_i \tag{3.78}$$

which, according to Lemma 21, contradicts the D-stationarity of $\mathbf{w}$. This completes the proof. □

Proof of Lemma 13

To prove this lemma, first we provide a lower bound on the probability of $\mathcal{G}(\Omega)$ being connected. Define C_k as the number of connected components with exactly k vertices in $\mathcal{G}(\Omega)$. Then, one can write:

$$\mathbb{P}(\mathcal{G}(\Omega) \text{ is connected}) = 1 - \mathbb{P}\left(\bigcup_{k=1}^{\lceil n/2 \rceil} \{C_k > 0\}\right) = 1 - \mathbb{P}(C_1 > 0) - \sum_{k=2}^{\lceil n/2 \rceil} \mathbb{P}(C_k > 0) \tag{3.79}$$

where $\lceil n/2 \rceil$ denotes the smallest integer that is greater than or equal to $n/2$. Next, we provide an upper bound on $\mathbb{P}(C_k > 0)$ for every $k \in \{2, \ldots, \lceil n/2 \rceil\}$. We have

$$\mathbb{P}(C_k > 0) \leq \mathbb{E}\{C_k\} = \sum_{\mathcal{X} \subseteq [1:n], |\mathcal{X}| = k} \mathbb{E}\{I_{\mathcal{X}}\} \tag{3.80}$$

where $I_{\mathcal{X}}$ is an indicator random variable taking the value 1 if the subgraph $\mathcal{G}_{\mathcal{X}}(\Omega)$ of $\mathcal{G}(\Omega)$ induced by the set of vertices in $\mathcal{X}$ is an isolated connected component of $\mathcal{G}(\Omega)$, and it takes the value 0 otherwise. On the other hand, note that $\mathcal{G}_{\mathcal{X}}(\Omega)$ is connected if and only if it contains a spanning tree. Therefore, one can write

$$\begin{aligned}\mathbb{E}\{I_{\mathcal{X}}\} &= \mathbb{P}(\mathcal{G}_{\mathcal{X}}(\Omega) \text{ has a spanning tree}) \\ &\leq \sum_{\mathcal{T}\subset\mathcal{K}_k} \mathbb{P}(\mathcal{T} \text{ belongs to } \mathcal{G}_{\mathcal{X}}(\Omega)) \\ &\leq k^{k-2}p^{k-1} \end{aligned} \tag{3.81}$$

where $\mathcal{K}_k$ is a complete graph over k vertices and $\mathcal{T}$ is a spanning tree. The last inequality is due to the fact that the number of different spanning trees in $\mathcal{K}_k$ is equal to k^{k-2} ([114]). Combining the above inequality with (3.80) results in

$$\begin{aligned}\mathbb{P}(C_k > 0) &\leq \binom{n}{k} k^{k-2}p^{k-1}(1-p)^{k(n-k)} \\ &\leq \left(\frac{ne}{k}\right)^k k^{k-2}e^{-pk(n-k)} \\ &\leq \frac{1}{k^2}e^{-pk(n-k)+k\log n+k} \\ &\leq \frac{1}{k^2}e^{-\frac{k(n-1)}{2}\left(p-\frac{2\log n+2}{n-1}\right)} \end{aligned} \tag{3.82}$$

where the second inequality is due to the relations $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$ and $(1-p)^{k(n-k)} \leq e^{-pk(n-k)}$. Furthermore, the last inequality is due to $k \leq (n+1)/2$. Now, upon choosing $p \geq \frac{(2\eta+2)\log n+2}{n-1}$ for some $\eta > 0$, one can write

$$\mathbb{P}(C_k > 0) \leq \frac{1}{k^2}e^{-\eta k\log n} = \frac{1}{k^2}(n^{-\eta})^k \tag{3.83}$$

Revisiting (3.79), one can also verify that

$$\mathbb{P}(C_1 > 0) \leq n(1-p)^{n-1} \leq e^{-p(n-1)+\log n} \leq n^{-\eta} \tag{3.84}$$

provided that $p \geq \frac{(\eta+1)\log n}{n-1}$, which is implied by $p \geq \frac{(2\eta+2)\log n+2}{n-1}$. Combining this bound with (3.79), one can write

$$\begin{aligned}\mathbb{P}(\mathcal{G}(\Omega) \text{ is connected}) &\geq 1 - n^{-\eta} - \sum_{k=2}^{\lceil n/2\rceil} \frac{1}{k^2}\left(n^{-\eta}\right)^k \\ &\geq 1 - n^{-\eta} - \frac{1}{4}\frac{n^{-2\eta}}{1-n^{-\eta}} \\ &\geq 1 - \left(1 + \frac{1}{4(n^{\eta}-1)}\right)n^{-\eta} \\ &\geq 1 - \frac{5}{4}n^{-\eta} \end{aligned} \tag{3.85}$$

where we have used the assumption $n \geq 2$ and $\eta \geq 1$. Finally, given the event that $\mathcal{G}(\Omega)$ is connected, it is non-bipartite if it has at least one self-loop. Therefore, the probability of $\mathcal{G}(\Omega)$ being non-bipartite is lower bounded by $1-(1-p)^n$. This implies that

$$\begin{aligned}
\mathbb{P}(\mathcal{G}(\Omega) \text{ is connected and non-bipartite}) &\geq \left(1-\frac{5}{4}n^{-\eta}\right)\left(1-(1-p)^n\right) \\
&\geq \left(1-\frac{5}{4}n^{-\eta}\right)\left(1-e^{-np}\right) \\
&\geq \left(1-\frac{5}{4}n^{-\eta}\right)\left(1-e^{-(n-1)p}\right) \\
&\geq \left(1-\frac{5}{4}n^{-\eta}\right)\left(1-e^{-2}n^{-(2\eta+2)}\right) \\
&\geq 1-\frac{3}{2}n^{-\eta} \qquad (3.86)
\end{aligned}$$

This completes the proof. □

Proof of Proposition 5:

To prove Proposition 5, we present another important result on Erdös-Rényi random graphs.

Lemma 23 ([74])**.** *Assuming that $np \to 0$ as $n \to \infty$, the following properties hold with probability approaching to one:*

- *$\mathcal{G}(n,p)$ is acyclic.*
- *The size of every component of $\mathcal{G}(n,p)$ is $O(\log n)$.*

Proof of Proposition 5: Assuming that $np \to 0$, Lemma 23 implies that $\mathcal{G}(\Omega)$ is the union of disjoint tree components, each with the size of at most $O(\log n)$. In what follows, we will show that, with probability approaching to one, $\mathcal{G}(\Omega)$ has at least a bipartite component without any self loops. This, together with Proposition 3, will immediately conclude the

proof. One can write

$$\begin{aligned}
\mathbb{P}(\mathcal{G}(\Omega) \text{ has a bipartite comp.}) &\overset{(a)}{\geq} \mathbb{P}(\mathcal{G}(\Omega) \text{ has a tree comp. without self loops}) \\
&\geq \mathbb{P}(\text{every comp. is a tree with size } O(\log n)) \\
&\times \mathbb{P}(\text{no self-loop in at least one comp}|\text{every comp. is a tree with size } O(\log n)) \\
&\overset{(b)}{=} \mathbb{P}(\text{every comp. is a tree with size } O(\log n)) \\
&\times \mathbb{P}(\text{no self-loop in at least one comp}|\text{every comp. has the size } O(\log n)) \\
&\geq \mathbb{P}(\text{every comp. is a tree with size } O(\log n)) \\
&\times (1 - \mathbb{P}(\text{every comp. has self-loops}|\text{every comp. has the size } O(\log n))) \\
&\geq \mathbb{P}(\underbrace{\text{every comp. is a tree with size } O(\log n)}_{\mathcal{A}}) \\
&\times (1 - \mathbb{P}(\underbrace{\text{there are at least } \Omega(n/\log n) \text{ self-loops}}_{\mathcal{B}})) \qquad (3.87)
\end{aligned}$$

where (a) is followed by the fact that every tree is bipartite, and (b) is followed by the fact that the self-loops are included in the graph independent of other edges. Based on Lemma 23, we have $\mathbb{P}(\mathcal{A}) \to 1$ as $n \to \infty$. Now, we only need to show that $\mathbb{P}(\mathcal{B}) \to 0$ as $n \to \infty$. One can easily verify that

$$\mathbb{P}(\mathcal{B}) \leq \binom{n}{\frac{n}{\log n}} p^{\frac{n}{\log n}} \leq (e \log n)^{\frac{n}{\log n}} \, p^{\frac{n}{\log n}} \qquad (3.88)$$

where the second inequality follows from the relation $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$. Replacing $p = o(1/n)$ gives rise to

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{B}) \leq \lim_{n\to\infty} (ep \log n)^{\frac{n}{\log n}} = 0 \qquad (3.89)$$

Together with (4.6), this implies that $\mathcal{G}(\Omega)$ will have a bipartite component without self loops with probability approaching 1. □

Proof of Lemma 14

We take an approach similar to the proof of Lemma 13. First, recall that $\{V_u, V_v\}$ with $V_u = \{1, \ldots, m\}$ and $V_v = \{m+1, \ldots, m+n\}$ is a vertex partitioning of the bipartite graph $\mathcal{G}(\bar{\Omega})$. Define $C_{k,l}$ as the number of connected components with exactly k vertices from V_u and l vertices from V_v. To simplify the presentation and without loss of generality, we assume

that m and n are even. One can write:

$$\mathbb{P}(\mathcal{G}(\bar{\Omega}) \text{ is connected}) = 1 - \mathbb{P}\left(\bigcup_{\substack{k=0\\k+l\neq 0}}^{\lceil m/2\rceil}\bigcup_{l=1}^{\lceil n/2\rceil}\{C_{k,l}>0\}\right)$$
$$\geq 1 - (\mathbb{P}(C_{1,0}>0)+\mathbb{P}(C_{0,1}>0)) - \sum_{k=1}^{\lceil m/2\rceil}\sum_{l=1}^{\lceil n/2\rceil}\mathbb{P}\left(C_{k,l}>0\right) \quad (3.90)$$

First, we provide an upper bound on $\mathbb{P}\left(C_{k,l}>0\right)$ for $k=1,\ldots,\lceil m/2\rceil$ and $l=1,\ldots,\lceil n/2\rceil$. Similar to the proof of Lemma 13, one can write

$$\mathbb{P}(C_{k,l}>0) \leq \mathbb{E}\{C_{k,l}\} = \sum_{\substack{\mathcal{X}_u\subseteq[1:m],|\mathcal{X}_u|=k\\ \mathcal{X}_v\subseteq[m+1:m+n],|\mathcal{X}_v|=l}} \mathbb{E}\{I_{\mathcal{X}_u,\mathcal{X}_v}\} \quad (3.91)$$

where $I_{\mathcal{X}_u,\mathcal{X}_v}$ is an indicator random variable taking the value 1 if the subgraph $\mathcal{G}_{\mathcal{X}_u,\mathcal{X}_v}(\bar{\Omega})$ of $\mathcal{G}(\bar{\Omega})$ induced by the set of vertices in $\mathcal{X}_u\cup\mathcal{X}_v$ is an isolated connected component of $\mathcal{G}(\bar{\Omega})$, and it takes the value 0 otherwise. On the other hand, we have

$$\begin{aligned}\mathbb{E}\{I_{\mathcal{X}_u,\mathcal{X}_v}\} &= \mathbb{P}(\mathcal{G}_{\mathcal{X}_u,\mathcal{X}_v}(\bar{\Omega}) \text{ has a spanning tree})\\ &\leq \sum_{\mathcal{T}\subset\mathcal{K}_{k,l}}\mathbb{P}(\mathcal{T} \text{ belongs to } \mathcal{G}_{\mathcal{X}_u,\mathcal{X}_v}(\bar{\Omega}))\\ &\leq k^{l-1}l^{k-1}p^{k+l-1}\end{aligned} \quad (3.92)$$

where $\mathcal{K}_{k,l}$ is a complete bipartite graph over two sets of vertices with the sizes k and l, and $\mathcal{T}$ is a spanning tree. The last inequality is due to the fact that the number of different spanning trees in $\mathcal{K}_{k,l}$ is equal to $k^{l-1}l^{k-1}$ ([114]). Therefore, one can write

$$\begin{aligned}\mathbb{P}(C_{k,l}>0) &\leq \binom{m}{k}\binom{n}{l}k^{l-1}l^{k-1}p^{k+l-1}(1-p)^{k(n-l)+l(m-k)}\\ &\leq \left(\frac{me}{k}\right)^k\left(\frac{ne}{l}\right)^l k^{l-1}l^{k-1}e^{-p(k(n-l)+l(m-k))}\\ &\leq \frac{1}{kl}\left(\frac{k}{l}\right)^{l-k}e^{-p(k(n-l)+l(m-k))+k\log m+l\log n+(k+l)}\\ &\leq \frac{1}{kl}e^{-p(k(n-l)+l(m-k))+(k+l)(\log(mn)+1)}\end{aligned} \quad (3.93)$$

where we used the relation $\left(\frac{k}{l}\right)^{l-k}\leq 1$ in the last inequality. Next, we show that the following inequality holds:

$$k(n-l)+l(m-k) \geq (k+l)\frac{(m-1)(n-1)}{m+n} \quad (3.94)$$

To this goal, note that

$$\begin{aligned}
&k(n-l)+l(m-k)\geq(k+l)\frac{(m-1)(n-1)}{m+n}\\
\Longleftrightarrow\ &k(m+n)(n-l)+l(m+n)(m-k)\geq(k+l)(m-1)(n-1)\\
\Longleftrightarrow\ &(k+l)mn+kn(n-2l)+lm(m-2k)\geq(k+l)(m-1)(n-1)\\
\Longleftrightarrow\ &kn(n-2l)+lm(m-2k)\geq-nk-ml-(n-1)l-(m-1)l
\end{aligned}\tag{3.95}$$

where the last inequality holds due to $l\leq(n+1)/2$ and $k\leq(m+1)/2$, which in turn implies that $kn(n-2l)+lm(m-2k)\geq-nk-ml$. Combining (3.94) and (3.93) leads to

$$\mathbb{P}(C_{k,l}>0)\leq\frac{1}{kl}e^{-(k+l)\frac{(m-1)(n-1)}{m+n}\left(p-\frac{(m+n)(\log(mn)+1)}{(m-1)(n-1)}\right)}\tag{3.96}$$

Upon choosing $p\geq\frac{(m+n)((1+\eta)\log(mn)+1)}{(m-1)(n-1)}$ for some $\eta\geq1$, one can write

$$\mathbb{P}(C_{k,l}>0)\leq\frac{1}{kl}\left((mn)^{-\eta}\right)^{(k+l)}\tag{3.97}$$

On the other hand, it is easy to verify that

$$\begin{aligned}
\mathbb{P}(C_{0,1}>0)&\leq n(1-p)^m\leq e^{-pm+\log n}\leq(mn)^{-\eta}\\
\mathbb{P}(C_{1,0}>0)&\leq m(1-p)^n\leq e^{-pn+\log m}\leq(mn)^{-\eta}
\end{aligned}\tag{3.98}$$

provided that $p\geq\frac{(1+\eta)\log(mn)}{m}$ and $p\geq\frac{(1+\eta)\log(mn)}{n}$, both of which are guaranteed to hold with the choice of $p\geq\frac{(m+n)((1+\eta)\log(mn)+1)}{(m-1)(n-1)}$. Combining (3.98), (3.97), and (3.90) results in

$$\begin{aligned}
\mathbb{P}(\mathcal{G}(\bar{\Omega})\text{ is connected})&\geq1-2(mn)^{-\eta}-\sum_{k=1}^{\lceil m/2\rceil}\sum_{l=1}^{\lceil n/2\rceil}\frac{1}{kl}\left((mn)^{-\eta}\right)^{(k+l)}\\
&\geq1-2(mn)^{-\eta}-4(mn)^{-2\eta}
\end{aligned}\tag{3.99}$$

where we have used the assumptions $m,n\geq2$ and $\eta\geq1$. This completes the proof. □

3.B Omitted Proofs of Section 3.5

Proof of Lemma 15

We present the proof for the symmetric case (the proof for the asymmetric case follows directly after symmetrization and the fact that the penalty on the norm difference is zero at the positive D-stationary points). First, we prove that $u_{\max}\leq2$. It suffices to show that $u_{\max}\leq\max\{2\beta,\sqrt{2n/\lambda}\}$. This, together with the choice of β and λ, implies $u_{\max}\leq2$. To this goal, we only need to verify that $u_{\max}>2\beta$ implies $u_{\max}\leq\sqrt{2n/\lambda}$. By contradiction,

suppose that $u_{\max} > \sqrt{2n/\lambda}$. In what follows, it will be shown that $\mathbf{u}$ has strictly positive and negative directional derivatives, thereby contradicting its D-stationarity. Consider a perturbation of $\mathbf{u}$ as $\hat{\mathbf{u}} = \mathbf{u} - \mathbf{e}_{\max}\epsilon$ for a sufficiently small $\epsilon > 0$, where $\mathbf{e}_{\max}$ is a vector with 1 at the location corresponding to $u_{\max}$ and 0 everywhere else. One can write

$$\begin{aligned} f_{\text{reg}}(\hat{\mathbf{u}}) - f_{\text{reg}}(\mathbf{u}) &\leq \left(\sum_{i=1}^{n} u_i\right)\epsilon + \lambda\left((u_{\max} - \epsilon - \beta)^4 - (u_{\max} - \beta)^4\right) \\ &= \left(\sum_{i=1}^{n} u_i\right)\epsilon - 4\lambda(u_{\max} - \beta)^3\epsilon + O(\epsilon^2) \\ &\overset{(a)}{\leq} \left(\sum_{i=1}^{n} u_i - \frac{\lambda}{2}u_{\max}^3\right)\epsilon + O(\epsilon^2) \\ &\leq \left(nu_{\max} - \frac{\lambda}{2}u_{\max}^3\right)\epsilon + O(\epsilon^2) \end{aligned} \tag{3.100}$$

where (a) is due to the fact that $u_{\max} \geq 2\beta$ implies $u_{\max} - \beta \geq u_{\max}/2$. (3.100) together with $u_{\max} > \sqrt{2n/\lambda}$, implies that $-\mathbf{e}_{\max}$ is a direction with a negative directional derivative. Similarly, it can be shown that $\mathbf{e}_{\max}$ is a direction with a positive directional derivative. This contradicts the D-stationarity of $\mathbf{u}$ and, hence, $u_{\max} \leq \max\{2\beta, \sqrt{2n/\lambda}\}$.

Next, we aim to show that $(c/2)u_{\min}^{*^2} \leq u_{\min}$. By contradiction, suppose that there exists an index i such that $(c/2)u_{\min}^{*^2} > u_i$. Now, since $u_i < 1$, we have $\mathbb{I}_{u_i \geq \beta} = 0$ due to the choice of β. Consider the terms in $f_{\text{reg}}(\mathbf{u})$ that involves u_i:

$$\sum_{j \in \Omega_i} |u_i u_j - X_{ij}| = \sum_{j \in G_i} |u_i u_j - u_i^* u_j^*| + \sum_{j \in B_i} |u_i u_j - (u_i^* u_j^* + S_{ij})| \tag{3.101}$$

Considering the fact that $u_{\max} \leq 2$, one can verify the following inequality for every $(i,j) \in G$:

$$u_i u_j < c u_{\min}^{*^2} \leq u_{\min}^{*^2} \leq u_i^* u_j^* \tag{3.102}$$

A similar inequality holds for $(i,j) \in B$:

$$u_i u_j < c u_{\min}^{*^2} \overset{(a)}{\leq} u_i^* u_j^* + S_{ij} \tag{3.103}$$

where we have used Assumption 1 for (a). Therefore, a positive and negative perturbation of u_i results in negative and positive directional derivatives at $\mathbf{u}$, thereby contradicting the D-stationarity of this point. □

Proof of Theorem 13

The next lemma is crucial in proving Theorem 13.

Lemma 24. *Suppose that the assumptions of Theorem 13 hold and define*

$$s(\mathbf{u}) = \underbrace{-\sum_{\substack{(i,j)\in\mathcal{B}\\ i,j\in T_1}} \frac{2u_iu_j}{u_n} + \sum_{\substack{(i,j)\in\mathcal{B}\\ i,j\in T_2}} \frac{2u_iu_j}{u_n} + \sum_{\substack{(i,j)\in\mathcal{B}\\ i\in T_1\cup T_2, j\in N}} \frac{u_iu_j}{u_n}}_{f_B(\mathbf{u})}$$
$$+ \underbrace{\sum_{\substack{(i,j)\in\mathcal{G}\\ i,j\in T_1}} \frac{2u_iu_j}{u_n} + \sum_{\substack{(i,j)\in\mathcal{G}\\ i,j\in T_2}} \frac{2u_iu_j}{u_n} + \sum_{\substack{(i,j)\in\mathcal{G}\\ i\in T_1\cup T_2, j\in N}} \frac{u_iu_j}{u_n}}_{f_G(\mathbf{u})} + \underbrace{\sum_{i\in T_2} \frac{4u_i(u_i-1)^3}{u_n}\mathbb{I}_{u_i\geq 1}}_{f_R(\mathbf{u})} \tag{3.104}$$

where the sets T_1 and T_2 are defined as (4.24) *and* (4.29)*, respectively. Then, for every D-stationary point $\mathbf{u} > 0$ such that $\mathbf{u} \neq \mathbf{u}^*$, the following inequalities hold with the choice of $\beta = 1$ and $\lambda = n/2$:*

- *$f_{\text{reg}}(\hat{\mathbf{u}}) - f_{\text{reg}}(\mathbf{u}) \leq -s(\mathbf{u})\epsilon + O(\epsilon^2)$ for $\hat{\mathbf{u}} = \mathbf{u} + \mathbf{d}\epsilon$ and a sufficiently small $\epsilon > 0$.*
- *$f_{\text{reg}}(\hat{\mathbf{u}}) - f_{\text{reg}}(\mathbf{u}) \geq s(\mathbf{u})\epsilon - O(\epsilon^2)$ for $\hat{\mathbf{u}} = \mathbf{u} - \mathbf{d}\epsilon$ and a sufficiently small $\epsilon > 0$.*

where $\mathbf{d}$ is defined as (3.20)*.*

Proof. To prove this lemma, first we show the validity of (4.20). By contradiction, suppose that (4.20) does not hold. Without loss of generality, assume that there exists an index i such that $u_i/u_i^* > u_n^*/u_n$ (the case with $u_i/u_i^* < u_1^*/u_1$ can be argued in a similar way). This implies that $u_iu_j > u_i^*u_j^*$ for every $(i,j) \in \Omega$. Define $\hat{\mathbf{u}} = \mathbf{u} - \mathbf{e}\epsilon$ for a sufficiently small $\epsilon > 0$, where $\mathbf{e}$ is a vector with $e_k = 1$ if $k = i$ and $e_k = 0$ otherwise. One can write

$$\begin{aligned} f_{\text{reg}}(\hat{\mathbf{u}}) - f_{\text{reg}}(\mathbf{u}) &\leq -\left(\sum_{j\in G_i} u_j\right)\epsilon + \left(\sum_{j\in B_i} u_j\right)\epsilon + \lambda\left((u_i-\epsilon-1)^4 - (u_i-1)^4\right)\mathbb{I}_{u_i\geq 1} \\ &\leq -\left(\sum_{j\in G_i} u_j\right)\epsilon + \left(\sum_{j\in B_i} u_j\right)\epsilon \\ &\leq -\frac{cu_{\min}^{*^2}}{2}\delta(\mathcal{G}(G)) + 2\Delta(\mathcal{G}(B)) \end{aligned} \tag{3.105}$$

where $G_i = \{j|(i,j) \in G\}$ and $B_i = \{j|(i,j) \in B\}$. The second inequality is due to the fact that $((u_i-\epsilon-1)^4 - (u_i-1)^4)\,\mathbb{I}_{u_i\geq 1}$ is non-negative and the third inequality follows from Lemma 15 and the definitions of $\delta(\mathcal{G}(G))$, $\Delta(\mathcal{G}(B))$. Based on the assumption of Theorem 13, we have

$$\frac{\delta(\mathcal{G}(G))}{\Delta(\mathcal{G}(B))} > \frac{48}{c^2}\kappa(\mathbf{u}^*)^4 = \frac{48}{c^2u_{\min}^{*^4}} > \frac{4}{cu_{\min}^{*^2}} \tag{3.106}$$

which implies $(-cu_{\min}^{*^2}/2)\delta(\mathcal{G}(G)) + 2\Delta(\mathcal{G}(B)) < 0$, and hence, $-\mathbf{e}$ is a direction with a negative directional derivative. Similarly, it can be shown that $\mathbf{e}$ is a direction with a positive

directional derivative. This contradicts the D-stationarity of $\mathbf{u}$ and hence (4.20) holds. Now, we will show the correctness of the first statement. Similar to the proof of Theorem 9, one can verify that

$$\sum_{(i,j)\in\Omega} |\hat{u}_i\hat{u}_j - X_{ij}| - \sum_{(i,j)\in\Omega} |u_iu_j - X_{ij}| \leq (f_B(\mathbf{u}) - f_G(\mathbf{u}))\epsilon + O(\epsilon^2) \tag{3.107}$$

Now, we only need to bound $R(\hat{\mathbf{u}})-R(\mathbf{u})$. To this goal, notice that if $i \in T_1$, then $u_i < u_i^* \leq 1$ due to the fact that $\mathbf{u} \neq \mathbf{u}^*$ and $u_i^*/u_i = u_n^*/u_n > 1$. Therefore, $\mathbb{I}_{u_i\geq 1} = 0$ for every $i \in T_1$. This implies that

$$\begin{aligned} R(\hat{\mathbf{u}}) - R(\mathbf{u}) &= \sum_{i\in T_2} \left(u_i - \frac{u_i}{u_n}\epsilon - 1\right)^4 \mathbb{I}_{u_i\geq 1} - \sum_{i\in T_2} (u_i-1)^4 \mathbb{I}_{u_i\geq 1} \\ &= -\sum_{i\in T_2} \frac{4u_i(u_i-1)^3}{u_n}\mathbb{I}_{u_i\geq 1}\epsilon + O(\epsilon^2) \end{aligned} \tag{3.108}$$

A similar approach can be taken to prove the second statement of the lemma. □

Lemma 25. *Suppose that $\mathcal{G}(\Omega)$ has no bipartite component and every entry of X is strictly positive. Then, for every D-min-stationary point $\mathbf{u}$ of* (P1-Sym)*, we have $\mathbf{u}[c] > 0$ or $\mathbf{u}[c] = 0$, where $\mathbf{u}[c]$ is a sub-vector of $\mathbf{u}$ induced by the c^{th} component of $\mathcal{G}(\Omega)$.*

Proof. The proof is similar to that of Lemma 12. □

Proof of Theorem 13: Similar to the proof of Theorem 9, it suffices to show that none of the points $\mathbf{u} > 0$ with $\mathbf{u} \neq \mathbf{u}^*$ can be D-stationary. By contradiction, suppose that this is not the case, i.e., there exists a D-stationary point $\mathbf{u} > 0$ such that $\mathbf{u} \neq \mathbf{u}^*$. Consider the functions $f_B(\mathbf{u})$ and $f_G(\mathbf{u})$ defined in Lemma 24. The main idea behind the proof is to show that the term $f_G(\mathbf{u})$ always dominates $f_B(\mathbf{u})$. This, together with the non-negativity of $f_R(\mathbf{u})$, shows that $s(\mathbf{u}) > 0$ and hence, $f'_{\text{reg}}(\mathbf{u}, \mathbf{d}) < 0$ and $f'_{\text{reg}}(\mathbf{u}, -\mathbf{d}) > 0$, which is a contradiction. One can bound each term in $f_B(\mathbf{u})$ and obtain

$$\begin{aligned} f_B(\mathbf{u}) \leq& \frac{1}{u_n}\left(2\cdot\frac{\Delta(\mathcal{G}(B))}{2}|T_1|u_{\max}^2 + 2\cdot\frac{\Delta(\mathcal{G}(B))}{2}|T_2|u_{\max}^2 + \frac{\Delta(\mathcal{G}(B))}{2}(|T_1|+|T_2|)u_{\max}^2\right)\epsilon + O(\epsilon^2) \\ \leq& \frac{3}{2u_n}\Delta(\mathcal{G}(B))(|T_1|+|T_2|)u_{\max}^2\epsilon + O(\epsilon^2) \\ \leq& \frac{6}{u_n}\Delta(\mathcal{G}(B))(|T_1|+|T_2|)\epsilon + O(\epsilon^2) \end{aligned} \tag{3.109}$$

where the last inequality follows from the fact that $u_{\max} \leq 2$ due to Lemma 15. Next, we derive a lower bound on $f_G(\mathbf{x})$:

$$\begin{aligned} f_G(\mathbf{x}) \geq & \frac{1}{u_n} \cdot \frac{\delta(\mathcal{G}(G))}{2}(|T_1| + |T_2|)u_{\min}^2\epsilon + O(\epsilon^2) \\ \geq & \frac{1}{u_n} \cdot \frac{\delta(\mathcal{G}(G))}{2}(|T_1| + |T_2|)\frac{c^2 u_{\min}^{*4}}{4}\epsilon + O(\epsilon^2) \\ = & \frac{c^2 u_{\min}^{*4}}{8u_n}\delta(\mathcal{G}(G))(|T_1| + |T_2|)\epsilon + O(\epsilon^2) \end{aligned} \tag{3.110}$$

where the first inequality is due to the fact that the minimum value for $f_G(\mathbf{u})$ happens when the neighbors of $T_1 \cup T_2$ in $\mathcal{G}(G)$ all belong to the set N and their corresponding values in $\mathbf{u}\mathbf{u}^\top$ are all equal to $u_{\min}^2$. Furthermore, the second inequality is due to Lemma 13 and the choice of β for $R(\mathbf{u})$. Therefore, one can write

$$\begin{aligned} f_B(\mathbf{x}) - f_G(\mathbf{x}) \leq & \left(\frac{6}{u_n}\Delta(\mathcal{G}(B)) - \frac{c^2 u_{\min}^{*4}}{8u_n}\delta(\mathcal{G}(G))\right)(|T_1| + |T_2|)\epsilon + O(\epsilon^2) \\ = & \frac{\Delta(\mathcal{G}(B))c^2 u_{\min}^{*4}}{8u_n}\left(\frac{48}{c^2}\kappa(\mathbf{u}^*)^4 - \frac{\delta(\mathcal{G}(G))}{\Delta(\mathcal{G}(B))}\right)(|T_1| + |T_2|)\epsilon + O(\epsilon^2). \end{aligned} \tag{3.111}$$

Therefore, the choice of $(48/c^2)\kappa(\mathbf{u}^*)^4 < \delta(\mathcal{G}(G))/\Delta(\mathcal{G}(B))$ implies that $f_B(\mathbf{x}) - f_G(\mathbf{x}) < 0$, thereby completing the proof. □

Proof of Lemma 16

The degree of each node is equal to the summation of n independent Bernoulli random variables, each with parameter p. Therefore, Chernoff bound yields that

$$\mathbb{P}(\deg(v) \geq (1+\delta)np) \leq e^{-np\delta^2/3} \tag{3.112a}$$

$$\mathbb{P}(\deg(v) \leq (1-\delta)np) \leq e^{-np\delta^2/3} \tag{3.112b}$$

for every vertex v and $0 \leq \delta \leq 1$, where $\deg(v)$ is the degree of vertex v in the graph. Therefore, a simple union bound leads to

$$\mathbb{P}(\Delta(\mathcal{G}(n,p)) \geq (1+\delta)np) \leq ne^{-np\delta^2/3} = e^{-np\delta^2/3+\log n} \tag{3.113a}$$

$$\mathbb{P}(\delta(\mathcal{G}(n,p)) \leq (1-\delta)np) \leq ne^{-np\delta^2/3} = e^{-np\delta^2/3+\log n} \tag{3.113b}$$

Setting $\delta = 1/2$ and assuming that $p \geq 12(1+\eta)\log n/n$ for some $\eta > 0$, one can write

$$\mathbb{P}\left(\Delta(\mathcal{G}(n,p)) \geq \frac{3np}{2}\right) \leq n^{-\eta} \tag{3.114a}$$

$$\mathbb{P}\left(\delta(\mathcal{G}(n,p)) \leq \frac{np}{2}\right) \leq n^{-\eta} \tag{3.114b}$$

Furthermore, $p < 12(1+\eta)\log n/n$ leads to

$$\begin{aligned}\mathbb{P}\left(\Delta(\mathcal{G}(n,p)) \geq 18(1+\eta)\log n\right) &\leq \mathbb{P}\left(\Delta\left(\mathcal{G}\left(n, \frac{12(1+\eta)\log n}{n}\right)\right) \geq 18(1+\eta)\log n\right) \\ &\leq \mathbb{P}\left(\Delta\left(\mathcal{G}\left(n, \frac{12(1+\eta)\log n}{n}\right)\right) \geq \frac{3np}{2}\right) \\ &\leq n^{-\eta} \end{aligned} \tag{3.115}$$

Combining (3.115) with (3.114a) and (3.114b) results in the desired inequalities. This completes the proof. □

Proof of Lemma 17

Define $S = \{1, ..., m\}$ and $T = \{m+1, ..., m+n\}$. Similar to the proof of Lemma 3.B, one can write the following concentration inequalities:

$$\mathbb{P}(\max_{v\in S}\{\deg(v)\} \geq (1+\delta)np) \leq me^{-np\delta^2/3} \tag{3.116a}$$

$$\mathbb{P}(\min_{v\in S}\{\deg(v)\} \leq (1-\delta)np) \leq me^{-np\delta^2/3} \tag{3.116b}$$

$$\mathbb{P}(\max_{v\in T}\{\deg(v)\} \geq (1+\delta)mp) \leq ne^{-mp\delta^2/3} \tag{3.116c}$$

$$\mathbb{P}(\min_{v\in T}\{\deg(v)\} \leq (1-\delta)mp) \leq ne^{-mp\delta^2/3} \tag{3.116d}$$

which imply

$$\mathbb{P}(\Delta(\mathcal{G}(m,n,p)) \geq (1+\delta)np) \leq me^{-np\delta^2/3} + ne^{-mp\delta^2/3} \leq 2e^{-mp\delta^2/3+\log n} \tag{3.117a}$$

$$\mathbb{P}(\delta(\mathcal{G}(m,n,p)) \leq (1-\delta)mp) \leq me^{-np\delta^2/3} + ne^{-mp\delta^2/3} \leq 2e^{-mp\delta^2/3+\log n} \tag{3.117b}$$

Setting $\delta = 1/2$ and assuming that $p \geq 12(1+\eta)\log n/m$ for some $\eta > 0$ results in

$$\mathbb{P}(\Delta(\mathcal{G}(m,n,p)) \geq \frac{3np}{2}) \leq 2n^{-\eta} \tag{3.118a}$$

$$\mathbb{P}(\delta(\mathcal{G}(m,n,p)) \leq \frac{mp}{2}) \leq 2n^{-\eta} \tag{3.118b}$$

Furthermore, if $p < 12(1+\eta)\log n/m$, one can write

$$\begin{aligned}\mathbb{P}\left(\Delta(\mathcal{G}(n,p)) \geq \frac{18(1+\eta)n\log n}{m}\right) &\leq \mathbb{P}\left(\Delta\left(\mathcal{G}\left(n, \frac{12(1+\eta)\log n}{m}\right)\right) \geq \frac{18(1+\eta)n\log n}{m}\right) \\ &\leq \mathbb{P}\left(\Delta\left(\mathcal{G}\left(n, \frac{12(1+\eta)\log n}{m}\right)\right) \geq \frac{3np}{2}\right) \\ &\leq 2n^{-\eta} \end{aligned} \tag{3.119}$$

This completes the proof. □

Part II

Network Optimization

Chapter 4

Convexification of Generalized Network Flow

This chapter is concerned with the minimum-cost flow problem over an arbitrary flow network. In this problem, each node is associated with some possibly unknown injection and each line has two unknown flows at its ends that are related to each other via a nonlinear function. Moreover, all injections and flows must satisfy certain box constraints. This problem, named generalized network flow (GNF), is highly non-convex due to its nonlinear equality constraints. Under the assumption of monotonicity and convexity of the flow and cost functions, a convex relaxation is proposed, which is shown to always obtain globally optimal injections. This relaxation may fail to find optimal flows because the mapping from injections to flows is not unique in general. We show that the proposed relaxation, named convexified GNF (CGNF), obtains a globally optimal flow vector if the optimal injection vector is a Pareto point. More generally, the network can be decomposed into two subgraphs such that the lines between the subgraphs are congested at optimality and that CGNF finds correct optimal flows over all lines of one of these subgraphs. We also fully characterize the set of all globally optimal flow vectors, based on the optimal injection vector found via CGNF. In particular, we show that this solution set is a subset of the boundary of a convex set, and may include an exponential number of disconnected components. A primary application of this work is in optimization over electrical power networks.

4.1 Introduction

The area of "network flows" plays a central role in operations research, computer science and engineering [105, 130]. This area is motivated by many real-word applications in assignment, transportation, communication networks, electrical power distribution, production scheduling, financial budgeting, and aircraft routing, to name only a few. Network flow problems have been studied extensively since 1962 [94, 140, 6, 29, 30, 23, 70, 197, 106, 32]. The minimum-cost flow problem aims to optimize the flows over a flow network that is

used to carry some commodity from suppliers to consumers. In a flow network, there is an injection of some commodity at every node, which leads to two flows over each line at its endpoints. The injection—depending on being positive or negative, corresponds to supply or demand at the node. The minimum-cost flow problem has been studied thoroughly for a lossless network, where the amount of flow entering a line equals the amount of flow leaving the line. However, since real-world flow networks could be lossy, the minimum-cost flow problem has also attracted much attention for generalized networks, also known as networks with gain [130, 38, 104]. In this type of network, each line is associated with a constant gain relating the two flows of the line through a linear function. From the optimization perspective, network flow problems are convex and can be solved efficiently, unless there are discrete variables involved [37].

There are important real-world flow networks that are lossy, where the loss is a nonlinear function of the flows. An example is electrical power networks for which the loss over each line (under fixed voltage magnitudes at both ends) is given by a parabolic function due to Kirchhoff's circuit laws [143]. The loss function could be much more complicated depending on the power electronic devices installed on the transmission line. To the best of our knowledge, there is no theoretical result in the literature on the design of efficient algorithms for network flow problems with nonlinear flow functions, except in very special cases. This chapter is concerned with this general problem, named Generalized Network Flow (GNF). Note that the term "GNF" has already been used in the literature for networks with linear losses, but it corresponds to arbitrary lossy networks in this work.

GNF aims to optimize the nodal injections subject to flow constraints for each line and box constraints for both injections and flows. A flow constraint is a nonlinear equation that relates the flows at both ends of a line. To solve GNF, this chapter makes the practical assumption that the cost and flow functions are all monotonic and convex. The GNF problem is still highly non-convex due to its equality constraints. However, a question arises as to whether there is an efficient algorithm for finding globally optimal injections and flows for GNF under the assumption that the GNF problem is feasible. In this work, we prove that the answer to this question is affirmative for optimal injections (and optimal total cost), but not necessarily optimal flows. More specifically, we provide a convex relaxation of GNF that yields globally optimal injections.

Observe that relaxing the nonlinear line flow equalities to convex inequalities gives rise to a convex relaxation of GNF. It can be easily seen that solving the relaxed problem may lead to a solution for which the new inequality flow constraints are not binding. One may speculate that this observation implies that the convex relaxation is not tight. However, the objective of this work is to show that as long as GNF is feasible, the convex relaxation is tight. We also generalize the above results to the case where, other than local constraints over a line or at a node, there are global constraints relating the flows of different lines or injections of different nodes.

Although the proposed convex relaxation always finds the optimal injections (and hence the optimal objective value), it may produce wrong flows leading to non-binding inequalities. The reason behind the failure of the convex relaxation in finding globally optimal flows is

that the mapping from flows to injections is not invertible. For example, it is known in the context of power systems that the power flow equations may not have a unique solution [8]. Having found the globally optimal injection vector through the proposed convex relaxation, we also study the possibility of finding optimal flows from the optimal injections. First, we prove that if the optimal injection vector is a Pareto point in its feasible region, the convex relaxation of GNF obtains globally optimal flows for GNF. Second, we show that whenever the optimal injection vector lies on the boundary of its feasible region, the flow network can be divided into two sub-networks such that: (i) the convex relaxation obtains optimal flows over one sub-network, (ii) the lines between the two sub-networks are all congested at optimality and the convex relaxation correctly identifies these lines. In other words, we relate the possible failure of the convex relaxation in finding optimal flows for the whole network to certain congested lines. Moreover, we fully characterize the set of all optimal flow vectors. In particular, we show that this set may be infinite, non-convex, and disconnected, but belongs to the boundary of a convex set.

Application of GNF in Power Systems

The operation of a power network depends heavily on various large-scale optimization problems such as state estimation, optimal power flow (OPF), security-constrained OPF, unit commitment, sizing of capacitor banks and network reconfiguration. These problems are highly non-convex due to the nonlinearities imposed by laws of physics [121, 236]. For example, each of the above problems has the power flow equations embedded in it, which are nonlinear equality constraints. The nonlinearity of OPF, as the most fundamental optimization problem for power systems, has been studied since 1962, leading to various heuristic and local-search algorithms [49, 64, 187, 188, 203, 18, 206, 129, 175]. These algorithms suffer from sensitivity and convergence issues, and more importantly they may converge to a local optimum that is noticeably far from a global solution.

Recently, it has been shown in [151, 169] that the semidefinite programming (SDP) relaxation is able to find a global or near-global solution of the OPF problem under a sufficient condition, which is satisfied for IEEE benchmark systems, Polish Grid with more than 3000 nodes, and many randomly generated power networks. The papers [240] and [236] prove that the satisfaction of this condition is due to the passivity of transmission lines and transformers. In particular, [236] shows that in the case where this condition is not satisfied (see [158] for counterexamples), OPF can always be solved globally in polynomial time (up to any finite precision) after two approximations: (i) relaxing angle constraints by adding a sufficient number of actual/virtual phase shifters to the network, (ii) relaxing power balance equalities to inequality constraints. OPF under Approximation (ii) was also studied in [35, 276, 154] for distribution networks. The paper [153] studies the optimization of active power flows over distribution networks under fixed voltage magnitudes and shows that the SDP relaxation works without having to use Approximation (ii) as long as a practical angle condition is satisfied.

The idea of convex relaxation developed in [150] and [151] can be applied to many other power problems, such as voltage regulation [145], energy storage [99], state estimation [262, 170], sensor placement [136], calculation of voltage stability margin [185], charging of electric vehicles [237], security constrained OPF with possibly variable tap-changers and capacitor banks [149, 169], dynamic energy management [143] and electricity market [152]. In the same vein, [120] and [141] combine a convex relaxation of the power flow equations with iterative approaches to reduce the complexity of the semidefinite programming and to address certain problems in power systems that include discrete variables, such as unit commitment and optimal transmission switching problems [81, 198]. Although the SDP relaxation has been shown to be exact in several real-world examples, [158] demonstrates that this relaxation may fail in some instances. To improve upon the SDP relaxation for such cases, [133] and [186] use a hierarchy of semidefinite relaxations, known as *Lasserre* hierarchy [148], which obtain global minima of the OPF problem at the expense of a higher computational complexity. The paper [171] proves that in the case where the SDP relaxation is not exact, it still has a low-rank solution whose rank is upper bounded by the treewidth of the power system plus one.

Energy-related optimization problems with embedded power flow equations can be regarded as nonlinear network flow problems, which are analogous to GNF. The results derived in this work for a general GNF problem lead to the generalization of the result of [154] to networks with virtual phase shifters. This proves that in order to use SDP relaxations for OPF over an arbitrary power network, it is not needed to approximate power balance equalities with inequality constraints (under a practical angle assumption).

4.2 Problem Formulation and Contributions

Consider an undirected graph (network) $\mathcal{G}$ with the vertex set $\mathcal{N} := \{1, 2, ..., m\}$ and the edge set $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$. For every $i \in \mathcal{N}$, let $\mathcal{N}(i)$ denote the set of the neighboring vertices of node i. Assume that every edge $(i, j) \in \mathcal{E}$ is associated with two unknown flows p_{ij} and p_{ji} belonging to $\mathbb{R}$. The parameters p_{ij} and p_{ji} can be regarded as the flows entering the edge (i, j) from the endpoints i and j, respectively. Define

$$p_i = \sum_{j \in \mathcal{N}(i)} p_{ij}, \qquad \forall i \in \mathcal{N} \tag{4.1}$$

The parameter p_i is called "nodal injection at vertex i" or simply "injection", which is equal to the sum of the flows leaving vertex i through the edges connected to this vertex. Given an edge $(i, j) \in \mathcal{E}$, we assume that the flows p_{ij} and p_{ji} are related to each other via a function $f_{ij}(\cdot)$ to be introduced later. To specify which of the flows p_{ij} and p_{ji} is a function of the other, we give an arbitrary orientation to every edge of the graph $\mathcal{G}$ and denote the resulting graph as $\overrightarrow{\mathcal{G}}$. Denote the directed edge (arc) set of $\overrightarrow{\mathcal{G}}$ as $\overrightarrow{\mathcal{E}}$. If an edge $(i, j) \in \mathcal{E}$ belongs to $\overrightarrow{\mathcal{E}}$, we then express p_{ji} as a function of p_{ij}.

Definition 16. *Define the vectors* $\mathbf{p}_n$, $\mathbf{p}_e$ *and* $\mathbf{p}_d$ *as follows:*

$$\mathbf{p}_n = \{p_i \mid \forall i \in \mathcal{N}\} \tag{4.2a}$$
$$\mathbf{p}_e = \{p_{ij} \mid \forall (i,j) \in \mathcal{E}\} \tag{4.2b}$$
$$\mathbf{p}_d = \{p_{ij} \mid \forall (i,j) \in \overrightarrow{\mathcal{E}}\} \tag{4.2c}$$

(the subscripts "n", "e" and "d" stand for nodes, edges and directed edges). The terms $\mathbf{p}_n$, $\mathbf{p}_e$ *and* $\mathbf{p}_d$ *are referred to as* injection vector, flow vector *and* semi-flow vector, *respectively (note that* $\mathbf{p}_e$ *contains two flows per each line, whereas* $\mathbf{p}_d$ *includes only one flow per line).*

Definition 17. *Given two arbitrary points* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, *the box* $\mathcal{B}(\mathbf{x}, \mathbf{y})$ *is defined as follows:*

$$B(\mathbf{x}, \mathbf{y}) = \{\mathbf{z} \in \mathbb{R}^n \mid \mathbf{x} \leq \mathbf{z} \leq \mathbf{y}\} \tag{4.3}$$

(note that $B(\mathbf{x}, \mathbf{y})$ *is non-empty only if* $\mathbf{x} \leq \mathbf{y}$*).*

Assume that each injection p_i and each flow p_{ij} must be within the pre-specified intervals $[p_i^{\min}, p_i^{\max}]$ and $[p_{ij}^{\min}, p_{ij}^{\max}]$, respectively, for every $i \in \mathcal{N}$ and $(i,j) \in \overrightarrow{\mathcal{E}}$. We use the shorthand notation $\mathcal{B}$ for the box $\mathcal{B}(\mathbf{p}_n^{\min}, \mathbf{p}_n^{\max})$, where $\mathbf{p}_n^{\min}$ and $\mathbf{p}_n^{\max}$ are the vectors of the lower bounds $p_i^{\min}$'s and the upper bounds $p_i^{\max}$'s, respectively.

This chapter is concerned with the following problem.

Generalized network flow (GNF) Problem:

$$\min_{\mathbf{p}_n \in \mathcal{B}, \mathbf{p}_e \in \mathbb{R}^{|\mathcal{E}|}} \quad \sum_{i \in \mathcal{N}} f_i(p_i) \tag{4.4a}$$
$$\text{subject to} \quad p_i = \sum_{j \in \mathcal{N}(i)} p_{ij}, \qquad \forall i \in \mathcal{N} \tag{4.4b}$$
$$p_{ji} = f_{ij}(p_{ij}), \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.4c}$$
$$p_{ij} \in [p_{ij}^{\min}, p_{ij}^{\max}], \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.4d}$$

where

1) $f_i(\cdot)$ *is convex and strictly increasing for every* $i \in \mathcal{N}$.

2) $f_{ij}(\cdot)$ *is convex and strictly decreasing for every* $(i,j) \in \overrightarrow{\mathcal{E}}$.

3) *The limits* $p_{ij}^{\min}$ *and* $p_{ij}^{\max}$ *are given for every* $(i,j) \in \overrightarrow{\mathcal{E}}$.

In the case where $f_{ij}(p_{ij})$ is equal to $-p_{ij}$ for all $(i,j) \in \overrightarrow{\mathcal{E}}$, the GNF problem reduces to the network flow problem for which every line is lossless. A few remarks can be made here:

- Given an edge $(i,j) \in \overrightarrow{\mathcal{E}}$, there is no explicit limit on p_{ji} in the above formulation of the GNF problem because any such constraint can be equivalently imposed on p_{ij}.

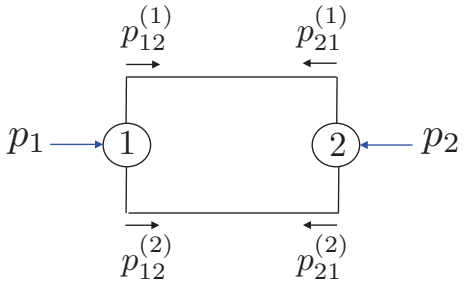


Figure 4.2.1: The graph $\mathcal{G}$ studied in Section 4.3.

- Given a node $i \in \mathcal{N}$, the assumption of $f_i(p_i)$ being monotonically increasing is motivated by the fact that increasing the injection p_i normally elevates the cost in practice.
- Given an edge $(i,j) \in \overrightarrow{\mathcal{E}}$, p_{ij} and $-p_{ji}$ can be regarded as the input and output flows of the line (i,j) traveling in the same direction. The assumption of $f_{ij}(p_{ij})$ being monotonically decreasing is motivated by the fact that increasing the input flow normally makes the output flow higher in practice (note that $-p_{ji} = -f_{ij}(p_{ij})$).

Definition 18. *Define $\mathcal{P}$ as the set of all vectors $\mathbf{p}_n$ for which there exists a vector $\mathbf{p}_e$ such that $(\mathbf{p}_n, \mathbf{p}_e)$ satisfies equations (4.4b), (4.4c) and (4.4d). The set $\mathcal{P}$ and $\mathcal{P} \cap \mathcal{B}$ are referred to as* injection region *and* box-constrained injection region*, respectively.*

Regarding Definition 18, the box-constrained injection region is indeed the projection of the feasible set of GNF onto the space for the injection vector $\mathbf{p}_n$. We express GNF geometrically as follows:

$$\text{Geometric GNF}: \quad \min_{\mathbf{p}_n \in \mathcal{P} \cap \mathcal{B}} \sum_{i \in \mathcal{N}} f_i(p_i) \tag{4.5}$$

Note that $\mathbf{p}_e$ has been eliminated in Geometric GNF. It is hard to solve this problem directly because the injection region $\mathcal{P}$ is non-convex in general. This non-convexity can be observed in Figure 4.2.2(a), which shows $\mathcal{P}$ for the two-node graph drawn in Figure 4.2.1. To address this non-convexity issue, the GNF problem will be convexified next.

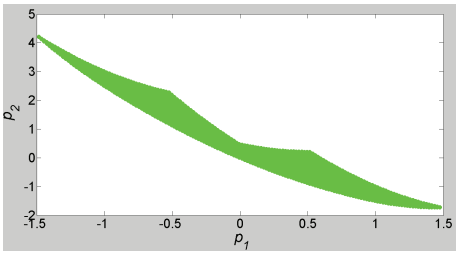
Convexified generalized network flow (CGNF) Problem:

$$\min_{\mathbf{P}_n \in \mathcal{B}, \mathbf{P}_e \in \mathbb{R}^{|\mathcal{E}|}} \quad \sum_{i \in \mathcal{N}} f_i(p_i) \tag{4.6a}$$

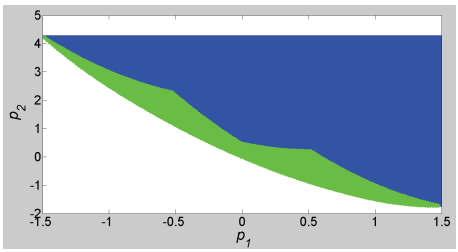
$$\textit{subject to} \quad p_i = \sum_{j \in \mathcal{N}(i)} p_{ij}, \qquad \forall i \in \mathcal{N} \tag{4.6b}$$

$$p_{ji} \geq f_{ij}(p_{ij}), \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.6c}$$

$$p_{ij} \in [p_{ij}^{\min}, p_{ij}^{\max}], \qquad \forall (i,j) \in \mathcal{E} \tag{4.6d}$$



(a) Injection region $\mathcal{P}$ for the GNF problem given in (4.8)



(b) the set $\mathcal{P}_c$ corresponding to the GNF problem given in (4.8)

Figure 4.2.2: The original and convexified injection regions.

where $(p_{ij}^{\min}, p_{ij}^{\max})$ *is defined as* $(f_{ji}(p_{ji}^{\max}), f_{ji}(p_{ji}^{\min}))$ *for every* $(i,j) \in \mathcal{E}$ *such that* $(j,i) \in \overrightarrow{\mathcal{E}}$.

CGNF has been obtained from GNF by relaxing equality (4.4c) to inequality (4.6c) and adding limits to p_{ij} for every $(j,i) \in \overrightarrow{\mathcal{E}}$. One can write:

$$\text{Geometric CGNF}: \quad \min_{\mathbf{p}_n \in \mathcal{P}_c \cap \mathcal{B}} \sum_{i \in \mathcal{N}} f_i(p_i) \tag{4.7}$$

where $\mathcal{P}_c$ denotes the set of all vectors $\mathbf{p}_n$ for which there exists a vector $\mathbf{p}_e$ such that $(\mathbf{p}_n, \mathbf{p}_e)$ satisfies equations (4.6b), (4.6c) and (4.6d). Two main results to be proved in this chapter are:

- **Geometry of injection region:** Given any two points $\mathbf{p}_n$ and $\tilde{\mathbf{p}}_n$ in the injection region, the box $\mathcal{B}(\mathbf{p}_n, \tilde{\mathbf{p}}_n)$ is entirely contained in the injection region. A similar result holds true for the box-constrained injection region.
- **Relationship between GNF and CGNF:** Using the above result on the geometry of the injection region, we show that if $(\mathbf{p}_n^*, \mathbf{p}_e^*)$ and $(\bar{\mathbf{p}}_n^*, \bar{\mathbf{p}}_e^*)$ denote two arbitrary solutions of GNF and CGNF, then $\mathbf{p}_n^* = \bar{\mathbf{p}}_n^*$. Hence, CGNF always finds the correct optimal injection vector for GNF. Moreover, $(\bar{\mathbf{p}}_n^*, \bar{\mathbf{p}}_e^*)$ is a solution of GNF as well if $\mathbf{p}_n^*$ is a Pareto point in $\mathcal{P}$. More generally, if $\mathbf{p}_n^*$ resides on the boundary of $\mathcal{P}$, but is not necessarily a Pareto point, CGNF finds the correct optimal flows for a non-empty subgraph of $\mathcal{G}$.

Furthermore, the above results are generalized to an extended GNF problem, where there are global constraints coupling the flows or injections of different parts of the network. In particular, it is proved that the technique developed for the GNF problem works for the extended GNF problem as well, provided that the coupling constraints are convex and preserve a box-preserving property. Note that this work implicitly assumes that every two nodes of $\mathcal{G}$ are connected via at most one edge. However, the results to be derived later

are all valid in the presence of multiple edges between two nodes. To avoid complicated notations, the proof will not be provided for this case. However, Section 4.3 will analyze a simple example with parallel lines.

In what follows, we first provide a detailed illustrative example to clarify the non-convexity issue and highlight some of the contributions of this chapter. The main results for GNF and CGNF problems are developed in Sections 4.4 and 4.5, respectively. The set of all optimal flow vectors is charaterized in Section 4.6. The generalization to the extended GNF problem is provided in Section 4.7. Finally, the application of the developed methodology in power systems is discussed in Section 4.8.

4.3 Illustrative Example

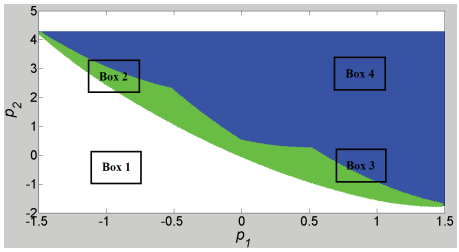
In this subsection, we study the particular graph $\mathcal{G}$ depicted in Figure 4.2.1. This graph has two vertices and two parallel edges. Let $(p_{12}^{(1)}, p_{21}^{(1)})$ and $(p_{12}^{(2)}, p_{21}^{(2)})$ denote the flows associated with the first and second edges of the graph, respectively. Consider the GNF problem

$$\begin{aligned}
\min \quad & f_1(p_1) + f_2(p_2) && \text{(4.8a)}\\
\text{s.t.} \quad & p_{21}^{(i)} = \left(p_{12}^{(i)} - 1\right)^2 - 1, \qquad \forall i \in \{1,2\} && \text{(4.8b)}\\
& -0.5 \le p_{12}^{(1)} \le 0.5, \quad -1 \le p_{12}^{(2)} \le 1, && \text{(4.8c)}\\
& p_1 = p_{12}^{(1)} + p_{12}^{(2)}, \quad p_2 = p_{21}^{(1)} + p_{21}^{(2)} && \text{(4.8d)}
\end{aligned}$$

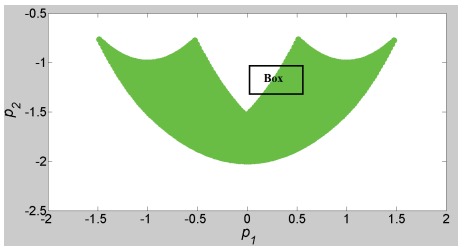
with the variables $p_1, p_2, p_{12}^{(1)}, p_{21}^{(1)}, p_{12}^{(2)}, p_{21}^{(2)}$, where $f_1(\cdot)$ and $f_2(\cdot)$ are arbitrary convex and monotonically increasing functions. The CGNF problem corresponding to this problem can be obtained by replacing (4.8b) with $p_{21}^{(i)} \ge (p_{12}^{(i)} - 1)^2 - 1$ and adding the limits $p_{21}^{(1)} \le 1.5^2 - 1$ and $p_{21}^{(2)} \le 2^2 - 1$. One can write:

$$\begin{aligned}
\text{Geometric GNF:} \quad & \min_{(p_1,p_2)\in\mathcal{P}} f_1(p_1) + f_2(p_2) && \text{(4.9a)}\\
\text{Geometric CGNF:} \quad & \min_{(p_1,p_2)\in\mathcal{P}_c} f_1(p_1) + f_2(p_2) && \text{(4.9b)}
\end{aligned}$$

where $\mathcal{P}$ and $\mathcal{P}_c$ are indeed the projections of the feasible sets of GNF and CGNF over the injection space for (p_1, p_2) (note that there is no box constraint on (p_1, p_2) at this point). The green area in Figure 4.2.2(a) shows the injection region $\mathcal{P}$. As expected, this set is non-convex. In contrast, the set $\mathcal{P}_c$ is a convex set containing $\mathcal{P}$. This set is shown in Figure 4.2.2(b), which includes two parts: (i) the green area that is the same as $\mathcal{P}$, (ii) the blue area that is the part of $\mathcal{P}_c$ that does not exist in $\mathcal{P}$. Thus, the transition from GNF to CGNF extends the injection region $\mathcal{P}$ to a convex set by adding the blue area. Notice that $\mathcal{P}_c$ has three boundaries: (i) a straight line on the top, (ii) a straight line on the right side, (iii) a lower curvy boundary. Since $f_1(\cdot)$ and $f_2(\cdot)$ are both monotonically increasing, the



(a) This figure shows the set $\mathcal{P}_c$ corresponding to the GNF problem given in (4.8) together with a box constraint $(p_1, p_2) \in \mathcal{B}$ for four different positions of $\mathcal{B}$



(b) this figure shows the injection region $\mathcal{P}$ for the GNF problem given in (4.8) but after changing (4.8b) to (4.10)

Figure 4.3.1: The injection regions with box constraints.

unique solution of Geometric CGNF must lie on the lower curvy boundary of $\mathcal{P}_c$. Since this lower boundary is in the green area, it is contained in $\mathcal{P}$. As a result, the unique solution of Geometric CGNF is a feasible point of $\mathcal{P}$ and therefore it is a solution of Geometric GNF. This means that CGNF finds the optimal injection vector for GNF.

To make the problem more interesting, we add the box constraint $(p_1, p_2) \in \mathcal{B}$ to GNF (and correspondingly to CGNF), where $\mathcal{B}$ is an arbitrary rectangular convex set in $\mathbb{R}^2$. The effect of this box constraint will be investigated in four different scenarios:

- Assume that $\mathcal{B}$ corresponds to Box 1 (including its interior) in Figure 4.3.1(a). In this case, $\mathcal{P} \cap \mathcal{B} = \mathcal{P}_c \cap \mathcal{B} = \phi$, implying that Geometric GNF and Geometric CGNF are both infeasible.

- Assume that $\mathcal{B}$ corresponds to Box 2 (including its interior) in Figure 4.3.1(a). In this case, the solution of Geometric CGNF lies on the lower boundary of $\mathcal{P}_c$ and therefore it is also a solution of Geometric GNF.

- Assume that $\mathcal{B}$ corresponds to Box 3 (including its interior) in Figure 4.3.1(a). In this case, the solutions of Geometric GNF and Geometric CGNF are identical and both correspond to the lower left corner of the box $\mathcal{B}$.

- Assume that $\mathcal{B}$ corresponds to Box 4 (including its interior) in Figure 4.3.1(a). In this case, $\mathcal{P} \cap \mathcal{B} = \phi$ but $\mathcal{P}_c \cap \mathcal{B} \neq \phi$. Hence, Geometric GNF is infeasible while Geometric CGNF has an optimal solution.

In summary, it can be argued that, independent of the position of the box $\mathcal{B}$ in $\mathbb{R}^2$, CGNF finds the optimal injection vector for GNF as long as GNF is feasible.

Now, suppose that the relationship between $p_{21}^{(i)}$ and $p_{12}^{(i)}$ is governed by

$$p_{21}^{(i)} = \left(p_{12}^{(i)}\right)^2 - 1, \qquad \forall i \in \{1, 2\} \tag{4.10}$$

instead of (4.8b). The injection region $\mathcal{P}$ in the case is depicted in Figure 4.3.1(b). As before, we impose a box constraint $(p_1, p_2) \in \mathcal{B}$ on GNF, where $\mathcal{B}$ is shown as "Box" in the figure. It is easy to show that the lower left corner of this box belongs to $\mathcal{P}_{\rm c}$ and hence it is a solution of Geometric CGNF. However, this corner point does not belong to Geometric GNF. More precisely, Geometric GNF is feasible in this case, while its solution does not coincide with that of Geometric CGNF. Hence, Geometric GNF and Geometric CGNF are no longer equivalent after changing (4.8b) to (4.10). This is a consequence of the fact that the function $(p-1)^2-1$ is decreasing in p over the interval $[-1, 1]$ while the function p^2-1 is not. This explains the necessity of the assumption of the monotonicity of $f_{ij}(\cdot)$ made earlier in the chapter.

4.4 Geometry of Injection Region

In order to study the relationship between GNF and CGNF, it is beneficial to explore the geometry of the feasible set of GNF. Hence, we investigate the geometry of the injection region $\mathcal{P}$ and the box-constrained injection region $\mathcal{P} \cap \mathcal{B}$ in this part.

Theorem 18. *Consider two arbitrary points $\hat{\mathbf{p}}_n$ and $\tilde{\mathbf{p}}_n$ in the injection region $\mathcal{P}$. The box $\mathcal{B}(\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n)$ is contained in $\mathcal{P}$.*

The proof of this theorem is based on four lemmas, and will be provided later in this subsection. To understand this theorem, consider the injection region $\mathcal{P}$ depicted in Figure 4.2.2(a) corresponding to the illustrative example given in Section 4.3. If any arbitrary box is drawn in $\mathbb{R}^2$ in such a way that its upper right corner and lower left corner both lie in the green area, then the entire box must lie in the green area completely. This can be easily proved in this special case and is true in general due to Theorem 18. However, this result does not hold for the injection region given in Figure 4.3.1(b) because the assumption of monotonicity of $f_{ij}(\cdot)$'s is violated in this case. The result of Theorem 18 can be generalized to the box-constrained injection region, as stated below.

Corollary 5. *Consider two arbitrary points $\hat{\mathbf{p}}_n$ and $\tilde{\mathbf{p}}_n$ belonging to the box-constrained injection region $\mathcal{P} \cap \mathcal{B}$. The box $\mathcal{B}(\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n)$ is contained in $\mathcal{P} \cap \mathcal{B}$.*

Proof: The proof follows immediately from Theorem 18. □

The rest of this subsection is dedicated to the proof of Theorem 18, which is based on a series of definitions and lemmas.

Definition 19. *Define $\mathcal{B}_d$ as the box containing all vectors $\mathbf{p}_d$ introduced in (4.2c) that satisfy the condition $p_{ij} \in [p_{ij}^{\min}, p_{ij}^{\max}]$ for every $(i,j) \in \overrightarrow{\mathcal{E}}$.*

Definition 20. *It is said that $\mathbf{p}_d$ is associated with $\mathbf{p}_n$, or vice versa, if $(\mathbf{p}_n, \mathbf{p}_d)$ is feasible for the GNF problem. Likewise, $\mathbf{p}_e$ is associated with $\mathbf{p}_n$ if $(\mathbf{p}_n, \mathbf{p}_e)$ is feasible for the CGNF problem.*

Definition 21. *Given two arbitrary points $\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d \in \mathcal{B}_d$, define $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ according to the following procedure:*

- *Let $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ be a matrix with $|\mathcal{N}|$ rows indexed by the vertices of $\mathcal{G}$ and with $|\overrightarrow{\mathcal{E}}|$ columns indexed by the edges in $\overrightarrow{\mathcal{E}}$.*
- *For every vertex $k \in \mathcal{N}$ and edge $(i,j) \in \overrightarrow{\mathcal{E}}$, set the $(k,(i,j))^{th}$ entry of $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ (the one in the intersection of row k and column (i,j)) as*

$$\begin{cases} 1 & \text{if} \quad k = i \\ \frac{f_{ij}(\bar{p}_{ij}) - f_{ij}(\tilde{p}_{ij})}{\bar{p}_{ij} - \tilde{p}_{ij}} & \text{if} \quad k = j \text{ and } \bar{p}_{ij} \neq \tilde{p}_{ij} \\ f'_{ij}(\bar{p}_{ij}) & \text{if} \quad k = j \text{ and } \bar{p}_{ij} = \tilde{p}_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

where $f'_{ij}(\bar{p}_{ij})$ denotes the right derivative of $f_{ij}(\bar{p}_{ij})$ if $\bar{p}_{ij} < p_{ij}^{max}$ and the left derivative of $f_{ij}(\bar{p}_{ij})$ if $\bar{p}_{ij} = p_{ij}^{max}$.

To illustrate Definition 21, consider the three-node graph $\overrightarrow{\mathcal{G}}$ depicted in Figure 4.4.1(a). The matrix $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ associated with this graph has the structure shown in Figure 4.4.1(b), where the "*" entries depend on the specific values of $\bar{\mathbf{p}}_d$ and $\tilde{\mathbf{p}}_d$. Consider an edge $(i,j) \in \overrightarrow{\mathcal{E}}$. The $(j,(i,j))^{\text{th}}$ entry of $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ is equal to

$$\frac{f_{ij}(\bar{p}_{ij}) - f_{ij}(\tilde{p}_{ij})}{\bar{p}_{ij} - \tilde{p}_{ij}}, \tag{4.12}$$

provided $\bar{p}_{ij} \neq \tilde{p}_{ij}$. As can be seen in Figure 4.4.1(c), this is equal to the slope of the line connecting the point $(\bar{p}_{ij}, \bar{p}_{ji})$ to the point $(\tilde{p}_{ij}, \tilde{p}_{ji})$ on the parameterized curve (p_{ij}, p_{ji}), where $p_{ji} = f_{ij}(p_{ij})$. Moreover, $f'_{ij}(\bar{p}_{ij})$ is the limit of this slope as the point $(\tilde{p}_{ij}, \tilde{p}_{ji})$ approaches $(\bar{p}_{ij}, \bar{p}_{ji})$. It is also interesting to note that $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ has one positive entry, one negative entry and $m-2$ zero entries in each column (note that the slope of the line connecting $(\bar{p}_{ij}, \bar{p}_{ji})$ to $(\tilde{p}_{ij}, \tilde{p}_{ji})$ is always negative). The next lemma explains how the matrix $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ can be used to relate the semi-flow vector to the injection vector.

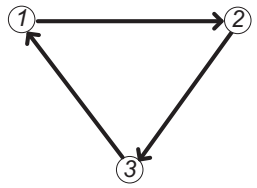
Lemma 26. *Consider two arbitrary injection vectors $\bar{\mathbf{p}}_n$ and $\tilde{\mathbf{p}}_n$ in $\mathcal{P}$, associated with the semi-flow vectors $\bar{\mathbf{p}}_d$ and $\tilde{\mathbf{p}}_d$ (defined in (4.2)). The relation*

$$\bar{\mathbf{p}}_n - \tilde{\mathbf{p}}_n = M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d) \times (\bar{\mathbf{p}}_d - \tilde{\mathbf{p}}_d) \tag{4.13}$$

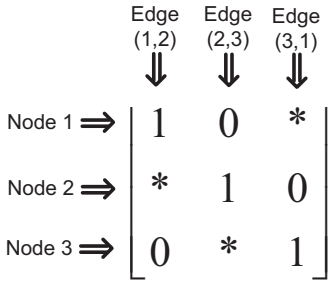
holds.

Proof: One can write

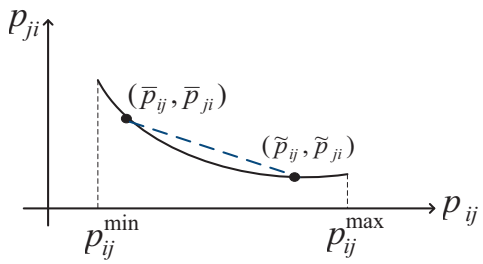
$$\bar{p}_i - \tilde{p}_i = \sum_{j \in \mathcal{N}(i)} (\bar{p}_{ij} - \tilde{p}_{ij}), \qquad \forall i \in \mathcal{N} \tag{4.14}$$



(a) A particular graph $\mathcal{G}$



(b) The matrix $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ corresponding to the graph $\mathcal{G}$ in Figure 4.4.1(a)



(c) The $(j,(i,j))^{\text{th}}$ entry of $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ (shown as "*") is equal to the slope of the line connecting the points $(\bar{p}_{ij}, \bar{p}_{ji})$ and $(\tilde{p}_{ij}, \tilde{p}_{ji})$

Figure 4.4.1: An illustrative example for Definition 21.

By using the relations

$$\bar{p}_{ji} = f_{ij}(\bar{p}_{ij}), \quad \tilde{p}_{ji} = f_{ij}(\tilde{p}_{ij}), \quad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.15}$$

it is straightforward to verify that (4.13) and (4.14) are equivalent. □

Lemma 26 can be regarded as a generalization of the conventional node-edge adjacency matrix used to describe the topology of the graph, which relates semi-flow vectors to injection vectors. The next lemma investigates an important property of the matrix $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$.

Lemma 27. *Given two arbitrary points $\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d \in \mathcal{B}_d$, assume that there exists a nonzero vector $\mathbf{x} \in \mathbb{R}^m$ such that $\mathbf{x}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d) \geq 0$. If $\mathbf{x}$ has at least one strictly positive entry, then there exists a nonzero vector $\mathbf{y} \in \mathbb{R}^m_+$ such that $\mathbf{y}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d) \geq 0$.*

Proof: Consider an index $i_0 \in \mathcal{N}$ such that $x_{i_0} > 0$. Define $\mathcal{V}(i_0)$ as the set of all vertices $i \in \mathcal{N}$ from which there exists a directed path to vertex i_0 in the graph $\overrightarrow{\mathcal{G}}$. Note that $\mathcal{V}(i_0)$ includes vertex i_0 itself. The first goal is to show that

$$x_i \geq 0, \qquad \forall i \in \mathcal{V}(i_0) \tag{4.16}$$

To this end, consider an arbitrary set of vertices $i_1, ..., i_k$ in $\mathcal{V}(i_0)$ such that $\{i_0, i_1..., i_k\}$ forms a direct path in $\overrightarrow{\mathcal{G}}$ as

$$i_k \to i_{k-1} \to \cdots i_1 \to i_0 \tag{4.17}$$

To prove (4.16), it suffices to show that $x_{i_1}, ..., x_{i_k} \geq 0$. For this purpose, one can expand the product $\mathbf{x}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ and use the fact that each column of $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ has $m-2$ zero entries to conclude that

$$x_{i_1} + \frac{f_{i_1 i_0}(\bar{p}_{i_1 i_0}) - f_{i_1 i_0}(\tilde{p}_{i_1 i_0})}{\bar{p}_{i_1 i_0} - \tilde{p}_{i_1 i_0}} x_{i_0} \geq 0 \tag{4.18}$$

Since x_{i_0} is positive and $f_{i_1 i_0}(\cdot)$ is a decreasing function, x_{i_1} turns out to be positive. Now, repeating the above argument for i_1 instead of i_0 yields that $x_{i_2} \geq 0$. Continuing this reasoning leads to $x_{i_1}, ..., x_{i_k} \geq 0$. Hence, inequality (4.16) holds. Now, define $\mathbf{y}$ as

$$y_i = \begin{cases} x_i & \text{if} \quad i \in \mathcal{V}(i_0) \\ 0 & \text{otherwise} \end{cases}, \qquad \forall i \in \mathcal{N} \tag{4.19}$$

In light of (4.16), $\mathbf{y}$ is a nonzero vector in $\mathbb{R}^m_+$. To complete the proof, it suffices to show that $\mathbf{y}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d) \geq 0$. Similar to the indexing procedure used for the columns of the matrix $M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$, we index the entries of the $|\overrightarrow{\mathcal{E}}|$ dimensional vector $\mathbf{y}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ according to the edges of $\overrightarrow{\mathcal{G}}$. Now, given an arbitrary edge $(\alpha, \beta) \in \overrightarrow{\mathcal{E}}$, the following statements hold true:

- If $\alpha, \beta \in \mathcal{V}(i_0)$, then the $(\alpha, \beta)^{\text{th}}$ entries of $\mathbf{y}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ and $\mathbf{x}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ (i.e., the entries corresponding to the edge (α, β)) are identical.
- If $\alpha \in \mathcal{V}(i_0)$ and $\beta \notin \mathcal{V}(i_0)$, then the $(\alpha, \beta)^{\text{th}}$ entry of $\mathbf{y}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ is equal to y_α.
- If $\alpha \notin \mathcal{V}(i_0)$ and $\beta \notin \mathcal{V}(i_0)$, then the $(\alpha, \beta)^{\text{th}}$ entry of $\mathbf{y}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d)$ is equal to zero.

Note that the case $\alpha \notin \mathcal{V}(i_0)$ and $\beta \in \mathcal{V}(i_0)$ cannot happen, because if $\beta \in \mathcal{V}(i_0)$ and $(\alpha, \beta) \in \overrightarrow{\mathcal{E}}$, then $\alpha \in \mathcal{V}(i_0)$ by the definition of $\mathcal{V}(i_0)$. It follows from the above results and the inequality $\mathbf{x}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d) \geq 0$ that $\mathbf{y}^T M(\bar{\mathbf{p}}_d, \tilde{\mathbf{p}}_d) \geq 0$. □

Definition 22. *Consider the graph $\mathcal{G}$ and an arbitrary flow vector $\mathbf{p}_e$. Given a subgraph $\mathcal{G}_s$ of the graph $\mathcal{G}$, define $\mathbf{p}_e(\mathcal{G}_s)$ as the flow vector associated with the edges of $\mathcal{G}_s$ that has been induced by $\mathbf{p}_e$. Define $\mathbf{p}_d(\mathcal{G}_s)$, $\mathbf{p}_n(\mathcal{G}_s)$ and $p_i(\mathcal{G}_s)$ as the semi-flow vector, injection vector and injection at node $i \in \mathcal{G}_s$ corresponding to $\mathbf{p}_e(\mathcal{G}_s)$, respectively. Define also $\mathcal{P}(\mathcal{G}_s)$ as the injection region associated with $\mathcal{G}_s$.*

The next lemma studies the injection region $\mathcal{P}$ in the case where $f_{ij}(\cdot)$'s are all piecewise linear.

Lemma 28. *Assume that the function $f_{ij}(\cdot)$ is piecewise linear for every $(i, j) \in \overrightarrow{\mathcal{E}}$. Consider two arbitrary points $\hat{\mathbf{p}}_n, \bar{\mathbf{p}}_n \in \mathcal{P}$ and a vector $\Delta\bar{\mathbf{p}}_n \in \mathbb{R}^m$ satisfying the relations*

$$\hat{\mathbf{p}}_n \leq \bar{\mathbf{p}}_n - \Delta\bar{\mathbf{p}}_n \leq \bar{\mathbf{p}}_n \tag{4.20}$$

There exists a strictly positive number ϵ^{max} with the property

$$\bar{\mathbf{p}}_n - \varepsilon\Delta\bar{\mathbf{p}}_n \in \mathcal{P}, \qquad \forall\varepsilon \in [0, \epsilon^{max}] \tag{4.21}$$

Proof: In light of (4.20), we have $\Delta\bar{\mathbf{p}}_n \geq 0$. If $\Delta\bar{\mathbf{p}}_n = 0$, then the lemma becomes trivial as ε can take any arbitrary value. So, assume that $\Delta\bar{\mathbf{p}}_n \neq 0$. Let $\hat{\mathbf{p}}_e$ and $\bar{\mathbf{p}}_e$ denote two flow vectors associated with the injection vectors $\hat{\mathbf{p}}_n$ and $\bar{\mathbf{p}}_n$, respectively. Denote the corresponding semi-flow vectors as $\hat{\mathbf{p}}_d$ and $\bar{\mathbf{p}}_d$. Given an edge $(i,j) \in \overrightarrow{\mathcal{E}}$, the curve

$$\big\{(p_{ij}, f_{ij}(p_{ij})) \mid p_{ij} \in [p_{ij}^{\min}, p_{ij}^{\max}]\big\} \tag{4.22}$$

is a Pareto set in $\mathbb{R}^2$ due to $f_{ij}(\cdot)$ being monotonically decreasing. Since $(\hat{p}_{ij}, \hat{p}_{ji})$ and $(\bar{p}_{ij}, \bar{p}_{ji})$ both lie on the above curve, one of the following cases occurs:

- *Case 1:* $\hat{p}_{ij} \geq \bar{p}_{ij}$ and $\hat{p}_{ji} \leq \bar{p}_{ji}$.
- *Case 2:* $\hat{p}_{ij} \leq \bar{p}_{ij}$ and $\hat{p}_{ji} \geq \bar{p}_{ji}$.

(this fact can be observed in Figure 4.4.1(c) for the points $(\bar{p}_{ij}, \bar{p}_{ji})$ and $(\tilde{p}_{ij}, \tilde{p}_{ji})$ instead of $(\hat{p}_{ij}, \hat{p}_{ji})$ and $(\bar{p}_{ij}, \bar{p}_{ji})$). With no loss of generality, assume that Case 1 occurs. Indeed, if Case 2 happens, it suffices to make two changes:

- Change the orientation of the edge (i,j) in the graph $\overrightarrow{\mathcal{G}}$ so that $(j,i) \in \overrightarrow{\mathcal{E}}$ instead of $(i,j) \in \overrightarrow{\mathcal{E}}$.
- Replace the constraint $p_{ji} = f_{ij}(p_{ij})$ in (4.4c) with $p_{ij} = f_{ij}^{-1}(p_{ji})$, where the existence, monotonicity and convexity of the inverse function $f_{ij}^{-1}(\cdot)$ is guaranteed by the convexity and decreasing property of $f_{ij}(\cdot)$.

Therefore, suppose that

$$\hat{p}_{ij} \geq \bar{p}_{ij}, \quad \hat{p}_{ji} \leq \bar{p}_{ji}, \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.23}$$

or

$$\hat{\mathbf{p}}_d \geq \bar{\mathbf{p}}_d \tag{4.24}$$

First, consider the case $\hat{\mathbf{p}}_d > \bar{\mathbf{p}}_d$. In light of Lemma 26, the assumption $\hat{\mathbf{p}}_n \leq \bar{\mathbf{p}}_n$ can be expressed as

$$M(\hat{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \times (\hat{\mathbf{p}}_d - \bar{\mathbf{p}}_d) = \hat{\mathbf{p}}_n - \bar{\mathbf{p}}_n \leq 0 \tag{4.25}$$

In order to guarantee the relation $\bar{\mathbf{p}}_n - \varepsilon\Delta\bar{\mathbf{p}}_n \in \mathcal{P}$, it suffices to seek a vector $\Delta\bar{\mathbf{p}}_d \in \mathbb{R}^{|\overrightarrow{\mathcal{E}}|}$ satisfying the equations

$$\bar{\mathbf{p}}_d - \varepsilon\Delta\bar{\mathbf{p}}_d \in \mathcal{B}_d \tag{4.26}$$

and

$$M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d - \varepsilon\Delta\bar{\mathbf{p}}_d) \times (\bar{\mathbf{p}}_d - (\bar{\mathbf{p}}_d - \varepsilon\Delta\bar{\mathbf{p}}_d)) = \bar{\mathbf{p}}_n - (\bar{\mathbf{p}}_n - \varepsilon\Delta\bar{\mathbf{p}}_n) \tag{4.27}$$

(see the proof of Lemma 26), or equivalently

$$\bar{\mathbf{p}}_d - \varepsilon\Delta\bar{\mathbf{p}}_d \in \mathcal{B}_d \tag{4.28a}$$
$$M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d - \varepsilon\Delta\bar{\mathbf{p}}_d) \times \Delta\bar{\mathbf{p}}_d = \Delta\bar{\mathbf{p}}_n \tag{4.28b}$$

Consider an arbitrary vector $\Delta\bar{\mathbf{p}}_d \in \mathbb{R}^{|\overrightarrow{\mathcal{E}}|}$ with all negative entries. In light of Definition 21, the inequality $\hat{\mathbf{p}}_d > \bar{\mathbf{p}}_d$ and the piecewise linear property of $f_{ij}(\cdot)$'s, there exists a positive number $\varepsilon^{\max}$ such that

$$\bar{\mathbf{p}}_d - \varepsilon\Delta\bar{\mathbf{p}}_d \in \mathcal{B}_d \tag{4.29a}$$
$$M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d - \varepsilon\Delta\bar{\mathbf{p}}_d) = M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \tag{4.29b}$$

for every $\varepsilon \in [0, \varepsilon^{\max}]$. To prove the lemma, it follows from (4.28) and (4.29) that it is enough to show the existence of a negative vector $\Delta\bar{\mathbf{p}}_d$ satisfying the relation

$$M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \times \Delta\bar{\mathbf{p}}_d = \Delta\bar{\mathbf{p}}_n \tag{4.30}$$

in which ε does not appear. Notice that since (4.30) is independent of ϵ, it can be chosen sufficiently small so that (4.29a) is satisfied automatically. To prove this by contradiction, assume that the above equation does not have a solution. By Farkas' Lemma, there exists a vector $\mathbf{x} \in \mathbb{R}^m$ such that

$$\mathbf{x}^T M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \geq 0, \qquad \mathbf{x}^T\Delta\bar{\mathbf{p}}_n > 0 \tag{4.31}$$

Since $\Delta\bar{\mathbf{p}}_n$ is nonnegative, the inequality $\mathbf{x}^T\Delta\bar{\mathbf{p}}_n > 0$ does not hold unless $\mathbf{x}$ has at least one strictly positive entry. Now, it follows from $\mathbf{x}^T M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \geq 0$ and Lemma 27 that there exists a nonzero vector $\mathbf{y} \in \mathbb{R}^m$ such that

$$\mathbf{y}^T M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \geq 0, \qquad \mathbf{y} \geq 0 \tag{4.32}$$

On the other hand, given an edge $(i,j) \in \overrightarrow{\mathcal{E}}$, since $\hat{p}_{ij} \geq \bar{p}_{ij}$ (due to (4.23)), the slope of the line connecting the points $(\hat{p}_{ij}, \hat{p}_{ji})$ and $(\bar{p}_{ij}, \bar{p}_{ji})$ is more than or equal to $f'_{ij}(\bar{p}_{ij})$ (this is implied by the fact that $f_{ij}(\cdot)$ is convex). This yields that

$$M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \leq M(\hat{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \tag{4.33}$$

Now, it follows from (4.24), (4.25), (4.32) and (4.33) that

$$0 \geq \mathbf{y}^T M(\hat{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \times (\hat{\mathbf{p}}_d - \bar{\mathbf{p}}_d) \geq \mathbf{y}^T M(\bar{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \times (\hat{\mathbf{p}}_d - \bar{\mathbf{p}}_d) \geq 0 \tag{4.34}$$

Thus,

$$0 = \mathbf{y}^T M(\hat{\mathbf{p}}_d, \bar{\mathbf{p}}_d) \times (\hat{\mathbf{p}}_d - \bar{\mathbf{p}}_d) = \mathbf{y}^T(\hat{\mathbf{p}}_n - \bar{\mathbf{p}}_n) \tag{4.35}$$

This is a contradiction because $\hat{\mathbf{p}}_n - \bar{\mathbf{p}}_n$ is strictly negative and the nonzero vector $\mathbf{y}$ is positive.

So far, the lemma has been proven in the case when $\hat{\mathbf{p}}_d > \bar{\mathbf{p}}_d$. To extend the proof to the case $\hat{\mathbf{p}}_d \geq \bar{\mathbf{p}}_d$, define $\mathcal{E}_{\rm r}$ as the set of every edge $(i,j) \in \mathcal{E}$ such that

$$\hat{p}_{ij} \neq \bar{p}_{ij} \tag{4.36}$$

(note that $\hat{p}_{ij} = \bar{p}_{ij}$ if and only if $\hat{p}_{ji} = \bar{p}_{ji}$). Define also $\mathcal{G}_{\rm r}$ as the unique subgraph of $\mathcal{G}$ induced by the edge set $\mathcal{E}_{\rm r}$. Let $\mathcal{N}_{\rm r}$ denote the vertex set of $\mathcal{G}_{\rm r}$, which may be different from $\mathcal{N}$. It is easy to verify that

$$\hat{\mathbf{p}}_d(\mathcal{G}_{\rm r}) > \bar{\mathbf{p}}_d(\mathcal{G}_{\rm r}), \tag{4.37a}$$
$$\hat{\mathbf{p}}_n(\mathcal{G}_{\rm r}) \leq \bar{\mathbf{p}}_n(\mathcal{G}_{\rm r}) - \Delta\bar{\mathbf{p}}_n(\mathcal{G}_{\rm r}) \leq \bar{\mathbf{p}}_n(\mathcal{G}_{\rm r}) \tag{4.37b}$$
$$\bar{p}_i - \hat{p}_i = \bar{p}_i(\mathcal{G}_{\rm r}) - \hat{p}_i(\mathcal{G}_{\rm r}), \qquad \forall i \in \mathcal{N}_{\rm r} \tag{4.37c}$$

Based on (4.37c), the relationship between $\Delta\bar{\mathbf{p}}_n$ and the new vector $\Delta\bar{\mathbf{p}}_n(\mathcal{G}_{\rm r})$ is as follows:

$$\Delta\bar{p}_i = \left\{ \begin{array}{cc} \Delta\bar{p}_i(\mathcal{G}_{\rm r}) & \text{if} \quad i \in \mathcal{N}_{\rm r} \\ 0 & \text{otherwise} \end{array} \right. , \qquad \forall i \in \mathcal{N} \tag{4.38}$$

In light of (4.37a) and (5.11), one can adopt the proof given earlier for the case $\hat{\mathbf{p}}_d > \bar{\mathbf{p}}_d$ to conclude the existence of a positive number $\epsilon^{\max}$ with the property

$$\bar{\mathbf{p}}_n(\mathcal{G}_{\rm r}) - \varepsilon\Delta\bar{\mathbf{p}}_n(\mathcal{G}_{\rm r}) \in \mathcal{P}(\mathcal{G}_{\rm r}), \qquad \forall \varepsilon \in [0, \epsilon^{\max}] \tag{4.39}$$

Given an arbitrary number $\varepsilon \in [0, \epsilon^{\max}]$, we use the shorthand notation $\mathbf{p}_n(\mathcal{G}_{\rm r})$ and $\mathbf{p}_n$ for $\bar{\mathbf{p}}_n(\mathcal{G}_{\rm r}) - \varepsilon\Delta\bar{\mathbf{p}}_n(\mathcal{G}_{\rm r})$ and $\bar{\mathbf{p}}_n - \varepsilon\Delta\bar{\mathbf{p}}_n$, respectively. Let $\mathbf{p}_e(\mathcal{G}_{\rm r})$ and $\mathbf{p}_e$ denote a flow vector corresponding to the injection vectors $\mathbf{p}_n(\mathcal{G}_{\rm r})$ and $\mathbf{p}_n$, respectively. One can expand the vector $\mathbf{p}_e(\mathcal{G}_{\rm r})$ into $\mathbf{p}_e$ for the graph $\mathcal{G}$ as follows:

- For every $(i,j) \in \mathcal{E}_{\rm r}$, the $(i,j)^{\rm th}$ entries of $\mathbf{p}_e$ and $\mathbf{p}_e(\mathcal{G}_{\rm r})$ (the ones corresponding to the edge (i,j)) are identical.
- For every $(i,j) \in \mathcal{E}\backslash\mathcal{E}_{\rm r}$, the $(i,j)^{\rm th}$ entry of $\mathbf{p}_e$ is equal to $\bar{p}_{ij}$ (or $\hat{p}_{ij}$).

It is straightforward to observe that $\mathbf{p}_n$ is associated with the designed vector $\mathbf{p}_e$ and, therefore, the feasibility of $\mathbf{p}_e$ implies that $\mathbf{p}_n$ belongs to $\mathcal{P}$. This completes the proof. □

The next lemma uses Lemma 28 to prove Theorem 18 in the case where $f_{ij}(\cdot)$'s are all piecewise linear.

Lemma 29. *Assume that the function $f_{ij}(\cdot)$ is piecewise linear for every $(i,j) \in \overrightarrow{\mathcal{E}}$. Given any two arbitrary points $\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n \in \mathcal{P}$, the box $\mathcal{B}(\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n)$ is a subset of the injection region $\mathcal{P}$.*

Proof: With no loss of generality, assume that $\hat{\mathbf{p}}_n \leq \tilde{\mathbf{p}}_n$ (because otherwise $\mathcal{B}(\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n)$ is empty). To prove the lemma by contradiction, suppose that there exists a point $\mathbf{p}_n \in \mathcal{B}(\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n)$ such that $\mathbf{p}_n \notin \mathcal{P}$. Consider the set

$$\left\{ \gamma \;\middle|\; \gamma \in [0,1],\ \tilde{\mathbf{p}}_n + \gamma(\mathbf{p}_n - \tilde{\mathbf{p}}_n) \in \mathcal{P} \right\} \tag{4.40}$$

Note that $\hat{\mathbf{p}}_n \leq \mathbf{p}_n \leq \tilde{\mathbf{p}}_n$, and that (4.40) describes the set of all γ's for which $\tilde{\mathbf{p}}_n + \gamma(\mathbf{p}_n - \tilde{\mathbf{p}}_n)$ belongs to the segment between $\mathbf{p}_n$ and $\tilde{\mathbf{p}}_n$. Denote the maximum of all those γ as $\gamma^{\max}$. The existence of this number is guaranteed because of two reasons: (1) when γ is equal to 0, the point $\tilde{\mathbf{p}}_n + \gamma(\mathbf{p}_n - \tilde{\mathbf{p}}_n)$ is equal to $\tilde{\mathbf{p}}_n$ and belongs to $\mathcal{P}$, (2) $\mathcal{P}$ is closed and compact. Furthermore, notice that $\tilde{\mathbf{p}}_n + \gamma(\mathbf{p}_n - \tilde{\mathbf{p}}_n)$ is equal to $\mathbf{p}_n$ at $\gamma = 1$. Since $\mathbf{p}_n \notin \mathcal{P}$ by assumption, we have $\gamma^{\max} < 1$. Denote $\tilde{\mathbf{p}}_n + \gamma^{\max}(\mathbf{p}_n - \tilde{\mathbf{p}}_n)$ as $\bar{\mathbf{p}}_n$. Hence, $\bar{\mathbf{p}}_n \in \mathcal{P}$ and $\hat{\mathbf{p}}_n \leq \mathbf{p}_n \leq \bar{\mathbf{p}}_n$ (recall that $\gamma^{\max} < 1$). Define $\Delta\bar{\mathbf{p}}_n$ as $\bar{\mathbf{p}}_n - \mathbf{p}_n$. One can write:

$$\hat{\mathbf{p}}_n \leq \bar{\mathbf{p}}_n - \Delta\bar{\mathbf{p}}_n \leq \bar{\mathbf{p}}_n, \qquad \hat{\mathbf{p}}_n, \bar{\mathbf{p}}_n \in \mathcal{P} \tag{4.41}$$

By Lemma 23, there exists a strictly positive number $\epsilon^{\max}$ with the property

$$\bar{\mathbf{p}}_n - \varepsilon\Delta\bar{\mathbf{p}}_n \in \mathcal{P}, \qquad \forall \varepsilon \in [0, \epsilon^{\max}] \tag{4.42}$$

or equivalently

$$\tilde{\mathbf{p}}_n + (\gamma^{\max} + \varepsilon(1 - \gamma^{\max}))(\mathbf{p}_n - \tilde{\mathbf{p}}_n) \in \mathcal{P}, \quad \forall \varepsilon \in [0, \epsilon^{\max}] \tag{4.43}$$

Notice that

$$\gamma^{\max} + \varepsilon(1 - \gamma^{\max}) > \gamma^{\max}, \quad \forall \varepsilon > 0 \tag{4.44}$$

Due to (4.43), this violates the assumption that $\gamma^{\max}$ is the maximum of the set given in (4.40). □

Lemma 29 will be deployed next to prove Theorem 18 in the general case.

Proof of Theorem 18: Consider an arbitrary approximation of $f_{ij}(\cdot)$ by a piecewise linear function for every $(i, j) \in \mathcal{E}$. As a counterpart of $\mathcal{P}$, let $\mathcal{P}_s$ denote the injection region in the piecewise-linear case. By Lemma 29, we have

$$\mathcal{B}(\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n) \subseteq \mathcal{P}_s \tag{4.45}$$

Since the piecewise linear approximation can be made in such a way that the sets $\mathcal{P}$ and $\mathcal{P}_s$ become arbitrarily close to each other, the above relation implies that the interior of $\mathcal{B}(\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n)$ is a subset of $\mathcal{P}$. On the other hand, $\mathcal{P}$ is a closed set. Hence, the box $\mathcal{B}(\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n)$ must entirely belong to $\mathcal{P}$. □

4.5 Convexified Generalized Network Flow

Using Theorem 18 developed in the preceding subsection, we study the relationship between GNF and CGNF below.

Definition 23. *Consider an arbitrary set $\mathcal{S} \in \mathbb{R}^n$ together with a point $\mathbf{x} \in \mathcal{S}$. The point $\mathbf{x}$ is called "Pareto" if there does not exist another point $\mathbf{y} \in \mathcal{S}$ that is less than or equal to $\mathbf{x}$ entry-wise. $\mathbf{x} \in \mathcal{S}$ is called an "interior point" if $\mathcal{S}$ contains a ball around this point. $\mathbf{x} \in \mathcal{S}$ is called a "boundary point" if it is not an interior point.*

To proceed with the results, the following mild assumption is required.

Assumption 2. *There exists a feasible point* $(\mathbf{p}_n, \mathbf{p}_e)$ *for the CGNF problem such that* $p_{ij} > p_{ij}^{min}$ *for every* $(i,j) \in \overrightarrow{\mathcal{E}}$ *and* $p_i < p_i^{max}$ *for every* $i \in \mathcal{N}$.

Theorem 19. *Assume that the GNF problem is feasible. Let* $(\mathbf{p}_n^*, \mathbf{p}_e^*)$ *and* $(\bar{\mathbf{p}}_n^*, \bar{\mathbf{p}}_e^*)$ *denote arbitrary globally optimal solutions of GNF and CGNF, respectively. The following relations hold:*

1) $\mathbf{p}_n^* = \bar{\mathbf{p}}_n^*$

2) $(\bar{\mathbf{p}}_n^*, \bar{\mathbf{p}}_e^*)$ *is a solution of GNF, provided that* $\mathbf{p}_n^*$ *is a Pareto point in* $\mathcal{P}$.

□

In what follows, we first prove Part 2 of Theorem 19 and illustrate it in some examples before proving Part 1.

Proof of Part 2 of Theorem 19: Define a new flow vector $\hat{\mathbf{p}}_e$ as

$$\hat{p}_{ij} = \bar{p}_{ij}^*, \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.46a}$$

$$\hat{p}_{ji} = f_{ij}(\bar{p}_{ij}^*), \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.46b}$$

Let $\hat{\mathbf{p}}_n$ denote the injection vector corresponding to $\hat{\mathbf{p}}_e$. Since $\hat{p}_{ji} = f_{ij}(\bar{p}_{ij}^*)$ for every $(i,j) \in \overrightarrow{\mathcal{E}}$, it can be concluded that $\hat{\mathbf{p}}_n \leq \bar{\mathbf{p}}_n^* = \mathbf{p}_n^*$ (the last equality follows from Part 1 of the theorem). Since $\mathbf{p}_n^*$ is assumed to be a Pareto point in $\mathcal{P}$, we must have $\hat{\mathbf{p}}_n = \bar{\mathbf{p}}_n^*$ and therefore $\hat{\mathbf{p}}_e = \bar{\mathbf{p}}_e^*$. This implies that $(\bar{\mathbf{p}}_n^*, \bar{\mathbf{p}}_e^*)$ is a feasible point for GNF and yet a global solution for CGNF. As a result, $(\bar{\mathbf{p}}_n^*, \bar{\mathbf{p}}_e^*)$ is a solution of GNF. □

Theorem 19 states that CGNF finds the optimal injections but not necessarily optimal flows for GNF. Note that Part 1 of the theorem implies that the globally optimal injection vector is unique. Two examples will be provided below to elaborate on Part 2 of Theorem 19.

Example 1: Consider the illustrative example explained in Section 4.3. It can be observed in Figure 4.2.2(b) that every point on the lower curvy boundary of the feasible set is a Pareto point. Therefore, if the box $\mathcal{B}$ defined by the lower and upper bound constraints on p_1 and p_2 intersects with any part of the lower boundary of the green area, CGNF always finds optimal flow vectors for GNF, leading to the equivalence of GNF and CGNF. □

Example 2: As stated before, a Pareto point lies on the boundary of the injection region. A question arises as to whether the condition "Pareto point" can be replaced by "boundary point" in Theorem 19. We will provide an example here to show that the optimal injection being a boundary point does not necessarily guarantee the equivalence of GNF and CGNF. To this end, consider the 4-node graph $\mathcal{G}$ depicted in Figure 4.5.1. This graph can be decomposed into two subgraphs $\mathcal{G}_1$ and $\mathcal{G}_2$, where each subgraph has the same topology as the 2-node graph studied in Example 1. Assume that the flow over the line $(2,3)$ is restricted

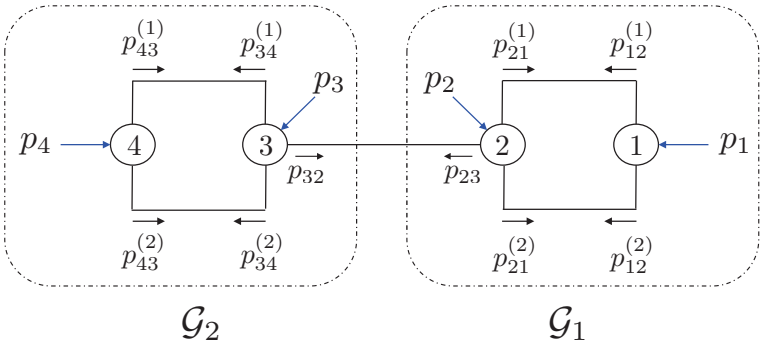
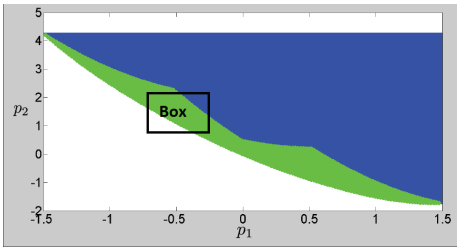
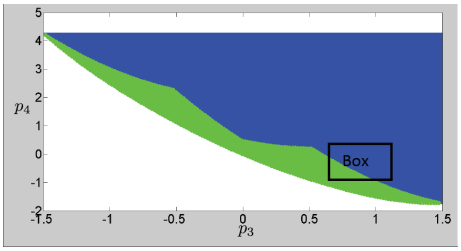


Figure 4.5.1: The 4-node graph $\mathcal{G}$ studied in Example 2.



(a) Injection region of the subgraph $\mathcal{G}_1$



(b) Injection region of the subgraph $\mathcal{G}_2$

Figure 4.5.2: The injection regions and box constraints in Example 2.

to zero, by imposing the constraints $p_{23}^{\min} = p_{23}^{\max} = p_{32}^{\min} = p_{32}^{\max} = 0$. This implies that $(2, 3)$ is redundant, whose removal splits the graph $\mathcal{G}$ into two disjoint subgraphs $\mathcal{G}_1$ and $\mathcal{G}_2$. Let $(\mathbf{p}_n^*, \mathbf{p}_e^*)$ be an arbitrary solution of GNF. The vector $\mathbf{p}_n^*$ can be broken down into two parts as

$$\mathbf{p}_n^* = [\mathbf{p}_n^*(\mathcal{G}_1)^T \quad \mathbf{p}_n^*(\mathcal{G}_2)^T]^T \tag{4.47}$$

where $\mathbf{p}_n^*(\mathcal{G}_1)$ and $\mathbf{p}_n^*(\mathcal{G}_2)$ denote the optimal values of the sub-vectors $[p_1 \;\; p_2]^T$ and $[p_3 \;\; p_4]^T$, respectively. Note that $\mathcal{P}(\mathcal{G}_1)$ and $\mathcal{P}(\mathcal{G}_2)$ could both resemble the green area in Figure 4.2.2(b). We make two assumptions here:

- *Assumption 1:* As demonstrated in Figure 4.5.2(a), the box constraints on p_1 and p_2 are such that $\mathbf{p}_n^*(\mathcal{G}_1)$ becomes a Pareto point located on the lower boundary of $\mathcal{P}(\mathcal{G}_1)$. In this case, it is guaranteed from Theorem 19 that if CGNF is solved just over $\mathcal{G}_1$, it finds feasible flows for this subgraph.

- *Assumption 2:* As demonstrated in Figure 4.5.2(b), the box constraints on p_3 and p_4 are such that $\mathbf{p}_n^*(\mathcal{G}_2)$ becomes an interior point of $\mathcal{P}(\mathcal{G}_2)$, corresponding to the lower

left corner of the box. In this case, assume that if CGNF is solved just over $\mathcal{G}_2$, it may not always find feasible flows for this subgraph (we will show it later in the chapter).

Since $(2,3)$ is not allowed to carry any flow, it is easy to show that CGNF solved over $\mathcal{G}$ finds feasible flows for the lines between nodes 1 and 2, but may result in wrong flows for the lines between nodes 3 and 4. Hence, CGNF and GNF are not equivalent. On the other hand, it is straightforward to inspect that $\mathcal{P}$ is the product of two regions as

$$\mathcal{P} = \mathcal{P}(\mathcal{G}_1) \times \mathcal{P}(\mathcal{G}_2) \tag{4.48}$$

Now, since $\mathbf{p}_n^*(\mathcal{G}_1)$ is on the boundary of $\mathcal{P}(\mathcal{G}_1)$ but $\mathbf{p}_n^*(\mathcal{G}_2)$ is in the interior of $\mathcal{P}(\mathcal{G}_2)$, it can be deduced that

- $\mathbf{p}_n^*$ is on the boundary of the injection region $\mathcal{P}$.
- $\mathbf{p}_n^*$ is not a Pareto point of the injection region $\mathcal{P}$.

In summary, although $\mathbf{p}_n^*$ is a boundary point for $\mathcal{G}$, CGNF is not equivalent to GNF. This is due to the connection of a well-behaved subgraph $\mathcal{G}_1$ to a problematic subgraph $\mathcal{G}_2$ via a redundant link with no flow. It will be shown in Corollary 6 that whenever $\mathbf{p}_n^*$ is on the boundary of its injection region, there exists a non-empty subgraph of $\mathcal{G}$ for which the correct (feasible and optimal) flows can be found via CGNF. □

Before presenting the proof of Part 1 of Theorem 19 in the general case, one special case will be studied for which the proof is less involved. Observe that since $(\bar{\mathbf{p}}_n^*, \bar{\mathbf{p}}_e^*)$ is a feasible point of CGNF, one can write

$$\bar{p}_i^* \geq p_i^{\min}, \qquad \forall i \in \mathcal{N} \tag{4.49}$$

The proof of Part 1 of Theorem 19 will be first derived in the special case

$$\bar{p}_i^* = p_i^{\min}, \qquad \forall i \in \mathcal{N} \tag{4.50}$$

Proof of Part 1 of Theorem 19 under Condition (4.50): $(\mathbf{p}_n^*, \mathbf{p}_e^*)$ being a feasible point of GNF implies that

$$p_i^* \geq p_i^{\min}, \qquad \forall i \in \mathcal{N} \tag{4.51}$$

Equations (4.50) and (4.51) lead to

$$\bar{\mathbf{p}}_n^* \leq \mathbf{p}_n^* \tag{4.52}$$

Define the vector $\tilde{\mathbf{p}}_n$ as

$$\tilde{p}_i = \sum_{(i,j)\in\vec{\mathcal{E}}} \bar{p}_{ij}^* + \sum_{(j,i)\in\vec{\mathcal{E}}} f_{ij}(\bar{p}_{ij}^*), \qquad \forall i \in \mathcal{N} \tag{4.53}$$

Notice that $\tilde{\mathbf{p}}_n$ belongs to $\mathcal{P}$, although it may not belong to $\mathcal{B}$. It can be inferred from the definition of CGNF that

$$\tilde{\mathbf{p}}_n \leq \bar{\mathbf{p}}_n^* \tag{4.54}$$

Since $\tilde{\mathbf{p}}_n, \mathbf{p}_n^* \in \mathcal{P}$, it follows from Theorem 18, (4.52) and (4.54) that $\bar{\mathbf{p}}_n^* \in \mathcal{P}$. On the other hand, $\bar{\mathbf{p}}_n^* \in \mathcal{B}$. Therefore, $\bar{\mathbf{p}}_n^* \in \mathcal{P} \cap \mathcal{B}$, implying that $\bar{\mathbf{p}}_n^*$ is a feasible point of Geometric GNF. Since the feasible set of Geometric CGNF includes that of Geometric GNF, $\bar{\mathbf{p}}_n^*$ must be a solution of Geometric GNF as well. The proof follows from equation (4.52) and the fact that $\mathbf{p}_n^*$ is another solution of Geometric GNF (recall that the objective function of this optimization problem is strictly increasing). □

Before proving Part 1 of Theorem 19 in the general case, some ideas need to be developed. Since $f_i(p_i)$ can be approximated by a differentiable function arbitrarily precisely, with no loss of generality, assume that $f_i(p_i)$ is differentiable for every $i \in \mathcal{N}$. Since CGNF is convex, one can take its Lagrangian dual.

Lemma 30. *Strong duality holds for the CGNF problem.*

Proof: To prove the lemma, it suffices to show that Slater's condition is satisfied or, alternatively, there exists a feasible solution for the CGNF problem satisfying (4.6c) with strict inequality. To this end, consider the feasible solution $(\mathbf{p}_n, \mathbf{p}_e)$ introduced in Assumption 2. It is easy to verify that there exists a strictly positive number ϵ such that $(\bar{\mathbf{p}}_n, \bar{\mathbf{p}}_e)$ is feasible for the CGNF with strict inequality in (4.6c), where $\bar{p}_{ij} = p_{ij}$ and $\bar{p}_{ji} = p_{ji} + \epsilon$ for every $(i,j) \in \overrightarrow{\mathcal{E}}$ and $\bar{\mathbf{p}}_n$ is associated with $\bar{\mathbf{p}}_e$. □

Let $\lambda_i^{\min}$ and $\lambda_i^{\max}$ denote optimal Lagrange multipliers corresponding to the constraints $p_i^{\min} \leq p_i$ and $p_i \leq p_i^{\max}$. Assume that $(\bar{\mathbf{p}}_n^*, \bar{\mathbf{p}}_e^*)$ is an optimal solution of the GNF problem. Using the duality theorem, it can be shown that changing the objective function to

$$\sum_{i \in \mathcal{N}} f_i(p_i) - \lambda_i^{\min}(p_i - p_i^{\min}) + \lambda_i^{\max}(p_i - p_i^{\max}) \tag{4.55}$$

would not affect the optimal solution [37]. Furthermore, it follows from the first-order optimality conditions that

$$(\bar{\mathbf{p}}_n^*, \bar{\mathbf{p}}_e^*) = \underset{\mathbf{p}_n \in \mathbb{R}^m, \mathbf{p}_e \in \mathcal{B}_e}{\arg\min} \quad \sum_{i \in \mathcal{N}} \lambda_i p_i \tag{4.56a}$$

$$\text{subject to} \quad p_i = \sum_{j \in \mathcal{N}(i)} p_{ij}, \qquad \forall i \in \mathcal{N} \tag{4.56b}$$

$$f_{ij}(p_{ij}) \leq p_{ji}, \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.56c}$$

$$p_{ij} \in [p_{ij}^{\min}, p_{ij}^{\max}], \quad \forall (i,j) \in \mathcal{E} \tag{4.56d}$$

where

$$\lambda_i = f_i'(\bar{p}_i^*) - \lambda_i^{\min} + \lambda_i^{\max}, \qquad \forall i \in \mathcal{N} \tag{4.57}$$

Hence,

$$(\bar{p}_{ij}^*, \bar{p}_{ji}^*) = \underset{(p_{ij},p_{ji})\in\mathbb{R}^2}{\arg\min} \quad \lambda_i p_{ij} + \lambda_j p_{ji} \tag{4.58a}$$

$$\text{subject to} \quad f_{ij}(p_{ij}) \leq p_{ji}, \tag{4.58b}$$

$$p_{ij} \in [p_{ij}^{\min}, p_{ij}^{\max}], \tag{4.58c}$$

$$p_{ji} \in [p_{ji}^{\min}, p_{ji}^{\max}] \tag{4.58d}$$

for every $(i,j) \in \overrightarrow{\mathcal{E}}$.

Definition 24. *Define $\mathcal{V}$ as the set of all indices $i \in \mathcal{N}$ for which $\lambda_i \leq 0$. Define $\bar{\mathcal{V}}$ as the set of all indices $i \in \mathcal{N}\backslash\mathcal{V}$ for which there exists a vertex $j \in \mathcal{V}$ such that $(i,j) \in \mathcal{G}$ (i.e., $\bar{\mathcal{V}}$ denotes the set of the neighbors of $\mathcal{V}$ in the graph $\mathcal{G}$).*

Since the objective function of the optimization problem (4.58) is linear, it is straightforward to verify that $f_{ij}(\bar{p}_{ij}^*) = \bar{p}_{ji}^*$ as long as $\lambda_i > 0$ or $\lambda_j > 0$. In particular,

$$f_{ij}(\bar{p}_{ij}^*) = \bar{p}_{ji}^*, \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}},\ \{i,j\} \not\subseteq \mathcal{V} \tag{4.59a}$$

$$\bar{p}_{ij}^* = p_{ij}^{\min}, \qquad \forall (i,j) \in \mathcal{E},\ i \in \bar{\mathcal{V}},\ j \in \mathcal{V} \tag{4.59b}$$

If $f_{ij}(\bar{p}_{ij}^*)$ were equal to $\bar{p}_{ji}^*$ for every $(i,j) \in \overrightarrow{\mathcal{E}}$, then the proof of Part 1 of Theorem 19 was complete. However, the relation $f_{ij}(\bar{p}_{ij}^*) < \bar{p}_{ji}^*$ might hold in theory if $(i,j) \in \overrightarrow{\mathcal{E}}$ and $\{i,j\} \subseteq \mathcal{V}$. Hence, is important to study this scenario.

Proof of Part 1 of Theorem 19 in the general case: For every given index $i \in \mathcal{V}$, the term λ_i is nonpositive by definition. On the other hand, $f_i'(\cdot)$ is strictly positive (since $f_i(\cdot)$ is monotonically increasing), and $\lambda_i^{\min}$ and $\lambda_i^{\max}$ are both nonnegative (since they are the Lagrange multipliers for inequality constraints). Therefore, it follows from (4.57) that $\lambda_i^{\min} > 0$, implying that

$$\bar{p}_i^* = p_i^{\min}, \qquad \forall i \in \mathcal{V} \tag{4.60}$$

Thus,

$$p_i^* \geq p_i^{\min} = \bar{p}_i^*, \qquad \forall i \in \mathcal{V} \tag{4.61}$$

Let $\mathcal{G}_s$ denote a subgraph of $\mathcal{G}$ with the vertex set $\mathcal{V} \cup \bar{\mathcal{V}}$ that includes those edges $(i,j) \in \mathcal{E}$ satisfying either of the following conditions:

- $\{i,j\} \subseteq \mathcal{V}$
- $i \in \mathcal{V}$ and $j \in \bar{\mathcal{V}}$.

Note that $\mathcal{G}_s$ includes all edges of $\mathcal{G}$ within the vertex subset $\mathcal{V}$ and those between the sets $\mathcal{V}$ and $\bar{\mathcal{V}}$, but this subgraph contains no edge between the vertices in $\bar{\mathcal{V}}$. The first objective is to show that

$$p_i^*(\mathcal{G}_s) \geq \bar{p}_i^*(\mathcal{G}_s), \qquad \forall i \in \mathcal{V} \cup \bar{\mathcal{V}} \tag{4.62}$$

To this end, two possibilities will be investigated:

- *Case 1)* Consider a vertex $i \in \mathcal{V}$. Given each edge $(i,j) \in \mathcal{E}$, vertex j must belong to $\mathcal{V} \cup \bar{\mathcal{V}}$, due to Definition 24. Hence, $p_i^*(\mathcal{G}_s) = p_i^*$ and $\bar{p}_i^*(\mathcal{G}_s) = \bar{p}_i^*$. Combining these equalities with (4.61) gives rise to $p_i^*(\mathcal{G}_s) \geq \bar{p}_i^*(\mathcal{G}_s)$.

- *Case 2)* Consider a vertex $i \in \bar{\mathcal{V}}$. Based on (4.59b), One can write:

$$\bar{p}_i^*(\mathcal{G}_s) = \sum_{j \in \mathcal{V} \cap \mathcal{N}(i)} \bar{p}_{ij}^* = \sum_{j \in \mathcal{V} \cap \mathcal{N}(i)} p_{ij}^{\min} \tag{4.63}$$

  Similarly,

$$p_i^*(\mathcal{G}_s) = \sum_{j \in \mathcal{V} \cap \mathcal{N}(i)} p_{ij}^* \geq \sum_{j \in \mathcal{V} \cap \mathcal{N}(i)} p_{ij}^{\min} \tag{4.64}$$

  Thus, $p_i^*(\mathcal{G}_s) \geq \bar{p}_i^*(\mathcal{G}_s)$.

So far, inequality (4.62) has been proven. Consider $\tilde{\mathbf{p}}_n$ introduced in (4.53). Similar to (4.54), it is straightforward to show that $\tilde{p}_i(\mathcal{G}_s) \leq \bar{p}_i^*(\mathcal{G}_s)$ for every $i \in \mathcal{V} \cup \bar{\mathcal{V}}$. Hence,

$$\tilde{\mathbf{p}}_n(\mathcal{G}_s) \leq \bar{\mathbf{p}}_n^*(\mathcal{G}_s) \leq \mathbf{p}_n^*(\mathcal{G}_s) \tag{4.65}$$

On the other hand, $\tilde{\mathbf{p}}_n(\mathcal{G}_s)$ and $\mathbf{p}_n^*(\mathcal{G}_s)$ are both in $\mathcal{P}(\mathcal{G}_s)$. Using (4.65) and Theorem 18 (but for $\mathcal{G}_s$ as opposed to $\mathcal{G}$), it can be concluded that $\bar{\mathbf{p}}_n^*(\mathcal{G}_s) \in \mathcal{P}(\mathcal{G}_s)$. Hence, there exists a flow vector $\hat{\mathbf{p}}_e(\mathcal{G}_s)$ associated with $\bar{\mathbf{p}}_n^*(\mathcal{G}_s)$, meaning that

$$\bar{p}_i^*(\mathcal{G}_s) = \sum_{j \in \mathcal{N}(i) \cap (\mathcal{V} \cup \bar{\mathcal{V}})} \hat{p}_{ij}(\mathcal{G}_s), \qquad \forall i \in \mathcal{V} \tag{4.66a}$$

$$\bar{p}_i^*(\mathcal{G}_s) = \sum_{j \in \mathcal{N}(i) \cap \mathcal{V}} \hat{p}_{ij}(\mathcal{G}_s), \qquad \forall i \in \bar{\mathcal{V}} \tag{4.66b}$$

$$\hat{p}_{ji}(\mathcal{G}_s) = f_{ij}(\hat{p}_{ij}(\mathcal{G}_s)), \qquad \forall (i,j) \in \overrightarrow{\mathcal{G}}_s \tag{4.66c}$$

Now, one can expand $\hat{\mathbf{p}}_e(\mathcal{G}_s)$ to $\hat{\mathbf{p}}_e$ as

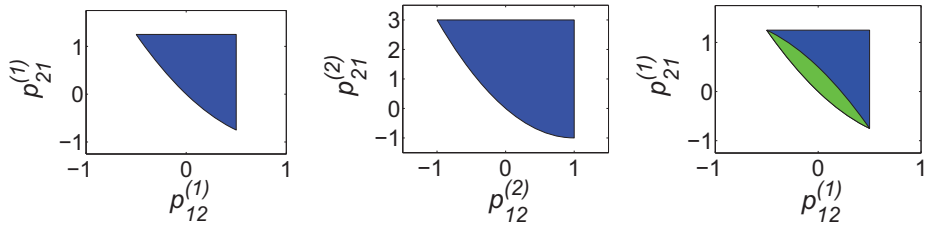
$$\hat{p}_{jk} = \begin{cases} \hat{p}_{jk}(\mathcal{G}_s) & \text{if} \quad (j,k) \in \mathcal{G}_s \\ \bar{p}_{jk}^* & \text{otherwise} \end{cases}, \quad \forall (j,k) \in \mathcal{E} \tag{4.67}$$

Let $\hat{\mathbf{p}}_n$ denote the injection vector associated with the flow vector $\hat{\mathbf{p}}_e$. Two observations can be made:

1) $\hat{\mathbf{p}}_n$ is equal to $\bar{\mathbf{p}}_n^*$.

2) Due to (4.59a), (4.66c) and (4.67), $(\hat{\mathbf{p}}_n, \hat{\mathbf{p}}_e)$ is a feasible point of GNF.

This means that $\bar{\mathbf{p}}_n^*$ is the unique optimal solution of Geometric CGNF and yet a feasible point of Geometric GNF. The rest of the proof is the same as the proof of Theorem 19 under Condition (4.50) (given earlier). $\square$



(a) (b) (c)

Figure 4.5.3: Figures (a) and (b) show the feasible sets $\mathcal{T}_{\rm c}^{(1)}$ and $\mathcal{T}_{\rm c}^{(2)}$, respectively. Figure (c) is aimed to show that CGNF may have an infinite number of solutions (all points in the yellow area may be the solutions of GNF).

Next example is provided to understand the reason why CGNF may fail to obtain a correct flow vector associated with the optimal injection vector.

Example 3: Consider again the illustrative example studied in Section 4.3, corresponding to the graph $\mathcal{G}$ depicted in Figure 4.2.1. Let $\mathcal{T}$ denote the projection of the feasible set of the GNF problem given in (4.8) over the flow space associated with the vector $(p_{12}^{(1)}, p_{21}^{(1)}, p_{12}^{(2)}, p_{21}^{(2)})$. It is easy to verify that $\mathcal{T}$ can be decomposed as the product of $\mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)}$, where

$$\mathcal{T}^{(1)} = \left\{ (p_{12}^{(1)}, p_{21}^{(1)}) \;\middle|\; p_{12}^{(1)} \in [-0.5, 0.5],\ p_{21}^{(1)} = \left(p_{12}^{(1)} - 1\right)^2 - 1 \right\}$$

and

$$\mathcal{T}^{(2)} = \left\{ (p_{12}^{(2)}, p_{21}^{(2)}) \;\middle|\; p_{12}^{(2)} \in [-1, 1],\ p_{21}^{(2)} = \left(p_{12}^{(2)} - 1\right)^2 - 1 \right\}$$

Likewise, define $\mathcal{T}_{\rm c}$ as the projection of the feasible set of the CGNF problem over its flow space. As before, $\mathcal{T}_{\rm c}$ can be written as $\mathcal{T}_{\rm c}^{(1)} \times \mathcal{T}_{\rm c}^{(2)}$, where $\mathcal{T}_{\rm c}^{(i)}$ is obtained from $\mathcal{T}^{(i)}$ by changing its equality

$$p_{21}^{(i)} = \left(p_{12}^{(i)} - 1\right)^2 - 1 \tag{4.68}$$

to the inequality

$$p_{21}^{(i)} \geq \left(p_{12}^{(i)} - 1\right)^2 - 1 \tag{4.69}$$

for $i = 1, 2$, and adding the limits $p_{21}^{(1)} \leq 1.5^2 - 1$ and $p_{21}^{(2)} \leq 2^2 - 1$. The sets $\mathcal{T}_{\rm c}^{(1)}$ and $\mathcal{T}_{\rm c}^{(2)}$ are drawn in Figures 4.5.3(a) and 4.5.3(b). Given $i \in \{1, 2\}$, note that $\mathcal{T}_{\rm c}^{(i)}$ has two flat boundaries and one curvy (lower) boundary that is the same as $\mathcal{T}^{(i)}$. Consider the flow

vector $(\bar{p}_{12}^{(1)}, \bar{p}_{21}^{(1)}, \bar{p}_{12}^{(2)}, \bar{p}_{21}^{(2)}) \in \mathcal{T}_{\mathrm{c}}$ defined as

$$\begin{aligned} \left(\bar{p}_{12}^{(1)}, \bar{p}_{21}^{(1)}\right) &= \left(0.5, (0.5-1)^2 - 1\right), \\ \left(\bar{p}_{12}^{(2)}, \bar{p}_{21}^{(2)}\right) &= \left(-0.5, (-0.5-1)^2 - 1\right) \end{aligned} \tag{4.70}$$

Define $\bar{p}_1 = \bar{p}_{12}^{(1)} + \bar{p}_{12}^{(2)}$ and $\bar{p}_2 = \bar{p}_{21}^{(1)} + \bar{p}_{21}^{(2)}$. It can be verified that for every point $(\tilde{p}_{12}^{(1)}, \tilde{p}_{21}^{(1)})$ in the green area of Figure 4.5.3(c), there exists a vector $(\tilde{p}_{12}^{(2)}, \tilde{p}_{21}^{(2)}) \in \mathcal{T}_{\mathrm{c}}^{(2)}$ such that

$$\bar{p}_1 = \tilde{p}_{12}^{(1)} + \tilde{p}_{12}^{(2)}, \quad \bar{p}_2 = \tilde{p}_{21}^{(1)} + \tilde{p}_{21}^{(2)} \tag{4.71}$$

This means that if $(\bar{p}_1, \bar{p}_2, \bar{p}_{12}^{(1)}, \bar{p}_{21}^{(1)}, \bar{p}_{12}^{(2)}, \bar{p}_{21}^{(2)})$ turns out to be an optimal solution of CGNF, then $(\bar{p}_1, \bar{p}_2, \tilde{p}_{12}^{(1)}, \tilde{p}_{21}^{(1)}, \tilde{p}_{12}^{(2)}, \tilde{p}_{21}^{(2)})$ becomes another solution of CGNF. As a result, although Geometric CGNF has a unique solution (optimal injection vector), CGNF may have an infinite number of solutions whose corresponding flow vectors do not necessarily satisfy the constraints of GNF. □

So far, we have shown that CGNF always finds the optimal injection vector and optimal objective value for the GNF problem. In addition, it finds the optimal flow vector if the injection vector is a Pareto point. Now, we consider the case where the optimal injection vector is not necessarily Pareto but lies on the boundary of the injection region. The objective is to prove that the network $\mathcal{G}$ can be decomposed into two subgraphs $\mathcal{G}_1$ and $\mathcal{G}_2$ such that: (i) the flows obtained from CGNF are optimal (feasible) for GNF for those lines inside $\mathcal{G}_1$ or between $\mathcal{G}_1$ and $\mathcal{G}_2$, (ii) the flows over the lines between $\mathcal{G}_1$ and $\mathcal{G}_2$ all hit their limits at optimality.

Definition 25. *Define $\mathcal{G}_1$ and $\mathcal{G}_2$ as the subgraphs of $\mathcal{G}$ induced by the vertex subsets $\mathcal{N}\backslash\mathcal{V}$ and $\mathcal{V}$, respectively.*

Theorem 20. *Assume that $f_i(\cdot)$ is strictly convex for every $i \in \mathcal{N}$. Let $(\mathbf{p}_n^*, \mathbf{p}_d^*)$ and $(\mathbf{p}_n^*, \bar{\mathbf{p}}_d^*)$ denote arbitrary globally optimal solutions of the GNF and CGNF problems, respectively. The following relations hold:*

$$p_{ij}^* = \bar{p}_{ij}^*, \qquad \forall (i,j) \in \mathcal{N}\backslash\mathcal{V} \tag{4.72a}$$

$$p_{ji}^* = \bar{p}_{ji}^* = p_{ji}^{\max}, \qquad \forall (i,j) \in (\mathcal{N}\backslash\mathcal{V} \times \mathcal{V}) \cap \mathcal{E} \tag{4.72b}$$

Proof: Since every solution of GNF is a solution of CGNF as well (due to Theorem 19), the points $(\mathbf{p}_n^*, \mathbf{p}_d^*)$ and $(\mathbf{p}_n^*, \bar{\mathbf{p}}_d^*)$ are both solutions of CGNF. Now, it follows from the duality theorem that $(\mathbf{p}_n^*, \mathbf{p}_d^*)$ and $(\mathbf{p}_n^*, \bar{\mathbf{p}}_d^*)$ are both minimizers of (4.56) and (4.58). Since the objective of (4.58) is linear and $f_i(\cdot)$ is strictly convex, it can be concluded that:

- The optimization problem (4.58) has a unique solution as long as $\lambda_i^* > 0$ or $\lambda_j^* > 0$.
- (p_{ij}, p_{ji}) becomes equal to $(p_{ij}^{\min}, p_{ji}^{\max})$ at optimality if $\lambda_i^* > 0$ and $\lambda_j^* \leq 0$.

- (p_{ij}, p_{ji}) becomes equal to $(p_{ij}^{\max}, p_{ji}^{\min})$ at optimality if $\lambda_j^* > 0$ and $\lambda_i^* \leq 0$.

Equations (4.72a) and (4.72b) follow immediately from the above properties. □

Corollary 6. *Let $(\mathbf{p}_n^*, \mathbf{p}_d^*)$ and $(\mathbf{p}_n^*, \bar{\mathbf{p}}_d^*)$ denote arbitrary globally optimal solutions of the GNF and CGNF problems, respectively. If there exists a vertex $i \in \mathcal{N}$ such that $\bar{p}_i^* > p_i^{\min}$, then $\mathbf{p}_d^*$ and $\bar{\mathbf{p}}_d^*$ must be identical in at least one entry.*

Proof: Consider a vertex $i \in \mathcal{N}$ such that $\bar{p}_i^* > p_i^{\min}$. It follows from (4.57) that λ_i^* is positive. Now, Definition 25 yields that the subgraph $\mathcal{G}_1$ is nonempty. The proof is an immediate consequence of Theorem 20. □

Definition 26. *Consider a solution $(\mathbf{p}_n^*, \mathbf{p}_d^*)$ of GNF. A line $(i, j) \in \mathcal{E}$ of the network $\mathcal{G}$ is called "congested" if p_{ij}^* is equal to $p_{ij}^{\max}$ or p_{ji}^* is equal to $p_{ji}^{\max}$.*

Corollary 7. *Let $(\mathbf{p}_n^*, \mathbf{p}_d^*)$ and $(\mathbf{p}_n^*, \bar{\mathbf{p}}_d^*)$ denote arbitrary globally optimal solutions of the GNF and CGNF problems, respectively. Assume that there exists a vertex $i \in \mathcal{N}$ such that $\bar{p}_i^* > p_i^{\min}$. If the network $\mathcal{G}$ has no congested line, then GNF and CGNF are equivalent, i.e., $(\mathbf{p}_n^*, \mathbf{p}_d^*) = (\mathbf{p}_n^*, \bar{\mathbf{p}}_d^*)$.*

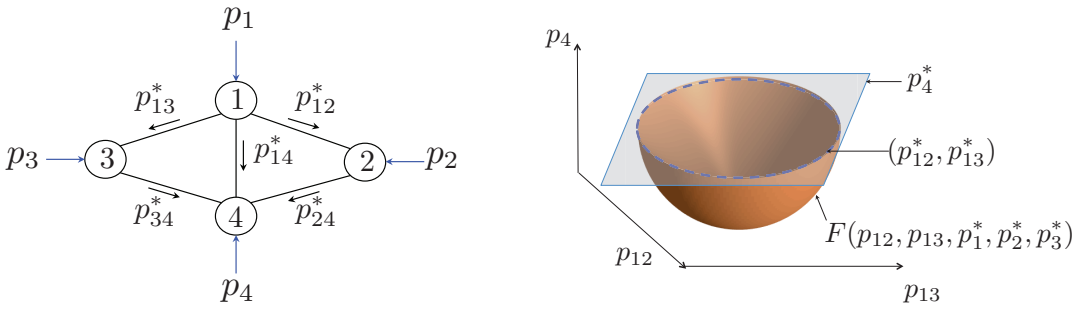
Proof: Due to the proof of Corollary 6, the set $\mathcal{N} \backslash \mathcal{V}$ is nonempty. On the other hand, since the network $\mathcal{G}$ has no congested line by assumption, it can be concluded from Theorem 20 that $(\mathcal{N} \backslash \mathcal{V} \times \mathcal{V}) \cap \mathcal{E}$ is an empty set. Therefore, $\mathcal{V}$ must be empty, which implies the equivalence of GNF and CGNF due to Theorem 20. □

4.6 Characterization of Optimal Flow Vectors

In this section, we aim to characterize the set of all optimal flow vectors for GNF, based on the optimal injection vector found using CGNF. In particular, we will show that this set could be nonconvex and disconnected. Before presenting the results, it is helpful to illustrate the key ideas in an example.

Example 4: Consider the graph $\mathcal{G}$ depicted in Figure 4.6.1(a), which consists of two cycles and four nodes. Let $(\mathbf{p}_n^*, \mathbf{p}_e^*)$ denote an arbitrary solution of GNF, where $\mathbf{p}_n^*$ is obtained from CGNF and $\mathbf{p}_e^*$ is to be found. The objective of this example is to demonstrate that all optimal flows in the network can be uniquely characterized in terms of two flows. Consider the unknown flows p_{12}^* and p_{13}^*. One can write

$$p_{24}^* = p_2^* - f_{12}(p_{12}^*) \tag{4.73a}$$
$$p_{34}^* = p_3^* - f_{13}(p_{13}^*) \tag{4.73b}$$
$$p_{14}^* = p_1^* - p_{12}^* - p_{13}^* \tag{4.73c}$$



(a) The 2-cycle graph studied in Example 4

(b) Visualization of the level-set problem used to find optimal flows for Example 4

Figure 4.6.1: The 2-cycle graph and its feasible region in Example 4.

It follows from the above equations that all flows in the network can be cast as functions of (p_{12}^*, p_{13}^*), and in addition $(p_{12}, p_{13}) = (p_{12}^*, p_{13}^*)$ is a solution to the level-set problem $F(p_{12}, p_{13}, p_1^*, p_2^*, p_3^*) = p_4^*$, where

$$\begin{aligned} F(p_{12}, p_{13}, p_1, p_2, p_3) = {} & f_{24}\left(p_2 - f_{12}(p_{12})\right) \\ & + f_{34}\left(p_3 - f_{13}(p_{13})\right) \\ & + f_{14}\left(p_1 - p_{12} - p_{13}\right) \end{aligned} \tag{4.74}$$

is a convex function with respect to (p_{12}, p_{13}) but not necessarily monotonic. On the other hand, the equations in (4.73) can be used to translate the box constraints on all flows to certain constraints only on p_{12}^* and p_{13}^*:

$$\tilde{p}_{12}^{\min} \le p_{12}^* \le \tilde{p}_{12}^{\max} \tag{4.75a}$$
$$\tilde{p}_{13}^{\min} \le p_{13}^* \le \tilde{p}_{13}^{\max} \tag{4.75b}$$
$$p_{14}^{\min} \le p_1^* - p_{12}^* - p_{13}^* \le p_{14}^{\max} \tag{4.75c}$$

for some numbers $\tilde{p}_{12}^{\min}, \tilde{p}_{12}^{\max}, \tilde{p}_{13}^{\min}, \tilde{p}_{13}^{\max}$. Let C_1 and C_2 denote the sets of all points (p_{12}^*, p_{13}^*) satisfying the level-set problem $F(p_{12}^*, p_{13}^*, p_1^*, p_2^*, p_3^*) = p_4^*$ and the reformulated flow constraints (4.75), respectively. The set of all optimal flow solutions (p_{12}^*, p_{13}^*) can be expressed as $\mathcal{C}_1 \cap \mathcal{C}_2$, where $\mathcal{C}_1$ is the boundary of a convex set (corresponding to $F(\cdot)$) and $\mathcal{C}_2$ is a polytope. As illustrated in Figure 4.6.1(b), $\mathcal{C}_1$ is the boundary of a convex set, and therefore its intersection with a polytope (e.g., a box) could form up to 4 disconnected components. In summary, the optimal flow vectors for GNF may constitute a nonconvex infinite set, consisting of as high as 4 disconnected components. □

By following the argument used in Example 5, it is straightforward to show that if the graph $\mathcal{G}$ is a tree, the optimal flow vector is unique and can be easily obtained from the optimal injection vector $\mathbf{p}_n^*$. Hence, the main challenge is to deal with mesh flow networks. To this end, consider an arbitrary spanning tree of the m-node graph $\mathcal{G}$, denoted as $\mathcal{G}_t$. Let

$\mathbf{p}_{dt}$ denote a sub-vector of the semi-flow vector $\mathbf{p}_d$ associated with those edges of $\mathcal{G}$ that do not exist in $\mathcal{G}_t$. Recall that $\overrightarrow{\mathcal{G}}$ was obtained through an arbitrary orientation of the edges of the graph $\mathcal{G}$. With no loss of generality, one can consider $\mathcal{G}_t$ as a rooted tree with node m as its root, where all arcs of $\overrightarrow{\mathcal{G}}$ are directed toward the root.

Lemma 31. *There exist convex functions $F_{ij} : \mathbb{R}^{|\mathcal{E}|} \to \mathbb{R}$ for all $(i,j) \in \overrightarrow{\mathcal{E}}$ such that the following statements hold:*

1) Given every arbitrary feasible solution $(\mathbf{p}_n, \mathbf{p}_e)$ of the GNF problem, the relations

$$p_{ji} = F_{ij}(\mathbf{p}_{dt}, p_1, p_2, ..., p_{m-1}), \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.76}$$

are satisfied.

2) The function $F(\mathbf{p}_{dt}, p_1, p_2, ..., p_{m-1})$ defined as

$$\sum_{j \in \mathcal{N}(m)} F_{jm}(\mathbf{p}_{dt}, p_1, p_2, ..., p_{m-1}) \tag{4.77}$$

is convex.

Proof: The proof is in line with the technique used in Example 4. The details are omitted for brevity. □

Definition 27. *Define $\mathcal{C}_1$ as the set of all vectors $\mathbf{p}_{dt}$ satisfying the level-set problem $F(\mathbf{p}_{dt}, p_1^*, p_2^*, ..., p_{m-1}^*) = p_m^*$. Also, define $\mathcal{C}_2$ as the set of all vectors $\mathbf{p}_{dt}$ satisfying the inequalities*

$$p_{ji}^{\min} \leq F_{ij}(\mathbf{p}_{dt}, p_1^*, p_2^*, ..., p_{m-1}^*) \leq p_{ji}^{\max}, \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.78}$$

Theorem 21. *A flow vector $\mathbf{p}_e^*$ is globally optimal for GNF if and only if*

$$\mathbf{p}_{dt}^* \in \mathcal{C}_1 \cap \mathcal{C}_2 \tag{4.79a}$$

$$p_{ji}^* = F_{ij}(\mathbf{p}_{dt}^*, p_1^*, p_2^*, ..., p_{m-1}^*), \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.79b}$$

$$p_{ij}^* = f_{ji}(p_{ji}^*), \qquad \forall (i,j) \in \overrightarrow{\mathcal{E}} \tag{4.79c}$$

Proof: The proof is based on Lemma 31 and the technique used in Example 4. The details are omitted for brevity. □

Theorem 21 states that: (i) the set of optimal flow vectors can be characterized in terms of the unique optimal injection vector as well as the flow sub-vector $\mathbf{p}_{dt}$, (ii) the set of optimal flow sub-vectors $\mathbf{p}_{dt}^*$ is the collection of all points in the intersection of $\mathcal{C}_1$ and $\mathcal{C}_2$. Moreover, in light of Lemma 31, $\mathcal{C}_1$ is the boundary of a convex set. Although $\mathcal{C}_2$ was shown to be a polytope in Examples 4 and 5, it is non-convex in general. Since $\mathcal{C}_1$ is the boundary of a convex set, it occurs that the intersection of $\mathcal{C}_2$ with $\mathcal{C}_1$ may lead to as high as $2^{|\mathcal{E}|-|\mathcal{N}|+1}$ disconnected components, all lying on the boundary of a convex set (note that $|\mathcal{E}| - |\mathcal{N}| + 1$ is the size of the vector $\mathbf{p}_{dt}$).

4.7 Extended Generalized Network Flow

In this subsection, we generalize the results developed for the GNF problem to the case where there are global convex constraints coupling the flows and/or injections of different parts of the network, in addition to the local constraints over individual lines and at separate nodes.

Definition 28. *Consider a set of convex constraints $g_i(\mathbf{p}_n, \mathbf{p}_e) \leq 0$ for $i = 1, 2, ..., k$, which are called* coupling constraints. *The* extended GNF problem *is defined as* (4.4) *subject to this set of coupling constraints. Denote $\mathcal{P}^e$ as the set of all vectors $\mathbf{p}_n$ for which there exists a vector $\mathbf{p}_e$ such that $(\mathbf{p}_n, \mathbf{p}_e)$ is feasible for the extended GNF problem. The above set of coupling constraints is referred to as* box-preserving *if its addition to the GNF problem preserves the box property of the injection region, meaning that the box $\mathcal{B}(\mathbf{p}_n, \tilde{\mathbf{p}}_n)$ is contained in $\mathcal{P}^e$ for every two points $\mathbf{p}_n$ and $\tilde{\mathbf{p}}_n$ in $\mathcal{P}^e$.*

Theorem 22. *Consider the extended GNF problem with the coupling constraints $g_i(\mathbf{p}_n, \mathbf{p}_e) \leq 0$ for every $i \in \{1, 2, ..., k\}$. This set of constraints is guaranteed to be box-preserving if either of the following conditions is satisfied:*

1) $\mathcal{G}$ is a tree and the function $g_i(\mathbf{p}_n, \mathbf{p}_e)$ is non-decreasing with respect to all entries of $\mathbf{p}_n$ and $\mathbf{p}_e$, for every $i \in \{1, 2, ..., k\}$.

2) The function $g_i(\mathbf{p}_n, \mathbf{p}_e)$ does not depend on $\mathbf{p}_e$ and is non-decreasing with respect to all entries of $\mathbf{p}_n$, for every $i \in \{1, 2, ..., k\}$.

Proof: The box-preserving property under Condition 2 follows from the fact that whenever the coupling constraints are non-decreasing functions of the injection vector, if $\mathbf{p}_n$ satisfies the constraints, any other injection vector $\tilde{\mathbf{p}}_n$ with the property $\tilde{\mathbf{p}}_n \leq \mathbf{p}_n$ also satisfies the constraints.

To prove the box-preserving property under Condition 1, it suffices to show that if $\mathcal{G}$ is a tree, every flow p_{ij} can be written as a non-decreasing function of $\mathbf{p}_n$ (then the proof follows from Condition 2 of the theorem). Consider $\mathcal{G}$ as a rooted tree with an arbitrary node at the root. Recall that $\overrightarrow{\mathcal{G}}$ was obtained through an arbitrary orientation of the edges of $\mathcal{G}$. Without loss of generality, assume that the directions of all edges are toward the root. Define h as the depth of $\mathcal{G}$ (maximum distance of every leaf from the root). Assume that a node with the distance t from the root is identified by i_t. First, we use induction to show that the flows going toward the root can be written as non-decreasing functions of the injection vector. We start with the farthest nodes from the root. For each node i_h, one can write $p_{i_h i_{h-1}} = p_{i_h}$, which is non-decreasing in terms of the injection vector. Now, for every flow $p_{i_t i_{t-1}}$ with $0 \leq t \leq h-1$, one can write

$$p_{i_t i_{t-1}} = p_{i_t} - \sum_{(j_{t+1}, i_t) \in \mathcal{E}} f_{j_{t+1}, i_t}(p_{j_{t+1} i_t}) \tag{4.80}$$

By the induction hypothesis, $p_{j_{t+1}i_t}$ can be written as a non-decreasing function of the injection vector. Therefore, (4.80) implies that the same statement holds for $p_{i_t i_{t-1}}$.

Now, we use another inductive argument to show that each flow going toward the leaves can be written as a non-decreasing function of the injection vector. We start from the root node. For every flow $p_{i_0 i_1}$, one can write

$$p_{i_0 i_1} = p_{i_0} - \sum_{\substack{(j_1,i_0)\in\mathcal{E} \\ j_1 \neq i_1}} f_{j_1,i_0}(p_{j_1 i_0}) \tag{4.81}$$

which implies that $p_{i_0 i_1}$ is a non-decreasing function of the injection vector (note that this property holds for $p_{j_1 i_0}$). For every flow $p_{i_{t-1}i_t}$ with $2 \leq t \leq h$, one can verify that

$$p_{i_{t-1}i_t} = p_{i_{t-1}} - f^{-1}_{i_{t-1}i_{t-2}}(p_{i_{t-2}i_{t-1}}) - \sum_{\substack{(j_1,i_0)\in\mathcal{E} \\ j_t \neq i_t}} f_{j_t,i_{t-1}}(p_{j_t i_{t-1}}) \tag{4.82}$$

The proof is completed by observing that

- $f^{-1}_{i_{t-1}i_{t-2}}(\cdot)$ is a decreasing function.
- $p_{i_{t-2}i_{t-1}}$ is a non-decreasing function of the injection vector due to the induction hypothesis.
- $p_{j_t i_{t-1}}$ is a non-decreasing function of the injection vector since its direction is toward the root. □

In the rest of this subsection, we assume that the set of coupling constraints in the extended GNF problem is box-preserving.

Corollary 8. *Consider two arbitrary points $\hat{\mathbf{p}}_n$ and $\tilde{\mathbf{p}}_n$ belonging to the box-constrained injection region $\mathcal{P}^e \cap \mathcal{B}$. The box $\mathcal{B}(\hat{\mathbf{p}}_n, \tilde{\mathbf{p}}_n)$ is contained in $\mathcal{P}^e \cap \mathcal{B}$.*

Proof: The proof follows immediately from the definition of $\mathcal{P}^e$ and Definition 28. □

Define the extended CGNF problem as CGNF subject to the additional constraints $g_i(\mathbf{p}_n, \mathbf{p}_e) \leq 0$ for $i = 1, 2, ..., k$. Note that this problem is convex.

Theorem 23. *Assume that the extended GNF problem is feasible. Let $(\mathbf{p}^*_n, \mathbf{p}^*_e)$ and $(\bar{\mathbf{p}}^*_n, \bar{\mathbf{p}}^*_e)$ denote arbitrary globally optimal solutions of the extended GNF and extended CGNF problems, respectively. The following relations hold:*

1) $\mathbf{p}^*_n = \bar{\mathbf{p}}^*_n$

2) $(\bar{\mathbf{p}}^*_n, \bar{\mathbf{p}}^*_e)$ *is a solution of the extended GNF problem, provided that $\mathbf{p}^*_n$ is a Pareto point in $\mathcal{P}^e$.*

Proof: The argument made in the proof of Theorem 19 can be adopted to prove this theorem. The details are omitted for brevity. □

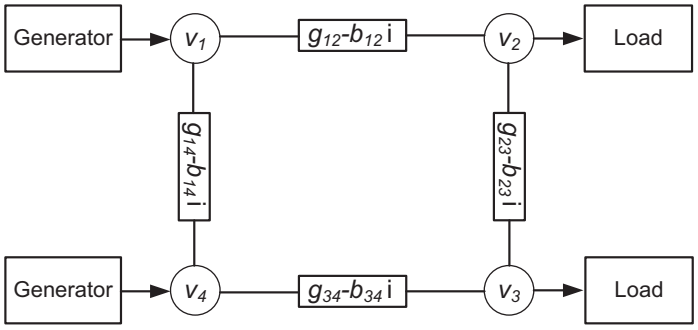


Figure 4.8.1: An example of electrical power network.

4.8 Optimal Power Flow in Electrical Power Networks

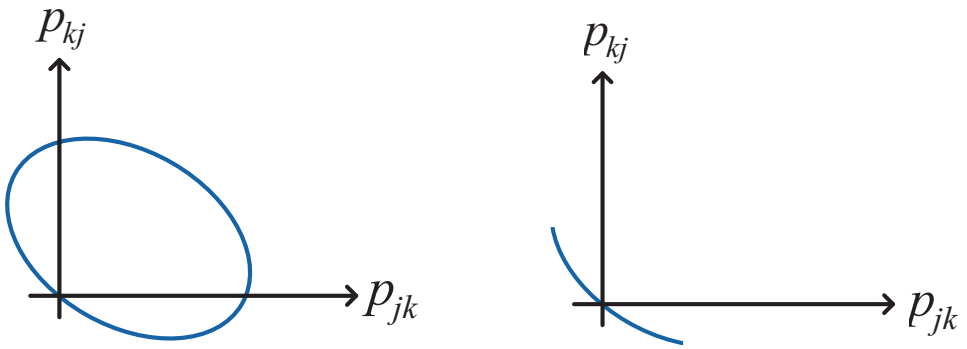
In this subsection, the results derived earlier for the GNF and extended GNF problems will be applied to power networks. Consider a group of generators (sources of energy), which are connected to a group of electrical loads (consumers) via an electrical power network (grid). This network comprises a set of lines connecting various nodes to each other (e.g., a generator to a load). Figure 4.8.1 exemplifies a four-node power network with two generators and two loads. Each load requests certain amount of energy, and the question of interest is to find the most economical power dispatch by the generators such that the demand and network constraints are satisfied. To formulate the problem, let $\mathcal{G}$ denote the flow network corresponding to the electrical power network, where

- The injection p_j at node $j \in \mathcal{N}$ represents either the active power produced by a generator and injected to the network or the active power absorbed from the network by an electrical load.
- The flow p_{jk} over each line $(j,k) \in \mathcal{E}$ represents the active power entering the line (j,k) from its j endpoint.

The problem of optimizing the flows in a power network is called "optimal power flow (OPF)".

Let v_i denote the complex (phasor) voltage at node $i \in \mathcal{N}$ of the power network. Denote the phase of v_i as θ_i. Given an edge $(j,k) \in \mathcal{G}$, we denote the admittance of the line between nodes j and k as $g_{jk} - \mathrm{i}b_{jk}$, where the symbol i denotes the imaginary unit. g_{jk} and b_{jk} are nonnegative numbers due to the passivity of the line. There are two active flows entering the line (j,k) from its both ends. These flows are given by the equations:

$$p_{jk} = |v_j|^2 g_{jk} + |v_j||v_k|b_{jk}\sin(\theta_{jk}) - |v_j||v_k|g_{jk}\cos(\theta_{jk}),$$
$$p_{kj} = |v_k|^2 g_{jk} - |v_j||v_k|b_{jk}\sin(\theta_{jk}) - |v_j||v_k|g_{jk}\cos(\theta_{jk})$$



(a) Feasible set for (p_{jk}, p_{kj})

(b) Feasible set for (p_{jk}, p_{kj}) after imposing lower and upper bounds on θ_{jk}

Figure 4.8.2: The feasible set of the active power flows in power systems.

where $\theta_{jk} = \theta_j - \theta_k$. First, consider the distribution system where the underlying network is a tree. For now, assume that $|v_j|$ and $|v_k|$ are fixed at their nominal values, while θ_{jk} is a variable to be designed. If θ_{jk} varies from $-\pi$ to π, then the feasible set of (p_{jk}, p_{kj}) becomes an ellipse, as illustrated in Figure 4.8.2(a). It can be observed that p_{kj} cannot be written as a function of p_{jk}. This observation is based on the implicit assumption that there is no limit on θ_{jk}. Suppose that θ_{jk} must belong to an interval $[-\theta_{jk}^{\max}, \theta_{jk}^{\max}]$ for some angle $\theta_{jk}^{\max}$. If the new feasible set for (p_{jk}, p_{kj}) resembles the partial ellipse drawn in Figure 4.8.2(b), then p_{kj} can be expressed as $f_{jk}(p_{jk})$ for a monotonically decreasing and convex function $f_{jk}(\cdot)$. This occurs if

$$\theta_{jk}^{\max} \leq \tan^{-1}\left(\frac{b_{jk}}{g_{jk}}\right) \tag{4.83}$$

It is interesting to note that the right side of the above inequality is equal to 45.0°, 63.4° and 78.6° for $\frac{b_{jk}}{g_{jk}}$ equal to 1, 2 and 5, respectively. Note that $\frac{b_{jk}}{g_{jk}}$ is normally greater than 5 (due to the specifications of the lines) and $\theta_{jk}^{\max}$ is normally less than 15° and very rarely as high as 30° due to stability and thermal limits (this angle constraint is forced either directly or through $p_{jk}^{\min}$ and $p_{jk}^{\max}$ in practice). Hence, Condition (4.83) is practical. Furthermore, each line of the power system can tolerate a certain amount of current in magnitude. One can verify that the magnitude of the current on the line (j, k), denoted by i_{jk}, satisfies the equation

$$|i_{jk}|^2 = |y_{jk}|\left(|v_j|^2 + |v_k|^2 - 2|v_j v_k|\cos(\theta_{jk})\right)$$

Therefore, an upper bound on $|i_{jk}|$ can be translated into a constraint on θ_{jk}, which can be reflected in $\theta_{jk}^{\max}$. By assuming that (4.83) is satisfied, there exists a monotonically decreasing, convex function $f_{jk}(\cdot)$ such that

$$p_{kj} = f_{jk}(p_{jk}), \quad \forall p_{jk} \in [p_{jk}^{\min}, p_{jk}^{\max}], \tag{4.84}$$

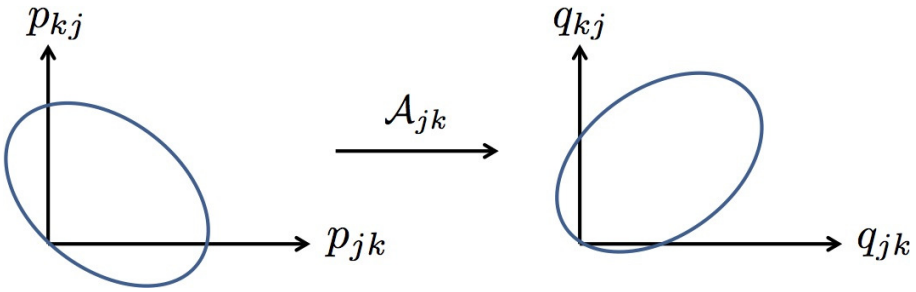


Figure 4.8.3: Linear transformation of active flows to reactive flows.

where $p_{jk}^{\min}$ and $p_{jk}^{\max}$ correspond to $\theta_{jk}^{\max}$ and $-\theta_{jk}^{\max}$, respectively.

Given two disparate edges (j,k) and (j',k'), the phase differences θ_{jk} and $\theta_{j'k'}$ may be varied independently in the distribution network. (4.84) implies that the problem of optimizing active flows reduces to GNF. In this case, Theorems 18 and 19 can be used to study the corresponding approximated OPF problem. As a result, the optimal injections for the approximated OPF can be found via the corresponding CGNF problem. This implies two facts about the conic relaxations studied in [151, 169, 240, 236, 158, 35, 276, 154, 153] for solving the OPF problem:

- The relaxations are exact without using the concept of load over-satisfaction (i.e., relaxing the flow constraints). This is a generalization of the results derived in the above papers (please refer to [153] for more details on this concept).
- Given the optimal injections, the optimal flows can be uniquely derived using the method delineated in the proof of Theorem 22.

In addition to active power, voltage magnitudes and reactive power are normally optimized in the OPF problem. In what follows, we generalize the above results to these cases.

Variable Reactive Power

In real-world power systems, different components of the network produce/consume reactive power. Since reactive power has a direct impact on the operation of the power system, this is often controlled in the OPF problem. To formulate the problem in this case, notice that each line has two reactive flows entering from its both endpoints. These equations can be described as

$$\begin{aligned} q_{jk} &= |v_j|^2 g_{jk} - |v_j||v_k| g_{jk} \sin(\theta_{jk}) - |v_j||v_k| b_{jk} \cos(\theta_{jk}), \\ q_{kj} &= |v_k|^2 g_{jk} + |v_j||v_k| g_{jk} \sin(\theta_{jk}) - |v_j||v_k| b_{jk} \cos(\theta_{jk}) \end{aligned} \tag{4.85}$$

Each bus at the network has a limited capacity to absorb/produce reactive power. Upon defining q_i as the reactive power injection at node i (which is equal to the summation of

outgoing reactive flows from node i), this limited capacity can be captured by the pre-specified constraints $q_i^{\min} \leq q_i \leq q_i^{\max}$. Therefore, reactive flows can be written as linear functions of active flows based on the formula

$$\begin{bmatrix} q_{jk} \\ q_{kj} \end{bmatrix} = \underbrace{\frac{1}{2b_{jk}g_{jk}} \begin{bmatrix} b_{jk}^2 - g_{jk}^2 & b_{jk}^2 + g_{jk}^2 \\ b_{jk}^2 + g_{jk}^2 & b_{jk}^2 - g_{jk}^2 \end{bmatrix}}_{\mathcal{A}_{jk}} \begin{bmatrix} p_{jk} \\ p_{kj} \end{bmatrix} \tag{4.86}$$

Figure 4.8.3 visualizes this linear transformation. Assume that $\mathcal{G}$ is a tree (corresponding to a distribution network). Using (4.86), one can write the reactive power constraints in terms of the active flows. It can be observed that as long as the practical condition $\frac{b_{jk}}{g_{jk}} \geq 1$ is satisfied for every line (j,k), the upper bound on the reactive power injection is a box-preserving convex constraint. This is due to the fact that each reactive power injection can be written as a linear and non-decreasing function of active flows (in light of (4.86)). This means that if the lower bounds on the reactive power injections are small enough (no matter what the upper bounds are), the OPF problem is reduced to the extended GNF problem with box-preserving coupling constraints. In this case, Theorem 23 can be invoked to conclude that the proposed convexification technique finds the optimal active-power injection vector. Similar to the previous case, once the optimal active-power injection vector is found, the optimal active and reactive flows can be uniquely extracted. It is worthwhile to mention that binding lower bounds on the reactive power injections may potentially destroy the exactness of the extended GNF problem since these constraints may not preserve the box property of the feasible region of the active-power injection vector.

Variable Voltage Magnitudes and Reactive Power

Consider the OPF problem with variable voltage magnitudes, namely $v_i^{\min} \leq |v_i| \leq v_i^{\max}$ for every node i in $\mathcal{G}$.

Definition 29. *Given an arbitrary line $(j,k) \in \mathcal{E}$, two numbers $u_j, u_k \in \mathbb{R}_+$, and an angle $\theta_{jk}^{\max} \in \mathbb{R}$, define $\mathcal{P}_{jk}(u_j, u_k, \theta_{jk}^{\max})$ as the set of all pairs (p_{jk}, p_{kj}) for which there exists an angle $-\theta_{jk}^{\max} \leq \theta_{jk} \leq \theta_{jk}^{\max}$ such that (4.85) holds after replacing $|v_j|$ and $|v_k|$ with u_j and u_k, respectively.*

We make the following assumptions:

- The set $\mathcal{P}_{jk}(u_j, u_k, \theta_{jk}^{\max})$ forms a monotonically decreasing curve in $\mathbb{R}^2$, for every line $(j,k) \in \mathcal{E}$ and the pair $(u_j, u_k) \in [v_j^{\min}, v_j^{\max}] \times [v_k^{\min}, v_k^{\max}]$.
- For every $[u_1, ..., u_{|\mathcal{N}|}] \in [v_1^{\min}, v_1^{\max}] \times ... \times [v_{|\mathcal{N}|}^{\min}, v_{|\mathcal{N}|}^{\max}]$, the OPF problem under the additional fixed-voltage-magnitude constraints $|v_i| = u_i, i = 1, ..., |\mathcal{N}|$ is feasible.

According to the first assumption, the upper bound on the angle difference between the two endpoints of each line must ensure that only the Pareto front of the ellipse describing

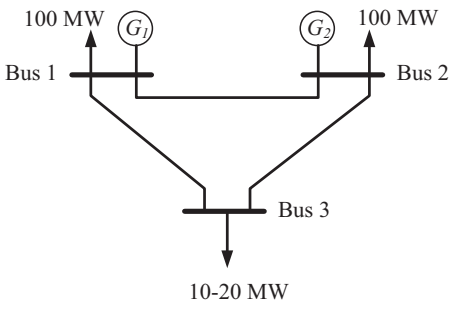


Figure 4.8.4: The three-bus power network studied in Section 4.8.

the relationship between p_{jk} and p_{kj} is feasible. Notice that for every fixed set of voltages, (4.83) ensures that the first assumption is satisfied. Furthermore, the second assumption is practical since for every node i, the limits $v_i^{\min}$ and $v_i^{\max}$ are normally chosen to be less than $5-10\%$ away from the nominal voltage magnitudes.

Observe that the OPF problem for distribution networks (or acyclic graphs $\mathcal{G}$) can be reduced to the GNF problem after fixing the magnitude of every voltage at its optimal value. Since the CGNF is exact in this case, it can be shown that there is a second-order cone programming (SOCP) relaxation of the OPF problem with variable voltage magnitudes that is exact. This conic relaxation can be regarded as the union of the CGNF problems with different fixed voltage magnitudes. The details can be found in [235]. Furthermore, this conic relaxation is exact even in presence of reactive power constraints if the inequality $\frac{b_{jk}}{g_{jk}} \geq 1$ holds for every line of the network. The main reason is that the problem reduces to the one studied in the preceding subsection after fixing the voltage magnitudes at their optimal values.

OPF for General Networks

Given two different edges (j,k) and (j',k'), the phase differences θ_{jk} and $\theta_{j'k'}$ may not be varied independently if the graph $\mathcal{G}$ is cyclic (because the sum of the phase differences over a cycle must be zero). This is not an issue if the graph $\mathcal{G}$ is acyclic (corresponding to distribution networks) or if there is a sufficient number of phase-shifting transformers in the network. If none of these cases is true, then one could add virtual phase shifters to the power network at the cost of approximating the OPF problem. The following simple example is provided to further elaborate on the effect of this approximation.

Consider the three-bus network illustrated in Figure 4.8.4 with the node set $\mathcal{N}=\{1,2,3\}$, the edge set $\mathcal{E}=\{(1,2),(2,3),(3,1)\}$, and the line admittances (y):

$$y_{12}=0.275-0.917\text{i}, \quad y_{23}=0.345-0.862\text{i}, \quad y_{31}=0.4-0.8\text{i}$$

In this network, the loads at buses 1 and 2 are fixed at the value 100MW, whereas the load at bus 3 is flexible and can accept any amount of power in the range [10MW,20MW]. For

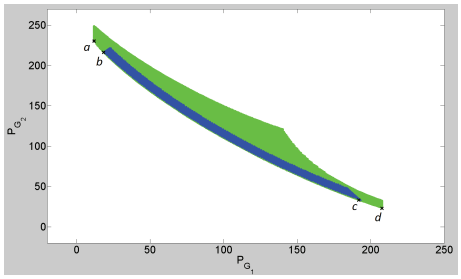


Figure 4.8.5: Feasible set $\mathcal{P}$ (blue area) and feasible set $\mathcal{P}_s$ (blue and green areas).

simplicity, assume that the voltages are fixed at their nominal values and we only consider the active powers in the system. Furthermore, suppose that $\theta_{12}^{\max} = 40°$, $\theta_{23}^{\max} = 50°$ and $\theta_{31}^{\max} = 20°$. Note that the angle constraint $|\theta_{jk}| \leq \theta_{jk}^{\max}$ can be regarded as the flow constraints $p_{jk}, p_{kj} \leq p_{jk}^{\max} = p_{kj}^{\max}$, where

$$p_{12}^{\max} = 71.29, \quad p_{23}^{\max} = 90.89, \quad p_{31}^{\max} = 37.21 \tag{4.87}$$

There are two generators in the system, whose active power outputs are denoted as P_{G_1} and P_{G_2}. Figure 4.8.5 represents the projection of the feasible set of OPF onto the space of the production vector (P_{G_1}, P_{G_2}) in two cases: (i) with no phase shifter, (ii) with a virtual phase shifter in the cycle. $\mathcal{P}$ is the feasible production region of (P_{G_1}, P_{G_2}). Define $\mathcal{P}_s$ as the projection of the feasible set of OPF problem onto the space for (P_{G_1}, P_{G_2}) after removing the angle constraint $\theta_{12} + \theta_{23} + \theta_{31} = 0$. The set $\mathcal{P}_s$ is depicted in Figure 4.8.5, which has two components: (i) the blue part $\mathcal{P}$, and (ii) the green part created by the elimination of the angle constraint. Four points have been marked on the Pareto front of $\mathcal{P}_s$ as a, b, c and d. Notice that the Pareto front of $\mathcal{P}_s$ has three segments:

- *Segment with the endpoints b and c:* This segment "almost" overlaps the Pareto front of $\mathcal{P}$. Indeed, there is a very little gap between this segment and the front of $\mathcal{P}$.
- *Segment with the endpoints a and b:* This segment extends the Pareto front of $\mathcal{P}$ from the top.
- *Segment with the endpoints c and d:* This segment extends the Pareto front of $\mathcal{P}$ from the bottom.

The gap between the Pareto front of $\mathcal{P}$ and a subset of the Pareto front of $\mathcal{P}_s$ with the endpoints b and c can be unveiled by performing some simulations. For instance, assume that $f_1(P_{G_1}) = P_{G_1}$ and $f_1(P_{G_2}) = 1.2P_{G_2}$. Two OPF problems will be solved next:

- *OPF without phase shifter:* The solution is $(P_{G_1}^{\text{opt}}, P_{G_2}^{\text{opt}}) = (144.27, 69.39)$ corresponding to the optimal cost \$227.53.

- *OPF with phase shifter:* The solution is $(P_{G_1}^{\text{opt}}, P_{G_2}^{\text{opt}}) = (145.56, 68.18)$ with $\theta_{12}^{\text{opt}} + \theta_{23}^{\text{opt}} + \theta_{31}^{\text{opt}} = 6.02^\circ$ corresponding to the optimal cost \$227.37.

Although the optimal value of the angle mismatch is not negligible, the optimal production $(P_{G_1}^{\text{opt}}, P_{G_2}^{\text{opt}})$ has very similar values in the above cases. In other words, the optimal injections obtained using the proposed convex problem are very close to the globally optimal solutions of OPF. Notice that the flows obtained from the convex problem could be completely wrong and one needs to pursue other techniques to find a set of optimal flows based on the obtained optimal injections.

The aforementioned case study offers a visual and intuitive explanation of the effect of virtual phase shifters on the optimal solution of the OPF problem and the Pareto front of the injection region. However, there is a large body of work suggesting that the inclusion of virtual phase shifters would have a small effect on the optimal solution of OPF in real-world systems [236, 79, 166, 167, 141]. Hence, the conclusion of this part is that the OPF problem with virtual phase shifters can be efficiently converted to an SOCP problem (under mild assumptions), which leads to an approximate solution for OPF (to be later rectified in a local-search solver) or can be strengthened via convex constraints accounting for omitted phase cycle effects. For example, the paper [141] proposes a strengthened SOCP to solve the OPF problem, which exhibits a great performance in many systems. The above result implies that the success of the method developed in [141] is due in part to the fact that the SOCP relaxation correctly convefixies the OPF problem with virtual phase shifters, and therefore it eliminates some of the non-convexity of the original problem.

Several works in the literature indicate that the convex relaxation of the OPF and its related problems, such as voltage regulation [145] and the state estimation [170], are exact in most practical instances. This chapter explains the reasoning behind the effectiveness of these methods by proposing a unified certificate on the exactness of these methods. In particular, it shows that these methods are successful under various conditions because the optimal solution belongs to the Pareto front of the feasible region and the proposed relaxations keep this Pareto front intact. One main application of this work is in the design of efficient algorithms for optimization over distribution networks, which is regarded as a key ingredient of future power systems, named Smart Grids. As a future work, the convexification of the GNF problem under a broader set of global coupling constraints (similar to the cycle effects in OPF) will be investigated. Another future direction is to study the GNF problem in the case where the injection and flow parameters are vectors of arbitrary dimensions, rather than scalars. This case naturally appears in multi-phased power systems, where the nodal injections (and line flows) are of dimension 1, 2 or 3. The machinery developed in this chapter suggests that the GNF problem for such networks could be convefixied through the notion of CGNF if certain monotonicity and box-preserving properties are satisfied. A detailed analysis of these types of networks is left as future work.

Chapter 5

An Efficient Method for Optimal Transmission Switching

This chapter studies the optimal transmission switching (OTS) problem for power systems, where certain lines are fixed (uncontrollable) and the remaining ones are controllable via on/off switches. The goal is to identify a topology of the power grid that minimizes the cost of the system operation while satisfying the physical and operational constraints. Most of the existing methods for the problem are based on first converting the OTS into a mixed-integer linear program (MILP) or mixed-integer quadratic program (MIQP), and then iteratively solving a series of its convex relaxations. The performance of these methods depends heavily on the strength of the MILP or MIQP formulations. In this chapter, it is shown that finding the strongest variable upper and lower bounds to be used in an MILP or MIQP formulation of the OTS based on the big-M or McCormick inequalities is NP-hard. Furthermore, it is proven that unless $P = NP$, there is no constant-factor approximation algorithm for constructing these variable bounds. Despite the inherent difficulty of obtaining the strongest bounds in general, a simple bound strengthening method is presented to strengthen the convex relaxation of the problem when there exists a connected spanning subnetwork of the system with fixed lines. With the proposed bound strengthening method, remarkable improvements in the runtime of the mixed-integer solvers and the optimality gaps of the solutions are achieved for medium- and large-scale real-world systems.

5.1 Introduction

In power systems, transmission lines have traditionally been considered uncontrollable infrastructure devices, except in the case of an outage or maintenance. However, due to the pressing needs to boost the sustainability, reliability and efficiency, power system directors call on leveraging the flexibility in the topology of the grid and co-optimizing the production and topology to improve the dispatch. In the last few years, Federal Energy Regulatory Commission (FERC) has held an annual conference on "Increasing Market and Planning

Efficiency through Improved Software" [88] to encourage research on the development of efficient software for enhancing the efficiency of the power systems via optimizing the flexible assets (e.g., transmission switches) in the system. Furthermore, The Energy Policy Act of 2005 explicitly addresses the "difficulties of siting major new transmission facilities" and calls for the utilization of better transmission technologies [89].

Unlike in the classical network flows, removing a line from a power network may improve the efficiency of the network due to physical laws. This phenomenon has been observed and harnessed to improve the power system performance by many authors. The notion of *optimally switching the lines of a transmission network* was introduced by O'Neill *et al.* [198]. Later on, it has been shown in a series of papers that the incorporation of controllable transmission switches in a grid could relieve network congestions [226], serve as a corrective action for voltage violation [17, 216, 111], reduce system loss [16, 93] and operational costs [116], improve the reliability of the system [117, 137] and enhance the economic efficiency of power markets [115]. We refer the reader to Hedman *et al.* [118] for a survey on the benefits of transmission switching in power systems. However, the identification of an optimal topology, namely optimal transmission switching (OTS) problem, is a non-convex combinatorial optimization problem that is proven to be NP-hard [157]. Therefore, brute-force search algorithms for finding an optimal topology are often inefficient. Most of the existing methods are based on heuristics and iterative relaxations of the problem. These methods include, but are not restricted to, Benders decomposition [116, 137], branch-and-bound and cutting-plane methods [92, 142], genetic algorithms [111], and line ranking [22, 97]. Recently, another line of work has been devoted to strong convexification techniques in solving mixed-integer problems for power systems [172, 86, 83].

In this work, the power flow equations are modeled using the well-known DC approximation, which is the backbone of the operation of power systems. Despite its shortcomings for the OTS in some cases [55], the DC approximation is often considered very useful for increasing the reliability, performance, and market efficiency of power systems [118]. The OTS consists of disjunctive constraints that are bilinear and nonconvex in the original formulation. However, all of these constraints can be written in a linear form using the so-called big-M or McCormick inequalities [25, 180]. This formulation of OTS is referred to as the *linearized OTS* in the sequel. A natural question arising in constructing the OTS formulation is: how can one find optimal values for the parameters of the big-M or McCormick inequalities? An optimal choice for these parameters is important for two reasons: 1) they would result in stronger convex relaxations of the problem, and hence, fewer iterations in branch-and-bound or cutting-plane methods, and 2) a conservative choice of these parameters would cause numerical and convergence issues [268]. Hedman *et al.* [117] point out that finding the optimal values for the parameters of the linearized OTS may be cumbersome, and, therefore, they impose restrictive constraints on the absolute angles of voltages at different buses at the expense of shrinking the feasible region. Other studies [92, 198, 116] have also used similar restrictive approaches to solve the linearized OTS.

In this work, it is proven that finding the optimal values for the parameters of the MILP or MIQP formulations of the OTS using either big-M or McCormick inequalities is NP-hard.

Moreover, it is shown that there does not exist any polynomial-time algorithm to approximate these parameters within any constant factor, unless $P = NP$. This new result adds a new dimension to the difficulty of the OTS; not only is solving the OTS as a mixed-integer nonlinear program difficult, but finding a good linearized reformulation of this problem is NP-hard as well. In order to maintain the reliability and security of the system, often a set of transmission lines are considered as fixed and the flexibility in the network topology is limited to the remaining lines. An implicit requirement is that the network should always remain connected in order to prevent islanding. One way to circumvent the islanding issue in the optimal transmission switching problem is to include additional security constraints in order to keep the underlying network connected at every feasible solution [139, 201]. However, this new set of constraints would lead to the over-complication of an already difficult problem. Therefore, in practice, many energy corporations, such as PJM and Exelon, consider only a selected subset of transmission lines as flexible assets in their network [1, 2].

In this chapter, it is proven that the OTS with a fixed connected spanning subnetwork is still NP-hard but one can find non-conservative values for the parameters of the big-M or McCormick inequalities in the linearized OTS without shrinking the feasible region or sacrificing the optimality of the obtained solution. In particular, a simple bound strengthening method is presented to strengthen the linearized formulation of the OTS. This method can be integrated as a preprocessing step into any numerical solver for the OTS. Despite its simplicity, it is shown through extensive case studies on the IEEE 118-bus system and different Polish networks that the incorporation of the proposed bound strengthening method leads to substantial speedup in the runtime of the solver. Furthermore, it is shown that while including additional constraints on the absolute values of the angles at different buses can improve the runtime of the solver, it may steer away from the optimality; this conservative approach can increase the operation cost by 7% for Polish networks.

5.2 Problem Formulation

Consider a power network with n_b buses, n_g generators, and n_l lines. This network can be represented by a directed graph, denoted by $G(\mathcal{B}, \mathcal{L})$, where $\mathcal{B}$ is the set of buses indexed from 1 to n_b and $\mathcal{L}$ is the set of lines whose directions are chosen arbitrarily and indexed as (i, j) to represent a connection between buses i and j. Denote $\mathcal{G} = \{1, 2, ..., n_g\}$ as the set of generators in the system. Furthermore, let $\mathcal{N}_g(i)$ be the indices of generators that are connected to bus i. Note that $\mathcal{N}_g(i)$ may be empty for a bus i. The variable p_i corresponds to the active-power production of generator $i \in \mathcal{G}$ and the variable θ_i is the voltage angle at bus $i \in \mathcal{B}$. For every $(i, j) \in \mathcal{L}$, the variable f_{ij} denotes the active flow from bus i to bus j. Consider the set of lines $\mathcal{S} \subseteq \mathcal{L}$ that are equipped with on/off switches and define the decision variable x_{ij} for every $(i, j) \in \mathcal{S}$ as the status of the line (i, j). Let n_s denote the cardinality of this set. We refer to the lines belonging to $\mathcal{S}$ as *flexible* lines and the remaining lines as *fixed* lines. Notice that the decision variables p_i, θ_i, and f_{ij} are continuous, whereas

x_{ij} is binary. For simplicity of notation, define the variable vectors

$$\begin{aligned}
&\mathbf{p} \triangleq [p_1, p_2, ..., p_{n_g}]^\top, \\
&\Theta \triangleq [\theta_1, \theta_2, ..., \theta_{n_b}]^\top, \\
&\mathbf{f} \triangleq [f_{i_1 j_1}, f_{i_2 j_2}, ..., f_{i_{n_l} j_{n_l}}]^\top, \\
&\mathbf{x} \triangleq [x_{i_1 j_1}, x_{i_2 j_2}, ..., x_{i_{n_s} j_{n_s}}]^\top,
\end{aligned} \tag{5.1}$$

where the lines in $\mathcal{L}$ are labeled as $(i_1, j_1), ..., (i_{n_l}, j_{n_l})$ such that the first n_s lines denote the members of $\mathcal{S}$. The objective function of the OTS is defined as $\sum_{i \in \mathcal{G}} g_i(p_i)$, where $g_i(p_i)$ takes the quadratic form $g_i(p_i) = a_i \times p_i^2 + b_i \times p_i$ with $a_i \neq 0$ or the linear form $g_i(p_i) = b_i \times p_i$, for some numbers $a_i, b_i \geq 0$. In this chapter, we consider both quadratic and linear objective functions, which may correspond to system loss and operational cost of generators. Every in-operation power system must satisfy operational constraints arising from physical and security limitations. The physical limitations include the unit and line capacities. Furthermore, the power system must satisfy the power balance equations. On the security side, there may be a cardinality constraint on the maximum number of flexible lines that can be switched off in order to avoid endangering the reliable operation of the system. Let the vector $\mathbf{d} = [d_1, d_2, ..., d_{n_b}]^\top$ collect the set of demands at all buses. Moreover, define $p_i^{\min}$ and $p_i^{\max}$ as the lower and upper bounds on the production level of generator i, and $f_{ij}^{\max}$ as the capacity of line $(i, j) \in \mathcal{L}$. Each line $(i, j) \in \mathcal{L}$ is associated with susceptance B_{ij}.

Using the above notations, the OTS is formulated as the following mixed-integer nonlinear problem:

$$\underset{\mathbf{f}, \mathbf{x}, \mathbf{\Theta}, \mathbf{p}}{\text{minimize}} \quad \sum_{i \in \mathcal{G}} g_i(p_i) \tag{5.2a}$$

$$\text{s.t.} \quad x_{ij} \in \{0, 1\}, \qquad \forall (i, j) \in \mathcal{S} \tag{5.2b}$$

$$p_k^{\min} \leq p_k \leq p_k^{\max}, \qquad \forall k \in \mathcal{G} \tag{5.2c}$$

$$-f_{ij}^{\max} x_{ij} \leq f_{ij} \leq f_{ij}^{\max} x_{ij}, \qquad \forall (i, j) \in \mathcal{S} \tag{5.2d}$$

$$-f_{ij}^{\max} \leq f_{ij} \leq f_{ij}^{\max}, \qquad \forall (i, j) \in \mathcal{L} \backslash \mathcal{S} \tag{5.2e}$$

$$B_{ij}(\theta_i - \theta_j) x_{ij} = f_{ij}, \qquad \forall (i, j) \in \mathcal{S} \tag{5.2f}$$

$$B_{ij}(\theta_i - \theta_j) = f_{ij}, \qquad \forall (i, j) \in \mathcal{L} \backslash \mathcal{S} \tag{5.2g}$$

$$\sum_{k \in \mathcal{N}_g(i)} p_k - d_i = \sum_{(i,j) \in \mathcal{L}} f_{ij} - \sum_{(j,i) \in \mathcal{L}} f_{ji}, \qquad \forall i \in \mathcal{B} \tag{5.2h}$$

$$\sum_{(i,j) \in \mathcal{S}} x_{ij} \geq r, \tag{5.2i}$$

where

- (5.2b) states that the status of each flexible line must be binary;

- (5.2c) imposes lower and upper bounds on the production level of generating units;

- (5.2d) and (5.2e) state that the flow over a flexible or fixed line must be within the line capacities when its switch is on, and it should be zero otherwise;

- (5.2f) and (5.2g) relate the flow over each line to the voltage angles of the two endpoints of the line if it is in service, and it sets the flow to zero otherwise;

- (5.2h) requires that the power balance equation be satisfied at every bus;

- (5.2i) states that at least r flexible lines must be switched on.

The reasoning behind incorporating the minimum cardinality constraint (5.2i) in the OTS is twofold:

- A small number of switching options is often essential to guarantee the practicality of different methods and a cardinality constraint on the maximum number of switchable lines is imposed to ensure this assumption [164, 22, 97].

- This lower bound is also used to guarantee the reliability of the network, especially when the switching is used as a post-contingency recourse action in the real-time operation of power systems [164, 117].

Define $\mathcal{F}$ as the feasible region of (5.2), i.e., the set of $\{\mathbf{f}, \mathbf{x}, \mathbf{\Theta}, \mathbf{p}\}$ satisfying (5.2b)- (5.2i).

Due to space restrictions, we consider only one time slot of the system operation. However, the techniques developed in this chapter can also be used for the OTS over multiple time slots with coupling constraints, such as ramping limits on the productions of the generators. As another generalization, one can consider a combined unit commitment and optimal transmission switching problem [116, 228, 229]. In this chapter, the term "optimal solution" refers to a globally optimal solution rather than a locally optimal solution.

5.3 Linearization of OTS

The aforementioned formulation of the OTS belongs to the class of mixed-integer nonlinear programs. The nonlinearity of this optimization problem is, in part, caused by the multiplication of the binary variable x_{ij} and the continuous variables θ_i and θ_j in (5.2f). However, since this nonlinear constraint has a disjunctive nature, one can use the big-M or McCormick reformulation technique to formulate it in a linear way. First, we consider the big-M method, and then show that the same result holds for the McCormick reformulation scheme in the OTS. One can re-write (5.2f) for each flexible line (i, j) in the form

$$B_{ij}(\theta_i - \theta_j) - M_{ij}(1 - x_{ij}) \leq f_{ij} \leq B_{ij}(\theta_i - \theta_j) + M_{ij}(1 - x_{ij}) \tag{5.3}$$

for a *large enough* scalar M_{ij}, which results in the *linearized OTS formulation*. The above inequality implies that if x_{ij} equals 1, then the line is in service and needs to satisfy the

physical constraint $f_{ij} = B_{ij}(\theta_i - \theta_j)$. On the other hand, if x_{ij} equals 0, then (5.3) (and hence (5.2f)) is redundant as it is dominated by (5.2d). The term "large enough" for M_{ij} is ambiguous, and indeed the design of an effective M_{ij} is a challenging task that will be studied below.

Definition 30. *For every $(i, j) \in \mathcal{S}$, it is said that M_{ij} is feasible for the OTS if it preserves the equivalence between* (5.3) *and* (5.2f) *in the OTS. The smallest feasible M_{ij} is denoted by M_{ij}^{opt}.*

Remark 9. *Note that the value of M_{ij}^{opt} is independent of the values of M_{rl}, for $(r, l) \in \mathcal{S}\backslash(i, j)$, in the linearized OTS formulation, as long as they are chosen to be feasible. In other words, given an instance of the OTS, the value of M_{ij}^{opt} is the same if M_{rl} satisfies $M_{rl} \geq M_{rl}^{\text{opt}}$ for every $(r, l) \in \mathcal{S}\backslash(i, j)$.*

The problem under investigation in this section is the following: given an instance of OTS, is there an efficient algorithm to compute M_{ij}^{opt} or a good approximation of that for every $(i, j) \in \mathcal{S}$? It is desirable to find the smallest feasible values for every $M_{ij}, (i, j) \in \mathcal{S}$, in (5.3) because of two reasons:

1. Commonly used methods for solving MILP or MIQP problems, such as cutting-plane and branch-and-bound algorithms, are based on iterative convex relaxations of the constraints. Therefore, while a sufficiently large value for M_{ij} does not change the feasible region of the OTS after replacing (5.2f) with (5.3), it may have a significant impact on the feasible region of its convex relaxation. Small values for M_{ij} yield stronger convex relaxations with smaller feasible sets.

2. Large values for M_{ij} may cause numerical issues for convex relaxation solvers.

For every $(i, j) \in \mathcal{S}$, define $\mathcal{F}_{ij}$ as the set of all points $\{\mathbf{f}, \mathbf{x}, \mathbf{\Theta}, \mathbf{p}\} \in \mathcal{F}$ such that $x_{ij} = 0$.

Lemma 32. *The equation*

$$M_{ij}^{\text{opt}} = B_{ij} \times \max_{\{\mathbf{f},\mathbf{x},\mathbf{\Theta},\mathbf{p}\}\in\mathcal{F}_{ij}} \{|\theta_i - \theta_j|\} \tag{5.4}$$

holds for every flexible line $(i, j) \in \mathcal{S}$.

Proof. Consider a number M_{ij} such that $M_{ij} \geq B_{ij} \times \max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$. Every set $\{\mathbf{f}, \mathbf{x}, \mathbf{\Theta}, \mathbf{p}\} \in \mathcal{F}$ satisfies (5.3) with the chosen M_{ij} and, hence, M_{ij} is feasible. Now, assume that $M_{ij} < B_{ij} \times \max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$. Based on the definition of the set $\mathcal{F}_{ij}$, this implies that there exists $\{\bar{\mathbf{f}}, \bar{\mathbf{x}}, \bar{\mathbf{\Theta}}, \bar{\mathbf{p}}\} \in \mathcal{F}$ such that $\bar{x}_{ij} = 0$, $\bar{f}_{ij} = 0$, and $M_{ij} < B_{ij}|\bar{\theta}_i - \bar{\theta}_j|$. Without loss of generality, suppose that $\bar{\theta}_i \geq \bar{\theta}_j$. Therefore, one can verify that

$$0 < B_{ij}(\bar{\theta}_i - \bar{\theta}_j) - M_{ij}(1 - \bar{x}_{ij}) \tag{5.5}$$

Combining with (5.3), this results in $\bar{f}_{ij} > 0$, contradicting the assumption $\bar{f}_{ij} = 0$. This completes the proof. □

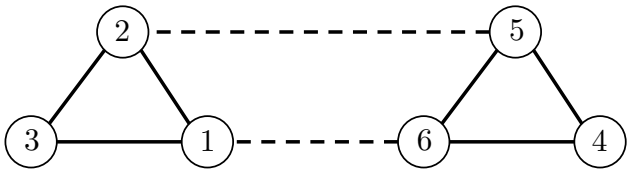


Figure 5.3.1: The topology of the network in Example 3. The solid and dashed edges denote the lines with ON and OFF switches, respectively.

Due to Lemma 32, the problem of finding M_{ij}^{opt} for every $(i,j) \in \mathcal{S}$ reduces to finding the $\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$.

Remark 10. *Note that, for a given $(i,j) \in \mathcal{S}$, the term $\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$ is finite if and only if the buses i and j are connected for every feasible point in $\mathcal{F}_{ij}$. This means that the linearization of the OTS is well-defined if and only if the power network remains connected at every feasible solution in $\mathcal{F}_{ij}$ for all $(i,j) \in \mathcal{S}$.*

The next example illustrates a scenario where the $\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$ is not finite.

Example 3. *Consider the network with 6 buses and 8 lines in Figure 5.3.1. Assume that the network is decomposed into two disjoint components (known as islands) with the buses $\{1,2,3\}$ and $\{4,5,6\}$ at a feasible point $\{\mathbf{f}, \mathbf{x}, \mathbf{\Theta}, \mathbf{p}\} \in \mathcal{F}_{16}$. Define $\tilde{\mathbf{\Theta}}$ as $\tilde{\theta}_i = \theta_i$ for $i \in \{1,2,3\}$ and $\tilde{\theta}_i = \theta_i + \tau$ for $i \in \{4,5,6\}$, where τ is an arbitrary scalar. It can be verified that $\{\mathbf{f}, \mathbf{x}, \tilde{\mathbf{\Theta}}, \mathbf{p}\} \in \mathcal{F}_{16}$ for every τ. Furthermore, $\tilde{\theta}_6 - \tilde{\theta}_1 = \theta_6 - \theta_1 + \tau$, which implies that $\max_{\mathcal{F}_{16}}\{|\tilde{\theta}_6 - \tilde{\theta}_1|\} \to +\infty$ as $\tau \to +\infty$.*

To avoid unbounded values for M_{ij}^{opt}, the existence of a connected spanning subnetwork connecting all the nodes in the network with fixed lines will be assumed in the next section. In what follows, it will be shown that, even if $\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$ is bounded for every $(i,j) \in \mathcal{S}$, one cannot devise an algorithm that efficiently finds $\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$ since it amounts to an NP-hard problem. Furthermore, the impossibility of any constant factor approximation of $\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$ in the linearized OTS is proven.

Theorem 24. *Consider an instance of the OTS together with a flexible line $(i,j) \in \mathcal{S}$, where $f_{kl}^{\max}$ is a given positive number for every $(k,l) \in \mathcal{L}\backslash\mathcal{S}$. Unless $P = NP$, it holds that:*

- *(Strong NP-hardness) there is no polynomial-time algorithm for finding $\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$;*
- *(Inapproximability) there is no polynomial-time constant-factor approximation algorithm for finding $\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$.*

Proof. To prove the strong NP-hardness of the problem, it suffices to show that there exists a polynomial reduction from the *longest path* problem in unweighted graphs–a well-known

strongly NP-hard problem [58]. The longest path problem is defined as follows: Given an undirected graph $G(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ stand for the sets of vertices and edges, respectively, what is the longest simple path between two particular vertices i and j in $\mathcal{V}$? Let the length of the longest path be denoted as p^{opt}. We construct an instance of the OTS in the following way: Consider $|\mathcal{V}|$ buses and, for every $(r, l) \in \mathcal{E}$, connect buses r and l through a line with an arbitrary orientation that is equipped with a switch (note that $\mathcal{S} = \mathcal{E}$ in this case). For each line $(r, l) \in \mathcal{E}$, its susceptance and flow capacity are set to 1. For every bus $s \notin \{i, j\}$ in the system, we set $d_s = p_s^{\min} = p_s^{\max} = 0$, which implies that there is no load or generator connected to bus s. Connect a generator with $p_i^{\min} = p_i^{\max} = 1$ to bus i. Furthermore, connect a load $d_j = 1$ to bus j. Finally, set $r = 0$.

The instance designed above is feasible if and only if there is a simple path between buses i and j in G. Furthermore, the size of the constructed instance of the OTS is polynomial in the size of the instance of the longest path problem. Denote the feasible region of the designed instance of the OTS as $\mathcal{F}$. Note that $M_{ij}^{\text{opt}} = \max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$ due to Lemma (32). Without loss of generality, we drop the absolute value in the remainder of the proof. According to the defined characteristics of the loads and generators in the system, for any feasible solution of the OTS, there should be at least one simple path from bus i to bus j consisting of only lines that are switched on. Therefore, for every $(\mathbf{f}^*, \Theta^*, \mathbf{x}^*, \mathbf{p}^*) \in \arg\max_{\mathcal{F}_{ij}}\{\theta_i - \theta_j\}$, there exists a path $\mathcal{P}^* = \{(i, v_1), (v_1, v_2), ..., (v_k, j)\}$ with $x_{rk}^* = 1$ for all $(r, k) \in \mathcal{P}^*$. This simple path is visualized in Figure 5.3.2. With no loss of generality, assume that the direction of the flow on the lines respect the directions in $\mathcal{P}^*$. Based on Figure 5.3.2, one can verify that

$$\theta_i^* - \theta_j^* = \sum_{(r,l)\in\mathcal{P}^*} (\theta_r^* - \theta_l^*) = \sum_{(r,l)\in\mathcal{P}^*} f_{rl}^* \leq \sum_{(r,l)\in\mathcal{P}^*} f_{rl}^{\max} \leq p^{\text{opt}} \tag{5.6}$$

Now, it is desirable to construct a feasible solution $(\bar{\mathbf{f}}, \bar{\Theta}, \bar{\mathbf{x}}, \bar{\mathbf{p}}) \in \mathcal{F}$ that includes a simple path with lines that are switched on from buses i to j whose length is p^{opt}. To this end, consider the instance of the longest path problem and suppose that $\mathcal{P}^{\text{opt}} = \{(i, u_1), (u_1, u_2), ..., (u_l, j)\}$ defines the longest simple path in G between nodes i and j. For every flexible line (i, j) in the corresponding instance of the OTS, we set $\bar{x}_{ij}$ to 1 if this line belongs to $\mathcal{P}^{\text{opt}}$ and set to 0 otherwise. Moreover, we set $\bar{\theta}_j$ to 0 and define $\bar{\theta}_k = p_{kj}^{\text{opt}}$ for every bus k in $\mathcal{P}^{\text{opt}}$, where p_{kj}^{opt} is the length of the unique path between buses k and j in $\mathcal{P}^{\text{opt}}$. This yields that $\bar{f}_{rl}$ is equal to 1 for every line (r, l) in $\mathcal{P}^{\text{opt}}$. Furthermore, for every flexible line (t, s) that does not belong to $\mathcal{P}^{\text{opt}}$, we set $\bar{f}_{ts}$ to 0. To satisfy (5.2h), set $\bar{p}_i = 1$. Therefore, a feasible solution $(\bar{\mathbf{f}}, \bar{\Theta}, \bar{\mathbf{x}}, \bar{\mathbf{p}})$ is constructed that satisfies the following property:

$$\theta_i^* - \theta_j^* \geq \bar{\theta}_i - \bar{\theta}_j = \bar{\theta}_i = p^{\text{opt}} \tag{5.7}$$

Inequality (5.7) together with (5.6) establishes the proof of the strong NP-hardness of finding $\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$. The inapproximability of the problem follows from the fact that, unless $P = NP$, there is no polynomial-time constant-factor approximation algorithm for determining the longest path between nodes i and j in G. □

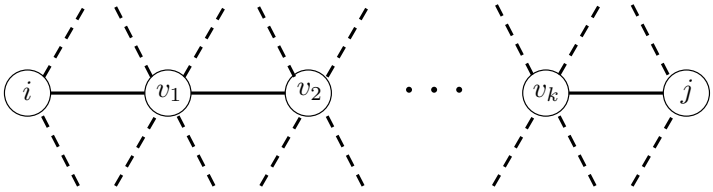


Figure 5.3.2: The visualization of the path $\mathcal{P}^*$ in the proof of Theorem 24. The solid edges denote the lines in $\mathcal{P}^*$ (with ON switches) and the dashed edges correspond to the remaining lines.

Theorem 24 together with Lemma 32 implies that finding M_{ij}^{opt} is both strongly NP-hard and inapproximable within any constant factor, hence providing a negative answer to the question raised in this section.

Remark 11. *The decision version of the OTS is known to be NP-complete [142]. One may speculate that the NP-hardness of finding the best M_{ij} for every $(i,j) \in \mathcal{S}$ may follow directly from that result. However, notice that there are some well-known problems with disjunctive constraints, such as the minimization of total tardiness on a single machine, which are known to be NP-hard [66] and yet there are efficient methods to find the optimal parameters of their big-M reformulation [119]. Theorem 24 shows that not only is finding the best M_{ij} for the OTS NP-hard, but one cannot hope for obtaining a strong linearized reformulation of the problem based on the big-M method.*

Note that one may choose to use McCormick inequalities [180] instead of the big-M method to obtain a linear reformulation of the bilinear constraint (5.2f). In what follows, it will be shown that the complexity of finding the optimal parameters of McCormick inequalities is the same as those in the big-M method for the OTS. The McCormick inequalities can be written in the following form for a flexible line (i,j):

$$f_{ij} \leq u_{ij|x_{ij}=1} x_{ij}, \tag{5.8a}$$
$$f_{ij} \geq l_{ij|x_{ij}=1} x_{ij}, \tag{5.8b}$$
$$f_{ij} \leq B_{ij}(\theta_i - \theta_j) - l_{ij|x_{ij}=0} x_{ij}, \tag{5.8c}$$
$$f_{ij} \geq B_{ij}(\theta_i - \theta_j) - u_{ij|x_{ij}=0} x_{ij}, \tag{5.8d}$$

where $u_{ij|x_{ij}=1}$ and $l_{ij|x_{ij}=1}$ are the respective upper and lower bounds for $B_{ij}(\theta_i - \theta_j)$ in the case where the line (i,j) is in service. Similarly, $u_{ij|x_{ij}=0}$ and $l_{ij|x_{ij}=0}$ are the respective upper and lower bounds for $B_{ij}(\theta_i - \theta_j)$ when the switch for the flexible line (i,j) is off. It can be

verified that the following equalities hold:

$$u_{ij|x_{ij}=1} = f_{ij}^{\max}, \tag{5.9a}$$

$$l_{ij|x_{ij}=1} = -f_{ij}^{\max}, \tag{5.9b}$$

$$u_{ij|x_{ij}=0} = B_{ij} \times \max_{\mathcal{F}_{ij}} \{\theta_i - \theta_j\}, \tag{5.9c}$$

$$l_{ij|x_{ij}=0} = B_{ij} \times \min_{\mathcal{F}_{ij}} \{\theta_i - \theta_j\}. \tag{5.9d}$$

Therefore, Theorem 24 immediately results in the NP-hardness and inapproximability of the pair $(l_{ij|x_{ij}=0}, u_{ij|x_{ij}=0})$.

5.4 Optimal Transmission Switching with a Fixed Connected Spanning Subgraph

In this section, we consider a power system with the property that the set of fixed lines contains a connected spanning tree of the power system. The objective is to show that a non-trivial upper bound on M_{ij} can be efficiently derived by solving a shortest path problem. Furthermore, it will be proven that this upper bound is tight in the sense that there exist instances of the OTS with a fixed connected spanning subgraph for which this upper bound equals M_{ij}^{opt}. Before presenting this result, it is desirable to state that the OTS is hard to solve even under the assumption of a fixed connected spanning subgraph.

Theorem 25. *The OTS with a fixed connected spanning subgraph is NP-hard.*

Proof. The proof is based on a reduction from subset sum problem [58] and a slight modification of the argument made in the proof of Theorem 3.1 in [142]. The details can be found in the Appendix. □

Remark 12. *Unlike Theorem 1, the statement of Theorem 25 does not imply the strong NP-hardness of the OTS problem with a fixed connected spanning subgraph since the subset sum problem is only weakly NP-hard. Instead, it implies that this problem may be efficiently solvable if the capacity and the susceptance of the lines are small. However, note that small upper bounds on the angle difference between two neighboring buses does not directly translate into small line capacities. To illustrate, assume that $|\theta_i - \theta_j|$ is upper bounded by 25 degrees (≈ 0.43 radians) for a fixed line (i, j), which means that the capacity of this line is equal to $0.43B_{ij}$. Therefore, despite having a small value for the angle difference, a large susceptance will lead to a large capacity, thereby rendering the OTS problem difficult to solve. Indeed, we have observed for Polish systems that the susceptance of some lines can be as large as $16,667$ per unit, which clearly cancels the positive effect of small angle differences.*

Consider a feasible point $\{\mathbf{f}, \mathbf{x}, \boldsymbol{\Theta}, \mathbf{p}\} \in \mathcal{F}$. For any line $(i, j) \in \mathcal{L}$, we have

$$B_{ij}(\theta_i - \theta_j) = B_{ij} \sum_{(r,l) \in \mathcal{P}_{ij}} (\theta_r - \theta_l), \tag{5.10}$$

where $\mathcal{P}_{ij}$ is an arbitrary path from node i to node j in the fixed spanning connected subgraph of G. Together with Lemma 32, this implies that

$$\begin{aligned} M_{ij}^{\text{opt}} &= B_{ij} \times \max_{\{\mathbf{f},\mathbf{x},\boldsymbol{\Theta},\mathbf{p}\}\in\mathcal{F}_{ij}} \{|\theta_i - \theta_j|\} \\ &= B_{ij}|\theta_i^{\text{opt}} - \theta_j^{\text{opt}}| \\ &\leq B_{ij} \sum_{(r,l)\in\mathcal{P}_{ij}} |\theta_r^{\text{opt}} - \theta_l^{\text{opt}}| \\ &\leq B_{ij} \sum_{(r,l)\in\mathcal{P}_{ij}} \frac{f_{rl}^{\max}}{B_{rl}}, \end{aligned} \tag{5.11}$$

where $\{\mathbf{f}^{\text{opt}}, \Theta^{\text{opt}}, \mathbf{x}^{\text{opt}}, \mathbf{p}^{\text{opt}}\} \in \arg\max_{\mathcal{F}_{ij}}\{|\theta_i - \theta_j|\}$. Note that (5.11) holds for every path $\mathcal{P}_{ij}$ in the fixed connected spanning subgraph of the network. We will use this observation in Theorem 26 to derive strong upper bounds for M_{ij}^{opt}. Denote the undirected weighted subgraph induced by the fixed lines in the power system as $G_{\mathcal{I}}(\mathcal{B}_{\mathcal{I}}, \mathcal{W}_{\mathcal{I}})$, where $\mathcal{B}_{\mathcal{I}} = \mathcal{B}$ and $\mathcal{W}_{\mathcal{I}}$ is the set of all tuples (i, j, w_{ij}) such that $(i, j) \in \mathcal{L}\backslash\mathcal{S}$ and w_{ij} is the weight corresponding to (i, j) defined as $f_{ij}^{\max}/B_{ij}$. Let $\mathcal{P}_{\mathcal{I};ij}$ and $p_{\mathcal{I};ij}$ be the set of edges in a shortest simple path between nodes i and j in $G_{\mathcal{I}}$ and its length, i.e., the summation of the weights of the edges in $\mathcal{P}_{\mathcal{I};ij}$, respectively.

Theorem 26. *For every flexible line $(i, j) \in \mathcal{S}$, the inequality*

$$M_{ij}^{\text{opt}} \leq B_{ij} \times p_{\mathcal{I},ij} \tag{5.12}$$

holds. Moreover, there exists an instance of the OTS for which this inequality is tight.

Proof. Based on (5.11), we have

$$M_{ij}^{\text{opt}} \leq B_{ij} \sum_{(r,l)\in\mathcal{P}_{\mathcal{I},ij}} \frac{f_{rl}^{\max}}{B_{rl}} = B_{ij} \sum_{(r,l)\in\mathcal{P}_{\mathcal{I},ij}} w_{rl} = B_{ij} \times p_{\mathcal{I},ij}. \tag{5.13}$$

Furthermore, a simple 3-bus system can be designed to show the tightness of the derived upper bound: consider a 3-bus network with the buses labeled as 1, 2, and 3. Assume that the lines $(1, 2)$ and $(2, 3)$ are fixed and the line $(1, 3)$ is flexible. Furthermore, suppose that the capacity and the susceptance of all lines are equal to 1. Upon connecting a generator with unit capacity ($p_1^{\max} = 1$ and $p_1^{\min} = 0$) to node 1 and a unit load to node 3, one can easily certify that $M_{13}^{\text{opt}} = 2$ which in turn equals to

$$B_{13}\left(\frac{f_{12}^{\max}}{B_{12}} + \frac{f_{23}^{\max}}{B_{23}}\right) = 2, \tag{5.14}$$

thereby verifying the tightness of (5.12) for this instance. □

Algorithm 3 Bound strengthening method for linearized OTS

input: $G_{\mathcal{I}}(\mathcal{B}_{\mathcal{I}}, \mathcal{W}_{\mathcal{I}})$ and $B = \{B_{ij} | (i,j) \in \mathcal{S}\}$
output: M_{ij} for every $(i,j) \in \mathcal{S}$
for $(i,j) \in \mathcal{S}$ **do**
    find $p_{\mathcal{I};ij}$ using Dijkstra's algorithm
    $M_{ij} \leftarrow B_{ij} \times p_{\mathcal{I};ij}$
end for

Theorem 26 proposes a bound strengthening scheme for every flexible line in the OTS that can be carried out as a simple preprocessing step before solving the OTS using any branch-and-bound method. The algorithm for the proposed bound strengthening method is described in Algorithm 3.

The worst-case complexity of performing this preprocessing step is $O(n_s n_b^2)$ since it is equivalent to performing n_s rounds of Dijkstra's algorithm on the weighted graph $G_{\mathcal{I}}$ (it can also be reduced to $O(n_s(n_l - n_s + n_b \log n_b))$ if the algorithm is implemented using a Fibonacci heap) [58]. This preprocessing step can be processed in an offline fashion before realizing the demand in the system. The impact of this preprocessing step on the runtime of the solver will be demonstrated on different cases in Section 5.5.

As mentioned in the Introduction, the existence of a fixed connected spanning subgraph in power systems is a practical assumption since power operators should guarantee the reliability of the system by ensuring the connectivity of the power network. Therefore, due to Theorem 26, one can design non-conservative values for M_{ij}'s in order to strengthen the convex relaxation of OTS.

Remark 13. *In practice, the angle difference between a pair of buses is tightly constrained if they are connected via a line. In other words, $|\theta_i - \theta_j|$ is constrained to be small if the line (i,j) is in service. One may conjecture that this can directly result in small values for M_{ij}^{opt}. In what follows, we will provide an easy and intuitive counterexample. Consider a 101-bus power system whose buses are labeled as $1, 2, ..., 101$. Define the set of lines as $\mathcal{L} = \{(i, i+1) | i = 1, 2, ..., 100\} \cup (101, 1)$ (note that the lines form a cycle). Furthermore, assume that all lines are fixed except for the line $(101, 1)$. Suppose that the upper bound on the angle difference between every two neighboring buses is set to 10 degrees. This implies that $|\theta_{101} - \theta_1|$ can be as large as 1000 degrees (17 radians) if $x_{101,1} = 0$ at a feasible solution of the OTS. Assume that the susceptance of the lines $(i, i+1)$ is 100 p.u. for every $i = 1, 2, ..., 100$ and the susceptance of the line $(101, 1)$ is 50 p.u.. Lemma 1 implies that $M_{ij}^{opt} \approx 1700$. Now, assume that there is a load in the amount of 17 p.u. at bus 101 and that a generator with the capacity 17 is connected to bus 1. One can easily verify that there exists a single feasible solution for the OTS in this case (independent of the objective function). Furthermore, any value for M_{ij} smaller than 1700 will cut this feasible solution and, hence, make the linearized OTS infeasible.*

Consider the cost function for the OTS. In practice, a quadratic objective function is often

used for production planning in order to model the cost of production, especially for thermal generators [265]. However, the nonlinearity introduced by a quadratic cost function makes the OTS particularly hard to solve. The main challenge of solving the MIQP is the fact that the optimal solution of its continuous relaxation often lies in the interior or on the boundary of its relaxed feasible region which may be infeasible for the original MIQP (as opposed to the extreme point solutions in MILP). More precisely, even obtaining the convex hull of the feasible region is not enough to guarantee the exactness of such continuous relaxations, since the optimal solution of the relaxed problem usually does not correspond to an extreme point in the convex hull if the objective function is quadratic. This introduces fractional solutions for the binary variables of the problem in most of the iterations of branch-and-bound methods which often leads to a high number of iterations. One way to partially remedy this problem is to reformulate the problem by introducing auxiliary variables such that a new linear function is minimized and the old quadratic objective function is moved to the constraints. This guarantees that the continuous relaxation of the reformulated problem will obtain an optimal solution that is an extreme point of the relaxed feasible region. This is a key reason behind the success of different conic relaxation and strengthening methods in MIQP [7, 13].

Assume that the objective function is quadratic in the form $\sum_{i=1}^{n_g} g_i(p_i)$, where $g_i(p_i) = a_i \times p_i^2 + b_i \times p_i$. Upon defining a new set of variables t_i for $i \in \mathcal{G}$, one can reformulate the objective function as $\sum_{i=1}^{n_g} \tilde{g}_i(p_i, t_i)$ where

$$\tilde{g}_i(p_i, t_i) = a_i \times t_i + b_i \times p_i. \tag{5.15}$$

subject to the additional convex constraints

$$p_i^2 \leq t_i, \qquad \forall i \in \mathcal{G} \tag{5.16}$$

To streamline the presentation, this problem is referred to as *conic formulation* of OTS whereas the previous formulation with quadratic objective function is called *quadratic formulation* henceforth.

5.5 Numerical Results

In this section, numerical studies on different test cases are conducted to evaluate the effectiveness of the proposed preprocessing method in solving the OTS. To this goal, we compare the proposed bound strengthening method to two different approaches:

- **Conservative approach:** In this method, the underlying structure of the power system is not exploited and a conservative value is chosen for every M_{ij}.
- **Restrictive approach:** In this method, additional constraints are imposed on the absolute value of the angles at all buses in order to obtain a small upper bound for M_{ij}'s. This comes at the expense of a shrinkage in the feasible region of the OTS and, hence, carries the risk of eliminating the globally optimal solution.

In the conservative approach, M_{ij} is chosen as $B_{ij} \sum_{(k,l)\in\mathcal{L}} f_{kl}^{\max}/B_{kl}$ for every $(i,j) \in \mathcal{S}$. This conservative value does not exploit the underlying structure of the network. There is also another upper bound on M_{ij} that does not take advantage of the underlying connectivity of the network. To describe the construction of this upper bound, for a given power network with n_b buses and n_l lines, let $\mathcal{T}$ collect the numbers $f_{kl}^{\max}/B_{kl}$ for all $(k,l) \in \mathcal{L}$ and set M_{ij} as the sum of the $n_b - 1$ largest elements in $\mathcal{T}$ multiplied by B_{ij}. First, note that this quantity is greater than or equal to $B_{ij} \times p_{\mathcal{I},ij}$ and, therefore, is a valid upper bound on M_{ij}^{opt} according to Theorem 26. Second, this number is clearly less conservative than the value $B_{ij} \sum_{(i,j)\in\mathcal{L}} f_{ij}^{\max}/B_{ij}$. However, we have observed in simulations that there is no improvement in the runtime of the solver using these upper bounds compared to the chosen values $B_{ij} \sum_{(i,j)\in\mathcal{L}} f_{ij}^{\max}/B_{ij}$. A detailed analysis of the the effect of these two upper bounds on the runtime of the solver can be found in Appendix.

Many studies on OTS in the literature use a restrictive approach and consider an additional set of constraints on the absolute value of the angles in the form of $|\theta_i| \leq \theta_i^{\max}$ in order to circumvent the issue of large values for M_{ij}'s [92, 117, 198, 116]. Under this new set of constraints, M_{ij} is upper bounded by $B_{ij}(\theta_i^{\max} + \theta_j^{\max})$. This quantity can be small if upper bounds for the absolute values of the angles are chosen to be small. However, imposing these types of constraints has no physical or safety justifications. Indeed, the stability and accuracy of the DC approximation is guaranteed by imposing strict constraints on the angle differences as opposed to the individual angles.

All of the test cases are chosen from the publicly available MATPOWER package [285, 56]. The simulations are run on a laptop computer with an Intel Core i7 quad-core 2.50 GHz CPU and 16GB RAM. The results reported in this section are for a serial implementation in MATLAB using the CVX framework and the GUROBI 6.00 solver with the default settings. The relative optimality gap threshold is defined as

$$\frac{z_{UB} - z_{LB}}{z_{UB}} \times 100,$$

where z_{UB} and z_{LB} are the objective value corresponding to the best found feasible solution and the best found lower bound, respectively. If the solver obtains a feasible solution for the OTS with the relative optimality gap of at most 0.1% within a time limit (to be defined later), it is said that an optimal solution is found.

Data Generation

First, we study the IEEE 118-bus system. There are 185 lines in this test case. In all of the considered instances, a randomly generated connected spanning subgraph of the network with 120 fixed lines is chosen and the remaining lines are considered flexible. To generate multiple instances of the OTS, the loads are multiplied by a load factor α chosen from the set $\{\alpha_1, \alpha_2, ..., \alpha_k\}$. Furthermore, a uniform line rating is considered for all lines in the system. We examine both linear and quadratic cost functions and perform the following comparisons:

- For the instances with a linear cost function, the total runtime of the solver is computed for the conservative and proposed bound strengthening methods (denoted by `L-C` and `L-P`, respectively) for different load factors and cardinality lower bounds.
- For the instances with a quadratic cost function, the runtime is computed for four different formulations: 1) the conic formulation with the proposed bound strengthening method (denoted by `C-P`), 2) the conic formulation with conservative approach (denoted by `C-C`), 3) the quadratic formulation with the proposed bound strengthening method (denoted by `Q-P`), and 4) the quadratic formulation with conservative approach (denoted by `Q-C`).

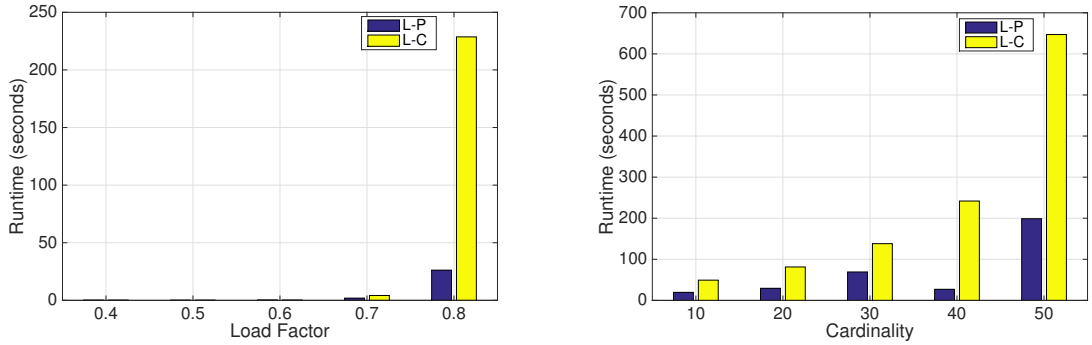
We also study six different large-scale Polish networks that are equipped with hundreds of switches. For each test case, a single load factor is considered for the OTS with linear and quadratic cost functions and the effect of the proposed bound strengthening method on the runtime and the optimality degree of the obtained solution is investigated compared to both conservative and restrictive approaches. Similar to the IEEE 118-bus case, we fix a randomly chosen connected spanning subgraph of the network with fixed lines. Similar to the previous works [92, 117, 198, 116], an upper bound of 0.6 radians (35 degrees) is chosen for the absolute value of the angles at every bus in the restrictive approach.

IEEE 118-bus System

In this subsection, the OTS is studied for the IEEE 118-bus system with 65 switches. Two types of cost functions are considered for this system:

Linear cost function: Figure 5.5.1a shows the runtime with respect to the various load factors. For all of these experiments, the lower bound on the cardinality of the ON switches is set to 45, i.e. $r = 45$ in (5.2i). It can be observed that, for small values of the load factor, the OTS is relatively easy to solve with a linear cost function and the solver can easily find the optimal solution within a fraction of second with or without the bound strengthening method. On the other hand, as the load factor increases, the OTS becomes harder to solve and the proposed bound strengthening method has a significant impact on the runtime. In particular, when the load factor equals 0.8, the strengthened formulation of the OTS is solved 8.73 faster.

In the second experiment, the performance of the solver is evaluated as a function of the lower bound on the number of the ON switches. As pointed out in [142], the OTS becomes computationally hard to solve with a relatively large cardinality lower bound. This can be a counter-intuitive observation; as this lower bound increases, the set of feasible solutions shrinks. However, a smaller feasible region does not necessarily result in fewer and faster branch-and-cut iterations. In fact, there are a number of cardinality-constrained NP-hard problems, such as *k-coverage* [123] or *subset selection in linear regression* [261], that become easy (and even trivial) when the cardinality constraint is removed from the formulation. Roughly speaking, this means that these types of constraints may shrink the feasible region,



(a) Runtime vs. load factor

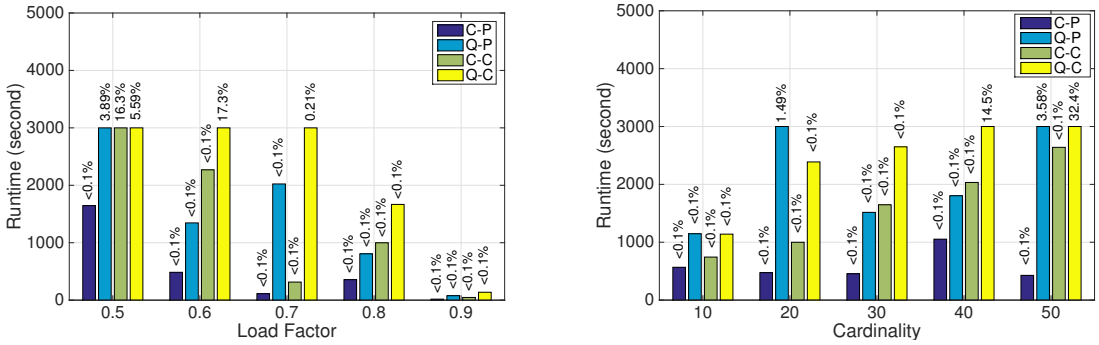
(b) Runtime vs. cardinality lower bound

Figure 5.5.1: The runtime of different formulations of OTS with a linear cost function with respect to different load factors and cardinality lower bounds. `L-C` and `L-P` correspond to the conservative and proposed bound strengthening methods, respectively.

but instead can make the enumeration process harder. This becomes more evident by noting that one of earliest results on the NP-hardness of the OTS assumes a cardinality constraint on the number of switches [33]. This behavior is observed in Figure 5.5.1b. However, note that the negative effect of increased lower bound diminishes when the bound strengthening step is performed. Specifically, the strengthened formulation is solved 2.66 times faster on average for the first two cardinality lower bounds (10 and 20) and 6.53 times faster on average for the last two cardinality lower bounds (40 and 50).

Quadratic cost function: When the cost function is quadratic, the runtime of the solver is drastically increased. Nevertheless, the modified formulation of the OTS combined with the proposed bound strengthening method reduces the runtime significantly. For all experiments, a time limit of $3,000$ seconds is imposed. For those instances that are not solved within the time limit, the relative optimality gap that is achieved by the solver at termination is reported. The runtime for different formulations of the OTS with respect to various load factors is depicted in Figure 5.5.2a. Similar to the previous case, the lower bound on the cardinality of the switches is set to 45 for different load factors. It can be observed that when the load factor equals 0.5, the solver can find the optimal solution within the time limit only for `Q-ET`. As the load factor increases, the average runtime decreases for all formulations. As it is clear from Figure 5.5.2a, `Q-ET` significantly outperforms other formulations for all load factors. Specifically, the runtime for `Q-ET` is at least 5.95, 2.96, and 13.58 times faster than `Q-OT`, `Q-EC`, and `Q-OC` on average, respectively. Notice that these values are the under-estimators of the actual speedups since the solver was terminated before finding the optimal solution in many cases.

Next, consider the runtime for different formulations with respect to the change in the cardinality lower bound of ON switches. It can be observed in Figure 5.5.2b that the solution times for `Q-OT`, `Q-EC`, and `Q-OC` increase as the lower bound increases. This observation



(a) Runtime vs. load factor

(b) Runtime vs. cardinality lower bound

Figure 5.5.2: The runtime of different formulations of OTS with a quadratic cost function with respect to different load factors and cardinality lower bounds. `C-P`, `C-C`, `Q-P`, and `Q-C` correspond to the conic formulation with the proposed bound strengthening method, the conic formulation with conservative approach, the quadratic formulation with the proposed bound strengthening method, and the quadratic formulation with conservative approach, respectively.

supports the argument made in [142] suggesting that a large lower bound on the cardinality of the ON switches would make the OTS harder to solve in general. However, notice that the cardinality constraint has a minor effect on the runtime of `Q-ET`. Notice that `Q-OC` has the worst runtime on average among different settings of the load factor and cardinality lower bound. This implies that the proposed reformulation of the objective function together with the bound strengthening step is crucial to efficiently solve the OTS with a quadratic objective function.

Polish Networks

In this part, the proposed bound strengthening method is applied to solve the OTS for Polish networks. As for the 118-bus system, the runtime is evaluated for both linear and quadratic cost functions. In all of the simulations, the cardinality lower bound on the number of ON switches is set to 0. The number of flexible lines varies from 70 to 400. The time limit is chosen as $14,400$ seconds (4 hours) for the solver. If the time limit is reached, the optimality gap of the best found feasible solution (if one exists) is reported. For the test cases with a quadratic cost function, only the modified formulation of the problem is considered because it significantly outperforms the original formulation.

Table 5.5.1 reports the performance and computational improvements when the bound strengthening method is incorporated into the formulation as a preprocessing step, compared to the conservative and restrictive approaches. This table includes the following columns:

- Cost Function: The type of the cost function used in the simulation;

			Restrictive		Proposed Method			Conservative		
Cases	Cost	# Cont.	# Binary	Time	Subopt	Pre. Time	Time	Optgap	Time	Optgap	Speedup
3120sp	Linear	3466	70	20	1.64%	< 1	477	$< 0.1\%$	3,623	$< 0.1\%$	7.60
	Quadratic	3466	70	90	1.25%	< 1	2,900	$< 0.1\%$	14,400	0.12%	4.97*
2383wp	Linear	2789	80	14	0.76%	< 1	418	$< 0.1\%$	931	$< 0.1\%$	2.23
	Quadratic	2789	80	44	0.52%	< 1	252	$< 0.1\%$	3,960	$< 0.1\%$	15.71
2736sp	Linear	3105	100	11	0.95%	< 1	80	$< 0.1\%$	188	$< 0.1\%$	2.35
	Quadratic	3105	100	15	1.14%	< 1	156	$< 0.1\%$	2,381	$< 0.1\%$	15.26
3012wp	Linear	3516	120	50	0.49%	1	2,447	$< 0.1\%$	14,400	0.11%	5.88*
	Quadratic	3516	120	162	0.23%	1	2,570	$< 0.1\%$	14,400	0.11%	5.60*
3375wp	Linear	4053	200	320	1.33%	< 1	98	$< 0.1\%$	77	$< 0.1\%$	0.79
	Quadratic	4053	200	490	2.70%	< 1	4,301	$< 0.1\%$	14,400	–	–
2746wop	Linear	3576	400	3,045	$< 0.1\%$	1	17	$< 0.1\%$	118	$< 0.1\%$	6.94
	Quadratic	3576	400	7,238	$< 0.1\%$	1	182	$< 0.1\%$	3,523	$< 0.1\%$	19.36
Average				**958**	**0.93%**	$< \mathbf{1}$	**1,158**	$< \mathbf{0.1\%}$	**6,033**	**0.1%***	**7.88***

Table 5.5.1: The performance of the solver with the proposed, conservative, and restrictive methods for Polish networks. The superscript * corresponds to the cases where the solver is terminated before finding the optimal solution due to the time limit.

- # Cont.: The number of continuous variables in the system;
- # Binary: The number of binary variables corresponding to the flexible lines in the system;
- Time: The runtime (in seconds) for solving the OTS using different formulations within the time limit;
- Subopt: The sub-optimality of the derived solution using restrictive approach. This value quantifies the distance between the cost obtained using the restrictive approach and the optimal value of the cost function found via the proposed bound strengthening method. In particular, it is defined as

$$\frac{z_R - z_{BS}}{z_{BS}} \times 100 \tag{5.17}$$

  where z_{R} and z_{BS} denote the optimal cost values of the restrictive and proposed methods, respectively. Note that the relative optimality gap threshold is still used to obtain the values of z_{R} and z_{BS};
- Pre. Time: The elapsed time of the proposed preprocessing step;
- Optgap: The relative optimality gap within the time limit. The solver is terminated when optgap is less than 0.1%;
- Speedup: The speedup in the runtime when the proposed bound strengthening method is used as a preprocessing step compared to the conservative approach.

It can be observed from Table 5.5.1 that the presented bound strengthening method can notably reduce the computation time compared to the conservative approach at no additional computational cost. In particular, the solver can be up to 19.36 times faster if the bound strengthening method is used to strengthen the formulation. Moreover, on average (excluding the case 3375wp with a quadratic cost function), the solution time is at least 7.88 times faster if the bound strengthening method is performed prior to solving the problem. For the case 3375wp with a quadratic cost function, the solver cannot obtain a feasible solution in $14,400$ seconds without bound strengthening. However, the solver can find an optimal solution within $4,301$ seconds after performing the proposed preprocessing step. The simplicity of the bound strengthening step is evident by the fact that this preprocessing step is carried out in less than 1 second in all of the experiments.

Furthermore, the solver cannot find a globally optimal solution of the OTS in most cases using the restrictive approach, due to the constraints imposed on the absolute values of the angles at all buses. In particular, the restrictive approach can increase the cost of the system operation by up to 2.70%. Furthermore, the proposed strengthened formulation results in 1% cost reduction on average, compared to the conventional restrictive approach. The runtime with the restrictive formulation is 17% less than the proposed method; however, the

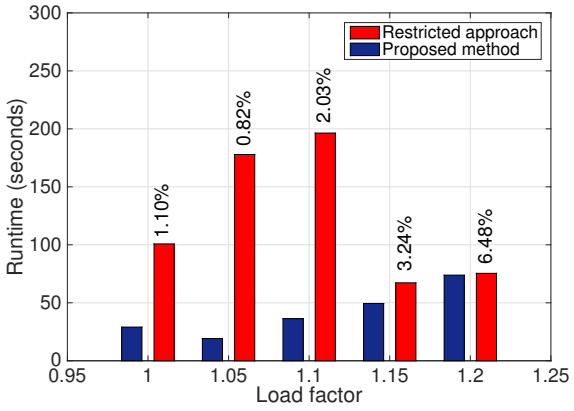


Figure 5.5.3: The runtime of the restrictive and proposed formulations, together with the sub-optimality level of the restrictive approach, for the system case3375wp under different load factors.

restrictive approach can only recover sub-optimal solutions for the OTS. In fact, the average runtime of the solver to obtain a solution with 1% (as opposed to 0.1%) relative optimality gap is only 508 seconds using the strengthened formulation.

To further elaborate on the effectiveness of the proposed strengthened formulation over the commonly used restrictive approaches, we study a modified version of the benchmark system 3375wp under different load scenarios. Similar to the previous case studies, a randomly chosen connected spanning subgraph of the network is fixed, and then 200 of the remaining lines are randomly selected and equipped with switches. We consider a linear objective for the generation cost, where the cost coefficient of each generator is chosen randomly from the interval $[20, 40]$. The load factors are chosen from the set $\{1, 1.05, 1.1, 1.15, 1.2\}$ and the line ratings are increased by 20% in order to guarantee the feasibility of the OTS for all load scenarios. It can be observed in Figure 5.5.3 that the runtime for the strengthened formulation is 66% less than that of the restrictive approach. Furthermore, it is evident that the restrictive approach results in sub-optimal solutions in all cases. In particular, the operational cost of the system with the load factor of 1.2 is increased by 6.48% when restrictive constraints are imposed on the absolute values of the angles at different buses. This clearly implies that the restrictive approach can significantly increase the operation cost in real-world networks and supports the premise of this work: *the proposed strengthened formulation strikes a good balance between the runtime of the solver and the objective of the derived solution.*

Appendix

5.A Proof of Theorem 25

In this section, the proof of Theorem 25 is provided. We show that the decision version of the OTS with a fixed spanning subgraph, which is introduced below, is NP-complete:

Decision version of OTS (D-OTS): Given an instance of the OTS and a scalar C, is there a feasible solution for the OTS problem with the cost less than or equal to C?

To prove the NP-completeness of D-OTS, we adopt the approach in [142] and introduce a reduction from the subset sum problem that is a well-known NP-complete problem [58].

Subset sum problem: Given a set of non-negative integers a_i for $i = 1, 2, \ldots, n$ and a positive integer b, is there a subset $\mathcal{I} \in \{1, 2, \ldots, n\}$ such that $\sum_{i \in \mathcal{I}} a_i = b$?

Given an instance of the subset sum problem, we produce an instance of the D-OTS and show that the subset sum problem is feasible if and only if the designed instance of the D-OTS is feasible. Consider a network with $n + 3$ buses and $2n + 2$ lines constructed according to the following procedure:

1. For every $i = 1, 2, \ldots, n$, connect bus i to buses $n + 1$ and $n + 2$ via two lines with the capacity a_i/b and the susceptance $2a_i$. Furthermore, suppose that the line $(i, n + 1)$ is fixed and the line $(i, n + 2)$ is flexible for every $i = 1, 2, \ldots, n$.

2. Connect bus $n+1$ to bus $n+3$ via a fixed line with capacity 1 and susceptance $b/(b+1)$.

3. Connect bus $n + 2$ to bus $n + 3$ via a fixed line with unit capacity and susceptance.

Figure 5.A.1 visualizes the constructed network. The cardinality lower bound r in (5.2i) is set to zero. A generator with capacity 2 is connected to bus $n + 1$ and there is a load in the amount of 2 at bus $n + 3$. Furthermore, assume that $g_{n+1}(p_{n+1})$ is zero. Finally, C (defined in the statement of D-OTS) is set to an arbitrarily chosen non-negative number. Based on this construction, the cost of every feasible solution for the OTS is zero. Therefore, addressing D-OTS reduces to verifying if the constructed instance of the OTS is feasible. First, we show that the feasibility of the subset sum problem implies the feasibility of the designed instance of the OTS. Consider a subset $\mathcal{I}$ such that $\sum_{i \in \mathcal{I}} a_i = b$. A feasible solution for the OTS is designed as follows:

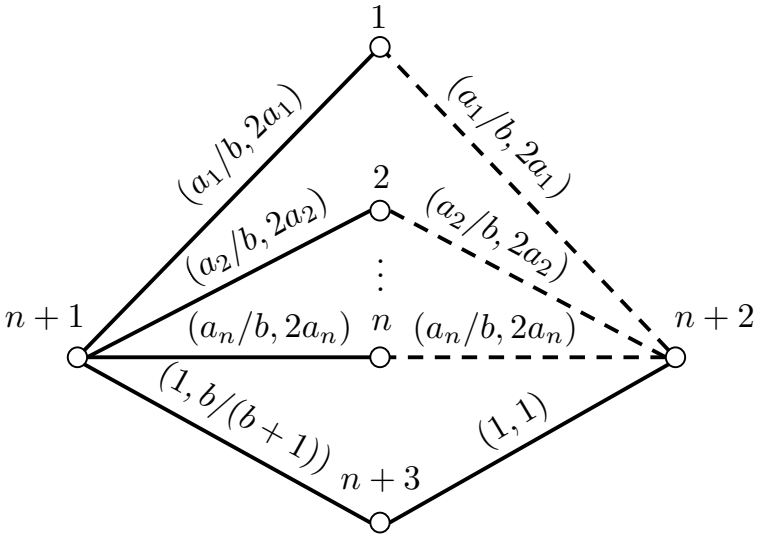


Figure 5.A.1: A visualization of the instance of D-OTS designed in the proof of Theorem 25. The solid and dashed edges denote the fixed and flexible lines, respectively. The first and the second arguments of the tuple on every line denote its capacity and susceptance, respectively.

- Set $\theta_{n+1} = 1 + \frac{1}{b}$, $\theta_{n+2} = 1$, $\theta_{n+3} = 0$, $\theta_i = 1 + \frac{1}{2b}$ for every $i \in \mathcal{I}$, and $\theta_i = 1 + \frac{1}{b}$ for every $i \notin \mathcal{I}$.
- Set $x_{i,n+2} = 1$ for every $i \in \mathcal{I}$ and $x_{i,n+2} = 0$ for every $i \notin \mathcal{I}$.

Based on the assigned values, one can easily verify that $p_{n+1} = 2$, $f_{n+1,n+3} = f_{n+2,n+3} = 1$, $f_{n+1,i} = f_{i,n+2} = \frac{a_i}{b}$ for every $i \in \mathcal{I}$, and $f_{n+1,i} = f_{i,n+2} = 0$ for every $i \notin \mathcal{I}$. Furthermore, all of the constraints in (5.2) are satisfied. This implies that the designed OTS is indeed feasible.

Next, suppose that OTS is feasible. Due to the assigned load at bus $n + 3$ and the capacity of each line, we should have $f_{n+1,n+3} = f_{n+2,n+3} = 1$. Upon setting $\theta_{n+3} = 0$, one can verify that $\theta_{n+2} = 1$ and $\theta_{n+1} = 1 + \frac{1}{b}$. On the other hand, due to the power balance constraint (5.2h) at bus $n + 2$, at least a flexible line should be in service. Denote the set of all flexible lines that are in service as $\mathcal{J}$. Given a bus i for which $(i, n + 2) \in \mathcal{J}$, one can verify that $\theta_i = 1 + \frac{1}{2b}$. To show this, note that any value for θ_i other than $1 + \frac{1}{2b}$ violates the power balance constraint (5.2h) at bus i. This, together with (5.2h) at bus $n + 2$, results in

$$\sum_{i \in \mathcal{J}} \frac{2a_i}{2b} = \sum_{i \in \mathcal{J}} \frac{a_i}{b} = 1 \tag{5.18}$$

which implies that the subset sum problem is feasible. This concludes the proof.

		Proposed Method		Conservative (1)			Conservative (2)		
Cases	Cost	Time	Optgap	Time	Optgap	Speedup	Time	Optgap	Speedup
3120sp	Linear	477	$< 0.1\%$	3,623	$< 0.1\%$	7.60	4,586	$< 0.1\%$	9.61
	Quadratic	2,900	$< 0.1\%$	14,400	0.12%	4.97*	14,400	0.13%	4.97*
2383wp	Linear	418	$< 0.1\%$	931	$< 0.1\%$	2.23	899	$< 0.1\%$	2.15
	Quadratic	252	$< 0.1\%$	3,960	$< 0.1\%$	15.71	3,080	$< 0.1\%$	12.22
2736sp	Linear	80	$< 0.1\%$	188	$< 0.1\%$	2.35	128	$< 0.1\%$	1.60
	Quadratic	156	$< 0.1\%$	2,381	$< 0.1\%$	15.26	6,166	$< 0.1\%$	39.52
3012wp	Linear	2,447	$< 0.1\%$	14,400	0.11%	5.88*	14,400	0.11%	5.88*
	Quadratic	2,570	$< 0.1\%$	14,400	0.11%	5.60*	14,400	0.11%	5.60*
3375wp	Linear	98	$< 0.1\%$	77	$< 0.1\%$	0.79	89	$< 0.1\%$	0.98
	Quadratic	4,301	$< 0.1\%$	14,400	–	–	14,400	0.15%	3.35*
2746wop	Linear	17	$< 0.1\%$	118	$< 0.1\%$	6.94	435	$< 0.1\%$	25.59
	Quadratic	182	$< 0.1\%$	3,523	$< 0.1\%$	19.36	316	$< 0.1\%$	1.74
Average		**1,158**	**$< 0.1\%$**	**6,033**	**0.1%***	**7.88***	**6,108**	**0.1%***	**9.43***

Table 5.B.1: Performance comparisons with two different conservative values for M_{ij}.

5.B Comparison Between Different Conservative Bounds

In this section, we compare the runtime of the solver when different conservative bounds are used for M_{ij}'s in the big-M reformulation of the OTS. The results for Polish networks are summarized in Table 5.B.1. In this table, Conservative (1) refers to the case where M_{ij}'s are chosen as $B_{ij} \sum_{(i,j)\in\mathcal{L}} f_{ij}^{\max}/B_{ij}$ and Conservative (2) corresponds to the case where the M_{ij} values are assigned according to the following procedure: for a given power network with n_b buses and n_l lines, let $\mathcal{T}$ collect the values of $f_{kl}^{\max}/B_{kl}$ for every line $(k,l) \in \mathcal{L}$ and set M_{ij} as the summation of the $n_b - 1$ largest elements in $\mathcal{T}$ multiplied by B_{ij}. It is observed in Table 5.B.1 that none of these conservative bounds can improve the runtime of the solver compared to the proposed strengthened bounds.

Part III

System Identification and Control

Chapter 6

Efficient Learning of Sparse Dynamical Systems

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

6.1 Introduction

Modern cyber-physical systems, such as the power grid, autonomous transportation systems, and distributed computing and sensing networks, are characterized by being large scale, spatially distributed, with complex and ever changing dynamics and interconnection topologies. The distributed optimal control literature addresses set-point tracking and regulation in this challenging setting by assuming known dynamics with a sparse interconnection topology. Indeed, this underlying sparsity structure is aggressively (and necessarily) exploited, with foundational results from the distributed control literature showing that both tractability [217] and scalability [260, 138, 87] in controller synthesis are only possible when

the underlying dynamical system is suitably sparse. However, in this large-scale, dynamic, and complex setting, it is unclear how the necessary dynamical system models are to be obtained. We expect data-driven methods to be needed to identify both the interconnection topology and dynamic behavior of these systems, as first-principle modeling becomes either intractable or impractical in these large-scale and dynamic settings.

This then raises a more fundamental question: how can data-driven methods be appropriately integrated into safety-critical control loops? This question has been addressed in the context of learning and controlling a small scale and dense unknown system, e.g., a single autonomous vehicle or robot [63, 62, 221, 207, 85, 4, 77]. These works recognize that if a learned model is to be integrated into a safety-critical control loop, then it is essential that the uncertainty associated with the learned model be explicitly quantified: in this way, the learned model and these uncertainty bounds can be integrated with tools from robust control to provide strong guarantees of system performance and stability. This chapter takes a first step towards extending these results to the large-scale distributed setting by providing a sample efficient and computationally tractable algorithm for the identification of sparse dynamical system models, as well as providing sharp estimates on the corresponding model uncertainty.

Main contributions: In particular, we show that large-scale sparse system models can be identified with complexity scaling quadratically with number of nonzero elements in the underlying dynamical system – for systems composed of a large number of subsystems that only interact with a small number of local neighbors, this computational saving can be significant. We further provide sharp bounds on the corresponding model uncertainty, paving the way for the use of these models in safety critical control loops. Finally, in contrast to previous work, we show that such models can be extracted from single trajectory of the system. In the context of large-scale systems, the system resets needed by methods relying on independent trajectories become prohibitively more expensive and impractical – indeed contrast resetting a robotic arm and a power distribution network, and the increase in difficulty becomes apparent. Note that we defer a detailed comparison of our results to prior work to Section 6.3.

6.2 Problem Formulation

Consider the linear time-invariant (LTI) system

$$x(t+1) = Ax(t) + Bu(t) + w(t) \tag{6.1}$$

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

The goal of this work is to estimate the underlying parameters of the dynamics, based on a limited number of *sample trajectories*, i.e., a sequence $\{(x^{(i)}(\tau), u^{(i)}(\tau))\}_{\tau=0}^{T}$ with $i = 1, 2, ..., d$, where d is the number of available sample trajectories and T is the length of each

sample trajectory. To simplify the notations, the superscript i is dropped from the sample trajectories when $d = 1$.

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

We now compare and contrast two approaches to collecting sample trajectories from a dynamical system (6.1):

Fixed d and variable T: In this method, the number of sample trajectories d is set to a fixed value (e.g., $d = 1$) and instead, a sufficiently long time horizon (also referred to as learning time) T is chosen to collect enough information about the dynamics. This approach is most suitable when the open-loop system is stable, or if a stabilizing controller is provided—note that this assumption of stability is necessary, as even a simple least-squares estimator may not be consistent if the system has unstable modes [221]. From a practical perspective, system instability may also impose limits on how large the learning time can be in order to ensure system safety, thereby restricting the amount of data that can be collected.

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

This work focuses on *sparse* system identification using a single trajectory, where it is assumed that the system is either stable, or equipped with an initial stabilizing controller, and our goal is to both identify the supports of the sparse system matrices (A, B) and estimate their values, using a single sample trajectory. As mentioned in [63], in many applications, the existence of an initial stabilizing controller for the unknown system (6.1) is not restrictive. In fact, [62] and [77] respectively introduce offline and adaptive procedures for designing such an initial stabilizing controller.

Indeed, one can cast the sparse system identification task as a *supervised learning* problem, where the goal is to fit the linear model (6.1)—parameterized by (A, B)—to a limited number of measurements $\{(x(\tau), u(\tau))\}_{\tau=0}^{T}$. Motivated by this observation, one can consider the following M-estimator:

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \frac{1}{2T} \sum_{t=0}^{T-1} \|x(t+1) - (Ax(t) + Bu(t))\|_2^2 + \lambda(\|A\|_1 + \|B\|_1). \tag{6.2}$$

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

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

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

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

To further analyze the properties of the above estimator, one can write (6.1) in a compact form. Let $\Psi^* = \begin{bmatrix} A & B \end{bmatrix}^\top$ denote the true parameters of the system. Furthermore, define

$$Y = \begin{bmatrix} x(1)^\top \\ \vdots \\ x(T)^\top \end{bmatrix}, X = \begin{bmatrix} x(0)^\top & u(0)^\top \\ \vdots & \vdots \\ x(T-1)^\top & u(T-1)^\top \end{bmatrix}, W = \begin{bmatrix} w(0)^\top \\ \vdots \\ w(T-1)^\top \end{bmatrix}. \tag{6.3}$$

The system identification problem is then reduced to estimating the unknown parameter Ψ^* given the *design matrix* X, and the *observation matrix* Y that is corrupted with the *noise matrix* W. We can therefore rewrite optimization problem (6.2) compactly as

$$\hat{\Psi} = \arg\min_{\Psi} \frac{1}{2T} \|Y - X\Psi\|_F^2 + \lambda \|\Psi\|_1 \tag{6.4}$$

which corresponds to the so-called *Lasso* estimator, initially popularized in statistics and machine learning to estimate the support parameter values of a sparse linear model [245]. The non-asymptotic properties of this estimator have been widely studied in the literature [256, 181, 281], all highlighting its sub-linear sample complexity under suitable technical conditions. In particular, they show that under the so-called *mutual incoherency* of the design matrix and the sparsity of the unknown parameters, the minimum number of observations for the accurate estimation of the Lasso scales logarithmically in the dimension of Ψ. Motivated by these results, one may speculate that the proposed estimator (6.2) benefits from a similar logarithmic sample complexity. However, the validity of the derived non-asymptotic estimation error bounds on the Lasso is contingent upon a number of assumptions on the independence between the design matrix X and the noise matrix W [256, 196]; such assumptions do not necessarily hold in the sparse system identification problem, partly due to the dependency between the states, the inputs and the disturbance noise. The problematic nature of this dependency becomes more evident by noting that the Lasso may not be consistent when the design and noise matrices are dependent [75].

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

Is the estimator (6.2) *consistent, and if so, what is its sample complexity?*

6.3 Statistical Guarantees

Despite the fact that in general, the Lasso may not be a consistent estimator when the design and noise matrices are dependent, we exploit the underlying structure of the system identification problem to control this dependency and provide an affirmative answer to the posed question. In other words, we show that not only is the proposed estimator (6.2) consistent, but that it also enjoys a logarithmic sample complexity in the state and input dimensions, under appropriate conditions. To this goal, we first provide a number of definitions.

Definition 31. *A zero-mean (centered) random variable x is **sub-Gaussian** with parameter b if its moment generating function satisfies*

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

for every t.

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

Definition 32. *Given a sub-Gaussian random variable x, its **sub-Gaussian norm**, denoted by $\|x\|_\psi$ is defined as the smallest $r > 0$ such that the inequality $\mathbb{E}\{x^2/r^2\} \leq 2$ is satisfied.*

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

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

[1] This is a standard result; see [214] and [255] for a simple proof.

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

Remark 14. *Most of the existing results on the sample complexity of the system identification problem assume a centered Gaussian distribution for the input noise [207, 85, 62]. Despite having desirable finite-time properties, these types of Gaussian inputs may jeopardize the safety of the dynamical system due to their unbounded range. Accordingly, in many control systems, the input is constrained to have a limited power. These types of constraints can be translated into ℓ_∞ or ℓ_2 bounds on the input signal. Due to the fact that such bounded random signals are sub-Gaussian, our results are readily applied to system identification problems with input constraints.*

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

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

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

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

$$(A + BK_0)Q^*(A + BK_0)^\top - Q^* + \sigma_w^2 I + \sigma_v^2 BB^\top = 0 \tag{6.5}$$

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

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

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

Assumption 3. *The following inequalities are satisfied*

A1 (Mutual incoherence)

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

A2 (Bounded eigenvalue)

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

A3 (Bounded infinity norm)

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

A4 (Nonzero gap)

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

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

Next, we present the main result of this chapter.

Theorem 27. *Assume that* $k\ge 2$ *and*

$$\lambda=c_1\cdot\frac{C}{1-\rho}\cdot\frac{\eta^2}{\gamma}\sqrt{\frac{\log((n+m)/\delta)}{T}} \tag{6.6}$$

$$T\ge c_2\cdot\frac{C^4}{(1-\rho)^4}\cdot\frac{D^2_{\max}}{\gamma^2C^2_{\min}\Psi^2_{\min}}\cdot k^2\log((n+m)/\delta), \tag{6.7}$$

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

1. *(Correct sparsity recovery)* (6.4) *has a unique solution and recovers the true sparsity pattern of* Ψ^*.
2. *(ℓ_∞-norm error) We have*

$$\|\hat{\Psi}-\Psi^*\|_\infty\le c_3\cdot\frac{C}{1-\rho}\cdot\frac{D_{\max}\eta^2}{\gamma}\sqrt{\frac{\log((n+m)/\delta)}{T}} \tag{6.8}$$

where c_3 *is a universal constant.*

Remark 15. *As mentioned before, the injection of a random input noise is essential to guarantee the identifiability of the parameters. This is also reflected in the above theorem: in order to guarantee a finite sample complexity for the proposed estimator, it is crucial to have $C_{\min} > 0$, which is only possible if $\sigma_v > 0$.*

A number of observations can be made based on Theorem 27. First, it implies that if γ, C, $D_{\max}$, $C_{\min}$, $\Psi_{\min}$, and ρ do not scale with the system dimension, then $T = \Omega(k^2 \log(n+m))$ is enough to guarantee the correct sparsity recovery and a small estimation error. Notice that for sparse systems, this quantity can be much smaller than the system dimension. Second, the sample complexity of the proposed estimator depends on $\frac{C}{1-\rho}$, which is a measure of the system stability. In particular, for highly stable systems, $\frac{C}{1-\rho}$ is small, resulting in an improved accuracy of the proposed estimator with smaller T. In contrast, when the system is close to its stability margin, $\frac{C}{1-\rho}$ will grow which negatively affects the estimation error as well as the lower bound on T. Another intuitive interpretation of $\frac{C}{1-\rho}$ is that it measures the amount of *dependency* between the states at different times: for highly stable systems where ρ is small, $(x(t), u(t))$ is only weakly dependent on $(x(\tau), u(\tau))$ for $\tau = 0, \dots, t-1$, thereby facilitating the estimation of the unknown parameters. We finally mention that this dependency is in contrast with the recent discoveries on the sample complexity of the least-squares estimator, which support the favorable effect of a large ρ on the accuracy of the estimator [233]. We leave investigating whether this seemingly contradictory observation is an artifact of our methodology (e.g., mixing the initial state to the stationary distribution), or is fundamental to the sparse system identification problem, to future work.

Remark 16. *In order to further enhance the accuracy of the proposed estimator, one can perform a least-squares estimation restricted to the nonzero elements of the estimated parameter, after obtaining its sparsity pattern via the proposed method. Although, theoretically, this post-model-selection estimation method may not improve the estimation error rate, it will incur less bias [24]. We will show in our simulations that the effect of this post-processing step can be significant in the accuracy of the estimation.*

Comparison to prior art

As mentioned before, another line of work focuses on unstructured system identification, where either the learning time T or the number of sample trajectories d is allowed to grow. In [62], the authors consider the sample complexity of the system identification problem with multiple sample trajectories via least-squares, where it is shown that the proposed estimator incurs a small error, provided that $d = \Omega(n+m)$. Revisiting (6.20) reveals that the proposed method outperforms the sample complexity of ordinary least-squares when k is significantly smaller than $n+m$, i.e., exploiting prior knowledge of the system sparsity leads to a reduction in sample complexity. In [221, 233, 4, 77], the authors consider unstructured system identification from a single sample trajectory under different assumptions on system stability and/or the initial state of the system. However, similar to [62], none of these works

take advantage of the underlying sparsity structures of the system matrices. As a result, they cannot correctly estimate the sparsity structure of (A, B) and suffer from poor dependencies on the system dimensions in the large-scale and structure setting.

Subsequently, a Lasso-type estimator is proposed in [85] to further exploit the underlying sparsity pattern of (A, B) with d sample trajectories, each with a zero initial state. In particular, it is shown that $d = \Omega\left(\frac{\kappa(\Sigma)^2}{\gamma^2 \Psi_{\min}^2} k \log(n+m)\right)$ is enough to ensure the correct sparsity recovery and a small estimation error with high probability, where $\kappa(\Sigma)$ is the condition number of the finite-time *controllability matrix* of the system. Comparing this quantity with (6.20), one can observe that the former has a better dependency on k. However, $\kappa(\Sigma)$ is highly dependent on the learning time T. In fact, it is easy to show that for unstable systems, $\kappa(\Sigma)$ may grow exponentially fast with respect to T. On the other hand, (6.20) is free of such dependency and instead, it is in terms of the stationary distributions of the state and input vectors.

Moreover, our work is a major extension to the results of [207], where the authors address a similar sparse system identification problem with a single sample trajectory. First, unlike the presented results, [207] only considers autonomous systems, i.e., systems (6.1) with B=0. Second, [207] only ensures the correct sparsity recovery of the true parameters. In contrast, we extend these results to obtain non-asymptotic bounds on the estimation error. As demonstrated in [62, 63], having these bounds is essential for the design of near-optimal and robustly stabilizing controllers. Third, [207] requires that the closed-loop system be contractive with respect to the spectral norm, i.e., that $|||(A + BK_0)||| < 1$, whereas we only require system stability. Notice that the former condition is much stronger, as in practice, stable systems are often not contractive in spectral norm. Finally, the validity of the non-asymptotic bounds introduced in [207] heavily relies on the Gaussian nature of the disturbance and input noises. As an extension to this result, our proposed method targets a larger class of uncertainties for the disturbance and input noises, thereby allowing for norm bounded disturbance and input signals.

Mutual incoherency

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

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

system:

$$\min_{\Psi} \ \|\Psi\|_0 \tag{6.9a}$$

$$\text{s.t.} \ \ X\Psi = Y - W \tag{6.9b}$$

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

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

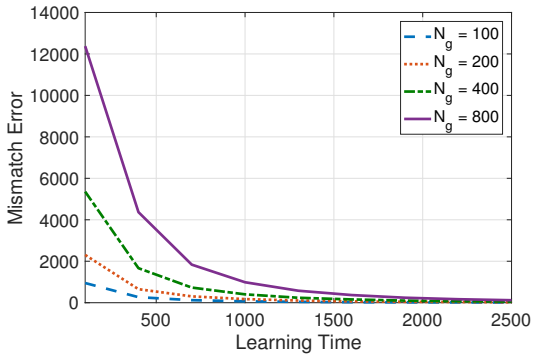
In other words, $\mu(A)$ measures the maximum correlation between distinct columns of A. Reminiscent of the classical results in the compressive sensing literature, it is well-known that the optimal solution Ψ^* of (6.9) is unique if the following *identifiability* condition

$$\|\Psi_{:,j}^*\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right) \tag{6.10}$$

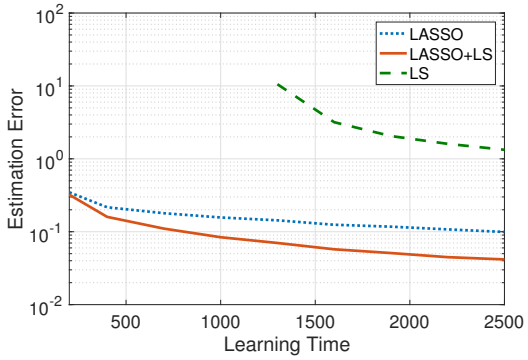
holds for $j = 1, 2, ..., n$ (see, e.g., Theorem 2.5 in [73]). Furthermore, this bound is tight, implying that there exists an instance of the problem for which the violation of $\|\Psi_{:,j}^*\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(X)}\right)$ for some j results in the non-uniqueness of the optimal solution. On the other hand, according to Lemma 35 (to be introduced later) and the Borel-Cantelli lemma, $\frac{1}{T}X^\top X$ converges to M^* almost surely, as $T \to \infty$. This implies that

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

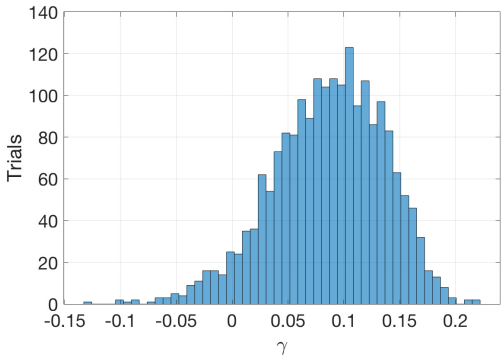
The above analysis reveals that the off-diagonal entries of M^* play a crucial role in the identifiability of the true parameters: as these elements become smaller relative to the diagonal entries, the oracle estimator can correctly identify the structure of Ψ for a wider range of sparsity levels. Similarly, our proposed mutual incoherence assumption is expected to be satisfied when the off-diagonals of M^* have small magnitudes, relative to the diagonal entries. This implies that Assumption A1 is a natural condition to impose in order to ensure the correct sparsity recovery of Ψ. Furthermore, in practice, M^* will be close to a diagonally dominant matrix with exponentially decaying off-diagonal entries, provided that the matrices A, B, and K_0 have sparse structures [234].



(a) Relative mismatch error



(b) Normalized estimation error



(c) The distribution of γ

Figure 6.4.1: (a) The mismatch error with respect to the learning time for different number of generators in the system. The values are averaged over 10 independent trials. (b) The normalized estimation error for Lasso (abbreviated as `LASSO`), Lasso + least-squares (abbreviated as `LASSO+LS`), and least-squares (abbreviated as `LS`) estimators with respect to the learning time. The values are averaged over 10 independent trials. (c) The distribution of mutual incoherence parameter γ for 2000 randomly generated instances of the problem.

6.4 Numerical Results

As a case study, we consider the frequency control problem for power systems, where the goal is to control the governing frequency of the entire network, based on the so-called *swing* equations. Assume that there exist N_g generators in the system. It is common to describe the per-unit swing equations using the well-known direct current (DC) approximation:

$$M_i\ddot{\theta}_i + D_i\dot{\theta}_i = P_{M_i} - P_{E_i}$$

where θ_i is the voltage angle at generator i, P_{M_i} is the mechanical power input at generator i, and P_{E_i} denotes the active power injection at the bus connected to generator i. Furthermore, M_i and D_i are the inertia and damping coefficients at generator i, respectively. Under the DC approximation, the relationship between active power injection and voltage is defined as follows:

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

where n is the number of generators in the network, $\mathcal{N}_i$ collects the neighbors of generator i, and B_{ij} is the susceptance of the line (i, j). After discretization with the sampling time dt, the system of swing equations is reduced to the following dynamical system:

$$x_i(t+1) = \left(A_{ii}x_i(t) + \sum_{j\in\mathcal{N}_i} A_{ij}x_j(t)\right) + B_{ii}u_i(t) + w_i(t)$$

where $x_i = \begin{bmatrix}\theta_i & \dot{\theta}_i\end{bmatrix}^\top$, $u_i(t) = P_{M_i}$, and

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

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

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

Furthermore, the sampling time dt is set to 0.1. We assume that the disturbance noise has a zero-mean Gaussian distribution with covariance $0.01I_{2\times2}$. Notice that the magnitude of the noise is comparable to those of the nonzero elements in A and B. Furthermore, the

mechanical input is set to $u_i(t) = -0.1(\theta_i + \dot{\theta}_i) + v_i(t)$, where $v_i(t)$ is a randomly generated input noise, distributed according to a zero-mean Gaussian distribution with variance 0.05. Notice that the first term in the input signal is used to ensure the closed-loop stability.

The reported results are for a serial implementation in MATLAB R2017b, and the function `lasso` is used to solve (6.2). It is worthwhile to note that the running time can be further reduced via parallelization; this is trivially possible due to the decomposable nature of the problem. The *mismatch error* is defined as the total number of false positives and false negatives in the sparsity pattern of the estimated parameters $(\hat{A}, \hat{B})$. Furthermore, *relative learning time* (RLT) is defined as the learning time normalized by the dimension of the system, and *relative mismatch error* (RME) is used to denote the mismatch error normalized by the total number of elements in A and B. In all of our experiments, the regularization coefficient λ is set to $\lambda = \sqrt{\frac{0.03 \log(n+m)}{T}}$. Note that this value does not require any additional fine-tuning and is at most a constant factor away from (6.6).

Figure 6.4.1a illustrates the mismatch error (averaged over 10 different trials) with respect to the learning time T and for different number of generators N_g that are chosen from $\{100, 200, 400, 800\}$. These correspond to the total system dimensions of $\{300, 600, 1200, 2400\}$. Note that the largest instance has more than 3.84 million unknown parameters. Not surprisingly, the learning time needed to achieve a small mismatch error increases as the dimension of the system grows. Conversely, a smaller value for RLT is needed to achieve infinitesimal RME for larger systems. In particular, when N_g is equal to 100, 200, 400, and 800, the minimum RLT to guarantee RME $\leq 0.1\%$ is equal to 3.83, 1.42, 0.50, and 0.16, respectively.

As mentioned before, the accuracy of the proposed estimator can be improved by additionally applying the least-squares over the nonzero elements of $(\hat{A}, \hat{B})$. Figure 6.4.1b illustrates the normalized 2-norm estimation error of this approach (abbreviated as `LASSO+LS`), compared to the proposed method without any post-processing step (abbreviated as `LASSO`), and the least-squares estimator (abbreviated as `LS`) when N_g is set to 200. It can be observed that both `LASSO+LS` and `LS` significantly outperform `LS`; in fact, `LS` is not even well-defined if the learning time is strictly less than the system dimensions. Furthermore, on average, the estimation error for `LASSO+LS` is 1.91 times smaller than that of `LASSO`.

Finally, only 32 out of 360 generated instances did not satisfy the proposed mutual incoherence condition. However, this violation did not have a significant effect on the accuracy of the proposed estimator. To further investigate the frequency of the instances that satisfy this condition, we plot the histogram of the mutual incoherence parameter γ for 2000 randomly generated instances with fixed $N_g = 200$. It can be seen in Figure 6.4.1c that the mutual incoherence condition is violated only for 5.15% of the instances.

Appendix

6.A Proof of the Main Theorem

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

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

and

$$\hat{\Psi}_{:,j} = \arg\min \frac{1}{2T} L(\Psi_{:,j}) + \lambda \|\Psi_{:,j}\|_1 \tag{6.11}$$

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

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

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

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

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

Lemma 33 (Proposition 4.1 [207])**.** *Suppose that the following conditions are satisfied:*

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

Then, (6.11) *recovers the true sparsity pattern of* $\Psi^*_{:,j}$.

The first step in proving Theorem 27 is to verify that the conditions of Lemma 33 hold with high probability. To this goal, first we write $x(t)$ and $u(t)$ in terms of $x(0)$, $w(\tau)$ and $v(\tau)$ for $\tau = 0, 1, \ldots, t$:

$$x(t) = (A + BK_0)^t x(0) + \sum_{\tau=0}^{t-1} (A + BK_0)^{t-\tau-1}(w(\tau) + Bv(\tau))$$

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

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

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

and

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

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

$$x(t) = \lim_{T_0 \to \infty} \sum_{\tau=-T_0-1}^{t-1} (A + BK_0)^{t-\tau-1}(w(\tau) + Bv(\tau))$$

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

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

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

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

Theorem 28 (Hanson-Wright inequality [218])**.** *Let $x = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix}$ be a random vector with independent zero-mean sub-Gaussian elements. Given a square and symmetric matrix P, the following inequality holds*

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

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

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

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

provided that $t \leq \left(\frac{\sum_k \lambda_k^2}{\max_k |\lambda_k|}\right)\|z\|_\psi^2$. The assumptions of Lemma 33 can be seen to hold directly as a consequence of the following two lemmas:

Lemma 34. *Let $i \in \{1, 2, ..., n+m\}$ and suppose that $\epsilon < \frac{3C\eta^2}{1-\rho}$. Then, there exists a universal constant c_4 such that*

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

Proof. See Appendix 6.B. □

Lemma 35. *Let $i, j \in \{1, 2, ..., n+m\}$ and suppose that $\epsilon \leq \frac{4C^2\eta^2}{(1-\rho)^2}$. Then, there exists a universal constant c_5 such that*

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

Proof. See Appendix 6.B. □

The following proposition shows that for a fixed column j, the proposed estimator (6.11) correctly recovers the sparsity pattern with high probability.

Proposition 6. *Assume that $k \geq 2$ and the following conditions are satisfied:*

$$\lambda = c_6 \cdot \sqrt{\frac{C^2\eta^4}{\gamma^2 T(1-\rho)} \log(n + m/\delta)} \tag{6.12}$$

$$T \geq c_7 \cdot \frac{C^4\eta^4 k^2}{\gamma^2 C_{\min}^2 \Psi_{\min}^2 (1-\rho)^4} \log(n + m/\delta) \tag{6.13}$$

for universal constants $c_6, c_7 \geq 0$. Then, (6.11) recovers the true sparsity pattern of $\Psi_{:,j}^*$ with probability of at least $1 - \delta$.

Proof. The Lemmas 34 and 35 can be used to prove statement. The details are provided in Appendix 6.B. □

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

Lemma 36. *Assume that*

$$\left|\left|\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\right|\right|_\infty \leq \frac{\min\{1, 2\eta^2\}}{2D_{\max}} \tag{6.14}$$

and (6.11) *recovers the correct sparsity pattern of* $\Psi^*_{:,j}$*. Then, the following inequality holds for* $E = \hat{\Psi}_{:,j} - \Psi^*_{:,j}$*:*

$$\begin{aligned} &E_{\mathcal{A}_j^c} = 0 \\ &\|E_{\mathcal{A}_j}\|_\infty \leq \left(2D^2_{\max}\left|\left|\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\right|\right|_\infty + D_{\max}\right)\left(\|G_{\mathcal{A}_j}\|_\infty + \lambda\right) \end{aligned} \tag{6.15}$$

Proof. See Appendix 6.B. □

The next lemma shows that the condition of Proposition 36 holds with high probability, provided that T is large enough.

Proposition 7. *Assume that*

$$T \geq c_8 \cdot \frac{D^2_{\max} C^4}{(1-\rho)^4} k^2 \log(k/\delta) \tag{6.16}$$

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

$$\left|\left|\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\right|\right|_\infty \leq \frac{\min\{1, 2\eta^2\}}{2D_{\max}} \tag{6.17}$$

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

$$\mathbb{P}\left(\left|\left|\left| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\right|\right|_\infty > \epsilon\right) \leq 2k^2 \exp\left(-c_5 \cdot \frac{(1-\rho)^4}{C^4\eta^4}\frac{T}{k^2}\epsilon^2\right) \tag{6.18}$$

provided that $\frac{\epsilon}{k} \leq \frac{4C^2\eta^2}{(1-\rho)^2}$. Setting $\epsilon = \frac{\min\{1,2\eta^2\}}{2D_{\max}}$ and recalling that $D_{\max}, C \geq 1$, one can verify that $\frac{\epsilon}{k} \leq \frac{4C^2\eta^2}{(1-\rho)^2}$ is satisfied. Furthermore, by choosing $c_8 = \frac{16}{c_5}$, one can certify that (6.16) is enough to ensure that the right hand side of the above inequality is upper bounded by δ, thereby completing the proof. □

Proof of Theorem 27: First note that (6.4) can be decomposed into n disjoint sub-problems over different columns of Ψ, each in the form of (6.11). Consider the following choices for λ and T:

$$\lambda = c_6 \cdot \sqrt{\frac{C^2\eta^4}{\gamma^2 T(1-\rho)^2}\log(4(n+m)/\delta)} \tag{6.19}$$

$$T \geq \max\left\{c_7, c_8, \frac{1}{c_4}, \frac{2}{c_5}\right\} \cdot \frac{C^4 D^2_{\max} k^2}{\gamma^2 C^2_{\min}\Psi^2_{\min}(1-\rho)^4}\log((n+m)/\delta) \tag{6.20}$$

where c_4, c_5, c_6, c_7, and c_6 are introduced in Lemmas 34, 35, and Propositions 6, 7. Based on the Proposition 6 and the above choices for λ and T, (6.11) recovers the sparsity pattern of $\Psi^*_{:,j}$ for a given column index j with probability of at least $1-\delta$. Furthermore, based on Proposition 7, the lower bound on T guarantees that the inequality

$$\left|\!\left|\!\left| Q_{\mathcal{A}_j,\mathcal{A}_j} - Q^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty \leq \frac{\min\{1, 2\eta^2\}}{2D_{\max}} \tag{6.21}$$

holds with probability of at least $1-\delta$. This, together with Proposition 36 results in

$$\|E_{:,j}\|_\infty \leq \left(2D^2_{\max}\left|\!\left|\!\left| Q_{\mathcal{A}_j,\mathcal{A}_j} - Q^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty + D_{\max}\right)\left(\|G_{\mathcal{A}_j}\|_\infty + \lambda\right) \tag{6.22}$$

with probability of at least $1-2\delta$. Now, it suffices to obtain concentration bounds for different terms of the above inequality. Based on (6.18) and Lemma 34, one can write

$$\mathbb{P}\left(\|G_{\mathcal{A}_j}\|_\infty > \epsilon_1\right) \leq \exp\left(\log(2k) - c_4 \cdot \frac{(1-\rho)^2}{C^2\eta^4} T\epsilon_1^2\right) \tag{6.23}$$

$$\mathbb{P}\left(\left|\!\left|\!\left| Q_{\mathcal{A}_j,\mathcal{A}_j} - Q^*_{\mathcal{A}_j,\mathcal{A}_j} \right|\!\right|\!\right|_\infty > \epsilon_2\right) \leq \exp\left(2\log(2k) - c_5 \cdot \frac{(1-\rho)^4}{C^4\eta^4}\frac{T}{k^2}\epsilon_2^2\right) \tag{6.24}$$

This implies that, with the following choices

$$\epsilon_1(\zeta_1) = \sqrt{\zeta_1 \cdot \frac{C^2\eta^4}{c_4 T(1-\rho)^2}\log(2k)} \tag{6.25}$$

$$\epsilon_2(\zeta_2) = \sqrt{\zeta_2 \cdot \frac{C^4\eta^4 k^2}{c_5 T(1-\rho)^4}\log(2k)} \tag{6.26}$$

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

$$\epsilon_1(\zeta_1) \leq \frac{3C\eta^2}{1-\rho}, \quad \epsilon_2(\zeta_2) \leq \frac{4C^2\eta^2}{(1-\rho)^2}k, \tag{6.27}$$

we have

$$\begin{aligned}\mathbb{P}\left(\|E_{:,j}\|_\infty \leq \left(2D^2_{\max}\epsilon_2(\zeta_2) + D_{\max}\right)\left(\epsilon_1(\zeta_1) + \lambda\right)\right) \geq 1 - \exp\left(-(\zeta_2 - 2)\log(2k)\right) \\ - \exp\left(-(\zeta_1 - 1)\log(2k)\right) - 2\delta\end{aligned} \tag{6.28}$$

Note that the last term on the right hand side is due to a simple union bound on the events that (6.21) holds and (6.11) recovers the correct sparsity pattern of $\Psi^*_{:,j}$. Now, upon defining

$$\zeta_1 = \frac{\log(2/\delta)}{\log(2k)} + 1 \tag{6.29}$$

$$\zeta_2 = \frac{\log(2/\delta)}{\log(2k)} + 2 \tag{6.30}$$

the inequalities in (6.27) are satisfied, provided that $T \geq \max\{\frac{1}{c_4}, \frac{2}{c_5}\} \cdot \log(4k/\delta)$. Furthermore, combining (6.29) and (6.30) with (6.28) results in

$$\mathbb{P}\left(\|E_{:,j}\|_\infty \leq \left(2D_{\max}^2\epsilon_2(\zeta_2) + D_{\max}\right)\left(\epsilon_1(\zeta_1) + \lambda\right)\right) \geq 1 - 3\delta \tag{6.31}$$

After plugging (6.29) and (6.30) into (6.26) and (6.25), the above inequality is reduced to

$$\begin{aligned}\|E_{:,j}\|_\infty \leq &\left(2D_{\max}^2\sqrt{\frac{2}{c_5}\cdot\frac{C^4\eta^4}{T(1-\rho)^4}k^2\log(4k/\delta)} + D_{\max}\right)\\ &\times\left(\sqrt{\frac{1}{c_4}\cdot\frac{C^2\eta^4}{T(1-\rho)^2}\log(4k/\delta)} + c_6\sqrt{\frac{C^2\eta^4}{\gamma^2T(1-\rho)^2}\log(4(n+m)/\delta)}\right)\end{aligned} \tag{6.32}$$

with probability of at least $1 - 3\delta$. Due to (6.20), one can write

$$D_{\max}^2\sqrt{\frac{2}{c_5}\cdot\frac{C^4\eta^4}{T(1-\rho)^4}k^2\log(4k/\delta)} \leq D_{\max} \tag{6.33}$$

Therefore,

$$\begin{aligned}\|E_{:,j}\|_\infty \leq& 3D_{\max}\left(\frac{1}{\sqrt{c_4}} + c_6\right)\sqrt{\frac{C^2\eta^4}{\gamma^2T(1-\rho)^2}\log(4(n+m)/\delta)}\\ =&\left(\frac{3}{\sqrt{c_4}} + 3c_6\right)\frac{D_{\max}C\eta^2}{\gamma(1-\rho)}\sqrt{\frac{\log(4(n+m)/\delta)}{T}}\end{aligned} \tag{6.34}$$

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

$$\|E\|_\infty \leq \left(\frac{3}{\sqrt{c_4}} + 3c_6\right)\frac{D_{\max}C\eta^2}{\gamma(1-\rho)}\sqrt{\frac{\log(4(n+m)/\delta)}{T}} \tag{6.35}$$

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

6.B Proof of Auxiliary Lemmas

Proof of Lemma 34

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

$$R_1(X(\tau)) = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ X(T_0) & X(T_0-1) & \dots & X(1) & X(0) & 0 & \dots & 0 & 0 \\ X(T_0+1) & X(T_0) & \dots & X(2) & X(1) & X(0) & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ X(T_0+T-1) & X(T_0+T-2) & \dots & X(T) & X(T-1) & X(T-2) & \dots & X(0) & 0 \end{bmatrix} \tag{6.36}$$

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

Lemma 37. *Let λ_k be the k^{th} eigenvalue of the matrix R_G defined as*

$$R_G = \begin{bmatrix} \tilde{R}_1\left(\left[(A+BK)^\tau\right]_{i\to j}\right)\eta^2 & \frac{1}{2}R_1\left(\left[(A+BK)^\tau B\right]_{i\to j}\right)\eta^2 \\ \frac{1}{2}R_1\left(\left[(A+BK)^\tau B\right]_{i\to j}\right)^T\eta^2 & 0 \end{bmatrix} \tag{6.37}$$

Then, the following relations hold

$$\max_k |\lambda_k| \le \frac{3}{2}\frac{C\eta^2}{1-\rho} \tag{6.38}$$

$$\sum_k^{(n+m)(T+T_0+1)} \lambda_k^2 \le \frac{9}{2}\frac{C^2\eta^4 T}{(1-\rho)^2} \tag{6.39}$$

Proof. Notice that

$$\|R_G\| \le \eta^2 \left\|\tilde{R}_1\left(\left[(A+BK)^\tau\right]_{i\to j}\right)\right\| + \frac{1}{2}\eta^2 \left\|R_1\left(\left[(A+BK)^\tau B\right]_{i\to j}\right)\right\| \tag{6.40}$$

Similar to the proof of Lemma A.3 in [207], one can verify that

$$\left\|\tilde{R}_1\left(\left[(A+BK)^\tau\right]_{i\to j}\right)\right\| \le \frac{C}{1-\rho} \tag{6.41}$$

$$\left\|R_1\left(\left[(A+BK)^\tau B\right]_{i\to j}\right)\right\| \le \frac{C}{1-\rho} \tag{6.42}$$

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

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

$$P_{ji} = \begin{bmatrix} 0_{(T_0+1)\times(T_0+1)} & 0_{(T_0+1)\times T} \\ 0_{T\times(T_0+1)} & I_{T\times T} \end{bmatrix} \otimes E_{ji} \tag{6.43}$$

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

Lemma 38. *Let λ_k be the k^{th} eigenvalue of the matrix $\tilde{R}_G$ defined as*

$$\tilde{R}_G = \begin{bmatrix} \tilde{R}_1\left([K(A+BK)^\tau]_{i\to j}\right)\eta^2 & \frac{1}{2}R_1\left([K(A+BK)^\tau B]_{i\to j}\right)\eta^2 + \frac{1}{2}P_{ji}\eta^2 \\ \frac{1}{2}R_1\left([K(A+BK)^\tau B]_{i\to j}\right)^T\eta^2 + \frac{1}{2}P_{ji}^T\eta^2 & 0 \end{bmatrix} \tag{6.44}$$

Then, the following relations hold

$$\max_k |\lambda_k| \leq \frac{2C\eta^2}{1-\rho} \tag{6.45}$$

$$\sum_k^{(n+m)(T+T_0+1)} \lambda_k^2 \leq \frac{16C^2\eta^4 T}{(1-\rho)^2} \tag{6.46}$$

Proof. The proof of the first statement follows directly from Lemma 37. Furthermore, it is easy to verify that the rank of $\tilde{R}_G$ is upper bounded by $4T$. This, together with the upper bound on the maximum eigenvalue completes the proof of the third statement. □

Proof of Lemma 34: One can easily verify that

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

Furthermore, note that the diagonal entries of both R_G and $\tilde{R}_G$ are zero and hence, $\mathbb{E}\left\{\frac{1}{T}z^T R_G z\right\} = \mathbb{E}\left\{\frac{1}{T}z^T \tilde{R}_G z\right\} = 0$. This, together with Hanson-Wright inequality and Lemmas 37 and 38 completes the proof. □

Proof of Lemma 35

Define the matrix

$$R_2(X(\tau)) = \begin{bmatrix} X(T_0) & X(T_0-1) & \ldots & X(1) & X(0) & 0 & \ldots & 0 & 0 \\ X(T_0+1) & X(T_0) & \ldots & X(2) & X(1) & X(0) & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ X(T_0+T-1) & X(T_0+T-2) & \ldots & X(T) & X(T-1) & X(T-2) & \ldots & X(0) & 0 \end{bmatrix} \tag{6.47}$$

and

$$
\begin{aligned}
H_{1i} &= R_2\left([(A+BK_0)^\tau]_{i,:}\right)\eta \in \mathbb{R}^{T\times n(T+T_0+1)}\\
H_{1j} &= R_2\left([(A+BK_0)^\tau]_{j,:}\right)\eta \in \mathbb{R}^{T\times n(T+T_0+1)}\\
H_{2i} &= R_2\left([(A+BK_0)^\tau B]_{i,:}\right)\eta \in \mathbb{R}^{T\times m(T+T_0+1)}\\
H_{2j} &= R_2\left([(A+BK_0)^\tau B]_{j,:}\right)\eta \in \mathbb{R}^{T\times m(T+T_0+1)}\\
H_{3i} &= R_2\left([K_0(A+BK_0)^\tau]_{i,:}\right)\eta \in \mathbb{R}^{T\times n(T+T_0+1)}\\
H_{3j} &= R_2\left([K_0(A+BK_0)^\tau]_{j,:}\right)\eta \in \mathbb{R}^{T\times n(T+T_0+1)}\\
H_{4i} &= R_2\left([K_0(A+BK_0)^\tau B]_{i,:}\right)\eta^2 + P_i\eta \in \mathbb{R}^{T\times m(T+T_0+1)}\\
H_{4j} &= R_2\left([K_0(A+BK_0)^\tau B]_{j,:}\right)\eta^2 + P_j\eta \in \mathbb{R}^{T\times m(T+T_0+1)}
\end{aligned}
\tag{6.48}
$$

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

$$
P_j = \begin{bmatrix} 0_{T\times(T_0+1)} & I_{T\times T}\end{bmatrix} \otimes e_j \tag{6.49}
$$

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

Lemma 39. *Let $\{k_1,k_2,k_3,k_4\} \in \{1,2,3,4\}^4$, where $k_1 \neq k_4$ and $k_2 \neq k_3$. Furthermore, let λ_k be the k^{th} eigenvalue of the following matrix*

$$
\begin{aligned}
R_M(k_1,k_2,k_3,k_4) = &\begin{bmatrix} \frac{1}{2}(H_{k_1i}^\top H_{k_3j} + H_{k_3j}^\top H_{k_1i}) & \frac{1}{2}(H_{k_1i}^\top H_{k_4j} + H_{k_3j}^\top H_{k_2i}) \\ \frac{1}{2}(H_{k_4j}^\top H_{k_1i} + H_{k_2i}^\top H_{k_3j}) & \frac{1}{2}(H_{k_2i}^\top H_{k_4j} + H_{k_4j}^\top H_{k_2i})\end{bmatrix}\\
&\in \mathbb{R}^{(n+m)(T+T_0+1)\times(n+m)(T+T_0+1)}
\end{aligned}
\tag{6.50}
$$

Then, the following relations hold

$$
\max_k |\lambda_k| \leq \frac{6C^2\eta^2}{(1-\rho)^2} \tag{6.51}
$$

$$
\sum_{k=1}^{(n+m)(T+T_0+1)} \lambda_k^2 \leq \frac{72C^4\eta^4}{(1-\rho)^4} \tag{6.52}
$$

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

$$
\begin{aligned}
&|||R_M(k_1,k_2,k_3,k_4)|||\\
&\leq \frac{1}{2}\max\{|||H_{k_1i}^\top H_{k_3j} + H_{k_3j}^\top H_{k_1i}|||, |||H_{k_2i}^\top H_{k_4j} + H_{k_4j}^\top H_{k_2i}|||\} + \frac{1}{2}|||H_{k_1i}^\top H_{k_4j} + H_{k_3j}^\top H_{k_2i}|||\\
&\leq \frac{1}{2}\max\{|||H_{k_1i}^\top|||\,|||H_{k_3j}||| + |||H_{k_3j}^\top|||\,|||H_{k_1i}|||, |||H_{k_2i}^\top|||\,|||H_{k_4j}||| + |||H_{k_4j}^\top|||\,|||H_{k_2i}|||\}\\
&\quad + \frac{1}{2}\left(|||H_{k_1i}^\top|||\,|||H_{k_4j}||| + |||H_{k_3j}^\top|||\,|||H_{k_2i}|||\right)
\end{aligned}
\tag{6.53}
$$

Furthermore, similar to the proof of Lemma A.4 in [207], one can verify that

$$\begin{aligned} |||H_{ri}|||, |||H_{rj}||| &\leq \frac{C}{1-\rho} && \text{if } \; r = 1, 2, 3 \\ |||H_{ri}|||, |||H_{rj}||| &\leq \frac{2C}{1-\rho} && \text{if } \; r = 4 \end{aligned}$$

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

$$R_M^{(1)} = \frac{1}{2} \begin{bmatrix} H_{k_1 i}^\top \\ H_{k_2 i}^\top \end{bmatrix} \begin{bmatrix} H_{k_3 j} & H_{k_4 j} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} H_{k_3 j}^\top \\ H_{k_4 j}^\top \end{bmatrix} \begin{bmatrix} H_{k_1 i} & H_{k_2 i} \end{bmatrix} \tag{6.54}$$

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

Lemma 40. *We have* $\mathbb{E}(M) = M^*$.

Proof. Define

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

The theorem can be proven by showing

$$\begin{aligned} &\frac{1}{T}\mathbb{E}(X_1 X_1^T) = Q^*, \\ &\frac{1}{T}\mathbb{E}(X_2 X_1^T) = KQ^*, \\ &\frac{1}{T}\mathbb{E}(X_2 X_2^T) = KQ^*K^T + \sigma_v^2 I, \end{aligned} \tag{6.55}$$

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

$$\frac{1}{T}\mathbb{E}(X_1 X_1^T) = \frac{1}{T}\sum_{\tau=0}^{T-1} \mathbb{E}(x(\tau)x(\tau)^T) \tag{6.56}$$

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

$$\mathbb{E}(x(1)x(1)^T) = (A + BK)Q^*(A + BK)^T + \sigma_w^2 I + \sigma_v^2 BB^T = Q^* \tag{6.57}$$

where the second inequality is due to (6.5). Similarly, one can show that $\mathbb{E}(x(\tau)x(\tau)^T) = Q^*$ for every $\tau \in \{2, 3, \ldots, T-1\}$ and hence,

$$\frac{1}{T}\mathbb{E}(X_1 X_1^T) = \frac{1}{T}\sum_{\tau=0}^{T-1} Q^* = Q^* \tag{6.58}$$

This completes the proof. □

Proof of Lemma 35: Due to Lemma 40 and upon taking $T_0 \to \infty$, we have

$$\mathbb{P}\{|M_{ij} - M_{ij}^*| > \epsilon\} = \mathbb{P}\{|M_{ij} - \mathbb{E}(M_{ij})| > \epsilon\} \tag{6.59}$$

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

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

Invoking the Hanson-Wright inequality and Lemma 39 for the aforementioned cases completes the proof. □

The proof of Proposition 6

We need the following lemma:

Lemma 41. *We have*

$$\|M^*\| \leq \frac{85 C^2 \eta^2}{1 - \rho} \tag{6.60}$$

Proof. One can easily verify that

$$Q^* = \sum_{\tau=0}^{\infty} \begin{bmatrix} \sigma_w (A + BK_0)^\tau & \sigma_v (A + BK_0)^\tau B \end{bmatrix} \begin{bmatrix} \sigma_w (A + BK_0)^\tau & \sigma_v (A + BK_0)^\tau B \end{bmatrix}^T \tag{6.61}$$

and hence

$$M^* = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_v^2 I \end{bmatrix} + \sum_{\tau=0}^{\infty} \begin{bmatrix} \sigma_w(A+BK_0)^\tau & \sigma_v(A+BK_0)^\tau B \\ \sigma_w K_0(A+BK_0)^\tau & \sigma_v K_0(A+BK_0)^\tau B \end{bmatrix} \begin{bmatrix} \sigma_w(A+BK_0)^\tau & \sigma_v(A+BK_0)^\tau B \\ \sigma_w K_0(A+BK_0)^\tau & \sigma_v K_0(A+BK_0)^\tau B \end{bmatrix}^T \tag{6.62}$$

Therefore, with the assumption $\sigma_w, \sigma_w \leq 1$ and the fact that $\sigma_u, \sigma_v \leq \sqrt{5}\eta$ (the proof of which is simple and can be found, e.g., in [214]), one can write

$$\begin{aligned} |||M^*||| &\leq 5\eta^2 + 5\eta^2 \sum_{\tau=0}^{\infty} \left|\left|\left| \begin{bmatrix} (A+BK_0)^\tau & (A+BK_0)^\tau B \\ K_0(A+BK_0)^\tau & K_0(A+BK_0)^\tau B \end{bmatrix} \right|\right|\right|^2 \\ &\leq 5\eta^2 + 5\eta^2 \sum_{\tau=0}^{\infty} (|||(A+BK_0)^\tau||| + |||K_0(A+BK_0)^\tau B||| + |||K_0(A+BK_0)^\tau||| \\ &\qquad + |||(A+BK_0)^\tau B|||)^2 \\ &\leq 5\eta^2 + 80\eta^2 \sum_{\tau=0}^{\infty} C^2\rho^{2\tau} \\ &\leq \frac{85C^2\eta^2}{1-\rho} \end{aligned} \tag{6.63}$$

This completes the proof. □

Based on this lemma, we will take a similar approach to the proof of Theorem 3.1 in [207] to prove the correct sparsity recovery of the system matrices.

Proof of Proposition 6: To prove this proposition, we need to show that the conditions of Lemma **??** holds with high probability. To ensure that the first condition on G implies the second one, it suffices to have

$$\frac{\lambda\gamma}{3} \leq \frac{\Psi_{\min} C_{\min}}{4k} - \lambda \tag{6.64}$$

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

$$\lambda \leq \frac{\Psi_{\min} C_{\min}}{8k} \tag{6.65}$$

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

$$\left|\left|\left| M_{:\mathcal{A}_j} - M^*_{:\mathcal{A}_j} \right|\right|\right|_\infty \leq \frac{\gamma C_{\min}}{12\sqrt{k}} \tag{6.66}$$

Based on the above analysis, it suffices to have

$$\mathbb{P}\left(\|G\|_\infty > \frac{\gamma\lambda}{3}\right) \leq \frac{\delta}{2} \tag{6.67a}$$

$$\mathbb{P}\left(\left|\left|\left|M_{:\mathcal{A}_j} - M^*_{:\mathcal{A}_j}\right|\right|\right|_\infty > \frac{\gamma C_{\min}}{12\sqrt{k}}\right) \leq \frac{\delta}{2} \tag{6.67b}$$

in order to ensure the exact recovery with probability of at least $1-\delta$. First, we derive conditions under which (6.67a) holds. Based on Lemma 34, one needs to ensure the following inequalities

$$2(n+m)\exp\left(-c_4 \cdot \frac{(1-\rho)^2}{C^2\eta^4}\frac{\gamma^2\lambda^2}{9}T\right) \leq \frac{\delta}{2} \tag{6.68a}$$

$$\lambda \leq \frac{\Psi_{\min} C_{\min}}{8k} \tag{6.68b}$$

$$\frac{\gamma\lambda}{3} \leq \frac{3C\eta^2}{1-\rho} \tag{6.68c}$$

where (6.68c) is a technical condition that is required by Lemma 34. It can be easily verified that (6.68a) is satisfied with the choice of

$$\lambda = \sqrt{\frac{9}{c_4} \cdot \frac{C^2\eta^4}{\gamma^2 T(1-\rho)^2}\log(4(n+m)/\delta)} \tag{6.69}$$

Based on the chosen value for λ and in order to satisfy (6.68b), we should have the following lower bound on T

$$T \geq \frac{576}{c_4} \cdot \frac{C^2\eta^4 k^2}{\Psi^2_{\min} C^2_{\min}\gamma^2(1-\rho)^2}\log(4(n+m)/\delta) \tag{6.70}$$

Similarly, to ensure the validity of (6.68c), we should have

$$T \geq \frac{1}{c_4} \cdot \log(4(n+m)/\delta) \tag{6.71}$$

Now, we will derive the conditions under which (6.67b) is satisfied using Lemma 35. To this goal, first we need to show that the following condition is satisfied:

$$0 < \epsilon < \frac{4C^2\eta^2}{(1-\rho)^2} \tag{6.72a}$$

which is reduced to

$$\frac{\gamma C_{\min}}{12\sqrt{k}} < \frac{4C^2\eta^2}{(1-\rho)^2}k \tag{6.73}$$

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

$$k^{3/2} > \frac{1}{48}\frac{\gamma C_{\min}(1-\rho)^2}{C^2\eta^2} \tag{6.74}$$

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

$$C_{\min} \leq \lambda_{\min}(M^*_{\mathcal{A}_j,\mathcal{A}_j}) \leq \lambda_{\max}(M^*) \leq \frac{85C^2\eta^2}{1-\rho} \tag{6.75}$$

where the last inequality is due to Lemma 41. Therefore,

$$\frac{1}{48}\frac{\gamma C_{\min}(1-\rho)^2}{C^2\eta^2} \leq \frac{85}{48} < 2 \tag{6.76}$$

which implies $k \geq 2$. Finally, to verify (6.67b) and according to Lemma 35, it suffices to have

$$2(n+m)k \exp\left(-c_5 \cdot \frac{(1-\rho)^4}{C^4\eta^4}\frac{\gamma^2 C^2_{\min}}{144k}T\right) \leq \frac{\delta}{2} \tag{6.77}$$

This implies that

$$T \geq \frac{144}{c_5} \cdot \frac{C^4\eta^4 k}{(1-\rho)^4\gamma^2 C^2_{\min}} \log(4(n+m)k/\delta) \tag{6.78}$$

Based on the above analysis, the inequalities (6.70), (6.71), and (6.78) impose lower bounds on T. Comparing these inequalities with (6.20), one can verify that the latter dominates all of them. This completes the proof. □

Proof of Lemma 36

To prove this lemma, first we introduce the KKT conditions for (6.11).

Lemma 42 (KKT conditions)**.** *$\hat{\Psi}_{:,j}$ is an optimal solution for* (6.11) *if and only if it satisfies*

$$M(\hat{\Psi}_{:,j} - \Psi^*_{:,j}) - G + \lambda S = 0 \tag{6.79}$$

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

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

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

Lemma 43. *Assuming that* (6.11) *recovers the correct sparsity pattern of $\Psi^*_{:,j}$, the following equalities hold for $E = \hat{\Psi}_{:,j} - \Psi^*_{:,j}$:*

$$E_{\mathcal{A}_j^c} = 0 \tag{6.80}$$

$$E_{\mathcal{A}_j} = (M_{\mathcal{A}_j,\mathcal{A}_j})^{-1}G_{\mathcal{A}_j} - \lambda(M_{\mathcal{A}_j,\mathcal{A}_j})^{-1}S_{\mathcal{A}_j} \tag{6.81}$$

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

$$M_{\mathcal{A}_j\mathcal{A}_j}E_{\mathcal{A}_j} - G_{\mathcal{A}_j} + \lambda S_{\mathcal{A}_j} = 0 \tag{6.82}$$

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

Proof of Lemma 36: Based on Lemma 43, one can write

$$\|E_{\mathcal{A}_j}\|_\infty \leq \underbrace{\left\|(M_{\mathcal{A}_j,\mathcal{A}_j})^{-1}G_{\mathcal{A}_j}\right\|_\infty}_{Z_1} + \underbrace{\lambda\left\|(M_{\mathcal{A}_j,\mathcal{A}_j})^{-1}S_{\mathcal{A}_j}\right\|_\infty}_{Z_2} \tag{6.83}$$

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

$$\begin{aligned} Z_2 &\leq \lambda\left\|\left((M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\right)S_{\mathcal{A}_j}\right\|_\infty + \lambda\left\|(M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1}S_{\mathcal{A}_j}\right\|_\infty \\ &\leq \lambda\left(\left|\!\left|\!\left|(M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\right|\!\right|\!\right|_\infty + \left|\!\left|\!\left|(M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\right|\!\right|\!\right|_\infty\right) \\ &\leq \lambda\left(\underbrace{\left|\!\left|\!\left|(Q_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\right|\!\right|\!\right|_\infty}_{\Delta} + D_{\max}\right) \end{aligned} \tag{6.84}$$

On the other hand, we have

$$\begin{aligned} (M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} =& (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\left(M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j}\right)(M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \\ =& (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \\ & - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\left(M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j}\right)\left((M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} + \left((M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\right)\right) \end{aligned} \tag{6.85}$$

and therefore

$$\Delta \leq \left|\!\left|\!\left|(M_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\right|\!\right|\!\right|_\infty \left|\!\left|\!\left|M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j}\right|\!\right|\!\right|_\infty \left(\left|\!\left|\!\left|(M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\right|\!\right|\!\right|_\infty + \Delta\right) \tag{6.86}$$

This leads to

$$\begin{aligned} \Delta &\leq \frac{D^2_{\max}}{1 - D_{\max}\left|\!\left|\!\left|M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j}\right|\!\right|\!\right|_\infty}\left|\!\left|\!\left|Q_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j}\right|\!\right|\!\right|_\infty \\ &\leq \frac{D^2_{\max}}{1 - \min\{1/2, \eta^2\}}\left|\!\left|\!\left|M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j}\right|\!\right|\!\right|_\infty \\ &\leq 2D^2_{\max}\left|\!\left|\!\left|M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j}\right|\!\right|\!\right|_\infty \end{aligned} \tag{6.87}$$

where the last inequality is due to the assumption (6.14). Combining the above inequality with (6.84) gives rise to

$$Z_2 \leq \lambda\left(2D^2_{\max}\left|\!\left|\!\left|M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j}\right|\!\right|\!\right|_\infty + D_{\max}\right) \tag{6.88}$$

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

$$\begin{aligned} Z_1 &\leq \left(\left\| (M_{\mathcal{A}_j,\mathcal{A}_j})^{-1} - (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right\|_\infty + \left\| (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right\|_\infty \right) \|G_{\mathcal{A}_j}\|_\infty \\ &\leq \left(\Delta + \left\| (M^*_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right\|_\infty \right) \|G_{\mathcal{A}_j}\|_\infty \\ &\leq \left(2D^2_{\max} \left\| M_{\mathcal{A}_j,\mathcal{A}_j} - M^*_{\mathcal{A}_j,\mathcal{A}_j} \right\|_\infty + D_{\max} \right) \|G_{\mathcal{A}_j}\|_\infty \end{aligned} \tag{6.89}$$

Putting together (6.89) and (6.88) completes the proof. □

Chapter 7

Efficient Learning of Distributed Control Policies

In this work, we propose a robust approach to design distributed controllers for unknown-but-sparse linear and time-invariant systems. By leveraging modern techniques in distributed controller synthesis and structured linear inverse problems as applied to system identification, we show that near-optimal distributed controllers can be learned with sub-linear sample complexity and computed with near-linear time complexity, both measured with respect to the dimension of the system. In particular, we provide sharp end-to-end guarantees on the stability and the performance of the designed distributed controller and prove that for sparse systems, the number of samples needed to guarantee robust and near optimal performance of the designed controller can be significantly smaller than the dimension of the system. Finally, we show that the proposed optimization problem can be solved to global optimality with near-linear time complexity by iteratively solving a series of small quadratic programs.

7.1 Introduction

Encouraged by the success of machine learning (ML) techniques applied to complex decision making problems [132] such as image classification [144], video and board games [183, 231, 230], and robotics [67, 200, 162], the use of ML for the control of autonomous systems interacting with physical environments has been an active area of research in recent years. While there is an increasing body of work studying the theoretical and practical aspects of deploying learning-enabled control policies in individual systems (e.g., self-driving cars, agile robots) [162, 67, 34, 241, 212], there has been little work studying the use of these techniques on *distributed systems*, that is to say systems composed of interconnected and often spatially-distributed subsystems. Examples of such distributed systems include intelligent transportation systems and cities, smart grids, and distributed sensor networks. Even when the individual components are well modeled, controlled, and understood, integrating them into a large-scale, interconnected, and heterogeneous system can make modeling and

control of the full system challenging, strongly motivating the use of machine-learning-based techniques.

Extending the application of data-driven techniques to large-scale and safety-critical systems requires overcoming several challenges. First, we must ensure that the new data-driven methods lead to autonomous systems that are safe, reliable, and robust, as many of our target application areas correspond to safety-critical infrastructure. Failure of such systems could be catastrophic in terms of both social, economic, and possible human losses. Second, any proposed learning and control algorithm must scale gracefully to large-scale and potentially spatially distributed systems. To address these challenges, we extend the approach taken in [62] for designing centralized control policies to the *distributed* optimal control of an unknown *distributed* dynamical system. We develop both deterministic and probabilistic guarantees for a novel robust distributed control synthesis approach. Our proposed method is scalable to large systems, and it allows us to provide the first end-to-end sample complexity guarantees for the distributed optimal control of an unknown system.

In particular, we consider the discrete-time stochastic linear time-invariant system

$$x(t+1) = A_\star x(t) + B_\star u(t) + w(t) \tag{7.1}$$

with the state $x(t) \in \mathbb{R}^n$, state matrix $A_\star \in \mathbb{R}^{n\times n}$, controllable input $u(t) \in \mathbb{R}^m$, input matrix $B_\star \in \mathbb{R}^{n\times m}$, and exogenous random noise $w(t) \in \mathbb{R}^n$ (also referred to as disturbance noise). The goal is to design a control policy $u(t) = f(\{x(\tau)\}_{\tau=0}^t, \{u(\tau)\}_{t=0}^\tau)$ that minimizes the following expected cost function:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{x(t)^\top Q x(t) + u(t)^\top R u(t)\right\} \tag{7.2}$$

subject to dynamics (7.1), where Q and R are positive-definite matrices. When the system matrices are known and there is no communication constraint on the control policy, this problem reduces to the well-known centralized linear-quadratic regulator (LQR) design for which the static linear policy $u(t) = Kx(t)$ is known to be optimal. The optimality of this control policy is contingent upon the full knowledge of the system matrices, as well as the absence of communication constraints on the structure of the controller. However, these conditions are not satisfied in general, as the system may be subject to unknown dynamics and spatiotemporal constraints as discussed below:

Unknown dynamics: As mentioned in Chapter 6, in many systems, the exact parameters of the dynamics are not known *a priori*. In particular, rather than having direct access to the system matrices $(A_\star, B_\star)$, we usually only have access to some estimates $(\hat{A}, \hat{B})$ obtained from first principles, domain knowledge, or a system identification technique. Further, in the distributed setting, the *sparsity structure* of these matrices may be unknown as well, due to dynamic interconnections between component sub-systems. As we describe in the sequel, identifying a structured model is the key to scaling robust and optimal control methods to large systems.

Spatiotemporal constraints: Large-scale distributed systems, such as power grids and distributed computing networks, are composed of smaller sub-systems that are locally interconnected according to a physical interaction topology. Exploiting the underlying sparsity of these systems, as induced by the local interactions between subsystems, is crucial in extending robust and optimal control methods to the distributed setting [257, 217] by allowing *local* sub-controllers to communicate and coordinate with each other. Furthermore, from a practical perspective, controllers that can be implemented using *finite impulse response (FIR)* components lead to simple and intuitive implementations [258, 259].

Contributions

In this work, we overcome the aforementioned difficulties by leveraging recent advances in control theory and machine learning. Namely, we develop a novel distributed robust control synthesis method using the System Level Synthesis (SLS) framework [257], and combine it with model error bounds obtained via the non-asymptotic analysis of regularized estimators as applied to sparse system identification [85, 84], leading to a method that is efficient *both in sample and computational complexities.*

Given the estimates $(\hat{A}, \hat{B})$ of the *true* system matrices $(A_\star, B_\star)$, we are interested in designing a *distributed* controller that can guarantee the stability of the true system with a small optimality gap in its cost function. In particular, given the estimates $\hat{A}$, $\hat{B}$ with an estimation error $\epsilon := \max\{\|\hat{A} - A_\star\|_2, \|\hat{B} - B_\star\|_2\}$, we propose a method to design a dynamic and linear state-feedback controller $\mathbf{K}$ that 1) admits a distributed implementation, respecting the spatiotemporal constraints imposed by the underlying communication topology, and 2) is robust against the model uncertainties; in particular, it stabilizes the closed-loop gain $A_\star + B_\star\mathbf{K}$ and admits a relative sub-optimality bound $J(A_\star, B_\star, \mathbf{K}) - J_\star \leq \alpha(\epsilon, L)J_\star$ for some positive sub-optimality factor $\alpha(\epsilon, L)$. Here, $J(A_\star, B_\star, \mathbf{K})$ is the value of the cost function (7.2) achieved by the controller $\mathbf{u} = \mathbf{K}\mathbf{u}$ acting on the true system, and $J_\star$ is the cost of the *oracle* distributed controller to be formally defined later. Furthermore, L is the enforced temporal length of the obtained system responses with the designed controller. We show that the sub-optimality factor $\alpha(\epsilon, L)$ can be decomposed into two terms:

$$\alpha(\epsilon, L) = \alpha_e(\epsilon) + \alpha_t(L) \tag{7.3}$$

where $\alpha_e(\epsilon)$ bounds the performance degradation caused by *model uncertainty*, and $\alpha_t(L)$ bounds the effect of *temporal truncation*, which quantifies the deviation of the designed controller from its oracle counterpart, when the system responses are restricted to the FIR filters with length L. We prove that the uncertainty and truncation errors decay linearly in ϵ and exponentially in L, respectively. Furthermore, by carefully examining the sparsity structure of the estimated system matrices and the controller, we show that under some conditions, these errors *do not* scale with the system dimensions, and instead, they are only dependent on the sparsity structures of the system dynamics and the controller, as well as other spectral characteristics of the system. By combining the derived bounds with

the recent high-dimensional system identification techniques [85, 84], we provide an end-to-end sub-optimality bound on the performance of the designed distributed controller in terms of the number of sample trajectories that are used for estimating the dynamics, as well as the required temporal length of the system responses. Finally, we provide an efficient algorithm with near-linear time complexity to solve the proposed optimization problem. The performance of the presented method is extensively evaluated in different case studies.

More notation: To streamline the presentation, we specialize and abuse some of the notations in this chapter. We use upper- and lower-case letters to denote matrices and vectors, respectively. Furthermore, we use boldface upper- and lower-case letters to denote transfer matrices and vector-valued signals, respectively. The symbols $\mathcal{H}_2$ and $\mathcal{H}_\infty$ are endowed with the standard definitions of the Hardy spaces, i.e., the class of holomorphic transfer functions on the open unit disk with bounded mean square and maximum norms, respectively. Accordingly, let $\mathcal{RH}_2$ and $\mathcal{RH}_\infty$ correspond to the restriction of these spaces to the set of real, rational, and proper functions. For a transfer matrix $\mathbf{M} \in \mathcal{RH}_\infty$, one can write $\mathbf{M} = \sum_{\tau=0}^{\infty} M(\tau) z^{-\tau}$, where $M(\tau)$ is the τ^{th} spectral component of $\mathbf{M}$. Given a matrix M, the symbol $\text{supp}(M)$ refers to a binary matrix that shares the same sparsity pattern as M. Finally, given a matrix M_0, the set $\mathcal{S}(M_0)$ is defined as $\{M \mid \text{supp}(M) = \text{supp}(M_0)\}$.

7.2 Related Work

Distributed Control Many dynamical systems, such as the power grid, intelligent transportation systems, and distributed computing networks, are large-scale, physically distributed, and interconnected. In such settings, control systems are composed of several sub-controllers, each equipped with their own sensors and actuators – these sub-controllers then exchange local sensor measurements and control actions via a communication network. This information exchange between sub-controllers is constrained by the underlying properties of the communication network, ultimately manifesting as information asymmetry among sub-controllers. This information asymmetry is what makes distributed optimal controller synthesis challenging [122, 174, 217, 19, 20, 194]—indeed, early negative results gave reason to suspect that the resulting distributed optimal control problems were intractable [263, 250].

However, in the early 2000s, a body of work [19, 209, 68, 20, 217, 174, 194] culminating with the introduction of quadratic invariance (QI) in the seminal paper [217], showed that for a large class of practically relevant systems, the resulting distributed optimal control problem is convex. The identification of QI as a useful condition for determining the tractability of a distributed optimal control problem led to an explosion of synthesis results in this area [161, 224, 146, 159, 223, 160, 177, 244, 147, 87]. These results showed that the robust and optimal control methods that were proven so powerful for centralized systems could be used in distributed settings. However, they also made clear that the synthesis and implementation of QI distributed optimal controllers did not scale gracefully with the size of the underlying system—indeed, the complexity of computing a QI distributed optimal controller is at least as expensive to compute as its centralized counterpart, and can be more difficult to implement.

This lack of scalability motivated the development of the SLS framework [257], which allowed for the convex synthesis of *localized* distributed optimal controllers [258, 259] that enjoyed *order constant* synthesis and implementation complexity. In this chapter, we build upon the SLS framework to synthesize an efficient learning-based distributed controller.

System Identification Estimating system models from input/output experiments has a well-developed theory dating back to the 1960s, particularly in the case of linear and time-invariant systems. Standard reference textbooks on the topic include [10, 165, 52, 109], all focusing on establishing *asymptotic* consistency of the proposed estimators.

On the other hand, contemporary results in statistical learning as applied to system identification seek to characterize *finite time and finite data* rates, leaning heavily on tools from stochastic optimization and concentration of measure. Such finite-time guarantees provide estimates of both system parameters and their uncertainty, which allows for a natural bridge to robust/optimal control. In [62], it was shown that under full state observation, if the system is driven by Gaussian noise, the ordinary least squares estimate of the system matrices constructed from independent data points achieves order optimal rates that are linear in the system dimension. This result was later generalized to the single trajectory setting for (i) marginally stable systems in [233], (ii) unstable systems in [221], and (iii) partially observed stable systems in [204, 222, 249, 232].

In this chapter, we leverage our results for the identification of sparse state-space parameters (Chapter 6), where rates are shown to be logarithmic in the ambient dimension, and polynomial in the number of nonzero elements to be estimated.

Machine Learning for Continuous Control We focus on classical and contemporary results most related to the approach taken in this chapter. The use of learning and adaptation in controller design goes back to Kalman: in particular, self-tuning adaptive control, as pioneered in [135, 11], proved to be successful, and was followed by a long sequence of contributions to adaptive control theory, deriving conditions for convergence, stability, robustness and performance under various assumptions. Contemporary approaches can be viewed as non-asymptotic refinements of these classical problems. The modern study of adaptive control, as applied to the LQR problem, was initiated in [4], which provided regret bounds for the optimal LQR control of an unknown system. The work [4] uses an Optimism in the Face of Uncertainty (OFU) based approach, where it maintains confidence ellipsoids of system parameters and selects those parameters that lead to the best closed-loop performance. This work was followed up by several refinements and extensions to different settings [219, 5, 202, 3, 63, 176, 211], and can all be viewed as model-based reinforcement learning algorithms. Another approach was taken in [12], where the authors proposed a learning-based model predictive control (MPC) approach to guarantee the robustness and high performance of an unknown system.

Closest to our work are the results in [62], where the LQR optimal control of an unknown system is studied in the centralized setting. In [62], the authors propose a two-step procedure.

First, they identify a coarse model of the matrices $(A_\star, B_\star)$ describing system behavior, as well as high-probability bounds on the corresponding model estimate uncertainty. They then use these model and uncertainty estimates to synthesize a robustly stabilizing controller, and analyze the end-to-end sample complexity of the resulting controller performance. We generalize this approach to distributed settings, by efficiently exploiting the structure of the system both during the identification and control synthesis phase. This in turn allows us to reduce both the sample and computational complexities of learning distributed controllers, as will be described in the sequel.

7.3 Preliminaries on System Level Synthesis

Given the true system matrices, the optimal centralized LQR controller can be computed by solving its corresponding Ricatti equation [28]. However, as described above, in general the resulting problem becomes highly difficult when solving for a structured controller since it amounts to an NP-hard problem [251]. To circumvent this inherent difficulty, [257] introduces the SLS framework, and shows how it can be used to synthesize distributed controllers by optimizing over their induced closed-loop *system responses*.

We motivate this approach via a simple example. Given a static state-feedback control policy K, the closed-loop map from the disturbance noise $\{w(0), w(1), \dots\}$ to the state $x(t)$ and the control input $u(t)$ at time t is given by

$$\begin{array}{rcl} x(t) & = & \sum_{\tau=0}^{t} (A_\star + B_\star K)^\tau w(t-\tau-1)\,, \\ u(t) & = & \sum_{\tau=0}^{t} K(A_\star + B_\star K)^\tau w(t-\tau-1)\,. \end{array} \tag{7.4}$$

where, with a slight abuse of notation, the initial state $x(0)$ is denoted by $w(-1)$. Letting $\Phi_x(t) := (A_\star + B_\star K)^{t-1}$ and $\Phi_u(t) := K(A_\star + B_\star K)^{t-1}$, we can rewrite (7.4) as

$$\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} = \sum_{\tau=0}^{t} \begin{bmatrix} \Phi_x(\tau) \\ \Phi_u(\tau) \end{bmatrix} w(t-\tau-1)\,, \tag{7.5}$$

where $\{\Phi_x(t), \Phi_u(t)\}$ are called the *system responses* induced by the controller K. The closed-loop system response elements can be defined for a *dynamic* controller in a similar vein. In particular, consider the control policy $\mathbf{u} = \mathbf{K}\mathbf{x}$ for some dynamic controller $\mathbf{K}$. Then, the closed-loop transfer matrices from the disturbance noise $\mathbf{w}$ to the state $\mathbf{x}$ and control action $\mathbf{u}$ satisfy

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} (zI - A - B\mathbf{K})^{-1} \\ \mathbf{K}(zI - A - B\mathbf{K})^{-1} \end{bmatrix} \mathbf{w}. \tag{7.6}$$

The following theorem parameterizes the set of stable closed-loop transfer matrices, as described in (7.6), that are achievable by any stabilizing controller $\mathbf{K}$.

Theorem 29 (State-Feedback Parameterization [257])**.** *The followings are true:*

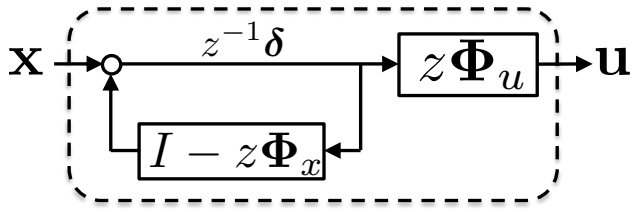


Figure 7.3.1: Internally stabilizing realization of the SLS controller specified in Theorem 29. Notice that sparsity structure imposed on the system responses $\{\boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u\}$ translates directly to the *internal sparsity structure* of the corresponding controller realization.

- *The affine subspace defined by*

$$\begin{bmatrix} zI - A & -B \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_x \\ \boldsymbol{\Phi}_u \end{bmatrix} = I, \ \boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u \in \frac{1}{z}\mathcal{RH}_\infty \tag{7.7}$$

  parameterizes all system responses (7.6) *from* $\mathbf{w}$ *to* $(\mathbf{x}, \mathbf{u})$ *that are achievable by an internally stabilizing state-feedback controller* $\mathbf{K}$*.*

- *For any transfer matrices* $\{\boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u\}$ *satisfying* (7.7)*, the controller* $\mathbf{K} = \boldsymbol{\Phi}_u \boldsymbol{\Phi}_x^{-1}$*, as implemented in Figure 7.3.1, is internally stabilizing and achieves the desired system response* (7.6)*.*

We now make two comments on the consequences of Theorem 29. First, note that $\{\boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u\} = \{(zI - A - B\mathbf{K})^{-1}, \mathbf{K}(zI - A - B\mathbf{K})^{-1}\}$ (as described in (7.6)) are elements of the affine subspace defined by (7.7) whenever $\mathbf{K}$ is a causal stabilizing controller. It is clear from (7.7) that any pair of transfer functions that satisfy (7.7) also obey

$$\Phi_x(t+1) = A_\star \Phi_x(t) + B_\star \Phi_u(t) \ , \ \ \Phi_x(1) = I \ , \ \ \forall t \geq 1 \ , \tag{7.8}$$

and hence, satisfy the state-space equation. Furthermore, the above theorem implies that there exists a dynamic controller $\mathbf{K}$ that achieves these system responses. The SLS framework therefore allows for any optimal control problem over linear systems to be cast as an optimization problem over elements $\{\Phi_x(t), \Phi_u(t)\}$, constrained to satisfy the affine equations (7.8). Comparing equations (7.4) and (7.5), we see that the former is non-convex in the controller $\mathbf{K}$, whereas the latter is convex in the elements $\{\Phi_x(t), \Phi_u(t)\}$, enabling solutions to previously difficult optimal control problems.

Second, notice that the realization of the controller $\mathbf{K} = \boldsymbol{\Phi}_u \boldsymbol{\Phi}_x^{-1}$ in Figure 7.3.1 implies that any sparsity structure imposed on the the system responses translates directly to the internal structure of the corresponding controller. Therefore, we can synthesize controllers that admit distributed realizations by imposing appropriate structural constraints on the system responses. For example, if we wish to limit communications between sub-controllers

that are first neighbors according to the topology defined by A, it suffices to impose additional *linear constraints* that the supports of the system responses $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_u$ be contained in the support of the matrix A. This concept of *locality* in system behavior and corresponding controller implementation is formalized and generalized in [258, 259], and is the key in scaling robust and optimal control methods to large-scale distributed systems.

It follows from Theorem 29 and the standard equivalence between infinite horizon LQR and $\mathcal{H}_2$ optimal control that, for a disturbance process $w_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_w^2 I)$, the standard LQR problem can be equivalently written as

$$\min_{\mathbf{\Phi}_x, \mathbf{\Phi}_u} \sigma_w^2 \left\| \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2}^2 \quad \text{s.t. equation (7.7).} \tag{7.9}$$

We drop the σ_w^2 in the objective function as it affects neither the optimal controller nor the sub-optimality guarantees.

Finally, we will make extensive use of a robust variant of Theorem 29.

Theorem 30 (Robust Stability [178])**.** *Suppose that the transfer matrices* $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\} \in \frac{1}{z}\mathcal{RH}_\infty$ *satisfy*

$$\begin{bmatrix} zI - A & -B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I + \mathbf{\Delta}. \tag{7.10}$$

Then, the controller $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$ *stabilizes the system described by* (A, B) *if and only if* $(I + \mathbf{\Delta})^{-1} \in \mathcal{RH}_\infty$*. Furthermore, the resulting system response is given by*

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} (I + \mathbf{\Delta})^{-1} \mathbf{w}. \tag{7.11}$$

7.4 A Tractable Formulation

Following the SLS framework, the following optimization serves as an alternative formulation of the optimal distributed control problem:

$$\min_{\mathbf{\Phi}_x, \mathbf{\Phi}_x} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \tag{7.12}$$

$$\text{s.t.} \begin{bmatrix} zI - A & -B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I, \tag{7.13}$$

$$\mathbf{\Phi}_x \in \frac{1}{z}\mathcal{RH}_\infty \cap \mathcal{C}_x, \tag{7.14}$$

$$\mathbf{\Phi}_u \in \frac{1}{z}\mathcal{RH}_\infty \cap \mathcal{C}_u, \tag{7.15}$$

where $\mathcal{C}_x := \{\mathcal{C}_x(\tau)\}_{\tau=1}^{\infty}$ and $\mathcal{C}_u := \{\mathcal{C}_u(\tau)\}_{\tau=1}^{\infty}$ capture the structural constraints on $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_u$, respectively. In particular, we have $\Phi_x(\tau) \in \mathcal{C}_x(\tau)$ and $\Phi_u(\tau) \in \mathcal{C}_u(\tau)$ for every $\tau \in \{1, \dots, \infty\}$. The optimization (7.12) is referred to as the *oracle optimization* and its corresponding optimal objective value is called the *oracle cost*. Notice that the formulation of the oracle optimization heavily relies on the availability of the true system matrices. Furthermore, although being convex, the oracle optimization is infinite dimensional as the system responses belong to the set of strictly proper functions. Despite these shortcomings of the oracle optimization, it can be used as a baseline to assess the performance of our proposed method. As a result, we regularly make use of this oracle optimization to measure the sub-optimality of our designed controller. Let $(\mathbf{\Phi}_x^\star, \mathbf{\Phi}_u^\star)$ denote the optimal solution of this optimization problem. According to Theorem 29, the corresponding oracle controller $\mathbf{K}^\star = \mathbf{\Phi}_u^\star \mathbf{\Phi}_x^{\star -1}$ uniformly asymptotically stabilizes the true system. This together with the fact that for LTI systems, uniform asymptotic stability is equivalent to exponential stability, implies that the system responses are exponentially stable [257]. Therefore, upon writing $\mathbf{\Phi}_x^\star = \sum_{t=1}^{\infty} \Phi_x^\star(t) z^{-t}$ and $\mathbf{\Phi}_u^\star = \sum_{t=1}^{\infty} \Phi_u^\star(t) z^{-t}$, there exist constants $C_\star \geq 1$ and $0 < \rho_\star < 1$ such that

$$\max\left\{\|\Phi_x^\star(t)\|_\infty, \|\Phi_u^\star(t)\|_\infty\right\} \leq C_\star \rho_\star^t \tag{7.16}$$

for every integer t.

In what follows, we introduce a surrogate to the oracle optimization that can be solved to robustly design a stabilizing distributed controller based on learned estimates $(\hat{A}, \hat{B})$, taking into account the resulted estimation error. Throughout the chapter, ϵ is used to refer to the spectral norm of the estimation error. In particular, upon defining $\Delta_A = \hat{A} - A_\star$ and $\Delta_B = \hat{B} - B_\star$, we have $\epsilon := \max\{\|\Delta_A\|_2, \|\Delta_B\|_2\}$. We now recall a robust stability result from [62]:

Lemma 44 ([62])**.** *Suppose that the controller $\hat{\mathbf{K}}$ stabilizes the system defined by the matrices $(\hat{A}, \hat{B})$ and that $(\hat{\mathbf{\Phi}}_x, \hat{\mathbf{\Phi}}_u)$ is its corresponding system response on $(\hat{A}, \hat{B})$. Then, controller $\hat{\mathbf{K}}$ stabilizes the system defined by the matrices $(A_\star, B_\star)$ if $\|\hat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} < 1$, where*

$$\hat{\mathbf{\Delta}} = \begin{bmatrix} \Delta_A & \Delta_B \end{bmatrix} \begin{bmatrix} \hat{\mathbf{\Phi}}_x \\ \hat{\mathbf{\Phi}}_u \end{bmatrix}. \tag{7.17}$$

Moreover, under this stability condition, one can write

$$J(A_\star, B_\star, \hat{\mathbf{K}}) = \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{\Phi}}_x \\ \hat{\mathbf{\Phi}}_u \end{bmatrix} \left(I + \hat{\mathbf{\Delta}}\right)^{-1} \right\|_{\mathcal{H}_2} \tag{7.18}$$

Following [62], we design a near-optimal distributed controller by solving the following robust counterpart of the oracle optimization problem (7.12) based on the estimated values

of $(\hat{A}, \hat{B})$ with a given estimation error ϵ:

$$\min_{\mathbf{\Phi}_x, \mathbf{\Phi}_x} \max_{\substack{\|\Delta_A\|_2 \le \epsilon, \\ \|\Delta_B\|_2 \le \epsilon}} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \left(I + \begin{bmatrix} \Delta_A & \Delta_B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right)^{-1} \right\|_{\mathcal{H}_2} \tag{7.19}$$

$$\text{s.t.} \begin{bmatrix} zI - \hat{A} & -\hat{B} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I, \tag{7.20}$$

$$\mathbf{\Phi}_x \in \frac{1}{z}\mathcal{RH}_\infty \cap \mathcal{C}_x, \quad \mathbf{\Phi}_u \in \frac{1}{z}\mathcal{RH}_\infty \cap \mathcal{C}_u, \tag{7.21}$$

The above optimization seeks to find a stabilizing distributed controller that minimizes the worst-case performance achieved on the true system, given the estimates $(\hat{A}, \hat{B})$, and the estimation error ϵ. Clearly, this problem is equivalent to its oracle analog if $\epsilon = 0$. However, notice that the above optimization is infinite-dimensional, since the variable system responses belong to the class of sparse and strictly proper transfer functions. Furthermore, unlike the oracle optimization, it is non-convex with respect to the system responses. To deal with its non-convexity, [62] introduces the following surrogate:

$$\min_{\gamma \in [0,1)} \frac{1}{1-\gamma} \min_{\mathbf{\Phi}_x, \mathbf{\Phi}_x} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \tag{7.22}$$

$$\text{s.t.} \begin{bmatrix} zI - \hat{A} & -\hat{B} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I, \left\| \begin{bmatrix} \frac{\epsilon_A}{\sqrt{\alpha}} \mathbf{\Phi}_x \\ \frac{\epsilon_B}{\sqrt{1-\alpha}} \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} \le \gamma, \tag{7.23}$$

$$\begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \in \frac{1}{z}\mathcal{RH}_\infty \cap \mathcal{C} \tag{7.24}$$

where $\epsilon_A = \|\hat{A} - A_\star\|_2$, $\epsilon_B = \|\hat{B} - B_\star\|_2$, and $\mathcal{C} = \{\begin{bmatrix} \mathbf{M}^\top & \mathbf{N}^\top \end{bmatrix}^\top | \mathbf{M} \in \mathcal{C}_x, \mathbf{N} \in \mathcal{C}_u\}$. Furthermore, γ is a variable that controls the trade-off between the performance of the designed controller and its robustness against the uncertainties in the estimated system matrices. It can be easily verified that the above optimization is jointly quasi-convex in γ and $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$. Therefore, upon restricting $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$ to FIR responses, it can be solved in polynomial time to an arbitrary accuracy. In the absence of sparsity constraints, [62] shows that the above problem gives rise to a robust controller that stabilizes the true system for sufficiently small ϵ_A and ϵ_B. Moreover, [62] characterizes the gap between the cost of the derived and optimal LQR controllers, and shows that the gap scales as $O(\epsilon_A + \epsilon_B)$. However, care must be taken when extending this approach to the distributed setting:

1. Sparsity constraints: The derived bound on the performance of the synthesized controller in [62] is only valid if there are no sparsity constraints on the system responses.

2. Computational complexity: As mentioned before, the above optimization is infinite dimensional and hence, intractable to solve. With the goal of reducing (7.22) to a finite-dimensional problem, [62] proposes to restrict (Φ_x, Φ_u) to FIR responses with length L. With this assumption, [62] shows that for a fixed γ, the inner optimization in (7.22) can be represented

as a semidefinite programming (SDP) with the size $L(n+m)+n$. Moreover, [62] introduces a gridding method to search for the optimal value of γ over the interval $[0, 1)$. Considering the expensive computational complexity of the available SDP solvers, (7.22) quickly becomes prohibitive to solve as the system dimension and/or the length of the FIR responses grow. In particular, using an interior point method [270] to solve the inner SDP for every γ, the proposed algorithm in [62] has the time complexity $\mathcal{O}\left((L(n+m))^{6.5}\frac{1}{\eta}\log\left(\frac{1}{\eta}\right)\right)$ to obtain an η-accurate solution.

3. Sample complexity: Combined with the proposed least-squares estimation method in [62], the minimum number of sample trajectories to accurately estimate the system matrices scales linearly in the system dimension. This linear dependency makes the accurate estimation impractical, if not impossible, as the system size scales up—this is because no *a priori* knowledge of sparsity in the underlying system is exploited.

In this chapter, we will remedy all of the aforementioned issues by introducing a scalable surrogate to the robust optimization problem (7.19) with provable optimality guarantees.

Tractable Surrogates

We now show how the underlying sparse structure of the system matrices $(A_\star, B_\star)$ and distributed controller can be exploited to develop a tractable and scalable convex surrogate to optimization problem (7.19).

Consider the sequence $\mathcal{C}_v := \{\mathcal{C}_v(\tau)\}_{\tau=1}^{\infty}$, where $\mathcal{C}_v(0) = \{X | X \in \mathcal{S}(I_n)\}$ and

$$\mathcal{C}_v(\tau) = \{X_1X_2 + X_3X_4 | X_1 \in \mathcal{S}(\hat{A}), X_2 \in \mathcal{C}_x(\tau), X_3 \in \mathcal{S}(\hat{B}), X_4 \in \mathcal{C}_u(\tau)\} \tag{7.25}$$

for every $\tau = 1, \ldots, \infty$. Assuming that $(\hat{A}, \hat{B})$ and $(A_\star, B_\star)$ share the same sparsity pattern, consider the following optimization problem:

$$\min_{\gamma\in[0,1)} \frac{1}{1-\gamma} \min_{\substack{V(0:L)\\ \Phi_x(1:L)\\ \Phi_u(1:L)}} \sqrt{\sum_{t=1}^{L} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \Phi_x(t) \\ \Phi_u(t) \end{bmatrix} \right\|_F^2} \tag{7.26a}$$

$$\text{s.t. } \Phi_x(1) = I + V(0) \tag{7.26b}$$

$$\Phi_x(t+1) = \hat{A}\Phi_x(t) + \hat{B}\Phi_u(t) + V(t) \qquad t = 1, \ldots, L-1 \tag{7.26c}$$

$$0 = \hat{A}\Phi_x(L) + \hat{B}\Phi_u(L) + V(L) \tag{7.26d}$$

$$\sum_{t=1}^{L} \left\| \begin{bmatrix} \bar{\epsilon}\Phi_x(t) \\ \bar{\epsilon}\Phi_u(t) \end{bmatrix}_{:,j} \right\|_1 \leq \alpha k_\phi^{-1/2}\gamma \qquad j = 1, \ldots, n \tag{7.26e}$$

$$\sum_{t=0}^{L} \|V_{:,j}(t)\|_1 \leq (1-\alpha)k_v^{-1}\gamma \qquad j = 1, \ldots, n \tag{7.26f}$$

$$\Phi_x(t) \in \mathcal{C}_x(t), \quad \Phi_u(t) \in \mathcal{C}_u(t) \qquad t = 1, \ldots, L \tag{7.26g}$$

$$V(t) \in \mathcal{C}_v(t) \qquad t = 0, \ldots, L \tag{7.26h}$$

Here, $\alpha \in (0,1)$ is a parameter to be tuned. Furthermore, $\bar{\epsilon}$ is an upper bound on the spectral norm of the true estimation error ϵ, i.e., $\bar{\epsilon} \geq \epsilon$. Later, we will show how to obtain such upper bound directly from the sample trajectories via bootstrapping. The scalar k_ϕ corresponds to the maximum number of nonzero elements in different rows and columns of $\begin{bmatrix} \boldsymbol{\Phi}_x^\top & \boldsymbol{\Phi}_u^\top \end{bmatrix}^\top$. Similarly, k_v denotes the maximum number of nonzero elements in different rows and columns of $\mathbf{V}$; we will explain later how to obtain k_v based on the imposed sparsity patterns of the system responses. Let a globally optimal solution of the above optimization be denoted by $(\boldsymbol{\Phi}_x^L, \boldsymbol{\Phi}_u^L, \mathbf{V}^L, \gamma^L)$. The inner optimization problem of (7.26) can be written as a parametric QP with respect to γ and is denoted by $\text{OPT}(\gamma)$, whose optimal objective value is referred to as $g(\gamma)$. It is easy to see that $g(\gamma)$ is defined over the domain $[\gamma_0, +\infty)$ for some $\gamma_0 \geq 0$, and is monotonically decreasing.

We will discuss a number of key properties of this problem. First, notice that the optimization is over only the first L components of the system responses, thus yielding a finite-dimensional approximation of the previous infinite-dimensional problem. The slack variables $V(0), V(1), \ldots, V(L)$ are used to capture the error incurred by this truncation. In Theorem 31, we show that the approximation error incurred by restricting our optimization to the first L system response elements decays exponentially with respect to L. Moreover, as will be shown in Lemma 45, the supports of the introduced slack variables are only slightly larger than those of the system responses. Therefore, if the computed system responses are sparse, so are the slack variables. This will in turn help reduce the number of variables in the problem, thereby resulting in a significant computational saving. Finally, a close comparison between (7.26) and (7.22) reveals that the constraint imposed on the $\mathcal{H}_\infty$-norm of the system responses in the latter is replaced by induced norm-1 constraints on the system response elements and the slack variables. Considering the fact that these constraints can

be represented as linear inequalities, we will later show how to efficiently decompose the proposed optimization problem into a series of small and independent QPs.

The next lemma characterizes the sparsity structure of the set $\mathcal{C}_v$. To simplify notation, k will be used to denote the maximum number of nonzero elements of every row and column of $\begin{bmatrix} A_\star & B_\star \end{bmatrix}$ and feasible $\begin{bmatrix} \Phi_x^\top(\tau) & \Phi_u^\top(\tau) \end{bmatrix}^\top, \tau = 1, \dots, L$. Furthermore, we will drop the scripts from a time-dependent sequence $\{M(\tau)\}_{\tau=t_1}^{t_2}$ whenever they are implied by the context.

Lemma 45. *The following statements hold:*

1. *The maximum number of nonzero elements in the rows or columns of every $M \in \mathcal{C}_v$ is upper bounded by $2k^2$.*
2. *The equality $\mathcal{C}_v(\tau) = \mathcal{S}(P_1 P_2 + P_3 P_4)$ is satisfied for every $\tau = 1, \dots, L$, where $P_1 = \operatorname{supp}(\hat{A})$ and $P_3 = \operatorname{supp}(\hat{B})$. Furthermore, P_2 and P_4 are binary matrices with the maximum number of nonzero elements that satisfy $P_2 \in C_x(\tau)$ and $P_4 \in C_u(\tau)$.*

Proof. The proofs of both statements are immediately implied by the sparsity patterns of $\hat{A}$, $\hat{B}$, and the elements of $\mathcal{C}_x(\tau)$ and $\mathcal{C}_u(\tau)$. □

Since P_1, P_2, P_3, and P_4 are sparse matrices, Lemma 45 implies that $\{\mathcal{C}_v(\tau)\}$ can be efficiently characterized by sparse matrix multiplication and summation.

Optimality gap

In this subsection, we analyze the performance of the controller derived from (7.26). The following is the first main theorem of this chapter.

Theorem 31. *Let $J_\star$ be the oracle cost and $(\gamma^L, \mathbf{\Phi}_x^L, \mathbf{\Phi}_u^L)$ be the optimal solution of (7.26). Suppose that $\hat{A}$ and $\hat{B}$ have the same sparsity structure as $A_\star$ and $B_\star$, and that*

$$\bar{\epsilon} < \frac{(1-\rho_\star)\min\{\alpha, 1-\alpha\}}{32 C_\star \rho_\star} k^{-2}, \quad L > \frac{2\log(k) + \log\left(\frac{4\sqrt{2}(\|A_\star\|_\infty + \|B_\star\|_\infty)}{1-\alpha}\right)}{1-\rho_\star}. \tag{7.27}$$

Then, the following statements hold:

1. *$\mathbf{K}^L = \mathbf{\Phi}_u^L \mathbf{\Phi}_x^{L^{-1}}$ stabilizes the true system.*
2. *We have*

$$\frac{J(A, B, \mathbf{K}^L) - J_\star}{J_\star} \leq \underbrace{\frac{16}{\min\{\alpha, 1-\alpha\}} \frac{C_\star \rho_\star}{(1-\rho_\star)} k^2 \bar{\epsilon}}_{\text{uncertainty error}} + \underbrace{\frac{2\sqrt{2}}{1-\alpha}(\|A_\star\|_\infty + \|B_\star\|_\infty) C_\star k^2 \rho_\star^L}_{\text{truncation error}} \tag{7.28}$$

Proof. See Appendix 7.A. □

Theorem 31 quantifies the effects of model uncertainty and spatiotemporal truncation on the optimality gap of the designed distributed controller. In particular, it shows that the uncertainty error is a linear function of $\bar{\epsilon}$, which is an available upper bound on the actual estimation error. On the other hand, even with $\bar{\epsilon} = \epsilon = 0$, one cannot guarantee a zero optimality gap for the designed controller due to the error incurred by the truncation of the system responses. Theorem 31 together with the fact that $0 \leq \rho_\star < 1$ implies that this truncation error decreases exponentially fast with respect to the FIR length L. Further, the smaller $\rho_\star$ is, i.e., the faster the optimal system response decays to zero, the faster the truncation error decays. Finally, if we assume that $\|A_\star\|_\infty$, $\|B_\star\|_\infty$, $C_\star$, and $\rho_\star$ do not scale with the system dimensions, then the derived bounds show that the uncertainty and truncation errors are independent of the system dimension and instead, they only scale with the number of nonzero elements in different rows or columns of the system matrices and responses. Note that $\|A_\star\|_\infty$, $\|B_\star\|_\infty$, $C_\star$, and $\rho_\star$ are defined in terms of the element-wise norm of the system matrices and responses; indeed, the assumption on independence of these quantities from the system dimension are milder and more practical than similar assumptions on their spectral norms, as is usually done in the literature.

7.5 Sample Complexity

Recently, special attention has been devoted to estimating state-space parameters of linear and time-invariant systems based on a limited number of input-output sample trajectories, defined as sequences $\{(x^{(i)}(\tau), u^{(i)}(\tau))\}_{\tau=0}^{T}$ with $i = 1, 2, ..., d$, where d is the number of available sample trajectories and T is the length of each sample trajectory. To simplify notation, the superscript i is dropped from the sample trajectories when $d = 1$. As mentioned in Chapter 6, in general, there are two different approaches to the identification of state-space parameters in the full observation setting: 1) Single sample trajectory, and 2) multiple sample trajectories.

As we seek *sparse* state-space parameters $(\hat{A}, \hat{B})$, we draw upon techniques from Chapter 6 and consider the following Lasso-type estimator:

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \frac{1}{2(t_2-t_1)d} \sum_{i=1}^{d} \sum_{t=t_1}^{t_2} \left\| x^{(i)}(t+1) - \big(Ax^{(i)}(t) + Bu^{(i)}(t)\big) \right\|_2^2 + \lambda(\|A\|_1 + \|B\|_1) \tag{7.29}$$

which is referred to as $\texttt{LASSO}(1:d, t_1:t_2)$ in the sequel. For simplicity of notation, let $\hat{\Psi} = \begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix}^\top$ and $\Psi_\star = \begin{bmatrix} A_\star & B_\star \end{bmatrix}^\top$ denote the estimated and true system matrices, respectively. In [85, 84], variants of the regression problem (7.29) are used to address the problem of sparse system identification with single and multiple sample trajectories.

Remark 17. *As mentioned before in Chapter 6, the system identification based on a single trajectory relies on the availability of an initial distributed controller K_0. Such initial controller may not be necessary if the system is internally stable or it may be obtained based on*

domain knowledge. Alternatively, we have developed a system identification technique in [85] that is based on multiple sample trajectories and hence, bypass the need for such initial controllers. Indeed, our optimization technique can be readily combined with the results of [85] to obtain end-to-end bounds on the sample complexity of the designed distributed controller based on multiple sample trajectories. Due to space restrictions and similarity of the results, we only focus on the system identification with single sample trajectory in this chapter.

Assume that $w(t) \overset{iid}{\sim} \mathcal{N}(0, \sigma_w^2 I)$ for some $\sigma_w > 0$ and the system is equipped with a known stabilizing and static localized controller K_0 with a sparse structure. As mentioned before, K_0 can be set to zero if the system is internally stable. Furthermore, suppose that $u(t) = K_0 x(t) + v(t)$ with $v(t) \overset{iid}{\sim} \mathcal{N}(0, \sigma_v^2 I)$ for some $\sigma_v > 0$.

As shown in Chapter 6, upon the stability of $A + BK_0$, the vector $\begin{bmatrix} x(t)^\top & u(t)^\top \end{bmatrix}^\top$ converges to a stationary distribution $\mathcal{N}(0, M_\star)$, where $M_\star$ is defined as

$$M_\star = \begin{bmatrix} P & PK_0^\top \\ K_0 P & K_0 P K_0^\top + \sigma_v^2 I \end{bmatrix} \tag{7.30}$$

and P satisfies the following Lyapunov equation:

$$(A_\star + B_\star K_0) P (A_\star + B_\star K_0)^\top - P + \sigma_w^2 I + \sigma_v^2 B_\star B_\star^\top = 0 \tag{7.31}$$

We assume that the initial state rests at its stationary distribution. As explained in Chapter 6, this assumption is mild since the state vector converges to its stationary distribution exponentially fast. The following proposition is a restatement of Theorem 27 from Chapter 6:`LASSO`$(1, 1:T-1)$.

Proposition 8 ([84])**.** *Suppose that $k \geq 2$ and the following conditions hold:*

$$\lambda = \mathcal{C}_{\mathrm{s};\lambda} \sqrt{\frac{\log((n+m)/\delta)}{T}}, \qquad T \geq \mathcal{C}_{\mathrm{s};T} k^2 \log((n+m)/\delta), \tag{7.32}$$

Then, under Assumption 3, `LASSO`$(1, 1:T-1)$ recovers the true sparsity pattern of $\Psi_\star$ and it incurs the element-wise estimation error

$$\|\hat{\Psi} - \Psi_\star\|_\infty \leq \mathcal{C}_{\mathrm{s;err}} \sqrt{\frac{\log((n+m)/\delta)}{T}} \tag{7.33}$$

with probability at least $1 - \delta$.

The system complexity constants $\mathcal{C}_{\mathrm{s};\lambda}$, $\mathcal{C}_{\mathrm{s};T}$, and $\mathcal{C}_{\mathrm{s;err}}$ depend on the spectral radius of the closed-loop gain $A + BK_0$, as well as other parameters of the system. The reader is referred to Assumption 3 and its corresponding discussion in Chapter 6.

Equipped with this proposition and Theorem 31, we present the following theorem that characterizes the sample complexity of the derived distributed controller in terms of the learning time and the FIR lengths of the system responses.

Theorem 32. *Suppose that $k \geq 2$, Assumption 3 holds, and* `LASSO`$(1, 1:T-1)$ *is used to obtain the estimates $(\hat{A}, \hat{B})$. Furthermore, suppose that $\bar{\epsilon} = \zeta \mathcal{C}_{\mathrm{s;err}} \sqrt{\frac{k^2 \log((n+m)/\delta)}{T}}$ for an arbitrary $\zeta \geq 1$ and that*

$$\lambda = \mathcal{C}_{\mathrm{s};\lambda} \sqrt{\frac{\log((n+m)/\delta)}{T}}, \tag{7.34}$$

$$T \geq \max\left\{ \left(\frac{32}{\min\{\alpha, 1-\alpha\}} \frac{C_\star \rho_\star \zeta \mathcal{C}_{\mathrm{s;err}}}{1-\rho_\star} \right)^2 k^6, \mathcal{C}_{\mathrm{s};T} k^2 \right\} \log((n+m)/\delta), \tag{7.35}$$

$$L \geq \frac{2\log(k) + \log\left(\frac{4\sqrt{2}(\|A_\star\|_\infty + \|B_\star\|_\infty)}{1-\alpha} \right)}{1-\rho_\star}. \tag{7.36}$$

where $\alpha \in (0,1)$ is an arbitrary and predefined parameter in (**??**)*. Then, the following statements hold with probability at least $1-\delta$:*

1. *$\mathbf{K}^L = \mathbf{\Phi}_u^L \mathbf{\Phi}_x^{L^{-1}}$ stabilizes the true system.*
2. *We have*

$$\begin{aligned} \frac{J(A, B, \mathbf{K}^L) - J_\star}{J_\star} \leq & \frac{16}{\min\{\alpha, 1-\alpha\}} \frac{C_\star \rho_\star \zeta \mathcal{C}_{\mathrm{s;err}}}{1-\rho_\star} k^3 \sqrt{\frac{\log((n+m)/\delta)}{T}} \\ & + \frac{2\sqrt{2}}{1-\alpha} (\|A_\star\|_\infty + \|B_\star\|_\infty) C_\star k \rho_\star^L \end{aligned} \tag{7.37}$$

Proof. Theorem 31 and Proposition 8 can be used to prove this theorem. First, note that (7.34) and (7.35) guarantee the validity of (7.32). Therefore, `LASSO`$(1, 1:T-1)$ can recover the correct sparsity pattern of the system matrices and the estimation error bound (7.33) holds with probability of at least $1-\delta$. This implies that

$$\epsilon = \|\hat{\Psi} - \Psi_\star\|_2 \leq k \|\hat{\Psi} - \Psi_\star\|_\infty \leq \zeta \mathcal{C}_{\mathrm{s;err}} \sqrt{\frac{k^2 \log((n+m)/\delta)}{T}} = \bar{\epsilon} \tag{7.38}$$

with the same probability. Combined with (7.35) and (7.36), this certifies the validity of (7.27). Therefore, (7.28) holds with probability of at least $1-\delta$. Replacing $\bar{\epsilon}$ with $\zeta \mathcal{C}_{\mathrm{s;err}} k \sqrt{\frac{\log((n+m)/\delta)}{T}}$ in (7.28) completes the proof. □

The above theorem characterizes the sample complexity of designing a distributed controller in terms of the lengths of the sample trajectory T, and the FIR filters L. Notice that, similar to Proposition 8, the statements of Theorem 32 hold with probability of $1-\delta$, where δ is the *probability of failure*. In particular, according to (7.35) and (7.37), in order to reduce the probability of failure by a factor of $c > 1$, one needs to increase the length

of the sample trajectory by a factor of $\log(c)$. Furthermore, under the assumption that δ, $C_\star$, ρ_*, $\|A_\star\|_\infty$, $\|A_\star\|_\infty$, and the system complexity constants do not scale with the system dimension, Theorem 32 implies that $T = \Omega(k^6 \log(n+m))$ is enough to guarantee that the optimality gap of the designed controller is on the order of $\mathcal{O}(k^3\sqrt{\log(n+m)/T} + k\rho_\star^L)$. Assuming that the dynamics and controller have sparse structures, i.e., $k \ll n+m$, the proposed bound improves upon the existing sample complexity bounds for learning optimal LQR controllers which scale linearly with the system dimension [62, 63].

Remark 18. *While the proposed method is best suited for designing controllers with sparse system responses, its performance can be compared against a more general oracle optimization* (7.12)*, where the constraint sets $\mathcal{C}_x$ and $\mathcal{C}_u$ are relaxed to* ***weakly sparse*** *structures. Under such circumstances, an optimal LQR controller can be a valid oracle controller, provided that its induced system responses are weakly sparse or, equivalently, they have spatially decaying structures; see [190, 191, 178]. Even though such generalizations are not discussed in this chapter, we note that the derived sub-optimality gap of the designed controller in Theorems 32 and 31 can be extended to this setting, with an additional non-vanishing term capturing the* ***model selection error****.*

7.6 Computational complexity

In this subsection, we propose an efficient algorithm for solving (7.26). It is easy to verify that the proposed optimization problem is jointly quasiconvex. In particular, it is convex with respect to $(\{\Phi_x(t)\}, \{\Phi_u(t)\}, \{V(t)\})$ (after fixing γ) and quasiconvex with respect to γ (after fixing $(\{\Phi_x(t)\}, \{\Phi_u(t)\}, \{V(t)\})$).

Lemma 46. *For every fixed and feasible $\bar{\gamma}$, $\mathrm{OPT}(\bar{\gamma})$ has a unique solution.*

Proof. Notice that $\{V(t)\}$ can be uniquely written in terms of $\{\Phi_x(t)\}$ and $\{\Phi_u(t)\}$. This, together with the fact that the objective is strictly convex, results in the uniqueness of the solution.

□

Lemma 46 and the quasiconvexity of $g(\gamma)$ do not necessarily result in the uniqueness of the solution for (7.26) since $g(\gamma)$ may contain spurious local minima in its *flat regions*. A naive approach to circumvent this issue is to discretize γ within the interval $[0, 1)$ with the points $\{\gamma_1, \ldots, \gamma_N\}$, compute $g(\gamma_i)$ for every $1 \leq i \leq N$, and select the solution with the lowest cost. However, notice that in this approach, the number of discrete points has undesirable dependency on the required accuracy of the solution: roughly speaking, one needs to evaluate and optimize over $\Omega(1/\epsilon)$ discrete points in order to get a solution whose cost is ϵ-away from the optimal cost. In the next proposition, we show that (7.26) is in fact unimodal with respect to γ and hence, it is free of spurious local minima (i.e. non-global

local minima).[1] The unimodal property of (7.26) with respect to γ implies that a simple application of the golden-section search method[2] on γ can find an ϵ-accurate solution by computing $g(\gamma_i)$ at no more than $O(\log(1/\epsilon))$ points.

Proposition 9. *Suppose that* (7.26) *is feasible. Furthermore, suppose that* γ_0 *is the smallest value such that* $0 \leq \gamma_0 < 1$ *and* $\mathrm{OPT}(\gamma_0)$ *is feasible. Then,* $\frac{g(\gamma)}{1-\gamma}$ *is unimodal in the interval* $[\gamma_0, 1)$.

Proof. See Appendix 7.A. □

For a fixed γ, problem $\mathrm{OPT}(\gamma)$ can be decomposed into n parallel sub-problems over the columns of

$$\begin{bmatrix}\Phi_x(1)^\top & \dots & \Phi_x(L)^\top & \Phi_u(1)^\top & \dots & \Phi_u(L)^\top & V(0)^\top & \dots V(L)^\top\end{bmatrix}^\top \tag{7.39}$$

In particular, define $\mathrm{OPT}_j(\gamma)$ as $\mathrm{OPT}(\gamma)$ after replacing the variable matrices $(\{\Phi_x(t)\}, \{\Phi_u(t)\}, \{V(t)\})$ with $(\{[\Phi_x(t)]_{:,j}\}, \{[\Phi_u(t)]_{:,j}\}, \{[V(t)]_{:,j}\})$, as in:

$$\min_{\substack{\{[V(t)]_{:,j}\}\\ \{[\Phi_x(t)]_{:,j}\}\\ \{[\Phi_u(t)]_{:,j}\}}} \sqrt{\sum_{t=1}^{L} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \Phi_x(t) \\ \Phi_u(t) \end{bmatrix}_{:,j} \right\|_F^2} \tag{7.40a}$$

$$\text{s.t. } [\Phi_x(1)]_{:,j} = I_{:,j} + [V(0)]_{:,j} \tag{7.40b}$$

$$[\Phi_x(t+1)]_{:,j} = \hat{A}[\Phi_x(t)]_{:,j} + \hat{B}[\Phi_u(t)]_{:,j} + [V(t)]_{:,j} \qquad t = 1, \dots, L-1 \tag{7.40c}$$

$$0 = \hat{A}[\Phi_x(L)]_{:,j} + \hat{B}[\Phi_u(L)]_{:,j} + [V(L)]_{:,j} \tag{7.40d}$$

$$\sum_{t=1}^{L} \left\| \begin{bmatrix} \bar{\epsilon}\Phi_x(t) \\ \bar{\epsilon}\Phi_u(t) \end{bmatrix}_{:,j} \right\|_1 \leq \alpha k_\phi^{-1/2} \gamma \qquad t = 1, \dots, L \tag{7.40e}$$

$$\sum_{t=0}^{L} \|[V(t)]_{:,j}\|_1 \leq (1-\alpha) k_v^{-1} \gamma \qquad t = 0, \dots, L \tag{7.40f}$$

$$[\Phi_x(t)]_{:,j} \in \mathcal{C}_{x;j}(t), \quad [\Phi_u(t)]_{:,j} \in \mathcal{C}_{u;j}(t) \qquad t = 1, \dots, L \tag{7.40g}$$

$$[V(t)]_{:,j} \in \mathcal{C}_{v;j}(t) \qquad t = 0, \dots, L \tag{7.40h}$$

[1] Note that another approach for eliminating the spurious local minima in the flat regions of a quasiconvex optimization problem is a reformulation based on its sublevel sets; see [37]. However, this method will destroy the decomposibility of (7.26); a feature that is at the core of near-linear solvability of (7.26), as will be shown later in this chapter.

[2] The golden-section search is an algorithm for finding the global minimum of a univariate and strictly unimodal function defined within a bounded interval. The method sequentially identifies and maintains an interval containing the global minimum with a geometrically diminishing length. The geometric shrinkage in the length of this interval implies that $O(\log(\eta^{-1}))$ number of function evaluations is enough to obtain the minimum of the function with η accuracy; see Section 10 in [208] for more details.

where $\mathcal{C}_{x;j}(t) = \{X_{:,j} : X \in \mathcal{C}_x(t)\}$, $\mathcal{C}_{u;j}(t) = \{X_{:,j} : X \in \mathcal{C}_u(t)\}$, and $\mathcal{C}_{v;j}(t) = \{X_{:,j} : X \in \mathcal{C}_v(t)\}$. Furthermore, let $g_j(\gamma)$ denote its optimal objective value. Then, $g(\gamma) = \sqrt{\sum_{j=1}^{n} g_j(\gamma)^2}$ and the optimal solution of $\mathrm{OPT}(\gamma)$ can be obtained by replacing the j^{th} column of (7.39) with the solution of the sub-problem $\mathrm{OPT}_j(\gamma)$ for every $j = 1, \ldots, n$.

The next lemma shows that the sub-problem $\mathrm{OPT}_j(\gamma)$ can be reformulated as a small QP whose size is independent of n.

Lemma 47. *The sub-problem* $\mathrm{OPT}_j(\gamma)$ *can be written as a QP over* $O(Lk^2)$ *variables subject to* $O(Lk^2)$ *constraints.*

Proof. For every $t = 0, \ldots, L$, let $(\Phi_x^{n_j}(t), \Phi_u^{n_j}(t), V^{n_j}(t))$ correspond to $(\Phi_x(t), \Phi_u(t), V(t))$ after removing the elements that are set to zero via the sparsity constraints (7.26g) and (7.26h). It is easy to see that $\mathrm{OPT}_j(\gamma)$ can be written in terms of $(\{\Phi_x^{n_j}(t)\}, \{\Phi_u^{n_j}(t)\}, \{V^{n_j}(t)\})$ with a total number of $O(Lk^2)$ variables. The rest of the proof is devoted to show how to reduce the number of constraints in $\mathrm{OPT}_j(\gamma)$ to $O(Lk^2)$. Let $\mathbf{\Phi}_x^{n_j}$, $\mathbf{\Phi}_u^{n_j}$, and $\mathbf{V}^{n_j}$ denote $\sum_{t=1}^{L} \Phi_x^{n_j}(t) z^{-t}$, $\sum_{t=1}^{L} \Phi_u^{n_j}(t) z^{-t}$, and $\sum_{t=0}^{L} V^{n_j}(t) z^{-t}$, respectively. The constraints (7.40b)-(7.40d) can be written compactly as

$$\begin{bmatrix} zI - \hat{A} & -\hat{B} & -I \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \\ \mathbf{V} \end{bmatrix}_{:,j} = I_{:,j} \iff \mathbf{M}_j \begin{bmatrix} \mathbf{\Phi}_x^{n_j} \\ \mathbf{\Phi}_u^{n_j} \\ \mathbf{V}^{n_j} \end{bmatrix} = I_{:,j} \tag{7.41}$$

Here, $\mathbf{M}_j$ is equal to $\sum_{t=0}^{L} M_j(t) z^{-t}$, where $M_j(t)$ is defined as $\begin{bmatrix} zI - \hat{A} & -\hat{B} & -I \end{bmatrix}$, after removing the columns that correspond to the zero elements of $\begin{bmatrix} \Phi_x(t)^\top & \Phi_u(t)^\top & V(t)^\top \end{bmatrix}^\top_{j,:}$ enforced by the sparsity constraints. The matrix $\mathbf{M}_j$ has at most n rows and $2k^2+k$ columns. On the other hand, every column of $\begin{bmatrix} zI - \hat{A} & -\hat{B} \end{bmatrix}$ has at most $k+1$ number of nonzero elements. Similarly, every column of $-I$ has exactly one nonzero element. Therefore, a simple calculation yields that $\mathbf{M}_j$ can have at most $3k^2 + k$ number of nonzero rows. This together with the definition of $\mathbf{M}_j$ implies that (7.40b)-(7.40d) can be reduced to $O(Lk^2)$ linear constraints. Finally, (7.40e) and (7.40f) can be trivially written as a set of $O(Lk^2)$ linear inequalities by introducing $O(Lk^2)$ slack variables. This completes the proof. □

It is worthwhile to mention that the above lemma is a generalization to the dimension reduction algorithm introduced in [259].

Remark 19. *Note that for every index* j*, the aforementioned reduced QP can be efficiently constructed in an offline fashion before running Algorithm 4 detailed below, provided that the estimated system matrices* $(\hat{A}, \hat{B})$ *and the sparsity constraints* (7.40g) *and* (7.40h) *are given in sparse matrix formats, such as Coordinate list [108]. While we do not discuss the structure of such representations, we note that the complexity of constructing these reduced QPs is dominated by that of Algorithm 4.*

Remark 20. *Without loss of generality, we assume that the proposed optimization* (7.26) *is finitely-representable on a Turing machine. In other words, the total number of digits required to write (or accurately approximate) the input data for* (7.26) *is a finite number* D*. This is a common assumption made for the complexity analysis of optimization problems; see e.g. [253].*

Definition 33. *An algorithm solves an optimization problem that is finitely-representable on a Turing machine to* η*-accuracy if the following statements hold:*

- *It returns a feasible solution if and only if the problem is feasible,*
- *Upon feasibility, it returns a feasible solution whose objective value is greater than the optimal objective value by no more than* η*.*

Algorithm 4 delineates the proposed method for solving (7.26). In particular, it uses a golden-section search method to optimize over the scalar variable γ, while solving multiple small QPs at each iteration to obtain $g(\gamma)$. At any iteration, $g(\gamma)$ is set to $+\infty$ if at least one of $\mathrm{OPT}_1(\gamma), \ldots, \mathrm{OPT}_n(\gamma)$ is infeasible. Suppose $g(\gamma)$ has the domain $[\gamma_0, +\infty)$ for some $\gamma_0 \geq 0$. It is easy to verify that a finite value for γ_0 always exists; however, $\gamma_0 < 1$ is required for (7.26) to be feasible.

Define $\underline{t}$ and $\overline{t}$ as the smallest and largest integers such that

$$\underline{\eta_1} = \left(\frac{2}{1+\sqrt{5}}\right)^{\underline{t}} \leq \eta_1, \quad \overline{\eta_1} = \left(\frac{2}{1+\sqrt{5}}\right)^{\overline{t}} > \eta_1 \tag{7.42}$$

Furthermore, define

$$\Delta_\gamma = \left(\frac{4}{1+\sqrt{5}} - 1\right)\overline{\eta_1} \tag{7.43}$$

Let $g_{\mathrm{ap}}(\gamma_c)$ and $g_{\mathrm{ap}}(\gamma_d)$ denote the objective values of the problems $\mathrm{OPT}(\gamma_c)$ and $\mathrm{OPT}(\gamma_c)$ when they are solved to η_2-accuracy. At each iteration, Algorithm (4) shrinks the interval $[\gamma_a, \gamma_b]$ by comparing the values of $\frac{g_{\mathrm{ap}}(\gamma_c)}{1-\gamma_c}$ and $\frac{g_{\mathrm{ap}}(\gamma_d)}{1-\gamma_d}$, while ensuring that $\gamma^L \in [\gamma_a, \gamma_b]$. However, notice that $g_{\mathrm{ap}}(\gamma_c)$ and $g_{\mathrm{ap}}(\gamma_d)$ are the approximations of $g(\gamma_c)$ and $g(\gamma_d)$, where the possible approximation error is due to the limited accuracy of the interior point method. The incurred error in the computation of $g(\gamma_c)$ and $g(\gamma_d)$ may be aggregated and result in wrong comparisons between their actual values, thereby violating $\gamma^L \in [\gamma_a, \gamma_b]$. To avoid such wrong comparisons, one needs to ensure that the approximation errors $g_{\mathrm{ap}}(\gamma_c) - g(\gamma_c)$ and $g_{\mathrm{ap}}(\gamma_d) - g(\gamma_d)$ are appropriately controlled at every iteration of the algorithm; this will be shown in the next theorem. In particular, we will show how to control the accuracy of the used interior point method for solving the sub-problems $\mathrm{OPT}_j(\gamma_c)$ and $\mathrm{OPT}_j(\gamma_d)$ in order to ensure $\gamma^L \in [\gamma_a, \gamma_b]$ at every iteration of the algorithm. Define the quantity

$$\Delta_g = \min_{\gamma \in [\gamma_0, \gamma^L - \Delta_\gamma] \cup [\gamma^L, 1-\Delta_\gamma)} \left| \frac{g(\gamma + \Delta_\gamma)}{1 - (\gamma + \Delta_\gamma)} - \frac{g(\gamma)}{1-\gamma} \right|. \tag{7.44}$$

Algorithm 4 Sequential Quadratic Programming

input: Estimates $\hat{A}$, $\hat{B}$, estimation error $\bar{\epsilon}$, and accuracy parameters η_1, and η_2
output: $\{\Phi_x(t)\}$, $\{\Phi_u(t)\}$, $\{V(t)\}$, and $g(\gamma)$
obtain $g(1)$ by solving n sub-problems $\mathrm{OPT}_1(1), \ldots, \mathrm{OPT}_n(1)$ to $\frac{\eta_2}{n}$-accuracy using interior point method.
if $g(1) = +\infty$ **then**
  return Infeasible
else
  set $\gamma_a \leftarrow 0$, $\gamma_b \leftarrow 1$, $\gamma_c \leftarrow 1 - \frac{2}{1+\sqrt{5}}$, and $\gamma_d \leftarrow \frac{2}{1+\sqrt{5}}$
  while $|\gamma_b - \gamma_a| > \eta_1$ **do**
    Solve $\mathrm{OPT}(\gamma_c)$ by solving n sub-problems $\mathrm{OPT}_1(\gamma_c), \ldots, \mathrm{OPT}_n(\gamma_c)$ to $\frac{\eta_2}{n}$-accuracy using interior point method. Let the corresponding objective value be denoted as $g_{\mathrm{ap}}(\gamma_c)$.
    Solve $\mathrm{OPT}(\gamma_d)$ by solving n sub-problems $\mathrm{OPT}_1(\gamma_d), \ldots, \mathrm{OPT}_n(\gamma_d)$ to $\frac{\eta_2}{n}$-accuracy using interior point method. Let the corresponding objective value be denoted as $g_{\mathrm{ap}}(\gamma_d)$.
    if $\frac{g_{\mathrm{ap}}(\gamma_c)}{1-\gamma_c} < \frac{g_{\mathrm{ap}}(\gamma_d)}{1-\gamma_d}$ **then**
      set $\gamma_b \leftarrow \gamma_d$
    else
      set $\gamma_a \leftarrow \gamma_c$
    end if
    $\gamma_c \leftarrow \gamma_b - \frac{2}{1+\sqrt{5}}(\gamma_b - \gamma_a)$ and $\gamma_d \leftarrow \gamma_a + \frac{2}{1+\sqrt{5}}(\gamma_b - \gamma_a)$
  end while
  $\bar{\gamma} \leftarrow (\gamma_a + \gamma_b)/2$
  obtain $(\{\bar{\Phi}_x(t)\}, \{\bar{\Phi}_u(t)\}, \{\bar{V}(t)\}, g(\bar{\gamma}))$ by solving n sub-problems $\mathrm{OPT}_1(\bar{\gamma}), \ldots, \mathrm{OPT}_n(\bar{\gamma})$ to $\frac{\eta_2}{n}$-accuracy using interior point method. Let the corresponding objective value be denoted as $g_{\mathrm{ap}}(\bar{\gamma})$.
  if $g_{\mathrm{ap}}(\bar{\gamma}) = +\infty$ **then**
    return Infeasible
  else
    return $(\{\bar{\Phi}_x(t)\}, \{\bar{\Phi}_u(t)\}, \{\bar{V}(t)\}, \bar{\gamma})$
  end if
end if

According to the Proposition 9, the function $\frac{g(\gamma)}{1-\gamma}$ is strictly monotone in the intervals $[\gamma_0, \gamma^L]$ and $[\gamma^L, 1)$ which implies that $\Delta_g > 0$.

Theorem 33. *Suppose that the input data for* (7.26) *can be represented with D digits, and that η_2 satisfies $D \leq C\log(1/\eta_2)$ for a universal constant C. Then, Algorithm 4 terminates in $O(L^{3.5}k^7 n\log(n)\log(1/\eta_1)\log(1/\eta_2))$ time. In particular:*

1. *If $\gamma_0 \leq 1 - \underline{\eta_1}/2$ and $\eta_2 \leq \min\left\{\frac{2}{1+\sqrt{5}}\Delta_g\underline{\eta_1}, \underline{\eta_1}^2\right\}$, then the algorithm returns a feasible solution with $|\bar{\gamma} - \gamma^L| \leq \underline{\eta_1}/2$. Furthermore,*

$$\frac{g_{\text{approx}}(\bar{\gamma})}{1-\bar{\gamma}} - \frac{g(\gamma^L)}{1-\gamma^L} \leq \left(\frac{g(\gamma_0)}{2(1-\gamma^L)^2\gamma^L} + 2\right)\underline{\eta_1} \tag{7.45}$$

   provided that $\underline{\eta_1} \leq 2(1-\gamma^L)^2$.

2. *If $\gamma_0 > 1 - \underline{\eta_1}/2$, then the algorithm declares infeasibility.*

Proof. See Appendix 7.A. □

Bootstrapping:

Recall that formulating the optimization problem (7.26) relies on the availability of the upper bound $\bar{\epsilon}$ on the actual estimation error $\epsilon = \max\{\|\hat{A} - A_\star\|_2, \|\hat{B} - B_\star\|_2\}$. It is evident from (7.26) that the performance (and even feasibility) of the proposed control design method heavily relies on the conservativeness of $\bar{\epsilon}$: a large value for $\bar{\epsilon}$ results in more restrictive constraints on the system responses. Although in some applications, an upper bound for ϵ may be readily available based on the domain knowledge, its value may be too conservative for practical purposes. A simple method to alleviate this issue is to resort to a bootstrap approach, where the goal is to *estimate the estimation error*, merely based on the available data samples. In particular, given the estimates $\hat{A}$ and $\hat{B}$, we draw sample trajectories from the empirical distribution induced by $(\hat{A}, \hat{B})$ in N rounds. Using these synthetically generated sample trajectories at each round i, we re-estimate the system dynamics $\hat{A}^{(i)}$ and $\hat{B}^{(i)}$. Finally, an upper bound on the estimation error is obtained by setting $\bar{\epsilon}$ as $100 \times (1-\delta)$ percentile of $\max\{\|\hat{A}^{(i)} - \hat{A}\|_2, \|\hat{B}^{(i)} - \hat{B}\|_2\}, i = 1, \ldots, N$, for some parameter $\delta > 0$. Roughly speaking, the obtained estimation error is an upper bound on the actual one with probability of at least $1 - \delta$. Similar bootstrap methods are widely used for estimating various characteristics of estimators, such as their bias, variance, etc. A more detailed analysis on bootstrap methods can be found in [71, 113, 225].

Algorithm 5 describes the proposed method for obtaining $\bar{\epsilon}$. In this algorithm, the matrix M is defined as (7.30), where P refers to the solution of the Lyapunov equation (7.31) after replacing the true system matrices with the estimated ones.

Algorithm 5 Bootstrapping

input: Initial state x_0, estimates $\hat{A}, \hat{B}$, initial controller K_0, distribution parameters η_w, η_v, M, confidence parameter δ, and number of rounds N
output: upper bound on the estimation error $\bar{\epsilon}$
for i in $\{1, \dots, N\}$ **do**
  $x(0) \sim \mathcal{N}(0, M)$
  for τ in $\{0, \dots, T-1\}$ **do**
    $u(\tau) \leftarrow K_0 x(\tau) + v(\tau)$, where $v(\tau) \sim \mathcal{N}(0, \eta_v^2 I)$
    $x(\tau+1) \leftarrow \hat{A}x(\tau) + \hat{B}u(\tau) + w(\tau)$ where $w(\tau) \sim \mathcal{N}(0, \eta_w^2 I)$
  end for
  Obtain $(\hat{A}^{(i)}, \hat{B}^{(i)})$ by solving $\texttt{LASSO}(1, 1:T-1)$ with $\left(\{x(\tau)\}_{\tau=0}^{T}, \{u(\tau)\}_{\tau=0}^{T-1}\right)$ as input
  $\bar{\epsilon}^{(i)} \leftarrow \max\{\|\hat{A}^{(i)} - \hat{A}\|, \|\hat{B}^{(i)} - \hat{B}\|\}$
end for
return $\bar{\epsilon}$ as the $100 \times (1-\delta)$ percentile of $\left\{\bar{\epsilon}^{(i)}\right\}_{i=1}^{N}$.

7.7 Numerical Results

To illustrate the effectiveness of the developed control design framework, we focus on a class of graph Laplacian systems with *chain* structures. Let the scalars $x_i(t)$, $u_i(t)$, and $w_i(t)$ denote the state, input, and the disturbance corresponding to the subsystem i. Consider the following dynamics:

$$\begin{aligned}
x_i(t+1) &= (D_i+1-2a_i)x_i(t)+a_i(x_{i-1}(t)+x_{i+1}(t))+b_i u_i(t)+w_i(t) && \text{if } 2 \le i \le n-1 \\
x_i(t+1) &= (D_i+1-a_i)x_i(t)+a_i x_{i-1}(t)+b_i u_i(t)+w_i(t) && \text{if } i = n \\
x_i(t+1) &= (D_i+1-a_i)x_i(t)+a_i x_{i+1}(t)+b_i u_i(t)+w_i(t) && \text{if } i = 1
\end{aligned} \tag{7.46}$$

where D_i and a_i are scalar numbers, and b_i is a binary number taking the value 1 only if subsystem i is directly controlled by an input signal; see Figure 7.7.1 for a simple realization of this model. We assume that $w(t) \sim \mathcal{N}(0, I)$ in all of our experiments. Inspired by the exponential decay of the truncation error with respect to the FIR length L in Theorem 31, we set the parameter α in (7.26) to 1.2^{-L} throughout our simulations. Similar to [258], we assume that the control structure is *local* and subject to *communication delays*, both of which can be translated to sparsity constraints on the system responses. In particular, given the locality parameter d, we are interested in designing a control structure with the property that the effect of a disturbance signal $w_i(t)$ hitting subsystem i is localized to a region defined by its d-hop neighbors. Recalling the definition of the system responses (7.6), one can easily verify that the local containment of the effect of disturbance noise within d-hop neighbors is equivalent to enforcing banded sparsity structures on $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_u$ with the bandwidth of

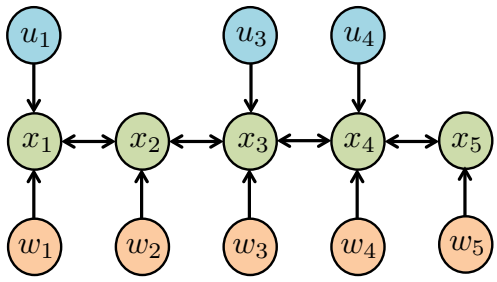


Figure 7.7.1: A realization of the graph Laplacian systems with chain structures. The number of state and input signals are equal to 5 and 3, respectively.

at most d. Furthermore, given the communication speed parameter c, the sub-controllers can interact c times faster than their corresponding subsystems. In particular, given the subsystems i and j with $b_i = b_j = 1$ and $|i - j| = k$, the control action $u_i(t)$ can use $x_j(\tau)$ and $u_j(\tau)$, provided that $\tau \leq (t - k)/c$. The local and communication constraints can be translated into sparsity constraints on the system responses. In particular, define

$$\mathcal{C}_x(t) = \mathcal{S}\left(\text{supp}(A)^{\min\{d-1,\max\{0,c(t-1)\}\}}\right) \tag{7.47}$$

$$\mathcal{C}_u(t) = \mathcal{S}\left(\text{supp}(B)^\top \cdot \text{supp}(A)^{\min\{d-1,\max\{0,c(t-1)\}\}}\right) \tag{7.48}$$

for every $t \in \{1, \ldots, L\}$. Then, the constraints $\Phi_x(t) \in \mathcal{C}_x(t)$ and $\Phi_u(t) \in \mathcal{C}_u(t)$ imply that the resulted controller satisfies the prescribed local and communication constraints. More details on these derivations can be found in [258]. As an example, Figure 7.7.2 shows the sparsity patterns of the system responses for $d = 5$ and $c = 2$.

All the simulations in this section are run on a laptop computer with an Intel Core i7 quad-core 2.50 GHz CPU and 16GB RAM. The reported results are for a serial implementation in MATLAB using the CVX framework and the MOSEK solver with default settings.

Stability analysis

In the first experiment, we consider a small-scale instance of the problem and study the robustness of the designed controller with respect to the uncertainties in the model. In particular, the considered system has 8 states, m of which are randomly chosen and equipped with input signals, for $m \in \{5, 6, 7, 8\}$. We choose $a_i = 1/3$ for every $i \in \{1, \ldots, 8\}$. In order to make the open-loop system marginally unstable, we set $D_i = 0.05$ for $i \in \{2, \ldots, 7\}$ and $D_1 = D_8 = 0.05 - 1/3$. We also assume 10% element-wise uncertainty in the estimated system matrices $\hat{A}$ and $\hat{B}$. In other words, $\hat{A}_{ij}$ is randomly chosen from the interval $[A_{\star ij} - 0.1|A_{\star ij}|, A_{\star ij} + 0.1|A_{\star ij}|]$ for every $(i, j) \in \{1, \ldots, 8\}^2$. Similarly, $\hat{B}_{kl}$ is randomly chosen from the interval $[B_{\star kl} - 0.1|B_{\star kl}|, B_{\star kl} + 0.1|B_{\star kl}|]$ for every $(k, l) \in \{1, \ldots, 8\} \times \{1, \ldots, m\}$. Finally, assume that the estimation error $\epsilon = \max\{\|\hat{A} - A_\star\|_2, \|\hat{B} - B_\star\|_2\}$ is known. Later, we will relax these assumptions and estimate $\hat{A}$, $\hat{B}$, and ϵ directly from the sample trajectories,

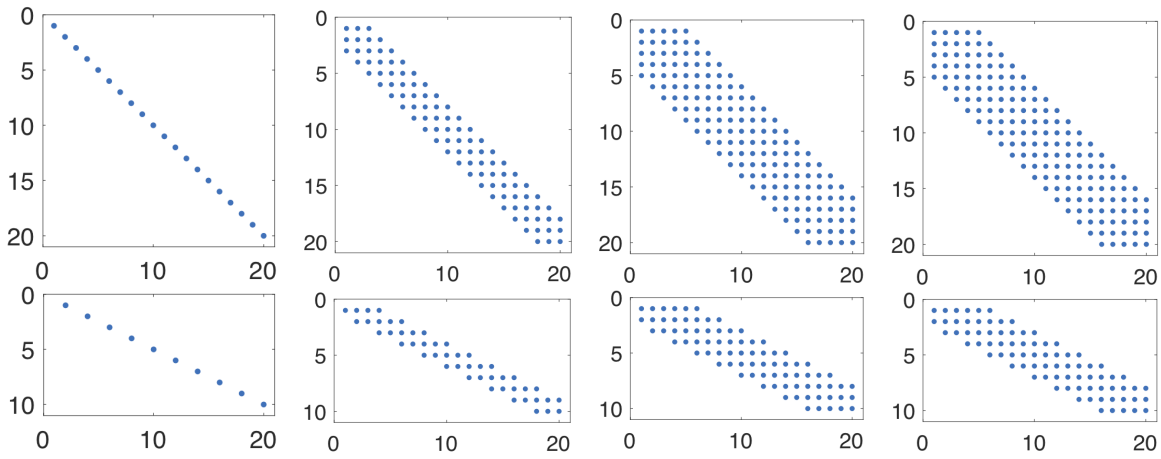


Figure 7.7.2: The sparsity pattern of the system responses $\{\Phi_x(t)\}_{t=1}^4$ and $\{\Phi_u(t)\}_{t=1}^4$ when $d = 5$ and $c = 4$. We assume that $n = 20$ and $b_i = 1$ for every other sub-system. The top row (from left to right) shows the sparsity patterns of $\Phi_x(1), \ldots, \Phi_x(4)$. The bottom row (from left to right) shows the sparsity patterns of $\Phi_u(1), \ldots, \Phi_u(4)$.

using the system identification and bootstrap methods that are introduced in Subsections 7.5 and 7.6. The FIR length L is set to 10. Finally, we set the locality parameter d and the communication speed parameter c to 3 and 2, respectively.

The goal in this simulation is to illustrate the robustness of the introduced distributed controller, compared to the *nominal* distributed (designed based on localized SLS approach in [258]) and centralized controllers (designed using Ricatti equations) that treat $\hat{A}$ and $\hat{B}$ as the true parameters of the system without taking into account their estimation errors.[3] For each input dimension $m \in \{5, 6, 7, 8\}$, we generate 100 independent instances of the problem and design the robust distributed, nominal distributed, and nominal centralized controllers. Figure 7.7.3 shows the ratio of the instances for which each controller stabilizes the system. As can be seen, the proposed robust distributed controller outperforms the nominal distributed controller when m is equal to 6,7, and 8. In particular, the nominal distributed controller either did not exist or failed to stabilize the true system for 100% and 98% of the instances when m is equal to 6 and 7, significantly underperforming compared to the robust distributed controller. Furthermore, the decrease in m deteriorated the performance of the nominal and robust distributed controllers. In particular, for $m = 5$, both controllers ceased to exist for all of the instances. This is indeed not a surprising observation: roughly speaking, designing a distributed controller with restrictive conditions on its locality and communication speed becomes harder as the input dimension decreases. On the other hand, the centralized controller stabilized the true system for 70% of the instances. Notice that

[3]Note that the nominal controller is also known as *certainty equivalent controller* in the literature; see [9, 176].

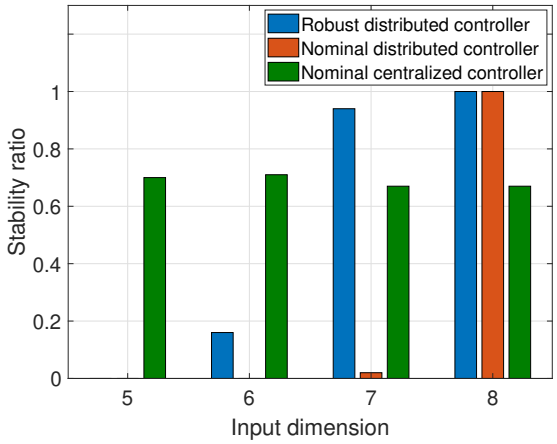


Figure 7.7.3: The ratio of the robust distributed, nominal distributed, and nominal centralized controllers that stabilize the true system.

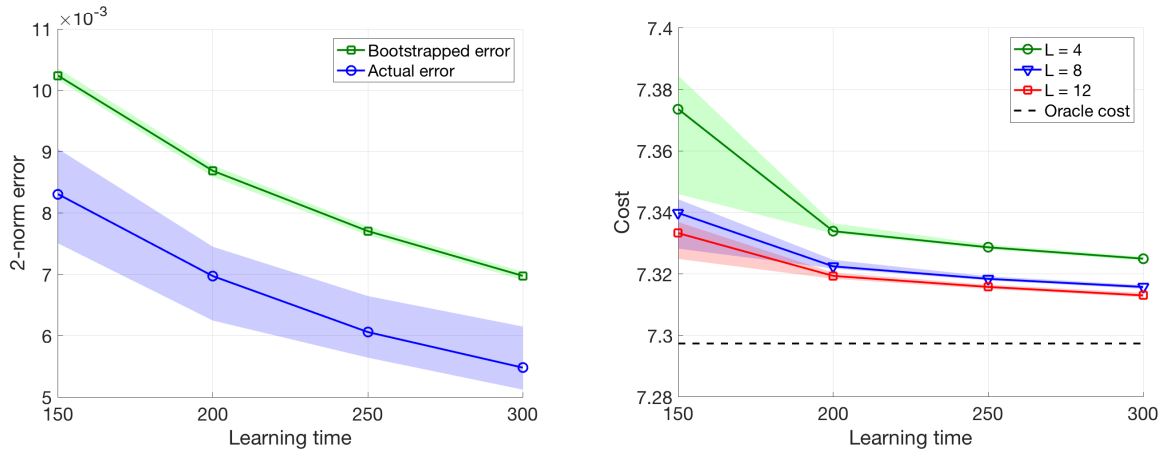
this controller is free of local and communication constraints and hence, its success rate is independent of the input dimension. Overall, the proposed robust distributed controller outperforms the nominal distributed and centralized controllers, provided that the input dimension is not too small.

Another benefit of the proposed controller compared to its nominal counterparts is its ability to identify whether there is "too much uncertainty" in the model. In particular, the infeasibility of the proposed optimization problem (7.26) implies that the estimation error in the model is too large to be accommodated by a robust controller; indeed, such information cannot be inferred by a nominal controller since it is oblivious to the uncertainties in the model.

End-to-end performance

Next, we showcase the end-to-end performance of the proposed robust distributed controller in larger systems. Given a graph Laplacian system, we assume that its dynamics are unknown and first identify the system matrices with a single sample trajectory using the proposed Lasso-based estimator (7.29). Then, we obtain an upper bound on the estimation error using the bootstrap method introduced in Algorithm 5. Finally, we design the robust distributed controller using Algorithm 4.

Consider the system dynamics (7.46) with $n = 40$, where each subsystem is equipped with an input signal (i.e. $B_\star = I$). Assume that $D_i = 0$ and $a_i = 0.2$ for every $i \in \{1, \dots, n\}$. We further multiply the resulting matrix $A_\star$ by 0.99 in order to make it marginally stable. To identify the dynamics, we excite the system with a sequence of randomly generated input signals $u(t) \sim \mathcal{N}(0, 0.1I)$ for $t = 0, 1, \dots, T$. The initial controller K_0 is set to zero



(a) Estimation errors

(b) Performance

Figure 7.7.4: (a) The true and bootstrapped estimation errors with respect to the learning time. (b) The end-to-end performance of the designed robust distributed controller with respect to learning time and for different FIR lengths. The shaded areas show the quartiles.

since the open-loop system is stable. After estimating the system dynamics, we obtain the bootstrapped estimation error using Algorithm 5 with the confidence parameter $\delta = 0.05$ and the number of rounds $N = 500$.

Figure 7.7.4a shows the true and bootstrapped estimation errors with respect to the learning time T. It can be seen that the bootstrapped error is a reliable upper bound on the true estimation error. Given the estimated system matrices and the bootstrapped error, we design the robust distributed controller using Algorithm 4. Figure 7.7.4b illustrates the end-to-end performance of the designed controller with respect to the learning time T and for different FIR lengths L, compared to the oracle cost[4]. It can be seen that the designed distributed controller performs similarly to the oracle one, even when learning time T is as short as 150, which is approximately equal to the number of nonzero elements in $(A_\star, B_\star)$. Furthermore, the performance of the controller improves as the estimation error shrinks or, equivalently, the learning time increases. Furthermore, there is a non-negligible improvement in the performance of the designed controller if the FIR length is increased from 4 to 8. However, the improvement in performance is marginal if the FIR length is increased from 8 to 12, indicating that the $L = 8$ is a reasonable choice for the designed distributed controller.

Finally, we evaluate the runtime of Algorithm 4 for different system dimensions. Consider the same dynamics for the system as before, with n changing from 20 to 150. Figure 7.7.5 shows the empirical runtime of the proposed algorithm. A log-log regression yields an em-

[4]To obtain the oracle cost, we solved the oracle optimization (7.12) to near-optimality after restricting the system responses to FIR filters with length 100. We empirically observed that a further increase in the FIR length has little to no effect on the controller cost

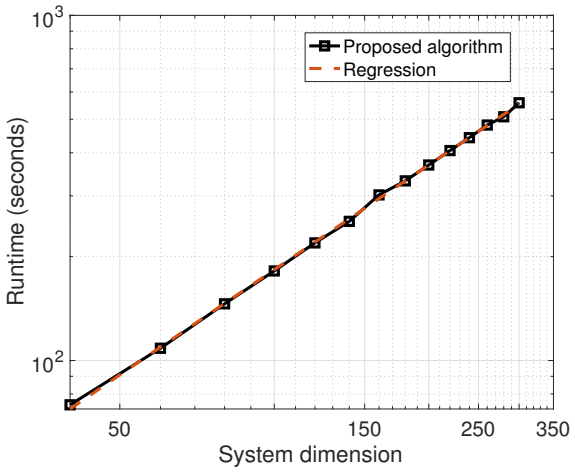


Figure 7.7.5: The empirical runtime of the algorithm with respect to the system dimension (i.e., $n + m$), along with its log-log regression.

pirical time complexity of $\mathcal{O}(n^{1.004})$ for the algorithm, being in line with the theoretical time complexity of the algorithm in Theorem 33. Finally, it is worthwhile to mention that Algorithm 4 is highly parallelizable. In particular, given a machine with n cores, the sub-problems in Algorithm 4 can be solved in parallel and, consequently, the complexity of the proposed algorithm becomes *independent* of the system dimension.

7.8 Summary

We propose a two-step procedure for designing robust distributed controllers for systems with unknown linear and time-invariant dynamics. Our method first actively probes the system to *learn* a model, and then designs a *robust* distributed controller by taking into account the uncertainty of the learned model. By taking advantage of recently-developed sparsity-promoting techniques in system identification, together with the localized System Level Synthesis (SLS) framework, we propose the first stabilizing and learning-based distributed controller with guaranteed sub-linear sample complexity and near-linear (constant order if we assume parallel computation) computational complexity. The graceful scalability of the proposed method makes it particularly useful for the control of large-scale and unknown systems with sparse interconnections.

Appendix

7.A Omitted Proofs

Proof of Theorem 31

To prove Theorem 31, we consider the following operator

$$\|\mathbf{G}\|_{\mathcal{E}_1} = \sup_{z\in\mathbb{T}} \|\mathbf{G}(z)\|_1 \tag{7.49}$$

for every $\mathbf{G} \in \mathcal{RH}_\infty$, where $\mathbb{T}$ is the complex unit circle.. The next lemma describes useful properties of the above operator.

Lemma 48. *The following statements hold:*

1. *(Semi-norm property) The operator $\|\cdot\|_{\mathcal{E}_1}$ is a well-defined semi-norm on $\mathcal{RH}_\infty$.*
2. *(Sub-multiplicativity) For $\mathbf{G}, \mathbf{H} \in \mathcal{RH}_\infty$, we have $\|\mathbf{GH}\|_{\mathcal{E}_1} \le \|\mathbf{G}\|_{\mathcal{E}_1}\|\mathbf{H}\|_{\mathcal{E}_1}$.*
3. *(Hölder's Inequality) For $\mathbf{G} \in \mathcal{RH}_\infty$, we have $\|\mathbf{G}\|_{\mathcal{H}_\infty} \le \sqrt{\|\mathbf{G}\|_{\mathcal{E}_1}\|\mathbf{G}^\top\|_{\mathcal{E}_1}}$.*
4. *For $\mathbf{G} \in \mathcal{RH}_\infty$, we have $\|\mathbf{G}\|_{\mathcal{H}_\infty} \le \sqrt{k}\|\mathbf{G}\|_{\mathcal{E}_1}$, where k is the maximum number of nonzero elements in different rows of $\mathbf{G}$.*
5. *For $\mathbf{G} \in \mathcal{RH}_\infty$, we have $\|\mathbf{G}\|_{\mathcal{E}_1} \le \sum_{t=0}^{\infty} \|G(t)\|_1$.*

Proof. The first statement follows immediately from the definition of $\|\cdot\|_{\mathcal{E}_1}$. Consider the following properties of the induced norms for matrices:

i. $\|\mathbf{G}(z)\mathbf{H}(z)\|_1 \le \|\mathbf{G}(z)\|_1\|\mathbf{H}(z)\|_1$ for every $z \in \mathbb{T}$.

ii. $\|\mathbf{G}(z)\|_2 \le \sqrt{\|\mathbf{G}(z)\|_1\|\mathbf{G}(z)^\top\|_1}$ for every $z \in \mathbb{T}$.

iii. $\|\mathbf{G}(z)\|_1 \le k\|\mathbf{G}(z)^\top\|_1$ for every $z \in \mathbb{T}$.

The second, third, and forth statements of the lemma are followed respectively from (i), (ii), and (iii) combined with (ii), respectively. To show the validity of the last statement, note that

$$\|\mathbf{G}\|_{\mathcal{E}_1} \leq \sup_{z\in\mathbb{T}} \left\| \sum_{t=0}^{\infty} G(t) z^{-t} \right\|_1 \leq \sup_{z\in\mathbb{T}} \sum_{t=0}^{\infty} \left\| G(t) z^{-t} \right\|_1 \leq \sum_{t=0}^{\infty} \|G(t)\|_1 \tag{7.50}$$

□

We provide the proof for Theorem 31 in two steps:

1. We derive conditions under which a feasible solution to (7.26) can be constructed based on the optimal solution of the oracle optimization.

3. We derive the gap between the cost of the designed feasible solution and the oracle cost in terms of $\bar{\epsilon}$ and L. The obtained gap will be used to derive an upper bound on the optimality gap of the synthesized distributed controller.

The following Lemma characterizes a feasible solution to (7.26) based on the system responses of the oracle controller.

Lemma 49. *Suppose that*

$$\bar{\epsilon} < \frac{(1-\rho_\star)\min\{\alpha, 1-\alpha\}}{16 C_\star \rho_\star} k^{-2}, \qquad L > \frac{2\log(k) + \log\left(\frac{2\sqrt{2}(\|A_\star\|_\infty + \|B_\star\|_\infty)}{1-\alpha}\right)}{1-\rho_\star} \tag{7.51}$$

and that $(\hat{A}, \hat{B})$ has the same sparsity as (A, B). Then,

$$\tilde{\Phi}_x(t) = \Phi_x^\star(t), \qquad t = 1, \ldots, L \tag{7.52a}$$

$$\tilde{\Phi}_u(t) = \Phi_u^\star(t), \qquad t = 1, \ldots, L \tag{7.52b}$$

$$\tilde{V}(t) = \begin{cases} 0 & \text{if} \quad t = 0 \\ -\Delta_A \Phi_x^\star(t) - \Delta_B \Phi_u^\star(t) & \text{if} \quad t = 1, \ldots, L-1 \\ -\hat{A}\Phi_x^\star(L) - \hat{B}\Phi_u^\star(L) & \text{if} \quad t = L \end{cases} \tag{7.52c}$$

$$\tilde{\gamma} = \frac{2C_\star\rho_\star}{1-\rho_\star}\left(\frac{1}{\alpha}k^{3/2} + \frac{2\sqrt{2}}{1-\alpha}k^2\right)\bar{\epsilon} + \frac{\sqrt{2}}{1-\alpha}\cdot(\|A_\star\|_\infty + \|B_\star\|_\infty)C_\star k^2 \rho_\star^L, \tag{7.52d}$$

is feasible for (7.26).

Proof. To show the feasibility of the proposed solution, first note that (7.51) results in

$$\frac{2C_\star\rho_\star}{1-\rho_\star}\left(\frac{1}{\alpha}k^{3/2} + \frac{2\sqrt{2}}{1-\alpha}k^2\right)\bar{\epsilon} < 1/2, \quad \frac{\sqrt{2}}{1-\alpha}\cdot(\|A_\star\|_\infty + \|B_\star\|_\infty)C_\star k^2\rho_\star^L < 1/2 \tag{7.53}$$

where, in the second inequality, we used the relation $-\log(\rho_*) \geq 1 - \rho_\star$. This implies that $\tilde{\gamma} < 1$. Furthermore, the definition of $(\tilde{\Phi}_x(t), \tilde{\Phi}_u(t), \tilde{V}(t))$ can be used to show that the

constraints (7.26b), (7.26c), (7.26d), (7.26g), (7.26h) are satisfied. It remains to show the feasibility of (7.26e) and (7.26f). One can write

$$\begin{aligned}\max_j \sum_{t=0}^{L} \|\tilde{V}_{:,j}(t)\|_1 \leq & (\|\hat{A}\|_\infty \|\Phi_x^\star(L)\|_1 + \|\hat{B}\|_\infty \|\Phi_u^\star(L)\|_1) + \sum_{t=1}^{L-1} \epsilon \left(\|\Phi_x^\star(t)\|_1 + \|\Phi_x^\star(t)\|_1\right) \\ \leq & (\|A_\star\|_\infty + \|B_\star\|_\infty + 2\bar{\epsilon}) k C_\star \rho_\star^L + \frac{2C_\star\rho_\star}{1-\rho_\star} k\bar{\epsilon} \\ \leq & (\|A_\star\|_\infty + \|B_\star\|_\infty) k C_\star \rho_\star^L + \frac{4C_\star\rho_\star}{1-\rho_\star} k\bar{\epsilon} \\ \leq & \frac{1-\alpha}{\sqrt{2}} k^{-1}\tilde{\gamma} \\ \leq & (1-\alpha) k_v^{-1/2}\tilde{\gamma} \end{aligned} \tag{7.54}$$

where, in the last inequality, we used the fact that $k_v \leq 2k^2$. Similarly, we have

$$\begin{aligned}\sum_{t=1}^{L} \left\| \begin{bmatrix} \bar{\epsilon}\Phi_x(t) \\ \bar{\epsilon}\Phi_u(t) \end{bmatrix}_{:,j} \right\|_1 &\leq \left(\sum_{t=1}^{L} \|\Phi_x^\star(t)\|_1 + \|\Phi_u^\star(t)\|_1 \right) \bar{\epsilon} \\ &\leq \frac{2C_\star\rho_\star}{1-\rho_\star} k\bar{\epsilon} \\ &\leq \alpha k^{-1/2}\tilde{\gamma} \\ &\leq \alpha k_\phi^{-1/2}\tilde{\gamma} \end{aligned} \tag{7.55}$$

where we used the fact that $k_\phi \leq k$. This completes the proof. □

Now we are ready to present the proof of Theorem 31.

Proof of Theorem 31: Let $\left(\gamma^L, \left\{\Phi_x^L(t)\right\}, \left\{\Phi_u^L(t)\right\}, \left\{V^L(t)\right\}\right)$ be the optimal solution of (7.26). Consider the transfer functions $\mathbf{\Phi}_x^L = \sum_{t=1}^{L} \Phi_x^L(t) z^{-t}$, $\mathbf{\Phi}_u^L = \sum_{t=1}^{L} \Phi_u^L(t) z^{-t}$, and $\mathbf{V}^L = \sum_{t=0}^{L} V^L(t) z^{-t}$. Define $\mathbf{\Delta}^L = \Delta_A \mathbf{\Phi}_x^L + \Delta_B \mathbf{\Phi}_u^L + \mathbf{V}^L$. One can easily verify that

$$\begin{bmatrix} zI - A_\star & -B_\star \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x^L \\ \mathbf{\Phi}_u^L \end{bmatrix} = I + \mathbf{\Delta}^L \tag{7.56}$$

Now, we show that $\|\mathbf{\Delta}^L\|_{\mathcal{H}_\infty} < 1$. To this end, we write

$$\begin{aligned}
\|\mathbf{\Delta}^L\|_{\mathcal{H}_\infty} &\leq \|\Delta_A \mathbf{\Phi}_x^L + \Delta_B \mathbf{\Phi}_u^L\|_{\mathcal{H}_\infty} + \|\mathbf{V}^L\|_{\mathcal{H}_\infty} \\
&\leq \left\| \begin{bmatrix} \frac{\Delta_A}{\bar{\epsilon}} & \frac{\Delta_B}{\bar{\epsilon}} \end{bmatrix} \right\|_2 \left\| \begin{bmatrix} \bar{\epsilon}\mathbf{\Phi}_x^L \\ \bar{\epsilon}\mathbf{\Phi}_u^L \end{bmatrix} \right\|_{\mathcal{H}_\infty} + \|\mathbf{V}^L\|_{\mathcal{H}_\infty} \\
&\overset{(a)}{\leq} \left(\left\| \begin{bmatrix} \bar{\epsilon}\mathbf{\Phi}_x^L \\ \bar{\epsilon}\mathbf{\Phi}_u^L \end{bmatrix} \right\|_{\mathcal{E}_1} \left\| \begin{bmatrix} \bar{\epsilon}\mathbf{\Phi}_x^L \\ \bar{\epsilon}\mathbf{\Phi}_u^L \end{bmatrix}^\top \right\|_{\mathcal{E}_1} \right)^{1/2} + \left(\|\mathbf{V}^L\|_{\mathcal{E}_1} \|\mathbf{V}^{L^\top}\|_{\mathcal{E}_1} \right)^{1/2} \\
&\overset{(b)}{\leq} k_\phi^{1/2} \left\| \begin{bmatrix} \bar{\epsilon}\mathbf{\Phi}_x^L \\ \bar{\epsilon}\mathbf{\Phi}_u^L \end{bmatrix} \right\|_{\mathcal{E}_1} + k_v^{1/2} \|\mathbf{V}^L\|_{\mathcal{E}_1} \\
&\overset{(c)}{\leq} k_\phi^{1/2} \max_j \left\{ \sum_{t=1}^{L} \left\| \begin{bmatrix} \epsilon\Phi_x^L(t) \\ \epsilon\Phi_x^L(t) \end{bmatrix}_{:,j} \right\|_1 \right\} + k_v^{1/2} \max_j \left\{ \|V_{:,j}^L(t)\|_1 \right\} \\
&\leq \alpha\gamma^L + (1-\alpha)\gamma^L \\
&= \gamma^L < 1
\end{aligned} \tag{7.57}$$

where (a), (b), and (c) are due to Lemma 48 and the fact that the maximum number of nonzero elements in different rows of $\begin{bmatrix} \Phi_x^L(t)^\top & \Phi_u^L(t)^\top \end{bmatrix}^\top$ and $V^L(t)$ is upper bounded by k_ϕ and k_v, respectively. Together with Theorem 30, this implies that the derived controller $\mathbf{K}^L = \mathbf{\Phi}_u^L \mathbf{\Phi}_x^{L^{-1}}$ stabilizes the true system. The rest of the proof is devoted to verifying the optimality gap for the designed controller $\mathbf{K}^L$. Based on (7.57) and Lemma 44, one can write

$$\begin{aligned}
J(A_\star, B_\star, \mathbf{K}^L) &= \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x^L \\ \mathbf{\Phi}_u^L \end{bmatrix} (I + \mathbf{\Delta}^L)^{-1} \right\|_{\mathcal{H}_2} \\
&\leq \frac{1}{1 - \|\mathbf{\Delta}^L\|_{\mathcal{H}_\infty}} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x^L \\ \mathbf{\Phi}_u^L \end{bmatrix} \right\|_{\mathcal{H}_2} \\
&\leq \frac{1}{1 - \gamma^L} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x^L \\ \mathbf{\Phi}_u^L \end{bmatrix} \right\|_{\mathcal{H}_2}
\end{aligned} \tag{7.58}$$

Now, consider the transfer functions $\tilde{\mathbf{\Phi}}_x = \sum_{t=1}^{L} \tilde{\Phi}_x(t) z^{-t}$ and $\tilde{\mathbf{\Phi}}_u = \sum_{t=1}^{L} \tilde{\Phi}_u(t) z^{-t}$, where $\tilde{\Phi}_x(t)$ and $\tilde{\Phi}_u(t)$ are defined in Lemma 49. One can write

$$\begin{aligned}
\frac{1}{1 - \gamma^L} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x^L \\ \mathbf{\Phi}_u^L \end{bmatrix} \right\|_{\mathcal{H}_2} &\leq \frac{1}{1 - \tilde{\gamma}} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{\Phi}}_x \\ \tilde{\mathbf{\Phi}}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \\
&\leq \frac{1}{1 - \tilde{\gamma}} J_\star
\end{aligned} \tag{7.59}$$

The first inequality is due to the feasibility of $(\tilde{\gamma}, \tilde{\mathbf{\Phi}}_x, \tilde{\mathbf{\Phi}}_u, \tilde{\mathbf{V}})$. The second equality is due to the fact that $(\tilde{\mathbf{\Phi}}_x, \tilde{\mathbf{\Phi}}_u)$ are the truncations of the system responses when $\mathbf{K}_\star$ acts on the true system to their first L time steps. This implies that

$$\frac{J(A, B, \mathbf{K}^L) - J_\star}{J_\star} \leq \frac{1}{1 - \tilde{\gamma}} - 1 \tag{7.60}$$

It remains to obtain an upper bound on the right hand side of the above inequality. We have

$$\frac{1}{1-\tilde{\gamma}}-1 \leq \frac{1}{1-\left(\underbrace{\frac{2C_\star\rho_\star}{1-\rho_\star}\left(\frac{1}{\alpha}k^{3/2}+\frac{2\sqrt{2}}{1-\alpha}k^2\right)}_{e_1}\bar{\epsilon}+\underbrace{\frac{\sqrt{2}}{1-\alpha}(\|A_\star\|_\infty+\|B_\star\|_\infty)C_\star k^2\rho_\star^L}_{e_2}\right)}-1$$
$$= \frac{e_1+e_2}{1-e_1-e_2} \tag{7.61}$$

Using (7.27), it is easy to verify that we have $e_1 \leq 1/4$ and $e_2 \leq 1/4$. This implies that

$$\frac{J(A,B,\mathbf{K}^L)-J_\star}{J_\star} \leq 2(e_1+e_2) \tag{7.62}$$

Plugging back the definitions of e_1 and e_2, together with some simple algebra completes the proof. □

Proof of Proposition 9

We need a number of lemmas in order to prove this proposition.

Lemma 50. *Given vectors a, b, and a positive definite matrix M, suppose that $a^\top Ma = -a^\top Mb = b^\top Mb$. Then, we have $a = -b$.*

Proof. $a^\top Ma = -a^\top Mb$ and $b^\top Mb = -b^\top Ma$ imply $a^\top M(a+b) = 0$ and $b^\top M(a+b) = 0$. Combining these equations leads to $(a+b)^\top M(a+b) = 0$. Due to the positive definiteness of M, we have $a = -b$. □

Lemma 51. *For every feasible γ, $g(\gamma)^2$ can be reformulated as the optimal solution of the following QP:*

$$\min_x \ \frac{1}{2}x^\top Mx \tag{7.63a}$$
$$\text{s.t. } H_1x \leq h_1+\gamma\mathbf{1} \tag{7.63b}$$
$$H_2x = 0 \tag{7.63c}$$

where

- *x is the vectorized concatenation of $(\{\Phi_x(t)\},\{\Phi_u(t)\})$.*
- *M is a positive definite matrix,*
- *H_1 and H_2 are matrices that only depend on $(\hat{A},\hat{B},\alpha,k)$ and $\mathcal{C}_v$.*

- h_1 *is a vector whose nonzero elements have absolute value greater than 1.*

- **1** *is a vector whose elements are equal to 1.*

Proof. The proof follows after writing the slack variables $\{V(t)\}_{t=0}^{L}$ in terms of $\{\Phi_x(t)\}_{t=1}^{L}$ and $\{\Phi_u(t)\}_{t=1}^{L}$ and linearizing ℓ_1 norm. The details are omitted for brevity. □

Proof of Proposition 9. According to Lemma 51, $g(\gamma)^2$ is equivalent to (7.63) which is a strictly convex QP. Therefore, based on the result of [27], the optimal solution of (7.63) is a continuous function of γ when it is feasible. Therefore, $g(\gamma)^2$ (and hence $g(\gamma)$) is continuous over the interval $[\gamma_0, 1)$. By contradiction, suppose that $\frac{g(\gamma)}{1-\gamma}$ is not unimodal. Then, the quasiconvexity of $\frac{g(\gamma)}{1-\gamma}$ in the interval $[\gamma_0, 1)$ implies that there must exist $\underline{\gamma}$ and $\bar{\gamma}$ such that $\gamma_0 \leq \underline{\gamma} < \bar{\gamma} < 1$ and $\frac{g(\gamma)}{1-\gamma}$ is constant in the interval $[\underline{\gamma}, \bar{\gamma}]$. This implies that $g(\gamma) = c(1-\gamma)$ and $g(\gamma)^2 = c^2(1-\gamma)^2$ for some c and every $\gamma \in [\underline{\gamma}, \bar{\gamma}]$. Define the active set $I(\gamma)$ as the set of the row indices of H_1 corresponding to the active inequalities, i.e., the set of indices i for which we have $(H_1)_{i,:}x = (h_1)_i + \gamma$. Let $H_1[I(\gamma)]$ be the submatrix of H_1 after removing the rows not belonging to $I(\gamma)$. Without loss of generality, we assume that the matrix $H[I(\gamma)] = \begin{bmatrix} H_2^\top & H_1[I(\gamma)]^\top \end{bmatrix}^\top$ is full row rank; otherwise, one can remove the dependent rows of $H[I(\gamma)]$ to reduce it to a full row rank matrix. Now, due to the continuity of $x(\gamma)$, there must exist $\underline{\bar{\gamma}}$ and $\bar{\bar{\gamma}}$ such that $\underline{\gamma} \leq \underline{\bar{\gamma}} < \bar{\bar{\gamma}} \leq \bar{\gamma}$ and $I(\gamma)$ remains the same for every $\gamma \in [\underline{\bar{\gamma}}, \bar{\bar{\gamma}}]$. Let $I(\underline{\bar{\gamma}})$ be denoted as I^* within this interval. Then, (7.63) is reduced to

$$\min_x \ \frac{1}{2}x^\top M x \tag{7.64}$$

$$\text{s.t. } H[I^*]x = h_3[I^*] + \gamma h_4[I^*] \tag{7.65}$$

for every $\gamma \in [\underline{\bar{\gamma}}, \bar{\bar{\gamma}}]$, where $h_3[I^*] = \begin{bmatrix} 0 & h_1[I^*]^\top \end{bmatrix}^\top$ and $h_4[I^*] = \begin{bmatrix} 0 & \mathbf{1}[I^*]^\top \end{bmatrix}^\top$. We consider two cases:

case 1: Suppose that I^* is empty. This implies that $h_4[I^*] = 0$ and therefore, $g(\gamma)$ is constant over the interval $[\underline{\bar{\gamma}}, \bar{\bar{\gamma}}]$ which is a contradiction.

case 2: Suppose that I^* is non-empty and hence, $h_4[I^*] \neq 0$. Due to the feasibility of the affine constraints, strong duality holds. Therefore, by solving the dual of (7.64), one can explicitly write the optimal value of (7.64) in the form of

$$\begin{aligned} g(\gamma)^2 =& \frac{1}{2}(h_3[I^*] + \gamma h_4[I^*])^\top \left(H[I^*]M^{-1}H[I^*]^\top\right)^{-1}(h_3[I^*] + \gamma h_4[I^*]) \\ =& \frac{1}{2}\left(h_4[I^*]^\top \left(H[I^*]M^{-1}H[I^*]^\top\right)^{-1} h_4[I^*]\right)\gamma^2 \\ &+ \left(h_3[I^*]^\top \left(H[I^*]M^{-1}H[I^*]^\top\right)^{-1} h_4[I^*]\right)\gamma \\ &+ \frac{1}{2}\left(h_3[I^*]^\top \left(H[I^*]M^{-1}H[I^*]^\top\right)^{-1} h_3[I^*]\right) \end{aligned} \tag{7.66}$$

Since we assumed that $g(\gamma)^2 = c^2(1-\gamma)^2$ for every $[\underline{\gamma}, \bar{\bar{\gamma}}]$, the following equalities must be satisfied:

$$\begin{aligned} h_4[I^*]^\top \left(H[I^*]M^{-1}H[I^*]^\top\right)^{-1} h_4[I^*] &= -\, h_3[I^*]^\top \left(H[I^*]M^{-1}H[I^*]^\top\right)^{-1} h_4[I^*] \\ &= h_3[I^*]^\top \left(H[I^*]M^{-1}H[I^*]^\top\right)^{-1} h_3[I^*] \end{aligned} \tag{7.67}$$

Note that $\left(H[I^*]M^{-1}H[I^*]^\top\right)^{-1}$ is positive definite due to the fact that $H[I^*]$ is full row rank. Therefore, Lemma 50 implies that $h_4[I^*] = -h_3[I^*]$. On the other hand, $h_4[I^*]$ has an element with value 1 due to the assumption that I^* is non-empty. Furthermore, according to Lemma 51, none of the elements of h_4 have magnitude equal to 1. This contradicts with $h_4[I^*] = -h_3[I^*]$ and completes the proof. □

Proof of Theorem 33

First, we show that the algorithm terminates in $O(L^{3.5}k^7 n \log(n) \log(1/\eta_1) \log(1/\eta_2))$ time. Without loss of generality, suppose that $g(1) < +\infty$. Then, the `while` loop will take at most $\lceil \log(1/\eta_1) \rceil$ iterations to satisfy $|\gamma_c - \gamma_d| \leq \eta_1$ and terminate. On the other hand, at each iteration, one needs to solve $\mathrm{OPT}_1(\gamma_c), \ldots, \mathrm{OPT}_n(\gamma_c)$ and $\mathrm{OPT}_1(\gamma_d), \ldots, \mathrm{OPT}_n(\gamma_d)$ by solving $2n$ instances of the reduced-QPs introduced in Lemma 47. Classical results on the interior methods show that each QP can be solved to $\frac{\eta_2}{n}$-accuracy in $O(L^{3.5}k^7 \log(n) \log(1/\eta_2))$ [37, 270]. Combining these time complexities, one can verify that the algorithm terminates in $O(L^{3.5}k^7 n \log(n) \log(1/\eta_1) \log(1/\eta_2))$.

Next, we prove the statements 1 and 2 of the theorem.

Proof of statement 2: Suppose that $\gamma_0 > 1 - \underline{\eta_1}/2$. Then, it is easy to verify that γ_a and γ_b will obtain the following values at the end of the `while` loop:

$$\gamma_a = 1 - \underline{\eta_1}, \quad \gamma_b = 1 \tag{7.68}$$

Therefore, $1 - \underline{\eta_1}/2$ will be assigned to $\bar{\gamma}$ after the line 18 of the algorithm. This implies that $\gamma_0 > \bar{\gamma}$ and $g(\bar{\gamma}) = +\infty$ due to the definition of γ_0.

Proof of statement 1: An argument similar to the proof of the first statement can be used to show that $g(\bar{\gamma}) < +\infty$ at the termination of the algorithm. Next, we show that we have $\gamma^L \in [\gamma_a, \gamma_b]$ at the end of the `while` loop. This trivially holds if the interior point method that is used to solve $\mathrm{OPT}_i(\gamma_c)$ and $\mathrm{OPT}_i(\gamma_d)$ could achieve zero optimality gap, i.e., $g_{\mathrm{ap}}(\gamma) = g(\gamma)$ at every iteration. As mentioned before, this may not be the case since the values of $g(\gamma)$ are available only up to a nonzero approximation error. By contradiction, suppose $\gamma^L \notin [\gamma_a, \gamma_b]$ at the end of the `while` loop. Together with the unimodal property of $\frac{g(\gamma)}{1-\gamma}$, this implies that one of the following events happens before the line 11 of the algorithm in at least one iteration of the `while` loop:

- $g(\gamma_c)$ and $g(\gamma_d)$ are finite, $\gamma^L \in [\gamma_d, \gamma_b]$, $\frac{g(\gamma_c)}{1-\gamma_c} \geq \frac{g(\gamma_d)}{1-\gamma_d}$, and $\frac{g_{\mathrm{ap}}(\gamma_c)}{1-\gamma_c} < \frac{g_{\mathrm{ap}}(\gamma_d)}{1-\gamma_d}$
- $g(\gamma_c)$ and $g(\gamma_d)$ are finite, $\gamma^L \in [\gamma_a, \gamma_c]$, $\frac{g(\gamma_c)}{1-\gamma_c} < \frac{g(\gamma_d)}{1-\gamma_d}$, and $\frac{g_{\mathrm{ap}}(\gamma_c)}{1-\gamma_c} \geq \frac{g_{\mathrm{ap}}(\gamma_d)}{1-\gamma_d}$

Suppose the first event occurs. In particular, assume that $g(\gamma_c)$ and $g(\gamma_d)$ are finite, $\gamma^L \in [\gamma_d, \gamma_b]$, and $\frac{g(\gamma_c)}{1-\gamma_c} \geq \frac{g(\gamma_d)}{1-\gamma_d}$. It is easy to see that $\gamma_d - \gamma_c > \Delta_\gamma$ due to the definition of Δ_γ in (7.43). On the other hand, notice that $[\gamma_c, \gamma_d] \subseteq [\gamma_0, \gamma^L]$ and hence, $\frac{g(\gamma)}{1-\gamma}$ is decreasing in $[\gamma_0, \gamma^L]$. Therefore, we have $\frac{g(\gamma_c)}{1-\gamma_c} \geq \frac{g(\gamma_d)}{1-\gamma_d} + \Delta_g$ due to the definition of Δ_g in (7.44). This leads to the following series of inequalities:

$$\frac{g_{\text{ap}}(\gamma_c)}{1-\gamma_c} \geq \frac{g(\gamma_c)}{1-\gamma_c} \geq \frac{g(\gamma_d)}{1-\gamma_d} + \Delta_g \geq \frac{g_{\text{ap}}(\gamma_d)}{1-\gamma_d} + \left(\Delta_g - \frac{\eta_2}{1-\gamma_d}\right) \tag{7.69}$$

where the first and last inequalities are due to the fact that $g_{\text{ap}} \geq g(\gamma_c)$ and $g_{\text{ap}}(\gamma_d) \leq g(\gamma_d)+\eta_2$, respectively. Furthermore, it is easy to verify that $\gamma_d \leq \left(1 - \frac{2}{1+\sqrt{5}}\right)\underline{\eta_1}$. Combining this inequality with the assumption $\eta_2 \leq \frac{2}{1+\sqrt{5}}\Delta_g\underline{\eta_1}$ leads to

$$\Delta_g - \frac{\eta_2}{1-\gamma_d} \geq \Delta_g - \frac{1+\sqrt{5}}{2}\frac{\eta_2}{\underline{\eta_1}} \geq 0 \tag{7.70}$$

Together with (7.69), these inequalities result in $\frac{g_{\text{ap}}(\gamma_c)}{1-\gamma_c} \geq \frac{g_{\text{ap}}(\gamma_d)}{1-\gamma_d}$ which is a contradiction. A similar argument can be made to show that the second event does not occur. Therefore, we have $\gamma^L \in [\gamma_a, \gamma_b]$ at the end of the `while` loop and therefore, $|\bar{\gamma} - \gamma^L| \leq \underline{\eta_1}/2$. It remains to show that (7.45) is valid, provided that $\underline{\eta_1} \leq (1-\gamma^L)^2$. One can write

$$\frac{g_{\text{ap}}(\bar{\gamma})}{1-\bar{\gamma}} - \frac{g(\gamma^L)}{1-\gamma^L} \leq \underbrace{\frac{g(\bar{\gamma})}{1-\bar{\gamma}} - \frac{g(\gamma^L)}{1-\gamma^L}}_{(a)} + \underbrace{\frac{\eta_2}{1-\bar{\gamma}}}_{(b)} \tag{7.71}$$

We provide separate upper bounds for (a) and (b). One can verify that the following relation holds for (b):

$$\frac{\eta_2}{1-\bar{\gamma}} \leq \frac{2\eta_2}{\underline{\eta_1}} \leq 2\underline{\eta_1} \tag{7.72}$$

where the first and second inequalities are due to $\bar{\gamma} \leq 1 - \underline{\eta_1}/2$ and the assumption $\eta_2 \leq {\underline{\eta_1}}^2$. Next, we provide an upper bound for (a). One can write

$$\begin{aligned}
\frac{g(\bar{\gamma})}{1-\bar{\gamma}} - \frac{g(\gamma^L)}{1-\gamma^L} &\leq g(\gamma_0)\left|\frac{1}{1-\gamma^L+(\gamma^L-\bar{\gamma})} - \frac{1}{1-\gamma^L}\right| \\
&\leq g(\gamma_0)\frac{|\gamma^L-\bar{\gamma}|}{(1-\gamma^L+(\gamma^L-\bar{\gamma}))(1-\gamma^L)} \\
&\leq g(\gamma_0)\frac{\underline{\eta_1}/2}{(1-\gamma^L-\underline{\eta_1}/2)(1-\gamma^L)}
\end{aligned} \tag{7.73}$$

where $\underline{\eta_1} \leq 2(1-\gamma^L)^2$ is used in the second inequality to ensure that the denominator is positive. On the other hand, we have

$$1 - \gamma^L - \underline{\eta_1}/2 \geq 1 - \gamma^L - (1-\gamma^L)^2 \geq (1-\gamma^L)\gamma^L \tag{7.74}$$

Combining this inequality with (7.73) results in

$$\frac{g(\bar{\gamma})}{1-\bar{\gamma}} - \frac{g(\gamma^L)}{1-\gamma^L} \leq \frac{g(\gamma_0)}{2(1-\gamma^L)^2\gamma^L}\underline{\eta_1} \tag{7.75}$$

This completes the proof. □

Chapter 8

Conclusions and Future Work

This dissertation is aimed to develop scalable and guaranteed computational methods for the efficient operation of complex and safety-critical systems. To this goal, we develop tools in data analytics, optimization, and control, which are the three pillars of reliable computation. Our results are categorized into three parts, namely *machine learning*, *network optimization*, and *system identification and control*. In each of these parts, we take advantage of the underlying structure of the real-world problems, such as their spectral or element-wise sparsity, to develop efficient and practical computational methods. In what follows, we briefly summarize our contributions and future directions.

8.1 Part I. Machine Learning

Graphical Lasso (GL) is a popular method for finding the conditional independence between the entries of a random vector. This technique aims at learning the sparsity pattern of the inverse covariance matrix from a limited number of samples, based on the regularization of a positive-definite matrix. Motivated by the computational complexity of solving the GL for large-scale problems, Chapter 2 of the dissertation provides conditions under which the GL behaves the same as the simple method of thresholding the sample covariance matrix. The conditions make direct use of the sample covariance matrix and are not based on the solution of the GL. More precisely, it is shown that the GL and thresholding techniques are equivalent if: (i) a certain matrix formed based on the sample covariance matrix is both sign-consistent and inverse-consistent, and (ii) the gap between the largest thresholded and the smallest un-thresholded entries of the sample covariance matrix is not too small. Although the GL is believed to be a difficult conic optimization problem, it is proved that it indeed has a closed-form solution in the case where the sparsity pattern of the solution is known to be acyclic. This result is then extended to general sparse graphs and an explicit formula is derived as an approximate solution of the GL, where the approximation error is also quantified in terms of the structure of the sparsity graph. The significant speedup and graceful scalability of the proposed explicit formula compared to other state-of-the-art

methods is showcased on different real-world and randomly generated data sets.

Chapter 3 of the dissertation deals with the non-negative rank-1 robust principal component analysis (RPCA), where the goal is to recover the true non-negative principal component of the data matrix exactly, using partial and potentially noisy measurements of the data matrix. The main difference between the RPCA and its classical counterpart is the sparse-but-arbitrarily-large values of the additive noise. The most commonly known methods for solving the RPCA are based on convex relaxations, where the problem is *convexified* at the expense of significantly increasing the number of variables. In this work, we show that the original non-convex and non-smooth ℓ_1 formulation of the positive rank-1 RPCA problem based on the well-known Burer-Monteiro approach has benign landscape, i.e., it does not have any spurious local solution and has a unique global solution that coincides with the true components. In particular, we provide strong deterministic and statistical guarantees for the benign landscape of the positive rank-1 RPCA and show that the absence of spurious local solutions is guaranteed to hold with a surprisingly large number of corrupted measurements. While the results on "no spurious local minima" are ubiquitous for smooth problems related to matrix completion and sensing, to the best of our knowledge, the results presented in this chapter are the first to prove the absence of local minima when the objective function is non-smooth. Finally, through extensive simulations, we provide strong evidence suggesting that the proposed results may hold for the general non-negative rank-r RPCA. The extension of our theoretical results to this generalized problem is left as a future work.

8.2 Part II. Network Optimization

Network flow problems play a central role in operations research, computer science and engineering. Due to the complexity of these problems, the main focus has long been on lossless flow networks and more recently on networks with linear loss functions. Chapter 4 of the dissertation studies the generalized network flow (GNF) problem, which aims to optimize the flows over a lossy flow network. It is assumed that each node is associated with an injection and that the two flows at the endpoints of each line are related to each other via an arbitrary convex monotonic function. The GNF problem is hard to solve due to the presence of nonlinear equality flow constraints. It is shown that although GNF is highly nonconvex, globally optimal injections can be found by means of a convexified generalized network flow (CGNF) problem. It is also proven that CGNF obtains globally optimal flows for GNF, as long as the optimal injection vector is a Pareto point. In the case where CGNF returns a wrong (infeasible) flow vector for GNF, the network can be decomposed into two subgraphs such that: (i) the flows found by CGNF for one of the subgraphs are all globally optimal, and (ii) the flows obtained by CGNF for the lines between the subgraphs are all correct and at their limits (i.e., the lines between the two subgraphs are congested). The set of all globally optimal flow vectors are characterized based on the optimal injection vector found using CGNF. This set may be infinite, non-convex, and disconnected, while it belongs to the boundary of a convex set. Finally, we generalize the GNF problem and

its convexification to include coupling convex constraints on the flows or the injections. An immediate application of this work is in power systems, where the goal is to optimize the power flows at nodes and over lines of a power grid. Recent work on the optimal power flow problem has shown that this non-convex problem can be solved via a convex relaxation after two approximations: relaxing angle constraints (by adding virtual phase shifters) and relaxing power balance equations to inequality flow constraints. The results on GNF prove that the second approximation (on power balance equations) is redundant under a practical angle assumption.

Chapter 5 of the dissertation is concerned with the optimal transmission switching in power systems. Finding an optimal topology of a power system subject to operational and security constraints is a daunting task. In this problem, certain lines are fixed/uncontrollable, whereas the remaining ones could be controlled via on/off switches. The objective is to co-optimize the topology of the grid and the parameters of the system (e.g., generator outputs). Common techniques for solving this problem are mostly based on mixed-integer linear or quadratic reformulations using the big-M or McCormick inequalities followed by iterative methods, such as branch-and-bound or cutting-plane algorithms. The performance of these methods partly relies on the strength of the convex relaxation of these reformulations. In this chapter, it is shown that finding the optimal parameters of a linear or convex reformulation based on big-M or McCormick inequalities is NP-hard. Furthermore, the inapproximability of these parameters up to any constant factor is proven. Despite the negative results on the complexity of the problem, a simple bound strengthening method is developed to significantly strengthen mixed-integer reformulations of the OTS, provided that there exists a connected spanning subgraph of the network with fixed lines. This bound strengthening method can be used as a preprocessing step even in an offline fashion, before forecasting the demand in the system. Through extensive computational experiments, it is verified that this simple preprocessing technique can significantly improve the runtime of the mixed-integer solvers without sacrificing optimality as is done in standard formulations with restricting constraints in many test cases, including the IEEE 118-bus system and Polish networks.

8.3 Part III. System Identification and Control

In chapter 6, we consider the problem of sparse system identification of linear time-invariant (LTI) systems, where the goal is to estimate the sparse structure of the system matrices based on a single sample trajectory of the dynamics. A Lasso-type estimator is introduced to identify the parameters of the system, while promoting their sparsity via a ℓ_1-regularization technique. By carefully examining the underlying properties of the system—such as its stability and mutual incoherency—we provide non-asymptotic bounds on the accuracy of the proposed estimator. In particular, we show that it correctly identifies the sparsity structure of the system matrices and enjoys a sharp upper bound on its estimation error, provided that the learning time exceeds a threshold. We further show that this threshold scales polynomially in the number of nonzero elements but logarithmically in the system

dimensions.

We extend these results in Chapter 7 and propose a two-step procedure for designing robust distributed controllers for systems with unknown linear and time-invariant dynamics. Our method first actively probes the system to *learn* a model, and then designs a *robust* distributed controller by taking into account the uncertainty of the learned model. By taking advantage of our developed sparsity-promoting techniques in system identification, together with the localized System Level Synthesis (SLS) framework, we propose the first stabilizing and learning-based distributed controller with guaranteed sub-linear sample complexity and near-linear (constant order if we assume parallel computation) computational complexity. The graceful scalability of the proposed method makes it particularly useful for the control of large-scale and unknown systems with sparse interconnections.

8.4 Future Directions

The work comprising this dissertation is a step towards building high-performance computational techniques for societal problems. To move forward, interdisciplinary research should be conducted with the goal of striking a balance between two major paradigms, namely *theory* and *application* of computational techniques. In what follows, we will discus some of the possible future research directions.

Distributed learning and control: Richer models. Most of the existing learning-based control techniques are focused on either the richness of their learned model (culminating in reinforcement learning) or the guaranteed robustness of the control actions (e.g. robust linear-quadratic controllers), with one coming at the expense of the other. However, most real-world systems, such as smart grids and automated transportation networks, are nonlinear and safety-critical, and they must be controlled in real-time. Moving forward, we need to develop efficient learning-based control frameworks for nonlinear dynamics, taking into account their safety constraints. In particular, we need to design efficient learning methods with guaranteed robustness that are applicable to richer system models and control paradigms, such as nonlinear and online (adaptive) learning-based control.

Global guarantees for data-driven nonsmooth optimization. In practice, local-search algorithms can efficiently recover globally-optimal solutions in some of the nonsmooth optimization problems in machine learning, such as robust low-rank matrix recovery. In contrast, undesired local minima are common and hard to avoid in a number of emerging nonsmooth problems, such as the training of deep nonlinear neural networks, as well as the robust state estimation of power systems with large-and-sparse noise values. A common feature of these problems is that *data leads the process of decision making*. A question therefore arises as to whether there exists a unifying framework to systematically study the effect of data on the global landscape of nonsmooth optimization problems. We consider answering this question as an enticing challenge for future research, as the existing techniques can only target a limited class of problems with specific structures. Furthermore, such insight can be used to

study how to reformulate a data-driven nonsmooth optimization so that its spurious local solutions disappear.

Massively-scalable algorithms: Bridging the gap between theory and practice. In recent years, the scale of real-life problems has significantly outpaced the ability of existing algorithms to operate in real-time. Despite being massive-scale, the real-world systems are structured in many ways: they may be modeled as tree-like graphs (e.g. power and transportation systems), have local structures (e.g. network of self-driving vehicles), or enjoy sparsity in their pattern, rank, etc. (e.g. low-rank representation of data in recommender systems). While such application-specific structures are well-known to domain experts, most of the current computational methods remain oblivious to them. We believe that exploiting the underlying structure of real-world problems is a key game changer in the pursuit of massively-scalable computational methods. To achieve this goal, we need "bilingual" researchers well-versed in both *theory* and *practice* to bridge the gap between these two major paradigms, within the realm of computational methods.

Mathematical tools for smart infrastructures. The integration of Internet of Things (IoT) sensors in urban infrastructure has taken us one step closer to the design of smart and autonomous cities, pinpointing the critical role of data analytics in their efficient operations. The infrastructure of the future must process the data in real-time and make reliable decisions. This calls for highly-efficient and data-driven computational methods that can automatically diagnose the errors caused by natural disasters, malicious activities, or the "human-in-the-loop". The lack of reliability in the operation of the existing infrastructures has been proven to be catastrophic in recent years. For instance, the major blackouts of 1977, 2003, and 2019 in Northeast United States and Canada are strong evidences highlighting the inability of existing state estimation techniques in power systems to reliably predict and prevent the cascading effect of a failure in the system. With the goal of addressing the emerging challenges in power systems, ARPA-E has recently announced its ambitious plan to revolutionize the operation of power grids by shifting towards data-driven approaches.[1] The final report on the 2003 blackout in the United States and Canada explicitly recommends to *"Evaluate and adopt better real-time tools for operators and reliability coordinators"* in order to ensure the safety of the power grid for the years to follow.[2]

This indeed calls for a novel, efficient, and trustworthy computational paradigm that can be easily used in tomorrow's interconnected systems; a goal that can be achieved by pushing the boundaries of science in both optimization and data analytics, and by conducting interdisciplinary research at the intersection of operations research, artificial intelligence, and computer science.

[1] https://arpa-e.energy.gov/?q=news-item

[2] http://eta-publications.lbl.gov/sites/default/files/2003-blackout-us-canada.pdf

Bibliography

[1] A. Ott, VP. *Private Communication*. Norristown, PA: PJM, 2008.

[2] S. Nauman, VP. *Private Communication*. Chicago, IL: Exelon, 2008.

[3] Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. "Model-Free Linear Quadratic Control via Reduction to Expert Prediction". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 3108–3117.

[4] 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*. 2011, pp. 1–26.

[5] Marc Abeille and Alessandro Lazaric. "Improved regret bounds for thompson sampling in linear quadratic control problems". In: *International Conference on Machine Learning*. 2018, pp. 1–9.

[6] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. "Network flows: theory, algorithms, and applications". In: *Prentice-Hall* (1993).

[7] M. Selim Aktürk, Alper Atamtürk, and Sinan Gürel. "A strong conic quadratic reformulation for machine-job assignment with controllable processing times". In: *Operations Research Letters* 37.3 (2009), pp. 187–191.

[8] A. Araposthatis, S. Sastry, and P. Varaiya. "Analysis of power-flow equation". In: *International Journal of Electrical Power & Energy Systems* 3 (1981), pp. 115–126.

[9] Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Corporation, 2013.

[10] Karl Johan Åström and Peter Eykhoff. "System identification—a survey". In: *Automatica* 7.2 (1971), pp. 123–162.

[11] Karl Johan Åström and Björn Wittenmark. "On self tuning regulators". In: *Automatica* 9.2 (1973), pp. 185–199.

[12] Anil Aswani et al. "Provably safe and robust learning-based model predictive control". In: *Automatica* 49.5 (2013), pp. 1216–1226.

[13] Alper Atamtürk and Vishnu Narayanan. "Cuts for conic mixed-integer programming". In: *Integer Programming and Combinatorial Optimization: 12th International IPCO Conference, Proceedings, Springer Berlin Heidelberg* (2007), pp. 16–29.

[14] Francis Bach. "Breaking the curse of dimensionality with convex neural networks". In: *Journal of Machine Learning Research* 18.19 (2017), pp. 1–53.

[15] Francis Bach et al. "Optimization with sparsity-inducing penalties". In: *Foundations and Trends® in Machine Learning* 4.1 (2012), pp. 1–106.

[16] R. Bacher and H. Glavitsch. "Loss reduction by network switching". In: *IEEE Transactions on Power Systems* 3.2 (1988), pp. 447–454.

[17] R. Bacher and H. Glavitsch. "Network topology optimization with security constraints". In: *IEEE Transactions on Power Systems* 1.4 (1986).

[18] R. Baldick. *Applied Optimization: Formulation and Algorithms for Engineering Systems*. Cambridge, 2006.

[19] Bassam Bamieh, Fernando Paganini, and Munther A. Dahleh. "Distributed Control of Spatially Invariant Systems". In: *Automatic Control, IEEE Transactions on* 47.7 (2002), pp. 1091–1107.

[20] Bassam Bamieh and Petros G Voulgaris. "A convex characterization of distributed control problems in spatially invariant systems with communication constraints". In: *Systems & Control Letters* 54.6 (2005), pp. 575–583.

[21] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data". In: *Journal of Machine Learning Research* 9 (2008), pp. 485–516.

[22] C. Barrows, S. Blumsack, and R Bent. "Computationally efficient optimal transmission switching: Solution space reduction". In: *IEEE Power and Energy Society General Meeting* (2012), pp. 1–8.

[23] M. S. Bazaraa, J. J. Jarvis, and H. D. Sherali. "Linear Programming and Network Flows". In: *John Wiley & Sons* (1990).

[24] Alexandre Belloni, Victor Chernozhukov, et al. "Least squares after model selection in high-dimensional sparse models". In: *Bernoulli* 19.2 (2013), pp. 521–547.

[25] P. Belotti et al. "Disjunctive inequalities: applications and extensions". In: *Wiley Encyclopedia of Operations Research and Management Science* (2011).

[26] Steven J. Benson, Yinyu Ye, and Xiong Zhang. "Solving large-scale sparse semidefinite programs for combinatorial optimization". In: *SIAM Journal on Optimization* 10.2 (2000), pp. 443–461.

[27] Arjan B Berkelaar, Kees Roos, and Tamás Terlaky. "The optimal set and optimal partition approach to linear and quadratic programming". In: *Advances in Sensitivity Analysis and Parametic Programming*. Springer, 1997, pp. 159–202.

[28] Dimitri P Bertsekas et al. *Dynamic programming and optimal control*. Vol. 1. 2. Athena scientific Belmont, MA, 1995.

[29] D. Bertsimas and M. Sim. "Robust discrete optimization and network flows". In: *Mathematical Programming* 98 (2003), pp. 49–71.

[30] D. Bertsimas and S. Stock-Paterson. "The Traffic Flow Management Rerouting Problem in Air Traffic Control: A Dynamic Network Flow Approach". In: *Transportation Science* 34 (2000), pp. 239–255.

[31] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. "Global optimality of local search for low rank matrix recovery". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3873–3881.

[32] D. Bienstock et al. "Minimum cost capacity installation for multicommodity network flows". In: *Mathematical Programming* 81 (1998), pp. 177–199.

[33] Daniel Bienstock and Sara Mattia. "Using mixed-integer programming to solve power grid blackout problems". In: *Discrete Optimization* 4.1 (2007), pp. 115–141.

[34] Mariusz Bojarski et al. "End to end learning for self-driving cars". In: *arXiv preprint arXiv:1604.07316* (2016).

[35] S. Bose et al. "Optimal Power Flow Over Tree Networks". In: *Proceedings of the Forth-Ninth Annual Allerton Conference* (2011), pp. 1342–1348.

[36] Léon Bottou, Frank E Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *SIAM Review* 60.2 (2018), pp. 223–311.

[37] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge, 2004.

[38] H. Brannlund et al. "Optimal Short Term Operation Planning of a Large Hydrothermal Power System Based on a Nonlinear Network Flow Concept". In: *IEEE Transactions on Power Systems* 1 (1986), pp. 75–81.

[39] Naama Brenner, William Bialek, and Rob de Ruyter Van Steveninck. "Adaptive rescaling maximizes information transmission". In: *Neuron* 26.3 (2000), pp. 695–702.

[40] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[41] 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.

[42] James V Burke, Adrian S Lewis, and Michael L Overton. "A robust gradient sampling algorithm for nonsmooth, nonconvex optimization". In: *SIAM Journal on Optimization* 15.3 (2005), pp. 751–779.

[43] *Caltrans Performance Management System (PeMS)*. 2017. URL: `http://pems.dot.ca.gov`.

[44] Emmanuel J Candes, Justin K Romberg, and Terence Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59.8 (2006), pp. 1207–1223.

[45] Emmanuel J Candes and Terence Tao. "Decoding by linear programming". In: *IEEE transactions on information theory* 51.12 (2005), pp. 4203–4215.

[46] Emmanuel J Candès et al. "Robust principal component analysis?" In: *Journal of the ACM (JACM)* 58.3 (2011), p. 11.

[47] Emmanuel Candes and Justin Romberg. "Sparsity and incoherence in compressive sampling". In: *Inverse Problems* 23.3 (2007), pp. 969–985.

[48] Arvind Caprihan, Godfrey D Pearlson, and Vincent D Calhoun. "Application of principal component analysis to distinguish patients with schizophrenia from healthy controls based on fractional anisotropy measurements". In: *Neuroimage* 42.2 (2008), pp. 675–682.

[49] J. Carpentier. "Contribution to the economic dispatch problem". In: *Bulletin Society Francaise Electriciens* (1962).

[50] Venkat Chandrasekaran et al. "Rank-sparsity incoherence for matrix decomposition". In: *SIAM Journal on Optimization* 21.2 (2011), pp. 572–596.

[51] Robin Chaney and Allen Goldstein. "An extension of the method of subgradients". In: *Nonsmooth Optimization* (1978), pp. 51–70.

[52] Han-Fu Chen and Lei Guo. *Identification and stochastic adaptive control*. Original work published 1991. Springer Science & Business Media, 2012.

[53] Yuejie Chi, Yue M Lu, and Yuxin Chen. "Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview". In: *arXiv preprint arXiv:1809.09573* (2018).

[54] Frank H Clarke. *Optimization and nonsmooth analysis*. Vol. 5. Siam, 1990.

[55] C. Coffrin et al. "Primal and dual bounds for optimal transmission switching". In: *IEEE Power Systems Computation Conference (PSCC)* (2014), pp. 1–8.

[56] Carleton Coffrin, Dan Gordon, and Paul Scott. "NESTA, the NICTA energy system test case archive". In: *arXiv preprint arXiv:1411.0359* (2014).

[57] Thomas Frederick Coleman and Yuying Li, eds. *Large-scale numerical optimization*. Vol. 46. SIAM, 1990.

[58] Thomas H Cormen. *Introduction to algorithms*. MIT press, 2009.

[59] Rita Cucchiara et al. "Detecting moving objects, ghosts, and shadows in video streams". In: *IEEE transactions on pattern analysis and machine intelligence* (2003).

[60] Ying Cui, Jong-Shi Pang, and Bodhisattva Sen. "Composite difference-max programs for modern statistical estimation problems". In: *arXiv preprint arXiv:1803.00205* (2018).

[61] Frank E Curtis and Michael L Overton. "A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization". In: *SIAM Journal on Optimization* 22.2 (2012), pp. 474–500.

[62] Sarah Dean et al. "On the sample complexity of the linear quadratic regulator". In: *arXiv preprint arXiv:1710.01688* (2017).

[63] Sarah Dean et al. "Regret bounds for robust adaptive control of the linear quadratic regulator". In: *Advances in Neural Information Processing Systems*. 2018, pp. 4192–4201.

[64] H. W. Dommel and W. F. Tinney. "Optimal Power Flow Solutions". In: *IEEE Transactions on Power Apparatus and Systems* (1968).

[65] David L Donoho. "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59.6 (2006), pp. 797–829.

[66] Jianzhong Du and Joseph Y. T. Leung. "Minimizing total tardiness on one machine is NP-hard". In: *Mathematics of Operations Research* 15.3 (1990), pp. 483–495.

[67] Yan Duan et al. "Benchmarking deep reinforcement learning for continuous control". In: *International Conference on Machine Learning*. 2016, pp. 1329–1338.

[68] Geir E. Dullerud and Raffaello D'Andrea. "Distributed Control of Heterogeneous Systems". In: *Automatic Control, IEEE Transactions on* 49.12 (2004), pp. 2113–2128.

[69] Susan Dumais et al. "Inductive learning algorithms and representations for text categorization". In: *Proceedings of the seventh international conference on Information and knowledge management*. ACM. 1998, pp. 148–155.

[70] J. Edmonds and R. M. Karp. "Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems". In: *Journal of the ACM* 19 (1972), pp. 248–264.

[71] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[72] Hilmi E. Egilmez, Eduardo Pavez, and Antonio Ortega. "Graph learning from data under laplacian and structural constraints". In: *IEEE Journal of Selected Topics in Signal Processing* 11.6 (2017), pp. 825–841.

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

[74] P Erdös and A Rényi. "On random graphs I". In: *Publ. Math. Debrecen* 6 (1959), pp. 290–297.

[75] Jianqing Fan and Yuan Liao. "Endogeneity in high dimensions". In: *Annals of statistics* 42.3 (2014), p. 872.

[76] Jianqing Fan and Jinchi Lv. "A selective overview of variable selection in high dimensional feature space". In: *Statistica Sinica* 20.1 (2010), p. 101.

[77] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. "Finite time identification in unstable linear systems". In: *Automatica* 96 (2018), pp. 342–353.

[78] Makan Fardad, Fu Lin, and Mihailo R. Jovanović. "Sparsity-promoting optimal control for a class of distributed systems". In: *American Control Conference* (2011), pp. 2050–2055.

[79] M. Farivar and S. H. Low. "Branch Flow Model: Relaxations and Convexification—Part II". In: *IEEE Transactions on Power Systems* 28.3 (2013), pp. 2565–2572.

[80] Giovanni Fasano et al. "A linesearch-based derivative-free approach for nonsmooth constrained optimization". In: *SIAM Journal on Optimization* 24.3 (2014), pp. 959–992.

[81] S. Fattahi et al. "Conic relaxations of the unit commitment problem". In: *Energy* 134 (2017), pp. 1079–1095.

[82] Salar Fattahi and Javad Lavaei. "On the convexity of optimal decentralized control problem and sparsity path". In: *American Control Conference (ACC), 2017*. IEEE. 2017, pp. 3359–3366.

[83] Salar Fattahi, Javad Lavaei, and Alper Atamtürk. "Promises of Conic Relaxations in Optimal Transmission Switching of Power Systems". In: *to appear in Proc. 56th IEEE Conference on Decision and Control* (2017).

[84] Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. "Learning Sparse Dynamical Systems from a Single Sample Trajectory". In: *arXiv preprint arXiv:1904.09396* (2019).

[85] Salar Fattahi, Richard Y Zhang, and Somayeh Sojoudi. "Sparse Inverse Covariance Estimation for Chordal Structures". In: *https://arxiv.org/abs/1711.09131* (2018).

[86] Salar Fattahi et al. "Conic Relaxations of the Unit Commitment Problem". In: *Energy* 134 (2017), pp. 1079–1095.

[87] Salar Fattahi et al. "Transformation of optimal centralized controllers into near-globally optimal static distributed controllers". In: *IEEE Transactions on Automatic Control* 64.1 (2019), pp. 63–77.

[88] Federal Energy Regulatory Commission. Sept. 2017. URL: https://www.ferc.gov/industries/electric/indus-act/market-planning.asp.

[89] Federal Energy Regulatory Commission. "Energy Policy Act of 2005". In: (2006). URL: https://www.ferc.gov/legal/fed-sta/epact-fact-sheet.pdf.

[90] Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. "Mind the duality gap: safer rules for the lasso". In: *arXiv preprint arXiv:1505.03410* (2015).

[91] Lino Figueiredo et al. "Towards the development of intelligent transportation systems". In: *IEEE Intelligent Transportation Systems* (2001), pp. 1206–1211.

[92] E. B. Fisher, R. P. O'Neill, and M. C. Ferris. "Optimal transmission switching". In: *IEEE Transactions on Power Systems* 23.3 (2008), pp. 1346–1355.

[93] S. Fliscounakis et al. "Topology influence on loss reduction as a mixed integer linear programming problem". In: *Power Tech, 2007 IEEE Lausanne* (2007), pp. 1987–1990.

[94] L. R. Ford and D. R. Fulkerson. "Flows in networks". In: *Princeton University Press* (1962).

[95] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Vol. 1. 3. Basel: Birkhäuser, 2013.

[96] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[97] J. D. Fuller, R. Ramasra, and A. Cha. "Fast heuristics for transmission-line switching". In: *IEEE Transactions on Power Systems* 27.3 (2012), pp. 1377–1386.

[98] Jochen Garcke, Michael Griebel, and Michael Thess. "Data mining with sparse grids". In: *Computing* 67.3 (2001), pp. 225–253.

[99] D. Gayme and U. Topcu. "Optimal power flow with large-scale storage integration". In: *IEEE Transactions on Power Systems* 28.2 (2013), pp. 709–717.

[100] Rong Ge, Chi Jin, and Yi Zheng. "No spurious local minima in nonconvex low rank problems: A unified geometric analysis". In: *arXiv preprint arXiv:1704.00708* (2017).

[101] Rong Ge, Jason D Lee, and Tengyu Ma. "Matrix completion has no spurious local minimum". In: *Advances in Neural Information Processing Systems*. 2016, pp. 2973–2981.

[102] Rong Ge et al. "Escaping from saddle points-online stochastic gradient for tensor decomposition". In: *Conference on Learning Theory*. 2015, pp. 797–842.

[103] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. "Safe feature elimination for the lasso and sparse supervised learning problems". In: *arXiv preprint arXiv:1009.4219* (2010).

[104] J. L. Goffin et al. "Solving nonlinear multicommodity flow problems by the analytic center cutting plane method". In: *Mathematical Programming* 76 (1996), pp. 131–154.

[105] A. V. Goldberg, E. Tardos, and R. E. Tarjan. "Network Flow Algorithms". In: *Flows, Paths and VLSI (Springer, Berlin)* (1990), pp. 101–164.

[106] D. Goldfarb and J. Hao. "Polynomial-time primal simplex algorithms for the minimum cost network flow problem". In: *Algorithmica* 8 (1992), pp. 145–160.

[107] AA Goldstein. "Optimization of Lipschitz continuous functions". In: *Mathematical Programming* 13.1 (1977), pp. 14–22.

[108] Gene H Golub and Charles F Van Loan. *Matrix computations*. Vol. 3. JHU press, 2012.

[109] Graham Clifford Goodwin and Robert L Payne. *Dynamic system identification: experiment design and data analysis*. Academic press, 1977.

[110] Alexander N Gorban et al. *Principal manifolds for data visualization and dimension reduction*. Vol. 58. Springer, 2008.

[111] G. Granelli et al. "Optimal network reconfiguration for congestion management by deterministic and genetic algorithms". In: *Electric Power Systems Research* 76.6 (2006), pp. 549–556.

[112] Maxim Grechkin et al. "Pathway Graphical Lasso". In: *AAAI* (2015), pp. 2617–2623.

[113] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

[114] Nora Hartsfield and Gerhard Ringel. *Pearls in graph theory: a comprehensive introduction*. Courier Corporation, 2013.

[115] K. W. Hedman, S. S. Oren, and R. P. O'Neill. "Optimal transmission switching: economic efficiency and market implications". In: *Journal of Regulatory Economics* 40.2 (2011), p. 111.

[116] K. W. Hedman et al. "Co-optimization of generation unit commitment and transmission switching with N-1 reliability". In: *IEEE Transactions on Power Systems* 25.2 (2010), pp. 1052–1063.

[117] K. W. Hedman et al. "Optimal transmission switching with contingency analysis". In: *IEEE Transactions on Power Systems* 24.3 (2009), pp. 1577–1586.

[118] Kory W Hedman, Shmuel S Oren, and Richard P O'Neill. "A review of transmission switching and network topology optimization". In: *Power and Energy Society General Meeting, 2011 IEEE*. IEEE. 2011, pp. 1–7.

[119] Oliver Herr and Asvin Goel. "Comparison of two integer programming formulations for a single machine family scheduling problem to minimize total tardiness". In: *Procedia CIRP* 19.174-179 (2014).

[120] H. Hijazi, C. Coffrin, and P. V. Hentenryck. "Convex quadratic relaxations for mixed-integer nonlinear programs in power systems". In: *Mathematical Programming Computation* (2013), pp. 1–47.

[121] I. A. Hiskens and R. J. Davy. "Exploring the power flow solution space boundary". In: *IEEE Transactions on Power Systems* 16.3 (2001), pp. 389–395.

[122] Yu-Chi Ho and K.-C. Chu. "Team decision theory and information structures in optimal control problems–Part I". In: *Automatic Control, IEEE Transactions on* 17.1 (1972), pp. 15–22.

[123] Dorit S Hochbaum. "Approximating covering and packing problems: set cover, vertex cover, independent set, and related problems". In: *Approximation algorithms for NP-hard problems*. PWS Publishing Co. 1996, pp. 94–143.

[124] Patrik O Hoyer. "Non-negative matrix factorization with sparseness constraints". In: *Journal of machine learning research* 5.Nov (2004), pp. 1457–1469.

[125] Cho-Jui Hsieh et al. "BIG & QUIC: Sparse inverse covariance estimation for a million variables". In: *Advances in neural information processing systems*. 2013, pp. 3165–3173.

[126] Cho-Jui Hsieh et al. "QUIC: quadratic approximation for sparse inverse covariance estimation". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 2911–2947.

[127] Daniel Hsu, Sham M Kakade, and Tong Zhang. "Robust matrix decomposition with sparse corruptions". In: *IEEE Transactions on Information Theory* 57.11 (2011), pp. 7221–7234.

[128] John Hull and Alan White. "Pricing interest-rate-derivative securities". In: *The Review of Financial Studies* 3.4 (1990), pp. 573–592.

[129] R. A. Jabr. "Optimal power flow using an extended conic quadratic formulation". In: *IEEE Transactions on Power Systems* 23.3 (2008), pp. 1000–1008.

[130] W. S. Jewell. "Optimal flow through networks with gains". In: *Operations Research* 10 (1962), pp. 476–499.

[131] Ian Jolliffe. "Principal component analysis". In: *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.

[132] Michael I Jordan and Tom M Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260.

[133] C. Josz et al. "Application of the moment-SOS approach to global optimization of the OPF problem". In: *IEEE Transactions on Power Systems* 30.1 (2015), pp. 463–470.

[134] Cedric Josz et al. "A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization". In: *Advances in neural information processing systems* (2018).

[135] Ri E Kalman. "Design of self-optimizing control system". In: *Trans. ASME* 80 (1958), pp. 468–478.

[136] V. Kekatos, G. B. Giannakis, and B. Wollenberg. "Optimal placement of phasor measurement units via convex relaxation". In: *IEEE Transactions on power systems* 27.3 (2012), pp. 1521–1530.

[137] M. Khanabadi, H. Ghasemi, and M. Doostizadeh. "Optimal transmission switching considering voltage security and N-1 contingency analysis". In: *IEEE Transactions on Power Systems* 28.1 (2013), pp. 542–550.

[138] Mohsen Kheirandishfard et al. "Convex Relaxation of Bilinear Matrix Inequalities Part II: Applications to Optimal Control Synthesis". In: *2018 IEEE Conference on Decision and Control (CDC)*. IEEE. 2018, pp. 75–82.

[139] Amin Khodaei and Mohammad Shahidehpour. "Transmission switching in security-constrained unit commitment". In: *IEEE Transactions on Power Systems* 25.4 (2010), pp. 1937–1945.

[140] M. Klein. "A primal method for minimal cost flows with applications to the assignment and transportation problems". In: *Management Science* 14 (1967), pp. 205–220.

[141] B. Kocuk, S. S. Dey, and X. A. Sun. "Strong SOCP relaxations for the optimal power flow problem". In: *Operations Research* 64.6 (2016), pp. 1177–1196.

[142] Burak Kocuk et al. "A cycle-based formulation and valid inequalities for DC power transmission problems with switching". In: *Operations Research* 64.4 (2016), pp. 922–938.

[143] M. Kraning et al. "Dynamic Network Energy Management via Proximal Message Passing". In: *Foundations and Trends in Optimization* 1.2 (2013), pp. 1–54.

[144] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[145] A. Y. S. Lam et al. "Optimal distributed voltage regulation in power distribution networks". In: *Submitted for publication* (2012).

[146] Andrew Lamperski and John C. Doyle. "Output Feedback $\mathcal{H}_2$ Model Matching for Decentralized Systems with Delays". In: *2013 IEEE American Control Conference (ACC)*. June 2013.

[147] Andrew Lamperski and Laurent Lessard. "Optimal Decentralized State-Feedback Control with Sparsity and Delays". In: *Automatica* 58 (2015), pp. 143–151.

[148] J. B. Lasserre. "Global optimization with polynomials and the problem of moments". In: *SIAM Journal on Optimization* 11.3 (2001), pp. 796–817.

[149] J. Lavaei. "Zero duality gap for classical OPF problem convexifies fundamental nonlinear power problems". In: *American Control Conference* (2011).

[150] J. Lavaei and S. H. Low. "Convexification of Optimal Power Flow Problem". In: *48th Annual Allerton Conference* (2010).

[151] J. Lavaei and S. H. Low. "Zero duality gap in optimal power flow problem". In: *IEEE Transactions on Power Systems* 27.1 (2012), pp. 92–107.

[152] J. Lavaei and S. Sojoudi. "Competitive Equilibria in Electricity Markets with Nonlinearities". In: *American Control Conference* (2012).

[153] J. Lavaei, D. Tse, and B. Zhang. "Geometry of power flows and optimization in distribution networks". In: *IEEE Transactions on Power Systems* 29.2 (2014), pp. 572–583.

[154] J. Lavaei, B. Zhang, and D. Tse. "Geometry of Power Flows in Tree Networks". In: *IEEE Power & Energy Society General Meeting* (2012).

[155] Laura Lazzeroni and Art Owen. "Plaid models for gene expression data". In: *Statistica sinica* (2002), pp. 61–86.

[156] Daniel D Lee and H Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (1999), p. 788.

[157] K. Lehmann, A. Grastien, and P. Van Hentenryck. "The complexity of DC-Switching problems". In: *arXiv preprint arXiv:1411.4369* (2014).

[158] B. Lesieutre et al. "Examining the Limits of the Application of Semidefinite Programming to Power Flow Problems". In: *49th Annual Allerton Conference on Communication, Control and Computing* (2011).

[159] L. Lessard, M. Kristalny, and A. Rantzer. "On structured realizability and stabilizability of linear systems". In: *American Control Conference (ACC), 2013*. June 2013, pp. 5784–5790.

[160] Laurent Lessard. "State-space solution to a minimum-entropy $\mathcal{H}_\infty$-optimal control problem with a nested information constraint". In: *2014 53rd IEEE Conference on Decision and Control (CDC)*. 2014. URL: `http://arxiv.org/pdf/1403.5020v2.pdf`.

[161] Laurent Lessard and Sanjay Lall. "Optimal Controller Synthesis for the Decentralized Two-Player Problem with Output Feedback". In: *2012 IEEE American Control Conference (ACC)*. June 2012.

[162] Sergey Levine et al. "End-to-end Training of Deep Visuomotor Policies". In: *Journal of Machine Learning Research* 17.1 (Jan. 2016), pp. 1334–1373.

[163] Xiao Li et al. "Nonconvex robust low-rank matrix recovery". In: *arXiv preprint arXiv:1809.09237* (2018).

[164] Xingpeng Li et al. "Real-Time Contingency Analysis With Corrective Transmission Switching". In: *IEEE Transactions on Power Systems* 32.4 (2017), pp. 2604–2617.

[165] Lennart Ljung. "System identification". In: *Wiley Encyclopedia of Electrical and Electronics Engineering* (1999), pp. 1–19.

[166] S. H. Low. "Convex relaxation of optimal power flow—Part I: Formulations and equivalence". In: *IEEE Transactions on Control of Network Systems* 1.1 (2014), pp. 15–27.

[167] S. H. Low. "Convex relaxation of optimal power flow—Part II: Exactness". In: *IEEE Transactions on Control of Network Systems* 1.2 (2014), pp. 177–189.

[168] Yao Ma et al. "Gradient Descent for Sparse Rank-One Matrix Completion for Crowd-Sourced Aggregation of Sparsely Interacting Workers". In: *International Conference on Machine Learning*. 2018, pp. 3341–3350.

[169] R. Madani, M. Ashraphijuo, and J. Lavaei. "Promises of Conic Relaxation for Contingency-Constrained Optimal Power Flow Problem". In: *IEEE Transactions on Power Systems* 31.2 (2016), pp. 1297–1307.

[170] R. Madani, J. Lavaei, and R. Baldick. "Convexification of Power Flow Equations for Power Systems in Presence of Noisy Measurements". In: *http://www.ieor.berkeley.edu/~lavaei/SE_J_2016.pdf* (2016).

[171] R. Madani et al. "Finding Low-rank Solutions of Sparse Linear Matrix Inequalities using Convex Optimization". In: *to appear in SIAM Journal on Optimization, available online at http://www.ieor.berkeley.edu/~lavaei/LMI_Low_Rank.pdf* (2017).

[172] Ramtin Madani, Alper Atamtürk, and Ali Davoudi. "A Scalable Semidefinite Relaxation Approach to Grid Scheduling". In: *arXiv preprint arXiv:1707.03541* (2017).

[173] Ramtin Madani et al. *Polynomial Optimization via Penalized Conic Relaxation*. 2018. URL: http://www.uta.edu/faculty/madanir/poly_conic.pdf.

[174] A. Mahajan et al. "Information structures in optimal decentralized control". In: *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*. 2012, pp. 1291–1306. DOI: 10.1109/CDC.2012.6425819.

[175] Yuri V. Makarov, Zhao Yang Dong, and David J. Hill. "On Convexity of Power Flow Feasibility Boundary". In: *IEEE Transactions on Power Systems* (2008).

[176] Horia Mania, Stephen Tu, and Benjamin Recht. "Certainty equivalent control of LQR is efficient". In: *arXiv preprint arXiv:1902.07826* (2019).

[177] Nikolai Matni. "Distributed Control Subject to Delays Satisfying an $\mathcal{H}_\infty$ Norm Bound". In: *2014 53rd IEEE Conference on Decision and Control (CDC)*. 2014. URL: http://arxiv.org/pdf/1402.1559.pdf.

[178] Nikolai Matni, Yuh-Shyang Wang, and James Anderson. "Scalable system level synthesis for virtually localizable systems". In: *IEEE Conference on Decision and Control*. 2017.

[179] Rahul Mazumder and Trevor Hastie. "Exact covariance thresholding into connected components for large-scale graphical lasso". In: *Journal of Machine Learning Research* 13 (2012), pp. 781–794.

[180] Garth P. McCormick. "Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems". In: *Mathematical Programming* 10.1 (1976), pp. 147–175.

[181] Nicolai Meinshausen and Peter Bühlmann. "High-dimensional graphs and variable selection with the lasso". In: *The annals of statistics* (2006), pp. 1436–1462.

[182] Peyman Milanfar. "A tour of modern image filtering: New insights and methods, both practical and theoretical". In: *IEEE Signal Processing Magazine* 30.1 (2013), pp. 106–128.

[183] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518 (Feb. 2015), pp. 529–533.

[184] Igor Molybog, Ramtin Madani, and Javad Lavaei. "Conic Optimization for Robust Quadratic Regression: Deterministic Bounds and Statistical Analysis". In: *IEEE 57th Conference on Decision and Control* (2018).

[185] D. K. Molzahn, B. C. Lesieutre, and C. L. DeMarco. "A sufficient condition for power flow insolvability with applications to voltage stability margins". In: *http://arxiv.org/pdf/1204.6285.pdf* (2012).

[186] D. Molzahn et al. "Solution of Optimal Power Flow Problems using Moment Relaxations Augmented with Objective Function Penalization". In: *IEEE Conference on Decision and Control (CDC)* (2015), pp. 31–38.

[187] J. A. Momoh, M. E. El-Hawary, and R. Adapa. "A review of selected optimal power flow literature to 1993. Part I: Nonlinear and quadratic programming approaches". In: *IEEE Transactions on Power Systems* (1999).

[188] J. A. Momoh, M. E. El-Hawary, and R. Adapa. "A review of selected optimal power flow literature to 1993. Part II: Newton, linear programming and interior point methods". In: *IEEE Transactions on Power Systems* (1999).

[189] Andrea Montanari and Emile Richard. "Non-negative principal component analysis: Message passing algorithms and sharp asymptotics". In: *IEEE Transactions on Information Theory* 62.3 (2016), pp. 1458–1484.

[190] Nader Motee and Ali Jadbabaie. "Optimal control of spatially distributed systems". In: *IEEE Transactions on Automatic Control* 53.7 (2008), pp. 1616–1629.

[191] Nader Motee and Qiyu Sun. "Sparsity measures for spatially decaying systems". In: *2014 American Control Conference*. IEEE. 2014, pp. 5459–5464.

[192] Shanmugavelayutham Muthukrishnan. "Data streams: Algorithms and applications". In: *Foundations and Trends® in Theoretical Computer Science* 1.2 (2005), pp. 117–236.

[193] Habibollah Nassiri and Rafegh Aghamohammadi. "A New Analytic Neuro-Fuzzy Model For Work Zone Capacity Estimation". In: *Transportation Research Board 96th Annual Meeting* 17.06061 (2017).

[194] A. Nayyar, A. Mahajan, and D. Teneketzis. "Decentralized Stochastic Control with Partial History Sharing: A Common Information Approach". In: *IEEE Transactions on Automatic Control* 58.7 (July 2013), pp. 1644–1658. ISSN: 0018-9286. DOI: 10.1109/TAC.2013.2239000.

[195] Eugene Ndiaye et al. "GAP safe screening rules for sparse multi-task and multi-class models". In: *Advances in Neural Information Processing Systems*. 2015, pp. 811–819.

[196] Sahand N Negahban et al. "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers". In: *Statistical Science* 27.4 (2012), pp. 538–557.

[197] K. E. Nygard, P. R. Chandler, and M. Pachter. "Dynamic Network Flow Optimization Models for Air Vehicle Resource Allocation". In: *American Control Conference* (2001).

[198] R. P. O'Neill et al. "Dispatchable transmission in RTO markets". In: *IEEE Transactions on Power Systems* 20.1 (2005), pp. 171–179.

[199] Matt Olfat and Anil Aswani. "Spectral Algorithms for Computing Fair Support Vector Machines". In: *International Conference on Artificial Intelligence and Statistics* (2018).

[200] OpenAI et al. "Learning Dexterous In-Hand Manipulation". In: *CoRR* abs/1808.00177 (2018). arXiv: 1808.00177. URL: http://arxiv.org/abs/1808.00177.

[201] James Ostrowski, Jianhui Wang, and Cong Liu. "Transmission switching with connectivity-ensuring constraints". In: *IEEE Transactions on Power Systems* 29.6 (2014), pp. 2621–2627.

[202] Y. Ouyang, M. Gagrani, and R. Jain. "Control of unknown linear systems with Thompson sampling". In: *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Oct. 2017, pp. 1198–1205. DOI: 10.1109/ALLERTON.2017.8262873.

[203] T. J. Overbye, Xu Cheng, and Yan Sun. "A Comparison of the AC and DC Power Flow Models for LMP Calculations". In: *Proceedings of the 37th Hawaii International Conference on System Sciences*. 2004.

[204] Samet Oymak and Necmiye Ozay. "Non-asymptotic identification of lti systems from a single trajectory". In: *arXiv preprint arXiv:1806.05722* (2018).

[205] Figen Oztoprak et al. "Newton-like methods for sparse inverse covariance estimation". In: *Advances in neural information processing systems*. 2012, pp. 755–763.

[206] K. S. Pandya and S. K. Joshi. "A survey of optimal power flow methods". In: *Journal of Theoretical and Applied Information Technology* (2008).

[207] José Pereira, Morteza Ibrahimi, and Andrea Montanari. "Learning networks of stochastic differential equations". In: *Advances in Neural Information Processing Systems*. 2010, pp. 172–180.

[208] William H Press et al. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[209] Xin Qi et al. "Structured optimal and robust control with multiple criteria: A convex solution". In: *Automatic Control, IEEE Transactions on* 49.10 (2004), pp. 1623–1640.

[210] Lishan Qiao, Songcan Chen, and Xiaoyang Tan. "Sparsity preserving projections with applications to face recognition". In: *Pattern Recognition* 43.1 (2010), pp. 331–341.

[211] Anders Rantzer. "Concentration bounds for single parameter adaptive control". In: *2018 Annual American Control Conference (ACC)*. IEEE. 2018, pp. 1862–1866.

[212] Benjamin Recht. "A tour of reinforcement learning: The view from continuous control". In: *Annual Review of Control, Robotics, and Autonomous Systems* 2 (2019), pp. 253–279.

[213] Bin Ren et al. "Non-negative Matrix Factorization: Robust Extraction of Extended Structures". In: *The Astrophysical Journal* 852.2 (2018), p. 104.

[214] Omar Rivasplata. "Subgaussian random variables: An expository note". In: *Internet publication, PDF* (2012).

[215] Benjamin Rolfs et al. "Iterative thresholding algorithm for sparse inverse covariance estimation". In: *Advances in Neural Information Processing Systems*. 2012, pp. 1574–1582.

[216] Jacqueline G. Rolim and Luiz Jairo B. Machado. "A study of the use of corrective switching in transmission systems". In: *IEEE Transactions on Power Systems* 14.1 (1999), pp. 336–341.

[217] Michael Rotkowitz and Sanjay Lall. "A characterization of convex problems in decentralized control". In: *IEEE Transactions on Automatic Control* 51.2 (2006), pp. 274–286.

[218] Mark Rudelson, Roman Vershynin, et al. "Hanson-Wright inequality and sub-gaussian concentration". In: *Electronic Communications in Probability* 18 (2013).

[219] Daniel J. Russo et al. "A Tutorial on Thompson Sampling". In: *Foundations and Trends on Machine Learning* 11.1 (July 2018), pp. 1–96. DOI: 10.1561/2200000070.

[220] AI Saltykov. "The number of components in a random bipartite graph". In: *Discrete Mathematics and Applications* 5.6 (1995), pp. 515–524.

[221] Tuhin Sarkar and Alexander Rakhlin. "How fast can linear dynamical systems be learned?" In: *arXiv preprint arXiv:1812.01251* (2018).

[222] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. "Finite-Time System Identification for Partially Observed LTI Systems of Unknown Order". In: *arXiv preprint arXiv:1902.01848* (2019).

[223] Carsten W. Scherer. "Structured $\mathcal{H}_\infty$-Optimal Control for Nested Interconnections: A State-Space Solution". In: *Systems and Control Letters* 62 (12 2013), pp. 1105–1113.

[224] Parikshit Shah and Pablo A Parrilo. "$\mathcal{H}_2$-optimal decentralized control over posets: A state space solution for state-feedback". In: *Decision and Control (CDC), 2010 49th IEEE Conference on*. 2010.

[225] Jun Shao and Dongsheng Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

[226] Wei Shao and Vijay Vittal. "BIP-based OPF for line and bus-bar switching to relieve overloads and voltage violations". In: *IEEE Power Systems Conference and Exposition* (2006), pp. 2090–2095.

[227] Ali Sharif Razavian et al. "CNN features off-the-shelf: an astounding baseline for recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 806–813.

[228] Jiaying Shi and Shmuel Oren. "Stochastic Unit Commitment with Topology Control Recourse for Power Systems with Large-Scale Renewable Integration". In: *IEEE Transactions on Power Systems* (2017).

[229] Jiaying Shi and Shmuel S Oren. "Wind power integration through stochastic unit commitment with topology control recourse". In: *Power Systems Computation Conference (PSCC)*. 2016, pp. 1–7.

[230] David Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144.

[231] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529 (Jan. 2016), pp. 484–489.

[232] Max Simchowitz, Ross Boczar, and Benjamin Recht. "Learning Linear Dynamical Systems with Semi-Parametric Least Squares". In: *arXiv preprint arXiv:1902.00768* (2019).

[233] Max Simchowitz et al. "Learning without mixing: Towards a sharp analysis of linear system identification". In: *arXiv preprint arXiv:1802.08334* (2018).

[234] Valeria Simoncini. "The Lyapunov matrix equation. Matrix analysis from a computational perspective". In: *arXiv preprint arXiv:1501.07564* (2015).

[235] S. Sojoudi and J. Lavaei. "Convexification of optimal power flow problem by means of phase shifters". In: *IEEE International Conference on Smart Grid Communications*. 2013, pp. 756–761.

[236] S. Sojoudi and J. Lavaei. "Physics of power networks makes hard optimization problems easy to solve". In: *IEEE Power & Energy Society General Meeting* (2012).

[237] S. Sojoudi and S. H. Low. "Optimal Charging of Plug-in Hybrid Electric Vehicles in Smart Grids". In: *IEEE Power & Energy Society General Meeting* (2011).

[238] Somayeh Sojoudi. "Equivalence of graphical lasso and thresholding for sparse graphs". In: *Journal of Machine Learning Research* 17.115 (2016), pp. 1–21.

[239] Somayeh Sojoudi and John Doyle. "Study of the brain functional network using synthetic data". In: *52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (2014), pp. 350–357.

[240] Somayeh Sojoudi and Javad Lavaei. "Exactness of semidefinite relaxations for nonlinear optimization problems with underlying graph structure". In: *SIAM Journal on Optimization* 24.4 (2014), pp. 1746–1778.

[241] Freek Stulp, Evangelos A Theodorou, and Stefan Schaal. "Reinforcement learning with sequences of motion primitives for robust manipulation". In: *IEEE Transactions on robotics* 28.6 (2012), pp. 1360–1370.

[242] Ju Sun, Qing Qu, and John Wright. "A geometric analysis of phase retrieval". In: *Foundations of Computational Mathematics* 18.5 (2018), pp. 1131–1198.

[243] Ju Sun, Qing Qu, and John Wright. "Complete dictionary recovery over the sphere I: Overview and the geometric picture". In: *IEEE Transactions on Information Theory* 63.2 (2017), pp. 853–884.

[244] Takashi Tanaka and Pablo A. Parrilo. "Optimal Output Feedback Architecture for Triangular LQG Problems". In: *2014 IEEE American Control Conference (ACC)*. June 2014.

[245] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[246] Robert Tibshirani et al. "Strong rules for discarding predictors in lasso-type problems". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.2 (2012), pp. 245–266.

[247] Kentaro Toyama et al. "Wallflower: Principles and practice of background maintenance". In: *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*. Vol. 1. IEEE. 1999, pp. 255–261.

[248] Eran Treister and Javier S Turek. "A block-coordinate descent approach for large-scale sparse inverse covariance estimation". In: *Advances in neural information processing systems*. 2014, pp. 927–935.

[249] Anastasios Tsiamis and George J Pappas. "Finite Sample Analysis of Stochastic System Identification". In: *arXiv preprint arXiv:1903.09122* (2019).

[250] John N. Tsitsiklis and Michael Athans. "On the complexity of decentralized decision making and detection problems". In: *IEEE Conference on Decision and Control (CDC)*. 1984.

[251] John Tsitsiklis and Michael Athans. "On the complexity of decentralized decision making and detection problems". In: *IEEE Transactions on Automatic Control* 30.5 (1985), pp. 440–446.

[252] Lieven Vandenberghe and Martin S. Andersen. "Chordal graphs and semidefinite optimization". In: *Foundations and Trends® in Optimization* 1.4 (2015), pp. 241–433.

[253] Stephen A. Vavasis. "Complexity theory: quadratic programming". In: *Encyclopedia of Optimization*. Ed. by Christodoulos A. Floudas and Panos M. Pardalos. Boston, MA: Springer US, 2001, pp. 304–307. ISBN: 978-0-306-48332-5. DOI: 10.1007/0-306-48332-7_65. URL: https://doi.org/10.1007/0-306-48332-7_65.

[254] Petra E. Vértes et al. "Simple models of human brain functional networks". In: *Proceedings of the National Academy of Sciences* 109.15 (2012), pp. 5868–5873.

[255] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[256] Martin J Wainwright. "Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso)". In: *IEEE transactions on information theory* 55.5 (2009), pp. 2183–2202.

[257] Yuh-Shyang Wang, Nikolai Matni, and John C Doyle. "A system level approach to controller synthesis". In: *arXiv preprint arXiv:1610.04815* (2016).

[258] Yuh-Shyang Wang, Nikolai Matni, and John C Doyle. "Localized LQR optimal control". In: *arXiv preprint arXiv:1409.6404* (2014).

[259] Yuh-Shyang Wang, Nikolai Matni, and John C Doyle. "Separable and localized system level synthesis for large-scale systems". In: *arXiv preprint arXiv:1701.05880* (2017).

[260] Yuh-Shyang Wang, Nikolai Matni, and John C Doyle. "Separable and Localized System-Level Synthesis for Large-Scale Systems". In: *IEEE Transactions on Automatic Control* 63.12 (2018), pp. 4234–4249.

[261] William J Welch. "Algorithmic complexity: three NP-hard problems in computational statistics". In: *Journal of Statistical Computation and Simulation* 15.1 (1982), pp. 17–25.

[262] Y. Weng et al. "Semidefinite programming for power system state estimation". In: *IEEE Power & Energy Society General Meeting* (2012).

[263] H. S. Witsenhausen. "A counterexample in stochastic optimum control". In: *SIAM Journal of Control* 6.1 (1968).

[264] Daniela M. Witten, Jerome H. Friedman, and Noah Simon. "New insights and faster computations for the graphical lasso". In: *Journal of Computational and Graphical Statistics* 20.4 (2011), pp. 892–900.

[265] Allen J. Wood and Bruce F. Wollenberg. *Power generation, operation, and control*. John Wiley & Sons, 2012.

[266] John Wright et al. "Robust face recognition via sparse representation". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31.2 (2009), pp. 210–227.

[267] John Wright et al. "Sparse representation for computer vision and pattern recognition". In: *Proceedings of the IEEE* 98.6 (2010), pp. 1031–1044.

[268] Margaret H Wright. "Ill-conditioning and computational error in interior methods for nonlinear programming". In: *SIAM Journal on Optimization* 9.1 (1998), pp. 84–111.

[269] Xindong Wu et al. "Data mining with big data". In: *IEEE Transactions on Knowledge and Data Engineering* 26.1 (2014), pp. 97–107.

[270] Y. Nesterov, A. S. Nemirovskii, and Y. Ye. "Interior-point polynomial algorithms in convex programming". In: *SIAM* 13 (1994).

[271] Eunho Yang, Aurélie C Lozano, and Pradeep K Ravikumar. "Elementary estimators for graphical models". In: *Advances in neural information processing systems*. 2014, pp. 2159–2167.

[272] Xinyang Yi et al. "Fast algorithms for robust PCA via gradient descent". In: *Advances in neural information processing systems*. 2016, pp. 4152–4160.

[273] Hongbin Yin et al. "Urban traffic flow prediction using a fuzzy-neural approach". In: *Transportation Research Part C: Emerging Technologies* 10.2 (2017), pp. 85–98.

[274] Ming Yuan and Yi Lin. "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94.1 (2007), pp. 19–35.

[275] Sangwoon Yun and Kim-Chuan Toh. "A coordinate gradient descent method for L_1-regularized convex minimization". In: *Computational Optimization and Applications* 48.2 (2011), pp. 273–307.

[276] B. Zhang and D. Tse. "Geometry of injection regions of power networks". In: *IEEE Transactions on Power Systems* 28.2 (2013), pp. 788–797.

[277] Richard Y. Zhang and Javad Lavaei. "Modified Interior-Point Method for Large-and-Sparse Low-Rank Semidefinite Programs". In: *56th IEEE Conference on Decision and Control* (2017).

[278] Richard Y Zhang et al. "How Much Restricted Isometry is Needed In Nonconvex Matrix Recovery?" In: *Advances in neural information processing systems* (2018).

[279] Richard Zhang, Salar Fattahi, and Somayeh Sojoudi. "Large-scale sparse inverse covariance estimation via thresholding and Max-Det matrix completion". In: *International Conference on Machine Learning*. 2018, pp. 5761–5770.

[280] Xiao Zhang et al. "A primal-dual analysis of global optimality in nonconvex low-rank matrix recovery". In: *International conference on machine learning*. 2018, pp. 5857–5866.

[281] Peng Zhao and Bin Yu. "On model selection consistency of Lasso". In: *Journal of Machine learning research* 7.Nov (2006), pp. 2541–2563.

[282] Qinqing Zheng and John Lafferty. "Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent". In: *arXiv preprint arXiv:1605.07051* (2016).

[283] Zihan Zhou et al. "Stable principal component pursuit". In: *2010 IEEE international symposium on information theory*. IEEE. 2010, pp. 1518–1522.

[284] Zhihui Zhu et al. "Global optimality in low-rank matrix optimization". In: *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2017, pp. 1275–1279.

[285] Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, and Robert John Thomas. "MATPOWER: Steady-state operations, planning and analysis tools for power systems research and education". In: *IEEE Transactions on power systems* 26.1 (2011), pp. 12–19.

[286] Hui Zou, Trevor Hastie, and Robert Tibshirani. "Sparse principal component analysis". In: *Journal of computational and graphical statistics* 15.2 (2006), pp. 265–286.