Title

Provably Efficient Algorithms for Non-convex Optimization with Benign Structures

Author

Zhang, Haixiang

Publication Date

2024

Peer reviewed|Thesis/dissertation

Provably Efficient Algorithms for Non-convex Optimization with Benign Structures

by

Haixiang Zhang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Applied Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Javad Lavaei, Co-chair
Professor Jon Wilkening, Co-chair
Professor Zeyu Zheng

Spring 2024

Provably Efficient Algorithms for Non-convex Optimization with Benign Structures

Abstract

Provably Efficient Algorithms for Non-convex Optimization with Benign Structures

by

Haixiang Zhang

Doctor of Philosophy in Applied Mathematics

University of California, Berkeley

Professor Javad Lavaei, Co-chair

Professor Jon Wilkening, Co-chair

This dissertation focuses on devising innovative approaches to understanding and computing the optimal decision within complex large-scale operational frameworks, such as service networks and power systems. These approaches aim to facilitate interpretable, scalable, and robust decision-making under uncertainty. However, achieving this objective is hindered by the computational hurdle presented by the fact that most practical problems are known to be $\mathcal{NP}$-hard in the worst case. Nevertheless, this dissertation has been primarily inspired by the observation that real-world problem instances often exhibit benign geometric properties, which can be exploited to reduce the computational complexity. By leveraging methodologies from mathematics, statistics, operations research, and machine learning, our goal is to identify and harness these underlying geometric properties to devise algorithms that are demonstrably efficient and resilient, thereby supporting decision-making in large-scale systems.

In the first part of the dissertation, we focus on the low-rank matrix optimization problem, which targets at recovering the underlying low-rank ground truth matrix from a small number of measurements. Various important applications in the fields of machine learning, signal processing, and power systems can be formulated as a low-rank matrix optimization problem. By utilizing the benign optimization landscape around the manifold of low-rank matrices, we establish state-of-the-art theoretical guarantees to the highly efficient non-convex formulation based on the Burer-Monteiro factorization. First, we significantly improve existing sufficient conditions in terms of the Restricted Isometry Property constant, leading to the guaranteed success of local search algorithms for more problem instances. Next, we propose a new complexity metric for the rank-1 generalized matrix completion problem. The new metric has the potential of unifying several existing metrics and provides both sufficient conditions and necessary conditions to the success of local search algorithms. This part serves as

a crucial step towards closing the gap between the empirical success and the theoretical understanding of the Burer-Monteiro factorization approach.

The second part of the dissertation is concerned with the discrete optimization via simulation problem. The design of scalable and robust simulation-optimization algorithms plays a vital role in the timely decision-making in large-scale systems with uncertainty, such as the bike-sharing system. Inspired by the marginal decreasing property, we utilize a special structure, named the $L^\natural$-convexity, to develop algorithms with scalability and optimality guarantees. More specifically, we first construct a subgradient estimator based on the Lovász extension and develop stochastic search algorithms using the subgradient information. We theoretically and empirically illustrate that the proposed algorithms are highly efficient for high-dimensional discrete optimization via simulation problems. Next, by combining the subgradient information with the discrete nature of the problem, we propose the stochastic localization algorithms, which exhibit an improved efficiency on large-scale applications.

In the third part of the dissertation, we focus on the AC power flow problem in power systems. The efficient and reliable control of large-scale power systems, e.g., the dispatch of electricity under safety-critical constraints, is contingent upon developments of advanced computational and analytical tools. In the first chapter, we consider the uniqueness of solutions to the power flow problem. Utilizing properties of the monotone regime and the network topology, stronger necessary conditions and sufficient conditions, based on the maximal girth and the maximal eye, are proposed to guarantee the uniqueness of solutions. We also provide the corresponding reduction algorithms to efficiently estimate these two network parameters. Given the ongoing emergence of intermittent renewable generation, the second chapter is devoted to the design of optimal power flow algorithms that are robust to large generation forecast errors, which plays an essential role in incorporating renewable energy generators into electrical networks. More concretely, we propose a novel distributionally robust optimization formulation for the chance-constrained optimal power flow problem and provide corresponding algorithms to effectively find the robust solution with the minimum generation cost.

To my mom and dad, for their unconditional love and support

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed or extended their valuable assistance throughout my journey in past five years. It is a pleasure to thank those who made the preparation and the completion of the study a possibility.

First, I would like to express my deepest appreciation to my advisor, Prof. Javad Lavaei, whose guidance, support and expertise have enabled me to complete my Ph.D. program. Javad continuously provided encouragement and was always willing and enthusiastic to assist in any way he could throughout the research project. I am extremely grateful to him for his invaluable support and infinite patience, and for providing many research opportunities, which helped me become an independent researcher.

I would also like to express my deepest appreciation to my dissertation committee member, mentor, and collaborator, Prof. Zeyu Zheng. This endeavor would not have been possible without the endless help and support from him. I cannot overstate value of the chance to collaborate with Zeyu, which introduced me to the field of simulation optimization and served as a new dimension in my research profile. I am also grateful for his editing help, late-night discussions, and moral support to both my academic and personal life.

I consider myself extremely fortunate to have had the opportunity to collaborate with two exceptional faculty members, Prof. Somayeh Sojoudi and Prof. Ross Baldick. Without their dedication, this thesis would not have been possible. Somayeh's insightful perspectives and constructive criticism have greatly enriched our collaboration projects. Her profound expertise in this field has not only made this journey inspiring but has also sparked numerous new research ideas. I am also thankful for the invaluable knowledge and tools I learned from her "Convex Optimization" course. Many thanks to Ross for initializing my research on power systems. His invaluable expertise and guidance have been instrumental in navigating the challenges encountered along the way. I am particularly appreciative of his support, which has been integral to the completion of this dissertation. Thanks should also go to my dissertation committee co-chair, Prof. Jon Wilkening, whose meticulous proofreading greatly improved my thesis.

The conclusions of this thesis would have not been as compelling without the people I have collaborated with over the past five years. In particular, I would like to thank Yingjie Bi and SangWoo Park. Their unwavering support and patience during the early stages of my Ph.D. journey have been indispensable. Their generous sharing of knowledge and expertise has guided and nurtured me, shaping me into a better researcher. I am deeply indebted to them for their guidance and encouragement. My gratitude extends to my other collaborators: Han Feng, Julie Mulvaney-Kemp, Baturalp Yalcin, Eli Brock, Ying Chen.

Through this acknowledgment, I express my heartfelt gratitude to my friends who have been my pillars of strength. Their presence has made my experience at Berkeley here truly enjoyable. A huge thank you is extended toward my group mates, Igor Molybog, Yuhao Ding, Donghao Ying, Jihun Kim, and Hyunin Lee, for the opportunity to learn from their research endeavors. I consider myself incredibly lucky to have such a friendly group of Applied Math

cohort students: Jiasu Wang, Yixiang Luo, and Raehyun Kim. I also wish to express my gratitude to the Department of Math and its staff for their considerate guidance. A special shoutout goes to Vicky Lee, for her generous assistance in starting my Ph.D. journey in the U.S. as an international student. I am truly blessed to have such amazing friends, and I extend my sincere thanks to all those who have provided invaluable input: Hansheng Jiang, Jiaming Wang, Yue Zhang, Jiahao Yao, Haotian Gu, Edric Wang, Yulong Dong, Ziye Ma, Mo Liu, Jingxu Xu, Haoting Zhang, Yiyuan Xiong, and Haitian Liu.

Without the ongoing support of a few special people in my life, completing a Ph.D. would have seemed unattainable. Words cannot express my gratitude to my partner, Ke Ding. I thank her for the indispensable patience and understanding she has shown me throughout the study. The most significant acknowledgement is reserved for my parents, Jian Zhang and Liufeng Shou. It is the love and support which they have provided to me that has enabled me to achieve everything that I have. It is with a feeling of immense accomplishment that I dedicate this thesis to my mom and dad.

Chapter 1

Introduction

The design of highly scalable algorithms with theoretical guarantees has been a long-standing challenge in a wide range of applications, including but not limited to artificial intelligence, power systems and service networks. After entering the era of big data, the size of datasets and the scale of systems are growing at an unprecedented rate. As a result, the development of computationally efficient algorithms has become a more essential and indispensable part to support the optimization and decision-making in large-scale systems.

For example, the recent years have witnessed the empirical success of deep neural networks in various fields. The deep neural network models can contain up to hundreds of millions of parameters and the training of such models requires solving non-convex optimization problems with the same number of variables, which is $\mathcal{NP}$-hard in general. By reducing the computational complexity with more efficient algorithms, it is possible to save a considerably large amount of the computational budget. In practice, it has been observed that simple local search algorithms, such as variants of the stochastic gradient descent algorithm, are able to find a solution with good generalization ability. Despite recent progress in the theoretical explanation of this phenomenon, there is still a huge gap between the theory and the practice. A few recent works hint that the landscape of the loss function of deep neural networks is high structured. As an important step towards closing the gap between theory and practice, it remains an open question whether we can identify the hidden structure to (i) theoretically explain the empirical success of local search algorithms and (ii) design faster optimization algorithms with guarantees to further reduce the computational complexity.

As another example, the efficient and resilient control of energy systems is contingent upon the development of fast algorithms that are (i) robust to adversarial noise in the data and (ii) guaranteed to find optimal or near-optimal decisions in real time for systems with millions of nodes. The power state estimation problem is one of the most important data-analysis problems for power systems and it is solved every 5-15 minutes in practice using heuristic methods. The lack of theoretical guarantees of these methods have caused catastrophic results in recent years, including several major blackouts in the US and Canada. In addition, with more advanced algorithms, higher-quality estimations are available and

better decisions can be made to improve the resiliency and efficiency of the system, which is critical for the sustainability of energy systems. This calls for scalable and reliable algorithms that can process massive amount of data in real time and resistant to the uncertainty from both measurement errors and adversarial attacks.

In this chapter, we first formulate the problems studied in the dissertation and the challenges that we need to overcome when solving the problem. Then, we provide an overview of our contributions in the dissertation and related publications associated with each chapter. Finally, we conclude the introduction chapter with the definition of notations used throughout the dissertation.

1.1 Optimization Algorithms with Theoretical Guarantees

This dissertation is primarily focused on addressing optimization problems formulated in the following manner:

$$\min_{x\in\mathbb{R}^d} f(x;\Theta) \quad \text{s.t. } x \in \mathcal{X}(\Theta), \tag{1.1}$$

where:

- The first variable $x \in \mathbb{R}^d$ is the targeted multivariate *optimization variable*. For example, if we target at optimally allocating staff across a service network, the optimization variable can denote the number of staff assigned to each service station. In the control of power systems, the variable may correspond to the bus voltage phasor vector. We note that in modern optimization applications, the number of optimization variables, or the problem dimension, d can be on the order of several millions.

- The second variable Θ summarizes the exogenous elements in problem (1.1), which directly or indirectly capture the problem parameters. In different problem settings, the variable Θ may take different forms. For instance, in the context of signal processing problems, it may include the measurement schemes and measurement results. When problem (1.1) involves certain randomness, the variable can correspond to the empirical distribution of a set of randomly generated samples or an unknown distribution that can generate independent samples of the distribution in an online setting. In general, we are not able to control this part of the problem in optimization algorithms.

- Function $f(x;\Theta)$ is the *objective function* that evaluates the quality of the optimization variable x. For example, it can be chosen as the generation cost of a power system. In a service system, the objective function may correspond to a utility function (e.g., the average waiting time of customers) that reflects the service quality and efficiency of the system under current decision x. In large-scale optimization problems, it is usually computationally heavy to evaluate the value and the derivatives of the objective

function, which calls for optimization algorithms that use as few as possible function evaluations to find the optimizer.

- Set $\mathcal{X}(\Theta) \subset \mathbb{R}^d$ is the *feasible set* of the optimization problem, namely, the set of all acceptable values of the optimization variable x. As an example, the feasible set should include the physical laws and safety-critical constraints in the operation of power systems. In addition, the feasible set can reflect the user's prior information or preference on the optimal decision, such as low rankness, sparsity, or robustness. Mathematically, the feasible set can consist of a set of equalities and inequalities that we can directly evaluate; or it may involve implicit constraints that we can only estimate from known information, such as the chance constraints in robust optimization.

In the fields of machine learning, operations research, signal processing, and power systems, a wide range of real-world problems can be cast into the form of problem (1.1). Optimization plays a major role in providing the computational tool for these applications. The ultimate goal of optimization algorithms is to find a global optimum to problem (1.1), in the sense that it achieves the universally minimal objective function value over all feasible solutions. Besides the optimality guarantee, there exist various other factors that practitioners may seek in algorithm design.

However, the goal of developing efficient algorithms with theoretical guarantees is hindered by the computational challenge that most practical problems are known to be $\mathcal{NP}$-hard in the worst case. More specifically, modern statistical and data analysis problems have posted new challenges for optimization, including but not limited to *massive-scale optimization problems, non-convexity, tight time budget, and adversarial noises*. In the following, we briefly discuss three important types of theoretical guarantees for optimization algorithms:

- *Convergence guarantees.* In most applications, it is important to theoretically guarantee the quality or the performance of the solution returned by optimization algorithms. For example, a low-quality solution may lead to an incorrect state estimation of power systems and result in dangerous control operations of the system. In the context of problem (1.1), it is necessary to design algorithms that provably converge to a global optimum or a sub-optimal solution with nearly optimal performance. In contrast to convex optimization problems, the optimization problem embedded in modern data analysis applications is often inherently non-convex. Conventionally, it is believed that the non-convexity of problem (1.1) fails the convergence of local search methods due to the existence of spurious local minima[1]. However, recent years have witnessed the empirical success of heuristic algorithms on a large number of non-convex optimization problems, e.g., low-rank matrix recovery, phase retrieval, deep learning, etc. Theoretical efforts, including this dissertation, have been made to establish convergence guarantees for these algorithms and shed light on future algorithm design.

[1] A spurious local minimum is a local minimum of problem (1.1) that is not globally optimal.

- *Efficiency guarantees.* To support real-time decision-making, it is important that the algorithms deliver solutions in a timely manner. For instance, in a service system, the operator may want to re-balance the resource allocation among service stations every a few hours. In this application, a high-quality decision needs to be made under a tight time budget. For an optimization algorithm, its computation efficiency can be characterized by the convergence rate and the per-iteration computation cost. In particular, we are interested in reducing the dependence of computation cost on the problem scale d and develop algorithms that are able to solve massive-scale optimization problems with more than millions of optimization variables.

  Besides the computation efficiency, the sample efficiency or the sample complexity is an equally important factor for optimization algorithms. In generally, the improvement on sample efficiency can lead to a better applicability of algorithms, especially for applications where collecting more samples is expensive if not impossible. For example, for the state estimation problem in power systems, the amount of available samples is closely related to electrical network infrastructures and collecting more samples means constructing more infrastructures, which may take years of planning and construction to complete. More specifically, this dissertation takes two approaches to improve the guarantees on sample efficiency: (i) we improve the minimum number of samples required to guarantee the success of local search algorithms, and (ii) we develop data-driven algorithms to utilize samples more efficiently.

- *Robustness guarantees.* In real-world dataset, there exist two common types of uncertainties: noises and adversarial attacks. The noises are relatively small-scale, universal and homoscedastic. For example, the MRI scans often involve small measurement errors; the customer arrival times in a service system may follow certain distribution. In contrast, adversarial attacks are relatively large-scale but sparse outliers in the dataset. For example, the measurement devices in power grids may suffer from natural disasters or cyber-attacks, which result in large deviation from the correct measurement values; the deep learning-backed autonomous driving algorithms can be misguided by adversarial attacks on traffic signs. Without robustness guarantees, algorithms are vulnerable to both types of uncertainties and may result in the violation of safety-critical constraints. In this dissertation, we design optimization algorithms that provably satisfy the quality and safety requirements with high probability under uncertainty of the dataset.

Throughout this dissertation, we illustrate that real-world problem instances usually satisfy certain benign geometric properties that can be utilized to develop algorithms with strong theoretical guarantees.

1.2 Summary of Contributions

Utilizing advanced tools from mathematics, statistics, operations research and machine learning, we aim at *identifying and utilizing the underlying geometric properties to design provably efficient algorithms that support the efficient and resilient decision-making in large-scale systems*. On a high level, the results in this dissertation consist of two major parts, i.e., the theoretical part and the algorithmic part:

1. *Theoretical work:* utilize statistical and geometric analysis techniques to establish solid theoretical guarantees that not only explain the empirical success of simple gradient-based optimization algorithms on non-convex optimization problems, but also identify the situations when those algorithms may fail to find the optimal solution.

2. *Algorithm design:* after formulating a practical problem as optimization and control problems, identify and utilize the special geometric structure of the non-convex optimization landscape to design scalable and robust algorithms with convergence guarantees.

The two parts are tightly connected with each other. The design of efficient and resilient optimization algorithms is critical to the real-time control and operation of large-scale systems, especially when the stability and robustness of the solution against uncertainty is necessary. For example, Chapters 6 and 7 analyze the uniqueness theory and the distributionally robust optimization of the AC power flow problem, respectively. The results provide important insights into the operation and planning of large-scale power systems under safety constraints. In the following, we discuss the contributions of each part of the dissertation.

Low-rank Matrix Optimization

In the low-rank matrix optimization, our goal is to recover an unknown low-rank matrix through a few measurements of its entries. This problem plays a central role in a wide range of applications, including but not limited to machine learning, signal processing, power systems and operations research. In practice, the optimization problem needs to be solved periodically and thus, it is necessary to design an efficient and robust algorithm. For the past few decades, the problem has attracted the attention of researchers from different fields and important progress has been made from both the theoretical and the experimental perspectives. Some of the early work focused on the semi-definite relaxation of the original non-convex problem and established exactness relaxation guarantees under various settings. However, despite the strong theoretical guarantees, the semi-definite relaxation approach is computationally challenging since solving the semi-definite program in the lifted space requires prohibitively huge computational efforts, especially when the problem size has surged in the big data era.

To deal with this challenge, Burer and Monteiro proposed a new factorization approach, which is more efficient than the semi-definite relaxation approach in terms of both the computation and memory complexity. Although the factorization approach requires solving a

non-convex optimization problem, it is shown that simple optimization algorithms, such as gradient descent and alternate minimization, exhibit fast and robust convergence to the ground truth solution. The goal of the first part of the dissertation is to utilize techniques from geometric analysis and statistics to provide theoretical characterizations on the optimization complexity of the non-convex optimization problem involved in the factorization approach. Namely, we estimate the chance that those algorithms can successfully find the ground truth solution from a random initial point.

Chapter 2. This paper considers the global geometry of general low-rank minimization problems via the Burer-Monteiro factorization approach. For the rank-1 case, we prove that there is no spurious second-order critical point for both symmetric and asymmetric problems if the rank-2 RIP constant δ is less than $1/2$. Combining with a counterexample with $\delta = 1/2$, we show that the derived bound is the sharpest possible. For the arbitrary rank-r case, the same property is established when the rank-$2r$ RIP constant δ is at most $1/3$. We design a counterexample to show that the non-existence of spurious second-order critical points may not hold if δ is at least $1/2$. In addition, for any problem with δ between $1/3$ and $1/2$, we prove that all second-order critical points have a positive correlation to the ground truth. Finally, the strict saddle property, which can lead to the polynomial-time global convergence of various algorithms, is established for both the symmetric and asymmetric problems when the rank-$2r$ RIP constant δ is less than $1/3$. The results of this paper significantly extend several existing bounds in the literature.

Chapter 3. In this chapter, we develop a new complexity metric for an important class of low-rank matrix optimization problems in both symmetric and asymmetric cases, where the metric aims to quantify the complexity of the non-convex optimization landscape of each problem and the success of local search methods in solving the problem. The existing literature has focused on two recovery guarantees. The RIP constant is commonly used to characterize the complexity of matrix sensing problems. On the other hand, the incoherence and the sampling rate are used when analyzing matrix completion problems. The proposed complexity metric has the potential to generalize these two notions and also applies to a much larger class of problems. To mathematically study the properties of this metric, we focus on the rank-1 generalized matrix completion problem and illustrate the usefulness of the new complexity metric on three types of instances, namely, instances with the RIP condition, instances obeying the Bernoulli sampling model, and a synthetic example. We show that the complexity metric exhibits a consistent behavior in the three cases, even when other existing conditions fail to provide theoretical guarantees. These observations provide a strong implication that the new complexity metric has the potential to generalize various conditions of optimization complexity proposed for different applications. Furthermore, we establish theoretical results to provide sufficient conditions and necessary conditions for the existence of spurious solutions in terms of the proposed complexity metric. This contrasts with the RIP and incoherence conditions that fail to provide any necessary condition.

Convex Discrete Optimization via Simulation

The second part of the dissertation focuses on the *discrete optimization via simulation* (DOvS) problem. The DOvS problem targets at finding an approximately optimal decision from a discrete feasible set via noisy evaluations to the objective function, which are generated by the simulations of stochastic system. A large number of important problems in the field of operations research, management science and economics can be formulated as the DOvS problem. Designing efficient DOvS algorithms is a vital part of time-sensitive decision-making in a large-scale system with uncertainty. Therefore, the DOvS problem has been a popular research topic over the past few decades and with very subtle differences, the DOvS problem is also known as the zeroth-order stochastic optimization in optimization literature and (contextual) multi-armed bandits in theoretical computer science literature. The challenges of solving the DOvS problem come from the following two parts:

1. First, the simulation of the system is usually time-consuming and cannot be accelerated by parallel computing. For example, the simulation of a queueing system involves generating the paths of time series, which can only been computed sequentially. With limited time and computation budgets, the DOvS algorithms aim at finding approximate solutions up to a specified precision using fewest possible number of simulations.

2. Additionally, the set of feasible decisions can be high-dimensional and large-scale. It requires a prohibitively large amount of computation resources to evaluate each feasible decision. As a result, without further assumptions on the problem, it is not possible to find a global solution.

In this dissertation, we focus on the second challenge and utilized the special structure of the objective function to avoid simulating each feasible decision. More specifically, we designed algorithms to screen out sub-optimal decisions and largely reduce the number of required simulations. The proposed algorithms outperform the state-of-the-art algorithms that ignore the special structure by a large margin. The main thrust of this part is the observation that in continuous optimization, the underlying structure of the problem (e.g., the convexity, the strict saddle property [244]) is able to reduce the optimization complexity from $\mathcal{NP}$-hard (general non-convex optimization) to polynomial-time solvable (e.g., convex optimization). Chapters 4 and 5 provide the first attempt in the area of discrete optimization to identify and utilize the hidden structure of the objective function, except the simple linear structure studied in mixed-integer programming literature.

In Chapters 4 and 5, we choose to analyze a specific structure called the $L^\natural$-*convexity* [168], which serves as one of the discrete counterparts of the classical convexity in continuous optimization. The choice of considering the $L^\natural$-convexity structure has several advantages. On the one hand, the $L^\natural$-convexity is satisfied by a wide range of applications, especially those in queueing networks and economics. In the one-dimensional case, the $L^\natural$-convexity reduces to the marginal decreasing property, which is a very common property of service networks. Moreover, the well-studied submodular function is a special case of the $L^\natural$-convex

function. On the other hand, the $L^\natural$-convexity is a sufficiently strong condition in the sense that utilizing the structure is able to reduce the optimization complexity from exponential to polynomial in terms of the problem dimension and scale. Therefore, our contributions to the convex DOvS problem (i.e., the DOvS problem with $L^\natural$-convexity) can significantly improve the efficiency of a variety of practical decision-making problems, in ways that both improve the quality of the solution and reduce the computation cost. We have designed two classes of algorithms for the convex DOvS problem, namely, the stochastic search algorithms (Chapter 4) and the stochastic localization algorithms (Chapter 5).

Chapter 4. We propose new sequential simulation-optimization algorithms for general convex optimization via simulation problems with high-dimensional discrete decision space. The performance of each choice of discrete decision variables is evaluated via stochastic simulation replications. If an upper bound on the overall level of uncertainties is known, our proposed simulation-optimization algorithms utilize the discrete convex structure and are guaranteed with high probability to find a solution that is close to the best within any given user-specified precision level. The proposed algorithms work for any general convex problem and the efficiency is demonstrated by proven upper bounds on simulation costs. The upper bounds demonstrate a polynomial dependence on the dimension and scale of the decision space. For some DOvS problems, a gradient estimator may be available at low costs along with a single simulation replication. By integrating gradient estimators, which are possibly biased, we propose simulation-optimization algorithms to achieve optimality guarantees with a reduced dependence on the dimension under moderate assumptions on the bias.

Chapter 5. We develop and analyze a set of new sequential simulation-optimization algorithms for large-scale multi-dimensional DOvS problems with a convexity structure. The "large-scale" notion refers to that the discrete decision variable has a large number of values to choose from on each dimension of the decision variable. The proposed algorithms are targeted to identify a solution that is close to the optimal solution given any precision level with any given probability. To achieve this target, utilizing the convexity structure, our algorithm design does not need to scan all the choices of the decision variable, but instead sequentially draws a subset of choices of the decision variable and uses them to "localize" potentially near-optimal solutions to an adaptively shrinking region. To show the power of the proposed methods based on the localization idea, we first consider one-dimensional large-scale problems. We develop the shrinking uniform sampling algorithm, which is proved to achieve the target with an optimal expected simulation cost under an asymptotic criterion. For multi-dimensional problems, we combine the idea of localization with subgradient information and propose a framework to design stochastic cutting-plane methods, whose expected simulation costs have a low dependence on the scale and the dimension of the problems. In addition, utilizing the discrete nature of the problems, we propose a stochastic dimension reduction algorithm, which does not require prior information about the Lipschitz constant of the objective function and its simulation costs are upper bounded by a value that is independent of

the Lipschitz constant. We implement the proposed algorithms on synthetic problems and queueing simulation optimization problems, and demonstrate better performances compared to benchmark methods especially for large-scale examples.

Power Systems

The third part of the dissertation focuses on the *AC power flow problem* in power systems. The power flow problem plays a crucial role in various aspects of power systems, e.g., the daily operations in contingency analysis and security-constrained dispatch of electricity markets. Designing scalable and reliable algorithms for the power flow problem is critical to the efficient and resilient operation of large-scale power systems.

Chapter 6. This chapter extends the uniqueness theory in [184] and establishes general necessary and sufficient conditions for the uniqueness of P-Θ power flow solutions in an AC power system using some properties of the monotone regime and the power network topology. We show that the necessary and sufficient conditions can lead to tighter sufficient conditions for the uniqueness in several special cases. Our results are based on the existing notion of maximal girth and our new notion of maximal eye. Moreover, we develop a series-parallel reduction method and search-based algorithms for computing the maximal eye and maximal girth, which are necessary for the uniqueness analysis. Reduction to a single line using the proposed reduction method is guaranteed for 2-vertex-connected Series-Parallel graphs. The relations between the parameters of the network before and after reduction are obtained. It is verified on real-world networks that the computation of the maximal eye can be reduced to the analysis of a much smaller power network, while the maximal girth is computed during the reduction process.

Chapter 7. Designing scalable and robust algorithms for the optimal power flow (OPF) problem is critical for the control of large-scale power systems under uncertainty. The chance-constrained AC OPF (CCOPF) problem provides a natural formulation of the trade-off between the operation cost and the constraint satisfaction rate. In Chapter 7, we propose a new data-driven algorithm for the CCOPF problem, which is based on the distributionally robust optimization (DRO). The proposed DRO approach achieves the *optimal efficiency* in the sense that (i) it finds the minimum-cost solution given the maximum rate of violating the constraints, and (ii) it uses an exact mixed-integer reformulation of chance constraints instead of inner approximations as used in existing literature. We apply the proposed algorithm to the semi-definite relaxation of the CCOPF problem and illustrate the advantage of our approach on IEEE benchmark power systems.

Related Publications

- **Chapter 2.**
  Main paper:

1. Haixiang Zhang, Yingjie Bi, and Javad Lavaei, "General Low-rank Matrix Optimization: Geometric Analysis and Sharper Bounds", *Neural Information Processing Systems*, 2021.

Related papers:

2. Yingjie Bi, Haixiang Zhang, and Javad Lavaei, "Local and Global Linear Convergence of General Low-rank Matrix Recovery Problems", *The AAAI Conference on Artificial Intelligence*, 2022.
3. Haixiang Zhang, Ying Chen, and Javad Lavaei, "Geometric Analysis of Matrix Sensing over Graphs", *Neural Information Processing Systems*, 2023.

- **Chapter 3.**
Main paper:

4. Haixiang Zhang, Baturalp Yalçın, Javad Lavaei, and Somayeh Sojoudi, "A New Complexity Metric for Nonconvex Rank-one Generalized Matrix Completion", *Mathematical Programming*, 2023.

Related paper:

5. Baturalp Yalçın, Haixiang Zhang, Javad Lavaei, and Somayeh Sojoudi, "Factorization Approach for Low-complexity Matrix Completion Problems: Exponential number of spurious solutions and failure of gradient methods", *International Conference on Artificial Intelligence and Statistics*, 2022.

- **Chapter 4.**
Main paper:

6. Haixiang Zhang, Zeyu Zheng, and Javad Lavaei, "Gradient-based Algorithms for Convex Discrete Optimization via Simulation", *Operations Research*, 2022.

Related paper:

7. Haixiang Zhang, Zeyu Zheng, and Javad Lavaei, "Stochastic $L^\natural$-convex Function Minimization", *Neural Information Processing Systems*, 2021.

- **Chapter 5.**
Main paper:

8. Haixiang Zhang, Zeyu Zheng, and Javad Lavaei, "Stochastic Localization Methods for Discrete Convex Simulation Optimization", *Operations Research*, 2023.

Related paper:

9. Haixiang Zhang, Zeyu Zheng, and Javad Lavaei, "Selection of the Best under Convexity", *technical report*, 2021.

- **Chapter 6.**
  Main paper:

  10. Haixiang Zhang, SangWoo Park, Javad Lavaei, and Ross Baldick, "Uniqueness of Power Flow Solutions using Graph-theoretic Notions", *IEEE Transactions on Control of Network Systems*, 2022.

- **Chapter 7.**
  Main paper:

  11. Haixiang Zhang*, Eli Brock*, Javad Lavaei, and Somayeh Sojoudi, "Distributionally Robust Joint Chance-constrained Optimal Power Flow using Relative Entropy", *submitted for journal publication*, 2024.

  Related paper:

  12. Haixiang Zhang*, Eli Brock*, Julie Mulvaney Kemp, Javad Lavaei, and Somayeh Sojoudi, "Distributionally Robust Optimization for Nonconvex QCQPs with Stochastic Constraints", *IEEE Conference on Decision and Control (CDC)*, 2023.

1.3 Notations

Sets. The number of elements in a finite set $\mathcal{S}$ is denoted as $|\mathcal{S}|$. We use $\overline{\mathcal{S}}$ to denote the closure of a set $\mathcal{S}$. For $N \in \mathbb{N}$, we define $[N] := \{1, 2, \ldots, N\}$. For a given set $\mathcal{S}$ and an integer $d \in \mathbb{N}$, the product set $\mathcal{S}^d$ is defined as $\{(x_1, x_2 \ldots, x_d) : x_i \in \mathcal{S}, i \in [d]\}$ in which $[d] = \{1, 2, \ldots, d\}$. For example, if $\mathcal{S} = [N]$, then $\mathcal{S}^d = \{(x_1, x_2 \ldots, x_d) : x_i \in [N], i \in [d]\}$. We use $\mathbb{N}, \mathbb{Z}, \mathbb{R}, \mathbb{C}$ to denote the set of all natural numbers, integers, real numbers and complex numbers, respectively. The set of n-dimensional integer, real and complex vectors are denoted as $\mathbb{Z}^n$, $\mathbb{R}^n$ and $\mathbb{C}^n$, respectively. Similarly, we use $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ to denote the set of m-by-n real and complex matrices, respectively. For every vector $x \in \mathbb{R}^n$, the sets of indices corresponding to zero and nonzero components of x are denoted as $\mathcal{I}_0(x)$ and $\mathcal{I}_1(x)$, respectively. In Chapters 4 and 5, let c be the indifference zone parameter in the PCS-IZ criterion. The symbol $\mathbf{j}$ denotes the unit imaginary number. For every complex number x, the real and imaginary parts of x are denoted as $\Re(x)$ and $\Im(x)$, respectively. The same notation applies componentwise to complex vectors and matrices. For a complex number x, $|x|$ denotes its magnitude.

Vectors, matrices, and tensors. Let $\mathbf{1}_n$ and $\mathbf{0}_n$ be the vectors with all elements equal to 1 and 0, respectively. Denote e_k as the k-th standard basis vector of $\mathbb{R}^n$. Let $\|\cdot\|$ be the 2-norm of vectors. The entry-wise ℓ_1-norm, operator 2-norm and the Frobenius norm of a matrix M are denoted as $\|M\|_1$, $\|M\|_2$ and $\|M\|_F$, respectively. The trace of matrix M is denoted as $\mathrm{tr}(M)$. The inner product between two matrices is defined as $\langle M, N\rangle := \mathrm{tr}(M^T N)$. For a given vector $\mathbf{v} \in \mathbb{R}^n$, matrix $\mathrm{diag}(\mathbf{v}) \in \mathbb{R}^{n\times n}$ is the diagonal

matrix with diagonal entries from $\mathbf{v}$. The unit sphere of matrices with non-negative entries denoted as $\mathbb{S}_{+,1}^{n^2-1}$ is the set of all symmetric matrices $X \in \mathbb{R}^{n\times n}$ such that $\|X\|_1 = 1$ and $X_{ij} \geq 0$ for all $i, j \in [n]$. Similarly, the unit sphere of vectors $\mathbb{S}_1^{n-1}$ is the set of all vectors $x \in \mathbb{R}^n$ such that $\|x\|_1 = 1$. The n-by-n identity matrix is denoted as $\mathcal{I}_n$. The identity tensor is denoted as $\mathcal{I}$. For any matrix $M \in \mathbb{R}^{n\times m}$, we denote its singular values by $\sigma_1(M) \geq \cdots \geq \sigma_k(M)$, where $k := \min\{n, m\}$. For any symmetric matrix $M \in \mathbb{R}^{n\times n}$, we denote its eigenvalues by $\lambda_1(M) \geq \cdots \geq \lambda_n(M)$. The minimal eigenvalue is denoted as $\lambda_{min}(\cdot)$. The notation $M \succeq 0$ means that the matrix M is symmetric and positive semi-definite. The set of symmetric and positive semi-definite matrices of size n-by-n is denoted as $\mathbb{S}_+^n$.

Operations. The notations $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and Hermitian transpose of a matrix, respectively. For two vectors $x, y \in \mathbb{R}^d$, we use $(x \wedge y)_i := \min\{x_i, y_i\}$ and $(x \vee y)_i := \max\{x_i, y_i\}$ to denote the component-wise minimum and maximum. Similarly, the ceiling function $\lceil\cdot\rceil$ and the flooring function $\lfloor\cdot\rfloor$ round each component to an integer when applied to vectors. For each vector $\mathbf{v} \in \mathbb{R}^n$, we say $\mathbf{v} \leq \mathbf{0}_n$ if $\mathbf{v}_k \leq 0$ for all $k \in [n]$. For any matrix U, we use $\mathcal{P}_U$ to denote the orthogonal projection onto the column space of U. For any matrices $A, B \in \mathbb{R}^{n\times m}$, we use $A \otimes B$ to denote the fourth-order tensor whose (i, j, k, ℓ) element is $A_{i,j}B_{k,\ell}$. The sub-matrix $R_{i:j,k:\ell}$ consists of the i-th to the j-th rows and the k-th to the ℓ-th columns of matrix R. The action of the Hessian $\nabla^2 f(M)$ on any two matrices K and L is given by $[\nabla^2 f(M)](K, L) := \sum_{i,j,k,\ell}[\nabla^2 f(M)]_{i,j,k,\ell}K_{ij}L_{k,\ell}$.

Big-O notation. The notations $a_n = O(b_n)$ and $a_n = \Theta(b_n)$ mean that there exist constants $c_1, c_2 > 0$ such that $a_n \leq c_2 b_n$ and $c_1 b_n \leq a_n \leq c_2 b_n$ hold for all $n \in \mathbb{Z}$, respectively. In Chapters 4 and 5, the notation $f = O(g)$ means that there exist constants $c_1, c_2 > 0$ independent of $N, d, \epsilon, \delta, c$ such that $f \leq c_1 g + c_2$. Similarly, the notation $f = \tilde{O}(g)$ means that there exist constants $c_1 > 0$ independent of $N, d, \epsilon, \delta, c$ and constant $c_2 > 0$ independent of δ such that $f \leq c_1 g + c_2$. The notation $f = \Theta(g)$ means that there exist constants $c_1, c_2, c_3 > 0$ independent of $N, d, \epsilon, \delta, c$ such that $c_3 g \leq f \leq c_1 g + c_2$. The notation $f = \tilde{\Theta}(g)$ means that there exist constants $c_1, c_3 > 0$ independent of $N, d, \epsilon, \delta, c$ and constants $c_2, c_4 > 0$ independent of δ such that $c_3 g + c_4 \leq f \leq c_1 g + c_2$. In Chapter 3, the objective function of an instance $\mathcal{MC}(C, u^*)$ is shown as $g(u; C, u^*) := \sum_{i,j\in[n]} C_{ij}(u_i u_j - u_i^* u_j^*)^2$.

Graphs and power systems. The unweighted undirected graph $\mathbb{G}$ with node set $\mathbb{V}$ and edge set $\mathbb{E}$ is denoted as $\mathbb{G} = (\mathbb{V}, \mathbb{E})$. Suppose that the edges of an undirected graph are weighted with the weights captured by a matrix $W \in \mathbb{R}^{|\mathbb{V}|\times|\mathbb{V}|}$, where W_{ij} is the weight of edge $\{i, j\}$. Then, the graph is represented as $(\mathbb{V}, \mathbb{E}, W)$. In Chapter 3, for every instance $\mathcal{MC}(C, u^*)$, we use $\mathbb{G}(C, u^*) = [\mathbb{V}(C, u^*), \mathbb{E}(C, u^*), \mathbb{W}(C, u^*)]$ to denote the associated weighted graph, which is defined in Section 3.2. For a directed graph $(\mathbb{V}, \mathbb{E}, A)$, the matrix $A \in \mathbb{R}^{|\mathbb{V}|\times|\mathbb{V}|}$ gives the orientation of each line, where $A_{ij} = 1$ (resp. $A_{ij} = -1$) represents the direction $i \to j$ (resp. $j \to i$). The undirected edge connecting two vertices k and ℓ is denoted by a set notation $\{k, \ell\}$, whereas (k, ℓ) denotes a directed edge coming out of vertex k and going into ℓ. For parallel edges, we use $\{k, \ell, i\}$ to represent different edges connecting k and ℓ, where $i \in \mathbb{Z}_+$ is the index of each parallel edge.

A power network $\mathbb{G} = (\mathbb{V}, \mathbb{E}, Y)$ consists of two parts: the underlying undirected graph $(\mathbb{V}, \mathbb{E})$ and the complex admittance matrix $Y \in \mathbb{C}^{n\times n}$, where n is the number of vertices in the underlying graph. The underlying graph is assumed to be a simple and connected graph. The set of vertices $\mathbb{V}$ and the set of edges $\mathbb{E}$ correspond to the set of buses and the set of lines of the power network. The series element of the equivalent Π-model of each line $\{k, \ell\}$ is modeled by admittance $Y_{k\ell} = G_{k\ell} - \mathbf{j}B_{k\ell}$, where $G_{k\ell}, B_{k\ell} \geq 0$. We denote $\mathbf{v} \in \mathbb{C}^n$ as the vector of complex bus voltages. The complex voltage at bus k can be written in the polar form using its magnitude and phase angle $v_k = |v_k|e^{\mathbf{j}\Theta_k}$ for all $k \in [n]$, where $|v_k| \in \mathbb{R}$ and $\Theta_k \in \mathbb{R}$ denote the voltage magnitude and phase angle, respectively. We denote $\Theta_{k\ell} := \Theta_k - \Theta_\ell \in [-\pi, \pi)$ as the phase difference modulus by 2π for all $\{k, \ell\} \in \mathbb{E}$. In the rest of the chapter, we use the corresponding values in $[-\pi, \pi)$ for phase differences.

Probability. The failing probability of simulation-optimization algorithms is denoted as δ. For a stochastic system labeled by its decision variable x, we denote ξ_x as the random object associated with the decision variable. We write $\xi_{x,1}, \xi_{x,2}, \ldots, \xi_{x,n}$ as independent and identically distributed (i.i.d.) copies of ξ_x. The empirical mean of the n independent evaluations for a decision variable labeled by x is denoted as $\hat{F}_n(x) := \frac{1}{n}\sum_{j=1}^{n} F(x, \xi_{x,j})$.

Part I

Low-rank Matrix Optimization

Chapter 2

Optimization Complexity Based on The RIP Constant

2.1 Introduction

A variety of modern signal processing and machine learning applications require solving optimization problems that involve a low-rank matrix variable. More specifically, given measurements to some unknown ground truth matrix $M^* \in \mathbb{R}^{n\times n}$ of rank $r \ll n$, the *low-rank matrix optimization* problem can be formulated as

$$\min_{M\in\mathbb{R}^{n\times n}} f(M; M^*) \quad \text{s.t.} \quad M \succeq 0, \quad \text{rank}(M) \leq r, \tag{2.1}$$

where $f(\cdot; M^*)$ is the loss function that penalizes the mismatch between the input matrix and M^*. The goal is to recover the matrix M^* via (2.1). Examples of this problem include matrix sensing [191, 247, 244], matrix completion [34, 35, 86], phase retrieval [200, 32, 210, 48], phase synchronization [204, 26] and robust principle component analysis [36, 75]; see the review papers [40, 50] for more applications.

To deal with the nonconvex rank constraint, there have been several works on the convex relaxations of problem (2.1). More concretely, one may replace the rank constraint with a nuclear norm regularizer [34, 191, 35, 36, 143]. The convex relaxation approach is proven to achieve the optimal sampling complexity for various statistical models. In the special case when $f(\cdot; M^*)$ is a linear function, the sketching method [243] can be applied to accelerate the computation. However, for most applications of problem (2.1), the convex relaxation approach needs to update a matrix variable in each iteration, which relies on the Singular Value Decomposition (SVD) of the matrix variable. This will lead to an $O(n^3)$ computational complexity in each iteration and an $O(n^2)$ space complexity, which are prohibitively high for large-scale problems; see the numerical comparison in [252].

To improve the computational efficiency, an alternative approach was proposed by Burer and Monteiro [28], which is named as the Burer-Monteiro factorization approach. The factorization approach is based on the fact that the mapping $U \mapsto UU^T$ is surjective onto the

manifold of positive semi-definite matrices of rank at most r, where $U \in \mathbb{R}^{n\times r}$. Therefore, problem (2.1) is equivalent to

$$\min_{U\in\mathbb{R}^{n\times r}} f(UU^T; M^*), \tag{2.2}$$

which is an unconstrained nonconvex problem. The Burer-Monteiro factorization provides a natural parameterization of the low-rank structure of the unknown solution, and reformulates problem (2.1) as an unconstrained optimization problem. In addition, the number of variables reduces from $O(n^2)$ or $O(nm)$ to as low as $O(rn)$ or $O(r(n+m))$ when $r \ll \min\{n, m\}$. However, the reformulated problem is highly non-convex, and $\mathcal{NP}$-hard to solve in the worst case. A major difficulty about nonconvex optimization problems is the existence of spurious local minima[1]. In general, common local search methods are only able to guarantee a point approximately satisfying the first-order and the second-order necessary optimality conditions. Therefore, local search methods with a random initialization will likely be stuck at spurious local minima and unable to converge to the global solution.

In recent years, simple iterative algorithms such as gradient descent and alternating minimization have achieved empirical success in various applications of problem (2.2), despite the aforementioned issue of nonconvex optimization problems. Intuitively, these problems share a specific non-convex structure, which makes it possible to utilize the structure and design efficient algorithms to find a global optimum under some conditions. Substantial progress has been made on the theoretical understandings of these algorithms, which generally focused on proving the absence of spurious local minima. For example, the alternating minimization algorithm was first studied in [118, 172, 173]. The (stochastic) gradient descent algorithm, which is in general easier to implement than the alternating minimization algorithm, was analyzed in [32, 219, 242, 48, 40]. Besides algorithmic analysis, a critical geometric property named the strict-saddle property [210] was established in [86, 210, 256, 244], which can guarantee the polynomial-time global convergence of various saddle-escaping algorithms [37, 124, 8].

Restricted Isometry Property and basic properties

In this chapter, we also consider the asymmetric version of problem (2.1), which eliminates the condition $M \succeq 0$ and allows M to be a non-square matrix. More specifically, given the natural numbers n, m and r, we consider the low-rank matrix optimization problems

$$\min_{M\in\mathbb{R}^{n\times n}} \quad f_s(M) \quad \text{s.t. } \operatorname{rank}(M) \le r, \quad M \succeq 0 \tag{2.3}$$

and

$$\min_{M\in\mathbb{R}^{n\times m}} \quad f_a(M) \quad \text{s.t. } \operatorname{rank}(M) \le r, \tag{2.4}$$

[1] A point U^0 is called a spurious local minimum if it is a local minimum of problem (2.2) and $U^0(U^0)^T \neq M^*$.

where the functions $f_s(\cdot)$ and $f_a(\cdot)$ are twice continuously differentiable. Problems (2.3)-(2.4) are referred to as the *symmetric* and the *asymmetric* problems, respectively. In addition, we call these problems *linear* if the objective function is induced by a linear measurement operator, i.e.,

$$f(M) = \tfrac{1}{2}\|\mathcal{A}(M) - b\|_F^2 \tag{2.5}$$

for some vector $b \in \mathbb{R}^p$ and linear operator $\mathcal{A}$ mapping each matrix M to a vector in $\mathbb{R}^p$, where $f(M)$ denotes either $f_s(M)$ or $f_a(M)$. Those problems not fitting into the above model are called *nonlinear*. One common example with non-linearity is the one-bit matrix sensing problem; please see [256, 145, 257] for more concrete discussions. Using the Burer-Monteiro factorization approach, the asymmetric problem (2.4) is equivalent to

$$\min_{U\in\mathbb{R}^{n\times r}, V\in\mathbb{R}^{m\times r}} \; h_a(U, V), \tag{2.6}$$

where $h_a(U, V) := f_a(UV^T)$. Similarly, the symmetric problem (2.3) is equivalent to

$$\min_{U\in\mathbb{R}^{n\times r}} \; h_s(U), \tag{2.7}$$

where $h_s(U) := f_s(UU^T)$.

A regularity condition, named the Restricted Isometry Property, is commonly used to guarantee the convergence of iterative local search algorithms. We state the following two definitions only in the context of the symmetric problem since the corresponding definitions for the asymmetric problem are similar.

Definition 1 ([191, 256])**.** Given natural numbers r and t, the function $f_s(\cdot)$ is said to satisfy the **Restricted Isometry Property** (RIP) of rank $(2r, 2t)$ for a constant $\delta \in [0, 1)$, denoted as δ-RIP$_{2r,2t}$, if for all matrices $M, K \in \mathbb{R}^{n\times n}$ such that $\mathrm{rank}(M) \le 2r, \mathrm{rank}(K) \le 2t$, it holds that

$$(1-\delta)\|K\|_F^2 \le \left[\nabla^2 f(M)\right](K, K) \le (1+\delta)\|K\|_F^2, \tag{2.8}$$

where $[\nabla^2 f(M)](\cdot, \cdot)$ is the curvature of the Hessian at point M.

The RIP condition appears in a variety of applications of the low-rank matrix optimization problem. For instance, in the case of linear measurements with a Gaussian model, [34] showed that $O(nr/\delta^2)$ samples are enough to ensure the δ-RIP$_{2r,2r}$ property with high probability. Please see the survey paper by [50] for more examples. In certain applications, even the RIP condition cannot be established over the whole low-rank manifold, we are able to establish similar strongly convex and smooth conditions on part of the manifold. If the iteration points of algorithms are constrained to or regularized (either explicitly or implicitly) towards those benign regions, the proof techniques in this chapter may still be applicable. Examples include the phase retrieval problem [157] and the matrix completion problem

[40]. However, the analysis of the case when the strong convexity does not hold is usually application-specific and cannot be generalized to general low-rank problems. Moreover, the RIP assumption is standard in the literature of general low-rank matrix optimization problem. Furthermore, if we drop the strong convexity assumption, we are unable to achieve linear convergence in general [21]. The work by [247] shows that the existence of RIP is enough to obtain guarantees on the local landscape of the problem and the size of this local region depends on the RIP constant that can be anything between 0 and 1 (however, the provided bounds on the RIP constant are not sharp). Although we aimed to obtain sharp bounds on the RIP constant for global landscape of the problem in this chapter, we believe that our analysis can be adopted to obtain sharp RIP bounds for local regions. We leave the precise derivation to a future work since it needs a number of lemma and we have space restrictions. We note that the RIP property is equivalent to the restricted strongly convex and smooth property defined in [227, 183, 257] with the condition number $(1+\delta)/(1-\delta)$. Intuitively, the RIP property implies that the Hessian matrix is close to the identity tensor when the perturbation is restricted to be low-rank. This intuition naturally leads to the following definition.

Definition 2 ([23])**.** Given a natural number r, the function $f_s(\cdot)$ is said to satisfy the **Bounded Difference Property** (BDP) of rank $2r$ for a constant $\kappa \geq 0$, denoted as κ-BDP$_{2r}$, if for all matrices $M, M', K, L \in \mathbb{R}^{n\times n}$ such that

$$\text{rank}(M),\ \ \text{rank}(M'),\ \ \text{rank}(K),\ \ \text{rank}(L) \leq 2r,$$

it holds that

$$\left|\left[\nabla^2 f_s(M) - \nabla^2 f_s(M')\right](K, L)\right| \leq \kappa \|K\|_F \|L\|_F.$$

It has been proven in [23, Theorem 1] that those functions satisfying the δ-RIP$_{2r,2r}$ property also satisfy the 4δ-BDP$_{2r}$ property. With the RIP property, there are basically two categories of algorithms that can solve the factorized problem in polynomial time. Algorithms in the first category require a careful initialization so that the initial point is already in a small neighbourhood of a global optimum, and a certain local regularity condition in the neighbourhood ensures that local search algorithms will converge linearly to a global optimum; see [219, 21, 183] for a detailed discussion. The other class of algorithms is able to converge globally from a random initialization. The convergence of these algorithms is usually established via the geometric analysis of the landscape of the objective function. One of the important geometric properties is the strict saddle property [210], which combined with the smoothness properties can guarantee the global polynomial-time convergence for various saddle-escaping algorithms [125, 124, 210, 110]. For the linear case, [87, 86] proved the strict saddle property for both problems (2.6)-(2.7) when the RIP constant is sufficiently small. More recently, [257] extended the results to the nonlinear asymmetric case. Moreover, a weaker geometric property, namely the non-existence of *spurious (non-global) second-order critical points*, has been established for both problems when the RIP constant is small [145, 95]. We note that second-order critical points are points that satisfy the first-order and

the second-order necessary optimality conditions, and thus the result of the non-existence of second-order critical points implies the non-existence of spurious local minima. Under certain regularity conditions, this weaker property is also able to guarantee the global convergence from a random initialization without an explicit convergence rate [138, 179]. Please refer to Table 2.1 for a summary of the state-of-the-art results.

Most of the aforementioned papers are based on the following assumption on the low-rank critical points of the functions $f_s(\cdot)$ and $f_a(\cdot)$:

Assumption 1. The function $f_a(\cdot)$ has a first-order critical point M_a^* such that $\text{rank}(M_a^*) \leq r$. Similarly, the function $f_s(\cdot)$ has a first-order critical point M_s^* that is symmetric, positive semi-definite and of rank at most r.

This assumption is inspired by the noiseless matrix sensing problem in the linear case for which the non-negative objective function becomes zero (the lowest value possible) at the true solution. This is a natural property of the matrix sensing problem for nonlinear measurement models as well. Under the above assumption and the RIP property, [256] proved that M_s^* and M_a^* are the unique global minima of problems (2.3)-(2.4).

Theorem 1 ([256])**.** *If the functions $f_s(\cdot)$ and $f_a(\cdot)$ satisfy the δ-RIP$_{2r,2r}$ property, then the critical points M_s^* and M_a^* are the* unique global minima *of problems* (2.3)-(2.4)*.*

Given a solution (U^*, V^*) to problem (2.6), we observe that (U^*P, V^*P^{-T}) is also a solution for any invertible $P \in \mathbb{R}^{r\times r}$. This redundancy may induce an extreme non-convexity on the landscape of the objective function. To reduce this redundancy, [219] considered the regularized problem

$$\min_{U\in\mathbb{R}^{n\times r}, V\in\mathbb{R}^{m\times r}} \rho(U, V), \tag{2.9}$$

where

$$\rho(U, V) := h_a(U, V) + \frac{\mu}{4} \cdot g(U, V)$$

with a constant $\mu > 0$ and the regularization term

$$g(U, V) := \|U^TU - V^TV\|_F^2.$$

The regularization term is introduced to balance the magnitudes of U^* and V^*. [256] showed that the regularization term does not introduce bias and thus problem (2.9) is equivalent to the original problem (2.4) in the sense that any first-order critical point (U, V) of problem (2) corresponds to a first-order critical point of problem (5) with balanced energy, i.e. $U^TU = V^TV$.

Theorem 2 ([256])**.** *Every first-order critical point (U^*, V^*) of problem* (2.9) *satisfies*

$$(U^*)^TU^* = (V^*)^TV^*.$$

Moreover, problems (2.4) *and* (2.9) *are equivalent.*

Detailed optimality conditions for problems (2.3)-(2.9) are provided in the appendix.

Problem Setups		No SSCPs		SSP Holds	
		Existing	Ours	Existing	Ours
Rank-1 Sym.	**Linear**	$\delta_{2,2} < \frac{1}{2}$ [247]	$\delta_{2,2} < \frac{1}{2}$	-	-
	Nonlinear	$\delta_{2,2} < \frac{2-O(\kappa)}{4+O(\kappa)}$ [23]	$\delta_{2,2} < \frac{1}{2}$	-	-
Rank-1 Asym.	**Linear & Nonlinear**	-	$\delta_{2,2} < \frac{1}{2}$	-	-
Rank-r Sym.	**Linear**	$\delta_{2r,2r} < \frac{1}{5}$ [87]	$\delta_{2r,2r} \leq \frac{1}{3}$	$\delta_{2r,2r} < \frac{1}{10}$ [86]	$\delta_{2r,2r} < \frac{1}{3}$
	Nonlinear	$\delta_{2r,4r} < \frac{1}{5}$ [145]	$\delta_{2r,2r} \leq \frac{1}{3}$	-	$\delta_{2r,2r} < \frac{1}{3}$
Rank-r Asym.	**Linear**	$\delta_{2r,2r} < \frac{1}{3}$ [95]	$\delta_{2r,2r} \leq \frac{1}{3}$	$\delta_{2r,2r} < \frac{1}{20}$ [86]	$\delta_{2r,2r} < \frac{1}{3}$
	Nonlinear	$\delta_{2r,2r} < \frac{1}{3}$ [95]	$\delta_{2r,2r} \leq \frac{1}{3}$	$\delta_{2r,4r} < \frac{\alpha(M_a^*)}{100}$ [257]	$\delta_{2r,2r} < \frac{1}{3}$

Table 2.1: Comparison of the state-of-the-art results and our results. Here $\delta_{2r,2t}$ and κ are the $\text{RIP}_{2r,2t}$ and BDP_{2r} constants of $f_s(\cdot)$ or $f_a(\cdot)$, respectively. Constant $\alpha(M_a^*) \in (0,1)$ only depends on M_a^*. "SSCP" and "SSP" refer to suprious second-order critical points and strict saddle property, respectively.

Contributions

In this chapter, we analyze the geometric properties of problems (2.7)-(2.9). Novel analysis methods are developed to obtain less conservative conditions for guaranteeing benign landscapes for both problems. We note that, unlike the linear measurements case, the RIP constant of nonlinear problems may not concentrate to 0 as the number of samples increases. Therefore, a sharper RIP bound leads to theoretical guarantees that hold under less stringent statistical requirements. In addition, even if the RIP constant concentrates to 0 when more samples are included, there may only be a limited number of samples available, either due to the constraints of specific applications or to the great expense of taking more samples. Hence, obtaining a sharper RIP bound is essential for many applications. We summarize our results in Table 2.1. More concretely, the contributions of this chapter are three-folds.

First, we derive necessary conditions and sufficient conditions for the existence of spurious second-order critical points for both symmetric and asymmetric problems. Using our necessary conditions, we show that the δ-$\text{RIP}_{2r,2r}$ property with $\delta \leq 1/3$ is enough to guarantee the non-existence of such points. This result provides a marginal improvement to the previous work [95], which developed the sufficient condition $\delta < 1/3$ for asymmetric problems, and is a major improvement over [87] and [145], which requires $\delta < 1/5$ for symmetric problems. With this non-existence property and under some common regularity conditions,

[138, 179] showed that the vanilla gradient descent method with a small enough step size and a random initialization almost surely converges to a global minimum. We note that the convergence rate was not studied and could theoretically be exponential in the worst case. In addition, by studying our necessary conditions, we show that every second-order critical point has a positive correlation to the global minimum when $\delta \in (1/3, 1/2)$. When $\delta = 1/2$, a counterexample with spurious second-order critical points is given by utilizing the sufficient conditions. We note that the sufficient conditions can greatly simplify the construction of counterexamples.

Second, we separately study the rank-1 case to further strengthen the bounds. In particular, we utilize the necessary conditions to prove that the δ-RIP$_{2,2}$ property with $\delta < 1/2$ is enough for the non-existence of spurious second-order critical points. Combining with a counterexample in the $\delta = 1/2$ case, we conclude that the bound $\delta < 1/2$ is the sharpest bound for the rank-1 case. Our results significantly extend the bounds in [247] derived for the linear symmetric case to the linear asymmetric case and the general nonlinear case. It also improves the bound in [23] by dropping the BDP constant.

Third, we prove that in the exact parametrization case, problems (2.7)-(2.9) both satisfy the strict saddle property [210] when the δ-RIP$_{2r,2r}$ property is satisfied with $\delta < 1/3$. This result greatly improves the bounds in [86, 257] and extends the result in [95] to approximate second-order critical points. With the strict saddle property and certain smoothness properties, a wide range of algorithms guarantee a global polynomial-time convergence with a random initialization; see [125, 124, 210, 110]. Due to the special non-convex structure of our problems and the RIP property, it is possible to prove the boundedness of the trajectory of the perturbed gradient descent method using a similar method as in [125]. Since the smoothness properties are satisfied over a bounded region, combined with the strict saddle property, it follows that the perturbed gradient descent method [125] achieves a polynomial-time global convergence when $\delta < 1/3$.

The remainder of this chapter is organized as follows. In Section 2.2, the Singular Value Projection algorithm is analyzed as an enlightening example for our main results. Sections 2.3 and 2.4 are devoted to the non-existence of spurious second-order critical points and the strict saddle property of the low-rank optimization problem in both symmetric and asymmetric cases, respectively.

2.2 Motivating Example: Singular Value Projection Algorithm

Before providing theoretical results for problems (2.7)-(2.9), we first consider the Singular Value Projection Method (SVP) algorithm (Algorithm 1) as a motivating example, which is proposed in [117]. The SVP algorithm is basically the projected gradient method of the original low-rank problems (2.3)-(2.4) via the truncated SVD. For the asymmetric problem

(2.4), the low-rank manifold is

$$\mathcal{M}_{asym} := \{M \in \mathbb{R}^{n\times m} \mid \text{rank}(M) \leq r\}$$

and the projection is given by only keeping components corresponding to the r largest singular values. For the symmetric problem (2.3), the low-rank manifold is

$$\mathcal{M}_{sym} := \{M \in \mathbb{R}^{n\times n} \mid \text{rank}(M) \leq r, \quad M^T = M, \quad M \succeq 0\}.$$

We assume without loss of generality that the gradient $\nabla f(\cdot)$ is symmetric; see Appendix 2.A for a discussion. The projection is given by only keeping components corresponding to the r largest eigenvalues and dropping all components with negative eigenvalues. Since both low-rank manifolds are non-convex, the projection solution may not be unique and we choose an arbitrary solution when it is not unique. We note that the above projections are orthogonal in the sense that

$$\|M_+ - M\|_F = \min_{K\in\mathcal{M}} \|K - M\|_F,$$

where M_+ is the projection of a matrix M. Henceforth, $\mathcal{M}$ stands for $\mathcal{M}_{sym}$ or $\mathcal{M}_{asym}$, which should be clear from the context. Although each truncated SVD operation can be computed within $O(nmr)$ operations, the constant hidden in the $O(\cdot)$ notation is considerably larger than 1. Thus, the truncated SVD operation is significantly slower than matrix multiplication, which makes the SVP algorithm impractical for large-scale problems. However, the analysis of the SVP algorithm, combining with the equivalence property given in [95], provides some insights into how to develop proof techniques for problems (2.7)-(2.9).

We extend the proof in [117] and show that Algorithm 1 converges linearly to the global minimum under the δ-RIP$_{2r,2r}$ property with $\delta < 1/3$.

Theorem 3. *If function $f_s(\cdot)$ (resp. $f_a(\cdot)$) satisfies the δ-RIP$_{2r,2r}$ property with $\delta < 1/3$ and the step size is chosen to be $\eta = (1+\delta)^{-1}$, then Algorithm 1 applied to problem* (2.3) *(resp.* (2.4)*) returns a solution M_T such that $M_T \in \mathcal{M}$ and $f(M_T) - f(M^*) \leq \epsilon$ within*

$$T := \left\lceil \frac{1}{\log[(1-\delta)/(2\delta)]} \cdot \log\left[\frac{f(M_0) - f(M^*)}{\epsilon}\right]\right\rceil$$

iterations, where $f(\cdot) := f_s(\cdot)$ (resp. $f(\cdot) := f_a(\cdot)$), M^* is the global minimum, M_0 is the initial point and $\lceil\cdot\rceil$ is the ceiling function.

The proof is almost identical to that in [117] except that we have replaced the quadratic function with the RIP bounds. However, the result of the proof provides a key inequality (2.17) for the subsequent proofs. We note that the above proof can be applied to other low-rank optimization problems with a suitable definition of the orthogonal projection. In [95], it is proved that the unique global minimum is the only fixed point of the SVP algorithm if the RIP constant δ is less than 1/3. However, the above paper has not proven the linear

Algorithm 1 Singular Value Projection (SVP) Algorithm

Input: Low-rank manifold $\mathcal{M}$, initial point M_0, number of iterations T, step size η, objective function $f(\cdot)$.

Output: Low-rank solution M_T.

1: **for** $t = 0, \ldots, T-1$ **do**
2: Update $\tilde{M}_{t+1} \leftarrow M_t - \eta \nabla f(M_t)$.
3: Set M_{t+1} to be the projection of $\tilde{M}_{t+1}$ onto $\mathcal{M}$ via truncated SVD.
4: **end for**
5: **return** M_T.

convergence (as done in Theorem 3). This difference leads to a strengthened inequality in the following analysis, which further serves as an essential step in proving the strict saddle property. The results in this section provide a hint that the landscape may be benign when the RIP constant is smaller than 1/3 and we may be able to establish linear convergence under this condition, which is the main topic of the remainder of this chapter.

2.3 No Spurious Second-order Critical Points

In this section, we develop necessary conditions and sufficient conditions for the existence of spurious second-order critical points of problems (2.7)-(2.9). Besides the non-existence of spurious local minima, the non-existence of spurious second-order critical points also guarantees the global convergence of many first-order algorithms with random initialization under certain regularity conditions [138, 179]. More precisely, we require the iterates of the algorithm to converge to a single point and the objective function to have a Lipschitz-continuous gradient. The first condition is satisfied by the gradient descent method applied to a large class of functions known as the KŁ-functions [14]. For the second condition, many objective functions that appear in applications, e.g., the ℓ_2-loss function, do not satisfy this condition. However, if the step size is small enough, the special non-convex structure of the Burer-Monteiro decomposition and the RIP property ensure that the trajectory of the gradient descent method stays in a compact set, where the Lipschitz condition is satisfied due to the second-order continuity of the functions $f_s(\cdot)$ and $f_a(\cdot)$. The proof of this claim is similar to Theorem 8 in [125] and is omitted here. Therefore, the non-existence of spurious second-order critical points can ensure the global convergence of the gradient descent method for many applications.

The non-existence of spurious second-order critical points has been proved in [86, 256] for problems with linear and nonlinear measurements, respectively. Recently, [95] proved a relation between the second-order critical points of problem (2.6) or (2.9) and the fixed points of the SVP algorithm on problem (2.4). Using this relation, they showed that problems (2.6) and (2.9) have no spurious second-order critical points when the δ-RIP$_{2r,2r}$ property is satisfied with $\delta < 1/3$. In this chapter, we take a different approach to show that $\delta \leq 1/3$ is

enough for the general case in both symmetric and asymmetric scenarios, and that $\delta < 1/2$ is enough for the rank-1 case. Moreover, we prove that there exists a positive correlation between every second-order critical point and the global minimum when $\delta \in (1/3, 1/2)$. We also show that there may exist spurious second-order critical points when $\delta = 1/2$ for both the symmetric and asymmetric problems, which extends the construction of such examples for the linear symmetric rank-1 problem in [248] to general cases. We first give necessary conditions and sufficient conditions for the existence of a function that satisfies the δ-RIP$_{2r,2r}$ condition and spurious second-order critical points below.

Theorem 4. *Let $\ell := \min\{m, n, 2r\}$. For a given $\delta \in [0, 1)$, there exists a function $f_a(\cdot)$ with the δ-RIP$_{2r,2r}$ property such that problems (2.6) and (2.9) have a spurious second-order critical point only if $1 - \delta < (1 + \delta)/2$ and there exists a constant $\alpha \in (1 - \delta, (1 + \delta)/2]$, a diagonal matrix $\Sigma \in \mathbb{R}^{r\times r}$, a diagonal matrix $\Lambda \in \mathbb{R}^{(\ell-r)\times(\ell-r)}$ and matrices $A \in \mathbb{R}^{r\times r}, B \in \mathbb{R}^{r\times r}, C \in \mathbb{R}^{(\ell-r)\times r}, D \in \mathbb{R}^{(\ell-r)\times r}$ such that*

If $CB^T = 0$ and $AD^T = 0$, then there exists a function $f_a(\cdot)$ with the δ-RIP$_{2r,2r}$ property such that problems (2.6) and (2.9) have a spurious second-order critical point.

$$
\begin{aligned}
&(1+\delta)\min_{1\le i\le r}\Sigma_{ii} \ge \max_{1\le i\le \ell-r}\Lambda_{ii}, \quad \Sigma \succ 0,\ \Lambda \succeq 0,\\
&\langle \Lambda, CD^T\rangle = \alpha\left[\mathrm{tr}(\Sigma^2) - 2\langle \Sigma, AB^T\rangle + \|AB^T\|_F^2 + \|AD^T\|_F^2 + \|CB^T\|_F^2 + \|CD^T\|_F^2\right], \quad (2.10)\\
&\mathrm{tr}(\Lambda^2) \le \alpha^{-1}(2\alpha - 1 + \delta^2)\cdot\langle \Lambda, CD^T\rangle, \quad \langle \Lambda, CD^T\rangle \ne 0.
\end{aligned}
$$

Remark 1. We note that there may exist simpler forms of the above conditions. For instance, we may solve α via the condition in the second line of (2.10) and substitute into other conditions. In addition, the requirement that $\alpha \in (1 - \delta, (1 + \delta)/2]$ may also be dropped without affecting the conditions. More specifically, the conditions in (2.10) are equivalent to

$$
\begin{aligned}
&(1+\delta)\min_{1\le i\le r}\Sigma_{ii} \ge \max_{1\le i\le \ell-r}\Lambda_{ii}, \quad \Sigma \succ 0,\ \Lambda \succeq 0, \quad \langle \Lambda, CD^T\rangle \ne 0,\\
&\mathrm{tr}(\Lambda^2) \le 2\cdot\langle \Lambda, CD^T\rangle - (1-\delta^2)\Big[\mathrm{tr}(\Sigma^2) - 2\langle \Sigma, AB^T\rangle\\
&\qquad\qquad + \|AB^T\|_F^2 + \|AD^T\|_F^2 + \|CB^T\|_F^2 + \|CD^T\|_F^2\Big].
\end{aligned}
$$

We state Theorem 4 in the current form since it helps with deriving corollaries more directly.

Intuitively, Λ and Σ correspond to the singular values of the second-order critical point and the gradient at the second-order critical point, respectively. Matrices A, B, C, D correspond to the SVD of the global optimum. The original problem of the non-existence of spurious second-order critical points can be viewed as a property of the set of functions satisfying the RIP property, which is a convex set in an infinite-dimensional functional space. The conditions in (2.10) reduce the infinite-dimensional problem to a finite-dimensional problem by utilizing the optimality conditions and the RIP property, which provides a basis for solving these conditions numerically. We note that the conditions in the third line of (2.10) are

novel and serve as an important step in developing strong theoretical guarantees. Although the conditions in (2.10) seem complicated, they lead to strong results on the non-existence of spurious second-order critical points. We provide two corollaries below to illustrate the power of the above theorem. The first corollary focuses on the rank-1 case. In this case, we can simplify condition (2.10) through suitable relaxations to obtain a sharper bound on δ that ensures the non-existence of spurious second-order critical points.

Corollary 1. *Consider the case $r = 1$, and suppose that the function $f_a(\cdot)$ satisfies the δ-RIP$_{2,2}$ property with $\delta < 1/2$. Then, problems* (2.6) *and* (2.9) *have no spurious second-order critical points.*

The following example shows that the counterexample in [247] designed for the symmetric case also works for the asymmetric rank-1 case.

Example 1. We note that Example 12 in [247] shows that problem (2.7) may have spurious second-order critical points when $\delta = 1/2$. In general, a second-order critical point for problem (2.7) is not a second-order critical point for problem (2.9), since the asymmetric manifold $\mathcal{M}_{asym}$ has a larger second-order critical cone than the symmetric manifold $\mathcal{M}_{sym}$. However, it can be verified that the same example also has a spurious second-order critical point in the asymmetric case. For completeness, we verify the claim in the appendix.

It follows from Corollary 1 and Example 1 that the bound $1/2$ is the *sharpest* bound for the rank-1 asymmetric case. The next corollary provides a marginal improvement to the state-of-the-art result for the general rank case, which derives the RIP bound $\delta < 1/3$ [95]. In addition, we prove that there exists a positive correlation between every second-order critical point and the global minimum when $\delta < 1/2$.

Corollary 2. *Given an arbitrary r, suppose that the function $f_a(\cdot)$ satisfies the δ-RIP$_{2r,2r}$ property. If $\delta \leq 1/3$, then both problems* (2.6) *and* (2.9) *have no spurious second-order critical points. In addition, if $\delta \in [0, 1/2)$, then every second-order critical point $\tilde{M}$ has a positive correlation with the ground truth M_a^*. Namely, there exists a universal function $C(\delta) : (0, 1/2) \mapsto (0, 1]$ such that*

$$\langle \tilde{M}, M_a^* \rangle \geq C(\delta) \cdot \|\tilde{M}\|_F \|M_a^*\|_F.$$

For the general rank-r case, we construct a counterexample with spurious second-order critical points when $\delta = 1/2$.

Example 2. Let $n = m = 2r$. Now, we use the sufficiency part of Theorem 4 to construct a counterexample. We choose

$$\delta := \frac{1}{2}, \quad \alpha := \frac{3}{5}, \quad \Sigma := \frac{1}{2} I_r, \quad \Lambda := \frac{3}{4} I_r, \quad A = B := 0_r, \quad C = D := I_r.$$

It can be verified that the conditions in (2.10) are satisfied and $CB^T = AD^T = 0$, which means that there exists a function $f_a(\cdot)$ satisfying the δ-RIP$_{2r,2r}$ property for which problems

(2.6) and (2.9) have spurious second-order critical points. We also give a direct construction with linear measurements in the appendix. This example illustrates that Theorem 4 can be used to systematically design instances of the problem with spurious second-order critical points.

Before closing this section, we note that similar conditions can be obtained for the symmetric problem (2.7). Although there exists a natural transformation of symmetric problems to asymmetric problems (see the appendix), the approach requires the objective function $f_s(\cdot)$ to have the δ-RIP$_{4r,2r}$ property, which provides sub-optimal RIP bounds compared to a direct analysis. We give the results of the direct analysis below and omit the proof due to the similarity to the asymmetric case.

Theorem 5. *Let $\ell := \min\{n, 2r\}$. For a given $\delta \in [0, 1)$, there exists a function $f_s(\cdot)$ with the δ-RIP$_{2r,2r}$ property such that problem (2.7) has a spurious second-order critical point only if $1 - \delta < (1 + \delta)/2$ and there exists a constant $\alpha \in (1 - \delta, (1 + \delta)/2]$, a diagonal matrix $\Sigma \in \mathbb{R}^{r\times r}$, a diagonal matrix $\Lambda \in \mathbb{R}^{(\ell-r)\times(\ell-r)}$ and matrices $A \in \mathbb{R}^{r\times r}, C \in \mathbb{R}^{(\ell-r)\times r}$ such that*

$$\begin{aligned} &(1+\delta)\min_{1\le i\le r}\Sigma_{ii} \ge \max_{1\le i\le \ell-r}\Lambda_{ii}, \quad \Sigma \succ 0, \\ &\langle \Lambda, CC^T\rangle = \alpha\left[\mathrm{tr}(\Sigma^2) - 2\langle \Sigma, AA^T\rangle + \|AA^T\|_F^2 + 2\|AC^T\|_F^2 + \|CC^T\|_F^2\right], \\ &\mathrm{tr}(\Lambda^2) \le \alpha^{-1}(2\alpha - 1 + \delta^2)\cdot\langle \Lambda, CC^T\rangle, \quad \langle \Lambda, CC^T\rangle \ne 0. \end{aligned} \tag{2.11}$$

If $AC^T = 0$, then there exists a function $f_s(\cdot)$ with the δ-RIP$_{2r,2r}$ property for which problem (2.7) has a spurious second-order critical point.

Compared to Theorem 4, the diagonal matrix Λ is not enforced to be positive semidefinite. The reason is that the eigenvalue decomposition is used instead of the singular value decomposition in the symmetric case, and therefore some eigenvalues can be negative. Similarly, we can obtain the non-existence and the positive correlation results for the symmetric problem.

Corollary 3. *If function $f_s(\cdot)$ satisfies the δ-RIP$_{2r,2r}$ property, then the following statements hold:*

- *If $\delta \le 1/3$, then there are no spurious second-order critical points;*
- *If $\delta < 1/2$, then there exists a positive correlation between every second-order critical point and the ground truth;*
- *If $\delta = 1/2$, then there exists a counterexample with spurious second-order critical points;*
- *If $\delta < 1/2$ and $r = 1$, then there are no spurious second-order critical points.*

We note that the last statement serves as a generalization of the results in [247] to the nonlinear measurement case, and improves upon the bound in [23] by dropping the BDP constant.

2.4 Global Landscape: Strict Saddle Property

Although the non-existence of spurious second-order critical points can ensure the global convergence under certain regularity conditions, it cannot guarantee a fast convergence rate in general. Saddle-point escaping algorithms may become stuck at approximate second-order critical points for exponentially long time. To guarantee the global polynomial-time convergence, the following strict saddle property is commonly considered in the literature:

Definition 3 ([210])**.** Consider an arbitrary optimization problem $\min_{x\in\mathcal{X}\subset\mathbb{R}^d} F(x)$ and let $\mathcal{X}^*$ denote the set of its global minima. It is said that the problem satisfies the (α,β,γ)-**strict saddle property** for $\alpha,\beta,\gamma>0$ if at least one of the following conditions is satisfied for every $x\in\mathcal{X}$:

$$\operatorname{dist}(x,\mathcal{X}^*)\le\alpha;\quad \|\nabla F(x)\|_F\ge\beta;\quad \lambda_{min}[\nabla^2 F(x)]\le-\gamma.$$

For the low-rank problems, we choose the distance to be the Frobenius norm in the factorization space. This distance is equivalent to the Frobenius norm in the matrix space in the sense that there exist constants $c_1(\mathcal{X}^*)>0$ and $c_2(\mathcal{X}^*)>0$ such that

$$c_1(\mathcal{X}^*)\cdot\|U-U^*\|_F\le\|UU^T-U^*(U^*)^T\|_F\le c_2(\mathcal{X}^*)\cdot\|U-U^*\|_F$$

holds for all $U\in\mathcal{X}$ as long as $\|U-U^*\|_F$ is small and $\mathcal{X}^*$ is bounded [219]. A similar relation holds for the asymmetric case.

In [125], it has been proved that the perturbed gradient descent method can find an ϵ-approximate second-order critical point in $\tilde{O}(\epsilon^{-2})$ iterations with high probability if the Hessian of the objective function is Lipschitz. Namely, the algorithm returns a point $x\in\mathcal{X}$ such that

$$\|\nabla F(x)\|_F\le O(\epsilon),\quad \lambda_{min}[\nabla^2 F(x)]\ge-O(\sqrt{\epsilon})$$

in $\tilde{O}(\epsilon^{-2})$ iterations with high probability. If we choose $\epsilon>0$ to be small enough such that $O(\epsilon)<\beta$ and $-O(\sqrt{\epsilon})>-\gamma$, then the strict saddle property ensures that the returned point satisfies $\operatorname{dist}(x,\mathcal{X}^*)\le\alpha$ with high probability. We note that the Lipschitz continuity of the Hessian can be similarly guaranteed by the boundedness of trajectories of the perturbed gradient method, which can be proved similarly as Theorem 8 in [125]. Since the smoothness properties are satisfied over a bounded region, we may apply the perturbed gradient descent method [125] to achieve the polynomial-time global convergence with random initialization.

In this section, we prove that problems (2.7) and (2.9) satisfy the strict saddle property with an arbitrary $\alpha>0$ in the exact parameterization case, i.e., when the global optimum has rank r.

Assumption 2. The global optimum M_a^* or M_s^* has rank r.

It has been proved in [257] that the regularized problem (2.9) satisfies the strict saddle property if the function $f_a(\cdot)$ has the δ-$\mathrm{RIP}_{2r,4r}$ property with

$$\delta<\frac{\sigma_r(M_a^*)^{3/2}}{100\|M_a^*\|_F\|M_a^*\|_2^{1/2}}.$$

Our results improve upon their bounds by allowing a larger problem-free RIP constant and requiring only the $\text{RIP}_{2r,2r}$ property (note that there are problems with $\text{RIP}_{2r,2r}$ property for which the $\text{RIP}_{2r,4r}$ property does not hold [23]). Our result can also be viewed as a robust version of the results in [95].

Theorem 6. *Suppose that the function $f_a(\cdot)$ satisfies the δ-RIP$_{2r,2r}$ property with $\delta < 1/3$. Given an arbitrary constant $\alpha > 0$, if μ is selected to belong to the interval $[(1-\delta)/3, 1-\delta)$, then there exist positive constants*

$$\epsilon_1 := \epsilon_1(\delta, r, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha), \quad \lambda_1 := \lambda_1(\delta, r, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha)$$

such that for every $\epsilon \in (0, \epsilon_1]$ and $\lambda \in (0, \lambda_1]$, problem (2.9) satisfies the (α, β, γ)-strict saddle property with

$$\beta := \min\left\{\mu(\epsilon/r)^{3/2}, \lambda\right\}, \quad \gamma := \mu\epsilon.$$

We note that the constraint $\mu \in [(1-\delta)/3, 1-\delta)$ is not optimal and it can be similarly proved that $\mu \in (\delta, 1-\delta)$ also guarantees the strict saddle property. The key step in the proof is to show that for every point (U, V) at which the gradient of $f_a(UV^T)$ is small, it holds that

$$\|\nabla f_a(UV^T)\|_2^2 \geq (1+\delta)^2\sigma_r^2(UV^T) + C \cdot (1-3\delta)[f_a(UV^T) - f_a(M_a^*)],$$

where $C > 0$ is a constant independent of (U, V). This inequality can be viewed as a major extension of the non-existence of spurious second-order critical points when $\delta < 1/3$ [95], which shows that every spurious second-order critical point (U, V) satisfies

$$\|\nabla f_a(UV^T)\|_2^2 > (1+\delta)\sigma_r^2(UV^T).$$

We emphasize that our proof requires a new framework and is not a standard revision of the existing methods, which is the reason why sharper bounds can be established. By replacing $\|\nabla f_a(M)\|_2$ with $-\lambda_{min}(\nabla f_s(M))$, the analysis for the asymmetric case can be extended to the symmetric case with minor modifications and the same bound follows.

Theorem 7. *Suppose that the function $f_s(\cdot)$ satisfies the δ-RIP$_{2r,2r}$ property with $\delta < 1/3$. Given an arbitrary constant $\alpha > 0$, there exists a positive constant*

$$\lambda_1 := \lambda_1(\delta, r, \sigma_r(M_s^*), \|M_s^*\|_F, \alpha)$$

such that for every $\lambda \in (0, \lambda_1]$, problem (2.7) satisfies the (α, β, γ)-strict saddle property with

$$\beta := \lambda, \quad \gamma := 2\lambda.$$

The above bound is the first theoretical guarantee of the strict saddle property for the nonlinear symmetric problem.

Appendix

2.A Optimality Conditions

In this section, we develop the optimality conditions for problems (2.3)-(2.9). We assume without loss of generality that $\nabla f_s(M)$ is symmetric for every $M \in \mathbb{R}^{n\times n}$. This is because we can always optimize the equivalent problem

$$\min_{M\in\mathbb{R}^{n\times n}} \ \frac{1}{2}\left[f_s(M) + f_s(M^T)\right] \quad \text{s. t. } \operatorname{rank}(M) \le r, \quad M^T = M, \quad M \succeq 0.$$

We first consider problems (2.3) and (2.4).

Theorem 8 ([145, 95])**.** *The matrix $\tilde{M} = \tilde{U}\tilde{U}^T$ with $\tilde{U} \in \mathbb{R}^{n\times r}$ is a first-order critical point of the constrained problem* (2.3) *if and only if*

$$\begin{cases} \nabla f_s(\tilde{M})\tilde{U} = 0 & \text{if } \operatorname{rank}(\tilde{M}) = r \\ \nabla f_s(\tilde{M}) \succeq 0 & \text{if } \operatorname{rank}(\tilde{M}) < r. \end{cases}$$

The matrix $\tilde{M} = \tilde{U}\tilde{V}^T$ with $\tilde{U} \in \mathbb{R}^{n\times r}$ and $\tilde{V} \in \mathbb{R}^{m\times r}$ is a first-order critical point of the constrained problem (2.4) *if and only if*

$$\begin{cases} [\nabla f_a(\tilde{M})]^T\tilde{U} = 0,\ \nabla f_a(\tilde{M})\tilde{V} = 0 & \text{if } \operatorname{rank}(\tilde{M}) = r \\ \nabla f_a(\tilde{M}) = 0 & \text{if } \operatorname{rank}(\tilde{M}) < r. \end{cases}$$

In [95], the authors proved that each second-order critical point of problem (2.6) or (2.9) is a fixed point of the SVP algorithm run on problem (2.4). We note that this relation can be extended to the symmetric and positive semi-definite case. This relation plays an important role in the analysis of Section 2.3.

Theorem 9 ([95])**.** *The matrix $\tilde{M} = \tilde{U}\tilde{U}^T$ with $\tilde{U} \in \mathbb{R}^{n\times r}$ is a fixed point of the SVP algorithm run on problem* (2.3) *with the step size $1/(1+\delta)$ if and only if*

$$\nabla f_s(\tilde{M})\tilde{U} = 0, \quad -\lambda_{min}(\nabla f_s(\tilde{M})) \le (1+\delta)\sigma_r(\tilde{U}).$$

The matrix $\tilde{M} = \tilde{U}\tilde{V}^T$ with $\tilde{U} \in \mathbb{R}^{n\times r}$ and $\tilde{V} \in \mathbb{R}^{m\times r}$ is a fixed point of the SVP algorithm run on problem (2.4) *with the step size $1/(1+\delta)$ if and only if*

$$[\nabla f_a(\tilde{M})]^T\tilde{U} = 0, \quad \nabla f_a(\tilde{M})\tilde{V} = 0, \quad \|\nabla f_a(\tilde{M})\|_2 \le (1+\delta)\sigma_r(\tilde{M}).$$

Next, we consider problems (2.6)-(2.9). Since the goal is to study only spurious local minima and saddle points, it is enough to focus on the second-order necessary optimality conditions. The following two theorems follow from basic calculations and we omit the proof.

Theorem 10. *The matrix $\tilde{U} \in \mathbb{R}^{n\times r}$ is a second-order critical point of problem* (2.7) *if and only if*

$$\nabla f_s(\tilde{U}\tilde{U}^T)\tilde{U} = 0$$

and

$$2\langle \nabla f_s(\tilde{U}\tilde{U}^T), \Delta\Delta^T\rangle + [\nabla^2 f_s(\tilde{U}\tilde{U}^T)](\tilde{U}\Delta^T + \Delta\tilde{U}^T, \tilde{U}\Delta^T + \Delta\tilde{U}^T) \geq 0$$

holds for every $\Delta \in \mathbb{R}^{n\times r}$.

Theorem 11. *The point $(\tilde{U}, \tilde{V})$ with $\tilde{U} \in \mathbb{R}^{n\times r}$ and $\tilde{V} \in \mathbb{R}^{m\times r}$ is a second-order critical point of problem* (2.6) *if and only if*

$$\nabla[f_a(\tilde{U}\tilde{V}^T)]^T\tilde{U} = 0, \quad \nabla f_a(\tilde{U}\tilde{V}^T)\tilde{V} = 0$$

and

$$2\langle \nabla f_a(\tilde{U}\tilde{V}^T), \Delta_U\Delta_V^T\rangle + [\nabla^2 f_a(\tilde{U}\tilde{V}^T)](\tilde{U}\Delta_V^T + \Delta_U\tilde{V}^T, \tilde{U}\Delta_V^T + \Delta_U\tilde{V}^T) \geq 0$$

holds for every $\Delta_U \in \mathbb{R}^{n\times r}$ and $\Delta_V \in \mathbb{R}^{m\times r}$. Moreover, the given point is a a second-order critical point of problem (2.9) *if and only if*

$$\nabla[f_a(\tilde{U}\tilde{V}^T)]^T\tilde{U} = 0, \quad \nabla f_a(\tilde{U}\tilde{V}^T)\tilde{V} = 0, \quad \tilde{U}^T\tilde{U} = \tilde{V}^T\tilde{V}$$

and

$$\begin{aligned} 2\langle \nabla f_a(\tilde{U}\tilde{V}^T), \Delta_U\Delta_V^T\rangle + [\nabla^2 f_a(\tilde{U}\tilde{V}^T)](\tilde{U}\Delta_V^T + \Delta_U\tilde{V}^T, \tilde{U}\Delta_V^T + \Delta_U\tilde{V}^T) \\ + \frac{\mu}{2}\|\tilde{U}^T\Delta_U + \Delta_U^T\tilde{U} - \tilde{V}^T\Delta_V - \Delta_V^T\tilde{V}\|_F^2 \geq 0 \end{aligned}$$

holds for every $\Delta_U \in \mathbb{R}^{n\times r}$ and $\Delta_V \in \mathbb{R}^{m\times r}$.

2.B Relation between the Symmetric and Asymmetric Problems

In this section, we study the relationship between problems (2.7)-(2.9). This relationship is more general than the topic of this chapter, namely the non-existence of spurious second-order critical points and the strict saddle property, and holds for any property that is characterized by the RIP constant δ and the BDP constant κ. Specifically, we show that

any property that holds for the symmetric problems (2.7) with (δ, κ) also holds for the regularized asymmetric problem (2.9) with another pair of constants $(\tilde{\delta}, \tilde{\kappa})$ decided by δ, κ, and vice versa.

We first consider the transformation from the asymmetric case to the symmetric case. The transformation to the symmetric case has been established in [86] for linear problem. Here, we show that the transformation can be revised and extended to the nonlinear measurements case.

Theorem 12. *Suppose that the function $f_a(\cdot)$ satisfies the δ-RIP$_{2r,2s}$ and the κ-BDP$_{2t}$ properties. If we choose $\mu := (1-\delta)/2$, then problem (2.9) is equivalent to a symmetric problem whose objective function satisfies the $2\delta/(1+\delta)$-RIP$_{2r,2s}$ and the $2\kappa/(1+\delta)$-BDP$_{2t}$ properties.*

Proof of Theorem 12. For any matrix $N \in \mathbb{R}^{(n+m)\times(n+m)}$, we divide the matrix into four blocks as

$$N = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix},$$

where $N_{11} \in \mathbb{R}^{n\times n}, N_{12} \in \mathbb{R}^{n\times m}, N_{22} \in \mathbb{R}^{m\times m}$. Then, we define a new function

$$\tilde{f}(N) := f_a(N_{12}) + f_a(N_{21}^T).$$

We observe that $\tilde{f}(WW^T) = 2h_a(U, V)$, where

$$W := \begin{bmatrix} U \\ V \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}.$$

For any $K \in \mathbb{R}^{(n+m)\times(n+m)}$, the Hessian of $\tilde{f}(\cdot)$ satisfies

$$[\nabla^2 \tilde{f}(N)](K, K) = [\nabla^2 f_a(N_{12})](K_{12}, K_{12}) + [\nabla^2 f_a(N_{21}^T)](K_{21}^T, K_{21}^T). \tag{2.12}$$

Similarly, we can define

$$\tilde{g}(N) := \|N_{11}\|_F^2 + \|N_{22}\|_F^2 - \|N_{12}\|_F^2 - \|N_{21}\|_F^2.$$

We can also verify that $\tilde{g}(WW^T) = g(U, V)$ and

$$[\nabla^2 \tilde{g}(N)](K, K) = 2\left(\|K_{11}\|_F^2 + \|K_{22}\|_F^2 - \|K_{12}\|_F^2 - \|K_{21}\|_F^2\right). \tag{2.13}$$

for every $K \in \mathbb{R}^{(n+m)\times(n+m)}$. The minimization problem (2.9) is then equivalent to

$$\min_{W\in\mathbb{R}^{(n+m)\times r}} F(WW^T) := \tilde{f}(WW^T) + \frac{\mu}{2} \cdot \tilde{g}(WW^T), \tag{2.14}$$

which is in the symmetric form as problem (2.7). For every $N, K \in \mathbb{R}^{(n+m)\times(n+m)}$ with $\operatorname{rank}(N) \le 2r$ and $\operatorname{rank}(K) \le 2s$, it results from relations (2.12) and (2.13) that

$[\nabla^2 F(N)](K, K)$

$$\begin{aligned}&\geq (1-\delta)\left(\|K_{12}\|_F^2+\|K_{21}\|_F^2\right)+\mu\left(\|K_{11}\|_F^2+\|K_{22}\|_F^2-\|K_{12}\|_F^2-\|K_{21}\|_F^2\right)\\&\geq \min\{1-\delta-\mu,\mu\}\cdot\|K\|_F^2\end{aligned}$$

and

$$\begin{aligned}&[\nabla^2 F(N)](K,K)\\&\qquad\leq (1+\delta)\left(\|K_{12}\|_F^2+\|K_{21}\|_F^2\right)+\mu\left(\|K_{11}\|_F^2+\|K_{22}\|_F^2-\|K_{12}\|_F^2-\|K_{21}\|_F^2\right)\\&\qquad\leq \max\{1+\delta-\mu,\mu\}\cdot\|K\|_F^2.\end{aligned}$$

Choosing $\mu := (1-\delta)/2$, we obtain

$$\frac{1-\delta}{2}\cdot\|K\|_F^2\leq[\nabla^2 F(N)](K,K)\leq\frac{1+3\delta}{2}\cdot\|K\|_F^2.$$

Hence, it follows that the function $2F(\cdot)/(1+\delta)$ satisfies the $2\delta/(1+\delta)$-RIP$_{2r,2s}$ property.

Moreover, for every $N,N',K,L\in\mathbb{R}^{(n+m)\times(n+m)}$ with

$$\operatorname{rank}(N),\operatorname{rank}(N'),\operatorname{rank}(K),\operatorname{rank}(L)\leq 2t,$$

it holds that

$$\begin{aligned}[\nabla^2\tilde{g}(N)](K,L)&=[\nabla^2\tilde{g}(N')](K,L)\\&=2\left(\langle K_{11},L_{11}\rangle+\langle K_{22},L_{22}\rangle-\langle K_{12},L_{12}\rangle-\langle K_{21},L_{21}\rangle\right)\end{aligned}$$

and

$$\begin{aligned}&\left|[\nabla^2 F(N)-\nabla^2 F(N')](K,L)\right|\\&\quad=\left|[\nabla^2 f(N_{12})-\nabla^2 f(N'_{12})](K_{12},L_{12})+[\nabla^2 f(N_{21}^T)-\nabla^2 f((N'_{21})^T)](K_{21}^T,L_{21}^T)\right|\\&\quad\leq\kappa\|K_{12}\|_F\|L_{12}\|_F+\kappa\|K_{21}\|_F\|L_{21}\|_F\leq\kappa\|K\|_F\|L\|_F,\end{aligned}$$

which implies that the function $\frac{2}{1+\delta}\cdot F(\cdot)$ satisfies the $2\kappa/(1+\delta)$-BDP$_{2r}$ property. Since problem (2.14) is equivalent to the minimization of $\frac{2}{1+\delta}\cdot F(WW^T)$, it is equivalent to a symmetric problem that satisfies the $2\delta/(1+\delta)$-RIP$_{2r,2s}$ and the $2\kappa/(1+\delta)$-BDP$_{2r}$ properties. □

We can see that both constants δ and κ are approximately doubled in the transformation. As an example, [22] showed that the symmetric linear problem has no spurious local minima if the δ-RIP$_{2r}$ property is satisfied with $\delta<1/5$. Using Theorem 12, we know that the asymmetric linear problem has no spurious local minima if the δ-RIP$_{2r}$ property is satisfied with $\delta<1/9$.

The transformation from a symmetric problem to an asymmetric one is more straightforward. We can equivalently solve the optimization problem

$$\min_{U,V\in\mathbb{R}^{n\times r}}\ f_s\left[\frac{1}{2}\left(UV^T+VU^T\right)\right]\tag{2.15}$$

or its regularized version with any parameter $\mu > 0$. It can be easily shown that the above problem has the same RIP and BDP constants as the original symmetric problem. We omit the proof for brevity.

Theorem 13. *Suppose that the function $f_s(\cdot)$ satisfies the δ-RIP$_{4r,2s}$ and the κ-BDP$_{4t}$ properties. For every $\mu > 0$, problem (2.7) is equivalent to an asymmetric problem and its regularized version with the δ-RIP$_{2r,2s}$ and the κ-BDP$_{2t}$ properties.*

Note that the transformation from a symmetric problem to an asymmetric problem will not increase the constants κ and δ but requires stronger RIP and BDP properties. Hence, a direct analysis on the symmetric case may establish the same property under a weaker condition. In addition to problem (2.15), we can also directly consider the problem $\min_{U,V} f_a(UV^T)$. However, in certain applications, the objective function is only defined for symmetric matrices and we can only use the formulation (2.15) to construct an asymmetric problem. In more restricted cases when the objective function is only defined for symmetric and positive semi-definite matrices, we can only apply the direct analysis to the symmetric case.

2.C Proofs for Section 2.2

Proof of Theorem 3

Proof of Theorem 3. We denote $f(\cdot) := f_s(\cdot)$ and $f(\cdot) := f_a(\cdot)$ for the symmetric and asymmetric case, respectively. Using the mean value theorem and the δ-RIP$_{2r,2r}$ property, there exists a constant $s \in [0, 1]$ such that

$$\begin{aligned} & f(M_{t+1}) - f(M_t) \\ &= \langle \nabla f(M_t), M_{t+1} - M_t \rangle + \frac{1}{2}[\nabla^2 f(M_t + s(M_{t+1} - M_t))](M_{t+1} - M_t, M_{t+1} - M_t) \\ &\leq \langle \nabla f(M_t), M_{t+1} - M_t \rangle + \frac{1+\delta}{2}\|M_{t+1} - M_t\|_F^2. \end{aligned}$$

We define

$$\phi_t(M) := \langle \nabla f(M_t), M - M_t \rangle + \frac{1+\delta}{2}\|M - M_t\|_F^2 = \frac{1+\delta}{2}\|M - \tilde{M}_{t+1}\|_F^2 + \textit{constant},$$

where the last constant term is independent of M. Since the projection is orthogonal, the projected matrix M_{t+1} achieves the minimal value of $\phi_t(M)$ over all matrices on the manifold $\mathcal{M}$. Therefore, we obtain

$$\begin{aligned} f(M_{t+1}) - f(M_t) &\leq \phi_t(M_{t+1}) \leq \phi_t(M^*) \\ &= \langle \nabla f(M_t), M^* - M_t \rangle + \frac{1+\delta}{2}\|M^* - M_t\|_F^2. \end{aligned} \tag{2.16}$$

On the other hand, we can similarly prove that the δ-RIP$_{2r,2r}$ property ensures

$$f(M^*) - f(M_t) \geq \langle \nabla f(M_t), M^* - M_t \rangle + \frac{1-\delta}{2}\|M^* - M_t\|_F^2,$$
$$f(M_t) - f(M^*) \geq \frac{1-\delta}{2}\|M^* - M_t\|_F^2.$$

Substituting the above two inequalities into (2.16), it follows that

$$\begin{aligned} f(M_{t+1}) - f(M_t) &\leq f(M^*) - f(M_t) + \delta\|M^* - M_t\|_F^2 \\ &\leq f(M^*) - f(M_t) + \frac{2\delta}{1-\delta}[f(M_t) - f(M^*)]. \end{aligned} \tag{2.17}$$

Therefore, using the condition that $\delta < 1/3$, we have

$$f(M_{t+1}) - f(M^*) \leq \frac{2\delta}{1-\delta}[f(M_t) - f(M^*)] := \alpha[f(M_t) - f(M^*)],$$

where $\alpha := 2\delta/(1-\delta) < 1$. Combining this single-step bound with the induction method proves the linear convergence of Algorithm 1. □

2.D Proofs for Section 2.3

Proof of Theorem 4

Proof of Theorem 4. We only consider the case when m and n are at least $2r$. In this case, we have $\ell = 2r$. Other cases can be handled similarly. For the notational simplicity, we denote $M^* := M_a^*$ in this proof.

Necessity. We first consider problem (2.6). Suppose that M^* and $\tilde{M}$ are the optimum and a spurious second-order critical point of problem (2.6), respectively. It has been proved in [95] that the spurious second-order critical point $\tilde{M}$ has rank r and is a fixed point of the SVP algorithm with the step size $(1+\delta)^{-1}$. Therefore, the point $\tilde{M}$ should be a minimizer of the projection step of the SVP algorithm. This implies that

$$\|\tilde{M} - [\tilde{M} - (1+\delta)^{-1}\nabla f_a(\tilde{M})]\|_F^2 \leq \|M^* - [\tilde{M} - (1+\delta)^{-1}\nabla f_a(\tilde{M})]\|_F^2,$$

which can be simplified to

$$\langle \nabla f_a(\tilde{M}), \tilde{M} - M^* \rangle \leq \frac{1+\delta}{2}\|\tilde{M} - M^*\|_F^2. \tag{2.18}$$

Let $\mathcal{U}$ and $\mathcal{V}$ denote the subspaces spanned by the columns and rows of $\tilde{M}$ and M^*, respectively. Namely, we have

$$\mathcal{U} := \{\tilde{M}v_1 + M^*v_2 \mid v_1, v_2 \in \mathbb{R}^m\}, \quad \mathcal{V} := \{\tilde{M}^T u_1 + (M^*)^T u_2 \mid u_1, u_2 \in \mathbb{R}^n\}.$$

Since the ranks of both matrices are bounded by r, the dimensions of $\mathcal{U}$ and $\mathcal{V}$ are bounded by $2r$. Therefore, we can find orthogonal matrices $U \in \mathbb{R}^{n\times 2r}$ and $V \in \mathbb{R}^{m\times 2r}$ such that

$$\mathcal{U} \subset \text{range}(U), \quad \mathcal{V} \subset \text{range}(V)$$

and write $\tilde{M}, M^*$ in the form

$$\tilde{M} = U \begin{bmatrix} \Sigma & 0_{r\times r} \\ 0_{r\times r} & 0_{r\times r} \end{bmatrix} V^T, \quad M^* = URV^T,$$

where $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix and $R \in \mathbb{R}^{2r\times 2r}$ has rank at most r. Recalling the first condition in Theorem 11, the column space and the row space of $\nabla f_a(\tilde{M})$ are orthogonal to the column space and the row space of $\tilde{M}$, respectively. Then, the δ-RIP$_{2r,2r}$ property gives

$$\begin{aligned} \exists \alpha \in [1-\delta, 1+\delta] \quad \text{s.t.} \quad & -\langle \nabla f_a(\tilde{M}), M^* \rangle = \langle \nabla f_a(\tilde{M}), \tilde{M} - M^* \rangle \\ & = \int_0^1 [\nabla^2 f_a(M^* + s(\tilde{M} - M^*))](\tilde{M} - M^*, \tilde{M} - M^*)\, ds \\ & = \alpha \|\tilde{M} - M^*\|_F^2 > 0. \end{aligned} \tag{2.19}$$

This means that

$$G := \mathcal{P}_U \nabla f_a(\tilde{M}) \mathcal{P}_V \neq 0,$$

where $\mathcal{P}_U$ and $\mathcal{P}_V$ are the orthogonal projections onto $\mathcal{U}$ and $\mathcal{V}$, respectively. Combining with inequality (2.18), we obtain $\alpha \leq (1+\delta)/2$. By the definition of G, we have

$$\langle \nabla f_a(\tilde{M}), M^* \rangle = \langle G, M^* \rangle.$$

Since both the column space and the row space of G are orthogonal to $\tilde{M}$, the matrix G has the form

$$G = U \begin{bmatrix} 0_{r\times r} & 0_{r\times r} \\ 0_{r\times r} & -\Lambda \end{bmatrix} V^T, \tag{2.20}$$

where $\Lambda \in \mathbb{R}^{r\times r}$. We may assume without loss of generality that $\Lambda_{ii} \geq 0$ for all i; otherwise, one can flip the sign of some of the last r columns of U. By another orthogonal transformation, we may assume without loss of generality that Λ is a diagonal matrix. Then, Theorem 9 gives

$$(1+\delta) \min_{1\leq i\leq r} \Sigma_{ii} = (1+\delta)\sigma_r(\tilde{M}) \geq \|\nabla f_a(\tilde{M})\|_2 \geq \|G\|_2 = \max_{1\leq i\leq (\ell - r)} \Lambda_{ii}. \tag{2.21}$$

In addition, condition (2.19) is equivalent to

$$\langle \Lambda, R_{r+1:2r,r+1:2r} \rangle = \alpha \|\tilde{M} - M^*\|_F^2 = \alpha \left[\text{tr}(\Sigma^2) - 2\langle \Sigma, R_{1:r,1:r} \rangle + \|R\|_F^2 \right]. \tag{2.22}$$

By the Taylor expansion, for every $Z \in \mathbb{R}^{n\times m}$, we have

$$\langle \nabla f_a(\tilde{M}), Z\rangle = \int_0^1 [\nabla^2 f_a(M^* + s(\tilde{M} - M^*))](\tilde{M} - M^*, Z)\, ds = (\tilde{M} - M^*) : \mathcal{H} : Z,$$

where the last expression is the tensor multiplication and $\mathcal{H}$ is the tensor such that

$$K : \mathcal{H} : L = \int_0^1 [\nabla^2 f_a(M^* + s(\tilde{M} - M^*))](K, L)\, ds, \quad \forall K, L \in \mathbb{R}^{n\times m}.$$

We define

$$\tilde{G} := G - \alpha(\tilde{M} - M^*).$$

By the definition of α, we know that $\langle \tilde{G}, \tilde{M} - M^*\rangle = 0$. Furthermore, using the definition of $\mathcal{H}$, we obtain

$$\begin{aligned}(\tilde{M} - M^*) : \mathcal{H} : (\tilde{M} - M^*) &= \alpha\|\tilde{M} - M^*\|_F^2,\\ (\tilde{M} - M^*) : \mathcal{H} : \tilde{G} = \tilde{G} : \mathcal{H} : (\tilde{M} - M^*) &= \|\tilde{G}\|_F^2.\end{aligned}$$

Suppose that

$$\tilde{G} : \mathcal{H} : \tilde{G} = \beta\|\tilde{G}\|_F^2$$

for some $\beta \in [1-\delta, 1+\delta]$. We consider matrices of the form

$$K(t) := t(\tilde{M} - M^*) + \tilde{G}, \quad \forall t \in \mathbb{R}.$$

Since $K(t)$ is a linear combination of $\tilde{M} - M^*$ and G, the column space of $K(t)$ is a subspace of $\mathcal{U}$, and thus $K(t)$ has rank at most $2r$ and the δ-RIP$_{2r,2r}$ property implies

$$(1-\delta)\|K(t)\|_F^2 \le K(t) : \mathcal{H} : K(t) \le (1+\delta)\|K(t)\|_F^2. \tag{2.23}$$

Using the facts that

$$\begin{aligned}\|K(t)\|_F^2 &= \|\tilde{M} - M^*\|_F^2 \cdot t^2 + \|\tilde{G}\|_F^2,\\ K(t) : \mathcal{H} : K(t) &= \alpha\|\tilde{M} - M^*\|_F^2 \cdot t^2 + 2\|\tilde{G}\|_F^2 \cdot t + \beta\|\tilde{G}\|_F^2,\end{aligned}$$

we can write the two inequalities in (2.23) as quadratic inequalities

$$\begin{aligned}&[\alpha - (1-\delta)]\|\tilde{M} - M^*\|_F^2 \cdot t^2 + 2\|\tilde{G}\|_F^2 \cdot t + [\beta - (1-\delta)]\|\tilde{G}\|_F^2 \ge 0,\\ &[(1+\delta) - \alpha]\|\tilde{M} - M^*\|_F^2 \cdot t^2 - 2\|\tilde{G}\|_F^2 \cdot t + [(1+\delta) - \beta]\|\tilde{G}\|_F^2 \ge 0.\end{aligned} \tag{2.24}$$

If $\alpha = 1-\delta$, then we must have $\|\tilde{G}\|_F = 0$ and thus $G = \alpha(\tilde{M} - M^*)$. Equivalently, we have $M^* = \tilde{M} - \alpha^{-1}G$. Since the column and row spaces of $G \neq 0$ are orthogonal to $\tilde{M}$, the rank of M^* is at least $\operatorname{rank}(\tilde{M}) + 1 = r + 1$, which is a contradiction. Since $\alpha \le (1+\delta)/2$, we have $\alpha < 1 + \delta$. Thus, we have proved that

$$1 - \delta < \alpha < 1 + \delta.$$

Checking the condition for quadratic functions to be non-negative, we obtain

$$\|\tilde{G}\|_F^2 \leq [\alpha - (1-\delta)][\beta - (1-\delta)] \cdot \|\tilde{M} - M^*\|_F^2,$$
$$\|\tilde{G}\|_F^2 \leq [(1+\delta) - \alpha][(1+\delta) - \beta] \cdot \|\tilde{M} - M^*\|_F^2.$$

Since

$$\alpha - (1-\delta) > 0, \quad (1+\delta) - \alpha > 0,$$

the above two inequalities are equivalent to

$$\frac{\|\tilde{G}\|_F^2}{\alpha - (1-\delta)} \leq [\beta - (1-\delta)] \cdot \|\tilde{M} - M^*\|_F^2,$$
$$\frac{\|\tilde{G}\|_F^2}{(1+\delta) - \alpha} \leq [(1+\delta) - \beta] \cdot \|\tilde{M} - M^*\|_F^2.$$

Summing up the two inequalities and dividing both sides by 2δ gives rise to

$$\frac{\|\tilde{G}\|_F^2}{\delta^2 - (1-\alpha)^2} \leq \|\tilde{M} - M^*\|_F^2. \tag{2.25}$$

We note that the above condition is also sufficient for the inequalities in (2.24) to hold by choosing $\beta = 2 - \alpha$. Using the relation $\|G\|_F^2 = \|\tilde{G}\|_F^2 + \alpha^2 \|\tilde{M} - M^*\|_F^2$, one can write

$$\mathrm{tr}(\Lambda^2) = \|G\|_F^2 \leq (2\alpha - 1 + \delta^2)\|\tilde{M} - M^*\|_F^2 = \alpha^{-1}(2\alpha - 1 + \delta^2)\langle \Lambda, R_{r+1:2r, r+1:2r}\rangle. \tag{2.26}$$

Now, using the fact that $\mathrm{rank}(M^*) \leq r$, we can write the matrix R as

$$R = \begin{bmatrix} A \\ C \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix}^T = \begin{bmatrix} AB^T & AD^T \\ CB^T & CD^T \end{bmatrix},$$

where $A, B, C, D \in \mathbb{R}^{r \times r}$. Then, conditions (2.22) and (2.26) become

$$\langle \Lambda, CD^T \rangle = \alpha \left[\mathrm{tr}(\Sigma^2) - 2\langle \Sigma, AB^T \rangle + \|AB^T\|_F^2 + \|AD^T\|_F^2 + \|CB^T\|_F^2 + \|CD^T\|_F^2 \right] \tag{2.27}$$

and

$$\mathrm{tr}(\Lambda^2) \leq \alpha^{-1}(2\alpha - 1 + \delta^2) \cdot \langle \Lambda, CD^T \rangle. \tag{2.28}$$

If $\langle \Lambda, CD^T \rangle = 0$, we have

$$\mathrm{tr}(\Sigma^2) - 2\langle \Sigma, AB^T \rangle + \|AB^T\|_F^2 + \|AD^T\|_F^2 + \|CB^T\|_F^2 + \|CD^T\|_F^2 = 0,$$

which implies that

$$AB^T = \Sigma, \quad AD^T = CB^T = CD^T = 0.$$

This contradicts the assumption that $\tilde{M} \neq M^*$. Combining this with conditions (2.21), (2.27) and (2.28), we arrive at the necessity part. For problem (2.9), Lemma 3 in [95] ensures that $\tilde{M}$ is still a fixed point of the SVP algorithm. Recalling the necessary conditions in Theorem 11, we know that the same necessary conditions also hold in this case.

Sufficiency. Now, we study the sufficiency part. We first consider problem (2.6). We choose two orthogonal matrices $U \in \mathbb{R}^{n\times 2r}, V \in \mathbb{R}^{m\times 2r}$ and define

$$\tilde{M} = U \begin{bmatrix} \Sigma & 0_{r\times r} \\ 0_{r\times r} & 0_{r\times r} \end{bmatrix} V^T, \quad M^* := U \left(\begin{bmatrix} A \\ C \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix}^T \right) V^T, \quad G := U \begin{bmatrix} 0_{r\times r} & 0_{r\times r} \\ 0_{r\times r} & -\Lambda \end{bmatrix} V^T.$$

Since $\langle \Lambda, CD^T \rangle \neq 0$, we have $\tilde{M} \neq M^*$. Then, we know that $\mathrm{rank}(\tilde{M}) \leq r$ and $\mathrm{rank}(M^*) \leq r$. We define

$$\tilde{G} := G - \alpha(\tilde{M} - M^*),$$

which satisfies $\langle \tilde{G}, \tilde{M} - M^* \rangle = 0$ by the condition in the second line of (2.10). If $\tilde{G} = 0$, then

$$\begin{bmatrix} 0_{r\times r} & 0_{r\times r} \\ 0_{r\times r} & -\Lambda \end{bmatrix} = \alpha \cdot \begin{bmatrix} \Sigma & 0_{r\times r} \\ 0_{r\times r} & 0_{r\times r} \end{bmatrix} - \alpha \cdot \begin{bmatrix} A \\ C \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix}^T = \alpha \cdot \begin{bmatrix} \Sigma & 0_{r\times r} \\ 0_{r\times r} & 0_{r\times r} \end{bmatrix} - \alpha \cdot \begin{bmatrix} AB^T & 0 \\ 0 & CD^T \end{bmatrix},$$

where the second step is because of $CB^T = 0$ and $AD^T = 0$. The above relation is equivalent to

$$\Sigma = AB^T, \quad \Lambda = \alpha \cdot CD^T.$$

Since $\Sigma \succ 0$, the matrix AB^T has rank r. Noticing that the decomposition of matrix M^* ensures that the rank of M^* is at most r, we have $CD^T = 0$, which is a contradiction to the condition that $\langle CD^T, \Lambda \rangle \neq 0$. Therefore, we have $\tilde{G} \neq 0$. We consider the rank-2 symmetric tensor

$$\begin{aligned} \mathcal{G}_1 :=& \frac{\alpha}{\|\tilde{M} - M^*\|_F^2} \cdot (\tilde{M} - M^*) \otimes (\tilde{M} - M^*) + \frac{2-\alpha}{\|\tilde{G}\|_F^2} \cdot \tilde{G} \otimes \tilde{G} \\ &+ \frac{1}{\|\tilde{M} - M^*\|_F^2} \left[(\tilde{M} - M^*) \otimes \tilde{G} + \tilde{G} \otimes (\tilde{M} - M^*) \right]. \end{aligned}$$

For every matrix $K \in \mathbb{R}^{n\times m}$, we have the decomposition

$$K = t(\tilde{M} - M^*) + s\tilde{G} + \tilde{K}, \quad \langle \tilde{M} - M^*, \tilde{K} \rangle = \langle \tilde{G}, \tilde{K} \rangle = 0,$$

where $t, s \in \mathbb{R}$ are two suitable constants. Then, using the definition of $\mathcal{G}_1$, we have

$$K : \mathcal{G}_1 : K = \alpha \|\tilde{M} - M^*\|_F^2 \cdot t^2 + 2\|\tilde{G}\|_F^2 \cdot ts + (2-\alpha)\|\tilde{G}\|_F^2 \cdot s^2.$$

By the conditions in the third line of (2.10), one can write

$$\|\tilde{G}\|_F^2 \leq [\alpha - (1-\delta)][(1+\delta) - \alpha] \cdot \|\tilde{M} - M^*\|_F^2,$$

which leads to

$$\begin{aligned} &[\alpha - (1-\delta)]\|\tilde{M} - M^*\|_F^2 \cdot t^2 + 2\|\tilde{G}\|_F^2 \cdot ts + [(1+\delta) - \alpha]\|\tilde{G}\|_F^2 \cdot s^2 \geq 0, \\ &[(1+\delta) - \alpha]\|\tilde{M} - M^*\|_F^2 \cdot t^2 - 2\|\tilde{G}\|_F^2 \cdot ts + [\alpha - (1-\delta)]\|\tilde{G}\|_F^2 \cdot s^2 \geq 0. \end{aligned}$$

The above two inequalities are equivalent to

$$(1-\delta)[\|\tilde{M}-M^*\|_F^2\cdot s^2+\|\tilde{G}\|_F^2\cdot t^2]\le K:\mathcal{G}_1:K\le(1+\delta)[\|\tilde{M}-M^*\|_F^2\cdot s^2+\|\tilde{G}\|_F^2\cdot t^2]. \tag{2.29}$$

By restricting to the subspace

$$\mathcal{S}:=\text{span}\{\tilde{M}-M^*,\tilde{G}\}=\{s(\tilde{M}-M^*)+t\tilde{G}\mid s,t\in\mathbb{R}\},$$

the tensor $\mathcal{G}_1$ can be viewed as a 2×2 matrix. Then, inequality (2.29) implies that the matrix has two eigenvalues λ_1 and λ_2 such that

$$1-\delta\le\lambda_1,\lambda_2\le 1+\delta.$$

Therefore, we can rewrite the tensor $\mathcal{G}_1$ restricted to $\mathcal{S}$ as

$$[\mathcal{G}_1]_{\mathcal{S}}=\lambda_1\cdot G_1\otimes G_1+\lambda_2\cdot G_2\otimes G_2,$$

where G_1,G_2 are linear combinations of $\tilde{M}-M^*,\tilde{G}$ and have the unit norm. Since the orthogonal complementary $\mathcal{S}^\perp$ is in the null space of $\mathcal{G}_1$, we have

$$\mathcal{G}_1=[\mathcal{G}_1]_{\mathcal{S}}=\lambda_1\cdot G_1\otimes G_1+\lambda_2\cdot G_2\otimes G_2.$$

Now, we choose matrices $G_3,\dots,G_N$ such that $G_1,\dots,G_N$ form an orthonormal basis of the linear vector space $\mathbb{R}^{n\times m}$, where $N:=nm$. We define another symmetric tensor by

$$\mathcal{H}:=\mathcal{G}_1+\sum_{i=3}^{N}(1+\delta)\cdot G_i\otimes G_i.$$

Then, inequality (2.29) implies that the quadratic form $K:\mathcal{H}:K$ satisfies the δ-RIP$_{2r,2r}$ property.

Therefore, we can choose the Hessian to be the constant tensor $\mathcal{H}$ and define the function $f_a(\cdot)$ as

$$f_a(K):=\frac{1}{2}(K-M^*):\mathcal{H}:(K-M^*),\quad\forall K\in\mathbb{R}^{n\times m}.$$

Combining with the definition of $\mathcal{H}$, we know

$$\nabla f_a(\tilde{M})=\mathcal{H}:(\tilde{M}-M^*)=G,\quad\nabla^2 f_a(\tilde{M})=\mathcal{H}.$$

We choose matrices $\bar{U}\in\mathbb{R}^{n\times r},\bar{V}\in\mathbb{R}^{m\times r}$ such that $\tilde{M}=\bar{U}\bar{V}^T$ and $\bar{U}^T\bar{U}=\bar{V}^T\bar{V}$. By the definitions of $\tilde{M}$ and G, we know that $\tilde{M}$ and G have orthogonal column and row spaces, i.e.,

$$\bar{U}^TG=0,\quad G\bar{V}=0.$$

This means that the first-order optimality conditions are satisfied at the point $(\bar{U}, \bar{V})$. For the second-order necessary optimality conditions, we consider the direction

$$\Delta := \begin{bmatrix} \Delta_U \\ \Delta_V \end{bmatrix} \in \mathbb{R}^{(n+m)\times r}.$$

We consider the decomposition

$$\Delta_U = \mathcal{P}_{\bar{U}}\Delta_U + \mathcal{P}_{\bar{U}}^{\perp}\Delta_U := \Delta_U^1 + \Delta_U^2, \quad \Delta_V = \mathcal{P}_{\bar{V}}\Delta_V + \mathcal{P}_{\bar{V}}^{\perp}\Delta_V := \Delta_V^1 + \Delta_V^2,$$

where $\mathcal{P}_{\bar{U}}, \mathcal{P}_{\bar{V}}$ are the orthogonal projection onto the column space of $\bar{U}, \bar{V}$, respectively. Then, using the conditions in the first line of (2.10), we have

$$\begin{aligned} \langle \nabla f_a(\tilde{M}), \Delta_U \Delta_V^T \rangle = \langle G, \Delta_U \Delta_V^T \rangle = \langle G, \Delta_U^2 (\Delta_V^2)^T \rangle \geq -\|G^T \Delta_U^2\|_F \|\Delta_V^2\|_F \\ \geq -(1+\delta)\sigma_r(\tilde{M})\|\Delta_U^2\|_F \|\Delta_V^2\|_F \geq -(1+\delta)\sigma_r(\tilde{M}) \cdot \frac{\|\Delta_U^2\|_F^2 + \|\Delta_V^2\|_F^2}{2}. \end{aligned} \tag{2.30}$$

We define

$$\Delta_1 := \bar{U}(\Delta_V^1)^T + \Delta_U^1 \bar{V}^T, \quad \Delta_2 := \bar{U}(\Delta_V^2)^T + \Delta_U^2 \bar{V}^T.$$

Then, we know that $\langle \Delta_1, \Delta_2 \rangle = 0$. Using the assumption that $CB^T = AD^T = 0$, we know that M^* has the form

$$M^* = U \begin{bmatrix} AB^T & 0 \\ 0 & CD^T \end{bmatrix} V^T = \mathcal{P}_{\bar{U}} M^* \mathcal{P}_{\bar{V}} + \mathcal{P}_{\bar{U}}^{\perp} M^* \mathcal{P}_{\bar{V}}^{\perp}. \tag{2.31}$$

Then, the special form (2.31) implies that

$$\langle M^*, \Delta_2 \rangle = \langle M^*, \bar{U}(\Delta_V^2)^T + \Delta_U^2 \bar{V}^T \rangle = \left\langle M^*, \bar{U}\Delta_V^T \mathcal{P}_{\bar{V}}^{\perp} + \mathcal{P}_{\bar{U}}^{\perp} \Delta_U \bar{V}^T \right\rangle = 0.$$

Using the definitions of $\tilde{M}$ and G, it can be concluded that

$$\langle \tilde{M}, \Delta_2 \rangle = 0, \quad \langle G, \Delta_2 \rangle = \langle G, \bar{U}(\Delta_V^2)^T + \Delta_U^2 \bar{V}^T \rangle = 0.$$

Since G_1, G_2 are linear combinations of $\tilde{M} - M^*$ and G, the last three relations lead to

$$\langle G_1, \Delta_2 \rangle = \langle G_2, \Delta_2 \rangle = 0.$$

Therefore, there exist constants $a_3, \ldots, a_N$ such that

$$\Delta_2 = \sum_{i=3}^{N} a_i G_i.$$

Suppose that the constants $b_1, \ldots, b_N$ satisfy

$$\Delta_1 = \sum_{i=1}^{N} b_i G_i.$$

Then, the fact $\langle \Delta_1, \Delta_2 \rangle = 0$ and the orthogonality of $G_1, \ldots, G_N$ imply that

$$\sum_{i=3}^{N} a_i b_i = 0.$$

We can calculate that

$$\begin{aligned}
&[\nabla^2 f_a(\tilde{M})](\bar{U}\Delta_V^T + \Delta_U \bar{V}^T, \bar{U}\Delta_V^T + \Delta_U \bar{V}^T) = (\Delta_1 + \Delta_2) : \mathcal{H} : (\Delta_1 + \Delta_2) \\
=&\lambda_1 \cdot b_1^2 + \lambda_2 \cdot b_2^2 + (1+\delta)\sum_{i=3}^{N}(a_i + b_i)^2 \geq (1+\delta)\sum_{i=3}^{N}(a_i + b_i)^2 \\
=&(1+\delta)\sum_{i=3}^{N}\left(a_i^2 + b_i^2\right) \geq (1+\delta)\sum_{i=3}^{N} a_i^2 = (1+\delta)\|\bar{U}(\Delta_V^2)^T + \Delta_U^2 \bar{V}^T\|_F^2,
\end{aligned}$$

where the third last step is due to $\sum_{i=3}^{N} a_i b_i = 0$. Noticing that $\langle \bar{U}(\Delta_V^2)^T, \Delta_U^2 \bar{V}^T \rangle = 0$, the above inequality gives that

$$\begin{aligned}
&[\nabla^2 f_a(\tilde{M})](\bar{U}\Delta_V^T + \Delta_U \bar{V}^T, \bar{U}\Delta_V^T + \Delta_U \bar{V}^T) \geq (1+\delta)\|\bar{U}(\Delta_V^2)^T\|_F^2 + (1+\delta)\|\Delta_U^2 \bar{V}^T\|_F^2 \\
\geq&(1+\delta)\sigma_r(\bar{U})^2\|\Delta_V^2\|_F^2 + (1+\delta)\sigma_r(\bar{V})^2\|\Delta_U^2\|_F^2 = (1+\delta)\sigma_r(\tilde{M})(\|\Delta_V^2\|_F^2 + \|\Delta_U^2\|_F^2),
\end{aligned}$$

where the last equality is because of $\sigma_r(\bar{U})^2 = \sigma_r(\bar{V})^2 = \sigma_r(\tilde{M})$ when $\bar{U}^T\bar{U} = \bar{V}^T\bar{V}$. Combining with inequality (2.30), one can write

$$\begin{aligned}
&[\nabla^2 h_a(U,V)](\Delta, \Delta) = 2\langle \nabla f_a(\tilde{M}), \Delta_U \Delta_V^T \rangle + [\nabla^2 f_a(\tilde{M})](\bar{U}\Delta_V^T + \Delta_U \bar{V}^T, \bar{U}\Delta_V^T + \Delta_U \bar{V}^T) \\
\geq& -(1+\delta)\sigma_r(\tilde{M})(\|\Delta_V^2\|_F^2 + \|\Delta_U^2\|_F^2) + (1+\delta)\sigma_r(\tilde{M})(\|\Delta_V^2\|_F^2 + \|\Delta_U^2\|_F^2) = 0.
\end{aligned}$$

This shows that $(\bar{U}, \bar{V})$ satisfies the second-order necessary optimality conditions, and therefore it is a spurious second-order critical point.

Now, we consider problem (2.9). Since the point $(\bar{U}, \bar{V})$ satisfies $\bar{U}^T\bar{U} = \bar{V}^T\bar{V}$, it is also a local minimum of the regularization term. Hence, the point $(\bar{U}, \bar{V})$ is also a spurious second-order critical point of the regularized problem (2.9). □

Proof of Corollary 1

Proof of Corollary 1. We assume that problem (2.6) has a spurious second-order critical point. By the necessity part of Theorem (2.10), there exist $\alpha \in (1-\delta, 1+\delta)$ and real numbers $\sigma, \lambda, a, b, c, d$ such that

$$\begin{aligned}
&(1+\delta)\sigma \geq \lambda > 0, \quad \alpha^{-1}(2\alpha - 1 + \delta^2)cd \cdot \lambda \geq \lambda^2 > 0, \\
&cd \cdot \lambda = \alpha[\sigma^2 - 2ab \cdot \sigma + (ab)^2 + (ad)^2 + (cb)^2 + (cd)^2].
\end{aligned} \tag{2.32}$$

We first relax the second line to

$$cd \cdot \lambda \geq \alpha[\sigma^2 - 2|ab| \cdot \sigma + (ab)^2 + 2|ab| \cdot |cd| + (cd)^2]. \tag{2.33}$$

Then, we denote $x := |ab|$ and consider the quadratic programming problem

$$\min_{x \geq 0} \; x^2 + 2(|cd| - \sigma) \cdot x,$$

whose optimal value is

$$-(\sigma - |cd|)_+^2,$$

where $(t)_+ := \max\{t, 0\}$. Substituting into inequality (2.33), we obtain

$$cd \cdot \lambda \geq \alpha[\sigma^2 - (\sigma - |cd|)_+^2 + (cd)^2]. \tag{2.34}$$

Then, we consider two different cases.

Case I. We first consider the case when $\sigma \geq |cd|$. In this case, the inequality (2.34) becomes

$$cd \cdot \lambda \geq 2\alpha \cdot \sigma |cd| = 2\alpha \cdot \sigma cd,$$

where the last equality is due to $cd > 0$. Therefore,

$$\lambda \geq 2\alpha \cdot \sigma.$$

The second inequality in (2.32) implies $\lambda \leq \alpha^{-1}(2\alpha - 1 + \delta^2) \cdot cd$. Combining with the above inequality and the assumption of this case, it follows that

$$\alpha^{-1}(2\alpha - 1 + \delta^2) \cdot \sigma \geq \alpha^{-1}(2\alpha - 1 + \delta^2) \cdot cd \geq 2\alpha \cdot \sigma,$$

which is further equivalent to

$$\alpha^{-1}(2\alpha - 1 + \delta^2) \geq 2\alpha \iff \delta^2 \geq 2\alpha^2 - 2\alpha + 1.$$

Since $2\alpha^2 - 2\alpha + 1 \geq 1/2$, we arrive at $\delta^2 \geq 1/2$, which is a contradiction to $\delta < 1/2$.

Case II. We then consider the case when $\sigma \leq |cd|$. In this case, the inequality (2.34) becomes

$$cd \cdot \lambda \geq \alpha[\sigma^2 + (cd)^2].$$

Combining with the second inequality in (2.32), we obtain $\lambda \leq \alpha^{-1}(2\alpha - 1 + \delta^2) \cdot (cd)$. Therefore,

$$\alpha^{-1}(2\alpha - 1 + \delta^2) \cdot (cd)^2 \geq cd \cdot \lambda \geq \alpha[\sigma^2 + (cd)^2].$$

Moreover, the first inequality in (2.32) gives

$$(1 + \delta)\sigma \cdot cd \geq cd \cdot \lambda \geq \alpha[\sigma^2 + (cd)^2].$$

By denoting $y := cd$, the above two inequalities become

$$\begin{aligned} \alpha^{-1}(2\alpha - 1 + \delta^2) \cdot y^2 &\geq \alpha[\sigma^2 + y^2], \\ (1+\delta)\sigma \cdot y &\geq \alpha[\sigma^2 + y^2]. \end{aligned} \tag{2.35}$$

By denoting $z := y/\sigma$, the first inequality in (2.35) implies

$$z^2 \geq \frac{\alpha^2}{\delta^2 - (1-\alpha)^2}. \tag{2.36}$$

Since $\delta < 1/2$, one can write

$$(1-\alpha)^2 + \alpha^2 \geq \frac{1}{2} > \frac{1}{4} > \delta^2,$$

which is equivalent to $\alpha^2 \geq \delta^2 - (1-\alpha)^2$. Therefore, inequality (2.36) implies that $z^2 \geq 1$ and

$$z^2 + \frac{1}{z^2} \geq \frac{\alpha^2}{\delta^2 - (1-\alpha)^2} + \frac{\delta^2 - (1-\alpha)^2}{\alpha^2}. \tag{2.37}$$

On the other hand, the second inequality in (2.35) implies

$$z + \frac{1}{z} \leq \frac{1+\delta}{\alpha} \quad \text{and thus} \quad z^2 + \frac{1}{z^2} + 2 \leq \frac{(1+\delta)^2}{\alpha^2}.$$

Combining with inequality (2.37), it follows that

$$\frac{\alpha^2}{\delta^2 - (1-\alpha)^2} + \frac{\delta^2 - (1-\alpha)^2}{\alpha^2} + 2 \leq \frac{(1+\delta)^2}{\alpha^2}. \tag{2.38}$$

By some calculation, the above inequality is equivalent to

$$(\delta^2 + 2\delta + 5) \cdot \alpha^2 + (2\delta^2 - 4\delta - 6) \cdot \alpha + 2(1+\delta)(1-\delta^2) \leq 0.$$

Checking the discriminant of the above quadratic function, we obtain

$$(2\delta^2 - 4\delta - 6)^2 - 8(\delta^2 + 2\delta + 5)(1+\delta)(1-\delta^2) \geq 0,$$

which is equivalent to

$$4(2\delta - 1)(\delta + 1)^4 \geq 0.$$

However, the above claim contradicts the assumption that $\delta < 1/2$.

In summary, the contradictions in the two cases imply that the condition (2.32) cannot hold, and therefore there does not exist spurious second-order critical points. □

Counterexample for the Rank-one Case

Example 3. Let $e_i \in \mathbb{R}^n$ be the i-th standard basis of $\mathbb{R}^n$. We define the tensor

$$\begin{aligned}\mathcal{H} := &\sum_{i,j=1}^{n}(e_ie_j^T)\otimes(e_ie_j^T) + \frac{1}{2}(e_1e_1^T)\otimes(e_2e_2^T) + \frac{1}{2}(e_2e_2^T)\otimes(e_1e_1^T)\\ &+ \frac{1}{4}\left[(e_1e_2^T)\otimes(e_1e_2^T) + (e_2e_1^T)\otimes(e_2e_1^T)\right] + \frac{1}{4}(e_1e_2^T)\otimes(e_2e_1^T) + \frac{1}{4}(e_2e_1^T)\otimes(e_1e_2^T)\end{aligned}$$

and the objective function

$$f_a(M) := (M - e_1e_1^T) : \mathcal{H} : (M - e_1e_1^T) \quad \forall M \in \mathbb{R}^{n\times n}.$$

The global minimizer of $f_a(\cdot)$ is the rank-1 matrix $M^* := e_1e_1^T$. It has been proved in [247] that the function $f_a(\cdot)$ satisfies the δ-RIP$_{2,2}$ property with $\delta = 1/2$. Moreover, we define

$$U := \frac{1}{\sqrt{2}}e_2, \quad V := U, \quad \tilde{M} := UU^T \neq M^*.$$

It has been proved in [247] that the first-order optimality condition is satisfied. To verify the second-order necessary condition, we can calculate that

$$\begin{aligned}[\nabla^2 h_a(U,U)](\Delta,\Delta) &= 2\langle\nabla f_a(\tilde{M}), \Delta_U\Delta_V^T\rangle + (U\Delta_V^T + \Delta_U U^T) : \mathcal{H} : (U\Delta_V^T + \Delta_U U^T)\\ &= -\frac{3}{2}(\Delta_U)_1(\Delta_V)_1 + \frac{5}{8}\left[(\Delta_U)_1^2 + (\Delta_V)_1^2\right] + \frac{1}{4}(\Delta_U)_1(\Delta_V)_1\\ &\quad + \frac{1}{2}\left[(\Delta_U)_2 + (\Delta_V)_2\right]^2 + \frac{1}{2}\sum_{i=3}^{n}\left[(\Delta_U)_i^2 + (\Delta_V)_i^2\right]\\ &= \frac{5}{8}\left[(\Delta_U)_1 - (\Delta_V)_1\right]^2 + \frac{1}{2}\left[(\Delta_U)_2 + (\Delta_V)_2\right]^2 + \frac{1}{2}\sum_{i=3}^{n}\left[(\Delta_U)_i^2 + (\Delta_V)_i^2\right],\end{aligned}$$

which is non-negative for every $\Delta \in \mathbb{R}^n$. Hence, we conclude that the point $\tilde{M}$ is a spurious second-order critical point of problem (2.6). Moreover, since we choose $V = U$, the point $\tilde{M}$ is a global minimizer of the regularizer $\|U^TU - V^TV\|_F^2$ and thus $\tilde{M}$ is also a spurious second-order critical point of problem (2.9).

Proof of Corollary 2

Proof of Corollary 2. We first consider the case when $\delta \leq 1/3$. We assume that there exists a spurious second-order critical point $\tilde{M}$. Then, by Theorem 4, we know that there exists a constant $\alpha \in (1-\delta, (1+\delta)/2]$. This means that

$$1 - \delta < \frac{1+\delta}{2},$$

which contradicts the assumption that $\delta \leq 1/3$.

Then, we consider the case when $\delta < 1/2$. With no loss of generality, assume that $\tilde{M} \neq M^*$ and $M^* \neq 0$; otherwise, the inequality in this theorem is trivially true. Define

$$m_{11} := \|\Sigma\|_F^2, \quad m_{12} := \langle \Sigma, AB^T \rangle, \quad m_{22} := \|AB^T\|_F^2 + \|AD^T\|_F^2 + \|CB^T\|_F^2 + \|CD^T\|_F^2.$$

By our construction in Theorem 4, we know that

$$m_{11} = \|\tilde{M}\|_F^2, \quad m_{12} = \langle \tilde{M}, M^* \rangle, \quad m_{22} = \|M^*\|_F^2.$$

Therefore, we only need to prove $m_{12} \geq C(\delta) \cdot \sqrt{m_{11} m_{22}}$ for some constant $C(\delta) > 0$. By the analysis in [95], we know that the second-order critical point $\tilde{M}$ must have rank r and thus $m_{11} \neq 0$. The remainder of the proof is split into two steps.

Step I. First, we prove that

$$\frac{(m_{11} + m_{22} - 2m_{12})^2}{m_{11}m_{22} - m_{12}^2} \leq \frac{(1+\delta)^2}{\alpha^2}, \quad \frac{(m_{11} - m_{12})^2}{m_{11}m_{22} - m_{12}^2} \leq \frac{\delta^2 - (1-\alpha)^2}{\alpha^2}. \tag{2.39}$$

We first rule out the case when $m_{11}m_{22} - m_{12}^2 = 0$. In this case, the equality condition of the Cauchy inequality shows that there exists a constant t such that

$$\tilde{M} = tM^*.$$

Since $\tilde{M} \neq 0$, the constant t is not 0. Using the mean value theorem, for any $Z \in \mathbb{R}^{n \times m}$, there exists a constant $c \in [0, 1]$ such that

$$\begin{aligned} \langle \nabla f_a(\tilde{M}), Z \rangle &= \nabla^2 f[M^* + c(\tilde{M} - M^*)](\tilde{M} - M^*, Z) \\ &= \nabla^2 f[M^* + c(\tilde{M} - M^*)][(t-1)M^*, Z]. \end{aligned}$$

The δ-RIP$_{2r,2r}$ property gives

$$\langle \nabla f_a(\tilde{M}), \tilde{M} \rangle = \nabla^2 f[M^* + c(\tilde{M} - M^*)][(t-1)M^*, tM^*] \geq t(t-1)(1-\delta)\|M^*\|_F^2.$$

If $t = 1$, we conclude that $\tilde{M} = M^*$, which contradicts the assumption that $\tilde{M} \neq M^*$. Therefore, it holds that

$$\langle \tilde{M}, \nabla f_a(\tilde{M}) \rangle \neq 0.$$

This contradicts the first-order optimality condition, which states that $\langle \tilde{M}, \nabla f_a(\tilde{M}) \rangle = 0$. Hence, we have proved that inequality (2.39) is well defined. We consider the decomposition

$$\begin{bmatrix} 0 & 0 \\ 0 & \Lambda \end{bmatrix} = c_1 \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} + c_2 \begin{bmatrix} A \\ C \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix}^T + K, \quad \left\langle K, \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \right\rangle = \left\langle K, \begin{bmatrix} A \\ C \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix}^T \right\rangle = 0.$$

Using the conditions in Theorem 4, it follows that

$$\left\langle \begin{bmatrix} 0 & 0 \\ 0 & \Lambda \end{bmatrix}, \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \right\rangle = 0, \quad \left\langle \begin{bmatrix} 0 & 0 \\ 0 & \Lambda \end{bmatrix}, \begin{bmatrix} A \\ C \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix}^T \right\rangle = \alpha(m_{11} - 2m_{12} + m_{22}).$$

The pair of coefficients (c_1, c_2) can be uniquely solved as

$$c_1 = -\alpha \cdot \frac{m_{11} + m_{22} - 2m_{12}}{m_{11}m_{22} - m_{12}^2} \cdot m_{12}, \quad c_2 = \alpha \cdot \frac{m_{11} + m_{22} - 2m_{12}}{m_{11}m_{22} - m_{12}^2} \cdot m_{11}.$$

Using the orthogonality of the decomposition, we have

$$\begin{aligned} \|\Lambda\|_F^2 &\geq \left\| c_1 \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} + c_2 \begin{bmatrix} A \\ C \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix}^T \right\|_F^2 = c_1^2 m_{11} + 2c_1c_2m_{12} + c_2^2 m_{22} \\ &= \alpha^2 \cdot \frac{m_{11}(m_{11} + m_{22} - 2m_{12})^2}{m_{11}m_{22} - m_{12}^2}. \end{aligned} \tag{2.40}$$

Using the last two lines of condition (2.10), one can write

$$\begin{aligned} &\alpha^2 \cdot \frac{m_{11}(m_{11} + m_{22} - 2m_{12})^2}{m_{11}m_{22} - m_{12}^2} \leq \|\Lambda\|_F^2 \\ &\leq (2\alpha - 1 + \delta^2) \left[\operatorname{tr}(\Sigma^2) - 2\langle \Sigma, AB^T \rangle + \|AB^T\|_F^2 + \|AD^T\|_F^2 + \|CB^T\|_F^2 + \|CD^T\|_F^2 \right] \\ &= (2\alpha - 1 + \delta^2)(m_{11} - 2m_{12} + m_{22}). \end{aligned}$$

Simplifying the above inequality, we arrive at the second inequality in (2.39). Now, the first inequality in condition (2.10) implies that

$$\|\Lambda\|_F^2 \leq (1 + \delta)^2 \|\Sigma\|_F^2 = (1 + \delta)^2 m_{11}.$$

Substituting inequality (2.40) into the left-hand side, it follows that

$$\alpha^2 \cdot \frac{m_{11}(m_{11} + m_{22} - 2m_{12})^2}{m_{11}m_{22} - m_{12}^2} \leq (1 + \delta)^2 m_{11},$$

which is equivalent to the first inequality in (2.39).

Step II. Next, we prove the existence of $C(\delta)$. We denote

$$\kappa := \frac{m_{12}}{\sqrt{m_{11}m_{22}}} \in (-1, 1).$$

and

$$C_1 := \frac{\delta^2 - (1 - \alpha)^2}{\alpha^2}, \quad C_2 := \frac{(1 + \delta)^2}{\alpha^2}, \quad t := \sqrt{\frac{m_{11}}{m_{22}}}.$$

Since $\tilde{M} \neq 0$, we have $t > 0$. The inequalities in (2.39) can be written as

$$(t-\kappa)^2 \leq (1-\kappa^2)C_1, \quad (t+1/t-2\kappa)^2 \leq (1-\kappa^2)C_2. \tag{2.41}$$

Using the assumption that $\delta < 1/2$, we can write

$$\delta^2 < \frac{1}{4} < (1-\alpha)^2 + \frac{1}{2}\alpha^2,$$

which leads to

$$C_1 = \frac{\delta^2 - (1-\alpha)^2}{\alpha^2} < \frac{1}{2}.$$

If $\kappa + \sqrt{(1-\kappa^2)C_1} \geq 1$, then

$$|\kappa| \geq \frac{1-C_1}{1+C_1} \geq \frac{1}{3} > 0. \tag{2.42}$$

If $\kappa < 0$, then it holds that

$$\kappa + \sqrt{(1-\kappa^2)C_1} \leq -\frac{1}{3} + \sqrt{\frac{1}{2}} < 1,$$

which contradicts the assumption. Therefore, we have $\kappa \geq 0$ and inequality (2.42) gives $\kappa \geq 1/3$.

Now, we assume that $\kappa + \sqrt{(1-\kappa^2)C_1} \leq 1$. Then, the first inequality in (2.41) gives

$$0 < t \leq \kappa + \sqrt{(1-\kappa^2)C_1} \leq 1,$$

which further leads to

$$t + \frac{1}{t} - 2\kappa \geq -\kappa + \sqrt{(1-\kappa^2)C_1} + \frac{1}{\kappa + \sqrt{(1-\kappa^2)C_1}}.$$

Combining with the second inequality in (2.41), we obtain

$$-\kappa + \sqrt{(1-\kappa^2)C_1} + \frac{1}{\kappa + \sqrt{(1-\kappa^2)C_1}} \leq \sqrt{(1-\kappa^2)C_2}.$$

The above inequality can be simplified to

$$\sqrt{1-\kappa^2}(1 + C_1 - \sqrt{C_1 C_2}) \leq \kappa\sqrt{C_2}.$$

We notice that the inequality $1 + C_1 - \sqrt{C_1 C_2} \leq 0$ is equivalent to inequality (2.38), which cannot hold when $\delta < 1/2$. Therefore, we have $1 + C_1 - \sqrt{C_1 C_2} > 0$ and $\kappa > 0$. Then, the above inequality is equivalent to

$$(1-\kappa^2)(1 + C_1 - \sqrt{C_1 C_2})^2 \leq \kappa^2 \cdot C_2.$$

Therefore, we have

$$\kappa^2 \geq \frac{(1+C_1-\sqrt{C_1C_2})^2}{(1+C_1-\sqrt{C_1C_2})^2+C_2} = 1-\frac{1}{1+\eta^2},$$

where we define

$$\eta := \frac{1+C_1-\sqrt{C_1C_2}}{\sqrt{C_2}}.$$

To prove the existence of $C(\delta)$ such that $\kappa \geq C(\delta) > 0$, we only need to show that η is lower bounded by a positive constant. With δ fixed, η can be viewed as a continuous function of α. Since $\eta = (1-\delta)/(1+\delta) > 0$ when $\alpha = 1-\delta$, the function/parameter η is defined for all α in the compact set $[1-\delta, (1+\delta)/2]$. Combining with the fact that $1+C_1-\sqrt{C_1C_2} > 0$, the function η is positive on a compact set, and thus there exists a positive lower bound $\bar{C}(\delta) > 0$.

In summary, we can define the function

$$C(\delta) := \min\left\{\frac{1}{3}, \bar{C}(\delta)\right\} > 0$$

such that $\kappa \geq C(\delta)$ for every spurious second-order critical point $\tilde{M}$. □

Counterexample for the General Rank Case with Linear Measurements

Example 4. Using the previous rank-1 example, we design a counterexample with linear measurement for the rank-r case. Let $n \geq 2r$ be an integer and $e_i \in \mathbb{R}^n$ be the i-th standard basis of $\mathbb{R}^n$. We define the tensor

$$\begin{aligned}\mathcal{H} :=& \frac{3}{2}\sum_{i,j=1}^{n}(e_ie_j^T)\otimes(e_ie_j^T) + \sum_{i=1}^{r}\Big\{-\frac{1}{2}\left[(e_{2i-1}e_{2i-1}^T)\otimes(e_{2i-1}e_{2i-1}^T)+(e_{2i}e_{2i}^T)\otimes(e_{2i}e_{2i}^T)\right]\\ &+\frac{1}{2}\left[(e_{2i-1}e_{2i-1}^T)\otimes(e_{2i}e_{2i}^T)+(e_{2i}e_{2i}^T)\otimes(e_{2i-1}e_{2i-1}^T)\right]\\ &-\frac{1}{4}\left[(e_{2i-1}e_{2i}^T)\otimes(e_{2i-1}e_{2i}^T)+(e_{2i}e_{2i-1}^T)\otimes(e_{2i}e_{2i-1}^T)\right]\\ &+\frac{1}{4}\left[(e_{2i-1}e_{2i}^T)\otimes(e_{2i}e_{2i-1}^T)+(e_{2i}e_{2i-1}^T)\otimes(e_{2i-1}e_{2i}^T)\right]\Big\}\end{aligned}$$

and the rank-r global minimum

$$U^* := \begin{bmatrix}e_1 & e_3 & \cdots & e_{2r-1}\end{bmatrix}, \quad M^* := U^*(U^*)^T = \sum_{i=1}^{r} e_{2i-1}e_{2i-1}^T.$$

The objective function is defined as

$$f_a(M) := (M-M^*) : \mathcal{H} : (M-M^*) \quad \forall M \in \mathbb{R}^{n\times n}.$$

We can similarly prove that the function $f_a(\cdot)$ satisfies the δ-RIP$_{2r,2r}$ property with $\delta = 1/2$. Moreover, we define

$$\tilde{U} := \frac{1}{\sqrt{2}}\begin{bmatrix} e_2 & e_4 & \cdots & e_{2r}\end{bmatrix}, \quad \tilde{M} := \tilde{U}\tilde{U}^T = \frac{1}{2}\sum_{i=1}^{r} e_{2i}e_{2i}^T \neq M^*.$$

The gradient of $f_a(\cdot)$ at point $\tilde{M}$ is

$$\nabla f_a(\tilde{M}) = -\frac{3}{4}\sum_{i=1}^{r} e_{2i-1}e_{2i-1}^T \in \mathbb{R}^{2r\times 2r}.$$

Since the column and row spaces of the gradient are orthogonal to those of $\tilde{M}$, the first-order optimality condition is satisfied. To verify the second-order necessary condition, we can similarly calculate that

$$\begin{aligned}
&[\nabla^2 h_a(\tilde{U},\tilde{U})](\Delta,\Delta)\\
&\quad =2\langle \nabla f_a(\tilde{M}), \Delta_U\Delta_V^T\rangle + (\tilde{U}\Delta_V^T + \Delta_V\tilde{U}^T) : \mathcal{H} : (\tilde{U}\Delta_V^T + \Delta_U\tilde{U}^T)\\
&\quad = -\frac{3}{2}\sum_{i=1}^{r}\left[\sum_{j=1}^{r}(\Delta_U)_{2i-1,j}\right]\left[\sum_{j=1}^{r}(\Delta_V)_{2i-1,j}\right] + \sum_{i=1}^{r}\Big\{\frac{5}{8}\left[(\Delta_U)_{2i-1,i}^2 + (\Delta_V)_{2i-1,i}^2\right]\\
&\qquad + \frac{1}{4}(\Delta_U)_{2i-1,i}(\Delta_V)_{2i-1,i} + \frac{1}{2}\left[(\Delta_U)_{2i,i} + (\Delta_V)_{2i,i}\right]^2\Big\}\\
&\qquad + \sum_{1\le i,j\le n, i\neq j}\frac{3}{4}\left[(\Delta_U)_{2j,i} + (\Delta_V)_{2i,j}\right]^2 + \sum_{1\le i,j\le n, i\neq j}\frac{3}{4}\left[(\Delta_U)_{2j-1,i}^2 + (\Delta_V)_{2j-1,i}^2\right]\\
&\quad = \sum_{i=1}^{r}\Big\{\frac{5}{8}\left[(\Delta_U)_{2i-1,i} - (\Delta_V)_{2i-1,i}\right]^2 + \frac{1}{2}\left[(\Delta_U)_{2i,i} + (\Delta_V)_{2i,i}\right]^2\Big\}\\
&\qquad + \sum_{1\le i,j\le n, i\neq j}\frac{3}{4}\left[(\Delta_U)_{2j,i} + (\Delta_V)_{2i,j}\right]^2 + \sum_{1\le i,j\le n, i\neq j}\frac{3}{4}\left[(\Delta_U)_{2j-1,i} - (\Delta_V)_{2j-1,i}\right]^2,
\end{aligned}$$

which is non-negative for every $\Delta \in \mathbb{R}^{n\times r}$. Hence, the point $\tilde{M}$ is a spurious second-order critical point of problem (2.6). Moreover, since we choose $\tilde{V} = \tilde{U}$, the point $\tilde{M}$ is a global minimizer of the regularizer $\|\tilde{U}^T\tilde{U} - \tilde{V}^T\tilde{U}\|_F^2$ and thus $\tilde{M}$ is also a spurious second-order critical point of problem (2.9).

2.E Proofs for Section 2.4

Proof of Theorem 6

In this subsection, we use the following notations:

$$M := UV^T,\ M^* := U^*(V^*)^T, \quad W := \begin{bmatrix}U\\V\end{bmatrix},\ W^* := \begin{bmatrix}U^*\\V^*\end{bmatrix}, \quad \hat{W} := \begin{bmatrix}U\\-V\end{bmatrix},\ \hat{W}^* := \begin{bmatrix}U^*\\-V^*\end{bmatrix},$$

where $M^* := M_a^*$ is the global optimum. We always assume that U^* and V^* satisfy $(U^*)^T U^* = (V^*)^T V^*$. When there is no ambiguity about W, we use W^* to denote the minimizer of $\min_{X \in \mathcal{X}^*} \|W - X\|_F$, where $\mathcal{X}^*$ is the set of global minima of problem (2.9). We note that the set $\mathcal{X}^*$ is the trajectory of a global minimum (U^*, V^*) under the orthogonal group:

$$\mathcal{X}^* = \{(U^*R, V^*R) \mid R \in \mathbb{R}^{r \times r}, R^T R = R R^T = I_r\}.$$

Therefore, the set $\mathcal{X}^*$ is a compact set and its minimum can be attained. With this choice, it holds that

$$\operatorname{dist}(W, \mathcal{X}^*) = \|W - W^*\|_F.$$

We first summarize some technical results in the following lemma.

Lemma 1 ([219, 256])**.** *The following statements hold for every $U \in \mathbb{R}^{n \times r}$, $V \in \mathbb{R}^{m \times r}$ and $W \in \mathbb{R}^{(n+m) \times r}$:*

- $4\|M - M^*\|_F^2 \geq \|WW^T - W^*(W^*)^T\|_F^2 - \|U^T U - V^T V\|_F^2$.
- $\|W^*(W^*)^T\|_F^2 = 4\|M^*\|_F^2$.
- *If* $\operatorname{rank}(W^*) = r$ *and* W^* *is the minimizer of* $\min_{X \in \mathcal{X}^*} \|W - X\|_F$*, then* $\|WW^T - W^*(W^*)^T\|_F^2 \geq 2(\sqrt{2} - 1)\sigma_r^2(W^*)\|W - W^*\|_F^2$.
- *If* $\operatorname{rank}(U^*) = r$ *and* U^* *is the minimizer of* $\min_{X \in \mathcal{X}^*} \|U - X\|_F$*, then* $\|UU^T - U^*(U^*)^T\|_F^2 \geq 2(\sqrt{2} - 1)\sigma_r^2(U^*)\|U - U^*\|_F^2$.

The proof of Theorem 6 follows from the following sequence of lemmas. We first identify two cases when the gradient is large. The following lemma proves that an unbalanced solution cannot be a first-order critical point.

Lemma 2. *Given a constant $\epsilon > 0$, if*

$$\|U^T U - V^T V\|_F \geq \epsilon,$$

then

$$\|\nabla \rho(U, V)\|_F \geq \mu(\epsilon/r)^{3/2}.$$

Proof. Using the relationship between the 2-norm and the Frobenius norm, we have

$$\|U^T U - V^T V\|_2 \geq r^{-1}\|U^T U - V^T V\|_F \geq \epsilon/r.$$

Let $q \in \mathbb{R}^r$ be an eigenvector of $U^T U - V^T V$ such that

$$\|q\|_2 = 1, \quad \left|q^T (U^T U - V^T V) q\right| = \|U^T U - V^T V\|_2.$$

We consider the direction

$$\Delta := \hat{W} q q^T.$$

Then, we can calculate that

$$\|\Delta\|_F^2 = \operatorname{tr}\left(\hat{W}qq^Tqq^T\hat{W}^T\right) = \operatorname{tr}\left(q^T\hat{W}^T\hat{W}q\right) = q^T(U^TU + V^TV)q.$$

In addition, we have

$$\begin{aligned}\langle \nabla h_a(U,V), \Delta\rangle &= \left\langle \begin{bmatrix} \nabla f_a(M)V \\ [\nabla f_a(M)]^T U \end{bmatrix}, \begin{bmatrix} Uqq^T \\ -Vqq^T \end{bmatrix} \right\rangle \\ &= \operatorname{tr}\left[V^T[\nabla f_a(M)]^T Uqq^T\right] - \operatorname{tr}\left[U^T\nabla f_a(M)Vqq^T\right] \\ &= q^T\left[V^T[\nabla f_a(M)]^T U\right]q - q^T\left[U^T\nabla f_a(M)V\right]q = 0.\end{aligned}$$

and

$$\begin{aligned}\left|\left\langle \frac{\mu}{4}\nabla g(U,V), \Delta\right\rangle\right| &= \mu\left|\left\langle \hat{W}\hat{W}^TW, Wqq^T\right\rangle\right| \\ &= \mu\left|\operatorname{tr}\left[(U^TU - V^TV)(U^TU + V^TV)qq^T\right]\right| \\ &= \mu\left|q^T(U^TU - V^TV)(U^TU + V^TV)q\right| \\ &= \mu\|U^TU - V^TV\|_2 \cdot q^T(U^TU + V^TV)q \\ &= \mu\|U^TU - V^TV\|_2 \cdot \sqrt{q^T(U^TU + V^TV)q}\cdot \|\Delta\|_F.\end{aligned}$$

Hence, Cauchy's inequality implies that

$$\|\nabla\rho(U,V)\|_F \geq \frac{|\langle \nabla\rho(U,V), \Delta\rangle|}{\|\Delta\|_F} = \mu\|U^TU - V^TV\|_2 \cdot \sqrt{q^T(U^TU + V^TV)q}.$$

Using the fact that

$$q^T(U^TU + V^TV)q \geq \left|q^T(U^TU - V^TV)q\right| = \|U^TU - V^TV\|_2,$$

we obtain

$$\|\nabla\rho(U,V)\|_F \geq \mu\|U^TU - V^TV\|_2^{3/2} \geq \mu(\epsilon/r)^{3/2}.$$

□

The next lemma proves that a solution with large norm cannot be a first-order critical point.

Lemma 3. *Given a constant $\epsilon > 0$, if*

$$\frac{1-\delta}{3} \leq \mu < 1-\delta, \quad \|WW^T\|_F^{3/2} \geq \max\left\{\left(\frac{1+\delta}{1-\mu-\delta}\right)^2 \|W^*(W^*)^T\|_F^{3/2}, \frac{4\sqrt{r}\lambda}{1-\mu-\delta}\right\},$$

then

$$\|\nabla\rho(U,V)\|_F \geq \lambda.$$

Proof. Choosing the direction $\Delta := W$, we can calculate that

$$\langle \nabla \rho(U,V), \Delta \rangle = 2\langle \nabla f_a(UV^T), UV^T \rangle + \mu \|U^T U - V^T V\|_F^2. \tag{2.43}$$

Using the δ-RIP$_{2r,2r}$ property, we have

$$[\nabla^2 f_a(N)](M,M) \geq (1-\delta)\|M\|_F^2, \quad [\nabla^2 f_a(N)](M^*,M) \leq (1+\delta)\|M\|_F \|M^*\|_F,$$

where $N \in \mathbb{R}^{n\times m}$ is every matrix with rank at most $2r$. Then, the first term can be estimated as

$$\begin{aligned}\langle \nabla f_a(UV^T), UV^T \rangle &= \int_0^1 [\nabla^2 f_a(M^* + s(M-M^*)][M-M^*, M]\, ds \\ &\geq (1-\delta)\|M\|_F^2 - (1+\delta)\|M^*\|_F \|M\|_F.\end{aligned}$$

The second term is

$$\mu \|U^T U - V^T V\|_F^2 = \mu \left(\|UU^T\|_F^2 + \|VV^T\|_F^2 \right) - 2\mu \|M\|_F^2.$$

Substituting into equation (2.43), it follows that

$$\begin{aligned}\langle \nabla \rho(U,V), \Delta \rangle &\geq \mu \left(\|UU^T\|_F^2 + \|VV^T\|_F^2 \right) + 2(1-\delta-\mu)\|M\|_F^2 - 2(1+\delta)\|M^*\|_F \|M\|_F \\ &\geq \mu \left(\|UU^T\|_F^2 + \|VV^T\|_F^2 \right) + 2(1-\delta-\mu)\|M\|_F^2 - 2c\|M\|_F^2 - \frac{(1+\delta)^2}{2c}\|M^*\|_F^2 \\ &\geq \min\{\mu, 1-\delta-\mu-c\}\, \|WW^T\|_F^2 - \frac{(1+\delta)^2}{2c}\|M^*\|_F^2,\end{aligned}$$

where $c \in (0, 1-\delta-\mu)$ is a constant to be designed later. Using equality that $(U^*)^T U^* = (V^*)^T V^*$, Lemma 1 gives

$$\|W^*(W^*)^T\|_F^2 = 4\|M^*\|_F^2.$$

As a result,

$$\langle \nabla \rho(U,V), \Delta \rangle \geq \min\{\mu, 1-\delta-\mu-c\}\, \|WW^T\|_F^2 - \frac{(1+\delta)^2}{8c}\|W^*(W^*)^T\|_F^2.$$

Now, choosing

$$c = \frac{1-\delta-\mu}{2}$$

and noticing that $\mu \geq (1-\delta-\mu)/2$, it yields that

$$\langle \nabla \rho(U,V), \Delta \rangle \geq \frac{1-\delta-\mu}{2}\|WW^T\|_F^2 - \frac{(1+\delta)^2}{4(1-\delta-\mu)}\|W^*(W^*)^T\|_F^2. \tag{2.44}$$

On the other hand,

$$\|\Delta\|_F = \|W\|_F \leq \sqrt{r}\|WW^T\|_F^{1/2}.$$

Combining with inequality (2.44) and using the assumption of this lemma, one can write

$$\begin{aligned}\|\nabla\rho(U,V)\|_F &\geq \frac{\langle\nabla\rho(U,V),\Delta\rangle}{\|\Delta\|_F}\\ &\geq \frac{1-\delta-\mu}{2\sqrt{r}}\|WW^T\|_F^{3/2} - \frac{(1+\delta)^2}{4\sqrt{r}(1-\delta-\mu)}\|W^*(W^*)^T\|_F^2\|WW^T\|_F^{-1/2}\\ &\geq \frac{1-\delta-\mu}{2\sqrt{r}}\|WW^T\|_F^{3/2} - \frac{(1+\delta)^2}{4\sqrt{r}(1-\delta-\mu)}\|W^*(W^*)^T\|_F^{3/2}\\ &\geq \frac{1-\delta-\mu}{4\sqrt{r}}\|WW^T\|_F^{3/2} \geq \lambda.\end{aligned}$$

□

Using the above two lemmas, we only need to focus on points such that

$$\|U^TU - V^TV\|_F = o(1), \quad \|WW^T\|_F = O(1).$$

The following lemma proves that if (U,V) is an approximate first-order critical point with a small singular value $\sigma_r(W)$, then the Hessian of the objective function at this point has a negative curvature.

Lemma 4. *Consider positive constants $\alpha, C, \epsilon, \lambda$ such that*

$$\epsilon^2 \leq (\sqrt{2}-1)\sigma_r^2(W^*)\cdot\alpha^2, \quad G > \mu\left(\epsilon + \frac{4H^2}{G^2}\right) + \frac{(1+\delta)H^2}{G^2}, \tag{2.45}$$

where $G := \|\nabla f_a(M)\|_2$ and $H := \lambda + \mu\epsilon C$. If

$$\|U^TU - V^TV\|_F^2 \leq \epsilon^2, \quad \|WW^T\|_F \leq C^2, \quad \|W - W^*\|_F \geq \alpha, \quad \|\nabla\rho(U,V)\|_F \leq \lambda$$

and

$$\sigma_r^2(W) \leq \frac{2}{1+\delta}\left[G - \mu\left(\epsilon + \frac{4H^2}{G^2}\right) - \frac{(1+\delta)H^2}{G^2}\right] - 2\tau \tag{2.46}$$

for some positive constant τ, then it holds that

$$\lambda_{min}(\nabla^2\rho(U,V)) \leq -(1+\delta)\tau.$$

Proof. We choose a singular vector q of W such that

$$\|q\|_2 = 1, \quad \|Wq\|_2 = \sigma_r(W).$$

Since $\|Wq\|_2 = \sqrt{\|Uq\|_2^2 + \|Vq\|_2^2}$, we have

$$\|Uq\|_2^2 + \|Vq\|_2^2 = \sigma_r^2(W).$$

We choose singular vectors u and v such that

$$\|u\|_2 = \|v\|_2 = 1, \quad \|\nabla f_a(M)\|_2 = u^T \nabla f_a(M) v.$$

We define the direction as

$$\Delta_U := -uq^T, \quad \Delta_V := vq^T, \quad \Delta := \begin{bmatrix} \Delta_U \\ \Delta_V \end{bmatrix}, \quad \hat{\Delta} := \begin{bmatrix} \Delta_U \\ -\Delta_V \end{bmatrix}.$$

For the Hessian of $h_a(\cdot,\cdot)$, we can calculate that

$$\langle \nabla f_a(M), \Delta_U \Delta_V^T \rangle = -\|\nabla f_a(M)\|_2 = -G \tag{2.47}$$

and the δ-RIP$_{2r,2r}$ property gives

$$\begin{aligned} [\nabla^2 f_a(M)](\Delta_U V^T + U\Delta_V^T, \Delta_U V^T + U\Delta_V^T) \\ &\le (1+\delta)\|\Delta_U V^T + U\Delta_V^T\|_F^2 = (1+\delta)\| - u(Vq)^T + (Uq)v^T\|_F^2 \\ &= (1+\delta)\left(\|Vq\|_F^2 + \|Uq\|_F^2\right) - 2(1+\delta)[q^T(U^T u)] \cdot [q^T(V^T v)] \\ &\le (1+\delta)\sigma_r^2(W) + 2(1+\delta) \cdot \|U^T u\|_F \|V^T v\|_F. \end{aligned} \tag{2.48}$$

Then, we consider the terms coming from the Hessian of the regularizer. First, we have

$$\begin{aligned} \langle \hat{\Delta}\hat{W}^T, \Delta W^T \rangle &\le \|U^T U - V^T V\|_F \cdot \|\Delta_U^T \Delta_U - \Delta_V^T \Delta_V\|_F \\ &\le \epsilon \cdot \left[\|\Delta_U^T \Delta_U\|_F + \|\Delta_V^T \Delta_V\|_F\right] = 2\epsilon. \end{aligned} \tag{2.49}$$

Next, we can estimate that

$$\begin{aligned} \langle \hat{W}\hat{\Delta}^T, \Delta W^T \rangle + \langle \hat{W}\hat{W}^T, \Delta\Delta^T \rangle &= \frac{1}{2}\|U^T\Delta_U + \Delta_U^T U - V^T\Delta_V - \Delta_V^T V\|_F^2 \\ &\le 4\left(\|U^T\Delta_U\|_F^2 + \|V^T\Delta_V\|_F^2\right) \\ &= 4\left(\|(U^T u)q^T\|_F^2 + \|(V^T v)q^T\|_F^2\right) \\ &= 4\left(\|U^T u\|_F^2 + \|V^T v\|_F^2\right). \end{aligned} \tag{2.50}$$

Using the assumption that $\|WW^T\|_F \le C^2$ and $\|U^T U - V^T V\|_F^2 \le \epsilon^2$, one can write

$$\|\hat{W}\hat{W}^T W\|_F^2 \le \|U^T U - V^T V\|_F^2 \cdot \|U^T U + V^T V\|_F \le \epsilon^2 \|WW^T\|_F \le \epsilon^2 C^2$$

and

$$\left\| \begin{bmatrix} \nabla f_a(UV^T)V \\ \nabla f_a(UV^T)^T U \end{bmatrix} \right\|_F = \|\nabla \rho(U,V) - \mu\hat{W}\hat{W}^T W\|_F \le \lambda + \mu\epsilon C = H. \tag{2.51}$$

The second relation implies that

$$\|\nabla f_a(UV^T)V\|_2 \le \|\nabla f_a(UV^T)V\|_F \le H, \quad \|U^T \nabla f_a(UV^T)\|_2 \le \|U^T \nabla f_a(UV^T)\|_F \le H. \tag{2.52}$$

By the definition of u and v, it holds that

$$\|v\|_2 = 1, \quad \|\nabla f_a(M)\|_2 u = \nabla f_a(M) v.$$

Therefore,

$$\|U^T u\|_F^2 = \frac{\|U^T \nabla f_a(M) v\|_F^2}{\|\nabla f_a(M)\|_2^2} \leq \frac{\|U^T \nabla f_a(M)\|_F^2 \|v\|_2^2}{\|\nabla f_a(M)\|_2^2} \leq \frac{H^2}{G^2}.$$

Similarly,

$$\|V^T v\|_F^2 \leq \frac{H^2}{G^2}.$$

Substituting into (2.48) and (2.50) yields that

$$[\nabla^2 f_a(M)](\Delta_U V^T + U\Delta_V^T, \Delta_U V^T + U\Delta_V^T) \leq (1+\delta)\sigma_r^2(W) + 2(1+\delta) \cdot \frac{H^2}{G^2} \tag{2.53}$$

and

$$\langle \hat{W}\hat{\Delta}^T, \Delta W^T \rangle + \langle \hat{W}\hat{W}^T, \Delta\Delta^T \rangle \leq 8 \cdot \frac{H^2}{G^2}. \tag{2.54}$$

Combining (2.47), (2.49), (2.53) and (2.54), it follows that

$$[\nabla^2 \rho(U,V)](\Delta,\Delta) \leq -2G + (1+\delta)\sigma_r^2(W) + 2\mu\epsilon + [8\mu + 2(1+\delta)] \cdot \frac{H^2}{G^2}.$$

Since $\|\Delta\|_F^2 = 2$, the above relation implies

$$\lambda_{min}(\nabla^2 \rho(U,V)) \leq -G + \frac{1+\delta}{2}\sigma_r^2(W) + \mu\epsilon + (4\mu + 1 + \delta) \cdot \frac{H^2}{G^2} \leq -(1+\delta)\tau.$$

□

Remark 2. The positive constants ϵ and λ in the proof of Lemma 4 can be chosen to be arbitrarily small with α, C fixed. Hence, we may choose small enough ϵ and λ such that the assumptions given in inequality (2.45) are satisfied. This lemma resolves the case when the minimal singular value $\sigma_r^2(W)$ is on the order of $\|\nabla f_a(M)\|_2/(2+2\delta)$. In the next lemma, we will show that this is the only case when $\delta < 1/3$.

The final step is to prove that condition (2.46) always holds provided that $\delta < 1/3$ and $\epsilon, \lambda, \tau = o(1)$.

Lemma 5. *Given positive constants $\alpha, C, \epsilon, \lambda$, if*

$$\|U^T U - V^T V\|_F^2 \leq \epsilon^2, \quad \max\{\|WW^T\|_F, \|W^*(W^*)^T\|_F\} \leq C^2,$$
$$\|W - W^*\|_F \geq \alpha, \quad \|\nabla \rho(U,V)\|_F \leq \lambda, \quad \delta < 1/3,$$

then the inequality $G \geq c\alpha$ holds for some constant $c > 0$ independent of $\alpha, \epsilon, \lambda, C$. Furthermore, there exist two positive constants

$$\epsilon_0(\delta, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C), \quad \lambda_0(\delta, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C)$$

such that

$$\sigma_r^2(W) \leq \frac{2}{1+\delta}\left[G - \mu\left(2\epsilon + \frac{4H^2}{G^2}\right) - \frac{(1+\delta)H^2}{G^2}\right] \tag{2.55}$$

whenever

$$\begin{aligned} 0 <&\epsilon \leq \epsilon_0(\delta, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C), \\ 0 <&\lambda \leq \lambda_0(\delta, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C). \end{aligned}$$

Here, G and H are defined in Lemma 4.

Proof. We first prove the existence of the constant c. Using Lemma 1, one can write

$$4\|M - M^*\|_F^2 \geq \|WW^T - W^*(W^*)^T\|_F^2 - \|U^TU - V^TV\|_F^2 \geq \|WW^T - W^*(W^*)^T\|_F^2 - \epsilon^2.$$

Using Lemma 1 and the assumption that $\|W - W^*\|_F \geq \alpha$, we have

$$\|M - M^*\|_F^2 \geq \frac{\sqrt{2}-1}{2}\sigma_r^2(W^*)\|W - W^*\|_F^2 - \frac{\epsilon^2}{4} \geq \frac{\sqrt{2}-1}{2}\sigma_r^2(W^*)\cdot\alpha^2 - \frac{\epsilon^2}{4}. \tag{2.56}$$

By the definition of ϵ, it follows that

$$\|M - M^*\|_F^2 \geq \frac{\sqrt{2}-1}{4}\sigma_r^2(W^*)\cdot\alpha^2 > 0.$$

Thus, the δ-RIP$_{2r,2r}$ property gives

$$\|\nabla f_a(M)\|_F \geq \frac{\langle \nabla f_a(M), M - M^*\rangle}{\|M - M^*\|_F} \geq (1-\delta)\|M - M^*\|_F \geq \sqrt{\frac{\sqrt{2}-1}{4}}\cdot\sigma_r(W^*)(1-\delta)\cdot\alpha.$$

Hence, we have

$$G = \|\nabla f_a(M)\|_2 \geq \sqrt{\frac{\sqrt{2}-1}{4r}}\cdot\sigma_r(W^*)(1-\delta)\cdot\alpha = c\alpha,$$

where we define

$$c := \sqrt{\frac{\sqrt{2}-1}{4r}}\cdot\sigma_r(W^*)(1-\delta).$$

Next, we prove inequality (2.55) by contradiction, i.e., we assume

$$\sigma_r^2(W) > \frac{2}{1+\delta}\left[G - \mu\left(2\epsilon + \frac{4H^2}{G^2}\right) - \frac{(1+\delta)H^2}{G^2}\right] \geq \frac{2c\alpha}{1+\delta} + \text{poly}(\epsilon, \lambda). \tag{2.57}$$

The remainder of the proof is divided into three steps.

Step I. We first develop a lower bound for $\sigma_r(M)$. We choose a vector $p \in \mathbb{R}^r$ such that

$$\|p\|_F = 1, \quad U^T U p = \sigma_r^2(U) \cdot p.$$

It can be shown that

$$\begin{aligned}\|(Wp)^T W\|_F = \|p^T U^T U + p^T V^T V\|_F &\leq 2\|p^T U^T U\|_F + \|p^T (V^T V - U^T U)\|_F \\ &\leq 2\sigma_r^2(U) + \|p^T\|_F \|V^T V - U^T U\|_F \leq 2\sigma_r^2(U) + \epsilon.\end{aligned}$$

On the other hand, since W has rank r, it holds that

$$\left\|(Wp)^T W\right\|_F \geq \sigma_r^2(W) \cdot \|p\|_F = \sigma_r^2(W).$$

Combining the above two estimates, we arrive at

$$2\sigma_r^2(U) \geq \sigma_r^2(W) - \epsilon > 0,$$

where the last inequality is from the assumption that ϵ, λ are small and $\sigma_r(W)$ is lower bounded by a positive value in (2.57). Using the inequality that $\sqrt{1-x} \geq 1 - x$ for every $x \in [0, 1]$, the above inequality implies that

$$\sigma_r(U) \geq \frac{1}{\sqrt{2}} \sigma_r(W) \cdot \sqrt{1 - \frac{\epsilon}{\sigma_r^2(W)}} \geq \frac{1}{\sqrt{2}} \sigma_r(W) - \frac{\epsilon}{\sqrt{2}\sigma_r(W)}. \tag{2.58}$$

Similarly, one can prove that

$$\sigma_r(V) \geq \frac{1}{\sqrt{2}} \sigma_r(W) - \frac{\epsilon}{\sqrt{2}\sigma_r(W)}.$$

When ϵ is small enough, we know that $\sigma_r(U), \sigma_r(V) \neq 0$ and both U, V have rank r. To lower bound the singular value $\sigma_r(M)$, we consider vectors x such that $\|x\|_2 = 1$ and lower bound $x^T V (U^T U) V^T x$. Since the range of $V(U^T U)V^T$ is a subspace of the range of V and the range of V has exactly dimension r, directions x that are in the orthogonal complement of the range of V correspond to exactly $m - r$ zero singular values. Hence, to estimate the r-th largest singular value of M, we only need to consider directions that are in the range of V. Namely, we only consider directions that have the form $x = Vy$ for some vector y. Then, we have

$$\begin{aligned}x^T V (U^T U) V^T x &= y^T (V^T V)(U^T U)(V^T V) y \\ &= y^T (V^T V)^3 y + y^T (V^T V)(U^T U - V^T V)(V^T V) y.\end{aligned}$$

First, we bound the second term by calculating that

$$\begin{aligned}\|V(V^T V - U^T U)V^T\|_2 \leq \|V\|_2^2 \|U^T U - V^T V\|_2 &\leq \|V^T V\|_F \|U^T U - V^T V\|_F \\ &\leq \|W^T W\|_F \|U^T U - V^T V\|_F \leq C^2 \epsilon.\end{aligned}$$

This implies that

$$y^T(V^TV)(U^TU - V^TV)(V^TV)y \geq -C^2\epsilon \cdot \|Vy\|_F^2.$$

Next, we assume that y has the decomposition

$$y = \sum_{i=1}^{r} c_i v_i,$$

where v_i is an eigenvector of V^TV associated with the eigenvalue $\sigma_i^2(V)$. Then, we can calculate that

$$y^T(V^TV)^3y = \sum_{i=1}^{r} c_i^2\sigma_i^6(V), \quad \|Vy\|_F^2 = \sum_{i=1}^{r} c_i^2\sigma_i^2(V) = 1.$$

Combining the above estimates leads to

$$\begin{aligned} x^TV(U^TU)V^Tx &\geq \left[\frac{\sum_{i=1}^r c_i^2\sigma_i^6(V)}{\sum_{i=1}^r c_i^2\sigma_i^2(V)} - C^2\epsilon\right] \cdot \|Vy\|_F^2 \\ &= \frac{\sum_{i=1}^r c_i^2\sigma_i^6(V)}{\sum_{i=1}^r c_i^2\sigma_i^2(V)} - C^2\epsilon \geq \sigma_r^4(V) - C^2\epsilon. \end{aligned}$$

This implies that

$$\begin{aligned} \sigma_r^2(M) \geq \sigma_r^4(V) - C^2\epsilon &\geq \left[\frac{1}{\sqrt{2}}\sigma_r(W) - \frac{\epsilon}{\sqrt{2}\sigma_r(W)}\right]^4 - C^2\epsilon \\ &\geq \frac{1}{4}\sigma_r^4(W) - \sigma_r^2(W)\epsilon - \sigma_r^{-2}(W)\epsilon^3 - C^2\epsilon \\ &\geq \frac{1}{4}\sigma_r^4(W) - \sigma_r^{-2}(W)\epsilon^3 - 2C^2\epsilon \\ &\geq \frac{1}{4}\sigma_r^4(W) - \frac{1+\delta}{G}\cdot\epsilon^3 - 2C^2\epsilon \\ &\geq \frac{1}{4}\sigma_r^4(W) - \frac{1+\delta}{c\alpha}\cdot\epsilon^3 - 2C^2\epsilon. \end{aligned} \tag{2.59}$$

where the second last inequality is due to (2.57) and the assumption that ϵ and λ are sufficiently small.

Step II. Next, we derive an upper bound for $\sigma_r(M)$. We define

$$\bar{M} := \mathcal{P}_r\left[M - \frac{1}{1+\delta}\nabla f_a(M)\right],$$

where $\mathcal{P}_r$ is the orthogonal projection onto the low-rank set via SVD. Since $M \neq M^*$ and $\delta < 1/3$, we recall that inequality (2.17) gives

$$\begin{aligned} -\phi(\bar{M}) &\geq \frac{1-3\delta}{1-\delta}[f_a(M) - f_a(M^*)] \geq \frac{1-3\delta}{2}\|M - M^*\|_F^2 \\ &\geq \frac{1-3\delta}{2}\left[\frac{\sqrt{2}-1}{2}\sigma_r^2(W^*)\alpha^2 - \frac{\epsilon^2}{4}\right] := K, \end{aligned}$$

where the second inequality follows from (2.56) and

$$-\phi(\bar{M}) = \langle \nabla f_a(M), M - \bar{M}\rangle - \frac{1+\delta}{2}\|M - \bar{M}\|_F^2.$$

Hence,

$$\langle \nabla f_a(M), M - \bar{M}\rangle - \frac{1+\delta}{2}\|M - \bar{M}\|_F^2 \geq K. \tag{2.60}$$

When we choose ϵ to be small enough, it holds that $K > 0$. For simplicity, we define

$$N := -\frac{1}{1+\delta}\nabla f_a(M).$$

Then, $\bar{M} = \mathcal{P}_r(M + N)$ and the left-hand side of (2.60) is equal to

$$\begin{aligned} &\langle \nabla f_a(M), M - \bar{M}\rangle - \frac{1+\delta}{2}\|M - \bar{M}\|_F^2 \\ &= (1+\delta)\langle N, \mathcal{P}_r(M+N) - M\rangle - \frac{1+\delta}{2}\|\mathcal{P}_r(M+N) - M\|_F^2 \\ &= \frac{1+\delta}{2}\left[\|N\|_F^2 - \|N + M - \mathcal{P}_r(M+N)\|_F^2\right] \\ &= \frac{1+\delta}{2}\left[\|N\|_F^2 - \|N+M\|_F^2 + \|\mathcal{P}_r(M+N)\|_F^2\right]. \end{aligned} \tag{2.61}$$

Similar to the proof of inequality (2.52), we can prove that

$$\|NV\|_F \leq \tilde{H} := \frac{H}{1+\delta}, \quad \|U^T N\|_F \leq \tilde{H}.$$

Then, we have

$$-\operatorname{tr}[N^T(UV^T)] \leq \|U^T N\|_F \|V\|_F \leq \tilde{H} \cdot \|W\|_F \leq \tilde{H} \cdot \sqrt{\sqrt{r}\|WW^T\|_F} \leq \sqrt[4]{r}C \cdot \tilde{H}.$$

Using the above relation, we obtain

$$\|N\|_F^2 - \|N+M\|_F^2 = -2\operatorname{tr}[N^T(UV^T)] - \|M\|_F^2 \leq 2\sqrt[4]{r}C \cdot \tilde{H} - \|M\|_F^2.$$

Suppose that $\mathcal{P}_U$ and $\mathcal{P}_V$ are the orthogonal projections onto the column spaces of U and V, respectively. We define

$$N_1 := \mathcal{P}_U N \mathcal{P}_V,\ N_2 := \mathcal{P}_U N(I-\mathcal{P}_V),\ N_3 := (I-\mathcal{P}_U)N\mathcal{P}_V,\ N_4 := (I-\mathcal{P}_U)N(I-\mathcal{P}_V).$$

Then, recalling the assumption (2.57) and inequality (2.58), it follows that

$$\begin{aligned}\|N_1\|_F = \|\mathcal{P}_U N \mathcal{P}_V\|_F &\le \sigma_r^{-1}(U)\|U^T\mathcal{P}_U N\mathcal{P}_V\|_F \le \sigma_r^{-1}(U)\|U^T N\|_F \le \frac{\sqrt{2}\sigma_r(W)}{\sigma_r^2(W)-\epsilon}\cdot\tilde{H}\\ &\le \left[\sqrt{\frac{1+\delta}{G}}+\mathrm{poly}(\epsilon,\lambda)\right]\cdot\tilde{H} \le \left[\sqrt{\frac{1+\delta}{c\alpha}}+\mathrm{poly}(\epsilon,\lambda)\right]\cdot\tilde{H} := \kappa\tilde{H}.\end{aligned}$$

Similarly, we can prove that

$$\|N_1+N_2\|_F = \|\mathcal{P}_U N\|_F \le \kappa\tilde{H}, \quad \|N_1+N_3\|_F = \|N\mathcal{P}_V\|_F \le \kappa\tilde{H},$$

which leads to

$$\|N_2\|_F \le 2\kappa\tilde{H}, \quad \|N_3\|_F \le 2\kappa\tilde{H}.$$

Using Weyl's theorem, the following holds for every $1 \le i \le r$:

$$|\sigma_i(M+N)-\sigma_i(M+N_4)| \le \|N_1+N_2+N_3\|_2 \le \|N_1+N_2+N_3\|_F \le 3\kappa\tilde{H}.$$

Therefore, we have

$$\begin{aligned}\|\mathcal{P}_r(M+N)\|_F^2 &= \sum_{i=1}^{r}\sigma_i^2(M+N)\\ &\ge \sum_{i=1}^{r}\sigma_i^2(M+N_4) - r\cdot 3\kappa\tilde{H}\cdot(\|M+N\|_2+\|M+N_4\|_2)\\ &\ge \sum_{i=1}^{r}\sigma_i^2(M+N_4) - 6r\kappa\tilde{H}\cdot(\|M\|_2+\|N\|_2)\\ &\ge \sum_{i=1}^{r}\sigma_i^2(M+N_4) - 6r\kappa\tilde{H}\cdot\left(\|M\|_F+\frac{G}{1+\delta}\right). \qquad (2.62)\end{aligned}$$

Using the assumption (2.57) and the inequality (2.59), one can write

$$\frac{G}{1+\delta} \le \frac{\sigma_r^2(W)}{2}+\mathrm{poly}(\sqrt{\epsilon},\lambda) \le \sigma_r(M)+\mathrm{poly}(\sqrt{\epsilon},\lambda) \le \|M\|_F+\mathrm{poly}(\sqrt{\epsilon},\lambda), \qquad (2.63)$$

where $\mathrm{poly}(\sqrt{\epsilon},\lambda)$ means a polynomial of $\sqrt{\epsilon}$ and λ. Therefore, we attain the bound

$$\|M\|_F+\|N\|_F \le 2\|M\|_F+\mathrm{poly}(\sqrt{\epsilon},\lambda) \le 2\cdot\frac{\|WW^T\|_F}{\sqrt{2}}+\mathrm{poly}(\sqrt{\epsilon},\lambda)$$

$$\leq \sqrt{2}C^2 + \text{poly}(\sqrt{\epsilon}, \lambda). \tag{2.64}$$

Substituting back into the previous estimate (2.62), it follows that

$$\|\mathcal{P}_r(M+N)\|_F^2 \geq \sum_{i=1}^{r} \sigma_i^2(M+N_4) - 6\sqrt{2}r\kappa\tilde{H}C^2 + \text{poly}(\sqrt{\epsilon}, \lambda) = \sum_{i=1}^{r} \sigma_i^2(M+N_4) + \text{poly}(\sqrt{\epsilon}, \lambda).$$

Now, since M and N_4 have orthogonal column and row spaces, the maximal r singular values of $M + N_4$ are simply the maximal r singular values of the singular values M and N_4, which we assume to be

$$\sigma_i(M),\ i = 1, \ldots, k \quad \text{and} \quad \sigma_i(N_4),\ i = 1, \ldots, r-k.$$

Now, it follows from (2.61) that

$$\begin{aligned}
&\frac{2}{1+\delta}\left[\langle \nabla f_a(M), M - \bar{M}\rangle - \frac{1+\delta}{2}\|M - \bar{M}\|_F^2\right] \\
&\quad = \|N\|_F^2 - \|N + M\|_F^2 + \|\mathcal{P}_r(M+N)\|_F^2 \\
&\quad \leq -\sum_{i=1}^{r} \sigma_i^2(M) + \sum_{i=1}^{k} \sigma_i^2(M) + \sum_{i=1}^{r-k} \sigma_i^2(N_4) + \text{poly}(\sqrt{\epsilon}, \lambda) + 2\sqrt[4]{r}C \cdot \tilde{H} \\
&\quad = -\sum_{i=k+1}^{r} \sigma_i^2(M) + \sum_{i=1}^{r-k} \sigma_i^2(N_4) + \text{poly}(\sqrt{\epsilon}, \lambda) \\
&\quad \leq -(r-k)\sigma_r^2(M) + (r-k)\|N_4\|_2^2 + \text{poly}(\sqrt{\epsilon}, \lambda) \\
&\quad \leq -(r-k)\sigma_r^2(M) + (r-k)\|N\|_2^2 + \text{poly}(\sqrt{\epsilon}, \lambda).
\end{aligned}$$

If $k = r$, then the above inequality and inequality (2.60) imply that

$$\text{poly}(\sqrt{\epsilon}, \lambda) \geq K = O(\alpha^2),$$

which contradicts the assumption that ϵ and λ are small. Hence, it can be concluded that $r - k \geq 1$. Combining with (2.60), we obtain the upper bound

$$\begin{aligned}
\sigma_r^2(M) &\leq -\frac{2}{1+\delta} \cdot \frac{K}{r-k} + \|N\|_2^2 + \frac{1}{r-k} \cdot \text{poly}(\sqrt{\epsilon}, \lambda) \\
&= -\frac{2}{1+\delta} \cdot \frac{K}{r} + \|N\|_2^2 + \text{poly}(\sqrt{\epsilon}, \lambda).
\end{aligned} \tag{2.65}$$

Step III. In the last step, we combine the inequalities (2.59) and (2.65), which leads to

$$\frac{1}{4}\sigma_r^4(W) - \frac{1+\delta}{c\alpha} \cdot \epsilon^3 - 2C^2\epsilon \leq -\frac{2}{1+\delta} \cdot \frac{K}{r} + \frac{1}{(1+\delta)^2}G^2 + \text{poly}(\sqrt{\epsilon}, \lambda).$$

This means that

$$\sigma_r^4(W) + \frac{8}{1+\delta} \cdot \frac{K}{r} \leq \frac{4}{(1+\delta)^2} G^2 + \text{poly}(\sqrt{\epsilon}, \lambda).$$

Since $K > 0$ has lower bounds that are independent of ϵ and λ, we can choose ϵ and λ to be small enough such that

$$\sigma_r^4(W) + \frac{4}{1+\delta} \cdot \frac{K}{r} \leq \frac{4}{(1+\delta)^2} G^2.$$

However, recalling the assumption (2.57), we have

$$\begin{aligned}
\sigma_r^4(W) &> \frac{4}{(1+\delta)^2} \left[G - \mu \left(2\epsilon + \frac{4H^2}{G^2} \right) - \frac{(1+\delta)H^2}{G^2} \right]^2 \\
&\geq \frac{4}{(1+\delta)^2} G^2 - \frac{16}{(1+\delta)^2} G \cdot \mu\epsilon + \text{poly}(\sqrt{\epsilon}, \lambda) \\
&\geq \frac{4}{(1+\delta)^2} G^2 - \frac{16}{(1+\delta)^2} \mu\epsilon \cdot \frac{1}{\sqrt{2}} (1+\delta) C^2 + \text{poly}(\sqrt{\epsilon}, \lambda) \\
&= \frac{4}{(1+\delta)^2} G^2 + \text{poly}(\sqrt{\epsilon}, \lambda),
\end{aligned}$$

where in the third inequality we use inequalities (2.63)-(2.64) to conclude that

$$G \leq (1+\delta)\|M\|_F + \text{poly}(\sqrt{\epsilon}, \lambda) \leq \frac{1}{\sqrt{2}} (1+\delta) C^2 + \text{poly}(\sqrt{\epsilon}, \lambda).$$

The above two inequalities cannot hold simultaneously when λ and ϵ are small enough. This contradiction means that the condition (2.55) holds by choosing

$$\begin{aligned}
0 < &\epsilon \leq \epsilon_0(\delta, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C), \\
0 < &\lambda \leq \lambda_0(\delta, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C),
\end{aligned}$$

for some small enough positive constants

$$\epsilon_0(\delta, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C), \quad \lambda_0(\delta, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C).$$

□

The only thing left is to piecing everything together.

Proof of Theorem 6. We first choose

$$C := \left[\left(\frac{1+\delta}{1-\mu-\delta} \right)^2 \|W^*(W^*)^T\|_F^{3/2} \right]^{1/3}.$$

Then, we select ϵ_1 and λ_1 as

$$\epsilon_1(\delta, r, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha) := \epsilon_0(\delta, r, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C),$$
$$\lambda_1(\delta, r, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha) := \min\left\{\lambda_0(\delta, r, \mu, \sigma_r(M_a^*), \|M_a^*\|_F, \alpha, C), \frac{(1-\mu-\delta)C^3}{4\sqrt{r}}\right\}.$$

Finally, we combine Lemmas 4-5 to get the bounds for the gradient and the Hessian. □

Proof of Theorem 7

In this subsection, we use similar notations:

$$M := UU^T, \quad M^* := U^*(U^*)^T,$$

where $M^* := M_s^*$ is the global optimum. We also assume that U^* is the minimizer of $\min_{X\in\mathcal{X}^*}\|U - X\|_F$ when there is no ambiguity about U. In this case, the distance is given by

$$\mathrm{dist}(U, \mathcal{X}^*) = \|U - U^*\|_F.$$

The proof of Theorem 7 is similar to that of Theorem 6. We first consider the case when $\|UU^T\|_F$ is large.

Lemma 6. *Given a constant $\epsilon > 0$, if*

$$\|UU^T\|_F^2 \geq \max\left\{\frac{2(1+\delta)}{1-\delta}\|U^*(U^*)^T\|_F^2, \left(\frac{2\lambda\sqrt{r}}{1-\delta}\right)^{4/3}\right\},$$

then

$$\|\nabla h_s(U)\|_F \geq \lambda.$$

Proof. Choosing the direction $\Delta := U$, we can calculate that

$$\langle \nabla h_s(U), \Delta\rangle = \langle \nabla f_s(UU^T), UU^T\rangle.$$

Using the δ-RIP$_{2r,2r}$ property, we have

$$\begin{aligned}\langle \nabla f_s(UU^T), UU^T\rangle &= \int_0^1 [\nabla^2 f_s(M^* + s(M - M^*)][M - M^*, M] \\ &\geq (1-\delta)\|M\|_F^2 - (1+\delta)\|M^*\|_F\|M\|_F \\ &\geq \frac{1-\delta}{2}\|M\|_F^2.\end{aligned}$$

Moreover,

$$\|\Delta\|_F = \|U\|_F \leq \sqrt{r}\|UU^T\|_F^{1/2}.$$

This leads to

$$\|\nabla h_s(U)\|_F \geq \frac{\langle \nabla h_s(U), \Delta\rangle}{\|\Delta\|_F} = \frac{\langle \nabla f_s(UU^T), UU^T\rangle}{\|U\|_F} \geq \frac{1-\delta}{2\sqrt{r}}\|UU^T\|_F^{3/2} \geq \lambda.$$

□

The next lemma is a counterpart of Lemma 4.

Lemma 7. *Consider positive constants α, C, λ such that*

$$\lambda \leq 2(\sqrt{r}C)^{-1}(\sqrt{2}-1)\sigma_r^2(U^*)\cdot\alpha^2, \quad G > \frac{(1+\delta)\lambda^2}{4G^2},$$

where $G := -\lambda_{min}(\nabla f_s(M))$. If

$$\|UU^T\|_F \leq C^2, \quad \|U-U^*\|_F \geq \alpha, \quad \|\nabla h_s(U)\|_F \leq \lambda,$$

then the inequality $G \geq c\alpha^2$ holds for some constant $c > 0$ independent of α, λ, C. Moreover, if there exists some positive constant τ such that

$$\sigma_r^2(U) \leq \frac{1}{1+\delta}\left[G - \frac{(1+\delta)\lambda^2}{4G^2}\right] - \tau, \tag{2.66}$$

then

$$\lambda_{min}(\nabla^2 h_s(U)) \leq -2(1+\delta)\tau.$$

Proof. We choose a singular vector q of U such that

$$\|q\|_2 = 1, \quad \|Uq\|_2 = \sigma_r(U).$$

We first prove the existence of the constant c. The δ-RIP$_{2r,2r}$ property gives

$$\langle \nabla f_s(M), M^* - M\rangle \leq -(1-\delta)\|M-M^*\|_F^2.$$

Using the assumption of this lemma, we have

$$\|\nabla f_s(M)U\|_2 \leq \|\nabla f_s(M)U\|_F = \frac{1}{2}\|\nabla h_s(U)\|_F \leq \frac{1}{2}\lambda, \tag{2.67}$$

which leads to

$$\langle \nabla f_s(M), M\rangle = \langle \nabla f_s(M)U, U\rangle \leq \|\nabla f_s(M)U\|_F\|U\|_F \leq \frac{1}{2}\lambda\cdot\sqrt{r}C.$$

Substituting into (2.67), it follows that

$$\langle \nabla f_s(M), M^* \rangle \leq -(1-\delta)\|M - M^*\|_F^2 + \frac{1}{2}\lambda \cdot \sqrt{r}C.$$

Using Lemma 1, we have

$$\|M - M^*\|_F^2 \geq 2(\sqrt{2}-1)\sigma_r^2(U^*)\|U - U^*\|_F^2 \geq 2(\sqrt{2}-1)\sigma_r^2(U^*) \cdot \alpha^2.$$

By the condition on λ, it follows that

$$\langle \nabla f_s(M), M^* \rangle \leq -(1-\delta)\|M - M^*\|_F^2 + \frac{1}{2}\lambda \cdot \sqrt{r}C \leq -(\sqrt{2}-1)(1-\delta)\sigma_r^2(U^*) \cdot \alpha^2. \quad (2.68)$$

The above inequality also indicates that $\lambda_{min}(\nabla f_s(M)) < 0$. Using the relations that

$$\nabla f_s(M) \succeq \lambda_{min}(\nabla f_s(M)) \cdot I_n, \quad M^* \succeq 0,$$

we arrive at

$$\langle \nabla f_s(M), M^* \rangle \geq \lambda_{min}(\nabla f_s(M)) \operatorname{tr}(M^*) \geq \sqrt{r}\|M^*\|_F \cdot \lambda_{min}(\nabla f_s(M)).$$

Combining the last inequality with (2.68), we obtain

$$\lambda_{min}(\nabla f_s(M)) \leq -(\sqrt{r}\|M^*\|_F)^{-1}(\sqrt{2}-1)(1-\delta)\sigma_r^2(U^*) \cdot \alpha^2 = -c\alpha^2$$

and thus $G \geq c\alpha^2$, where

$$c := (\sqrt{r}\|M^*\|_F)^{-1}(\sqrt{2}-1)(1-\delta)\sigma_r^2(U^*)$$

Next, we prove the upper bound on the minimal eigenvalue. We choose an eigenvector u such that

$$\|u\|_2 = 1, \quad \lambda_{min}(\nabla f_s(M)) = u^T \nabla f_s(M) u.$$

The direction is chosen to be

$$\Delta := uq^T.$$

For the Hessian of $h_s(\cdot,\cdot)$, we can calculate that

$$\langle \nabla f_s(M), \Delta\Delta^T \rangle = \lambda_{min}(\nabla f_s(M)) = -G \quad (2.69)$$

and the δ-RIP$_{2r,2r}$ property gives

$$\begin{aligned}
[\nabla^2 f_s(M)](\Delta U^T + U\Delta^T, \Delta U^T + U\Delta^T) & \\
&\leq (1+\delta)\|\Delta U^T + U\Delta^T\|_F^2 = (1+\delta)\|u(Uq)^T + (Uq)u^T\|_F^2 \\
&= 2(1+\delta)\|Uq\|_F^2 + 2(1+\delta)[q^T(U^T u)]^2 \\
&\leq 2(1+\delta)\sigma_r^2(U) + 2(1+\delta) \cdot \|U^T u\|_F^2. \quad (2.70)
\end{aligned}$$

By letting the vector $\tilde{v}$ be

$$\|\tilde{v}\|_2 = 1, \quad \lambda_{min}(\nabla f_s(M))u = \nabla f_s(M)\tilde{v},$$

the inequality (2.67) implies that

$$\|U^T u\|_F^2 = \frac{\|U^T \nabla f_s(M)\tilde{v}\|_F^2}{\lambda_{min}^2(\nabla f_s(M))} = \frac{\|U^T \nabla f_s(M)\tilde{v}\|_2^2}{\lambda_{min}^2(\nabla f_s(M))} \leq \frac{\|U^T \nabla f_s(M)\|_2^2 \|\tilde{v}\|_2^2}{\lambda_{min}^2(\nabla f_s(M))} \leq \frac{\lambda^2}{4G^2}.$$

Substituting into (2.70), we obtain

$$[\nabla^2 f_s(M)](\Delta U^T + U\Delta^T, \Delta U^T + U\Delta^T) \leq 2(1+\delta)\sigma_r^2(U) + (1+\delta)\cdot\frac{\lambda^2}{2G^2}. \tag{2.71}$$

Combining (2.69) and (2.71), it follows that

$$[\nabla^2 h_s(U)](\Delta, \Delta) \leq -2G + 2(1+\delta)\sigma_r^2(U) + (1+\delta)\cdot\frac{\lambda^2}{2G^2}.$$

Since $\|\Delta\|_F^2 = 1$, the above inequality implies

$$\lambda_{min}(\nabla^2 h_s(U)) \leq -2G + 2(1+\delta)\sigma_r^2(U) + (1+\delta)\cdot\frac{\lambda^2}{2G^2} \leq -(1+\delta)\tau.$$

□

We finally give the counterpart of Lemma 5, which states that the condition (2.66) always holds when $\delta < 1/3$.

Lemma 8. *Given positive constants $\alpha, C, \epsilon, \lambda$, if*

$$\max\{\|UU^T\|_F, \|U^*(U^*)^T\|_F\} \leq C^2, \ \|U - U^*\|_F \geq \alpha, \ \|\nabla h_s(U)\|_F \leq \lambda, \ \delta < 1/3,$$

then there exists a positive constant $\lambda_0(\delta, W^*, \alpha, C)$ such that

$$\sigma_r^2(U) \leq \frac{1}{1+\delta}\left[G - \frac{(1+\delta)\lambda^2}{4G^2} - \lambda\right] \tag{2.72}$$

whenever

$$0 < \lambda \leq \lambda_0(\delta, \sigma_r(M_s^*), \|M_s^*\|_F, \alpha, C).$$

Proof. We prove by contradiction, i.e., we assume

$$\sigma_r^2(U) > \frac{1}{1+\delta}\left[G - \frac{(1+\delta)\lambda^2}{4G^2} - \lambda\right] \geq \frac{c\alpha^2}{1+\delta} + \text{poly}(\lambda). \tag{2.73}$$

To follow the proof of Lemma 5, we also divide the argument into three steps, although the first step is superficial.

Step I. We first give a lower bound for $\lambda_r(M)$. In the symmetric case, this step is straightforward, since we always have

$$\lambda_r^2(M) = \sigma_r^4(U). \tag{2.74}$$

Step II. Next, we derive an upper bound for $\lambda_r(M)$. We define

$$\bar{M} := \mathcal{P}_r\left[M - \frac{1}{1+\delta}\nabla f_s(M)\right],$$

where $\mathcal{P}_r$ is the orthogonal projection onto the low-rank manifold (we do not drop negative eigenvalues in this proof). Since $M \neq M^*$ and $\delta < 1/3$, we recall that inequality (2.17) gives

$$\begin{aligned} -\phi(\bar{M}) &\geq \frac{1-3\delta}{1-\delta}[f_s(M) - f_s(M^*)] \geq \frac{1-3\delta}{2}\|M - M^*\|_F^2 \\ &\geq (1-3\delta)\cdot(\sqrt{2}-1)\sigma_r^2(W^*)\alpha^2 := K > 0, \end{aligned}$$

where the second inequality comes from Lemma 1 and

$$-\phi(\bar{M}) = \langle \nabla f_s(M), M - \bar{M}\rangle - \frac{1+\delta}{2}\|M - \bar{M}\|_F^2.$$

Hence,

$$\langle \nabla f_s(M), M - \bar{M}\rangle - \frac{1+\delta}{2}\|M - \bar{M}\|_F^2 \geq K. \tag{2.75}$$

For simplicity, we define

$$N := -\frac{1}{1+\delta}\nabla f_s(M).$$

Then, $\bar{M} = \mathcal{P}_r(M+N)$ and the left-hand side of (2.75) is equal to

$$\begin{aligned} &\langle \nabla f_s(M), M - \bar{M}\rangle - \frac{1+\delta}{2}\|M - \bar{M}\|_F^2 \\ &\qquad = (1+\delta)\langle N, \mathcal{P}_r(M+N) - M\rangle - \frac{1+\delta}{2}\|\mathcal{P}_r(M+N) - M\|_F^2 \\ &\qquad = \frac{1+\delta}{2}\left[\|N\|_F^2 - \|N + M - \mathcal{P}_r(M+N)\|_F^2\right] \\ &\qquad = \frac{1+\delta}{2}\left[\|N\|_F^2 - \|N+M\|_F^2 + \|\mathcal{P}_r(M+N)\|_F^2\right]. \end{aligned} \tag{2.76}$$

Similar to the proof of inequality (2.67), we can prove that

$$\|U^T N\|_F \leq \tilde{H} := \frac{\lambda}{2(1+\delta)}.$$

Then, we have

$$-\operatorname{tr}[N^T(UU^T)] \le \|U^TN\|_F\|U\|_F \le \tilde{H}\cdot\|U\|_F \le \tilde{H}\cdot\sqrt{\sqrt{r}\|UU^T\|_F} \le \sqrt[4]{r}C\cdot\tilde{H}.$$

Using the above relation, one can write

$$\|N\|_F^2 - \|N+M\|_F^2 = -2\operatorname{tr}[N^T(UU^T)] - \|M\|_F^2 \le 2\sqrt[4]{r}C\cdot\tilde{H} - \|M\|_F^2.$$

Suppose that $\mathcal{P}_U$ is the orthogonal projections onto the column space of U. We define

$$N_1 := \mathcal{P}_U N\mathcal{P}_U,\ N_2 := \mathcal{P}_U N(I-\mathcal{P}_U),\ N_3 := (I-\mathcal{P}_U)N\mathcal{P}_U,\ N_4 := (I-\mathcal{P}_U)N(I-\mathcal{P}_U).$$

Then, it follows from (2.73) that

$$\begin{aligned}\|N_1\|_F = \|\mathcal{P}_U N\mathcal{P}_U\|_F &\le \sigma_r^{-1}(U)\|U^T\mathcal{P}_U N\mathcal{P}_U\|_F \le \sigma_r^{-1}(U)\|U^TN\|_F \le \sigma_r^{-1}(U)\cdot\tilde{H}\\ &\le \left[\sqrt{\frac{1+\delta}{G}} + \operatorname{poly}(\lambda)\right]\cdot\tilde{H} \le \left[\sqrt{\frac{1+\delta}{c\alpha^2}} + \operatorname{poly}(\lambda)\right]\cdot\tilde{H} := \kappa\tilde{H}.\end{aligned}$$

Similarly, we can prove that

$$\|N_1+N_2\|_F = \|\mathcal{P}_U N\|_F \le \kappa\tilde{H}, \quad \|N_1+N_3\|_F = \|N\mathcal{P}_V\|_F \le \kappa\tilde{H},$$

which leads to

$$\|N_2\|_F \le 2\kappa\tilde{H}, \quad \|N_3\|_F \le 2\kappa\tilde{H}.$$

Using Weyl's theorem, the following holds for every $1 \le i \le r$:

$$|\lambda_i(M+N) - \lambda_i(M+N_4)| \le \|N_1+N_2+N_3\|_2 \le \|N_1+N_2+N_3\|_F \le 3\kappa\tilde{H}.$$

Therefore, we have

$$\begin{aligned}\|\mathcal{P}_r(M+N)\|_F^2 &= \sum_{i=1}^r \lambda_i^2(M+N)\\ &\ge \sum_{i=1}^r \lambda_i^2(M+N_4) - r\cdot 3\kappa\tilde{H}\cdot(\|M+N\|_2 + \|M+N_4\|_2)\\ &\ge \sum_{i=1}^r \lambda_i^2(M+N_4) - 6r\kappa\tilde{H}\cdot(\|M\|_2 + \|N\|_2)\\ &\ge \sum_{i=1}^r \lambda_i^2(M+N_4) - 6r\kappa\tilde{H}\cdot\left(\|M\|_F + \frac{G}{1+\delta}\right). \qquad (2.77)\end{aligned}$$

Similar to the asymmetric case, we can prove that

$$\frac{G}{1+\delta} \le \|M\|_F + \operatorname{poly}(\lambda).$$

holds under the assumption (2.73). Therefore, we obtain the bound

$$\|M\|_F + \|N\|_F \leq 2\|M\|_F + \text{poly}(\lambda) \leq 2C^2 + \text{poly}(\lambda).$$

Substituting back into the previous estimate (2.77), it follows that

$$\|\mathcal{P}_r(M+N)\|_F^2 \geq \sum_{i=1}^{r} \lambda_i^2(M+N_4) + \text{poly}(\lambda).$$

Now, since M and N_4 have orthogonal column and row spaces, the maximal r eigenvalues of $M+N_4$ are simply the maximal r eigenvalues of the eigenvalues of M and N_4, which we assume to be

$$\lambda_i(M),\ i=1,\dots,k \quad \text{and} \quad \lambda_i(N_4),\ i=1,\dots,r-k.$$

Now, it follows from (2.76) that

$$\begin{aligned}
\frac{2}{1+\delta}\left[\langle \nabla f_s(M), M-\bar{M}\rangle - \frac{1+\delta}{2}\|M-\bar{M}\|_F^2\right] & \\
&= \|N\|_F^2 - \|N+M\|_F^2 + \|\mathcal{P}_r(M+N)\|_F^2 \\
&\leq -\sum_{i=1}^{r}\lambda_i^2(M) + \sum_{i=1}^{k}\lambda_i^2(M) + \sum_{i=1}^{r-k}\lambda_i^2(N_4) + \text{poly}(\lambda) + 2\sqrt[4]{r}C\cdot\tilde{H} \\
&= -\sum_{i=k+1}^{r}\lambda_i^2(M) + \sum_{i=1}^{r-k}\lambda_i^2(N_4) + \text{poly}(\lambda). \qquad (2.78)
\end{aligned}$$

Using the assumption (2.73) and the fact that λ is small, we know that $\lambda_i(N_4) > 0$ for all $i \in \{1,\dots,k\}$. Therefore,

$$-\sum_{i=k+1}^{r}\lambda_i^2(M) + \sum_{i=1}^{r-k}\lambda_i^2(N_4) \leq -(r-k)\lambda_r^2(M) + (r-k)\lambda_{max}(N_4)^2.$$

Substituting into (2.78) gives rise to

$$\begin{aligned}
\frac{2}{1+\delta}\left[\langle \nabla f_s(M), M-\bar{M}\rangle - \frac{1+\delta}{2}\|M-\bar{M}\|_F^2\right] & \\
&\leq -(r-k)\lambda_r^2(M) + (r-k)\lambda_{max}(N_4)^2 + \text{poly}(\lambda) \\
&\leq -(r-k)\lambda_r^2(M) + (r-k)\lambda_{max}(N)^2 + \text{poly}(\lambda).
\end{aligned}$$

If $k=r$, then the above inequality and inequality (2.75) imply that

$$\text{poly}(\lambda) \geq K = O(\alpha^2),$$

which contradicts the assumption that λ is small. Hence, we conclude that $r - k \geq 1$. Combining with (2.75), we obtain the upper bound

$$\begin{aligned}\lambda_r^2(M) &\leq -\frac{2}{1+\delta} \cdot \frac{K}{r-k} + \lambda_{max}(N)^2 + \frac{1}{r-k} \cdot \text{poly}(\lambda) \\ &= -\frac{2}{1+\delta} \cdot \frac{K}{r} + \lambda_{max}(N)^2 + \text{poly}(\lambda). \end{aligned} \tag{2.79}$$

Step III. In the last step, we combine the relations (2.74) and (2.79), which leads to

$$\sigma_r^4(U) \leq -\frac{2}{1+\delta} \cdot \frac{K}{r} + \frac{1}{(1+\delta)^2} G^2 + \text{poly}(\lambda).$$

This means that

$$\sigma_r^4(U) + \frac{2}{1+\delta} \cdot \frac{K}{r} \leq \frac{1}{(1+\delta)^2} G^2 + \text{poly}(\lambda).$$

Since $K > 0$ has lower bounds that are independent of λ, we can choose λ to be small enough such that

$$\sigma_r^4(U) + \frac{1}{1+\delta} \cdot \frac{K}{r} \leq \frac{1}{(1+\delta)^2} G^2.$$

However, considering the assumption (2.73), we have

$$\begin{aligned}\sigma_r^4(U) &\geq \frac{1}{(1+\delta)^2} \left[G - \frac{(1+\delta)\lambda^2}{4G^2} - \lambda \right]^2 = \frac{1}{(1+\delta)^2} G^2 - 2\lambda \cdot G + \text{poly}(\lambda) \\ &\geq \frac{1}{(1+\delta)^2} G^2 - 2\lambda \cdot (1+\delta) C^2 + \text{poly}(\lambda) = \frac{1}{(1+\delta)^2} G^2 + \text{poly}(\lambda),\end{aligned}$$

where the second inequality is due to $G \leq (1+\delta)C^2$, which can be proved similar to the asymmetric case. The above two inequalities cannot hold simultaneously when λ is small enough. This contradiction means that the condition (2.72) holds by choosing

$$0 < \lambda \leq \lambda_0(\delta, \sigma_r(M_s^*), \|M_s^*\|_F, \alpha, C),$$

for a small enough positive constant $\lambda_0(\delta, \sigma_r(M_s^*), \|M_s^*\|_F, \alpha, C)$.

□

Proof of Theorem 7. We first choose

$$C := \left[\frac{2(1+\delta)}{1-\delta} \|U^*(U^*)^T\|_F^2 \right]^{1/4}.$$

Then, we select λ_1 as

$$\lambda_1(\delta, r, \sigma_r(M_s^*), \|M_s^*\|_F, \alpha) := \min \left\{ \lambda_0(\delta, r, \sigma_r(M_s^*), \|M_s^*\|_F, \alpha, C), \frac{(1-\delta)C^3}{2\sqrt{r}} \right\}.$$

Finally, we combine Lemmas 6-8 to get the bounds for the gradient and the Hessian. □

Chapter 3

A New Complexity Metric for Rank-one Generalized Matrix Completion

3.1 Introduction

To explain the empirical success of the Burer-Monteiro factorization (2.2), multiple *complexity metrics* were proposed to characterize the behavior of local search methods. A small complexity metric implies that the landscape of problem (2.2) is benign and thus, local search methods with random initialization converge to global solutions with high probability. Otherwise, if the complexity metric takes a large value, problem (2.2) may have spurious local minima, which will imply the failure of most local search methods. However, the existing so-called "complexity metrics" for problem (2.2) are only able to guarantee a benign landscape when the complexity is small and fail to prove the existence of spurious local minima when the complexity is large. To differentiate with true complexity metrics, we use the term *recovery guarantees* to reflect such weaker properties. In addition, the existing recovery guarantees were designed separately for different applications. As a result, several different bounds were proposed to characterize the optimization complexity of problem (2.2).

For example, in the context of matrix sensing problems, we have shown in Chapter 2 that the *Restrict Isometry Property* (RIP) is effective in characterizing the optimization complexity of the Burer-Monteiro factorization. As an important class of matrix sensing problems, the *linear matrix sensing problem* (2.5) can be equivalently formulated as

$$\min_{U\in\mathbb{R}^{n\times r}} \frac{1}{m}\sum_{i=1}^{m}\langle A_i, UU^T - M^*\rangle^2, \tag{3.1}$$

where $m \in \mathbb{N}$ is the number of measurements modeled by the known measurement matrices $A_i \in \mathbb{R}^{n\times n}$ for all $i \in [m]$. In the special case when each matrix A_i is an independently identically distributed Gaussian random matrix, the δ-RIP$_{2r,2s}$ condition holds with high

probability if $m = O(nr\delta^{-2})$ [33]. The RIP constant δ plays a critical role in bounding the optimization complexity of problem (2.2). In Chapter 2 and the related paper [24], we showed that the strict-saddle property holds for problem (2.2) if the δ-RIP$_{2r,2r}$ condition holds with $\delta < 1/2$ and the ground truth matrix satisfies $\text{rank}(M^*) = r$. On the other hand, counterexamples have been constructed in Chapter 2 to illustrate that the strict-saddle property can fail under the δ-RIP$_{2r,2r}$ condition with $\delta \geq 1/2$.

Despite these strong theoretical results under the RIP assumption, there exists a large number of applications that do not satisfy the RIP condition. One of those applications without the RIP condition is the *matrix completion problem*. Given a set of indices $\Omega \subset [n] \times [n]$, the matrix completion problem aims at recovering the low-rank matrix M^* from the available entries M^*_{ij} for $(i,j) \in \Omega$. With the least squares loss function, the matrix completion problem can be formulated as

$$\min_{U \in \mathbb{R}^{n \times r}} \sum\nolimits_{(i,j) \in \Omega} \left[(UU^T)_{ij} - M^*_{ij} \right]^2. \tag{3.2}$$

The matrix completion problem (3.2) is a special case of the matrix sensing problem (3.1), where each measurement matrix A_i has exactly one nonzero entry. However, the RIP$_{2r,2r}$ condition does not hold for problem (3.2) unless all entries of M^* are observed, namely, when $\Omega = [n] \times [n]$. As an alternative to the RIP condition, the optimization complexity of problem (3.2) is closely related to the incoherence of M^*.

Definition 4 ([34])**.** Given a constant $\mu \in [1, n]$, the ground truth matrix M^* is said to be **μ-incoherent** if

$$\|e_i^T V^*\|_F \leq \sqrt{\mu r / n}, \quad \forall i \in [n], \tag{3.3}$$

where $V^* \Lambda^* (V^*)^T$ is the truncated SVD of M^* and e_i is the i-th standard basis of $\mathbb{R}^n$.

Intuitively, if the ground truth M^* is highly sparse, it is likely that only zero entries of M^* are observed and there is no chance to learn the other entries of the matrix M^*. A relatively small incoherence of M^* avoids this extreme case. The most popular statistical model of the measurements for problem (3.2) is the Bernoulli model, where each entry of M^* is observed independently with probability $p \in (0, 1]$. Assuming the Bernoulli model, the incoherence of M^* and the sampling probability p can jointly characterize the complexity of the matrix completion problem. For example, the scaled gradient descent algorithm with a spectral initialization [216] converges linearly given the condition $p \geq O(\mu r^2 \kappa^2 \max(\mu\kappa^2, \log n)/n)$, where $\kappa := \sigma_1(M^*)/\sigma_r(M^*)$ is the condition number of M^*. In addition, under the assumption that $p \geq O(\mu^4 r^6 \kappa^6 \log n / n)$, the global convergence was established in [86] through the strict-saddle property of a regularized version of problem (3.2). We note that the dependence on the condition number κ may be unnecessary as shown in [97] and that the condition number is equal to 1 in the rank-1 case. On the other hand, the information-theoretical lower bound in [34] shows that $p \geq \Theta(\mu r \log(n/\delta)/n)$ is necessary for the exact completion with probability at least $1 - \delta$. Therefore, the complexity of problem (3.2) is closely related to the

incoherence of M^* and the sampling probability p. In the remainder of this chapter, we refer to the conditions on the incoherence of M^* and sampling rate p as *incoherence conditions* when there is no confusion in the context.

To be more rigorous, the RIP condition and the incoherence condition may have a subtle difference in their nature. As a counterpart of the incoherence condition in other low-rank matrix optimization problems, one should consider conditions in terms of the sampling complexity. On the other hand, the RIP condition is a deterministic condition on the loss function and is not related to the underlying random model. However, there is a wide range of problems that satisfy the RIP condition when the sample complexity is sufficiently large. By considering the properties of the RIP condition, we are able to analyze a large number of low-rank matrix optimization problems simultaneously. Therefore, we use the RIP condition instead of conditions based on the sample complexity as a notion of the computational complexity for those problems.

The main issue with the notions of RIP and incoherence is that they require stringent conditions to guarantee the success of local search methods for recovering M^*. Whenever these conditions are violated, local search methods may still work successfully, which questions whether these customized notions designed for special cases of the problem truly capture the complexity of the problem in general. Hence, it is natural to ask:

> *Does there exist a complexity metric with two properties: (i) it is consistent with existing recovery guarantees designed for different applications, e.g., the RIP constant δ and the incoherence μ combined with the sampling rate p, (ii) even when the customized conditions for different applications are violated, it still quantifies the optimization complexity of the problem in the sense that the smaller the value of this metric is, the higher the success of local search methods with random initialization is in finding the ground truth M^*?*

In this chapter, we provide a partial answer to the question by developing a powerful complexity metric. To analyze the usefulness of this new metric, we focus on the rank-1 generalized matrix completion problem

$$\min_{u\in\mathbb{R}^n} \sum\nolimits_{i,j\in[n]} C_{ij}(u_iu_j - M^*_{ij})^2, \tag{3.4}$$

where the ground truth M^* is symmetric and has rank at most 1. The weights are $C_{ij} \geq 0$ for all $i, j \in [n]$. Without loss of generality, we can assume that the matrix $C := (C_{ij})_{i,j\in[n]}$ is symmetric since otherwise one can replace C with $(C + C^T)/2$, which will not change the optimization landscape. We use $\mathcal{MC}(C, u^*)$ to denote the instance of problem (3.4) with the weight matrix C and the ground truth $M^* = u^*(u^*)^T$, for all $C \in \mathbb{R}^{n\times n}$ and $u^* \in \mathbb{R}^n$. The matrix completion problem (3.2) is a special case of the generalized matrix completion problem (3.4), where $C_{ij} = 1$ if $(i, j) \in \Omega$ and $C_{ij} = 0$ otherwise.

Moreover, problem (3.4) is a special case of the matrix sensing problem (3.1), where each measurement only captures one entry of M^*. However, the problem (3.4) still contains difficult instances of the matrix sensing problem from the perspective of the RIP condition.

In Section 3.3, we show that there exists an instance of problem (3.4) that satisfies the $1/2$-RIP$_{2,2}$ condition but has spurious local minima. This counterexample implies that the optimal RIP bound in [247, 244] still holds for problem (3.4) and thus, problem (3.4) contains difficult instances of the matrix sensing problem. Moreover, we show in Section 3.3 that some of the results developed for problem (3.4) can be extended to general problem (2.2).

Now, we provide an intuition into the design of our complexity metric for problem (3.4). For a given problem instance of (3.4), if there exist global solutions u^1, u^2 such that $u^1(u^1)^T \neq u^2(u^2)^T$, it is impossible to decide which global solution corresponds to M^* from the observations. Intuitively, no matter what optimization algorithm we choose and how much computational effort is exerted, there is a chance that we could not recover M^* by solving problem (3.4). This observation motivates us to define the complexity metric to be the inverse of the infimum of the distance between any given instance and the set of instances with multiple global solutions. Since problem (3.4) is parameterized by the weight matrix C and the global solution M^*, we are able to define the metric through norms in Euclidean spaces and their Cartesian products. In addition, in the rank-1 case, (random) graph theory serves as an important tool in characterizing the solvability of problem (3.4). These two advantages enable a more thorough analysis of the new complexity metric. The formal definition of the metric is provided in Section 3.2. In this chapter, we exhibit several pieces of evidence to show that the proposed metric can serve as an alternative to the RIP constant and the incoherence, which are summarized below:

1. For problem instances that satisfy the δ-RIP$_{2,2}$ condition, we provide an upper bound on the complexity metric. The upper bound is tightened with extra information about the incoherence of M^*. Similarly, for matrix completion problems obeying the Bernoulli sampling model, an upper bound on the complexity metric in terms of the incoherence of M^* is derived.

2. We then construct a class of parameterized instances of problem (3.4), where the RIP condition fails to provide useful guarantees. A lower bound on the complexity metric is developed to prove that instances whose complexity metric is larger than the lower bound have an exponential number of spurious local minima. In addition, an upper bound that is consistent with the aforementioned two upper bounds is established to guarantee the absence of spurious local minima if the complexity metric is below this bound. The consistency of the upper bounds between different types of models provides strong evidence that the new complexity metric is able to provide theoretical guarantees for different applications, even when the RIP condition or the incoherence condition fails.

3. We prove the existence of a non-trivial upper bound on the complexity metric. For all problem instances whose complexity metric is below this upper bound, problem (3.4) has no spurious local minima and M^* can be successfully found via local search methods with random initialization. In addition, under a standard bounded-away-

from-zero assumption, we show that all instances with a larger complexity metric will possess spurious local minima.

4. We extend all results for the symmetric generalized matrix completion problem to the asymmetric case, where low-rank matrices is decomposed in to UV^T for some $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ in problem (2.2).

Based on the aforementioned results, we make some key conjectures and discuss the potential extensions of the proposed metric to more general cases of the low-rank matrix optimization problem (2.1).

Related works

Following the famous *Netflix prize*, the theoretical analysis of problem (2.1) has attracted a lot of attention in recent years; see the review papers [46, 50]. Early attempts mainly focused on the construction of convex relaxations to rank-constrained problems [34, 35, 191, 36], where the RIP condition and the incoherence condition were introduced. Recently, several modified RIP conditions were proposed to better characterize the landscapes of other classes of problems, e.g., the ℓ_1/ℓ_2-RIP condition [146], the sign-RIP condition [158], and the approximation and sharpness condition [38].

Although the convex relaxation is usually guaranteed to recover the exact ground truth with almost the optimal sample complexity, the associated algorithms operate in the space of matrix variables and, thus, are computationally inefficient for large-scale problems [252]. Similar issues are observed for algorithms based on the Singular Value Projection [117] and Riemannian optimization algorithms [230, 229, 106, 4, 156]. The analysis of the convex relaxation approach in the noisy case is recently conducted by bridging the convex and the nonconvex approaches [49, 47].

To deal with the difficulties in solving large-scale problems, an efficient alternative model (2.2) using the Burer-Monteiro factorization is considered. Despite the nonconvexity, a growing number of works demonstrated that problem (2.2) has benign landscapes and, therefore, is amenable for efficient optimization. Theoretical analysis stems from the alternating minimization method [118, 172, 96, 97, 173, 1]. The alternating minimization method has the advantage that the number of iterations has only logarithmic dependence on the condition number of the ground truth [97]. More recently, this advantage is also achieved by the scaled (sub)gradient descent algorithm [216, 217, 218, 245].

The gradient descent algorithm has also gained a significant attention due to its simplicity in implementation. In general, there are two ways to apply the gradient descent algorithm. First, the gradient descent algorithm can serve as the local refinement method after a suitable initialization [32, 219, 213, 242, 7, 40]. On the other hand, the gradient descent algorithm is proved to converge globally for the phase retrieval problem [48]. More generally, under the strict-saddle property, a number of saddle-escaping algorithms [124, 37, 8] converge to the global solution in polynomial time; see e.g., [211, 87, 86, 256, 210, 247, 39, 244, 23, 24,

161]. Moreover, the gradient descent algorithm is proved to have the implicit regularization phenomenon in the over-parameterization case [147, 53, 209].

In the remainder of this chapter, we first define the proposed complexity metric and derive basic properties of the metric in Section 3.2. In Section 3.3, we analyze this metric under existing conditions, including the RIP condition and the incoherence condition. Section 3.4 is devoted to the theoretical guarantees provided by the new complexity metric on the general instances of problem (3.4). The results for the rank-1 asymmetric generalized matrix completion problem are provided in Appendix 3.E. Some of the proofs are provided in the appendix.

3.2 New Complexity Metric and Basic Properties

In this section, we first provide the formal definition of the new complexity and investigate the properties of the proposed metric. More specifically, we show that we are able to utilize the graph theory to estimate the complexity metric and calculate the minimum possible value of the proposed complexity metric in closed form. Before proceeding to the definitions, we note that the problem (3.4) is "scale-free" in the sense that the instance $\mathcal{MC}(\eta_1 C, \eta_2 u^*)$ has the same landscape as $\mathcal{MC}(C, u^*)$ up to a scaling, where $C \in \mathbb{R}^{n\times n}$, $u^* \in \mathbb{R}^n$ and $\eta_1, \eta_2 > 0$ are constants. Therefore, we may normalize the parameters C and u^* without loss of generality, as follows:

Assumption 3. Assume that $C \in \mathbb{S}_{+,1}^{n^2-1}$ and $u^* \in \mathbb{S}_1^{n-1}$, i.e., $\|C\|_1 = \|u^*\|_1 = 1$.

The above assumption excludes the degenerate cases when $C = 0$ or $M^* = 0$. If $C = 0$, the objective function is always 0 and it is impossible to recover the ground truth. For the case when $M^* = 0$, we can prove that either $u = 0$ is the only stationary point or the instance $\mathcal{MC}(C, 0)$ has multiple different global solutions. In the first situation, the results in [138] imply that randomly initialized gradient descent algorithm will converge to 0 with probability 1. In the second situation, the instance is information-theoretically unsolvable. We provide a more detailed analysis in the appendix and assume that Assumption 3 holds in the remainder of the chapter.

The definition of the complexity metric is closely related to the set of instances with multiple "essentially different" global solutions. More specifically, the set of degenerate instances is defined as

$$\mathcal{D} := \{(C, u^*) \mid C \in \mathbb{S}_{+,1}^{n^2-1}, u^* \in \mathbb{S}_1^{n-1},\ \exists u \in \mathbb{R}^n \quad \text{s. t. } g(u; C, u^*) = 0,\ uu^T \neq u^*(u^*)^T\}.$$

Since there exist multiple global solutions to problem (3.4) if $(C, u^*) \in \mathcal{D}$, it is information-theoretically impossible to find the ground truth for any instance in $\mathcal{D}$. Intuitively, we say that the *optimization complexity* of all instances in $\mathcal{D}$ is infinity. Motivated by the above observation, we introduce the new complexity metric.

Definition 5 (Complexity Metric)**.** Given arbitrary parameters $C \in \mathbb{S}_{+,1}^{n^2-1}$, $u^* \in \mathbb{S}_1^{n-1}$ and $\alpha \in [0,1]$, the complexity of the instance $\mathcal{MC}(C, u^*)$ is defined as

$$\mathbb{D}_\alpha(C, u^*) := \left[\inf_{(\tilde{C}, \tilde{u}^*) \in \mathcal{D}} \alpha \|C - \tilde{C}\|_1 + (1-\alpha)\|u^* - \tilde{u}^*\|_1 \right]^{-1}. \tag{3.5}$$

Since the set $\mathcal{D}$ is bounded, the infimum in the definition is finite. The term inside the inverse operation can be viewed as a weighted distance between the point (C, u^*) and the set $\mathcal{D}$. In addition, we take the convention that $1/0 = +\infty$ and thus, $\mathbb{D}_\alpha(C, u^*) = +\infty$ for all $(C, u^*) \in \mathcal{D}$. In this chapter, we choose the entry-wise ℓ_1-norm in (3.5) for the simplicity of calculations. We believe that similar theory can still be derived for other choices of the norm. We note that a similar complexity was proposed in [193, 192] for conic optimization and to the best of authors' knowledge, there is no similar complexity metric for nonconvex optimization problems.

For the parameter α, we will discuss two potential choices in this section, namely α^* and $\alpha^\diamond$. In the case when $\alpha = \alpha^*$, the range of the complexity metric has the largest size. Intuitively, by choosing $\alpha = \alpha^*$, the difference between the complexities of two instances will be maximized and thus, it is easier to compare the complexities of different instances. On the other hand, when we choose $\alpha = \alpha^\diamond$, the complexity metric attains its minimum possible value if and only if the 0-RIP$_{2,2}$ condition holds. This is consistent with the intuition that instances with the RIP constant 0 are the easiest to solve. We note that both α^* and $\alpha^\diamond$ satisfy $1 - \alpha = \Theta(1/n)$. Moreover, in Section 3.3, we show that the parameter α strikes a balance between the RIP constant of the instance and the incoherence of the ground truth. It is still an open question what the optimal choice of parameter α is, which may depend on the class of problems under consideration. It may be needed to jointly consider the complexity metric with several different choices of α to determine the solvability of the instance.

Basic Properties of the New Complexity Metric

We first provide a more concrete characterization of the set $\mathcal{D}$. In the rank-1 case, we are able to exactly describe the set $\mathcal{D}$ using graph-theoretic notations. We introduce the associated graphs of any instance of the problem. Given an instance $\mathcal{MC}(C, u^*)$, the weighted graph $\mathbb{G}(C, u^*) = [\mathbb{V}(C, u^*), \mathbb{E}(C, u^*), \mathbb{W}(C, u^*)]$ is defined by

$$\begin{aligned} &\mathbb{V}(C, u^*) := [n], \quad \mathbb{E}(C, u^*) := \{\{i, j\} \mid C_{ij} > 0, i, j \in [n]\}, \\ &[\mathbb{W}(C, u^*)]_{ij} := C_{ij}, \quad \forall i, j \in [n] \quad \text{s.t. } \{i, j\} \in \mathbb{E}(C, u^*). \end{aligned}$$

To include the information of u^*, we define

$$\begin{aligned} &\mathcal{I}_1(C, u^*) := \{i \in [n] \mid u_i^* \neq 0\}, \quad \mathcal{I}_0(C, u^*) := [n] \backslash \mathcal{I}_1(C, u^*), \\ &\mathcal{I}_{00}(C, u^*) := \{i \in \mathcal{I}_0(C, u^*) \mid \{i, j\} \notin \mathbb{E}(C, u^*),\ \forall j \in \mathcal{I}_1(C, u^*)\}. \end{aligned}$$

Intuitively, the sets $\mathcal{I}_1(C, u^*)$ and $\mathcal{I}_0(C, u^*)$ contain the locations of the nonzero and zero components of u^*. The subset $\mathcal{I}_{00}(C, u^*)$ corresponds to indices in $\mathcal{I}_0(C, u^*)$ that are not

connected to any index in $\mathcal{I}_1(C,u^*)$. We denote the subgraph of $\mathbb{G}(C,u^*)$ induced by the index set $\mathcal{I}_1(C,u^*)$ as $\mathbb{G}_1(C,u^*) = [\mathcal{I}_1(C,u^*), \mathbb{E}_1(C,u^*), \mathbb{W}_1(C,u^*)]$, where $\mathbb{E}_1(C,u^*)$ and $\mathbb{W}_1(C,u^*)$ are the edge set and weight set of this subgraph. The following theorem provides an equivalent definition of $\mathcal{D}$ in terms of $\mathcal{I}_{00}(C,u^*)$ and $\mathbb{G}_1(C,u^*)$.

Theorem 14. *Given $C \in \mathbb{S}^{n^2-1}_{+,1}$ and $u^* \in \mathbb{S}^{n-1}_1$, it holds that $(C,u^*) \notin \mathcal{D}$ if and only if*

1. *$\mathbb{G}_1(C,u^*)$ is connected and not bipartite;*
2. *$\{i,i\} \in \mathbb{E}(C,u^*)$ for all $i \in \mathcal{I}_{00}(C,u^*)$.*

Proof. We first construct counterexamples for the necessity part and then prove the uniqueness of the global minimum (up to a sign flip) for the sufficiency part. For the notational simplicity, we fix the point (C,u^*) and omit them in the notations.

Necessity. In this part, our goal is to construct a solution $u \in \mathbb{R}^n$ such that

$$u_iu_j = u^*_iu^*_j, \quad \forall \{i,j\} \in \mathbb{E}; \quad uu^T \neq u^*(u^*)^T.$$

We denote $M^* := u^*(u^*)^T$ and analyze three different cases below.

Case I. First, we consider the case when $\mathbb{G}_1$ is disconnected, which means that there exist two non-empty subsets $\mathcal{I}$ and $\mathcal{J}$ such that

$$\mathcal{I} \cup \mathcal{J} = \mathcal{I}_1, \quad \mathcal{I} \cap \mathcal{J} = \emptyset; \quad \{i,j\} \notin \mathbb{E}_1, \quad \forall i \in \mathcal{I},\ \forall j \in \mathcal{J}.$$

We define the vector $u \in \mathbb{R}^n$ as

$$u_i := 0, \quad \forall i \in \mathcal{I}_0; \quad u_i = u^*_i, \quad \forall i \in \mathcal{I}; \quad u_i = -u^*_i, \quad \forall i \in \mathcal{J}.$$

The above definition leads to

$$u_iu_j = \begin{cases} -M^*_{ij} & \text{if } i \in \mathcal{I} \text{ and } j \in \mathcal{J} \\ M^*_{ij} & \text{otherwise.} \end{cases}$$

Since $u^*_i \neq 0$ for all $i \in \mathcal{I}_1$, it follows that $u_iu_j = -M^*_{ij} \neq M^*_{ij}$ for all $\{i,j\}$ such that $i \in \mathcal{I}$ and $j \in \mathcal{J}$.

Case II. Next, we consider the case when $\mathbb{G}_1$ is bipartite, which means that there exist two non-empty subsets $\mathcal{I}$ and $\mathcal{J}$ such that

$$\mathcal{I} \cup \mathcal{J} = \mathcal{I}_1, \quad \mathcal{I} \cap \mathcal{J} = \emptyset; \quad \{i,j\} \notin \mathbb{E}_1, \quad \forall i,j \in \mathcal{I}_1 \quad \text{s.t. } i,j \in \mathcal{I} \text{ or } i,j \in \mathcal{J}.$$

In this case, we define the vector $u \in \mathbb{R}^n$ as

$$u_i := 0, \quad \forall i \in \mathcal{I}_0; \quad u_i := u^*_i/2, \quad \forall i \in \mathcal{I}; \quad u_i := 2u^*_i, \quad \forall i \in \mathcal{J}.$$

Now, we have

$$u_i u_j = \begin{cases} M^*_{ij}/4 & \text{if } i,j \in \mathcal{I} \\ 4M^*_{ij} & \text{if } i,j \in \mathcal{J} \\ M^*_{ij} & \text{otherwise.} \end{cases}$$

Since $M^*_{ij} \neq 0$ for all $i,j \in \mathcal{J}$, we have that $u_i u_j = 4M^*_{ij} \neq M^*_{ij}$ for all $i,j \in \mathcal{J}$.

Case III. Finally, we check the case when there exists a node $i_0 \in \mathcal{I}_{00}$ such that $\{i_0, i_0\} \notin \mathbb{E}$. In this case, we define the vector $u \in \mathbb{R}^n$ as

$$u_{i_0} := 1, \quad u_i := u^*_i, \quad \forall i \in [n] \backslash \{i_0\}.$$

Now, we have

$$u_{i_0} u_{i_0} = 1 \neq 0 = M^*_{i_0 i_0}, \quad u_i u_j = M^*_{ij}, \quad \forall \{i,j\} \in \mathbb{E}.$$

Combining the above three cases completes the proof of the necessity part.

Sufficiency. We prove that any global solution $u \in \mathbb{R}^n$ to problem (3.4) satisfies $uu^T = M^*$, where $M^* := u^*(u^*)^T$. Since u is a global solution, it follows that

$$u_i u_j = M^*_{ij}, \quad \forall i,j \in [n] \quad \text{s.t. } \{i,j\} \in \mathbb{E}.$$

Since the graph $\mathbb{G}_1$ is not bipartite, there exists a cycle with an odd number of edges in $\mathbb{G}_1$. We denote the length of the cycle as $2k+1$, where k is a non-negative integer. Moreover, we denote the edges of the cycle as

$$\{i_0, i_1\}, \{i_1, i_2\}, \ldots, \{i_{2k}, i_0\}.$$

Since $\{i_0, \ldots, i_{2k}\} \subset \mathcal{I}_1$, we know that

$$u_i u_j = M^*_{ij} \neq 0, \quad \forall i,j \in [n] \quad \text{s.t. } \{i,j\} \in \{\{i_\ell, i_{\ell+1}\}, \ell \in \{0, \ldots, 2k\}\},$$

where $i_{2k+1} := i_0$. Hence, we can calculate that

$$u_0^2 = \prod_{\ell=0}^{2k} (u_{i_\ell} u_{i_{\ell+1}})^{(-1)^\ell} = \prod_{\ell=0}^{2k+1} M^{(-1)^\ell}_{i_\ell i_{\ell+1}} = (u^*_{i_0})^2.$$

Without loss of generality, assume that $u_{i_0} = u^*_{i_0}$ since otherwise we can consider the solution $-u$ if $u_{i_0} = -u^*_{i_0}$. With the value of u_{i_0} correctly recovered, it follows that

$$u_{i_1} = \frac{u_{i_0} u_{i_1}}{u_{i_0}} = \frac{u^*_{i_0} u^*_{i_1}}{u^*_{i_0}} = u^*_{i_1}.$$

Similarly, we can utilize the connectivity of $\mathbb{G}_1$ to iteratively obtain $u_i = u^*_i$ for all $i \in \mathcal{I}_1$.

The remaining part is to show that $u_i = 0$ for all $i \in \mathcal{I}_0$. For every node $i \in \mathcal{I}_0 \backslash \mathcal{I}_{00}$, there exists a node $j \in \mathcal{I}_1$ such that $\{i, j\} \in \mathbb{E}$. This implies that

$$u_j = u_j^* \neq 0, \quad u_i u_j = M_{ij}^* = 0,$$

Hence, it holds that $u_i = 0$. For every node $i \in \mathcal{I}_{00}$, the assumption in the theorem requires that $\{i, i\} \in \mathbb{E}$, which leads to

$$u_i^2 = M_{ii}^* = 0.$$

In this case, we also obtain $u_i = 0$. □

Since the set $\mathcal{D}$ is bounded, the infimum in the definition (3.5) can be attained by using the closure of $\mathcal{D}$, namely

$$\mathbb{D}_\alpha(C, u^*) = \left[\min_{(\tilde{C}, \tilde{u}^*) \in \overline{\mathcal{D}}} \alpha \|C - \tilde{C}\|_1 + (1 - \alpha) \|u^* - \tilde{u}^*\|_1 \right]^{-1}. \tag{3.6}$$

The alternative definition (3.6) simplifies the verification of parameters that attain the infimum. In addition, with the help of Theorem 14, we can exactly characterize the closure $\overline{\mathcal{D}}$, which has a slightly simpler form than $\mathcal{D}$.

Theorem 15. *We have the following relation:*

$$\begin{aligned} \overline{\mathcal{D}} = \{(C, u^*) \mid C \in \mathbb{S}_{+,1}^{n^2-1}, u^* \in \mathbb{S}_1^{n-1}, \mathbb{G}_1(C, u^*) \text{ is disconnected or bipartite}\} \\ \cup \{(C, u^*) \mid C \in \mathbb{S}_{+,1}^{n^2-1}, u^* \in \mathbb{S}_1^{n-1}, \mathcal{I}_{00}(C, u^*) \text{ is not empty}\}. \end{aligned}$$

Let the set in the right-hand side of the above equation be called $\mathcal{D}'$. The proof of Theorem 15 is based on a standard technique that first shows $\bar{\mathcal{D}} \subset \mathcal{D}'$ and then shows $\mathcal{D}' \subset \bar{\mathcal{D}}$. The details can be found in Appendix 3.B. Using the results in Theorems 14 and 15, we provide an estimate on the scale of the new metric. Since $\mathcal{D}$ is a bounded set, there exists an upper bound on the minimum possible value of the complexity metric, which is defined below:

$$\mathbb{D}_\alpha^{min} := \min_{C \in \mathbb{S}_{+,1}^{n^2-1}, u^* \in \mathbb{S}_1^{n-1}} \mathbb{D}_\alpha(C, u^*).$$

The next theorem provides the expression of $\mathbb{D}_\alpha^{min}$.

Theorem 16. *Suppose that $n \geq 5$. Then, it holds that*

$$\mathbb{D}_\alpha^{min} = \begin{cases} \frac{n}{4\alpha} & \text{if } \alpha \leq \frac{n^2-3n-2}{n^2-5n+4} \\ \frac{n^2}{2(1-\alpha)(n-2)n+4\alpha} & \text{if } \frac{n}{n+2} \leq \alpha \leq \frac{n}{n+1} \\ \frac{n(n+1)}{2(1-\alpha)(n-2)(n+1)+4} & \text{if } \alpha \geq \frac{n}{n+1}. \end{cases}$$

In the regime $(n^2 - 3n - 2)/(n^2 - 5n + 4) \leq \alpha \leq n/(n + 2)$, we have the estimate

$$\mathbb{D}_\alpha^{min} \in \left[\frac{n}{4\alpha}, \frac{n^2}{4\alpha(n-1)} \right].$$

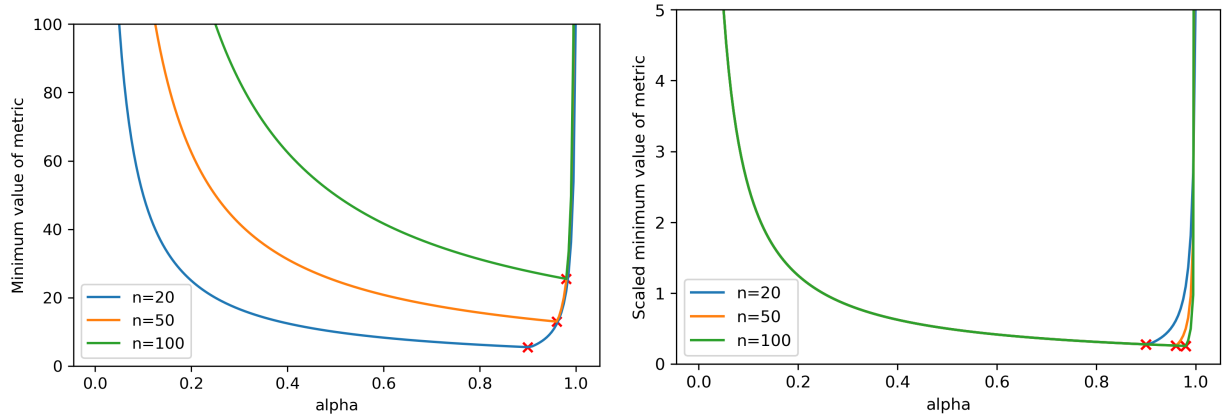


Figure 3.2.1: Comparison of $\mathbb{D}_\alpha^{min}$ for $n = 20, 50, 100$. The red "$\times$" sign refers to the value at α^*. In the right plot, the complexity metric is scaled by n^{-1}.

The proof of Theorem 16 can be found in Appendix 3.B. Now, we provide the proof of Theorem 16. The results of Theorem 16 imply that in the regime where $\alpha \geq \Theta(1)$ and $1-\alpha \geq \Theta(n^{-1})$, we have $\mathbb{D}_\alpha^{min} = O(n)$. This suggests that $n^{-1}\mathbb{D}_\alpha(C, u^*)$ may be a dimension-free complexity metric; see more examples supporting this claim in Section 3.3. In addition, the minimum possible value of the complexity is attained at

$$\alpha^* := (n^2 - 5n + 4)/(n^2 - 3n - 2).$$

Hence, the set of possible values of the complexity metric attains the maximum size by choosing $\alpha = \alpha^*$. This observation hints that α^* may be the optimal choice of α since it may enable the metric to differentiate instances with different complexities to the maximum degree. Using the exact formulation of $g(\alpha, c)$ in Lemma 11, we plot the minimum possible value of the complexity metric both without scaling and after scaling by n^{-1} in Figure 3.2.1. From the numerical results, we can see that the complexity scales with n if α is smaller than α^*, which is consistent with Theorem 16. If α is larger than α^*, the complexity metric for different values of n approximately lies on the same curve.

In the following theorem, we show that if $\alpha = \alpha^*$, the instances that attain the minimum value of the complexity metric are unique up to sign flips to components of the global solution.

Theorem 17. *Suppose that $n \geq 5$ and the instance $\mathcal{MC}(C, u^*)$ satisfies*

$$\mathbb{D}_{\alpha^*}(C, u^*) = n/(4\alpha^*).$$

Then, it holds that

$$|u_i^*| = 1/n, \quad C_{ii} = 0, \quad \forall i \in [n]; \quad C_{ij} = 1/[n(n-1)], \quad \forall i, j \in [n],\ i \neq j.$$

The proof of Theorem 17 can be found in Appendix 3.B. The above theorem states that if we choose the weight $\alpha = \alpha^*$, the "easiest" instance is unique up to a change in the signs of the components of the global solution u^*. In the next theorem, we show that a similar property as α^* holds if we set α to be

$$\alpha^\diamond := n/(n+2).$$

Theorem 18. *Suppose that $n \geq 5$ and the instance $\mathcal{MC}(C, u^*)$ satisfies*

$$\mathbb{D}_{\alpha^\diamond}(C, u^*) = \mathbb{D}_\alpha^{min} = n(n+2)/[4(n-1)].$$

Then, it holds that

$$|u_i^*| = 1/n, \quad \forall i \in [n]; \quad C = n^{-2} I_n.$$

Since the proof is similar to that of Theorem 17, we omit it for brevity. The above theorem implies that the weight matrix C of the "easiest" instances is a constant multiple of the identity matrix I_n, which satisfies the δ-RIP$_{2,2}$ condition with $\delta = 0$. This is consistent with the common sense that the RIP constant δ being 0 is the optimal situation. Hence, Theorem 18 suggests that the choice $\alpha^\diamond = n/(n+2)$ may potentially be the optimal choice of α. On the other hand, we will prove in Section 3.4 that the "easiest" instances in Theorems 17 and 18 all have a benign landscape in the sense that they satisfy the strict-saddle property [210], which guarantees the polynomial-time global convergence of various algorithms. If the weight α is different from α^* and $\alpha^\diamond$, there may exist multiple "essentially" different instances attaining the minimum complexity.

3.3 Connections to Existing Results

In this section, we provide estimates of the proposed complexity metric on two well-studied problem instances and a synthetic problem. More specifically, we consider matrix sensing problems satisfying the RIP condition and matrix completion problems under the Bernoulli sampling model. In addition, we construct a class of instances parameterized by a single parameter. We estimate the threshold of the parameter that separates instances with a desirable optimization landscape from those with a bad landscape. The results in the synthetic example show that our proposed complexity metric has the potential to provide guarantees on the optimization landscape when the RIP condition fails.

Matrix Sensing Problem: RIP Condition

We first consider instances of problem (3.4) that satisfy the δ-RIP$_{2,2}$ condition, where $\delta \in [0, 1)$ is the RIP constant. However, the constraint that $C \in \mathbb{S}_{+,1}^{n^2-1}$ is inconsistent with the RIP condition (2.8) in the sense that the entries of C are averagely on the scale of n^{-2}, but the RIP condition requires that the entries of C be on the scale of $O(1)$. Therefore, we generalize the definition of the RIP condition to deal with the inconsistent scaling:

Definition 6. Given natural numbers r and s, the function $f(\cdot; M^*)$ is said to satisfy the **Restricted Isometry Property** (RIP) of rank $(2r, 2s)$ for a constant $\delta \in [0, 1)$, denoted as δ-RIP$_{2r,2s}$, if there exist constants $c_1, c_2 \geq 0$ such that $c_2/c_1 = (1+\delta)/(1-\delta)$ and

$$c_1\|K\|_F^2 \leq \left[\nabla^2 f(M; M^*)\right](K, K) \leq c_2\|K\|_F^2 \tag{3.7}$$

holds for all matrices $M, K \in \mathbb{R}^{n\times n}$ such that $\mathrm{rank}(M) \leq 2r, \mathrm{rank}(K) \leq 2s$.

The above definition of the RIP condition is scale-free in the sense that for any constant $c > 0$, the function $cf(\cdot; M^*)$ satisfies the δ-RIP$_{2r,2s}$ condition if and only if $f(\cdot; M^*)$ satisfies the same condition.

Since the instances satisfying the RIP condition have a benign optimization landscape, we expect that the complexity metric is upper-bounded for those instances. By suitably generalizing the definitions of $\mathbb{D}_\alpha(C, u^*)$ and $\mathcal{D}$, we provide an upper bound for problem (2.2) under the RIP condition. Note that the ground truth M^* is not necessarily rank-1 in this part. Instead, we assume that $M^* = U^*(U^*)^T$ is rank-r, where U^* belongs to $\mathbb{R}^{n\times r}$. For problem (2.2), each instance is defined by the loss function $f(\cdot;\cdot)$ and the ground truth M^*. We assume that the M^* is a global optimum of the loss function, namely,

$$f(M^*; M^*) = \min_{K\in\mathbb{R}^{n\times n}} f(K; M^*), \quad \forall M^* \in \mathbb{R}^{n\times n} \quad \text{s. t. } M^* \succeq 0, \ \mathrm{rank}(M^*) = r. \tag{3.8}$$

In the special case when $f(\cdot;\cdot)$ is the weighted ℓ_2-loss function in (3.4), the above condition implies that $C_{ij} \geq 0$ for all $i, j \in [n]$. Similar to the normalization constraint $C \in \mathbb{S}_{+,1}^{n^2-1}$, we assume that objective function $f(\cdot; M^*)$ is normalized in the sense that

$$\sum\nolimits_{i,j\in[n]} [f(M^* + E_{ij}; M^*) - f(M^*; M^*)] = 1. \tag{3.9}$$

For the normalization constraint $u^* \in \mathbb{S}_1^{n-1}$, we assume that the global truth M^* satisfies

$$\|U^*\|_1 = 1. \tag{3.10}$$

The set of degenerate instances is given by

$$\mathcal{D} := \Big\{ (f, M^*) \;\Big|\; f(\cdot;\cdot) \text{ and } M^* \text{ satisfy (3.8)-(3.10)}, \\ \exists M \neq M^* \quad \text{s. t. } f(M; M^*) = f(M^*; M^*), \ M^* \succeq 0, \ \mathrm{rank}(M^*) = r \Big\}.$$

The "entry-wise ℓ_1-norm" between two arbitrary functions $h^1(\cdot)$ and $h^2(\cdot)$ with the domain $\mathbb{R}^{n\times n}$ is defined as the restricted ℓ_∞-Lipschitz constant of $h^1 - h^2$. Namely, we define $\|h^1 - h^2\|_1$ to be

$$\|h^1 - h^2\|_1 := \sup_{K,L\in\mathbb{R}^{n\times n}} \frac{|(h^1(K) - h^2(K)) - (h^1(L) - h^2(L))|}{\max_{i,j\in[n]}(K_{ij} - L_{ij})^2}$$

$$\text{s.t. } K \neq L, \quad \text{rank}(K-L) \leq 2r.$$

For every constant $\alpha \in [0,1]$, the distance between two instances (f, M^*) and $(\tilde{f}, \tilde{M}^*)$ is defined as

$$\text{dist}_\alpha \left[(f, M^*), (\tilde{f}, \tilde{M}^*) \right] := \alpha \| f(\cdot; M^*) - \tilde{f}(\cdot; \tilde{M}^*) \|_1 + (1-\alpha) \| U^* - \tilde{U}^* \|_1,$$

where $U^*, \tilde{U}^* \in \mathbb{R}^{n \times r}$ satisfy $U^*(U^*)^T = M^*$ and $\tilde{U}^*(\tilde{U}^*)^T = \tilde{M}^*$. Finally, the complexity metric is given by

$$\mathbb{D}_\alpha(f, M^*) := \left[\inf_{(\tilde{f}, \tilde{M}^*) \in \mathcal{D}} \text{dist}_\alpha \left[(f, M^*), (\tilde{f}, \tilde{M}^*) \right] \right]^{-1}. \tag{3.11}$$

We note that the definitions of $\mathcal{D}$ and $\mathbb{D}_\alpha(f, M^*)$ are consistent with those of instance (3.4). The following theorem provides an upper bound on the complexity metric of any instance satisfying the $\text{RIP}_{2,2}$ condition.

Theorem 19. *Let $\alpha \in [0,1]$ and $\delta \in [0,1)$ be two constants. Suppose that the function $f(\cdot; M^*)$ satisfies the δ-RIP$_{2r,2r}$ condition and the normalization constraint* (3.9)*, where r is the rank of M^*. Then, it holds that*

$$\mathbb{D}_\alpha(f, M^*) \leq \frac{n^2(1+\delta)}{\alpha(1-\delta)}.$$

Proof. We fix the instance (f, M^*) and assume that $(\tilde{f}, \tilde{M}^*) \in \mathcal{D}$. Suppose that the matrix $M \neq \tilde{M}^*$ satisfies

$$\tilde{f}(M; \tilde{M}^*) = \tilde{f}(\tilde{M}^*; \tilde{M}^*).$$

We first consider the case when $M \neq M^*$. In this case, we can estimate that

$$\begin{aligned}
&\| f(\cdot; M^*) - \tilde{f}(\cdot; \tilde{M}^*) \|_1 \qquad (3.12)\\
&\geq \frac{\left| \left[f(M; M^*) - \tilde{f}(M; \tilde{M}^*) \right] - \left[f(M^*; M^*) - \tilde{f}(M^*; \tilde{M}^*) \right] \right|}{\max_{i,j \in [n]} (M_{ij} - M^*_{ij})^2} \\
&= \frac{\left| [f(M; M^*) - f(M^*; M^*)] + \left[\tilde{f}(M^*; \tilde{M}^*) - \tilde{f}(M; \tilde{M}^*) \right] \right|}{\max_{i,j \in [n]} (M_{ij} - M^*_{ij})^2} \\
&= \frac{\left| [f(M; M^*) - f(M^*; M^*)] + \left[\tilde{f}(M^*; \tilde{M}^*) - \tilde{f}(\tilde{M}^*; \tilde{M}^*) \right] \right|}{\max_{i,j \in [n]} (M_{ij} - M^*_{ij})^2} \\
&\geq \frac{f(M; M^*) - f(M^*; M^*)}{\max_{i,j \in [n]} (M_{ij} - M^*_{ij})^2} \geq \frac{(c_1/2) \cdot \| M - M^* \|_F^2}{\max_{i,j \in [n]} (M_{ij} - M^*_{ij})^2} \geq \frac{c_1}{2},
\end{aligned}$$

where c_1 is the constant in the RIP condition of $f(\cdot; M^*)$. The second inequality is due to

$$f(M; M^*) - f(M^*; M^*) \geq 0, \quad \tilde{f}(M^*; \tilde{M}^*) - \tilde{f}(\tilde{M}^*; \tilde{M}^*) \geq 0.$$

The second last inequality follows from the global optimality of M^* and the second inequality after inequality (12) in [244], namely,

$$f(M; M^*) \geq f(M^*; M^*) + \frac{c_1}{2}\|M - M^*\|_F^2, \quad \forall M \in \mathbb{R}^{n\times n}, \ \operatorname{rank}(M) \leq r.$$

Now, we provide a lower bound on c_1. Using the normalization constraint (3.9) and the stationarity of M^*, it holds that

$$1 = \sum_{i,j\in[n]} [f(M^* + E_{ij}; M^*) - f(M^*; M^*)] \leq \frac{c_2}{2} \cdot \sum_{i,j\in[n]} \|E_{ij}\|_F^2 = \frac{c_2 n^2}{2},$$

which implies that $c_2 \geq 2n^{-2}$. Using the relation $c_2/c_1 = (1+\delta)/(1-\delta)$, we obtain that

$$c_1 \geq \frac{2(1-\delta)}{n^2(1+\delta)}.$$

By substituting into inequality (3.12), it follows that

$$\|f(\cdot; M^*) - \tilde{f}(\cdot; \tilde{M}^*)\|_1 \geq \frac{1-\delta}{n^2(1+\delta)}.$$

which leads to $\operatorname{dist}_\alpha[(f, M^*), (\tilde{f}, \tilde{M}^*)] \geq \alpha(1-\delta)/[n^2(1+\delta)]$. Now, the desired bound on $\mathbb{D}_\alpha(f, M^*)$ follows from taking the inverse. In the case when $M = M^*$, we can replace M with $\tilde{M}^*$ and the proof can be done in the same way. □

We note that the upper bound on $\mathbb{D}_\alpha(C, u^*)$ is increasing in δ, which is consistent with the intuition that a smaller δ will lead to a better optimization landscape. Moreover, in the case when $\alpha(1-\delta) = \Theta(1)$, the upper bound is on the order of $O(n^2)$, which is $O(n)$ larger than the minimum possible complexity metric in Theorem 16. Now, we provide a remedy to the aforementioned issue for problem (3.4). With the knowledge about the incoherence of the global solution, we can improve the upper bound on the complexity metric.

Theorem 20. *Suppose that the instance $\mathcal{MC}(C, u^*)$ satisfies the δ-RIP$_{2,2}$ condition and u^* has incoherence μ. Then, it holds that*

$$\mathbb{D}_\alpha(C, u^*) \leq \max\left\{\frac{n(1+\delta)}{4\alpha(1-\delta)}, \frac{1}{2(1-\alpha)\mu}\right\} \times \min\left\{\left(\frac{1}{\mu} - \frac{1}{n}\right)^{-1}, 3\mu\right\}.$$

The proof of Theorem 20 can be found in Appendix 3.C. From the above theorem, we can use the weight α to control the balance between the RIP constant δ and the incoherence μ. If we choose $1-\alpha = \Theta(n^{-1})$, then the complexity can be upper-bounded by

$$\mathbb{D}_\alpha(C, u^*) = \mu n \cdot \max\left\{O\left(\frac{1+\delta}{1-\delta}\right), O\left(\frac{1}{\mu}\right)\right\} = O\left(\mu n \cdot \frac{1+\delta}{1-\delta}\right).$$

In addition, if it holds that $\mu = O(1)$ and $(1-\delta)^{-1} = O(1)$, then the complexity is upper-bounded by $O(n)$, which matches the minimum possible complexity in Theorem 16 up to a constant. Although the complexity metric may have a large value for extreme instances (i.e., instances with a large incoherence), the complexity of regular instances achieves the optimal value up to a constant. Furthermore, we conjecture in Section 3.4 that the condition $\mathbb{D}_\alpha(C, u^*) = O(n\mu/\alpha)$ is sufficient to guarantee the success of local search methods. Assuming that this conjecture is true, then the condition $(1-\delta)^{-1} = O(1)$ alone is sufficient to guarantee that the optimization landscapes are benign regardless of the value of the incoherence μ. This is consistent with the existing results on the RIP condition. We conclude the discussion of instances with the RIP condition by showing that the dependence of δ in Theorem 20 is tight up to a constant.

Theorem 21. *Suppose that $n \geq 4$, $\alpha \in [0,1]$, $\mu \in [1,n]$ and $\delta \in [0,1)$. Let $\ell := \lceil n/\mu \rceil$. Then, there exists an instance $\mathcal{MC}(C, u^*)$ such that $\mathcal{MC}(C, u^*)$ satisfies the δ-RIP$_{2,2}$ condition, u^* has incoherence μ and*

$$\mathbb{D}_\alpha(C, u^*) \geq \frac{n(1+\delta)}{4\alpha(1-\delta)} \cdot \min\left\{\frac{n\mu}{\mu\ell - \mu}, \mu\right\}.$$

The proof of Theorem 21 can be found in Appendix 3.C.

Matrix Completion Problem: Bernoulli Model and Incoherence Condition

Next, we consider instances $\mathcal{MC}(C, u^*)$ of problem (3.4) where the global solution u^* is μ-incoherent and the random weight matrix C obeys the Bernoulli model. Similar to the RIP condition, we need to generalize the definition of the Bernoulli model under the normalization constraint.

Definition 7. Given the sampling rate $p \in (0,1]$, a random matrix $C \in \mathbb{S}^{n^2-1}_{+,1}$ is said to obey the **Bernoulli model** if

$$C_{ij} = \frac{\delta_{ij}}{\sum_{k,\ell \in [n]} \delta_{k\ell}}, \quad \forall i, j \in [n],$$

where $\{\delta_{k\ell} | k, \ell \in [n]\}$ are independent Bernoulli random variables with the parameter p.

We note that the above model is well defined only when $\sum_{i,j} \delta_{ij} > 0$, which happens with probability $1 - (1-p)^{n^2} \geq 1 - \exp(-n^2 p)$. This probability is sufficiently large if $n^2 p \gg 1$. In [35], the authors showed that $p \geq \Theta(\mu \log n/n)$ is necessary and under this condition, the success probability is at least $1 - O(n^{-\mu n})$. Therefore, we only focus on the case when the event $\sum_{i,j} \delta_{ij} > 0$ happens. In the existing literature [34, 87, 40], the instances obeying the Bernoulli model are proven to have no spurious local minima. We show that our complexity metric is able to characterize this property by proving an upper bound on the complexity metric.

Theorem 22. *Given $\mu \in [1, n]$ and $p \in (0, 1]$, suppose that the weight matrix C obeys the Bernoulli model with the parameter p and that u^* has incoherence μ. If $\eta > 2$ is a constant and the sampling rate satisfies*

$$p \geq \min\left\{1, \frac{16(1+\eta\mu)\log n + 16}{n}\right\},$$

it holds with probability at least $1 - 3n^{-\eta/2+1}$ that

$$\mathbb{D}_\alpha(C, u^*) \leq \max\left\{\frac{3n}{4\alpha}, \frac{1}{2(1-\alpha)\mu}\right\} \times \min\left\{\left(\frac{1}{\mu} - \frac{1}{n}\right)^{-1}, 3\mu\right\}.$$

The proof of Theorem 22 can be found in Appendix 3.C. By Theorem 22, if $1 - \alpha = \Theta(n^{-1}\mu^{-1})$, then the complexity of instances obeying the Bernoulli model is on the order of $\Theta[n^2\mu/(n-\mu)]$. If the incoherence $\mu = O(1)$, the complexity is on the order of $O(n)$, which matches the minimum possible complexity up to a constant. Therefore, the proposed metric can also serve as a good indicator for the matrix completion problem with the Bernoulli model. Finally, we note that the bound $p \geq \Theta(\mu \log n/n)$ is optimal up to a constant [35]; see also the discussions in Appendix E of [75].

Finally, we note that problem (3.4) may still have spurious local minima when the sampling probability p and the incoherence μ satisfy the condition in Theorem 22. In the existing literature, the global convergence of randomly initialized local search methods is established for problem (3.4) only under an extra regularizer or an extra constraint on the incoherence of u. That being said, our proposed complexity metric correctly reflects the commonsense that the matrix completion problem is generally easier to solve when the incoherence is small or when the sampling rate p is large. When the complexity is small, it is possible to apply local search methods to find the ground truth. The local search methods may be different for different classes of low-rank matrix optimization problems. In addition, the new complexity metric has the advantage that it is able to simultaneously capture the RIP condition, the incoherence condition and potentially other existing complexity metrics.

One-parameter Class of Instances

In Sections 3.3 and 3.3, we provided several upper bounds on the complexity metric. In this part, we consider a class of instances that are parameterized by a single parameter

$\epsilon \in [0,1]$. Intuitively, when the parameter grows from 0 to 1, the optimization landscape of the instance becomes more benign. Unlike the previous results in this section, the analysis of the small parameter case provides necessary conditions for the existence of spurious local minima. More specifically, we fix $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ to be an unweighted undirected graph without self-loops, where the node set is $\mathbb{V} = [n]$. We consider the maximal independent set of $\mathbb{G}$, which is defined as follows:

Definition 8. For an undirected graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, a set $\mathcal{S} \subset \mathbb{V}$ is called an *independent set* if no two nodes in $\mathcal{S}$ are adjacent. The set $\mathcal{S}$ is called a *maximal independent set* if it is an independent set with the maximum number of nodes [1].

Suppose that $\mathcal{S} \subset [n]$ is a maximal independent set of $\mathbb{G}$. For every $\epsilon \in [0,1]$, the instance $\mathcal{MC}(C^\epsilon, u^*)$ is defined by

$$\begin{aligned} &C^\epsilon_{ij} := \epsilon/Z_\epsilon, \quad \forall i,j \in \mathcal{S} \quad \text{s.t. } i \neq j; \quad C^\epsilon_{ij} := 1/Z_\epsilon, \quad \text{if } \{i,j\} \in \mathbb{E}; \\ &C^\epsilon_{ii} := 1/Z_\epsilon, \quad \forall i \in [n], \quad C^\epsilon_{ij} := 0, \quad \text{otherwise}, \\ &u^*_i := 1/m, \quad \forall i \in \mathcal{S}; \quad u^*_i := 0, \quad \forall i \notin \mathcal{S}, \end{aligned} \tag{3.13}$$

where $m := |\mathcal{S}|$ and $Z_\epsilon := 2|\mathbb{E}| + n + m(m-1)\epsilon$ is the normalization constant. In the remainder of this subsection, we assume without loss of generality that $\mathcal{S} = [m]$.

First, we study for what values of ϵ the instance $\mathcal{MC}(C^\epsilon, u^*)$ has benign landscape or has spurious local minima. The following theorem guarantees that the threshold $\epsilon = \Theta(m^{-1}) = \Theta(\mu/n)$ separates the regimes where the instance possesses and does not possess spurious local minima, where $\mu := n/m$ denotes the incoherence of u^*.

Theorem 23. *If $\epsilon \geq \Theta(m^{-1})$, the instance $\mathcal{MC}(C^\epsilon, u^*)$ does not have spurious second-order critical points[2] (SSCPs), namely, all second-order critical points are global minima associated with the ground truth solution M^*. If $\epsilon = O(m^{-1})$, the instance $\mathcal{MC}(C^\epsilon, u^*)$ has at least $O(2^{m/2})$ spurious local minima.*

The proof of Theorem 23 can be found in Appendix 3.C. In the case when $m = 2$, the proof of Theorem 23 (more specifically, Theorem 31) states that the instance $\mathcal{MC}(C^\epsilon, u^*)$ has spurious local minima if $\epsilon < 1/3$. The condition $\epsilon = 1/3$ corresponds to the δ-RIP$_{2,2}$ condition holding with $\delta = 1/2$. Therefore, the RIP constant $\delta \leq 1/2$ is necessary for the instance $\mathcal{MC}(C^\epsilon, u^*)$ to have no spurious local minima. Combined with the results in [244, 24], we can see that the one-parameter group $\mathcal{MC}(C^\epsilon, u^*)$ also contains difficult instances of the general problem (2.2).

Furthermore, we note that the constants in the proof of Theorem 23 are not optimal. We conjecture that the instance $\mathcal{MC}(C^\epsilon, u^*)$ has spurious solutions if $\epsilon < (m+1)^{-1} + o(m^{-1})$

[1] We note that this definition is different from the common definition of maximum independent set, which only requires that a maximum independent set is not a proper subset of an independent set.

[2] A point $u \in \mathbb{R}^n$ is called a spurious second-order critical point if it satisfies the first-order and the second-order necessary optimality conditions and $uu^T \neq u^*(u^*)^T$.

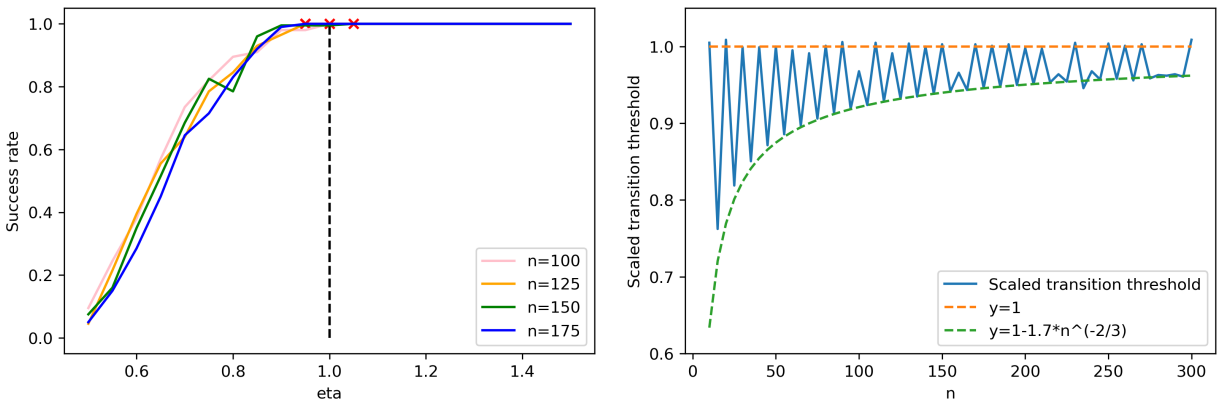


Figure 3.3.1: The left plot shows the transitions of the success rate of the gradient descent algorithm when $n = 100, 125, 150, 175$. The red "$\times$" sign refers to the *transition threshold*, i.e., the smallest value of η that attains 100% success rate. In the right plot, the transition thresholds of η are compared with the curves $y = 1$ and $y = 1 - 1.7(n+1)^{-2/3}$.

and does not have spurious solutions if $\epsilon > (m+1)^{-1} + o(m^{-1})$. We numerically verify this conjecture in the special case when $m = n$. In numerical examples, we consider the scaled parameter $\eta := (n+1)\epsilon$. For each instance, we implement the randomly initialized gradient descent algorithm for 200 times and check the number of implements for which the distance between the last iterate and $\pm u^*$ has Frobenius norm at most 10^{-5}. The results are plotted in Figure 3.3.1. In the left plot, we can see that in most cases, the success rate grows with the parameter η, which is proportional to ϵ. This indicates that the optimization landscape becomes more benign when ϵ is larger. In addition, the transition thresholds of η are very close to 1 (to be more accurate, the thresholds of η are between 0.95 and 1.05). This observation is consistent with our conjecture. In the right plot, we compare the transition thresholds of η against the constant number 1. We observe that the thresholds are approximately located between 1 and $1 - 1.7(n+1)^{-2/3}$, which implies that the original thresholds of ϵ are between $(n+1)^{-1}$ and $(n+1)^{-1} - 1.7(n+1)^{-5/3}$. Hence, the thresholds become close to $(n+1)^{-1}$ when n is large, which is also consistent with our conjecture. Moreover, we can see that the threshold of η is not monotone in n and is slightly smaller when n is odd.

Finally, we transform the estimates on the parameter ϵ to the complexity metric.

Theorem 24. *Suppose that $n \geq m \geq 36$, $\alpha \in [0,1]$ and $\epsilon \in [0,1]$. Then, the following statements hold true:*

1. *If*

$$\mathbb{D}_\alpha(C^\epsilon, u^*) \leq \left[\frac{36\alpha}{n^2} + \min\left\{72\alpha \cdot \frac{m}{n^2}, 2(1-\alpha)\right\}\right]^{-1},$$

then the instance $\mathcal{MC}(C^\epsilon, u^*)$ has no spurious local minima;

2. If

$$\mathbb{D}_\alpha(C^\epsilon, u^*) \geq \frac{18}{17} \max\left\{\frac{13n^2}{2\alpha}, \frac{1}{2(1-\alpha)}\right\},$$

then the instance $\mathcal{MC}(C^\epsilon, u^*)$ has spurious local minima.

The proof of Theorem 24 can be found in Appendix 3.C. In the case when $1-\alpha \geq \Theta(m/n^2)$, the upper bound on $\mathbb{D}_\alpha(C^\epsilon, u^*)$ is on the order of $O(n\mu/\alpha)$, where $\mu := n/m$ is the incoherence of u^*. This result is consistent with the upper bounds in Sections 3.3 and 3.3. In addition, the RIP constant is $1 - O(1/m)$ if $\epsilon = O(1/m)$, which shows that the proposed complexity metric can provide better guarantees on the optimization complexities than the RIP constant. On the other hand, the lower bound in Theorem 24 is on the order of $O(n^2/\alpha)$ in the case when $1-\alpha \geq \Theta(n^{-2})$.

In summary, we have provided a consistent upper bound on the complexity metric that is on the order of $\Theta(n\mu/\alpha)$ for all three examples ($\Theta[n\mu/\alpha \cdot (1+\delta)/(1-\delta)]$ for the RIP case) if we choose $1-\alpha = O(n^{-1})$. These theoretical results provide strong evidence that our proposed complexity metric is able to capture the properties of the optimization landscape for several different models, even when other existing conditions fail to provide theoretical guarantees; see the comparison of the condition and our complexity metric in Section 3.3. In Section 3.4, we make some conjectures based on these observations and provide a partial theoretical explanation.

3.4 Theoretical Results for General Instances

In this section, we provide a theoretical analysis for the proposed complexity metric (3.6) on the general problem (3.4). Intuitively, we expect the problem (3.4) to have a benign landscape when the complexity metric is small and vice versa. We first prove that the proposed complexity metric is able to provide a sufficient condition on the absence of SSCPs of problem (3.4). Then, we construct another complexity metric that lower-bounds the metric (3.5) and show that the alternative complexity metric is able to provide necessary conditions on the absence of SSCPs.

Recalling the analysis in Section 3.3, one might have the following questions: Suppose that $1-\alpha \geq \Theta(n^{-1})$ and the solution u^* is μ-incoherent. Can we find two constants $\delta, \Delta > 0$ such that

1. If $\mathbb{D}_\alpha(C, u^*) \leq \delta\mu n/\alpha$, the instance $\mathcal{MC}(C, u^*)$ has no SSCPs;
2. If $\mathbb{D}_\alpha(C, u^*) \geq \Delta n^2/\alpha$, the instance $\mathcal{MC}(C, u^*)$ has SSCPs?

Suppose that the first property in the above question holds. The results in Section 3.3 imply that the proposed complexity metric guarantees the absence of SSCPs when the RIP constant is $O[(\delta-1)/(\delta+1)]$, which is independent of μ. In addition, the matrix completion problem under the Bernoulli model does not have SSCPs when $p \geq O(\mu \log n/n)$, which

matches the lower bound in [35]. In Section 3.4, we prove a weaker version of the first property in the case when α is equal to α^* or $\alpha^\diamond$, which are defined in Section 3.2. We note that both α^* and $\alpha^\diamond$ satisfy the condition that $1-\alpha = \Theta(n^{-1})$. On the other hand, in Section 3.4, we refute the second property in the above question by constructing counterexamples. This observation implies that similar to the RIP constant and the incoherence, the proposed complexity metric cannot provide necessary conditions on the absence of spurious local solutions. However, if we substitute the degenerate set $\mathcal{D}$ with a slightly smaller set, we prove that the complexity metric is able to provide a necessary condition.

Small Complexity Case

We first consider instances with a small complexity metric. In the case when α is equal to α^* or $\alpha^\diamond$, we prove that $\mathbb{D}_\alpha(C, u^*) \leq \delta n/\alpha$ serves as a sufficient condition for the absence of SSCPs, where $\delta > 0$ is an absolute constant. Since the incoherence μ is at least 1, the aforementioned condition is weaker than the first property in the aforementioned question. By Theorem 16, the minimum possible value of the complexity metric is on the order of $O(n/\alpha)$. In this subsection, we show that the constant δ can be chosen such that $\delta n/\alpha$ is strictly larger than the minimum possible complexity. The following theorem deals with the case when $\alpha = \alpha^*$.

Theorem 25. *Suppose that $n \geq 5$ and $\alpha = \alpha^*$. Then, there exists a constant $\delta > 1/4$ such that for every instance $\mathcal{MC}(C, u^*)$ satisfying*

$$\mathbb{D}_{\alpha^*}(C, u^*) \leq \delta n/\alpha^*,$$

the instance $\mathcal{MC}(C, u^*)$ does not have any SSCPs.

Since the minimum possible complexity metric is $n/(4\alpha^*)$, the upper bound in Theorem 25 is *non-trivial* in the sense that there exist instances satisfying the inequality. By Theorem 17, the minimum complexity metric $n/(4\alpha^*)$ is only attained by instances in $\mathcal{M}$, where

$$\mathcal{M} := \left\{ (C, u^*) \;\middle|\; |u_i^*| = \frac{1}{n},\ C_{ii} = 0,\ \forall i \in [n],\ C_{ij} = \frac{1}{n(n-1)},\ \forall i, j \in [n],\ i \neq j \right\}.$$

In the next lemma, we prove the strict-saddle property [210] of the ℓ_1-norm for instances in $\mathcal{M}$, which can be viewed as a robust version of the absence of SSCPs.

Lemma 9. *Suppose that $n \geq 2$ and $(C^0, u^0) \in \mathcal{M}$. Then, there exist a positive constant η_0 and two positive-valued functions $\beta(\eta)$ and $\gamma(\eta)$ such that for all $\eta \in (0, \eta_0]$ and $u \in \mathbb{R}^n$, at least one of the following properties holds:*

1. $\min\{\|u - u^*\|_1, \|u + u^*\|_1\} \leq \eta$;
2. $\|\nabla g(u; C, u^*)\|_\infty \geq \beta(\eta)$;

3. $\lambda_{min}[\nabla^2 g(u; C, u^*)] \leq -\gamma(\eta)$.

We then show that after a sufficiently small perturbation to any point $(C^0, x^0) \in \mathcal{M}$, the new instance does not have any SSCPs.

Lemma 10. *Suppose that $n \geq 3$. There exists a small positive constant ϵ such that for every pair $(C^0, u^0) \in \mathcal{M}$ and $(\tilde{C}, \tilde{u}^*)$ satisfying*

$$\alpha^* \|\tilde{C} - C^0\|_1 + (1 - \alpha^*)\|\tilde{u}^* - u^0\|_1 < \epsilon,$$

the instance $\mathcal{MC}(\tilde{C}, \tilde{u}^*)$ does not have SSCPs.

The proofs of the last two lemmas involve several standard calculations and can be found in Appendices 3.D and 3.D. Now, we prove the existence of a non-trivial upper bound on the metric.

Proof of Theorem 25. Let ϵ be the constant in Lemma 10. We consider the compact set

$$\mathcal{C} := \Big\{ (C, u^*) \;\Big|\; \|C\|_1 = \|u^*\|_1 = 1, \\ \alpha^* \|\tilde{C} - C^0\|_1 + (1 - \alpha^*)\|\tilde{u}^* - u^0\|_1 \geq \epsilon, \quad \forall (C^0, u^0) \in \mathcal{M} \Big\}.$$

Since the minimum possible complexity metric $n/(4\alpha^*)$ is only attained by points in $\mathcal{M}$, it holds that

$$\mathbb{D}_{\alpha^*}(\mathcal{C}) := \max_{(C,u^*)\in\mathcal{C}} \mathbb{D}_{\alpha^*}(C, u^*) > n/(4\alpha^*).$$

Therefore, choosing

$$\delta := (\alpha^*/n) \cdot \mathbb{D}_{\alpha^*}(\mathcal{C}) > 1/4,$$

we have

$$\mathbb{D}_{\alpha^*}(C, u^*) \leq \delta n/\alpha^* \implies (C, u^*) \notin \mathcal{C} \implies \text{the instance } \mathcal{MC}(C, u^*) \text{ has no SSCPs.}$$

This completes the proof. □

The case when $\alpha = \alpha^\diamond$ can be analyzed in a similar way. We note that the strict-saddle property of the instances in Theorem 18 has been established in [125]. Hence, we present the results in the following theorem and omit the proof.

Theorem 26. *Suppose that $n \geq 5$ and $\alpha = \alpha^\diamond$. Then, there exists a constant $\delta > 1/4$ such that for every pair (C, u^*) satisfying*

$$\mathbb{D}_{\alpha^\diamond}(C, u^*) \leq \delta n(n + 2)/(n + 1),$$

the instance $\mathcal{MC}(C, u^*)$ does not have any SSCPs.

Similar to Theorem 25, since the minimum possible complexity metric is attained with $\delta = 1/4$, the upper bound in Theorem 26 is non-trivial.

Large Complexity Case

In this subsection, we first refute the second property in the question that we asked in the beginning of Section 3.4 and then refine its statement to make it hold true. We note that the RIP condition and the incoherence condition cannot provide necessary conditions for the absence of SSCPs either. Namely, there exist instances that satisfy the δ-RIP$_{2,2}$ condition with δ as high as 1 which do not have SSCPs. Similarly, in the case when the incoherence of the global solution is n, it is still possible to have an instance of the matrix completion problem without any SSCPs. In other words, although small values for the RIP constant and incoherence guarantee the absence of spurious solutions, these notions cannot capture the complexity of the problem since there are low-complexity problems with large values for these parameters. We first show that our new metric suffers from the same shortcoming, but we then propose a simple refinement to address this issue.

Example 5. Suppose that the weight matrix and the ground truth are

$$C^\delta := \frac{1}{1+3\delta}\begin{bmatrix}1 & \delta\\ \delta & \delta\end{bmatrix}, \quad u^* := \begin{bmatrix}1\\ 0\end{bmatrix},$$

where $\delta \geq 0$ is a constant. One can verify that $\pm u^*$ are the only local minima to the instance $\mathcal{MC}(C^\delta, u^*)$ for all $\delta > 0$. However, in the case when $\delta = 0$, the instance $\mathcal{MC}(C^0, u^*)$ has the set of global solutions

$$\pm\begin{bmatrix}1\\ c\end{bmatrix}, \quad \forall c \in \mathbb{R}.$$

Moreover, we consider the case when both components of u^* are measured, where the instance $\mathcal{MC}(\tilde{C}^\epsilon, \tilde{u}^\epsilon)$ is defined by

$$\tilde{C}^\epsilon := \frac{1}{1+\epsilon}\begin{bmatrix}1 & 0\\ 0 & \epsilon\end{bmatrix}, \quad \tilde{u}^\epsilon := \frac{1}{1+\epsilon}\begin{bmatrix}1\\ \epsilon\end{bmatrix},$$

where ϵ is a positive constant. One can verify that the pair $(\tilde{C}^\epsilon, \tilde{u}^\epsilon)$ belongs to $\mathcal{D}$ for all $\epsilon > 0$. Setting δ and ϵ to be small enough, the instances $\mathcal{MC}(C^\delta, u^*)$ and $\mathcal{MC}(\tilde{C}^\epsilon, \tilde{u}^\epsilon)$ can be arbitrarily close to each other in the sense that

$$\alpha\|C^\delta - \tilde{C}^\epsilon\|_1 + (1-\alpha)\|u^* - \tilde{u}^\epsilon\|_1 = O(\alpha\delta + \epsilon).$$

Therefore, the complexity metric of $\mathcal{MC}(C^\delta, u^*)$ can be arbitrarily large. This example shows that instances without SSCPs can be arbitrarily close to those in $\mathcal{D}$, which have non-unique global solutions.

Nevertheless, we derive a lower bound on the complexity metric (3.6) by constructing a subset of $\mathcal{D}$, which allows obtaining a necessary condition. Intuitively, if an instance has multiple global minima, these global minima are still locally optimal after a sufficiently small perturbation to the instance. To ensure the "robustness" of the local optimality, we require

the positive-definiteness of the Hessian matrix. For each instance $\mathcal{MC}(C, u^*)$, let $\mathbb{G}_{1k}(C, u^*)$ for all $k \in [n_1]$ be the connected components of $\mathbb{G}_1(C, u^*)$, where n_1 is the number of connected components. Moreover, we use $\mathcal{I}_{1k}(C, u^*)$ to denote the node set of $\mathbb{G}_{1k}(C, u^*)$ for all $k \in [n_1]$. We define the following subset of $\mathcal{D}$:

$$\begin{aligned}\mathcal{SD} := \{(C, u^*) \in \mathcal{D} \mid \mathbb{G}_{1k}(C, u^*) \text{ is not bipartite for all } k \in [n_1],\\ \mathbb{G}_1(C, u^*) \text{ is disconnected}, \quad \mathcal{I}_{00}(C, u^*) = \emptyset\}.\end{aligned}$$

The following theorem provides a characterization of the Hessian matrix at global solutions for pairs in $\mathcal{SD}$.

Theorem 27. *Suppose that $(C, u^*) \in \mathcal{SD}$. Then, the Hessian matrix is positive definite at all global solutions of the instance $\mathcal{MC}(C, u^*)$.*

The proof of Theorem 27 can be found in Appendix 3.D. Using the positive-definiteness of the Hessian matrix, we are able to apply the implicit function theorem to guarantee the existence of spurious local minima in a neighbourhood of each instance in $\mathcal{SD}$; see Appendix 3.D for more details. The global guarantee can be established by considering closed subsets of $\mathcal{SD}$. For every constant $\epsilon \geq 0$, we consider the closed subset $\mathcal{SD}_\epsilon$, which is defined as

$$\mathcal{SD}_\epsilon := \big\{(C, u^*) \in \mathcal{SD} \mid C_{ij} \in \{0\} \cup [\epsilon, 1], \quad \forall i, j \in [n], \quad |u_i^*| \in \{0\} \cup [\epsilon, 1], \quad \forall i \in [n]\big\}.$$

Basically, the extra condition in the definition of $\mathcal{SD}_\epsilon$ requires that the nonzero components of C and u^* be at least ϵ. We can verify that the set $\mathcal{SD}_\epsilon$ is a compact set and for every $\epsilon_n \to 0$, it holds that

$$\lim_{n\to\infty} \cup_{i=1}^n \mathcal{SD}_{\epsilon_i} = \mathcal{SD}_0 = \mathcal{SD}.$$

Now, we define the alternative complexity metric

$$\mathbb{D}_{\alpha,\epsilon}(C, u^*) := \left[\min_{(\tilde{C}, \tilde{u}^*) \in \mathcal{SD}_\epsilon} \alpha\|C - \tilde{C}\|_1 + (1-\alpha)\|u^* - \tilde{u}^*\|_1\right]^{-1}. \tag{3.14}$$

Since $\mathcal{SD}_\epsilon$ is a subset of $\mathcal{D}$, it holds that

$$\mathbb{D}_{\alpha,\epsilon}(C, u^*) \leq \mathbb{D}_\alpha(C, u^*).$$

Similar to Theorem 15, we can prove the following relation:

$$\begin{aligned}\overline{\mathcal{SD}} = \{(C, u^*) \mid C \in \mathbb{S}_{+,1}^{n^2-1}, u^* \in \mathbb{S}_1^{n-1}, \mathbb{G}_1(C, u^*) \text{ is disconnected}\}\\ \cup \{(C, u^*) \mid C \in \mathbb{S}_{+,1}^{n^2-1}, u^* \in \mathbb{S}_1^{n-1}, \mathcal{I}_{00}(C, u^*) \text{ is not empty}\}.\end{aligned}$$

Hence, the closure of $\mathcal{SD}$ is a proper subset of $\overline{\mathcal{D}}$. Combining with the fact that $\mathcal{SD}_\epsilon$ is a subset of $\mathcal{SD}$, the metric $\mathbb{D}_{\alpha,\epsilon}(C, u^*)$ is not equivalent to $\mathbb{D}_\alpha(C, u^*)$. Using the compactness of $\mathcal{SD}_\epsilon$, the following theorem provides a necessary condition for the existence of spurious local minima.

Theorem 28. *Suppose that $\epsilon > 0$ is a constant. Then, there exists a large constant $\Delta(\epsilon) > 0$ such that for every instance $\mathcal{MC}(C, u^*)$ satisfying*

$$\mathbb{D}_{\alpha,\epsilon}(C, u^*) \geq \Delta(\epsilon),$$

the instance $\mathcal{MC}(C, u^*)$ has spurious local minima.

Proof. For every pair $(C, u^*) \in \mathcal{SD}_\epsilon$, Lemma 22 implies that there exists an open neighborhood of (C, u^*) such that the desired properties hold. Now, we consider the union of such open neighborhoods over all points $(C, u^*) \in \mathcal{SD}_\epsilon$, which is an open cover of $\mathcal{SD}_\epsilon$. Using the Heine-Borel covering theorem, there exists an open sub-cover of $\mathcal{SD}_\epsilon$. Therefore, we obtain the existence of $\Delta(\epsilon)$. □

We note that the maximum possible value of $\mathbb{D}_{\alpha,\epsilon}(C, u^*)$ is $+\infty$, which is attained by instances in $\mathcal{SD}_\epsilon$. Therefore, there exist instances satisfying the condition of Theorem 28 and the lower bound is *non-trivial*. Using Theorem 28, the slightly modified complexity metric is able to provide a necessary condition on the absence of SSCPs. This result implies that our complexity metric is able to provide conditions that are much better than the RIP condition and the incoherence condition that fail to provide necessary conditions.

Finally, we conjecture that the second property in the question we asked in the beginning of the section holds for any fixed weight matrix. More specifically, we define

$$\mathbb{D}_C(u^*) := \left(\min_{(C,\tilde{u}^*) \in \overline{\mathcal{D}}} \|u^* - \tilde{u}^*\|_1 \right)^{-1}. \tag{3.15}$$

We have the following conjecture:

Conjecture 1. *Suppose that $\epsilon \in [0, 1]$. Then, there exists a large constant $\Gamma(\epsilon) > 0$ such that for every instance $\mathcal{MC}(C, u^*)$ satisfying*

$$C_{ij} \in \{0\} \cup [\epsilon, 1], \quad \mathbb{D}_C(u^*) \geq \Gamma(\epsilon),$$

the instance $\mathcal{MC}(C, u^*)$ has spurious local minima.

We note that the metric $\mathbb{D}_C(u^*)$ is equal to 0 if $\mathcal{MC}(C, u^*)$ satisfies the δ-RIP$_{2,2}$ condition with $\delta \in [0, 1)$.

Appendix

3.A Analysis of the Degenerate Case

In this section, we provide a detailed analysis on instances with $u^* = 0$. The optimization problem of the instance $\mathcal{MC}(C, 0)$ can be written as

$$\min_{u \in \mathbb{R}^n} \sum\nolimits_{i,j \in [n]} C_{ij} u_i^2 u_j^2. \tag{3.16}$$

We prove that problem (3.16) either has multiple global solutions or has no SSCPs.

Theorem 29. *If $C_{ii} > 0$ for all $i \in [n]$, the instance $\mathcal{MC}(C, 0)$ has no SSCPs. Otherwise if $C_{ii} = 0$ for some $i \in [n]$, the instance $\mathcal{MC}(C, 0)$ has nonzero global solutions.*

Proof. We first consider the case when $C_{ii} > 0$ for all $i \in [n]$. Let $u^0 \in \mathbb{R}^n$ be a second-order critical point. By the first-order optimality conditions, it holds that

$$\frac{1}{4}\nabla_i g(u^0; C, 0) = C_{ii}(u_i^0)^3 + \sum_{j \in [n], j \neq i} C_{ij} u_i^0 (u_j^0)^2 = 0, \quad \forall i \in [n].$$

Multiplying u_i^0 on both sides, we have

$$0 = C_{ii}(u_i^0)^4 + \sum_{j \in [n], j \neq i} C_{ij}(u_i^0)^2 (u_j^0)^2 \geq C_{ii}(u_i^0)^4 \geq 0,$$

which implies that $C_{ii}(u_i^0)^4 = 0$. Since $C_{ii} > 0$, it follows that

$$u_i^0 = 0, \quad \forall i \in [n].$$

Hence, $u^0 = 0$ is the unique second-order critical point.

Next, we consider the case when there exists an index i_0 such that $C_{i_0 i_0} = 0$. In this case, define $u^0 \in \mathbb{R}^n$ by

$$u_{i_0}^0 = 1, \quad u_i^0 = 0, \quad \forall i \in [n] \backslash \{i_0\}.$$

Then, we have

$$\left[u^0 (u^0)^T\right]_{i_0 i_0} = 1, \quad \left[u^0 (u^0)^T\right]_{ij} = 0, \quad \text{otherwise.}$$

Since the (i_0, i_0) entry is not observed, the point u^0 leads to the same measurements as $u^* = 0$. Therefore, u^0 is a nonzero global solution to the instance $\mathcal{MC}(C, 0)$.

□

3.B Proofs in Section 3.2

Proof of Theorem 15

Proof. We denote the set on the right-hand side as $\mathcal{D}'$. We first prove that

$$\overline{\mathcal{D}} \supset \mathcal{D}'. \tag{3.17}$$

Suppose that $(C, u^*) \in \mathcal{D}'$. If $\mathbb{G}_1(C, u^*)$ is disconnected or bipartite, the instance $\mathcal{MC}(C, u^*)$ already belongs to $\mathcal{D}$ and, therefore, belongs to the closure $\overline{\mathcal{D}}$. We only need to consider the case when $\mathcal{I}_{00}(C, u^*)$ is not empty. For every constant $\epsilon > 0$, we construct a new global solution $\tilde{u}^*$ as follows:

$$\tilde{u}_i^* := \begin{cases} u_i^* + \epsilon & \text{if } i \in \mathcal{I}_{00}(C, u^*) \\ u_i^* & \text{otherwise.} \end{cases}$$

Let $\tilde{M}^* := \tilde{u}^*(\tilde{u}^*)^T$. For the instance $\mathcal{MC}(C, \tilde{u}^*)$, we have

$$\mathcal{I}_1(C, \tilde{u}^*) = \mathcal{I}_1(C, u^*) \cup \mathcal{I}_{00}(C, u^*).$$

By the definition of $\mathcal{I}_{00}(C, u^*)$, the nodes in $\mathcal{I}_1(C, u^*)$ and $\mathcal{I}_{00}(C, u^*)$ are disconnected. Therefore, the new subgraph $\mathbb{G}_1(C, \tilde{u}^*)$ is disconnected and the new instance $\mathcal{MC}(C, \tilde{u}^*)$ belongs to $\mathcal{D}$. By letting $\epsilon \to 0$, it follows that (C, u^*) is a limit point of $\mathcal{D}$ and belongs to $\overline{\mathcal{D}}$. This completes the proof of the relation (3.17).

Then, we prove the other direction $\overline{\mathcal{D}} \subset \mathcal{D}'$. By Theorem 14, we have $\mathcal{D} \subset \mathcal{D}'$. Hence, it remains to prove that the set $\mathcal{D}'$ is closed. Equivalently, we prove that $(\mathcal{D}')^c$ is open, where $(\mathcal{D}')^c$ is the complementary set with respect to $\mathbb{R}^{n\times n} \times \mathbb{R}^n$. Suppose that $(C, u^*) \in (\mathcal{D}')^c$. If $\|C\|_1 \neq 1$ or $\|u^*\|_1 \neq 1$, changing C and u^* by a small perturbation will not make $\|C\|_1 = \|u^*\|_1 = 1$. Now, we only consider the case when $\|C\|_1 = \|u^*\|_1 = 1$. Since $(C, u^*) \in (\mathcal{D}')^c$, the subgraph $\mathbb{G}_1(C, u^*)$ is connected and not bipartite and the set $\mathcal{I}_{00}(C, u^*) = \emptyset$. Denote

$$\epsilon := \min\left\{ \min_{C_{ij}>0} C_{ij}, \min_{u_i^* \neq 0} |u_i^*| \right\} > 0.$$

Suppose that we add a sufficiently small perturbation to the point (C, u^*) such that each component of C and u^* is changed by at most $\epsilon/2$. Then, all nonzero components of C and u^* are still nonzero after the perturbation. Therefore, the edges of the subgraph $\mathbb{G}_1(C, M^*)$ are not deleted after the perturbation and, thus, the subgraph is still connected and not bipartite. Similarly, after perturbation, each node in $\mathcal{I}_0(C, M^*)$ either becomes nonzero or is connected to $\mathbb{G}_1(C, M^*)$, which implies that $\mathcal{I}_{00}(C, M^*)$ is still an empty set. Therefore, the perturbed instance still belongs to $(\mathcal{D}')^c$. Hence, the set $(\mathcal{D}')^c$ is open and we obtain the relation $\overline{\mathcal{D}} \subset \mathcal{D}'$. □

Proof of Theorem 16

The proof of Theorem 16 relies on the following two lemmas, which transform the computation of $\mathbb{D}_\alpha^{min}$ into a one-dimensional optimization problem. The first lemma upper-bounds the maximum possible distance.

Lemma 11. *Suppose that $n \geq 2$. It holds that*

$$\left(\mathbb{D}_\alpha^{min}\right)^{-1} \leq \max_{c\in\left[0,\frac{1}{n(n-1)}\right]} g(\alpha, c),$$

where the function $g(\alpha, c)$ is defined by

$$\begin{aligned} g(\alpha,c) := \min\Bigg\{ & 2(1-\alpha)\cdot\frac{n-2}{n} + 4\alpha c, \quad 4\alpha(n-1)c, \\ & 2(1-\alpha)\cdot\frac{n-4}{n} + 2\alpha\left(\frac{4}{n} - 4(n-2)c\right), \\ & 2(1-\alpha)\cdot\frac{n-3}{n} + 2\alpha\left(\frac{3}{n} - (3n-5)c\right), \\ & 2(1-\alpha)\cdot\frac{n-2}{n} + 2\alpha\left(\frac{2}{n} - 2(n-1)c\right), \\ & 2(1-\alpha)\cdot\frac{n-1}{n} + 2\alpha\left(\frac{1}{n} - (n-1)c\right)\Bigg\}. \end{aligned}$$

Proof. Denote the distance between (C, u^*) and $\mathcal{D}$ as

$$\mathbb{T}_\alpha(C, u^*) := \min_{(\tilde{C},\tilde{u}^*)\in\overline{\mathcal{D}}} \alpha\|C - \tilde{C}\|_1 + (1-\alpha)\|u^* - \tilde{u}^*\|_1.$$

We fix the pair (C, u^*) and let

$$\eta := \frac{1}{n(n-1)} \sum_{i,j\in[n], i\neq j} C_{ij} \in \left[0, \frac{1}{n(n-1)}\right].$$

Using the condition $\|C\|_1 = 1$, it follows that

$$\theta := \frac{1}{n}\sum_{i\in[n]} C_{ii} = \frac{1}{n}\left(1 - \sum_{i,j\in[n],i\neq j} C_{ij}\right) = \frac{1}{n} - (n-1)\eta \in [0, n^{-1}].$$

Our goal is to prove that

$$\mathbb{T}_\alpha(C, u^*) \leq g(\alpha, \eta).$$

In the remainder of the proof, we upper-bound the distance $\mathbb{T}_\alpha(C, u^*)$ by constructing some instances in $\overline{\mathcal{D}}$.

We first consider those instances in $\overline{\mathcal{D}}$ with a disconnected subgraph $\mathbb{G}_1$. For every $k \in \{2, \ldots, n\}$, let $\mathcal{I}_1$ be a subset of $[n]$ satisfying $|\mathcal{I}_1| = k$ and $\mathcal{I}_0 := [n] \backslash \mathcal{I}_1$. Suppose that $\epsilon > 0$ is a sufficiently small constant. For every $i_0 \in \mathcal{I}_1$, we consider the pair $(\tilde{C}, \tilde{u}^*)$, where

$$\tilde{u}_i^* = 0, \quad \forall i \in \mathcal{I}_0; \quad \tilde{u}_i^* = (1-\epsilon) u_i^* + \epsilon \cdot \frac{\left\|u_{\mathcal{I}_1}^*\right\|_1}{|\mathcal{I}_1|} + \frac{\left\|u_{\mathcal{I}_0}^*\right\|_1}{|\mathcal{I}_1|}, \quad \forall i \in \mathcal{I}_1 \tag{3.18}$$

and

$$\tilde{C}_{i_0 j} = \tilde{C}_{j i_0} = 0, \quad \forall j \in \mathcal{I}_1 \backslash \{i_0\}; \quad \tilde{C}_{ij} = C_{ij} + \frac{2}{n^2 - 2(k-1)} \sum\nolimits_{j \in \mathcal{I}_1 \backslash \{i_0\}} C_{i_0 j}, \quad \text{otherwise.}$$

By choosing a sufficiently small ϵ, it can be shown that

$$\mathcal{I}_1(\tilde{C}, \tilde{u}^*) = \mathcal{I}_1; \quad \mathcal{I}_0(\tilde{C}, \tilde{u}^*) = \mathcal{I}_0.$$

The node i_0 is disconnected from other nodes in $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ and, therefore, $(\tilde{C}, \tilde{u}^*) \in \overline{\mathcal{D}}$. The distance between u^* and $\tilde{u}^*$ is

$$\|u^* - \tilde{u}^*\|_1 = 2 \left\|u_{\mathcal{I}_0}^*\right\|_1 + 2\epsilon \left\|u_{\mathcal{I}_1}^*\right\|_1 \leq 2 \left\|u_{\mathcal{I}_0}^*\right\|_1 + 2\epsilon. \tag{3.19}$$

In addition, the distance between C and $\tilde{C}$ can be calculated as

$$\|C - \tilde{C}\|_1 = 4 \sum\nolimits_{j \in \mathcal{I}_1 \backslash \{i_0\}} C_{i_0 j}. \tag{3.20}$$

Combining inequalities (3.19) and (3.20), we have

$$\mathbb{T}_\alpha(C, u^*) \leq 2(1-\alpha) \left\|u_{\mathcal{I}_0}^*\right\|_1 + 4\alpha \sum\nolimits_{j \in \mathcal{I}_1 \backslash \{i_0\}} C_{i_0 j} + 2\epsilon. \tag{3.21}$$

Taking the average of inequality (3.21) over i_0 for $\mathcal{I}_1$, we have

$$\mathbb{T}_\alpha(C, u^*) \leq 2(1-\alpha) \left\|u_{\mathcal{I}_0}^*\right\|_1 + 4\alpha(k-1) \sum\nolimits_{i,j \in \mathcal{I}_1, i \neq j} C_{ij} + 2\epsilon. \tag{3.22}$$

Then, we take the average of (3.22) over $\mathcal{I}_1$ for all k-element subsets of $[n]$, which leads to

$$\mathbb{T}_\alpha(C, u^*) \leq 2(1-\alpha) \cdot \frac{n-k}{n} + 4\alpha(k-1)\eta + 2\epsilon.$$

By setting $\epsilon \to 0$, we obtain that

$$\mathbb{T}_\alpha(C, u^*) \leq 2(1-\alpha) \cdot \frac{n-k}{n} + 4\alpha(k-1)\eta. \tag{3.23}$$

Since inequality (3.23) is linear in k, the minimum of the right-hand side over $k \in \{2, \ldots, n\}$ is attained by either 2 or n. Hence, it holds that

$$\mathbb{T}_\alpha(C, u^*) \leq \min\left\{2(1-\alpha)\cdot\frac{n-2}{n} + 4\alpha\eta, 4\alpha(n-1)\eta\right\}. \tag{3.24}$$

Using a similar analysis, we can obtain inequality (3.23) by considering instances in $\overline{\mathcal{D}}$ whose $\mathcal{I}_{00}$ is non-empty.

Finally, we check those instances in $\overline{\mathcal{D}}$ whose $\mathbb{G}_1$ is bipartite. Let $\mathcal{I}_1$ be a subset of $[n]$ satisfying $|\mathcal{I}_1| = 4$, and let $\mathcal{I}_0 = [n]\backslash\mathcal{I}_1$. We define $\tilde{u}^*$ in the same way as (3.18). For every subset $\mathcal{I}_{11} \subset \mathcal{I}_1$ such that $|\mathcal{I}_{11}| = 2$, the new weight matrix is defined as

$$\tilde{C}_{ii} = 0, \quad \forall i \in \mathcal{I}_1; \quad \tilde{C}_{ij} = 0, \quad \forall i,j \in \mathcal{I}_{11}; \quad \tilde{C}_{ij} = 0, \quad \forall i,j \in \mathcal{I}_1\backslash\mathcal{I}_{11};$$
$$\tilde{C}_{ij} = C_{ij} + \frac{2}{n^2-8}\left(\sum_{i\in\mathcal{I}_1} C_{ii} + \sum_{i,j\in\mathcal{I}_{11}, i\neq j} C_{ij} + \sum_{i,j\in\mathcal{I}_1\backslash\mathcal{I}_{11}, i\neq j} C_{ij}\right).$$

The distance between C and $\tilde{C}$ is

$$\|C - \tilde{C}\|_1 = 2\left(\sum_{i\in\mathcal{I}_1} C_{ii} + \sum_{i,j\in\mathcal{I}_{11}, i\neq j} C_{ij} + \sum_{i,j\in\mathcal{I}_1\backslash\mathcal{I}_{11}, i\neq j} C_{ij}\right)$$

Therefore, the maximum distance is bounded by

$$\begin{aligned}\mathbb{T}_\alpha(C, u^*) \leq\ & 2(1-\alpha)\left\|u^*_{\mathcal{I}_0}\right\|_1 \\ & + 2\alpha\left(\sum_{i\in\mathcal{I}_1} C_{ii} + \sum_{i,j\in\mathcal{I}_{11}, i\neq j} C_{ij} + \sum_{i,j\in\mathcal{I}_1\backslash\mathcal{I}_{11}, i\neq j} C_{ij}\right) + 2\epsilon.\end{aligned} \tag{3.25}$$

By taking the average of (3.25) over $\mathcal{I}_{11}$ for all 2-element subsets of $\mathcal{I}_1$, it follows that

$$\mathbb{T}_\alpha(C, u^*) \leq 2(1-\alpha)\left\|u^*_{\mathcal{I}_0}\right\|_1 + 2\alpha\left(\sum_{i\in\mathcal{I}_1} C_{ii} + \frac{1}{3}\sum_{i,j\in\mathcal{I}_1, i\neq j} C_{ij}\right) + 2\epsilon. \tag{3.26}$$

Furthermore, we take the average of (3.26) over $\mathcal{I}_1$ for all 4-element subsets of $[n]$, which gives

$$\mathbb{T}_\alpha(C, u^*) \leq 2(1-\alpha)\cdot\frac{k}{n} + 2\alpha\left(4\theta + 4\eta\right) + 2\epsilon.$$

By letting $\epsilon \to 0$, we conclude that

$$\mathbb{T}_\alpha(C, u^*) \leq 2(1-\alpha)\cdot\frac{4}{n} + 2\alpha\left(4\theta + 4\eta\right). \tag{3.27}$$

By applying a similar technique to subsets of $[n]$ with $1, 2, 3$ elements, the distance can be bounded as

$$\begin{aligned} \mathbb{T}_\alpha(C, u^*) &\leq 2(1-\alpha) \cdot \frac{3}{n} + 2\alpha \left(3\theta + 2\eta\right), \\ \mathbb{T}_\alpha(C, u^*) &\leq 2(1-\alpha) \cdot \frac{2}{n} + 2\alpha \cdot 2\theta, \\ \mathbb{T}_\alpha(C, u^*) &\leq 2(1-\alpha) \cdot \frac{1}{n} + 2\alpha \cdot \theta. \end{aligned} \tag{3.28}$$

By combining inequalities (3.23), (3.27) and (3.28) and recalling the relation that $\theta = 1/n - (n-1)\eta$, it follows that

$$\mathbb{T}_\alpha(C, u^*) \leq g(\alpha, \eta).$$

Now, we take the maximum over $C \in \mathbb{S}^{n^2-1}_{+,1}$ and $u^* \in \mathbb{S}^{n-1}_1$, which is equivalent to taking the maximum over $\eta \in \left[0, \frac{1}{n(n-1)}\right]$ in the right-hand side. This yields that

$$\max_{\|C\|_1 = \|u^*\|_1 = 1} \mathbb{T}_\alpha(C, u^*) \leq \max_{c \in \left[0, \frac{1}{n(n-1)}\right]} g(\alpha, c).$$

This completes the proof. □

We denote $g_i(\alpha, c)$ be the i-th term in the above minimization for all $i \in \{1, \ldots, 6\}$. The next lemma proves the other direction.

Lemma 12. *Suppose that $n \geq 2$. It holds that*

$$\left(\mathbb{D}^{min}_\alpha\right)^{-1} \geq \max_{c \in \left[0, \frac{1}{n(n-1)}\right]} g(\alpha, c),$$

where the function $g(\alpha, c)$ is defined in Lemma 11.

Proof. Let $\eta \in \left[0, \frac{1}{n(n-1)}\right]$ and define the pair (C, u^*) according to

$$u^*_i := \frac{1}{n}, \; C_{ii} := \frac{1}{n} - (n-1)\eta, \quad \forall i \in [n]; \quad C_{ij} := \eta, \quad \forall i, j \in [n] \quad \text{s. t. } i \neq j.$$

Our goal is to prove that

$$\mathbb{T}_\alpha(C, u^*) \geq g(\alpha, \eta).$$

Suppose that $(\tilde{C}, \tilde{u}^*) \in \overline{\mathcal{D}}$ attains the distance $\mathbb{T}_\alpha(C, u^*)$, namely,

$$\mathbb{T}_\alpha(C, u^*) = \alpha \|C - \tilde{C}\|_1 + (1-\alpha)\|u^* - \tilde{u}^*\|_1.$$

We analyze three different cases.

Case I. We first consider the case when $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ is disconnected. Denote $k := |\mathcal{I}_1(\tilde{C}, \tilde{u}^*)|$. The distance between u^* and $\tilde{u}^*$ is lower-bounded by

$$\|u^* - \tilde{u}^*\|_1 \geq 2\|u^*_{\mathcal{I}_0(\tilde{C},\tilde{u}^*)} - \tilde{u}^*_{\mathcal{I}_0(\tilde{C},\tilde{u}^*)}\|_1 = 2\|u^*_{\mathcal{I}_0(\tilde{C},\tilde{u}^*)}\|_1 = \frac{2(n-k)}{n}. \tag{3.29}$$

Since there are k nodes in $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$, we need to eliminate at least $k-1$ edges that are not self-loops to make the graph disconnected. Therefore, at least $2(k-1)$ non-diagonal weights of $\tilde{C}$ are 0 and the distance between C and $\tilde{C}$ is at least

$$\|C - \tilde{C}\|_1 \geq 2 \cdot 2(k-1)\eta = 4(k-1)\eta. \tag{3.30}$$

Combining inequalities (3.29) and (3.30), we obtain that

$$\mathbb{T}_\alpha(C, u^*) \geq 2(1-\alpha) \cdot \frac{n-k}{n} + 4\alpha(k-1)\eta. \tag{3.31}$$

Case II. For the case when $\mathcal{I}_{00}(\tilde{C}, \tilde{u}^*)$ is not empty, similar estimations as *Case I* can be derived and inequality (3.31) also holds true.

Case III. Finally, we consider the case when $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ is bipartite. Denote the size of set $\mathcal{I}_1(\tilde{C}, \tilde{u}^*)$ as $k := |\mathcal{I}_1(\tilde{C}, \tilde{u}^*)|$. If $k \geq 5$, we need to eliminate at least $k-1$ edges that are not self-loops to make the graph bipartite. Thus, we can follow the same proof as *Case I* to arrive at inequality (3.31). If $k = 4$, we need to eliminate at least 2 edges that are not self-loops and 4 self-loops to make the graph bipartite. Therefore, at least 4 non-diagonal weights and 4 diagonal weights of $\tilde{C}$ are 0, and the distance between C and $\tilde{C}$ is at least

$$\|C - \tilde{C}\|_1 \geq 2\left[4\eta + 4\left(\frac{1}{n} - (n-1)\eta\right)\right] = 2\left[\frac{4}{n} - (4n-8)\eta\right]. \tag{3.32}$$

Combining inequalities (3.29) and (3.32) yields that

$$\mathbb{T}_\alpha(C, u^*) \geq 2(1-\alpha) \cdot \frac{n-4}{n} + 2\alpha\left[\frac{4}{n} - (4n-8)\eta\right]. \tag{3.33}$$

The cases when $k = 1, 2, 3$ can be analyzed similarly, leading to

$$\mathbb{T}_\alpha(C, u^*) \geq 2(1-\alpha) \cdot \frac{n-3}{n} + 2\alpha\left[\frac{3}{n} - (3n-5)\eta\right], \tag{3.34}$$

$$\mathbb{T}_\alpha(C, u^*) \geq 2(1-\alpha) \cdot \frac{n-2}{n} + 2\alpha\left[\frac{2}{n} - (2n-2)\eta\right],$$

$$\mathbb{T}_\alpha(C, u^*) \geq 2(1-\alpha) \cdot \frac{n-1}{n} + 2\alpha\left[\frac{1}{n} - (n-1)\eta\right].$$

By combining *Cases I-III*, it follows that

$$\mathbb{T}_\alpha(C, u^*) \geq g(\alpha, \eta).$$

Choosing η to be the maximizer

$$\eta^* := \underset{c \in \left[0, \frac{1}{n(n-1)}\right]}{\arg\max}\ g(\alpha, c),$$

we have

$$\mathbb{T}_\alpha(C, u^*) \geq \max_{c \in \left[0, \frac{1}{n(n-1)}\right]} g(\alpha, c).$$

Taking the maximum over $C \in \mathbb{S}_{+,1}^{n^2-1}$ and $u^* \in \mathbb{S}_1^{n-1}$ gives rise to the desired conclusion. □

Proof of Theorem 16. By the results of Lemmas 11 and 12, we only need to compute

$$\max_{c \in \left[0, \frac{1}{n(n-1)}\right]} g(\alpha, c).$$

Let $\kappa := (1-\alpha)/\alpha \in [0, +\infty]$. We study three cases below.

Case I. We first consider the case when $\kappa \geq 2(n-3)/[(n-4)(n-1)]$. We prove that $g(\alpha, c) = g_2(\alpha, c)$. Since $g_2(\alpha, c)$ has a larger gradient than $g_1(\alpha, c)$ and the function $g_i(\alpha, c)$ is decreasing in c for $i = 3, 4, 5, 6$, we only need to show that

$$g_i\left(\alpha, \frac{1}{n(n-1)}\right) \geq g_2\left(\alpha, \frac{1}{n(n-1)}\right), \quad \forall i \in \{1, 3, 4, 5, 6\}. \tag{3.35}$$

The above inequality with $i = 1$ is equivalent to $\kappa \geq 2/(n-1)$, which is guaranteed by the assumption that $\kappa \geq 2(n-3)/[(n-4)(n-1)]$. For $i \in \{3, 4, 5, 6\}$, the inequality (3.35) is equivalent to

$$\kappa \geq \max\left\{\frac{2(n-3)}{(n-1)(n-4)}, \frac{2(n-2)}{(n-1)(n-3)}, \frac{2}{n-2}, \frac{2}{n-1}\right\} = \frac{2(n-3)}{(n-1)(n-4)}.$$

Therefore, it holds that

$$g(\alpha, c) = g_2(\alpha, c) = 4\alpha(n-1)c.$$

whose maximum is attained at $c = [n(n-1)]^{-1}$ and

$$\max_{C, u^*} \mathbb{T}_\alpha(C, u^*) = g_2\left(\alpha, \frac{1}{n(n-1)}\right) = \frac{4\alpha}{n}.$$

Case II. Then, we consider the case when $\kappa \leq 2/n$. In this case, we prove that the maximum is achieved by the intersection point between $g_1(\alpha, c)$ (an increasing function in c) and $\min\{g_5(\alpha, c), g_6(\alpha, c)\}$ (a decreasing function in c). The intersection points between $g_1(\alpha, C)$ and the other five functions are

$$\frac{\kappa}{2n}, \quad \frac{2-\kappa}{n(2n-3)}, \quad \frac{3-\kappa}{3n(n-1)}, \quad \frac{1}{n^2}, \quad \frac{1+\kappa}{n(n+1)}.$$

In the regime $\kappa \leq 1/n$, we have

$$\frac{\kappa}{2n} \leq \frac{1+\kappa}{n(n+1)} \leq \frac{1}{n^2} \leq \min\left\{\frac{2-\kappa}{n(2n-3)}, \frac{3-\kappa}{3n(n-1)}\right\},$$

which implies that the maximum is attained at $c = (1+\kappa)/[n(n+1)]$. Hence, the maximum distance is

$$\max_{C,u^*} \mathbb{T}_\alpha(C, u^*) = g_1\left(\alpha, \frac{1+\kappa}{n(n+1)}\right) = \frac{2(1-\alpha)(n-2)(n+1)+4}{n(n+1)}.$$

In the regime $1/n \leq \kappa \leq 2/n$, we have

$$\frac{\kappa}{2n} \leq \frac{1}{n^2} \leq \frac{1+\kappa}{n(n+1)} \leq \min\left\{\frac{2-\kappa}{n(2n-3)}, \frac{3-\kappa}{3n(n-1)}\right\},$$

which implies that the maximum is attained at $c = 1/n^2$. Hence, the maximum distance is

$$\max_{C,u^*} \mathbb{T}_\alpha(C, u^*) = g_1\left(\alpha, \frac{1}{n^2}\right) = \frac{2(1-\alpha)(n-2)n+4\alpha}{n^2}.$$

Case III. We finally consider the case when $2/n \leq \kappa \leq 2(n-3)/[(n-4)(n-1)]$. In this regime, the intersection point between $g_2(\alpha, c)$ and $g_5(\alpha, c)$ is

$$\frac{\kappa(n-2)+2}{4n(n-1)} \leq \frac{\kappa}{2n}.$$

This implies that $g_2(\alpha, c)$ intersects with $g_5(\alpha, c)$ before $g_1(\alpha, c)$. Therefore, the maximum is attained at one of the intersects between $g_2(\alpha, c)$ and $g_i(\alpha, c)$ for $i = 3, 4, 5, 6$. By calculating the four intersects, the optimal c that achieves the maximum is given by

$$c^*(\kappa) := \min\left\{\frac{\kappa(n-4)+4}{n(6n-10)}, \frac{\kappa(n-3)+3}{n(5n-7)}, \frac{\kappa(n-2)+2}{4n(n-1)}, \frac{\kappa(n-1)+1}{3n(n-1)}\right\},$$

which is an increasing function in κ. If $\kappa = 2/n$, we can estimate that

$$c^*(\kappa) = \min\left\{\frac{2(n-4)/n+4}{n(6n-10)}, \frac{2(n-3)/n+3}{n(5n-7)}, \frac{2(n-2)/n+2}{4n(n-1)}, \frac{2(n-1)/n+1}{3n(n-1)}\right\} \tag{3.36}$$

$$= \min\left\{\frac{3n-4}{n^2(3n-5)}, \frac{5n-6}{n^2(5n-7)}, \frac{1}{n^2}, \frac{3n-2}{n^2(3n-3)}\right\} = \frac{1}{n^2}.$$

Similarly, if $\kappa = 2(n-3)/[(n-4)(n-1)]$, it holds that

$$c^*(\kappa) = \frac{1}{n(n-1)}. \tag{3.37}$$

Combining (3.36) and (3.37), we have

$$c^*(\kappa) \in \left[\frac{1}{n^2}, \frac{1}{n(n-1)}\right], \quad \forall \kappa \in \left[\frac{2}{n}, \frac{2(n-3)}{(n-4)(n-1)}\right].$$

Therefore, the maximum distance satisfies the bound

$$\max_{C,u^*} \ \mathbb{T}_\alpha(C,u^*) = g_2\left[\alpha, c^*(\kappa)\right] \in \left[\frac{4\alpha(n-1)}{n^2}, \frac{4\alpha}{n}\right].$$

This completes the proof. □

Proof of Theorem 17

Proof. By the assumption that the complexity metric of (C,u^*) is finite, we have that $(C,u^*) \notin \overline{\mathcal{D}}$. It follows from Theorem 15 that the subset $\mathcal{I}_{00}(C,u^*)$ is empty and that $\mathbb{G}_1(C,u*)$ is connected and not bipartite. Let $k := |\mathcal{I}_1(C,u^*)|$. For each node $i_0 \in \mathcal{I}_1(C,u^*)$, we define the new weight matrix $\tilde{C}$ as

$$\begin{aligned}
&\tilde{C}_{i_0 j} = \tilde{C}_{j i_0} = 0, \quad \forall j \in \mathcal{I}_1(C,u^*)\backslash\{i_0\};\\
&\tilde{C}_{ij} = C_{ij} + \frac{2}{n^2 - 2(k-1)} \sum\nolimits_{j \in \mathcal{I}_1(C,u^*)\backslash\{i_0\}} C_{i_0 j}, \quad \text{otherwise.}
\end{aligned}$$

The subgraph $\mathbb{G}_1(\tilde{C},u^*)$ is disconnected and, therefore, we have $(\tilde{C},u^*) \in \overline{\mathcal{D}}$. It follows that

$$\frac{4\alpha^*}{n} = [\mathbb{D}_{\alpha^*}(C,u^*)]^{-1} \le \alpha^* \|C - \tilde{C}\|_1 = 4\alpha^* \sum\nolimits_{j \in \mathcal{I}_1(C,u^*)\backslash\{i_0\}} C_{i_0 j}. \tag{3.38}$$

For each node $i_0 \in \mathcal{I}_0(C,u^*)$, a similar construct of $\tilde{C}$ leads to

$$\frac{4\alpha^*}{n} = [\mathbb{D}_{\alpha^*}(C,u^*)]^{-1} \le 4\alpha^* \sum_{j \in \mathcal{I}_1(C,u^*)} C_{i_0 j}. \tag{3.39}$$

By summing inequality (3.38) over i_0 for all nodes in $\mathcal{I}_1(C,u^*)$ and summing inequality (3.39) over i_0 for all nodes in $\mathcal{I}_0(C,u^*)$, it follows that

$$4\alpha^* \le 4\alpha^* \left[\sum_{i,j \in \mathcal{I}_1(C,u^*), i \ne j} C_{ij} + \sum_{i \in \mathcal{I}_1(C,u^*), j \in \mathcal{I}_0(C,u^*)} C_{ij}\right] \le 4\alpha^* \sum_{i,j \in [n], i \ne j} C_{ij} \le 4\alpha^*, \tag{3.40}$$

where all inequalities should hold with equality. Since the last inequality in (3.40) holds with equality, we obtain that

$$C_{ii} = 0, \quad \forall i \in [n].$$

It follows from the equality of inequalities (3.38) and (3.39) that

$$\sum_{j \in \mathcal{I}_1(C,u^*)\setminus\{i\}} C_{ij} = \frac{1}{n}, \ \forall i \in \mathcal{I}_1(C,u^*); \qquad \sum_{j \in \mathcal{I}_1(C,u^*)} C_{ij} = \frac{1}{n}, \ \forall i \in \mathcal{I}_0(C,u^*). \tag{3.41}$$

Using the condition that $\|C\|_1 = 1$, the above equalities imply that all weights of C are limited to edges with a node in $\mathcal{I}_1(C,u^*)$. Namely, we have

$$\sum\nolimits_{j \in \mathcal{I}_0(C,u^*)} C_{ij} = 0, \quad \forall i \in \mathcal{I}_1(C,u^*). \tag{3.42}$$

If $\mathcal{I}_o(C,u^*)$ is not empty, the above equality contradicts the second equality in (3.41). Hence, the point (C,u^*) satisfies that $\mathcal{I}_0(C,u^*) = \emptyset$. By a similar analysis of the bipartite instance in Lemma 11, for every 4-element subset $\{i,j,k,\ell\}$ of $[n]$, it holds that

$$2(1-\alpha^*)(1-|u_i^*|-|u_j^*|-|u_k^*|-|u_\ell^*|) + 4\alpha^*(C_{ij}+C_{k\ell}) = 4\alpha^*/n.$$

Taking the average of the above equality over $\{i,j,k,\ell\}$ for all 4-element subsets of $[n-1]$, we obtain that

$$2(1-\alpha^*)\left(1 - \frac{3\|u_{1:n-1}^*\|_1}{n-1}\right) + 4\alpha^* \frac{2}{(n-1)(n-2)} \|C_{1:n-1,1:n-1}\|_1 = \frac{4\alpha^*}{n}.$$

Using the first equality in (3.41) and the symmetry of C, it holds that $\|C_{1:n-1,1:n-1}\|_1 = 1 - 2/n$. Substituting into the above equality, we know

$$2(1-\alpha^*)\left(1 - \frac{3\|u_{1:n-1}^*\|_1}{n-1}\right) = 4\alpha^* \cdot \frac{n-3}{n(n-1)}.$$

By recalling that $\alpha^* = (n-1)(n-4)/(n^2-3n-2)$, the above inequality leads to

$$\|u_{1:n-1}^*\|_1 = (n-1)/n,$$

which is equivalent to $|u_n^*| = 1/n$. By the same proof technique, we conclude that

$$|u_i^*| = 1/n, \quad \forall i \in [n].$$

By substituting back into equality (3.42), it holds for all 4-element subsets $\{i,j,k,\ell\} \subset [n]$ that

$$C_{ij} + C_{k\ell} = \frac{2}{n(n-1)},$$

which implies that

$$C_{ij} = \frac{1}{n(n-1)}, \quad \forall i,j \in [n] \quad \text{s.t. } i \neq j.$$

□

3.C Proofs in Section 3.3

Proof of Theorem 20

Before proving the estimation of the complexity metric, we prove two properties of μ-incoherent vectors.

Lemma 13. *Given any constant $\mu \in [1, n]$, suppose that u^* has incoherence μ and $\|u^*\|_1 = 1$. Then, the following properties hold:*

1. *u^* has at least n/μ nonzero components;*
2. *$|u_i^*| \le \mu/n$ for all $i \in [n]$.*

Proof. Assume without loss of generality that

$$|u_i^*| > 0, \quad \forall i \in [\ell]; \quad u_i^* = 0, \quad \forall i \in \{\ell+1, \dots, n\}.$$

By the definition (3.3), we have

$$(u_i^*)^2 \le \frac{\mu}{n}\|u^*\|_2^2 = \frac{\mu}{n}\sum\nolimits_{i\in[\ell]}(u_i^*)^2, \quad \forall i \in [\ell].$$

Summing over $i \in [\ell]$, we obtain that

$$\sum\nolimits_{i\in[\ell]}(u_i^*)^2 \le \frac{\ell\mu}{n}\sum\nolimits_{i\in[\ell]}(u_i^*)^2,$$

which implies that $\ell \ge n/\mu$. Let

$$c_i := |u_i^*|/\|u^*\|_2, \quad \forall i \in [\ell].$$

The assumption that the incoherence is equal to μ implies that

$$c_i \in (0, \sqrt{\mu/n}], \quad \forall i \in [\ell]. \tag{3.43}$$

In addition, it holds that

$$\begin{aligned} \|u^*\|_2^2 &= \sum\nolimits_{i\in[\ell]}(u_i^*)^2 = \sum\nolimits_{i\in[\ell]} c_i^2\|u^*\|_2^2, \\ 1 = \|u^*\|_1 &= \sum\nolimits_{i\in[\ell]}|u_i^*| = \sum\nolimits_{i\in[\ell]} c_i\|u^*\|_2, \end{aligned}$$

which implies that

$$\sum\nolimits_{i\in[\ell]} c_i^2 = 1, \quad \sum\nolimits_{i\in[\ell]} c_i = \|u^*\|_2^{-1}.$$

Combined with (3.43), it follows that

$$\|u^*\|_2^{-1} = \sum\nolimits_{i\in[\ell]} c_i \ge \sqrt{\frac{n}{\mu}} \cdot \sum\nolimits_{i\in[\ell]} c_i^2 = \sqrt{\frac{n}{\mu}}.$$

Therefore,

$$|u_i^*| = c_i\|u^*\|_2 \leq \sqrt{\mu/n} \cdot \sqrt{\mu/n} = \mu/n.$$

□

The following lemma lower-bounds the perturbation of the weight matrix C.

Lemma 14. *Suppose that the instance $\mathcal{MC}(C, u^*)$ satisfies the δ-RIP$_{2,2}$ condition and the weight matrix $\tilde{C} \in \mathbb{S}_{+,1}^{n^2-1}$ has N zero entries, where $\delta \in [0, 1)$ and $N \in [n^2]$. Then, it holds that*

$$\|C - \tilde{C}\|_1 \geq 2\sum\nolimits_{(i,j)\in\mathcal{N}} C_{ij} \geq \frac{2(1-\delta)N}{(1+\delta)n^2 - 2\delta N},$$

where $\mathcal{N}$ is the set of indices of zero entries of $\tilde{C}$.

Proof. The δ-RIP$_{2,2}$ condition implies that

$$\frac{\min_{i,j} C_{ij}}{\max_{i,j} C_{ij}} \geq \frac{1-\delta}{1+\delta}.$$

Therefore, considering the average of entries in $\mathcal{N}$ and that of entries not in $\mathcal{N}$, we have

$$\frac{\frac{1}{N}\sum_{(i,j)\in\mathcal{N}} C_{ij}}{\frac{1}{n^2-N}\sum_{(i,j)\notin\mathcal{N}} C_{ij}} \geq \frac{1-\delta}{1+\delta},$$

which further leads to

$$\sum_{(i,j)\in\mathcal{N}} C_{ij} \geq \frac{1-\delta}{1+\delta} \cdot \frac{N}{n^2-N} \sum_{(i,j)\notin\mathcal{N}} C_{ij} = \frac{1-\delta}{1+\delta} \cdot \frac{N}{n^2-N} \left(1 - \sum_{(i,j)\in\mathcal{N}} C_{ij}\right).$$

The above inequality is equivalent to

$$\sum\nolimits_{(i,j)\in\mathcal{N}} C_{ij} \geq \frac{(1-\delta)N}{(1+\delta)n^2 - 2\delta N}.$$

Hence, the distance between C and $\tilde{C}$ is lower-bounded as

$$\|C - \tilde{C}\|_1 \geq 2\sum\nolimits_{(i,j)\in\mathcal{N}} C_{ij} \geq \frac{2(1-\delta)N}{(1+\delta)n^2 - 2\delta N}.$$

This completes the proof. □

Now, we prove the main theorem.

Proof of Theorem 20. Suppose that $\mathcal{MC}(\tilde{C}, \tilde{u}^*) \in \overline{\mathcal{D}}$ is the instance such that

$$[\mathbb{D}_\alpha(C, u^*)]^{-1} = \alpha\|C - \tilde{C}\|_1 + (1-\alpha)\|u^* - \tilde{u}^*\|_1.$$

In the following, we split the proof into two steps.

Step I. We first fix $\tilde{u}^*$ and consider the closest matrix $\tilde{C}$ to C such that $(\tilde{C}, \tilde{u}^*) \in \overline{\mathcal{D}}$. Let $k := |\mathcal{I}_1(\tilde{C}, \tilde{u}^*)|$. Without loss of generality, we assume that

$$\mathcal{I}_1(\tilde{C}, \tilde{u}^*) = \{1, \dots, k\}, \quad \mathcal{I}_0(\tilde{C}, \tilde{u}^*) = \{k+1, \dots, n\}.$$

We first consider the case when $k \geq 2$. If $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ is disconnected, at least $2(k-1)$ entries of $\tilde{C}$ are 0. If $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ are bipartite, at least $k^2/2 \geq 2(k-1)$ entries of $\tilde{C}$ are 0. If $\mathcal{I}_{00}(\tilde{C}, \tilde{u}^*)$ is non-empty, at least $2k$ entries of $\tilde{C}$ are 0. Otherwise if $k = 1$, at least one entry of $\tilde{C}$ should be 0 to make $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ bipartite. In summary, at least $N(k)$ entries of $\tilde{C}$ are 0, where

$$N(k) := \max\{2(k-1), 1\}.$$

Using the results in Lemma 14, the distance between C and $\tilde{C}$ is at least

$$\|C - \tilde{C}\|_1 \geq \frac{2(1-\delta)N(k)}{(1+\delta)n^2 - 2\delta N(k)}. \tag{3.44}$$

We note that the distance is monotonously increasing as a function of k.

Step II. Now, we consider the optimal choice of $\tilde{u}^*$ based on the lower bound in (3.44). Let

$$\ell := |\mathcal{I}_1(C, u^*)|, \quad k := |\mathcal{I}_1(\tilde{C}, \tilde{u}^*)|.$$

Since the distance between C and $\tilde{C}$ is a monotonously increasing function of k, the minimum distance between (C, u^*) and $(\tilde{C}, \tilde{u}^*)$ cannot be attained by $k > \ell$. Therefore, we focus on the case when $k \leq \ell$. Without loss of generality, we assume that

$$|u_1^*| \geq |u_2^*| \geq \cdots \geq |u_\ell^*| > 0; \quad |u_i^*| = 0, \quad \forall i \geq \ell + 1.$$

Then, the distance between u^* and $\tilde{u}^*$ satisfies

$$\|u^* - \tilde{u}^*\|_1 \geq 2\sum\nolimits_{i=k+1}^{\ell} |u_i^*|. \tag{3.45}$$

Denote the distance between (C, u^*) and $(\tilde{C}, \tilde{u}^*)$ by

$$d_\alpha := \alpha\|C - \tilde{C}\|_1 + (1-\alpha)\|u^* - \tilde{u}^*\|_1.$$

Step II-1. We first consider the case when $\mu \leq 2n/3$. Combining inequalities (3.44) and (3.45), we obtain a lower bound on d_α:

$$d_\alpha \geq \min_{k \in [\ell]} \left[\frac{2\alpha(1-\delta)N(k)}{n^2(1+\delta) - 2\delta N(k)} + 2(1-\alpha)\sum\nolimits_{i=k+1}^{\ell} |u_i^*| \right].$$

For every $k \in [\ell]$, the term inside the above minimization can be lower-bounded by

$$\frac{2\alpha(1-\delta)N(k)}{n^2(1+\delta) - 2\delta N(k)} + 2(1-\alpha)\sum\nolimits_{i=k+1}^{\ell} |u_i^*| \geq \frac{2\alpha(1-\delta)\cdot 2(k-1)}{n^2(1+\delta)} + 2(1-\alpha)\sum\nolimits_{i=k+1}^{\ell} |u_i^*|$$

$$= \frac{4\alpha(1-\delta)}{n^2(1+\delta)} \cdot (k-1) + 2(1-\alpha)\sum\nolimits_{i=k+1}^{\ell} |u_i^*|.$$

The minimum of the right-hand side over $k \in [\ell]$ can be solved in closed form and is equal to

$$\sum\nolimits_{i=2}^{\ell} \min\left\{\frac{4\alpha(1-\delta)}{n^2(1+\delta)}, 2(1-\alpha)|u_i^*|\right\}.$$

Using the second property in Lemma 13, we have

$$\begin{aligned}\min\left\{\frac{4\alpha(1-\delta)}{n^2(1+\delta)}, 2(1-\alpha)|u_i^*|\right\} &\geq \min\left\{\frac{4\alpha(1-\delta)}{n^2(1+\delta)} \cdot \frac{n|u_i^*|}{\mu}, 2(1-\alpha)|u_i^*|\right\} \\ &= \min\left\{\frac{4\alpha(1-\delta)}{\mu n(1+\delta)}, 2(1-\alpha)\right\} \cdot |u_i^*|.\end{aligned}$$

Taking the summation over $k \in \{2, \ldots, \ell\}$, we can conclude that

$$d_\alpha \geq \sum\nolimits_{k=2}^{\ell} \min\left\{\frac{4\alpha(1-\delta)}{\mu n(1+\delta)}, 2(1-\alpha)\right\} \cdot |u_i^*| = \min\left\{\frac{4\alpha(1-\delta)}{\mu n(1+\delta)}, 2(1-\alpha)\right\} \cdot \sum\nolimits_{k=2}^{\ell} |u_i^*|. \tag{3.46}$$

Using the second property in Lemma 13 and $\|u^*\|_1 = 1$, it follows that

$$\sum\nolimits_{k=2}^{\ell} |u_i^*| \geq 1 - \frac{\mu}{n}.$$

Substituting back into inequality (3.46), we have

$$d_\alpha \geq \min\left\{\frac{4\alpha(1-\delta)}{\mu n(1+\delta)}, 2(1-\alpha)\right\} \cdot \left(1 - \frac{\mu}{n}\right).$$

Step II-2. Next, we consider the case when $\mu \geq 2n/3$. By Theorem 19, the distance is at least

$$d_\alpha \geq \frac{2\alpha(1-\delta)}{n^2(1+\delta) - 2\delta} \geq \frac{2\alpha(1-\delta)}{(3/2)\mu \cdot n(1+\delta)} \geq \min\left\{\frac{4\alpha(1-\delta)}{\mu n(1+\delta)}, 2(1-\alpha)\right\} \cdot \frac{1}{3},$$

where the second inequality is due to the assumption that $\mu \geq 2n/3$.

By combining *Steps II-1* and *II-2*, the distance is lower-bounded by

$$\begin{aligned}d_\alpha &\geq \min\left\{\frac{4\alpha(1-\delta)}{\mu n(1+\delta)}, 2(1-\alpha)\right\} \times \max\left\{1 - \frac{\mu}{n}, \frac{1}{3}\right\} \\ &= \min\left\{\frac{4\alpha(1-\delta)}{n(1+\delta)}, 2(1-\alpha)\mu\right\} \times \max\left\{\frac{1}{\mu} - \frac{1}{n}, \frac{1}{3\mu}\right\}\end{aligned}$$

The proof is completed by using the relation between d_α and $\mathbb{T}_\alpha(C, u^*)$. □

Proof of Theorem 21

Proof. The proof is split into two different cases.

Case I. We first consider the case when $\mu \leq n/2$. We construct the weight matrix $\tilde{C}$ as

$$\tilde{C}_{1i} = \tilde{C}_{i1} = 0, \quad \forall i \in \{2, \ldots, \ell\}; \quad \tilde{C}_{ij} = \frac{1}{n^2 - 2(\ell - 1)}, \quad \text{otherwise.}$$

For the instance $\mathcal{MC}(\tilde{C}, u^*)$, node 1 is disconnected from nodes $\{2, \ldots, \ell\}$ and thus, the subgraph $\mathbb{G}_1(\tilde{C}, u^*)$ is disconnected. This implies that $(\tilde{C}, u^*) \in \overline{\mathcal{D}}$. The matrix C is defined as

$$\begin{aligned} C_{1i} = C_{i1} &= \frac{1 - \delta}{(1 + \delta)n^2 - 4\delta(\ell - 1)}, \quad \forall i \in \{2, \ldots, \ell\}; \\ C_{ij} &= \frac{1 + \delta}{(1 + \delta)n^2 - 4\delta(\ell - 1)}, \quad \text{otherwise.} \end{aligned}$$

We can verify that the weight matrix C ensures that $\mathcal{MC}(C, u^*)$ satisfies the δ-RIP$_{2,2}$ condition. The complexity of $\mathcal{MC}(C, u^*)$ is lower-bounded by

$$\begin{aligned} \mathbb{D}_\alpha(C, u^*) &\geq \left(\alpha \|C - \tilde{C}\|_1\right)^{-1} = \frac{(1 + \delta)n^2 - 4\delta(\ell - 1)}{4\alpha(\ell - 1)(1 - \delta)} \\ &\geq \frac{(1 + \delta)(n^2 - 2n)}{4\alpha(\ell - 1)(1 - \delta)} = \frac{n(1 + \delta)}{4\alpha(1 - \delta)} \cdot \frac{n - 2}{\ell - 1} \geq \frac{n(1 + \delta)}{4\alpha(1 - \delta)} \cdot \frac{n\mu}{2(n\ell - 1)}, \end{aligned}$$

where the second last inequality follows from $4\delta \leq 2(1 + \delta)$ and the last inequality is due to $n \geq 4$.

Case II. Next, we consider the case when $\mu \geq n/2$. Theorem 19 implies that there exists an instance $\mathcal{MC}(C, u^*)$ such that

$$\mathbb{D}_\alpha(C, u^*) = \frac{n^2(1 + \delta) - 2\delta}{2\alpha(1 - \delta)} \geq \frac{(n^2 - 1)(1 + \delta)}{2\alpha(1 - \delta)} \geq \frac{n(1 + \delta)}{2\alpha(1 - \delta)} \cdot \frac{n}{2},$$

where the first inequality results from $2\delta \leq 1 + \delta$ and the second inequality is in light of $n \geq 4$. Using the condition that $\mu \leq n$, it follows that

$$\mathbb{D}_\alpha(C, u^*) \geq \frac{n(1 + \delta)}{4\alpha(1 - \delta)} \cdot \mu.$$

Combining *Cases I* and *II* completes the proof. □

Proof of Theorem 22

We first establish several lemmas before providing the proof of Theorem 22. The first lemma is the Chernoff bound for the sum of Bernoulli random variables, which is a result of Proposition 2.14 in [224].

Lemma 15. *Suppose that $X_1, \ldots, X_m$ are i.i.d. Bernoulli random variables with the parameter p. Then, it holds that*

$$\mathbb{P}\left(\sum_{i\in[m]} X_i \leq \frac{mp}{2}\right) \leq \exp\left(\frac{-mp}{8}\right), \quad \mathbb{P}\left(\sum_{i\in[m]} X_i \geq \frac{3mp}{2}\right) \leq \exp\left(\frac{-mp}{10}\right).$$

The next lemma provides an upper bound on the total number of nonzero entries.

Lemma 16. *Suppose that $n \geq 3$. With probability at least $1 - \exp(-np/10)$, there are at most $3n^2p/2$ nonzero entries in C. With the same probability, it holds that*

$$C_{ij} \geq \frac{2}{3n^2p}, \quad \forall i, j \in [n] \quad \text{s. t. } C_{ij} > 0.$$

Proof. For the $n(n-1)$ non-diagonal entries of C, Lemma 15 implies that there are at most $(3/2)\cdot n(n-1)p$ nonzero entries with probability at least $1 - \exp(-n(n-1)p/20)$. For the n diagonal entries of C, the same lemma implies that there are at most $(3/2)\cdot np$ nonzero entries with probability at least $1 - \exp(-np/10)$. Combining both parts concludes that there are at most $(3/2)\cdot n^2p$ nonzero entries in C with probability at least

$$1 - \exp(-n(n-1)p/20) - \exp(-np/10) \geq 1 - 2\exp(-np/10),$$

where the last inequality is due to $n \geq 3$. The lower bound on C_{ij} follows from the normalization constraint. □

For every fixed global solution $\tilde{u}^*$, the next lemma estimates the distance between $(C, \tilde{u}^*)$ and $\mathcal{D}$.

Lemma 17. *Suppose that $\tilde{u}^*$ is a given vector and the random matrix C obeys the Bernoulli model. In addition, suppose that $\eta > 2$ is a constant and*

$$\|\tilde{u}^*\|_0 \geq \frac{n}{2\mu}, \quad p \geq \min\left\{1, \frac{16(1+\eta\mu)\log n + 16}{n}\right\},$$

where $\|\tilde{u}^*\|_0$ is the number of nonzero entries of $\tilde{u}^*$. For every instance $(\tilde{C}, \tilde{u}^*) \in \overline{\mathcal{D}}$, it holds with probability at least $1 - 3n^{-\eta/2}$ that

$$\|C - \tilde{C}\|_1 \geq \frac{4(\|\tilde{u}^*\|_0 - 1)}{3n^2}.$$

Proof. For all $i, j \in [n]$, we define Bernoulli random variables X_{ij} to be 1 if $C_{ij} > 0$ and 0 otherwise. Then, X_{ij} are independent identically distributed Bernoulli random variables with the parameter p. Let $N := \sum_{i,j} X_{ij}$ be the number of nonzero weights in C. By the definition of the Bernoulli model, all nonzero entries of C are equal to N^{-1}. Since the global solution $\tilde{u}^*$ is fixed, we assume without loss of generality that

$$\mathcal{I}_1(C, \tilde{u}^*) = [\ell], \quad \mathcal{I}_0(C, \tilde{u}^*) = \{\ell + 1, \dots, n\}.$$

We fix $\tilde{C}$ to be a weight matrix such that $(\tilde{C}, \tilde{u}^*) \in \overline{\mathcal{D}}$ and investigate three cases.

Case I. We first consider the case when $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ is disconnected. Suppose that $\tilde{\mathcal{I}}_{11}$ and $\tilde{\mathcal{I}}_{12}$ are a division of $[\ell]$ such that the nodes in $\tilde{\mathcal{I}}_{11}$ are not connected with the nodes in $\tilde{\mathcal{I}}_{12}$. In addition, we denote $k := |\tilde{\mathcal{I}}_{11}|$ and assume that $k \leq \ell/2$. Since the nodes in $\tilde{\mathcal{I}}_{11}$ are disconnected from the nodes in $\tilde{\mathcal{I}}_{12}$, at least

$$2\sum\nolimits_{i \in \tilde{\mathcal{I}}_{11}, j \in \tilde{\mathcal{I}}_{12}} X_{ij}$$

nonzero entries in C are equal to 0 in $\tilde{C}$. Therefore, we have

$$\|C - \tilde{C}\|_1 \geq \frac{1}{N} \cdot 4 \sum_{i \in \tilde{\mathcal{I}}_{11}, j \in \tilde{\mathcal{I}}_{12}} X_{ij} = \frac{4}{N} \sum_{i \in \tilde{\mathcal{I}}_{11}, j \in \tilde{\mathcal{I}}_{12}} X_{ij}.$$

Using Lemma 15, it holds that

$$\sum_{i \in \tilde{\mathcal{I}}_{11}, j \in \tilde{\mathcal{I}}_{12}} X_{ij} \geq \frac{1}{2} \cdot |\tilde{\mathcal{I}}_{11}||\tilde{\mathcal{I}}_{12}|p = \frac{k(\ell - k)p}{2}$$

with probability at least $1 - \exp(-k(\ell - k)p/8)$. Since $k(\ell - k) \geq \ell - 1$, one can write:

$$\|C - \tilde{C}\|_1 \geq \frac{4}{N} \sum_{i \in \tilde{\mathcal{I}}_{11}, j \in \tilde{\mathcal{I}}_{12}} X_{ij} \geq \frac{4}{N} \cdot \frac{(\ell - 1)p}{2} = \frac{2(\ell - 1)p}{N} \tag{3.47}$$

with the same probability. Considering the union bound over all weight matrices $\tilde{C}$ for which $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ is disconnected, inequality (3.47) holds with probability at least

$$\begin{aligned} 1 - \sum_{k=1}^{\lfloor \ell/2 \rfloor} \binom{\ell}{k} \exp\left[-\frac{k(\ell - k)p}{8}\right] &\geq 1 - \sum_{k=1}^{\lfloor \ell/2 \rfloor} \left(\frac{\ell e}{k}\right)^k \exp\left[-\frac{k(\ell - k)p}{8}\right] \\ &= 1 - \sum_{k=1}^{\lfloor \ell/2 \rfloor} \exp\left[k + k \log\left(\frac{\ell}{k}\right) - \frac{k(\ell - k)p}{8}\right], \end{aligned}$$

where the inequality uses the relation $\binom{\ell}{k} \leq (\ell e/k)^k$. Using the relation that $k \leq \ell/2$, we can estimate that

$$\begin{aligned}
&\exp\left[k + k\log\left(\frac{\ell}{k}\right) - \frac{k(\ell-k)p}{8}\right] \leq \exp\left[k + k\log\ell - \frac{k\ell p}{16}\right] \\
=&\exp\left[-\frac{k\ell}{16}\left(p - \frac{16(1+\log\ell)}{\ell}\right)\right] \leq \exp\left[-\frac{k\ell}{16}\left(p - \frac{16(1+\log n)}{n}\right)\right] \\
\leq&\exp\left[-\frac{k\ell}{16}\cdot\frac{16\eta\mu\log n}{n}\right] = \exp\left(-\frac{\eta\mu k\ell\log n}{n}\right) = n^{-\frac{\eta\mu\ell}{n}\cdot k} \leq n^{-\frac{\eta}{2}\cdot k},
\end{aligned}$$

where the second last inequality is from the assumption on p and the last inequality is from $\ell \geq n/(2\mu)$. By taking the summation over $k = 1, \ldots, \lfloor \ell/2 \rfloor$, it follows that

$$1 - \sum_{k=1}^{\lfloor \ell/2 \rfloor} \exp\left[k + k\log\left(\frac{\ell}{k}\right) - \frac{k(\ell-k)p}{8}\right] \geq 1 - \sum_{k=1}^{\lfloor \ell/2 \rfloor} n^{-\frac{\eta}{2}\cdot k} \geq 1 - \frac{n^{-\frac{\eta}{2}}}{1 - n^{-\frac{\eta}{2}}} \geq 1 - 2n^{-\eta/2},$$

where the last inequality is due to $n^{-\eta/2} \geq n^{-1} \geq 1/2$. Therefore, inequality (3.47) holds with probability at least $1 - 2n^{-\eta/2}$. Using the lower bound of N in Lemma 16, the distance between C and $\tilde{C}$ is at least

$$\frac{2}{3n^2 p} \cdot 2(\ell - 1)p = \frac{4(\ell-1)}{3n^2}$$

with probability at least

$$1 - 2n^{-\eta/2} - \exp(-np/10) \geq 1 - 2n^{-\eta/2} - n^{-4\mu\eta/5} \geq 1 - 3n^{-\eta/2}.$$

Case II. For the case when $\mathcal{I}_{00}(\tilde{C}, \tilde{u}^*)$ is non-empty, the analysis is the same as *Case I.* and it holds that

$$\|C - \tilde{C}\|_1 \geq \frac{2}{3n^2 p} \cdot 2(\ell - 1)p = \frac{4(\ell-1)}{3n^2}$$

with probability at least $1 - 3n^{-\eta/2}$.

Case III. Finally, we consider the case when $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ is bipartite. In this case, we show that there exists a set of indices $\mathcal{I} \subset [n]^2$ with at least $\max\{\ell^2/2, 1\}$ elements such that

$$\tilde{C}_{ij} = 0, \quad \forall (i,j) \in \mathcal{I}.$$

The proof of the above claim can be found in the proof of Theorem 20 and we omit it here. If $\ell \geq 2$, we have $\ell^2/2 \geq 2(\ell - 1)$ and the proof is the same as *Case I.* Otherwise if $\ell = 1$, the inequality

$$\|C - \tilde{C}\|_1 \geq \frac{4(\ell-1)}{3n^2} = 0$$

always holds.

By combining the above three cases, it holds with probability at least $1 - 9n^{-\eta/2}$ that

$$\|C - \tilde{C}\|_1 \geq \frac{4(\ell - 1)}{3n^2}.$$

□

Now, we are ready to prove Theorem 22.

Proof of Theorem 22. Suppose that the instance $\mathcal{MC}(\tilde{C}, \tilde{u}^*) \in \overline{\mathcal{D}}$ attains the maximum in (3.6). Denote

$$d_\alpha := \alpha\|C - \tilde{C}\|_1 + (1 - \alpha)\|u^* - \tilde{u}^*\|_1.$$

Let

$$k := |\mathcal{I}_1(C, u^*)|, \quad \ell := |\mathcal{I}_1(\tilde{C}, \tilde{u}^*)|.$$

Similar to Theorem 20, our goal is to decide the optimal global solution $\tilde{u}^*$. By Lemma 17, the high-probability lower bound of $\|C - \tilde{C}\|_1$ is increasing in ℓ. Hence, the optimal choice of ℓ is not larger than k. We then analyze two cases.

Case I. We first consider the case when $\ell < n/(2\mu)$. Since $\ell \geq 1$, it follows that $\mu < n/2$. By Lemma 13, at least $k - \ell > n/(2\mu)$ nonzero entries in u^* are equal to 0 in $\tilde{u}^*$. Hence, the distance between u^* and $\tilde{u}^*$ satisfies

$$\|u^* - \tilde{u}^*\|_1 \geq 2\left(1 - \frac{n}{2\mu} \cdot \frac{\mu}{n}\right) \geq 1.$$

Therefore, it holds that

$$\begin{aligned}\mathbb{D}_\alpha(C, u^*) = d_\alpha^{-1} &= \left[\alpha\|C - \tilde{C}\|_1 + (1 - \alpha)\|u^* - \tilde{u}^*\|_1\right]^{-1} \\ &\leq \frac{1}{1 - \alpha} \leq \frac{1}{2(1 - \alpha)} \cdot \left(1 - \frac{\mu}{n}\right)^{-1} = \frac{1}{2(1 - \alpha)\mu} \cdot \left(\frac{1}{\mu} - \frac{1}{n}\right)^{-1}.\end{aligned}$$

Case II. Next, we focus on the case when $\ell \geq n/(2\mu)$. By Lemma 17, it holds with probability at least $1 - 3n^{-\eta/2}$ that

$$\|C - \tilde{C}\|_1 \geq \frac{4(\ell - 1)}{3n^2}. \tag{3.48}$$

By considering the union bound over $\ell \in \mathcal{L} := \{\lceil n/(2\mu)\rceil, \ldots, k\}$, the probability that inequality (3.48) holds for all $\ell \in \mathcal{L}$ is at least

$$1 - \left(\ell - \frac{n}{2\mu}\right) \cdot 3n^{-\eta/2} \geq 1 - 3n^{-\eta/2+1}.$$

In the remainder of this proof, we assume that inequality (3.48) holds for all $\ell \in \mathcal{L}$. In addition, we assume without loss of generality that

$$|u_1^*| \geq |u_2^*| \geq \cdots \geq |u_k^*| > 0; \quad |u_i^*| = 0, \quad \forall i \geq k+1.$$

By the assumption of this case, at least $k - \ell$ nonzero entries in u^* are equal to 0 in $\tilde{u}^*$. Then, we can estimate that

$$d_\alpha \geq \min_{n/(2\mu) \leq \ell \leq k} \left[\frac{4\alpha(\ell-1)}{3n^2} + 2(1-\alpha) \sum_{i=\ell+1}^{k} |u_i^*| \right] \geq \min_{1 \leq \ell \leq k} \left[\frac{4\alpha(k-1)}{3n^2} + 2(1-\alpha) \sum_{i=\ell+1}^{k} |u_i^*| \right].$$

The above minimization problem can be solved in closed form, which leads to

$$d_\alpha \geq \sum\nolimits_{\ell=1}^{k} \min \left\{ \frac{4\alpha}{3n^2}, 2(1-\alpha)|u_i^*| \right\}.$$

By the second property in Lemma 13, we have

$$\begin{aligned} d_\alpha &\geq \sum_{i=2}^{k} \min \left\{ \frac{4\alpha}{3\mu n}|u_i^*|, 2(1-\alpha)|u_i^*| \right\} = \min \left\{ \frac{4\alpha}{3\mu n}, 2(1-\alpha) \right\} \sum_{i=2}^{k} |u_i^*| \\ &\geq \min \left\{ \frac{4\alpha}{3\mu n}, 2(1-\alpha) \right\} \cdot \left(1 - \frac{\mu}{n}\right) = \min \left\{ \frac{4\alpha}{3n}, 2(1-\alpha)\mu \right\} \cdot \left(\frac{1}{\mu} - \frac{1}{n} \right). \end{aligned}$$

The desired upper bound follows from $\mathbb{D}_\alpha(C, u^*) = d_\alpha^{-1}$.

By combining the above two cases, the distance d_α satisfies

$$\mathbb{D}_\alpha(C, u^*) \leq \max \left\{ \frac{3n}{4\alpha}, \frac{1}{2(1-\alpha)\mu} \right\} \cdot \left(\frac{1}{\mu} - \frac{1}{n} \right)^{-1} \tag{3.49}$$

with probability at least $1 - 3n^{-\eta/2+1}$.

In the case when $\mu \geq n/16$, the sampling probability p is equal to 1 and the instance $\mathcal{MC}(C, u^*)$ satisfies the $\mathrm{RIP}_{2,2}$ condition with $\delta = 0$. Hence, we can utilize the upper bound in Theorem 20 to obtain

$$\mathbb{D}_\alpha(C, u^*) \leq \max \left\{ \frac{n}{4\alpha}, \frac{1}{2(1-\alpha)\mu} \right\} \times \min \left\{ \left(\frac{1}{\mu} - \frac{1}{n} \right)^{-1}, 3\mu \right\}.$$

Combing with the upper bound in (3.49), we conclude the proof of the theorem. □

Reduction of Problem (3.13)

Before discussing the properties of problem instances in Section 3.3, we prove that the SSCPs of the instance $\mathcal{MC}(C^\epsilon, u^*)$ are closely related to those of the m-dimensional problem

$$\min_{x \in \mathbb{R}^m} \quad \sum\nolimits_{i \in [m]} (x_i^2 - 1)^2 + \epsilon \sum\nolimits_{i,j \in [m], i \neq j} (x_i x_j - 1)^2. \tag{3.50}$$

Lemma 18. *If problem* (3.50) *has no SSCPs, then the instance* $\mathcal{MC}(C^\epsilon, u^*)$ *has no SSCPs. In addition, given a number* $N \in \mathbb{N}$*, suppose that problem* (3.50) *has* N *SSCPs with nonzero components at which the objective function has a positive definite Hessian matrix. Then, the instance* $\mathcal{MC}(C^\epsilon, u^*)$ *has at least* N *spurious local minima.*

Proof. To prove the first part of the theorem, we assume that problem (3.50) has no SSCPs. Suppose that $u^0 \in \mathbb{R}^n$ is a second-order critical point of the instance $\mathcal{MC}(C^\epsilon, u^*)$. Calculating the gradient of $g(u; C, u^*)$ with respect to u_i for any index $i \geq m$ leads to

$$Z_\epsilon \nabla_i g(u^0; C^\epsilon, u^*) = 4(u_i^0)^3 + 4\sum\nolimits_{j\in[n],\{i,j\}\in\mathbb{E}} u_i^0 (u_j^0)^2 = 0,$$

where $\nabla_i g(\cdot; C^\epsilon, u^*)$ is i-th component of the gradient. By multiplying u_i^0 on both sides, it follows that

$$4(u_i^0)^4 + 4(u_i^0)^2 \sum\nolimits_{j\in[n],\{i,j\}\in\mathbb{E}} (u_j^0)^2 = 0,$$

which implies that $u_i^0 = 0$ for all $i \in \{m+1, \ldots, n\}$. Calculating the gradient and the Hessian matrix with respect to $u_{1:m}$ yields that

$$\begin{aligned}
Z_\epsilon \nabla_i g(u^0; C^\epsilon, u^*) &= 4\epsilon \sum\nolimits_{j\in[m], j\neq i} u_j^0 (u_i^0 u_j^0 - 1/m^2) + 4u_i^0[(u_i^0)^2 - 1/m^2], \quad \forall i \in [m];\\
Z_\epsilon \nabla_{ii}^2 g(u^0; C^\epsilon, u^*) &= 12(u_i^0)^2 - 4/m^2 + 4\epsilon \sum\nolimits_{j\in[m], j\neq i} (u_j^0)^2, \quad \forall i \in [m];\\
Z_\epsilon \nabla_{ij}^2 g(u^0; C^\epsilon, u^*) &= 4\epsilon(2u_i^0 u_j^0 - 1), \quad \forall i, j \in [m] \quad \text{s.t. } i \neq j,
\end{aligned}$$

where $\nabla_{ij} g(\cdot; C^\epsilon, u^*)$ is the (i,j)-th component of the Hessian matrix. By defining $x^0 \in \mathbb{R}^m$ as $x_i^0 := m u_i^0$ for all $i \in [m]$, the above gradient and Hessian matrix turn out to be the same as those of problem (3.50). Since the first m entries of $\nabla g(u^0; C^\epsilon, u^*)$ are 0 and the first m-by-m principle sub-matrix of $\nabla^2 g(u^0; C^\epsilon, u^*)$ is positive semi-definite, the point x^0 is a second-order critical point of problem (3.50). In addition, the point u^0 is a global optimum if and only if $|u_i^0| = 1/m$ for all $i \in [m]$, which is further equivalent to $x_i^0 = 1$ for all $i \in [m]$ and x^0 is the global solution to problem (3.50). Therefore, the point x^0 is a SSCP if u^0 is a SSCP, which is a contradiction to the assumption that problem (3.50) has no SSCPs. Therefore, the point u^0 is a global minimum of the instance $\mathcal{MC}(C^\epsilon, u^*)$.

For the second part of the theorem, suppose that x^0 is a SSCP of problem (3.50), where the Hessian matrix is positive definite and $x_i^0 \neq 0$ for all $i \in [m]$. We construct $u^0 \in \mathbb{R}^n$ by setting $u_i^0 := m^{-1} x_i^0$ for all $i \in [m]$ and $u_i^0 = 0$ for all $i \in \{m+1, \ldots, n\}$. By similar calculations, we can prove that the Hessian matrix at u^0 is a block diagonal matrix with two blocks, where the first block is $H(x; \epsilon)$ and the second block is a diagonal matrix with positive diagonal entries. Moreover, the gradient at u^0 is equal to 0. Hence, u^0 is a SSCP with a positive definite Hessian matrix. The construction shows that the mapping from x^0 to u^0 is injective. □

Proof of Theorem 23

To simplify the notations in the following proofs, we denote the gradient and the Hessian matrix of the objective function of problem (3.50) by

$$\begin{aligned}
\mathrm{g}_i(x;\epsilon) &:= 4\left[x_i^3 - x_i + \epsilon\sum\nolimits_{j\neq i} x_j(x_i x_j - 1)\right], \quad \forall i \in [m];\\
H_{ii}(x;\epsilon) &:= 4\left[3x_i^2 - 1 + \epsilon\sum\nolimits_{j\neq i} x_j^2\right], \quad \forall i \in [m];\\
H_{ij}(x;\epsilon) &:= 4\epsilon(2x_i x_j - 1), \quad \forall i,j \in [m] \quad \text{s.t. } i \neq j.
\end{aligned}$$

The following theorem guarantees that the instance $\mathcal{MC}(C^\epsilon, u^*)$ does not have spurious local minima when $\epsilon \geq O(m^{-1})$.

Theorem 30. *If $\epsilon > 18/m$, the instance $\mathcal{MC}(C^\epsilon, u^*)$ does not have SSCPs, namely, all second-order critical points are global minima associated with the ground truth solution M^*.*

Proof. By Lemma 18, we only need to prove that problem (3.50) has no SSCPs. The conclusion holds when $\epsilon = 1$ since the δ-RIP$_{2,2}$ condition holds with $\delta = 0$ and the results in [247] guarantee that there is no SSCP. In the remainder of the proof, we assume that $\epsilon \in [0, 1)$. Suppose that $x^0 \in \mathbb{R}^m$ is a second-order critical point of problem (3.50). Denote

$$S_k := \sum\nolimits_{i=1}^m (x_i^0)^k, \quad \forall k \in \mathbb{N}.$$

Using the first-order optimality conditions, we have

$$\begin{aligned}
0 &= \frac{1}{4}\sum\nolimits_{i\in[m]} g_i(x^0;\epsilon) = (1-\epsilon)S_3 - (1-\epsilon)S_1 - m\epsilon S_1 + \epsilon S_1 S_2, \qquad (3.51)\\
0 &= \frac{1}{4}\sum\nolimits_{i\in[m]} x_i^0 g_i(x^0;\epsilon) = (1-\epsilon)S_4 - (1-\epsilon)S_2 - \epsilon S_1^2 + \epsilon S_2^2.
\end{aligned}$$

Using the second-order necessary optimality conditions, the curvatures of the objective function along the directions

$$c_+ := (x_1^0 - 1, \ldots, x_m^0 - 1) \quad \text{and} \quad c_- := (x_1^0 + 1, \ldots, x_m^0 + 1)$$

are given by

$$\begin{aligned}
c_+^T H(x;\epsilon)c_+/4 &= 3(1-\epsilon)(S_4 - 2S_3 + S_2) + [\epsilon S_2 - (1-\epsilon)](S_2 - 2S_1 + m)\\
&\quad + 2\epsilon(S_2^2 - 2S_2S_1 + S_1^2) - \epsilon(S_1^2 - 2nS_1 + m^2) \geq 0,\\
c_-^T H(x;\epsilon)c_-/4 &= 3(1-\epsilon)(S_4 + 2S_3 + S_2) + [\epsilon S_2 - (1-\epsilon)](S_2 + 2S_1 + m)\\
&\quad + 2\epsilon(S_2^2 + 2S_2S_1 + S_1^2) - \epsilon(S_1^2 + 2nS_1 + m^2) \geq 0.
\end{aligned}$$

Using the relations in (3.51), we can write S_3 and S_4 in terms of S_1 and S_2, which leads to

$$[m\epsilon + 5(1-\epsilon)]S_2 + 4\epsilon S_1^2 - 4[m\epsilon + (1-\epsilon)] \cdot |S_1| - [m^2\epsilon + m(1-\epsilon)] \geq 0. \qquad (3.52)$$

Let c be a positive number such that

$$S_1^2 = cS_2.$$

Using Hölder's inequality, we have $c \in [1, m]$. We note that in the case when $S_2 = 0$, it holds that $S_1 = 0$ and we can choose c to be any constant in $[1, m]$. Then, inequality (3.52) can be written as

$$[m\epsilon + 5(1-\epsilon) + 4\epsilon c]S_2 - 4[m\epsilon + (1-\epsilon)]\sqrt{c} \cdot \sqrt{S_2} - [m^2\epsilon + m(1-\epsilon)] \geq 0. \tag{3.53}$$

Inequality (3.53) is a quadratic inequality in $\sqrt{S_2}$ and thus, it can be solved in closed form, namely, inequality (3.53) is equivalent to

$$\begin{aligned} &\sqrt{S_2} \\ \geq &\frac{4[m\epsilon + (1-\epsilon)]\sqrt{c} + \sqrt{4[m\epsilon + (1-\epsilon)][8m\epsilon c + 4(1-\epsilon)c + m^2\epsilon + 5m(1-\epsilon)]}}{2[m\epsilon + 5(1-\epsilon) + 4\epsilon c]} \\ = &m\sqrt{m\epsilon + (1-\epsilon)} \cdot \left[\sqrt{[8m\epsilon + 4(1-\epsilon)]c + m^2\epsilon + 5m(1-\epsilon)} - \sqrt{4[m\epsilon + (1-\epsilon)]c}\right]^{-1}. \end{aligned} \tag{3.54}$$

Consider the function

$$e(c) := \sqrt{[8m\epsilon + 4(1-\epsilon)]c + m^2\epsilon + 5m(1-\epsilon)} - \sqrt{4[m\epsilon + (1-\epsilon)]c}, \quad \forall c \in [1, m],$$

which is the negative of a unimodal function[3]. Hence, the maximum value of $e(c)$ on $[1, m]$ is attained at 1 or m. Let

$$C := m\epsilon > 18.$$

We calculate that

$$\begin{aligned} e(m) &= \sqrt{9m[m\epsilon + (1-\epsilon)]} - \sqrt{4m[m\epsilon + (1-\epsilon)]} \\ &= \sqrt{m[m\epsilon + (1-\epsilon)]} \leq \sqrt{m(C+1)} \leq \sqrt{2mC}, \\ e(1) &= \sqrt{8m\epsilon + 4(1-\epsilon) + m^2\epsilon + 5m(1-\epsilon)} - \sqrt{4[m\epsilon + (1-\epsilon)]} \\ &\leq \sqrt{8C + 4 + mC + 5m} \leq \sqrt{2(m+8)C}. \end{aligned}$$

Hence, we have

$$e(c) \leq \sqrt{2(m+8)C}, \quad \forall c \in [1, m].$$

By combining with (3.54), it follows that

$$\sqrt{S_2} \geq m\sqrt{C + (1-\epsilon)} \cdot \left[\sqrt{2(m+8)C}\right]^{-1} \geq m\sqrt{C} \cdot \left[\sqrt{2(m+8)C}\right]^{-1} = \frac{m}{\sqrt{2(m+8)}},$$

[3]We say a function $f : \mathbb{R} \mapsto \mathbb{R}$ is a *unimodal function* if there exists a constant $c \in \mathbb{R}$ such that f is increasing on $(-\infty, c]$ and decreasing on $[c, +\infty)$.

which further leads to

$$S_2 \geq \frac{m^2}{2(m+8)} \geq \frac{m}{18}. \tag{3.55}$$

Therefore, we obtain that

$$\frac{\epsilon}{1-\epsilon} S_2 - 1 \geq \frac{\epsilon m}{18} - 1 > 0.$$

Using the first-order optimality condition, each component x_i^0 is the solution to the third-order polynomial equation

$$g_i(x;\epsilon) = x_i^3 + \left[\frac{\epsilon}{1-\epsilon} S_2 - 1\right] x_i - \frac{\epsilon}{1-\epsilon} S_1 = 0, \quad \forall i \in [m]. \tag{3.56}$$

Since the first-order coefficient $\epsilon/[(1-\epsilon)S_2] - 1$ is positive, the derivative of the polynomial is positive and the equation has a unique real root x_0. Hence, we know

$$x_1^0 = \cdots = x_m^0 = x_0.$$

The equation in (3.56) now becomes

$$x_0^3 + \left[\frac{\epsilon}{1-\epsilon} \cdot m x_0^2 - 1\right] x_0 - \frac{\epsilon}{1-\epsilon} \cdot m x_0 = \left[\frac{m\epsilon}{1-\epsilon} + 1\right](x_0^3 - x_0) = 0,$$

which gives $x_0 \in \{-1, 0, 1\}$. If $x_0 \in \{-1, 1\}$, then the point x^0 is a global optimum. Otherwise if $x_0 = 0$, it follows that $x^0 = 0$ and $S_2 = 0$, which contradicts (3.55). Combining the two cases, we conclude that problem (3.50) does not have SSCPs, which implies that the instance $\mathcal{MC}(C^\epsilon, u^*)$ also has no SSCPs. □

Then, we consider the regime of ϵ where the instance $\mathcal{MC}(C^\epsilon, u^*)$ has spurious solutions. The following theorem studies the case when m is an even number.

Theorem 31. *Suppose that m is an even number. If $\epsilon < 1/(m+1)$, then the instance $\mathcal{MC}(C^\epsilon, u^*)$ has at least $2^{m/2}$ spurious local minima.*

Proof. By Lemma 18, we only need to show that problem (3.50) has at least $\binom{m}{m/2}$ SSCPs whose associated Hessian matrices are positive definite and whose components are nonzero. We consider a point $x^0 \in \mathbb{R}^m$ such that

$$(x_i^0)^2 = \frac{1-\epsilon}{1+(m-1)\epsilon} > 0, \quad \forall i \in [m]; \quad \sum\nolimits_{i\in[m]} x_i^0 = 0.$$

The above equations have a solution since m is an even number. By a direct calculation, we can verify that the gradient $g(x^0;\epsilon)$ is equal to 0. We only need to show that the Hessian matrix $H(x^0;\epsilon)$ is positive definite, namely

$$c^T H(x^0;\epsilon) c > 0, \quad \forall c \in \mathbb{R}^m \backslash \{0\}.$$

The above condition is equivalent to

$$\left[(3+(m-3)\epsilon)\left(x_1^0\right)^2-1+\epsilon\right]\sum\nolimits_{i\in[m]}c_i^2-\epsilon\left(\sum\nolimits_{i\in[m]}c_i\right)^2 + 2\epsilon\left(x_1^0\right)^2\left(\sum\nolimits_{i\in[m]}\mathrm{sign}(x_i^0)c_i\right)^2>0,\quad \forall c\in\mathbb{R}^n\backslash\{0\}.$$

Under the normalization constraint $\|c\|_2=1$, the Cauchy inequality implies that the minimum of the left-hand side is attained by

$$c_1=\cdots=c_m=1/\sqrt{m}.$$

Therefore, the Hessian is positive definite if and only if

$$(3+(m-3)\epsilon)\left(x_1^0\right)^2-1+\epsilon>m\epsilon.$$

By substituting $(x_1^0)^2=(1-\epsilon)/[1+(m-1)\epsilon]$, the above condition is equivalent to

$$2-(m+4)\epsilon-(m-2)(m+1)\epsilon^2>0.$$

Using the condition that $(m+1)\epsilon<1$, we obtain that

$$2-(m+4)\epsilon-(m-2)(m+1)\epsilon^2>1-3\epsilon-(m-2)\epsilon=1-(m+1)\epsilon>0,$$

where the first inequality is from the fact that $m\geq 2$, which follows from the assumption that $m>0$ is an even number.

To estimate the number of SSCPs, we observe that $m/2$ components of x^0 have a positive sign and the other $m/2$ components have a negative sign. Hence, there are at least

$$\binom{m}{m/2}$$

spurious SSCPs. The estimate on the combinatorial number is in light of the inequality $\binom{n}{k}\geq(n/k)^k$. □

The estimation of the odd number case is similar and we present the result in the following theorem.

Theorem 32. *Suppose that m is an odd number. If $\epsilon<1/[13(m+1)]$, then the instance $\mathcal{MC}(C^\epsilon,u^*)$ has at least $[2m/(m+1)]^{(m+1)/2}$ spurious local minima.*

Proof. We pursue a similar way as in Theorem 31 to construct spurious solutions. By Lemma 18, we only need to show that problem (3.50) has at least $\binom{m}{(m-1)/2}$ SSCPs whose Hessian matrices are positive definite and whose components are nonzero. Let $k:=(m-1)/2\in\mathbb{Z}$. We first choose a subset

$$\mathcal{I}\subset[m],\quad |\mathcal{I}|=k.$$

Then, we consider the point $x \in \mathbb{R}^m$, where

$$u_i = y_1, \quad \forall i \in \mathcal{I}, \quad u_i = y_2, \quad \forall i \notin \mathcal{I},$$

where y_1 and y_2 are real numbers such that

$$\begin{aligned} &(1+k\epsilon)(1+2k\epsilon)[(1-\epsilon)y_2^2]^3 - 2(1+k\epsilon)(1+(k-1)\epsilon)[(1-\epsilon)y_2^2]^2 \\ &+ (1+k\epsilon)(1+(k-1)\epsilon)(2k^2\epsilon^2+2k\epsilon^2-k\epsilon-\epsilon+1)[(1-\epsilon)y_2^2] \\ &\qquad - k^2\epsilon^2(1+(k-1)\epsilon)(1-\epsilon)^2 = 0, \\ y_1 = \frac{y_2}{k\epsilon} &\cdot \frac{(1+k\epsilon)[(1-\epsilon)y_2^2] - (k^2\epsilon^2+(k-1)\epsilon+1)}{[(1-\epsilon)y_2^2]+(1+(k-1)\epsilon)}. \end{aligned} \tag{3.57}$$

We first assume the existence of the constants y_1 and y_2. After some direct calculations, one can show that the conditions in (3.57) imply the first-order optimality condition of the instance $\mathcal{MC}(C^\epsilon, u^*)$, i.e.,

$$\begin{aligned} y_1^3 - y_1 + \epsilon[(k-1)y_1^2 + (k+1)y_2^2]y_1 - \epsilon[(k-1)y_1 + (k+1)y_2] = 0, \\ y_2^3 - y_2 + \epsilon[ky_1^2 + ky_2^2]y_2 - \epsilon[ky_1 + ky_2] = 0. \end{aligned}$$

Therefore, the point x is a first-order critical point of the instance $\mathcal{MC}(C^\epsilon, u^*)$. In addition, the following relations result from the condition (3.57):

$$\begin{aligned} (1-\epsilon)y_1y_2(y_1+y_2) &= -\epsilon[ky_1 + (k+1)y_2], \\ (1-\epsilon)(y_1^2+y_1y_2+y_2^2-1) &= -\epsilon[ky_1^2+(k+1)y_2^2]. \end{aligned} \tag{3.58}$$

Now, we prove the existence of y_1, y_2 and estimate their values. We note that the first equation in (3.57) is a third-order polynomial equation for $(1-\epsilon)y_2^2$, which has at least one real root. To show that the equation has a positive root, we observe that the coefficient of the third-order term is $(1+k\epsilon)(1+2k\epsilon) > 0$ and the value at zero is $-k^2\epsilon^2(1+(k-1)\epsilon)(1-\epsilon)^2 < 0$. Therefore, the polynomial equation for $(1-\epsilon)y_2^2$ has at least one positive root and y_2 is well defined. We provide a more accurate estimate to y_1 and y_2, namely, we show that there exists a solution (y_1, y_2) to equations (3.57) such that

$$y_1 \in [-2, -3/5], \quad y_2 \in [1/2, 1].$$

Define the polynomial function

$$\begin{aligned} g(z) :=& (1+k\epsilon)(1+2k\epsilon)z^3 - 2(1+k\epsilon)(1+(k-1)\epsilon)z^2 \\ &+ (1+k\epsilon)(1+(k-1)\epsilon)(2k^2\epsilon^2+2k\epsilon^2-k\epsilon-\epsilon+1)z - k^2\epsilon^2(1+(k-1)\epsilon)(1-\epsilon)^2. \end{aligned}$$

We first estimate $g(1-(2k+1)\epsilon)$ as follows:

$$g(1-(2k+1)\epsilon)$$

$$
\begin{aligned}
=&(1+k\epsilon)[1-(2k+1)\epsilon]\Big[(1+2k\epsilon)[1-(2k+1)\epsilon]^2-2[1+(k-1)\epsilon][1-(2k+1)\epsilon]\\
&+[1+(k-1)\epsilon][1-(k+1)\epsilon+2k(k+1)\epsilon^2]\Big]-k^2\epsilon^2(1+(k-1)\epsilon)(1-\epsilon)^2\\
=&(1+k\epsilon)[1-(2k+1)\epsilon]\Big[k^2\epsilon^2+2k^2(5k+4)\epsilon^3\Big]-k^2\epsilon^2(1+(k-1)\epsilon)(1-\epsilon)^2\\
\geq&k^2\epsilon^2(1+k\epsilon)[1-(2k+1)\epsilon][1+2(5k+4)\epsilon]-k^2\epsilon^2(1+k\epsilon)(1-\epsilon)^2\\
=&k^2\epsilon^2(1+k\epsilon)[(8k+9)\epsilon-[2(2k+1)(5k+4)+1]\epsilon^2]\\
\geq&k^2\epsilon^2(1+k\epsilon)[(8k+8)\epsilon-20(k+1)^2\epsilon^2]>0,
\end{aligned}
$$

where the last inequality is due to $(k+1)\epsilon=(n+1)\epsilon/2<2/5$. Next, we estimate $g(1-(3k/2+1)\epsilon)$ as follows:

$$
\begin{aligned}
&g(1-(3k/2+1)\epsilon)\\
=&(1+k\epsilon)[1-(k+1)\epsilon]\Big[(1+2k\epsilon)[1-(3k/2+1)\epsilon]^2-2[1+(k-1)\epsilon][1-(3k/2+1)\epsilon]\\
&+[1+(k-1)\epsilon][1-(k+1)\epsilon+2k(k+1)\epsilon^2]\Big]-k^2\epsilon^2(1+(k-1)\epsilon)(1-\epsilon)^2\\
=&(1+k\epsilon)[1-(3k/2+1)\epsilon]\Big[k^2\epsilon^2/4+k^2(13k/2+6)\epsilon^3\Big]-k^2\epsilon^2(1+(k-1)\epsilon)(1-\epsilon)^2\\
=&k^2\epsilon^2(1+k\epsilon)[1-(3k/2+1)\epsilon][1/4+(13k/2+6)\epsilon]-k^2\epsilon^2(1+(k-1)\epsilon)(1-\epsilon)^2\\
\leq&k^2\epsilon^2(1+k\epsilon)[1-(3k/2+1)\epsilon][1/4+(13k/2+6)\epsilon]-k^2\epsilon^2\cdot[(1+k\epsilon)/2]\cdot(1-\epsilon)^2\\
\leq&k^2\epsilon^2(1+k\epsilon)\Big[[1-(3k/2+1)\epsilon][1/4+(13k/2+6)\epsilon]-(1-\epsilon)^2/2\Big]\\
=&k^2\epsilon^2(1+k\epsilon)\Big[-1/4+(49k/8+27/4)\epsilon-(39k^2/4+31k/2+13/2)\epsilon^2\Big]\\
\leq&k^2\epsilon^2(1+k\epsilon)\Big[-1/4+27(k+1)/4\epsilon-39(k+1)^2\epsilon^2/4\Big]<0,
\end{aligned}
$$

where the last inequality is in light of $(k+1)\epsilon=(n+1)\epsilon/2<1/26$. Combining the above two estimates, we conclude that there exists a solution y_2 to the first equation in (3.57) such that

$$
(1-\epsilon)y_2^2\in[1-(2k+1)\epsilon,1-(3k/2+1)\epsilon]. \tag{3.59}
$$

Hence,

$$
y_2\leq\sqrt{\frac{1-(3k/2+1)\epsilon}{1-\epsilon}}\leq 1 \tag{3.60}
$$

and

$$
y_2\geq\sqrt{\frac{1-(2k+1)\epsilon}{1-\epsilon}}\geq\sqrt{1-(2k+1)\epsilon}\geq\frac{1}{2}. \tag{3.61}
$$

Now, we use the second equation in (3.57) to estimate y_1, which leads to

$$\begin{aligned}&\frac{(1+k\epsilon)[(1-\epsilon)y_2^2]-(k^2\epsilon^2+(k-1)\epsilon+1)}{k\epsilon}\\ \geq&\frac{(1+k\epsilon)[1-(2k+1)\epsilon]-(k^2\epsilon^2+(k-1)\epsilon+1)}{k\epsilon}\\ =&-2-(3k+1)\epsilon\end{aligned}$$

and

$$\begin{aligned}&\frac{(1+k\epsilon)[(1-\epsilon)y_2^2]-(k^2\epsilon^2+(k-1)\epsilon+1)}{k\epsilon}\\ \leq&\frac{(1+k\epsilon)[1-(3k/2+1)\epsilon]-(k^2\epsilon^2+(k-1)\epsilon+1)}{k\epsilon}\\ =&-\frac{3}{2}-\left(\frac{5k}{2}+1\right)\epsilon.\end{aligned}$$

On the other hand, we have

$$\frac{y_2}{[(1-\epsilon)y_2^2]+(1+(k-1)\epsilon)}=\frac{1}{(1-\epsilon)(y_2+y_2^{-1})+k\epsilon}\leq\frac{1}{2(1-\epsilon)+k\epsilon}.$$

Using the bound in (3.59), it holds that

$$y_2\geq\sqrt{\frac{1-(2k+1)\epsilon}{1-\epsilon}}\geq\frac{1-(2k+1)\epsilon}{1-\epsilon}=\frac{1}{2}-\frac{1-(4k+1)\epsilon}{2(1-\epsilon)}\geq\frac{1}{2}.$$

Therefore,

$$\frac{y_2}{[(1-\epsilon)y_2^2]+(1+(k-1)\epsilon)}=\frac{1}{(1-\epsilon)(y_2+y_2^{-1})+k\epsilon}\geq\frac{1}{2.5(1-\epsilon)+k\epsilon}.$$

Combining the above inequalities and the second equation in (3.57) yields that

$$y_1\geq\frac{-2-(3k+1)\epsilon}{2(1-\epsilon)+k\epsilon}\geq-\left(1+\frac{5\epsilon}{1-\epsilon}\right)\geq-2 \tag{3.62}$$

and

$$y_1\leq\frac{-3/2-(5k/2+1)\epsilon}{2.5(1-\epsilon)+k\epsilon}\leq-\frac{1.5}{2.5}=-\frac{3}{5}, \tag{3.63}$$

where the last inequality in (3.62) results from $\epsilon\leq 1/(3(k+1))\leq 1/6$. In summary, inequalities (3.60)-(3.63) lead to

$$y_1\in[-2,-3/5],\quad y_2\in[1/2,1].$$

We then prove that $y_1 + 2y_2 > y_2 \geq 0.5$, which is equivalent to

$$\frac{y_2}{k\epsilon} \cdot \frac{(1+k\epsilon)[(1-\epsilon)y_2^2] - (k^2\epsilon^2 + (k-1)\epsilon + 1)}{[(1-\epsilon)y_2^2] + (1+(k-1)\epsilon)} + y_2 > 0.$$

Since $y_2 > 0$, we only need to prove that

$$\begin{aligned} 0 &< (1+k\epsilon)[(1-\epsilon)y_2^2] - (k^2\epsilon^2 + (k-1)\epsilon + 1) + k\epsilon\Big[[(1-\epsilon)y_2^2] + (1+(k-1)\epsilon)\Big] \\ &= (1+2k\epsilon)[(1-\epsilon)y_2^2] - (k^2\epsilon^2 + (k-1)\epsilon + 1) + k\epsilon(1+(k-1)\epsilon). \end{aligned}$$

Using inequality (3.59), it suffices to show that

$$\begin{aligned} & (1+2k\epsilon)[1-(3k/2+1)\epsilon)] + k\epsilon(1+(k-1)\epsilon) > 1 + (k-1)\epsilon + k^2\epsilon^2 \\ \iff & \frac{1}{2}k\epsilon > 3k\left(k+\frac{3}{2}\right)\epsilon^2 \iff 3(2k+3)\epsilon < 1 \Leftarrow 6(k+1)\epsilon < 1, \end{aligned}$$

where the last inequality holds since $(k+1)\epsilon = (n+1)\epsilon/2 < 1/6$.

Now, we verify the second-order sufficient optimality condition. For every $c \in \mathbb{R}^m \backslash \{0\}$, we calculate that

$$\begin{aligned} c^T H(x;\epsilon)c = & \sum_{i\in\mathcal{I}} \left[3y_1^2 - 1 + \epsilon((k-1)y_1^2 + (k+1)y_2^2)\right] c_i^2 \\ & + \sum_{i\notin\mathcal{I}} \left[3y_2^2 - 1 + \epsilon(ky_1^2 + ky_2^2)\right] c_i^2 + \sum_{i,j\in\mathcal{I}, i\neq j} \epsilon\left(2y_1^2 - 1\right) c_i c_j \\ & + \sum_{i,j\notin\mathcal{I}, i\neq j} \epsilon\left(2y_2^2 - 1\right) c_i c_j + 2\sum_{i\in\mathcal{I}, j\notin\mathcal{I}} \epsilon\left(2y_1 y_2 - 1\right) c_i c_j \\ = & \left[3y_1^2 - 1 + \epsilon((k-1)y_1^2 + (k+1)y_2^2) - \epsilon\left(2y_1^2 - 1\right)\right] \sum_{i\in\mathcal{I}} c_i^2 \\ & + \left[3y_2^2 - 1 + \epsilon(ky_1^2 + ky_2^2) - \left(2y_2^2 - 1\right)\right] \sum_{i\notin\mathcal{I}} c_i^2 \\ & + \epsilon\left(2y_1^2 - 1\right)\left(\sum_{i\in\mathcal{I}} c_i\right)^2 + \epsilon\left(2y_2^2 - 1\right)\left(\sum_{i\notin\mathcal{I}} c_i\right)^2 \\ & + 2\epsilon\left(2y_1 y_2 - 1\right)\left(\sum_{i\in\mathcal{I}} c_i\right)\left(\sum_{i\notin\mathcal{I}} c_i\right). \end{aligned}$$

Using the Cauchy inequality, the above expression is positive if and only if

$$\left[3y_1^2 - 1 + \epsilon((k-1)y_1^2 + (k+1)y_2^2) - \epsilon\left(2y_1^2 - 1\right)\right] \cdot \frac{1}{k}\left(\sum_{i\in\mathcal{I}} c_i\right)^2$$

$$
\begin{aligned}
&+\left[3y_2^2-1+\epsilon(ky_1^2+ky_2^2)-\left(2y_2^2-1\right)\right]\cdot\frac{1}{k+1}\left(\sum_{i\notin\mathcal{I}}c_i\right)^2\\
&+\epsilon\left(2y_1^2-1\right)\left(\sum_{i\in\mathcal{I}}c_i\right)^2+\epsilon\left(2y_2^2-1\right)\left(\sum_{i\notin\mathcal{I}}c_i\right)^2\\
&+2\epsilon\left(2y_1y_2-1\right)\left(\sum_{i\in\mathcal{I}}c_i\right)\left(\sum_{i\notin\mathcal{I}}c_i\right)>0.
\end{aligned}
$$

We denote

$$A:=\sum\nolimits_{i\in\mathcal{I}}c_i,\quad B:=\sum\nolimits_{i\notin\mathcal{I}}c_i.$$

Then, the second-order sufficient condition is equivalent to

$$
\begin{aligned}
&\left[3y_1^2-1+\epsilon((k-1)y_1^2+(k+1)y_2^2)-\epsilon\left(2y_1^2-1\right)\right]\cdot\frac{1}{k}A^2\\
&+\left[3y_2^2-1+\epsilon(ky_1^2+ky_2^2)-\left(2y_2^2-1\right)\right]\cdot\frac{1}{k+1}B^2+2\epsilon(y_1A+y_2B)^2-\epsilon(A+B)^2>0.
\end{aligned}
$$

The above inequality is a quadratic inequality in A and B, which can be rewritten as

$$
\begin{aligned}
&\left[\frac{1}{k}\left[3y_1^2-1+\epsilon((k-1)y_1^2+(k+1)y_2^2)-\epsilon\left(2y_1^2-1\right)\right]+\epsilon(2y_1^2-1)\right]A^2\\
&+2\epsilon(2y_1y_2-1)AB\\
&+\left[\frac{1}{k+1}\left[3y_2^2-1+\epsilon(ky_1^2+ky_2^2)-\epsilon\left(2y_2^2-1\right)\right]+\epsilon(2y_2^2-1)\right]B^2>0.
\end{aligned}
$$

Therefore, the positivity condition can be verified through the discriminant, namely,

$$
\begin{aligned}
\epsilon^2(2y_1y_2-1)^2<&\left[\frac{1}{k}\left[3y_1^2-1+\epsilon((k-1)y_1^2+(k+1)y_2^2)-\epsilon\left(2y_1^2-1\right)\right]+\epsilon(2y_1^2-1)\right]\\
&\cdot\left[\frac{1}{k+1}\left[3y_2^2-1+\epsilon(ky_1^2+ky_2^2)-\epsilon\left(2y_2^2-1\right)\right]+\epsilon(2y_2^2-1)\right].
\end{aligned}
$$

Using the second property in (3.58), the above condition can be simplified into

$$
\begin{aligned}
&-(1-\epsilon)^2(y_2-y_1)^2(2y_1+y_2)(y_1+2y_2)+(k+1)\epsilon(1-\epsilon)(2y_2^2-1)(y_1-y_2)(2y_1+y_2)\\
&\quad+k\epsilon(1-\epsilon)(2y_1^2-1)(y_2-y_1)(y_1+2y_2)>k(k+1)\epsilon^2(y_1-y_2)^2.
\end{aligned}
$$

Since $y_2>y_1$, it suffices to have

$$
\begin{aligned}
&-(1-\epsilon)^2(y_2-y_1)(2y_1+y_2)(y_1+2y_2)-(k+1)\epsilon(1-\epsilon)(2y_2^2-1)(2y_1+y_2)\\
&\qquad\qquad+k\epsilon(1-\epsilon)(2y_1^2-1)(y_1+2y_2)>k(k+1)\epsilon^2(y_2-y_1).
\end{aligned}
$$

We can estimate that

$$\begin{aligned}
&-(1-\epsilon)^2(y_2-y_1)(2y_1+y_2)(y_1+2y_2)-(k+1)\epsilon(1-\epsilon)(2y_2^2-1)(2y_1+y_2)\\
&\qquad\qquad +k\epsilon(1-\epsilon)(2y_1^2-1)(y_1+2y_2)-k(k+1)\epsilon^2(y_2-y_1)\\
&\geq [1-(k+1)\epsilon]^2\cdot 1.1\cdot 0.2\cdot 0.5-(k+1)\epsilon\cdot 1\cdot 0.5\cdot 2\\
&\qquad\qquad -(k+1)\epsilon\cdot 1\cdot 0.64\cdot 1.4-(k+1)^2\epsilon^2\cdot 3\\
&=0.11[1-(k+1)\epsilon]^2-1.896(k+1)\epsilon-3(k+1)^2\epsilon^2\\
&=0.11-2.116(k+1)\epsilon-2.89(k+1)^2\epsilon^2\\
&\geq 0.11-2.116(k+1)\epsilon-2.89(k+1)^2\epsilon^2>0,
\end{aligned}$$

where the last inequality is due to $(k+1)\epsilon=(n+1)\epsilon/2<1/26$. Thus, we have shown that the Hessian matrix is positive definite and the point x is a SSCP.

To count the number of spurious solutions, we notice that the subset $\mathcal{I}$ has $\binom{m}{(m+1)/2}$ different choices. Hence, the total number of SSCPs is at least $\binom{m}{(m+1)/2}$. The estimate on the combinatorial number follows from $\binom{n}{k}\geq(n/k)^k$. □

By combining Theorems 30-32, we complete the proof of Theorem 23.

Proof of Theorem 24

The proof of Theorem 24 relies on the following lemma, which calculates the complexity metric of the instance $\mathcal{MC}(C^\epsilon,u^*)$. The proof of Lemma 19 is similar to that of Theorem 16.

Lemma 19. *Suppose that $n\geq m\geq 5$, $\alpha\in[0,1]$ and $\epsilon\in[0,1]$. The complexity metric $\mathbb{D}_\alpha(C^\epsilon,u^*)$ has the closed form*

$$[\mathbb{D}_\alpha(C^\epsilon,u^*)]^{-1}=\min\left\{\frac{2\alpha}{Z_\epsilon}+\frac{2(1-\alpha)(m-1)}{m},\frac{4\alpha\epsilon}{Z_\epsilon}+\frac{2(1-\alpha)(m-2)}{m},\frac{4\alpha(m-1)\epsilon}{Z_\epsilon}\right\}.$$

Moreover, $\mathbb{D}_\alpha(C^\epsilon,u^*)$ is strictly decreasing in ϵ on $[0,1/2]$.

Proof. We fix ϵ, α and m in the proof. Let $\mathcal{MC}(\tilde{C},\tilde{u}^*)$ be an instance that attains the minimum in (3.6) and $\ell:=|\mathcal{I}_1(\tilde{C},\tilde{u}^*)|$. Denote

$$d_\alpha:=\alpha\|C-\tilde{C}\|_1+(1-\alpha)\|u^*-\tilde{u}^*\|_1.$$

Then, we investigate three different cases.

Case I. Suppose that $\mathbb{G}_1(\tilde{C},\tilde{u}^*)$ is disconnected. In this case, at least $2(\ell-1)$ non-diagonal entries of $\tilde{C}$ are equal to 0. This implies that

$$\|C^\epsilon-\tilde{C}\|_1\geq 4(\ell-1)\cdot(\epsilon/Z_\epsilon). \tag{3.64}$$

Case II. The case when $\mathcal{I}_{00}(\tilde{C}, \tilde{u}^*)$ is non-empty can be analyzed similarly as *Case I* and the inequality (3.64) holds. We omit the proof for brevity.

Case III. Finally, we consider the case when $\mathbb{G}_1(\tilde{C}, \tilde{u}^*)$ is bipartite. If $\ell \geq 5$, at least $2(\ell-1)$ non-diagonal entries of $\tilde{C}$ are equal to 0 and inequality (3.64) holds. If $\ell = 4$, at least 4 non-diagonal entries and 4 diagonal entries of $\tilde{C}$ are equal to 0. Hence, we have

$$\|C^\epsilon - \tilde{C}\|_1 \geq 8 \cdot \frac{\epsilon}{Z_\epsilon} + 8 \cdot \frac{1}{Z_\epsilon} = \frac{8\epsilon + 8}{Z_\epsilon} \geq \frac{12\epsilon}{Z_\epsilon}. \tag{3.65}$$

Similarly, it follows from analyzing the cases with $\ell = 1, 2, 3$ that

$$\begin{aligned} &\|C^\epsilon - \tilde{C}\|_1 \geq (4\epsilon + 6)/(Z_\epsilon) \geq 8\epsilon/Z_\epsilon, \\ &\|C^\epsilon - \tilde{C}\|_1 \geq 4/Z_\epsilon \geq 4\epsilon/Z_\epsilon, \\ &\|C^\epsilon - \tilde{C}\|_1 \geq 2/Z_\epsilon. \end{aligned} \tag{3.66}$$

Combining inequalities (3.64), (3.65) and (3.66), we know that

$$\|C^\epsilon - \tilde{C}\|_1 \geq N(\ell)/Z_\epsilon, \tag{3.67}$$

where $N(\ell) := 4(\ell-1)\epsilon$ if $\ell \geq 2$ and $N(1) := 2$.

Now, we consider the optimal choice of $\tilde{u}^*$. Since the distance in (3.67) is increasing in ℓ, it is not optimal to choose $\ell > m$. For every $\ell \in [m]$, at least $m - \ell$ of the first m entries of $\tilde{u}^*$ are 0. Hence, we have the lower bound

$$\|u^* - \tilde{u}^*\|_1 \geq 2(m-\ell) \cdot m^{-1}. \tag{3.68}$$

Combining inequalities (3.67) and (3.68), we have

$$d_\alpha \geq \frac{N(\ell) \cdot \alpha}{Z_\epsilon} + \frac{2(1-\alpha)(m-\ell)}{m}.$$

Taking the minimum over $\ell \in [m]$ leads to

$$d_\alpha \geq \min_{\ell \in [m]} \left[\frac{N(\ell) \cdot \alpha}{Z_\epsilon} + \frac{2(1-\alpha)(m-\ell)}{m} \right].$$

We note that the above inequality indeed attains equality with a suitable choice of $\tilde{C}$ and $\tilde{u}^*$. For all $\ell \geq 2$, we can set $\tilde{u}^*_i = 0$ for all $i \in \{\ell+1, m\}$ and make node 1 disconnected from nodes $\{2, \ldots, \ell\}$. If $\ell = 1$, we can remove the self-loop at node 1. Therefore, it holds that

$$d_\alpha = \min_{\ell \in [m]} \left[\frac{\alpha N(\ell)}{Z_\epsilon} + \frac{2(1-\alpha)(m-\ell)}{m} \right].$$

The minimum in the above equality is attained at one of the points $1, 2, m$, which gives

$$d_\alpha = \min\left\{\frac{2\alpha}{Z_\epsilon} + \frac{2(1-\alpha)(m-1)}{m}, \frac{4\alpha\epsilon}{Z_\epsilon} + \frac{2(1-\alpha)(m-2)}{m}, \frac{4\alpha(m-1)\epsilon}{Z_\epsilon}\right\}.$$

Since each component in the minimization is an increasing function in ϵ, the distance d_α is also increasing in ϵ. Results for $\mathbb{D}_\alpha(C^\epsilon, u^*)$ follow accordingly by taking the inverse of d_α.

Since the closed form expression of $\mathbb{D}_\alpha(C^\epsilon, u^*)$ is the minimum of three monotone functions in ϵ, the complexity metric is the negative of a unimodal function. For every $\epsilon \le 1/(2m)$, we can prove that

$$\frac{2\alpha}{Z_\epsilon} + \frac{2(1-\alpha)(m-1)}{m} > \min\left\{\frac{4\alpha\epsilon}{Z_\epsilon} + \frac{2(1-\alpha)(m-2)}{m}, \frac{4\alpha(m-1)\epsilon}{Z_\epsilon}\right\}.$$

Therefore, in the regime $[0, 1/2]$, the complexity metric $\mathbb{D}_\alpha(C^\epsilon, u^*)$ is the minimum of two strictly decreasing functions and, thus, is also strictly decreasing in ϵ. □

Combining Theorem 23 and Lemma 19, we are able to estimate the range of the complexity metric.

Proof of Theorem 24. By defining constants $\delta := 1/26$ and $\Delta := 18$, Theorem 23 implies that

1. If $\epsilon < \delta/m$, the instance $\mathcal{MC}(C^\epsilon, u^*)$ has spurious local minima;
2. If $\epsilon > \Delta/m$, the instance $\mathcal{MC}(C^\epsilon, u^*)$ has no spurious local minima.

Then, we study two different cases.

Case I. We first consider the case when $m\epsilon$ is large. Since $\epsilon < \Delta/m \le 1/2$, the threshold is located in the regime where $\mathbb{D}_\alpha(C^\epsilon, u^*)$ is strictly decreasing. Hence, it suffices to show that

$$\left[\frac{2\alpha\Delta}{n^2} + \min\left\{4\alpha\Delta\cdot\frac{m}{n^2}, 2(1-\alpha)\right\}\right]^{-1}$$

is a lower bound on $\mathbb{D}_\alpha(C^\epsilon, u^*)$ when $\epsilon = \Delta/m$. By Lemma 19, it holds that

$$\begin{aligned}
[\mathbb{D}_\alpha(C^\epsilon, u^*)]^{-1} =& \min\left\{\frac{2\alpha}{Z_\epsilon} + \frac{2(1-\alpha)(m-1)}{m}, \frac{4\alpha\epsilon}{Z_\epsilon} + \frac{2(1-\alpha)(m-2)}{m}, \frac{4\alpha(m-1)\epsilon}{Z_\epsilon}\right\} \\
\le& \min\left\{\frac{4\alpha\epsilon}{Z_\epsilon} + \frac{2(1-\alpha)(m-2)}{m}, \frac{4\alpha(m-1)\epsilon}{Z_\epsilon}\right\} \\
=& \frac{4\alpha\epsilon}{Z_\epsilon} + (m-2)\min\left\{\frac{4\alpha\epsilon}{Z_\epsilon}, \frac{2(1-\alpha)}{m}\right\} \le \frac{4\alpha\epsilon}{Z_\epsilon} + m\min\left\{\frac{4\alpha\epsilon}{Z_\epsilon}, \frac{2(1-\alpha)}{m}\right\}.
\end{aligned}$$

Since the graph $\mathbb{G}$ does not contain any independence set with $m+1$ nodes, Turán's theorem [5] implies that the graph $\mathbb{G}$ has at least $n^2/(2m)$ edges, namely,

$$|\mathbb{E}| \geq n^2/(2m).$$

We note that the above bound is asymptotically tight and is attained by the Turán graph. Hence, we obtain that

$$Z_\epsilon = 2|\mathbb{E}| + n + m(m-1)\epsilon \geq 2|\mathbb{E}| \geq n^2/m.$$

By substituting into the estimate of $\mathbb{D}_\alpha(C^\epsilon, u^*)$, it follows that

$$\begin{aligned}[\mathbb{D}_\alpha(C^\epsilon, u^*)]^{-1} &\leq \frac{4\alpha\epsilon \cdot m}{n^2} + m \min\left\{\frac{4\alpha\epsilon \cdot m}{n^2}, \frac{2(1-\alpha)}{m}\right\} \\ &= \frac{2\alpha\Delta}{n^2} + \min\left\{4\alpha\Delta \cdot \frac{m}{n^2}, 2(1-\alpha)\right\}.\end{aligned}$$

Case II. Next, we consider the case when ϵm is small. Similar to *Case I*, it suffices to show that

$$\frac{18}{17}\max\left\{\frac{n^2}{4\alpha\delta}, \frac{1}{2(1-\alpha)}\right\}$$

is an upper bound for $\mathbb{D}_\alpha(C^\epsilon, u^*)$ when $\epsilon = \delta/m$. Since $\delta < 1/2$, we have

$$2\alpha/Z_\epsilon > 4\alpha\epsilon/Z_\epsilon.$$

By Lemma 19, it holds that

$$\begin{aligned}&[\mathbb{D}_\alpha(C^\epsilon, u^*)]^{-1} \\ =&\min\left\{\frac{2\alpha}{Z_\epsilon} + \frac{2(1-\alpha)(m-1)}{m}, \frac{4\alpha\epsilon}{Z_\epsilon} + \frac{2(1-\alpha)(m-2)}{m}, \frac{4\alpha(m-1)\epsilon}{Z_\epsilon}\right\} \\ =&\min\left\{\frac{4\alpha\epsilon}{Z_\epsilon} + \frac{2(1-\alpha)(m-2)}{m}, \frac{4\alpha(m-1)\epsilon}{Z_\epsilon}\right\} \\ =&\frac{4\alpha\epsilon}{Z_\epsilon} + (m-2)\min\left\{\frac{4\alpha\epsilon}{Z_\epsilon}, \frac{2(1-\alpha)}{m}\right\} \geq \frac{17}{18}\min\left\{\frac{4\alpha\epsilon m}{Z_\epsilon}, 2(1-\alpha)\right\},\end{aligned}$$

where the last inequality is from $m \geq 36$. Since $\epsilon \leq 1$, the definition of Z_ϵ implies that $Z_\epsilon \leq n^2$. By substituting into the estimate of $\mathbb{D}_\alpha(C^\epsilon, u^*)$, it follows that

$$[\mathbb{D}_\alpha(C^\epsilon, u^*)]^{-1} \geq \frac{17}{18}\min\left\{\frac{4\alpha\epsilon m}{n^2}, 2(1-\alpha)\right\} = \frac{17}{18}\min\left\{\frac{4\alpha\delta}{n^2}, 2(1-\alpha)\right\}.$$

By combining *Cases I* and *II*, we complete the proof. □

3.D Proofs in Section 3.4

Proof of Lemma 9

Proof. Without loss of generality, we assume that

$$u_i^0 = 1/n, \quad \forall i \in [n].$$

We first consider the scaled problem instance

$$\min_{x\in\mathbb{R}^n} \sum\nolimits_{i,j\in[n],i\neq j}(x_i x_j - 1)^2. \tag{3.69}$$

We denote the gradient and the Hessian matrix of problem (3.69) as $g(x) \in \mathbb{R}^n$ and $H(x) \in \mathbb{R}^{n\times n}$, respectively. Then, we can calculate that

$$\begin{aligned}
&\frac{1}{4}g_i(x) = -x_i^3 + (\|x\|_2^2 + 1)x_i - \sum\nolimits_{k\in[n]} x_k, \quad \forall i \in [n];\\
&\frac{1}{4}H_{ii}(x) = \sum\nolimits_{k\in[n],k\neq i} x_k^2, \quad \frac{1}{4}H_{ij}(x) = 2x_i x_j - 1, \quad \forall i, j \in [n].
\end{aligned}$$

Let c be a small positive constant and define $\epsilon := c/n$. Suppose that $x \in \mathbb{R}^n$ satisfies

$$\|g(x)\|_\infty < 4\epsilon. \tag{3.70}$$

Then, we study three different cases.

Case I. We first consider the case when $\sum_{i\in[n]} x_i > 2\epsilon$. For all $i \in [n]$, the condition (3.70) implies that

$$\frac{1}{4}|g_i(x)| = \left|\left(\sum\nolimits_{j\in[n],j\neq i} x_j^2 + 1\right)x_i - \sum\nolimits_{j\in[n]} x_j\right| < \epsilon. \tag{3.71}$$

If $x_i \leq \epsilon$, it holds that

$$\left(\sum\nolimits_{j\in[n],j\neq i} x_j^2 + 1\right)x_i - \sum\nolimits_{j\in[n]} x_j \leq x_i - \sum\nolimits_{j\in[n]} x_j < -\epsilon,$$

which contradicts (3.71). Hence,

$$x_i > \epsilon, \quad \forall i \in [n].$$

Define three index sets

$$\mathcal{I}_1 := \{i \in [n] \mid x_i \geq 1 + \epsilon\},\ \mathcal{I}_2 := \{i \in [n] \mid x_i \leq 1 - \epsilon\},\ \mathcal{I}_3 := [n]\backslash(\mathcal{I}_1 \cup \mathcal{I}_2).$$

Choosing the perturbation direction $q \in \mathbb{R}^n$ to be

$$q_i = -x_i, \quad \forall i \in \mathcal{I}_1; \quad q_i = x_i, \quad \forall i \in \mathcal{I}_2; \quad q_i = 0, \quad \forall i \in \mathcal{I}_3,$$

we can calculate that

$$\begin{aligned} \frac{1}{4} q^T g(x) = & \sum_{i,j \in \mathcal{I}_1, i \neq j} -x_i x_j (x_i x_j - 1) + \sum_{i,j \in \mathcal{I}_2, i \neq j} x_i x_j (x_i x_j - 1) \\ & + \sum_{i \in \mathcal{I}_1, j \in \mathcal{I}_3} -x_i x_j (x_i x_j - 1) + \sum_{i \in \mathcal{I}_2, j \in \mathcal{I}_3} x_i x_j (x_i x_j - 1). \end{aligned} \tag{3.72}$$

We then consider four sub-cases.

Case I-1. We first assume that $|\mathcal{I}_1| \geq 2$. In this case, we have

$$\begin{aligned} \sum_{i,j \in \mathcal{I}_1, i \neq j} -x_i x_j (x_i x_j - 1) &\leq \sum_{i,j \in \mathcal{I}_1, i \neq j} -x_i x_j [(1+\epsilon)^2 - 1] \leq -2\epsilon \sum_{i,j \in \mathcal{I}_1, i \neq j} x_i x_j \\ &\leq -2\epsilon (|\mathcal{I}_1| - 1) \|x_{\mathcal{I}_1}\|_1 \leq -2 \|x_{\mathcal{I}_1}\|_1 \cdot \epsilon, \\ \sum_{i,j \in \mathcal{I}_2, i \neq j} x_i x_j (x_i x_j - 1) &= \sum_{i,j \in \mathcal{I}_2, i \neq j} -x_i \cdot [x_j - x_i x_j^2] \leq \sum_{i,j \in \mathcal{I}_2, i \neq j} x_i \cdot [x_j - (1-\epsilon) x_j^2] \\ &\leq \sum_{i \in \mathcal{I}_2} x_i \max(|\mathcal{I}_2| - 1, 0) \cdot \epsilon [1 - (1-\epsilon)\epsilon] \\ &= -\max(|\mathcal{I}_2| - 1, 0) \|x_{\mathcal{I}_2}\|_1 \cdot \epsilon + O(n\epsilon^2), \\ \sum_{i \in \mathcal{I}_1, j \in \mathcal{I}_3} -x_i x_j (x_i x_j - 1) &= \sum_{i \in \mathcal{I}_1, j \in \mathcal{I}_3} -\frac{1}{4}(2 x_i x_j - 1)^2 + \frac{1}{4} \\ &\leq |\mathcal{I}_1||\mathcal{I}_3| \left[-\frac{1}{4} [2(1+\epsilon)(1-\epsilon) - 1]^2 + \frac{1}{4} \right], \\ &= |\mathcal{I}_1||\mathcal{I}_3| \left(\epsilon^2 - \epsilon^4 \right) = O(n^2 \epsilon^2), \\ \sum_{i \in \mathcal{I}_2, j \in \mathcal{I}_3} x_i x_j (x_i x_j - 1) &= \sum_{i \in \mathcal{I}_2, j \in \mathcal{I}_3} \frac{1}{4}(2 x_i x_j - 1)^2 - \frac{1}{4} \\ &\leq |\mathcal{I}_2||\mathcal{I}_3| \left[\frac{1}{4} [2(1+\epsilon)(1-\epsilon) - 1]^2 - \frac{1}{4} \right] \leq 0. \end{aligned} \tag{3.73}$$

Choosing ϵ to be small enough and substituting the above four estimates into (3.72), we obtain that

$$\begin{aligned} \frac{1}{4} q^T g(x) &\leq -2\epsilon \|x_{\mathcal{I}_1}\|_1 - \max(|\mathcal{I}_2| - 1, 0) \epsilon \|x_{\mathcal{I}_2}\|_1 + O(n^2 \epsilon^2) \\ &\leq -\left[\|x_{\mathcal{I}_1}\|_1 + \max(|\mathcal{I}_2| - 1, 0) \|x_{\mathcal{I}_2}\|_1 \right] \cdot \epsilon. \end{aligned}$$

If $|\mathcal{I}_2| \geq 2$, it follows from Hölder's inequality that

$$\|g(x)\|_\infty \geq \frac{4(\|x_{\mathcal{I}_1}\|_1 + \|x_{\mathcal{I}_2}\|_1)\cdot\epsilon}{\|q\|_1} = \frac{\|x_{\mathcal{I}_1}\|_1 + \|x_{\mathcal{I}_2}\|_1}{\|x_{\mathcal{I}_1}\|_1 + \|x_{\mathcal{I}_2}\|_1}\cdot 4\epsilon = 4\epsilon.$$

which is a contradiction to (3.70). Otherwise if $|\mathcal{I}_2| \leq 1$, it also follows from Hölder's inequality that

$$\|g(x)\|_\infty \geq \frac{4\|x_{\mathcal{I}_1}\|_1\epsilon}{\|q\|_1} = \frac{4\|x_{\mathcal{I}_1}\|_1}{\|x_{\mathcal{I}_1}\|_1 + \|x_{\mathcal{I}_2}\|_1}\cdot\epsilon \geq \frac{\|x_{\mathcal{I}_1}\|_1}{\|x_{\mathcal{I}_1}\|_1 + 1}\cdot 4\epsilon \geq 2\epsilon.$$

In summary, in this sub-case, we have

$$\|g(x)\|_\infty \geq 2\epsilon.$$

Case I-2. Now, we consider the case when $|\mathcal{I}_1| = 1$ and $|\mathcal{I}_2| \geq 2$. Assume without loss of generality that $\mathcal{I}_1 = \{1\}$. A similar calculation as (3.73) leads to

$$\frac{1}{4}q^T g(x) \leq -\max(|\mathcal{I}_2| - 1, 0)\epsilon\|x_{\mathcal{I}_2}\|_1 + O(n^2\epsilon^2) \leq -\frac{1}{2}\|x_{\mathcal{I}_2}\|_1\cdot\epsilon.$$

If $x_1 \leq 2\epsilon^{-1}$, Hölder's inequality gives

$$\|g(x)\|_\infty \geq \frac{4\epsilon\|x_{\mathcal{I}_2}\|_1}{2\|q\|_1} = 2\epsilon\cdot\frac{\|x_{\mathcal{I}_2}\|_1}{\|x_{\mathcal{I}_1}\|_1 + \|x_{\mathcal{I}_2}\|_1} \geq 2\epsilon\cdot\frac{2\epsilon}{2\epsilon^{-1} + 2\epsilon} \geq 2\epsilon\cdot\frac{\epsilon^2}{2} = \epsilon^3.$$

Now, we assume that $x_1 > 2\epsilon^{-1}$. The first component of the gradient is

$$\begin{aligned}\frac{1}{4}g_1(x) &= \sum\nolimits_{j\in[n],j\neq 1}(x_j^2 x_i - x_j) \geq \sum\nolimits_{j\in[n],j\neq 1}(\epsilon^2 x_i - \epsilon)\\ &= (n-1)\epsilon^2\cdot x_1 - (n-1)\epsilon > (n-1)\epsilon > \epsilon,\end{aligned}$$

which contradicts (3.70). In summary, in this sub-case, we have

$$\|g(x)\|_\infty \geq \epsilon^3.$$

Case I-3. In this case, we assume $|\mathcal{I}_1| = 1$ and $|\mathcal{I}_2| \leq 1$. In addition, we assume $\mathcal{I}_1 = \{1\}$. If $x_1 \geq (1-\epsilon)^{-1} + \epsilon$, the third estimate in (3.73) becomes

$$\begin{aligned}\sum\nolimits_{j\in\mathcal{I}_3} - x_1 x_j(x_1 x_j - 1) &\leq \sum\nolimits_{j\in\mathcal{I}_3} - x_1(1-\epsilon)[x_1(1-\epsilon) - 1]\\ &\leq -(1-\epsilon)^2\epsilon x_1 \leq -\frac{1}{2}\|x_{\mathcal{I}_1}\|_1\cdot\epsilon.\end{aligned}$$

Then, using a similar analysis and by applying Hölder's inequality, it follows that

$$\frac{1}{4}q^T g(x) \leq -\frac{1}{2}\|x_{\mathcal{I}_1}\|_1\epsilon \quad \text{and} \quad \|g(x)\|_\infty \geq 2\epsilon\cdot\frac{\|x_{\mathcal{I}_1}\|_1}{\|x_{\mathcal{I}_1}\|_1 + \|x_{\mathcal{I}_2}\|_1} > \epsilon.$$

Otherwise, if $x_1 < (1-\epsilon)^{-1} + \epsilon$,

$$|x_1 - 1| < \frac{\epsilon}{1-\epsilon} + \epsilon < 3\epsilon.$$

Hence,

$$\|x - x^0\|_1 \leq 3\epsilon + (n-1)\epsilon = (n+2)\epsilon.$$

In summary, in this sub-case, we have

$$\|g(x)\|_\infty < \epsilon/4 \quad \text{or} \quad \|x - x^0\|_1 \leq (n+2)\epsilon.$$

Case I-4. Finally, we assume $|\mathcal{I}_1| = 0$. If $|\mathcal{I}_2| \geq 2$, we can use a similar analysis as *Case I-2* to conclude that

$$\frac{1}{4} q^T g(x) \leq -\frac{1}{2} \|x_{\mathcal{I}_2}\|_1 \cdot \epsilon + O(n\epsilon^2)$$

and thus

$$\|g(x)\|_\infty \geq \epsilon.$$

Next, we consider the case when $|\mathcal{I}_2| = 1$ and we assume $\mathcal{I}_2 = \{1\}$. The fourth term in (3.73) can be estimated as

$$\sum\nolimits_{i \in \mathcal{I}_2, j \in \mathcal{I}_3} x_i x_j (x_i x_j - 1) = \left(\sum\nolimits_{j=2}^n x_j^2\right) x_1^2 - \left(\sum\nolimits_{j=2}^n x_j\right) x_1.$$

Since $x_j \in [1-\epsilon, 1+\epsilon]$ for all $j \in \{2, \ldots, n\}$, it holds that

$$\frac{\sum_{j=2}^n x_j}{\sum_{j=2}^n x_j^2} \geq \frac{1}{1+\epsilon} > 1 - \epsilon.$$

Therefore,

$$\begin{aligned}
\sum\nolimits_{i \in \mathcal{I}_2, j \in \mathcal{I}_3} x_i x_j (x_i x_j - 1) &= \left(\sum\nolimits_{j=2}^n x_j^2\right) x_1^2 - \left(\sum\nolimits_{j=2}^n x_j\right) x_1 \\
&\leq \left(\sum\nolimits_{j=2}^n x_j^2\right) (1-\epsilon)^2 - \left(\sum\nolimits_{j=2}^n x_j\right) (1-\epsilon) \\
&= \sum\nolimits_{j=2}^n \left[(1-\epsilon)^2 x_j^2 - (1-\epsilon) x_j\right] \\
&\leq \sum\nolimits_{j=2}^n \left[(1-\epsilon)^2 (1+\epsilon)^2 - (1-\epsilon)(1+\epsilon)\right] \\
&\leq -(n-1)\epsilon^2 + O(n\epsilon^3).
\end{aligned}$$

Thus, it holds that

$$\frac{1}{4} q^T g(x) \leq -(n-1)\epsilon^2 + O(n\epsilon^3) \geq -\epsilon^2.$$

Hölder's inequality implies that

$$\|g(x)\|_\infty \geq \frac{4\epsilon^2}{\|q\|_1} = \frac{4\epsilon^2}{x_1} \geq \frac{4\epsilon^2}{1-\epsilon} \geq 4\epsilon^2.$$

The only remaining case is when $|\mathcal{I}_2| = 0$. In this case, we have

$$x_i \in [1-\epsilon, 1+\epsilon], \quad \forall i \in [n].$$

Therefore, it holds that

$$\|x - x^0\|_1 \leq n\epsilon.$$

In summary, in this sub-case, we have

$$\|g(x)\|_\infty \geq 4\epsilon^2 \quad \text{or} \quad \|x - x^0\|_1 \leq n\epsilon.$$

Combining *Cases I-1* to *I-4* yields that

$$\|g(x)\|_\infty \geq \epsilon^3 \quad \text{or} \quad \|x - x^0\|_1 \leq (n+4)\epsilon$$

in *Case I*.

Case II. For the case when $\sum_{i\in[n]} x_i < -2\epsilon$, one can obtain the same conclusions as *Case I* by the symmetry of the landscape.

Case III. We finally consider the case when $\sum_{i\in[n]} x_i \in [-2\epsilon, 2\epsilon]$. Considering the assumption (3.70), we have

$$\frac{1}{4}g_i(x) = \left(\sum\nolimits_{j\in[n],j\neq i} x_j^2 + 1\right) x_i - \sum\nolimits_{j\in[n]} x_j \in [-\epsilon, \epsilon], \quad \forall i \in [n].$$

Combined with the assumption that $\sum_{i\in[n]} x_i \in [-2\epsilon, 2\epsilon]$, it follows that

$$\left(\sum\nolimits_{j\in[n],j\neq i} x_j^2 + 1\right) x_i \in [-3\epsilon, 3\epsilon].$$

Furthermore, since $\sum_{j\in[n],j\neq i} x_j^2 + 1 \geq 1$, we have

$$x_i \in [-3\epsilon, 3\epsilon], \quad \forall i \in [n].$$

We consider the descent direction $p \in \mathbb{R}^n$, where

$$p_i = 1/\sqrt{n}, \quad \forall i \in [n].$$

Then, we can calculate that

$$\begin{aligned}\frac{1}{4}p^T H(x)p &= \sum\nolimits_{i,j\in[n]j\neq i} \left[x_j^2 p_i^2 + (2x_i x_j - 1)p_i p_j\right] = \frac{1}{n}\sum\nolimits_{i,j\in[n]j\neq i} \left[x_j^2 + (2x_i x_j - 1)\right] \\ &= \frac{1}{n}\left[(n-1)\sum\nolimits_{i\in[n]} x_i^2 + 2\sum\nolimits_{i,j\in[n],i\neq j} x_i x_j - n(n-1)\right]\end{aligned}$$

$$\leq \frac{1}{n}\left[(n-1)\cdot 9n\epsilon^2 + 2n(n-1)\cdot 9\epsilon^2 - n(n-1)\right]$$
$$= 27(n-1)\epsilon^2 - (n-1) \leq -n/2,$$

where the last inequality is because ϵ is sufficiently small.

Combined *Cases I-III*, we have proved that under assumption (3.70), it holds that

$$\min\{\|x - x^0\|_1, \|x + x^0\|_1\} \leq (n+4)\epsilon \quad \text{or} \quad \|g(x)\|_\infty \geq \epsilon^3 \quad \text{or} \quad \lambda_{min}[H(x)] \leq -2n.$$

Letting $\epsilon := \eta/(n+4) \ll 1$, we know that the property stated in the theorem holds for problem (3.69) with

$$\beta(\eta) = \frac{\eta^3}{(n+4)^3}, \quad \gamma(\eta) = 2n.$$

In addition, we have $\eta_0 = O(1)$, $\beta(\eta) = O(n^{-3}\eta^3)$ and $\gamma(\eta) = O(n)$. Transforming back to the instance (C^0, u^0), the property stated in the theorem holds with

$$\eta_0 = O(n^{-0.5}), \quad \beta(\eta) = O(n^{-6.5}\eta^3), \quad \gamma(\eta) = O(n^{-2}).$$

This completes the proof. □

Proof of Lemma 10

Proof. Similar to Lemma 9, it is equivalent to prove the results for the scaled instance $\mathcal{MC}(n(n-1)\tilde{C}, n\tilde{u}^*)$. With a little abuse of notations, we use $(\tilde{C}\tilde{u}^*)$ to denote the scaled pair of parameters. Denote

$$\delta := \max\left\{\frac{n(n-1)\epsilon}{\alpha^*}, \frac{n\epsilon}{1-\alpha^*}\right\}.$$

Then, the condition stated in the lemma implies that

$$\tilde{C}_{ij} \in [1-\delta, 1+\delta], \quad \forall i,j \in [n] \quad \text{s.t. } i \neq j; \quad \tilde{C}_{ii} \in [0, \delta], \quad \tilde{u}_i^* \in [1-\delta, 1+\delta], \quad \forall i \in [n].$$

Let $R > 0$ be a large enough constant. Suppose that $u \in \mathbb{R}^n$ is a stationary point of the instance $(\tilde{C}, \tilde{u}^*)$ such that $\|u\|_2 = R$. Denote the gradient and the Hessian matrix of the instance $\mathcal{MC}(\tilde{C}, \tilde{u}^*)$ at u as $g(u) \in \mathbb{R}^n$ and $H(u) \in \mathbb{R}^{n\times n}$, respectively. Then, it holds that

$$\frac{1}{4}g_i(u) = \sum\nolimits_{j\in[n]} \tilde{C}_{ij}u_j(u_iu_j - \tilde{u}_i^*\tilde{u}_j^*) = 0, \quad \forall i \in [n]. \tag{3.74}$$

We assume without loss of generality that

$$u_1 = \max_{i\in[n]}|u_i| \geq R/\sqrt{n} > 0.$$

If $u_i = 0$ for all $i \in [n]\backslash\{1\}$, we have

$$\frac{1}{4}g_2(u) = \left(\tilde{C}_{21}u_1^2 + \sum\nolimits_{j\geq 2}\tilde{C}_{2j}u_j^2\right)u_2 - \left(\tilde{C}_{21}\tilde{u}_1^*\tilde{u}_2^*u_1 + \sum\nolimits_{j\geq 2}\tilde{C}_{2j}\tilde{u}_j^*\tilde{u}_2^*u_j\right)$$

$$= -\tilde{C}_{21}\tilde{u}_1^*\tilde{u}_2^*u_1 \leq -(1-\delta)\cdot(1-\delta)^2\cdot R < 0,$$

where the last inequality is in light of $\tilde{C}_{21} > 1-\delta$ and $\tilde{u}_i^* > 1-\delta$. This contradicts the stationarity of point x and thus

$$\sum\nolimits_{j\geq 2} u_j^2 > 0.$$

Moreover, since $\tilde{C}_{1j} > 1-\delta$ for all $j \in [n]\backslash\{1\}$, we have

$$\sum\nolimits_{j\in[n]} \tilde{C}_{1j}u_j^2 \geq \sum\nolimits_{j\geq 2} \tilde{C}_{1j}u_j^2 > (1-\delta)\sum\nolimits_{j\geq 2} u_j^2 > 0.$$

Similarly, for all $i \in [n]\backslash\{1\}$, it holds that

$$\sum\nolimits_{j\in[n]} \tilde{C}_{ij}u_j^2 \geq \sum\nolimits_{j\in[n],j\neq i} \tilde{C}_{ij}u_j^2 > (1-\delta)\sum\nolimits_{j\in[n],j\neq i} u_j^2 > (1-\delta)u_i^2 > 0.$$

Solving (3.74) for all $i \in [n]$, we conclude that

$$u_i = \frac{\sum_{j\in[n]} \tilde{C}_{ij}\tilde{u}_i^*\tilde{u}_j^*u_j}{\sum_{j\in[n]} \tilde{C}_{ij}u_j^2}. \tag{3.75}$$

Assuming that

$$u_1 < R - \frac{2n}{(1-\delta)R},$$

it follows that

$$\sum\nolimits_{j\in[n]} \tilde{C}_{1j}u_j^2 > (1-\delta)\sum\nolimits_{j\geq 2} u_j^2 \geq 4n - \frac{4n^2}{(1-\delta)R^2}. \tag{3.76}$$

In addition, we can calculate that

$$\begin{aligned}\sum\nolimits_{j\in[n]} \tilde{C}_{1j}\tilde{u}_1^*\tilde{u}_j^*u_j &\leq \sum\nolimits_{j\in[n]} \tilde{C}_{1j}\tilde{u}_1^*\tilde{u}_j^*|u_j| \\ &< (1+\delta)\cdot(1+\delta)^2\sum\nolimits_{j\in[n]}|u_j| \leq 2\|u\|_1 \leq 2\sqrt{n}R,\end{aligned} \tag{3.77}$$

where the second last inequality is because δ is a sufficiently small constant. Combining inequalities (3.76)-(3.77), we have

$$u_1 = \frac{\sum_{j\in[n]} \tilde{C}_{1j}\tilde{u}_1^*\tilde{u}_j^*u_j}{\sum_{j\in[n]} \tilde{C}_{1j}u_j^2} < \frac{2\sqrt{n}R}{4n - 4n^2/[(1-\delta)^2R^2]}.$$

Choosing $R \geq 4n \geq 2(1-\delta)^{-1}n$, the above inequality leads to

$$u_1 < \frac{2\sqrt{n}R}{4n - 4n^2/[(1-\delta)^2R^2]} < \frac{2\sqrt{n}R}{2n} = \frac{R}{\sqrt{n}},$$

which contradicts the assumption that $u_1 \geq R/\sqrt{n}$. Therefore,

$$u_1 \geq R - \frac{2n}{(1-\delta)R}.$$

Using the condition that $\|x\|_2 = R$, it holds that

$$\sum\nolimits_{j\geq 2} u_j^2 \leq \frac{2n}{1-\delta} - \frac{4n^2}{(1-\delta)^2 R^2} < \frac{2n}{1-\delta}.$$

For all $i \in [n]\backslash\{1\}$, the relation (3.75) implies that

$$\begin{aligned}
u_i = \frac{\sum_{j\in[n]} \tilde{C}_{ij}\tilde{u}_i^*\tilde{u}_j^* u_j}{\sum_{j\in[n]} \tilde{C}_{ij}u_j^2} &= \frac{\tilde{C}_{1i}\tilde{u}_i^*\tilde{u}_1^* u_1 + \sum_{j\geq 2} \tilde{C}_{ij}\tilde{u}_i^*\tilde{u}_j^* u_j}{\sum_{j\in[n]} \tilde{C}_{ij}u_j^2} \\
&\geq \frac{(1-\delta)\cdot(1-\delta)^2(R - 2n/[(1-\delta)R]) - (1+\delta)\cdot(1+\delta)^2\sqrt{n\cdot\sum_{j\geq 2} u_j^2}}{\sum_{j\in[n]} \tilde{C}_{ij}u_j^2} \\
&\geq \frac{(1-\delta)\cdot(1-\delta)^2(R - 2n/[(1-\delta)R]) - (1+\delta)\cdot(1+\delta)^2\sqrt{n\cdot 2n/(1-\delta)}}{\sum_{j\in[n]} \tilde{C}_{ij}u_j^2} \\
&\geq \frac{1/2\cdot(R-1) - 2n\sqrt{2/(1-\delta)}}{\sum_{j\in[n]} \tilde{C}_{ij}u_j^2} \geq \frac{R/2 - 1/2 - 4n}{\sum_{j\in[n]} \tilde{C}_{ij}u_j^2} > 0,
\end{aligned}$$

where the last inequality is due to choosing $R > 8n+1$ and the second last inequality results from the fact that δ is sufficiently small. Using the same relation, it follows that

$$\begin{aligned}
u_i = \frac{\sum_{j\in[n]} \tilde{C}_{ij}\tilde{u}_i^*\tilde{u}_j^* u_j}{\sum_{j\in[n]} \tilde{C}_{ij}u_j^2} &\geq \frac{\tilde{C}_{1i}\tilde{u}_i^*\tilde{u}_1^* u_1}{\sum_{j\in[n]} \tilde{C}_{ij}u_j^2} \geq \frac{(1-\delta)(1-\delta)^2 u_1}{(1+\delta)\cdot R^2} \\
&\geq \frac{1}{4R^2}\cdot\left(R - \frac{2n}{(1-\delta)R}\right) \geq \frac{1}{8R},
\end{aligned}$$

where the last inequality is due to choosing $R \geq 8n \geq 4(1-\delta)^{-1}n$. Furthermore, using the relation (3.75) with $i = 1$, we have

$$\begin{aligned}
\sum\nolimits_{j\geq 2} u_j &\geq \frac{1}{(1+\delta)(1+\delta)^2}\sum\nolimits_{j\geq 2} \tilde{C}_{1j}\tilde{u}_i^*\tilde{u}_j^* u_j \\
&= \frac{1}{(1+\delta)(1+\delta)^2}\cdot u_1\left[\sum\nolimits_{j\geq 2} \tilde{C}_{1j}u_j^2 + \tilde{C}_{11}[u_1^2 - (\tilde{u}_1^*)^2]\right] \\
&\geq \frac{1}{(1+\delta)(1+\delta)^2}\cdot u_1\left[(1-\delta)\sum\nolimits_{j\geq 2} u_j^2 + \tilde{C}_{11}[(R - 2n/[(1-\delta)R])^2 - (1+\delta)^2]\right] \\
&\geq \frac{1-\delta}{(1+\delta)(1+\delta)^2}\cdot u_1\left(\sum\nolimits_{j\geq 2} u_j^2\right)
\end{aligned}$$

$$\geq \frac{1-\delta}{(1+\delta)(1+\delta)^2}\left(R-\frac{2n}{(1-\delta)R}\right)\cdot\sum\nolimits_{j\geq 2}u_j^2 \geq \frac{1}{4}\left(R-\frac{2n}{(1-\delta)R}\right)\cdot\sum\nolimits_{j\geq 2}u_j^2.$$

Since $\sum_{j\geq 2} u_j \leq \sqrt{n(\sum_{j\geq 2} u_j^2)}$, it follows that

$$\sqrt{n\left(\sum\nolimits_{j\geq 2}u_j^2\right)} \geq \frac{1}{4}\left(R-\frac{2n}{(1-\delta)R}\right)\cdot\sum\nolimits_{j\geq 2}u_j^2,$$

which further implies that

$$\sum\nolimits_{j\geq 2}u_j^2 \leq \frac{16n}{(R-2n/[(1-\delta)R])^2} \leq \frac{16n}{(R-1)^2} \leq \frac{1}{4},$$

where the last inequality is because of choosing $R \geq 1+8\sqrt{n}$. Now, we consider the descent direction $q \in \mathbb{R}^n$, where

$$q_1 = -u_1; \quad q_i = u_i, \quad \forall i \in [n]\backslash\{1\}.$$

Similar to the proof of Lemma 9, we can calculate that

$$\begin{aligned}
\frac{1}{4}\langle g(u), q\rangle &= \sum\nolimits_{i,j\geq 2, i\neq j}\tilde{C}_{ij}u_iu_j\left(u_iu_j-\tilde{u}_i^*\tilde{u}_j^*\right)-\tilde{C}_{11}u_1^2[u_1^2-(\tilde{u}_1^*)^2]+\sum\nolimits_{i\geq 2}\tilde{C}_{ii}u_i^2[u_i^2-(\tilde{u}_i^*)^2]\\
&\leq \sum\nolimits_{i,j\geq 2, i\neq j}\tilde{C}_{ij}u_iu_j\left(u_iu_j-\tilde{u}_i^*\tilde{u}_j^*\right)+\sum\nolimits_{i\geq 2}\tilde{C}_{ii}u_i^2[u_i^2-(\tilde{u}_i^*)^2]\\
&\leq \sum\nolimits_{i,j\geq 2, i\neq j}\tilde{C}_{ij}u_iu_j\left[1/4-(1-\delta)^2\right]+\sum\nolimits_{i\geq 2}\tilde{C}_{ii}u_i^2[1/4-(1-\delta)^2]\\
&\leq \sum\nolimits_{i,j\geq 2, i\neq j}(1-\delta)\cdot(8R)^{-2}\cdot(1/4-1/2)+\sum\nolimits_{i\geq 2}\delta\cdot(8R)^{-2}\cdot(1/4-1/2)<0,
\end{aligned}$$

which contradicts the assumption that x is a stationary point. Therefore, the above analysis implies that the instance $(\tilde{C}, \tilde{u}^*)$ has no stationary point in the region $\{u \in \mathbb{R}^n \mid \|u\|_2 > 8n+1\}$.

Now, We focus on the compact region $\{u \in \mathbb{R}^n \mid \|u\|_2 \leq 8n+1\}$. Since the gradient and the Hessian matrix are continuous functions of (C, u^*), the ℓ_∞-norm of the gradient and the eigenvalues of the Hessian matrix are also continuous functions of (C, u^*). Intuitively, a small perturbation to (C, u^*) would not significantly change the norms of the gradient and the Hessian matrix. Thus, the strict-saddle property still holds after a small perturbation. More rigorously, let $(C^0, u^0) \in \mathcal{M}$ and $\eta \in (0, \eta_0]$. In the region

$$\mathcal{R}_\eta := \{u \in \mathbb{R}^n \mid \|u\|_2 \leq 8n+1, \|u-u^0\|_1 \geq \eta, \|u+u^0\|_1 \geq \eta\},$$

at least one of the following properties holds:

$$\|\nabla g(u; C^0, u^0)\|_\infty \geq \beta(\eta), \quad \lambda_{min}[\nabla^2 g(u; C^0, u^0)] \leq -\gamma(\eta).$$

Since $\mathcal{R}_\eta$ is a compact set and we constrain (C, u^*) by $\|C\|_1 = 1$ and $\|u^*\|_1 = 1$, the functions

$$\|\nabla g(u; C, u^*)\|_\infty \quad \text{and} \quad \lambda_{min}[\nabla^2 g(x; C, u^*)]$$

are Lipschitz continuous in (C, u^*). Suppose that the Lipschitz constants are L_g and L_H under the weighted ℓ_1-norm, namely

$$\left|\|\nabla g(u; C, u^*)\|_\infty - \|\nabla g(u; \tilde{C}, \tilde{u}^*)\|_\infty\right| \le L_g \left[\alpha^* \|\tilde{C} - C\|_1 + (1-\alpha^*)\|\tilde{u}^* - u^*\|_1\right],$$
$$\left|\lambda_{min}[\nabla^2 g(u; C, u^*)] - \lambda_{min}[\nabla^2 g(u; \tilde{C}, \tilde{u}^*)]\right| \le L_H \left[\alpha^* \|\tilde{C} - C\|_1 + (1-\alpha^*)\|\tilde{u}^* - u^*\|_1\right],$$
$$\forall x \in \mathcal{R}_\eta,\ (C, u^*) \quad \text{s.t. } \|C\|_1 = \|u^*\|_1 = 1.$$

Let

$$\epsilon := \min\left\{\frac{\beta(\eta)}{2L_g}, \frac{\gamma(\eta)}{2L_H}\right\}.$$

Then, for every pair $(\tilde{C}, \tilde{u}^*)$ satisfying

$$\alpha^* \|\tilde{C} - C^0\|_1 + (1-\alpha^*)\|\tilde{u}^* - u^0\|_1 < \epsilon,$$

at least one of the following properties holds for all $x \in \mathcal{R}_\eta$:

$$\|\nabla g(u; \tilde{C}, \tilde{u}^*)\|_\infty \ge \beta(\eta)/2, \quad \lambda_{min}[\nabla^2 (u; \tilde{C}, \tilde{u}^*)] \le -\gamma(\eta)/2.$$

This implies that the strict-saddle property holds for the the perturbed instance $\mathcal{MC}(\tilde{C}, \tilde{u}^*)$. Letting $\eta \to 0$, it follows that $\pm\tilde{u}^*$ are the only points satisfying the second-order necessary optimality conditions, and thus $\mathcal{MC}(\tilde{C}, \tilde{u}^*)$ does not have SSCPs. □

Proof of Theorem 27

The proof of Theorem 27 directly follows from the next two lemmas.

Lemma 20. *Suppose that $(C, u^*) \in \mathcal{SD}$ and that u^0 is a global solution to $\mathcal{MC}(C, u^*)$. Then, for all $k \in [n_1]$, it holds that $u_i^0 u_j^0 = u_i^* u_j^*$ for all $i, j \in \mathcal{I}_{1k}$. In addition, $u_i^0 = 0$ for all $i \in \mathcal{I}_0(C, u^*)$.*

Proof. Denote $M^* := u^*(u^*)^T$. We first consider nodes in $\mathcal{G}_{1k}$ for some $k \in [n_1]$. Since the subgraph is not bipartite, there exists a cycle with an odd length $2\ell+1$, which we denote as

$$\{i_1, \ldots, i_{2\ell+1}\}.$$

Then, we have

$$(u_{i_1}^0)^2 = \prod_{s=1}^{2\ell+1} (u_{i_s}^0 u_{i_{s+1}}^0)^{(-1)^{s-1}} = \prod_{s=1}^{2\ell+1} (M_{i_s i_{s+1}}^*)^{(-1)^{s-1}} = \prod_{s=1}^{2\ell+1} (u_{i_s}^* u_{i_{s+1}}^*)^{(-1)^{s-1}} = (u_{i_1}^*)^2,$$

which implies that the conclusion holds for $i = j = i_1$. Using the connectivity of $\mathbb{G}_{1k}(C, u^*)$, we know

$$u_i^0 u_j^0 = u_i^* u_k^*, \quad \forall i, j \in \mathcal{I}_{1k}(C, u^*).$$

Then, we consider nodes in $\mathcal{I}_0(C, u^*)$. Since $\mathcal{I}_{00}(C, u^*)$ is empty, for every node $i \in \mathcal{I}_0(C, u^*)$, there exists another node $j \in \mathcal{I}_1(C, u^*)$ such that $C_{ij} > 0$. Hence, we have

$$u_i^0 = M_{ij}^*/u_j^0 = 0.$$

This completes the proof. □

The following lemma provides a necessary and sufficient condition for instances with a positive definite Hessian matrix at global solutions, which is stronger than what Theorem 27 requires.

Lemma 21. *Suppose that $u^0 \in \mathbb{R}^n$ is a global minimizer of the instance $\mathcal{MC}(C, u^*)$ such that the conditions in Lemma 20 hold. Then, the Hessian matrix is positive definite at u^0 if and only if*

1. *$\mathbb{G}_{1i}(C, u^*)$ is not bipartite for all $i \in [n_1]$;*
2. *$\mathcal{I}_{00}(C, u^*) = \emptyset$.*

Proof. We first construct counterexamples for the necessity part and then prove the positive definiteness of the Hessian matrix for the sufficiency part.

Necessity. We construct counterexamples by discussing two different cases.

Case I. We first consider the case when there exists $k \in [n_1]$ such that $\mathbb{G}_{1k}(C, u^*)$ is bipartite. Suppose that $\mathbb{G}_{1i}(C, u^*) = \mathbb{G}_{1k1} \cup \mathbb{G}_{1k2}$ is a partition of $\mathbb{G}_{1k}(C, u^*)$. Let the sets $\mathcal{I}_{1k}, \mathcal{I}_{1k1}$ and $\mathcal{I}_{1k2}$ be the node sets of the corresponding graphs. Define $q \in \mathbb{R}^n$ as

$$q_i := u_i^0, \quad \forall i \in \mathcal{I}_{1k1}; \quad q_i := -u_i^0, \quad \forall i \in \mathcal{I}_{1k2}; \quad q_i := 0, \quad \forall i \notin \mathcal{I}_{1k}.$$

Then, the curvature of the Hessian along the direction q is

$$\begin{aligned}
&\frac{1}{4}[\nabla^2 g(u^0; C, u^*)](q, q) \\
&= \sum_{i \in \mathcal{I}_{1k1}, j \in \mathcal{I}_{1k2}} C_{ij} \left[(u_i^0)^2 q_j^2 + (u_j^0)^2 q_i^2\right] + 2 \sum_{i \in \mathcal{I}_{1k1}, j \in \mathcal{I}_{1k2}} C_{ij}(2u_i^0 u_j^0 - u_i^* u_j^*) q_i q_j \\
&= \sum_{i \in \mathcal{I}_{1k1}, j \in \mathcal{I}_{1k2}} C_{ij} \left[(u_i^0)^2 q_j^2 + (u_j^0)^2 q_i^2\right] + 2 \sum_{i \in \mathcal{I}_{1k1}, j \in \mathcal{I}_{1k2}} C_{ij} u_i^0 u_j^0 q_i q_j \\
&= \sum_{i \in \mathcal{I}_{1k1}, j \in \mathcal{I}_{1k2}} 2C_{ij}(u_i^0 u_j^0)^2 - 2 \sum_{i \in \mathcal{I}_{1k1}, j \in \mathcal{I}_{1k2}} C_{ij}(u_i^0 u_j^0)^2 = 0.
\end{aligned}$$

We note that there is no self-loop in $\mathcal{G}_{1k}(C, u^*)$ and, thus, the diagonal entries of the weight matrix are equal to 0. Therefore, the Hessian matrix has a zero curvature along q and is not positive definite.

Case II. We consider the case when $\mathcal{I}_{00}(C, u^*) \neq \emptyset$. Suppose that $k \in \mathcal{I}_{00}(C, u^*)$. Define the vector $q \in \mathbb{R}^n$ as

$$q_k := 1; \quad q_i := 0, \quad \forall i \neq k.$$

The curvature of the Hessian along the direction q is

$$\begin{aligned}\frac{1}{4}[\nabla^2 g(u^0; C, u^*)](q, q) &= C_{kk}\left[2(u_k^0)^2 - (u_k^*)^2\right] q_k^2 + \sum_{j \in \mathcal{I}_0(C,u^*), j \neq k} C_{kj}(u_j^0)^2 q_k^2 \\ &= C_{kk}(u_k^0)^2 + \sum_{j \in \mathcal{I}_0(C,u^*), j \neq k} C_{kj}(u_j^0)^2 = 0.\end{aligned}$$

Therefore, the Hessian matrix is not positive-definite at u^0.

Sufficiency. Next, we consider the sufficiency part, namely, we prove that the Hessian matrix is positive definite under the two conditions stated in the theorem. Suppose that there exists a nonzero vector $q \in \mathbb{R}^n$ such that

$$[\nabla^2 g(u^0; C, u^*)](q, q) = 0.$$

Then, after straightforward calculations, we arrive at

$$\begin{aligned} u_i^0 q_j + u_j^0 q_i &= 0, \quad \forall i, j \quad \text{s. t. } C_{ij} > 0,\ i \neq j; \\ [2(u_i^0)^2 - (u_i^*)^2] q_i^2 = (u_i^0 q_i)^2 &= 0, \quad \forall i \quad \text{s. t. } C_{ii} > 0. \end{aligned}$$

The two conditions can be written compactly as

$$u_i^0 q_j + u_j^0 q_i = 0, \quad \forall i, j \quad \text{s. t. } C_{ij} > 0. \tag{3.78}$$

Consider the index set $\mathcal{I}_{1k}(C, u^*)$ for some $k \in [n_1]$. The equality (3.78) implies that

$$q_i/u_i^0 + q_j/u_j^0 = 0, \quad \forall i, j \in \mathcal{I}_{1k}(C, u^*). \tag{3.79}$$

Since the graph $\mathbb{G}_{1k}(C, u^*)$ is not bipartite, there exists a cycle with an odd length $2\ell + 1$, which we denote as

$$\{i_1, i_2, \ldots, i_{2\ell+1}\}.$$

Denoting $i_{2\ell+2} := i_1$, we can calculate that

$$2\frac{q_{i_1}}{u_{i_1}^0} = \sum_{s=1}^{2\ell+1} (-1)^{s-1} \left(\frac{q_{i_s}}{u_{i_s}^0} + \frac{q_{i_{s+1}}}{u_{i_{s+1}}^0} \right) = 0,$$

which leads to $q_{i_1} = 0$. Using the connectivity of $\mathcal{G}_{1k}$ and the relation (3.79), it follows that

$$q_i = 0, \quad \forall i \in \mathcal{I}_{1k}(C, u^*).$$

Moreover, the same conclusion holds for all $k \in [n_1]$ and, thus, we conclude that

$$q_i = 0, \quad \forall i \in \mathcal{I}_1(C, u^*).$$

Since $\mathcal{I}_{00}(C, u^*) = \emptyset$, for every node $i \in \mathcal{I}_0(C, u^*)$, there exists another node $j \in \mathcal{I}_1(C, u^*)$ such that $C_{ij} > 0$. Considering the relation (3.79), we obtain that

$$q_j = -u_j^0 q_i / u_i^0 = 0.$$

In summary, we have proved that $q_i = 0$ for all $i \in [n]$, which contradicts the assumption that $q \neq 0$. Hence, the Hessian matrix at u^0 is positive definite. □

Application of the implicit function theorem

Using the positive-definiteness of the Hessian matrix, we are able to apply the implicit function theorem to certify the existence of spurious local minima.

Lemma 22. *Suppose that $\alpha \in [0, 1]$ and consider a pair $(C, u^*) \in \mathcal{SD}$. Then, there exists a small constant $\delta(C, x^*) > 0$ such that for every instance $\mathcal{MC}(\tilde{C}, \tilde{u}^*)$ satisfying*

$$\alpha \|\tilde{C} - C\|_1 + (1 - \alpha)\|\tilde{u}^* - u^*\|_1 < \delta(C, u^*),$$

the instance $\mathcal{MC}(\tilde{C}, \tilde{u}^*)$ has spurious local minima.

Proof. By Theorem 27, there exists a global solution u^0 to the instance $\mathcal{MC}(C, u^*)$ such that

$$u^0(u^0)^T \neq u^*(u^*)^T, \quad \nabla^2 g(u^0; C, u^*) \succ 0.$$

Consider the system of equations:

$$\nabla g(u; C, u^*) = 0.$$

Since the Jacobi matrix of $\nabla g(u; C, u^*)$ with respect to u is the Hessian matrix $\nabla^2 g(u; C, u^*)$ and (u^0, C, u^*) is a solution, the implicit function theorem guarantees that there exists a small constant $\delta(C, u^*) > 0$ such that in the neighborhood

$$\mathcal{N} := \left\{ (\tilde{C}, \tilde{u}^*) \;\middle|\; \alpha \|\tilde{C} - C\|_1 + (1 - \alpha)\|\tilde{u}^* - u^*\|_1 < \delta(C, u^*) \right\},$$

there exists a function $u(\tilde{C}, \tilde{u}^*) : \mathcal{N} \mapsto \mathbb{R}^n$ such that

1. $u(C, u^*) = u^0$;
2. $u(\cdot, \cdot)$ is a continuous function in $\mathcal{N}$;
3. $\nabla g[u(\tilde{C}, \tilde{u}^*); \tilde{C}, \tilde{u}^*] = 0$.

Using the continuity of the Hessian matrix and $u(\cdot, \cdot)$, we can choose $\delta(C, u^*)$ to be small enough such that

$$u(\tilde{C}, \tilde{u}^*)[u(\tilde{C}, \tilde{u}^*)]^T \neq \tilde{u}^*(\tilde{u}^*)^T, \quad \nabla^2 g \left[u(\tilde{C}, \tilde{u}^*); \tilde{C}, \tilde{u}^* \right] \succ 0, \quad \forall (\tilde{C}, \tilde{u}^*) \in \mathcal{N}.$$

Therefore, the point $u(\tilde{C}, \tilde{u}^*)$ is a spurious local minimum of the instance $\mathcal{MC}(\tilde{C}, \tilde{u}^*)$. □

3.E Analysis for the Asymmetric Case

In this section, we extend the analysis of the *symmetric* weighted matrix completion problem (3.4) to the *asymmetric* weighted matrix completion problem, which is defined as

$$\min_{u\in\mathbb{R}^m, v\in\mathbb{R}^n} \sum\nolimits_{i\in[m], j\in[n]} C_{ij}(u_i v_j - M^*_{ij})^2, \tag{3.80}$$

where $M^* \in \mathbb{R}^{m\times n}$ is the ground truth matrix and $C \in \mathbb{R}^{m\times n}$ is the weight matrix. We note that in the asymmetric case, we do not assume that the weight matrix C is symmetric. Similar to the symmetric case, we assume that $M^* = u^*(v^*)^T$ has rank-1, where $u^* \in \mathbb{R}^m$ and $v^* \in \mathbb{R}^n$. We denote each instance of problem (3.80) as $\mathcal{MC}(C, u^*, v^*)$, where C is the weight matrix and $u^*(v^*)^T$ is the ground truth matrix. Moreover, since the degenerate instances where $C = 0$ or $M^* = 0$ can be easily analyzed separately, we utilize the "scale-free" property of problem (3.80) and extend the normalization assumption (Assumption 3) to the asymmetric case:

Assumption 4. Assume that $C \in \mathbb{S}^{nm-1}_{+,1}$, $u^* \in \mathbb{S}^{m-1}_1$ and $v^* \in \mathbb{S}^{n-1}_1$, i.e., $\|C\|_1 = \|u^*\|_1 = \|v^*\|_1 = 1$.

Define the objective function of problem (3.80) as

$$h(u, v; C, u^*, v^*) := \sum\nolimits_{i\in[m], j\in[n]} C_{ij}(u_i v_j - u^*_i v^*_j)^2.$$

Then, the set of degenerate instances is defined as

$$\begin{aligned}\mathcal{D}_{asym} := \{(C, u^*, v^*) \mid\ & C \in \mathbb{S}^{n^2-1}_{+,1}, u^* \in \mathbb{S}^{n-1}_1, v^* \in \mathbb{S}^{m-1}_1,\\ & \exists u \in \mathbb{R}^m,\ v \in \mathbb{R}^n \quad \text{s.t. } h(u, v; C, u^*, v^*) = 0,\ uv^T \neq u^*(v^*)^T\}.\end{aligned}$$

Using graphical notations, we can establish an exact characterization for the set $\mathcal{D}_{asym}$. The weighted graph $\mathbb{G}(C, u^*, v^*) = [\mathbb{V}(C, u^*, v^*), \mathbb{E}(C, u^*, v^*), \mathbb{W}(C, u^*, v^*)]$ is defined by

$$\begin{aligned}&\mathbb{V}(C, u^*, v^*) := [m + n],\\ &\mathbb{E}(C, u^*, v^*) := \{\{i, j + m\} \mid C_{ij} > 0, i \in [m], j \in [n]\},\\ &[\mathbb{W}(C, u^*, v^*)]_{i,j+m} := C_{ij}, \quad \forall i \in [m], j \in [n] \quad \text{s.t. } \{i, j + m\} \in \mathbb{E}(C, u^*, v^*).\end{aligned}$$

To include the information of u^* and v^*, we define

$$\begin{aligned}&\mathcal{I}^u_1(C, u^*, v^*) := \{i \in [m] \mid u^*_i \neq 0\}, \quad \mathcal{I}^u_0(C, u^*, v^*) := [m]\backslash\mathcal{I}^u_1(C, u^*, v^*),\\ &\mathcal{I}^v_1(C, u^*, v^*) := \{j + m \mid j \in [m], v^*_j \neq 0\},\\ &\mathcal{I}^v_0(C, u^*, v^*) := \{m + 1, \ldots, m + n\}\backslash\mathcal{I}^v_1(C, u^*, v^*),\\ &\mathcal{I}^u_{00}(C, u^*, v^*) := \{i \in \mathcal{I}^u_0(C, u^*, v^*) \mid \{i, j + m\} \notin \mathbb{E}(C, u^*, v^*),\ \forall j \in \mathcal{I}^v_1(C, u^*, v^*)\},\\ &\mathcal{I}^v_{00}(C, u^*, v^*) := \{j + m \in \mathcal{I}^v_0(C, u^*, v^*) \mid \{i, j + m\} \notin \mathbb{E}(C, u^*, v^*),\ \forall i \in \mathcal{I}^u_1(C, u^*, v^*)\}.\end{aligned}$$

The sub-graph $\mathbb{G}_1(C, u^*, v^*)$ is induced by $\mathcal{I}^u_1(C, u^*, v^*) \cup \mathcal{I}^v_1(C, u^*, v^*)$. The following theorem provides necessary and sufficient conditions for instances in $\mathcal{D}_{asym}$ and $\overline{\mathcal{D}}_{asym}$.

Theorem 33. *Given $C \in \mathbb{S}_{+,1}^{nm-1}$, $u^* \in \mathbb{S}_1^{m-1}$ and $v^* \in \mathbb{S}_1^{n-1}$, it holds that (C, u^*, v^*) does not belong to $\mathcal{D}_{asym}$ if and only if*

1. *$\mathbb{G}_1(C, u^*, v^*)$ is connected;*
2. *$\mathcal{I}_{00}^u(C, u^*, v^*) = \emptyset$ and $\mathcal{I}_{00}^v(C, u^*, v^*) = \emptyset$.*

Moreover, the following relation holds:

$$\begin{aligned}\overline{\mathcal{D}}_{asym} =& \{(C, u^*, v^*) \mid C \in \mathbb{S}_{+,1}^{nm-1}, u^* \in \mathbb{S}_1^{m-1}, v^* \in \mathbb{S}_1^{n-1}, \mathbb{G}_1(C, u^*, v^*) \text{ is disconnected}\} \\ &\cup \{(C, u^*, v^*) \mid C \in \mathbb{S}_{+,1}^{nm-1}, u^* \in \mathbb{S}_1^{m-1}, v^* \in \mathbb{S}_1^{n-1}, \mathcal{I}_{00}^u(C, u^*, v^*) \text{ is not empty}\} \\ &\cup \{(C, u^*, v^*) \mid C \in \mathbb{S}_{+,1}^{nm-1}, u^* \in \mathbb{S}_1^{m-1}, v^* \in \mathbb{S}_1^{n-1}, \mathcal{I}_{00}^v(C, u^*, v^*) \text{ is not empty}\}.\end{aligned}$$

The proof of Theorem 33 is based on a slight modification of the proof of Theorem 14 and therefore, we omit the proof here. Similarly, the proofs of all subsequent theorems in this section follow directly from those of the symmetric case and are omitted for brevity. The new complexity metric for the asymmetric problem is given by

$$\begin{aligned}\mathbb{D}_\alpha^{asym}(C, u^*, v^*) &:= \left[\inf_{(\tilde{C}, \tilde{u}^*, \tilde{v}^*) \in \mathcal{D}_{asym}} \alpha\|C - \tilde{C}\|_1 + (1-\alpha)(\|u^* - \tilde{u}^*\|_1 + \|v^* - \tilde{v}^*\|_1)\right]^{-1} \\ &= \left[\min_{(\tilde{C}, \tilde{u}^*, \tilde{v}^*) \in \overline{\mathcal{D}}_{asym}} \alpha\|C - \tilde{C}\|_1 + (1-\alpha)(\|u^* - \tilde{u}^*\|_1 + \|v^* - \tilde{v}^*\|_1)\right]^{-1}.\end{aligned}$$

Connection to Existing Results

Now, we derive upper bounds on the complexity metric under several different existing conditions. We first develop an upper bound on the complexity metric under the RIP condition, which is stated in the following theorem.

Theorem 34. *Suppose that $\delta \in [0, 1)$ is a constant and the instance $\mathcal{MC}(C, u^*, v^*)$ satisfies the δ-RIP$_{2,2}$ condition. Then, it holds that*

$$\mathbb{D}_\alpha^{asym}(C, u^*, v^*) \leq \frac{mn(1+\delta) - 2\delta}{2\alpha(1-\delta)}.$$

The maximum complexity is attained by the instance $\mathcal{MC}(C^\delta, u^\delta, v^\delta)$, where

$$\begin{aligned}&C_{11}^\delta = \frac{1-\delta}{(1+\delta)mn - 2\delta}; \quad C_{ij}^\delta = \frac{1+\delta}{(1+\delta)mn - 2\delta}, \ \forall (i,j) \in [m] \times [n] \backslash \{(1,1)\}; \\ &u_1^\delta = 1; \quad u_i^\delta = 0, \quad \forall i \geq 2, \quad v_1^\delta = 1; \quad v_j^\delta = 0, \quad \forall j \geq 2.\end{aligned}$$

We note that the upper bound in Theorem 34 is $O(\min\{m, n\})$ larger than the smallest possible complexity, which is $O(\max\{m, n\})$. Following the same path as in the symmetric case, we improve the upper bound using the incoherence information. We first give the definition of the incoherence in the asymmetric case.

Definition 9 ([118])**.** Given constants $\mu_1 \in [1, m]$ and $\mu_2 \in [1, n]$, the ground truth matrix $M^* \in \mathbb{R}^{m\times n}$ is said to be (μ_1, μ_2)**-incoherent** if

$$\|(e_i^m)^T U^*\|_F \le \sqrt{\mu_1 r/m}, \quad \forall i \in [m], \quad \|(e_j^n)^T V^*\|_F \le \sqrt{\mu_2 r/n}, \quad \forall j \in [n],$$

where $U^*\Sigma^*(V^*)^T$ is the truncated SVD of M^*, e_i^m is the i-th standard basis of $\mathbb{R}^m$ and e_j^n is the j-th standard basis of $\mathbb{R}^n$. Moreover, the ground truth matrix $M^* \in \mathbb{R}^{m\times n}$ is said to be μ**-incoherent** if it is (μ_1, μ_2)-incoherent with some $\mu_1, \mu_2 \le \mu$.

As a counterpart of Theorem 20, the upper bound can be improved to $O[\mu \max\{m, n\}]$.

Theorem 35. *Suppose that the instance $\mathcal{MC}(C, u^*, v^*)$ satisfies the δ-RIP$_{2,2}$ condition and $u^*(v^*)^T$ is (μ_1, μ_2)-incoherent. Then, it holds that*

$$\mathbb{D}_\alpha^{asym}(C, u^*, v^*) \le \max\left\{\frac{\max\{\rho_1, \rho_2\} mn(1+\delta)}{4\alpha(1-\delta)}, \frac{1}{2(1-\alpha)}\right\} \times \min\left\{(1 - \max\{\rho_1, \rho_2\})^{-1}, 3\right\},$$

where $\rho_1 := \mu_1/m$ and $\rho_2 := \mu_2/n$. Moreover, suppose that the instance $\mathcal{MC}(C, u^*, v^*)$ satisfies the δ-RIP$_{2,2}$ condition and $u^*(v^*)^T$ is μ-incoherent. Then, it holds that

$$\mathbb{D}_\alpha^{asym}(C, u^*, v^*) \le \max\left\{\frac{\max\{m, n\}(1+\delta)}{4\alpha(1-\delta)}, \frac{1}{2(1-\alpha)\mu}\right\} \times \min\left\{\left(\frac{1}{\mu} - \frac{1}{\min\{m, n\}}\right)_+^{-1}, 3\mu\right\},$$

where we define $x_+ := \max\{x, 0\}$ and $1/0 = +\infty$.

If we choose $1 - \alpha = \Theta(n^{-1})$, the complexity can be upper-bounded by

$$\mathbb{D}_\alpha^{asym}(C, u^*, v^*) = O\left(\mu \max\{m, n\} \cdot \frac{1+\delta}{1-\delta}\right).$$

In the case when $1 - \delta = \Theta(1)$ and $\mu = O(1)$, the upper bound is on the same order (i.e., $O(\max\{m, n\})$) as the minimum possible complexity.

Next, we consider the case when components of M^* are observed under the Bernoulli model with parameter p.

Theorem 36. *Given $\mu \in [1, n]$ and $p \in (0, 1]$, suppose that the weight matrix C obeys the Bernoulli model with the parameter p and that u^* has incoherence μ. If $\eta > 2$ is a constant and the sampling rate satisfies*

$$p \ge \min\left\{1, \frac{(m+n)[16(1+\eta\mu)\log(mn) + 16]}{mn}\right\},$$

then it holds with probability at least $1 - O[(mn)^{-\eta/2+1}]$ *that*

$$\mathbb{D}_{\alpha}^{asym}(C, u^*, v^*) \leq \max\left\{\frac{3\max\{m,n\}}{4\alpha}, \frac{1}{2(1-\alpha)\mu}\right\} \times \min\left\{\left(\frac{1}{\mu} - \frac{1}{\min\{m,n\}}\right)_+^{-1}, 3\mu\right\}.$$

In the case when $1 - \alpha = \Theta(n^{-1})$ and $\mu = O(1)$, the upper bound is on the order of $O(\max\{m, n\})$, which is also the same as the minimum possible complexity.

Theoretical Results

Now, we extend the theoretical results in Section 3.4 to the asymmetric case. We first prove that if the complexity metric is on the order of $O(\max\{m, n\})$, there does not exist spurious second-order critical point. This result is established in the case when we choose $\alpha = \alpha_{asym}^*$, where α_{asym}^* is the minimizer of the minimum possible complexity metric:

$$\mathbb{D}_{\alpha}^{min,asym} := \min_{C \in \mathbb{S}_{+,1}^{nm-1}, u^* \in \mathbb{S}_1^{m-1}, v^* \in \mathbb{S}_1^{n-1}} \mathbb{D}_{\alpha}^{asym}(C, u^*, v^*).$$

The following theorem provides a characterization of the complexity metric when $\alpha = \alpha_{asym}^*$.

Theorem 37. *It holds that*

$$\alpha_{asym}^* = 1 - \frac{1}{\max\{m,n\}+1}, \quad \mathbb{D}_{\alpha_{asym}^*}^{min,asym} = \frac{\max\{m,n\}}{2\alpha_{asym}^*}.$$

Moreover, the complexity metric $\mathbb{D}_{\alpha_{asym}^*}^{asym}(C, u^*, v^*)$ *is equal to* $\mathbb{D}_{\alpha_{asym}^*}^{min,asym}$ *if and only if*

$$C_{ij} = \frac{1}{mn}, \quad \forall i \in [m],\ j \in [n], \quad u_i^* = \frac{1}{m}, \quad \forall i \in [m], \quad v_j^* = \frac{1}{n}, \quad \forall j \in [n].$$

The next theorem states that the optimization landscape is benign when the complexity is close to $\mathbb{D}_{\alpha_{asym}^*}^{min,asym}$.

Theorem 38. *Suppose that* $\alpha = \alpha_{asym}^*$. *Then, there exists a constant* $\delta > 1/2$ *such that for every instance* $\mathcal{MC}(C, u^*, v^*)$ *satisfying*

$$\mathbb{D}_{\alpha_{asym}^*}^{asym}(C, u^*, v^*) \leq \delta \max\{m,n\}/\alpha_{asym}^*,$$

the instance $\mathcal{MC}(C, u^*, v^*)$ *does not have any SSCPs.*

Next, we consider instances with a large complexity. We note that the landscape of problem (3.80) is "scale-invariant". Namely, if (u, v) is a stationary point of problem (3.80), the scaled point $(c_1 u, c_1^{-1} v)$ is also a stationary point of problem (3.80) for all constants $c_1 \neq 0$. To deal with this problem, consider a regularized version of problem (3.80):

$$\min_{u \in \mathbb{R}^m, v \in \mathbb{R}^n} \sum\nolimits_{i \in [m], j \in [n]} C_{ij}(u_i v_j - M_{ij}^*)^2 + \lambda(u^T u - v^T v)^2, \tag{3.81}$$

where $\lambda > 0$ is the regularization parameter. We denote instances of problem (3.81) as $\mathcal{MC}_{reg}(C, u^*, v^*)$. It is proved in [256] that problems (3.80) and (3.81) are equivalent in the sense that they have the same set of local minima under scaling; see [244] for a more detailed discussion. We note that adding the regularizer to problem (3.80) will not affect the existence of SSCPs, and we consider problem (3.81) since it is desirable to construct degenerate instances with a positive definite Hessian matrix at global minima. Similar to the symmetric case, we define the following subset of $\mathcal{D}_{asym}$:

$$\begin{aligned}\mathcal{SD}_{asym} := \{(C, u^*, v^*) \in \mathcal{D}_{asym} \mid \mathbb{G}_1(C, u^*, v^*) \text{ is disconnected,} \\ \mathcal{I}^u_{00}(C, u^*, v^*) = \mathcal{I}^v_{00}(C, u^*, v^*) = \emptyset\}.\end{aligned}$$

The following theorem proves that the Hessian matrix is positive definite at global solutions for instances in $\mathcal{SD}_{asym}$.

Theorem 39. *Suppose that $(C, u^*, v^*) \in \mathcal{SD}_{asym}$. Then, the Hessian matrix of the objective function of problem* (3.81) *is positive definite at all global solutions of the instance $\mathcal{MC}(C, u^*, v^*)$.*

The next step is to consider a closed subset of $\mathcal{SD}_{asym}$, which is defined as

$$\begin{aligned}\mathcal{SD}_{asym,\epsilon} := \big\{(C, u^*, v^*) \in \mathcal{SD}_{asym} \mid C_{ij} \in \{0\} \cup [\epsilon, 1], \quad \forall i \in [m],\ j \in [n], \\ |u^*_i| \in \{0\} \cup [\epsilon, 1], \quad \forall i \in [m], \quad |v^*_j| \in \{0\} \cup [\epsilon, 1], \quad \forall j \in [n]\big\}.\end{aligned}$$

Define the alternative complexity metric as

$$\mathbb{D}^{asym}_{\alpha,\epsilon}(C, u^*, v^*) := \left[\min_{(\tilde{C},\tilde{u}^*,\tilde{v}^*)\in\mathcal{SD}_{asym,\epsilon}} \alpha\|C - \tilde{C}\|_1 + (1-\alpha)(\|u^* - \tilde{u}^*\|_1 + \|v^* - \tilde{v}^*\|_1)\right]^{-1}.$$

The new metric $\mathbb{D}^{asym}_{\alpha,\epsilon}$ is a lower bound on the original metric $\mathbb{D}^{asym}_{\alpha}$. The following theorem provides a sufficient condition on the existence of spurious local minima for problems (3.80) and (3.81).

Theorem 40. *Suppose that $\epsilon > 0$ is a constant. Then, there exists a large constant $\Delta(\epsilon) > 0$ such that for every instance $\mathcal{MC}(C, u^*, v^*)$ satisfying*

$$\mathbb{D}^{asym}_{\alpha,\epsilon}(C, u^*, v^*) \geq \Delta(\epsilon),$$

both instances $\mathcal{MC}(C, u^*, v^*)$ and $\mathcal{MC}_{reg}(C, u^*, v^*)$ have spurious local minima.

Part II

Convex Discrete Optimization via Simulation

Chapter 4

Gradient-based Simulation-optimization Methods

4.1 Introduction

Many decision making problems in operations research and management science involve large-scale complex stochastic systems. The objective function in the decision making problems often involve expected system performances that need to be evaluated by discrete-event simulation or general stochastic simulation. The decision variables in many of these problems are naturally discrete-valued and multi-dimensional. This class of problems is called *Discrete Optimization via Simulation*, *Discrete Simulation Optimization*, or *Simulation/Stochastic Optimization with Integer Decision Variables* (see [169, 103, 189]). Typically for discrete optimization via simulation problems, continuous approximations are either not naturally available or may incur additional errors that are themselves difficult to accurately quantify; see [169]. this chapter is centered around designing and proving theoretical guarantees for simulation-optimization algorithms to solve discrete optimization via simulation problems with multi-dimensional decision space.

For complex stochastic systems, even one replication of simulation can be time consuming or costly; see also [236, 212, 239, 237] for related discussions. An accurate evaluation of the expected performance associated with a single decision needs many independent replications of simulation. Running simulations for all feasible choices of decision variables in a high-dimensional discrete space to find the optimal is computationally prohibitive. The use of parallel computing (e.g. [155]) may alleviate the computation burden, but to find the best decision in high-dimensional problems can still be challenging. Fortunately, for a number of applications, the objective function exhibits convexity in the discrete decision variables, or the problem can be transformed into a convex one. One such example with convex structure comes from a bike-sharing system [205, 121, 77]. This problem involves around 750 stations and 25,000 docks. The goal is to find the optimal allocation of bikes and docks, which are naturally discrete decision variables. The performance of each allocation is evaluated by the

dissatisfaction function, which is defined as the total number of failures to rent or return a bike in a whole day. In the presence of non-stationary exogenous random demands and travel patterns, the evaluation of the dissatisfaction function for a given allocation needs to be done by simulation. This simulation is costly as it may need to simulate the full operation of the system over the entire day.

When the decision space is large, it is often computationally impractical to run simulations for all choices of the decision variables, creating a challenge in finding the optimal or near-optimal choice of decision variables. To circumvent this challenge, problem structure such as convexity or local convexity of the objective function may need to be exploited to lower costs and improve the efficiency to find an optimal or near-optimal choice. See [105] for a more detailed overview. In [77], the expected dissatisfaction function is proved to be "convex" under a linear transformation if the stochastic arrival processes are exogenous. For this problem, running stochastic simulations for the entire discrete and high-dimensional decision space is computationally prohibitive. It is therefore of interest to explore how the convexity structure of the objective function may help solve the simulation-optimization problem. In fact, many performance functions in the operations research and management science domain exhibit convexity in discrete decision variables. For example, the expected customer waiting time in a multi-server queueing network was proved to be convex in the routing policy and staffing decisions; see [10] and [232]. [198] discusses a wide range of stochastic systems including queueing systems, reliability systems and branching systems and show the convexity of key expected performance measures as a function of the associated decision variable. In addition, a large variety of problems in economics, computer vision and network flow optimization exhibit convexity with discrete decision variables [168].

Even in the presence of convexity, the nominal task in discrete optimization via simulation – correctly finding the best decision with high enough probability, which is often referred to as the *Probability of Correct Selection* (PCS) guarantee – can still be computationally prohibitive. For a convex problem without convenient assumptions such as strong and strict convexity, there may be a large number of choices of decision variables that render very close objective value compared to the optimal. In this case, the simulation efforts to identify the exact optimal choice of decision variables can be huge and practically unnecessary. Our focus, alternatively, is to find a good choice of decision variables that is assured to render ϵ-close objective value compared to the optimal with high probability, where ϵ is any arbitrarily small user-specified precision level. This guarantee is also called the *Probability of Good Selection* (PGS) or *Probably Approximately Correct* (PAC) in the literature. this chapter adopts the notion of PGS as a guarantee for simulation-optimization algorithms design. We refer to [67, 68] for thorough discussions on settings when the use of PGS is preferable compared to the use of PCS. In this chapter, we propose simulation-optimization algorithms that achieve the PGS guarantee for general discrete convex problems, without knowing any further information such as strong convexity, etc. Knowing strong convexity or a specific parametric function form of the objective function, of course, will further enhance the simulation-optimization algorithms. However, such fine structural information may not be available a priori for large-scale simulation optimization problems. The design of our simulation-optimization

algorithms utilizes the convex structure and the intuition is that the convex structure of optimization landscapes can provide *global information* through *local evaluations*. Global information helps the algorithm avoid evaluating all feasible choices of decision variables, which therefore avoids spending simulation efforts that are proportional to the number of choices of decision variables and are exponentially dependent on the dimension in general. Our proposed simulation-optimization algorithms are based on stochastic gradient methods and discrete steepest descent methods, which need to be designed as fundamentally different from continuous optimization algorithms. For high-dimensional problems, gradient-based methods are preferred compared to strongly polynomial methods like cutting-plane methods, because the simulation costs of gradient-based methods usually have a slower growth rate when the dimension increases.

In order to compare algorithms that all return a solution that achieves the PGS optimality guarantee, we use the metric of expected simulation cost. Intuitively here but with exact definition to follow in the main body of this chapter, the expected simulation cost is described by the expected number of simulation replications that are run over the decision space, in order to achieve a solution with the PGS guarantee. We prove upper bounds on the expected simulation cost for our proposed simulation-optimization algorithms that achieve the PGS guarantee. The proven upper bounds show a low-order polynomial dependence on the decision space dimension d. Note that the upper bounds hold for any arbitrary convex problem. As a comparison, if the convex structure is not present or utilized, the expected simulation cost to achieve the PGS guarantee can easily be exponential in the dimension d. We also provide lower bounds on the expected simulation costs that are needed for any possible simulation-optimization algorithm. The lower and upper bounds of expected simulation costs imply the limit of algorithm performance and provide directions to improving existing simulation-optimization algorithms. In general, we refer readers to [160] and [254] for more detailed discussions on the use of simulation costs and upper/lower bounds on the order of simulation costs to analyze and compare algorithms.

Main Results and Contributions

We design gradient-based simulation-optimization algorithms that achieve the PGS guarantee for high-dimensional and large-scale discrete convex problems with a known upper bound on the level of overall uncertainties. We consider the decision space to be

$$\{(x_1, x_2 \ldots, x_d) \mid x_i \in \{1, 2, \ldots, N\},\ i \in \{1, 2, \ldots, d\}\},$$

which has in total N^d possible choices of decision variables. The discrete convexity in high dimension that preserves the mid-point convexity (namely, the mid-point has an objective value smaller than the average of objective values at the two endpoints) is called $L^\natural$-convexity [168]. From the optimization perspective, our work addresses the stochastic version of discrete convex analysis in [168]. From the simulation optimization perspective, this chapter provides simulation-optimization algorithms with optimality guarantee and polynomial de-

pendence of simulation costs on dimension, for high-dimensional discrete convex simulation optimization problems.

We categorize our simulation-optimization algorithms to two classes. One class is the *zeroth-order algorithm*, for which the simulation is a black-box and one run of simulation can only provide an evaluation of a single decision. The other class is the *first-order algorithm*, for which the neighboring choices of decision variables can be simultaneously evaluated (possibly results in a biased finite difference gradient estimator) within a single simulation run for a given choice of decision variables. We develop simulation-optimization algorithms with the PGS guarantee as a major focus, but we also provide algorithms with the PCS-IZ guarantee for cases when the indifference zone (IZ) parameter is known. See [101] for detailed discussions on the PCS-IZ guarantee. We summarize our results in Table 4.1.1, where algorithm performance is demonstrated by the expected simulation cost. In this table, we omit terms in the expected simulation cost that do not depend on the failing probability δ, i.e., the probability that the solution does not satisfy the specified precision. Therefore, when δ is very small, the dominating term in the expected computation cost is what we list in Table 4.1.1. This comparison scheme is also considered in [128]. That being said, we provide all terms in the upper bounds for expected simulation costs in corresponding theorems.

Algorithms	**PGS**	**PCS-IZ** (known IZ parameter c)
Zeroth-order Alg. (Gaussian Noise)	$\tilde{O}(d^2N^2\epsilon^{-2}\log(1/\delta))$ (Lower bound: $\tilde{O}(d\epsilon^{-2}\log(1/\delta))$)	$\tilde{O}(d^2\log(N)c^{-2}\log(1/\delta))$
Zeroth-order Alg. (Assumption 10)	$\tilde{O}(dN^2\epsilon^{-2}\log(1/\delta))$	$\tilde{O}(d\log(N)c^{-2}\log(1/\delta))$
Lower Bound	$\tilde{O}(d\epsilon^{-2}\log(1/\delta))$	$\tilde{O}(dc^{-2}\log(1/\delta))$
Biased First-order Alg. (Assumption 9)	$\tilde{O}(N^3\epsilon^{-2}\log(1/\delta))$ (requires additional memory cost)	$\tilde{O}(Nc^{-2}\log(1/\delta))$

Table 4.1.1: Upper bounds and lower bounds on expected simulation cost for the proposed simulation-optimization algorithms that achieve the PGS and the PCS-IZ guarantees. Constants and terms that do not depend on δ are omitted in the $\tilde{O}(\cdot)$ notation. In comparison, the expected simulation cost without $L^\natural$-convexity is $\tilde{O}(N^d\epsilon^{-2}\log(1/\delta))$. Here, d and N are the problem dimension and scale; the feasible set is $\{1,\ldots,N\}^d$; constants ϵ and δ are the precision and failing probability of algorithms.

For zeroth-order algorithms, the Lovász extension [152] is introduced to define a convex

linear interpolation of the original discrete function. Using properties of the Lovász extension [80], it is equivalent to optimize the interpolated continuous function. Therefore, the projected stochastic subgradient descent method can be used to find PGS solutions. Moreover, the truncation of stochastic subgradients is essential in reducing the expected simulation costs and we prove that the dependence on the dimension d is reduced from $O(d^3)$ to $O(d^2)$ using truncation. In stochastic optimization literature, it is common to assume the stochastic subgradient is bounded when deriving high-probability bounds, and we also provide a theoretical guarantee under the boundedness assumption. When the boundedness assumption can be verified, the dependence on dimension can be further reduced to $O(d)$. When the indifference zone parameter c is known, an accelerated algorithm is proposed and is proved to reduce the dependence on the scale N from $O(N^2)$ to $O(\log(N))$. Finally, an information-theoretical lower bound is derived to show the limit of simulation-optimization algorithms.

For first-order algorithms, we have available gradient information, at a cost as a constant multiplying the cost of one simulation run, for which the constant does not depend on the dimension. This gradient information is regarded as a subgradient estimator. In practice, the subgradient estimator can be biased, and there is no convergence guarantee for any optimization algorithm in general. However, under a moderate assumption on the bias, we are still able to develop simulation-optimization algorithms that achieve the PGS guarantee through a stochastic version of the steepest descent method. The associated simulation cost does not scale up with d, but the memory cost and the number of arithmetic operations can be much larger than those of simulation-optimization algorithms designed for the unbiased gradient estimators. Finally, utilizing the indifference zone, the expected simulation cost can be reduced from $O(N^3)$ to $O(N)$ in terms of dependence on N.

Literature Review

The problem of selecting the best or a good choice of decision variables through simulation has been widely studied in the simulation literature. The problem is often called *ranking-and-selection* (R&S). We refer to [101] as a recent review of this literature. There have been two approaches to categorize the R&S literature. One approach is differentiating the frequentist view and the Bayesian view when describing the probability models and procedures in R&S; see [131] and [52]. The other approach differentiates the fixed-confidence procedures and the fixed-budget procedures; see [112] and [101]. In particular, the probability of correct selection (PCS) of the best choice of decision variables has been a widely used guarantee for both types of procedures. Generally in the R&S problems, there is no structural information such as convexity that is considered.

A large number of R&S procedures based on the PCS guarantee adopt the indifference zone formulation, called PCS-IZ. The PCS-IZ guarantee is built upon the assumption that the expected performance of the best choice of decision variables is at least $c > 0$ better than all other choices of decision variables. This IZ parameter c is typically assumed to be known, while [74], as a notable exception, provides selection guarantees without the knowledge of

the indifference-zone parameter. In practice, for some problem settings, this IZ parameter may be unknown a priori. When many choices of decision variables have close performance compared to the best, it is practically inefficient to select the exact best. In this case, choices of decision variables that are close enough to the best are referred to as "good choices" and any one of them can be satisfying. This naturally gives rise to a notion of PGS. [67, 68] have thoroughly discussed settings when the use of PGS is preferable to the use of PCS-IZ.

Discussions on discrete optimization via simulation can be found in [78], [169], [212], [182, 181], [103] and [44] among others. [107, 108] have discussed model reference adaptive search algorithms in order to ensure global convergence. [102, 104, 236] propose and study algorithms based on the convergent optimization via most-promising-area stochastic search (COMPASS) that can be used to solve general simulation optimization problems with discrete decision variables. The proposed algorithms are computationally efficient and are proven to converge with probability one to optimal points. [149] studies simulation optimization problems over multidimensional discrete sets where the objective function adopts multimodularity, which is equivalent to the submodularity under a linear transform; see two equivalent definitions of multimodular functions in [11] and [168]. They propose algorithms that converge almost surely to the global optimal. [226] discusses stochastic optimization problems with integer-ordered decision variables.

When a simulation problem involves a response surface to estimate or optimize over, gradient information may be constructed and used to enhance simulation. [43] constructs gradient estimator to enhance simulation metamodeling. [188] proposes a new approach called gradient extrapolated stochastic kriging that exploits the extrapolation structure. [79] discusses the use of Monte Carlo gradient estimators to enhance regression. See also [135] for a review of Monte Carlo gradient estimators. [71] discusses the use of possibly biased gradient estimators in continuous stochastic optimization, by assuming that the bias is uniformly bounded. [228] considers a setting in which the response surface is a quadratic function and gradient information is available and discusses optimal budget allocation to maximize the probability of correct selection.

Discrete optimization via simulation is also formulated as the best-arm identification problem, or the pure-exploration multi-armed bandits problem. The best-arm identification literature usually does not consider the problem structure nor the high-dimensional nature of an arm. More recent works focus on general distribution families and utilize techniques from the information theory. Informational upper bounds and lower bounds for exponential bandit models are established by the change of measure technique in [129, 128]. In [85], a transportation inequality is proved and a general non-asymptotic lower bound can be formulated thorough the solution of a max-min optimization problem. [3] shows that restrictions on the distribution family are necessary and generalizes the algorithm to models with milder restriction than exponential family.

Discrete optimization via simulation problems fall into the more general class of problems called discrete stochastic optimization. In contrast to continuous optimization, most works on discrete stochastic optimization [82, 94, 83, 132, 196] do not consider the convex structure. The main obstacle to the development of discrete convex optimization lies in the lack of a

suitable definition of the discrete convex structure. A natural definition of the discrete convex functions would be functions that are extensible to continuous convex functions. However, for that class of functions, the local optimality does not imply the global optimality and therefore it is not suitable for the purpose of optimization. An example with spurious local minima is given in Section 4.2. Later, [76] proposes a stronger condition, named the integral convexity, that ensures the local optimality is equivalent to the global optimality. On the other hand, after [152] shows the equivalence between the submodularity of a function and the convexity of its Lovász extension, submodular functions are viewed as the discrete analogy of convex functions in the field of combinatorial optimization. The Fenchel-type min-max duality theorem [81] and the subgradient [80] of submodular functions provide a good framework of applying gradient-based method to the submodular function minimization (SFM) problem. The SFM problem has wide applications in computer vision, economics, game theory and is well-studied in literature [140, 15, 246]. In contrast, the stochastic SFM problem is less understood and [114] gives the only result on stochastic SFM problem, where they provide upper and lower bounds for finding solutions with small error bound in expectation. In [168], a generalization of submodular functions, called the $L^\natural$-convex functions, are defined through the translation submodularity. The $L^\natural$-convex functions are equivalent to functions that are both submodular and integrally convex on integer lattice. In addition, the $L^\natural$-convex function has a convex extension that shares similar properties as the Lovász extension and therefore gradient-based methods are also applicable for $L^\natural$-convex functions minimization.

Two most recent papers [228] and [70] also discuss the use of the convexity structure in simulation. [228] consider a discrete simulation optimization problem with a specific polynomial functional form for the objective function, and focus on how to strategically use gradient information to accelerate the selection of the best. However, in general simulation optimization problems, when the decision variables are discrete, the gradient with respect to the decision variable may not be appropriately defined. Instead, the difference of performance between two neighboring choices of decision variables contains gradient-like information. [120] uses this information to guide the search for the optimal choices of decision variables. In addition, they focused on the fixed-budget problem with an approximately quadratic objective function and a one-dimensional decision space, which is different from our problem setting. [70] utilize the convexity structure to select a feasible region that contains the optimal given existing simulation samples at different choices; see also [69]. Because they do not consider an optimization problem and their goal is not to find an optimal or near-optimal solution, the focus of [70] is different from ours. For example, they do not provide simulation-optimization algorithms that can find an optimal or near-optimal decision, nor do they analyze simulation costs and their dependence on problem scale. On the other hand, the method and analysis provided by [70] and [69] can serve effectively as a module to help solve other general simulation problems, such as multi-objective simulation optimization, which is not the focus of our work.

4.2 Model and Framework

The model in consideration contains a complex stochastic system whose performance depends on discrete decision variables that belong to a discrete feasible set $\mathcal{X} \subset \mathbb{Z}^d$. From a modeling perspective, in a stochastic system, the system performance may depend on three elements: the decision variable $x \in \mathbb{Z}^d$, a random object ξ_x supported on a proper space $(\mathsf{Y}, \mathcal{B}_\mathsf{Y})$ that summarizes all the associated random quantities and processes involved in the system when the decision x is taken, and a deterministic function $F : \mathcal{X} \times \mathsf{Y} \to \mathbb{R}$ that takes the value of decision variables and a realization of the randomness as inputs and outputs the associated system performance. Specifically, the deterministic function F captures the full operations logic of the stochastic system, which can be complicated. The objective function with decision variable x is given by

$$f(x) := \mathbb{E}[F(x, \xi_x)].$$

We consider scenarios when $f(x)$ does not adopt a closed-form representation and can only be evaluated by averaging over simulation samples of $F(x, \xi_x)$. More specifically, we write $\xi_{x,1}, \xi_{x,2}, \ldots, \xi_{x,n}$ as independent and identically distributed (iid) copies of ξ_x. We use $\hat{F}_n(x) := \frac{1}{n}\sum_{j=1}^{n} F(x, \xi_{x,j})$ to denote the empirical mean of the n independent evaluations for the choice of decision variables x. The selection of the optimal choice of decision variables is through the selection of a choice of decision variable x that renders the best objective value $f(x)$. Denote x^* as any choice of decision variable that renders the optimal objective value, such that

$$f(x^*) = \min_{x \in \mathcal{X}} \; f(x). \tag{4.1}$$

Note that we fix the use of minimum operation to represent the optimal. Our general goal is to develop simulation-optimization algorithms that select a good choice of decision variable x, such that

$$f(x) - f(x^*) \leq \epsilon,$$

where $\epsilon > 0$ is the given user-specified precision level. In this chapter, we consider this selection problem in a large decision space with high dimension.

Because f does not have a closed-form representation and has to be evaluated by simulation, we take the view that no further structure information is available in addition to the convex structure. For instance, for a real-world model, f may have a very flat landscape around the minimum, which may not be known a priori. In this case, there may be a number of choices of decision variables that render objective value that is at most ϵ apart from the optimal. This also motivates our goal to select a good choice of decision variables instead of the best, because too much computational resource may be needed to identify exactly the best, when the landscape around the minimum is flat. Therefore, our general goal is to develop simulation-optimization algorithms that are expected to robustly work for any convex model without knowing further specific structure.

Because the precision level ϵ cannot be delivered almost surely with finite computational budget for simulation, we consider a selection optimality guarantee called *Probability of Good Selection*; see [67, 68, 101].

- *Probability of good selection* (PGS). With probability at least $1 - \delta$, the solution x returned by an algorithm has objective value at most ϵ larger than the optimal objective value.

This PGS guarantee is also called the probably approximately correct selection (PAC) guarantee in the literature [73, 128, 159]. While our focus is to design algorithms that satisfy the PGS optimality guarantee, we also consider the optimality guarantee of *Probability of Correct Selection with Indifference Zone* as a comparison.

- *Probability of correct selection with indifference zone* (PCS-IZ). The problem is assumed to have a unique solution that renders the optimal objective value. The optimal value is assumed to be at least $c > 0$ smaller than other objective values. The gap width c is called the *indifference zone parameter* in [17]. The PCS-IZ guarantee requires that with probability at least $1 - \delta$, the solution x returned by an algorithm is the unique optimal solution.

By choosing $\epsilon < c$, algorithms satisfying the PGS guarantee can be directly applied to satisfy the PCS-IZ guarantee. On the other hand, counterexamples in [67] show that algorithms satisfying the PCS-IZ guarantee may fail to satisfy the PGS guarantee. This phenomenon is further explained from the hypothesis-testing perspective in [101]. The failing probability δ in either PGS or PCS-IZ is typically chosen to be very small to ensure a high probability result. Hence, we assume in the following of this chapter that δ is small enough and focus on the asymptotic expected simulation cost.

To facilitate the construction of simulation-optimization algorithms that can deliver the PGS guarantee for general convex problems, we specify the composition of simulation-optimization algorithms in the next subsection. In addition, we assume that the probability distribution for the simulation output $F(x, \xi_x)$ is sub-Gaussian.

Assumption 5. The distribution of $F(x, \xi_x)$ is sub-Gaussian with known parameter σ^2 for any $x \in \mathcal{X}$.

The sub-Gaussian distributional assumption part in Assumption 5 is standard in simulation optimization literature; see for example the discussions in [253]. One special case is that the probability distribution for the simulation output at a choice of decision variables x is Gaussian with variance σ_x^2. However, it is indeed possible that these variances for different x's are unknown in advance, therefore posing a challenge. In that regard, one may consider using the system structure to provide a generic upper bound $\sigma^2 \geq \max_{x \in \mathcal{X}} \sigma_x^2$, particularly when the maximum possible level of uncertainties associated with a system is available. In practice, if the decision maker knows in advance what specific extreme choices of decision variables lead to the highest achievable variance of the system, that would be significantly

valuable to find the upper bound. In general, when the variances are not known in advance, such a generic upper bound can sometimes be loose and therefore is conservative. In this chapter, we take the view that an upper bound (maybe a loose one) is known in advance, and focus on the algorithm design to search for a good solution that has light dependence on the dimension. Note that our analysis under Assumption 5 can be naturally extended to models whose randomness distribution satisfies certain concentration inequalities. For example, when the randomness is sub-exponential (which may have heavier tails than Gaussian), one can apply the Hoeffding-Azuma inequality for sub-exponential tailed martingales to achieve provably efficient algorithms.

Simulation-optimization Algorithms

In this subsection, we define different classes of simulation-optimization algorithms. We hope to design simulation-optimization algorithms that can deliver certain optimality guarantee, say, PGS, for any convex model without knowing further structure. A broad range of sequential simulation-optimization algorithms consist of three parts.

- The *sampling rule* determines which choice of decision variables to simulate next, based on the history of simulation observations up to current time.
- The *stopping rule* controls the end of the simulation phase and is a stopping time according to the filtration up to current time. We assume that the stopping time is finite almost surely.
- The *recommendation rule* selects the choice of decision variables that satisfies the optimality guarantee based on the history of simulation observations.

The *model* of problem (4.1) consists of the decision set $\mathcal{X}$, the space of randomness $(\mathsf{Y}, \mathcal{B}_{\mathsf{Y}})$ and the function $F(\cdot, \cdot)$. Next, we define the class of simulation-optimization algorithms that can deliver solutions satisfying certain optimality guarantee for a given set of models.

Definition 10. Suppose the optimality guarantee $\mathcal{O}$ and the set of models $\mathcal{M}$ is given. A simulation-optimization algorithm is called an $(\mathcal{O}, \mathcal{M})$-algorithm, if for any model $M \in \mathcal{M}$, the algorithm returns a solution to M that satisfies the optimality guarantee $\mathcal{O}$.

We define the set of all models such that the objective function $f(\cdot)$ is convex (defined in the next subsection) on the discrete set $\mathcal{X}$ as $\mathcal{MC}(\mathcal{X})$, or simply $\mathcal{MC}$. Using this definition, a $(\text{PGS}, \mathcal{MC})$-algorithm is one that guarantees the finding of a solution that satisfies the PGS guarantee for any convex model without knowing further structure.

Simulation Costs

In the development of simulation-optimization algorithms that satisfy a certain optimality guarantee, especially for large-scale problems, the performance of different algorithms can

be compared based on the their computational costs to achieve the same optimality guarantee. We take the view that the simulation cost of generating replications of $F(x, \xi_x)$ is the dominant contributor to the computational cost associated with a simulation-optimization algorithm. See also [155], [176], and [160]. Therefore, we quantify the computational cost as the total number of evaluations of $F(x, \xi_x)$ for all $x \in \mathcal{X}$. In some simulation problems but not all, we may also have access to noisy and possibly biased estimates of $f(\cdot)$ near point x along with an evaluation of $F(x, \xi_x)$. The simulation cost in this case is discussed in Section 4.6. For all simulation-optimization algorithms proposed in this chapter, we provide upper bounds on the expected simulation cost to achieve a certain optimality guarantee. Note that these upper bounds do not rely on the specific structure of the problem in addition to convexity. The expected simulation cost serves as a measurement to compare different algorithms and provide insights on how the computational cost depends on the scale and dimension of the problem.

Now, we define the expected simulation cost for a given set of models $\mathcal{M}$ and given optimality guarantee $\mathcal{O}$.

Definition 11. Given the optimality guarantee $\mathcal{O}$ and a set of models $\mathcal{M}$, the *expected simulation cost* is defined as

$$T(\mathcal{O}, \mathcal{M}) := \inf_{\mathbf{A} \text{ is } (\mathcal{O},\mathcal{M})} \sup_{M \in \mathcal{M}} \mathbb{E}[\tau],$$

where $\mathbf{A}$ is a simulation-optimization algorithm and τ is the stopping time of the algorithm $\mathbf{A}$, which is also the number of simulation evaluations of $F(\cdot, \cdot)$.

The notion of simulation cost in this chapter is largely focused on

$$T(\epsilon, \delta, \mathcal{MC}) := T((\epsilon, \delta)\text{-}PGS, \mathcal{MC}), \quad T(\delta, \mathcal{MC}_c) := T((c, \delta)\text{-}PCS\text{-}IZ, \mathcal{MC}_c).$$

Note that the (ϵ, δ)-PGS refers to the PGS optimality guarantee with user-specified precision level $\epsilon > 0$ and confidence level $1 - \delta$. The notion (c, δ)-PCS-IZ refers to the PCS-IZ optimality guarantee with confidence level $1 - \delta$ and IZ parameter c. The class of models $\mathcal{MC}$ include all convex models while $\mathcal{MC}_c$ include all convex models with IZ parameter c. In addition, we mention that all upper bounds derived in this chapter are actually almost sure bounds of the simulation cost, while lower bounds only hold in expectation.

Discrete Convex Functions in Multi-dimensional Space

In contrast to the continuous case, the discrete convexity has various definitions, e.g., convex extensible functions and submodular functions. Although these concepts coincide for the one-dimensional case, they have essential differences in the multi-dimensional case. In this chapter, we consider $L^\natural$-convex functions [168], which are defined by the mid-point convexity (defined later in this subsection) for discrete variables. Considerably many discrete optimization via simulation problems have the $L^\natural$-convex structure. For example, the

expected customer waiting time in a multi-server queueing network is proved to be a separated convex function [10, 232] and therefore is $L^\natural$-convex. In addition, the dissatisfaction function of bike-sharing system is shown to be multimodular in [77], which is $L^\natural$-convex under a linear transformation. More examples of $L^\natural$-convex functions are given in [168]. On the other hand, the minimization of a $L^\natural$-convex function is equivalent to the minimization of its linear interpolation, which is continuous and convex. Combined with the closed-form subgradient, $L^\natural$-convex functions provide a good framework for studying discrete convex simulation optimization problems.

Before we give the definition of $L^\natural$-convexity, we first show that it is not suitable to define discrete convex functions just as functions that have a convex extension. The main problem of this definition based on extension is that the "local optimality" may not be equivalent to the global optimality, which is one of the important properties used in convex optimization. In the discrete case, we say a point $\bar{x}$ is a *local minimum* of $f(\cdot)$ if $f(\bar{x}) \leq f(x)$ for all feasible x such that $\|x - \bar{x}\|_\infty \leq 1$. Without this property, algorithms may get stuck at spurious local minima and fail to satisfy the optimality guarantee. We give an example to illustrate the failure.

Example 6. We consider the case when $N = 4$ and $d = 2$. The objective function is given as

$$f(x, y) := 4|2x + y - 8| + |x - 2y + 6|.$$

The function $f(x, y)$ is a convex function on the set $[1, 4]^2$ and the unique global minimizer is $(2, 4)$. When restricted to the integer lattice $\{1, 2, 3, 4\}^2$, the global minimizer is still $(2, 4)$. We consider the point $(3, 2)$ with objective value $f(3, 2) = 5$. In the local neighborhood $\{2, 3, 4\} \times \{1, 2, 3\}$, which contains points that have ℓ_∞-distance at most 1 from $(3, 2)$, the objective values are

$$\begin{aligned} &f(2, 1) = 18,\ f(3, 1) = 11,\ f(4, 1) = 12,\ f(2, 2) = 12,\\ &f(4, 2) = 14,\ f(2, 3) = 6,\ f(3, 3) = 7,\ f(4, 3) = 16. \end{aligned}$$

Thus, the point $(3, 2)$ is a spurious local minimizer of the discrete function. This shows that local optimality cannot imply global optimality.

On the other hand, the $L^\natural$-convexity ensures that local optimality implies global optimality. Similar to the continuous case, $L^\natural$-convex functions can be characterized by the mid-point convexity property.

Definition 12. A set $\mathcal{S} \subset \mathbb{Z}^d$ is called a $L^\natural$-convex set, if it holds that

$$x, y \in \mathcal{S} \implies \lfloor (x + y)/2 \rfloor, \lceil (x + y)/2 \rceil \in \mathcal{S}.$$

A function $f(x) : \mathcal{X} \mapsto \mathbb{R}$ is called a $L^\natural$-convex function, if $\mathcal{X}$ is a $L^\natural$-convex set and the discrete mid-point convexity holds:

$$f(x) + f(y) \geq f(\lceil (x + y)/2 \rceil) + f(\lfloor (x + y)/2 \rfloor), \quad \forall x, y \in \mathcal{X}.$$

The set of models such that $f(x)$ is $L^\natural$-convex on $\mathcal{X}$ is denoted as $\mathcal{MC}(\mathcal{X})$, or simply $\mathcal{MC}$. The set of models such that $f(x)$ is $L^\natural$-convex with indifference zone parameter c is denoted as $\mathcal{MC}_c(\mathcal{X})$, or simply $\mathcal{MC}_c$.

We assume that the objective function is $L^\natural$-convex in the remainder of this chapter.

Assumption 6. The objective function $f(x)$ is a $L^\natural$-convex function on the $L^\natural$-convex set $\mathcal{X}$.

Before proceeding to the properties, we provide a few examples of $L^\natural$-convex sets and $L^\natural$-convex functions.

Example 7. Examples of $L^\natural$-convex sets include the whole space $\mathbb{Z}^d$ and the hypercube $[N_1] \times [N_2] \times \cdots \times [N_d]$, where d and N_i are positive integers for all $i \in [d]$. Another important example of $L^\natural$-convex sets is the linearly transformed capacity-constrained hypercube; see the derivation in Section 4.7. Specifically, for positive integers d, N and $M \leq N$, the following set is $L^\natural$-convex:

$$\left\{ x \in \mathbb{Z}^d \mid x_1 \in [N],\ x_{i+1} - x_i \in [N],\ \forall i \in [d-1],\ x_d \leq M \right\}.$$

Examples of $L^\natural$-convex functions include the indicator function of any $L^\natural$-convex set, linear functions and separably convex functions, namely, functions having the form

$$f(x) = \sum_{i=1}^{d} f^i(x_i),$$

where $f^i(\cdot)$ is a convex function for all $i \in [d]$. See [168] for more examples.

In the following lemma, we list several properties of $L^\natural$-convex functions.

Lemma 23. *Suppose that the function $f(x) : \mathcal{X} \mapsto \mathbb{R}$ is $L^\natural$-convex. The following properties hold.*

- *There exists a convex function $\tilde{f}(x)$ on the convex hull $\operatorname{conv}(\mathcal{X})$ such that $\tilde{f}(x) = f(x)$ for all $x \in \mathcal{X}$.*
- *Local optimality is equivalent to global optimality:*

$$f(x) \leq f(y), \quad \forall y \in \mathcal{X} \iff f(x) \leq f(y), \quad \forall y \in \mathcal{X} \quad \text{s.t. } \|y - x\|_\infty = 1.$$

- *Translation submodularity holds:*

$$f(x) + f(y) \geq f((x - \alpha\mathbf{1}) \vee y) + f(x \wedge (y + \alpha\mathbf{1})),$$
$$\forall x, y \in \mathcal{X},\ \alpha \in \mathbb{N} \quad \text{s.t. } (x - \alpha\mathbf{1}) \vee y,\ x \wedge (y + \alpha\mathbf{1}) \in \mathcal{X}.$$

The $L^\natural$-convexity can be viewed as a combination of submodularity and integral convexity [168, Theorem 7.20]. Intuitively, the submodularity ensures the existence of a piecewise linear convex interpolation in the local neighborhood of each point, while the integral convexity ensures that the piecewise linear convex interpolations can be pieced together to form a convex function on $[1, N]^d$. In addition, we can calculate a subgradient of the convex extension with $O(d)$ function value evaluations. Hence, $L^\natural$-convex functions provides a good framework for extending continuous convex optimization theory to the discrete case.

4.3 Simulation-optimization Algorithms and Expected Simulation Costs for a Special Case

In this section and the following section, we propose simulation-optimization algorithms that achieve the PGS guarantee for any simulation optimization problem with a $L^\natural$-convex objective function. We prove upper bounds on the expected simulation costs. To better present the dependence of expected simulation costs on the scale and dimension of the problem, we assume that the feasible set is the hypercube $[N]^d$ in complexity analysis.

Assumption 7. The feasible set of decision variables is $\mathcal{X} = [N]^d$, where $N \geq 2$ and $d \geq 1$.

In large-scale simulation problems, either N, or d, or both N and d can be large. We note that if the feasible set $\mathcal{X}$ is a general $L^\natural$-convex set, the construction of the convex extension and the analysis are still valid by replacing N with $\max_{x,x' \in \mathcal{X}} \|x - x'\|_\infty$. Moreover, our algorithms are directly applicable to the case where $\mathcal{X}$ is a general $L^\natural$-convex set, which is also the minimal requirement on the feasible set for the definition of $L^\natural$-convexity. In this section, we start with a special case where the decision space is $\{0, 1\}^d$ for a large d. We defer the discussions for general N to Section 4.4. The simulator may have a general complex and discontinuous structure that no unbiased gradient estimator is available within the replication of simulation. For scenarios when a single replication of simulation can also generate gradient information at very low costs, we propose and analyze simulation-optimization algorithms in Section 4.6.

The general idea of designing simulation-optimization algorithms in the multi-dimensional case is to construct subgradients of the convex extension with $O(d)$ function value evaluations on the neighboring choices of a decision. Hence, the stochastic subgradient descent (SSGD) method can be used to solve problem (4.1). Compared with the bi-section method and general cutting plane methods, gradient-based methods have two advantages in our case. First, as pseudo-polynomial algorithms, gradient-based methods usually have lighter dependence on the problem dimension d compared to strongly polynomial or weakly polynomial algorithms. For example, the deterministic integer-valued submodular function minimization (SFM) problem can be solved with $\tilde{O}(d)$, $\tilde{O}(d^2)$, $\tilde{O}(d^3)$ function value evaluations using pseudo-polynomial [15], weakly polynomial and strongly polynomial [140] algorithms, respectively. Usually, gradient-based methods have extra polynomial dependence on the Lipschitz

constant of the objective function, in exchange for the reduced dependence on d. However, for a large group of problems, the Lipschitz constant may be estimated a priori. Moreover, we can design algorithms whose expected simulation cost does not critically rely on the Lipschitz constant, in the sense that the Lipschitz constant only appears in a smaller order term in the expected simulation cost. Hence, gradient-based methods are preferred for high-dimensional problems. On the other hand, ordinary cutting plane methods are not robust to noise and problem-specific stabilization techniques should be designed for stochastic problems [197], or complicated robust scheme should be constructed [170, 2]. Considering these two advantages of gradient-based methods, we focus on the SSGD method in designing our simulation-optimization algorithms and make the assumption that an upper bound of the ℓ_∞-Lipschitz constant is known a priori.

Assumption 8. An upper bound on the ℓ_∞-Lipschitz constant of $f(x)$ is known to be L a priori. Namely, we know beforehand that

$$|f(x) - f(y)| \leq L, \quad \forall x, y \in \mathcal{X} \quad \text{s. t. } \|x - y\|_\infty \leq 1.$$

We remark that this constant L, in the general decision-making contexts, reflects the impact on the objective function by a small change in the value of the high-dimensional decision variable. For example, in bike-sharing applications, this L may reflect the impact of allocating one more bike to a station. Whether the objective function being revenue or number of dissatisfied customers, the upper bound on the impact of allocating one more bike can be quantified. The estimation of L usually relies on the domain knowledge about the problem. For example, the user dissatisfaction function in the bike-sharing application takes values in $\{0, 1, \ldots, M\}$, where M is the expected number of users each day. Then, an estimate of the Lipschitz constant is $L \leq M$.

When the decision space is $\mathcal{X} = \{0, 1\}^d$, $L^\natural$-convex functions are equivalent to submodular functions and therefore problem (4.1) is equivalent to the stochastic submodular function minimization (stochastic SFM) problem. To prepare the design of simulation algorithms, we first define the Lovász extension of submodular functions and give an explicit subgradient of the Lovász extension at each point.

Definition 13. Suppose that function $f(x) : \{0, 1\}^d \mapsto \mathbb{R}$ is a submodular function, i.e., it holds that

$$f(x) + f(y) \geq f(x \wedge y) + f(x \vee y), \quad \forall x, y \in \{0, 1\}^d.$$

For any $x \in [0, 1]^d$, we say a permutation $\alpha_x : [d] \mapsto [d]$ is a *consistent permutation* of x, if

$$x_{\alpha_x(1)} \geq x_{\alpha_x(2)} \geq \cdots \geq x_{\alpha_x(d)}.$$

We define $S^{x,0} := (0, \ldots, 0)$. For each $i \in \{1, \ldots, d\}$, the *i-th neighbouring point* of x is defined as

$$S^{x,i} := \sum_{j=1}^{i} e_{\alpha_x(j)} \in \mathcal{X},$$

where vector e_k is the k-th unit vector of $\mathbb{R}^d$. We define the *Lovász extension* $\tilde{f}(x) : [0,1]^d \mapsto \mathbb{R}$ as

$$\tilde{f}(x) := f\left(S^{x,0}\right) + \sum_{i=1}^{d} \left[f\left(S^{x,i}\right) - f\left(S^{x,i-1}\right)\right] x_{\alpha_x(i)}. \tag{4.2}$$

We note that the value of the Lovász extension does not rely on the consistent permutation we choose. A numerical illustration of the Lovász extension is provided in the appendix. We list several well-known properties of the Lovász extension and refer their proofs to [152, 80]. We note that the subdifferential at point $x \in [0,1]^d$ is defined as the set

$$\partial \tilde{f}(x) = \left\{ g \in \mathbb{R}^d : \langle g, x - y \rangle \geq \tilde{f}(x) - \tilde{f}(y),\ \forall y \in [0,1]^d \right\}.$$

Lemma 24. *Suppose that Assumptions 5-8 hold. Then, the following properties of $\tilde{f}(x)$ hold.*

(i) For any $x \in \mathcal{X}$, it holds $\tilde{f}(x) = f(x)$.

(ii) The minimizers of $\tilde{f}(x)$ satisfy $\arg\min_{x\in[0,1]^d} \tilde{f}(x) = \arg\min_{x\in\{0,1\}^d} f(x)$.

(iii) Function $\tilde{f}(x)$ is a convex function on $[0,1]^d$.

(iv) A subgradient $g \in \partial \tilde{f}(x)$ is given by

$$g_{\alpha_x(i)} := f\left(S^{x,i}\right) - f\left(S^{x,i-1}\right), \quad \forall i \in [d]. \tag{4.3}$$

(v) Subgradients of $\tilde{f}(x)$ satisfy

$$\|g\|_1 \leq 3L/2, \quad \forall g \in \partial \tilde{f}(x),\ x \in [0,1]^d.$$

To apply the SSGD method to design simulation-optimization algorithms for problem (4.1), we need to resolve the following two questions:

- How to design an unbiased subgradient estimator?
- How to round an approximate solution in $[0,1]^d$ to an approximate solution in $\mathcal{X} = \{0,1\}^d$?

For the first question, we consider the subgradient estimator at point x as

$$\hat{g}_{\alpha_x(i)} := F\left(S^{x,i}, \xi_i^1\right) - F\left(S^{x,i-1}, \xi_{i-1}^2\right), \quad \forall i \in [d], \tag{4.4}$$

where ξ_i^j are mutually independent for $i \in [d]$ and $j \in [2]$. By definition, we know the components of $\hat{g}$ are mutually independent and the simulation cost of each $\hat{g}$ is $2d$. Using the subgradient defined in (4.3), we have

$$\mathbb{E}\left[\hat{g}_{\alpha_x(i)}\right] = \mathbb{E}\left[F\left(S^{x,i}, \xi_i\right) - F\left(S^{x,i-1}, \xi_{i-1}\right)\right] = f\left(S^{x,i}\right) - f\left(S^{x,i-1}\right) = g_{\alpha_x(i)}, \quad \forall i \in [d],$$

which means that $\hat{g}$ is an unbiased estimator of g.

Next, we consider the second question. We define the relaxed problem as

$$f^* := \min_{x\in[0,1]^d} \tilde{f}(x). \tag{4.5}$$

Properties (i) and (ii) of Lemma 24 imply that the original problem (4.1) is equivalent to the relaxed problem (4.5). In the deterministic case, suppose we already have an ϵ-optimal solution to problem (4.5), i.e., a point $\bar{x}$ in $[0,1]^d$ such that $\tilde{f}(\bar{x}) \leq f^* + \epsilon$. Then, we rewrite the Lovász extension in (4.2) as

$$\tilde{f}(\bar{x}) = \left[1 - \bar{x}_{\alpha_{\bar{x}}(1)}\right] f\left(S^{\bar{x},0}\right) + \sum_{i=1}^{d-1} \left[\bar{x}_{\alpha_{\bar{x}}(i)} - \bar{x}_{\alpha_{\bar{x}}(i+1)}\right] f\left(S^{\bar{x},i}\right) + \bar{x}_{\alpha_{\bar{x}}(d)} f\left(S^{\bar{x},d}\right), \tag{4.6}$$

which is a convex combination of $f\left(S^{\bar{x},0}\right), \ldots, f\left(S^{\bar{x},d}\right)$. Hence, there exists an ϵ-optimal solution among the neighboring points of $\bar{x}$. This means that we can solve a sub-problem with $d+1$ points to get the ϵ-optimal solution among neighboring points. For the stochastic case, a similar rounding process can be designed and we give the pseudo-code in Algorithm 2. The rounding process for the (c,δ)-PCS-IZ guarantee follows by choosing $\epsilon = c/2$.

Algorithm 2 Rounding process to a feasible solution

Input: Model $\mathcal{X}, \mathcal{B}_Y, F(x,\xi_x)$, optimality guarantee parameters ϵ, δ, $(\epsilon/2, \delta/2)$-PGS solution $\bar{x}$ to problem (4.5).

Output: An (ϵ,δ)-PGS solution x^* to problem (4.1).

1: Compute a consistent permutation of $\bar{x}$, denoted as α.
2: Compute the neighbouring points of $\bar{x}$, denoted as $S^0, \ldots, S^d$.
3: Simulate at S^i until the $1 - \delta/(4d)$ confidence half-width of $\hat{F}_n(S^i)$ is smaller than $\epsilon/4$ for all i.
4: Return the optimal point $x^* \leftarrow \arg\min_{S\in\{S^0,\ldots,S^d\}} \hat{F}_n(S)$.

The following theorem proves the correctness and estimates the simulation cost of Algorithm 2. Note that all the upper bound results on simulation costs in this chapter are proved to hold both almost surely and in expectation. We do not differentiate the use of *simulation costs* and *expected simulation costs* in upper bound results.

Theorem 41. *Suppose that Assumptions 5-8 hold. The solution returned by Algorithm 2 satisfies the (ϵ,δ)-PGS guarantee. The simulation cost of Algorithm 2 is at most*

$$O\left[\frac{d}{\epsilon^2}\log\left(\frac{d}{\delta}\right) + d\right] = \tilde{O}\left[\frac{d}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

We note that the simulation cost in the $\tilde{O}$ notation gives the asymptotic simulation cost when δ is small enough. After resolving these two problems, we can first use the SSGD method to find an approximate solution to problem (4.5) and then round the solution to get an approximate solution to problem (4.1). Hence, the focus of the remainder of this section is to provide upper bounds of simulation cost to the SSGD method. The main difficulty of giving sharp upper bounds lies in the fact that the Lovász extension is neither smooth nor strongly-convex. This property of the Lovász extension prohibits the application of Nesterov acceleration and common variance reduction techniques.

Now, we propose the projected and truncated SSGD method for the (ϵ, δ)-PGS guarantee. The orthogonal projection onto the convex hull $\mathrm{conv}(\mathcal{X})$, which is defined as

$$\mathcal{P}_{\mathcal{X}}(x) := \operatorname*{arg\,min}_{y \in \mathrm{conv}(\mathcal{X})} \|y - x\|_2, \quad \forall x \in \mathbb{R}^d,$$

is applied after each iteration to ensure the feasibility of iteration point. Since the convex hull is a convex set, the projection is well-defined. For the case when the feasible set is $\{0,1\}^d$, the projection is given by

$$\mathcal{P}_{\mathcal{X}}(x) := (x \wedge \mathbf{1}) \vee \mathbf{0}, \quad \forall x \in \mathbb{R}^d.$$

In addition to the projection, componentwise truncation of stochastic subgradient is critical in reducing expected simulation costs. The truncation operator with threshold $M > 0$ is defined as

$$\mathcal{T}_M(g) := (g \wedge M\mathbf{1}) \vee (-M\mathbf{1}), \quad \forall g \in \mathbb{R}^d.$$

The pseudo-code of projected and truncated SSGD method is listed in Algorithm 3.

Algorithm 3 Projected and truncated SSGD method for the PGS guarantee

Input: Model $\mathcal{X}, \mathcal{B}_{\Upsilon}, F(x, \xi_x)$, optimality guarantee parameters ϵ, δ, number of iterations T, step size η, truncation threshold M.

Output: An (ϵ, δ)-PGS solution x^* to problem (4.1).

1: Choose an initial point $x^0 \in \mathcal{X}$.
2: **for** $t = 0, \ldots, T-1$ **do**
3: Generate a stochastic subgradient $\hat{g}^t$ at x^t.
4: Truncate the stochastic subgradient $\tilde{g}^t \leftarrow \mathcal{T}_M(\hat{g}^t)$.
5: Update $x^{t+1} \leftarrow \mathcal{P}_{\mathcal{X}}(x^t - \eta \tilde{g}^t)$.
6: **end for**
7: Compute the averaging point $\bar{x} \leftarrow \left(\sum_{t=0}^{T-1} x^t\right)/T$.
8: Round $\bar{x}$ to an integral point by Algorithm 2.

The analysis of Algorithm 3 fits into the classical convex optimization framework. With a suitable choice of the step size, the truncation threshold and the number of iterations, Algorithm 3 returns an (ϵ, δ)-PGS solution and the expected simulation cost has $O(d^2)$ dependence on the dimension.

Theorem 42. *Suppose that Assumptions 5-8 hold and the subgradient estimator in* (4.4) *is used. If we choose*

$$T = \tilde{\Theta}\left[\frac{d}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right], \quad M = \tilde{\Theta}\left[\sqrt{\log\left(\frac{dT}{\epsilon}\right)}\right], \quad \eta = \frac{1}{M\sqrt{T}},$$

then Algorithm 3 returns an (ϵ, δ)*-PGS solution. Furthermore, we have*

$$T(\epsilon, \delta, \mathcal{MC}) = O\left[\frac{d^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right) + \frac{d^3}{\epsilon^2}\log\left(\frac{d^2}{\epsilon^3}\right) + \frac{d^3L^2}{\epsilon^2}\right] = \tilde{O}\left[\frac{d^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

Although independent of δ, we note that the last two terms in the expected simulation cost may be comparable to the first term when δ is not that small. We can prove that, without the truncation step (i.e., $M = \infty$), the expected simulation becomes

$$\tilde{O}\left[\frac{d^3}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

Hence, the truncation of stochastic subgradient is necessary for reducing the asymptotic expected simulation cost. In addition, we note that the Lipschitz constant L is required in determining the truncation threshold M; see Lemma EC.3 for more details. While the the error of the normal SSGD method only contains the optimization residual and the variance terms, the residual of the truncated SSGD method has an extra bias term. We note that the bias term can be made arbitrarily small with high probability by choosing large enough M and utilizing the tail bound for sub-Gaussian random variables, and therefore the total error can be controlled similarly as the normal SSGD method. By choosing $\epsilon = c/2$, Algorithm 3 returns a (c, δ)-PCS-IZ solution and the expected simulation cost for the PCS-IZ guarantee is

$$T(\delta, \mathcal{MC}_c) = \tilde{O}\left[\frac{d^2}{c^2}\log\left(\frac{1}{\delta}\right)\right].$$

We note that the expected simulation cost for both guarantees does not critically depend on the Lipschitz constant L. As an alternative to estimator (4.4), one may consider generating a stochastic subgradient by randomly choosing a subset of components and only estimating the chosen components of subgradients. However, using this estimator, we cannot achieve better simulation cost and the expected simulation cost may be critically dependent on L.

Before finishing the discussion of stochastic SFM problem, we note that the expected simulation cost in Theorem 42 may be improved if we further assume the stochastic subgradient is bounded almost surely. We provide a detailed analysis in the appendix.

4.4 Simulation-optimization Algorithms and Expected Simulation Costs for the General Case

In this section, we extend to the general $L^\natural$-convex function minimization problem with decision space $[N]^d$ for general large N and d. We design simulation-optimization algorithms

that achieve the PGS guarantee and prove upper bounds on the simulation costs.

As an extension to the methodology in Section 4.3, we first show that the Lovász extension in the neighborhood of each point can be pieced together to form a convex function on $\text{conv}(\mathcal{X}) = [1, N]^d$. We define the local neighborhood of each point $y \in [N-1]^d$ as the hypercube

$$\mathcal{C}_y := y + [0, 1]^d,$$

where the Minkowski sum of a point $y \in \mathbb{R}^d$ and a set $\mathcal{C} \subset \mathbb{R}^d$ is defined as

$$y + \mathcal{C} := \{y + x \mid x \in \mathcal{C}\}.$$

We denote the objective function $f(x)$ restricted to $\mathcal{C}_y \cap \mathcal{X}$ as $f_y(x)$. For point $x \in \mathcal{C}_y$, we denote α_x as a consistent permutation of $x - y$ in $\{0, 1\}^d$, and for each $i \in \{0, 1, \ldots, d\}$, the corresponding i-th neighboring point of x is defined as

$$S^{x,i} := y + \sum_{j=1}^{i} e_{\alpha_x(j)}.$$

By the translation submodularity property of $L^\natural$-convex functions, we know function $f_y(x)$ is a submodular function on $y + \{0, 1\}^d$ and its Lovász extension in $\mathcal{C}_y$ can be calculated as

$$\tilde{f}_y(x) := f\left(S^{x,0}\right) + \sum_{i=1}^{d} \left[f\left(S^{x,i}\right) - f\left(S^{x,i-1}\right)\right] x_{\alpha_x(i)}.$$

Now, we piece together the Lovász extension in each hypercube by defining

$$\tilde{f}(x) := \tilde{f}_y(x), \quad \forall x \in [1, N]^d,\ y \in [N-1]^d \quad \text{s. t. } x \in \mathcal{C}_y. \tag{4.7}$$

The next theorem verifies the well-definedness and the convexity of $\tilde{f}$.

Theorem 43. *The function $\tilde{f}(x)$ in (4.7) is well-defined and is convex on $\mathcal{X}$.*

A numerical verification of the results of Theorem 43 is provided in the appendix. Properties of the Lovász extension in Lemma 24 can be naturally extended to the convex extension $\tilde{f}(x)$.

Lemma 25. *Suppose that Assumptions 5-8 hold. Then, the following properties of $\tilde{f}(x)$ hold.*

- *For any $x \in \mathcal{X}$, it holds $\tilde{f}(x) = f(x)$.*
- *The minimizers of $\tilde{f}$ satisfy $\arg\min_{y \in [1,N]^d} \tilde{f}(y) = \arg\min_{y \in [N]^d} f(y)$.*
- *For a point $x \in \mathcal{C}_y$, a subgradient $g \in \partial \tilde{f}(x)$ is given by*

$$g_{\alpha_x(i)} := f\left(S^{x,i}\right) - f\left(S^{x,i-1}\right), \quad \forall i \in [d]. \tag{4.8}$$

- *Subgradients of function $\tilde{f}(x)$ satisfy*

$$\|g\|_1 \leq 3L/2, \quad \forall g \in \partial \tilde{f}(x),\ x \in \mathcal{X}.$$

Similar to the proof of Theorem 43, the subgradient given in (4.8) does not depend on the hypercube and the consistent permutation we choose. The subgradient estimator defined in (4.4) is still valid in the general case. Thus, changing the orthogonal projection to be

$$\mathcal{P}_{\mathcal{X}}(x) := (x \wedge N\mathbf{1}) \vee \mathbf{1}, \quad \forall x \in \mathbb{R}^d,$$

Algorithm 3 can be applied to the general case and we get the counterpart to Theorem 42.

Theorem 44. *Suppose that Assumptions 5-8 hold and the subgradient estimator in* (4.4) *is used. If we choose*

$$T = \tilde{\Theta}\left[\frac{dN^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right], \quad M = \tilde{\Theta}\left[\sqrt{\log\left(\frac{dNT}{\epsilon}\right)}\right], \quad \eta = \frac{N}{M\sqrt{T}},$$

then Algorithm 3 returns an (ϵ, δ)-PGS solution. Furthermore, we have

$$T(\epsilon, \delta, \mathcal{MC}) = O\left[\frac{d^2N^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right) + \frac{d^3N^2}{\epsilon^2}\log\left(\frac{d^2N}{\epsilon^3}\right) + \frac{d^3N^2L^2}{\epsilon^2}\right] = \tilde{O}\left[\frac{d^2N^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

We reiterate that the results also apply to the general $L^\natural$-convex set case by replacing the scale N with $\max_{x,x' \in \mathcal{X}} \|x - x'\|_\infty$. Similarly, the expected simulation costs in Theorem 44 can be improved under the bounded stochastic subgradient assumption and we defer the discussion to the appendix. For the PCS-IZ guarantee, we can choose $\epsilon = c/2$ and Algorithm 3 will return a (c, δ)-PCS-IZ solution. Hence, the above asymptotic simulation costs also hold for the PCS-IZ guarantee. However, with the priori knowledge about the indifference zone parameter, we can design an acceleration scheme similar to [239], which is based on the Weak Sharp Minimum condition. The acceleration scheme reduces the dependence on N from $O(N^2)$ to $O(\log(N))$ and we provide details in the appendix.

4.5 Lower Bound on Expected Simulation Cost

We derive lower bounds on the expected simulation cost for any simulation-optimization algorithm that can achieve the PGS guarantee. In this section, we prove that the expected simulation cost is lower bounded by $O(d\epsilon^{-2}\log(1/\delta))$. We acknowledge that the lower bound may not be tight, but the proven lower bound results suggest the limits for all simulation-optimization algorithms to achieve the PGS guarantee for general simulation optimization problems with convex structure.

To prove lower bounds, basically, we construct several convex models that are "similar" to each other but they have distinct optimal solutions, where the difference between two

models is characterized by the Kullback–Leibler (KL) divergence between their distributions. Hence, any simulation-optimization algorithms need a large number of simulation runs to differentiate these models. More rigorously, the information-theoretical inequality in [128] provides a systematic way to prove lower bounds of zeroth-order algorithms. Given a zeroth-order algorithm and a model M, we denote $N_x(\tau)$ as the number of times that $F(x, \xi_x)$ is sampled when the algorithm terminates, where τ is the stopping time of the algorithm. Then, it follows from the definition that

$$\mathbb{E}_M[\tau] = \sum_{x \in \mathcal{X}} \mathbb{E}_M \left[N_x(\tau) \right],$$

where $\mathbb{E}_M$ is the expectation when the model M is given. Similarly, we can define $\mathbb{P}_M$ as the probability when the model M is given. The following lemma was proved in [128] and is the major tool for deriving lower bounds in this chapter.

Lemma 26. *For any two models M_1, M_2 and any event $\mathcal{E} \in \mathcal{F}_\tau$, we have*

$$\sum_{x \in \mathcal{X}} \mathbb{E}_{M_1} \left[N_x(\tau) \right] \mathrm{KL}(\nu_{1,x}, \nu_{2,x}) \geq d(\mathbb{P}_{M_1}(\mathcal{E}), \mathbb{P}_{M_2}(\mathcal{E})), \tag{4.9}$$

where $d(x, y) := x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$, $\mathrm{KL}(\cdot, \cdot)$ is the KL divergence and $\nu_{k,x}$ is the distribution of model M_k at point x for $k = 1, 2$.

The information-theoretical inequality (4.9) is our major tool for deriving lower bounds. We first reduce the construction of $L^\natural$-convex functions to the construction of submodular functions. Then, using the family of submodular functions defined in [90], we can construct $d + 1$ submodular functions that have different optimal solutions and have the same value except on $d + 1$ potential solutions. Hence, the algorithm has to simulate enough samples on the $d + 1$ potential solutions to decide the optimal solution and the simulation cost is proportional to d.

Theorem 45. *Suppose that Assumptions 5-7 hold. We have*

$$T(\epsilon, \delta, \mathcal{MC}) \geq \Theta \left[\frac{d}{\epsilon^2} \log \left(\frac{1}{\delta} \right) \right].$$

We note that the lower bound above is also true when Assumption 8 holds with $L \geq \epsilon/N$. In addition, a similar construction to Theorem 45 leads to a lower bound on the expected simulation cost for the PCS-IZ guarantee.

Theorem 46. *Suppose that Assumptions 5-7 hold. We have*

$$T(\delta, \mathcal{MC}_c) \geq \Theta \left[\frac{d}{c^2} \log \left(\frac{1}{\delta} \right) \right].$$

4.6 Simulation-optimization Algorithms with Biased Gradient Information

In large-scale discrete optimization via simulation, during a simulation run for performance evaluation at a given value of the d-dimensional decision variable x, it is sometimes possible that the neighboring values of decision variables (those very close to x) can be evaluated simultaneously within the same simulation run for x at marginal costs. See [121] and [120] for a bike sharing discrete optimization via simulation problem that adopts this feature. When the decision variable x is in continuous space, this simultaneous simulation approach is called the *Infinitesimal Perturbation Analysis* (IPA) or the *Forward/Backward Automatic Differentiation*, in which a gradient estimator at x can be obtained within the same simulation run for evaluation of x. In continuous decision space, such gradient estimators can be unbiased under Lipschitz continuity regularity conditions, though no general guarantees on unbiasedness exist when continuity fails. In contrast, for discrete optimization via simulation problems, in particular for those where discrete decision variables do not easily relax to continuous variables, the difference of function value on x and function value on the neighboring points of x can be viewed as an approximate directional derivative. This approximate gradient information (i.e., the difference of objective function values) is very difficult, if not impossible, to estimate without bias using only a single simulation run. In general, the system dynamics and logic are different for two different discrete decision variables even when they differ in only one coordinate. Therefore, in the simulation run for some choice of the decision variable x, the simultaneous evaluation for neighboring choices of the decision variable may incur a bias. See Chapter 4 of [120] for a detailed discussion in the bike-sharing optimization as an example. Despite the bias, the availability of such gradient information can potentially be beneficial when d is large, because only one simulation run is needed to evaluate a biased version of a d-dimension gradient estimator. The gradient estimator can be usually obtained at a marginal cost that does not depend on the dimension d, which is much lower than the cost of constructing a finite difference gradient estimator.

In this section, we provide simulation-optimization algorithms to achieve the PGS guarantee for discrete convex simulation optimization problems, when the gradient information is available (but possibly biased) within a simulation run at a cost that does not depend on dimension. We call this class of simulation-optimization algorithms, which utilize the available gradient information, *first-order algorithms*. We will show how the use of the gradient information reduces the expected simulation cost and how the bias existing in the gradient information affects the results. We first rigorously define the gradient information that can be obtained in simulation with different choices of decision variables. The gradient information that can be obtained within one simulation run is generally biased and has correlated components. The existence of correlation may increase the difficulty of analyzing the performance of simulation-optimization algorithms. Moreover, the correlation could contribute to a larger overall variance of the norm of the subgradient estimator, which may adversely affect the simulation-optimization algorithm.

On the bias side, if the bias in the subgradient estimator can be arbitrarily large, the sign of a subgradient estimator can even be flipped (see an example in [71]). In those cases, there is in general no guarantee for gradient-based algorithms even for convex problems. Examples in [6] also show that the biased gradient-based methods may not converge to the optimum or even dramatically diverge. To circumvent this challenge, some existing works on biased gradient-based methods require the objective function to be smooth and have additional benign geometrical properties, e.g., the strongly convexity or the Polyak-Łojasiewicz (PL) condition [60, 41, 6, 109]. Since the convex extension of a general $L^\natural$-convex function is a piecewise linear function and is neither smooth nor strongly convex, these methods which require benign structure cannot be applied to our case.

In the special case when the biased subgradient estimator of $f(x)$ is the unbiased subgradient estimator of another function $h(x)$, we can view $h(x)$ as a perturbed version of $f(x)$. We define the Lovász extension of $h(x)$ in the same way and equivalently minimize the Lovász extension via the SSGD method. However, since function $h(x)$ may not be $L^\natural$-convex, its Lovász extension is a non-smooth and non-convex function and there is no guarantee on the complexity of the SSGD method [58, 56]. In [246], the authors proposed a stochastic normalized subgradient descent method with sample complexity $O(\epsilon^{-4})$ for finding a point with a subgradient with norm smaller than ϵ. Under the assumption of weak convexity, algorithms with sample complexity of $O(\epsilon^{-2})$ have been proved in [57, 249, 163]. On the other hand, to achieve the same sample complexity as convex optimization, it is proved that the perturbation $h(x) - f(x)$ should has order $O(1/d)$ for all feasible x [18, 126, 164]. However, the existence of the perturbed function $h(x)$ does not always hold and therefore we may not use the above methods.

The above discussion shows that some regularity assumptions on the bias are necessary for the applicability of gradient information to achieve the PGS guarantee. Now, we describe a formal definition of biased subgradient estimator along with the assumption on bias. The key in the assumption is to regulate the relative magnitude of the bias, so that in expectation the bias does not flip the sign of any components of the true subgradient at any choices of decision variables, i.e., the magnitude of any component of the bias is bounded by the magnitude of this component of the true subgradient. The use of common random variables whenever available in general can contribute to the validity of this assumption. As a comparison, [71] regulate the norm of the bias to provide guarantees for continuous stochastic optimization problems. To prepare notation, the set of neighboring choices of decision variable $x \in \mathcal{X}$ is defined as

$$\mathcal{N}_x := \{x \pm e_{\mathcal{S}} : \mathcal{S} \subset [d]\} \cap \mathcal{X}.$$

where e_i is the i-th unit vector of $\mathbb{R}^d$ and $e_{\mathcal{S}}$ is the indicator vector $\sum_{i\in\mathcal{S}} e_i$. The following assumption describes the case that allows the gradient information to have bias and correlation among different directions.

Assumption 9 (Subgradient estimator with bias and correlation.)**.** Given the bias ratio $a \in [0, 1)$, for any point $x \in \mathcal{X}$, there exists a deterministic function $H_x(y, \eta_y) : \mathcal{N}_x \times \mathrm{Z} \mapsto \mathbb{R}$

such that

$$|\mathbb{E}[H_x(y,\eta_y)] - [f(y) - f(x)]| \leq a \cdot |f(y) - f(x)|, \quad \forall y \in \mathcal{N}_x, \tag{4.10}$$

where $\mathcal{N}_x$ is the set of neighboring points of x and $(\mathrm{Z}, \mathcal{B}_\mathrm{Z})$ is a proper space that summarizes the randomness of $G(x,\eta_x)$. Moreover, the marginal distribution for each $H_x(y,\eta_y)$ is sub-Gaussian with parameter $\tilde{\sigma}^2$ and the simulation cost of evaluating $H_x(y,\eta_y)$ *for all* $y \in \mathcal{N}_x$ is at most γ multiplying the simulation cost of evaluating $F(x,\xi_x)$.

Under Assumption 9, $\mathbb{E}[H_x(y,\eta_y)]$ has the same sign as $f(y) - f(x)$ and, using Theorem 7.14 in [168], point $x \in \mathcal{X}$ is a minimizer of $f(x)$ if and only if

$$\mathbb{E}[H_x(y,\eta_y)] \geq 0, \quad \forall y \in \mathcal{N}_x.$$

Therefore, it is still possible to check the global optimality by merely comparing the differences with neighboring points. A similar optimality condition can be established for the PGS guarantee. Using the above observation, we give an algorithm for the PGS guarantee using the biased subgradient estimator $H_x(y,\eta_y)$. The algorithm can be seen as a stochastic version of the steepest descent method in [168] and is listed in Algorithm 4.

Algorithm 4 Adaptive stochastic steepest descent method for the PGS guarantee

Input: Model $\mathcal{X}, \mathcal{B}_\mathrm{Y}, F(x,\xi_x)$, optimality guarantee parameters ϵ, δ, biased subgradient estimator $H_x(y,\eta_y)$, bias ratio a.

Output: An (ϵ,δ)-PGS solution x^* to problem (4.1).

Choose the initial point $x^{0,0} \leftarrow (N/2,\ldots,N/2)^T$.
Set the initial confidence half-width threshold $h_0 \leftarrow (1-a)L/12$.
Set maximal number of epochs $E \leftarrow \lceil \log_2(NL/\epsilon) \rceil$.
Set maximal number of iterations $T \leftarrow (1+a)/(1-a) \cdot 6N$.
for $e = 0, 1, \ldots, E-1$ **do**
  for $t = 0, 1, \ldots, T-1$ **do**
    repeat simulate $H_{x^{e,t}}(y,\eta_y)$ for all $y \in \mathcal{N}_{x^{e,t}}$
      Compute the empirical mean $\hat{H}_{x^{e,t}}(y)$ using all simulated samples for all $y \in \mathcal{N}_{x^{e,t}}$.
      Compute the $1 - \delta/(ET)$ one-sided confidence interval

$$\left[\hat{H}_{x^{e,t}}(y) - h_y, \infty\right), \quad \forall y \in \mathcal{N}_{x^{e,t}}.$$

    until the confidence half-width $h_y \leq h_e$ for all $y \in \mathcal{N}_{x^{e,t}}$
    if $\hat{H}_{x^{e,t}}(y) \leq -2h_e$ for some $y \in \mathcal{N}_{x^{e,t}}$ **then** ▷ This takes 2^{d+1} arithmetic operations.
      Update $x^{e,t+1} \leftarrow y$.
    else if $\hat{H}_{x^{e,t}}(y) > -2h_e$ for all $y \in \mathcal{N}_{x^{e,t}}$ **then**

14: **break**
15: **end if**
16: **end for**
17: Set $x^{e+1,0} \leftarrow x^{e,t}$ and $h_{e+1} \leftarrow h_e/2$.
18: **end for**
19: Return $x^{E,0}$.

The following theorem verifies the correctness of Algorithm 4 and estimates its simulation cost.

Theorem 47. *Suppose that Assumptions 5-9 hold. Algorithm 4 returns an (ϵ, δ)-PGS solution and we have*

$$T(\epsilon, \delta, \mathcal{MC}) = O\left[\frac{\gamma N^3}{(1-a)^3\epsilon^2}\log\left(\frac{1}{\delta}\right) + \frac{\gamma N}{1-a}\log\left(\frac{N}{\epsilon}\right)\right] = \tilde{O}\left[\frac{\gamma N^3}{(1-a)^3\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

We note that Algorithm 4 requires 2^{d+1} arithmetic operations for each iteration. Even though they share the same simulation logic, the memory cost may not be negligible, which may also incur additional computational cost of keeping track of large-scale vectors. There is then a trade-off between simulation costs and memory in general, which we do not exactly model in this chapter. To avoid exponentially many arithmetic operations and memory occupation in the steepest descent method, the comparison-based zeroth-order method in [2] can be extended to our case and reduce the number of arithmetic operations to a polynomial in d. In addition, we may consider using the following stochastic coordinate steepest descent method as a simple and fast implementation of Algorithms 4 and 6. Let x^t be the current iteration point and we update by two steps.

1. Simulate $H_{x^t}(y, \eta_y)$ for all $y \in \{x^t \pm e_i,\ i \in [d]\}$ until the confidence interval is small enough.

2. If for some $y \in \{x^t \pm e_i,\ i \in [d]\}$, we know $f(y) < f(x^t)$ holds with high probability, then update $x^{t+1} = y$; otherwise if $f(y) \geq f(x^t) - O(\epsilon)$ holds for all $y \in \{x^t \pm e_i,\ i \in [d]\}$ with high probability, then we terminate the iteration and return x^t as the solution.

We can see that the number of arithmetic operations for each iteration is $O(d)$. Moreover, an analogous method utilizing $O(d)$ neighboring points in constructing gradient is shown to have good empirical performance in [120]. However, theoretically, without extra assumptions on the problem structure, the stopping criterion $f(y) \geq f(x) - O(\epsilon)$ for all $y \in \{x^t \pm e_i,\ i \in [d]\}$ cannot ensure the approximate optimality of solution x. We give a counterexample to show that $f(y) \geq f(x)$ for all $y \in \{x^t \pm e_i,\ i \in [d]\}$ cannot ensure the optimality of solution x.

Example 8. We consider the case when $d = 2$ and $N = 3$. Define the objective function as

$$f(x, y) := 2|x - y| - |x + y - 2|, \quad \forall (x, y) \in \{1, 2, 3\}^2.$$

We can verify that $f(x, y)$ is a $L^\natural$-convex function and its minimizer is $(3, 3)$ with optimal value -4. Considering point $(2, 2)$, we can calculate that

$$f(2,2) = -2,\ f(1,2) = 1,\ f(3,2) = -1,\ f(2,1) = 1,\ f(2,3) = -1.$$

Hence, the guarantee is satisfied at $(2, 2)$ but the point is not a minimizer of $f(x)$.

Finally, in the case when the indifference zone parameter c is known, we can prove that choosing $\epsilon = Nc$ is enough for the (c, δ)-PCS-IZ guarantee. We provide the algorithm and its complexity analysis in the appendix.

4.7 Numerical Experiments

In this section, we implement our proposed simulation-optimization algorithms that are guaranteed to find high-confidence high-precision PGS solutions. We first consider the optimal allocation problem of a queueing system, where we show the advantage of using the truncation step. Next, we consider an artificially constructed $L^\natural$-convex function, where more details about the objective function landscape are available for the evaluation of the performance.

Optimal Allocation Problem

In the optimal allocation problem, we consider the 24-hour operation of a service system with a single stream of incoming customers. The customers arrive according to a a doubly stochastic non-homogeneous Poisson process with intensity function

$$\Lambda(t) := 0.5\lambda N \cdot (1 - |t - 12|/12), \quad \forall t \in [0, 24],$$

where λ is a positive constant and N is a positive integer. Each customer requests a service with service time independent and identically distributed according to the log-normal distribution with mean $1/\lambda$ and variance 0.1. We divide the 24-hours operation into d time slots with length $24/d$ for some positive integer d. For the i-th time slot, there are $x_i \in [N]$ of homogeneous servers that work independently in parallel and the number of servers cannot be changed during the slot. Assume that the system operates based on a first-come-first-serve routine, with unlimited waiting room in each queue, and that customers never abandon.

The decision maker's objective is to select the staffing level $x := (x_1, \ldots, x_d)$ such that the total waiting time of all customers is minimized. Namely, letting $f(x)$ be the expected total waiting time under the staffing plan x, then the optimization problem can be written as

$$\min_{x \in [N]^d} f(x). \tag{4.11}$$

It has been proved in [10] that the function $f(\cdot)$ is multimodular. We define the linear transformation

$$g(y) := (y_1, y_2 - y_1, \ldots, y_d - y_{d-1}), \quad \forall y \in \mathbb{R}^d.$$

Then, [168] has proved that

$$h(y) := f \circ g(y) = f(y_1, y_2 - y_1, \ldots, y_d - y_{d-1})$$

is a $L^\natural$-convex function on the $L^\natural$-convex set

$$\mathcal{Y} := \{y \in [Nd]^d \mid y_1 \in [N],\ y_{i+1} - y_i \in [N],\ i = 1, \ldots, N-1\}.$$

The optimization problem (4.11) has the trivial solution $x_1 = \cdots = x_d = N$. However, in reality, it is also necessary to keeping the staffing cost low. There are two different approaches to achieve this goal. First, we can constrain the total number of servers $\sum_{i=1}^d x_i$ to be at most K, where $K \leq Nd$ is a positive integer and the optimization problem can be written as

$$\min_{y \in \mathcal{Y}} h(y) \quad \text{s. t. } y_d \leq K. \tag{4.12}$$

On the other hand, we can add a regularization term $R(x_1, \ldots, x_d) := C/d \cdot \sum_{i=1}^d x_i = C/d \cdot y_d$ to the objective function, where $C > 0$ is a constant. The optimization problem can be written as

$$\min_{y \in \mathcal{Y}} h(y) + C/d \cdot y_d. \tag{4.13}$$

We refer problems (4.12) and (4.13) as the constrained and the regularized problems, respectively. Our algorithms can be extended to this case by considering the Lovász extension $\tilde{h}(y)$ on the set

$$\tilde{\mathcal{Y}} := \{y \in [1, Nd]^d \mid y_1 \in [1, N],\ y_{i+1} - y_i \in [1, N],\ i = 1, \ldots, N-1\}.$$

We compare the performance of the projected SSGD method (Algorithm 3) with truncation ($M < \infty$) and without truncation ($M = \infty$) on both problems. In the truncation-free case, the step size is chosen to be $\eta = O(N\sqrt{d/T})$. We first fix the dimension (number of time slots) to be $d = 4$ and compare the performance when the scale $N \in \{10, 20, 30, 40, 50\}$, and we then fix the scale to be $N = 10$ and compare the performance when the dimension $d \in \{4, 8, 12, 16, 20, 24\}$. The parameters of the problem are chosen as $\lambda = 4$, $C = 50$ and $K = \lfloor Nd/3 \rfloor$, and the optimality guarantee parameters are $\epsilon = N/2$ and $\delta = 10^{-6}$. For each problem setup, we average the simulation costs of 10 independent implementations to estimate the expected simulation cost. Moreover, early stopping is used to terminate algorithms early when little progress is made after some iterations. More concretely, we maintain the empirical mean of stochastic objective function values up to the current iteration and terminate the algorithm if the empirical mean does not decrease by $\epsilon/\sqrt{N}$ after $O(d\epsilon^{-2}\log(1/\delta))$ consecutive iterations.

Params.		Regularized				Constrained			
		Truncated		Not truncated		Truncated		Not truncated	
d	N	Cost	Obj.	Cost	Obj.	Cost	Obj.	Cost	Obj.
4	10	2.99e5	2.10e2	6.56e5	2.11e2	3.00e5	4.76e1	4.99e5	4.97e1
4	20	1.21e5	3.53e2	2.61e5	3.53e2	1.14e5	5.23e1	1.77e5	5.38e1
4	30	8.85e4	4.75e2	1.68e5	4.76e2	7.38e4	5.24e1	1.23e5	5.21e1
4	40	6.25e4	5.91e2	1.34e5	6.07e2	5.28e4	5.31e1	9.24e4	5.28e1
4	50	5.34e4	7.07e2	1.08e5	7.07e2	4.66e4	5.64e1	6.61e4	5.51e1
8	10	1.19e6	1.75e2	3.80e6	1.76e2	1.20e6	3.11e1	2.23e6	3.02e1
12	10	2.68e6	1.59e2	9.48e6	1.59e2	2.69e6	1.87e1	5.36e6	1.86e1
16	10	6.35e6	1.49e2	1.31e7	1.50e2	4.78e6	1.49e1	1.08e7	1.41e1
20	10	9.91e6	1.43e2	2.09e7	1.46e2	9.43e6	1.17e1	1.70e7	1.28e1
24	10	1.50e7	1.35e2	3.09e7	1.41e2	1.36e7	9.43e0	2.10e7	1.17e1

Table 4.7.1: Simulation costs and objective function values of Algorithm 3 on the optimal allocation problem.

We first implement both algorithms on the trivial problem (4.11) for 10 times. Since the optimal solution is known, it is possible to verify whether the solutions returned by algorithms are at most ϵ worse than the optimum, at a confidence that is larger than $1-\delta$. In the experiment, we run sufficiently large number of simulation replications to verify the ϵ-optimality at the selected solution with confidence higher than $1-\delta'$, where $\delta' \ll \delta$.

Next, we consider the performance of algorithms on problems (4.12) and (4.13). We summarize the simulation costs and the objective values in Table 4.7.1. We can see that both algorithms return a similar objective value and the simulation cost grows when d becomes larger. The growth rate is approximately quadratic. The simulation cost becomes smaller when N gets larger, since we allow a larger sub-optimality gap ($N/2$) when N is larger. We note that the feasible set of both problems is not a hypercube, and thus the dependence of simulation costs on d and N is not exactly quadratic as indicated by our theory. In addition, we can see that the truncation plays an important role in reducing the simulation cost, especially when the dimension is high.

Separable Convex Function Minimization

We consider the problem of minimizing a stochastic $L^\natural$-convex function whose expectation is a separable convex function parameterized by a vector $c \in \mathbb{R}^d$ and the optimal solution

$x^* \in \mathbb{R}^d$:

$$f_{c,x^*}(x) := \sum_{i=1}^{d} c_i g(x_i^*; x_i),$$

where $c_i \in [0.75, 1.25]$, $x_i^* \in \{1, \ldots, \lfloor 0.3N \rfloor\}$ for all $i \in [d]$ and

$$g(y^*; y) := \begin{cases} \sqrt{\frac{y^*}{y}} - 1 & \text{if } y \leq y^* \\ \sqrt{\frac{N+1-y^*}{N+1-y}} - 1 & \text{if } y > y^* \end{cases}, \quad \forall y, y^* \in [N].$$

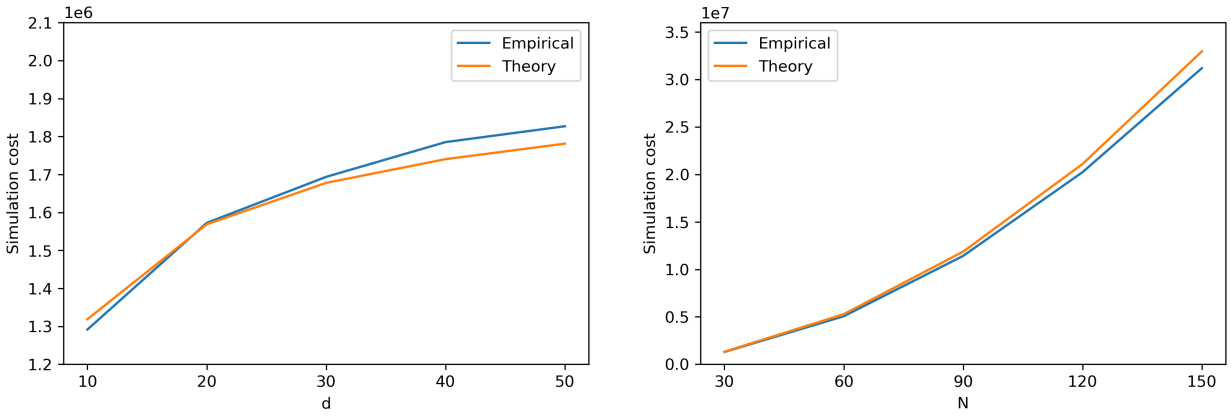
It is observed that the function $f_{c,x^*}(x)$ is a separable convex functions and therefore is $L^\natural$-convex. Moreover, the function $f_{c,x^*}(x)$ has the optimum x^* associated with the optimal value 0. For stochastic evaluations, we add Gaussian noise with mean 0 and variance 1 to each point $x \in \mathcal{X}$. Due to the $O[(y^*)^{-3/2}]$ growth rate, the landscape of $g(y^*; y)$ is flat around x^*. The advantage of this numerical example is that the expected objective function has a closed form, and we are able to verify the ϵ-optimality of the solutions returned by the proposed algorithms.

To analyze the effect of the dimension and the scale on the expected simulation cost, we first fix $d = 10$ and compare the performance when $N = 30, 60, 90, 120, 150$; then we fix $N = 30$ and compare the performance when $d = 10, 20, 30, 40, 50$. The optimality guarantee parameters are chosen as $\epsilon = (d!)^{1/d}/5$ and $\delta = 10^{-6}$. In the one-dimensional case, this choice of ϵ ensures that the ϵ-sub-level set of the objective function approximately covers $N/4$ choices of decisions. We note that this choice of ϵ is only for comparisons between different (d, N) and our results can be extended to other choices of ϵ. We compute the average simulation cost of 100 independently generated models to estimate the expected simulation cost. Similar early stopping criteria are also applied.

Figure 4.7.1 shows the results of fixed d and fixed N. Since the choice of ϵ is dependent on d, the relation between the simulation costs and d is not clear. Therefore, we compare the simulation costs to the theoretical bound (up to a constant)

$$T(d, N) := N^2 d^2 \epsilon^{-2} \log 1/\delta.$$

More specifically, we compare the simulation costs to $0.87T(d, N)$ in this experiment, which corresponds to the "Theory" curve in the figure. We can observe from the plotting that the growth of simulation costs matches our theory very well. This implies that our estimation on the performance of the truncated SSGD algorithm is tight on this example. Moreover, the optimality gap between the returned solutions and the optimal solution is smaller than ϵ for all experiments, which implies that the algorithm succeeds with high probability.



(a) (b)

Figure 4.7.1: The expected simulation costs of the separable convex minimization problem. **(a)** Expected simulation costs with $N = 30$. **(b)** Expected simulation costs with $d = 10$.

Appendix

4.A More Numerical Experiments

Illustrations of the Lovász Extension

In this subsection, we show the Lovász extension of a two-dimensional function on $[3]^2 = \{1, 2, 3\}^2$. We consider the quadratic function

$$f(x) := x^T \begin{bmatrix} 0.101 & -0.068 \\ -0.068 & 0.146 \end{bmatrix} x, \quad \forall x \in \mathbb{R}^2.$$

By the results in [168, Section 7.3], we know the function $f(\cdot)$ is a $L^\natural$-convex function. We compare the landscapes of the original objective and the Lovász extension in Figure 4.A.1. We can see that the Lovász extension is a piecewise linear and convex function, which is consistent with the results in Section 4.4 and [168].

4.B Proofs in Section 4.3

Proof of Theorem 41

Proof of Theorem 41. We denote the optimal value of $f(x)$ as f^*. Since point $\bar{x}$ satisfies the $(\epsilon/2, \delta/2)$-PGS guarantee, we have

$$\tilde{f}(\bar{x}) - f^* \leq \epsilon/2$$

holds with probability at least $1 - \delta/2$. We assume this event happens in the following of this proof. Let $S^0, S^1, \ldots, S^d$ be the neighboring points of $\bar{x}$. Using the expression of the Lovász extension in (4.6), we know there exists an $\epsilon/2$-optimal solution among $S^0, S^1, \ldots, S^d$. We denote the $\epsilon/2$-optimal solution and the solution returned by Algorithm 2 as S^* and $\hat{S}$, respectively. By the definition of confidence intervals, we know

$$\left|\hat{F}_n(S_i) - f(S_i)\right| \leq \epsilon/4, \quad \forall i \in \{0, \ldots, d\}, \quad \left|\hat{F}_n(\hat{S}) - f(\hat{S})\right| \leq \epsilon/4$$

holds uniformly with probability at least $1 - \delta/2$. Under this event, we know

$$f(\hat{S}) - f^* \leq \hat{F}_n(\hat{S}) - f^* + \epsilon/4 \leq \hat{F}_n(S^*) - f^* + \epsilon/4 \leq f(S^*) - f^* + \epsilon/2 \leq \epsilon,$$

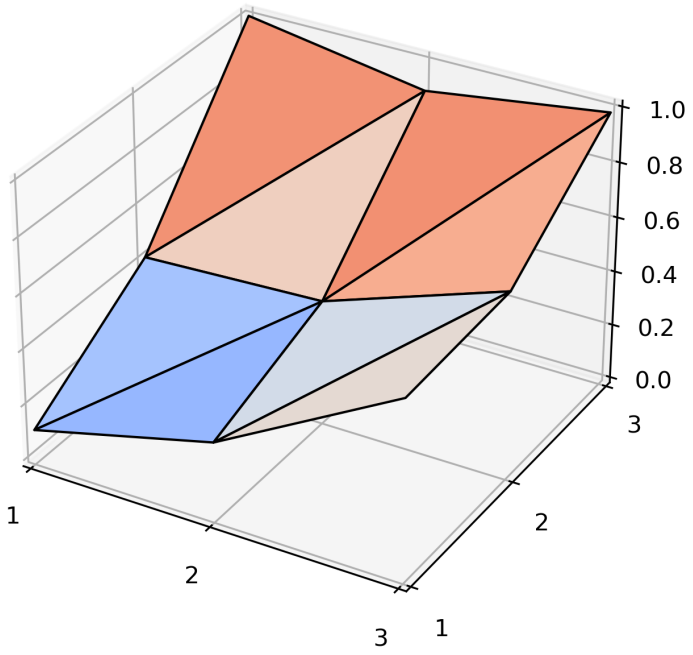


Figure 4.A.1: The Lovász extension of the objective function.

which implies that $x^* \in \mathcal{X}$ is an ϵ-optimal solution and the probability is at least $1 - \delta/2 - \delta/2 = 1 - \delta$. Hence, we know x^* is an (ϵ, δ)-PGS solution to problem (4.1).

Now, we estimate the simulation cost of Algorithm 2. By Hoeffding bound, simulating

$$\frac{32}{\epsilon^2} \log\left(\frac{8d}{\delta}\right)$$

times on each neighboring point is enough to achieve $1 - \delta/(4d)$ confidence half-width $\epsilon/4$. Hence, the simulation cost of Algorithm 2 is at most

$$\frac{32(d+1)}{\epsilon^2} \log\left(\frac{8d}{\delta}\right) = O\left[\frac{d}{\epsilon^2} \log\left(\frac{d}{\delta}\right)\right] = \tilde{O}\left[\frac{d}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right].$$

□

Proof of Theorem 42

The following Azuma's inequality for martingales with sub-Gaussian tails plays as a major role for deriving high-probability bounds, i.e., the number of required samples to ensure the algorithms succeed with high probability.

Lemma 27 (Azuma's inequality for sub-Gaussian tails [199])**.** *Let $X_0, \ldots, X_{T-1}$ be a martingale difference sequence. Suppose there exist constants $b_1 \geq 1, b_2 > 0$ such that, for any $t \in \{0, \ldots, T-1\}$,*

$$\mathbb{P}(|X_t| \geq a \mid X_1, \ldots, X_{t-1}) \leq 2b_1 \exp(-b_2 a^2), \quad \forall a \geq 0. \tag{4.14}$$

Then for any $\delta > 0$, it holds with probability at least $1 - \delta$ that

$$\frac{1}{T} \sum_{t=0}^{T-1} X_t \leq \sqrt{\frac{28 b_1}{b_2 T} \log\left(\frac{1}{\delta}\right)}.$$

Since the stochastic subgradient $\hat{g}^t$ is truncated, the stochastic subgradient used for updating, namely $\tilde{g}^t$, is not unbiased. We define the bias at each step as

$$b_t := \mathbb{E}\left[\tilde{g}^t \mid x^0, x^1, \ldots, x^t\right] - g^t, \quad \forall t \in \{0, 1, \ldots, T-1\}.$$

First, we bound the ℓ_1-norm of the bias.

Lemma 28. *Suppose that Assumptions 5-8 hold. If we have*

$$M \geq 2\sigma \cdot \sqrt{\log\left(\frac{4\sigma d T}{\epsilon}\right)} = \Theta\left[\sqrt{\log\left(\frac{dT}{\epsilon}\right)}\right], \quad T \geq \frac{2\epsilon}{\sigma},$$

then it holds

$$\|b^t\|_1 \leq \frac{\epsilon}{2T}, \quad \forall t \in \{0, 1, \ldots, T-1\}.$$

Proof. Let α_t be a consistent permutation of x^t and $S^{t,i}$ be the corresponding i-th neighboring points. We only need to prove

$$\left|b^t_{\alpha_t(i)}\right| \leq \frac{\epsilon}{2dT}, \quad \forall i \in [d].$$

We define two random variables

$$Y_1 := F\left(S^{t,i}, \xi_i^1\right) - f\left(S^{t,i}\right), \quad Y_2 := F\left(S^{t,i-1}, \xi_{i-1}^2\right) - f\left(S^{t,i-1}\right).$$

By Assumption 5, both Y_1 and Y_2 are independent and sub-Gaussian with parameter σ^2. Hence, we know

$$\begin{aligned}
b^t_{\alpha_t(i)} &= \mathbb{E}\left[\tilde{g}^t_{\alpha_t(i)} - g^t_{\alpha_t(i)}\right] \\
&= \mathbb{E}\left[(Y_1 + Y_2) \cdot \mathbf{1}_{-M \leq Y_1 + Y_2 \leq M}\right] + \mathbb{E}\left[M \cdot \mathbf{1}_{Y_1 + Y_2 > M}\right] + \mathbb{E}\left[-M \cdot \mathbf{1}_{Y_1 + Y_2 < -M}\right] \\
&= \mathbb{E}\left[(M - Y_1 - Y_2) \cdot \mathbf{1}_{Y_1 + Y_2 > M}\right] + \mathbb{E}\left[-(M + Y_1 + Y_2) \cdot \mathbf{1}_{Y_1 + Y_2 < -M}\right],
\end{aligned}$$

where the second step is from $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = 0$. Taking the absolute value on both sides, we get

$$\begin{aligned}\left|b^t_{\alpha_t(i)}\right| &\leq \mathbb{E}\left[(Y_1 + Y_2 - M) \cdot \mathbf{1}_{Y_1+Y_2>M}\right] + \mathbb{E}\left[-(M + Y_1 + Y_2) \cdot \mathbf{1}_{Y_1+Y_2<-M}\right] \\ &= \mathbb{E}\left[(Y - M) \cdot \mathbf{1}_{Y>M}\right] + \mathbb{E}\left[-(Y + M) \cdot \mathbf{1}_{Y<-M}\right],\end{aligned} \tag{4.15}$$

where we define the random variable $Y := Y_1 + Y_2$. Since Y_1, Y_2 are independent, random variable Y is sub-Gaussian with parameter $2\sigma^2$. Let $F(y) := \mathbb{P}[Y \leq y]$ be the distribution function of Y. Then, we have

$$\mathbb{E}\left[(Y - M) \cdot \mathbf{1}_{Y>M}\right] = \int_M^\infty (y - M)\, dF(y) = \int_M^\infty (1 - F(y))\, dy. \tag{4.16}$$

By the Hoeffding bound, we know

$$1 - F(y) = \mathbb{P}[Y > y] \leq \exp\left(-y^2/4\sigma^2\right), \quad \forall y \geq 0.$$

Using the upper bound for Q-function in [25], it holds that

$$\int_M^\infty 1 - F(y)\, dy \leq \int_M^\infty \exp\left(-y^2/4\sigma^2\right)\, dy \leq \frac{2\sigma^2}{M} \exp\left(-\frac{M^2}{4\sigma^2}\right).$$

By the choice of M, we know

$$M \geq 2\sigma\sqrt{\log(8d)} \geq 2\sigma \quad \text{and} \quad \sigma \exp(-M^2/4\sigma^2) \leq \frac{\epsilon}{4dT}.$$

which implies that

$$\int_M^\infty 1 - F(y)\, dy \leq \frac{2\sigma^2}{M} \exp(-M^2/4\sigma^2) \leq \frac{\epsilon}{4dT}.$$

Substituting the above inequality into (4.16), we have

$$\mathbb{E}\left[(Y - M) \cdot \mathbf{1}_{Y>M}\right] \leq \frac{\epsilon}{4dT}.$$

Considering $-Y$ in the same way, we can prove

$$\mathbb{E}\left[-(Y + M) \cdot \mathbf{1}_{Y<-M}\right] \leq \frac{\epsilon}{4dT}.$$

Substituting the last two estimates into inequality (4.15), we know

$$\left|b^t_{\alpha_t(i)}\right| \leq \frac{\epsilon}{2dT}.$$

□

Next, we show that $\langle g^t + b^t - \tilde{g}^t, x^t - x^* \rangle$ forms a martingale sequence and use Azuma's inequality to bound the deviation, where x^* is a minimizer of $f(x)$.

Lemma 29. *Suppose that Assumptions 5-8 hold and let x^* be a minimizer of $f(x)$. The sequence*

$$X_t := \left\langle g^t + b^t - \tilde{g}^t, x^t - x^* \right\rangle \quad t = 0, 1, \ldots, T-1$$

forms a martingale difference sequence. Furthermore, if we have

$$M = \max\left\{ L, 2\sigma \cdot \sqrt{\log\left(\frac{4\sigma d T}{\epsilon}\right)} \right\} = \tilde{\Theta}\left[\sqrt{\log\left(\frac{dT}{\epsilon}\right)} \right], \quad T \geq \frac{2\epsilon}{\sigma},$$

then it holds

$$\frac{1}{T}\sum_{t=0}^{T-1} X_t \leq \sqrt{\frac{224 d \sigma^2}{T} \log\left(\frac{1}{\delta}\right)}$$

with probability at least $1 - \delta$.

Proof. Let $\mathcal{F}_t$ be the filtration generated by $x_0, x_1, \ldots, x_t$. By the definition of b^t, we know

$$\mathbb{E}\left[g^t + b^t - \tilde{g}^t \mid \mathcal{F}_t \right] = 0,$$

which implies that

$$\mathbb{E}\left[X_t \mid \mathcal{F}_t \right] = \left\langle \mathbb{E}\left[g^t + b^t - \tilde{g}^t \mid \mathcal{F}_t \right], x^t - x^* \right\rangle = 0.$$

Hence, the sequence $\{X_t\}$ is a martingale difference sequence. Next, we estimate the probability $\mathbb{P}[|X_t| \geq a \mid \mathcal{F}_t]$. We have the bound

$$\begin{aligned} |X_t| &= \left| \left\langle g^t + b^t - \tilde{g}^t, x^t - x^* \right\rangle \right| \leq \left\| g^t + b^t - \tilde{g}^t \right\|_1 \left\| x^t - x^* \right\|_\infty \\ &\leq \left\| g^t + b^t - \tilde{g}^t \right\|_1 \leq \left\| g^t - \tilde{g}^t \right\|_1 + \left\| b^t \right\|_1. \end{aligned}$$

Since M satisfies the condition in Lemma 28, we know $\|b^t\|_1 \leq \epsilon/2T$. Recalling Assumption 8, we get $|g_i^t| \leq L$ for all $i \in [d]$. By the truncation rule and the assumption $M \geq L$, we have

$$\left| \tilde{g}_i^t - g_i^t \right| = \left| (\hat{g}_i^t \wedge M) \vee (-M) - g_i^t \right| \leq \left| \hat{g}^t - g^t \right|, \quad \forall i \in [d].$$

Hence, we get

$$|X_t| \leq \frac{\epsilon}{2T} + \left\| \hat{g}^t - g^t \right\|_1. \tag{4.17}$$

Define random variables $Y_i := |\hat{g}_i^t - g_i^t|$ for all $i \in [d]$. By Assumption 5, Y_i is sub-Gaussian with parameter σ^2. Hence, we have

$$Y := \left\| \hat{g}^t - g^t \right\|_1 = \sum_{i=1}^{d} Y_i$$

is sub-Gaussian with parameter $d\sigma^2$. First, we consider the case when $a \geq \epsilon/T$. Using inequality (4.17), it follows that

$$\mathbb{P}\left[|X_t| \geq a \mid \mathcal{F}_\sigma\right] \leq \mathbb{P}\left[\frac{\epsilon}{2T} + Y \geq a\right] \leq \mathbb{P}\left[Y \geq a - \frac{\epsilon}{2T}\right] \leq \mathbb{P}\left[Y \geq \frac{a}{2}\right] \leq 2\exp\left(-\frac{a^2}{8d\sigma^2}\right), \tag{4.18}$$

where the last inequality is from Hoeffding bound. In this case, we know condition (4.14) holds with

$$b_1 = 1, \quad b_2 = \frac{1}{8d\sigma^2}.$$

Now, we consider the case when $a < \epsilon/T$. In this case, by the assumption that $T \geq 2\epsilon/\sigma$, we have

$$2b_1 \exp\left(-b_2 a^2\right) > 2\exp\left(-\frac{1}{8d\sigma^2} \cdot \frac{\epsilon^2}{T^2}\right) \geq 2\exp\left(-\frac{1}{32d}\right) \geq 2\exp\left(-\frac{1}{32}\right) > 1.$$

Hence, it holds

$$\mathbb{P}\left[|X_t| \geq a \mid \mathcal{F}_\sigma\right] \leq 1 < 2b_1 \exp\left(-b_2 a^2\right).$$

Combining with inequality (4.18), we know condition (4.14) holds with b and c defined above. Using Lemma 27, we know

$$\frac{1}{T}\sum_{t=0}^{T-1} X_t \leq \sqrt{\frac{224d\sigma^2}{T}\log\left(\frac{1}{\delta}\right)}$$

holds with probability at least $1-\delta$. □

Then, we prove a lemma similar to the Lemma in [260].

Lemma 30. *Suppose that Assumptions 5-8 hold and let x^* be a minimizer of $f(x)$. If we choose*

$$\eta = \frac{1}{M\sqrt{T}},$$

then we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \langle \tilde{g}^t, x^t - x^* \rangle \leq \frac{dM}{\sqrt{T}}.$$

Proof. We define $\tilde{x}^{t+1} := x^t - \eta\tilde{g}^t$ as the next point before the projection onto $[0,1]^d$. Recalling the non-expansion property of orthogonal projection, we get

$$\begin{aligned}\|x^{t+1} - x^*\|_2^2 &= \|\mathcal{P}_{\mathcal{X}}\left(\tilde{x}^{t+1} - x^*\right)\|_2^2 \leq \|\tilde{x}^{t+1} - x^*\|_2^2 = \|x^t - x^* - \eta\tilde{g}^t\|_2^2 \\ &= \|x^t - x^*\|_2^2 + \eta^2\|\tilde{g}^t\|_2^2 - 2\eta\langle\tilde{g}^t, x^t - x^*\rangle,\end{aligned}$$

and equivalently,

$$\langle \tilde{g}^t, x^t - x^* \rangle = \frac{1}{2\eta} \left[\left\| x^t - x^* \right\|_2^2 - \left\| x^{t+1} - x^* \right\|_2^2 \right] + \frac{\eta}{2} \cdot \left\| \tilde{g}^t \right\|_2^2 .$$

Summing over $t = 0, 1, \ldots, T-1$, we have

$$\begin{aligned} \sum_{t=0}^{T-1} \langle \tilde{g}^t, x^t - x^* \rangle &= \frac{\|x^0 - x^*\|_2^2 - \left\| x^T - x^* \right\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=0}^{T-1} \left\| \tilde{g}^t \right\|_2^2 \\ &\leq \frac{d \|x^0 - x\|_\infty^2}{2\eta} + \frac{\eta}{2} \sum_{t=0}^{T-1} \left\| \tilde{g}^t \right\|_2^2 \leq \frac{d}{2\eta} + \frac{\eta}{2} \sum_{t=0}^{T-1} \left\| \tilde{g}^t \right\|_2^2 . \end{aligned}$$

By the definition of truncation, it follows that $\|\tilde{g}^t\|_2^2 \leq dM^2$. Choosing

$$\eta := \frac{1}{M\sqrt{T}},$$

it follows that

$$\sum_{t=0}^{T-1} \langle \tilde{g}^t, x^t - x^* \rangle \leq \frac{d}{2\eta} + \frac{\eta}{2} \sum_{t=0}^{T-1} \left\| \tilde{g}^t \right\|_2^2 \leq \frac{d}{2\eta} + \frac{\eta T d M^2}{2} = dM\sqrt{T}.$$

□

Finally, using Lemmas 28, 29 and 30, we can finish the proof of Theorem 42.

Proof of Theorem 42. Denote f^* as the optimal value of $\tilde{f}(x)$. Using the convexity of $\tilde{f}(x)$, we know

$$\begin{aligned} \tilde{f}(\bar{x}) - f^* &\leq \frac{1}{T} \sum_{t=0}^{T-1} \left[\tilde{f}(x^t) - f^* \right] \leq \frac{1}{T} \sum_{t=0}^{T-1} \left\langle g^t, x^t - x^* \right\rangle \\ &= \frac{1}{T} \sum_{t=0}^{T-1} \left[\left\langle g^t + b^t - \tilde{g}^t, x^t - x^* \right\rangle + \left\langle \tilde{g}^t, x^t - x^* \right\rangle - \left\langle b^t, x^t - x^* \right\rangle \right] . \end{aligned} \tag{4.19}$$

We choose

$$T := \frac{3584 d \sigma^2}{\epsilon^2} \log\left(\frac{2}{\delta}\right) = \Theta\left[\frac{d}{\epsilon^2} \log\left(\frac{1}{\delta}\right) \right].$$

Recalling Assumption 5, we know δ is small enough and therefore we have the following estimates:

$$L^2 \leq M^2 = \tilde{\Theta}\left[\log\left(\frac{dT}{\epsilon}\right) \right] = \tilde{O}\left[\log\left(\frac{d^2}{\epsilon^3}\right) + \log\log\left(\frac{1}{\delta}\right) \right] \leq \frac{\epsilon^2 T}{64 d^2}, \quad T \geq \max\left\{ \frac{2\epsilon}{\sigma}, 4 \right\}.$$

Hence, the conditions in Lemmas 28 and 29 are satisfied. By Lemma 28, we know

$$-\frac{1}{T}\sum_{t=0}^{T-1}\left\langle b^t, x^t - x^*\right\rangle \leq \frac{1}{T}\sum_{t=0}^{T-1}\left\|b^t\right\|_1\left\|x^t - x^*\right\|_\infty \leq \frac{\epsilon}{2T} \leq \frac{\epsilon}{8}. \tag{4.20}$$

By Lemma 29, it holds

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\langle g^t + b^t - \tilde{g}^t, x^t - x^*\right\rangle \leq \sqrt{\frac{224d\sigma^2}{T}\log\left(\frac{2}{\delta}\right)} \leq \frac{\epsilon}{4} \tag{4.21}$$

with probability at least $1-\delta$, where the last inequality is from our choice of T. By Lemma 30, we know

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\langle \tilde{g}^t, x^t - x^*\right\rangle \leq \frac{dM}{\sqrt{T}} \leq \frac{\epsilon}{8}. \tag{4.22}$$

Substituting inequalities (4.20), (4.21) and (4.22) into inequality (4.19), we get

$$\tilde{f}(\bar{x}) - f^* \leq \frac{\epsilon}{2}$$

holds with probability at least $1-\delta/2$. By the results of Theorem 41, we know Algorithm 3 returns an (ϵ,δ)-PGS solution.

Finally, we estimate the simulation cost of Algorithm 3. For each iteration, we need to generate a stochastic subgradient using (4.4) and the simulation cost is $2d$. Hence, the total simulation cost of all iterations is

$$2d \cdot T = \tilde{\Theta}\left[\frac{d^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

By Theorem 41, the simulation cost of rounding process is at most

$$\tilde{O}\left[\frac{d}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

Thus, we know the total simulation cost of Algorithm 3 is at most

$$\tilde{O}\left[\frac{d^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

□

Analysis of the Bounded Stochastic Subgradient Case

In this subsection, we consider the special case when the stochastic subgradient is assumed to have a bounded ℓ_1-norm.

Assumption 10. There exist a constant G and an unbiased subgradient estimator $\hat{g}$ such that

$$\mathbb{P}\left(\|\hat{g}\|_1 \leq G\right) = 1.$$

Moreover, the simulation cost of generating each $\hat{g}$ is at most β simulations.

We note that G and β may depend on d and N. In the field of stochastic optimization, this assumption is common when analyzing the high-probability convergence of stochastic subgradient methods [98, 239]. We first give examples where Assumption 10 holds.

Example 9. We consider the case when the randomness of each choice of decision variables shares the same measure space, i.e., there exists a measure space $(\mathrm{Z}, \mathcal{B}_{\mathrm{Z}})$ such that ξ_x can be any element in the measure space for all $x \in \mathcal{X}$. Moreover, for any fixed $\xi \in \mathcal{B}$, the function $F(\cdot, \xi)$ is also $L^\natural$-convex (or submodular when $N = 2$) and has ℓ_∞-Lipschitz constant $\tilde{L}$. Then, we consider the subgradient estimator

$$\hat{g}_{\alpha_x(i)} := F\left(S^{x,i}, \xi\right) - F\left(S^{x,i-1}, \xi\right), \quad \forall i \in [d]. \tag{4.23}$$

The simulation cost of estimator (4.23) is $d+1$. In addition, property (v) of Lemma 24 gives

$$\|\hat{g}\|_1 \leq 3\tilde{L}/2.$$

Therefore, in this situation, the Assumption 10 holds with $G = 3\tilde{L}/2$ and $\beta = d + 1$.

When the distribution at each choice of decision variables is the Bernoulli, we show that Assumption 10 also holds.

Example 10. We consider the case when the distribution at each point $x \in \mathcal{X}$ is Bernoulli, namely, we have

$$\mathbb{P}[F(x, \xi_x) = 1] = 1 - \mathbb{P}[F(x, \xi_x) = 0] = f(x) \in [0, 1], \quad \forall x \in \mathcal{X}.$$

We note that the Bernoulli distribution is a special case of sub-Gaussian distributions. In this case, the ℓ_∞-Lipschitz constant is 1 and property (v) in Lemma 24 gives $\|g\|_1 \leq 3/2$ for any subgradient g. We consider the subgradient estimator (4.4). At point x, if index i is chosen, then we know that

$$\|\hat{g}\|_1 = d \cdot \left|F\left(S^{x,i}, \xi_i^1\right) - F\left(S^{x,i-1}, \xi_{i-1}^2\right)\right| \leq d.$$

Hence, Assumption 10 holds with $G = d$ and $\beta = 2$.

Next, we estimate the expected simulation cost of Algorithm 3 under Assumption 10. Since the stochastic subgradient is bounded, the truncation step is unnecessary in Algorithm 3. The simulation cost of Algorithm 3 is estimated in the following theorem. The proof is similar to Lemma 10 in [98] and, since the feasible set is the hypercube $[0,1]^d$, we use ℓ_∞-norm instead of ℓ_2-norm to bound distances between points.

Theorem 48. *Suppose that Assumptions 5-8 and 10 hold. If we skip the truncation step in Algorithm 3 (i.e., set $M = \infty$) and choose*

$$T = \tilde{\Theta}\left[\frac{(L+G)^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right], \quad \eta = \sqrt{\frac{d}{TG^2}},$$

then Algorithm 3 returns an (ϵ,δ)-PGS solution. Furthermore, we have

$$T(\epsilon,\delta,\mathcal{MC}) = O\left[\frac{\beta(L+G)^2+d}{\epsilon^2}\log\left(\frac{1}{\delta}\right) + \frac{d^2G^2}{\epsilon^2}\right] = \tilde{O}\left[\frac{\beta(L+G)^2+d}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

In the case of Example 9, we have $\beta = d+1, G = 3\tilde{L}/2$ and then the asymptotic simulation cost of Algorithm 3 is at most

$$\tilde{O}\left[\frac{d(L+\tilde{L})^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

If both Lipschitz constants are independent of d and N, the asymptotic simulation cost becomes

$$\tilde{O}\left[\frac{d}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right],$$

which is $O(d)$ better than the general case without Assumption 10. In addition, in the case of Example 10, we have $G = d$ and $\beta = 2$. Hence, the asymptotic simulation cost is at most

$$\tilde{O}\left[\frac{d^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

Finally, we note that if we substitute ϵ with $c/2$, all upper bounds of simulation cost under Assumption 10 also hold for the PCS-IZ guarantee.

Proof of Theorem 48

In this subsection, we provide a proof to Theorem 4.B. Since the stochastic gradient is bounded, we apply the Azuma's inequality for martingale difference sequences with bounded tails.

Lemma 31 (Azuma's inequality with bounded tails)**.** *Let $X_0, \ldots, X_{T-1}$ be a martingale difference sequence. Suppose there exists a constant b such that for any $t \in \{0, \ldots, T-1\}$,*

$$\mathbb{P}(|X_t| \leq b) = 1.$$

Then for any $\delta > 0$, it holds with probability at least $1 - \delta$ that

$$\frac{1}{T}\sum_{t=0}^{T-1} X_t \leq b\sqrt{\frac{2}{T}\log\left(\frac{1}{\delta}\right)}. \tag{4.24}$$

The proof of Theorem 48 follows a similar way as Theorem 42. We first bound the noise term by Azuma's inequality.

Lemma 32. *Suppose that Assumptions 5-10 hold and let x^* be a minimizer of $f(x)$. Then, it holds*

$$\frac{1}{T}\sum_{t=0}^{T-1} \left\langle g^t - \hat{g}^t, x^t - x^* \right\rangle \leq \left(\frac{3L}{2} + G\right)\sqrt{\frac{2}{T}\log\left(\frac{1}{\delta}\right)}$$

with probability at least $1 - \delta$.

Proof. Same as the proof of Lemma 29, the fact that $\hat{g}^t$ is unbiased implies that

$$X_t := \left\langle g^t - \hat{g}^t, x^t - x^* \right\rangle \quad t = 0, 1, \ldots, T-1$$

is a martingale difference sequence. By Assumption 10 and property (v) in Lemma 24, we know

$$|X_t| = \left|\left\langle g^t - \hat{g}^t, x^t - x^* \right\rangle\right| \leq \left\|g^t - \hat{g}^t\right\|_1 \left\|x^t - x^*\right\|_\infty \leq \left\|g^t - \hat{g}^t\right\|_1 \leq 3L/2 + G,$$

which implies that the condition (4.24) holds with $b = 3L/2 + G$. Using Lemma 31, we get the conclusion of this lemma. □

The following lemma bounds the error of the algorithm and is similar to Theorem 3.2.2 in [171].

Lemma 33. *Suppose that Assumptions 5-10 hold and let x^* be a minimizer of $f(x)$. If we choose*

$$\eta = \sqrt{\frac{d}{TG^2}},$$

then we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \left\langle \hat{g}^t, x^t - x^* \right\rangle \leq \sqrt{\frac{dG^2}{T}}.$$

Proof. We define $\tilde{x}^{t+1} := x^t - \eta\hat{g}^t$ as the next point before the projection onto $[0,1]^d$. Recalling the non-expansion property of orthogonal projection, we get

$$\begin{aligned}\|x^{t+1} - x^*\|_2^2 &= \|\mathcal{P}_{\mathcal{X}}\left(\tilde{x}^{t+1} - x^*\right)\|_2^2 \leq \|\tilde{x}^{t+1} - x^*\|_2^2 = \|x^t - x^* - \eta\hat{g}^t\|_2^2 \\ &= \|x^t - x^*\|_2^2 + \eta^2\|\tilde{g}^t\|_2^2 - 2\eta\langle\hat{g}^t, x^t - x^*\rangle,\end{aligned}$$

and equivalently,

$$\langle\hat{g}^t, x^t - x^*\rangle = \frac{1}{2\eta}\left[\left\|x^t - x^*\right\|_2^2 - \left\|x^{t+1} - x^*\right\|_2^2\right] + \frac{\eta}{2}\cdot\left\|\hat{g}^t\right\|_2^2.$$

Using Assumption 10, we know $\|\hat{g}^t\|_2^2 \leq \|\hat{g}^t\|_1^2 \leq G^2$ and therefore

$$\langle\hat{g}^t, x^t - x^*\rangle = \frac{1}{2\eta}\left[\left\|x^t - x^*\right\|_2^2 - \left\|x^{t+1} - x^*\right\|_2^2\right] + \frac{\eta G^2}{2}.$$

Summing over $t = 0, 1, \ldots, T-1$, we have

$$\begin{aligned}\sum_{t=0}^{T-1}\langle\hat{g}^t, x^t - x^*\rangle &= \frac{\|x^0 - x^*\|_2^2 - \left\|x^T - x^*\right\|_2^2}{2\eta} + T\cdot\frac{\eta G^2}{2} \\ &\leq \frac{d\left\|x^0 - x\right\|_\infty^2}{2\eta} + \frac{\eta T G^2}{2} \leq \frac{d}{2\eta} + \frac{\eta T G^2}{2}.\end{aligned}$$

Choosing

$$\eta := \sqrt{\frac{d}{TG^2}},$$

it follows that

$$\sum_{t=0}^{T-1}\langle\tilde{g}^t, x^t - x^*\rangle \leq G\sqrt{dT}.$$

□

Now, we prove Theorem 48 using Lemmas 32 and 33.

Proof of Theorem 48. According to to the proof of Theorem 42, we have

$$\begin{aligned}\tilde{f}(\bar{x}) - f^* &\leq \frac{1}{T}\sum_{t=0}^{T-1}\left[\tilde{f}(x^t) - f^*\right] \leq \frac{1}{T}\sum_{t=0}^{T-1}\left\langle g^t, x^t - x^*\right\rangle \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\left\langle\hat{g}^t, x^t - x^*\right\rangle + \frac{1}{T}\sum_{t=0}^{T-1}\left\langle g^t - \hat{g}^t, x^t - x^*\right\rangle.\end{aligned} \tag{4.25}$$

By Lemmas 32 and 33, it holds

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\langle\hat{g}^t, x^t - x^*\right\rangle \leq \left(\frac{3L}{2} + G\right)\sqrt{\frac{2}{T}\log\left(\frac{2}{\delta}\right)}, \quad \frac{1}{T}\sum_{t=0}^{T-1}\left\langle g^t - \hat{g}^t, x^t - x^*\right\rangle \leq \sqrt{\frac{dG^2}{T}}$$

with probability at least $1 - \delta/2$. Choosing

$$T = \left(\frac{3L}{2} + G\right)^2 \cdot \frac{32}{\epsilon^2} \log\left(\frac{2}{\delta}\right) = \Theta\left[\frac{(L+G)^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right],$$

we know

$$T \geq \frac{16dG^2}{\epsilon^2}$$

when δ is small enough. Hence, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \left\langle \hat{g}^t, x^t - x^* \right\rangle \leq \frac{\epsilon}{4}, \quad \frac{1}{T}\sum_{t=0}^{T-1} \left\langle g^t - \hat{g}^t, x^t - x^* \right\rangle \leq \frac{\epsilon}{4}$$

holds with probability at least $1 - \delta/2$. Substituting into inequality (4.25), we have

$$\tilde{f}(\bar{x}) - f^* \leq \frac{\epsilon}{2}$$

holds with probability at least $1 - \delta/2$. By the results of Theorem 41, we know Algorithm 3 returns an (ϵ, δ)-PGS solution.

Finally, we estimate the simulation cost of Algorithm 3. For each iteration, the simulation cost is decided by the generation of a stochastic subgradient, which is at most β by Assumption 10. Hence, the total simulation cost of all iterations is

$$O\left[\beta T\right] = \tilde{O}\left[\frac{\beta(L+G)^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right].$$

By Theorem 41, the simulation cost of rounding process is at most

$$\tilde{O}\left[\frac{d}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right].$$

Thus, we know the total simulation cost of Algorithm 3 is at most

$$\tilde{O}\left[\frac{\beta(L+G)^2 + d}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right].$$

□

4.C Proofs in Section 4.4

Proof of Theorem 43

Proof of Theorem 43. To prove the function is well-defined, we only need to show that for any two different points $y, z \in [N-1]^d$ such that $\mathcal{C}_y \cap \mathcal{C}_z \neq \emptyset$, we have $\tilde{f}_y(x) = \tilde{f}_z(x)$ for all

$x \in \mathcal{C}_y \cap \mathcal{C}_z$. We first consider the case when $\|y - z\|_1 = 1$. Without loss of generality, we assume

$$y = (1, 1 \ldots, 1), \quad z = (2, 1, \ldots, 1).$$

In this case, we know that

$$\mathcal{C}_y \cap \mathcal{C}_z = \{(2, x_2, \ldots, x_d) : x_2, \ldots, x_d \in [0, 1]\}.$$

Suppose that point $x \in \mathcal{C}_y \cap \mathcal{C}_z$. We first calculate $\tilde{f}_y(x)$. We can define the "local coordinate" of x in $\mathcal{C}_y$ as

$$x - y = (1, x_2 - 1, \ldots, x_d - 1).$$

Let α_1 be a consistent permutation of x in $\mathcal{C}_y$ and and $S^{1,i}$ be the corresponding i-th neighbouring point. Since $(x - y)_1 = 1$ is not smaller than any other components, we can assume $\alpha_1(1) = 1$ and calculate $\tilde{f}_y(x)$ as

$$\begin{aligned}
\tilde{f}_y(x) &= [1 - (x - y)_{\alpha_1(1)}] f(S^{1,0}) \\
&\quad + \sum_{i=1}^{d-1} [(x - y)_{\alpha_1(i)} - (x - y)_{\alpha_1(i+1)}] f(S^{1,i}) + (x - y)_{\alpha_1(d)} f(S^{1,d}) \\
&= \sum_{i=1}^{d-1} [(x - y)_{\alpha_1(i)} - (x - y)_{\alpha_1(i+1)}] f(S^{1,i}) + (x - y)_{\alpha_1(d)} f(S^{1,d}) \\
&= \sum_{i=1}^{d-1} [x_{\alpha_1(i)} - x_{\alpha_1(i+1)}] f(S^{1,i}) + \left[x_{\alpha_1(d)} - 1\right] f(S^{1,d}).
\end{aligned} \tag{4.26}$$

Next, we consider $\tilde{f}_z(x)$ and define the "local coordinate" of x in $\mathcal{C}_z$ is

$$x - z = (0, x_2 - 1, \ldots, x_d - 1).$$

We define the permutation α_2 as

$$\alpha_2(i) = \alpha_1(i + 1), \quad \forall i \in [d - 1], \quad \alpha_2(d) = \alpha_1(1) = 1.$$

By the definition of α_1, we know

$$\begin{aligned}
&(x - z)_{\alpha_2(i)} = (x - y)_{\alpha_1(i+1)} \geq (x - y)_{\alpha_1(i+2)} = (x - z)_{\alpha_2(i+1)}, \quad \forall i \in [d - 2], \\
&(x - z)_{\alpha_2(d-1)} \geq 0 = (x_z)_{\alpha_2(d)}.
\end{aligned}$$

Hence, we know α_2 is a consistent permutation of x in $\mathcal{C}_z$ and let $S^{2,i}$ be the corresponding i-th neighbouring point of x in $\mathcal{C}_z$. Similar to the first case, the Lovász extension $\tilde{f}_z(x)$ can be calculated as

$$\tilde{f}_y(x) = [1 - (x - z)_{\alpha_2(1)}] f(S^{2,0}) \tag{4.27}$$

$$+\sum_{i=1}^{d-1}\ [(x-z)_{\alpha_2(i)}-(x-z)_{\alpha_2(i+1)}]f(S^{2,i})+(x-z)_{\alpha_2(d)}f(S^{2,d})$$

$$=[1-(x-z)_{\alpha_2(1)}]f(S^{2,0})+\sum_{i=1}^{d-1}\ [(x-z)_{\alpha_2(i)}-(x-z)_{\alpha_2(i+1)}]f(S^{2,i})$$

$$=[2-x_{\alpha_2(1)}]f(S^{2,0})+\sum_{i=1}^{d-1}\ [x_{\alpha_2(i)}-x_{\alpha_2(i+1)}]f(S^{2,i})+f(S^{2,d-1}).$$

Recalling the fact that $z=y+e_1$, for any $i\in[d-1]$, we have

$$S^{2,i}=z+\sum_{j=1}^{i}\ e_{\alpha_2(j)}=y+e_1+\sum_{j=1}^{i}\ e_{\alpha_1(j+1)}=y+\sum_{j=1}^{i+1}\ e_{\alpha_1(i)}=S^{1,i+1}.$$

Substituting into equation (4.27), we know

$$\begin{aligned}\tilde{f}_y(x)&=[2-x_{\alpha_2(1)}]f(S^{2,0})+\sum_{i=1}^{d-1}\ [x_{\alpha_2(i)}-x_{\alpha_2(i+1)}]f(S^{2,i})+f(S^{2,d-1})\\&=[2-x_{\alpha_2(1)}]f(S^{1,1})\\&\quad+\sum_{i=1}^{d-2}\ [x_{\alpha_2(i)}-x_{\alpha_2(i+1)}]f(S^{1,i+1})+[x_{\alpha_2(d-1)}-x_{\alpha_2(d)}]f(S^{1,d})+f(S^{1,d})\\&=[x_{\alpha_1(1)}-x_{\alpha_1(2)}]f(S^{1,1})\\&\quad+\sum_{i=1}^{d-2}\ [x_{\alpha_1(i+1)}-x_{\alpha_1(i+2)}]f(S^{1,i+1})+\left[x_{\alpha_1(d)}-2\right]f(S^{1,d})+f(S^{1,d})\\&=\sum_{i=1}^{d-1}\ [x_{\alpha_1(i)}-x_{\alpha_1(i+1)}]f(S^{1,i})+\left[x_{\alpha_1(d)}-1\right]f(S^{1,d}),\end{aligned}$$

which is equal to $\tilde{f}_y(x)$ by equation (4.26).

Then, we consider the case when $\|y-z\|_1>1$. Since $\mathcal{C}_y\cap\mathcal{C}_z\neq\emptyset$, we know $\|y-z\|_\infty=1$. Without loss of generality, we consider the case when

$$y=(1,1,\ldots,1),\quad z=y+\sum_{j=1}^{k}\ e_j,$$

where constant $k\in[d]$. In this case, we know

$$\mathcal{C}_y\cap\mathcal{C}_z=\left\{x\in\mathbb{R}^d:x_j=2,\ \forall j\leq k,\ x_j\in[0,1],\ \forall j\geq k+1\right\}.$$

We define

$$y_i:=y+\sum_{j=1}^{i}\ e_j,\quad \forall i\in\{0,1,\ldots,k\}.$$

Then, it follows that

$$\|y_i - y_{i-1}\|_1 = 1, \quad \forall i \in [k], \quad y_0 = y, \quad y_k = z$$

and

$$x \in \mathcal{C}_y \cap \mathcal{C}_z \subset \mathcal{C}_{y_i} \cap \mathcal{C}_{y_{i-1}} = \left\{x \in \mathbb{R}^d : x_i = 2,\ x_j \in [0,1],\ \forall j \in [d] \backslash \{i\}\right\}, \quad \forall i \in [k].$$

Hence, by the results for the case when $\|y - z\|_1 = 1$, we know

$$\tilde{f}_y(x) = \tilde{f}_{y_0}(x) = \tilde{f}_{y_1}(x) = \cdots = \tilde{f}_{y_k}(x) = \tilde{f}_z(x),$$

which means $\tilde{f}(x)$ is well-defined.

Finally, we prove the convexity of $\tilde{f}(x)$. Since the Lovász extension is the support function of submodular functions [80, section 6.3], the function $\tilde{f}_y(x)$ is the support function of $f(x)$ inside hypercube $\mathcal{C}_y$. In addition, Theorem 7.20 in [168] implies that the $L^\natural$-convex function $f(x)$ is integrally convex. Hence, we know that the support function of $f(x)$ on $\mathcal{X}$ is equal to $\tilde{f}_y(x)$ in each hypercube $\mathcal{C}_y$. By the definition of $\tilde{f}(x)$ in (4.7), the function $\tilde{f}(x)$ is the support function of $f(x)$ on $\mathcal{X}$. Since support functions are convex, we know $\tilde{f}(x)$ is convex. □

Proof of Theorem 44

Proof of Theorem 44. The proof can be done in the same way as Theorem 42 and we only give a sketch of the proof. We use the same notation as the proof of Theorem 28.

- If we have

$$M \geq 2\sigma \cdot \sqrt{\log\left(\frac{8\sigma dT}{\epsilon}\right)} = \tilde{\Theta}\left[\sqrt{\log\left(\frac{dT}{\epsilon}\right)}\right], \quad T \geq \frac{2\epsilon}{\sigma},$$

  then the proof of Lemma 28 implies that

$$\left\|b^t\right\|_1 \leq \frac{\epsilon}{2T}, \quad \forall t \in \{0, 1, \ldots, T-1\}.$$

- If we have

$$M = \max\left\{L, 2\sigma \cdot \sqrt{\log\left(\frac{8\sigma dT}{\epsilon}\right)}\right\} = \tilde{\Theta}\left[\sqrt{\log\left(\frac{dNT}{\epsilon}\right)}\right], \quad T \geq \frac{2N\epsilon}{\sigma},$$

  then the proof of Lemma 29 shows that

$$\frac{1}{T}\sum_{t=0}^{T-1} X_t \leq \sqrt{\frac{224dN^2\sigma^2}{T}\log\left(\frac{1}{\delta}\right)}$$

  holds with probability at least $1 - \delta$.

- If we choose
$$\eta = \frac{N}{M\sqrt{T}},$$
then the proof of Lemma 30 implies that
$$\frac{1}{T}\sum_{t=0}^{T-1}\langle \tilde{g}^t, x^t - x^*\rangle \leq \frac{dNM}{\sqrt{T}}.$$

Hence, choosing
$$T = \tilde{\Theta}\left[\frac{dN^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right], \quad M = \tilde{\Theta}\left[\sqrt{\log\left(\frac{dNT}{\epsilon}\right)}\right], \quad \eta = \frac{N}{M\sqrt{T}}$$
and using the inequality (4.19), we know the averaging point $\bar{x}$ is an $(\epsilon/2, \delta/2)$-PGS solution. Combining with Theorem 41, Algorithm 3 returns an (ϵ, δ)-PGS solution. Since the simulation cost of each iteration is $2d$, the total simulation cost of Algorithm 3 is at most
$$\tilde{O}\left[\frac{d^2N^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right] + \tilde{O}\left[\frac{d}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right] = \tilde{O}\left[\frac{d^2N^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

□

Similarly, we can estimate the asymptotic simulation cost under Assumption 10.

Theorem 49. *Suppose that Assumptions 5-8 and 10 hold. If we skip the truncation step in Algorithm 3 (or equivalently set $M = \infty$) and choose*
$$T = \tilde{\Theta}\left[\frac{(L+G)^2N^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right], \quad \eta = \sqrt{\frac{dN^2}{TG^2}},$$
then Algorithm 3 returns an (ϵ, δ)-PGS solution. Furthermore, we have
$$\begin{aligned} T(\epsilon, \delta, \mathcal{MC}) &= O\left[\frac{\beta(L+G)^2N^2 + d}{\epsilon^2}\log\left(\frac{1}{\delta}\right) + \frac{G^2d^2N^2}{\epsilon^2}\right] \\ &= \tilde{O}\left[\frac{\beta(L+G)^2N^2 + d}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right]. \end{aligned}$$

The above theorem can be proved in the same way as Theorem 48 and we omit the proof. We note that the step size η does not depend on N in this case.

Algorithms for the PCS-IZ Case

We first prove that the existence of indifference zone is equivalent to the so-called weak sharp minima condition of the convex extension. Moreover, we use the ℓ_∞ norm in place of the ℓ_2 norm since the feasible set is a hypercube.

Definition 14. We say a function $f(x) : \mathcal{X} \mapsto \mathbb{R}$ satisfies the *Weak Sharp Minimum* (WSM) condition, if the function $f(x)$ has a unique minimizer x^* and there exists a constant $\kappa > 0$ such that

$$\|x - x^*\|_\infty \leq \kappa \left(f(x) - f^*\right), \quad \forall x \in \mathcal{X},$$

where $f^* := f(x^*)$.

The WSM condition was first defined in [29], and is also called the polyhedral error bound condition in recent literature [241]. In addition, the WSM condition is a special case of the global growth condition in [239] with $\theta = 1$. The WSM condition can be used to leverage the distance between intermediate solutions and (c, δ)-PCS-IZ solutions. The next theorem verifies that the WSM condition is equivalent to the existence of indifference zone.

Theorem 50. *Suppose that function $f(x) : \mathcal{X} \mapsto \mathbb{R}$ is a $L^\natural$-convex function and $\tilde{f}(x)$ is the convex extension on $[1, N]^d$. Given a constant $c > 0$, function $f(x) \in \mathcal{MC}_c$ if and only if $\tilde{f}(x)$ satisfies the WSM condition with $\kappa = c^{-1}$.*

Proof of Theorem 50. We first prove the sufficiency part and then consider the necessity part.

Sufficiency. Suppose there exists a constant $\kappa > 0$ such that the function $\tilde{f}(x)$ satisfies the WSM condition with κ. Considering any point $x \in \mathcal{X}\backslash\{x^*\}$, we know $\|x - x^*\|_\infty \geq 1$ and, by the WSM condition,

$$f(x) - f^* = \tilde{f}(x) - f^* \geq \kappa^{-1}\|x - x^*\|_\infty \geq \kappa^{-1}.$$

Thus, we know the indifference zone parameter for $f(x)$ is at least κ^{-1} and $f(x) \in \mathcal{MC}_{\kappa^{-1}}$.

Necessity. Suppose there exists a constant $c > 0$ such that

$$f(x) - f^* \geq c, \quad \forall x \in \mathcal{X}\backslash\{x^*\}.$$

We first consider point $x \in [1, N]^d$ such that $\|x - x^*\|_\infty \leq 1$. In this case, we know there exists a hypercube $\mathcal{C}_y$ containing both x and x^*. By the definition of Lovász extension, we know that

$$\tilde{f}(x) = [1 - x_{\alpha_x(1)}]f\left(S^{x,0}\right) + \sum_{i=1}^{d-1} [x_{\alpha_x(i)} - x_{\alpha_x(i+1)}]f\left(S^{x,i}\right) + x_{\alpha_x(d)}f\left(S^{x,d}\right) = \sum_{i=0}^{d} \lambda_i f\left(S^{x,i}\right),$$

where we define

$$\lambda_i := x_{\alpha_x(i)} - x_{\alpha_x(i+1)}, \quad \forall i \in [d-1], \quad \lambda_0 := 1 - x_{\alpha_x(1)}, \quad \lambda_d := x_{\alpha_x(d)}.$$

Recalling the definition of consistent permutation, we get

$$\sum_{i=0}^{d} \lambda_i = 1, \quad \lambda_i \geq 0, \quad \forall i \in \{0, \ldots, d\}$$

and $\tilde{f}(x)$ is a convex combination of $f\left(S^{x,0}\right), \ldots, f\left(S^{x,d}\right)$. In addition, we can calculate that

$$\left(\sum_{i=0}^{d} \lambda_i S^{x,i}\right)_{\alpha_x(k)} = \sum_{i=0}^{d} \lambda_i \cdot S^{x,i}_{\alpha_x(k)} = \sum_{i=0}^{d} \lambda_i \cdot \mathbf{1}(i \geq k) = \sum_{i=k}^{d} \lambda_i = x_{\alpha_x(k)},$$

which implies that

$$x = \sum_{i=0}^{d} \lambda_i S^{x,i}.$$

If $x^* \notin \left\{S^{x,0}, \ldots, S^{x,d}\right\}$, the assumption that indifference zone parameter is c gives

$$\tilde{f}(x) - f^* = \sum_{i=0}^{d} \lambda_i \left[f\left(S^{x,i}\right) - f^*\right] \geq \sum_{i=0}^{d} \lambda_i \cdot c = c.$$

Combining with $\|x - x^*\|_\infty \leq 1$, we have

$$\|x - x^*\|_\infty \leq c^{-1} \cdot \left[\tilde{f}(x) - f^*\right].$$

Otherwise if $x^* = S^{x,i}$ for some $i \in \{0, \ldots, d\}$. Then, we know

$$\tilde{f}(x) - f^* = \sum_{i=0}^{d} \lambda_i \left[f\left(S^{x,i}\right) - f^*\right] \geq \sum_{i \neq k} \lambda_i \cdot c = (1 - \lambda_k)c$$

and

$$\begin{aligned}\|x - x^*\|_\infty &= \left\|\sum_{i=0}^{d} \lambda_i S^{x,i} - x^*\right\|_\infty = \left\|\sum_{i=0}^{d} \lambda_i \left(S^{x,i} - x^*\right)\right\|_\infty = \left\|\sum_{i \neq k} \lambda_i \left(S^{x,i} - x^*\right)\right\|_\infty \\ &\leq \sum_{i \neq k} \lambda_i \left\|S^{x,i} - x^*\right\|_\infty \leq \sum_{i \neq k} \lambda_i = 1 - \lambda_k,\end{aligned}$$

where the last inequality is because $S^{x,i}$ and x^* are in the same hypercube $\mathcal{C}_y$. Combining the above two inequalities, it follows that

$$\|x - x^*\|_2 \leq c^{-1} \cdot \left[\tilde{f}(x) - f^*\right],$$

which means that the WSM condition holds with $\kappa = c^{-1}$. Now we consider point $x \in [1, N]^d$ such that $\|x - x^*\|_\infty \geq 1$. We define

$$\tilde{x} := x^* + \frac{x - x^*}{\|x - x^*\|_\infty}$$

to be the point on the segment $\overline{xx^*}$ such that $\|\tilde{x} - x^*\|_\infty = 1$. By the convexity of $\tilde{f}(x)$ and the WSM condition for point $\tilde{x}$, we know

$$\tilde{f}(x) - f^* \geq \frac{\|x - x^*\|_\infty}{\|\tilde{x} - x^*\|_\infty} \left[\tilde{f}(\tilde{x}) - f^*\right] = \frac{\tilde{f}(\tilde{x}) - f^*}{\|\tilde{x} - x^*\|_\infty} \cdot \|x - x^*\|_\infty \geq c^{-1} \cdot \|x - x^*\|_\infty,$$

which shows that the WSM condition holds with $\kappa = c^{-1}$. Hence, the WSM condition holds for all points in $[1, N]^d$ with $\kappa = c^{-1}$. □

Using the WSM condition, we can accelerate Algorithm 3 by dynamically shrinking the search space. To describe the shrinkage of search space, we define the ℓ_∞-neighbourhood of point x as

$$\mathcal{N}(x, a) := \{y \in [1, N]^d : \|y - x\|_\infty \leq a\}$$

and the orthogonal projection onto $\mathcal{N}(x, a)$ as

$$\mathcal{P}_{x,a}(y) := (y \wedge (x + a)\mathbf{1}) \vee (x - a)\mathbf{1}, \quad \forall x \in \mathbb{R}^d.$$

Now we give the adaptive SSGD algorithm for the PCS-IZ guarantee.

Algorithm 5 Adaptive SSGD method for the PCS-IZ guarantee

Input: Model $\mathcal{X}, \mathcal{B}_\Upsilon, F(x, \xi_x)$, optimality guarantee parameter δ, indifference zone parameter c.

Output: An (c, δ)-PCS-IZ solution x^* to problem (4.1).

1: Set the initial guarantee $\epsilon_0 \leftarrow cN/4$.
2: Set the number of epochs $E \leftarrow \lceil \log_2(N) \rceil + 1$.
3: Set the initial search space $\mathcal{Y}_0 \leftarrow [1, N]^d$.
4: **for** $e = 0, \ldots, E - 1$ **do**
5: Use Algorithm 3 to get an $(\epsilon_e, \delta/(2E))$-PGS solution x_e in $\mathcal{Y}_e$.
6: Update guarantee $\epsilon_{e+1} \leftarrow \epsilon_e/2$.
7: Update the search space $\mathcal{Y}_{e+1} \leftarrow \mathcal{N}(x_e, 2^{-e-2}N)$.
8: **end for**
9: Round x_{E-1} to an integral point satisfying the (c, δ)-PCS-IZ guarantee by Algorithm 2.

Basically, the algorithm finds a $(c/2, \delta)$-PGS solution and, with the assumption that the indifference zone parameter is c, the solution satisfies the (c, δ)-PCS-IZ guarantee. We prove that the expected simulation cost of Algorithm 5 has only $O(\log(N))$ dependence on N.

Theorem 51. *Suppose that Assumptions 5-8 hold. Then, Algorithm 5 returns a (c, δ)-PCS-IZ solution. Furthermore, we have*

$$\begin{aligned} T(\delta, \mathcal{MC}_c) &= O\left[\frac{d^2\log(N)}{c^2}\log\left(\frac{1}{\delta}\right) + \frac{d^3\log(N)}{c^2}\log\left(\frac{d^2N}{\epsilon^3}\right) + \frac{d^3\log(N)L^2}{c^2}\right] \\ &= \tilde{O}\left[\frac{d^2\log(N)}{c^2}\log\left(\frac{1}{\delta}\right)\right]. \end{aligned}$$

Proof of Theorem 51. We first prove the correctness of Algorithm 5. Let x^* be the minimizer of $f(x)$ and $f^* := f(x^*)$. We use the induction method to prove that, for each epoch e, it holds

$$\tilde{f}(x_e) - f^* \leq \epsilon_e$$

with probability at least $1-(e+1)\delta/(2E)$. For epoch 0, the solution x_0 is $(\epsilon_0, \delta/(2E))$-PGS and we know

$$\tilde{f}(x_0) - f^* \leq \epsilon_0$$

holds with probability at least $1-\delta/(2E)$. We assume that the above event happens for the $(e-1)$-th epoch with probability at least $1-e\cdot\delta/(2E)$ and consider the case when this event happens. By Theorem 50, function $\tilde{f}(x)$ satisfies the WSM condition with $\kappa = c^{-1}$. Hence, the intermediate solution x_{e-1} satisfies

$$\|x_{e-1} - x^*\|_\infty \leq c^{-1}\left[\tilde{f}(x_{e-1}) - f^*\right] \leq c^{-1}\epsilon_{e-1} = c^{-1}\cdot 2^{-e+1}\epsilon_0 = 2^{-e-1}N,$$

which implies that $x^* \in \mathcal{N}(x_{e-1}, 2^{-e-1}N) = \mathcal{N}_e$ and therefore $x^* \in \mathcal{N}_e$. For the epoch e, it holds

$$\tilde{f}(x_e) - f^* = \tilde{f}(x_e) - \min_{x\in\mathcal{N}_e}\ \tilde{f}(x) \leq \epsilon_e$$

with probability at least $1-\delta/(2E)$. Hence, the above event happens with probability at least $1-\delta/(2E) - e\cdot\delta/(2E) = 1-(e+1)\delta/(2E)$ for epoch e. By the induction method, we know the claim holds for all epochs. Considering the last epoch, we know

$$\tilde{f}(x_{E-1}) - f^* \leq \epsilon_{E-1} = 2^{-E+1}\epsilon_0 = 2^{-\lceil\log_2(N)\rceil-2}\cdot cN \leq 2^{-\log_2(N)-2}\cdot cN = c/4$$

holds with probability at least $1-\delta/2$. Thus, we know x_{E-1} satisfies the $(c/4, \delta/2)$-PGS guarantee. By Theorem 41, the integral solution returned by Algorithm 5 satisfies the $(c/2, \delta)$-PGS guarantee. Since the indifference zone parameter is c, the solution satisfying the $(c/2, \delta)$-PGS guarantee must satisfies the (c, δ)-PCS-IZ guarantee.

Next, we estimate the asymptotic simulation cost of Algorithm 5. By Theorem 42, the simulation cost of epoch e is at most

$$\tilde{O}\left[\frac{d^2\left(2^{-e}N\right)^2}{\epsilon_e^2}\log\left(\frac{E}{\delta}\right)\right] = \tilde{O}\left[\frac{d^2\left(2^{-e}N\right)^2}{\left(2^{-e-2}\cdot cN\right)^2}\log\left(\frac{E}{\delta}\right)\right] = \tilde{O}\left[\frac{d^2}{c^2}\log\left(\frac{1}{\delta}\right)\right].$$

Summing over $e = 0, 1, \ldots, E-1$, we know the total simulation cost of E epochs is at most

$$\tilde{O}\left[E \cdot \frac{d^2}{c^2} \log\left(\frac{1}{\delta}\right)\right] = \tilde{O}\left[\frac{d^2 \log(N)}{c^2} \log\left(\frac{1}{\delta}\right)\right].$$

By Theorem 41, the simulation cost of the rounding process is at most

$$\tilde{O}\left[\frac{d}{c^2} \log\left(\frac{1}{\delta}\right)\right].$$

Combining the two parts, we know the asymptotic simulation cost of Algorithm 5 is at most

$$\tilde{O}\left[\frac{d^2 \log(N)}{c^2} \log\left(\frac{1}{\delta}\right)\right].$$

□

Similarly, we can estimate the asymptotic simulation cost under Assumption 10 and we omit the proof.

Theorem 52. *Suppose that Assumptions 5-8 and 10 hold. Then, Algorithm 5 returns a (c, δ)-PCS-IZ solution. Furthermore, we have*

$$T(\delta, \mathcal{MC}_c) = \tilde{O}\left[\frac{\beta(L+G)^2 \log(N) + d}{c^2} \log\left(\frac{1}{\delta}\right)\right].$$

4.D Proofs in Section 4.5

Proof of Theorem 45

Proof of Theorem 45. In this proof, we change the feasible set to $\mathcal{X} = \{0, 1, \ldots, N\}^d$, where $N \geq 1$. We split the proof into three steps.

Step 1. We first show that the construction of $L^\natural$-convex functions can be reduced to the construction of submodular functions. Equivalently, we show that any submodular function defined on $\{0,1\}^d$ can be extended to a $L^\natural$-convex function on $\mathcal{X}$ with the same convex extension after scaling. Let $g(x)$ be a submodular function defined on $\{0,1\}^d$ and $\tilde{g}(x)$ be the Lovász extension of $g(x)$. We first extend the domain of the Lovász extension to $[0, N]^d$ by scaling, i.e.,

$$\tilde{f}(x) := \tilde{g}(x/N), \quad \forall x \in [0, N]^d.$$

Then, we define the discretization of $\tilde{f}(x)$ by restricting to the integer lattice

$$f(x) := \tilde{f}(x), \quad \forall x \in \mathcal{X}.$$

We prove that $f(x)$ is a $L^\natural$-convex function. By Proposition 7.25 in [168], we know the Lovász extension $\tilde{g}(x)$ is a polyhedral L-convex function. Since the scaling operation does not change the L-convexity, we know $\tilde{f}(x)$ is also polyhedral L-convex. Hence, by Theorem 7.29 in [168], the function $\tilde{f}(x)$ satisfies the SBF$^\natural[\mathbb{R}]$ property, namely,

$$\tilde{f}(p) + \tilde{f}(q) \geq \tilde{f}[(p - \alpha\mathbf{1}) \vee q] + \tilde{f}(p \wedge (q + \alpha\mathbf{1})), \quad \forall p, q \in [0, N]^d,\ \alpha \geq 0.$$

Restricting to the integer lattice, we know the SBF$^\natural[\mathbb{Z}]$ property holds for $f(x)$, namely,

$$f(p) + f(q) \geq f[(p - \alpha\mathbf{1}) \vee q] + f(p \wedge (q + \alpha\mathbf{1})), \quad \forall p, q \in \{0, \ldots, N\}^d,\ \alpha \in \mathbb{N}.$$

Finally, Theorem 7.7 in [168] shows that the $L^\natural$-convexity is equivalent to the SBF$^\natural[\mathbb{Z}]$ property and therefore we know that $f(x)$ is a $L^\natural$-convex function.

Step 2. Next, we construct $d+1$ submodular functions on $\{0,1\}^d$ and extend them to $\mathcal{X}$ by the process defined in Step 1. The construction is based on the family of submodular functions defined in [90]. We denote $\mathcal{I} := \{0\} \cup [d]$. For each $i \in \mathcal{I}$, we define point $x^i \in \{0,1\}^d$ as

$$x^i := \sum_{j=1}^{i} e_j,$$

where e_j is the j-th unit vector of $\mathbb{R}^d$. Index $j(x)$ is defined as the maximal index j such that

$$x_i = 1, \quad \forall i \in [j].$$

If $x_1 = 0$, then we define $j(x) = 0$. Given $c : \mathcal{I} \mapsto \mathbb{R}$, we define a function on $\{0,1\}^d$ as

$$g^c(x) := \begin{cases} -c(i) & \text{if } x = x^i \text{ for some } i \in \mathcal{I} \\ (\|x\|_1 - j(x)) \cdot (d + 2 - j(x)) & \text{otherwise.} \end{cases}$$

By Lemma 6 in [90], the function $g^c(x)$ is submodular if $c(i) \in \{0,1\}$. Using the fact that convex combinations of submodular functions are still submodular, we know that $g^c(x)$ is submodular for any c such that $c(i) \in [0,1]$. Then, for each $i \in \mathcal{I}$, we construct

$$c^i(0) := \frac{1}{2}, \quad c^i(j) := \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases}, \quad \forall j \in [d].$$

We denote $g^i(x) := g^{c^i}(x)$ and let $f^i(x)$ be the extension of $6\epsilon \cdot g^i(x)$ on $\mathcal{X}$ by the process in Step 1. By the result in Step 1, we know that $f^i(x)$ is $L^\natural$-convex.

Next, we prove that $f^0(x)$ has disjoint set of ϵ-optimal solutions with $f^i(x)$ for any $i \in [d]$. For each $f^i(x)$, we define the set of ϵ-optimal solutions as

$$\mathcal{X}^i_\epsilon := \{x \in \mathcal{X} : f^i(x) - \min_y\ f^i(y) \leq \epsilon\}.$$

We first consider $\mathcal{X}_\epsilon^0$. By the definition of $g^0(x)$, we know that

$$f^0(x^0) = g^0(x^0) = -3\epsilon, \quad f^0(x) = g^0(x/N) \geq 0, \quad \forall x \in \{0, N\}^d \backslash \{x^0\}, \tag{4.28}$$

which implies that

$$\mathcal{X}_\epsilon^0 = \{x \in \mathcal{X} : f^0(x) \leq -2\epsilon\}.$$

Since $f^0(x)$ is defined by the scaled Lovász extension of $g^0(x)$, we have

$$f^0(x) = N^{-1} \cdot \left[(N - x_{\alpha(1)}) f^0(S^0) + \sum_{i=1}^{d-1} (x_{\alpha(i)} - x_{\alpha(i+1)}) f^0(S^i) + x_{\alpha(d)} f^0(S^d) \right], \tag{4.29}$$

where α is a consistent permutation of x/N and $S^i := N \cdot S^{x/N,i} \in \{0, N\}^d$ is the i-th neighbouring points of x in the hypercube $\{0, N\}^d$. Using the relation in (4.28) and the fact $S^0 = x^0$, we get

$$f^0(x) \geq N^{-1} \cdot (N - x_{\alpha(1)}) f(S_0) = N^{-1} \cdot (N - x_{\alpha(1)}) f(x^0) = -3\epsilon N^{-1} \cdot (N - x_{\alpha(1)}).$$

Hence, for any point $x \in \mathcal{X}_\epsilon^0$, we have $N - x_{\alpha(1)} = N - \max_i \ x_i \geq 2N/3$ and therefore

$$\mathcal{X}_\epsilon^0 \subset \{x \in \mathcal{X} : N - \max_i \ x_i \geq 2N/3\} = \{x \in \mathcal{X} : \max_i \ x_i \leq N/3\}. \tag{4.30}$$

Next, we consider $\mathcal{X}_\epsilon^i$ with $i \in [d]$. By the definition of $g^i(x)$, we have

$$f^i(x^0) = g^i(x^0) = -3\epsilon, \quad f^i(x) = g^i(x) \geq -6\epsilon, \quad \forall x \in \{0, N\}^d \backslash \{x^0\},$$

which implies that

$$\mathcal{X}_\epsilon^i = \{x \in \mathcal{X} : f^i(x) \leq -5\epsilon\}.$$

Since the consistent permutation and neighboring points only depend on the coordinate of x, we know

$$\begin{aligned} f^i(x) &= N^{-1} \cdot \left[(N - x_{\alpha(1)}) f^i(S^0) + \sum_{i=1}^{d-1} (x_{\alpha(i)} - x_{\alpha(i+1)}) f^i(S^i) + x_{\alpha(d)} f^i(S^d) \right] \\ &\geq N^{-1} \cdot \left[-3\epsilon(N - x_{\alpha(1)}) - 6\epsilon \sum_{i=1}^{d-1} (x_{\alpha(i)} - x_{\alpha(i+1)}) - 6\epsilon \cdot x_{\alpha(d)} \right] \\ &= N^{-1} \cdot \left[-3\epsilon(N - x_{\alpha(1)}) - 6\epsilon \cdot x_{\alpha(1)} \right] = -3\epsilon N^{-1} \cdot (N + x_{\alpha(1)}). \end{aligned} \tag{4.31}$$

Hence, the set $\mathcal{X}_\epsilon^i$ satisfies

$$\mathcal{X}_\epsilon^i \subset \left\{ x \in \mathcal{X} : N + \max_i \ x_i \geq 5N/3 \right\} = \left\{ x \in \mathcal{X} : \max_i \ x_i \geq 2N/3 \right\}. \tag{4.32}$$

Combining the relations (4.30) and (4.32), we know $\mathcal{X}_\epsilon^0 \cap \mathcal{X}_\epsilon^i = \emptyset$ for all $i \in [d]$.

Step 3. Finally, we give a lower bound of $T_0(\epsilon, \delta, \mathcal{MC})$. For each $i \in \mathcal{I}$, we define M_i as the model such that the objective function is $f^i(x)$ and the distribution at each point is Gaussian with variance σ^2. Same as the one-dimensional case, given a zeroth-order algorithm and a model M, we denote $N_x(\tau)$ as the number of times that $F(x, \xi_x)$ is simulated when the algorithm terminates. By definition, we have

$$\mathbb{E}_M[\tau] = \sum_{x \in \mathcal{X}} \mathbb{E}_M \left[N_x(\tau) \right],$$

where $\mathbb{E}_M$ is the expectation when the model M is given. Similarly, we can define $\mathbb{P}_M$ as the probability when the model M is given. Suppose $\mathcal{A}$ is an $[(\epsilon, \delta)$-PGS, $\mathcal{MC}]$-algorithm and let $\mathcal{E}$ be the event that the solution returned by $\mathcal{A}$ is in the set $\mathcal{X}_\epsilon^0$. Since $\mathcal{X}_\epsilon^0 \cap \mathcal{X}_\epsilon^i = \emptyset$ for all $i \in [d]$, we know

$$\mathbb{P}_{M_0}[\mathcal{E}] \geq 1 - \delta, \quad \mathbb{P}_{M_i}[\mathcal{E}] \leq \delta, \quad \forall i \in [d].$$

Using the information-theoretical inequality (4.9), it holds

$$\sum_{x \in \mathcal{X}} \mathbb{E}_{M_0} \left[N_x(\tau) \right] \mathrm{KL}(\nu_{0,x}, \nu_{i,x}) \geq d(\mathbb{P}_{M_0}(\mathcal{E}), \mathbb{P}_{M_i}(\mathcal{E})) \geq d(1 - \delta, \delta) \geq \log \left(\frac{1}{2.4\delta} \right), \tag{4.33}$$

where $d(x, y) := x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$, $\mathrm{KL}(\cdot, \cdot)$ is the KL divergence and $\nu_{i,x}$ is the distribution of $F^i(x, \xi_x)$. Since the distributions $\nu_{i,x}$ are Gaussian with variance σ^2, the KL divergence can be calculated as

$$\mathrm{KL}(\nu_{0,x}, \nu_{i,x}) = 2\sigma^{-2} \left(f^0(x) - f^i(x) \right)^2.$$

Now we estimate $f^0(x) - f^i(x)$ for all $i \in [d]$. By equations (4.29) and (4.31), we get

$$\begin{aligned} f^0(x) - f^i(x) = N^{-1} \Bigg[& (N - x_{\alpha(1)}) \left(f^0(S^0) - f^i(S^0) \right) \\ & + \sum_{j=1}^{d-1} (x_{\alpha(j)} - x_{\alpha(j+1)}) \left(f^0(S^j) - f^i(S^j) \right) + x_{\alpha(d)} \left(f^0(S^d) - f^i(S^d) \right) \Bigg], \end{aligned} \tag{4.34}$$

where α is a consistent permutation of x/N and S^i is the i-th neighboring point of x in hypercube $\{0, N\}^d$. By the definition of $f^0(x)$ and $f^i(x)$, we have

$$f^0(x) - f^i(x) = \begin{cases} 6\epsilon & \text{if } x = x^i \\ 0 & \text{otherwise.} \end{cases}$$

Since $\|x^i\|_1 = i$ and $\|S^j\|_1 = j$ for all $j \in \mathcal{I}$, we know

$$f^0(S^i) - f^i(S^i) \leq 6\epsilon, \quad f^0(S^j) - f^i(S^j) = 0, \quad \forall j \in \mathcal{I} \backslash \{i\}.$$

Substituting into equation (4.34), it follows that

$$f^0(x) - f^i(x) \leq \begin{cases} \left(6\epsilon \cdot (x_{\alpha(i)} - x_{\alpha(i+1)})\right)^2 & \text{if } i \in [d-1] \\ \left(6\epsilon \cdot x_{\alpha(d)}\right)^2 & \text{if } i = d. \end{cases}$$

Hence, the KL divergence is bounded by

$$\mathrm{KL}(\nu_{0,x}, \nu_{i,x}) = 2\sigma^{-2}\left(f^0(x) - f^i(x)\right)^2 \leq \begin{cases} 72\sigma^{-2}N^{-2}\epsilon^2\left((x_{\alpha(i)} - x_{\alpha(i+1)})\right)^2 & \text{if } i \in [d-1] \\ 72\sigma^{-2}N^{-2}\epsilon^2 x_{\alpha(d)}^2 & \text{if } i = d. \end{cases}$$

Substituting the KL divergence into inequality (4.33) and summing over $i = 1, \ldots, d$, we get

$$\sum_{x \in \mathcal{X}} \mathbb{E}_{M_0}\left[N_x(\tau)\right] \cdot 72\sigma^{-2}N^{-2}\epsilon^2 \left[\sum_{i=1}^{d-1}(x_{\alpha(i)} - x_{\alpha(i+1)})^2 + x_{\alpha(d)}^2\right] \geq d \log\left(\frac{1}{2.4\delta}\right). \tag{4.35}$$

Since α is the consistent permutation of x, we know

$$0 \leq x_{\alpha(i)} - x_{\alpha(i+1)} \leq N, \quad \forall i \in [d-1]$$

and therefore

$$\sum_{i=1}^{d-1}(x_{\alpha(i)} - x_{\alpha(i+1)})^2 + x_{\alpha(d)}^2 \leq N \cdot \left(\sum_{i=1}^{d-1}(x_{\alpha(i)} - x_{\alpha(i+1)}) + x_{\alpha(d)}\right) = N \cdot x_{\alpha(1)} \leq N^2.$$

Combining with inequality (4.35), we get

$$\sum_{x \in \mathcal{X}} \mathbb{E}_{M_0}\left[N_x(\tau)\right] \cdot 72\epsilon^2\sigma^{-2} \geq d \log\left(\frac{1}{2.4\delta}\right),$$

which implies that

$$\mathbb{E}_{M_0}[\tau] = \sum_{x \in \mathcal{X}} \mathbb{E}_{M_0}\left[N_x(\tau)\right] \geq \frac{d\sigma^2}{72\epsilon^2} \log\left(\frac{1}{2.4\delta}\right).$$

□

Proof of Theorem 46

Proof of Theorem 46. We consider the submodular functions $g^0(x), \ldots, g^d(x)$ constructed in the proof of Theorem 45. We want to construct objective functions $f^0(x), \ldots, f^d(x)$ on $\mathcal{X} = [N]^d$ such that

$$f^i(x) = \begin{cases} 6c \cdot g^i(x-1) + h(x) & \text{if } x \in [2]^d \\ h(x) & \text{if } x \in [N]^d \backslash [2]^d, \end{cases} \quad \forall i \in \{0, \ldots, d\},$$

where $(x-1)_j := x_j - 1$ for all $j \in [d]$ and $h(x)$ is a suitably designed function. Similar to the proof of Theorem 45, we apply the information-theoretical inequality (4.9) to pairs $f^0(x)$ and $f^i(x)$ for all $i \in [d]$. Since the objective function values for $f^0(x)$ and $f^i(x)$ are equal for all $x \in [N]^d \backslash [2]^d$, the terms with respect to those x will disappear and we only need to analyze the terms with $x \in [2]^d$. Now, using the same analysis and notations as Theorem 45, we get the desired lower bound

$$\mathbb{E}_{M_0}[\tau] = \sum_{x \in \mathcal{X}} \mathbb{E}_{M_0}\left[N_x(\tau)\right] \geq \frac{d\sigma^2}{72c^2} \log\left(\frac{1}{2.4\delta}\right).$$

Therefore, it remains to chose a suitable function $h(x)$ such that $f^i(x)$ are $L^\natural$-convex on the whole feasible set $\mathcal{X}$. We define

$$M := \max_{x \in \{0,1\}^d,\ i \in \{0,\ldots,d\}} 6c \cdot |g^i(x)|.$$

The extended function $f^i(x)$ is defined by

$$h(x) := 4M \sum_{j=1}^{d} (x_j - 1)(x_j - 2) + 2M \max_j x_j + 2M \sum_{j=1}^{d} \mathbf{1}(x_j = 1), \quad x \in [N]^d,$$

where $\mathbf{1}(\cdot)$ is the indicator function. The function $h(x)$ is the sum of two $L^\natural$-convex functions [168] and thus is a $L^\natural$-convex function. We prove that for each $i \in [d]$, the function $f^i(x)$ is $L^\natural$-convex, namely, it satisfies the discrete mid-point convexity. Suppose that $x, y \in [N]^d$ are two feasible points. We consider three different cases.

Case I. We first consider the case when $x, y \in [2]^d$. In this case, the fact that $[2]^d$ is a $L^\natural$-convex set implies that

$$\left\lceil \frac{x+y}{2} \right\rceil, \left\lfloor \frac{x+y}{2} \right\rfloor \in [2]^d.$$

Since the function $6c \cdot g^i(x) + h(x)$ is $L^\natural$-convex, the discrete mid-point convexity holds for x and y.

Case II. We consider the case when $x, y \notin [2]^d$. Since the function $\sum_j \mathbf{1}(x_j = 1)$ is $L^\natural$-convex, it satisfies the discrete mid-point convexity and we can safely ignore its effect in this case. If $\lfloor (x+y)/2 \rfloor, \lceil (x+y)/2 \rceil \notin [2]^d$, then the $L^\natural$-convexity of $h(x)$ implies the discrete mid-point convexity of points x and y. Now, we consider the case when $\lfloor (x+y)/2 \rfloor, \lceil (x+y)/2 \rceil \in [2]^d$. Since at least one component of x and y is larger than 2, it holds that

$$f^i(x) \geq 4M \cdot (3-1)(3-2) + 3M = 11M, \quad f^i(y) \geq 11M.$$

Hence, we get

$$f^i(x) + f^i(y) \geq 22M \geq f^i\left(\left\lceil \frac{x+y}{2} \right\rceil\right) + f^i\left(\left\lfloor \frac{x+y}{2} \right\rfloor\right).$$

The only remaining case is when

$$\left\lceil \frac{x+y}{2} \right\rceil \notin [2]^d, \quad \left\lfloor \frac{x+y}{2} \right\rfloor \in [2]^d.$$

In this case, we have

$$\left\lfloor \frac{x_j+y_j}{2} \right\rfloor \leq 2, \quad \forall j \in [d], \quad \max_j \left\lceil \frac{x_j+y_j}{2} \right\rceil \geq 3,$$

which implies that

$$x_j + y_j \leq \max_j (x_j + y_j) = 5, \quad \forall j \in [d]$$

and

$$\max_j x_j \geq 3, \quad \max_j y_j \geq 3, \quad \max_j \left\lceil \frac{x_j+y_j}{2} \right\rceil = 3, \quad \max_j \left\lfloor \frac{x_j+y_j}{2} \right\rfloor = 2.$$

Let

$$\mathcal{J}_x := \{j \in [d] : \ x_j \geq 3\}, \quad \mathcal{J}_y := \{j \in [d] : \ y_j \geq 3\}, \quad \mathcal{J} := \{j \in [d] : \ x_j + y_j = 5\}. \tag{4.36}$$

The analysis above gives

$$\mathcal{J} \subset \mathcal{J}_x \cup \mathcal{J}_y, \quad \mathcal{J}_x \cap \mathcal{J}_y = \emptyset.$$

Hence, we know

$$\sum_j (x_j - 1)(x_j - 2) + \sum_j (y_j - 1)(y_j - 2) \geq 2|\mathcal{J}_x| + 2|\mathcal{J}_y|,$$
$$\sum_j \left(\left\lceil \frac{x_j+y_j}{2} \right\rceil - 1 \right) \left(\left\lceil \frac{x_j+y_j}{2} \right\rceil - 2 \right) + \sum_j \left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor - 1 \right) \left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor - 2 \right) = 2|\mathcal{J}|.$$

Combining with inequality (4.36), we get

$$h(x) + h(y) - h\left(\left\lceil \frac{x_j+y_j}{2} \right\rceil \right) - h\left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor \right) \geq 8M(|\mathcal{J}_x| + |\mathcal{J}_y| - |\mathcal{J}|) + 2M \geq 2M.$$

Therefore, it holds that

$$\begin{aligned} f^i(x) + f^i(y) = h(x) + h(y) &\geq h\left(\left\lceil \frac{x_j+y_j}{2} \right\rceil \right) + h\left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor \right) + 2M \\ &\geq h\left(\left\lceil \frac{x_j+y_j}{2} \right\rceil \right) + h\left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor \right) + 6c \cdot g^i(x-1) \\ &= f^i\left(\left\lceil \frac{x_j+y_j}{2} \right\rceil \right) + f^i\left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor \right). \end{aligned}$$

Case III. Finally, we consider the case when $x \in [2]^d$ and $y \notin [2]^d$. If

$$\left\lceil \frac{x+y}{2} \right\rceil, \left\lfloor \frac{x+y}{2} \right\rfloor \in [2]^d,$$

we know

$$f^i(y) + f^i(x) \geq 11M - M > 6M \geq f^i\left(\left\lceil \frac{x_j+y_j}{2} \right\rceil\right) + f^i\left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor\right).$$

Next, for the case where

$$\left\lceil \frac{x+y}{2} \right\rceil, \left\lfloor \frac{x+y}{2} \right\rfloor \notin [2]^d,$$

we get

$$\max_j \frac{x_j+y_j}{2} \geq 3,$$

which implies that

$$\max_j y_j \geq 4.$$

Considering the component j such that $y_j \geq 4$, it follows that

$$\begin{aligned}
&\left(\left\lceil \frac{x_j+y_j}{2} \right\rceil - 1\right)\left(\left\lceil \frac{x_j+y_j}{2} \right\rceil - 2\right) + \left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor - 1\right)\left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor - 2\right) \\
\leq &\left(\frac{x_j+y_j+1}{2} - 1\right)\left(\frac{x_j+y_j+1}{2} - 2\right) + \left(\frac{x_j+y_j}{2} - 1\right)\left(\frac{x_j+y_j}{2} - 2\right) \\
\leq &\left(\frac{y_j+3}{2} - 1\right)\left(\frac{y_j+3}{2} - 2\right) + \left(\frac{y_j+2}{2} - 1\right)\left(\frac{y_j+2}{2} - 2\right) = \frac{1}{2}y_j^2 - \frac{1}{2}y_j - \frac{1}{4}.
\end{aligned}$$

Combining with the $L^\natural$-convexity of functions $(x_k - 1)(y_k - 2)$ for each $k \in [d]$ and $\max_k x_k$, we get

$$\begin{aligned}
&h(x) + h(y) - h\left(\left\lceil \frac{x+y}{2} \right\rceil\right) - h\left(\left\lfloor \frac{x+y}{2} \right\rfloor\right) \\
\geq &h_j(x) + h_j(y) - h_j\left(\left\lceil \frac{x+y}{2} \right\rceil\right) - h_j\left(\left\lfloor \frac{x+y}{2} \right\rfloor\right) \\
\geq &4M(x_j - 1)(x_j - 2) + 4M(y_j - 1)(y_j - 2) \\
&- 4M\left(\left\lceil \frac{x_j+y_j}{2} \right\rceil - 1\right)\left(\left\lceil \frac{x_j+y_j}{2} \right\rceil - 2\right) - 4M\left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor - 1\right)\left(\left\lfloor \frac{x_j+y_j}{2} \right\rfloor - 2\right) \\
\geq &0 + 4M\left[y_j^2 - 3y_j + 2 - \left(\frac{1}{2}y_j^2 - \frac{1}{2}y_j - \frac{1}{4}\right)\right] = M(2y_j^2 - 10y_j + 9) \geq M.
\end{aligned}$$

Therefore, we have

$$f^i(x) + f^i(y) = h(x) + h(y) + 6c \cdot g^i(x - 1) \geq h(x) + h(y) - M$$

$$\geq h\left(\left\lceil\frac{x+y}{2}\right\rceil\right)+h\left(\left\lfloor\frac{x+y}{2}\right\rfloor\right)+M-M$$
$$\geq h\left(\left\lceil\frac{x+y}{2}\right\rceil\right)+h\left(\left\lfloor\frac{x+y}{2}\right\rfloor\right)=f^i\left(\left\lceil\frac{x+y}{2}\right\rceil\right)+f^i\left(\left\lfloor\frac{x+y}{2}\right\rfloor\right).$$

Now, we consider the last case where

$$\left\lceil\frac{x+y}{2}\right\rceil\notin[2]^d,\quad \left\lfloor\frac{x+y}{2}\right\rfloor\in[2]^d.$$

Similar to Case II, we can prove that

$$x_j+y_j\leq 5,\quad \forall j\in[d].$$

If it holds that

$$h(y)>h\left(\left\lceil\frac{x+y}{2}\right\rceil\right),$$

we can utilize that fact that $y,\lceil\frac{x+y}{2}\rceil\in\mathbb{Z}^d$ to prove

$$h(y)\geq h\left(\left\lceil\frac{x+y}{2}\right\rceil\right)+2M,$$

which leads to

$$\begin{aligned}f^i(x)+f^i(y)&\geq h(x)+h(y)-M\geq 0+h\left(\left\lceil\frac{x+y}{2}\right\rceil\right)+2M-M\\&\geq h\left(\left\lceil\frac{x+y}{2}\right\rceil\right)+6c\cdot g^i\left(\left\lfloor\frac{x+y}{2}\right\rfloor\right)\\&=h\left(\left\lceil\frac{x+y}{2}\right\rceil\right)+0+6c\cdot g^i\left(\left\lfloor\frac{x+y}{2}\right\rfloor\right)\\&\geq h\left(\left\lceil\frac{x+y}{2}\right\rceil\right)+h\left(\left\lfloor\frac{x+y}{2}\right\rfloor\right)+6c\cdot g^i\left(\left\lfloor\frac{x+y}{2}\right\rfloor\right)\\&=f^i\left(\left\lceil\frac{x+y}{2}\right\rceil\right)+f^i\left(\left\lfloor\frac{x+y}{2}\right\rfloor\right).\end{aligned}$$

Therefore, we focus on the case when

$$h(y)\leq h\left(\left\lceil\frac{x+y}{2}\right\rceil\right). \tag{4.37}$$

First, using the facts that $x\in[2]^d$ and $y\notin[2]^d$, it is easy to prove that

$$\max_j y_j\geq\max_j\left\lceil\frac{x_j+y_j}{2}\right\rceil=3,\quad \mathbf{1}(y_j=1)\geq\mathbf{1}\left(\left\lceil\frac{x_j+y_j}{2}\right\rceil=1\right),\quad\forall j\in[d]. \tag{4.38}$$

Moreover, using the condition that $x_j \in [2]$, it holds that

$$\left| y_j - \frac{3}{2} \right| \geq \left| \left\lceil \frac{x_j + y_j}{2} \right\rceil - \frac{3}{2} \right|, \quad \forall j \in [d],$$

which implies that

$$\sum_j (y_j - 1)(y_j - 2) \geq \sum_j \left(\left\lceil \frac{x_j + y_j}{2} \right\rceil - 1 \right) \left(\left\lceil \frac{x_j + y_j}{2} \right\rceil - 2 \right).$$

Combining with inequalities in (4.38), we get

$$h(y) \geq h\left(\left\lceil \frac{x + y}{2} \right\rceil \right).$$

In addition, the equality of the above inequality holds in combination with our assumption in (4.37). The equality conditions imply that

$$\max_j y_j = 3, \quad \mathbf{1}(y_j = 1) = \mathbf{1}\left(\left\lceil \frac{x_j + y_j}{2} \right\rceil = 1 \right), \quad \left| y_j - \frac{3}{2} \right| = \left| \left\lceil \frac{x_j + y_j}{2} \right\rceil - \frac{3}{2} \right|, \quad \forall j \in [d].$$

The above three conditions imply that

$$y = \left\lceil \frac{x_j + y_j}{2} \right\rceil.$$

Utilizing the identity

$$x + y = \left\lceil \frac{x_j + y_j}{2} \right\rceil + \left\lfloor \frac{x_j + y_j}{2} \right\rfloor,$$

we know

$$x = \left\lfloor \frac{x_j + y_j}{2} \right\rfloor.$$

In this case, the discrete mid-point convexity holds evidently.

□

4.E Proofs in Section 4.6

Proof of Theorem 47

First, the following lemma shows that the lower bound of $\mathbb{E}[H_x(y, \eta_y)]$ in $\mathcal{N}_x$ implies a global lower bound of $f(x)$.

Lemma 34. *Suppose that Assumptions 5-9 hold. If we have*

$$\mathbb{E}[H_x(y, \eta_y)] \geq -b, \quad \forall y \in \mathcal{N}_x$$

for some constant $b \geq 0$, then it holds

$$f(y) \geq f(x) - \frac{2N}{1-a} \cdot b, \quad \forall y \in \mathcal{X}.$$

Proof of Lemma 34. The proof follows the same framework as Theorem 50. We first consider points $y \in \mathcal{N}_x$. By the condition of this lemma and inequality (4.10), we have

$$f(y) - f(x) \geq (1-a)^{-1} \cdot \mathbb{E}[H_x(y, \eta_y)] \geq -(1-a)^{-1}(1-a)^{-1}b.$$

Next, we consider point $y \in \mathcal{X}$ such that $\|y - x\|_\infty \leq 1$. Then, there exists two disjoint sets $\mathcal{S}_1, \mathcal{S}_2 \subset [d]$ such that

$$y = x + e_{\mathcal{S}_1} - e_{\mathcal{S}_2},$$

where $e_{\mathcal{S}} := \sum_{i \in \mathcal{S}} e_i$ is the indicator vector of $\mathcal{S}$. Using the translation submodularity of $f(x)$, we have

$$f(y) \geq f(x + e_{\mathcal{S}_1}) - f(x) + f(x - e_{\mathcal{S}_2}) - f(x) \geq -2(1-a)^{-1}b.$$

Now, let $\tilde{f}(x)$ be the convex extension of $f(x)$ defined in (4.7) and consider $y \in [1, N]^d$ such that $\|y - x\|_\infty \leq 1$. We consider the hypercube $\mathcal{C}_z$ that contains both x and y and denote $S^{y,i}$ as the i-th neighboring point of y in $\mathcal{C}_z$. Recalling the expression (4.6), we know $f(y)$ is a convex combination of $f(S^{y,0}), \ldots, f(S^{y,d})$. Since the neighboring point $S^{y,i} \in \mathcal{X}$ satisfies $\|S^{y,i} - x\|_\infty \leq 1$, we know

$$\tilde{f}(y) \geq \min_{i \in \{0\} \cup [d]} f(S^{y,i}) \geq -2(1-a)^{-1}b.$$

Finally, we consider points $y \in [1, N]^d$. We define

$$\tilde{y} := x + \frac{y - x}{\|y - x\|_\infty}.$$

Then, we know $\|\tilde{y} - x\|_\infty = 1$ and $\tilde{f}\tilde{y}) \geq -2(1-a)^{-1}b$. By the convexity of $\tilde{f}(x)$,

$$\tilde{f}(y) - f(x) \geq \frac{\|y - x\|_\infty}{\|\tilde{y} - x\|_\infty} \left[\tilde{f}(\tilde{y}) - f(x)\right] \geq -N \cdot 2(1-a)^{-1}b = -\frac{2N}{1-a} \cdot b.$$

□

Hence, to find an (ϵ, δ)-PGS solution, it suffices to find point x such that

$$\mathbb{E}[H_x(y, \eta_y)] \geq -\frac{(1-a)\epsilon}{2N}, \quad \forall y \in \mathcal{N}_x$$

holds with probability at least $1 - \delta$.

Proof of Theorem 47. Let x^* be a minimizer of $f(x)$. We use the induction method to prove that

$$f(x^{e,0}) - f(x^*) \leq 2^{-e} \cdot NL, \quad \forall e \in \{0, 1, \ldots, E\} \tag{4.39}$$

holds with probability at least $1 - e \cdot \delta/E$. Using Assumption 8, we have

$$f(x^{0,0}) - f(x^*) \leq L \cdot \|x^{0,0} - x^*\|_\infty \leq NL,$$

which means the induction assumption holds for epoch 0. Suppose the induction assumption is true for epochs $0, 1, \ldots, e-1$. Now we consider epoch e. We assume the event

$$f(x^{e-1,0}) - f(x^*) \leq 2^{-e+1} \cdot NL$$

happens in the following proof, which has probability at least $1 - (e-1)\delta/E$. We suppose epoch e terminates after T_e iterations and discuss by two different cases.

Case I. We first consider the case when $T_e \leq T - 1$. This event happens only if epoch $e-1$ is terminated by the condition in Line 13, i.e.,

$$\hat{H}_{x^{e-1,T_e-1}}(y) > -2h_{e-1}, \quad \forall y \in \mathcal{N}_{x^{e-1,T_e-1}}.$$

By the definition of confidence intervals, it follows that

$$\min_{y \in \mathcal{N}_{x^{e-1,T_e-1}}} \mathbb{E}[H_{x^{e-1,T_e-1}}(y, \eta_y)] \geq -3h_{e-1} = -3 \cdot 2^{-e+1} h_0 = -2^{-e-1} \cdot (1-a)L$$

holds with probability at least $1 - \delta/(ET)$. Then, considering the results of Lemma 34, we know

$$f(x^{e,0}) - f(x^*) = f(x^{e-1,T_e-1}) - f(x^*) \leq \frac{2N}{1-a} \cdot 2^{-e-1} \cdot (1-a)L = 2^{-e} \cdot NL$$

happens with the same probability. Combining with the induction assumption for epoch $e-1$, the above event happens with probability at least $1 - (e-1)\delta/E - \delta/(ET) \geq 1 - e \cdot \delta/E$ and the induction assumption holds for epoch e.

Case II. Next, we consider the case when $T_e = T$. We estimate the object function decrease for each iteration $t = 0, 1, \ldots, T-1$. By the definition of confidence intervals, it holds

$$\mathbb{E}[H_{x^{e-1,t}}(y, \eta_y)] \leq -h_{e-1}$$

with probability at least $1 - \delta/(ET)$, where $y = x^{e-1,t+1}$ is the next iteration point. Recalling inequality (4.10), we know

$$f(x^{e-1,t+1}) - f(x^{e-1,t}) \leq -(1+a)^{-1} h_{e-1}$$

happens with probability at least $1-\delta/(ET)$. We assume the above event happens for all $t=1,2,\ldots,T$, which has probability at least $1-T\cdot\delta/(ET)=1-\delta/E$. Then, we have

$$\begin{aligned} f(x^{e,0})-f(x^{e-1,0}) &= f(x^{e-1,T})-f(x^{e-1,0}) = \sum_{t=1}^{T} f(x^{e-1,t})-f(x^{e-1,t-1}) \\ &\leq -T\cdot(1+a)^{-1}h_{e-1} = -2^{-e}\cdot NL \end{aligned}$$

holds with the same probability. Combining with the induction assumption for epoch $e-1$, we know

$$f(x^{e,0})-f(x^*)\leq 2^{-e}\cdot NL$$

happens with probability at least $1-(e-1)\delta/E-\delta/E = 1-e\cdot\delta/E$. This means the induction assumption holds for epoch e.

Combining the above two cases, we know the induction assumption is true for epoch e. By the induction method, we know inequality (4.39) holds for epoch E, i.e.,

$$f(x^{E,0})-f(x^*)\leq 2^{-E}\cdot NL = 2^{-\lceil\log_2(NL/\epsilon)\rceil}\cdot NL \leq 2^{-\log_2(NL/\epsilon)}\cdot NL = \epsilon$$

with probability at least $1-E\cdot\delta/E = 1-\delta$. Hence, Algorithm 4 returns an (ϵ,δ)-PGS solution.

Next, we estimate the simulation cost of Algorithm 4. For each iteration in epoch e, Hoeffding bound implies that simulating $H_x(y,\eta_y)$ for

$$\frac{2\tilde{\sigma}^2}{h_e^2}\log\left(\frac{2ET}{\delta}\right) = 2^{2e}\cdot\frac{288\tilde{\sigma}^2}{(1-a)^2L^2}\log\left(\frac{2ET}{\delta}\right)$$

times is sufficient to ensure that the $1-\delta/(ET)$ confidence half-width is at most T_e. Since the simulation cost of each evaluation of all $H_x(y,\eta_y)$ is γ, the simulation cost of epoch e is at most

$$\gamma\cdot T\cdot 2^{2e}\cdot\frac{288\tilde{\sigma}^2}{(1-a)^2L^2}\log\left(\frac{2ET}{\delta}\right) = 2^{2e}\cdot\frac{1728(1+a)\tilde{\sigma}^2\gamma N}{(1-a)^3L^2}\log\left(\frac{2ET}{\delta}\right).$$

Summing over $e=0,1,\ldots,E-1$, we get the bound of total simulation cost as

$$\begin{aligned} &\sum_{e=0}^{E-1} 2^{2e}\cdot\frac{1728(1+a)\tilde{\sigma}^2\gamma N}{(1-a)^3L^2}\log\left(\frac{2ET}{\delta}\right) = \left(4^E-1\right)\cdot\frac{576(1+a)\tilde{\sigma}^2\gamma N}{(1-a)^3L^2}\log\left(\frac{2ET}{\delta}\right) \\ \leq& 4^{\lceil\log_2(NL/\epsilon)\rceil}\cdot\frac{576(1+a)\tilde{\sigma}^2\gamma N}{(1-a)^3L^2}\log\left(\frac{2ET}{\delta}\right) \leq 4^{\log_2(NL/\epsilon)+1}\cdot\frac{576(1+a)\tilde{\sigma}^2\gamma N}{(1-a)^3L^2}\log\left(\frac{2ET}{\delta}\right) \\ =& \frac{4N^2L^2}{\epsilon^2}\cdot\frac{576(1+a)\tilde{\sigma}^2\gamma N}{(1-a)^3L^2}\log\left(\frac{2ET}{\delta}\right) = \frac{2304(1+a)\tilde{\sigma}^2\gamma N^3}{(1-a)^3\epsilon^2}\log\left(\frac{2ET}{\delta}\right). \end{aligned}$$

When δ is small enough, the asymptotic simulation cost is at most

$$\frac{2304(1+a)\tilde{\sigma}^2\gamma N^3}{(1-a)^3\epsilon^2}\log\left(\frac{2ET}{\delta}\right) = \tilde{O}\left[\frac{\gamma N^3}{(1-a)^3\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

□

First-order Algorithms for the PCS-IZ Case

We first give the stochastic steepest descent method for the PCS-IZ guarantee in Algorithm 6.

Algorithm 6 Adaptive stochastic steepest descent method for PCS-IZ guarantee

Input: Model $\mathcal{X}, \mathcal{B}_Y, F(x, \xi_x)$, optimality guarantee parameter δ, indifference zone parameter c, biased subgradient estimator $H_x(y, \eta_y)$, bias ratio a.

Output: A (c, δ)-PCS-IZ solution x^* to problem (4.1).

1: Set the initial confidence half-width threshold $h \leftarrow (1-a)c/12$.
2: Set maximal number of iterations $T \leftarrow (1+a)/(1-a) \cdot 12N$.
3: Use Algorithm 4 to find an $(Nc, \delta/2)$-PGS solution.
4: **for** $t = 0, 1, \ldots, T-1$ **do**
5: **repeat** simulate $H_{x^t}(y, \eta_y)$ for all $y \in \mathcal{N}_{x^t}$
6: Compute the empirical mean $\hat{H}_{x^t}(y)$ using all simulated samples for all $y \in \mathcal{N}_{x^t}$.
7: Compute the $1 - \delta/(2T)$ confidence interval

$$\left[\hat{H}_{x^t}(y) - h_y, \hat{H}_{x^t}(y) + h_y\right], \quad \forall y \in \mathcal{N}_{x^t}.$$

8: **until** the confidence half-width $h_y \leq h$ for all $y \in \mathcal{N}_{x^t}$
9: **if** $\hat{H}_{x^t}(y) \leq -2h$ for some $y \in \mathcal{N}_{x^t}$ **then** ▷ This step takes 2^{d+1} arithmetic operations.
10: Update $x^{t+1} \leftarrow y$.
11: **else if** $\hat{H}_{x^t}(y) > -2h$ for some $y \in \mathcal{N}_{x^t}$ **then**
12: **break**
13: **end if**
14: **end for**
15: Return x^t.

The following theorem verifies the correctness of Algorithm 6 and estimates its asymptotic simulation cost.

Theorem 53. *Suppose that Assumptions 5-8, 9 hold. Algorithm 6 returns an (c, δ)-PCS-IZ solution and we have*

$$\begin{aligned} T(\delta, \mathcal{MC}_c) &= O\left[\frac{\gamma N}{(1-a)^3 c^2} \log\left(\frac{1}{\delta}\right) + \frac{\gamma N}{1-a} \max\left\{\log\left(\frac{1}{c}\right), 1\right\}\right] \\ &= \tilde{O}\left[\frac{\gamma N}{(1-a)^3 c^2} \log\left(\frac{1}{\delta}\right)\right]. \end{aligned}$$

Proof of Theorem 53. If the algorithm terminates before the T-th iteration, then the condition at Line 11 is satisfied for the last iteration point, which we denote as x^t. Let

$$y^t := \arg\min_{y \in \mathcal{N}_{x^t}} f(y).$$

Then, by the definition of confidence intervals, it holds

$$\mathbb{E}[H_{x^{t-1}}(y^t, \eta_{y^t})] \geq -3h$$

with probability at least $1 - \delta/(2T) \geq 1 - \delta$. By inequality (4.10), we know

$$\min_{y \in \mathcal{N}_{x^t}} \; f(y) - f(x^t) = f(y^t) - f(x^t) \geq -\frac{3h}{1-a} = -\frac{c}{4}$$

holds wit the same probability. We assume the event happens in the following proof. For any point $y \in \mathcal{X}$ such that $\|y - x^t\|_\infty \leq 1$, there exists two disjoint sets $\mathcal{S}_1, \mathcal{S}_2 \subset [d]$ such that

$$y = x^t + e_{\mathcal{S}_1} - e_{\mathcal{S}_2},$$

where $e_{\mathcal{S}} := \sum_{i \in \mathcal{S}} e_i$ is the indicator vector of $\mathcal{S}$. Then, using the $L^\natural$-convexity of $f(x)$, we know

$$f(y) - f(x^t) \geq f(x^t + e_{\mathcal{S}_1}) - f(x^t) + f(x^t - e_{\mathcal{S}_2}) - f(x^t) \geq -\frac{c}{2}.$$

Let $\tilde{f}(x)$ be the convex extension of $f(x)$ defined in (4.7). Recalling expression (4.6), we know

$$\tilde{f}(y) - f(x^t) \geq -\frac{c}{2}, \quad \forall y \in [1, N]^d \quad \text{s. t. } \|y - x^t\|_\infty \leq 1. \tag{4.40}$$

We assume that x^t is not the minimizer of $f(x)$, which we denote as x^*. Since the indifference zone parameter is c, we know

$$f(y) - f(x^*) \geq c, \quad \forall y \in \mathcal{X} \backslash \{x^*\}. \tag{4.41}$$

Similarly, using expression (4.6), we get

$$\tilde{f}(y) - f(x^*) \geq c, \quad \forall y \in [1, N]^d \quad \text{s. t. } \|y - x^*\|_\infty \leq 1.$$

If $\|x^t - x^*\|_\infty \leq 1$, then there exists a point x^* such that $\|x^* - x^t\|_\infty \leq 1$ and

$$f(x^*) - f(x^t) \leq -c,$$

which contradicts with inequality (4.40). Otherwise if $\|x^t - x^*\|_\infty \geq 2$, we define

$$x^{t,1} := x^t + \frac{x^* - x^t}{\|x^t - x^*\|_\infty}, \quad x^{t,2} := x^* + \frac{x^t - x^*}{\|x^* - x^t\|_\infty}.$$

Then, it holds

$$\|x^t - x^{t,1}\|_\infty = 1, \quad \|x^* - x^{t,2}\|_\infty = 1$$

and $x^{t,1}, x^{t,2}$ are closer to x^t, x^*, respectively. By inequalities (4.40) and (4.41), we get

$$\tilde{f}(x^{t,1}) - f(x^t) \geq -\frac{c}{2}, \quad \tilde{f}(x^*) - f(x^{t,2}) \leq -c.$$

However, the convexity of $\tilde{f}(x)$ on the segment $\overline{x^t x^*}$ implies that

$$-\frac{c}{2} \leq \tilde{f}(x^{t,1}) - f(x^t) \leq \tilde{f}(x^*) - f(x^{t,2}) \leq -c,$$

which is a contradiction. Hence, we know $x^t = x^*$ is the minimizer of $f(x)$. This event happens with probability at least $1 - \delta$ and therefore x^t is a (c, δ)-PCS-IZ solution.

Otherwise, we assume the algorithm terminates after T iterations. We use the induction method to prove that

$$f(x^t) - f(x^0) \leq -t \cdot \frac{(1-a)c}{12(1+a)}$$

happens with probability at least $1 - t \cdot \delta/(2T)$. For the initial point x^0, this claim holds trivially. Suppose the induction assumption is true for $x^0, x^1, \ldots, x^{t-1}$. For the $(t-1)$-th iteration, by the definition of confidence intervals, it holds

$$\mathbb{E}[H_{x^{t-1}}(x^t, \eta_{x^t})] \leq -h$$

with probability at least $1 - \delta/(2T)$. Using inequality (4.10), we know

$$f(x^t) - f(x^{t-1}) \leq -\frac{h}{1+a} = -\frac{(1-a)c}{12(1+a)}$$

holds with the same probability. Using the induction assumption for x^{t-1}, we have

$$f(x^t) - f(x^0) \leq -(t-1) \cdot \frac{(1-a)c}{12(1+a)} - \frac{(1-a)c}{12(1+a)} = -t \cdot \frac{(1-a)c}{12(1+a)}$$

holds with probability at least $1 - (t-1)\delta/(2T) - \delta/(2T) = 1 - t \cdot \delta/(2T)$. Hence, the induction assumption holds for x^t and, by the induction method, holds for all iterations. Since the algorithm terminates after T iterations, the last point x^T satisfies

$$f(x^T) - f(x^0) \leq -T \cdot \frac{(1-a)c}{12(1+a)} = -cN$$

with probability at least $1 - T \cdot \delta/(2T) = 1 - \delta/2$. Recalling the initial point x^0 is a $(cN, \delta/2)$-PGS solution, we know x^T is the optimal point with probability at least $1 - \delta$ and therefore is a (c, δ)-PCS-IZ solution.

Finally, we estimate the simulation cost of Algorithm 6. By Theorem 47, the simulation cost for generating the initial point is

$$\tilde{O}\left[\frac{\gamma N}{(1-a)^3 c^2} \log\left(\frac{1}{\delta}\right)\right].$$

For each iteration, Hoeffding bound implies that simulating

$$\frac{2\tilde{\sigma}^2}{h^2} \log\left(\frac{4T}{\delta}\right) = \frac{288\tilde{\sigma}^2}{(1-a)^2 c^2} \log\left(\frac{4T}{\delta}\right)$$

times is enough for the $1 - \delta/(2T)$ confidence half-width to be smaller than h. Hence, the total simulation for iterations is at most

$$T \cdot \gamma \cdot \frac{288\tilde{\sigma}^2}{(1-a)^2c^2} \log\left(\frac{4T}{\delta}\right) = \frac{1152\gamma\tilde{\sigma}^2(1+a)N}{(1-a)^3c^2} \log\left(\frac{4T}{\delta}\right) = O\left[\frac{\gamma N}{(1-a)^3c^2} \log\left(\frac{1}{\delta}\right)\right].$$

Combining the simulation costs for initialization and iterations, we know the asymptotic simulation cost of Algorithm 6 is at most

$$\tilde{O}\left[\frac{\gamma N}{(1-a)^3c^2} \log\left(\frac{1}{\delta}\right)\right].$$

□

Chapter 5

Stochastic Localization Simulation-optimization Methods

5.1 Introduction

In Chapter 4, we propose subgradient-based stochastic search algorithms for problems with a high-dimensional decision space. Roughly speaking, these algorithms scale well to high-dimensional problems, but are computationally expensive for large-scale problems. In practice, however, many problem settings have a large scale but a relatively low dimension or even a single dimension; see the examples in the first paragraph of Section 5.3. In this chapter, the focus is on designing algorithms that work well for *large-scale* discrete optimization via simulation problems with a convex objective function. The notion of "large-scale" refers to a large number of choices for the discrete decision variable on each dimension. Optimization problems with such features naturally arise in many operations research and management science applications, including queueing networks, supply chain networks, sharing economy operations, financial markets, etc.; see [198], [232], [10], [205, 121, 77] for example. Particularly in the area of supply chain management, a significant amount of models are proved to be discrete convex: lost-sales inventory systems with positive lead time [261]; serial inventory systems [111]; single-stage inventory systems with positive order lead time [180]; capacitated inventory systems with remanufacturing [88]; more applications are discussed in [45]. Overall, in these papers, the authors consider various decision-making settings and prove convexity for commonly used objective functions in the corresponding settings. In these applications, the convexity is proved, but finer structure such as strong convexity often does not hold or is very difficult to prove. In addition, there may be many choices of decision variables whose associated objective values are close to the optimal objective value, and the gap between optimal and sub-optimal solutions is hard to measure or estimate a priori. For the algorithms designed in this chapter, we take the view that this gap information is not available and the algorithms are designed to work for arbitrarily small unknown gap.

Similar with Chapter 4, we develop provably efficient simulation-optimization algorithms

that guarantee the (ϵ, δ)-*Probability of Good Selection* ((ϵ, δ)-PGS) criterion ; see Section 4.2 of Chapter 4, [159] and [101]. Although the asymptotic regime $\delta \ll 1$ is of more interest in many theoretical works, this chapter provides bounds on the simulation cost that hold for all $\epsilon \geq 0$ and $\delta \in (0, 1]$. To quantify the computational cost for the proposed algorithms that are guaranteed to find ϵ-optimal solutions with high probability, we take the view that the simulation cost is the dominant contributor to the computational cost; see also [160]. The simulation cost of an algorithm is measured as the total number of simulation replications run at all possible decisions visited by the algorithm until it stops. When designing algorithms to solve large-scale discrete optimization via simulation problems, the dependence of the simulation cost on the problem size (or, the number of alternatives/solutions/systems in the area of ranking and selection) is crucial to understand; see also discussions in Section 4.2 of Chapter 4 and [254].

Moreover, the subgradient descent algorithms in Chapter 4 require prior knowledge about the upper bounds on the Lipschitz constant L and the variance σ^2, and the simulation cost of the subgradient descent algorithm has a polynomial dependence on these upper bounds; see the comparison of results in Table 5.5.3. For many real-world discrete simulation via optimization problems, the Lipschitz constant and the variance are unknown and hard to estimate. As a result, both upper bounds are likely to be over-estimated, which will lead to worse simulation costs. In this chapter, algorithms that do not rely on prior information about L and σ^2 are proposed, which resolve the aforementioned issues. Moreover, the algorithms proposed in this chapter have a logarithmic dependence or no dependence on the upper bound L. Intuitively, (discrete) convex functions grow at a super-linear rate when the input goes to infinity, i.e., $\lim_{\|x\|\to\infty} |f(x)|/\|x\| = \infty$. In this case, the Lipschitz constant will be large in the regions where $\|x\|$ is large. Therefore, the upper bound on the Lipschitz constant L will be large for large-scale optimization via simulation problems and reducing the dependence on the Lipschitz constant is important. Another idea is to adaptively adjust the stepsize (or parameters that play a similar role), which is common for stochastic optimization for machine learning problems, e.g., ADAM, AdaGrad and RMSProp algorithms. However, the convergence of those algorithms is established for smooth objective functions. In our case, the Lovász extension is a non-smooth function, which prohibits the application of most adaptive methods. To the best of our knowledge, the only techniques in literature that considered a similar setting are the R-SPLINE [226] the ADALINE algorithms [189], which only provided an asymptotic convergence result.

Contributions

The major methodology in algorithm design in this chapter can be classified as *stochastic localization methods*, in the sense that we "localize" potentially near-optimal solutions in a subset and adaptively shrink the subset (denoted as $\mathcal{S}$ in our proposed algorithms) at each step. At iteration k, the stochastic localization algorithms guarantee that the set $\mathcal{S}_k$ contains an ϵ-optimal solution with high probability. At the beginning of the algorithm ($k = 0$), the set $\mathcal{S}_0$ is equal to the feasible set $\mathcal{X}$, which can be viewed as a "global" neighbourhood of

the optimal solution x^*. After several iterations, the size of $\mathcal{S}_k$ is reduced by a lot and can be viewed as a "local" neighbourhood of x^*. We describe this process as the *localization* of an approximately optimal solution. The design of algorithms relies on and addresses the challenge from the fact that the feasible set is a discrete set. Intuitively, if the feasible set has a finite number of discrete points, the subset of potentially near-optimal solutions can only be shrunk for a finite number of times, and the number of localization operations cannot exceed the size of the feasible set. The proposed algorithms generally do not rely on prior estimates of the Lipschitz constant and the variance. In addition, the simulation cost of achieving the PGS guarantee does not depend on the Lipschitz constant. We note that the expected simulation cost has an inevitable dependence on the variance σ^2. To avoid requiring prior knowledge about the variance in the Gaussian case, after designing algorithms that do not require information about the Lipschitz constant, we propose in the appendix an adaptive scheme to address the challenge of unknown variances. The idea of localization also appears in prior literature of discrete optimization [91] or more specifically discrete optimization via simulation, such as empirical stochastic branch-and-bound [238], nested partition [201] and COMPASS [102, 236]. The line search procedure in R-SPLINE [226] and ADALINE [189] is able to capture the local convexity of the objective function. However, existing works do not utilize the global information implied by the convexity structure and do not provide complexity analysis of the proposed algorithms. In contrast, we propose specially-designed algorithms for discrete convex objective functions and provide an estimate of the simulation costs; see our comparisons to the Industrial-strength COMPASS algorithm in 5.H.

To show the usefulness of the localization operation, we first consider an important case of discrete simulation via optimization problems, where the decision space is the "one-dimensional" set $\{1, 2, \ldots, N\}$. Here, N is an arbitrary positive integer that represents the problem scale. Without the convexity structure, the problem setting is mathematically equivalent to the problem of *ranking and selection*; see [101] for a comprehensive review. In this chapter, the objective function is assumed to be discrete convex on the decision space, but no other structure information such as strong convexity or the knowledge of a minimal gap between the optimal and sub-optimal solutions is known. Utilizing the idea of *localization*, we overcome the shortcoming of the subgradient descent algorithm that its simulation cost has a quadratic dependence on the problem scale. We propose two localization algorithms. As a natural generalization of the classical bi-section algorithm, we design the tri-section sampling (TS) algorithm to find an (ϵ, δ)-PGS solution. We prove that, when δ is small, $O(\log(N)\epsilon^{-2}\log(1/\delta))$ serves as an upper bound on the simulation cost for the TS algorithm for any one-dimensional convex problem, which represents the same logarithmic dependence on the scale as the bi-section algorithm. Note that when the convexity structure is not exploited, the optimal dependence on N can be linear. We then design the shrinking uniform sampling (SUS) algorithm that beats the TS algorithm. The SUS algorithm is proved to enjoy the upper bound on the simulation cost as $O[\epsilon^{-2}(\log(N) + \log(1/\delta))]$ when δ is small. Using the asymptotic criterion in [128], namely $\delta \to 0$ with other parameters fixed, the SUS algorithm asymptotically achieves the optimal performance and, therefore, is the first algorithm to achieve a matching upper bound on simulation costs for ranking and selection

problems with general convex structure. This theoretical superiority of the SUS algorithm is also verified in numerical experiments. We remark that our major contribution is the SUS algorithm rather than the TS algorithm, though the analysis provided for these two algorithms may be separately useful in broader settings.

Next, we turn to the settings of large-scale multi-dimensional problems with the "d-dimensional" discrete decision space $\{1, 2, \ldots, N\} \times \{1, 2, \ldots, N\} \times \ldots \times \{1, 2, \ldots, N\}$. We note that the scale N can easily be relaxed to be different in each dimension in our algorithm design (e.g., after linear constraints are applied on the decision space), but we unify the use of N in each dimension in the analysis, so as to clearly demonstrate the impact of the scale N. In a multi-dimensional decision space, a common definition of discrete convexity, which guarantees that a local optimum is globally optimal, is the $L^\natural$-convexity [168]; see [66, 77] for examples of $L^\natural$-convex functions. We observe that even though the TS algorithm and the SUS algorithm designed for one-dimensional problems can be extended to the multi-dimensional case, the dependence of their simulation cost on the dimension d can be large, even up to an exponential order of dependence, which may prohibit their practical use in high-dimensional problems. This motivates us to consider alternative approaches to design stochastic localization algorithms that have a low dependence on the dimension d.

In this chapter, we combine the idea of localization with the *subgradient information* in the multi-dimensional case. The subgradient information is constructed by taking simulation samples and plays a crucial role in reducing the dependence of simulation cost on the dimension d. The cutting-plane methods [221, 20, 140, 123] are based on a similar idea and are known to have a lower order or no dependence on the Lipschitz constant. However, the cutting-plane methods are not robust to noise. Therefore, we develop a novel framework to design stochastic cutting-plane (SCP) algorithms based on deterministic cutting-plane algorithms, with the goal of achieving the PGS guarantee. A novel stochastic separation oracle is designed and analyzed. A straightforward application of the proposed framework leads to SCP algorithms that have an $O(d^3)$ dependence on the dimension and a logarithmic dependence on L, which improves the quadratic dependence of the subgradient-based algorithms in Chapter 4.

Utilizing the discrete nature of the problem, we further develop the dimension reduction algorithm whose simulation cost is upper bounded by a constant that is independent of Lipschitz constant L and has an $O(d^4)$ dependence on the dimension. This is the first algorithm for discrete optimization via simulation that utilizes the convex structure of the objective to reduce the simulation cost and does not require knowledge about the Lipschitz constant L. In contrast, the subgradient-based search algorithms developed in Chapter 4 has a higher order dependence on L and requires the knowledge about the Lipschitz constant, although it has a lower dependence ($O(d^2)$) on the dimension compared to the dimension reduction algorithm. Our developed SCP algorithms may particularly be preferable when the Lipschitz parameter L for a given problem is large or hard to estimate. When a prior estimate of the Lipschitz constant is unavailable, we need to estimate the Lipschitz constant through the stochastic oracle and this leads to two major difficulties. First, the objective function can only be evaluated with noise. This means that we need to simulate $F(x, \xi_x)$

a number of times to get a considerably accurate estimate of the Lipschitz constant and this process can be time-consuming. Second, we need to check a large number of points to estimate the Lipschitz constant. Even in the local neighbourhood $\{y \in \mathbb{Z}^d \mid \|y - x\|_\infty \leq 1\}$, we need to evaluate at least $O(2^d)$ points to get an estimate of the Lipschitz constant. The simulation cost will be prohibitively large even when d is as low as 50. The idea of gradually reducing the problem dimension was proposed in parallel in [122], where the author made the algorithm more practical by reducing the number of arithmetic operations to be polynomial. We numerically verify that the dimension reduction algorithm has a better performance than the subgradient descent algorithm (Algorithm 3) both on the synthetic and the queueing simulation optimization examples, especially for the large-scale case.

In terms of dependence on the scale N, we theoretically show that the subgradient descent algorithm and the SCP algorithms all present an $O(N^2)$ dependence on N for their simulation costs. However, the SCP algorithms empirically perform better than the subgradient descent algorithm in applications where N is large. On the other hand, the SUS algorithm, when extended to multi-dimensional problems, still present no dependence on N under the asymptotic criterion [128], but however incurs an exponential dependence on d. These analyses can assist practitioners to choose which algorithm to use depending on the knowledge or partial knowledge on d, N and L in the specific problems.

Finally, we propose a novel algorithm that is able to adaptively estimate the variance of the randomness at each feasible decision in the case when the noise is Gaussian. Adaptive variants are highly important since the variance is not known in many real-life applications. The design of the algorithm is based on the property that the lower tail for χ^2-random variables is sub-Gaussian [224]. The adaptive algorithm is suitable for the case when an upper bound on the variance is hard to estimate and over-estimation is inevitable. In addition, the adaptive algorithm provides an approach to improve the simulation cost in the case when location-dependent upper bounds of the variance σ_x^2 is available for all feasible decisions x. This is because the uniform upper bound $\sigma^2 = \max_{x \in \mathcal{X}} \sigma_x^2$ is in general attained by extreme choices of the decision variable and may be much larger than the variance of a large proportion of feasible decisions. In contrast to common two-stage procedures for the unknown variance case in ranking and selection literature, the proposed adaptive algorithm does not require simulating all choices of the decision variable (which requires $O(N^d)$ simulations) to get an upper bound on the variance. Moreover, using the novel algorithm, the simulation cost is at most increased by a constant factor compared to the known variance case.

The remainder of this chapter is outlined as follows. Section 5.2 introduces the model, framework, optimality criterion, and simulation costs. Section 5.3 discusses the algorithms and performance analysis developed for one-dimensional large-scale problems. Section 5.4 discusses the algorithms and performance analysis developed for multi-dimensional large-scale problems. Section 5.5 provides numerical experiments to compare the proposed algorithms to benchmark methods. The adaptive algorithm for estimating the variance is provided in the appendix.

5.2 Model and Framework

Since we use the same framework as Chapter 4, we only discuss essential notations and omit the detailed discussion of the framework. We consider a complex stochastic system that involves discrete decision variables in a d-dimensional subspace $\mathcal{X} = [N_1] \times [N_2] \times \cdots \times [N_d]$ in which the N_i's are positive integers. The objective function $f(x)$ for $x \in \mathcal{X}$ is given by

$$f(x) := \mathbb{E}[F(x, \xi_x)],$$

in which ξ_x is a random object belonging to the probability space $(\mathsf{Y}, \mathcal{B}_\mathsf{Y})$ and $F : \mathcal{X} \times \mathsf{Y} \to \mathbb{R}$ is a measurable function. Specifically, the function F captures the full operations logic in the stochastic system and measures the performance of the system. For example, in a queueing system, ξ_x is the arrival times and the service times of customers, and $F(\cdot, \xi_x)$ is the average waiting time of all customers under the situation described by ξ_x. We consider scenarios when the objective function $f(x)$ is not in closed-form and needs to be evaluated by averaging over simulation replications of $F(x, \xi_x)$. The random objects ξ_x's can be different for different choices of decision variables. We assume that the probability distribution for the stochastic simulation output $F(x, \xi_x)$ is sub-Gaussian.

Assumption 5. The distribution of $F(x, \xi_x)$ is sub-Gaussian with known parameter σ^2 for any $x \in \mathcal{X}$.

We note that a special case of Assumption 5 is when the distribution follows the Gaussian distribution. In that case, the parameter σ^2 can be chosen as the upper bound on the variance of the distribution. For more general distributions with a finite variance, the mean estimator in [139] can be used in place of the empirical mean estimator and the results in this chapter can be directly generalized. We assume that Assumption 5 holds in the remainder of the chapter except Section 5.G, where we consider the Gaussian noise case when the variance is unknown. We propose a novel algorithm to adaptively estimate the variance σ^2 in the Gaussian case.

Same as Chapter 4, we focus on identifying the optimal decision, i.e., finding the decision that has the minimal objective value:

$$\min_{x \in \mathcal{X}} \ f(x). \tag{5.1}$$

Our general goal is to design algorithms that guarantee the selection of a good decision that yields a close-to-optimal performance with high probability. Formally, this criterion is defined as *Probability of Good Selection*. We reiterate the definition of this criterion:

- (ϵ, δ)-*Probability of good selection* (PGS). The solution x returned by an algorithm has an objective value at most ϵ larger than the optimal objective value with probability at least $1 - \delta$.

In this chapter, we focus on the regime where δ is small enough and estimate the asymptotic expected simulation cost.

In addition, we assume that the objective function and the feasible set are both $L^\natural$-convex. We describe in detail in Section 4.2 of Chapter 4 the exact definition and properties of $L^\natural$-convexity.

Assumption 6. The objective function $f(x)$ is a $L^\natural$-convex function on the $L^\natural$-convex set $\mathcal{X}$.

For optimization via simulation problems, we take the view that the simulation cost of generating replications of $F(x, \xi_x)$ is the dominant contributor to the computational cost is widely held; see [155, 176, 159, 160]. Therefore, for the purpose of comparing different simulation-optimization algorithms that satisfy certain optimality guarantee, the performance of each algorithm is measured by the *expected simulation cost* (see Definition 11). The main focus of this chapter is to develop provably efficient simulation-optimization algorithms for the (ϵ, δ)-PGS guarantee and provide an upper bound on the expected simulation cost to achieve that guarantee. Therefore, the notion of simulation cost in this chapter is largely focused on

$$T(\epsilon, \delta, \mathcal{MC}) := T((\epsilon, \delta)\text{-}PGS, \mathcal{MC}),$$

where the class of models $\mathcal{MC}$ include all convex models. We mention that the upper bounds derived in this chapter also hold almost surely, while the lower bounds only hold in expectation. Furthermore, our proposed algorithms do not require additional structures of the selection problem in addition to convexity.

To better present the dependence of the expected simulation cost on the scale and dimension of the problem, we assume that $N_1 = N_2 = \cdots = N_d$.

Assumption 7. The feasible set of decision variables is $\mathcal{X} = [N]^d$, where $N \geq 2$ and $d \geq 1$.

With Assumption 7 in hand, we will present the dependence of the expected simulation cost on N and d. We note that the results in this chapter can be naturally extended to the case when each dimension has a different number of feasible choices of decision variables. Furthermore, if the objective function f is defined on a $L^\natural$-convex set (i.e., the indicator function of the set is a $L^\natural$-convex function, which we will define later), the algorithms proposed in this chapter can be directly extended with small modifications. A typical example of a $L^\natural$-convex set is the capacity-constrained set

$$\left\{(x_1, \ldots, x_d) \;\middle|\; x_i \in [N],\ \forall i \in [d],\ \textstyle\sum_i x_i \leq C\right\}$$

under a linear transform, where $C > 0$ is the capacity constraint; see Section 5.5 for more details.

5.3 Simulation-optimization Algorithms and Complexity Analysis: One-dimensional Case

We first consider a special class of optimization via simulation problems where the dimension of the decision variable is one, but there are a large number of choices of decision variable. This class of one-dimensional problems, despite of the less generality compared to multi-dimensional large-scale problems, have applications when the one-dimensional decision variable is a choice of overall resource level. For example, large delivery companies often need to decide the total number of trucks that should be recruited for operations in a self-contained region. A service system may need to decide the total number of staff members needed to host a special event. Such decisions often involve a trade-off between service satisfaction and resource costs. The convexity in the objective function often comes from the marginal decay of contribution to service satisfaction as the resource level increases; see the optimal allocation example and Figure 5.5.1 in Section 5.5 for more details.

In the one-dimensional case, the feasible set is $\mathcal{X} = [N] = \{1, 2, \ldots, N\}$. The $L^\natural$-convexity for a function f reduces to the ordinary continuous convexity through the discrete mid-point convexity property, namely,

$$f(x+1) + f(x-1) \geq 2f(x), \quad \forall x \in \{2, \ldots, N-1\}.$$

If the function $f(x)$ is convex on $\mathcal{X}$, it has a convex linear interpolation on the continuous interval $[1, N]$, defined as

$$\tilde{f}(x) := [f(x_0+1) - f(x_0)] \cdot (x - x_0) + f(x_0), \quad \forall x \in [x_0, x_0+1],\ x_0 \in [N-1]. \tag{5.2}$$

In this section, we propose simulation-optimization algorithms that are guaranteed to find solutions that satisfy the PGS guarantee, provided that the objective function has a convex structure. For every developed simulation-optimization algorithm, we provide an upper bound on the expected simulation cost to achieve the PGS guarantee. We also provide a lower bound on the expected simulation cost that reflects the best achievable performance for any algorithm. Under the asymptotic criterion in [128], one of our proposed algorithms can attain the best achievable asymptotic performance.

In contrast to the multi-dimensional case in Chapter 4, where the subgradient descent algorithm achieves satisfying performance, the subgradient descent algorithm is not efficient for large-scale one-dimensional problems. This is because of the $O(N^2)$ dependence in the simulation cost. In addition, the subgradient descent algorithm relies on the Lipschitz constant of the objective function, which is shown to be unnecessary for discrete problems in this section. Utilizing the localization operation, the algorithms proposed in this section do not have the aforementioned issues. Therefore, the algorithms in this section provide better alternatives to the subgradient descent algorithm for one-dimensional problems. The analysis of the one-dimensional case also shows the limitation of subgradient-based search methods and provides a hint on how to improve algorithms for multi-dimensional problems.

Tri-section Sampling Algorithm and Upper Bound on Expected Simulation Cost

We first propose the *tri-section sampling* (TS) algorithm for the PGS guarantee. The idea of the TS algorithm is from the classical bi-section method and the golden section method. A similar TS algorithm is proposed in [2] for stochastic continuous convex optimization, which controls the regret instead of the objective value. However, their algorithm does not utilize the prior information that the optimal solution is an integral point and thus the simulation cost has a polynomial dependence on the Lipschitz constant. In addition, although an algorithm that minimizes the regret can be used to minimize the objective function value, the resulting simulation cost may be larger than that of specialized optimization algorithms and has an inferior dependence on the dimension d in the multi-dimensional case. The pseudo-code of the proposed TS algorithm is listed in Algorithm 7. The 3-quantiles of an interval $[L, U]$ are $(2L+U)/3$ and $(L+2U)/3$. Since we are looking for integral solutions, we round these quantiles to integers. In the procedure of Algorithm 7, one step is to compute confidence intervals that satisfy certain confidence guarantees. We now provide one feasible approach to construct such confidence intervals, which is based on Hoeffding's inequality for sub-Gaussian random variables. Define

$$h(n, \sigma, \alpha) := \sqrt{\frac{2\sigma^2}{n} \log\left(\frac{2}{\alpha}\right)}.$$

Recall that σ^2 is the upper bound on the sub-Gaussian parameters of all choices of decision variables. With this function $h(\cdot)$ in hand, whenever n independent simulations of the decision x are available, one can construct a $(1-\alpha)$-confidence interval for $f(x)$ as

$$\left[\hat{F}_n(x) - h(n, \sigma, \alpha), \hat{F}_n(x) + h(n, \sigma, \alpha)\right].$$

This is because the distribution of the empirical mean $\hat{F}_n(x)$ is sub-Gaussian with parameter σ^2/n and Hoeffding's inequality gives

$$\mathbb{P}\left[|\hat{F}_n(x) - f(x)| > h(n, \sigma, \alpha)\right] \leq 2\exp\left(-\frac{nh(n, \sigma, \alpha)^2}{2\sigma^2}\right) = \alpha.$$

If the variance σ_x^2 of a single choice of decision variable x is known, the confidence interval may be sharpened by replacing σ with σ_x; see Section 5.G. We note that the analysis in this chapter can be generalized to more general distributions, such as the sub-exponential distributions, by replacing $h(n, \sigma, \alpha)$ with other concentration bounds.

Algorithm 7 Tri-section sampling algorithm for the PGS guarantee

Input: Model $\mathcal{X} = [N], (\mathsf{Y}, \mathcal{B}_{\mathsf{Y}}), F(x, \xi_x)$, optimality guarantee parameters ϵ, δ.
Output: An (ϵ, δ)-PGS solution x^* to problem (5.1).

Set upper and lower bounds of the current interval $x_L \leftarrow 1, x_U \leftarrow N$.
Set maximal number of comparisons $T_{max} \leftarrow \log_{1.5}(N) + 2$.
while $x_U - x_L > 2$ **do** ▷ Iterate until there are at most 3 points.
    Compute 3-quantiles of the interval $q_{1/3} \leftarrow \lfloor 2x_L/3 + x_U/3 \rfloor$ and $q_{2/3} \leftarrow \lceil x_L/3 + 2x_U/3 \rceil$.
    Simulate n independent copies of $F(q_{1/3}, \xi_{1/3})$ and $F(q_{2/3}, \xi_{2/3})$, where n is the smallest integer such that $h[n, \sigma, 1 - \delta/(2T_{max})] \leq \epsilon/8$.
    Compute the empirical means $\hat{F}_n(q_{1/3}), \hat{F}_n(q_{2/3})$.
    if $\hat{F}_n(q_{1/3}) - \epsilon/8 \geq \hat{F}_n(q_{2/3}) + \epsilon/8$ **then**
        Update $x_L \leftarrow q_{1/3}$.
    else if $\hat{F}_n(q_{1/3}) + \epsilon/8 \leq \hat{F}_n(q_{2/3}) - \epsilon/8$ **then**
        Update $x_U \leftarrow q_{2/3}$.
    else
        Update $x_L \leftarrow q_{1/3}$ and $x_U \leftarrow q_{2/3}$.
    end if
end while
Simulate $\tilde{n}$ independent copies of $F(x, \xi_x)$ for each $x \in \{x_L, \ldots, x_U\}$, where $\tilde{n}$ is the smallest integer such that $h[\tilde{n}, \sigma, 1 - \delta/(2T_{max})] \leq \epsilon/2$. ▷ Now $x_U - x_L \leq 2$.
Return the point in $\{x_L, \ldots, x_U\}$ with the minimal empirical mean.

Intuitively, the algorithm iteratively shrinks the size of the set containing a potentially near-optimal choice of decision variables. We provide an example of the TS algorithm in Figure 5.3.1. In this example, we suppose that the current set is [10] and then, the two 3-quantiles are 4 and 7. Without loss of generality, we assume that $\epsilon_0 := f(7) - f(4) \geq 0$ and the global minimum is in the left set [4]. We consider two different cases. First, if we know that $\epsilon_0 > 0$ holds with high probability, no solution in $\{7, \ldots, 10\}$ can be a global optimum and we can shrink the set to [6]. On the other hand, if we know that $\epsilon_0 = O(\epsilon)$ holds with high probability, we can construct a linear lower bound for the objective function in [4] and $\{7, \ldots, 10\}$. In the both sets, the decrease of the lower bound is at most ϵ_0. Therefore, we have the relation $\min_{x \in \{4,5,6,7\}} f(x) \leq \min_{x \in [10]} f(x) + \epsilon_0$, which implies that $\{4, \ldots, 7\}$ contains ϵ_0-optimal solutions with high probability and we can shrink the set to $\{4, \ldots, 7\}$ in the next iteration.

The algorithm shrinks the length of the current interval by at least 1/3 for each iteration. Thus, the total number of iterations is at most $O(\log_{1.5}(N))$ to shrink the set until there are at most 3 points. Then, the algorithm solves a sub-problem with at most 3 points. We can prove that Algorithm 7 achieves the PGS guarantee for any given convex problem without knowing further structural information, i.e., Algorithm 7 is an $[(\epsilon, \delta)\text{-PGS}, \mathcal{MC}]$-algorithm. By estimating the simulation cost of the algorithm, an upper bound on the expected simulation cost to achieve the PGS guarantee follows.

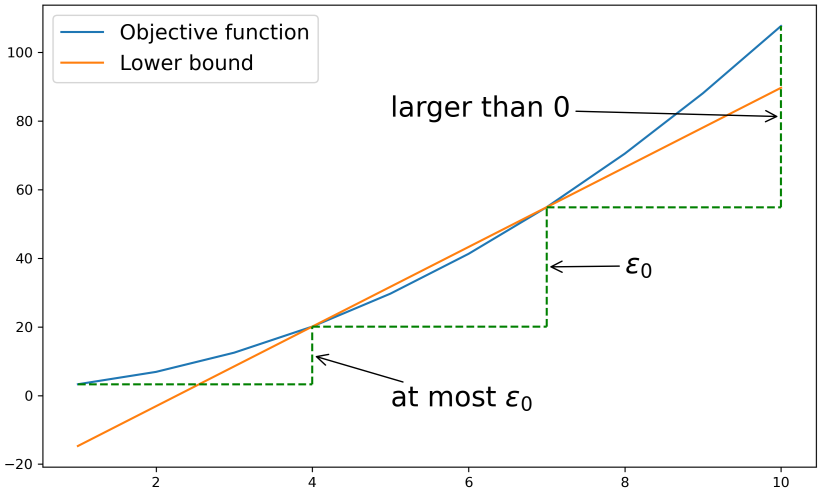


Figure 5.3.1: An example of the iteration of the TS algorithm.

Theorem 54. *Suppose that Assumptions 5-7 hold. Algorithm 7 is an* $[(\epsilon, \delta)$*-PGS,* $\mathcal{MC}]$*-algorithm. Furthermore, we have*

$$T(\epsilon, \delta, \mathcal{MC}) = O\left[\frac{\log(N)}{\epsilon^2}\log\left(\frac{\log(N)}{\delta}\right) + \log(N)\right] = \tilde{O}\left[\frac{\log(N)}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

We provide an explanation on the additional $\log(N)$ term. We note that in practice, the number of simulation samples taken in each iteration must be an integer, while the simulation cost is treated as a real number in our complexity analysis. Hence, the practical simulation cost of each iteration should be the smallest integer larger than the theoretical simulation cost, which introduces an extra $O(1)$ term. Then, the total expected simulation cost of Algorithm 7 should contain an extra $O(\log(N))$ term, which is not related to δ and is relatively small compared to the main term when δ is small.

Remark 3. The term in the $\tilde{O}(\cdot)$ notation reflects the asymptotic simulation cost when $\delta \to 0$. The asymptotic simulation cost is commonly used in multi-armed bandits literature to compare the computational complexities of different algorithms [136, 30, 127, 119, 42, 128]. In practice, the failing probability δ is usually not small enough to enter the asymptotic regime and thus the simulation cost of algorithms may deviate from the asymptotic simulation cost. Therefore, we provide both the non-asymptotic and the asymptotic simulation costs for all algorithms.

Shrinking Uniform Sampling Algorithm and Upper Bound on Expected Simulation Cost

We have shown that the expected simulation cost of TS algorithm for the PGS guarantee has a $\log(N)$ dependence on N. Then, one may naturally ask: is there any algorithm for the PGS guarantee whose simulation cost has a better dependence on N? The answer is affirmative. In this subsection, the *shrinking uniform sampling* (SUS) algorithm for the PGS guarantee is proposed, which is proven to have a simulation cost as $O[\epsilon^{-2}(\log(N)+\log(1/\delta))]$, which grows as $\epsilon^{-2}\log(1/\delta)$ in the asymptotic regime $\delta \to 0$. Similarly, utilizing the idea of localization, the SUS algorithm maintains a set of active points and shrinks the set in each iteration until there are at most 2 points. However, instead of only sampling at 3-quantiles points of the current interval, the SUS algorithm samples all points in the current active set but with much fewer simulations. We give the pseudo-code in Algorithm 8.

Algorithm 8 Shrinking uniform sampling algorithm for the PGS guarantee

Input: Model $\mathcal{X} = [N], (\mathsf{Y}, \mathcal{B}_{\mathsf{Y}}), F(x, \xi_x)$, optimality guarantee parameters ϵ, δ.
Output: An (ϵ, δ)-PGS solution x^* to problem (5.1).

Set the active set $\mathcal{S} \leftarrow \mathcal{X}$.
Set the step size $s \leftarrow 1$, maximal number of comparisons $T_{max} \leftarrow N$.
Set number of samples $n_x \leftarrow 0$ simulated at x for all $x \in \mathcal{X}$.
while the size of $\mathcal{S}$ is at least 3 **do** ▷ Iterate until $\mathcal{S}$ has at most 2 points.
    for $x \in \mathcal{S}$ **do**
        Simulate independent copies of $F(x, \xi_x)$ such that $h[n_x, \sigma, 1-\delta/(2T_{max})] \le |\mathcal{S}| \cdot \epsilon/80$.
    end for
    Compute the empirical mean (using all simulated samples) $\hat{F}_{n_x}(x)$ for all $x \in \mathcal{S}$.
    if $\hat{F}_{n_x}(x) + h[n_x, \sigma, 1-\delta/(2T_{max})] \le \hat{F}_{n_y}(y) - h[n_y, \sigma, 1-\delta/(2T_{max})]$ for some $x, y \in \mathcal{S}$ **then**
        ▷ **Type-I Operation**
        if $x < y$ **then**
            Remove all points $z \in \mathcal{S}$ with the property $z \ge y$ from $\mathcal{S}$.
        else
            Remove all points $z \in \mathcal{S}$ with the property $z \le y$ from $\mathcal{S}$.
        end if
    else ▷ **Type-II Operation**
        Update the step size $s \leftarrow 2s$.
        Update $\mathcal{S} \leftarrow \{x_{min}, x_{min} + s, \ldots, x_{min} + ks\}$, where $x_{min} = \min_{x \in \mathcal{S}} x$ and $k = \lceil |\mathcal{S}|/2 \rceil - 1$.
    end if
end while ▷ Now $\mathcal{S}$ has at most 2 points.
Simulate $\tilde{n}$ independent copies of $F(x, \xi_x)$ for each $x \in \{x_L, \ldots, x_U\}$, where $\tilde{n}$ is the smallest integer such that $h[\tilde{n}, \sigma, 1-\delta/(2T_{max})] \le \epsilon/4$.

21: Return the point in $\mathcal{S}$ with minimal empirical mean.

There are two kinds of shrinkage operations in Algorithm 8, which we denote as Type-I and Type-II Operations. Intuitively, Type-I Operations are implemented when we can compare and differentiate the function values of two points with high probability, and Type-II Operations are implemented when all points have similar function values. In the latter case, we prove that there exists a neighboring point to the optimum that has a function value at most $\epsilon/2$ larger than the optimum. Hence, we can discard every other point in $\mathcal{S}$ (the set in the algorithm that contains a potential good selection) with at least one $\epsilon/2$-optimal point remaining in the active set. We give a rough estimate to the expected simulation cost of Algorithm 8. We assign an order to points in $\mathcal{X}$ by the time they are discarded from $\mathcal{S}$. Points discarded in the same iteration are ordered randomly. Then, for the last k-th discarded point x_k, there are at least k points in $\mathcal{S}$ when x_k is discarded. By the second termination condition in Line 17, the confidence half-width at x_k is at least $k\epsilon/80$. If the Hoeffding bound is used, simulating $\tilde{O}(\epsilon^{-2}k^{-2}\log(1/\delta))$ times is enough to achieve the confidence half-width. Recalling the fact that $\sum_k k^{-2} < \pi^2/6 = O(1)$, if we sum the simulation cost over $k \in [N]$, the total expected simulation cost is bounded by $\tilde{O}(\epsilon^{-2}\log(1/\delta))$ and is independent of N. We note that we are able to reuse the samples in the previous rounds since we use the union bound to bound the total failing probability across iterations, which does not require the independence between samples in different iterations. In addition, we mention that the bound $|\mathcal{S}|\epsilon/80$ in lines 10 and 17 is not optimal and we choose this bound since the proof is simpler using this upper bound, and the expected simulation cost is only a constant factor worse than that of the case when the optimal bound is chosen. The following theorem proves that Algorithm 8 indeed achieves the PGS guarantee for any convex problem and provides a rigorous upper bound on the expected simulation cost $T(\epsilon, \delta, \mathcal{MC})$.

Theorem 55. *Suppose that Assumptions 5-7 hold. Algorithm 8 is an* $[(\epsilon, \delta)$*-PGS,* $\mathcal{MC})]$*-algorithm. Furthermore, we have*

$$T(\epsilon, \delta, \mathcal{MC}) = O\left[\frac{1}{\epsilon^2}\log\left(\frac{N}{\delta}\right) + N\right] = \tilde{O}\left[\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

If we consider the asymptotic regime $\delta \ll 1$ (which is considered in [128]), the expected simulation cost of the SUS algorithm grows as $\epsilon^{-2}\log(1/\delta)$. This dependence is asymptotic and holds in the sense that the required failing probability δ tends to be very small. When δ is moderately large, the cost can depend on N. We demonstrate in the numerical experiments this asymptotic independence.

Lower Bound on Expected Simulation Cost

In this subsection, we consider the lower bounds on the expected simulation costs for all of the simulation-optimization algorithms that satisfy certain optimality guarantee for general

convex problems. The lower bounds show the fundamental limit behind the simulation-optimization algorithms for general selection problems with a convex structure. By comparing those lower bounds with the upper bounds established for specific simulation-optimization algorithms, we can conclude that the SUS algorithm is optimal up to a constant factor. The lower bound on $T(\epsilon, \delta, \mathcal{MC})$, i.e., the expected simulation cost for achieving the PGS guarantee, is derived in Corollary 5, which is also presented in the following corollary.

Corollary 4. *Suppose that Assumptions 5-7 hold. We have*

$$T(\epsilon, \delta, \mathcal{MC}) \geq \Theta\left[\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

Combining with the upper bounds derived in this section, we conclude that the TS algorithm has a $\log(N)$ order gap for the PGS guarantee, while the SUS algorithm is optimal up to a constant in the asymptotic regime $\delta \ll 1$. However, the space complexities of the TS algorithm and the SUS algorithms are $O(\log(N))$ and $O(N)$, respectively. The space complexity of an algorithm refers to the amount of memory used by a program to execute the algorithm. This observation implies that we need to consider the trade-off between the simulation cost and the space complexity when choosing the best algorithm.

Before concluding this section, we note that the subgradient descent algorithms in Chapter 4 requires the knowledge of the Lipschitz constant and has a simulation cost as

$$\tilde{O}\left[N^2\epsilon^{-2}\log(1/\delta)\right],$$

which is $O(N^2)$ larger than that of the TS and the SUS algorithms. This observation implies that subgradient-based search methods may not be able to fully utilize the discrete nature and the convex structure of problem (5.1), especially for low-dimensional problems. Therefore, the proposed algorithms in this section provide a non-trivial improvement for solving one-dimensional convex optimization via simulation problems and hint a potential improvement direction (namely, localization methods) for multi-dimensional problems.

5.4 Simulation-optimization Algorithms and Complexity Analysis: Multi-dimensional Case

In this section, we propose simulation-optimization algorithms to achieve the PGS guarantee for convex discrete optimization via simulation problems with multi-dimensional decision variables. The decision space is considered as $\mathcal{X} = [N]^d$. In the multi-dimensional case, the discrete convexity of f is defined by the $L^\natural$-convexity (Definition 12). The $L^\natural$-convexity can lead to the property that the discrete convex function has a convex extension along with an explicit subgradient defined on the convex hull of $\mathcal{X}$. We refer the readers to Section 4.2 for the definition and properties of $L^\natural$-convex functions.

We outline the intuition underlying the algorithm design of this section before discussing the details. Since we have observed the power of localization from the one-dimensional case,

the major approach is to design multi-dimensional algorithms based on the same idea. The first idea of applying the localization technique is to extend the TS algorithm to the multi-dimensional case. A direct generalization of the TS algorithm results in the zeroth-order stochastic ellipsoid method [2] and the zeroth-order random walk method [148], whose computational complexities have $O(d^{33})$ and $O(d^{14})$ dependence on the dimension, respectively. On the other hand, we show in the appendix that the SUS method can be naturally extended to the multi-dimensional case. The multi-dimensional SUS algorithm also has an expected simulation cost independent of the scale N using the asymptotic criterion in [128] (i.e., when δ is sufficiently small). However, the expected simulation cost has an exponential dependence on the dimension d and, therefore, the SUS algorithm is only suitable for low-dimensional problems.

We thus take an alternative approach and combine the localization operation with the *subgradient information*, which is known to be useful for high-dimensional problems. In this chapter, we design stochastic cutting-plane methods, which utilize properties of $L^\natural$-convex functions and the Lovász extension to evaluate unbiased stochastic subgradients at each point via finite difference. More specifically, we develop a new framework to design stochastic cutting-plane methods and thus reduce the dependence of the simulation cost on d. A straightforward application of our proposed framework leads to stochastic cutting-plane methods whose simulation cost has a $O(d^3)$ dependence on d. In addition, stochastic cutting-plane methods have only a logarithmic dependence on the Lipschitz constant L, while the subgradient-based algorithm (Algorithm 3) has a higher-order dependence on L. Further utilizing the *discrete nature* of problem (5.1), we develop the dimension reduction algorithm, whose simulation cost is upper bounded by a constant that is independent of the Lipschitz constant. In addition, the dimension reduction algorithm does not require any prior knowledge about the Lipschitz constant, which makes it suitable for the case when prior knowledge about the objective function is limited.

Stochastic Cutting-plane Methods: Stochastic Separation Oracles

Now, we consider designing simulation-optimization algorithms with simulation costs having a polynomial dependence on the problem parameters d and N. In addition, we reiterate that the goal is to design algorithms that do not require the information about the Lipschitz constant L and the simulation cost is upper bounded by a constant that is independent of L. Intuitively, the subgradient information is useful for high-dimensional problems, while the localization operation is good at utilizing the discrete nature of the problem and getting rid of the dependence on the Lipschitz constant. Therefore, one may expect subgradient-based localization methods to satisfy the aforementioned requirements. Using the definitions and tools introduced in Section 4.2, we are able to design the desired algorithm in two steps. In this subsection, we first introduce the definition of stochastic separation oracles and give a novel framework to design stochastic cutting-plane methods via deterministic cutting-plane methods. Straightforward extensions of deterministic cutting-

plane methods require prior knowledge about L and the simulation cost has a logarithmic dependence on L. Hence, the following assumption is required.

Assumption 11. The ℓ_∞-Lipschitz constant L is known a priori. Namely, we have

$$|f(x) - f(y)| \leq L, \quad \forall x, y \in \mathcal{X}, \quad \text{s.t. } \|x - y\|_\infty \leq 1.$$

In the next subsection, we incorporate the stochastic cutting-plane methods with the dimension reduction operation. The resulting algorithm, named as the dimension reduction algorithm, does not require prior information about L and the simulation cost is upper bounded by a constant that is independent of L. We note that the design of the dimension reduction algorithm is the main objective of this section and stochastic cutting-plane methods mainly serve as an example of our novel framework.

In each iteration of a cutting-plane algorithm, a cutting hyperplane is generated to shrink the subset of potentially optimal choices of decision variables. In other words, the cutting hyperplane is used to localize the optimal solution. When the volume is small enough, the Lipschitz continuity implies that the all points in the polytope have their objective values close to the optimal value. In general, cutting-plane methods have a higher order of dependence on the dimension than subgradient-based search methods [140, 123]. Hence, we expect that the simulation cost of cutting-plane methods will have a higher-order dependence on the problem dimension compared to that of subgradient-based search methods. As a counterpart of separation oracles, we introduce the stochastic separation oracle, named as the (ϵ, δ)-separation oracle, to characterize the accuracy of separation oracles in the stochastic case.

Definition 15. A *(ϵ, δ)-separation oracle ((ϵ, δ)-$\mathcal{SO}$)* is a function on $[1, N]^d$ with the property that for any input $x \in [1, N]^d$, it outputs a stochastic vector $\hat{g}_x \in \mathbb{R}^d$ such that the inequality

$$f(y) \geq f(x) - \epsilon, \quad \forall y \in [1, N]^d \cap H$$

holds with probability at least $1-\delta$, where the half space H is defined as $\{z : \langle \hat{g}_x, z - x \rangle \geq 0\}$.

Before we state algorithms, we give a concrete example of (ϵ, δ)-$\mathcal{SO}$ oracles and provide an upper bound on the expected simulation cost of evaluating each oracle. We define the averaged subgradient estimator $\hat{g}^n$ as

$$\hat{g}^n_{\alpha_x(i)} := \hat{F}_n\left(S^{x,i}\right) - \hat{F}_n\left(S^{x,i-1}\right), \quad \forall i \in [d], \tag{5.3}$$

where α_x is a consistent permutation of x, $n \geq 1$ is the number of samples, and $\hat{F}_n$ is the empirical mean of n independent evaluations of F. The following lemma gives a lower bound on n to guarantee that $\hat{g}^n$ is an (ϵ, δ)-$\mathcal{SO}$ oracle.

Lemma 35. *Suppose that Assumptions 5-7 hold. If we choose n such that*

$$n = \Theta\left[\frac{dN^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right],$$

then $\hat{g}^n$ is an (ϵ, δ)-$\mathcal{SO}$ oracle. Moreover, the expected simulation cost of generating an (ϵ, δ)-$\mathcal{SO}$ oracle is at most

$$O\left[\frac{d^2N^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)+d\right]=\tilde{O}\left[\frac{d^2N^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

We note that the condition in Lemma 35 provides a sufficient condition of the $\mathcal{SO}$ oracle. In practice, the value of n can be much smaller than the bound in Lemma 35; see numerical examples in Section 5.5. To show the usefulness of the stochastic separation oracle, we extend Vaidya's cutting-plane method [221] to a stochastic cutting-plane method that can find high-precision solutions with high probability in the stochastic case. Vaidya's cutting-plane method maintains a polytope that contains the optimal points and iteratively reduces the volume of polytope by generating a separation oracle at the approximate volumetric center. We provide the pseudo-code of deterministic Vaidya's method in 5.D for the self-contained purpose. Other deterministic cutting-plane methods based on reducing the volume of a polytope can also be extended to the stochastic case using our novel framework, and we consider Vaidya's method mainly for its simplicity.

It is desirable to prove that by substituting the separation oracles with stochastic separation oracles, Vaidya's cutting-plane method can be used to find high-precision solutions with high probability. The pseudo-code of the stochastic cutting-plane method is given in Algorithm 9.

Algorithm 9 Stochastic cutting-plane method for the PGS guarantee

Input: Model $\mathcal{X}, (\Upsilon, \mathcal{B}_\Upsilon), F(x, \xi_x)$, optimality guarantee parameters ϵ and δ, Lipschitz constant L, (ϵ, δ)-$\mathcal{SO}$ oracle $\hat{g}$.

Output: An (ϵ, δ)-PGS solution x^* to problem (5.1).

Set the initial polytope $P \leftarrow [1, N]^d$.
Set the constant $\rho \leftarrow 10^{-7}$. ▷ Constant ρ corresponds to ϵ in [221].
Set the number of iterations $T_{max} \leftarrow \lceil 2d/\rho \cdot \log[dNL/(\rho\epsilon)] \rceil$.
Initialize the set of points used to query separation oracles $\mathcal{S} \leftarrow \emptyset$.
Initialize the volumetric center $z \leftarrow (N+1)/2 \cdot (1, 1, \ldots, 1)^T$.
for $T = 1, 2, \ldots, T_{max}$ **do**
    Decide adding or removing a cutting plane by Vaidya's method.
    if add a cutting plane **then**
        Evaluate an $(\epsilon/8, \delta/4)$-$\mathcal{SO}$ oracle $\hat{g}_z$ at z.
        if $\hat{g}_z = 0$ **then**
            Round z to an integral solution by Algorithm 2 and return the rounded solution.
        end if
        Add the current point z to $\mathcal{S}$.
    else if remove a cutting plane **then**
        Remove corresponding point z from $\mathcal{S}$.

16: **end if**
17: Update the approximate volumetric center z by a Newton-type method.
18: **end for** ▷ There are at most $O(d)$ points in $\mathcal{S}$ by Vaidya's method.
19: Find an $(\epsilon/4, \delta/4)$-PGS solution $\hat{x}$ of problem $\min_{x\in\mathcal{S}} f(x)$.
20: Round $\hat{x}$ to an integral solution by Algorithm 2.

We note that if the approximate volumetric center z is not in $[1, N]^d$, then we choose a violated constraint $x_i \geq 1$ or $x_i \leq N$ and return e_i or $-e_i$ as the separating vector, respectively. For arithmetic operations, each iteration of Algorithm 9 requires $O(d)$ inversions and multiplications of $d \times d$ matrices. Each inversion and multiplication can be finished within $O(d^\omega)$ arithmetic operations, where $\omega < 2.373$ is the matrix exponent [9]. Hence, Algorithm 9 needs $O(d^{\omega+1})$ arithmetic operations for each iteration. The calculation of the number of iterations T_{max} is provided in Section 5.F. The correctness and the expected simulation cost of Algorithm 9 are studied in the following theorem.

Theorem 56. *Suppose that Assumptions 5-7, and 11 hold. Algorithm 9 returns an* (ϵ, δ)*-PGS solution and we have*

$$\begin{aligned} T(\epsilon, \delta, \mathcal{MC}) &= O\left[\frac{d^3N^2}{\epsilon^2}\log\left(\frac{dLN}{\epsilon}\right)\log\left(\frac{1}{\delta}\right) + d^2\log\left(\frac{dLN}{\epsilon}\right)\right] \\ &= \tilde{O}\left[\frac{d^3N^2}{\epsilon^2}\log\left(\frac{dLN}{\epsilon}\right)\log\left(\frac{1}{\delta}\right)\right]. \end{aligned}$$

Remark 4. We note that another popular deterministic cutting-plane method, the random walk-based cutting-plane method [20], can also be extended to the stochastic case and achieves a better expected simulation cost

$$\tilde{O}\left[\frac{d^3N^2}{\epsilon^2}\log\left(\frac{LN}{\epsilon}\right)\log\left(\frac{1}{\delta}\right)\right]$$

at the expense of $\tilde{O}[d^6 + \log^2(1/\delta)]$ arithmetic operations in each iteration. We provide the pseudo-code in 5.D for the self-contained purpose. Here, the $O[\log^2(1/\delta)]$ factor is required to ensure the high-probability approximation to the centroid. Moreover, we note that the fast implementation of Vaidya's method in [123] reduces number of arithmetic operations in each iteration to $O(d^2)$.

Remark 5. Stochastic cutting-plane methods can also be applied to problems that are defined on $[N]^d$ with linear constraints $\{x \in \mathbb{Z}^d : Ax \leq b\}$, since we can choose the initial polytope to be $\mathcal{X} := [1, N]^d \cap \{Ax \leq b\}$. The results in this section still hold if we replace N with $\max_{x,y\in\mathcal{X}} \|x - y\|_\infty$.

From Theorem 56, we can see that the upper bound on the expected simulation cost only has a logarithmic dependence on the Lipschitz constant L, which is better than the

quadratic dependence of the subgradient-based algorithms in Chapter 4. In the next subsection, we further improve the dependence and develop a dimension reduction algorithm, whose simulation cost is upper bounded by a constant that is independent of the Lipschitz constant L.

Stochastic Cutting-plane Methods: Dimension Reduction Algorithm

In this subsection, we develop the dimension reduction algorithm, which does not require the knowledge about the Lipschitz constant L and whose simulation cost is upper bounded by a constant that is independent of L. The idea behind the dimension reduction algorithm is based on the following observation: if a convex body $P \subset \mathbb{R}^d$ has a volume $\text{vol}(P)$ smaller than $(d!)^{-1} = O[\exp(-(d+1/2)\log(d)+d)]$, then all integral points inside P must lie on a hyperplane. Otherwise, if there exist $d+1$ integral points $x_0, \ldots, x_d \in P$ that are not on the same hyperplane, then the convex body P contains the polytope $\text{conv}\{x_0, \ldots, x_d\}$, which has the volume

$$\frac{1}{d!}\left|\det(x_1 - x_0, \ldots, x_d - x_0)\right| \geq \frac{1}{d!},$$

where $\text{conv}(\cdot)$ is the convex hull and $\det(\cdot)$ is the determinant of matrices. This leads to a contradiction since we assume that $\text{vol}(P) < (d!)^{-1}$. Hence, we may use Vaidya's method or the random walk method to reduce the volume of the search polytope P to $O[\exp(-(d+1/2)\log(d)+d)]$, and then we reduce the problem dimension by projecting the polytope onto the hyperplane that all remaining points lie on. After $d-1$ dimension reductions, we have a one-dimensional convex problem and algorithms in Section 5.3 can be applied. This idea is summarized in Algorithm 10.

Algorithm 10 Dimension reduction algorithm for the PGS guarantee

Input: Model $\mathcal{X}, (\mathsf{Y}, \mathcal{B}_{\mathsf{Y}}), F(x, \xi_x)$, optimality guarantee parameters ϵ and δ, (ϵ, δ)-$\mathcal{SO}$ oracle $\hat{g}$.

Output: An (ϵ, δ)-PGS solution x^* to problem (5.1).

1: Set the initial polytope $P \leftarrow [1, N]^d$.
2: Initialize the set of points used to query separation oracles $\mathcal{S} \leftarrow \emptyset$.
3: **for** $d' = d, d-1, \ldots, 2$ **do** ▷ The current dimension d' is gradually reduced.
4: Initialize Vaidya's cutting-plane method.
5: **while** the volume of P is larger than $(d'!)^{-1}$ **do**
6: Take one step of Vaidya's cutting-plane method with $(\epsilon/4, \delta/4)$-$\mathcal{SO}$ oracle. ▷ Vaidya's cutting-plane method decides a suitable cutting plane H.
7: Add the point where the stochastic separation oracle is called to $\mathcal{S}$.
8: Shrink the volume of P using the cutting plane H.
9: **end while**
10: Find the hyperplane H that contains all integral points in P.

▷ If P contains no integral points, then an arbitrary hyperplane works.
11: Project P onto the hyperplane H. ▷ Reduce the dimension by 1.
12: **end for**
13: Find an $(\epsilon/4, \delta/4)$-PGS solution of the last one-dim problem and add the solution to $\mathcal{S}$.
14: Find the $(\epsilon/4, \delta/4)$-PGS solution $\hat{x}$ of problem $\min_{x\in\mathcal{S}} \ f(x)$.
15: Round $\hat{x}$ to an integral solution by Algorithm 2.

We note that the application of Vaidya's method in Line 6 refers to implementing the cutting-plane algorithm for one iteration. Namely, only a single cutting hyperplane will be generated. Importantly, the implementation of Vaidya's method in this step does not require the knowledge about the Lipschitz constant, since the Lipschitz constant is only used to calculate the total number of steps in Algorithm 9. In addition, Vaidya's cutting-plane method can be replaced with other deterministic cutting-plane methods. Furthermore, many cutting-plane methods, including Vaidya's method and random-walk-based method, guarantee that the volume of the polytope P is decreased at a constant rate. For example, the random walk-based cutting-plane method reduces the volume at the rate $1 - e^{-1}$ and after k iterations, the volume of P is at most $(1 - e^{-1})^k N^d$. Thus, we can terminate the cutting-plane method when the upper bound is lower than $(d!)^{-1}$.

We note that the search of hyperplane H in Line 10 does not require evaluations of the function $F(x, \xi_x)$ and therefore, it will not affect the simulation cost of Algorithm 10. The simplest algorithm to find the hyperplane H is to enumerate all hyperplanes generated by the points in $[N]^d$ and check if the condition $P \cap \mathbb{Z}^d \subset H$ is satisfied. This naive algorithm terminates in finite time but may require an exponential number of arithmetic operations. In [122], the author reduced the problem of finding a hyperplane H to the problem of finding an approximate solution to the Shortest Vector Problem in lattices. When the volume of the current polytope P is small enough, it is proved that a set of Lenstra–Lenstra–Lovász (LLL)-reduced basis [141] contains the normal vector of the hyperplane H, namely, the vector $c \in \mathbb{Z}^d$ such that

$$\langle c, x - y\rangle = 0, \quad \forall x, y \in P \cap \mathbb{Z}^d.$$

The LLL algorithm [141], which only requires a polynomial number of arithmetic operations, can be applied to find the desired LLL-reduced basis. We show that their results can be extended to the stochastic case and combined with the framework in Section 5.4 to generate an algorithm that only requires a polynomial number of arithmetic operations.

Intuitively, the dimension reduction algorithm implements the stochastic cutting-plane method at each dimension from d to 1. Therefore, the total simulation cost is on the same order as the summation of i^3 for $i \in [d]$, which is on the order of $O(d^4)$. More rigorously, we provide the correctness and the simulation cost of Algorithm 10 in the following theorem.

Theorem 57. *Suppose that Assumptions 5-7 hold. Algorithm 10 returns an (ϵ, δ)-PGS solution and we have*

$$T(\epsilon, \delta, \mathcal{MC}) = O\left[\frac{d^3 N^2 (d + \log(N))}{\epsilon^2} \log\left(\frac{1}{\delta}\right) + d^2(d + \log(N))\right]$$

$$= \tilde{O}\left[\frac{d^3N^2(d+\log(N))}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

We note that the idea of gradually reducing the dimension is proposed in our work and [122] independently, although the author of [122] has made the algorithm more practical. More specifically, if we allow exponentially many arithmetic operations, the LLL algorithm is not necessary. In that case, we can reduce the number of separation oracles to $O(d^2)$ and the computational complexity can be reduced to $\tilde{O}[d^4N^2\epsilon^{-2}\log(1/\delta)]$. From Theorem 57, we can see that the expected simulation cost of Algorithm 10 is upper bounded by a constant that is independent of the Lipschitz constant L. In addition, Algorithm 10 does not require any prior estimation of the upper bound of the Lipschitz constant. Since the estimated Lipschitz constant is likely to be much larger than the real Lipschitz constant, the error in the estimation will lead to a significant increase in the the simulation cost. Therefore, the dimension reduction algorithm is suitable when an estimate of the upper bound of the Lipschitz constant is difficult to obtain or the upper bound of the Lipschitz constant is large.

5.5 Numerical Experiments

In this section, we implement our proposed simulation-optimization algorithms that are guaranteed to find high-confidence high-precision solutions. Through these numerical experiments, we show that the localization methods proposed in this manuscript outperform benchmark algorithms on large-scale problems. We also compare our proposed algorithms to benchmark algorithms that do not utilize the $L^\natural$-convexity, e.g. the Industrial-strength COMPASS algorithm [236] and the R-SPLINE algorithm [226]; see Section 5.H for additional numerical results and discussions.

First, we consider the problem of finding the optimal allocation of a total number of N staffs to two queues so that the average waiting time for all of the arrivals from the two queues is minimized. Given the optimality parameters ϵ and δ, we empirically show that the TS algorithm and the SUS algorithm have respectively $O(\log N)$ and $O(1)$ dependence on the scale N, which supports our theoretical results. In addition, we construct a synthetic one-dimensional convex function with a similar landscape to show that the returned solution satisfies the high-probability guarantee. Second, we construct a multi-dimensional stochastic function, whose expectation is a separable convex function, i.e., functions of the form $f(x) = \sum_{i=1}^{d} f^i(x_i)$ for convex functions $f^1(x), \ldots, f^d(x)$, to test and compare the subgradient descent algorithm in Chapter 4 with the stochastic localization methods proposed in this chapter for different values of the scale N and dimension d, especially for large N. Similar to the one-dimensional case, we consider functions with a closed-form to check the coverage rate of the proposed algorithms. Finally, the multi-dimensional resource allocation problem in service systems is considered to compare the performance of proposed algorithms on practical problems.

Staffing Two Queues under Resource Constraints

Consider a service system that operates over a time horizon $[0, T]$ with two streams of customers arriving at the system. One example is that the system receives service requests from both online app-based customers and offline walk-in customers, and each stream needs dedicated servers assigned. The first stream of customers arrives according to a doubly stochastic non-homogeneous Poisson process $N_1 := (N_1(t) : t \in [0, T])$, with the customer service times being independent and identically distributed according to a distribution S_1. The second stream of customers obeys the same model with the process $N_2 := (N_2(t) : t \in [0, T])$ and distribution S_2. The two streams of customers form two separate queues and their arrival processes can be correlated. Suppose that the decision maker needs to staff the two queues separately. There are in total a number of $N + 1$ homogeneous servers that work independently in parallel. Each server can handle the service requested by customers from either stream, one at a time. Suppose that no change on the staffing plan can be made once the system starts working. Assume that the system operates based on a first-come-first-serve routine, with unlimited waiting room in each queue, and that customers never abandon.

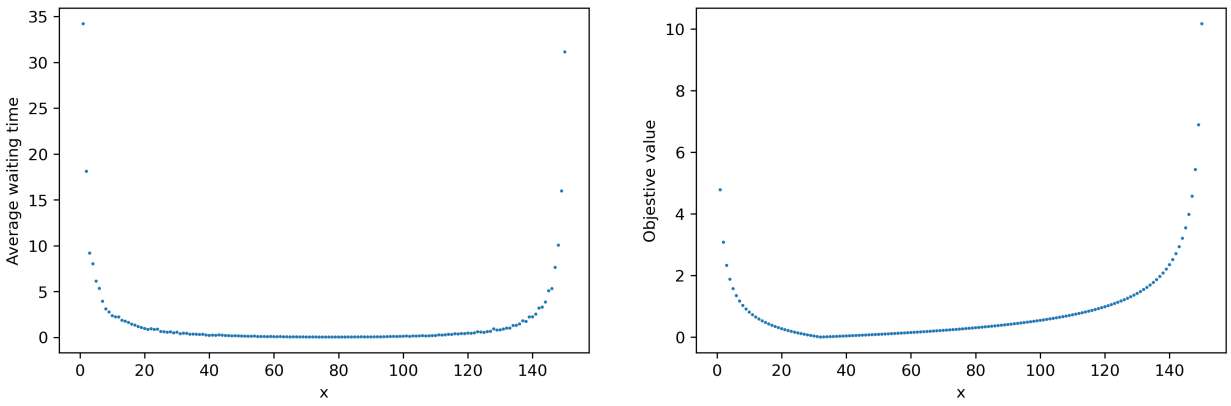
The decision maker's objective is to select the staffing level $x \in [N]$ for the first queue and the staffing level $N + 1 - x$ for the second queue, in order to minimize the expected average waiting time for all customers from the two streams over the time horizon $[0, T]$. In the numerical example, we consider $N \in \{10, 20, \ldots, 150\}$ and $T = 2$. The arrival processes N_1 and N_2 are non-homogeneous processes with random intensity functions $\Gamma_1 \cdot \lambda_1(t)$ and $\Gamma_2 \cdot \lambda_2(t)$, in which

$$\lambda_1(t) := 75 + 25\sin(0.3t), \quad \lambda_2(t) := 80 + 40\sin(0.2t).$$

Positive-valued random variables Γ_1 and Γ_2 are defined as

$$\Gamma_1 := X + Z, \quad \Gamma_2 := Y - Z,$$

where X, Y are independent uniform random variables on $[0.75, 1.25]$ and Z is an independent uniform random variable on $[-0.5, 0.5]$. The service time distribution S_1 is log-normal distributed with mean 0.75 and variance 0.1. The service time distribution S_2 is gamma distributed with mean 0.65 and variance 0.1. Figure 5.5.1 plots an empirical average waiting time as a function of the discrete decision variable x. It can be observed that the landscape around the optimum is extremely flat and such property may cause challenges for algorithms that aim to exactly select the optimal solution (i.e., the PCS guarantee). In practice, the decision maker may be indifferent about a very small difference in the averaging waiting time performance, when the small difference does not impact much on customers' satisfaction. Instead, algorithms that are designed for the (ϵ, δ)-PGS guarantee do not suffer from the extremely flat landscape around the global optimum.



(a) (b)

Figure 5.5.1: The landscapes of objective functions in the one-dimensional case. **(a)** The empirical average waiting time with $N = 150$. **(b)** The landscape of the synthetic convex function with scale $N = 150$ and optimum $x^* = 31$.

Separable Convex Function Minimization

We consider the problem of minimizing a stochastic function whose expectation is a separable $L^\natural$-convex function of the form

$$f_{c,x^*}(x) := \sum_{i=1}^{d} c_i g(x_i; x_i^*),$$

where $c_i \in [0.75, 1.25]$, $x_i^* \in \{1, \ldots, \lfloor 0.3N \rfloor\}$ for all $i \in [d]$ and

$$g(x; x^*) := \begin{cases} \sqrt{\frac{x^*}{x}} - 1 & \text{if } x \leq x^* \\ \sqrt{\frac{N+1-x^*}{N+1-x}} - 1 & \text{if } x > x^* \end{cases}, \quad \forall x, x^* \in [N],$$

It can be observed that the function $f_{c,x^*}(x)$ is the sum of separable convex functions and therefore is $L^\natural$-convex. Moreover, the function $f_{c,x^*}(x)$ has the optimum x^* associated with the optimal value 0. The objective function has a similar landscape as the average waiting time; see Figure 5.5.1. For stochastic evaluations, we add Gaussian noise with mean 0 and variance 1. The advantage of this numerical example is that the expected objective function has a closed form, and we are able to exactly compute the optimality gap of the solutions returned by the proposed algorithms.

Resource Allocation Problem in Service Systems

We consider the 24-hour operation of a service system with a single stream of incoming customers. The customers arrive according to a doubly stochastic non-homogeneous Poisson process with the intensity function

$$\Lambda(t) := 0.5\lambda N \cdot (1 - |t - 12|/12), \quad \forall t \in [0, 24],$$

where λ is a positive constant and N is a positive integer. Each customer requests a service with the service time independent and identically distributed according to the log-normal distribution with mean $1/\lambda$ and variance 0.1. We divide the 24-hour operation into d time slots with length $24/d$ for some positive integer d. For the i-th time slot, there are $x_i \in [N]$ of homogeneous servers that work independently in parallel and the number of servers cannot be changed during the slot. Assume that the system operates based on a first-come first-serve routine, with an unlimited waiting room in each queue, and that customers never abandon.

The decision maker's objective is to select the staffing level $x := (x_1, \ldots, x_d)$ such that the total waiting time of all customers is minimized. Namely, by letting $f(x)$ be the expected total waiting time under the staffing plan x, the optimization problem can be written as

$$\min_{x \in [N]^d} f(x). \tag{5.4}$$

It has been proved in [10] that the function $f(\cdot)$ is multimodular. We define the linear transformation

$$g(y) := (y_1, y_2 - y_1, \ldots, y_d - y_{d-1}) \quad \forall y \in \mathbb{R}^d.$$

Then, [168] has proved that

$$h(y) := f \circ g(y) = f(y_1, y_2 - y_1, \ldots, y_d - y_{d-1})$$

is a $L^\natural$-convex function on the $L^\natural$-convex set

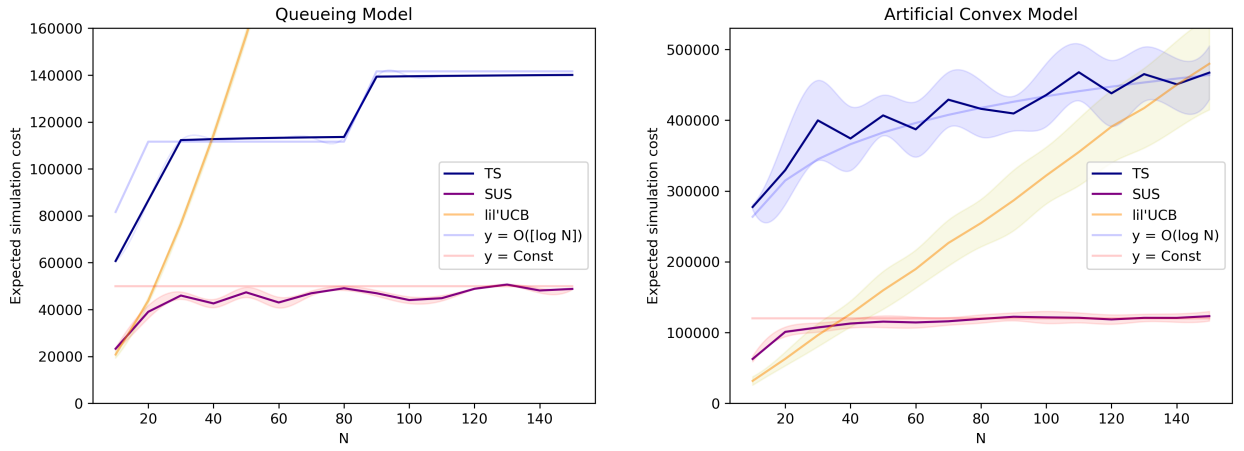
$$\mathcal{Y} := \{y \in [Nd]^d \mid y_1 \in [N],\ y_{i+1} - y_i \in [N],\ i = 1, \ldots, d-1\}.$$

The optimization problem (5.4) has the trivial solution $x_1 = \cdots = x_d = N$. However, in reality, it is also necessary to keep the staffing cost low. Therefore, we add the staffing cost term $R(x_1, \ldots, x_d) := C/d \cdot \sum_{i=1}^d x_i = C/d \cdot y_d$ to the objective function, where C is a positive constant. The optimization problem can be written as

$$\min_{y \in \mathcal{Y}} h(y) + C/d \cdot y_d. \tag{5.5}$$

The proposed algorithms can be extended to this problem by considering the Lovász extension $\tilde{h}(y)$ on the set

$$\tilde{\mathcal{Y}} := \{y \in [1, Nd]^d \mid y_1 \in [1, N],\ y_{i+1} - y_i \in [1, N],\ i = 1, \ldots, d-1\}.$$



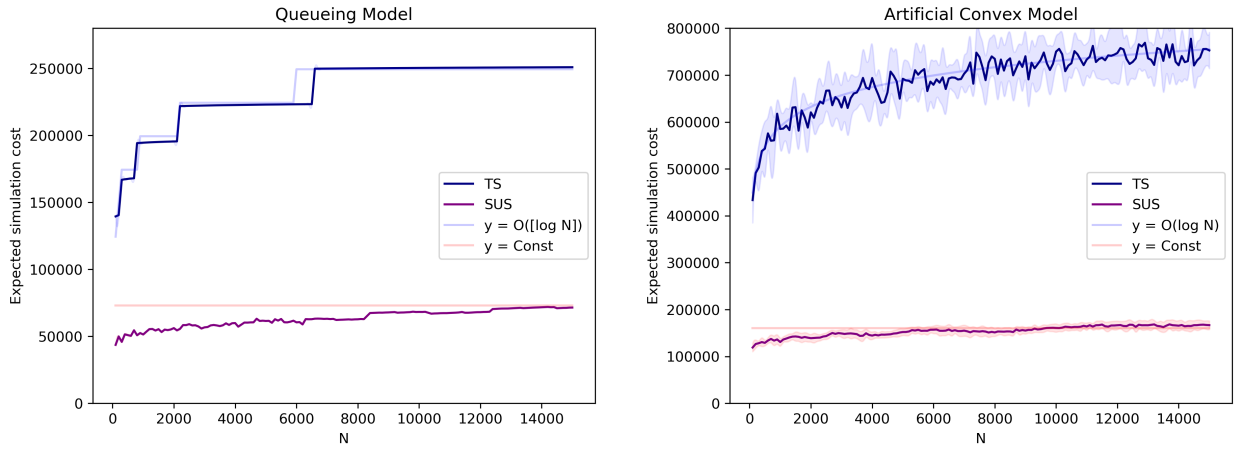
(a) (b)

Figure 5.5.2: The expected simulation cost of TS, SUS and lil'UCB algorithms in the one-dimensional case. **(a)** Optimal allocation problem. **(b)** One-dimensional separable convex function minimization.

Numerical Results: Tri-section Sampling Algorithm and Shrinking Uniform Sampling Algorithm

We first compare the performance of the TS algorithm and the SUS algorithm on the optimal allocation problem in Section 5.5 and the closed-form convex function minimization problem in Section 5.5. As a comparison to the existing algorithms, we also implement the state-of-the-art algorithm for the best arm identification problem, namely the lil'UCB algorithm [119]. The best arm identification problem is equivalent to problem (5.1) without any convexity structure. We consider problems with dimension $d = 1$ and scale $N \in \{10, 20, \ldots, 150\}$. The expected simulation cost is computed by averaging 400 independent solving processes. For the optimal allocation problem, we set the optimality parameters for the PGS guarantee as $\epsilon = 1$ and $\delta = 10^{-6}$. An upper bound on the variance is estimated as $\sigma^2 = 10$. For the convex function minimization problem, we generate each c_i from the uniform distribution on $[0.75, 1.25]$ and x_i^* from the discrete uniform distribution on $\{1, 2, \ldots, \lfloor 0.3N \rfloor\}$. The optimality parameters are chosen as $\epsilon = 0.2$ and $\delta = 10^{-6}$ and the variance is set to be $\sigma^2 = 1$.

It is observed that both algorithms satisfy the given PGS guarantee on the synthetic convex function minimization problem, namely, the ϵ-optimality is satisfied for all implementations. We then plot the estimated expected simulation costs in Figure 5.5.2. For the optimal allocation problem, the expected simulation costs of the TS and SUS algorithms approximately have $O(\lceil \log N \rceil)$ and almost $O(1)$ dependence on the scale N, respectively. The expected simulation cost of the SUS algorithm is almost independent of N and this



(a) (b)

Figure 5.5.3: The expected simulation cost of TS and SUS algorithms for problems with a larger scale. **(a)** Optimal allocation problem. **(b)** One-dimensional separable convex function minimization.

verifies our theoretical analysis. For the synthetic convex function minimization problem, same as the queueing example, the estimated expected simulation costs of the TS algorithm and SUS algorithm have $O(\log N)$ and almost $O(1)$ dependence on N, respectively. This again verifies our theoretical analysis. Moreover, both algorithms outperform the lil'UCB algorithm, when N is large. The numerical results show that our proposed algorithms can efficiently solve large-scale one-dimensional convex problems.

Furthermore, we also compare the TS and the SUS algorithms on problems with a larger scale. Specifically, we consider both problems with $N \in \{100, 200, \ldots, 15000\}$ under the same setting. The results are plotted in Figure 5.5.3 and we can see that the outcomes of the algorithms comply with our theoretical results.

Numerical Results: Subgradient Descent and Localization Methods

We next compare the performances of the truncated stochastic subgradient descent algorithm (Algorithm 3) and stochastic localization methods proposed in this chapter. We first consider the separable convex function minimization problem, where we can compute the optimality gap and verify the ϵ-optimality. The dimension and scale of the separable convex model are chosen as $d \in \{2, 6, 10, 15\}$ and $N \in \{50, 500, 5000\}$. The optimality guarantee parameters are chosen as $\epsilon = d$ and $\delta = 10^{-6}$, respectively. The empirical choice of ϵ ensures that any ϵ-optimal solution x^0 satisfies $\|x^0 - x^*\|_1 \leq d^{5/6} \ll N$. We compute the average

simulation cost of 100 independently generated models to estimate the expected simulation cost. Moreover, early stopping conditions are designed to terminate algorithms early when little progress is made at any iteration. For the subgradient descent algorithm, we maintain the empirical mean of stochastic objective function values up to the current iteration and terminate the algorithm if the empirical mean does not decrease by $O(\epsilon/\sqrt{N})$ after $O[d\epsilon^{-2}\log(1/\delta)]$ consecutive iterations. For stochastic cutting-plane methods, we terminate the algorithm if the empirical mean of the objective function of the last 5 iterations does not decrease by ϵ/d. For the dimension reduction method, we terminate the algorithm early if the polytope is empty. Furthermore, we have observed that using $(N\epsilon/4, \delta/4)$-$\mathcal{SO}$ oracles in localization methods is sufficient for producing high-probability guarantees on this example.

We summarize the results in Table 5.5.1. We define the coverage rate to be the percentage of implementations that produce an ϵ-optimal solution. The coverage rates of the algorithms are all equal to 100% and thus, the PGS guarantee is likely to be satisfied by all of the algorithms. We note that, similar to the one-dimensional case, the standard deviation of the simulation cost is smaller than 10% of the estimated simulation cost in all settings. The performances of localization methods are better than the subgradient descent algorithm in all settings especially for the large-scale instances. The simulation cost of the random walk-based cutting-plane method is better than the Vaidya's cutting-plane method, which may be a result of the extra $\log(d)$ term in the simulation cost; see the discussion in Remark 4. The dimension reduction method has the best performance on examples with $N = 500$, 5000 and has the advantage of not requiring any knowledge about the Lipschitz constant. From the experimental results, we can see that the empirical performances of proposed algorithms are sometimes better than their theoretical guarantees. One possible explanation for the better empirical performance is that during the implementation of the stochastic localization methods, the diameter of the feasible set (i.e., the set $\mathcal{S}$ in Algorithms 9 and 10) will decrease and become much smaller than N after a few iterations. In contrast, for the theoretical analysis, we need to consider the worst case and assume that the diameter is still N for the shrunken set. Therefore, the number of simulations to generate an accurate stochastic separation oracle is overestimated in Lemma 35 and the simulation costs of the stochastic localization algorithms have a better dependence on the problem scale N in practice.

We then consider the multi-dimensional resource allocation problem. We first fix the dimension (number of time slots) to be $d = 4$ and compare the performance with the scale $N \in \{10, 20, 30, 40, 50\}$, and we then fix the scale to be $N = 10$ and compare the performance with the dimension $d \in \{4, 8, 12, 16, 20, 24\}$. The parameters of the problem are chosen as $\lambda = 1$ and $C = 10$, and the optimality guarantee parameters are $\epsilon = N/2$ and $\delta = 10^{-6}$. We also compare the algorithms with a smaller precision parameter $\epsilon = N/10 + 1$ in Section 5.H. An upper bound on the variance is estimated as $\sigma^2 = 30\sqrt{N}$. For each problem setup, we average the results of 10 independent implementations to estimate the expected simulation cost and the objective value of the returned solution. The results are summarized in Table 5.5.2. Similarly, the standard deviation of the simulation cost is smaller than 10% of the estimated simulation cost in all settings. It is observed that the dimension reduction method achieves the best performance in all cases, although its simulation costs

Params.		Search Methods	Localization Methods (Chapter 5)		
		SubGD (Chapter 4)	Vaidya's	Random Walk	Dim Reduction
d	N	Cost	Cost	Cost	Cost
2	50	1.08e3	2.74e2	1.66e2	**1.56e2**
2	500	2.54e4	6.54e2	2.32e2	**2.08e2**
2	5000	3.97e5	1.13e4	5.29e2	**4.66e2**
6	50	5.00e3	4.13e2	**3.36e2**	4.05e2
6	500	4.75e4	1.34e3	1.65e3	**6.45e2**
6	5000	2.72e6	8.15e4	4.75e3	**8.25e2**
10	50	8.46e3	7.98e2	**7.70e2**	8.34e2
10	500	6.32e4	6.57e3	2.16e3	**1.48e3**
10	5000	7.76e6	2.42e5	8.03e3	**2.02e3**
15	50	1.23e4	1.50e3	**1.91e3**	2.18e3
15	500	2.83e5	2.66e4	1.06e4	**3.19e3**
15	5000	1.85e7	1.96e6	1.55e5	**4.85e3**

Table 5.5.1: Simulation cost of different algorithms on separable convex functions.

have a faster growth rate than other methods. The stochastic cutting-plane methods also outperform the subgradient descent algorithm when the dimension is 4. The truncated stochastic subgradient descent algorithm returns the smallest objective values except the case when $(d, N) = (4, 50)$, and the objective values returned by other algorithms are not much larger than the truncated stochastic subgradient descent algorithm. This is possible since we are searching for ϵ-optimal solutions and an optimality gap smaller than $\epsilon = N/2$ is acceptable.

In summary, based on the results from numerical results, the SUS algorithm and the dimension reduction method provide a more efficient choice for large-scale convex discrete optimization via simulation problems, and they have the advantage that no prior information about the objective function is required except the $L^\natural$-convexity.

Params.		Search Methods		Localization Methods (Chapter 5)					
		SubGD (Chapter 4)		Vaidya's		Random Walk		Dim Reduction	
d	N	Cost	Obj.	Cost	Obj.	Cost	Obj.	Cost	Obj.
4	10	3.06e5	2.13e1	9.89e4	2.19e1	6.92e4	2.47e1	**2.42e4**	2.40e1
4	20	1.08e5	3.41e1	3.64e4	3.42e1	2.45e4	3.73e1	**1.40e4**	3.44e1
4	30	7.79e4	4.59e1	1.94e4	4.65e1	1.33e4	5.10e1	**9.21e3**	4.59e1
4	40	5.06e4	5.73e1	1.24e4	5.86e1	8.68e3	6.35e1	**6.31e3**	5.75e1
4	50	4.50e4	6.91e1	9.24e3	6.98e1	6.22e3	7.49e1	**4.03e3**	6.67e1
8	10	1.20e6	2.01e1	7.27e5	2.12e1	5.53e5	2.17e1	**1.48e5**	2.12e1
12	10	2.69e6	1.90e1	2.49e6	2.07e1	1.86e6	2.13e1	**6.10e5**	2.01e1
16	10	4.78e6	1.83e1	6.64e6	2.02e1	4.43e6	2.04e1	**1.59e6**	1.91e1
20	10	7.45e6	1.78e1	1.38e7	2.01e1	8.65e6	2.04e1	**3.21e6**	1.81e1
24	10	1.43e7	1.71e1	2.42e7	1.99e1	1.49e7	2.04e1	**8.54e6**	1.76e1

Table 5.5.2: Simulation cost and objective value of different algorithms on the resource allocation problem.

Algorithms	**Expected Simulation Cost**
TS (Section 5.3)	$\tilde{O}(\log(N)\epsilon^{-2}\log(1/\delta))$
SUS (Section 5.3)	$\tilde{O}(\epsilon^{-2}\log(1/\delta))$ (best achievable performance)
Subgradient-based (Chapter 4)	$\tilde{O}(d^2N^2L^2\epsilon^{-2}\log(1/\delta))$
Stochastic Cutting-plane (Section 5.4)	$\tilde{O}(d^3N^2\epsilon^{-2}\log(dNL/\epsilon)\log(1/\delta))$
Dimension Reduction (Section 5.4)	$\tilde{O}(d^3N^2(d+\log(N))\epsilon^{-2}\log(1/\delta))$
Shrinking Uniform Sampling (Section 5.C)	$\tilde{O}(M^d\epsilon^{-2}\log(1/\delta))$

Table 5.5.3: Upper bounds on the expected simulation cost for algorithms that achieve the PGS guarantee. Here, M is an absolute constant. All constants except $d, N, \epsilon, \delta, c, M$ are omitted in the $\tilde{O}(\cdot)$ notation. In comparison, the expected simulation cost without convexity is $O(N^d\epsilon^{-2}\log(1/\delta))$.

Appendix

5.A Algorithms and Complexity Analysis for the PCS-IZ Guarantee

In this section, we provide modified simulation-optimization algorithms for the PCS-IZ guarantee. We assume that the objective value of any sub-optimal choice of decision variables is at least c larger than the optimal objective value, where the indifference zone parameter $c > 0$ is known a priori. We reiterate the definition of the PCS-IZ guarantee for the completeness.

- *Probability of correct selection with indifference zone* (PCS-IZ). (See [101]) The problem is assumed to have a unique solution that renders the optimal objective value. The optimal objective value is assumed to be at least $c > 0$ smaller than the objective values at sub-optimal choices of decisions. The gap width c is called the **indifference zone parameter** in [17]. The PCS-IZ guarantee requires that the solution returned by an algorithm be the optimal solution with probability at least $1 - \delta$.

Let $\mathcal{MC}_c$ be the set that includes all convex models with the indifference zone parameter c. Then, the expected simulation cost for the PCS-IZ criterion is denoted as

$$T(\delta, \mathcal{MC}_c) := T((c, \delta)\text{-}PCS\text{-}IZ, \mathcal{MC}_c).$$

Modified Tri-section Sampling Algorithm for the PCS-IZ Guarantee

We first consider the one-dimensional case. When the prior information about the indifference zone parameter c is available, we can modify the TS algorithm to achieve a better simulation cost. The modified algorithm also consists of two parts: the shrinkage of intervals and a sub-problem with at most 3 points. The improvement is achieved by a weaker condition for the comparison of objective values at two 3-quantiles. We give the modified algorithm in Algorithm 11 and omit those lines that are the same as Algorithm 7.

Algorithm 11 Tri-section sampling algorithm for the PCS-IZ guarantee

Input: Model $\mathcal{X} = [N], (\mathsf{Y}, \mathcal{B}_{\mathsf{Y}}), F(x, \xi_x)$, optimality guarantee parameter δ, indifference zone parameter c.

Output: An (c, δ)-PCS-IZ solution x^* to problem (5.1).

1: Set upper and lower bounds of current interval $x_L \leftarrow 1, x_U \leftarrow N$.
2: Set maximal number of comparisons $T_{max} \leftarrow \log_{1.5}(N) + 2$.
3: **while** $x_U - x_L > 2$ **do** ▷ Iterate until there are at most 3 decisions.
...
5: Simulate n independent copies of $F(q_{1/3}, \xi_{1/3})$ and $F(q_{2/3}, \xi_{2/3})$, where n is the smallest integer such that $h[n, \sigma, 1 - \delta/(2T_{max})] \leq (q_{2/3} - q_{1/3}) \cdot c/5$.
...
14: **end while**
15: Simulate $\tilde{n}$ independent copies of $F(x, \xi_x)$ for all $x \in \{x_L, \ldots, x_U\}$, where $\tilde{n}$ is the smallest integer such that $h[\tilde{n}, \sigma, 1 - \delta/(2T_{max})] \leq c/3$. ▷ Now $x_U - x_L \leq 2$.
16: Return the point in $\{x_L, \ldots, x_U\}$ with minimal empirical mean.

The following theorem proves the correctness and the expected simulation cost of the modified TS algorithm.

Theorem 58. *Suppose that Assumptions 5-7 hold. The modified TS algorithm is a* $[(c, \delta)$*-PCS-IZ,* $\mathcal{MC}_c]$*-algorithm. Furthermore, we have*

$$T(\delta, \mathcal{MC}_c) = O\left[\frac{1}{c^2}\log\left(\frac{\log(N)}{\delta}\right) + \log(N)\right] = \tilde{O}\left[\frac{1}{c^2}\log\left(\frac{1}{\delta}\right)\right].$$

By Theorem 58, the expected simulation cost for the PCS-IZ guarantee is asymptotically independent of the number of points N, when the failing probability δ is sufficiently small. If the SUS algorithm is used for the PCS-IZ guarantee, the algorithm also achieves an $\tilde{O}(c^{-2}\log(1/\delta))$ expected simulation cost by setting the optimality parameter $\epsilon = c/2$. This is because the objective values of sub-optimal solutions are larger than that of the optimal solution by at least c and because the solution satisfying the $(c/2, \delta)$-PGS guarantee also satisfies the (c, δ)-PCS-IZ guarantee. Hence, for both the modified TS algorithm and the SUS algorithm, the asymptotic simulation cost has an upper bound that is independent of N. However, we note that the space complexity of the modified TS algorithm is only $\tilde{O}(\log(N))$, whereas the SUS algorithm requires $O(N)$ memory space. Therefore, the modified TS algorithm is preferred for the PCS-IZ guarantee.

Modified Stochastic Cutting-plane Methods for the PCS-IZ Guarantee

In the multi-dimensional case, we develop modified stochastic cutting-plane methods for the PCS-IZ guarantee. Using the same adaptive acceleration scheme as in Chapter 4, the indifference zone parameter can help reduce the dependence of the simulation cost on

the problem scale N. We give the pseudo-code of the accelerated stochastic cutting-plane method in Algorithm 12.

Algorithm 12 Stochastic cutting-plane method for the PCS-IZ guarantee

Input: Model $\mathcal{X}, (\mathsf{Y}, \mathcal{B}_{\mathsf{Y}}), F(x, \xi_x)$, optimality guarantee parameter δ, indifference zone parameter c, Lipschitz constant L, (ϵ, δ)-$\mathcal{SO}$ oracle $\hat{g}$.

Output: An (c, δ)-PCS-IZ solution x^* to problem (5.1).

1: Set the initial guarantee $\epsilon_0 \leftarrow cN/4$.
2: Set the number of epochs $E \leftarrow \lceil \log_2(N) \rceil + 1$.
3: Set the initial searching space $\mathcal{Y}_0 \leftarrow [1, N]^d$.
4: **for** $e = 0, \ldots, E-1$ **do**
5: Use Algorithm 9 to get an $(\epsilon_e, \delta/(2E))$-PGS solution x_e in $\mathcal{Y}_e$.
6: Update guarantee $\epsilon_{e+1} \leftarrow \epsilon_e/2$.
7: Update the searching space $\mathcal{Y}_{e+1} \leftarrow \mathcal{N}(x_e, 2^{-e-2}N)$.
8: **end for**
9: Round x_{E-1} to an integral point by Algorithm 2.

We can prove the correctness and estimate the expected simulation cost of the accelerated algorithm in the same way as Theorem 51 Thus, we omit the proof.

Theorem 59. *Suppose that Assumptions 5-7, and 11 hold. The accelerated stochastic cutting-plane method returns a (c, δ)-PCS-IZ solution and we have*

$$\begin{aligned} T(c, \delta, \mathcal{MC}_c) &= O\left[\frac{d^3 \log(N)}{\epsilon^2} \log(\frac{dLN}{\epsilon}) \log\left(\frac{1}{\delta}\right) + d^2 \log(N) \log(\frac{dLN}{\epsilon})\right] \\ &= \tilde{O}\left[\frac{d^3 \log(N)}{\epsilon^2} \log(\frac{dLN}{\epsilon}) \log\left(\frac{1}{\delta}\right)\right]. \end{aligned}$$

By substituting Algorithm 9 with Algorithm 10 in the above algorithm, the acceleration scheme can be applied to Algorithm 10 to reduce the number of required simulation runs when the indifference zone parameter c is known. We give the reduced expected simulation cost for achieving the PCS-IZ guarantee and omit the proof.

Theorem 60. *Suppose that Assumptions 5-7 hold. The accelerated dimension reduction method returns an (c, δ)-PCS-IZ solution and we have*

$$\begin{aligned} T(c, \delta, \mathcal{MC}_c) &= O\left[\frac{d^3 \log(N)(d + \log(N))}{\epsilon^2} \log\left(\frac{1}{\delta}\right) + d^2 \log(N)(d + \log(N))\right] \\ &= \tilde{O}\left[\frac{d^3 \log(N)(d + \log(N))}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right]. \end{aligned}$$

Lower Bound on Expected Simulation Cost

In this subsection, we derive the lower bounds on the expected simulation costs for all of the simulation-optimization algorithms that satisfy certain optimality guarantee for general convex problems. In the proof for the lower bound result, we construct two convex models that have similar distributions at each point but have distinct optimal solutions. Then, the information-theoretical inequality in [128] can be used to provide a lower bound on the simulation costs for all algorithms.

We first present the results in [128] for completeness. Given a simulation-optimization algorithm and a model $\mathbb{M}$, we define random variable $N_x(\tau)$ to be the number of times that $F(x, \xi_x)$ is sampled when the algorithm terminates, where τ is the stopping time of the algorithm. Then, it follows from the definition that

$$\mathbb{E}_M[\tau] = \sum_{x \in \mathcal{X}} \mathbb{E}_{\mathbb{M}} \left[N_x(\tau) \right],$$

where $\mathbb{E}_{\mathbb{M}}$ is the expectation when the model $\mathbb{M}$ is given. Similarly, we can define $\mathbb{P}_{\mathbb{M}}$ as the probability when the model $\mathbb{M}$ is given. We denote the filtration up to the stopping time τ as $\mathcal{F}_\tau$. The following lemma is proved in [128] and is the major tool for deriving lower bounds in this chapter.

Lemma 36 ([128]). *For any two models $\mathbb{M}_1, \mathbb{M}_2$ and any event $\mathcal{E} \in \mathcal{F}_\tau$, we have*

$$\sum_{x \in \mathcal{X}} \mathbb{E}_{\mathbb{M}_1} \left[N_x(\tau) \right] \mathrm{KL}(\nu_{1,x}, \nu_{2,x}) \geq d(\mathbb{P}_{\mathbb{M}_1}(\mathcal{E}), \mathbb{P}_{\mathbb{M}_2}(\mathcal{E})), \tag{5.6}$$

where $d(x, y) := x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$, $\mathrm{KL}(\cdot, \cdot)$ is the Kullback–Leibler divergence (KL divergence), and $\nu_{k,x}$ is the distribution of model $\mathbb{M}_k$ at point x for $k = 1, 2$.

We first give a lower bound for the PCS-IZ guarantee.

Theorem 61. *Suppose that Assumptions 5-7 hold. We have*

$$T(\delta, \mathcal{MC}_c) \geq \Theta \left[\frac{1}{c^2} \log \left(\frac{1}{\delta} \right) \right].$$

The lower bound on $T(\epsilon, \delta, \mathcal{MC})$, i.e., the expected simulation cost for achieving the PGS guarantee, can be derived in a similar way by substituting c with 2ϵ in the construction of two models.

Corollary 5. *Suppose that Assumptions 5-7 hold. We have*

$$T(\epsilon, \delta, \mathcal{MC}) \geq \Theta \left[\frac{1}{\epsilon^2} \log \left(\frac{1}{\delta} \right) \right].$$

Combining with the upper bounds derived in Sections 5.3, 5.3 and 5.A, we conclude that the TS algorithm and the SUS algorithm are optimal up to a constant for the PCS-IZ guarantee in the asymptotic regime $\delta \ll 1$. Considering the space complexity, the TS algorithm is preferred for the PCS-IZ guarantee, while for the PGS guarantee we need to consider the trade-off between the simulation cost and the space complexity when choosing the best algorithm.

Proof of Theorem 58

We first estimate the simulation cost of each iteration and the sub-problem.

Lemma 37. *Suppose that Assumptions 5-7 hold. The simulation cost for each iteration of Algorithm 11 is at most* $100\sigma^2 c^{-2}(q_{2/3} - q_{1/3})^{-2} \log[4T_{max}/\delta]$*, where* $T_{max} := \log_{1.5}(N) + 2$*. The simulation cost of the sub-problem is at most* $54\sigma^2 c^{-2} \log[4T_{max}/\delta]$*.*

Proof. The proof is similar to the proof of Lemma 40 and we only give a sketch of the proof. By the definition of $h(n, \sigma, \alpha)$, simulating

$$n = \frac{50\sigma^2}{(q_{2/3} - q_{1/3})^2 c^2} \log\left(\frac{4T_{max}}{\delta}\right)$$

times on quantiles $q_{1/3}$ and $q_{2/3}$ is enough to ensure that the confidence half-width is at most $(q_{2/3} - q_{1/3}) \cdot c/5$. It implies that the last condition in Line 8 is satisfied and the simulation cost of each iteration is at most $100\sigma^2 c^{-2}(q_{2/3} - q_{1/3})^{-2} \log[4T_{max}/\delta]$. For the last sub-problem, simulating

$$\tilde{n} = \frac{18\sigma^2}{c^2} \log\left(\frac{4T_{max}}{\delta}\right)$$

times for each point is enough to ensure that the confidence half-width is at most $c/3$. Since there are at most 3 points in the sub-problem, the simulation cost for the sub-problem is at most $54\sigma^2 c^{-2} \log[4T_{max}/\delta]$. □

Using Lemma 37, we can estimate the total simulation cost of Algorithm 11.

Lemma 38. *Suppose that Assumptions 5-7 hold. The expected simulation cost of Algorithm 11 is bounded by*

$$\frac{459\sigma^2}{c^2} \log\left(\frac{4T_{max}}{\delta}\right) = O\left[\frac{1}{c^2} \log\left(\frac{1}{\delta}\right)\right],$$

where $T_{max} := \log_{1.5}(N) + 2$*.*

Proof. We denote the upper bound and the lower bound at the beginning of the k-th iteration as x_{U_k} and x_{L_k}, respectively. By Lemma 37, the simulation cost for the k-th iteration is at most $100\sigma^2 c^{-2}(q^k_{2/3} - q^k_{1/3})^{-2} \log[4T_{max}/\delta]$, where $q^k_{1/3}$ and $q^k_{2/3}$ are the 3-quantiles for the k-th

iteration. By the definition of 3-quantiles, it follows that $q^k_{2/3} - q^k_{1/3} \geq (x_{U_k} - x_{L_k})/3$ and therefore

$$\frac{100\sigma^2}{(q^k_{2/3} - q^k_{1/3})^2 c^2} \log\left(\frac{4T_{max}}{\delta}\right) \leq \frac{900\sigma^2}{(x_{U_k} - x_{L_k})^2 c^2} \log\left(\frac{4T_{max}}{\delta}\right). \tag{5.7}$$

Hence, we only need to bound the sum $\sum_{k=1}^{T}(x_{U_k} - x_{L_k})^{-2}$, where T is the number of iterations of Algorithm 11. By inequality (5.12), we know

$$x_{U_k} - x_{L_k} \geq \frac{3}{2}(x_{U_{k+1}} - x_{L_{k+1}}) - 1, \quad \forall k \in \{1, 2, \ldots, T-1\}.$$

We can rewrite the above inequality as $x_{U_k} - x_{L_k} - 2 \geq 3/2 \cdot (x_{U_{k+1}} - x_{L_{k+1}} - 2)$. Since T is the last iteration, it holds that $x_{U_T} - x_{L_T} \geq 4$ and therefore

$$x_{U_k} - x_{L_k} - 2 \geq \left(\frac{3}{2}\right)^{T-k} (x_{U_T} - x_{L_T} - 2) \geq 2 \cdot \left(\frac{3}{2}\right)^{T-k}.$$

Summing over $k = 1, 2, \ldots, T$, we get the bound

$$\begin{aligned}\sum_{k=1}^{T}(x_{U_k} - x_{L_k})^{-2} &\leq \sum_{k=1}^{T}\left(2 \cdot \left(\frac{3}{2}\right)^{T-k} + 2\right)^{-2} \leq \sum_{k=1}^{T}\frac{1}{4} \cdot \left(\frac{3}{2}\right)^{-2(T-k)} \\ &= \frac{9}{20}\left[1 - \left(\frac{4}{9}\right)^{T}\right] \leq \frac{9}{20}.\end{aligned}$$

Combining with inequality (5.7), the simulation cost for T iterations is at most

$$\frac{900\sigma^2}{c^2} \log\left(\frac{4T_{max}}{\delta}\right) \cdot \sum_{k=1}^{T}(x_{U_k} - x_{L_k})^{-2} \leq \frac{405\sigma^2}{c^2} \log\left(\frac{4T_{max}}{\delta}\right).$$

Considering the simulation cost of the sub-problem, the total simulation cost of Algorithm 11 is at most

$$\frac{405\sigma^2}{c^2} \log\left(\frac{4T_{max}}{\delta}\right) + \frac{54\sigma^2}{c^2} \log\left(\frac{4T_{max}}{\delta}\right) = \frac{459\sigma^2}{c^2} \log\left(\frac{4T_{max}}{\delta}\right).$$

□

Finally, we verify the correctness of Algorithm 11 and get an upper bound on $T(\delta, \mathcal{MC}_c)$.

Proof of Theorem 58. Similar to the proof of Theorem 54, we use the induction method to prove that Event-I happens for the k-th iteration with probability at least $1-(k-1)\delta/T_{max}$. For the first iteration, the solution to problem (5.1) is in $\mathcal{X} = \{1, 2, \ldots, N\}$ with probability 1. We assume that the claim is true for the first $k-1$ iterations, and consider the k-th

iteration. If one of the first two conditions holds when the current iteration terminates, then, by the same analysis as the proof of Theorem 54, we know that Event-I happens for the k-th iteration with probability at least $1-(k-1)\delta/T_{max}$. Hence, we only need consider the case when only the last condition holds when the current iteration terminates. Since the first two conditions do not hold, we know

$$\left|\hat{F}_n(q_{1/3}) - \hat{F}_n(q_{2/3})\right| \leq (q_{2/3} - q_{1/3}) \cdot 2c/5. \tag{5.8}$$

In addition, it holds that

$$\left|f(q_{1/3}) - \hat{F}_n(q_{1/3})\right| \leq (q_{2/3} - q_{1/3}) \cdot c/5, \quad \left|f(q_{2/3}) - \hat{F}_n(q_{2/3})\right| \leq (q_{2/3} - q_{1/3}) \cdot c/5$$

with probability at least $1-\delta/T_{max}$. Combining with inequality (5.8), we know that

$$\left|f(q_{1/3}) - f(q_{2/3})\right| \leq (q_{2/3} - q_{1/3}) \cdot 4c/5 < (q_{2/3} - q_{1/3}) \cdot c \tag{5.9}$$

holds with probability at least $1-\delta/T_{max}$. We assume that the above event and Event-I for the $(k-1)$-th iteration both hold, which has a joint probability of at least $1-\delta/T_{max}-(k-2)\delta/T_{max} = 1-(k-1)\delta/T_{max}$. If the solution to problem (5.1) is not in $\{q_{1/3}, \ldots, q_{2/3}\}$, then function $f(x)$ is monotone on $\{q_{1/3}, \ldots, q_{2/3}\}$ and

$$\left|f(q_{1/3}) - f(q_{2/3})\right| = \sum_{x=q_{1/3}}^{q_{2/3}-1} \left|f(x) - f(x+1)\right|.$$

Since the indifference zone parameter is c and the function $f(x)$ is convex, the function value difference between any two neighbouring points is at least c, which implies that

$$\sum_{x=q_{1/3}}^{q_{2/3}-1} |f(x) - f(x+1)| \geq (q_{2/3}) - q_{1/3}) \cdot c.$$

However, the above inequality contradicts inequality (5.9) and thus the solution to problem (5.1) is in $\{q_{1/3}, \ldots, q_{2/3}\}$. Hence, Event-I happens for the k-th iteration with probability at least $1-(k-1)\delta/T_{max}$.

Suppose that there are T iterations in Algorithm 11. Since the updating rule of intervals is not changed, Lemma 39 gives $T \leq T_{max}-1$. By the induction method, the solution to problem (5.1) is in $\{x_{L_{T+1}}, \ldots, x_{U_{k+1}}\}$ with probability at least $1-T\cdot\delta/T_{max} \geq 1-\delta+\delta/T_{max}$. Using the same analysis as Theorem 54, the point returned by the sub-problem is at most $2c/3$ larger than the optimal value with probability at least $1-\delta$. By the assumption that the indifference zone parameter is c, all feasible points have function values at least c larger than the optimal value. This implies that the solution returned by Algorithm 11 is optimal with probability at least $1-\delta+\delta/T_{max}-\delta/T_{max} \geq 1-\delta$ and Algorithm 11 is a $[(c,\delta)$-PCS-IZ, $\mathcal{MC}_c]$-algorithm. □

Proof of Theorem 61

Proof of Theorem 61. We construct the two models $\mathbb{M}_1, \mathbb{M}_2 \in \mathcal{MC}$ as

$$\nu_{1,x} := \mathcal{N}\left[cx, \sigma^2\right], \quad \nu_{2,x} := \mathcal{N}\left[c\left(|x-2|+2\right), \sigma^2\right], \quad \forall x \in \mathcal{X}.$$

Given a $[(c,\delta)$-PCS-IZ, $\mathcal{MC}_c]$-algorithm, the algorithm returns point 1 with probability at least $1-\delta$ when applied to model $\mathbb{M}_1$, and returns point 2 with probability at least $1-\delta$ when applied to model $\mathbb{M}_2$. We choose $\mathcal{E}$ as the event that the algorithm returns point 1 as the solution. Then, we know

$$\mathbb{P}_{\mathbb{M}_1}(\mathcal{E}) \geq 1-\delta, \quad \mathbb{P}_{\mathbb{M}_2}(\mathcal{E}) \leq \delta.$$

Using the monotonicity of function $d(x,y)$, we get

$$d(\mathbb{P}_{\mathbb{M}_1}(\mathcal{E}), \mathbb{P}_{\mathbb{M}_2}(\mathcal{E})) \geq d(1-\delta, \delta) \geq \log(1/2.4\delta). \tag{5.10}$$

Since the distributions $\nu_{1,x}$ and $\nu_{2,x}$ are Gaussian with variance σ^2, the KL divergence can be calculated as

$$\mathrm{KL}(\nu_{1,x}, \nu_{2,x}) = \frac{\left[cx - c\left(|x-2|+2\right)\right]^2}{2\sigma^2} = \begin{cases} 2c^2\sigma^{-2} & \text{if } x = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the summation can be calculated as

$$\sum_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{M}_1}\left[N_x(\tau)\right] \mathrm{KL}(\nu_{1,x}, \nu_{2,x}) = \frac{2c^2}{\sigma^2} \cdot \mathbb{E}_{\mathbb{M}_1}\left[N_1(\tau)\right]. \tag{5.11}$$

Substituting (5.10) and (5.11) into inequality (5.6), we know

$$\frac{2c^2}{\sigma^2} \cdot \mathbb{E}_{\mathbb{M}_1}\left[N_1(\tau)\right] \geq \log(1/2.4\delta),$$

which implies that

$$\mathbb{E}_{\mathbb{M}_1}[\tau] \geq \mathbb{E}_{\mathbb{M}_1}\left[N_1(\tau)\right] \geq \frac{\sigma^2}{2c^2} \cdot \log\left(\frac{1}{2.4\delta}\right) = \Theta\left[\frac{1}{c^2}\log\left(\frac{1}{\delta}\right)\right].$$

□

5.B Proofs in Section 5.3

Proof of Theorem 54

We first estimate the simulation cost of Algorithm 7. The following lemma gives an upper bound on the total number of iterations.

Lemma 39. *Suppose that Assumptions 5-7 hold. The number of iterations of Algorithm 7 is at most* $\log_{1.5}(N) + 1$.

Proof. If the total number of points N is at most 3, then there is no iteration. In the following proof, we assume $N \geq 4$. We first calculate the shrinkage of interval length after each iteration. We denote the upper and the lower bound at the beginning of the k-th iteration as x_{U_k} and x_{L_k}, respectively. Then, we know there are $n_k := x_{U_k} - x_{L_k} + 1$ points in the k-th iteration and the algorithm starts with $x_{L_1} = 1, x_{U_1} = N$. We define the 3-quantiles $q_{1/3} := \lfloor 2x_{L_k}/3 + x_{U_k}/3 \rfloor$ and $q_{2/3} := \lceil x_{L_k}/3 + 2x_{U_k}/3 \rceil$. By the updating rule, the next interval is

$$[x_{L_k}, q_{2/3}] \quad \text{or} \quad [q_{1/3}, x_{U_k}] \quad \text{or} \quad [q_{1/3}, q_{2/3}].$$

By discussing three cases when $n_k \in 3\mathbb{Z}$, $n_k \in 3\mathbb{Z} + 1$ and $n_k \in 3\mathbb{Z} + 2$, we know the next interval has at most $2n_k/3 + 1$ points, i.e.,

$$n_{k+1} \leq 2n_k/3 + 1. \tag{5.12}$$

Rewriting the inequality, we get the relation $n_{k+1} - 3 \leq 2(n_k - 3)/3$. Combining with the fact that $n_1 = N$, it follows that

$$n_k \leq \left(\frac{2}{3}\right)^{k-1} (N - 3) + 3.$$

Suppose Algorithm 7 terminates after T iterations. Then, it holds that $n_T \geq 4$ and $n_{T+1} \leq 3$. Hence, we know

$$4 \leq n_T \leq \left(\frac{2}{3}\right)^{T-1} (N - 3) + 3,$$

which implies

$$T \leq \log_{1.5}(N - 3) + 1 < \log_{1.5}(N) + 1.$$

□

In the next lemma, we estimate the simulation cost of each iteration.

Lemma 40. *Suppose that Assumptions 5-7 hold. The simulation cost of each iteration of Algorithm 7 is at most* $256\sigma^2\epsilon^{-2} \log(4T_{max}/\delta)$, *where* $T_{max} := \log_{1.5}(N) + 2$. *The simulation cost of the sub-problem is at most* $24\sigma^2\epsilon^{-2} \log(4T_{max}/\delta)$.

Proof. We first estimate the simulation cost at each iteration. If we choose $n = n_{\epsilon,\delta}$, the confidence half-width is $\epsilon/4$ and the condition $h[n, \sigma, 1-\delta/(2T_{max})] \leq \epsilon/8$ is satisfied. Hence, the simulation cost of each iteration is at most $2n_{\epsilon,\delta}$.

For the last sub-problem, we choose

$$\tilde{n} = \tilde{n}_{\epsilon,\delta} := \frac{8\sigma^2}{\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right)$$

and it follows that

$$h[\tilde{n}_{\epsilon,\delta}, \sigma, 1 - \delta/(2T_{max})] \leq \epsilon/2.$$

Hence, simulating $\tilde{n}_{\epsilon,\delta}$ times on each point is sufficient and the simulation cost is at most $3\tilde{n}_{\epsilon,\delta}$. □

Combining Lemmas 39 and 40, we get the total simulation cost of Algorithm 7.

Lemma 41. *Suppose that Assumptions 5-7 hold. The expected simulation cost of Algorithm 7 is at most*

$$\frac{256T_{max}\sigma^2}{\epsilon^2}\log\left(\frac{4T_{max}}{\delta}\right) = O\left[\frac{\log(N)}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right],$$

where $T_{max} := \log_{1.5}(N) + 2$.

Proof. By Lemmas 39 and 40, the total simulation cost of the first part is at most

$$[\log_{1.5}(N) + 1] \cdot \frac{256\sigma^2}{\epsilon^2}\log\left(\frac{4T_{max}}{\delta}\right) \leq [T_{max} - 1] \cdot \frac{256\sigma^2}{\epsilon^2}\log\left(\frac{4T_{max}}{\delta}\right)$$

and the simulation cost of the second part is at most

$$\frac{24\sigma^2}{\epsilon^2}\log\left(\frac{4T_{max}}{\delta}\right).$$

Combining two parts, we know the total simulation cost is at most

$$[T_{max} - 1] \cdot \frac{256\sigma^2}{\epsilon^2}\log\left(\frac{4T_{max}}{\delta}\right) + \frac{24\sigma^2}{\epsilon^2}\log\left(\frac{4T_{max}}{\delta}\right) \leq \frac{256T_{max}\sigma^2}{\epsilon^2}\log\left(\frac{4T_{max}}{\delta}\right).$$

□

Finally, we verify the correctness of Algorithm 7 and get an upper bound on $T(\epsilon, \delta\mathcal{MC})$.

Proof of Theorem 54. We denote $T_{max} := \log_{1.5}(N) + 2$. We also denote the upper and the lower bound at the beginning of the k-th iteration as x_{U_k} and x_{L_k}, respectively. We use the induction method to prove that, for the k-th iteration, at least one of the following two events happens with probability at least $1 - (k-1) \cdot \delta/T_{max}$:

- **Event-I.** A solution to problem (5.1) is in $\{x_{L_k}, \ldots, x_{U_k}\}$,
- **Event-II.** For any $x \in \{x_{L_k}, \ldots, x_{U_k}\}$, it holds $f(x) \leq \min_{y\in\mathcal{X}} f(y) + \epsilon$.

When $k = 1$, all solutions to problem (5.1) are in $\mathcal{X} = \{x_{L_1}, \ldots, x_{U_1}\}$ and Event-I happens with probability 1. Suppose the claim is true for the first $k-1$ iterations. We consider the k-th iteration. For the $(k-1)$-th iteration, if Event-II happens with probability at least $1 - (k-2) \cdot \delta/T_{max}$, then Event-II happens for the k-th iteration with the same probability.

This is because the interval $\{x_{L_k}, \ldots, x_{U_k}\}$ is a subset of $\{x_{L_{k-1}}, \ldots, x_{U_{k-1}}\}$ and all points in the new interval satisfy the condition of Event-II.

Hence, we only need to consider the case when only Event-I for the $(k-1)$-th iteration happens with probability at least $1-(k-2)\delta/T_{max}$. We assume Event-I happens and consider conditional probabilities in the following of the proof. We denote

$$q_{1/3} := \lfloor 2x_{L_{k-1}}/3 + x_{U_{k-1}}/3 \rfloor, \quad q_{2/3} := \lfloor x_{L_{k-1}}/3 + 2x_{U_{k-1}}/3 \rfloor$$

and discuss by two different cases.

Case I. Suppose one of the first two conditions holds when the current iteration terminates. Since the two conditions are symmetrical, we assume without loss of generality that the first condition holds. Then, the new interval is $[q_{1/3}, x_{U_{k-1}}]$ and, by the definition of confidence interval, we know

$$f(q_{1/3}) \geq f(q_{2/3})$$

holds with probability at least $1 - \delta/T_{max}$. We assume the above event and Event-I for the $(k-1)$-th iteration both happen, which has joint probability at least $1 - \delta/T_{max} - (k-2)\delta/T_{max} = 1 - (k-1)\delta/T_{max}$. By the convexity of $f(x)$, it holds that

$$f(x) \geq f(q_{1/3}) + \frac{x - q_{1/3}}{q_{1/3} - q_{2/3}} \left[f(q_{1/3}) - f(q_{2/3}) \right] \geq f(q_{1/3}), \quad \forall x \in \{x_{L_{k-1}}, \ldots, q_{1/3}\}.$$

Hence, the minimum of $f(x)$ in $\{x_{L_{k-1}}, \ldots, x_{U_{k-1}}\}$ is attained by a point in $\{q_{1/3}, \ldots, x_{U_{k-1}}\}$. Combining with the assumption that there exists a solution to problem (5.1) in

$$\{x_{L_{k-1}}, \ldots, x_{U_{k-1}}\},$$

we know that there exists a solution to problem (5.1) in $\{q_{1/3}, \ldots, x_{U_{k-1}}\}$. Thus, Event-I for the k-th iteration happens with probability at least $1 - (k-1)\delta/T_{max}$.

Case II. Suppose only the last condition holds when the current iteration terminates. Since the first two conditions do not hold, we have

$$\left| \hat{F}_n(q_{1/3}) - \hat{F}_n(q_{2/3}) \right| \leq \epsilon/4. \tag{5.13}$$

In addition, by the definition of confidence interval, it holds

$$\left| f(q_{1/3}) - \hat{F}_n(q_{1/3}) \right| \leq \epsilon/8, \quad \left| f(q_{2/3}) - \hat{F}_n(q_{2/3}) \right| \leq \epsilon/8$$

with probability at least $1 - \delta/T_{max}$. Combining with inequality (5.13), we know

$$\left| f(q_{1/3}) - f(q_{2/3}) \right| \leq \epsilon/2 \tag{5.14}$$

holds with probability at least $1 - \delta/T_{max}$. We assume that the above event and Event-I for the $(k-1)$-th iteration both happen, which has joint probability at least $1 - \delta/T_{max} - (k-2)\delta/T_{max} = 1 - (k-1)\delta/T_{max}$. We prove that if Event-I for the k-th iteration does not happen, then Event-II for the k-th iteration happens. Under the condition that Event-I does not happen, we assume without loss of generality that solutions to problem (5.1) are in $\{x_{L_{k-1}}, \ldots, q_{1/3} - 1\}$. Using the convexity of function $f(x)$, we know

$$f(x) \geq f(q_{1/3}) - \frac{q_{1/3} - x}{q_{2/3} - q_{1/3}} \left[f(q_{1/3}) - f(q_{2/3}) \right], \quad \forall x \in \{x_{L_{k-1}}, \ldots, q_{1/3}\}.$$

Choosing

$$x \in \left(\arg\min_{y \in \mathcal{X}} \ f(y) \right) \cap \{x_{L_{k-1}}, \ldots, x_{U_{k-1}}\} \neq \emptyset,$$

we get

$$\begin{aligned} \min_{y \in \mathcal{X}} \ f(y) &\geq f(q_{1/3}) - \frac{q_{1/3} - x}{q_{2/3} - q_{1/3}} \left[f(q_{1/3}) - f(q_{2/3}) \right] \\ &\geq f(q_{1/3}) - \frac{q_{1/3} - x_{L_{k-1}}}{q_{2/3} - q_{1/3}} \cdot \epsilon/2 \geq f(q_{1/3}) - \epsilon/2, \end{aligned}$$

where the last inequality is from the definition of 3-quantiles. Combining with inequality (5.14), we get

$$\min_{y \in \mathcal{X}} \ f(y) \geq f(q_{2/3}) - \epsilon.$$

By the convexity of $f(x)$, it holds that

$$\max_{x \in \{q_{1/3}, \ldots, q_{2/3}\}} \ f(x) = \max\{f(q_{1/3}), f(q_{2/3})\} \leq \min_{y \in \mathcal{X}} \ f(y) + \epsilon,$$

which means Event-II for the k-th iteration happens.

Combining the two cases, we know the claim holds for the k-th iteration. Suppose there are T iterations in Algorithm 7. By Lemma 39, we have $T \leq T_{max} - 1$. By the induction method, the last interval $\{x_{L_{T+1}}, \ldots, x_{U_{T+1}}\}$ satisfies the condition in Event-I or Event-II with probability at least $1 - T \cdot \delta/T_{max} \geq 1 - \delta + \delta/T_{max}$. If Event-II happens with probability at least $1 - \delta + \delta/T_{max}$, then regardless of the point chosen in the sub-problem, the solution returned by the algorithm has value at most ϵ larger than the optimal value with probability at least $1 - \delta + \delta/T_{max} \geq 1 - \delta$. Hence, the solution satisfies the (ϵ, δ)-PGS guarantee. Otherwise, we assume Event-I happens with probability at least $1 - \delta + \delta/T_{max}$. Then, a solution to problem (5.1) is in $\{x_{L_{T+1}}, \ldots, x_{U_{T+1}}\}$. We choose

$$x^* \in \left(\arg\min_{x \in \mathcal{X}} \ f(x) \right) \cap \{x_{L_{T+1}}, \ldots, x_{U_{T+1}}\}$$

and suppose the algorithm returns

$$x^{**} \in \arg\min_{x \in \{x_{L_{T+1}}, \ldots, x_{U_{T+1}}\}} \hat{F}_n(x).$$

By the definition of confidence interval, it holds

$$f(x^{**}) \le \hat{F}_n(x^{**}) + \epsilon/2, \quad f(x^*) \ge \hat{F}_n(x^*) - \epsilon/2$$

with probability at least $1 - \delta/T_{max}$. Under the above event, we get

$$f(x^{**}) \le \hat{F}_n(x^{**}) + \epsilon/2 \le \hat{F}_n(x^*) + \epsilon/2 \le f(x^*) + \epsilon.$$

Recalling that Event-I happens with probability at least $1 - \delta + \delta/T_{max}$, the point x^{**} satisfies the above relation with probability at least $1 - \delta$ and therefore satisfies the (ϵ, δ)-PGS guarantee. Combining with the first case, we know Algorithm 7 is an $[(\epsilon, \delta)\text{-PGS}, \mathcal{MC}]$-algorithm. □

Proof of Theorem 55

We first estimate the simulation cost of Algorithm 8.

Lemma 42. *Suppose that Assumptions 5-7 hold. The expected simulation cost for Algorithm 8 is at most*

$$\frac{25600\sigma^2}{\epsilon^2} \log\left[\frac{4N}{\delta}\right] = O\left[\frac{1}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right].$$

Proof. Denote $T_{max} := N$. Suppose that there are T iterations in Algorithm 8. We denote $\mathcal{S}_k$ as the active set at the beginning of the k-th iteration. Since each iteration reduces the size of $\mathcal{S}_k$ by at least 1, it follows that

$$|\mathcal{S}_k| \ge |\mathcal{S}_{T+1}| + T + 1 - k, \quad \forall k \in [T+1].$$

By the same analysis as Lemma 40, we know that for the k-th iteration, simulating

$$n(|\mathcal{S}_k|) := \frac{12800\sigma^2}{|\mathcal{S}_k|^2\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right)$$

times is sufficient to achieve $1 - \delta/(2T_{max})$ confidence half-width $|\mathcal{S}_k|/80 \cdot \epsilon$. By the condition on Line 6, each point discarded during the k-th iteration is simulated at most $n(|\mathcal{S}_k|)$ times. Hence, the total number of simulations on points discarded during the k-th iteration is at most

$$\begin{aligned}(|\mathcal{S}_k| - |\mathcal{S}_{k+1}|) \cdot n(|\mathcal{S}_k|) &= \frac{|\mathcal{S}_k| - |\mathcal{S}_{k+1}|}{|\mathcal{S}_k|^2} \cdot \frac{12800\sigma^2}{\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right) \\ &\le \left(\frac{1}{|\mathcal{S}_{k+1}|} - \frac{1}{|\mathcal{S}_k|}\right) \cdot \frac{12800\sigma^2}{\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right),\end{aligned}$$

where the inequality is because of $|\mathcal{S}_k| \ge |\mathcal{S}_{k+1}|$. Summing over $k = 1, 2, \ldots, T$, we get the number of simulations on all discarded points during iterations is at most

$$\sum_{k=1}^{T} \left(\frac{1}{|\mathcal{S}_{k+1}|} - \frac{1}{|\mathcal{S}_k|}\right) \cdot \frac{12800\sigma^2}{\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right) = \left(\frac{1}{|\mathcal{S}_{T+1}|} - \frac{1}{|\mathcal{S}_1|}\right) \cdot \frac{12800\sigma^2}{\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right)$$

$$\leq \left(1 - \frac{1}{N}\right) \cdot \frac{12800\sigma^2}{\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right) \leq \frac{12800\sigma^2}{\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right).$$

For points in the last active set $\mathcal{S}_{T+1}$, the number of simulations is bounded by

$$|\mathcal{S}_{T+1}| \cdot n(|\mathcal{S}_{T+1}|) = \frac{12800\sigma^2}{|\mathcal{S}_{T+1}|\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right) \leq \frac{12800\sigma^2}{\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right).$$

Combining two parts, we know that the simulation cost of Algorithm 8 is at most

$$\frac{25600\sigma^2}{\epsilon^2} \log\left(\frac{4T_{max}}{\delta}\right).$$

□

Then, we prove the correctness of Algorithm 8. The following lemma plays a critical role in verifying the correctness of Type-II Operations.

Lemma 43. *Suppose function $h(x)$ is convex on $[1, M+a]$, where integer $M \geq 3$ and constant $a \in [0,1]$. Then, the restriction of function $h(x)$ to $[M]$, which we denote as $\tilde{h}(x)$, is also convex. Furthermore, given a constant $\epsilon > 0$, if it holds that*

$$\max_{x\in[M]} \tilde{h}(x) - \min_{x\in[M]} \tilde{h}(x) \leq M/20 \cdot \epsilon, \tag{5.15}$$

then we know

$$\min_{y\in[M']} \tilde{h}(2y-1) - \min_{x\in[1,M+a]} h(x) \leq \epsilon/2.$$

where we define $M' := \lceil M/2 \rceil$.

Proof. Since the mid-point convexity of $h(x)$ implies the discrete mid-point convexity of $\tilde{h}$, we know $\tilde{h}(x)$ is also convex. We prove the second claim in three steps.

Step 1. We first prove that

$$\min_{x\in[1,M]} h(x) - \min_{x\in[1,M+a]} h(x) \leq \epsilon/10. \tag{5.16}$$

Suppose x^* is a minimizer of $h(x)$ on $[1, M+a]$. If $x^* \in [1, M]$, then inequality (5.16) holds trivially. We assume that $x^* \in (M, M+a]$. By the convexity of $h(x)$, we have

$$h(M) - h(x^*) \leq \frac{x^* - M}{M-1} \cdot [h(M) - h(1)] \leq \frac{1}{M-1} \cdot [h(M) - h(1)] \leq \frac{M/20 \cdot \epsilon}{M/2} = \epsilon/10.$$

Hence, we know

$$\min_{x\in[1,M]} h(x) - \min_{x\in[1,M+a]} h(x) = \min_{x\in[1,M]} h(x) - h(x^*) \leq h(M) - h(x^*) \leq \epsilon/10.$$

Step 2. Next, we prove that

$$\min_{x\in[M]} \tilde{h}(x) - \min_{x\in[1,M]} h(x) \leq \epsilon/5. \tag{5.17}$$

Let x^* be a minimizer of $\tilde{h}(x)$ on $[M]$. By inequality (5.15), we know

$$\max_{x\in[M]} \tilde{h}(x) = \max\{\tilde{h}(1), \tilde{h}(M)\} \leq \tilde{h}(x^*) + M/20 \cdot \epsilon.$$

By the convexity of $h(x)$, there exists a minimizer $x^{**} \in [1, M]$ of $h(x)$ in $(x^* - 1, x^* + 1)$. If $x^{**} = x^*$, then $\min \tilde{h}(x) = \min h(x)$ and inequality (5.17) holds. Hence, we assume that $x^{**} \neq x^*$ and, without loss of generality, $x^{**} \in (x^*, x^* + 1)$. Since $(M - x^* - 1) + (x^* - 1) = M - 2$, we know $\max\{M - x^* - 1, x^* - 1\} \geq \lceil (M-2)/2 \rceil$. We first consider the case when

$$M - x^* - 1 \geq \lceil (M-2)/2 \rceil.$$

By the convexity of $h(x)$, we have

$$\begin{aligned} h(x^* + 1) - h(x^{**}) &\leq \frac{x^* + 1 - x^{**}}{M - x^* - 1} \cdot [h(M) - h(x^* + 1)] \\ &\leq \frac{1}{\lceil (M-2)/2 \rceil} \cdot [h(M) - h(x^*)] \leq \frac{M/20 \cdot \epsilon}{\lceil (M-2)/2 \rceil}. \end{aligned}$$

By simple calculations, we get $M/4 \leq \lceil (M-2)/2 \rceil$ for all $M \geq 3$ and therefore

$$h(x^*) - h(x^{**}) \leq h(x^* + 1) - h(x^{**}) \leq \epsilon/5,$$

which means inequality (5.17) holds. Now we consider the case when

$$x^* - 1 \geq \lceil (M-2)/2 \rceil.$$

Similarly, by the convexity of $h(x)$, we have

$$\begin{aligned} h(x^*) - h(x^{**}) &\leq \frac{x^{**} - x^*}{x^* - 1} \cdot [h(x^*) - h(1)] \leq \frac{1}{\lceil (M-2)/2 \rceil} \cdot [h(x^*) - h(1)] \\ &\leq \frac{M/20 \cdot \epsilon}{\lceil (M-2)/2 \rceil} \leq \epsilon/5. \end{aligned}$$

Combining the two cases, we know inequality (5.17) holds.

Step 3. Finally, we prove that

$$\min_{y\in[M']} \tilde{h}(2y - 1) - \min_{x\in[M]} \tilde{h}(x) \leq \epsilon/5. \tag{5.18}$$

Let x^* be a minimizer of $\tilde{h}(x)$. If x^* is an odd number, then $\min_{y\in[M']} \tilde{h}(2y-1) = \min_{x\in[M]} \tilde{h}(x)$ and inequality (5.18) holds. Otherwise, we assume $x^* = 2y^*$ is an even number. Then, by the convexity of $\tilde{h}(x)$, there exists a minimizer of $\tilde{h}(2y-1)$ in $\{y^*, y^*+1\}$. Without loss of generality, we assume y^*+1 is a minimizer of $\tilde{h}(2y-1)$. Since $(M-x^*)+(x^*-1) = M-1$, we have $\max\{M-x^*, x^*-1\} \geq \lceil (M-1)/2 \rceil$. We first consider the case when

$$M - x^* \geq \lceil (M-1)/2 \rceil.$$

By the convexity of $\tilde{h}(x)$, we have

$$\tilde{h}(2y^*+1) - \tilde{h}(2y^*) \leq \frac{1}{M-2y^*} \cdot \left[\tilde{h}(M) - \tilde{h}(2y^*)\right] \leq \frac{M/20 \cdot \epsilon}{\lceil (M-1)/2 \rceil}.$$

We can verify that $M/4 \leq \lceil (M-1)/2 \rceil$ for all $M \geq 3$. Hence, it holds that

$$\tilde{h}(2y^*+1) - \tilde{h}(2y^*) \leq \epsilon/5.$$

Then, we consider the case when

$$x^* - 1 \geq \lceil (M-1)/2 \rceil.$$

Similarly, using the convexity of $\tilde{h}(x)$, we have

$$\tilde{h}(2y^*-1) - \tilde{h}(2y^*) \leq \frac{1}{2y^*-1} \cdot \left[\tilde{h}(2y^*) - \tilde{h}(1)\right] \leq \frac{M/20 \cdot \epsilon}{\lceil (M-1)/2 \rceil} \leq \epsilon/5,$$

which implies that

$$\tilde{h}(2y^*+1) - \tilde{h}(2y^*) \leq \tilde{h}(2y^*-1) - \tilde{h}(2y^*) \leq \epsilon/5.$$

Combining the two cases, we know inequality (5.18) holds.

By inequalities (5.16), (5.17) and (5.18), we have

$$\min_{y\in[M']} \tilde{h}(2y-1) - \min_{x\in[1,M+a]} h(x) \leq \epsilon/10 + \epsilon/5 + \epsilon/5 = \epsilon/2.$$

□

We denote $\mathcal{S}_k$ and d_k as the active set and the step size at the beginning of the k-th iteration, respectively. We define the upper bound and the lower bound for the k-th iteration as

$$x_{x_{L_1}} := 1, \quad x_{L_{k+1}} := \begin{cases} y + d_k & \text{if the second case of Type-I Operation happens} \\ x_{L_k} & \text{otherwise,} \end{cases}$$

$$x_{x_{U_1}} := N, \quad x_{U_{k+1}} := \begin{cases} y - d_k & \text{if the first case of Type-I Operation happens} \\ x_{U_k} & \text{otherwise.} \end{cases}$$

Although not explicitly defined in the algorithm, the interval $\{x_{L_k}, \ldots, x_{U_k}\}$ plays a similar role as in the TS algorithm and characterizes the set of possible solutions. In the following lemma, we prove that the active set $\mathcal{S}_k$ is a good approximation to the interval $\{x_{L_k}, \ldots, x_{U_k}\}$. We note that the following lemma is deterministic.

Lemma 44. *For any iteration k, we have*

$$x_{L_k} = \min\ \mathcal{S}_k \quad \textit{and} \quad x_{U_k} \leq \max\ \mathcal{S}_k + d_k. \tag{5.19}$$

Proof. We use the induction method to prove the result. When $k = 1$, we know $x_{L_1} = 1, x_{U_1} = N, \mathcal{S} = [N]$ and $d_1 = 1$. Hence, the relations in (5.19) hold. We assume these relations hold for the first $k-1$ iterations. We discuss by two different cases.

Case I. Type-I Operation is implemented during the $(k-1)$-th iteration. If the first case of Type-I Operation happens, then we know $x_{L_k} = x_{L_{k-1}}$ and $x_{U_k} = y - d_{k-1}$. By the updating rule, the step size d_{k-1} is not changed and all points in $\mathcal{S}_{k-1}$ that are at least y are discarded from $\mathcal{S}_{k-1}$. Hence, it follows that $\max\ \mathcal{S}_k = x_{U_k}$ and the inequality $x_{U_k} \leq \max\ \mathcal{S}_k + d_k$ holds. Moreover, since both x_{L_k} and $\min\ \mathcal{S}_{k-1}$ are not changed, the equality $x_{L_k} = \min\ \mathcal{S}_k$ still holds.

Otherwise if the second case of Type-I Operation happens, then we know $x_{L_k} = y + d_{k-1}$ and $x_{U_k} = x_{U_{k-1}}$. Similarly, we can prove that $x_{L_k} = \min\ \mathcal{S}_k$. Moreover, since $d_k = d_{k-1}$ and $\max\ \mathcal{S}_{k-1} + d_{k-1} = \max\ \mathcal{S}_k + d_{k-1}$, it holds

$$x_{U_k} = x_{U_{k-1}} \leq \max\ \mathcal{S}_{k-1} + d_{k-1} = \max\ \mathcal{S}_k + d_k.$$

Case II. Type-II Operation is implemented during the $(k-1)$-th iteration. In this case, bounds $x_{L_{k-1}}$ and $x_{U_{k-1}}$ are not changed. By the update rule, we know the step size $d_k = 2d_{k-1}$ and

$$\min\ \mathcal{S}_k = \min\ \mathcal{S}_{k-1}, \quad \max\ \mathcal{S}_k \in \{\max\ \mathcal{S}_{k-1} - d_{k-1}, \max\ \mathcal{S}_{k-1}\}. \tag{5.20}$$

Thus, the equality $x_{L_k} = \min\ \mathcal{S}_k$ still holds. By the induction assumption, we know that

$$x_{U_k} \leq x_{U_{k-1}} \leq \max\ \mathcal{S}_{k-1} + d_{k-1}.$$

Combining with the latter relation in (5.20), we get

$$x_{U_k} \leq \max\ \mathcal{S}_{k-1} + d_{k-1} \leq \max\ \mathcal{S}_k + 2d_{k-1} = \max\ \mathcal{S}_k + d_k.$$

Combining the two cases, we know the relations in (5.19) hold for the k-th iteration. By the induction method, the relations hold for all iterations. □

Finally, utilizing Lemmas 43 and 44, we can prove the correctness of Algorithm 8 and get a better upper bound on $T_0(\epsilon, \delta, \mathcal{MC})$.

Proof of Theorem 55. Denote $T_{max} := N$. We use the induction method to prove that, for any iteration k, the two events

- $\min_{x \in \mathcal{S}_k}\ f(x) \leq \min_{x \in \mathcal{X}}\ f(x) + \epsilon/2$.

- $\min_{x\in\{x_{L_k}, x_{L_k}+1,\ldots,x_{U_k}\}} f(x) = \min_{x\in\mathcal{X}} f(x)$.

happen jointly with probability at least $1-(k-1)\delta/T_{max}$. When $k=1$, we know $\mathcal{S}_1=\mathcal{X}$ and $x_{L_1}=1, x_{U_1}=N$. Hence, the two events happen with probability 1. Suppose the claim is true for the first $k-1$ iterations. We assume the two events happen for the $(k-1)$-th iteration and consider conditional probabilities in the following proof. We discuss by two different cases.

Case I. Type-I Operation is implemented in the $(k-1)$-th iteration. In this case, there exists $x, y \in \mathcal{S}_{k-1}$ such that $\hat{F}_{n_x}(x) + h_x \le \hat{F}_{n_y}(y) - h_y$. By the definition of confidence intervals, we know $f(x) \le f(y)$ holds with probability at least $1-\delta/T_{max}$. We assume event $f(x)\le f(y)$ happens jointly with the claim for the $(k-1)$-th iteration, which has probability at least $1-(k-2)\delta/T_{max}-\delta/T_{max} = 1-(k-1)\delta/T_{max}$. If $x<y$, then using the convexity of $f(x)$, we know

$$f(z) \ge f(y) \ge f(x), \quad \forall z\in[N] \quad \text{s.t. } z\ge y,$$

which means all discarded points have function values at least $f(x)$. Hence, the minimums in the claim are not changed, i.e., we have

$$\min_{x\in\mathcal{S}_k} f(x) = \min_{x\in\mathcal{S}_{k-1}} f(x) \le \min_{x\in\mathcal{X}} f(x) + \epsilon/2$$

and

$$\min_{x\in\{x_{L_k}, x_{L_k}+1,\ldots,x_{U_k}\}} f(x) = \min_{x\in\{x_{L_{k-1}}, x_{L_{k-1}}+1,\ldots,x_{U_{k-1}}\}} f(x) = \min_{x\in\mathcal{X}} f(x).$$

The two events happen with probability at least $1-(k-1)\delta/T_{max}$. If $y<x$, the proof is the same and therefore the claim holds for the k-th iteration.

Case II. Type-II Operation is implemented in the $(k-1)$-th iteration. Since $x_{L_{k-1}}$ and $x_{U_{k-1}}$ are not changed, the first event happens for the k-th iteration. Hence, we only need to verify that the second event happens with high probability. Let x^* and x^{**} be a minimizer and a maximizer of $f(x)$ on $\mathcal{S}_{k-1}$, respectively. By the condition of Type-II Operations, we know

$$|\hat{F}_n(x^*) - \hat{F}_n(x^{**})| \le h_{x^*} + h_{x^{**}}$$

and

$$h_x \le |\mathcal{S}_{k-1}|/80\cdot\epsilon, \quad \forall x\in\mathcal{S}_{k-1}.$$

By the definition of confidence intervals, it holds

$$|f(x^*) - \hat{F}_n(x^*)| \le h_{x^*}, \quad |f(x^{**}) - \hat{F}_n(x^{**})| \le h_{x^{**}}$$

with probability at least $1-\delta/T_{max}$. Under the above event, we have

$$|f(x^*) - f(x^{**})| \le |f(x^*) - \hat{F}_n(x^*)| + |\hat{F}_n(x^*) - \hat{F}_n(x^{**})| + |f(x^{**}) - \hat{F}_n(x^{**})| \quad (5.21)$$

$$\leq 2(h_{x^*}) + h_{x^{**}} \leq M/20 \cdot \epsilon.$$

We assume the above event happens jointly with with the claim for the $(k-1)$ -th iteration, which has probability at least $1-(k-1)\delta/T_{max}$. By the induction assumption, the original problem (5.1) is equivalent to

$$\min_{x\in\{x_{L_{k-1}}, x_{L_{k-1}}+1,\ldots,x_{U_{k-1}}\}} f(x).$$

Moreover, if we denote $\tilde{f}$ as the linear interpolation of $f(x)$ defined in (5.2), then the above problem is equivalent to

$$\min_{x\in\{x_{L_{k-1}}, x_{L_{k-1}}+1,\ldots,x_{U_{k-1}}\}} f(x) = \min_{x\in[x_{L_{k-1}}, x_{U_{k-1}}]} \tilde{f}(x). \tag{5.22}$$

We define the constant $\tilde{M} := (x_{U_{k-1}} - x_{L_{k-1}})/d_{k-1} + 1$ and the linear transformation

$$T(x) := x_{L_{k-1}} + d_{k-1}(x-1), \quad \forall x \in [1, \tilde{M}].$$

The inverse image $T^{-1}\left([x_{L_{k-1}}, x_{U_{k-1}}]\right)$ is $[1, \tilde{M}]$. Defining the composite function

$$\tilde{g}(x) := \tilde{f}(T(x)), \quad \forall x \in [1, \tilde{M}],$$

we know that the problem (5.22) is equivalent to

$$\min_{x\in[1,\tilde{M}]} \tilde{g}(x). \tag{5.23}$$

The inverse image $T^{-1}(\mathcal{S}_{k-1})$ is $[M]$, where $M := |\mathcal{S}_{k-1}|$ is the number of points in $\mathcal{S}_{k-1}$. Lemma 44 implies that $a := \tilde{M} - M \in [0,1]$. Recalling inequality (5.21), we get

$$\max_{x\in[M]} \tilde{g}(x) - \min_{x\in[M]} \tilde{g}(x) \leq M/20 \cdot \epsilon.$$

Now, we can apply Lemma 43 to get

$$\min_{y\in[M']} \tilde{g}(2y-1) - \min_{x\in[1,M+a]} \tilde{g}(x) = \min_{y\in[M']} \tilde{g}(2y-1) - \min_{x\in[1,\tilde{M}]} \tilde{g}(x) \leq \epsilon/2,$$

where $M' := \lceil M/2 \rceil$. We note that the interval $[1, M+a]$ corresponds to the interval $[1, N]$ before scaling the x-axis. Since problem (5.23) is equivalent to problem (5.22) and further equivalent to problem (5.1), it holds that

$$\min_{y\in[M']} \tilde{g}(2y-1) - \min_{x\in[1,\tilde{M}]} \tilde{g}(x) = \min_{y\in[M']} \tilde{g}(2y-1) - \min_{x\in\mathcal{X}} f(x) \leq \epsilon/2.$$

By the definition of $\mathcal{S}_k$, we know $T(2[M']-1)$ is $\mathcal{S}_k$ and therefore

$$\min_{y\in[M']} \tilde{g}(2y-1) - \min_{x\in\mathcal{X}} f(x) = \min_{x\in\mathcal{S}_k} f(x) - \min_{x\in\mathcal{X}} f(x) \leq \epsilon/2,$$

which implies that the second case happens for the k-th iteration with probability at least $1-(k-1)\delta/T_{max}$.

Combining the above two cases, we know the claim holds for all iterations. Suppose there are T iterations in Algorithm 8. Since each iteration will decrease the active set $\mathcal{S}$ by at least 1, we get $T \leq N-1$. Then after the T iterations, we have

$$\min_{x\in\mathcal{S}_{T+1}} f(x) \leq \min_{x\in\mathcal{X}} f(x) + \epsilon/2 \tag{5.24}$$

holds with probability at least $1 - T\cdot\delta/T_{max} \geq 1-\delta+\delta/T_{max}$. For the sub-problem, using the same analysis as Theorem 54, the point returned by Algorithm 8 satisfies the $(\epsilon/2, \delta/T_{max})$-PGS guarantee. Combining with the relation (5.24), we know the algorithm returns a solution satisfying the (ϵ,δ)-PGS guarantee. □

5.C Multi-dimensional Shrinking Uniform Sampling Algorithm

In this section, we give the multi-dimensional version of the SUS algorithm designed in Section 5.3. Similar to the one-dimensional case, the asymptotic simulation cost of the multi-dimensional algorithm is upper bounded by a constant that does not depend on the problem scale N and the Lipschitz constant of the objective function. Hence, the multi-dimensional algorithm provides a matching simulation cost to the one-dimensional case. However, the expected simulation cost is exponentially dependent on the dimension d. Therefore, the multi-dimensional SUS algorithm is mainly theoretical and only suitable for low-dimensional problems.

The main idea of the generalization to multi-dimensional problems is to view optimization algorithms as (usually biased) estimators to the optimal value, which is elaborated in the following definition.

Definition 16. Given a constant $S > 0$, we say that an algorithm is *sub-Gaussian with dimension d and parameter S* if for any d-dimensional $L^\natural$-convex problem, any $\epsilon > 0$ and small enough $\delta > 0$, the algorithm returns an ϵ-optimal solution $\hat{x}$ along with an estimate $\hat{f}^*$ to the optimal value f^* that satisfies $|\hat{f}^* - f^*| \leq \epsilon$ with probability at least $1-\delta$ using at most

$$T(\epsilon,\delta) := \tilde{O}\left[\frac{2S}{\epsilon^2}\log(\frac{2}{\delta})\right]$$

simulation runs.

For example, Theorem 55 shows that the one-dimensional SUS algorithm (Algorithm 8) returns an $(\epsilon/2, \delta/2)$-PGS solution with $\tilde{O}[\epsilon^{-2}\log(1/\delta)]$ simulations. Then, we can simulate the function value at the solution for $O[\epsilon^{-2}\log(1/\delta)]$ times such that the $1-\delta/2$ confidence half-width becomes smaller than $\epsilon/4$. Then, the empirical mean of function values at the

solution is at most ϵ distant from f^* with probability at least $1-\delta$. Hence, we know that Algorithm 8 is sub-Gaussian with dimension 1. We denote its associated parameter as S. We note that if we treat algorithms as estimators, the estimators are generally "*biased*" (but consistent). This fact implies that the empirical mean of several estimates to the optimal value does not produce a better optimality guarantee, while the empirical mean of several unbiased estimators usually has a tighter deviation bound.

Now, we inductively construct sub-Gaussian algorithms for multi-dimensional problems. We first define the marginal objective function as

$$f^{d-1}(x) := \min_{y\in[N]^{d-1}} f(y,x). \tag{5.25}$$

Observe that each evaluation of $f^{d-1}(x)$ requires solving a $(d-1)$-dimensional $L^\natural$-convex sub-problem. Hence, if we have an algorithm for $(d-1)$-dimensional $L^\natural$-convex problems, we only need to solve the one-dimensional problem

$$\min_{x\in[N]} f^{d-1}(x) = \min_{x\in[N]} \min_{y\in[N]^{d-1}} f(y,x) = \min_{x\in\mathcal{X}} f(x) \tag{5.26}$$

Moreover, we can prove that problem (5.26) is also a convex problem.

Lemma 45. *If function $f(x)$ is $L^\natural$-convex, then function $f^{d-1}(x)$ is $L^\natural$-convex on $[N]$.*

Based on the observations, we can use sub-Gaussian algorithms for $(d-1)$-dimensional problems and Algorithm 8 to construct sub-Gaussian algorithms for d-dimensional problems. We give the pseudo-code in Algorithm 13.

Algorithm 13 Multi-dimensional shrinking uniform sampling algorithm

Input: Model $\mathcal{X}, (\mathsf{Y}, \mathcal{B}_\mathsf{Y}), F(x,\xi_x)$, optimality guarantee parameters ϵ and δ, sub-Gaussian algorithm $\mathcal{A}$ with dimension $d-1$.

Output: An (ϵ,δ)-PGS solution x^* to problem (5.1).

1: Set the active set $\mathcal{S} \leftarrow [N]$.
2: Set the step size $s \leftarrow 1$ and the maximal number of comparisons $T_{max} \leftarrow N$.
3: Set $N_{cur} \leftarrow +\infty$.
4: **while** the size of $\mathcal{S}$ is at least 3 **do** ▷ Iterate until $\mathcal{S}$ has at most 2 points.
5: **if** $|\mathcal{S}| \le N_{cur}/2$ **then** ▷ Update the confidence interval.
6: Record current active set size $N_{cur} \leftarrow |\mathcal{S}|$.
7: Set the confidence half-width $h \leftarrow N_{cur}\cdot\epsilon/160$.
8: For each $x \in \mathcal{S}$, use algorithm $\mathcal{A}$ to get an estimate to $f^{d-1}(x)$ such that

$$\left|\hat{f}^{d-1}(x) - f^{d-1}(x)\right| \le h$$

holds with probability at least $1-\delta/(2T_{max})$.
9: **end if**

10: **if** $\hat{f}^{d-1}(x) + h \leq \hat{f}^{d-1}(y) - h$ for some $x, y \in \mathcal{S}$ **then** ▷ **Type-I Operation**
11: **if** $x < y$ **then**
12: Remove all points $z \in \mathcal{S}$ with the property $z \geq y$ from $\mathcal{S}$.
13: **else**
14: Remove all points $z \in \mathcal{S}$ with the property $z \leq y$ from $\mathcal{S}$.
15: **end if**
16: **else** ▷ **Type-II Operation**
17: Update the step size $s \leftarrow 2s$.
18: Update $\mathcal{S} \leftarrow \{x_{min}, x_{min} + s, \ldots, x_{min} + ks\}$, where $x_{min} = \min_{x \in \mathcal{S}} x$ and $k = \lceil |\mathcal{S}|/2 \rceil - 1$.
19: **end if**
20: **end while** ▷ Now $\mathcal{S}$ has at most 2 points.
21: For each $x \in \mathcal{S}$, use Algorithm $\mathcal{A}$ to obtain an estimate to $f^{d-1}(x)$ such that

$$\left| \hat{f}^{d-1}(x) - f^{d-1}(x) \right| \leq \epsilon/4$$

holds with probability at least $1 - \delta/(2T_{max})$.
22: Return $x^* \leftarrow \arg\min_{x \in \mathcal{S}} \hat{f}^{d-1}(x)$.

We note that the bound $N_{cur} \cdot \epsilon/160$ in Line 7 is tighter than the condition in Algorithm 8. This is also because the "algorithm estimators" are usually biased. Therefore, taking the mean of several algorithm estimators does not necessarily reduce the width of confidence intervals, and the algorithm needs to generate a new estimator with a better confidence interval to guarantee the desired PGS criterion (instead of taking the mean of two less accurate estimators to improve the precision). We prove that Algorithm 13 is sub-Gaussian with dimension d and estimate its parameter.

Theorem 62. *Suppose that Assumptions 5-7 hold, and that Algorithm $\mathcal{A}$ is sub-Gaussian with dimension $d-1$ and parameter S. Then, Algorithm 13 is a sub-Gaussian algorithm with dimension d and parameter MC, where $M > 0$ is an absolute constant.*

If we treat $F(x, \xi_x)$ as a sub-Gaussian algorithm with dimension 0 and parameter σ^2, then Theorem 62 implies that there exists a sub-Gaussian algorithm with dimension 1 and parameter $\sigma^2 M$. However, the parameter S of Algorithm 8 is usually smaller than $\sigma^2 M$ and therefore Algorithm 8 is preferred in the one-dimensional case. Using the results of Theorem 62 and the fact that Algorithm 8 is sub-Gaussian with dimension 1, we can inductively construct sub-Gaussian algorithms with any dimension d.

Theorem 63. *Suppose that Assumptions 5-7 hold and Algorithm 8 is sub-Gaussian with dimension 1 and parameter S. There exists an $[(\epsilon, \delta)$-PGS, $\mathcal{MC}]$-algorithm that is sub-Gaussian with parameter $M^{d-1}S$, where M is the constant in Theorem 62. Hence, we have*

$$T(\epsilon, \delta, \mathcal{MC}) = \tilde{O}\left[\frac{M^d}{\epsilon^2} \log\left(\frac{1}{\delta} \right) \right].$$

Furthermore, by choosing $\epsilon = c/2$, it holds that

$$T(\delta, \mathcal{MC}_c) = \tilde{O}\left[\frac{M^d}{c^2}\log\left(\frac{1}{\delta}\right)\right].$$

We note that although the upper bound in Theorem 62 is independent of the Lipschitz constant L and independent of N when $\delta \ll 1$, the dependence on d is exponential. Hence, Algorithm 13 is largely theoretical and only suitable for low-dimensional problems, e.g., problems with $d \leq 3$. On the other hand, if the dimension d is treated as a fixed constant, Algorithm 13 attains the optimal asymptotic performance under the asymptotic criterion in [128]. We also mention that Algorithm 13 does not make a full use of the properties of $L^\natural$-convex functions. Actually, Algorithm 13 is an (ϵ, δ)-PGS algorithm for those functions that are convex in each direction.

Proof of Lemma 45

Proof of Lemma 45. Let $k \in \{2, 3, \ldots, N-1\}$. By the definition of $f^{d-1}(x)$, there exists vectors $y_{k-1}, y_{k+1} \in [N]^{d-1}$ such that

$$f^{d-1}(k-1) = f(y_{k-1}, k-1), \quad f^{d-1}(k+1) = f(y_{k+1}, k+1).$$

By the $L^\natural$-convexity of $f(x)$, we have

$$\begin{aligned} f^{d-1}(k-1) + f^{d-1}(k+1) &= f(y_{k-1}, k-1) + f(y_{k+1}, k+1) \\ &\geq f\left(\left\lceil \frac{y_{k-1}+y_{k+1}}{2} \right\rceil, k\right) + f\left(\left\lfloor \frac{y_{k-1}+y_{k+1}}{2} \right\rfloor, k\right) \\ &\geq 2 \min_{y \in [N]^{d-1}} f(y, k) = 2f^{d-1}(k), \end{aligned}$$

which means the discrete mid-point convexity holds at point k. Since we can choose k arbitrarily, we know function $f^{d-1}(x)$ is convex on $[N]$. □

Proof of Theorem 62

Proof of Theorem 62. We first verify the correctness of Algorithm 13. The algorithm is the same as Algorithm 8 except the condition for implementing Type-II Operations. Hence, if we can prove that, when Type-II Operations are implemented, it holds

$$h \leq |\mathcal{S}| \cdot \epsilon/80, \tag{5.27}$$

then the proof of Theorem 55 can be directly applied to this case. If the confidence interval is updated at the beginning of current iteration, then we have

$$h = |\mathcal{S}| \cdot \epsilon/160 < |\mathcal{S}| \cdot \epsilon/80.$$

Otherwise, if the confidence interval is not updated in the current iteration. Then, we have $|\mathcal{S}| > N_{cur}/2$ and therefore

$$h = N_{cur} \cdot \epsilon/160 < 2|\mathcal{S}| \cdot \epsilon/160 = |\mathcal{S}| \cdot \epsilon/80.$$

Combining the two cases, we have inequality (5.27) and the correctness of Algorithm 13.

Next, we estimate the simulation cost of Algorithm 13. Denote the active sets when we update the confidence interval as $\mathcal{S}_1, \ldots, \mathcal{S}_m$, where $m \geq 1$ is the number of times when the confidence interval is updated. Then, we know $|\mathcal{S}_1| = N$ and $|\mathcal{S}_m| \geq 3$. By the condition for updating the confidence interval, it holds

$$|\mathcal{S}_{k+1}| \leq |\mathcal{S}_k|/2, \quad \forall k \in [m-1],$$

which implies

$$|\mathcal{S}_k| \geq 2^{m-k}|\mathcal{S}_m| \geq 3 \cdot 2^{m-k}, \quad \forall k \in [m].$$

Since the algorithm $\mathcal{A}$ is sub-Gaussian with parameter S, for each $x \in \mathcal{S}_k$, the simulation cost for generating $\hat{f}^{d-1}(x)$ is at most

$$\frac{2S}{h^2}\log(\frac{2T_{max}}{\delta}) = \frac{2S}{160^{-2}|\mathcal{S}_k|^2\epsilon^2}\log(\frac{2T_{max}}{\delta}) = |\mathcal{S}_k|^{-2} \cdot \frac{51200S}{\epsilon^2}\log(\frac{2T_{max}}{\delta}).$$

Hence, the total simulation cost for the k-th update of confidence intervals is at most

$$|\mathcal{S}_k| \cdot |\mathcal{S}_k|^{-2} \cdot \frac{51200S}{\epsilon^2}\log(\frac{2T_{max}}{\delta}) = |\mathcal{S}_k|^{-1} \cdot \frac{51200S}{\epsilon^2}\log(\frac{2T_{max}}{\delta}) \leq 2^{k-m}/3 \cdot \frac{51200S}{\epsilon^2}\log(\frac{2T_{max}}{\delta}).$$

Summing over all iterations, we have the simulation cost of all iterations of Algorithm 13 is at most

$$\sum_{k=1}^{m} 2^{k-m}/3 \cdot \frac{51200S}{\epsilon^2}\log(\frac{2T_{max}}{\delta}) = \left(2 - 2^{1-m}\right) \cdot \frac{51200S}{3\epsilon^2}\log(\frac{2T_{max}}{\delta}) < \frac{102400S}{3\epsilon^2}\log(\frac{2T_{max}}{\delta}).$$

Now we consider the simulation cost of the last subproblem. Since the algorithm 13 is sub-Gaussian with parameter S, the simulation cost of the subproblem is at most

$$2 \cdot \frac{2S}{(\epsilon/4)^2}\log(\frac{2T_{max}}{\delta}) = \frac{64S}{\epsilon^2}\log(\frac{2T_{max}}{\delta}).$$

Hence, the total simulation cost of Algorithm 13 is at most

$$\frac{102400S}{3\epsilon^2}\log(\frac{2T_{max}}{\delta}) + \frac{64S}{\epsilon^2}\log(\frac{2T_{max}}{\delta}) < 17099 \cdot \frac{2S}{\epsilon^2}\log(\frac{2T_{max}}{\delta}).$$

When δ is small enough, we can choose $M = 17100$ and the asymptotic simulation cost of Algorithm 13 is at most

$$\frac{2MC}{\epsilon^2}\log(\frac{2T_{max}}{\delta}),$$

which implies that Algorithm 13 is sub-Gaussian with dimension d and parameter MC. $\square$

5.D Deterministic cutting-plane methods

In this section, we give the pseudo-codes of Vaidya's cutting-plane method [221] and the random walk-based cutting-plane method [20] for the self-contained purpose.

Vaidya's Cutting-plane Method

We first give the pseudo-code for Vaidya's cutting-plane method [221]. We note that examples of Newton-type methods include the original Newton method, quasi-Newton methods and the cubic-regularized Newton method.

Algorithm 14 Vaidya's cutting-plane method

Input: Model $\mathcal{X}, f(x)$, optimality guarantee parameter ϵ, Lipschitz constant L, $\mathcal{SO}$ oracle $\hat{g}$.

Output: An ϵ-solution x^* to problem (5.1).

Set the initial polytope $P \leftarrow [1, N]^d$.
Set the constant $\rho \leftarrow 10^{-7}$. ▷ Constant ρ corresponds to ϵ in [221].
Set the number of iterations $T_{max} \leftarrow \lceil 2d/\rho \cdot \log[dNL/(\rho\epsilon)] \rceil$.
Initialize the set of points used to query separation oracles $\mathcal{S} \leftarrow \emptyset$.
Initialize the volumetric center $z \leftarrow (N+1)/2 \cdot (1, 1, \ldots, 1)^T$.
for $T = 1, 2, \ldots, T_{max}$ **do**
    Decide adding or removing a cutting plane by checking $\sigma_i(z)$ for $i \in [T]$ [221].
    if add a cutting plane **then**
        Evaluate the $\mathcal{SO}$ oracle $\hat{g}$ at z.
        if $\hat{g} = 0$ **then**
            Return z as the approximate solution.
        end if
        Add the current point z to $\mathcal{S}$.
    else if remove a cutting plane **then**
        Remove corresponding point z from $\mathcal{S}$.
    end if
    Update the approximate volumetric center z by a Newton-type method.
end for ▷ There are at most $O(d)$ points in $\mathcal{S}$ by Vaidya's method.
Return the solution $\hat{x}$ to problem $\min_{x \in \mathcal{S}} f(x)$.

Random Walk-based Cutting-plane Method

Next, we list the pseudo-code for the random walk-based cutting-plane method in [20].

Algorithm 15 Deterministic random walk-based cutting-plane method

Input: Model $\mathcal{X}, f(x)$, optimality guarantee parameter ϵ, Lipschitz constant L, $\mathcal{SO}$ oracle $\hat{g}$.

Output: An ϵ-solution x^* to problem (5.1).

Set the initial polytope $P \leftarrow [1, N]^d$.
Set the number of iterations $T_{max} \leftarrow O[d \log(dLN/\epsilon)]$.
Set the number of samples required to calculate the center $M \leftarrow O(d)$.
Initialize the set of points used to query separation oracles $\mathcal{S} \leftarrow \emptyset$.
Initialize the volumetric center $z \leftarrow (N+1)/2 \cdot (1, 1, \ldots, 1)^T$.
for $T = 1, 2, \ldots, T_{max}$ **do**
    Evaluate the $\mathcal{SO}$ oracle $\hat{g}$ at z.
    Add the current point z to $\mathcal{S}$.
    Add the cutting plane using $\hat{g}$ to P.
    if $P = \emptyset$ **then** ▷ This step requires solving a linear feasibility problem
        break
    end if
    Uniformly sample M points from the new polytope P via random walk.
    Update the approximate volumetric center z to the average of all sampled points.
end for
Return the solution $\hat{x}$ to problem $\min_{x \in \mathcal{S}} \ f(x)$.

5.E Dimension Reduction Algorithm with LLL Algorithm

In this section, we provide a more detailed description for the dimension reduction algorithm that utilizes the LLL algorithm. More specifically, the LLL algorithm approximately solves the Shortest Vector Problem (SVP) in lattice to find the hyperplane H in Algorithm 10. Given a lattice Λ and a positive semi-definite matrix Σ that is full-rank on the span of Λ, the SVP problem is given by

$$\arg\min_{v \in \Lambda} \ v^T \Sigma v.$$

In the statement of the algorithm, we define $[x]$ to be the nearest integer to $x \in \mathbb{R}$.

Algorithm 16 Dimension reduction algorithm for the PGS guarantee

Input: Model $\mathcal{X}, (\mathsf{Y}, \mathcal{B}_{\mathsf{Y}}), F(x, \xi_x)$, optimality guarantee parameters ϵ and δ, (ϵ, δ)-$\mathcal{SO}$ oracle $\hat{g}$.

Output: An (ϵ, δ)-PGS solution x^* to problem (5.1).

Set the initial polytope $P \leftarrow [1, N]^d$.

Set the initial subspace $W \leftarrow \mathbb{R}^d$ and the initial lattice $\Lambda \leftarrow \mathbb{Z}^d$.
Initialize the set of points used to query separation oracles $\mathcal{S} \leftarrow \emptyset$.
for $d' = d, d-1, \ldots, 2$ **do** ▷ The current dimension d' is gradually reduced.
    Compute the volumetric center z and covariance matrix Σ by Algorithm 15.
    repeat Get an approximate solution v of SVP with Λ and Σ. ▷ Use the LLL algorithm.
        Take one step of Algorithm 15 with $(\epsilon/4, \delta/4)$-$\mathcal{SO}$ oracle.
        ▷ Algorithm 15 decides a suitable cutting plane H.
        Add the point where the stochastic separation oracle is called to $\mathcal{S}$.
        Shrink the volume of P using the cutting plane H.
        Update the volumetric center z and covariance matrix Σ by Algorithm 15.
    until $v^T \Sigma v \leq (10n)^{-2}$
    Find the vector $\tilde{v} \in \mathbb{Z}^d$ such that v is the orthogonal projection of $\tilde{v}$ on hyperplane $-z + W$.
    ▷ The hyperplane $-z + W$ passes through the origin.
    Construct hyperplane $H \leftarrow \{y \in \mathbb{R}^d \mid \langle y, v \rangle = \langle z, v - \tilde{v} \rangle + [\langle z, \tilde{v} \rangle]\}$.
    Project P, W and Λ onto the hyperplane H. ▷ Reduce the dimension by 1.
    Update the volumetric center z and covariance matrix Σ by Algorithm 15.
end for
Find an $(\epsilon/4, \delta/4)$-PGS solution of the last one-dim problem and add the solution to $\mathcal{S}$.
Find the $(\epsilon/4, \delta/4)$-PGS solution $\hat{x}$ of problem $\min_{x \in \mathcal{S}} f(x)$.
Round $\hat{x}$ to an integral solution by Algorithm 2.

5.F Proofs in Section 5.4

Proof of Lemma 35

Proof of Lemma 35. By the assumption that $F(x, \xi_x) - f(x)$ is sub-Gaussian with parameter σ^2 for any x, we know that $\hat{g}_{\alpha_x(i)} - g_{\alpha_x(i)}$ is the difference of two independent sub-Gaussian random variables and therefore

$$\hat{g}_{\alpha_x(i)} - g_{\alpha_x(i)} \sim \text{subGaussian}\left(2\sigma^2\right), \quad \forall i \in [d],$$

where g is the subgradient of $f(x)$ defined in (4.3). Then, using the properties of sub-Gaussian random variables, it holds that

$$\hat{g}^n_{\alpha_x(i)} - g_{\alpha_x(i)} \sim \text{subGaussian}\left(\frac{2\sigma^2}{n}\right), \quad \forall i \in [d].$$

Recalling that components of $\hat{g}^n$ are mutually independent, we know

$$\langle \hat{g}^n - g, y - x \rangle = \sum_i \left(\hat{g}^n_{\alpha_x(i)} - g_{\alpha_x(i)}\right) \cdot (y - x)_{\alpha_x(i)} \sim \text{subGaussian}\left(\frac{2\sigma^2}{n} \cdot \|y - x\|_2^2\right).$$

Since $\|y-x\|_2^2 \leq dN^2$, we know

$$\langle \hat{g}^n - g, y - x\rangle \sim \text{subGaussian}\left(\frac{2dN^2\sigma^2}{n}\right).$$

By the Hoeffding bound, it holds

$$|\langle \hat{g}^n - g, y - x\rangle| \leq \sqrt{\frac{4dN^2\sigma^2}{n}\log\left(\frac{2}{\delta}\right)}$$

with probability at least $1-\delta$. If we choose

$$n = \left\lceil \frac{4dN^2\sigma^2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)\right\rceil \leq \frac{4dN^2\sigma^2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)+1,$$

it follows that

$$|\langle \hat{g}^n - g, y - x\rangle| \leq \epsilon. \tag{5.28}$$

Since $f(x)$ is a convex function and g is a subgradient at point x, we have $f(y) \geq f(x) + \langle g, y-x\rangle$ for all $y \in [1,N]^d$. Combining with inequality (5.28) gives

$$f(y) \geq f(x) + \langle \hat{g}^n, y-x\rangle + \langle g-\hat{g}^n, y-x\rangle \geq f(x) + \langle \hat{g}^n, y-x\rangle - \epsilon, \quad \forall y \in [1,N]^d$$

holds with probability at least $1-\delta$. Then, considering the half space $H = \{y : \langle \hat{g}^n, y-x\rangle \leq 0\}$, it holds

$$f(y) \geq f(x) + \langle \hat{g}^n, y-x\rangle - \epsilon \geq f(x) - \epsilon, \quad \forall y \in [1,N]^d \cap H^c$$

with the same probability. Taking the minimum over $[1,N]^d \cap H^c$, it follows that the averaged stochastic subgradient provides an (ϵ,δ)-$\mathcal{SO}$ oracle. Finally, the expected simulation cost of each oracle evaluation is at most

$$d\cdot n \leq \frac{4d^2N^2\sigma^2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)+d = \tilde{O}\left[\frac{d^2N^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

□

Proof of Theorem 56

Before we provide the proof of Theorem 56, we show the calculation of the number of iterations T_{max}. With a slight abuse of notations, we use the same notations as [221] only in this calculation. Before the first iteration, we have the volumetric center as

$$\omega = \frac{N+1}{2}\cdot(1,\ldots,1)^T \in \mathbb{R}^d.$$

Therefore, we can calculate that

$$H(\omega) = \frac{8}{(N-1)^2} \cdot I_d, \quad \rho^0 = \frac{d}{2} \log\left(\frac{8}{(N-1)^2}\right),$$

where I_d is the $d \times d$ identity matrix. By [221], the volume of the polytope at the beginning of the t-th iteration satisfies

$$\begin{aligned} \log(\pi^t) &\leq d \log\left(\frac{2d}{\rho}\right) - \rho^0 - \frac{\rho}{2} \cdot t = d \log\left(\frac{2d}{\rho}\right) - \frac{d}{2} \log\left(\frac{8}{(N-1)^2}\right) - \frac{\rho}{2} \cdot t \\ &\leq d \log\left(\frac{2d}{\rho}\right) + d \log\left(\frac{N}{2}\right) - \frac{\rho}{2} \cdot t = d \log\left(\frac{Nd}{\rho}\right) - \frac{\rho}{2} \cdot t. \end{aligned} \tag{5.29}$$

The target set consists of points in the set

$$P(\epsilon) := \left\{ x \in [1, N]^d \ : \ \|x - x^*\|_1 \leq \epsilon / L \right\},$$

where x^* is the optimal solution of problem (5.1) and L is the Lipschitz constant of $f(\cdot)$. By a simple analysis, we know that the volume of $P(\epsilon)$ satisfies

$$\mathrm{vol}(P(\epsilon)) \geq \left(\frac{\epsilon}{L}\right)^d.$$

Therefore, we can terminate the algorithm when

$$\log(\pi^{T_{max}}) \leq d \log\left(\frac{\epsilon}{L}\right).$$

Combining with inequality (5.29), we know

$$T_{max} \geq \frac{2d}{\rho} \cdot \log\left(\frac{NdL}{\rho\epsilon}\right)$$

is sufficient for ϵ-approximate solutions.

Proof of Theorem 56. We first prove the correctness of Algorithm 9. If $\hat{g} = 0$ for some iteration, the half space $H = \mathbb{R}^d$ and the definition of $(\epsilon/8, \delta/4)$-$\mathcal{SO}$ implies that

$$f(y) \geq f(z) - \epsilon/8, \quad \forall y \in [1, N]^d$$

holds with probability at least $1 - \delta/4$, where z is the point that the separation oracle is called. Hence, we know z is an $(\epsilon/8, \delta/4)$-PGS solution and obviously satisfies the $(\epsilon/2, \delta/2)$-PGS guarantee. Then, by Theorem 41, the integral solution after the round process is an (ϵ, δ)-PGS solution.

In the following of the proof, we assume $\hat{g} \neq 0$ for all iterations. Let $x^* \in \mathcal{X}$ be a minimizer of problem (5.1). We consider the set

$$Q := \left(x^* + \left[-\frac{\epsilon}{8L}, \frac{\epsilon}{8L}\right]^d\right) \cap [1, N]^d.$$

We can verify that set Q is not empty and has volume at least $(\epsilon/(8L))^d$. Moreover, for any $x \in Q$, it holds

$$f(x) \leq f(x^*) + L\|x - x^*\|_\infty \leq f(x^*) + \frac{\epsilon}{8}.$$

By the analysis in [221], the volume of the polytope P is smaller than $(\epsilon/(8L))^d$ after

$$T_{max} := O\left[d \log\left(\frac{8dLN}{\epsilon}\right)\right]$$

iterations. Hence, after T_{max} iterations, the volume of P is smaller than the volume of Q and it must hold $Q \backslash P \neq \emptyset$. Since $Q \subset [1, N]^d$, the constraint $1 \leq x_i \leq N$ is not violated for all $i \in [d]$. Thus, if we choose $x \in Q \backslash P$, there exists a cutting plane $-\hat{g}^T y \geq \beta$ in P such that

$$-\hat{g}^T x < \beta \leq -\hat{g}^T z,$$

where z is the point that the $(\epsilon/8, \delta/4)$-$\mathcal{SO}$ oracle $\hat{g}$ is evaluated and β is the value chosen by Vaidya's method. This implies that x is not in the half space

$$H := \{y : \hat{g}^T y \leq \hat{g}^T z\}.$$

Then, by the definition of $(\epsilon/8, \delta/4)$-$\mathcal{SO}$ oracle and the claim that $x \in [1, N]^d \cap H^c$, we know

$$f(x) \geq f(z) - \epsilon/8$$

holds with probability at least $1 - \delta/4$. On the other hand, the condition $x \in P$ leads to

$$f(x) \leq f(x^*) + \epsilon/8.$$

Combining the last two inequalities gives that

$$\min_{y \in \mathcal{S}} \ f(y) \leq f(z) \leq f(x^*) + \epsilon/4$$

holds with probability at least $1 - \delta/4$. Hence, the $(\epsilon/4, \delta/4)$-PGS solution $\hat{x}$ of problem $\min_{y \in \mathcal{S}} \ f(y)$ satisfies

$$f(\hat{x}) \leq f(x^*) + \epsilon/2$$

with probability at least $1 - \delta/2$. Equivalently, the solution $\hat{x}$ is an $(\epsilon/2, \delta/2)$-PGS solution. Using Theorem 41, the integral solution returned by Algorithm 9 is an (ϵ, δ)-PGS solution.

Now, we estimate the expected simulation cost of Algorithm 9. By Lemma 35, the simulation cost of each $(\epsilon/8, \delta/4)$-$\mathcal{SO}$ oracle is at most

$$O\left[\frac{d^2 N^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right) + d\right].$$

Since at most one separation oracle is evaluated in each iteration, the total simulation cost of T_{max} iterations is at most

$$O\left[\left(\frac{d^2 N^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right) + d\right) \cdot d \log\left(\frac{8dLN}{\epsilon}\right)\right] = \tilde{O}\left[\frac{d^3 N^2}{\epsilon^2} \log(\frac{dLN}{\epsilon}) \log\left(\frac{1}{\delta}\right)\right].$$

By the property of Vaidya's method, there are $O(d)$ cutting planes in the polytope P. Then, using the same analysis as in Chapter 4, the expected simulation cost of finding an $(\epsilon/4, \delta/4)$-PGS solution of the sub-problem $\min_{y\in\mathcal{S}} f(y)$ is at most

$$\tilde{O}\left[\frac{d^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

We note that the evaluation of $(\epsilon/8, \delta/4)$-$\mathcal{SO}$ oracles at points in $\mathcal{S}$ provides enough simulations for the sub-problem and therefore the simulation cost of this part can be avoided. Finally, the expected simulation cost of the rounding process is bounded by

$$\tilde{O}\left[\frac{d}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

Combining the three parts, the total expected simulation cost of Algorithm 9 is at most

$$\tilde{O}\left[\frac{d^3N^2}{\epsilon^2}\log(\frac{dLN}{\epsilon})\log\left(\frac{1}{\delta}\right)\right].$$

□

Proof of Theorem 57

Proof of Theorem 57. We first verify the correctness of Algorithm 10. If the optimal solution has been removed during the dimension reduction process, we claim that the optimal solutions are removed from the search set by some cutting plane. This is because the dimension reduction steps will not remove integral points from the current search set [122]. Then, by the same proof as Theorem 56, it holds

$$\min_{x\in\mathcal{S}} f(x) \le \min_{x\in\mathcal{X}} f(x) + \epsilon/4 \tag{5.30}$$

with probability at least $1-\delta/4$. Otherwise if the optimal solution has not been removed from the search set throughout the dimension reduction process, we know the last one-dimensional problem contains the optimal solution. Hence, the $(\epsilon/4, \delta/4)$-PGS solution to the one-dimensional problem is also an $(\epsilon/4, \delta/4)$-PGS solution to the original problem. Since the PGS solution is also added to the set $\mathcal{S}$, we also have relation (5.30) holds with probability at least $1-\delta/4$. Then, the $(\epsilon/4, \delta/4)$-PGS solution $\bar{x}$ to problem $\min_{x\in\mathcal{S}} f(x)$ satisfies

$$f(\bar{x}) \le \min_{x\in\mathcal{X}} f(x) + \epsilon/2$$

with probability at least $1-\delta/2$, or equivalently $\bar{x}$ is an $(\epsilon/2, \delta/2)$-PGS solution to problem (5.1). Using the results of Theorem 41, the solution returned by Algorithm 10 is an (ϵ, δ)-PGS solution.

Next, we estimate the expected simulation cost of Algorithm 10. By the results in [122], $(\epsilon/4, \delta/4)$-$\mathcal{SO}$ oracles are called at most $O[d(d + \log(N))]$ times. Hence, the size of $\mathcal{S}$ is at most $O[d(d + \log(N))]$. By the estimates in Lemma 35, the total simulation cost of the dimension reduction process is at most

$$O\left[\frac{d^3N^2(d + \log(N))}{\epsilon^2}\log\left(\frac{1}{\delta}\right) + d^2(d + \log(N))\right] = \tilde{O}\left[\frac{d^3N^2(d + \log(N))}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

Moreover, the one-dimensional convex problem has at most N feasible points and Theorem 55 implies that the expected simulation cost for this problem is at most

$$\tilde{O}\left[\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

Since the size of $\mathcal{S}$ is at most $O[d(d + \log(N))]$, the sub-problem for the set $\mathcal{S}$ takes at most

$$O\left[\frac{d^2(d + \log(N))}{\epsilon^2}\log\left(\frac{1}{\delta}\right) + d^2(d + \log(N))\right] = \tilde{O}\left[\frac{d^2(d + \log(N))}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right]$$

simulation runs. Finally, Theorem 41 shows that the expected simulation cost of the rounding process is at most

$$\tilde{O}\left[\frac{d}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

In summary, the total expected simulation cost of Algorithm 10 is at most

$$\tilde{O}\left[\frac{d^3N^2(d + \log(N))}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right].$$

□

5.G Adaptive Sub-Gaussian Parameter Estimator

In this section, we provide a simple adaptive mean estimator to adaptively estimate the variance of each choice of decision variable under the assumption that the distribution of the randomness is Gaussian. The estimator can be used to further enhance our proposed algorithm and we hope the procedure to be useful for other optimization via simulation problems and algorithms that do not know the variances a priori. Using the adaptive estimator, the prior knowledge about the upper bound on the variance σ^2 is not necessary. In addition, for the multi-dimensional localization algorithms proposed in this chapter, the simulation cost for the unknown variance case is at most a constant factor larger than the case when an upper bound on the variance is known a priori. Therefore, the algorithm using the adaptive estimator, or the adaptive algorithm, is able to improve the performance of our proposed algorithms if an estimate of the upper bound σ^2 is much larger than the true variance. In

this case, the original algorithms will implement an unnecessarily large number of simulation runs to shrink the confidence interval, while the adaptive algorithm is able to automatically learn the true variance and thus save the computational cost. Another situation where the adaptive algorithm is useful is when the variance of the system varies a lot at different choices of decision variable. In this case, the upper bound of the variance is usually attained at extreme choices of decision variable and is much larger than the variance of a majority of feasible choices. For example, in queueing systems, the distribution of arrival times usually follows the Poisson distribution (or the generalizations of Poisson distributions, such as the jump distribution). The mean and the variance of the Poisson distribution are equal and, thus, the variance is large at solutions where the mean is large. The queueing system contains the Poisson process as a part and will also exhibit certain heteroscedasticity in the variance. Using the upper bound at all points leads to a conservative mean estimator; see our experiment results in Table 5.G.1.

We now state the proposed adaptive mean estimator. To increase the generality of our results, we make a weaker assumption than the Gaussian case.

Assumption 12. The distribution of $F(x, \xi_x) - f(x)$ belongs to the family of sub-Gaussian distributions $\mathcal{F}_\kappa$, where $\kappa > 0$ is a known constant. For any random variable X whose distribution belongs to $\mathcal{F}_\kappa$, it holds that

$$\kappa\sigma_X^2 \leq \mathrm{Var}(X), \tag{5.31}$$

where σ_X^2 is the sub-Gaussian parameter of the distribution.

We note that the inverse inequality $\mathrm{Var}(F(x, \xi_x)) \leq \sigma_x^2$ always holds for all sub-Gaussian distributions. However, there does not exist a universal constant $\kappa > 0$ such that inequality (5.31) holds for all sub-Gaussian distributions. Therefore, Assumption 12 cannot be implied by Assumption 5 and additional prior knowledge about the distribution is required for the estimation of κ. We provide three special cases when the value of κ can be estimated:

1. Suppose that the distribution of $F(x, \xi_x)$ is Gaussian. In this case, the constant $\kappa = 1$, i.e., we have the relation $\sigma_x^2 = \mathrm{Var}[F(x, \xi_x)]$.

2. Suppose that the distribution of $F(x, \xi_x)$ is the uniform distribution in the interval $[f(x) - a_x, f(x) + a_x]$, where the values of $f(x)$ and a_x are unknown. In this case, the variance and the sub-Gaussian parameter of $F(x, \xi_x) - f(x)$ are $a_x^2/3$ and a_x, respectively. Therefore, the parameter κ is equal to $1/3$ in this case.

3. Suppose that the distribution of $F(x, \xi_x)$ is the Bernoulli distribution with the parameters (n_x, p_x), where the value of p_x is unknown. In this case, the variance and the sub-Gaussian parameter of $F(x, \xi_x) - f(x)$ are $p_x(1 - p_x)$ and $\frac{1-2p_x}{2\log[(1-p_x)/p_x]}$ [177], respectively. Therefore, we can show that either the sub-Gaussian parameter is at most $1/2$ or the parameter κ is $(e - 1)/(2e^2 - 4e)$. The analysis of the Bernoulli case can be directly extended to a binomial distribution with the parameters (n_x, p_x), where

an upper bound on the parameter n_x is known (for example, we know a priori that $F(x, \xi_x)$ belongs to $\{0, 1, \ldots, M\}$ for some integer $M > 0$).

Under the above assumption, we propose the adaptive mean estimator.

Definition 17. Let $\epsilon > 0$ be the precision and $\delta \in (0, 1]$ be the failing probability. We construct the adaptive mean estimator of $f(x)$ in two steps:

1. Sample $2n$ independent evaluations $F(x, \xi_i)$ for $i \in [2n]$, where $n := \lceil 256\kappa^{-2} \log(2/\delta) \rceil$. Compute the variance estimator

$$\hat{\mathrm{Var}}_x := \frac{1}{n} \sum_{i=1}^{n} [F(x, \xi_{2i-1}) - F(x, \xi_{2i})]^2$$

and the parameter estimator

$$\hat{\sigma}_x^2 := \frac{1}{\kappa} \hat{\mathrm{Var}}_x.$$

2. Let $m := \max\{\lceil 2\epsilon^{-2} \hat{\sigma}_x^2 \log(2/\delta) \rceil, 2n\}$ and sample $m - 2n$ independent evaluations $F(x, \xi_{2n+i})$ for $i \in [m - 2n]$ and compute the empirical mean

$$\hat{F}(x; \delta) := \frac{1}{m} \sum_{i=1}^{m} F(x, \xi_i).$$

The construction of the adaptive mean estimator has two steps. In the first step, we estimate an upper bound for the sub-Gaussian parameter, and in the second step, we use the estimated upper bound to calculate the required number of simulation so that the sub-Gaussian parameter is less than a known constant. The purpose of the choice of $\hat{\mathrm{Var}}$ is to utilize the Bernstein bound on the lower tail of sum of squared sub-Gaussian random variables, i.e., $\mathbb{P}\left(\sum_{i=1}^{n} Z_i^2 \leq a\sigma^2\right)$, where $Z_1, \ldots, Z_n$ are independent sub-Gaussian random variables with parameter σ^2 and $a > 0$ is a constant. If we use the common estimator of variance $n^{-1} \sum_i (Z_i - \bar{Z})^2$, where $\bar{Z}$ is the empirical mean, the random variables $Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z}$ are not necessarily independent and the Bernstein bound cannot be applied. We note that the adaptive mean estimator is an online estimator. To be more concrete, if a smaller precision $\epsilon' < \epsilon$ is required, it suffices to add

$$\lceil 2(\epsilon')^{-2} \hat{\sigma}_x^2 \log(2/\delta) \rceil - \lceil 2\epsilon^{-2} \hat{\sigma}_x^2 \log(2/\delta) \rceil$$

more evaluations into the empirical mean in step 2. The following theorem verifies that $\hat{F}(\cdot; \delta)$ is an unbiased mean estimator for $f(x)$ and its tail is sub-Gaussian with a small failing probability.

Theorem 64. *Suppose that Assumption 12 holds. Let $\delta \in (0, 1]$ be the failing probability. For all $\epsilon \geq 0$, the adaptive mean estimator satisfies*

$$\mathbb{P}\left[|\hat{F}(x;\delta) - f(x)| \geq \epsilon\right] \leq \delta, \tag{5.32}$$

In addition, the expected simulation cost of the adaptive mean estimator is

$$O[(\kappa^{-2} + \kappa^{-1}\epsilon^{-2}\sigma_x^2)\log(1/\delta)].$$

Remark 6. We note that the estimator $\hat{\sigma}_x^2$ in Definition 17 is the sum of squared sub-Gaussian random variables and may have a heavy tail. Thus, the simulation cost of $\hat{F}(x;\delta)$ may also have a heavy tail. To deal with this issue, we can utilize the Median-of-Mean (MoM) estimator [142] of the random variable

$$G_x := \kappa^{-1}[F(x,\xi_1) - F(x,\xi_2)]^2,$$

where ξ_1 and ξ_2 are independent. Choosing $b := \lceil 2^{15}\kappa^{-2}\rceil$ and $K := \lceil 2\log(2/\delta)\rceil$ in the MoM estimator, we define the alternative estimator $(\hat{\sigma}_x^{MoM})^2$ by

$$(\hat{\sigma}_x^{MoM})^2 := \kappa^{-1}\text{median}\left\{\frac{1}{b}\sum_{j=1}^{b} G_{x,ib+j},\ i \in [K]\right\}$$

where $G_{x,1}, \ldots, G_{x,Kb}$ are independent samples of G_x. The estimator $(\hat{\sigma}_x^{MoM})^2$ also has simulation cost $O[\kappa^{-2}\log(1/\delta)]$. Using Proposition 12 in [142], we know that the estimator $(\hat{\sigma}_x^{MoM})^2$ satisfies

$$\mathbb{P}\left[\sigma_x^2 \leq (\hat{\sigma}_x^{MoM})^2 \leq (2+\kappa^{-1})\sigma_x^2\right] \geq 1 - \delta/2.$$

Using this MoM estimator, we are able to bound the simulation cost of $\hat{F}(x;\delta)$ in high probability.

If the sub-Gaussian parameter σ_x^2 is known, the Hoeffding bound shows that

$$O[\epsilon^{-2}\sigma_x^2\log(1/\delta)]$$

samples are sufficient to generate an estimator for inequality (5.32). Therefore, the relative efficiency of the adaptive mean estimator is

$$\frac{\epsilon^{-2}\sigma_x^2}{\kappa^{-2} + \epsilon^{-2}\kappa^{-1}\sigma_x^2} = \frac{1}{\kappa^{-1} + \kappa^{-2}\epsilon^2\sigma_x^{-2}}.$$

If the precision ϵ is small or the parameter σ_x^2 is large, the adaptive mean estimator is only a constant (κ) time less efficient than the known variance case.

Now, we estimate the expected simulation cost of our proposed simulation-optimization algorithms combined with the adaptive estimator. Intuitively, we need to implement the first step in Definition 17 once for all simulated choices of decision variable. Suppose that a

simulation-optimization algorithm simulates $N(\epsilon, \delta)$ different choices of decision variable in expectation and the expected simulation cost is $T(\epsilon, \delta)$. Then, the expected simulation cost of the adaptive simulation-optimization algorithm is

$$O\left[\kappa^{-1} T(\epsilon, \delta) + \kappa^{-2} N(\epsilon, \delta) \log\left(N(\epsilon, \delta)/\delta\right)\right].$$

For the localization algorithms, we usually have $T(\epsilon, \delta) = O[N(\epsilon, \delta) \log\left(N(\epsilon, \delta)/\delta\right)]$. Therefore, the expected simulation cost of the localization algorithms is $O[T(\epsilon, \delta)]$. More concretely, we have the following corollary.

Corollary 6. *Suppose that Assumptions 6-12 hold. The following estimates hold:*

- *The expected simulation cost of the adaptive TS algorithm (Algorithm 7) is*

$$O\left[(\kappa^{-2} + \kappa^{-1}\sigma^2\epsilon^{-2}) \log(N) \log\left(\frac{\log(N)}{\delta}\right)\right] = \tilde{O}\left[(\kappa^{-2} + \kappa^{-1}\sigma^2\epsilon^{-2}) \log(N) \log\left(\frac{1}{\delta}\right)\right].$$

- *The expected simulation cost of the adaptive SUS algorithm (Algorithm 8) is*

$$O\left[(\kappa^{-2} N + \kappa^{-1}\sigma^2\epsilon^{-2}) \log\left(\frac{N}{\delta}\right)\right] = \tilde{O}\left[(\kappa^{-2} N + \kappa^{-1}\sigma^2\epsilon^{-2}) \log\left(\frac{1}{\delta}\right)\right].$$

- *The expected simulation cost of the adaptive stochastic cutting-plane algorithm (Algorithm 9) is*

$$\begin{aligned} &O\Big[\left(\kappa^{-2} + \frac{\kappa^{-1}\sigma^2 dN^2}{\epsilon^2}\right) \cdot d^2 \log\left(\frac{dLN}{\epsilon}\right) \log\left(\frac{1}{\delta}\right) \\ &\qquad + \kappa^{-2} d^2 \log\left(\frac{dLN}{\epsilon}\right) \log\left(d^2 \log\left(\frac{dLN}{\epsilon}\right)\right)\Big] \\ =&\tilde{O}\left[\left(\kappa^{-2} + \frac{\kappa^{-1}\sigma^2 dN^2}{\epsilon^2}\right) \cdot d^2 \log\left(\frac{dLN}{\epsilon}\right) \log\left(\frac{1}{\delta}\right)\right]. \end{aligned}$$

- *The expected simulation cost of the adaptive dimension reduction algorithm (Algorithm 10) is*

$$\begin{aligned} &O\Big[\left(\kappa^{-2} + \frac{\kappa^{-1}\sigma^2 dN^2}{\epsilon^2}\right) d^2(d + \log(N)) \log\left(\frac{1}{\delta}\right) \\ &\qquad + \kappa^{-2} d^2(d + \log(N)) \log\left(d^2(d + \log(N))\right)\Big] \\ =&\tilde{O}\left[\left(\kappa^{-2} + \frac{\kappa^{-1}\sigma^2 dN^2}{\epsilon^2}\right) d^2(d + \log(N)) \log\left(\frac{1}{\delta}\right)\right]. \end{aligned}$$

We can see that the expected simulation cost of the stochastic cutting-plane method and the dimension reduction algorithm is only increased by a factor. Therefore, the adaptive mean estimator is useful in dropping the requirement of known parameter σ for the multi-dimensional case. For the one-dimensional case, the expected simulation cost of the trisection algorithm is also increased by a constant factor. On the other hand, the cost of the adaptive SUS algorithm is larger than the original version, especially when $\epsilon^{-2} \ll N$. Therefore, in the one-dimensional unknown variance case, we need to consider the size of ϵ to decide whether to use the TS algorithm or the SUS algorithm, More specifically, if $\epsilon = O(\sqrt{\log N/N})$, then the SUS algorithm is preferred; otherwise the TS algorithm is preferred.

Now, we briefly discuss how to apply the dimension reduction algorithm and the adaptive sub-Gaussian parameter estimator to the two examples in Section 5.5, and we present the effects of the adaptive sub-Gaussian parameter estimator on the optimal allocation problem. We first consider the estimation of the parameter κ. For the synthetic example, the noise is Gaussian and, thus, we have $\kappa = 1$. For the queueing example, the dimension reduction algorithm will simulate $\Theta[\hat{\sigma}_x^2 dN^2/\epsilon^2 \log(1/\delta)]$ independent samples at each iteration point x (more rigorously, the neighbouring points of point x) to estimate the expected value $f(x)$, where $\hat{\sigma}_x^2$ is chosen to be $30\sqrt{N} \gg 1$ in the numerical experiments. Define $n := dN^2/\epsilon^2 \log(1/\delta)$ and

$$G(x, \bar{\xi}_x) := \frac{1}{n} \sum_{i=1}^{n} F(x, \xi_{x,i}),$$

where $\xi_{x,1}, \ldots, \xi_{x,n}$ are independently sampled and $\bar{\xi}_x := (\xi_{x,1}, \ldots, \xi_{x,n})$. Then, the sub-Gaussian parameter of $F(x, \xi_x) - f(x)$ should be at most n times larger than the sub-Gaussian parameter of $G(x, \bar{\xi}_x) - f(x)$; the same relation also holds for the variances. Hence, we can instead estimate the constant κ of random variable $G(x, \bar{\xi}_x)$. Using the Central Limit Theorem (CLT), the asymptotic behavior of the empirical mean of $F(x, \xi_x)$ is Gaussian. Indeed, with our choice of n, the distribution of $G(x, \bar{\xi}_x)$ is already very close to the Gaussian distribution. To verify this claim, we plot the Quantile-Quantile plot for 200 independent samples of $G(x_0, \bar{\xi}_{x_0})$ and quantiles of the Gaussian distribution, where $x_0 = (\frac{N+1}{2}, N + 1, \ldots, \frac{d(N+1)}{2})$. The results of $(d, N) = (4, 10), (4, 50), (24, 10)$ are plotted in Figure 5.G.1 and we can see that the distributions in all cases are very close to the Gaussian distribution. Therefore, the parameter κ is approximately equal to 1 for $G(x, \bar{\xi}_x)$. Hence, we use the first $50n \leq 30\sqrt{N}n$ samples of $F(x, \xi_x)$ to generate 25 pairs of evaluations of $G(x, \bar{\xi}_x)$. Here, the sample size 25 is derived from the estimate that $\kappa^{-2} \log[d^2(d + \log N)/\delta] = \log[d^2(d + \log N)/\delta] \leq 25$. The sub-Gaussian parameter of $F(x, \xi_x) - f(x)$ is then estimated by

$$\hat{\sigma}_x^2 := \frac{n}{25} \sum_{i=1}^{25} \left[G(x, \bar{\xi}_{x,2i-1}) - G(x, \bar{\xi}_{x,2i}) \right]^2.$$

Finally, since the failing probability in Theorem 64 is bounded by the union bound, we only need to take $\max\{\hat{\sigma}_x^2 - 50, 0\}n$ extra samples of $F(x, \xi_x)$ to generate the mean estimator with the desired confidence level.

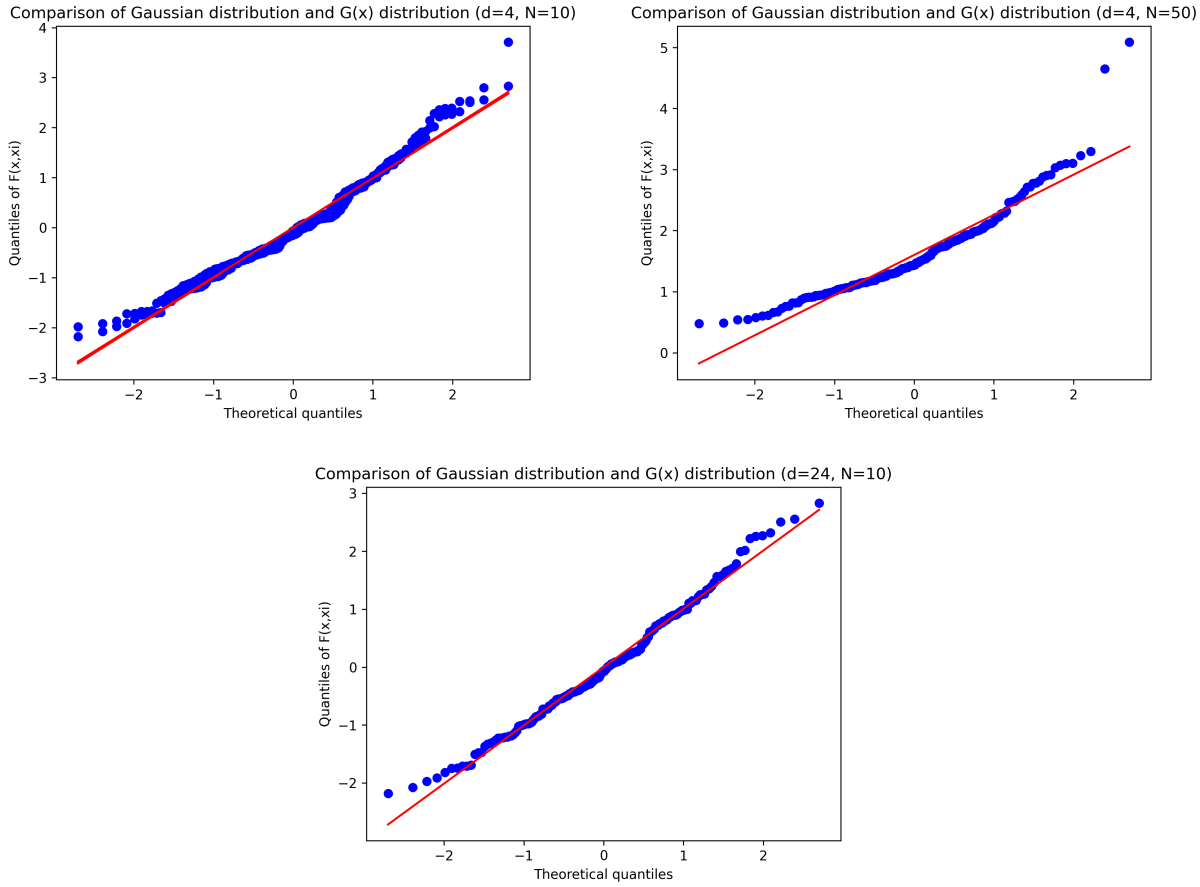


Figure 5.G.1: The Quantile-Quantile plots of the distribution of $G(x, \bar{\xi}_x)$ in the cases when $(d, N) = (4, 10), (4, 50), (24, 10)$.

We test the dimension reduction algorithm with the adaptive sub-Gaussian parameter estimator on the optimal allocation example under the same setting as in Section 5.5 and the results are summarized in Table 5.G.1. From the numerical results, we can see that the adaptive sub-Gaussian parameter estimator is able to reduce the expected simulation cost in most cases. This is because the maximum variance is usually attained by decisions around the global minimum. Thus, during the early stage of the optimization process, the true sub-Gaussian parameter is relatively small and is overestimated by our estimation $\hat{\sigma}^2 = 30\sqrt{N}$. Using the adaptive estimator, we are able to estimate the sub-Gaussian parameter and reduce the expected simulation cost.

Params.		No adaptive estimator		With adaptive estimator	
d	N	Cost	Obj.	Cost	Obj.
4	10	2.42e4	2.40e1	**1.22e4**	2.42e1
4	20	1.40e4	3.44e1	**1.06e4**	3.42e1
4	30	9.21e3	4.59e1	**8.21e3**	4.15e1
4	40	6.31e3	5.75e1	**4.31e3**	5.75e1
4	50	**4.03e3**	6.67e1	6.03e3	6.70e1
8	10	1.48e5	2.12e1	**1.55e4**	2.21e1
12	10	6.10e5	2.01e1	**4.72e4**	2.11e1
16	10	1.59e6	1.91e1	**3.65e5**	1.92e1
20	10	3.21e6	1.81e1	**7.83e5**	1.88e1
24	10	8.54e6	1.76e1	**1.81e6**	1.81e1

Table 5.G.1: Simulation cost of the dimension reduction algorithm with and without the adaptive variance estimator on the resource allocation problem.

Proof of Theorem 64

We first prove that $\hat{\sigma}$ serves as an upper bound on the sub-Gaussian parameter. The proof is based on the property that the lower tail of a squared sub-Gaussian random variable is sub-Gaussian.

Lemma 46. *Let $\delta \in (0, 1]$ be the failing probability. The parameter estimator $\hat{\sigma}^2$ in Definition 17 satisfies*

$$\mathbb{P}\left(\hat{\sigma}^2 \leq \sigma_x^2\right) \leq \delta/2.$$

Proof of Lemma 46. By the definition of $\hat{\text{Var}}$ and $\hat{\sigma}^2$, we only need to prove that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left[F(x,\xi_{2i-1}) - F(x,\xi_{2i})\right]^2 \leq \kappa\sigma_x^2\right) \leq \delta/2. \tag{5.33}$$

By the independence between ξ_{2i-1} and ξ_{2i}, the random variable $F_i := F(x,\xi_{2i-1}) - F(x,\xi_{2i})$ is zero-mean and sub-Gaussian with parameter $2\sigma_x^2$ for all $i \in [n]$. Using the fact that $-F_i^2 \leq 0$ almost surely and the one-sided Bernstein's inequality, we have

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\left(-F_i^2 + \mathbb{E}(F_i^2)\right) \geq \kappa\sigma_x^2\right] \leq \exp\left[-\frac{n\kappa^2\sigma_x^4}{2/n\sum_{i=1}^{n}\mathbb{E}(F_i^4)}\right].$$

Since $\{F_i, i \in [n]\}$ are i.i.d. and zero-mean random variables, the above inequality is equivalent to

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} F_i^2 - \mathrm{Var}(F_1) \le -\kappa\sigma_x^2\right] \le \exp\left[-\frac{n\kappa^2\sigma_x^4}{2\mathbb{E}(F_1^4)}\right].$$

Now, recalling the assumption in (5.31) and $\mathrm{Var}(F_1) = 2\mathrm{Var}(F(x,\xi_1))$, we get

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} F_i^2 - 2\kappa\sigma_x^2 \le -\kappa\sigma_x^2\right] \le \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} F_i^2 - \mathrm{Var}(F_1) \le -\kappa\sigma_x^2\right] \le \exp\left[-\frac{n\kappa^2\sigma_x^4}{2\mathbb{E}(F_1^4)}\right]. \tag{5.34}$$

To estimate the fourth moment of F_1, we calculate that

$$\begin{aligned}\mathbb{E}(F_1^4) &= \int_0^\infty t\mathbb{P}(F_1^4 \ge t)\, dt = \int_0^\infty 4s^3\mathbb{P}(F_1^4 \ge s^4)\, ds = \int_0^\infty 4s^3\mathbb{P}(|F_1| \ge s)\, ds \\ &\le \int_0^\infty 4s^3 \cdot 2\exp[-s^2/(8\sigma_x^2)]\, ds = 128\sigma_x^4,\end{aligned}$$

where the second equality is from the substitution $t = s^4$ and the last inequality is from the fact that F_1 is $2\sigma_x^2$-sub-Gaussian. Substituting into inequality (5.34), we get

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} F_i^2 \le \kappa\sigma_x^2\right] \le \exp\left(-\frac{n\kappa^2\sigma_x^4}{256\sigma_x^4}\right) = \exp\left(-\frac{n\kappa^2}{256}\right) \le \frac{\delta}{2},$$

where the last inequality is from the choice of n. The above inequality is equivalent to inequality (5.33) and the proof is done. □

With the help of Lemma 46, we now prove the theorem.

Proof of Theorem 64. By Lemma 46, we know that the event $\mathcal{E} := \{\sigma_x^2 \le \hat{\sigma}^2\}$ happens with probability at least $1 - \delta/2$. By the Hoeffding's inequality and the definitions of n and m, we have

$$\begin{aligned}\mathbb{P}\left[|\hat{F}(x;\delta) - f(x)| \ge \epsilon, \mathcal{E} \mid \hat{\sigma}^2\right] &\le \exp\left[-\frac{m\epsilon^2}{2\sigma_x^2}\right] \le \exp\left[-\frac{2\epsilon^{-2}\hat{\sigma}^2 \cdot \log(2/\delta)\epsilon^2}{2\sigma_x^2}\right] \\ &\le \exp\left[-\log(2/\delta)\epsilon^2\right] = \frac{\delta}{2}.\end{aligned}$$

Taking expectation over $\hat{\sigma}^2$ leads to

$$\mathbb{P}\left[|\hat{F}(x;\delta) - f(x)| \ge \epsilon, \mathcal{E}\right] \le \frac{\delta}{2}.$$

Therefore, we get

$$\mathbb{P}\left[|\hat{F}(x;\delta) - f(x)| \ge \epsilon\right] = \mathbb{P}\left[|\hat{F}(x;\delta) - f(x)| \ge \epsilon, \mathcal{E}\right] + \mathbb{P}\left[|\hat{F}(x;\delta) - f(x)| \ge \epsilon, \mathcal{E}^c\right]$$

$$\leq \mathbb{P}\left[|\hat{F}(x;\delta) - f(x)| \geq \epsilon, \mathcal{E}\right] + \mathbb{P}\left[\mathcal{E}^c\right] \leq \delta,$$

where $\mathcal{E}^c$ is the complementary set of $\mathcal{E}$.

The estimation of the expected simulation cost is from the fact that $\hat{\text{Var}}$ is an unbiased estimator of $\text{Var}[f(x,\xi_x)]$. Therefore, we get the bound

$$\mathbb{E}[m] \leq \max\{\lceil 2\epsilon^{-2}\mathbb{E}(\hat{\sigma}^2)\log(2/\delta)\rceil, \lceil 512\kappa^{-2}\log(2/\delta)\rceil\} \leq \lceil (512\kappa^{-2} + 2\kappa^{-1}\epsilon^{-2}\sigma_x^2)\log(2/\delta)\rceil.$$

□

5.H Additional Numerical Experiments

Comparison to Industrial Strength COMPASS

In this subsection, we compare the dimension reduction algorithm (Algorithm 10), the subgradient descent algorithm (Algorithm 3), and the Industrial Strength COMPASS (ISC) algorithm [236] on multi-dimensional optimization via simulation problems. For the dimension reduction algorithm and the subgradient descent algorithm, we use the results in Section 5.5. For the ISC algorithm, our experiments are based on the source codes provided by the authors of [236]. We implemented the ISC codes on a C++17 compiler and used the *LpSolve* package with version 5.5.0.3. Since the ISC algorithm does not support the PGS criterion, we empirically set the value of the CLEANUP_DELTA parameter so that the ISC algorithm finds a comparable solution to other algorithms. To be more concrete, we choose CLEANUP_DELTA to be $d/10$ in the separable convex function minimization problem and $N/100$ in the optimal allocation problem. Other parameters are set to their default values. We test the performance of the ISC algorithm on 100 independent experiments for the separable convex function minimization and 5 independent experiments for the optimal allocation problem.

The results are summarized in Tables 5.H.1 and 5.H.2. For the separable convex function minimization problem, the coverage rate in all settings is equal to 100%, and the proposed dimension reduction methods is always better than the ISC algorithm. For the optimal allocation problem, we can see that the proposed dimension reduction methods is better when the scale is large (e.g., when $N \geq 20$) or when the dimension is large (e.g., when $d \geq 20$). This is consistent with the main contribution of this chapter, i.e., efficient simulation-optimization algorithms for large-scale problems. However, the ISC algorithm achieves better simulation costs for high-dimensional problems with a relatively small scale or dimension. There are two possible reasons for this phenomenon.

First, the ISC algorithm only finds "locally optimal solutions". Specifically, in [236], a solution $x \in \mathcal{X}$ is called a locally optimal solution if $f(y) \geq f(x)$ for all $y \in \mathcal{X}$ such that $\|y-x\|_1 \leq 1$. For $L^\natural$-convex functions, a locally optimal solution is not necessarily a globally optimal solution. Instead, a solution $x \in \mathcal{X}$ is a global optimum of a $L^\natural$-convex function if and only if $f(y) \geq f(x)$ for all $y \in \mathcal{X}$ such that $\|y - x\|_\infty \leq 1$. Therefore, to check the

Params.		SubGD	Dim Reduction	ISC	R-SPLINE
d	N	Cost	Cost	Cost	Cost
2	50	1.08e3	1.56e2	8.73e2	**4.50e1**
2	500	2.54e4	2.08e2	9.00e2	**9.00e1**
2	5000	3.97e5	4.66e2	1.05e3	**1.84e2**
6	50	5.00e3	4.05e2	4.06e3	**1.60e2**
6	500	4.75e4	6.45e2	8.86e3	**3.20e2**
6	5000	2.72e6	8.25e2	1.20e4	**8.25e2**
10	50	8.46e3	8.34e2	3.69e4	**3.32e2**
10	500	6.32e4	1.48e3	6.71e4	**8.88e2**
10	5000	7.76e6	2.02e3	1.08e5	**1.62e3**
15	50	1.23e4	2.18e3	2.19e5	**8.72e2**
15	500	2.83e5	3.19e3	4.26e5	**1.94e3**
15	5000	1.85e7	4.85e3	1.16e6	**3.40e3**

Table 5.H.1: Simulation cost of different algorithms on separable convex functions.

global optimality of a solution, the algorithm needs to estimate the objective values of $3^d - 1$ neighbouring points. This requires considerably more computational efforts compared to the ISC algorithm that only checks $2d$ neighbouring points to verify the local optimality. Even in the case when $d = 8$, this will lead to 400 times more simulations. For the optimal allocation example, the ISC algorithm simulates the neighbouring points of each potential solution at least 20 times to guarantee the targeted confidence level of statistical tests, which will lead to 1.31×10^5 extra simulations for each potential solution. On the other hand, our proposed algorithms are able to find globally optimal solutions for $L^\natural$-convex functions and, thus, our proposed algorithms provides a stronger theoretical guarantee.

Second, the implementation of the ISC algorithm is highly optimized to reduce the simulation costs in large-scale industrial applications. As a comparison, our implementation of the algorithms proposed in this chapter are not optimized to achieve the best performance, since the purpose of our codes is to compare the performance of our algorithms and verify our theoretical analysis. We believe that the simulation costs of our proposed algorithms can be further reduced by using an improved implementation. For example, we can use different estimated variances at different solutions to reduce the simulation costs (since the simulation costs can be lower at solutions whose simulation output has a lower variance).

In summary, the ISC algorithm achieves a better empirical performance on some experiments but provides a weaker theoretical guarantee. Our proposed algorithms, on the other hand, may be inferior in certain cases but have a better performance in the large-scale case and provide a stronger theoretical guarantee.

Params.		SubGD		Dim Reduction		ISC		R-SPLINE	
d	N	Cost	Obj.	Cost	Obj.	Cost	Obj.	Cost	Obj.
4	10	3.06e5	2.13e1	2.42e4	2.40e1	1.01e5	2.18e1	**1.21e4**	2.23e1
4	20	1.08e5	3.41e1	**1.40e4**	3.44e1	6.47e4	3.41e1	2.80e4	3.43e1
4	30	7.79e4	4.59e1	**9.21e3**	4.59e1	7.27e4	4.79e1	1.66e4	4.71e1
4	40	5.06e4	5.73e1	**6.31e3**	5.75e1	8.95e4	5.63e1	1.07e4	5.73e1
4	50	4.50e4	6.91e1	**4.03e3**	6.67e1	9.79e4	6.66e1	6.86e3	6.75e1
8	10	1.20e6	2.01e1	**1.48e5**	2.12e1	1.54e5	2.13e1	5.33e5	2.17e1
12	10	2.69e6	1.90e1	6.10e5	2.01e1	**3.02e5**	2.09e1	1.71e6	5.84e1
16	10	4.78e6	1.83e1	1.59e6	1.91e1	**1.16e6**	1.88e1	4.29e6	7.13e1
20	10	7.45e6	1.78e1	**3.21e6**	1.81e1	6.51e6	1.78e1	9.95e6	8.05e1
24	10	1.43e7	1.71e1	**8.54e6**	1.76e1	2.42e7	2.67e1	2.48e7	8.42e1

Table 5.H.2: Simulation cost and objective value of different algorithms on the resource allocation problem.

Comparison to R-SPLINE

In this subsection, we compare the dimension reduction algorithm (Algorithm 10), the subgradient descent algorithm (Algorithm 3), and the R-SPLINE algorithm [226]. For the R-SPLINE algorithm, we choose the maximal number of retrospective iterations to be the maximal budget, the initial sample size to be 10 and the initial spline budget to be 10. Other parameters are chosen to be their default values. For the separable convex function optimization problem, we require that the returned solutions of all experiments have objective function values at most d; for the optimal allocation problem, we require that the average of the estimated objective function value of the returned solution not be larger than that of the dimension reduction algorithm. Since the R-SPLINE algorithm only supports the fixed-budget optimization via simulation, we set the budget of the R-SPLINE algorithm to be the minimum multiple of $\lfloor 0.1B \rfloor$ such that the aforementioned condition is satisfied, where B is the average simulation cost of the dimension reduction algorithm in this setting. In addition, we cap the maximum budget at $5B$. We implement 100 independent experiments for the separable convex function minimization and 20 independent experiments for the optimal allocation problem. The R-SPLINE codes are implemented using MATLAB 2020a.

The results are summarized in Tables 5.H.1 and 5.H.2. For the separable convex function minimization problem, the R-SPLINE algorithm achieves the best performance. Similar to the ISC algorithm, this is because the R-SPLINE algorithm only checks the local optimality by estimating the objective function values of $2d$ neighbouring points. In the separable convex function minimization problem, the locally optimal solution happens to be the globally

optimal solution and, thus, the R-SPLINE algorithm can achieve the best performance. We note that the growth of the simulation cost of the R-SPLINE algorithm is faster than that of the dimension reduction algorithm when the scale N becomes larger. Hence, we expect that the dimension reduction algorithm will be better when the problem scale is very large.

For the optimal allocation problem, the R-SPLINE algorithm has a difficulty in reaching a global solution when the dimension is larger than 4. This is also because the R-SPLINE only checks the local optimality condition and may get stuck at locally optimal solutions that are not global optimal. By comparing with the performance of the dimension reduction algorithm, we can see that our proposed algorithms can provide a stronger theoretical guarantee for $L^\natural$-convex functions.

Numerical Results with Small Precision Parameter

In this subsection, we consider the optimal allocation problem and compare the simulation cost of different algorithms with a smaller precision parameter ϵ. In Section 5.5, we choose ϵ to be $N/2$, which is at least 20% of the optimal objective function value. To show the performance of algorithms when the precision parameter is small, we consider the case when $\epsilon = N/10 + 1$. This choice of the precision parameter is approximately 10% of the optimal objective function value. With this smaller choice of ϵ and dimension $d \geq 8$, the simulation cost may be prohibitively large (longer than 24 hours) on a personal computer. Therefore, we focus on the case when $d = 4$ and $N = 10, \ldots, 50$. The results are summarized in Table 5.H.3. Compared with the results of large precision parameter in Table 5.5.2, we can see that the algorithms perform similarly and the dimension reduction algorithm also achieves the best performance. With a smaller precision parameter, the objective function value of the solution returned by different algorithms is closer to each other compared with the large precision parameter case. This indicates that our algorithms may have achieved a much better optimality gap than ϵ.

Params.		**Search Methods**		**Localization Methods (this chapter)**					
		SubGD		Vaidya's		Random Walk		Dim Reduction	
d	N	Cost	Obj.	Cost	Obj.	Cost	Obj.	Cost	Obj.
4	10	5.61e6	2.13e1	7.41e5	2.18e1	9.88e5	2.13e1	**1.93e5**	2.25e1
4	20	3.53e6	3.42e1	4.67e5	3.41e1	5.27e5	3.41e1	**1.53e5**	3.42e1
4	30	2.43e6	4.52e1	3.48e5	4.59e1	3.81e5	4.53e1	**1.24e5**	4.51e1
4	40	1.80e6	5.62e1	2.17e5	5.68e1	2.76e5	5.63e1	**1.06e5**	5.65e1
4	50	1.66e6	6.67e1	1.53e5	6.74e1	1.68e5	6.81e1	**9.81e4**	6.66e1

Table 5.H.3: Simulation cost and objective value on the allocation problem with smaller precision parameter.

Part III

Power Systems

Chapter 6

Uniqueness of Power Flow Solutions Using Graph-theoretic Notions

6.1 Introduction

The *AC power flow problem* plays a crucial role in various aspects of power systems, e.g., the daily operations in contingency analysis and security-constrained dispatch of electricity markets. In essence, the goal of the AC power flow problem is to solve for the complex voltage of each bus that determines the power system set-point. However, the nonlinear nature of the AC power flow equations makes it difficult to analytically solve the equations, if not impossible. Moreover, the uniqueness of the AC power flow solution is not guaranteed, even when either voltage magnitudes or phase angle differences are limited to the "physically realizable" regime [184, 51, 99, 175]. Hence, unexpected operating points may appear for some system conditions and can jeopardize the normal operations of power systems. Conditions that ensure the existence of a unique "physically realizable" power flow solution are important but not fully understood.

For a special case of the AC power flow problem, the uniqueness property of the P-Θ power flow problem [113] has been studied in [184]. In the P-Θ power flow problem, the magnitude of the complex voltage at each node is given and the objective is to find a set of voltage phases such that the power flow equations are satisfied. The "physically realizable" constraint requires that the angular difference across every line lies within the stability limit of $\pi/2$ for lossless networks. Sufficient conditions (on the angular differences) that depend on the topological properties of the power network are established in [184]. Specifically, the authors proposed the notion of monotone regime and an upper bound on the angular differences based on the power network topology, which together can ensure the uniqueness of solutions. However, due to the nonlinear property of sinusoidal functions and the low-rank structure of angular differences, it is unclear to what extent the sufficient conditions given in [184] are necessary.

The goal of this chapter is to provide more general necessary and sufficient conditions

for the uniqueness, using the notion of maximal eye defined in Section 6.3 and the notion of maximal girth introduced in [184]. The paper also designs algorithms to compute these graph-theoretic parameters.

Main results

In this chapter, we extend the uniqueness theory of P-Θ power flow problem proposed in [184]. We focus on the uniqueness of the power flow problem in a stronger sense and derive general necessary and sufficient conditions that *depend only on the choice of the monotone regime and network topology*. Under certain circumstances, the general conditions can be simplified to obtain tighter sufficient conditions. In addition, some algorithms for computing the maximal eye and the maximal girth of undirected graphs are proposed. A reduction method is designed to reduce the size of graphs and accelerate the computation process. More specifically, the contributions of this chapter are three-fold:

- We extend the uniqueness theory of the P-Θ problem to a stronger sense. The new uniqueness property is named strong uniqueness. and a constant called the maximal eye is developed to classify all network topologies that ensure the strong uniqueness. Numerical results show that the maximal eye gives more reasonable conditions compared to its counterpart for the weak uniqueness defined in [184] and is known as the maximal girth.

- We propose general necessary and sufficient conditions for both the strong and the weak uniqueness. The conditions are derived by Farka's Lemma, which are associated with the dual to the negation of the uniqueness problem. Sufficient conditions for the strong and the weak uniqueness are derived directly from the general conditions. In the special case when the power network is a single cycle or is lossless, necessary and sufficient conditions that do not contain sinusoidal functions are derived.

- Finally, we develop a reduction method, named the ISPR method, that can accelerate the computation of the maximal eye and the maximal girth. The ISPR method is proved to reduce 2-vertex-connected Series-Parallel graphs to a single line, independent of the choice of the slack bus. The relationship between the maximal eye (girth) of graphs before and after the reduction is unveiled. When applying the ISPR method to real-world examples, the maximal eye is usually not changed over the reduction process, while the maximal girth is computed during the reduction process. We also design search-based algorithms for computing the maximal eye and the maximal girth, which are able to compute the exact value for graphs with up to 100 nodes before reduction in a reasonable amount of time.

In summary, this chapter constitutes a substantial generalization of the uniqueness theory in [184]. A stronger notion of uniqueness is proposed and general necessary and sufficient conditions are proposed. These two combined provides a tool for analyzing large-scale power

networks and enables a deeper understanding of the uniqueness of the P-Θ power flow problem.

Related Work

The study of solutions to the power flow problem has a long history dating back to [134], which gave an example showing the general non-uniqueness of solutions for the power flow problem. Then, the number of solutions of the power flow problem was estimated in [13], which also characterized the stability region for the power flow problem. However, these early works only considered lossless transmission networks consisting of PV buses.

The fully coupled AC power flow equations are extremely difficult to analyze and the theoretical results that can be obtained are often highly conservative or complicated to interpret. One approach to overcoming this difficulty is to study two decoupled power flow problems (the P-Θ problem and the Q-V problem) as in [113]. The intuition comes from the fact that the sensitivity of real power with respect to the change in angle differences outweighs the sensitivity with respect to the change in voltage magnitudes when angle differences are small and voltage magnitudes are close to 1 p.u. (the opposite relationship holds for reactive power). This simplification should be differentiated from the DC approximations, which greatly simplifies the AC power flow equations by linearizing the equations and discarding all of the non-linearities in the problem. Note that the P-Theta problem is still highly nonlinear. Under the assumption that resistive losses are negligible, conditions for the existence and uniqueness of both real power-phase (P-Θ) problem, and reactive power-voltage (Q-V) problem were derived in [214, 113].

In another line of work, the topology structure of the power network was also considered to derive stronger conditions for the uniqueness. The number of solutions was estimated for radial networks in [51, 165], and later for general networks. Moreover, a more recent work [64] gave several algorithms to compute the unique high-voltage solution. [59] established upper bounds on the number of linearly stable fixed point solutions for locally coupled Kuramoto models, which can be applied towards a lossless power flow problem. In this chapter, we consider the P-Θ problem [113] for general lossy power networks and utilize the topology information. We refer to [184] for a more detailed review of the existing literature.

The fixed-point technique is often used for proving the existence and uniqueness of equations. For the power flow problem, the fixed-point technique was first utilized in [234] and was further developed by several works [195, 225, 202, 203, 19, 55]. Another more recently applied approach is to treat the P-Θ power flow problem as a rank-1 matrix sensing problem and solve its convex relaxation counterpart [162, 251]. The work [65] also considered the domain of voltages over which the power flow operator is monotone. However, the relation between the rank-1-constrained problem and its convexification is not clear for general power networks.

The work [116] presented a unifying framework for network problems on the n-torus. The framework applies to the AC power flow problem when the power networks are lossless. The idea of considering the regime when the power flow on each line is monotone was extended

to lossy power networks in [184]. The regime where the power flow on a line increases monotonically with the angle difference across the line – called the monotone regime in this chapter – was proposed. In [184], it was also shown that the solution of P-Θ problem is unique under the assumption that angle differences across the lines are bounded by some limit related to the maximal girth of the network, which is defined in [185]. We refer the reader to the survey paper [62] for an overview.

The existing algorithms in the literature cannot be directly used to compute maximal eye (introduced in Section 6.3) or maximal girth. A related problem is computing the maximal chordless cycle as an upper bound to these parameters. The computation of maximal chordless cycles was proved to be $\mathcal{NP}$-complete in [84]. Efficient algorithms for enumerating chordless cycles were proposed in [61, 220] and both take linear time to enumerate a single chordless cycle. The algorithms for enumerating maximal chordless cycles can be easily modified to compute the minimal chordless cycle containing a given edge. Series-parallel reduction method was introduced as an alternative definition of Generalized Series-Parallel (GSP) graphs in [133]. Under the assumption that the slack bus is the last bus to be reduced, all GSP graphs can be reduced to a single line [184]. However, whether the series-parallel reduction method can still reduce GSP graphs without the assumption on the slack bus is not known. In this chapter, we show that 2-vertex-connected[1] Series-Parallel graphs can be reduced to a single line without the assumption.

Organization

The remainder of this chapter is organized as follows. Section 6.2 gives the necessary background knowledge about the P-Θ power flow problem and the existing uniqueness theory for the P-Θ problem. The notions of strong uniqueness and weak uniqueness are also introduced. In Section 6.3, we propose the general analysis framework of the uniqueness theory that only depends on the monotone regime and the topological structure. We show that necessary and sufficient conditions can be fully characterized by a feasibility problem, which has fewer variables than the P-Θ problem. Sufficient conditions for uniqueness are derived and it is shown that the uniqueness conditions in [184] follow as a natural corollary. Then, we consider three special cases in Section 6.4 by assuming specific topological structures for the underlying graph or a specific monotone regime. In these special cases, the necessary and sufficient conditions are simplified and the intricate sinusoidal functions are avoided in the verification of those conditions. Furthermore, the sufficient conditions proposed in Section 6.3 are proved to be tight when no information beyond the monotone regime and the topological structure is available. Finally, a reduction method and search-based algorithms for computing the maximal girth and maximal eye are given in Section 6.5. We provide numerical illustrations in Section 6.6. Proofs are delineated in the appendix.

[1] A graph is called 2-vertex-connected if it is connected after the deletion of any single vertex.

6.2 Preliminaries

P-Θ Problem Formulation

As mentioned in the introduction, we focus our attention to the P-Θ problem, which describes the relationship between the voltage phasor angles and the real power injections. We first make the following assumptions.

Assumption 13. The slack bus and the reference bus are bus 1. All other buses except the slack bus are PV buses.

Recall that the following injection operator describes the P-Θ problem, where the shunt elements are assumed to be purely reactive.

Definition 18. Given $\mathbb{G} = (\mathbb{V}, \mathbb{E}, Y)$, define $\hat{P}_k : \{0\} \times \mathbb{R}^{n-1} \to \mathbb{R}$ as the map from the vector of phasor angles to the real power injection at bus k:

$$\hat{P}_k(\Theta) := \Re\{(Yv)_k^H v_k\}, \quad \forall \Theta \in \{0\} \times \mathbb{R}^{n-1}.$$

Moreover, define the injection operator $\hat{P} : \{0\} \times \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ as

$$\hat{P}(\Theta) := [\hat{P}_2(\Theta), \ldots, \hat{P}_n(\Theta)].$$

The goal of the P-Θ problem is, given $P \in \mathbb{R}^{n-1}$, to find the voltage phasor angles $\Theta \in \{0\} \times \mathbb{R}^{n-1}$ such that

$$\hat{P}(\Theta) = P. \tag{6.1}$$

Monotone Regime and Allowable Sets

We are interested in the uniqueness property of the solution to problem (6.1). In general, the number of solutions to problem (6.1) is hard to estimate because of the periodic behavior of sinusoidal functions, especially when there is no symmetrical structure in the power network. Thus, we limit the phase angle vectors to the monotone regime, within which the real power flow from bus k to bus ℓ increases monotonically with respect to the phase difference $\Theta_{k\ell}$ for each line $\{k, \ell\} \in \mathbb{E}$. The monotone regime is defined in [184] as follows.

Definition 19. The **monotone regime** of a power network $(\mathbb{V}, \mathbb{E}, Y)$ is the set

$$\{\Theta \in \mathbb{R}^n \mid \Theta_1 = 0, \Theta_{k\ell} \in [-\gamma_{k\ell}, \gamma_{k\ell}], \forall \{k, \ell\} \in \mathbb{E}\},$$

where $\gamma_{k\ell} := \tan^{-1}(B_{k\ell}/G_{k\ell}) \in [0, \pi/2]$ for all $\{k, \ell\} \in \mathbb{E}$.

Due to the periodicity of sinusoidal functions, the solution to the P-Θ problem is trivially non-unique if there is no constraint on the phase angles. In this chapter, we consider the case when the voltage phase angles are within the monotone regime. It is noted in [206] that $\gamma_{k\ell}$ is generally larger than $2\pi/5$ while $\Theta_{k\ell}$ is rarely larger than $\pi/6$ due to stability and thermal limits. The constraint that the angular difference across every line lies within the stability limit of $[-\gamma_{k\ell}, \gamma_{k\ell}]$ is equivalent to the steady-state stability limit if each line is considered individually. As shown in [184], the phase angle vectors of leaf buses except the slack bus are uniquely determined by the phase angle vectors of non-leaf buses in the monotone regime. Hence, we assume that all vertices in the underlying graph except vertex 1 have degree at least 2.

Assumption 14. The graph $(\mathbb{V}, \mathbb{E})$ is connected. All vertices except vertex 1 in the graph $(\mathbb{V}, \mathbb{E})$ have degree at least 2.

We focus on finding a neighborhood of a solution in which there is no other solution to the P-Θ problem. The neighborhood is defined as follows.

Definition 20. The set of **allowable perturbations** is defined as

$$\mathcal{W} := \{\omega_{k\ell} \geq 0 \mid \forall \{k, \ell\} \in \mathbb{E}\}.$$

Suppose that Θ is a solution to the P-Θ problem in the monotone regime. Then, the set of **neighboring phases** is defined as

$$\begin{aligned}\mathcal{N}(\mathbb{G}, \Theta, \mathcal{W}) := \{\tilde{\Theta} \in \mathbb{R}^n \mid \tilde{\Theta}_1 = 0,\\ \tilde{\Theta}_{k\ell} \in [-\gamma_{k\ell}, \gamma_{k\ell}] \cap [\Theta_{k\ell} - \omega_{k\ell}, \Theta_{k\ell} + \omega_{k\ell}], \forall \{k, \ell\} \in \mathbb{E}\}.\end{aligned}$$

We note that $\tilde{\Theta}_{k\ell}$ refers to the value of $\tilde{\Theta}_k - \tilde{\Theta}_\ell$ modulo 2π.

Without loss of generality, we assume that $\omega_{k\ell} \leq 2\gamma_{k\ell}$ for all $\{k, \ell\} \in \mathbb{E}$, since the width of the monotone regime is $2\gamma_{k\ell}$, setting $\omega_{k\ell} > 2\gamma_{k\ell}$ will not enlarge the set of neighboring phases compared to setting $\omega_{k\ell} = 2\gamma_{k\ell}$.

Assumption 15. The perturbation width satisfies $\omega_{k\ell} \leq 2\gamma_{k\ell}$ for all $\{k, \ell\} \in \mathbb{E}$.

It is desirable to analyze the uniqueness of the solution in the neighborhood $\mathcal{N}(\mathbb{G}, \Theta, \mathcal{W})$. In [184], the authors considered the *set of allowable angles*, which is defined as

$$\{\tilde{\Theta} \in \mathbb{R}^n | \tilde{\Theta}_1 = 0, \tilde{\Theta}_{k\ell} \in [-\omega_{k\ell}/2, \omega_{k\ell}/2], \forall \{k, \ell\} \in \mathbb{E}\}.$$

Note that the set of allowable angles is a special case of the set of allowable perturbations, since any two phase vectors in the set of allowable angles are in the corresponding sets of neighboring phases of each other. In this chapter, we use the *set of allowable perturbations* but the sufficient conditions we derive can be naturally applied to using the *set of allowable angles*.

Notions of Weak and Strong Uniqueness

Informally, we say that the P-Θ problem (6.1) has a unique solution Θ under the allowable perturbation set $\mathcal{W}$, if there exists at most one solution in the set $\mathcal{N}(\mathbb{G}, \Theta, \mathcal{W})$. We give two different definitions of uniqueness. Firstly, we introduce the uniqueness in the weak sense.

Definition 21. We say that a solution Θ to the P-Θ problem (6.1) is **weakly unique** with the given set of allowable perturbations $\mathcal{W}$, if for any solution $\tilde{\Theta} \in \mathcal{N}(\mathbb{G}, \Theta, \mathcal{W})$, there exists a line $\{k, \ell\} \in \mathbb{E}$ such that $\Theta_{k\ell} = \tilde{\Theta}_{k\ell}$.

In other words, two solutions are different according to Definition 21 if and only if they have different phase differences for every line. Next, we extend the definition of weak uniqueness to a stronger sense that is also more useful and usual.

Definition 22. We say that a solution Θ to the P-Θ problem (6.1) is **strongly unique** with the given set of allowable perturbations $\mathcal{W}$, if for any solution $\tilde{\Theta} \in \mathcal{N}(\mathbb{G}, \Theta, \mathcal{W})$ and any $\{k, \ell\} \in \mathbb{E}$, we have $\Theta_{k\ell} = \tilde{\Theta}_{k\ell}$.

In other words, two solutions are different according to Definition 22 if and only if the phase differences are different on at least one line.

6.3 Uniqueness Theory for General Graphs

In this section, we derive necessary and sufficient conditions on the set of allowable perturbations $\mathcal{W}$ such that the solution to problem (6.1) becomes strongly or weakly unique. In particular, we aim to analyze *the impact of the power system topology and the size of the monotone regime* on the uniqueness property. Namely, given the topological structure and the monotone regime, we aim to find conditions on $\mathcal{W}$ such that the uniqueness of solutions holds. To achieve this, we need to derive conditions under which all power networks with the same topological structure and monotone regime have unique solutions. To formalize the problem, we fix the underlying graph $(\mathbb{V}, \mathbb{E})$ and the angles specifying the monotone regime $\Gamma := \{\gamma_{k\ell} \in (0, \pi/2] \mid \{k, \ell\} \in \mathbb{E}\}$. We define the set of possible admittances with the same monotone regime as

$$\mathcal{S}(\gamma) := \{(C\cos(\gamma), C\sin(\gamma)) \mid C > 0\},, \quad \forall \gamma \in [0, \pi/2].$$

The set of complex admittance matrices with the same monotone regime is defined as

$$\begin{aligned}\mathcal{Y}(\mathbb{V}, \mathbb{E}, \Gamma) := \{&Y \text{ is an admittance matrix } | \\ &Y_{k\ell} = G_{k\ell} - \mathbf{j}B_{k\ell}, (G_{k\ell}, B_{k\ell}) \in \mathcal{S}(\gamma_{k\ell}), \{k, \ell\} \in \mathbb{E}\}.\end{aligned}$$

Then, we define the set of power networks with the same topological structure and same monotone regime as

$$\mathcal{G}(\mathbb{V}, \mathbb{E}, \Gamma) := \{\mathbb{G} = (\mathbb{V}, \mathbb{E}, Y) \mid Y \in \mathcal{Y}(\mathbb{V}, \mathbb{E}, \Gamma)\},$$

or simply $\mathcal{G}$ if there is no confusion about $\mathbb{V}$, $\mathbb{E}$ and Γ. Hence, the problem under study in this chapter can be stated as follows:

- What are the necessary conditions and sufficient conditions on the allowable perturbations $\mathcal{W}$ such that the solution to problem (6.1) is unique within the set of allowable perturbations for any power network $\mathbb{G} \in \mathcal{G}$?

The necessary conditions and the sufficient conditions provide two sides on the uniqueness theory. The sufficient conditions give a guarantee for the uniqueness of solutions for any single power network with the given topological structure and monotone regime, while the necessary conditions bound the optimal conditions we can derive only using the knowledge of topological structure and monotone regime. We first give an equivalent characterization of strong and weak uniqueness.

Lemma 47. (Necessary and Sufficient Conditions for Uniqueness) *Given the set of power networks $\mathcal{G}(\mathbb{V}, \mathbb{E}, \Gamma)$ and the set of allowable perturbations $\mathcal{W}$, the following two statements are equivalent:*

1) For any power network $\mathbb{G} \in \mathcal{G}(\mathbb{V}, \mathbb{E}, \Gamma)$ and any power injection $P \in \mathbb{R}^{|\mathbb{V}|-1}$ such that problem (6.1) is feasible in the monotone regime, the solution to problem (6.1) in the monotone regime is strongly unique in $\mathcal{N}(\mathbb{G}, \Theta, \mathcal{W})$.

2) For any power network $\mathbb{G} \in \mathcal{G}(\mathbb{V}, \mathbb{E}, \Gamma)$ and any two phase angle vectors Θ^1, Θ^2 in the monotone regime with the property $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$, there exists a vector $\mathbf{y} \in \mathbb{R}^{|\mathbb{V}|}$ such that $y_1 = 0$ and

$$\begin{aligned} \sin(\gamma_{k\ell} + \Theta^1_{k\ell}/2 + \Theta^2_{k\ell}/2) \cdot y_k & \\ \geq \sin(\gamma_{k\ell} - \Theta^1_{k\ell}/2 - \Theta^2_{k\ell}/2) \cdot y_\ell, & \\ \forall \{k, \ell\} \in \mathbb{E} \quad \text{s.t. } \Theta^1_{k\ell} - \Theta^2_{k\ell} > 0, & \end{aligned} \tag{6.2}$$

where at least one of the inequalities above is strict.

The equivalence between statements 1 and 2 still holds true even after replacing strong uniqueness with weak uniqueness in statement 1, provided that the phase angle vector Θ^2 in statement 2 is required to satisfy $\Theta^1_{k\ell} \neq \Theta^2_{k\ell}$ for all $\{k, \ell\} \in \mathbb{E}$.

Intuitively, the above lemma studies the uniqueness of solutions through its dual form. The existence of multiple solutions can be formulated as a linear feasibility problem. Then, the strong duality of linear programming allows us to equivalently consider the dual of the feasibility problem. The dual form is preferred since the dual problem has fewer variables and its solution is easier to construct. We then derive several sufficient conditions using Lemma 47. We first show that we only need to verify statement 2 in Lemma 47 for two phase angle vectors Θ^1 and Θ^2 that induce a (weakly) feasible orientation, which we will define below. We define the orientation induced by two phase angle vectors.

Definition 23. Suppose that Θ^1 and Θ^2 are two phase angle vectors of the graph. Then, we define the **induced orientation** of $\Delta := \Theta^1 - \Theta^2$ as $A_{k\ell} := \text{sign}(\Delta_{k\ell})$ for all $\{k, \ell\} \in \mathbb{E}$, where the sign function $\text{sign}(\cdot)$ is defined as

$$\text{sign}(x) := \begin{cases} +1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

In the definition of induced orientations, we assign one of the three directions $+1, -1, 0$ to each edge. The first two directions are "normal" directions for directed graphs. An edge with direction $+1$ or -1 is called a **normal edge**. Edges with direction 0 are viewed as an undirected edge and reachable in both directions. In addition, edges with direction 0 are not considered when computing the in-degree and the out-degree. We only need to consider orientations induced by two different phase angle vectors Θ^1, Θ^2 such that $\hat{P}(\Theta^1) = \hat{P}(\Theta^2)$. However, a precise characterization of those orientations is difficult and we consider a larger set that contains those orientations.

Definition 24. An orientation assigned to an undirected graph is called a **feasible orientation** if all edges are normal and each vertex except vertex 1 has nonzero in-degree and out-degree.

According to the analysis in [184], the induced orientation of two solutions Θ^1 and Θ^2 in the monotone regime that are different according to Definition 21 must be a feasible orientation. Then, we give the definition of weakly feasible orientations as the counterpart for strong uniqueness.

Definition 25. An orientation assigned to an undirected graph is called a **weakly feasible orientation** if two properties are satisfied: (i) there exists at least one normal edge, and (ii) the in-degree and the out-degree of any vertex except vertex 1 are both zero or both nonzero.

Edges with direction 0 are lines with the same angular difference for the two phase angle vectors Θ^1 and Θ^2. By the same discussion as in Section 6.2, we can view a weakly feasible orientation as a feasible orientation for the sub-graph that only has normal edges. The next lemma shows that we only need to consider weakly feasible orientations or feasible orientations when checking the conditions in statement 2 of Lemma 47.

Lemma 48. *If two different phase angle vectors $\Theta^1 - \Theta^2$ in the monotone regime satisfy $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$ and the induced orientation of $\Theta^1 - \Theta^2$ is not weakly feasible, then there exists a vector $\mathbf{y} \in \mathbb{R}^{|\mathbb{V}|}$ such that statement 2 of Lemma 47 holds. The result holds true for the weak uniqueness property as well, provided that the induced orientation of Θ^1, Θ^2 is not a feasible orientation.*

Combining Lemmas 47 and 48, we obtain sufficient conditions for strong uniqueness and weak uniqueness.

Theorem 65. (Sufficient Conditions for Uniqueness) *Given the set of allowable perturbations $\mathcal{W}$, suppose that for any two different phase angle vectors Θ^1 and Θ^2 in the monotone regime satisfying $\Theta_2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$, the induced orientation of $\Theta^1 - \Theta^2$ is not a weakly feasible orientation. Then, the solution to problem* (6.1) *is strongly unique for all power networks in $\mathcal{G}$. The result holds true for the weak uniqueness as well, provided that the induced orientation of $\Theta^1 - \Theta^2$ is not a feasible orientation.*

The sufficient condition given above is a generalization of Theorem 4 in [184], which ensures the weak uniqueness of solutions in the set of allowable phases. Using Theorem 65, we can derive a corollary similar to Theorem 4 in [184].

Corollary 7. *Consider an arbitrary set of allowable perturbations $\mathcal{W}$. The solution to problem (6.1) in the monotone regime is strongly unique for any power network $\mathbb{G} \in \mathcal{G}$ if for any weakly feasible orientation of the underlying graph $(\mathbb{V}, \mathbb{E})$, there exists a directed cycle $(k_1, \ldots, k_t)$ containing at least one normal edge such that the allowable perturbations satisfy the inequality*

$$\sum_{\{k_i, k_{i+1}\} \text{ is normal}} \omega_{k_i k_{i+1}} < 2\pi,$$

where $k_{t+1} := k_1$. The same result holds true for the weak uniqueness if we substitute weakly feasible orientations with feasible orientations.

Now, we consider a special case where all constants $\omega_{k\ell}$ in the set of allowable perturbations are equal, i.e., there exists a constant $\omega \geq 0$ such that the set of allowable perturbation is

$$\mathcal{W}_\omega := \{\omega_{k\ell} = \omega, \forall \{k, \ell\} \in \mathbb{E}\}.$$

The problem we consider in this case is:

- What is the sufficient condition on ω such that the solution to problem (6.1) is unique with the allowable perturbation set $\mathcal{W}_\omega$?

We derive an upper bound on the constant ω to guarantee the uniqueness. We first define the maximal eye and the maximal girth of an undirected graph.

Definition 26. Consider an undirected graph $(\mathbb{V}, \mathbb{E})$. For any weakly feasible orientation assigned to the graph $(\mathbb{V}, \mathbb{E})$, we define the minimal length of directed cycles that contain at least one normal edge as the **size of eye** of this orientation, where edges with direction 0 are considered as bi-directional edges. We define the **maximal eye** of the graph $(\mathbb{V}, \mathbb{E})$ as the maximum of the size of eye over all possible weakly feasible orientations. We denote the maximal eyes of the graph $(\mathbb{V}, \mathbb{E})$, a power network $\mathbb{G}$ and a group of power networks $\mathcal{G}$ as $\mathrm{e}(\mathbb{V}, \mathbb{E})$, $\mathrm{e}(\mathbb{G})$ and $\mathrm{e}(\mathcal{G})$, respectively.

Remark 7. There always exists a directed cycle containing normal edges when the underlying graph is under a weakly feasible orientation. To understand this, we first choose an arbitrary

normal edge $(k_1, k_2) \in \mathbb{E}$. Since the vertex k_2 has nonzero in-degree, it also has nonzero out-degree. Hence, there exists another vertex k_3 such that $(k_2, k_3) \in \mathbb{E}$. Continuing this procedure will result in the existence of a vertex k_t such that $v_t = k_s$ for some $s < t$. This generates a directed cycle $(k_s, k_{s+1}, \ldots, k_{t-1})$ containing only normal edges. Hence, the size of eye is well-defined.

The counterpart of the maximal eye, known as the maximal girth, is defined in [184] and we restate the definition below.

Definition 27. Consider an undirected graph $(\mathbb{V}, \mathbb{E})$. For any feasible orientation assigned to the underlying graph $(\mathbb{V}, \mathbb{E})$, we define the minimal size of directed cycles as the **girth** of this feasible orientation. We define the **maximal girth** of the graph $(\mathbb{V}, \mathbb{E})$ as the maximum of the girth over all feasible orientations. We denote the maximal girths of the graph $(\mathbb{V}, \mathbb{E})$, a power network $\mathbb{G}$ and a group of power networks $\mathcal{G}$ as $\mathrm{g}(\mathbb{V}, \mathbb{E})$, $\mathrm{g}(\mathbb{G})$ and $\mathrm{g}(\mathcal{G})$, respectively.

Remark 8. Similar to the discussion in Remark 7, there exists at least one directed cycle when the graph is under a feasible orientation. The maximal eye can be equivalently defined as the maximum of the maximal girth over all sub-graphs that do not have degree-1 vertices.

We provide an upper bound for ω using the maximal eye and the maximal girth, which follows from Corollary 7.

Corollary 8. *If the inequality*

$$\omega_{k\ell} < \frac{2\pi}{\mathrm{e}(\mathcal{G})}, \quad \forall \{k, \ell\} \in \mathbb{E}, \tag{6.3}$$

is satisfied, then the solution to problem (6.1) in the monotone regime is strongly unique for any power network $\mathbb{G} \in \mathcal{G}$. The same result holds true for weak uniqueness, provided that $\mathrm{e}(\mathcal{G})$ *in* (6.3) *is substituted by* $\mathrm{g}(\mathcal{G})$.

In Section 6.5, we design search-based algorithms to calculate the maximal eye and the maximal girth. However, computing the maximal eye or the maximal girth is challenging for graphs with more than 100 nodes. Hence, we seek upper bounds and lower bounds for the maximal eye and the maximal girth. In this chapter, we obtain a simple upper bound for both the maximal girth and the maximal eye. We define $\kappa(\mathbb{G})$ and $\kappa(\mathcal{G})$ as the sizes of the longest chordless cycles of the underlying graph of the power network $\mathbb{G}$ and any power network in the power network class $\mathcal{G}$, respectively. The upper bound on the maximal girth and eye will be provided below.

Theorem 66. *For any power network $\mathbb{G}$, it holds that*

$$\mathrm{g}(\mathbb{G}) \leq \mathrm{e}(\mathbb{G}) \leq \kappa(\mathbb{G}) \tag{6.4}$$

and that $\mathrm{g}(\mathcal{G}) \leq \mathrm{e}(\mathcal{G}) \leq \kappa(\mathcal{G})$.

We note that although computing the longest chordless cycle is $\mathcal{NP}$-complete [84], the computation of the longest chordless cycle is faster than the computation of the maximal eye and the maximal girth in practice.

6.4 Uniqueness Theory for Three Special Cases

In this section, we consider three special cases. For each case, the power network either has a special topological structure or a special monotone regime. In the first two cases, the underlying graph of the power network is a single cycle or a 2-vertex-connected Series-Parallel (SP) graph. When the underlying graph is a single cycle, the sufficient conditions in Corollary 7 are also necessary. If the underlying graph is a 2-vertex-connected SP graph, we prove that the sufficient conditions for the weak uniqueness in Corollary 8 also ensure the strong uniqueness. In the last case, the power network is assumed to be lossless. In this case, the monotone regime of each line reaches the maximum possible size $[-\pi/2, \pi/2]$. Sinusoidal functions can then be avoided in statement 2 of Lemma 47, and therefore the verification of conditions is easier.

Single Cycles

We first consider the case when the underlying graph $(\mathbb{V}, \mathbb{E})$ is a single cycle. We first show that the weak uniqueness is equivalent to the strong uniqueness in this case.

Lemma 49. *Suppose that the underlying graph is a single cycle with the edges* $(1,2)$, $(2,3)$, $\ldots$, $(n,1)$. *Then, given the set of allowable perturbations* $\mathcal{W}$, *the solution to problem* (6.1) *in the monotone regime is weakly unique if and only if it is strongly unique.*

Next, we prove that the sufficient conditions derived in Corollary 7 are also necessary for a single cycle with non-trivial monotone regime.

Theorem 67. *Suppose that the underlying graph is a single cycle with the edges* $(1,2)$, $(2,3)$, $\ldots$, $(n,1)$, *and that the set of allowable perturbations satisfies* $0 < \omega_{i,i+1} \leq \gamma_{i,i+1}$ *for all* $i \in [n]$, *where* $\gamma_{n,n+1} := \gamma_{n,1}$ *and* $\omega_{n,n+1} := \omega_{n,1}$. *The solution to problem* (6.1) *in the monotone regime is strongly unique for any power network* $\mathbb{G} \in \mathcal{G}(\mathbb{V}, \mathbb{E}, \Gamma)$ *and any power injection* $P \in \mathbb{R}^{n-1}$ *that makes problem* (6.1) *feasible* if and only if *the set of allowable perturbations* $\mathcal{W}$ *satisfies*

$$\sum_{i=1}^{n} \omega_{i,i+1} < 2\pi,$$

where $\omega_{n,n+1} := \omega_{n,1}$.

In contrast to requiring $\omega_{i,i+1} > 0$ in the above theorem, the condition that $\omega_{i,i+1} = 0$ for some i is sufficient but not necessary for the uniqueness of solutions. Under this condition, two solutions Θ^1 and Θ^2 in the monotone regime such that $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$ must satisfy $\Theta^1_{i,i+1} = \Theta^2_{i,i+1}$. Hence, any solution is strongly unique with this set of allowable perturbations. However, by Theorem 67, this condition is not necessary for the uniqueness of solutions.

Series-Parallel Graphs

In this subsection, we consider another special class of graphs, namely, the 2-vertex-connected SP graphs. The objective is to find an upper bound on the constant ω to guarantee that the solution to problem (6.1) is unique. Corollary 8 shows that the solution is strongly unique if $\omega < 2\pi/\mathrm{e}(\mathbb{G})$ and is weakly unique if $\omega < 2\pi/\mathrm{g}(\mathbb{G})$. However, for a 2-vertex-connected SP graph, we can prove a stronger theorem. We first prove that the maximal eye is equal to the maximal girth for a 2-vertex-connected SP graph. The main tool is the ear decomposition of an undirected graph [72].

Definition 28. An **ear** of an undirected graph $(\mathbb{V}, \mathbb{E})$ is a simple path or a single cycle. An **ear decomposition** of an undirected graph $(\mathbb{V}, \mathbb{E})$, denoted as $\mathcal{D} := (L_0, \ldots, L_{r-1})$, is a partition of $\mathbb{E}$ into an ordered sequence of ears such that one or two endpoints of each ear L_k are contained in an earlier ear, i.e., an ear L_ℓ with $\ell < k$, and the internal vertices of each ear do not belong to any earlier ear. We call $\mathcal{D}$ a **proper ear decomposition** if each ear L_k is a simple path for all $k = 1, \ldots, r-1$. A **tree ear decomposition** is a proper ear decomposition in which the first ear is a single edge and for each subsequent ear L_k, there is a single ear L_ℓ with $\ell < k$, such that both endpoints of L_k lie on L_ℓ. A **nested ear decomposition** is a tree ear decomposition such that, within each ear L_ℓ, the set of pairs of endpoints of other ears L_k that lie within L_ℓ forms a set of nested intervals.

The following theorem provides an equivalent characterization of 2-vertex-connected SP graphs through the ear decomposition.

Theorem 68 ([130])**.** *A 2-vertex-connected graph is series-parallel if and only if it has a nested ear decomposition.*

With the help of the nested ear decomposition, we will prove that the maximal girth is equal to the maximal eye for 2-vertex-connected SP graphs. The intuition behind the proof is that we first choose two vertices as the "source" and the "sink" for the power flow network. For each edge with direction 0, we first consider the directed path that contains this edge and goes from the "source" to the "sink" and then assign a normal direction (± 1) to this edge according to the directed path. This step ensures that the first inequality in (6.4) holds as equality.

Lemma 50. *Suppose that $(\mathbb{V}, \mathbb{E})$ is a 2-vertex-connected SP graph. Then, the following equality holds true:*

$$\mathrm{g}(\mathbb{V}, \mathbb{E}) = \mathrm{e}(\mathbb{V}, \mathbb{E}).$$

Therefore, combining the above lemma with Corollary 8, we obtain a stronger sufficient condition for 2-vertex-connected SP graphs. This result implies that the sufficient conditions for the weak uniqueness in Corollary 8 also guarantee the strong uniqueness.

Theorem 69. *Suppose that the underlying graph* $(\mathbb{V}, \mathbb{E})$ *is a* 2*-vertex-connected SP graph. The solution to problem* (6.1) *is strongly unique for any power network* $\mathbb{G} \in \mathcal{G}$ *in the monotone regime if*

$$\omega < \frac{2\pi}{\mathrm{g}(\mathcal{G})}.$$

Lossless Networks

Finally, we consider the case when the power network is lossless, namely, when $\gamma_{k\ell} = \pi/2$ for all $\{k, \ell\} \in \mathbb{E}$. In this case, we prove that the strong uniqueness holds if and only if there does not exist another solution in the set of neighboring phases such that the induced orientation has strictly more strongly connected components than weakly connected components. This results makes it possible to avoid nonlinear sinusoidal functions in statement 2 of Lemma 47, and therefore the uniqueness of solutions becomes easier to verify. We first define the sub-graph induced by two phase angle vectors.

Definition 29. Suppose that Θ^1 and Θ^2 are two different phase angle vectors, and that the orientation A is the induced orientation of $\Theta^1 - \Theta^2$. Then, the **induced sub-graph** of $\Theta^1 - \Theta^2$ is constructed as a directed sub-graph of $(\mathbb{V}, \mathbb{E}, A)$ by first deleting all edges with direction 0 and then deleting all degree-1 vertices.

In what follows, we establish a necessary and sufficient condition for the uniqueness of the solution that does not contain sinusoidal functions.

Theorem 70. *Consider a that the set of allowable perturbations* $\mathcal{W}$*. If the monotone regime satisfies* $\gamma_{k\ell} = \pi/2$ *for all* $\{k, \ell\} \in \mathbb{E}$*, then the following two statements are equivalent:*

1) *For any power network* $\mathbb{G} \in \mathcal{G}(\mathbb{V}, \mathbb{E}, \Gamma)$ *and any power injection* $P \in \mathbb{R}^{|\mathbb{V}|-1}$ *such that problem* (6.1) *is feasible, the solution to problem* (6.1) *in the monotone regime is strongly unique in* $\mathcal{N}(\mathbb{G}, \Theta, \mathcal{W})$*.*

2) *For any power network* $\mathbb{G} \in \mathcal{G}(\mathbb{V}, \mathbb{E}, \Gamma)$ *and any two phase angle vectors* Θ^1 *and* Θ^2 *in the monotone regime with the property* $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$*, the induced sub-graph of* $\Theta^1 - \Theta^2$ *has strictly more strongly connected components than weakly connected components.*

The equivalence between statements 1 and 2 still holds true even after replacing strong uniqueness with weak uniqueness in statement 1, provided that the phase angle vectors Θ^2 *in statement 2 is required to satisfy* $\Theta^1_{k\ell} \neq \Theta^2_{k\ell}$ *for all* $\{k, \ell\} \in \mathbb{E}$*.*

The result of the above theorem is stronger than the sufficient conditions in Theorem 65. This is because any (weakly) *infeasible* orientation has strictly more strongly connected components than weakly connected components. Hence, the sufficient conditions in Theorem 65 ensure that all induced orientations are (weakly) *infeasible*. Then, statement 2 of this theorem holds true and the solution becomes strongly (weakly) unique.

6.5 Iterative Series-Parallel Reduction

In the preceding sections, we have shown that the maximal eye and the maximal girth play important roles in the uniqueness theory. However, computing the maximal eye or maximal girth is cumbersome for large graphs. Hence, we develop an iterative reduction method to design a reduced graph, and then prove the relationship between the maximal eye or the maximal girth of the original graph and those of the reduced graph. Next, we test the performance of those algorithms on real-world problems. Search-based algorithms for computing the maximal eye and the maximal girth are given in the appendix.

Iterative Series-Parallel Reduction Method

In this subsection, we propose an iterative reduction method, named as the Iterative Series-Parallel Reduction (ISPR) method, that can reduce the size of the underlying graph for computing the maximal eye and maximal girth. The ISPR method is different from the Series-Parallel Reduction (SPR) method introduced in [184] in two aspects. First, the purpose of the ISPR method is to accelerate the computation of the maximal eye and the maximal girth, while the focus of SPR method is to facilitate the verification of uniqueness conditions. Second, we prove that all 2-vertex-connected SP graphs can be reduced to a single edge (K_2) without the assumption in [184] that the slack bus is the last to be reduced.

Before introducing the ISPR method, we extend the definition of the maximal eye and the maximal girth to weighted graphs with "multiple slack buses". This generalized class of graphs appear during the reduction process. By defining the length of a cycle as the sum of the weights of the edges on the cycle, the maximal eye and the maximal girth can be generalized to weighted graphs. Next, we define (weakly) feasible orientations for graphs with "multiple slack buses", namely, the slack nodes.

Definition 30. For a weighted undirected graph $(\mathbb{V}, \mathbb{E}, W)$, a subset of vertices $\mathbb{V}_s \subseteq \mathbb{V}$ is called the set of **slack nodes**. An orientation A assigned to the graph is called a **weakly feasible orientation** if each edge has one of the directions $\{+1, -1, 0\}$ and each vertex not in $\mathbb{V}_s$ either has nonzero in-degree and nonzero out-degree, or has zero in-degree and zero out-degree. An orientation A assigned to the graph is called a **feasible orientation** if each edge has one of the directions $\{+1, -1\}$ and each vertex not in $\mathbb{V}_s$ has nonzero in-degree and nonzero out-degree.

Now, we can define the maximal eye for graphs with slack nodes by taking the maximum of the size of eye over weakly feasible orientations. The maximal girth can be defined in a similar way. For power networks, the only slack node is the slack bus of the power network. Hence, the extended definitions of the maximal eye and the maximal girth are consistent with their original definitions. The ISPR method is based on three types of operations:

- **Type I Operation.** Replacement of a set of parallel edges with a single edge that connects their common endpoints. The weight of the new single edge is the minimum over the weights of the deleted parallel edges.

- **Type II Operation.** Replacement of the two edges incident to a degree-2 vertex with a single edge, if the vertex has exactly two neighboring vertices and is not a slack node. The weight of the new edge is the sum of the weights of the two deleted edges.

- **Type III Operation.** Deletion of a vertex that has only a single neighboring vertex. If the deleted vertex is a slack node, or if the deleted vertex has degree at least 2 for the problem of computing the maximal girth, then we define its neighboring vertex as a slack node.

The update scheme of weights and slack nodes is designed to control the change of the maximal eye or the maximal girth. The ISPR method successively reduces the size of the graph by applying Type I-III Operations; the pseudo-code of the ISPR method is given in Algorithm 17. We note that after the reduction process, there is no parallel edge or pendant (degree-1) vertex in the reduced graph. Ignoring the weights of the edges and the set of the slack nodes, the operations in the ISPR method can cover the operations in the classical series-parallel reduction [133], which are defined as

- **Type I' Operation.** Replacement of parallel edges with a single edge that connects their common endpoints.

- **Type II' Operation.** Replacement of the two edges incident to a degree-2 vertex with a single edge.

- **Type III' Operation.** Deletion of a pendant vertex.

Hence, the ISPR method can be viewed as a generalization of the classical series-parallel reduction. We first consider the change of the maximal eye after each operation.

Lemma 51. *Given a weighted undirected graph* $(\mathbb{V}, \mathbb{E}, W)$*, let* e *denote its maximal eye. Assume that one of Type I-III Operations is implemented on the graph. By denoting the new graph and its maximal eye as* $(\tilde{\mathbb{V}}, \tilde{\mathbb{E}}, \tilde{W})$ *and* $\tilde{\mathrm{e}}$*, the following statements hold:*

- *If Type I Operation is implemented, then*

$$\tilde{\mathrm{e}} \leq \mathrm{e} \leq \max\{\tilde{\mathrm{e}}, W_{max} + W_{min}\},$$

  where W_{max} *and* W_{min} *are the maximal and minimal weights of the deleted parallel edges, respectively.*

- *If Type II Operation is implemented, then* $\mathrm{e} = \tilde{\mathrm{e}}$*.*

- *If Type III Operation is implemented and the deleted vertex has degree* 1*, then* $\mathrm{e} = \tilde{\mathrm{e}}$*.*

Algorithm 17 Iterative Series-Parallel Reduction method

Input: Undirected unweighted graph $(\mathbb{V}, \mathbb{E})$, slack bus k
Output: Reduced undirected weighted graph $(\mathbb{V}_R, \mathbb{E}_R, W_R)$, two constants α_1, α_2 defined in Theorems 71 and 72, set of slack nodes $\mathbb{V}_s$

```
Set the initial weight for each edge to be 1.
Set the initial set of slack nodes as V_s ← {k}.
while at least one operation is implementable do
    if Type I Operations are implementable then
        Implement Type I Operation.
        Update values α_1, α_2 according to their definitions in Theorems 71 and 72.
        continue
    end if
    if Type II Operations are implementable then
        Implement Type II Operation.
        continue
    end if
    if Type III Operations are implementable then
        Implement Type III Operation.
        Update values α_1, α_2 according to their definitions in Theorems 71 and 72.
        Update the set of slack nodes V_s.
        continue
    end if
end while
Return reduced graph (V_R, E_R, W_R), set of slack nodes V_s and values α_1, α_2.
```

- *If Type III Operation is implemented and the deleted vertex has degree larger than* 1*, then*

$$\mathrm{e} = \max\{\tilde{\mathrm{e}}, W_{max} + W_{min}\},$$

where W_{max} and W_{min} are the maximal and minimal weights of the deleted parallel edges, respectively.

Using the above lemma, we have the following theorem.

Theorem 71. *Given a power network with the underlying graph $(\mathbb{V}, \mathbb{E})$, let* e *denote the maximal eye of the graph. Denote the graph after reduction and its maximal eye as $(\mathbb{V}_R, \mathbb{E}_R, W_R)$ and* e_R*, respectively. Then, we have*

$$\max\{\mathrm{e}_R, \alpha_2\} \leq \mathrm{e} \leq \max\{\mathrm{e}_R, \alpha_1, \alpha_2\},$$

where α_1 and α_2 are the maximum of $W_{max} + W_{min}$ over Type I and Type III Operations, respectively. Here, W_{max}, W_{min} are defined in Lemma 51. If Type I or Type III Operations is never implemented, then we set α_1 or α_2 to 0*.*

Similarly, we can prove the relation between the maximal girth of the original graph and that of the reduced graph. We first show the change of the maximal girth after each operation.

Lemma 52. *Given a weighted undirected graph $(\mathbb{V}, \mathbb{E}, W)$, let* g *denote its maximal girth. Assume that one of Type I-III Operations is implemented on the graph. By denoting the new graph and its maximal girth of new graph as $(\tilde{\mathbb{V}}, \tilde{\mathbb{E}}, \tilde{W})$ and* $\tilde{\text{g}}$*, the following statements hold:*

- *If Type I Operation is implemented, then*

$$\tilde{\text{g}} \leq \text{g} \leq \max\{\tilde{\text{g}}, W_{max} + W_{min}\},$$

  where W_{max} and W_{min} are the maximal and minimal weights of the deleted parallel edges, respectively.

- *If Type II Operation is implemented, then* $\text{g} = \tilde{\text{g}}$.

- *If Type III Operation is implemented and the deleted vertex has degree* 1*, then* $\text{g} = \tilde{\text{g}}$.

- *If Type III Operation is implemented, the deleted vertex is a slack node and has degree larger than* 1*, then*

$$\tilde{\text{g}} \leq \text{g} \leq \max\{\tilde{\text{g}}, W_{max} + W_{min}\},$$

  where W_{max} and W_{min} are the maximal and minimal weights of the deleted parallel edges, respectively.

- *If Type III Operation is implemented, the deleted vertex is not a slack node and has degree larger than* 1*, then*

$$\text{g} = \min\{\tilde{\text{g}}, W_{max} + W_{min}\},$$

  where W_{max} and W_{min} are the maximal and minimal weights of the deleted parallel edges, respectively.

By the above lemma, the relationship between the maximal girth of the original graph and that of the reduced graph will be discovered below.

Theorem 72. *Given a power network with the underlying graph $(\mathbb{V}, \mathbb{E})$, let* g *denote its the maximal girth. By denoting the graph after reduction and its maximal girth as $(\mathbb{V}_R, \mathbb{E}_R, W_R)$ and* g_R*, we have*

$$\min\{\text{g}_R, \alpha_2\} \leq \text{g} \leq \min\{\max\{\text{g}_R, \alpha_1\}, \alpha_2\},$$

where α_1 is the maximum of $W_{max} + W_{min}$ over Type I Operations and the second case of Type III Operations, and α_2 is the minimum of $W_{max} + W_{min}$ over the third case of Type III Operations. Here, W_{max}, W_{min} are defined in Lemma 51. If operations for computing α_1 or α_2 are never implemented, then we set α_1 to 0 *or α_2 to $+\infty$.*

Based on the numerical results in Tables 6.6.1 and 6.6.2 for large power networks, the values of α_1 and α_2 in Theorems 71 and 72 are usually smaller than e_R and g_R. Hence, we have the approximation

$$\mathrm{e} \approx \mathrm{e}_R, \quad \mathrm{g} \approx \alpha_2. \tag{6.5}$$

The above relations imply that for large power networks, computing the maximal eye is equivalent to computing the maximal eye of a reduced graph, while the maximal girth is already computed during the reduction process. Finally, we prove that 2-vertex-connected SP graphs can be reduced to a single edge by the ISPR method.

Theorem 73. *If the underlying graph $(\mathbb{V}, \mathbb{E})$ of a power network is a 2-vertex-connected SP graph, then the ISPR method reduces the underlying graph to a single edge.*

For an undirected graph without slack nodes, the classical series-parallel reduction (Type I'-III' Operations) can reduce the graph to a single edge if and only if the graph is a Generalized Series-Parallel (GSP) graph [133]. We note that 2-vertex-connected SP graphs are a special class of GSP graphs and it is unclear whether the reduction guarantee for theISPR method can be extended to any GSP graphs in the presence of slack nodes.

6.6 Numerical results

In this section, we verify the theoretical results of this chapter and test the performance of the proposed algorithms. First, we show that, using the ISPR method, the computation of the maximal eye can be reduced to a smaller graph, while the computation of the maximal girth is finished during the process of reduction. Then, we show that Corollary 8 gives a valid sufficient condition for strong uniqueness. We use IEEE power networks in MATPOWER [259] to perform experiments. Finally, the proximity between the P-Θ problem and the AC power flow problem is numerically illustrated.

Computation of the Maximal Eye and the Maximal Girth

We first consider the computation of the maximal eye. The results are listed in Table 6.6.1. Here, we use '-' to denote the case when this value does not exist, and use 'TLE' (Time Limit Exceeded) to denote the case when the algorithm does not find any leaf node in two days. The lower bounds for the maximal eye are derived by stopping the algorithm before it terminates. It can be observed that the ISPR method can largely reduce the size of the graph, and therefore can accelerate the computing process. Moreover, the values of α_1 and α_2 are small compared to the maximal eye of the reduced graph. Hence, the approximation in equation (6.5) holds and the maximal eye of the original graph is equal to the maximal eye of the reduced graph. Although the algorithm achieves acceleration compared to the brute-force search method, we are only able to compute the maximal eye for graphs with up

Power Network	**Original Size**	**Reduced Size**	α_1	α_2	e_R
Case 14	(14,20)	(2,1)	6	3	0
Case 30	(30,41)	(8,13)	4	3	8
Case 39	(39,46)	(8,12)	4	5	8
Case 57	(57,78)	(22,39)	4	-	23
Case 118	(118,179)	(44,83)	5	-	13
Case 300	(300,409)	(109,196)	8	4	$\geq$10
Case 1354	(1354,1710)	(263,500)	9	8	TLE
Case 2383	(2383,2886)	(499,949)	11	5	TLE

Table 6.6.1: Comparison of graph sizes before and after the ISPR method for maximal eye. Number of vertices and edges before and after the ISPR method for maximal eye along with values computed during the reduction process.

to 118 vertices. Note that since graph problems have exponential complexities, solving them for graphs having as low as 200 nodes is still beyond the current computational capabilities. However, this does not undermine the usefulness of the introduced graph parameters, since it is shown in this chapter that those parameters accurately decide whether the power flow problem has a unique solution.

Next, we consider the computation of the maximal girth. We use the same algorithms and the results are listed in Table 6.6.2. In this case, it can be observed that α_2 is equal to 3 for large power networks. This is because the underlying graphs of large power networks considered in the table have "pendant triangles". Pendant triangles are triangles that have only one vertex connected to the rest of the graph. Furthermore, the approximation in Theorem 72 holds and the maximal girth of the original graph is equal to $\alpha_2 = 3$. Hence, the maximal girth can be computed during the reduction process. This shows that the conditions for the weak uniqueness is significantly loose and requiring $\omega_{k\ell}$ to be at most $2\pi/3$ for all edges $\{k, \ell\}$ is enough. However, for 2-vertex-connected SP graphs, we have shown that the maximal girth is equal to the maximal eye and the requirement for the weak uniqueness is the same as that for the strong uniqueness.

Verification of Corollary 8

In this subsection, we validate the results in Corollary 8, i.e., showing that there does not exist a different solution in the monotone regime with the set of allowable perturbations being $\mathcal{W}_{2\pi/\mathrm{e}(\mathcal{G})}$.

A random power flow set point is generated by first choosing a random vector of voltages. The voltage magnitudes and angles are randomly sampled from a uniform distribution

Power Network	**Original Size**	**Reduced Size**	α_1	α_2	g_R
Case 14	(14,20)	(2,1)	6	3	0
Case 30	(30,41)	(9,14)	4	3	3
Case 39	(39,46)	(10,14)	4	3	3
Case 57	(57,78)	(22,39)	4	-	23
Case 118	(118,179)	(44,83)	5	-	4
Case 300	(300,409)	(110,197)	8	3	$\geq$7
Case 1354	(1354,1710)	(271,509)	9	3	$\geq$3
Case 2383	(2383,2886)	(500,950)	11	3	$\geq$3

Table 6.6.2: Number of vertices and edges before and after the ISPR method for maximal girth along with values computed during the reduction process.

ranging from user-set min/max values:

$$|v_i^0| \sim \mathcal{U}(V_{min}, V_{max}) \text{ for all } i \in \mathbb{V},$$
$$|\Theta_i^0| \sim \mathcal{U}(\Theta_{min}, \Theta_{max}) \text{ for all } i \in \mathbb{V},$$

where $\mathcal{U}(a, b)$ is the uniform distribution on $[a, b]$. The voltage angles are rejected and discarded if they do not belong to the monotone regime. A new random sample is chosen until the angles belong to the the monotone regime. Finally, once we have a voltage profile belonging to the monotone regime, we use the information to calculate the real power injections, namely P^0. The values of $|v^0|$ and P^0 are provided as an input to the power flow algorithm. Note that Θ^0 is always a solution to the P-Θ problem $\hat{P}(\Theta) = P^0$. In this sense, we refer to Θ^0 the ground truth solution. There are usually other solutions and the goal of this experiment is to analyze where those other solutions are situated with respect to the ground truth solution.

In order to explore different parts of the solution space, we randomly sample an initial point around the ground truth Θ^0 and feed it into MATPOWER. The current setting is to consider a normal distribution around the ground truth, with some specified standard deviation. Intuitively, if the random initial point is close enough to the ground truth solution, then the algorithm will converge to the ground truth solution. However, if we start the algorithm with a suitably far initial point, then the power flow algorithms may converge to a different solution. Note that initializing too far away can lead to convergence issues of the algorithm.

Next, we define a metric that can capture the distance between two solutions to the P-Θ problem. Consider a solution of the P-Θ problem, Θ^i, where i corresponds to the random initialization number ($i \in \mathcal{R} := \{1, \ldots, 10,000\}$). Let Θ_k^i denote the voltage angle at bus k for the i-th experiment. We define dist(Θ^i) to be the distance between the particular

Power Networks	dist^m	$2\pi/\text{e}$
Case 14	∞	∞
Case 30	71.8	45
Case 39	53.8	45
Case 57	37.8	15.7
Case 118	66.1	27.7

Table 6.6.3: Distance measure for different test cases.

solution Θ^i and the ground truth solution, characterized in terms of their angle differences:

$$\text{dist}(\Theta^i) := \max_{\{k,\ell\}\in\mathbb{E}} |\Theta^i_{k\ell} - \Theta^0_{k\ell}|.$$

Now, define $\text{dist}^m(\mathbb{G})$ to be the smallest nonzero distance among all solutions in the monotone region for a given power system $\mathbb{G}$. More concretely, we let the symbol $\mathcal{M}$ represent the set of indices i such that the solution Θ^i belongs to the monotone region defined in the paper and define

$$\text{dist}^m(\mathbb{G}) := \min_{i\in\mathcal{M}\cap\mathcal{R}} \text{dist}(\Theta^i) \quad \text{s.t. } \text{dist}(\Theta^i) \neq 0.$$

As a specific scenario, we consider the case when all the line properties are the same and the voltage magnitudes are fixed to be one. In other words, $V_{max} = V_{min} = 1$. Furthermore, the lines are close to being lossless. We note that when we experimented with significantly lossy lines, different solutions were not found within the monotone region. This is because the monotone regime is small when the lines are very lossy.

The values of $\text{dist}(\Theta^i)$ and dist^m are calculated for different networks and are summarized in Table 6.6.3. The results in the table are two-folds. First, the distance dist^m provides an upper bound on the allowable perturbations such that the solution is strongly unique. On the other hand, the results in the last column are the theoretical lower bound on the allowable perturbations to guarantee the strong uniqueness. We can see that the numerical results verify our theoretical findings, although there exists a gap between the maximal possible allowable perturbations that ensure the uniqueness and the bounds obtained from our theoretical results.

Implication for AC Power Flow Problem

The P-Θ problem discussed in this chapter makes the assumption that all buses are PV buses. In order to show the connection between the P-Θ problem and the full AC power flow problem, we numerically demonstrate the proximity of the power flow solutions under the two problem settings. A random set point is generated, as we did in the previous subsection, by producing a random voltage profile $(|v^0|, \Theta^0)$ and computing the corresponding real/reactive

powers. Then, this set point is utilized as input parameters to solve for the AC power flow problem (without the assumption that all buses are PV buses) with random initial points around $|v^0|, \Theta^0$. Note that $|v^0|, \Theta^0$ (call it the reference solution) obviously comprise one solution to the AC power flow problem, but there are potentially other solutions that satisfy the AC power flow equations. Let us define the distance between two solutions as we did in the previous subsection. Figure 6.6.1 shows that none of the other solutions are within the allowable perturbation bound obtained in Corollary 8 when compared to the reference solution. Furthermore, the voltage magnitude distance shows that these are unrealistic solutions, since voltage magnitudes are usually maintained to be within 5 percent of the nominal value. Similar experiments conducted for various set points and all the power networks mentioned in Table 6.6.3 lead to the same results.

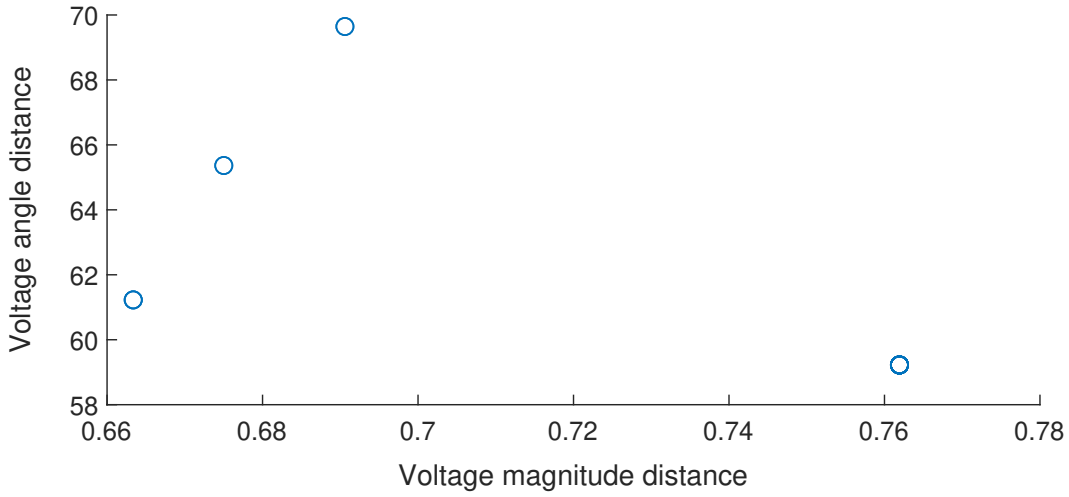


Figure 6.6.1: Distance between AC power flow solutions for IEEE-39.

Appendix

6.A Algorithms for Computing the Maximal Girth and Eye

In the appendix, we propose search-based algorithms for computing the maximal eye and the maximal girth. Our approach is based on the Depth-First Search (DFS) method and utilized the pruning technique to accelerate the computing process. We first describe a common sub-procedure that will be used in both algorithms. The sub-procedure computes the minimal directed chordless cycle containing a given edge. Given a truncation length $T \geq 1$, the sub-procedure returns the truncation length if there does not exist a directed chordless cycle that contains the given edge and has length at most T. The sub-procedure is also based on the DFS method with pruning and borrows the idea of *blocking* from [215] to accelerate the searching process. The pseudo-code of the sub-procedure is listed in Algorithm 18.

The search space of the sub-procedure is the set of directed chordless paths with length at most T. When the current directed chordless path is a directed chordless cycle, the length of the cycle is recorded and the minimal length of known directed chordless cycles is updated. By searching over all chordless paths, we find the length of the minimal directed chordless cycle. The DFS method is initialized with the given edge, denoted as (k, ℓ), and extends the directed chordless path by adding a neighbouring vertex of the end point other than k to the path. The pruning technique becomes effective and delete the end point other than k from the path if one of the following cases occurs:

- The length of the directed chordless path is larger than T or the known minimal length of directed chordless cycles;
- All neighbouring vertices have been searched or will introduce a chord if added to the path.

Using the idea of blocking, one can efficiently check whether adding a vertex to the path will introduce a chord. This approach is based on the following observation: if the path $(k_1, \ldots, k_t)$ is chordless, then any vertex k_s can only be in the neighborhood of k_{s-1}, k_{s+1}. We construct an array and, for each vertex, we record the number of vertices on the path that are in the neighborhood of the vertex. The array is updated whenever the path is

updated. If there are at least two vertices on the path in the neighbourhood of a vertex not on the path, then adding the vertex to the path will introduce a chord. Hence, the cost of checking this condition for each potential vertex not on the path is a single evaluation of an array.

Next, we propose the algorithms for computing the maximal eye and the maximal girth. Since the algorithm of maximal girth is similar to the algorithm for maximal eye, we only discuss the algorithm for computing the maximal eye. The algorithm is also based on the DFS method with pruning, and the pseudo-code is provided in Algorithm 20. We first order all edges and gradually assign one of the directions $\{0, -1, +1\}$ to each edge following the ordering of the edges. The search space consists of the orientations for the first several edges (intermediate states) and the orientations for the entire graph (final states). One can verify that all intermediate states and final states form a trinomial[2] tree, since each orientation for the first $k < |\mathbb{E}|$ edges leads to three different orientations for the first $k + 1$ edges. Then, the algorithm for computing the maximal eye searches in the same way as the classical DFS method on a directed tree. For each node, we consider the sub-graph consisting of those edges that have been assigned a direction. We compute the length of the minimal directed chordless cycle in the sub-graph, which contains the last edge in the sub-graph, using the sub-procedure (Algorithm 18). The truncation length can be decided as follows. Since a DFS method is implemented on a trinomial tree, there exists a directed path from the root node of the trinomial tree to the current node. The truncation length can be chosen as the minimal length computed on the preceding nodes of the path. When the search reaches a leaf node, we obtain an orientation for the entire graph, and the size of the eye becomes the minimal length on the path to the root node. By searching over all leaf nodes, we find the maximal eye. Similarly, one can use the pruning technique to reduce the search space. The current node is pruned if it can not be extended to a weakly feasible orientation for the entire graph, or the size of the eye of the sub-graph is smaller than the known maximal size of the eye.

Algorithm 18 Truncated Minimal Chordless Cycle

Input: Directed weighted graph $(\mathbb{V}, \mathbb{E}, W)$, selected edge (k, ℓ), truncation length T
Output: Length of minimal chordless cycle c

Construct the neighbourhood of each vertex $N : \mathbb{V} \mapsto 2^{\mathbb{V}}$.
Initialize blocked array $block[i] \leftarrow 0$ for all vertices $i \in \mathbb{V}$.
Set the length of minimal cycle recorded $c \leftarrow T$.
Set current length $L_{cur} \leftarrow W_{k\ell}$.
Set the path $P \leftarrow [k, \ell]$.
Set $block[k] \leftarrow 1, block[\ell] \leftarrow 1$.
if $L_{cur} \geq T$ **then** ▷ Already longer than truncation length
  return c

[2]A directed tree is called a trinomial tree if there is a root node and each non-leaf node has exactly three descendant nodes.

end if
while the length of P is at least 2 **do**
Get the endpoint $i \leftarrow P[-1]$.
Increase *block* for vertices in $N[j]$ by 1.
Get the minimal vertex $j \in N[i]$ such that $block[j] \leq 1$ and $L_{cur} + W_{P[-1]j} < v$.
if no such vertex j exists **then**
▷ Recursion: no next unblocked vertex
Find the maximal index h such that $P[h] \notin \{k, \ell, i\}$ and $P[h+1]$ is not the maximal vertex in $N[P[h]]$.
if no such h exists **then** ▷ Search finished
break
else
Remove $P[h+1], \ldots, P[-1]$ from path P.
Decrease *block* of $N[P[h]], \ldots, N[P[-1]]$ by 1.
Add the next smallest vertex in $N[P[h]]$ to P.
Update L_{cur} to be the length of path P.
continue
end if
else ▷ Add a new vertex
Add vertex j to P and update L_{cur}.
if $k \in N[j]$ **then** ▷ find a cycle
Calculate length $c_{cur} \leftarrow L_{cur} + W_{jk}$.
if $c_{cur} > 0$ **then**
Update $c \leftarrow \min\{c, c_{cur}\}$.
end if
Recursion similarly as above.
else
continue
end if
end if
end while
return c

Algorithm 19 Algorithm for Computing the Maximal Girth

Input: Undirected weighted graph $(\mathbb{V}, \mathbb{E}, W)$, slack bus k
Output: Maximal girth g

Set the maximal girth $\mathrm{g} \leftarrow 0$.
Assign an order to the set of edges $\mathbb{E}$ and denote edges as

$$\{k_1, \ell_1\}, \ \ldots, \ \{k_m, \ell_m\}.$$

Initialize the set of edges $\mathbb{E}_0 \leftarrow \{\{k_1, \ell_1\}\}$.
Initialize the set of orientations $A_{k_1,\ell_1} \leftarrow -1$.

loop
  Check the feasibility with current orientation.
  if feasibility fails **then** ▷ Recursion
    Get the maximal index j such that $A_{k_j,\ell_j} \neq 1$.
    if no such j exists **then** ▷ Terminate the algorithm
      break
    else
      Remove $\{k_{j+1}, \ell_{j+1}\}, \ldots, \{k_m, \ell_m\}$ from $\mathbb{E}_0$.
      Change orientation $A_{k_j,\ell_j} \leftarrow -A_{k_j,\ell_j}$.
      continue
    end if
  end if
  Compute the girth g_{cur} under $\mathbb{E}_0$ and A using Algorithm 18. The truncation length is set to be the girth of the precedent state.
  if $\mathrm{g}_{cur} < \mathrm{g}$ **then** ▷ Smaller than known girth
    Recursion in the same way.
  end if
  Get the next edge $\{k_i, \ell_i\}$ that is not in $\mathbb{E}_0$.
  if no such edge **then** ▷ Leaf node reached
    Update $\mathrm{g} \leftarrow \max\{\mathrm{g}, \mathrm{g}_{cur}\}$.
    Recursion in the same way.
  else
    Add the next edge $\{k_i, \ell_i\}$ that is not in $\mathbb{E}_0$.
    Assign $A_{k_j,\ell_j} \leftarrow -1$.
    continue
  end if
end loop
return g

Algorithm 20 Algorithm for Computing the Maximal Eye

Input: Undirected weighted graph $(\mathbb{V}, \mathbb{E}, W)$, slack bus k
Output: Maximal eye e

Set the maximal eye e $\leftarrow 0$.
Assign an order to the set of edges $\mathbb{E}$ and denote edges as

$$\{k_1, \ell_1\}, \ \ldots, \ \{k_m, \ell_m\}.$$

Initialize the set of edges $\mathbb{E}_0 \leftarrow \{\{k_1, \ell_1\}\}$.
Initialize the set of orientations $A_{k_1,\ell_1} \leftarrow -1$.
loop
  Check the weak feasibility with current orientation.
  if weak feasibility fails **then** ▷ Recursion
    Get the maximal index j such that $A_{k_j,\ell_j} \neq 1$.
    if no such j exists **then** ▷ Terminate the loop
      break
    else
      Remove $\{k_{j+1}, \ell_{j+1}\}, \ldots, \{k_m, \ell_m\}$ from $\mathbb{E}_0$.
      Change orientation $A_{k_j,\ell_j} \leftarrow A_{k_j,\ell_j} + 1$.
      continue
    end if
  end if
  Compute the size of eye e_{cur} under $\mathbb{E}_0$ and A using Algorithm 18. The truncation length is set to be the size of eye of the precedent state.
  if $\mathrm{e}_{cur} < \mathrm{e}$ **then** ▷ Smaller than known size of eye
    Recursion in the same way.
  end if
  Get the next edge $\{k_i, \ell_i\}$ that is not in $\mathbb{E}_0$.
  if no such edge **then** ▷ Leaf node reached
    Update $\mathrm{e} \leftarrow \max\{\mathrm{e}, \mathrm{e}_{cur}\}$.
    Recursion in the same way.
  else
    Add the next edge $\{k_i, \ell_i\}$ that is not in $\mathbb{E}_0$.
    Assign $A_{k_j,\ell_j} \leftarrow -1$.
    continue
  end if
end loop
return e

6.B Proof for General Graphs

Proof of Lemma 47

Proof. We only prove the strong uniqueness part since the proof for weak uniqueness is similar. For a given power network, we define the real power flow along the line $\{k, \ell\} \in \mathbb{E}$ from bus k in the direction of bus ℓ as

$$\tilde{p}_{k\ell}(\Theta) := -G_{k\ell}|v_k||v_\ell|\cos(\Theta_{k\ell}) + B_{k\ell}|v_k||v_\ell|\sin(\Theta_{k\ell}).$$

By definition, it follows that

$$\hat{P}_k(\Theta) = \sum_{\ell:\{k,\ell\}\in\mathbb{E}} \tilde{p}_{k\ell}(\Theta), \quad \forall k \in \mathbb{V}.$$

Proof of sufficiency. We first show by contradiction that statement 2 of the lemma is sufficient for statement 1. In particular, suppose that statement 2 holds, but the solution is not strongly unique for some graph $\mathbb{G} \in \mathcal{G}$ and some real power injection P while problem (6.1) is feasible. Then, there exist two different phase angle vectors Θ^1, Θ^2 such that $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$ and $\hat{P}(\Theta^1) = \hat{P}(\Theta^2)$. For each line $\{k, \ell\} \in \mathbb{E}$, there exists a constant $C_{k\ell} > 0$ such that

$$B_{k\ell} = C_{k\ell}\sin(\gamma_{k\ell}), \quad G_{k\ell} = C_{k\ell}\cos(\gamma_{k\ell}).$$

We calculate the change of power flow from k to ℓ as

$$\begin{aligned}
&\tilde{p}_{k\ell}(\Theta^1) - \tilde{p}_{k\ell}(\Theta^2) \\
&= -G_{k\ell}|v_k||v_\ell|[\cos(\Theta^1_{k\ell}) - \cos(\Theta^2_{k\ell})] \\
&\qquad + B_{k\ell}|v_k||v_\ell|[\sin(\Theta^1_{k\ell}) - \sin(\Theta^2_{k\ell})] \\
&= -C_{k\ell}\cos(\gamma_{k\ell})|v_k||v_\ell|[\cos(\Theta^1_{k\ell}) - \cos(\Theta^2_{k\ell})] \\
&\qquad + C_{k\ell}\sin(\gamma_{k\ell})|v_k||v_\ell|[\sin(\Theta^1_{k\ell}) - \sin(\Theta^2_{k\ell})] \\
&= (-\cos(\gamma_{k\ell})[\cos(\Theta^1_{k\ell}) - \cos(\Theta^2_{k\ell})] \\
&\qquad + \sin(\gamma_{k\ell})[\sin(\Theta^1_{k\ell}) - \sin(\Theta^2_{k\ell})]) \cdot |v_k||v_\ell|C_{k\ell} \\
&= 2[\cos(\gamma_{k\ell})\sin(\Theta^1_{k\ell}/2 + \Theta^2_{k\ell}/2) \\
&\qquad + \sin(\gamma_{k\ell})\cos(\Theta^1_{k\ell}/2 + \Theta^2_{k\ell}/2)] \\
&\qquad \cdot \sin(\Delta_{k\ell}/2)|v_k||v_\ell|C_{k\ell} \\
&= 2\sin(\gamma_{k\ell} + \Theta^1_{k\ell}/2 + \Theta^2_{k\ell}/2) \cdot \operatorname{sign}(\sin(\Delta_{k\ell}/2)) \\
&\qquad \cdot |\sin(\Delta_{k\ell}/2)||v_k||v_\ell|C_{k\ell} \\
&:= \delta_{k\ell} \cdot |\sin(\Delta_{k\ell}/2)v_k v_\ell|C_{k\ell},
\end{aligned}$$

where

$$\Delta_{k\ell} := \Theta^1_{k\ell} - \Theta^2_{k\ell}, \tag{6.6}$$

$$\delta_{k\ell} := 2\sin(\gamma_{k\ell} + \Theta^1_{k\ell}/2 + \Theta^2_{k\ell}/2)\text{sign}(\sin(\Delta_{k\ell}/2)).$$

Note that the third equality in (6.6) is due to the following triangular identities:

$$\begin{aligned}\cos(\eta) - \cos(\varphi) &= -2\sin[(\eta-\varphi)/2]\sin[(\eta+\varphi)/2],\\ \sin(\eta) - \sin(\varphi) &= 2\sin[(\eta-\varphi)/2]\cos[(\eta+\varphi)/2].\end{aligned}$$

Since $\hat{P}_k(\Theta^1) = \hat{P}_k(\Theta^2)$ for all $k \neq 1$, we obtain

$$\begin{aligned}\hat{P}_k(\Theta^1) - \hat{P}_k(\Theta^2) &= \sum_{\ell:\{k,\ell\}\in\mathbb{E}} \left[\tilde{p}_{k\ell}(\Theta^1) - \tilde{p}_{k\ell}(\Theta^2)\right]\\ &= \sum_{\ell:\{k,\ell\}\in\mathbb{E}} \delta_{k\ell}\cdot|\sin(\Delta_{k\ell}/2)v_k v_\ell|C_{k\ell} = 0\end{aligned}$$

for all $k \neq 1$. Let the set $\mathbb{E}_0$ be the subset of edges such that $\Delta_{k\ell} \neq 0$ for all $\{k,\ell\} \in \mathbb{E}_0$; we assign an order to elements in $\mathbb{E}_0$. Define the matrix $M \in \mathbb{R}^{|\mathbb{V}|\times|\mathbb{E}_0|}$ and the vector $\mathbf{g} \in \mathbb{R}^{|\mathbb{E}_0|}$ as

$$M_{ki} := \delta_{k\ell}, \quad M_{\ell i} := \delta_{\ell k}, \quad \mathbf{g}_i := |\sin(\Delta_{k\ell}/2)v_k v_\ell|C_{k\ell},$$

where $\{k,\ell\}$ is the i-th edge in the set $\mathbb{E}_0$. Since $\Delta_{k\ell} \neq 0$ for all $\{k,\ell\} \in \mathbb{E}_0$ and $\Delta_{k\ell} \leq 2\gamma_{k\ell} \leq \pi$, it holds that

$$|\sin(\Delta_{k\ell}/2)| > 0, \quad \forall\{k,\ell\} \in \mathbb{E}_0.$$

Then, the vector $\mathbf{g}$ is a solution to the linear feasibility problem

$$\text{find } \mathbf{x} \in \mathbb{R}^{|\mathbb{E}_0|} \quad \text{s.t. } (M\mathbf{x})_{2:|V|} = 0, \quad \mathbf{x} > 0.$$

where $(y)_{i:j} := (y_i, y_{i+1}, \ldots, y_j)$ includes the i-th to the j-th entries of the vector y and inequality $x > 0$ means that $x_k > 0$ holds for all entries of the vector x. The notation $x \geq 1$ is defined in the same way. The above feasibility problem is equivalent to

$$\text{find } \mathbf{x} \in \mathbb{R}^{|\mathbb{E}_0|} \quad \text{s.t. } (M\mathbf{x})_{2:|V|} = 0, \quad \mathbf{x} \geq 1.$$

Then, by Farka's Lemma, the dual feasibility problem

$$\text{find } \mathbf{y} \in \mathbb{R}^{|\mathbb{V}|} \quad \text{s.t. } M^T\mathbf{y} \geq 0, \quad \mathbf{1}^T M^T\mathbf{y} > 0, \quad y_1 = 0$$

is infeasible. However, the conditions in the dual problem are the same as the conditions in statement 2 of Lemma 47. This contradicts the claim in statement 2 that there exists a vector $\mathbf{y}$ satisfying these conditions. Thus, statement 1 must hold true.

Proof of necessity. Next, we again show by contradiction that statement 2 of the lemma is necessary for statement 1. Assume that statement 1 holds true, and that there exist two different phase angle vectors Θ^1, Θ^2 in the monotone regime such that $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$ while there does not exist $\mathbf{y}$ satisfying the conditions in statement 2. We define $\mathbb{E}_0$ as the set of edges such that $\Delta_{k\ell} \neq 0$, where $\Delta_{k\ell} := \Theta^1_{k\ell} - \Theta^2_{k\ell}$ for all $\{k, \ell\} \in \mathbb{E}_0$. We construct the matrix $M \in \mathbb{R}^{|\mathbb{V}|\times|\mathbb{E}_0|}$ as

$$M_{ki} := \delta_{k\ell}, \quad M_{\ell i} := \delta_{\ell k},$$

where $\{k, \ell\}$ is the i-th edge in the set $\mathbb{E}_0$ and

$$\delta_{k\ell} := \sin(\gamma_{k\ell} + \Theta^1_{k\ell}/2 + \Theta^2_{k\ell}/2)\text{sign}(\sin(\Delta_{k\ell}/2)).$$

By the same analysis, the conditions in statement 2 turn out to be equivalent to the feasibility of the linear feasibility problem

$$\text{find } \mathbf{y} \in \mathbb{R}^{|\mathbb{V}|} \quad \text{s. t. } M^T\mathbf{y} \geq 0, \quad \mathbf{1}^T M^T\mathbf{y} > 0, \quad y_1 = 0.$$

By our assumption, the above problem is infeasible. By Farka's Lemma, there exists a solution $\mathbf{g} \in \mathbb{R}^{|\mathbb{E}_0|}$ to the feasibility problem

$$\text{find } \mathbf{x} \in \mathbb{R}^{|\mathbb{E}_0|} \quad \text{s. t. } (M\mathbf{x})_{2:|V|} = 0, \quad \mathbf{x} \geq 1$$

and also to the feasibility problem

$$\text{find } \mathbf{x} \in \mathbb{R}^{|\mathbb{E}_0|} \quad \text{s. t. } (M\mathbf{x})_{2:|V|} = 0, \quad \mathbf{x} > 0.$$

We define the matrix $C \in \mathbb{R}^{|\mathbb{V}|\times|\mathbb{V}|}$ as

$$C_{k\ell} := |\sin(\Delta_{k\ell}/2)v_k v_\ell|^{-1}\mathbf{g}_i, \quad \forall\{k, \ell\} \in \mathbb{E}_0,$$

where $\{k, \ell\}$ is the i-th edge in the set $\mathbb{E}_0$, and

$$C_{k\ell} := 1, \quad \forall\{k, \ell\} \in \mathbb{E}\backslash\mathbb{E}_0, \quad C_{k\ell} := 0, \quad \forall\{k, \ell\} \notin \mathbb{E}.$$

By the definition, it follows that $C_{k\ell} > 0$ for all $\{k, \ell\} \in \mathbb{E}$. We construct a graph $\mathbb{G}$ with the complex admittance matrix

$$Y_{k\ell} := C_{k\ell}\cos(\gamma_{k\ell}) - \mathbf{j}C_{k\ell}\sin(\gamma_{k\ell}), \quad \forall\{k, \ell\} \in \mathbb{E}.$$

Then, for all $k \neq 1$, we have

$$\hat{P}_k(\Theta^1) - \hat{P}_k(\Theta^2) = \sum_{\ell:\{k,\ell\}\in\mathbb{E}} \left[\tilde{p}_{k\ell}(\Theta^1) - \tilde{p}_{k\ell}(\Theta^2)\right]$$

$$= \sum_{\ell:\{k,\ell\}\in\mathbb{E}} \delta_{k\ell} \cdot |\sin(\Delta_{k\ell}/2)v_k v_\ell| C_{k\ell} = (M\mathbf{g})_k = 0.$$

This implies that Θ^1 and Θ^2 are both solutions to problem (6.1) in the monotone regime when the real power injection is

$$P := \hat{P}(\Theta^1).$$

This contradicts statement 1 that the solution is strongly unique for any real power injection. Hence, the conditions in statement 2 must be satisfied. □

Proof of Lemma 48

Proof. We only prove the strong uniqueness part since the proof for weak uniqueness is similar. Since the induced orientation A is not a weakly feasible orientation, there exists a vertex $i \neq 1$ such that it has nonzero out-degree and zero in-degree, or it has nonzero in-degree and zero out-degree. Without loss of generality, assume that the vertex i has nonzero out-degree and zero in-degree. We prove that the i-th unit vector $\mathbf{y} := \mathbf{e}_i$ satisfies the conditions in statement 1 of Lemma 47. It is straightforward that $y_1 = 0$. We only need to show that the inequalities in (6.2) hold and at least one of them is strict. We consider any edge (k, ℓ) such that $\Delta_{k\ell} > 0$. First, if $k \neq i$ and $\ell \neq i$, then both sides of the inequality (6.2) are zero. Next, if $k \neq i$ and $\ell = i$, then the condition $\Delta_{ki} > 0$ implies that $A_{ki} = +1$, which contradicts the assumption that i has zero in-degree. Finally, if $k = i$ and $\ell \neq i$, the goal is to prove that

$$\begin{aligned}\sin(\gamma_{i\ell} + \Theta^1_{i\ell}/2 + \Theta^2_{i\ell}/2) \cdot y_i \\ > \sin(\gamma_{i\ell} - \Theta^1_{i\ell}/2 - \Theta^2_{i\ell}/2) \cdot y_\ell.\end{aligned}$$

Since $y_i = 1$ and $y_\ell = 0$, the above inequality is equivalent to

$$\sin(\gamma_{i\ell} + \Theta^1_{i\ell}/2 + \Theta^2_{i\ell}/2) > 0.$$

Recalling the assumption that $\Theta^1_{i\ell}$ and $\Theta^2_{i\ell}$ are in the monotone regime $[-\gamma_{i\ell}, \gamma_{i\ell}]$, one can write

$$\gamma_{i\ell} + \Theta^1_{i\ell}/2 + \Theta^2_{i\ell}/2 \in [0, 2\gamma_{i\ell}] \subset [0, \pi].$$

Hence, it is enough to show that

$$\gamma_{i\ell} + \Theta^1_{i\ell}/2 + \Theta^2_{i\ell}/2 \in (0, 2\gamma_{k\ell}) \subset (0, \pi).$$

If $\gamma_{i\ell} + \Theta^1_{i\ell}/2 + \Theta^2_{i\ell}/2 = 0$, then it holds that

$$\Theta^1_{i\ell} = \Theta^2_{i\ell} = -\gamma_{i\ell}.$$

This contradicts the inequality $\Delta_{i\ell} = \Theta^1_{i\ell} - \Theta^2_{i\ell} > 0$. If $\gamma_{i\ell} + \Theta^1_{i\ell}/2 + \Theta^2_{i\ell}/2 = 2\gamma_{k\ell}$, then it holds that

$$\Theta^1_{i\ell} = \Theta^2_{i\ell} = \gamma_{i\ell},$$

which also contradicts the inequality $\Delta_{i\ell} > 0$. Combining the two cases, we obtain that $\sin(\gamma_{i\ell} + \Theta^1_{i\ell}/2 + \Theta^2_{i\ell}/2) > 0$ and the inequality

$$\begin{aligned}\sin(\gamma_{i\ell} + \Theta^1_{i\ell}/2 + \Theta^2_{i\ell}/2) \cdot y_i \\ > \sin(\gamma_{i\ell} - \Theta^1_{i\ell}/2 - \Theta^2_{i\ell}/2) \cdot y_\ell.\end{aligned}$$

holds strictly. It follows that $\mathbf{y} = \mathbf{e}_i$ satisfies the conditions in statement 2 of Lemma 47. □

Proof of Theorem 65

Proof. Suppose that Θ^1 and Θ^2 are in the monotone regime and Θ^2 is in the neighbourhood of Θ^1. Then, the condition in the theorem implies that $\Theta^1 - \Theta^2$ is not weakly feasible. By Lemma 48, there exists a vector $y \in \mathbb{R}^{|\mathbb{V}|}$ such that statement 2 in Lemma 47 is satisfied. Since this is true for every Θ^2, Lemma 47 implies that the solution to problem 6.1 is strongly unique. □

Proof of Corollary 7

Proof. We only prove the strong uniqueness part since the proof for weak uniqueness is similar. Suppose that Θ^1 and Θ^2 are two solutions to problem (6.1) in the monotone regime such that $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$. Using the results of Theorem 65, we only need to show that the induced orientation of $\Theta^1 - \Theta^2$ is not weakly feasible. Assume conversely that the induced orientation A is a weakly feasible orientation. Then, by hypothesis, there exists a directed cycle $(k_1, \ldots, k_t)$ containing at least one normal edge such that

$$\sum_{k_i k_{i+1} \text{ is normal}} \omega_{k_i k_{i+1}} < 2\pi, \tag{6.7}$$

where $k_{t+1} := k_1$. We denote $\Delta_{k\ell} := \Theta^1_{k\ell} - \Theta^2_{k\ell}$ and it follows that

$$\begin{aligned} 0 < \Delta_{k_i k_{i+1}} \leq \omega_{k_i k_{i+1}} \quad & \forall i \text{ s. t. } \{k_i, k_{i+1}\} \text{ is normal}, \\ \Delta_{k_i k_{i+1}} = 0 \quad & \forall i \text{ s. t. } \{k_i, k_{i+1}\} \text{ is not normal}, \end{aligned} \tag{6.8}$$

where the right part of the first inequality is because $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$. Combining inequalities (6.7) and (6.8) yields that

$$\begin{aligned} 0 < \sum_{i=1}^{t} \Delta_{k_i k_{i+1}} &= \sum_{k_i k_{i+1} \text{ is normal}} \Delta_{k_i k_{i+1}} \\ &\leq \sum_{k_i k_{i+1} \text{ is normal}} \omega_{k_i k_{i+1}} < 2\pi. \end{aligned} \tag{6.9}$$

However, by the definition of $\Delta_{k\ell}$ and $\Theta_{k\ell}$, one can write

$$\begin{aligned} \sum_{i=1}^{t} \Delta_{k_i k_{i+1}} &= \sum_{i=1}^{t} \Theta^1_{k_i k_{i+1}} - \sum_{i=1}^{t} \Theta^2_{k_i k_{i+1}} \\ = \sum_{i=1}^{t} \left[\Theta^1_{k_i} - \Theta^1_{k_{i+1}}\right] &- \sum_{i=1}^{t} \left[\Theta^2_{k_i} - \Theta^2_{k_{i+1}}\right] = 0, \end{aligned}$$

where the second last equality is the congruence relation module 2π and the last equality is because $(k_1, \ldots, k_t)$ is a cycle. This contradicts equation (6.9). Thus, the induced orientation is not a weakly feasible orientation and the strong uniqueness holds. □

Proof of Corollary 8

Proof. Suppose that condition (6.3) is satisfied. Then, for every directed cycle $(k_1, \ldots, k_t)$ containing at least one normal edge, it must hold that $t \leq \mathrm{e}(\mathcal{G})$. Therefore, using condition (6.3), we know

$$\sum_{\{k_i, k_{i+1}\} \text{ is normal}} \omega_{k_i k_{i+1}} < 2\pi.$$

By Corollary 7, we obtain the desired conclusion. □

Proof of Theorem 66

Proof. To prove the first inequality, we only need to notice that any feasible orientation is also a weakly feasible orientation and the size of eye is equal to the girth when all edges are normal.

Then, we consider the second inequality. Assume conversely that the maximal eye is attained by a directed cycle with chords in the weakly feasible orientation A. Without loss of generality, assume that the directed cycle $(1, \ldots, t)$ attains the maximal eye *with fewest chords*, where $t \geq \mathrm{e}(\mathbb{G})$ and $\{1, i\} \in \mathbb{E}$ is a chord for some $i \in \{3, \ldots, t-1\}$. We consider four different cases:

1. $A_{1,i} = 0$: Consider the directed cycle

   $$(1, i, i+1, \ldots, t),$$

   which has at most $\mathrm{e}(\mathbb{G})$ normal edges and strictly fewer chords than $(1, \ldots, t)$. This contradicts the assumption that the cycle $(1, \ldots, t)$ is a directed cycle that attains the size of eye with fewest chords.

2. $A_{1,i} = +1$: and there exists at least one normal edge among $\{1, 2\}, \ldots, \{i-1, i\}$: The directed cycle

   $$(1, i, i+1, \ldots, t)$$

   has at most $\mathrm{e}(\mathbb{G})$ normal edges and strictly fewer chords than $(1, \ldots, t)$. This also contradicts the assumption on $(1, \ldots, t)$.

3. $A_{1,i} = +1$ and edges $\{1, 2\}, \ldots, \{i-1, i\}$ are not normal: Consider the directed cycle

   $$(1, i, i-1, \ldots, 2),$$

   which has exactly one normal edge and strictly fewer chords. By the definition of the maximal eye, we know $\mathrm{e}(\mathbb{G}) \geq 1$ and the cycle $(1, i, i-1, \ldots, 2)$ has at most $\mathrm{e}(\mathbb{G}) \geq 1$ normal edges. Hence, this contradicts the assumption on $(1, \ldots, t)$.

4. $A_{1,i} = -1$: Consider the orientation $\tilde{A}$ defined as

$$\tilde{A}_{k\ell} := -A_{k\ell}, \quad \forall \{k, \ell\} \in \mathbb{E}$$

and use the discussion in the first three cases.

Combining the above four cases concludes that the maximal eye of the power network $\mathbb{G}$ must be attained by a chordless cycle. Hence, the maximal eye is upper bounded by the longest chordless cycle. □

6.C Proof for Three Special Cases

Proof of Lemma 49

Proof. By the definition of strong uniqueness and weak uniqueness, if a solution to problem (6.1) is strongly unique, than it is also weakly unique. We only need to consider the other direction. Assume conversely that there exists a solution Θ^1 in the monotone regime that is weakly unique but not strongly unique. Then, there exists another solution $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$ that is different from Θ^1 according to Definition 22. Then, the phase difference of some line is different for the two solutions. Considering the power injection balance at each bus, we know that the phase difference is different at all lines. This means that the two solutions Θ^1 and Θ^2 are different according to Definition 21, which contradicts the assumption that Θ^1 is weakly unique. □

Proof of Theorem 67

Proof. The sufficient part is proved in Corollary 7 and we only prove the necessary part. In this proof, bus $n+1$ is defined as bus 1. We assume that

$$\sum_{i=1}^{n} \omega_{i,i+1} \geq 2\pi$$

We construct a power network $\mathbb{G} \in \mathcal{G}$ and power injection P such that there exist two different solutions Θ^1, Θ^2 in the monotone regime and $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$. Without loss of generality, assume that

$$\sum_{i=1}^{n} \omega_{i,i+1} = 2\pi.$$

This is because the construction for

$$\tilde{W} := \left\{ \frac{2\pi}{\sum_{j=1}^{n} \omega_{j,j+1}} \cdot \omega_{i,i+1} : i \in [n] \right\}.$$

also works for the original $W = \{\omega_{i,i+1} : i \in [n]\}$ if $\sum_{j=1}^{n} \omega_{j,j+1} \geq 2\pi$. We define two phase angle vectors as

$$\Theta_1^1 := 0, \quad \Theta_i^1 := \sum_{j=2}^{i} \omega_{j,j+1}, \quad \forall i \in \{2, \ldots, n\},$$
$$\Theta_i^2 := 0, \quad \forall i \in [n].$$

Then, it follows that

$$\Theta_{i,i+1}^1 = \omega_{i,i+1}, \quad \Theta_{i,i+1}^2 = 0, \quad \forall i \in [n],$$

which means that Θ^1 and Θ^2 are both in the monotone regime. Since $\omega_{i,i+1}, \gamma_{i,i+1} \in (0, \pi/2]$, we know that $\gamma_{i,i+1} + \omega_{i,i+1} \in (0, \pi]$ and therefore, by the monotonicity of $\cos(\cdot)$ in $[0, \pi]$, we have

$$\cos(\gamma_{i,i+1} + \omega_{i,i+1}) < \cos(\gamma_{i,i+1}).$$

For each line $\{i, i+1\}$, we define the positive constant

$$C_{i,i+1} := |v_i v_{i+1}|^{-1} \left[-\cos(\gamma_{i,i+1} + \omega_{i,i+1}) + \cos(\gamma_{i,i+1})\right]^{-1}$$

and the complex admittance

$$B_{i,i+1} := \sin(\gamma_{i,i+1}) C_{i,i+1}, \quad G_{i,i+1} := \cos(\gamma_{i,i+1}) C_{i,i+1}.$$

We use $\tilde{p}_{i,i+1}(\Theta)$ to denote the real power flow from bus i to bus $i+1$ given the phase angle vectors Θ. Then, we can calculate that

$$\begin{aligned}
&\tilde{p}_{i,i+1}(\Theta^1) - \tilde{p}_{i,i+1}(\Theta^2) \\
=& -G_{i,i+1}|v_i v_{i+1}|[\cos(\Theta_{i,i+1}^1) - \cos(\Theta_{i,i+1}^2)] \\
&\qquad + B_{k\ell}|v_i v_{i+1}|[\sin(\Theta_{i,i+1}^1) - \sin(\Theta_{i,i+1}^2)] \\
=& -\cos(\gamma_{i,i+1}) C_{i,i+1}|v_i v_{i+1}|[\cos(\omega_{i,i+1}) - 1] \\
&\qquad + \sin(\gamma_{i,i+1}) C_{i,i+1}|v_i v_{i+1}| \sin(\omega_{i,i+1}) \\
=& C_{i,i+1}|v_i v_{i+1}| \cdot [-\cos(\gamma_{i,i+1} + \omega_{i,i+1}) + \cos(\gamma_{i,i+1})] \\
=& 1.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\hat{P}_i(\Theta^1) - \hat{P}_i(\Theta^2) \\
=& \left[\tilde{p}_{i-1,i}(\Theta^1) - \tilde{p}_{i-1,i}(\Theta^2)\right] - \left[\tilde{p}_{i,i+1}(\Theta^1) - \tilde{p}_{i,i+1}(\Theta^2)\right] \\
=& 1 - 1 = 0.
\end{aligned}$$

If we choose $P := \hat{P}(\Theta^1)$, then Θ^1 and Θ^2 are two different solutions to problem (6.1) in the monotone regime such that $\Theta^2 \in \mathcal{N}(\mathbb{G}, \Theta^1, \mathcal{W})$ and that the strong uniqueness does not hold.

□

Proof of Lemma 50

Proof. For the notational simplicity, we denote the maximal eye and the maximal girth of the graph $(\mathbb{V}, \mathbb{E}, W)$ as e and g, respectively. Since the graph is 2-vertex-connected, there does not exist a degree-1 vertex. By Lemmas 51 and 52, Type II Operations do not change the maximal eye and the maximal girth of the graph. Moreover, the graph has a nested ear decomposition $\{L_0, L_1, \ldots, L_{r-1}\}$ by Theorem 68. Hence, we can assume that there is no degree-2 vertex except the slack bus. Assume conversely that graph $(\mathbb{V}, \mathbb{E}, W)$ is the 2-vertex-connected SP graph with minimal number of ears such that $\mathrm{e} > \mathrm{g}$. We will show that there must exist another graph with fewer ears in the ear decomposition and $\mathrm{e} > \mathrm{g}$. This will lead to a contradiction with our assumption that this graph has the minimal number of ears. If the graph has at most two ears, then the graph is a single line of a cycle and we know $\mathrm{e} = \mathrm{g}$. Hence, there exist at least three ears in the graph $(\mathbb{V}, \mathbb{E}, W)$.

Step 1. In this step, we prove that the graph has a pair of parallel edges that contains a leaf ear, which we will define below. Since a nested ear decomposition is also a tree decomposition, we can assign a directed tree structure to ears in the decomposition. Here, we call an ear L_k a **descendant ear** of L_ℓ if L_k is a descendant node of L_ℓ on the directed tree, or equivalently, both endpoints of ear L_k are on L_ℓ and at least one of them is different from the endpoints of L_ℓ. We also call ear L_ℓ the **precedent ear** of L_k. For any ear L_ℓ, we say that ear L_k is a **smallest descendant ear** of L_ℓ if L_k is a descendant ear of L_ℓ and there does not exist another ear L_i such that L_i is also a descendant ear of L_ℓ and the interval formed by the endpoints of L_i on L_ℓ is a strict subset of the interval formed by the endpoints of L_k. We note that each ear may have multiple smallest descendant ears. We say that an ear is a **leaf ear** if it is the smallest descendant ear of some ear and has no descendant ear. We denote the set of leaf ears as $\mathcal{L}$. Considering the directed tree structure of the ear decomposition, we know that the set $\mathcal{L}$ is not empty.

Suppose that L_k is a leaf ear with the endpoints k_1, k_2 and that L_ℓ is the precedent ear of L_k. Since we have deleted all degree-2 vertices except the slack bus, ear L_k is either a single line $\{k_1, k_2\}$ or two edges $\{k_1, k_3\}$ and $\{k_2, k_3\}$ connecting the endpoints to the slack bus k_3. Similarly, the path connecting the two endpoints of L_k on the precedent ear L_ℓ, which we denote as P_k, is either a single line or contains the slack bus. Considering the ear L_k and the path P_k, there are two cases: two parallel edges with endpoints $\{k_1, k_2\}$, or one is a single line and the other is two edges with the slack bus. If the first case occurs, we have a pair of parallel edges containing a leaf ear. Now, we consider the second case. If we exchange the two paths, i.e., let P_k be a leaf ear and L_k be a path on the precedent ear, then the structure of nested ear decomposition is not changed. Hence, without loss of generality, assume that L_k is a single line and P_k contains the slack bus. If there exists an ear L_j different from L_ℓ that also contains leaf ears, then by the uniqueness of slack bus, the first case occurs for leaf ears on ear L_j.

Hence, we simply need to consider the case when L_ℓ is the only ear that contains leaf ears. We consider the root ear L_0. By the definition of tree ear decomposition, we know

that L_0 is a single line; let ℓ_1, ℓ_2 be the two endpoints of L_0. Since all vertices except the slack bus have degree at least 3 and the slack bus is not an endpoint of ears, both ℓ_1 and ℓ_2 have degree at least 3. This implies that the root ear L_0 has at least 2 descendant ears and all descendant ears have endpoints ℓ_1, ℓ_2. Let $L_{k_1}, L_{k_2}, \ldots, L_{k_m}$ be the descendant ears of L_0. For each L_{k_i}, we define a sub-graph of $(\mathbb{V}, \mathbb{E}, W)$ consisting of ear L_0 and ears that are descendant nodes of L_{k_i} in the directed tree of ears. We can verify that each sub-graph also has a nested ear decomposition and therefore contains at least one leaf ear, which implies that ear L_ℓ belongs to all sub-graphs. On the other hand, due to the tree structure, the intersection of two different sub-graphs is ear L_0 and is not a leaf ear. Hence, the leaf ears in different sub-graphs are different and $L_\ell = L_0$. It follows that all descendant ears of L_0 are leaf ears and they form at least a pair of parallel edges containing a leaf ear.

Step 2. In this step, we construct a nested ear decomposition of the graph $(\mathbb{V}, \mathbb{E}, W)$ such that there exists a pair of parallel edges that contains the root ear L_0 and that all edges are ears in the ear decomposition. According to Step 1, there exists a pair of parallel edges that contains a leaf ear. We denote the leaf ear in the pair of parallel edges as L_k. We consider the (undirected) cycle containing L_0 and L_k. Suppose that the cycle has a non-empty edge intersection with ears $L_{k_0}, \ldots, L_{k_t}$, where $k_0 = 0, k_t = k$ and $L_{k_{s+1}}$ is a descendant ear of L_{k_s} for $s = 0, 1, \ldots, t-1$. Notice that the endpoints of each ear L_{k_s} are on the cycle. Now, we construct a new nested ear decomposition $\tilde{L}_0, \ldots, \tilde{L}_{m-1}$ such that $L_k = \tilde{L}_0$ is the root ear. We define $\tilde{L}_0 := L_k$ and $\tilde{L}_k$ as the remaining part of the cycle. For ears L_{k_s} with $1 \le s \le t-1$, we define $\tilde{L}_{k_s}$ as the ear L_{k_s} with edges on the cycle deleted. For ears that do not intersect with the cycle, we define $\tilde{L}_i := L_i$. It is desirable to show that with the new set of ears still forms a nested ear decomposition. To this end, we analyze three cases:

- **Case I.** First, it can be verified that ears $\tilde{L}_{k_1}, \ldots, \tilde{L}_{k_{t-1}}$ are nested ears on $\tilde{L}_{k_t}$. Hence, ears $\tilde{L}_{k_0}, \ldots, \tilde{L}_{k_t}$ still form a nesting structure.
- **Case II.** Next, we consider an ear $\tilde{L}_i = L_i$ that is not changed and has both endpoints on L_{k_s} for some $s \in \{0, 1, \ldots, t-1\}$. Since $L_{k_{s+1}}$ is a descendant ear on L_{k_s}, by the definition of nested ear decomposition, we know that the endpoints of $\tilde{L}_i$ are either both on $\tilde{L}_{k_s}$ or both on $\tilde{L}_{k_t}$. For the first case, L_i is an ear on $\tilde{L}_{k_s}$ and ears on $\tilde{L}_{k_s}$ have the same nesting structure as L_{k_s}. For the second case, both endpoints of L_k locate on $\tilde{L}_{k_t}$ and are nested between the endpoints of $\tilde{L}_{k_s}$ and $\tilde{L}_{k_{s-1}}$. We note that for the case when $s = 0$, both endpoints are equal to the endpoints of L_0 and they form the smallest possible interval on $\tilde{L}_{k_t}$. Hence, ears on $\tilde{L}_{k_t}$ also have a nested structure.
- **Case III.** Finally, we consider ears that are not changed and do not have endpoints on L_{k_s} for any $s = 0, \ldots, t$. These ears still form a nested structure on the original precedent ear and the nested ear decomposition structure is not changed.

Combining the above three cases concludes that the new set of ears is also a nested ear decomposition. Moreover, the topological structure of the graph is not changed. Hence, in

the new ear decomposition, the root ear $\tilde{L}_0 = L_k$ has parallel edges. Finally, we observe that the parallel edges of the root ear are also ears in the ear decomposition.

Step 3. Suppose that the maximal eye is achieved by the weakly feasible orientation A. In this step, we show that we can modify A such that each edge with direction 0 is incident to a degree-0 vertex and the size of eye is not changed. Here, the degree is calculated for the directed graph with orientation A and all edges with orientation 0 are not counted towards the degree. We define a partition of vertices as

$$\begin{aligned}\mathbb{V}_1 &:= \{k \in \mathbb{V} \mid \deg(k) > 0 \text{ or } k \text{ is the slack bus}\},\\ \mathbb{V}_2 &:= \{k \in \mathbb{V} \mid \deg(k) = 0 \text{ and } k \text{ is not the slack bus}\}\end{aligned}$$

and a partition of edges as

$$\begin{aligned}\mathbb{E}_1 &:= \{\{k,\ell\} \in \mathbb{E} \mid A_{k\ell} \in \{+1,-1\}\},\\ \mathbb{E}_2 &:= \{\{k,\ell\} \in \mathbb{E} \mid A_{k\ell} = 0,\ k \in \mathbb{V}_1 \text{ and } \ell \in \mathbb{V}_1\},\\ \mathbb{E}_3 &:= \{\{k,\ell\} \in \mathbb{E} \mid A_{k\ell} = 0,\ k \in \mathbb{V}_2 \text{ or } \ell \in \mathbb{V}_2\}.\end{aligned}$$

Then, the objective is to show that there exists a weakly feasible orientation such that the size of eye is still e and the set $\mathbb{E}_2$ is empty. For any edge $\{k,\ell\} \in \mathbb{E}_2$, we can arbitrarily assign direction $+1$ or -1 to the edge and the orientation is still weakly feasible. This is because for vertices in $\mathbb{V}_1$, the requirement on in-degree and out-degree is satisfied by other edges. More specifically, if the degree of k or ℓ is nonzero, then by the definition of weakly feasible orientation, the vertex already has nonzero in-degree and out-degree. Otherwise, if k or ℓ is the slack bus, then the in-degree and out-degree can be arbitrary. Thus, we can arbitrarily assign directions $+1$ or -1 to all edges in $\mathbb{E}_2$ and the new orientation is still weakly feasible. We define a new orientation as

$$\begin{aligned}\tilde{A}_{k\ell} &:= \begin{cases}+1 & \text{if } k > \ell\\ -1 & \text{otherwise}\end{cases}, \quad \forall \{k,\ell\} \in \mathbb{E}_2,\\ \tilde{A}_{k\ell} &:= A_{k\ell}, \quad \forall \{k,\ell\} \in \mathbb{E}_1 \cup \mathbb{E}_3.\end{aligned}$$

We prove that with orientation $\tilde{A}$, the size of eye is not changed. Let $(k_1,\ldots,k_t)$ be a directed cycle in the graph with orientation $\tilde{A}$. If some edges of this cycle are in $\mathbb{E}_1 \cup \mathbb{E}_3$, then this cycle also exists in the graph with A. By assigning directions ± 1 to edges with direction 0, the lengths of the cycles are not decreased and therefore the length of $(k_1,\ldots,k_t)$ is at least e under the orientation $\tilde{A}$. If all edges of this cycle are in $\mathbb{E}_2$, then we choose the minimal index in $\{k_1,\ldots,k_t\}$, which is assumed to be k_1 without loss of generality. By the definition of $\tilde{A}$, the edge $\{k_1,k_2\}$ has orientation $\tilde{A}_{k_1k_2} = -1$, which contradicts the fact that $(k_1,\ldots,k_t)$ is a directed cycle with $\tilde{A}$. Combining the above two cases, it can be inferred that the size of eye with orientation $\tilde{A}$ is at least e. On the other hand, e is defined to be the maximal eye. Hence, the size of eye with orientation $\tilde{A}$ is equal to e.

Step 4. In this step, we prove that the maximal eye is equal to the maximal girth. Suppose that the maximal eye is achieved by the weakly feasible orientation A and orientation A satisfies the conditions in Steps 2-3. We consider the set of parallel edges containing the root ear, which we denote as $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$ for some $t \geq 2$. We analyze two different cases:

- **Case I.** If there exists at least one parallel edge having direction 0, then by the conditions in Step 3, we know that at least one of the endpoints k, ℓ has degree 0. This means that all parallel edges have direction 0. We construct another graph $(\tilde{\mathbb{V}}, \tilde{\mathbb{E}}, \tilde{W})$, where the parallel edges $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$ are substituted by a single edge $\{k, \ell\}$ and the weight of the new edge is the minimal weight among all parallel edges, i.e.,

$$\tilde{W}_{k\ell} := \min_{s \in [t]} \; W_{k,\ell,s}.$$

  Other edges are the same as those in the original graph. We construct a weakly feasible orientation $\tilde{A}$ for the new graph. For the edge $\{k, \ell\}$, we define

$$\tilde{A}_{k\ell} := 0.$$

  For other edges, we define

$$\tilde{A}_{k_1\ell_1} := A_{k_1\ell_1} \qquad \forall \{k_1, \ell_1\} \in \mathbb{E} \backslash \{\{k, \ell, 1\}, \ldots, \{k, \ell, t\}\}.$$

  Since the orientations $\tilde{A}$ and A have the same degree at each node, $\tilde{A}$ also becomes weakly feasible. Moreover, the size of eye of the graph with $\tilde{A}$ is also equal to e, which implies that the maximal eye of the new graph $\tilde{\mathrm{e}}$ is at least e. Since the new graph $(\tilde{\mathbb{V}}, \tilde{\mathbb{E}}, \tilde{W})$ has $t - 1$ fewer ears, the induction assumption implies that the maximal girth of the new graph $\tilde{\mathrm{g}}$ satisfies

$$\tilde{\mathrm{g}} = \tilde{\mathrm{e}} \geq \mathrm{e}.$$

  Hence, we can choose a feasible orientation $\tilde{A}^g$ such that the girth is equal to $\tilde{\mathrm{g}}$. Now, we extend the feasible orientation $\tilde{A}^g$ to be a feasible orientation of the original graph $(\mathbb{V}, \mathbb{E}, W)$. We define

$$A^g_{k_1\ell_1} := \tilde{A}^g_{k_1\ell_1} \qquad \forall \{k_1, \ell_1\} \in \mathbb{E} \backslash \{\{k, \ell, 1\}, \ldots, \{k, \ell, t\}\}$$

  and

$$A^g_{k,\ell,s} := \tilde{A}^g_{k\ell}, \quad \forall s \in [t].$$

  Since the in-degree and out-degree at points k, ℓ are still nonzero for the orientation A^g, it can be concluded that A^g is a feasible orientation for the original graph. Moreover, the girth of the original graph with orientation A^g is equal to $\tilde{\mathrm{g}}$. It follows that the maximal girth g is at least $\tilde{\mathrm{g}} \geq \mathrm{e}$. This contradicts the assumption that $\mathrm{e} > \mathrm{g}$.

- **Case II.** Next, we consider the case when all parallel edges $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$ are normal edges. In this case, the goals is to construct a feasible orientation with the same size of eye by assigning directions to edges with direction 0. We first construct a feasible orientation $\tilde{A}$. Assume that $L_0 = \{k, \ell, 1\}$ is the root ear, and define

$$\begin{aligned} \tilde{A}_{k,\ell,1} &:= A_{k,\ell,1}, \\ \tilde{A}_{k,\ell,s} &:= -A_{k,\ell,1}, \quad \forall s \in \{2, \ldots, t\}. \end{aligned}$$

  Then, we inductively define the directions of other ears using the directed tree structure of ears. For any ear L_k that has been assigned a direction, we assign its descendant ear L_ℓ with the parallel direction as the path formed by the endpoints of L_ℓ on L_k. In this way, the orientation $\tilde{A}$ is defined for all ears and the definition is unique because of the directed tree structure. Considering the structure of the nested ear decomposition, we also know that all directed cycles in orientation $\tilde{A}$ must contain the root ear. In addition, the orientation $\tilde{A}$ is feasible. This is because all internal vertices of ears have nonzero in-degree and nonzero out-degree. The only vertices that are not internal vertices of ears are the endpoints of the root ear. For the endpoints of the root ear, they also have nonzero in-degree nonzero and out-degree by the definition of directions on parallel edges. Hence, the constructed orientation $\tilde{A}$ is feasible.

  We then define an orientation that combines orientations A and $\tilde{A}$ as follows:

$$A^g_{k\ell} := \begin{cases} A_{k\ell} & \text{if } A_{k\ell} \in \{+1, -1\} \\ \tilde{A}_{k,\ell} & \text{if } A_{k\ell} = 0, \end{cases}, \quad \forall \{k, \ell\} \in \mathbb{E}.$$

  We prove that A^g is a feasible orientation and the girth of orientation A^g is at least e. For any vertex k that has a nonzero degree in orientation A, the vertex k has nonzero in-degree and out-degree by the definition of weakly feasible orientation. Hence, the vertex k also has nonzero in-degree and out-degree in the new orientation. If the vertex has degree 0 in the orientation A, then all edges incident to the vertex k has the same direction as in $\tilde{A}$. Since the orientation $\tilde{A}$ is feasible, the vertex k has nonzero in-degree and nonzero out-degree in the new orientation A^g. Combining the two cases, it can be concluded that the orientation A^g is feasible. Now, we estimate the girth of orientation A^g. We consider any directed cycle C in A^g. If the cycle C has normal edges in the original orientation A, then the length of cycle C is not decreased in the new orientation and therefore is at least e. If the cycle C does not have normal edges in the original orientation A, then all edges of C have the same direction as in $\tilde{A}$ and therefore is also a cycle in in $\tilde{A}$. This implies that the root ear L_0 is on the cycle C. However, the root ear is a normal edge in orientation $\tilde{A}$ and this contradicts the assumption that none of the edges of the cycle C are normal. Thus, the girth of A^g is at least e. On the other hand, the girth of a feasible orientation is bounded by the maximal girth g. This contradicts the assumption that e $>$ g.

Combining the above two cases and using the induction method, it can be concluded that the maximal eye of a 2-vertex-connected SP graph is equal to its maximal girth.

□

Proof of Theorem 70

Proof. We only prove the strong uniqueness part since the proof for the weak uniqueness is similar. We only need to show that statement 2 of this theorem holds if and only if statement 2 of Lemma 47 holds.

Proof of sufficiency. We assume conversely that there exist two sets of phase angle vectors Θ^1 and Θ^2 satisfying statement 2 of Lemma 47 such that the induced sub-graph of $\Theta^1 - \Theta^2$ denoted as $(\mathbb{V}_0, \mathbb{E}_0, A_0)$ has the same number of strongly connected components and weakly connected components. Let $\mathbf{y}$ be a vector that satisfies conditions in statement 2 of Lemma 47. We prove that if vertices k and ℓ are in the same connected component, then $y_k = y_\ell$. By the definition of strongly connected components, there exist directed paths from k to ℓ and from ℓ to k. We first consider the directed path from k to ℓ, which we denote as $(k, k_1, \ldots, k_t, \ell)$. Considering the edge $\{k, k_1\}$ and inequality (6.2), one can write

$$\begin{aligned} &\sin(\pi/2 + \Theta^1_{k,k_1}/2 + \Theta^2_{k,k_1}/2) \cdot y_k \\ \geq &\sin(\pi/2 - \Theta^1_{k,k_1}/2 - \Theta^2_{k,k_1}/2) \cdot y_{k_1}. \end{aligned} \tag{6.10}$$

By the same analysis in Lemma 48, the condition $\Delta_{k,k_1} > 0$ implies that $\Theta^1_{k,k_1}/2 + \Theta^2_{k,k_1}/2 \in (-\pi/2, \pi/2)$, which leads to

$$\begin{aligned} &\sin(\pi/2 + \Theta^1_{k,k_1}/2 + \Theta^2_{k,k_1}/2) \\ = &\sin(\pi/2 - \Theta^1_{k,k_1}/2 - \Theta^2_{k,k_1}/2) > 0. \end{aligned} \tag{6.11}$$

Combining the relations in (6.10) and (6.11), we obtain $y_k \geq y_{k_1}$. Considering edges $\{k_1, k_2\}, \ldots, \{k_n, \ell\}$ and using the same analysis, we have

$$y_k \geq y_{k_1} \geq y_{k_2} \geq \cdots \geq y_{k_t} \geq y_\ell,$$

and therefore $y_k \geq y_\ell$. Similarly, the existence of a directed path from y_ℓ to y_k implies that $y_\ell \geq y_k$. Combining the two directions, we obtain $y_k = y_\ell$. If we further assume $\{k, \ell\} \in \mathbb{E}_0$ and $\Delta_{k\ell} > 0$, then the relation in (6.11) implies that inequality (6.2) holds with equality for $\{k, \ell\}$. By the definition of weakly connected components, there does not exist any edge in $\mathbb{E}_0$ connecting different connected components. Hence, the endpoints of all edges in $\mathbb{E}_0$ are in the same connected component and therefore inequality (6.2) holds with equality for all $\{k, \ell\} \in \mathbb{E}_0$ such that $\Delta_{k\ell} > 0$. Finally, by the definition of induced sub-graph, $\mathbb{E}_0$ contains all edges $\{k, \ell\} \in \mathbb{E}$ such that $\Delta_{k\ell} > 0$. It follows that inequality (6.2) holds with equality for all $\{k, \ell\} \in \mathbb{E}$ such that $\Delta_{k\ell} > 0$. This contradicts statement 2 of Lemma 47 that there exists at least one strict inequality in the set of inequalities (6.2). Hence, statement 2 of this theorem holds.

Proof of necessity. Assume that the conditions in statement 2 of this theorem hold. We denote the strongly connected components as $\mathcal{C}_1, \ldots, \mathcal{C}_m$. Now, we define a tree structure for the set $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$. For two different strongly connected components $\mathcal{C}_s$ and $\mathcal{C}_t$, if there exists a directed path from $\mathcal{C}_s$ to $\mathcal{C}_t$, we define a directed edge from $\mathcal{C}_t$ to $\mathcal{C}_s$. Considering all strongly connected components pairs, we obtain a directed graph with the vertex set $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$. By the definition of strongly connected components, we know that there does not exist directed cycle in this directed graph and therefore this directed graph is a directed tree. Using the directed tree structure, we can choose m real numbers $c_1, \ldots, c_m$ such that if there exists a directed path from $\mathcal{C}_t$ to $\mathcal{C}_s$, then it holds that $c_t > c_s$. Moreover, if vertex 1 belongs to some strongly connected component $\mathcal{C}_s$, then we can shift c_t for all $t \in [m]$ such that $c_s = 0$ and the relation between all c_t's is not changed. If vertex 1 does not belong to any strongly connected component, we do not change the value of c_t.

We construct a vector $\mathbf{y} \in \mathbb{R}^{|\mathbb{V}|}$ by

$$y_k := \begin{cases} c_s & \text{if } k \text{ is in } \mathcal{C}_s \\ 0 & \text{if } k \in \mathbb{V} \backslash \mathbb{V}_0. \end{cases}$$

Note that the set of strongly connected components gives a disjoint partition of the set $\mathbb{V}_0$. Hence, the vector $\mathbf{y}$ is well-defined. By the choice of $\{c_1, \ldots, c_m\}$, the vector $\mathbf{y}$ satisfies $y_1 = 0$. Suppose that the edge $\{k, \ell\}$ belongs to $\mathbb{E}$ and $\Delta_{k\ell} > 0$. We verify that inequality (6.2) holds for $\{k, \ell\}$, namely,

$$\begin{aligned} &\sin(\pi/2 + \Theta^1_{k\ell}/2 + \Theta^2_{k\ell}/2) \cdot y_k \\ \geq &\sin(\pi/2 - \Theta^1_{k\ell}/2 - \Theta^2_{k\ell}/2) \cdot y_\ell. \end{aligned}$$

Recalling that the relation (6.11) holds for all $\{k, \ell\}$ such that $\Delta_{k\ell} > 0$, we only need to verify

$$y_k \geq y_\ell, \quad \forall \{k, \ell\} \in \mathbb{E}_0 \quad \text{s. t. } \Delta_{k\ell} > 0. \tag{6.12}$$

By the definition of induced sub-graph, the condition $\Delta_{k\ell} > 0$ implies that $\{k, \ell\} \in \mathbb{E}_0$. Thus, vertices k and ℓ must belong to certain strongly connected components. If k and ℓ belong to the same strongly connected component $\mathcal{C}_s$, then $y_k = y_\ell = c_s$ and inequality (6.12) holds. Otherwise, we assume that k and ℓ belong to two different strongly connected components $\mathcal{C}_s$ and $\mathcal{C}_t$, respectively. Since (k, ℓ) is a directed path from $\mathcal{C}_s$ to $\mathcal{C}_t$, one can write

$$y_k = c_s > c_t = y_\ell$$

and inequality (6.12) holds strictly. By the assumption that there are strictly more strongly connected components than weakly connected components, there exists at least one edge $\{k, \ell\} \in \mathbb{E}_0$ such that k and ℓ belong to different strongly connected components. Without loss of generality, assume that $\Delta_{k\ell} > 0$. Then, the inequality (6.12), or equivalently the inequality (6.2), holds strictly for $\{k, \ell\}$. This shows that $\mathbf{y}$ is a vector that satisfies conditions in statement 2 of Lemma 47.

□

6.D Proof for Iterative Series-Parallel Reduction Method

Proof of Lemma 51

Proof. We prove the four claims separately.

Type I Operation. We first consider the inequality on the right. We denote the two endpoints as k, ℓ and the parallel edges connecting them as $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$ for some $t \geq 2$. Without loss of generality, assume that the weights of parallel edges satisfy

$$W_{min} = W_{k,\ell,1} \leq \cdots \leq W_{k,\ell,t} = W_{max}.$$

Suppose that the maximal eye of graph $(\mathbb{V}, \mathbb{E}, W)$ is achieved by the weakly feasible orientation A. If there exist different directions among these parallel edges when orientation A is assigned, then we choose the first edge $\{k, \ell, 1\}$ and another edge $\{k, \ell, s\}$ such that the direction of $\{k, \ell, s\}$ is different from the direction of $\{k, \ell, 1\}$. Hence, $\{k, \ell, 1\}$ and $\{k, \ell, s\}$ form a directed cycle and two edges have different directions. Then, at least one edge is a normal edge, i.e., an edge with direction $+1$ or -1. The weight of the cycle is bounded by $W_{k,\ell,1} + W_{k,\ell,s} \leq W_{max} + W_{min}$. Thus, it holds that $\mathrm{e} \leq W_{max} + W_{min}$ in this case. Otherwise, assume that all parallel edges have the same direction when orientation A is assigned. Considering a directed cycle that contains the edge $\{k, \ell, s\}$ for some $s \in \{2, \ldots, t\}$, we can substitute the edge $\{k, \ell, s\}$ with edge $\{k, \ell, 1\}$ and the length of the directed cycle is not increased. Hence, if we delete edges $\{k, \ell, 2\}, \ldots, \{k, \ell, t\}$, the size of eye is not changed. On the other hand, the deletion of edges $\{k, \ell, 2\}, \ldots, \{k, \ell, t\}$ is equivalent to the Type I Operation on the set of parallel edges $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$. Hence, we obtain $\mathrm{e} = \tilde{\mathrm{e}}$ in this case. Combining the two cases, it follows that $\mathrm{e} \leq \max\{\tilde{\mathrm{e}}, W_{max} + W_{min}\}$.

We now prove the inequality on the left. Suppose that the maximal eye of the new graph $(\tilde{\mathbb{V}}, \tilde{\mathbb{E}}, \tilde{W})$ is achieved by the weakly feasible orientation $\tilde{A}$. By the definition of Type I Operations, the weight $\tilde{W}_{k,\ell}$ is equal to the weight $W_{k,\ell,1}$. We consider the inverse operation of Type I Operation. Namely, we add parallel edges $\{k, \ell, s\}$ with weight $W_{k,\ell,s}$ to the new graph and define the direction $\tilde{A}_{k,\ell,s} := \tilde{A}_{k,\ell,1}$ for all $s \in \{2, \ldots, t\}$. Then, the orientation $\tilde{A}$ becomes a weakly feasible orientation for the original graph. By the discussion for the inequality on the right, the inverse operation will not change the size of eye. Therefore, we have a weakly feasible orientation for $(\mathbb{V}, \mathbb{E}, W)$ and the size of eye is $\tilde{\mathrm{e}}$, which implies that $\mathrm{e} \geq \tilde{\mathrm{e}}$.

Type II Operation. We consider the case when a Type II Operation is implemented. We denote the deleted degree-2 vertex as k. By the definition of Type II Operations, vertex k has two neighbouring vertices and we denote the two neighbouring vertices as $\ell_1 \neq \ell_2$. If A is a weakly feasible orientation for $(\mathbb{V}, \mathbb{E}, W)$, then the direction $A_{\ell_1,k}$ must be equal to the direction A_{k,ℓ_2}. Hence, treating the two edges $\{\ell_1, k\}$ and $\{k, \ell_2\}$ as a single edge with

weight $W_{\ell_1,k} + W_{k,\ell_2}$ will not change the size of eye. Noticing that the claim is true for any weakly feasible orientation A, we know that $\mathrm{e} = \tilde{\mathrm{e}}$.

Type III Operation with a pendant vertex. Removing a pendant vertex will not affect the maximal eye, since any directed cycle does not contain pendant vertices. Thus, we conclude that $\mathrm{e} = \tilde{\mathrm{e}}$.

Type III Operation with a non-pendant vertex. Finally, we consider the case when the deleted vertex has degree at least 2. We denote the deleted vertex as k and denote the only neighbouring vertex as ℓ. The parallel edges connecting k and ℓ are denoted as $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$ for some $t \geq 2$. Similar to the Type I Operation case, assume that the weights of parallel edges satisfy

$$W_{min} = W_{k,\ell,1} \leq \cdots \leq W_{k,\ell,t} = W_{max}.$$

We can split the deletion of vertex k into two operations. We first substitute parallel edges $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$ with a single edge $\{k, \ell\}$ with weight $W_{k,\ell,1}$. Then, we delete the pendant vertex k. The two operations can be viewed as Type I and Type III Operations, respectively. Using the results in the first case and the third case, one can write

$$\tilde{\mathrm{e}} \leq \mathrm{e} \leq \max\{\tilde{\mathrm{e}}, W_{max} + W_{min}\}.$$

Hence, it remains to prove that $\mathrm{e} \geq W_{max} + W_{min}$. We can construct a weakly feasible orientation such that size of eye is $W_{max} + W_{min}$. Specifically, we define

$$A_{k,\ell,s} := +1, \quad \forall s \in \{1, \ldots, t-1\}, \quad A_{k,\ell,t} := -1$$

and all other edges are assigned the direction 0. Then, vertices k and ℓ have nonzero in-degree and out-degree, while other vertices have zero in-degree and out-degree. Hence the orientation A is weakly feasible. Now, consider directed cycles with at least one normal edge. Since parallel edges $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$ are the only normal edges, the directed cycle must contain at least one of these parallel edges. Using the facts that ℓ is the only neighbouring vertex of k and directed cycles do not have repeated vertices, vertices k and ℓ are the only two vertices of the directed cycle. Hence, the size of eye should be the the minimal length of such directed cycles, which is $W_{k,\ell,1} + W_{k,\ell,t} = W_{max} + W_{min}$. Thus, it follows that $\mathrm{e} \geq W_{max} + W_{min}$.

Combining the two parts yields that $\mathrm{e} = \max\{\tilde{\mathrm{e}}, W_{max} + W_{min}\}$. □

Proof of Theorem 71

Proof. We first consider the upper bound. Using Lemma 51, the upper bound on the maximal eye can be either the maximal eye of the reduced graph or $W_{max} + W_{min}$, where W_{max} and

W_{min} are defined in Lemma 51. Since $\max\{\alpha_1, \alpha_2\}$ is the maximal value of $W_{max} + W_{min}$ appeared during the reduction, we know the upper bound is given by $\max\{\mathrm{e}_R, \alpha_1, \alpha_2\}$.

Then, we consider the lower bound. If Type-III operation is implemented, the maximal eye is changed to be the maximum of the maximal eye of the new graph and $W_{max} + W_{min}$. In addition, since the maximal eye is not decreased after each reduction, we get the lower bound $\max\{\mathrm{e}_R, \alpha_2\}$.

□

Proof of Lemma 52

Proof. The first three claims can be proved in the same way as Lemma 51 and we only prove the last two claims. We denote the deleted vertex as k and its only neighboring vertex as ℓ. The parallel edges connecting k and ℓ are denoted as $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$ for some $t \geq 2$. Without loss of generality, assume that the weights of parallel edges satisfy

$$W_{min} = W_{k,\ell,1} \leq \cdots \leq W_{k,\ell,t} = W_{max}.$$

Type III Operation for slack node. We first consider the case when the deleted vertex is a slack node. By discussing whether parallel edges $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$ have the same direction as in the first claim in Lemma 51, it holds that $\mathrm{g} \leq \max\{\tilde{\mathrm{g}}, W_{max} + W_{min}\}$.

We prove the other inequality $\tilde{\mathrm{g}} \leq \mathrm{g}$ by constructing a feasible orientation A such that the girth is $\tilde{\mathrm{g}}$. Suppose that the maximal girth of the new graph $(\tilde{\mathbb{V}}, \tilde{\mathbb{E}}, \tilde{W})$ is achieved by the feasible orientation $\tilde{A}$. We define directions for deleted parallel edge such that the orientation $\tilde{A}$ becomes a feasible orientation of the original graph $(\mathbb{V}, \mathbb{E}, W)$. We note that, by the definition of Type III Operations, the vertex ℓ is a slack node in the new graph and it may not satisfy the condition on in-degree and out-degree. If the vertex ℓ in the new graph with orientation $\tilde{A}$ has nonzero in-degree, then we define

$$\tilde{A}_{k,\ell,s} := -1, \quad \forall s \in \{1, \ldots, t\}.$$

Then, the vertex ℓ has both nonzero in-degree and nonzero out-degree. Since the vertex k is a slack node, the orientation $\tilde{A}$ becomes a feasible orientation for the original graph $(\mathbb{V}, \mathbb{E}, W)$. By the construction of $\tilde{A}$, the vertex k only has nonzero in-degree and therefore there does not exist any directed cycle containing parallel edges $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$. It follows that the girth is not changed and is equal to $\tilde{\mathrm{g}}$. If the vertex ℓ in the new graph with orientation $\tilde{A}$ has nonzero out-degree, then we can similarly define

$$\tilde{A}_{k,\ell,s} := +1, \quad \forall s \in \{1, \ldots, t\}.$$

The orientation $\tilde{A}$ also becomes a feasible orientation for the original graph and the girth is $\tilde{\mathrm{g}}$. Combining the two cases concludes that $\mathrm{e} \geq \tilde{\mathrm{g}}$.

Type III Operation for non-slack node. We then consider the case when the deleted vertex is not a slack node. Suppose that the maximal girth of the original graph $(\mathbb{V}, \mathbb{E}, W)$ is achieved by the feasible orientation A. Since the vertex k has nonzero in-degree and nonzero out-degree, there must exist different directions among parallel edges $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$. Hence, by the same analysis as the first claim in Lemma 51, it holds that $\mathrm{g} \leq W_{max} + W_{min}$. Now, we consider restricting the orientation A to the new graph$(\tilde{\mathbb{V}}, \tilde{\mathbb{E}}, \tilde{W})$. Since the vertex ℓ is a slack node in the new graph and the orientation A is not changed for other vertices, the orientation A becomes a feasible orientation for the new graph. Then, by the definition of the maximal girth, there exists a directed cycle in the new graph with length at most $\tilde{\mathrm{g}}$. Hence, we conclude that $\mathrm{g} \leq \tilde{\mathrm{g}}$. Combining the two inequalities, it follows that $\mathrm{g} \leq \min\{\tilde{\mathrm{g}}, W_{max} + W_{min}\}$.

Now, it remains to prove $\mathrm{g} \geq \min\{\tilde{\mathrm{g}}, W_{max} + W_{min}\}$. Suppose that the maximal girth of the new graph $(\tilde{\mathbb{V}}, \tilde{\mathbb{E}}, \tilde{W})$ is achieved by the feasible orientation $\tilde{A}$. We extend the orientation $\tilde{A}$ to be an orientation for the original graph by defining

$$A_{k,\ell,s} := +1, \quad \forall s \in \{1, \ldots, t-1\}, \quad A_{k,\ell,t} = -1.$$

Since both vertices k, ℓ have nonzero in-degree and nonzero out-degree and the orientation at other vertices is not changed, the orientation A becomes a feasible orientation for the original graph. Now, we calculate the girth of the original graph. For any directed cycle that does not contain parallel edges $\{k, \ell, 1\}, \ldots, \{k, \ell, t\}$, it is also a directed cycle in the new graph and has length at least $\tilde{\mathrm{g}}$. For any directed cycle that contains at least one of those parallel edges, vertices k and ℓ are the only two vertices of the directed cycle, since there does not exist repeated vertices on directed cycles. Hence, the length of the directed cycle is at least $W_{k,\ell,1} + W_{k,\ell,t} = W_{max} + W_{min}$. Combining the two cases yields that the girth is at least $\min\{\tilde{\mathrm{g}}, W_{max} + W_{min}\}$ and therefore $\mathrm{g} \geq \min\{\tilde{\mathrm{g}}, W_{max} + W_{min}\}$. □

Proof of Theorem 72

Proof. The proof is similar to that of Theorem 71 and we omit it.

□

Proof of Theorem 73

Proof. We prove that Type I-II Operations are enough for reducing a 2-vertex-connected SP graph to a single edge. Since Type I-II Operations do not introduce new slack nodes, there exists at most one slack node in the graph throughout the reduction process. By the assumption that the graph is a 2-vertex-connected SP graph, Theorem 68 implies that there exists a nested ear decomposition $(L_0, \ldots, L_{r-1})$ of the graph. We use the induction method on the number of ears in the ear decomposition. If there are only one ear or two ears in the ear decomposition, then the result holds trivially. We assume that any 2-vertex connected SP graphs with at most $r-1$ ears in the ear decomposition can be reduced to a single edge with Type I-II Operations.

Now, we consider the case when there are r ears in the ear decomposition. We first implement Type II Operations until there is no degree-2 vertices except the slack bus. Since Type II Operations will not change the structure of the nested ear decomposition, the new graph still has a nested ear decomposition with at most r ears in the decomposition. By the first step in the proof of Theorem 50, there exists a set of parallel edges containing the root ear or a leaf ear. We analyze two different cases:

Case I. Assume that there exists a set of parallel edges containing a leaf ear. We denote the leaf ear as $L_s = \{k, \ell\}$. Let L_t be the precedent ear of L_t. Then, then set of parallel ears consists of the segment $\overline{k\ell}$ on ear L_t and leaf ears on L_t. We can apply a Type I Operation to substitute the set of parallel edges with a single edge. We can view the new edge as the segment $\overline{k\ell}$ on ear L_t. Then, at least leaf ear is deleted and the new graph has a nested ear decomposition with at most $r - 1$ ears. By the induction assumption, the new graph can be reduced to a single edge with Type I-II Operations. Thus, the original graph can be reduced to a single edge with Type I-II Operations.

Case II. Assume that there exists a set of parallel edges containing the root ear. Then by the same construction in the second step in the proof of Theorem 50, we can change the root ear to a leaf ear. Hence, we obtain a set of parallel edges containing a leaf ear and we can apply the discussion in Case I.

Combining the two cases, it follows that the result is true when there are r ears in the ear decomposition. By the induction method, the result is true for any $r \geq 1$ and the ISPR method can reduce a 2-vertex-connected SP graph to a single edge.

□

Chapter 7

Distributionally Robust Optimization for Chance-Constrained Optimal Power Flow

7.1 Introduction

Developing resilient algorithms for the *Optimal Power Flow* (OPF) problem is fundamental to efficient and reliable decision-making in large-scale energy systems. The OPF problem consists of minimizing some objective, including but not limited to generation costs, subject to the physics of the power network as well as additional constraints on power quality, safety, and reliability. Independent system operators solve OPF at several timescales, from hours to minutes ahead of the dispatch time, in order to manage the market and match supply to demand. Traditionally, the primary source of uncertainty in OPF was stochastic loads. This uncertainty was handled through forecasts which were accurate enough that mismatches between supply and demand could be handled in real-time without a significant deviation from nominal network and market conditions. However, given the ongoing emergence of intermittent renewable generation, more sophisticated methods will be necessary to ensure that decisions can be made as efficiently as possible while being robust to large forecast errors.

The randomness in the constraints prohibits the application of optimization algorithms for deterministic problems and most stochastic optimization algorithms, which are often applicable to optimization problems that only contain randomness in the objective function. The robust optimization approach was proposed in [115] and [151] to find the worst-case solution, namely, the optimal decision that satisfies all constraints for all possible realizations of the randomness in the system. The robust optimization approach produces the most conservative solution and results in a high operational cost.

To improve the efficiency of the operation of power systems, it is often preferable to allow a small user-specified probability of violating the constraints in the OPF solution in exchange for a much better operational cost (small violations will later be handled via a

real-time control mechanism). *Chance-constrained OPF* (CCOPF) is a natural formulation for balancing the trade-off between efficiency and robustness [194]. In CCOPF, system operators attempt to find the minimum-cost solution which violates the constraints with a probability at most equal to a pre-defined parameter named the violation probability. Chance-constrained methods avoid the conservativeness associated with robust optimization, which insures an operating point that is feasible for all possible realizations of a system's forecast errors. Please refer to [194] and [223] for popular formulations of CCOPF.

A challenge for CCOPF is that the true underlying distribution of the random parameters is generally unknown and must be inferred from historical data. A conventional approach is the sample average approximation [178], which is easily applicable but may lead to a high-variance estimate of the true distribution. The scenario approach lower-bounds the number of samples required to achieve a given degree of confidence in the probability of satisfying the chance constraints [31] and is employed for CCOPF in [194] and [223]. However, the scenario approach is sample-intensive, may be overly conservative, and is often computationally complex. Additionally, more sample-efficient methods allow for samples over larger time horizons (i.e., a day instead of an hour) to be aggregated into a single realization of a random vector, which could reduce bias if forecast errors follow temporal patterns.

Distributionally robust optimization (DRO) alleviates the issue of unknown true distributions by enforcing the chance constraints for all distributions in an *ambiguity set* centered, in the sense of some characteristic metric of probability distributions, around the empirical distribution [190]. The idea is that, given enough samples, the true distribution is highly likely to fall inside the ambiguity set. A number of papers have applied distributionally robust optimization to OPF or related problems in energy systems. The authors of [154, 250, 231, 235] employ *moment-based* ambiguity sets containing probability distributions with the first and second moments close to those of the empirical distribution. Li *et al.* [144] add a unimodality assumption to the moment-based sets to reduce the conservatism. Moment-based ambiguity sets often yield exact tractable reformulations of the chance-constrained program, but they lose information about the true distribution revealed through other features of the data. *Metric-based* ambiguity sets, by contrast, are constructed using measures of distance between probability distributions, most often the Wasserstein metric, and are more expressive. The metric-based approach has the advantage that various statistical consistency and convergence guarantees can be established for DRO estimators [166, 222]. To reformulate the chance constraints as tractable constraints, inner approximations of Wasserstein metric-based ambiguity sets, such as hyper-cubes [63] and polytopes [255], have previously been studied. However, these inner approximations are overly conservative in practice and lead to pessimistic estimations.

All of the aforementioned DRO approaches are designed for *disjoint* chance constraints, in which each constraint individually must be satisfied with a given probability. The chance constraints in CCOPF are formulated disjointly for each two-sided constraint [255, 235, 63] or separately for each upper and lower bound [154, 144, 250]. *Joint* chance constraints, by contrast, require that a solution be feasible, that is, satisfies *all* constraints simultaneously, with a given probability. Given the same violation probability, joint chance constraints are

clearly stronger than disjoint chance constraints. Joint chance constraints can be guaranteed by applying the Boole inequality to appropriately scaled disjoint chance constraints; see [16]. However, this approach is highly conservative and does not exploit the potential correlation between random variables in different constraints. Intuitively, when the randomness between constraints is highly correlated, joint chance constraints can be satisfied at a cost that is only slightly higher than that of the chance constraint of a single stochastic constraint. Yang *et al.* [240] build on the Boole inequality approach and achieve an inner approximation of a moment-based ambiguity set for the joint case.

The particularly interesting line of work [93, 92, 186, 12] is inspired by [166], which provides a reformulation of Wasserstein metric-based DRO problems using conditional value-at-risk (CVaR). The two-part work [93]-[92] is the first to apply the CVaR reformulation to OPF by penalizing constraint violations in the objective function; however, this is not a chance-constrained approach and cannot guarantee the satisfaction of the constraints in any well-defined sense. Poolla *et al.* [186] approximate the joint chance constraints using the Boole inequality and reformulate them using CVaR. To achieve the reformulation, the authors use an inner approximation of the ambiguity set via a hyper-rectangle in the parameter space. Arab *et al.* [12] improve on [186] by using an ellipsoidal approximation, which reduces the conservativeness by exploiting the correlation between random variables. While the ellipse approximation improves on the hyper-rectangle approximation, the method in [12] remains overly conservative as a consequence of mismatch between the inner approximation and the ambiguity set; see Section 7.4 for numerical illustrations. To address the above issues, we build upon our conference paper [27] tailored to a class of non-convex problems using DRO to study the CCOPF problem. Compared to [27], we develop strong theoretical results in the context of power systems for both the joint and disjoint cases, and we numerically illustrate the performances of our approach on benchmark IEEE power systems.

In this chapter, we expand upon the existing findings related to the DRO approach for CCOPF. Inspired by [222], we use a relative entropy-based ambiguity set in our DRO formulation and establish stronger theoretical guarantees than those in existing literature. Then, we apply the approximation of the chance-constraints from Roald *et al.* [194] and the semi-definite relaxation from Low *et al.* [153] to reformulate the problem as a mixed-integer program, which can be handled by existing optimization solvers. Moreover, we implement the algorithms on benchmark OPF problem instances, showcasing the advantages of our new formulation. We summarize our contributions in the following:

- Instead of the commonly used Wasserstein metric, our DRO formulation utilizes a relative entropy-based ambiguity set. We prove that the relative entropy-based formulation admits the *least conservative* DRO solution in the sense that the solution achieves the minimum possible generation cost under a certain asymptotic bound on out-of-sample performance.
- We provide the first *exact* reformulation of *joint* distributionally robust chance constraints over the ambiguity set. By comparison, existing works construct an approximation set of the ambiguity set and/or only consider disjoint chance constraints, which

	Chance-constrained	Joint	Metric-based	Exact Reformulation
[154]	✓	✗	✗	✓
[255]	✓	✗	✓	✗
[235]	✓	✗	✗	✓
[63]	✓	✗	✓	✗
[250]	✓	✗	✗	✓
[144]	✓	✗	✗	✓
[240]	✓	✓	✗	✗
[93, 92]	✗	-	✓	✓
[186]	✓	✓	✓	✗
[12]	✓	✓	✓	✗
Chapter 7	✓	✓	✓	✓

Table 7.1.1: Comparison of relevant chance-constrained OPF literature.

makes it challenging to control the trade-off between the efficiency and robustness of the solution. In our formulation, the balance can be effectively controlled by an input parameter. In addition, our reformulation always leads to a feasible problem, while existing approaches cannot guarantee the feasibility.

- We empirically compare the performance of our DRO approach with the state-of-the-art approach in [12] on the IEEE 14- and 118-bus test cases. We show that our approach is able to find more reliable and efficient solutions satisfying the joint chance constraints, while the approximation algorithm in [12] leads to overly conservative solutions.

Table 7.1.1 summarizes the relevant existing literature on DRO for power systems (most, but not all, of the listed papers focus on OPF) and illustrates our contributions. It is worth mentioning that all works in Table 7.1.1 except [12] use the common linearized DC approximation of the nonlinear power flow equations, though this approximation is not always coupled to the specific handling of chance constraints. In comparison, we consider the full ACOPF problem in this chapter.

The remainder of the chapter is organized as follows. In Section 7.2, we first introduce the AC OPF problem and the corresponding joint chance constraint. Reformulations of the chance-constrained AC OPF problem, including the distributionally robust optimization approach, are derived in Section 7.3. Finally, in Section 7.4, we implement the proposed algorithm to verify the theory and illustrate the superior empirical performances compared with existing algorithms. The proofs are provided in the appendix.

7.2 AC OPF Problem and Chance Constraints

In this section, we first introduce the notation for system variables and parameters in power flow equations. Then, we formulate the deterministic AC OPF problem as a *quadratically constrained quadratic program* (QCQP). Finally, we consider the stochastic OPF problem, where the system status is subject to random power injections. We formally define the joint and disjoint chance constraints and formulate the chance-constrained OPF problem.

QCQP Formulation of Deterministic AC OPF

Here, we present the deterministic AC OPF problem, namely, the AC OPF problem without unforecasted power injections, as a QCQP. Our formulation and most of our notation is based on [137]. Consider a power system with the set of buses $\mathcal{N} \coloneqq [n]$. Define the system variables and parameters:

- $\mathbf{V} \in \mathbb{C}^n$: Vector of complex bus voltages.
- $\mathbf{I} \in \mathbb{C}^n$: Vector of complex nodal current injections.
- $\mathbf{Y} \in \mathbb{C}^{n \times n}$: Network admittance matrix, constructed such that $\mathbf{I} = \mathbf{YV}$.
- $P^D, Q^D \in \mathbb{R}^n$: Vectors of active and reactive loads, respectively. If bus k has no load, the k-th components are zero.
- $\overline{P}^G, \underline{P}^G \in \mathbb{R}^n$: Vectors of upper and lower active generation limits, respectively. If bus k has no generator, the k-th components are zero.
- $\overline{Q}^G, \underline{Q}^G \in \mathbb{R}^n$: Vectors of upper and lower reactive generation limits, respectively. If bus k has no generator or dispatchable reactive compensation device, the k-th components are zero.
- $\overline{V}, \underline{V} \in \mathbb{R}^n$: vectors of upper and lower voltage magnitude limits, respectively.
- c_{kd}: The d-th degree coefficient of the quadratic cost function for the k-th generator, where $d \in \{0, 1, 2\}$.

Conventional OPF formulations consider fixed loads and dispatchable generators. To simplify the notation, the formulation used in this chapter incorporates renewable generators (without curtailment) into the load vectors as negative loads. Conversely, centrally dispatchable demand responses can be incorporated into the generator limits through appropriate adjustments.

To formulate the OPF problem as an optimization problem with real variables, we define the real vector

$$\mathbf{X} \coloneqq \begin{bmatrix} \Re\{\mathbf{V}\} \\ \Im\{\mathbf{V}\} \end{bmatrix} \in \mathbb{R}^{2n}.$$

Additionally, for each $k \in \mathbb{N}$, we define the following matrices:

$$\begin{aligned} Y_k &\coloneqq e_k e_k^T Y, \\ \mathbf{Y}_k &\coloneqq \frac{1}{2} \begin{bmatrix} \Re\{Y_k + Y_k^T\} & \Im\{Y_k^T - Y_k\} \\ \Im\{Y_k - Y_k^T\} & \Re\{Y_k + Y_k^T\} \end{bmatrix}, \\ \overline{\mathbf{Y}}_k &\coloneqq -\frac{1}{2} \begin{bmatrix} \Im\{Y_k + Y_k^T\} & \Re\{Y_k - Y_k^T\} \\ \Re\{Y_k^T - Y_k\} & \Im\{Y_k + Y_k^T\} \end{bmatrix}, \end{aligned}$$

$$\mathbf{M}_k \coloneqq \begin{bmatrix} e_k e_k^T & 0 \\ 0 & e_k e_k^T \end{bmatrix}.$$

While the OPF problem can accommodate different choices of the objective function, we focus on a common total generation cost $f : \mathbb{R}^{2n\times 2n} \to \mathbb{R}$ as follows:

$$f(\mathbf{W}) \coloneqq \sum\nolimits_{k\in\mathcal{N}} \Big[c_{k0} + c_{k1}\left(\langle \mathbf{W}, \mathbf{Y}_k\rangle + P_k^D\right) + c_{k2}\left(\langle \mathbf{W}, \mathbf{Y}_k\rangle + P_k^D\right)^2 \Big].$$

We can now write the deterministic AC OPF problem as a real-valued QCQP in terms of $\mathbf{X}$:

$$\begin{aligned}
\min_{\mathbf{X}\in\mathbb{R}^{2n}} \quad & f(\mathbf{X}\mathbf{X}^T) \\
\text{s.t.} \quad & \underline{P}_k^G - P_k^D \le \langle \mathbf{X}\mathbf{X}^T, \mathbf{Y}_k\rangle \le \overline{P}_k^G - P_k^D, && \text{(7.1a)} \\
& \underline{Q}_k^G - Q_k^D \le \langle \mathbf{X}\mathbf{X}^T, \overline{\mathbf{Y}}_k\rangle \le \overline{Q}_k^G - Q_k^D, && \text{(7.1b)} \\
& \underline{V}_k^2 \le \langle \mathbf{X}\mathbf{X}^T, \mathbf{M}_k\rangle \le \overline{V}_k^2, && \text{(7.1c)} \\
& \forall k \in \mathcal{N},
\end{aligned}$$

where constraints (7.1a) and (7.1b) are the real and reactive power balance equations, respectively. These are *hard* constraints imposed by the laws of physics and cannot be violated. Moreover, for buses without generators, the lower and upper bounds in (7.1a) and (7.1b) are equal. Constraint (7.1c) limits the voltage magnitude at each bus. This is a *soft* constraint imposed by regulation or operator preference. It is physically possible to violate these constraints, and small violations may be tolerated if they are sparse and/or low in magnitude.

In our analysis, we neglect line flow limits to streamline the presentation and mathematical derivation. However, since such constraints are in the form of (7.1a), our method can readily handle a more detailed formulation of AC OPF.

System Response to Unforecasted Power Injections

We now turn to rigorously formulating an approximate OPF problem based on the analysis in [194]. For the purposes of solving the OPF problem, the complex voltage vector serves as the decision variable as it fully specifies the operating point of the system; that is, given the complex voltage at each bus, the current and power injections can be computed. However, in practice, system operators cannot directly actuate voltage magnitudes and angles at all buses in the system. Instead, they control voltage magnitudes and active power injections at generator buses. Combined with the active and reactive demand from loads, the system naturally resolves to the complex voltage profile obtained as the OPF solution if the forecast is accurate.

We consider a random active power injection vector $\xi \in \mathbb{R}^n$, realized after the OPF decision is made. The random vector ξ represents the forecast error associated with either

loads or intermittent renewable energy generators, such as wind turbines or solar panels. For simplicity, we will assume that the active power injection induces a proportional reactive power injection according to a constant power factor $\cos\phi$.

If the system operator leaves non-slack generator setpoints unchanged after the realization of the random variable, then the following variables are held constant:

1. Voltage magnitude and active power injection from generators at the set of generator buses $\mathcal{PV}$.
2. Active and reactive loads at the set of load buses $\mathcal{PQ}$.
3. Voltage magnitude and angle at the slack bus $P\theta$.

Under this response mechanism, the full aggregate active power imbalance induced by ξ is offset by the slack bus. Instead, we assume that an Automatic Generation Control (AGC) scheme is used to distribute the burden among the generators. The imbalance is divided among buses according to participation factors $\alpha \in \mathbb{R}^n$, where $\mathbf{1}_n^T\alpha = 1$ and $\alpha_k = 0$ for all $k \notin \mathcal{PV}$. In summary, the *known* change of the post-contingency system state is given by:

$$\begin{aligned} \Delta P_k &= \xi_k - \alpha_k \mathbf{1}_n^T \xi, \quad \forall k \in \mathcal{PQ} \cup \mathcal{PV}, \\ \Delta Q_k &= \gamma \xi_k, \quad \forall k \in \mathcal{PQ}, \\ \Delta |V_k| &= 0, \quad \forall k \in \mathcal{PV} \cup \{P\theta\}, \quad \Delta\theta_{P\theta} = 0, \end{aligned} \tag{7.2}$$

where we use symbol Δ to denote the change of corresponding system state variable, γ is equal to $\sqrt{\cos^{-2}\phi - 1}$ and θ_k is the voltage angle at bus k. After applying these changes, we can determine $2n$ of the post-contingency power flow variables and the other $2n$ variables are determined by solving the power flow equations.

Remark 9. In [12], the participation factors in α account for the mismatch from forecast errors *and* the difference in resistive losses induced by the forecast errors. For simplicity, we do not adopt this approach as the difference in losses is small relative to the errors. In practice, the participation factor at the slack bus may be artificially lowered to offset the burden associated with resistive losses.

Given voltage profile $\mathbf{X} \in \mathbb{R}^{2n}$ and forecast error $\xi \in \mathbb{R}^n$, denote the change in active power injections, reactive power injections and squared voltage magnitude, respectively, as

$$\Delta P(\mathbf{X}, \xi), \quad \Delta Q(\mathbf{X}, \xi), \quad \Delta |V|^2(\mathbf{X}, \xi).$$

Hence, the constraints in the OPF problem (7.1) become

$$\begin{aligned} \underline{P}_k^G - P_k^D \le \langle \mathbf{X}\mathbf{X}^T, \mathbf{Y}_k \rangle - \Delta P_k(\mathbf{X}, \xi) + \xi \le \overline{P}_k - P_k^D, \\ \underline{Q}_k^G - Q_k^D \le \langle \mathbf{X}\mathbf{X}^T, \overline{\mathbf{Y}}_k \rangle - \Delta Q_k(\mathbf{X}, \xi) + \gamma\xi \le \overline{Q}_k - Q_k^D, \\ \underline{V}_k^2 \le \langle \mathbf{X}\mathbf{X}^T, M_k \rangle + \Delta |V|_k^2(\mathbf{X}, \xi) \le \overline{V}_k^2, \end{aligned}$$

$$\forall k \in \mathcal{N}. \tag{7.3}$$

Notice that we include the forecast error explicitly in the active and reactive power balance equations. This is necessary because the bounds also change with the forecast error. For the notational simplicity, we write the constraints in (7.3) in the compact form:

$$\mathcal{A}(\mathbf{X}\mathbf{X}^T) + \Delta(\mathbf{X}, \xi) \leq \mathbf{0}_{6n}, \tag{7.4}$$

where $\mathcal{A} : \mathbb{R}^{2n\times 2n} \mapsto \mathbb{R}^{6n}$ is an affine operator, $\Delta : \mathbb{R}^{2n} \times \mathbb{R}^n \mapsto \mathbb{R}^{6n}$ characterizes the response to the random power injections, and the inequality is enforced componentwise.

Chance-constrained OPF

The randomness of ξ prohibits the application of most stochastic optimization algorithms to problems that involve constraint (7.4). As an alternative formulation, the chance constraints provide a practical way to enforce the stochastic constraint and quantify the satisfaction rate. Intuitively, the chance constraint requires that the stochastic constraint of (7.4) be satisfied with high probability. Let $\mathbb{P}_0(\cdot)$ be the probability with respect to the distribution of ξ and $\epsilon \in (0, 1]$ be the desired maximum violation probability. The *joint chance constraint* is defined as

$$\mathbb{P}_0 \left[\mathcal{A}(\mathbf{X}\mathbf{X}^T) + \Delta(\mathbf{X}, \xi) \leq \mathbf{0}_{6n}\right] \geq 1 - \epsilon. \tag{7.5}$$

Now, we can formulate the joint chance-constrained OPF (CCOPF) problem:chap6-

$$\min_{\mathbf{X}\in\mathbb{R}^{2n}} f(\mathbf{X}\mathbf{X}^T) \quad \text{s. t.} \quad \text{chance constraint (7.5)}, \tag{7.6}$$

The parameter ϵ effectively controls the trade-off between the reliability and efficiency of the solution to problem (7.6). With a smaller ϵ, the chance constraint becomes more restrictive and the operational cost becomes higher; and vice versa. Another similar chance constraint, named the *disjoint chance constraint*, can be written as

$$\mathbb{P}_0 \left[\mathcal{A}_k(\mathbf{X}\mathbf{X}^T) + \Delta_k(\mathbf{X}, \xi) \leq 0\right] \geq 1 - \epsilon, \quad \forall k \in [6n]. \tag{7.7}$$

In the main manuscript, we focus on the joint chance constraint and leave the analysis of the disjoint case, as well as their generalizations, to the appendix.

7.3 Reformulations of CCOPF

Although the CCOPF problem (7.6) is mathematically well-defined, the presence of uncertainty presents two challenges. First, since the change in the power system status is implicitly decided by the active power injection ξ through power flow equations, the function $\Delta(\cdot, \cdot)$ cannot be written in closed form. In Section 7.3, we derive a linear approximation of

$\Delta(\cdot,\cdot)$ and develop an efficient fixed-point iteration algorithm to find approximate solutions to the original non-linear formulation.

Second, the true distribution of ξ is unknown in most applications. Hence, it is not possible to enforce or verify the chance constraint (7.5). In Section 7.3, we propose DRO-based reformulations of CCOPF, which only rely on historical samples of ξ. We show that the chance constraint is satisfied by the DRO solutions with high probability in terms of the sample complexity.

Linearization and Fixed-point Iteration Algorithm

To avoid the computation cost of solving $\Delta(\cdot,\cdot)$ via power flow equations, we construct linear approximations to the implicit function and design an iterative algorithm that converges to a reliable approximation solution in practice. First, we utilize the prior information that the forecast errors are relatively small in practice and approximate $\Delta(\mathbf{X},\xi)$ with the first-order Taylor expansion around point $\xi = \mathbf{0}_n$. Namely, we have

$$\Delta(\mathbf{X},\xi) \approx \Delta(\mathbf{X},\mathbf{0}_n) + D_\Delta(\mathbf{X})\xi = D_\Delta(\mathbf{X})\xi,$$

where $D_\Delta(\mathbf{X}) \in \mathbb{R}^{6n\times n}$ is the Jacobian of $\Delta(\cdot,\cdot)$ with respect to the second input at point $(\mathbf{X},\mathbf{0}_n)$. Given vector $\mathbf{X}$, the Jacobian can be computed in closed form; see the appendix for the derivation of $D_\Delta(\mathbf{X})$. Then, the approximate joint chance constraint is given by

$$\mathbb{P}_0\left[\mathcal{A}(\mathbf{X}\mathbf{X}^T) + D_\Delta(\mathbf{X})\xi \le \mathbf{0}_{6n}\right] \ge 1-\epsilon. \tag{7.8}$$

The approximate disjoint chance constraint is defined in a similar way and we focus on the joint chance constraint in the remainder of this subsection. We note that this linearization approach is commonly used in CCOPF literature [194, 12].

Moreover, as proposed in [194], we further decouple the interaction between $\mathbf{X}$ and ξ through the fixed-point iteration. To be more specific, in the t-th iteration of the algorithm, the Jacobian $D_\Delta(\mathbf{X}_t)$ is fixed and we compute the new point $\mathbf{X}_{t+1}$ by solving problem (7.9). Note that we apply DRO-based algorithms in Section 7.3 to find solution $\mathbf{X}_{t+1}$ that satisfies the chance constraint with high probability. Then, the Jacobian at point $(\mathbf{X}_{t+1},\mathbf{0}_n)$ is computed and used in the next iteration.

The pseudo-code of the heuristic algorithm is provided in Algorithm 21. Intuitively, if the initialization is close to the solution, the fixed-point iteration enjoys fast convergence. In most applications, the forecast errors are small relative to the forecasted power injections and thus, our approximation scheme is considerably accurate in the following sense:

1. The first-order approximation $\Delta(\mathbf{X},\xi) \approx D_\Delta(\mathbf{X})\xi$ is acceptable under a wide range of operating conditions.

2. The robust solution to the chance-constrained problem is expected to be not too far from to the deterministic solution (i.e., the solution with $\xi = 0$).

Algorithm 21 Fixed-point iteration for joint CCOPF problem.

1: **Input:** tolerance μ, maximum violation probability ϵ.
2: **Output:** robust solution $\mathbf{X}$.
3: **Initialization:**

$$\mathbf{X}_0 \leftarrow \arg\min_{\mathbf{X}\in\mathbb{R}^{2n}} f(\mathbf{X}\mathbf{X}^T) \quad \text{s. t. } \mathcal{A}(\mathbf{X}\mathbf{X}^T) \leq \mathbf{0}_{6n}.$$

4: **for** $t = 0, 1, \ldots$ **do**
5: Update $\mathbf{X}_{t+1}$ to be the optimizer of

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{R}^{2n}} \quad & f(\mathbf{X}\mathbf{X}^T) \\ \text{s. t. } & \mathbb{P}_0\big[\mathcal{A}(\mathbf{X}\mathbf{X}^T) + D_\Delta(\mathbf{X}_t)\xi \leq \mathbf{0}_{6n}\big] \geq 1-\epsilon. \end{aligned} \tag{7.9}$$

▷ Solved by DRO-based algorithms.

6: **If** $\|\mathbf{X}_{t+1} - \mathbf{X}_t\| \leq \mu$, **return** $\mathbf{X}_{t+1}$.
7: **end for**

As a consequence, although there is no convergence guarantee, the fixed-point iteration exhibits efficient and robust convergence in practice; see the numerical experiments in Section 7.4 and [194]. We also numerically illustrate the approximation quality for benchmark power systems instances in the appendix.

Distributionally Robust Optimization Approach

In this subsection, we develop exact reformulations of the approximate chance constraints in Section 7.3, including but not limited to the joint chance constraint (7.8), based on DRO techniques. To preserve the generality of our results, we consider the general objective function and constraint function

$$g(\mathbf{X}) : \mathbb{R}^d \mapsto \mathbb{R}, \quad h(\mathbf{X}, \xi) : \mathbb{R}^d \times \mathbb{R}^n \mapsto \mathbb{R}^m,$$

where random vector $\xi \in \mathbb{R}^n$ obeys the distribution $\mathbb{P}_0$, and integers d and m are the size of input variable $\mathbf{X}$ and the number of constraints, respectively. In this subsection, we consider the optimization problem with stochastic constraints:

$$\min_{\mathbf{X}\in\mathbb{R}^d} g(\mathbf{X}) \quad \text{s.t. } h(\mathbf{X}, \xi) \leq \mathbf{0}_m. \tag{7.10}$$

Note that our theory can be extended to the case when the randomness ξ also incurs in the objective function g or the feasible set is a convex subset of $\mathbb{R}^d$. We focus on the simpler problem (7.10) since our target is to solve the CCOPF problem (7.9). We make the following assumption:

Assumption 16. The support of $\mathbb{P}_0$ belongs to a compact set $\Xi \subset \mathbb{R}^n$. Both functions $g(\cdot)$ and $h(\cdot,\cdot)$ are continuous. In addition, for every positive integer S and all realizations $\xi_1, \ldots, \xi_S \in \mathbb{R}^n$, problem

$$\min_{\mathbf{X}\in\mathbb{R}^d} g(\mathbf{X}) \quad \text{s.t. } h(\mathbf{X}, \xi_j) \leq \mathbf{0}_m, \ \forall j \in [S]$$

is feasible and has a finite optimal value.

In our formulation of the CCOPF problem (7.9), Assumption 16 is satisfied unless, for instance, the reserve capacity of the conventional generators is insufficient to compensate for some realizations of the forecast error. In this case, CCOPF would be infeasible.

In practice, the true distribution $\mathbb{P}_0$ is unknown and only limited historical samples may be available. Suppose that there are S independently and identically distributed samples, $\xi^1, \ldots, \xi^S$, generated from the distribution $\mathbb{P}_0$. We define the empirical distribution of ξ as

$$\hat{\mathbb{P}}_S := \frac{1}{S}\sum\nolimits_{k\in[S]} \delta_{\xi^k},$$

where δ_ξ is the Dirac measure at ξ. The goal of the DRO approach is to find robust solutions that satisfy the chance constraint with high probability using empirical distribution $\hat{\mathbb{P}}_S$. Define the ambiguity set

$$\mathcal{D}_r(\mathbb{P}) \coloneqq \{\mathbb{P}' \in \mathcal{P} \mid I(\mathbb{P}, \mathbb{P}') \leq r\}, \quad \forall \mathbb{P} \in \mathcal{P},$$

where $I(\cdot,\cdot)$ is the relative entropy [54], $r > 0$ is the radius and $\mathcal{P}$ is the family of Borel distributions with support in Ξ. The robustness of DRO solutions is guaranteed by the satisfaction of chance constraints under all distributions in the ambiguity set $\mathcal{D}_r(\hat{\mathbb{P}}_S)$. Other distributional metrics, such as the Wasserstein metric, are considered in CCOPF literature [186, 12]. In this chapter, however, we use the relative entropy due to the strong optimality guarantees it can provide; see Theorems 75-76 and [54, 222]. Intuitively, the large deviation theory guarantees that the relative entropy between the true data-generation distribution and the empirical distribution can be bounded by a value that depends on the sample size [54]. Hence, the true distribution is contained in the ambiguity set with high probability and the relative entropy-based ambiguity set is the "smallest" ambiguity set with such property [222].

For problem (7.10), the joint chance constraint is given by

$$\mathbb{P}_0[h(\mathbf{X}, \xi) \leq \mathbf{0}_m] \geq 1 - \epsilon. \tag{7.11}$$

Remark 10. More generally, our results can be extended to the case when the joint constraints are defined by a convex cone

$$\mathbb{P}_0\left[\omega^T h(\mathbf{X}, \xi) \leq 0, \quad \forall \omega \in \mathcal{W}\right] \geq 1 - \epsilon, \tag{7.12}$$

where $\mathcal{W}$ is the convex cone spanned by weight vectors[1] $\omega_1, \ldots, \omega_L$. Constraint (7.12) reduces to the cardinal case (7.11) when $L = m$ and $\omega_\ell = \mathbf{e}_\ell$ for all $\ell \in [m]$.

[1] A vector $\omega \in \mathbb{R}^m$ is called a weight vector if $\omega \geq \mathbf{0}_m$ and $\mathbf{1}_m^T\omega = 1$.

Define the α-quantile

$$q_\alpha(F, \mathbb{P}) := \sup \{q \mid \mathbb{P}[F(\xi) \leq q] \leq \alpha\}$$

for all $\alpha \in [0,1]$, function $F(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}$ and distribution $\mathbb{P} \in \mathcal{P}$. Then, the chance constraint (7.11) can be equivalently written as

$$\begin{aligned} & \mathbb{P}_0[h_\ell(\mathbf{X}, \xi) \leq 0, \quad \forall \ell \in [m]] \geq 1 - \epsilon \\ \iff & \mathbb{P}_0\left[\bar{h}_{\mathbf{X}}(\xi) \leq 0\right] \geq 1 - \epsilon \\ \iff & q_{1-\epsilon}\left(\bar{h}_{\mathbf{X}}, \mathbb{P}_0\right) \leq 0, \end{aligned} \tag{7.13}$$

where we define

$$\bar{h}_{\mathbf{X}}(\xi) \coloneqq \max_{\ell \in [m]} h_\ell(\mathbf{X}, \xi).$$

Adopting language from [222], we first introduce the distributionally robust predictor of the α quantile.

Definition 31 (Distributionally Robust Predictor)**.** For all $\epsilon \in [0,1]$, $r > 0$, $\mathbf{X} \in \mathbb{R}^d$ and $\mathbb{P} \in \mathcal{P}$, the distributionally robust predictor is defined as

$$\hat{q}_{1-\epsilon,r,\mathbb{P}}(\mathbf{X}) := \sup_{\mathbb{P}' \in \mathcal{D}_r(\mathbb{P})} q_{1-\epsilon}\left(\bar{h}_{\mathbf{X}}, \mathbb{P}'\right).$$

For notational simplicity, when there is no confusion about $\hat{\mathbb{P}}_S$, we denote predictor $\hat{q}_{1-\epsilon,r,\hat{\mathbb{P}}_S}$ as $\hat{q}_{1-\epsilon,r,S}$.

Intuitively, the distributionally robust predictor is the *worst-case* α-quantile over all distributions in the relative entropy ball $\mathcal{D}_r(\mathbb{P})$. In the following lemma, we prove that the distributionally robust predictor is either a quantile of $\bar{h}_{\mathbf{X}}$ under the empirical distribution $\hat{\mathbb{P}}_S$ or the maximum value

$$h^*_{\mathbf{X}} \coloneqq \max_{\xi \in \Xi} \bar{h}_{\mathbf{X}}(\xi).$$

Lemma 53. *For all $\epsilon \in [0,1]$ and $r > 0$, there exists an integer $k(\epsilon, r, S) \in [S+1]$ such that*

$$\hat{q}_{1-\epsilon,r,S}(\mathbf{X}) = \bar{h}_{k(\epsilon,r,S),\hat{\mathbb{P}}_S}(\mathbf{X}), \quad \forall \mathbf{X} \in \mathbb{R}^d,$$

where $\bar{h}_{k,\hat{\mathbb{P}}_S}(\mathbf{X})$ is the k-th smallest value in $\{\bar{h}_{\mathbf{X}}(\xi^j), j \in [S]\} \cup \{h^*_{\mathbf{X}}\}$. When there is no confusion, we denote for the notational simplicity

$$k \coloneqq k(\epsilon, r, S) \quad \textit{and} \quad \bar{h}_k(\mathbf{X}) \coloneqq \bar{h}_{k(\epsilon,r,S),\hat{\mathbb{P}}_S}(\mathbf{X}).$$

In the case when $k = S+1$, the evaluation of $\bar{h}_{S+1}(\mathbf{X})$ requires the knowledge of $h^*_{\mathbf{X}}$, which may be unknown in practice. Hence, we focus on the case when $k \in [S]$, which can be guaranteed by choosing suitable values of ϵ and r. Furthermore, the value of k can be computed by solving a convex optimization problem; see problem (7.28) in the appendix.

Then, we define the distributionally robust prescriptor.

Definition 32 (Distributionally Robust Prescriptor)**.** For all $\epsilon \in [0,1]$ and $r > 0$, the distributionally robust prescriptor $\hat{\mathbf{X}}_{1-\epsilon,r,\mathbb{P}}$ is a quasi-continuous function of $\mathbb{P}$ that solves

$$\min_{\mathbf{X}\in\mathbb{R}^d} g(\mathbf{X}) \quad \text{s.t. } \hat{q}_{1-\epsilon,r,\mathbb{P}}(\mathbf{X}) \leq 0. \tag{7.14}$$

Similarly, when there is no confusion about $\hat{\mathbb{P}}_S$, we denote prescriptor $\hat{\mathbf{X}}_{1-\epsilon,r,\hat{\mathbb{P}}_S}$ as $\hat{\mathbf{X}}_{1-\epsilon,r,S}$.

By Lemma 53, the feasible set of problem (7.14) is a subset of $\{\mathbf{X} \in \mathbb{R}^d \mid \bar{h}_k(\mathbf{X}) \leq 0\}$. Thus, combining with Assumption 16, problem (7.14) has a finite optimal value and the distributionally robust prescriptor $\hat{\mathbf{X}}_{1-\epsilon,r,S}$ is well-defined.

Now, we provide a mixed-integer reformulation of (7.14) to compute the distributionally robust prescriptor. Choosing $C > 0$ to be a sufficiently large constant, we show that the distributionally robust prescriptor is a solution to

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{R}^d, \mathbf{b}\in\mathbb{Z}^S} \quad & g(\mathbf{X}) \\ \text{s.t. } & h_\ell(\mathbf{X}, \xi^j) \leq C\mathbf{b}_j, \quad \forall \ell \in [m],\ j \in [S], \\ & \mathbf{1}_S^T \mathbf{b} \leq S - k, \quad \mathbf{b}_j \in \{0,1\}, \quad \forall j \in [S]. \end{aligned} \tag{7.15}$$

Intuitively, the constraints in (7.15) enforce the joint chance constraint under the empirical distribution $\hat{\mathbb{P}}_S$. Namely, the constraint $h(\mathbf{X}, \xi^j) \leq \mathbf{0}_m$ is satisfied by at least k samples. In the next theorems, we show that the chance constraint under the true distribution $\mathbb{P}_0$ can also be guaranteed by choosing k to be slightly larger than $(1-\epsilon)S$.

Theorem 74. *The solution to (7.15) is a distributionally robust prescriptor.*

For the CCOPF problem (7.9), the problem (7.15) is equivalent to a mixed-integer QCQP. In Section 7.3, we apply the semi-definite relaxation to the QCQP and when the relaxation is exact, problem (7.9) is equivalent to a *mixed-integer semi-definite program* (MISDP), which can be handled by off-the-shelf convex optimization solvers.

Finally, we establish the theoretical properties of the distributionally robust prescriptor. First, we prove that the distributionally robust prescriptor satisfies joint chance constraint (7.13) with high probability in terms of the sample complexity S.

Theorem 75. *For all $\epsilon \in [0,1]$ and $r > 0$, it holds that*

$$\mathbb{P}_\infty\left[q_{1-\epsilon}\left(\bar{h}_{\hat{\mathbf{X}}_{1-\epsilon,r,S}}, \mathbb{P}_0\right) \leq 0\right] \geq 1 - \exp\left[-rS + o(S)\right], \tag{7.16}$$

where $\mathbb{P}_\infty$ is the probability measure of the sample path space of ξ under distribution $\mathbb{P}_0$. Furthermore, we have

$$\mathbb{P}_\infty\left[h_\ell\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}, \xi\right) \leq 0, \quad \forall \ell \in [m]\right] \geq 1 - \epsilon - \exp\left[-rS + o(S)\right]. \tag{7.17}$$

In the regime when the support Ξ is a finite set, we can apply the strong large deviation principle [222] and derive the following finite-sample bound in the same way as Theorem 75:

$$\mathbb{P}_\infty\left[q_{1-\epsilon}\left(\bar{h}_{\hat{\mathbf{X}}_{1-\epsilon,r,S}}, \mathbb{P}_0\right) \le 0\right] \ge 1 - (S+1)^d e^{-rS}. \tag{7.18}$$

Moreover, we show that the distributionally robust prescriptor achieves the minimum operational cost over all decisions that asymptotically satisfy the joint chance constraint (7.13).

Theorem 76. *Suppose that prescriptor $\tilde{\mathbf{X}}_{1-\epsilon,r,\mathbb{P}} \in \mathbb{R}^d$ is a quasi-continuous function of $\mathbb{P}$ and satisfies constraint (7.16). Then, we have*

$$\mathbb{P}_\infty\left[g\left(\tilde{\mathbf{X}}_{1-\epsilon,r,S}\right) < g\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}\right)\right] = 0,$$

where we denote $\tilde{\mathbf{X}}_{1-\epsilon,r,S} \coloneqq \tilde{\mathbf{X}}_{1-\epsilon,r,\hat{\mathbb{P}}_S}$.

Compared with existing DRO formulations [186, 12], our formulation provides stronger guarantees in the following two senses. First, the DRO solution $\hat{\mathbf{X}}_{1-\epsilon,r,S}$ achieves the minimum possible generation cost over all robust solutions that satisfy the joint chance constraint (7.16). This optimality property arises from the choice of the relative entropy for the ambiguity set, and such property cannot be established by other distributional metrics, although the Wasserstein metric can provide similar high-probability bounds [166]. Second, the mixed-integer reformulation (7.15) is *exact*. In contrast, existing literature considered parameterized approximations to the ambiguity set, such as the hyper-rectangle [186] and the ellipsoid [12]. In practice, however, there is no guarantee that the ambiguity set is of the specified shape and thus, the approximate DRO solution is usually overly conservative; see the comparison results in Section 7.4.

In practice, it is preferable for the user to first specify k and then compute the optimal ϵ and r to maximize the right-hand side of (7.17). Given $k \in [S]$ and $\epsilon \in [1 - k/S, 1]$, the maximal radius r such that $k(\epsilon, r, S) = k$ is given by

$$r = -\frac{k}{S}\log\left(\frac{S(1-\epsilon)}{k}\right) - \frac{S-k}{S}\log\left(\frac{S\epsilon}{S-k}\right),$$

where we define $0\log 0 = 0$. Therefore, when the sample size S is sufficiently large, we ignore the $o(S)$ term on the right-hand side of (7.17) and solve the maximization problem

$$\epsilon^*_{k,S} \coloneqq \operatorname*{arg\,max}_{\epsilon\in[1-k/S,1]} \; 1 - \epsilon - \frac{S^S}{k^k(S-k)^{S-k}}(1-\epsilon)^k\epsilon^{S-k}, \tag{7.19}$$

where we define $0^0 = 1$. The solution to the above problem maximizes the right-hand side of (7.17) and can be found by the bi-section algorithm.

Semi-definite Relaxation

In the last part of this section, we deal with the non-convexity of problem (7.9) induced by the quadratic parameterization $\mathbf{X}\mathbf{X}^T$. In the context of CCOPF problem, even with a fixed integer vector $\mathbf{b}$, problem (7.15) is still a non-convex QCQP and can be $\mathcal{NP}$-hard to solve in the worst case. To achieve the efficient and reliable operation of large-scale power systems, various techniques have been proposed to reduce the optimization complexity by utilizing the special structures of real-world power circuits. In literature, the semi-definite relaxation is widely applied to transform the non-convex QCQP to a semi-definite program (SDP); see [137] and [153] for semi-definite relaxations of OPF. More specifically, after making the change-of-variables $\mathbf{W} \coloneqq \mathbf{X}\mathbf{X}^T$ and dropping the rank constraint $\text{rank}(W) = 1$, we can apply the distrubutionally robust reformulation (7.15) to solve problem (7.9) by

$$\begin{aligned} \min_{\mathbf{W}\in\mathbb{S}^{2n}_{+},\mathbf{b}\in\mathbb{Z}^S} \quad & f(\mathbf{W}) \\ \text{s.t. } & \mathcal{A}(\mathbf{W}) + D_{\Delta}(\mathbf{X}_t)\xi^j \le C\mathbf{b}_j \cdot \mathbf{1}_{6n}, \quad \forall j \in [S], \\ & \mathbf{1}_S^T\mathbf{b} \le S - k, \quad \mathbf{b}_j \in \{0,1\}, \quad \forall j \in [S]. \end{aligned} \tag{7.20}$$

Note that the SDP part of (7.20) can be further written in the standard form using the Schur complement [137]. We denote problem (7.20) as the distributionally robust CCOPF (DRCCOPF) problem. With the rank constraint dropped, problem (7.20) is a MISDP, which can be solved efficiently by various solvers (e.g., YALMIP [150]).

The graphical structures of practical power networks, reflected in the algebraic properties of operator $\mathcal{A}$, guarantee the exactness of the semi-definite relaxation [153, 208, 207]. In this chapter, we make the assumption that the network admits an exact semi-definite relaxation, and our method can be readily extended to other formulations including DC OPF. Under the exact relaxation assumption, we are able to recover the rank-1 solution from the MISDP solution [137]. Therefore, in Step 5 of Algorithm 21, we first solve the relaxation (7.20) to find $\mathbf{W}_{t+1}$ and then generate the rank-1 solution $\mathbf{X}_{t+1}$ for the next iteration.

To further reduce the computational complexity of solving problem (7.20), we develop a heuristic algorithm that finds approximate solutions by solving a small number of SDPs. Intuitively, the proposed algorithm searches for the optimal integer vector $\mathbf{b}$ in a greedy way and avoids the mixed-integer part in problem (7.20). With a given k, the algorithm removes the "most restrictive sample" among the $k+1$ samples selected in the case of $k+1$. The heuristic algorithm is able to find nearly optimal approximate solutions for benchmark power systems and requires a much shorter running time; see Section 7.4 and the appendix for more details.

7.4 Demonstration on IEEE Test Cases

In this section, we apply our results to solve the joint CCOPF problem (7.6) on the IEEE 14- and 118-bus test cases. All system parameters are taken from the case data in MATPOWER [259, 258]. We make the following modifications to the system parameters:

1. As suggested in [137], a small resistance of 10^{-4} per-unit is added to each transformer to insure an exact semi-definite relaxation.

2. Wind generators are installed at n_W buses randomly chosen from load buses with nonzero active loads. The forecasted output of a generator is equal to a proportion η of the pre-installation load at that bus.

Note that wind generators could also have been installed at slack or generator buses. In our model, wind generator forecast errors account for the entire random power injection; that is, we assume loads are deterministic. Wind output forecast errors are taken from hour-ahead forecast errors from the National Renewable Energy Laboratory's Wind Integration National Dataset, which contains simulated forecast and output data for over 120,000 sites in the United States over seven years [100]. Specifically, each load bus chosen for a wind generator is assigned to a randomly selected site in Alameda County, California. The forecast errors are then scaled appropriately, assuming that the forecasted output is half of the installed capacity of the turbine. Turbines are selected from a single county and thus, their outputs are correlated; exploiting the correlations between random variables in different constraints is an important feature of joint chance-constrained methods.

For each network, we implement Algorithm 21 and use the DRO formulation (7.20) to solve (7.9). We use $\hat{S}$ training samples drawn independently from the full path of S samples. For comparison, we also solve (7.9) using the method from [12], which approximates a Wasserstein metric-based ambiguity set using a minimum-volume ellipsoid in the parameter space. To the best of our knowledge, this is the least conservative approximation of a metric-based ambiguity set for joint chance constraints in the literature (with ours being the first exact reformulation). The MISDP solver, the greedy algorithm and the method from [12] is called DRCCOPF-KL, DRCCOPF-G and DRCCOPF-E, respectively. For benchmarking, we also compare with the robust optimization (RO) approach, where all constraints satisfied for all S available samples, and the deterministic OPF approach, where each wind generator simply outputs its forecasted value. As the forecast errors are approximately zero-mean in practice, deterministic OPF essentially enforces the constraints in expectation. The RO approach gives the most conservative solution.

We simulate the IEEE 14- and 118-bus systems with parameters $\eta = 0.9$, $\cos\phi = 1$, and $n_W = 6$ for the 14-bus system and $n_W = 45$ for the 118-bus system. We generate $S = 2185$ samples and $\hat{S} = 200$ samples are used as training samples. Our hour-ahead forecast errors were taken from June 1 to August 31, 2007. To evaluate the performance under different robustness requirements, we performed a sweep over k from 180 to 200. Recall that for a given k and $\hat{S}$, the optimal ϵ can be computed from (7.19).

All simulations are performed in MATLAB 2023b. The MISDP problem (7.20) is solved using YALMIP [150] and all convex problems are solved using CVX [89]. We note that all simulation results in this section describe the performance of the voltage setpoint recovered from the solution to the MISDP (7.20) *after* applying Corollary 1 of [137].

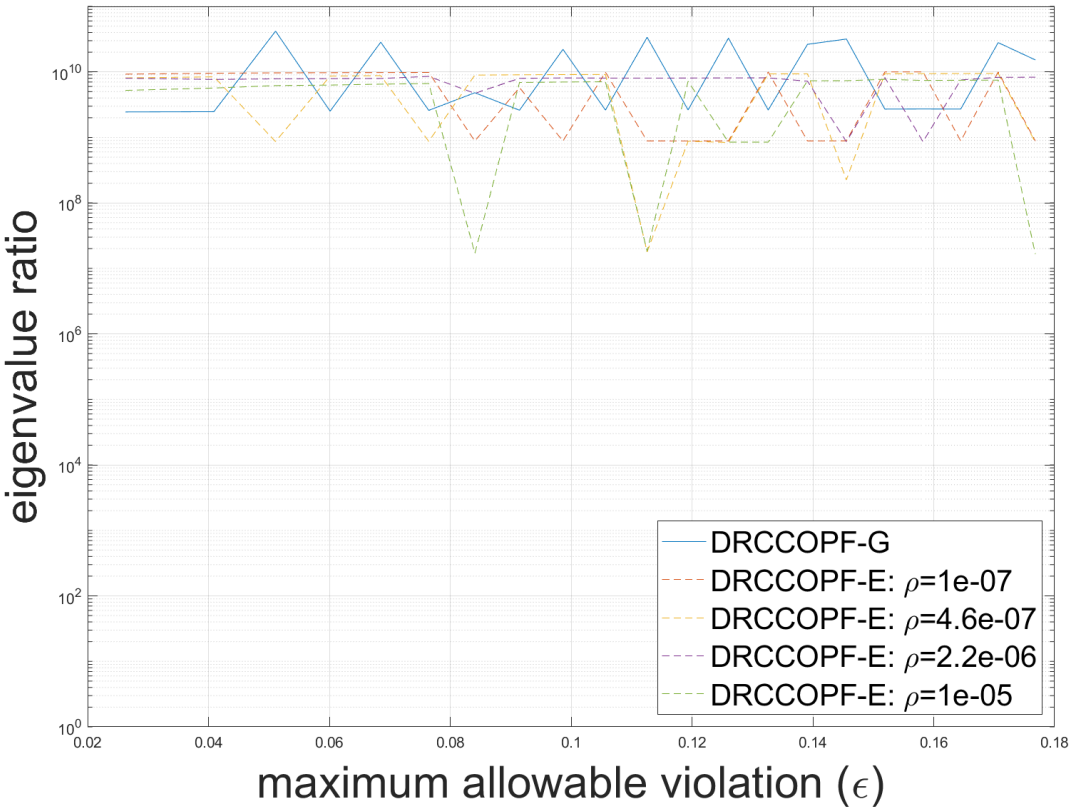


Figure 7.4.1: Ratio between the second and third largest eigenvalues for IEEE 118-bus system.

Solution Recovery

First, we analyze the exactness of the semi-definite relaxation in Section 7.3. Molzahn *et al.* [167] proposed the ratio between the second and third largest eigenvalues of the solution to the relaxed problem as a metric for exactness. If the eigenvalue ratio is high, the true power injections and voltage magnitudes at each bus will be close to those computed from the solution to the relaxed problem. We compute the eigenvalue ratio of solutions generated by DRCCOPF-G and DRCCOPF-E for the 118-bus system. As seen in Figure 7.4.1, both approaches produce solutions with eigenvalue ratios higher than, in the logarithmic sense, to the benchmark value of 10^7 from [167]. However, DRCCOPF-G often leads to slightly higher-quality solutions with ratios closer to 10^9 or 10^10. This is a promising feature of the method proposed here.

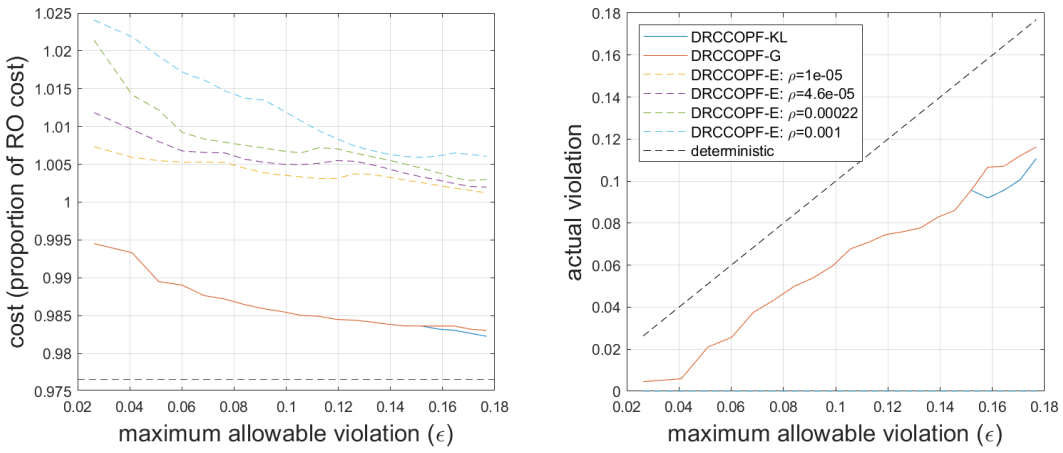


Figure 7.4.2: Performance comparison for 14-bus system.

Efficiency and Robustness

Now, we compare the generation cost (efficiency) of different solutions under the same maximum allowable violation rate (robustness). For the DRCCOPF-G and DRCCOPF-KL approach, the violation rate can be effectively controlled by parameter k, while the DRCCOPF-E approach controls the robustness by parameters ϵ and ρ. We use the RO approach as the benchmark, which corresponds to the most conservative setting, and other costs are given as a proportion of the RO cost for reference. As another benchmark, we also illustrate the cost of deterministic OPF approach. All reasonable robust methods should always have higher costs than the deterministic approach, which, as a result, represents a lower bound on the achievable efficiency.

The results of the 14- and 118-bus systems are plotted in Figures 7.4.2 and 7.4.3, respectively. The left subplots compare the cost of the solution of DRCCOPF-G with that of DRCCOPF-E for several different Wasserstein radii ρ. The right subplots give the realized constraint violation rates for both methods on all of the S available samples. Note that the deterministic OPF method violates the constraints at a rate of more than 97%. This is a result of *joint* chance constraints, since violating even a single constraint constitutes a violation. Notice that DRCCOPF-E never violates the joint chance-constraint for any radius. For the 14-bus case, we also compare with the solutions of DRCCOPF-KL. We can observe that the solutions of DRCCOPF-KL and DRCCOPF-G exhibit very similar behaviors, which imply that the greedy algorithm finds nearly optimal solutions. Therefore, we focus on the DRCCOPF-G approach in the following discussion.

First, DRCCOPF-E is qualitatively more conservative than DRCCOPF-G in the sense that it fails to exploit the tolerance provided by non-zero values of ϵ to achieve lower objective values. In fact, for small ϵ and large ρ, DRCCOPF-E is even more inefficient than RO as a consequence of its approximation of the ambiguity set. By contract, distributionally robust

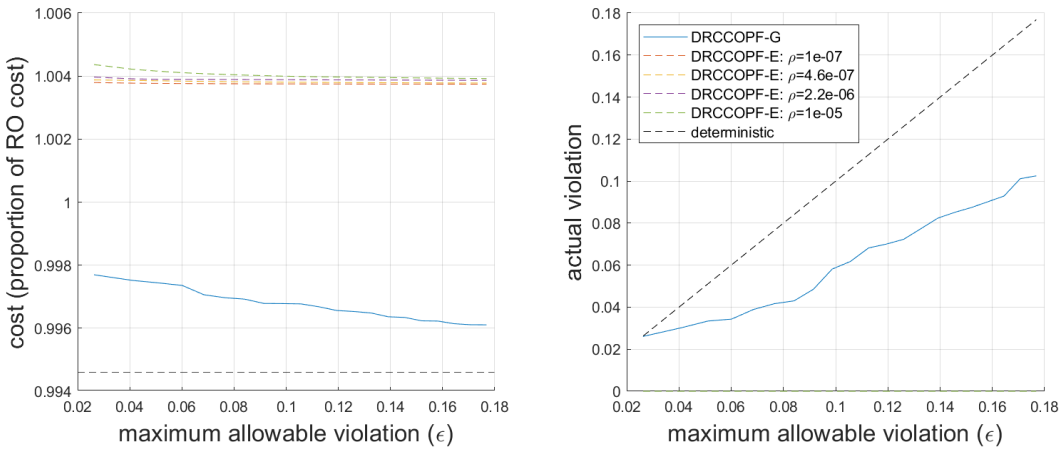


Figure 7.4.3: Performance comparison for 118-bus system.

methods are generally less conservative than RO. As shown in our results, the DRCCOPF-G approach closes a significant fraction of the gap between the robust and deterministic optimizations by approaching but never exceeding the prescribed maximum violation rate.

Moreover, the running time of DRCCOPF-G is comparable with that of DRCCOPF-E. For the 118-bus system, DRCCOPF-G takes an average of 412 seconds to find the solution using a single 3.79-GHz CPU, while the DRCCOPF-E finishes in 235 seconds on average. However, the greedy search structure of DRCCOPF-G approach allows the application of parallel computing techniques, which will significantly improve the computational efficiency of the algorithm.

Furthermore, the DRCCOPF-G approach always generates a feasible solution as long as the deterministic problem is feasible for every sample (that is, as long as Assumption 16 is satisfied). This is because DRCCOPF-G basically requires satisfaction of the constraints for a subset of k samples. DRCCOPF-E and other existing approximate methods, by contrast, compute uncertainty margins indirectly using training samples, possibly rendering the problem infeasible. Indeed, we have observed that DRCCOPF-E is only feasible for a limited range of parameters and fail to find a solution for relatively large values of η (more renewable penetration) and small values of ϵ.

Demonstration on Selected Constraints

Finally, we select a few constraints to highlight the difference in performance. We focus on the generator outputs of three selected buses in the 14-bus system and plot the distribution of post-contingency generator outputs of DRCCOPF-G and DRCCOPF-E solutions. The results are shown in Figure 7.4.4, where the deterministic setpoint and lower bound are included for reference. Since the upper bound is not violated, it is not included in the histograms. From the results, we can see that DRCCOPF-G is significantly closer to the

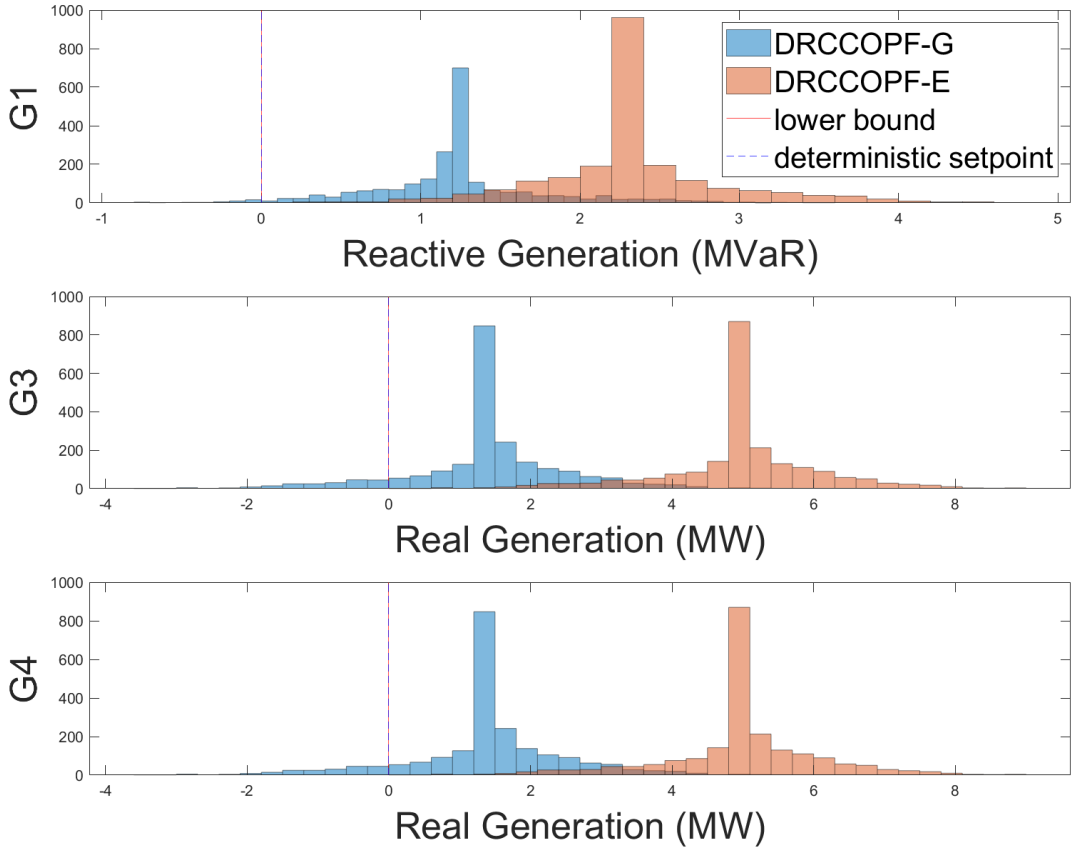


Figure 7.4.4: Distribution of selected generator outputs for DRCCOPF-G and DRCCOPF-E for the 14-bus system.

lower bound for real or reactive power for three selected generators, while DRCCOPF-E produces overly conservative power outputs.

Appendix

7.A Heuristic Algorithm for MISDP

In this section, we describe a heuristic algorithm for the MISDP problem (7.20). Comparing with the internal algorithm of the YALMIP solver, the proposed algorithm achieves a much better computational complexity and is able to find solutions of the same quality, in terms of the objective function value and the constraint satisfaction rate.

Intuitively, the heuristic algorithm searches for the optimal integer vector $\mathbf{b}$ in a greedy way by gradually reducing the value of k. The algorithm starts with the most conservative case when $k = S$. In this case, the only feasible vector $\mathbf{b}$ is the zero vector $\mathbf{0}_S$ and problem (7.20) reduces to a SDP problem, which can be efficiently solved by a variety of optimization solvers. Then, for each integer $k_0 < S$, the algorithm searches for the optimal $\mathbf{b}$ for the case $k = k_0$ from that for the case $k = k_0 + 1$. More specifically, suppose that we have obtained an approximate solution $(\mathbf{W}^{k_0+1}, \mathbf{b}^{k_0+1})$ for the case when $k = k_0 + 1$. For each index $\ell \in [S]$ such that $\mathbf{b}_\ell^{k_0+1} = 0$, we construct the vector $\mathbf{b}^{k_0,\ell} \in \mathbb{R}^S$ by

$$\mathbf{b}_\ell^{k_0,\ell} = 1, \quad \mathbf{b}_j^{k_0,\ell} = \mathbf{b}_j^{k_0+1}, \quad \forall j \in [S]\backslash\{\ell\}, \tag{7.21}$$

and we solve the following SDP problem:chap6-

$$\begin{aligned} \min_{\mathbf{W}\in\mathbb{S}_+^{2n}} \quad & f(\mathbf{W}) \\ \text{s. t. } & \mathcal{A}(\mathbf{W}) + D_\Delta \xi^j \le C\mathbf{b}_j^{k_0,\ell} \cdot \mathbf{1}_{6n}, \quad \forall j \in [S]. \end{aligned} \tag{7.22}$$

Let f_ℓ^* be the optimal objective function value of the above SDP problem. Then, the greedy algorithm chooses the most restrictive sample ξ^{ℓ^*} by

$$\ell^* := \arg\min\nolimits_{\ell\in[S]} f_\ell^*, \quad \text{s. t. } \mathbf{b}_\ell^{k_0+1} = 0.$$

After removing the most restrictive sample, the approximate solution for the case $k = k_0$ is given by $(\mathbf{W}^{k_0}, \mathbf{b}^{k_0,\ell^*})$, where $\mathbf{W}^{k_0}$ is the solution to problem (7.22) with $\ell = \ell^*$. To further reduce the computational cost, we only need to consider indices $\ell \in [S]$ such that $\mathbf{b}_\ell^{k_0+1} = 0$ and problem (7.20) has active constraints with sample ξ^ℓ. Namely, the following componentwise inequality holds with equality for some components:

$$\mathcal{A}(\mathbf{W}^{k_0+1}) + D_\Delta \xi^\ell \le C\mathbf{b}_\ell^{k_0+1} \cdot \mathbf{1}_{6n}. \tag{7.23}$$

Algorithm 22 Greedy algorithm for MISDP problem (7.20).

Input: samples $\xi^1, \ldots, \xi^S$, integer k_0.
Output: robust solution $\mathbf{X}$.
Initialization: let $\mathbf{b}^S \leftarrow \mathbf{0}_S$.
for $k = S-1, S-2, \ldots, k_0$ **do**
    Let $\mathcal{L} \subset [S]$ be the set of indices such that:
    1. $\mathbf{b}_\ell^{k+1} = 0$;
    2. inequality (7.23) does not hold strictly with sample ξ^ℓ.
    for $\ell \in \mathcal{L}$ **do**
        Construct vector $\mathbf{b}^{k_0,\ell} \in \mathbb{R}^S$ by (7.21).
        Apply Algorithm 21 to solve problem (7.22).
        Let f_ℓ^* be the optimal objective value.
    end for
    Let $\ell^* \leftarrow \arg\min_{\ell \in \mathcal{L}} f_\ell^*$ and $\mathbf{b}^k \leftarrow \mathbf{b}^{k,\ell^*}$.
end for
Apply Algorithm 21 to solve problem (7.21) with $\mathbf{b} = \mathbf{b}^{k_0}$.
Return the solution $\mathbf{X}$ to the above problem.

This is because otherwise if the above inequality holds strictly, the optimal objective value will be the same with sample ξ^ℓ (i.e., $\mathbf{b} = \mathbf{b}^{k_0+1}$) and without sample ξ^ℓ (i.e., $\mathbf{b} = \mathbf{b}^{k_0,\ell}$). Therefore, the value f_ℓ^* will not be the minimum among all choices of ℓ. The pesudo-code of the greedy is provided in Algorithm 22. We note that Algorithm 22 operates in the "top-down" style in the sense that it gradually decreases the value of k from S. Similarly, we can develop the "bottom-up" version of the greedy algorithm, which gradually increases the value of k from 0. In practice, the top-down algorithm is preferred since a large value of k is usually chosen to ensure a high constraint satisfaction rate. Therefore, the top-down algorithm requires fewer iterations to reach the targeted value of k.

Since the greedy algorithm approximates the solution to MISDP (7.20) with a small number of SDP problems, the running time of the greedy algorithm is much better than that of the off-the-shelf solvers, e.g., YALMIP. For example, using a single 3.79-GHz CPU, the greedy algorithm takes an average of 412 seconds to solve the 118-bus case for each $k \in \{180, \ldots, 200\}$, while the YALMIP solver takes more than two hours for a single iteration of Algortihm 21, which includes solving a single MISDP problem instance. In addition, we compare the solutions generated by the greedy algorithm and the YALMIP solver for the 14-bus system. The YALMIP solutions are shown to be optimal up to a small gap between the lower and upper bounds. The results are plotted in Figure 7.4.2. We can see that the greedy algorithm is able to find solutions of almost the same quality as the global optima. To be more concrete, the objective function values and constraint satisfaction rates of the two solutions are very close.

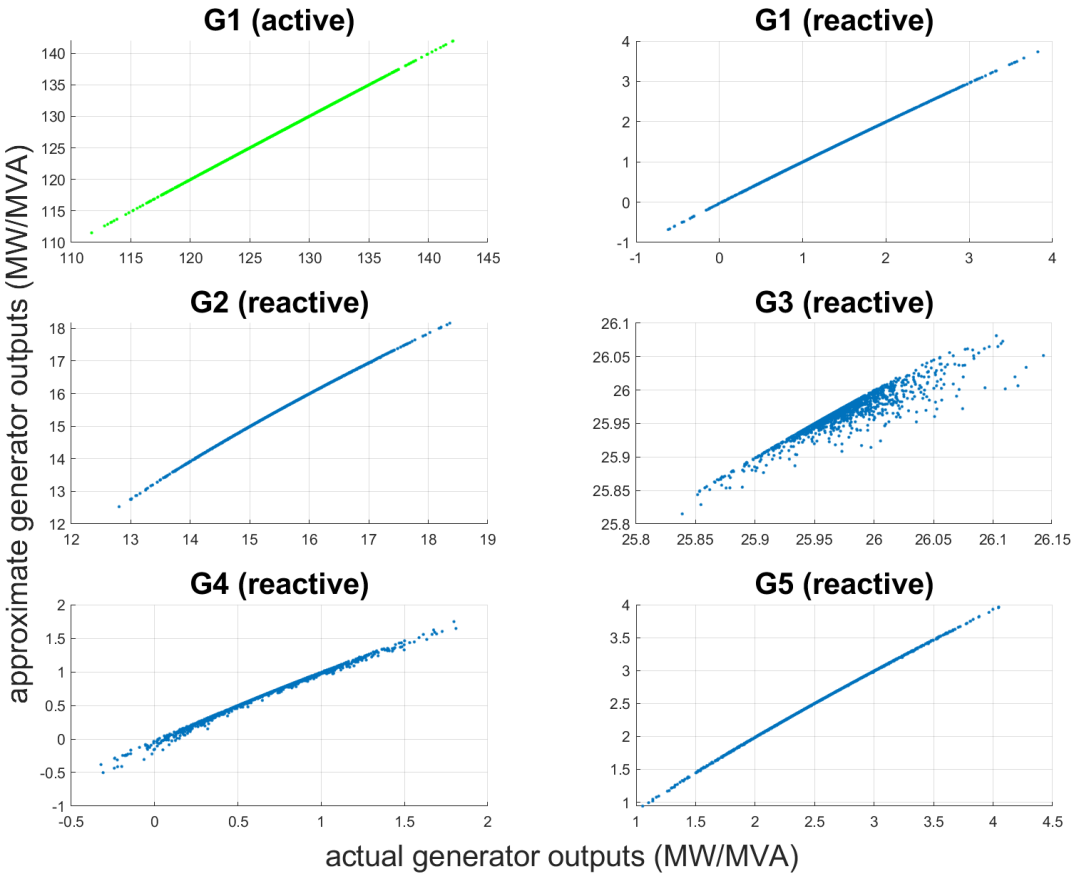


Figure 7.B.1: Actual and approximate post-contingency generator outputs.

In summary, although there is no theoretical optimality guarantee, the heuristic greedy algorithm is able to find near-optimal solutions in an efficient and robust way for benchmark power systems.

7.B Linearization Accuracy

To evaluate the accuracy of the first-order approximation presented in Section 7.3 for our test case, we compare the actual and first-order approximate post-contingency system responses for the 14-bus system. The responses are computed for all available samples. The operating point is obtained by DRCCOPF-KL with $k = \hat{S} = 100$. Actual system responses are computed by running the MATPOWER power flow solver. The generator outputs (on separate subplots) and squared voltage magnitudes (on a single plot, with a different color for each bus) are given in Figures 7.B.1 and 7.B.2, respectively. As shown by the figures,

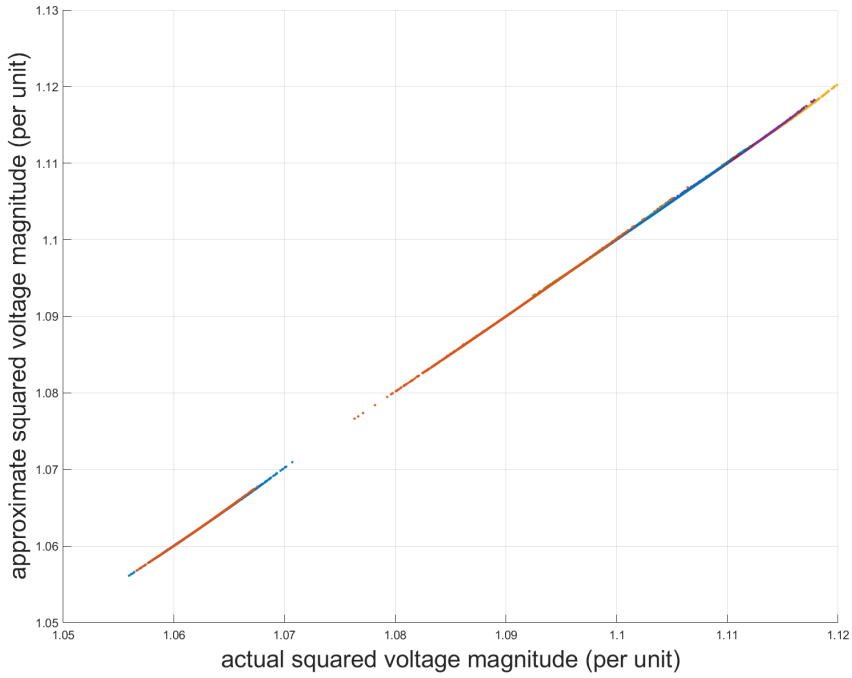


Figure 7.B.2: Actual and approximate post-contingency squared voltage magnitudes.

the approximation is quite accurate and appears appropriate for our problem instance. In Section 7.4, the approximate system response will be used to compute violation rates.

7.C Derivation of Sensitivity Factor

In this section, we derive the sensitivity factor $D_\Delta(\mathbf{X})$, which is defined as the Jacobian of $\Delta(\mathbf{X}, \xi)$ with respect to ξ. More specifically, by the definition of $D_\Delta(\mathbf{X})$ and the constraints (7.3), we have

$$D_\Delta(\mathbf{X}) = \begin{bmatrix} \partial_\xi \Delta P(\mathbf{X}, \xi) - \mathbf{I}_n \\ -\partial_\xi \Delta P(\mathbf{X}, \xi) + \mathbf{I}_n \\ \partial_\xi \Delta Q(\mathbf{X}, \xi) - \gamma \mathbf{I}_n \\ -\partial_\xi \Delta Q(\mathbf{X}, \xi) + \gamma \mathbf{I}_n \\ \partial_\xi \Delta |V|^2(\mathbf{X}, \xi) \\ -\partial_\xi \Delta |V|^2(\mathbf{X}, \xi) \end{bmatrix} \in \mathbb{R}^{6n \times n}. \tag{7.24}$$

Therefore, the problem reduces to the calculation of the partial derivatives of ΔP, ΔQ and $\Delta |V|^2$ with respect to ξ.

We begin with the first-order approximation of the power flow equations around an operating point $\mathbf{X}$:

$$\mathbf{J}\begin{bmatrix}\Delta\Theta\\ \Delta|V|\end{bmatrix} = \begin{bmatrix}\Delta P\\ \Delta Q\end{bmatrix}, \tag{7.25}$$

where $\Theta \in \mathbb{R}^n$ is the vector of voltage angles and $\mathbf{J} \in \mathbb{R}^{2n\times 2n}$ is the Jacobian matrix, which can be computed by the implicit function theorem. For convenience, we provide the expression of $\mathbf{J}$:

$$\mathbf{J} = 2(\mathbf{I}_n \otimes \mathbf{X}^T)\begin{bmatrix}\mathbf{Y}_1\\ \vdots\\ \mathbf{Y}_n\\ \overline{\mathbf{Y}}_1\\ \vdots\\ \overline{\mathbf{Y}}_n\end{bmatrix}\cdot\begin{bmatrix}-\operatorname{diag}(|V|)\operatorname{diag}(\sin\Theta) & \operatorname{diag}(\cos\Theta)\\ \operatorname{diag}(|V|)\operatorname{diag}(\cos\Theta) & \operatorname{diag}(\sin\Theta)\end{bmatrix},$$

where $\otimes$ denotes the Kronecker product and the magnitude, sine, and cosine operators are elementwise. Note that $\mathbf{J}$ depends on $\mathbf{X}$; we do not write this dependence to avoid clutter.

Then, applying the forecast error and AGC response (7.2), the equations (7.25) can be rewritten as:

$$\mathbf{J}\begin{bmatrix}\Delta\Theta\\ \Delta|V|\end{bmatrix} = \begin{bmatrix}\mathbf{I}_n - \alpha\mathbf{1}_n^T\\ \gamma\mathbf{I}_n\end{bmatrix}\xi + \begin{bmatrix}\delta P\\ \delta Q\end{bmatrix}, \tag{7.26}$$

where δP and δQ are the additional changes in active and reactive power injections after non-slack generator setpoints are changed, respectively. Applying knowledge of different bus types (slack, generator, and load), many components in equations (7.26) are zero and we only need solve the sub-systems

$$\mathbf{J}_1\begin{bmatrix}\Delta\Theta_{\mathcal{PV}\cup\mathcal{PQ}}\\ \Delta|V|_{\mathcal{PQ}}\end{bmatrix} = \underbrace{\begin{bmatrix}(\mathbf{I}_n - \alpha\mathbf{1}_n^T)_{\mathcal{PV}\cup\mathcal{PQ}}\\ (\gamma\mathbf{I}_n)_{\mathcal{PQ}}\end{bmatrix}}_{:=\mathbf{G}_1}\xi,$$

$$\mathbf{J}_2\begin{bmatrix}\Delta\Theta_{\mathcal{PV}\cup\mathcal{PQ}}\\ \Delta|V|_{\mathcal{PQ}}\end{bmatrix} = \underbrace{\begin{bmatrix}(\mathbf{I}_n - \alpha\mathbf{1}_n^T)_{\{P\theta\}}\\ (\gamma\mathbf{I}_n)_{\{P\theta\}\cup\mathcal{PV}}\end{bmatrix}}_{:=\mathbf{G}_2}\xi + \begin{bmatrix}\delta P_{\{P\theta\}}\\ \delta Q_{\{P\theta\}\cup\mathcal{PV}}\end{bmatrix},$$

where $(\cdot)_{\mathcal{S}}$ denotes the rows indexed by members of $\mathcal{S}$. Solving the above sub-systems gives

$$\begin{bmatrix}\Delta\Theta_{\mathcal{PV}\cup\mathcal{PQ}}\\ \Delta|V|_{\mathcal{PQ}}\end{bmatrix} = \mathbf{J}_1^{-1}\mathbf{G}_1\xi,$$

$$\begin{bmatrix}\delta P_{\{P\theta\}}\\ \delta Q_{\{P\theta\}\cup\mathcal{PV}}\end{bmatrix} = (\mathbf{J}_2\mathbf{J}_1^{-1}\mathbf{G}_1 - \mathbf{G}_2)\xi.$$

Equivalently, we get the following rows of sensitivity factors

$$\begin{aligned}(\partial_\xi \delta P)_{\{P\theta\}} &= (\mathbf{J}_2\mathbf{J}_1^{-1}\mathbf{G}_1 - \mathbf{G}_2)_1,\\ (\partial_\xi \delta Q)_{\{P\theta\}\cup\mathcal{PV}} &= (\mathbf{J}_2\mathbf{J}_1^{-1}\mathbf{G}_1 - \mathbf{G}_2)_{2:1+|\{P\theta\}|+|\mathcal{PV}|},\\ (\partial_\xi \delta |V|)_{\mathcal{PQ}} &= (\mathbf{J}_1^{-1}\mathbf{G}_1)_{|\mathcal{PV}|+|\mathcal{PQ}|+1:|\mathcal{PV}|+2|\mathcal{PQ}|},\end{aligned}$$

where on the right-hand side, we use the MATLAB-style of row indexing. All other rows are zero.

Finally, combining the above results with (7.26), we can compute the sensitivity factors of the constraint functions on active power, reactive power, and squared voltage magnitude, respectively:

$$\begin{aligned}\partial_\xi \Delta P(\mathbf{X},\xi) &= \mathbf{I}_n - \alpha \mathbf{1}_n^T + \partial_\xi \delta P(\mathbf{X},\xi),\\ \partial_\xi \Delta Q(\mathbf{X},\xi) &= \gamma \mathbf{I}_n + \partial_\xi \delta Q(\mathbf{X},\xi),\\ \partial_\xi \Delta |V|^2(\mathbf{X},\xi) &= 2|V| \circ \partial_\xi \Delta |V|(\mathbf{X},\xi),\end{aligned}$$

where $\circ$ denotes the elementwise product. Substituting the partial derivatives into the expression (7.24), we get the sensitivity factor $D_\Delta(\mathbf{X})$.

7.D Proof of Lemma 53

Proof of Lemma 53. We first show that in the definition of predictor $\hat{q}_{1-\epsilon,r,S}$, the supremum can be restricted to the set of distributions in $\mathcal{D}_r(\hat{\mathbb{P}}_S)$ that are absolutely continuous with respect to $\hat{\mathbb{P}}_S$ except on the set

$$\Xi^*(\mathbf{X}) := \left\{\xi \mid \bar{h}_{\mathbf{X}}(\xi) = h^*_{\mathbf{X}}\right\}.$$

The proof is the same as that of Lemma 2 of [222] except the bound on the expectation, i.e., the second last inequality in the proof. To deal with this issue, we only need to prove that for all $\mathbf{X} \in \mathbb{R}^d$, $p \in [0,1]$, $\xi^* \in \Xi^*(\mathbf{X})$, and $\mathbb{P}_c, \mathbb{P}_\perp \in \mathcal{P}$ such that $\mathbb{P}_c \ll \hat{\mathbb{P}}_S$ and $\mathbb{P}_\perp \perp \mathbb{P}_c$[2], it holds that

$$q_{1-\epsilon}\left(\bar{h}_{\mathbf{X}}, \mathbb{P}'\right) \geq q_{1-\epsilon}\left(\bar{h}_{\mathbf{X}}, \mathbb{P}''\right), \tag{7.27}$$

where

$$\mathbb{P}' \coloneqq p \cdot \mathbb{P}_c + (1-p)\cdot \delta_{\xi^*}, \quad \mathbb{P}'' \coloneqq p \cdot \mathbb{P}_c + (1-p)\cdot \mathbb{P}_\perp.$$

Let $F'(h)$ and $F''(h)$ be the cumulative distribution function of $\bar{h}_{\mathbf{X}}(\xi)$ under distribution $\mathbb{P}'$ and $\mathbb{P}''$, respectively. By the definition of quantile, to prove inequality (7.27), it is sufficient to show that

$$F'(h) \geq F''(h), \quad \forall h \in \mathbb{R},$$

[2]For distributions $\mathbb{P}, \mathbb{P}' \in \mathcal{P}$, we use $\mathbb{P} \ll \mathbb{P}'$ and $\mathbb{P} \perp \mathbb{P}'$ to denote the case when $\mathbb{P}$ is absolutely continuous and singular with respect to $\mathbb{P}'$, respectively.

which is equivalent to

$$\mathbb{E}_{\xi\sim\mathbb{P}'}\left[\mathbf{1}(\bar{h}_{\mathbf{X}}(\xi)\leq h)\right]\geq\mathbb{E}_{\xi\sim\mathbb{P}''}\left[\mathbf{1}(\bar{h}_{\mathbf{X}}(\xi)\leq h)\right],\quad\forall h\in\mathbb{R},$$

where $\mathbf{1}(\gamma(\nu,\xi)\leq\gamma)$ is the indicator function. This can be proved in the same way as the proof in [222]. As a result, there exists a distribution that attains $\hat{q}_{1-\epsilon,r,\hat{\mathbb{P}}_S}(\mathbf{X})$ and has support in $\{\xi^j, j\in[S]\}\cup\Xi^*(\mathbf{X})$, which implies the existence of an integer $k\in[S+1]$ such that

$$\hat{q}_{1-\epsilon,r,\hat{\mathbb{P}}_S}(\mathbf{X})=\bar{h}_{k,\hat{\mathbb{P}}_S}(\mathbf{X}).$$

Next, we prove that integer k does not depend on $\mathbf{X}$ and $\hat{\mathbb{P}}_S$. Let $\mathbb{P}_{1-\epsilon,r,S}$ be the aforementioned worst-case distribution that attains $\hat{q}_{1-\epsilon,r,S}(\mathbf{X})$. Assume without loss of generality that

$$\bar{h}_{\mathbf{X}}(\xi^1)\leq\cdots\leq\bar{h}_{\mathbf{X}}(\xi^S).$$

Define vector $\mathbf{p}\in\mathbb{R}^{S+1}$ as

$$\mathbf{p}_j:=\mathbb{P}_{1-\epsilon,r,S}(\xi^j),\quad\forall j\in[S],\quad\mathbf{p}_{S+1}:=\mathbb{P}_{1-\epsilon,r,S}\left[\Xi^*(\mathbf{X})\right].$$

Then, by problem (33) in [222], the integer k is the solution to

$$\begin{aligned}\max_{k\in[S],\mathbf{p}\in\mathbb{R}^{S+1}}\ & k\\ \text{s.t.}\ & \sum\nolimits_{j\in[k]}\mathbf{p}_j\leq 1-\epsilon,\ \mathbf{1}_{S+1}^T\mathbf{p}=1,\ \mathbf{p}\geq\mathbf{0}_{S+1},\\ & -\frac{1}{S}\sum\nolimits_{j\in[S]}\log(S\mathbf{p}_j)\leq r,\end{aligned}\tag{7.28}$$

which is independent of $\mathbf{X}$ and $\hat{\mathbb{P}}_S$. Intuitively, k is the largest integer such that the probability $\mathbb{P}_{1-\epsilon,r,S}$ on the smallest k samples is at most $1-\epsilon$ and the relative entropy constraint is not violated. □

7.E Proof of Theorem 74

Proof of Theorem 74. The formulation (7.15) is based on the big-M method [233]. If the variable $\mathbf{b}_j=1$, since the constant C is sufficiently large, there is no constraint on $h_\ell(\mathbf{X},\xi^j)$. Otherwise if the variable $\mathbf{b}_j=0$, the first constraint becomes

$$h_\ell(\mathbf{X},\xi^j)\leq 0,\quad\forall\ell\in[m],$$

which is equivalent to the condition $\bar{h}_{\mathbf{X}}(\xi^j)\leq 0$. With a given $\mathbf{X}\in\mathbb{R}^d$, the constraint $\mathbf{1}_S^T\mathbf{b}\leq S-k$ requires that the above condition holds for at least k samples. To achieve the minimum over $\mathbf{X}$, the condition $\mathbf{b}_j=0$ should hold for the k indices that correspond to the k smallest values in $\{\bar{h}_{\mathbf{X}}(\xi^j), j\in[S]\}$. In other words, the constraints in (7.15) are equivalent to

$$\bar{h}_k(\mathbf{X})\leq 0.$$

Combining with Lemma 53, we get the desired result. □

7.F Proof of Theorem 75

Proof of Theorem 75. By the definition of the prescriptor $\hat{\mathbf{X}}_{1-\epsilon,r,S}$, we have

$$\hat{q}_{1-\epsilon,r,S}\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}\right) \leq 0.$$

By a similar technique to the proof of Lemma 53, the results of Theorem 10 of [222] also holds for the predictor $\hat{q}_{1-\epsilon,r,S}$ and we have

$$\limsup_{S\to\infty} \frac{1}{S} \log \mathbb{P}_\infty \left[\hat{q}_{1-\epsilon,r,S}\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}\right) < q_{1-\epsilon}\left(\bar{h}_{\hat{\mathbf{X}}_{1-\epsilon,r,S}}, \mathbb{P}_0\right) \right] \leq -r.$$

Combining the above two inequalities, we get

$$\mathbb{P}_\infty \left[q_{1-\epsilon}\left(\bar{h}_{\hat{\mathbf{X}}_{1-\epsilon,r,S}}, \mathbb{P}_0\right) \leq 0 \right] \geq 1 - \exp\left[-rS + o(S)\right].$$

By the definition of the quantile and applying the union bound, it follows that

$$\mathbb{P}_\infty \left[h_\ell\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}, \xi\right) \leq 0, \quad \forall \ell \in [m] \right] \geq 1 - \epsilon - \exp\left[-rS + o(S)\right].$$

which is the desired result of this theorem. □

7.G Proof of Theorem 76

Proof of Theorem 76. We first construct a set where the distributionally robust predictor $\hat{q}_{1-\epsilon,r,S}$ takes positive value. Assume conversely that

$$p_S \coloneqq \mathbb{P}_\infty \left[g\left(\tilde{\mathbf{X}}_{1-\epsilon,r,S}\right) < g\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}\right) \right] > 0.$$

Since the prescriptor $\hat{\mathbf{X}}_{1-\epsilon,r,S}$ attains the minimal objective value under the constraint $\bar{h}_k(\mathbf{X}) \leq 0$, we have

$$\mathbb{P}_\infty \left[\bar{h}_k\left(\tilde{\mathbf{X}}_{1-\epsilon,r,S}\right) > 0 \right] \geq p_S.$$

Since $\mathbb{P}_\infty(\mathbf{b} > z)$ is a right-continuous function of $z \in \mathbb{R}$ for every random variable $\mathbf{b}$, there exists a sufficiently small constant $\tau > 0$ such that

$$\mathbb{P}_\infty \left[\bar{h}_k\left(\tilde{\mathbf{X}}_{1-\epsilon,r,S}\right) > \tau \right] \geq p_S/2 > 0.$$

Consider the set

$$\mathcal{X}_S \coloneqq \left\{ \left(\tilde{\mathbf{X}}_{1-\epsilon,r,S}, \hat{\mathbb{P}}_S\right) \mid \bar{h}_k\left(\tilde{\mathbf{X}}_{1-\epsilon,r,S}\right) > \tau \right\} \subset \mathbb{R}^d \times \mathcal{P}.$$

Since $\tilde{\mathbf{X}}_{1-\epsilon,r,S}$ is a quasi-continuous function of the empirical distribution $\hat{\mathbb{P}}_S$, the set $\mathcal{X}_S$ is a non-empty quasi-open set [174, Prop. 1.2.4] under the product topology of the Euclidean topology on $\mathbb{R}^d$ and the weak topology on $\mathcal{P}$. Therefore, the interior of $\mathcal{X}_S$, denoted as $\mathcal{X}_S^\circ$, is non-empty.

Now, we construct a data-driven predictor $\tilde{q}_{1-\epsilon,r,\mathbb{P}}$ that is continuous and does not dominate the distributionally robust predictor $\hat{q}_{1-\epsilon,r,\mathbb{P}}$. For every point $(\mathbf{X},\mathbb{P}) \in \mathcal{X}_S$, we define

$$\mathrm{d}(\mathbf{X},\mathbb{P}) \coloneqq \min\left\{\mathrm{dist}\left[(\mathbf{X},\mathbb{P}),\mathcal{X}_S^c\right],\tau\right\},$$

where $\mathcal{X}_S^c \coloneqq (\mathbb{R}^d \times \mathcal{P}) \backslash \mathcal{X}_S$ is the complementary set of $\mathcal{X}_S$ and the distance function is induced by the Euclidean 2-norm on $\mathbb{R}^d$ and the Prokhorov metric [187] on $\mathcal{P}$. Since the distance function is continuous, the function $\mathrm{d}(\cdot,\cdot)$ is also continuous and takes positive values on $\mathcal{X}_S^\circ$. We define

$$\tilde{q}_{1-\epsilon,r,\mathbb{P}}(\mathbf{X}) \coloneqq \hat{q}_{1-\epsilon,r,\mathbb{P}}(\mathbf{X}) - \mathrm{d}(\mathbf{X},\mathbb{P}), \quad \forall (\mathbf{X},\mathbb{P}) \in \mathbb{R}^d \times \mathcal{P}.$$

It follows from the definition of d and $\mathcal{X}_S$ that

$$0 \le \tilde{q}_{1-\epsilon,r,\mathbb{P}}(\mathbf{X}) \le \hat{q}_{1-\epsilon,r,\mathbb{P}}(\mathbf{X}), \quad \forall (\mathbf{X},\mathbb{P}) \in \mathcal{X}_S, \tag{7.29}$$

where the second inequality holds strictly on $\mathcal{X}_S^\circ$. Note that the predictor $\tilde{q}_{1-\epsilon,r,S} \coloneqq \tilde{q}_{1-\epsilon,r,\hat{\mathbb{P}}_S}$ is a data-driven predictor since it only relies on the empirical distribution $\hat{\mathbb{P}}_S$.

Finally, we show that $\tilde{q}_{1-\epsilon,r,\mathbb{P}}$ is feasible for problem (5) in [222], namely,

$$\limsup_{S\to\infty} \frac{1}{S} \log \mathbb{P}_\infty \left[\tilde{q}_{1-\epsilon,r,S}(\mathbf{X}) < q_{1-\epsilon}\left(\bar{h}_{\mathbf{X}},\mathbb{P}_0\right)\right] \le -r. \tag{7.30}$$

Since condition (7.30) is satisfied by $\hat{q}_{1-\epsilon,r,S}$ and

$$\tilde{q}_{1-\epsilon,r,S}(\mathbf{X}) = \hat{q}_{1-\epsilon,r,S}(\mathbf{X}), \quad \forall \mathbf{X} \in \mathbb{R}^d \text{ s.t. } \mathbf{X} \neq \tilde{\mathbf{X}}_{1-\epsilon,r,S},$$

we only need to show

$$\limsup_{S\to\infty} \frac{1}{S} \log \mathbb{P}_\infty \left[\tilde{q}_{1-\epsilon,r,S}\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}\right) < q_{1-\epsilon}\left(\bar{h}_{\hat{\mathbf{X}}_{1-\epsilon,r,S}},\mathbb{P}_0\right)\right] \le -r. \tag{7.31}$$

Since prescriptor $\tilde{\mathbf{X}}_{1-\epsilon,r,S}$ satisfies condition (7.16), it holds that

$$\limsup_{S\to\infty} \frac{1}{S} \log \mathbb{P}_\infty \left[q_{1-\epsilon}\left(\bar{h}_{\hat{\mathbf{X}}_{1-\epsilon,r,S}},\mathbb{P}_0\right) < 0\right] \le -r.$$

Combining with the property (7.29), we get the desired result (7.31).

In summary, we have constructed a predictor $\tilde{q}_{1-\epsilon,r,S}\left(\hat{\mathbf{X}}_{1-\epsilon,r,\mathbb{P}}\right)$ that is continuous and feasible for problem (5) in [222], but it does not dominate the distributionally robust predictor $\hat{q}_{1-\epsilon,r,S}\left(\hat{\mathbf{X}}_{1-\epsilon,r,\mathbb{P}}\right)$. However, this is contradictory with Theorem 10 in [222], which claims that the distributionally robust predictor is the strong solution to problem (5). □

7.H Disjoint Chance Constraint

In this section, we extend the theory in Section 7.3 to disjoint chance constraints. Using the same notation as Section 7.3, the disjoint chance constraint (7.7) can be written as

$$\mathbb{P}_0\left[h_\ell(\mathbf{X},\xi)\le 0\right]\ge 1-\epsilon,\quad \forall \ell\in[m]. \tag{7.32}$$

With the same violation probability ϵ, the disjoint chance constraint is less restrictive than the joint counterpart (7.11). On the other hand, if we choose ϵ to be ϵ/m in (7.32), Boole's inequality leads to

$$\mathbb{P}_0\left[h(\mathbf{X},\xi)>\mathbf{0}_m\right]\le \sum\nolimits_{\ell\in[m]}\mathbb{P}_0\left[h_\ell(\mathbf{X},\xi)>0\right]\le\epsilon,$$

which implies that the joint chance constraint holds with violation probability ϵ. More generally, with the disjoint chance constraint, we are able to bound the probability that at least s constraints are violated for all $s\in[m]$. More specifically, define the indicator function

$$\mathbf{1}_\ell(\mathbf{X},\xi):=\begin{cases}1 & \text{if } h_\ell(\mathbf{X},\xi)>0\\ 0 & \text{otherwise,}\end{cases}\quad \forall \ell\in[m].$$

Then, it holds that

$$\mathbb{E}_0\left[\mathbf{1}_\ell(\mathbf{X},\xi)\right]=\mathbb{P}_0\left[h_\ell(\mathbf{X},\xi)>0\right]\le\epsilon,$$

where $\mathbb{E}_0$ is the expectation under the true distribution $\mathbb{P}_0$. Using Markov's inequality, we get

$$\begin{aligned}&\mathbb{P}_0\left[h_\ell(\mathbf{X},\xi)>0 \text{ for at least } s \text{ indices } \ell\right]=\mathbb{P}_0\left[\sum\nolimits_{\ell\in[m]}\mathbf{1}_\ell(\mathbf{X},\xi)\ge s\right]\\ \le&\frac{\mathbb{E}_0\left[\sum_{\ell\in[m]}\mathbf{1}_\ell(\mathbf{X},\xi)\right]}{s}\le\frac{m\epsilon}{s}.\end{aligned}$$

In the following, we consider two different generalizations of chance constraint (7.32), which we denote as the *finite* case and the *infinite* case.

We first define the finite case of disjoint chance constraint. Given L weight vectors $\omega_1,\ldots,\omega_L\in\mathbb{R}^m$, the disjoint chance constraint is defined as

$$\mathbb{P}_0\left[\omega_\ell^T h(\mathbf{X},\xi)\le 0\right]\ge 1-\epsilon,\quad \forall \ell\in[L]. \tag{7.33}$$

The cardinal case (7.32) is a special case with $L=m$ and $\omega_\ell=\mathbf{e}_\ell$ for all $\ell\in[m]$. Basically, the reformulation of the finite case can be derived in a similar way as that of the joint chance constraint. Therefore, we omit the proofs for the disjoint chance constraint and use the same notation as Section 7.3. Choosing $C>0$ to be a sufficiently large constant, the distributionally robust prescriptor $\hat{\mathbf{X}}_{1-\epsilon,r,S}$ is a solution to

$$\min_{\mathbf{X}\in\mathbb{R}^d,\mathbf{B}\in\mathbb{Z}^{S\times L}}\ g(\mathbf{X}) \tag{7.34}$$

$$\begin{aligned}\text{s.t. } & \omega_\ell^T h(\mathbf{X}, \xi^j) \le C\mathbf{B}_{j,\ell}, \\ & \mathbf{1}_S^T \mathbf{B}_{:,\ell} \le S - k, \quad \mathbf{B}_{j,\ell} \in \{0,1\}, \quad \forall \ell \in [L],\ j \in [S],\end{aligned}$$

where the integer $k \in [\ell]$ is defined in Lemma 53 as a function of ϵ, r and S. Problem (7.34) is the disjoint counterpart of problem (7.15) and can be formulated as a MISDP for the CCOPF problem if the semi-definite relaxation is exact. Similarly, we can prove that the distributionally robust prescriptor achieves the optimal cost among solutions that satisfy disjoint chance constraint (7.33) with high probability.

Theorem 77. *For all $\epsilon \in [0,1]$ and $r > 0$, it holds that*

$$\mathbb{P}_\infty\left[q_{1-\epsilon}\left(\omega_\ell^T h(\hat{\mathbf{X}}_{1-\epsilon,r,S}, \cdot), \mathbb{P}_0\right) \le 0\right] \ge 1 - \exp\left[-rS + o(S)\right], \quad \forall \ell \in [L], \tag{7.35}$$

which leads to

$$\mathbb{P}_\infty\left[\omega_\ell^T h\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}, \xi\right) \le 0\right] \ge 1 - \epsilon - \exp\left[-rS + o(S)\right], \quad \forall \ell \in [L].$$

Furthermore, suppose that prescriptor $\tilde{\mathbf{X}}_{1-\epsilon,r,\mathbb{P}} \in \mathbb{R}^d$ is a quasi-continuous function of $\mathbb{P}$ and satisfies constraint (7.35). Then, we have

$$\mathbb{P}_\infty\left[g\left(\tilde{\mathbf{X}}_{1-\epsilon,r,S}\right) < g\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}\right)\right] = 0,$$

where we denote $\tilde{\mathbf{X}}_{1-\epsilon,r,S} \coloneqq \tilde{\mathbf{X}}_{1-\epsilon,r,\hat{\mathbb{P}}_S}$.

Next, we extend the disjoint chance constraint to a more general case. Instead of a finite number of weight vectors, the infinite case is defined by a set of weight vectors $\mathcal{W}$, which can contain an infinite number of elements. The infinite case of disjoint chance constraint is then formulated as

$$\mathbb{P}_0\left[\omega^T h(\mathbf{X}, \xi) \le 0\right] \ge 1 - \epsilon, \quad \forall \omega \in \mathcal{W}. \tag{7.36}$$

Hence, the finite case can be viewed as a special example of the infinite case, where the set $\mathcal{W}$ only contains a finite number of weight vectors. As an example of the infinite case, the set $\mathcal{W}$ can be the set of all weight vectors:

$$\mathcal{W} = \left\{\omega \in \mathbb{R}^m \mid \mathbf{1}_m^T \omega = 1, \omega \ge \mathbf{0}_m\right\}.$$

In this case, the constraint (7.36) enforces that all convex combinations of stochastic constraints are satisfied with high probability. More generally, in certain applications, the constraints can be divided into several groups. We can choose the set $\mathcal{W}$ to be the union of weight vectors of a subset of indices:

$$\mathcal{W} = \bigcup_{k \in [L]} \left\{\omega \in \mathbb{R}^m \mid \mathbf{1}_m^T \omega = 1, \omega \ge \mathbf{0}_m, \omega_\ell = 0, \forall \ell \notin \mathcal{I}_k\right\},$$

where $\mathcal{I}_k \subset [m]$ are disjoint subsets. Similar to the finite case, the chance constraint (7.36) can be reformulated as a MISDP. However, the MISDP contains an infinite number of constraints and thus, is considerably more challenging to solve. More specifically, for each $\omega \in \mathcal{W}$, the constraint requires that there exists a vector $\mathbf{b}^\omega \in \mathbb{Z}^S$ such that

$$\omega^T h(\mathbf{X}, \xi^j) \leq C\mathbf{b}_j^\omega, \quad \mathbf{1}_S^T \mathbf{b}^\omega \leq S - k, \quad \mathbf{b}_j^\omega \in \{0, 1\}, \quad \forall j \in [S]. \tag{7.37}$$

To deal with this challenge, we develop an iterative algorithm to approximate the constraint (7.37). In the t-th iteration, we use a finite set of weight vectors $\mathcal{W}_t$ to approximate the set $\mathcal{W}$. The algorithm proceeds in two stages:

1. With a fixed set $\mathcal{W}_t$, the algorithm generates an approximate distributionally robust prescriptor $\hat{\mathbf{X}}_t$ by solving problem (7.34) with weight vectors in $\mathcal{W}_t$;
2. With a fixed solution $\hat{\mathbf{X}}_t$, the algorithm finds the weight vector ω_t that violates constraint (7.37) by the largest margin. If there does not exist such weight vectors, we know that the constraint (7.37) is satisfied and the algorithm is terminated. Otherwise, we add vector ω_t to set $\mathcal{W}_t$.

The pseudo-code of the aforementioned algorithm is provided in Algorithm 23. For the CCOPF problem, if the set $\mathcal{W}$ is a polyhedral, problem (7.38) becomes a MISDP and problem (7.39) becomes a MIP. In this case, the algorithm runs efficiently in practice and exhibits good empirical performances; see more details in Section III of [27]. If the set $\mathcal{W}$ has certain special structure, the initial set $\mathcal{W}_1$ can be chosen based on the prior information about $\mathcal{W}$. For example, if $\mathcal{W}$ is a polyhedral, we can initialize $\mathcal{W}_1$ to contain all extreme points of the polyhedral. In the general case when the set $\mathcal{W}$ is not a polyhedral or even non-convex, problem (7.39) can be more challenging to solve.

Problem (7.39) is also based on the big-M method. If the variable $\mathbf{b}_j = 1$, since the constant C is sufficiently large, there is no constraint on s. Otherwise if the variable $\mathbf{b}_j = 0$, the constraint requires that

$$\omega^T h\left(\hat{\mathbf{X}}_t, \xi^j\right) \geq s.$$

This means that s should be the minimal value of the left-hand side over all indices j such that $\mathbf{b}_j = 0$. With a given $\omega \in \mathbb{R}^m$, to maximize the value of s, variable $\mathbf{b}_j$ is equal to one for indices with the k largest values of the left-hand side. Then, the optimal value of s should be the k-th largest value of the left-hand side over all samples. If we further minimize over the weight vector ω, the condition (7.37) holds if and only if the optimal value s_t is non-positive. In addition, if $s_t > 0$, the corresponding vector ω_t provides a weight vector such that condition (7.37) is violated by the largest margin.

Since problem (7.38) usually involves cone constraints, such as the semi-definite constraint in the CCOPF case, Algorithm 23 does not fit into the framework of classical cutting-plane methods, e.g., [221]. Therefore, the convergence of Algorithm 23 cannot be directly derived from those of existing cutting-plane methods and we leave the theoretical analysis to future works.

Algorithm 23 Algorithim for the infinite case of disjoint chance constraints.

Input: Set of weight vectors $\mathcal{W}$, empirical distribution $\hat{\mathbb{P}}_S$, number of iterations t_{max}, parameters ϵ, r.
Output: Approximate prescriptor $\hat{\mathbf{X}}_{1-\epsilon,r,S}$.
Compute k by solving (7.28).
Initialize $\mathcal{W}_1 \leftarrow \emptyset$.
▷ Alternatively, initialize with a finite subset of $\mathcal{W}$.
for $t = 1, 2, \ldots, t_{max}$ **do**
    Let $\hat{\mathbf{X}}_t$ be a solution to:

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{R}^d, \mathbf{B}\in\mathbb{Z}^{S\times L_t}} \quad & g(\mathbf{X}) \\ \text{s. t. } & \omega_\ell^T h(\mathbf{X}, \xi^j) \le C\mathbf{B}_{j,\ell}, \\ & \mathbf{1}_S^T \mathbf{B}_{:,\ell} \le S - k, \quad \mathbf{B}_{j,\ell} \in \{0,1\}, \quad \forall \ell \in [L_t], j \in [S], \end{aligned} \tag{7.38}$$

    where we define $L_t = |\mathcal{W}_t|$ and $\mathcal{W}_t = \{\omega_1, \ldots, \omega_{L_t}\}$.
    Let $(s_t, \omega_t, \mathbf{b}_t)$ be a solution to:

$$\begin{aligned} \max_{s\in\mathbb{R}, \omega\in\mathcal{W}, \mathbf{b}\in\mathbb{Z}^S} \quad & s, \\ \text{s.t. } & \omega^T h\left(\hat{\mathbf{X}}_t, \xi^j\right) \ge s + C\mathbf{b}_j, \\ & \mathbf{1}_S^T \mathbf{b} \le S - k, \quad \mathbf{b}_j \in \{0,1\}, \quad \forall j \in [S]. \end{aligned} \tag{7.39}$$

    if solution $s_t \le 0$ **then** ▷ condition (7.37) is satisfied.
        break
    end if
    Update $\mathcal{W}_{t+1} \leftarrow \mathcal{W}_t \cup \{\omega_t\}$.
end for
Return the last iterate of $\hat{\mathbf{X}}_t$ as $\hat{\mathbf{X}}_{1-\epsilon,r,S}$.

In this chapter, we assume that the minimum-cost solution $\hat{\mathbf{X}}_{1-\epsilon,r,S}$ can be found, namely, it is a solution to the following optimization problem:chap6-

$$\begin{aligned} \min_{\mathbf{X}\in\mathbb{R}^d, \mathbf{b}^w\in\mathbb{R}^S} \quad & g(\mathbf{X}) \\ \text{s. t.} \quad & \text{constraint (7.37) is satisfied for all } \omega \in \mathcal{W}. \end{aligned}$$

The next theorem claims that the solution $\hat{\mathbf{X}}_{1-\epsilon,r,S}$ satisfies a similar optimality condition as the finite case.

Theorem 78. *For all $\epsilon \in [0,1]$ and $r > 0$, it holds that*

$$\mathbb{P}_\infty\left[q_{1-\epsilon}\left(\omega^T h(\hat{\mathbf{X}}_{1-\epsilon,r,S}, \cdot), \mathbb{P}_0\right) \le 0\right] \ge 1 - \exp\left[-rS + o(S)\right], \quad \forall \omega \in \mathcal{W}, \tag{7.40}$$

which leads to

$$\mathbb{P}_\infty\left[\omega^T h\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}, \xi\right) \leq 0\right] \geq 1 - \epsilon - \exp\left[-rS + o(S)\right], \quad \forall \omega \in \mathcal{W}.$$

Furthermore, suppose that prescriptor $\tilde{\mathbf{X}}_{1-\epsilon,r,\mathbb{P}} \in \mathbb{R}^d$ *is a quasi-continuous function of* $\mathbb{P}$ *and satisfies constraint (7.40). Then, we have*

$$\mathbb{P}_\infty\left[g\left(\tilde{\mathbf{X}}_{1-\epsilon,r,S}\right) < g\left(\hat{\mathbf{X}}_{1-\epsilon,r,S}\right)\right] = 0,$$

where we denote $\tilde{\mathbf{X}}_{1-\epsilon,r,S} \coloneqq \tilde{\mathbf{X}}_{1-\epsilon,r,\hat{\mathbb{P}}_S}$.

We omit the proof due to its similarity to the proof of Theorems 75 and 76.

Chapter 8

Conclusions and Future Directions

This dissertation aims at exploring and developing practical solutions by providing scalable, reliable and resilient optimization algorithms. The solutions to the computational challenges will significantly change both everyday technologies and the frontiers of advanced research areas. It may take life-long work to address the challenges in these fields and this dissertation serves as important initial attempts towards achieving this goal. With the introduction of advanced analysis and computation techniques, we develop and analyze algorithms that are able to effectively utilize the underlying benign structure of real-world problems. Our contributions are classified into three parts, namely, *low-rank matrix optimization*, *convex discrete optimization via simulation*, and *power systems*. In the following, we summarize the contributions of each part and discuss potential future research directions.

8.1 Low-rank Matrix Optimization

In Chapter 2, we analyze the geometric properties of low-rank optimization problems via the non-convex factorization approach. We prove novel necessary conditions and sufficient conditions for the non-existence of spurious second-order critical points in both symmetric and asymmetric cases. We show that these conditions lead to sharper bounds and greatly simplify the construction of counterexamples needed to study the sharpness of the bounds. The developed bounds significantly generalize several of the existing results. In the rank-1 case, the bound is proved to be the sharpest possible. In the general rank case, we show that there exists a positive correlation between second-order critical points and the global minimum for problems whose RIP constants are higher than the developed bound but lower than the fundamental limit obtained by the counterexamples. Finally, the strict saddle property is proved with a weaker requirement on the RIP constant for asymmetric problems. This chapter develops the first strict saddle property in the literature for nonlinear symmetric problems.

In Chapter 3, we propose a new complexity metric for an important class of the low-rank matrix optimization problems, which has the potential to generalize major existing recovery

guarantees and is applicable to a much broader set of problems. The proposed complexity metric aims to measure the complexity of the non-convex optimization landscape of each problem and quantifies the likelihood of local search methods in successfully solving each instance of the problem under a random initialization. We focus on the rank-1 generalized matrix completion problem (3.4) to mathematically prove the usefulness of the new metric from three aspects. Namely, we show that the complexity metric has a small value if the instance satisfies the RIP condition or the incoherence condition. The results in these two scenarios are consistent with the existing results on the RIP condition and the incoherence condition. In addition, we analyze a one-parameter class of instances to illustrate that the proposed metric captures the true complexity of this class as the parameter varies and has consistent behavior with the aforementioned two scenarios. This consistency implies that our proposed complexity metric is able to characterize the optimization landscapes of different applications, which the RIP condition and the incoherence condition fail to capture. Finally, we provide strong theoretical results on the generalized matrix completion problem by showing that a small value for the proposed complexity metric guarantees the absence of spurious solutions, whereas a large value for a slightly modified complexity metric guarantees the existence of spurious solutions. This also shows the superiority of this metric over the RIP condition and the incoherence condition since those notions cannot offer any necessary conditions on having spurious solutions.

Given the gap between the empirical and the theoretical results, there are numerous future research directions to pursue for low-rank matrix optimization problem. The existing geometric analysis of the low-rank optimization problem mostly relies on restrictive assumptions, which only hold approximately, if are not violated, in practice. Aiming at developing new tools and algorithms, we need to extend the theory to settings that are more pertinent to real-world applications.

For example, most existing works assumed that the rank of the underlying low-rank matrix is known or can be estimated from prior information. However, in practice, the rank is usually unknown and an over-parameterized factorization model is commonly used. Instead of overfitting, it is observed that the gradient descent algorithm exhibits robust convergence to the global solution with the lowest rank starting from a small initialization. This phenomenon is known as the *implicit regularization* of the gradient descent algorithm and also appears in other machine learning models, such as linear regression and neural networks. The theoretical explanation of the implicit regularization is far from complete. Due to the similarity with linear neural networks, understanding the implicit regularization phenomenon of the factorization approach serves as a very important intermediate step towards that of deep neural networks. Moreover, establishing the convergence guarantees through the implicit regularization justifies the application of the over-parameterized model.

Moreover, the ℓ_2-loss function is considered extensively in literature. It may be beneficial to use other loss functions, such as the ℓ_1-loss function and the Sigmoid function, especially when the underlying data generation distribution is not from Gaussian. For example, when the measurements contain a small number of outliers, the ℓ_1-loss function is more robust to the outliers. However, since the ℓ_1-loss function is not smooth, most existing analysis

techniques cannot be directly applied and the optimization landscape is less well-understood. Establishing the convergence results for general loss functions, especially non-smooth loss functions, is another important future research direction.

In summary, the field of low-rank matrix optimization still contains lots of important open questions. The research progress on those open questions will extend our understanding on non-convex optimization algorithms and have significant impact on a wide range of application areas.

8.2 Convex Discrete Optimization via Simulation

In Chapter 4, we propose computationally efficient simulation-optimization algorithms for large-scale simulation optimization problems that have high-dimensional discrete decision space in the presence of a convex structure. For a user-specified precision level, the proposed simulation-optimization algorithms are guaranteed to find a choice of decision variables that is close to the optimal within the precision level with desired high probability. We provide upper bounds on simulation costs for the proposed simulation-optimization algorithms. In Chapter 4, we mainly focus on algorithm design and theoretical guarantees. In future work, we seek to design better simulation-optimization algorithms that provide simulation costs with matching upper and lower bounds.

In Chapter 5, algorithms based on the idea of localization are proposed for large-scale convex discrete optimization via simulation problems. The simulation-optimization algorithms are theoretically guaranteed to identify a solution whose corresponding objective value is close to the optimal objective value up to a given precision with high probability. Moreover, the efficiency of the developed algorithms is evaluated by obtaining upper bounds on the expected simulation cost. We summarize the performances of our algorithms in Table 5.5.3. Specifically, in the one-dimensional case, we propose the SUS method, which has an expected simulation cost as $O[\epsilon^{-2}(\log(N) + \log(1/\delta))]$, which attains the best achievable performance under the asymptotic criterion [128], i.e., when $\delta \to 0$. For the multi-dimensional case, we combine the idea of localization with subgradient information. The dimension reduction algorithm is designed using a new framework to extend deterministic cutting-plane methods. The expected simulation cost is proven to be upper bounded by a constant that is independent of the Lipschitz constant. In addition, the dimension reduction algorithm does not require prior knowledge about the Lipschitz constant. Finally, an adaptive algorithm (described in 5.G) is designed to avoid the requirement that the variance of the noise should be estimated a priori. Numerical results on both synthetic and queueing models demonstrate that the proposed algorithms have better performances compared to benchmark methods especially when the problem scale is large. In summary, the stochastic localization algorithms are preferred when either (i) the problem scale is large or (ii) the Lipschitz constant is large or difficult to estimate. On the other hand, if the problem scale is moderate but the dimension is high, the subgradient-based search methods are preferred.

Looking ahead, there are various ways to extend the results on the convex Discrete Optimization via Simulation (DOvS) problem. First, we can consider other structures of the objective function. In continuous non-convex optimization, several different geometric structures were identified and proved to be able to guarantee the polynomial-time convergence of saddle-avoiding algorithms. One of the most famous structure is the strict saddle property [244], which is satisfied by a number of non-convex optimization problems. In the case of discrete optimization, we can study other geometric structures, besides the $L^\natural$-convexity, that can also be utilized to develop new algorithms and reduce the optimization complexity of the DOvS problem.

Moreover, the algorithm design can be improved with more sophisticated computation techniques. For example, the parallel DOvS algorithms can greatly reduce the computation time of large-scale problems, when the computation cluster is available. In many DOvS problems, the *common random number* technique can be utilized to reduce the joint simulation cost of several similar decisions. More advanced statistical computing methods can also be applied to improve the statistical guarantees of the developed algorithms.

In summary, the area of utilizing hidden structures in discrete optimization is newly chopped and has lots of future directions to pursue. The research progress in this area will lead to non-trivial impacts on many application fields of the DOvS problem.

8.3 Power Systems

In Chapter 6, we extend the uniqueness theory of P-Θ power flow solutions developed in [184] for an AC power system. The notion of strong uniqueness is introduced to characterize the uniqueness in the common sense. We propose a general necessary and sufficient condition for the uniqueness of the solution, which depends only on the monotone regime and the network topology. These conditions can be greatly simplified in certain scenarios. When the underlying graph of the power network is a single cycle, sufficient conditions in [184] are proved to be necessary. For 2-vertex-connected SP graphs, we show that the maximal eye is equal to the maximal girth, which means that the sufficient condition for the weak uniqueness also implies the strong uniqueness. When the power network is lossless, we derive a necessary and sufficient condition that does not contain sinusoidal functions and its sufficient part is stronger than the general sufficient conditions. A reduction method, named the ISPR method, is proposed to reduce the size of power network and accelerate the computation of the maximal eye and the maximal girth. The ISPR method is proved to reduce a 2-vertex-connected SP graph to a single edge and the relation between the graphs before and after the reduction is analyzed. Some algorithms based on the DFS method with pruning are designed to compute the maximal eye and maximal girth.

In Chapter 7, we focus on the distributionally robust approach for the CCOPF problem. We propose a new DRO formulation based on the relative entropy, which achieves the optimal generation cost given the maximum violation rate. In addition, we provide an exact reformulation of the joint chance constraint, which guarantees the feasibility of the reformu-

lated problem and leads to significantly better efficiency compared with existing approaches based on inner approximation. Finally, numerical results on IEEE benchmark power systems are exhibited to show the superior performance of our approach compared to existing state-of-the-art approaches.

As a potential extension, we can study the OPF problem from the perspective of low-rank matrix optimization. We consider the case when the power systems suffer from adversarial outliers. In the adversarial setting, the ℓ_1-loss function is known to be robust to outliers in several applications, e.g., LASSO regression and robust principle component analysis. We are interested in establishing similar theoretical guarantees for the factorization model with ℓ_1-loss function.

Bibliography

[1] Alekh Agarwal et al. "Learning sparsely used overcomplete dictionaries via alternating minimization". In: *SIAM Journal on Optimization* 26.4 (2016), pp. 2775–2799.

[2] Alekh Agarwal et al. "Stochastic convex optimization with bandit feedback". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1035–1043.

[3] Shubhada Agrawal, Sandeep Juneja, and Peter Glynn. "Optimal δ-Correct Best-Arm Selection for Heavy-Tailed Distributions". In: *Algorithmic Learning Theory*. 2020, pp. 61–110.

[4] Kwangjun Ahn and Felipe Suarez. "Riemannian perspective on matrix factorization". In: *arXiv preprint arXiv:2102.00937* (2021).

[5] Martin Aigner. "Turán's graph theorem". In: *The American Mathematical Monthly* 102.9 (1995), pp. 808–816.

[6] Ahmad Ajalloeian and Sebastian U Stich. "Analysis of SGD with Biased Gradient Estimators". In: *arXiv preprint arXiv:2008.00051* (2020).

[7] Tayo Ajayi et al. "Provably convergent acceleration in factored gradient descent with applications in matrix sensing". In: *arXiv preprint arXiv:1806.00534* (2018).

[8] Zeyuan Allen-Zhu and Yuanzhi Li. "Neon2: Finding local minima via first-order oracles". In: *Advances in Neural Information Processing Systems* 31 (2018).

[9] Josh Alman and Virginia Vassilevska Williams. "A refined laser method and faster matrix multiplication". In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2021, pp. 522–539.

[10] Eitan Altman, Bruno Gaujal, and Arie Hordijk. *Discrete-event control of stochastic networks: Multimodularity and regularity*. springer, 2003.

[11] Eitan Altman, Bruno Gaujal, and Arie Hordijk. "Multimodularity, convexity, and optimization properties". In: *Mathematics of Operations Research* 25.2 (2000), pp. 324–347.

[12] Alireza Arab and Joseph Euzebe Tate. "Distributionally Robust Optimal Power Flow via Ellipsoidal Approximation". In: *IEEE Transactions on Power Systems* (2022). Publisher: IEEE.

[13] A. Araposthatis, S. Sastry, and P. Varaiya. "Analysis of power-flow equation". In: *International Journal of Electrical and Power Energy Systems* 3 (July 1981), pp. 115–126.

[14] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods". In: *Mathematical Programming* 137.1 (2013), pp. 91–129.

[15] Brian Axelrod, Yang P Liu, and Aaron Sidford. "Near-optimal approximate discrete and continuous submodular function minimization". In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2020, pp. 837–853.

[16] Kyri Baker and Andrey Bernstein. "Joint Chance Constraints in AC Optimal Power Flow: Improving Bounds Through Learning". In: *IEEE Transactions on Smart Grid* 10.6 (Nov. 2019), pp. 6376–6385. ISSN: 1949-3053, 1949-3061.

[17] Robert E Bechhofer. "A single-sample multiple decision procedure for ranking means of normal populations with known variances". In: *The Annals of Mathematical Statistics* (1954), pp. 16–39.

[18] Alexandre Belloni et al. "Escaping the local minima via simulated annealing: Optimization of approximately convex functions". In: *Conference on Learning Theory*. 2015, pp. 240–265.

[19] Andrey Bernstein et al. "Load-Flow in Multiphase Distribution Networks: Existence, Uniqueness, Non-Singularity and Linear Models". In: *IEEE Transactions on Power Systems* 33.6 (Apr. 2018), pp. 5832–5843.

[20] Dimitris Bertsimas and Santosh Vempala. "Solving convex programs by random walks". In: *Journal of the ACM (JACM)* 51.4 (2004), pp. 540–556.

[21] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. "Dropping convexity for faster semi-definite optimization". In: *Conference on Learning Theory*. PMLR. 2016, pp. 530–582.

[22] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. "Global optimality of local search for low rank matrix recovery". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 3880–3888.

[23] Yingjie Bi and Javad Lavaei. "On the absence of spurious local minima in nonlinear low-rank matrix recovery problems". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 379–387.

[24] Yingjie Bi, Haixiang Zhang, and Javad Lavaei. "Local and Global Linear Convergence of General Low-rank Matrix Recovery Problems". In: *Proceedings of 36th AAAI Conference on Artificial Intelligence (AAAI), Vancouver, Canada*. 2022, pp. 1–9.

[25] P Borjesson and C-E Sundberg. "Simple approximations of the error function Q (x) for communications applications". In: *IEEE Transactions on Communications* 27.3 (1979), pp. 639–643.

[26] Nicolas Boumal. "Nonconvex phase synchronization". In: *SIAM Journal on Optimization* 26.4 (2016), pp. 2355–2377.

[27] Eli Brock et al. "Distributionally Robust Optimization for Nonconvex QCQPs with Stochastic Constraints". In: *2023 62th IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 1–7.

[28] 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.

[29] James V Burke and Michael C Ferris. "Weak sharp minima in mathematical programming". In: *SIAM Journal on Control and Optimization* 31.5 (1993), pp. 1340–1359.

[30] Apostolos N Burnetas and Michael N Katehakis. "Optimal adaptive policies for sequential allocation problems". In: *Advances in Applied Mathematics* 17.2 (1996), pp. 122–142.

[31] Marco C. Campi, Simone Garatti, and Maria Prandini. "The scenario approach for systems and control design". en. In: *Annual Reviews in Control* 33.2 (Dec. 2009), pp. 149–157. ISSN: 13675788.

[32] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. "Phase retrieval via Wirtinger flow: Theory and algorithms". In: *IEEE Transactions on Information Theory* 61.4 (2015), pp. 1985–2007.

[33] Emmanuel J Candes and Yaniv Plan. "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements". In: *IEEE Transactions on Information Theory* 57.4 (2011), pp. 2342–2359.

[34] Emmanuel J Candès and Benjamin Recht. "Exact matrix completion via convex optimization". In: *Foundations of Computational mathematics* 9.6 (2009), pp. 717–772.

[35] Emmanuel J Candès and Terence Tao. "The power of convex relaxation: Near-optimal matrix completion". In: *IEEE Transactions on Information Theory* 56.5 (2010), pp. 2053–2080.

[36] Emmanuel J Candès et al. "Robust principal component analysis?" In: *Journal of the ACM (JACM)* 58.3 (2011), pp. 1–37.

[37] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. "Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results". In: *Mathematical Programming* 127.2 (2011), pp. 245–295.

[38] Vasileios Charisopoulos et al. "Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence". In: *Foundations of Computational Mathematics* 21.6 (2021), pp. 1505–1593.

[39] Ji Chen and Xiaodong Li. "Model-free Nonconvex Matrix Completion: Local Minima Analysis and Applications in Memory-efficient Kernel PCA." In: *J. Mach. Learn. Res.* 20.142 (2019), pp. 1–39.

[40] Ji Chen, Dekai Liu, and Xiaodong Li. "Nonconvex Rectangular Matrix Completion via Gradient Descent Without $\ell_{2,\infty}$ Regularization". In: *IEEE Transactions on Information Theory* 66.9 (2020), pp. 5806–5841.

[41] Jie Chen and Ronny Luss. "Stochastic gradient descent with biased but consistent gradient estimators". In: *arXiv preprint arXiv:1807.11880* (2018).

[42] Lijie Chen, Anupam Gupta, and Jian Li. "Pure exploration of multi-armed bandit under matroid constraints". In: *Conference on Learning Theory*. 2016, pp. 647–669.

[43] Xi Chen, Bruce E Ankenman, and Barry L Nelson. "Enhancing stochastic kriging metamodels with gradient estimators". In: *Operations Research* 61.2 (2013), pp. 512–528.

[44] Xi Chen, Enlu Zhou, and Jiaqiao Hu. "Discrete optimization via gradient-based adaptive stochastic search methods". In: *IISE Transactions* 50.9 (2018), pp. 789–805.

[45] Xin Chen and Menglong Li. "Discrete convex analysis and its applications in operations: A survey". In: *Production and Operations Management* (2020).

[46] Yudong Chen and Yuejie Chi. "Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization". In: *IEEE Signal Processing Magazine* 35.4 (2018), pp. 14–31.

[47] Yuxin Chen et al. "Bridging convex and nonconvex optimization in robust PCA: Noise, outliers and missing data". In: *The Annals of Statistics* 49.5 (2021), pp. 2948–2971.

[48] Yuxin Chen et al. "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval". In: *Mathematical Programming* 176.1 (2019), pp. 5–37.

[49] Yuxin Chen et al. "Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization". In: *SIAM journal on optimization* 30.4 (2020), pp. 3098–3121.

[50] Yuejie Chi, Yue M Lu, and Yuxin Chen. "Nonconvex optimization meets low-rank matrix factorization: An overview". In: *IEEE Transactions on Signal Processing* 67.20 (2019), pp. 5239–5269.

[51] Hsiao-Dong Chiang and Mesut E. Baran. "On the Existence and Uniqueness of Load Flow Solution for Radial Distribution Power Networks". In: *IEEE Transactions on Circuits and Systems* CAS-37.3 (Mar. 1990), pp. 410–416.

[52] Stephen E Chick. "Subjective probability and Bayesian methodology". In: *Handbooks in Operations Research and Management Science* 13 (2006), pp. 225–257.

[53] Hung-Hsu Chou et al. "Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank". In: *Applied and Computational Harmonic Analysis* 68 (2024), p. 101595.

[54] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[55] Bai Cui and Xu Andy Sun. "Solvability of power flow equations through existence and uniqueness of complex fixed point". In: (2019). available online at `https://arxiv.org/pdf/1904.08855.pdf`.

[56] Aris Daniilidis and Dmitriy Drusvyatskiy. "Pathological subgradient dynamics". In: *SIAM Journal on Optimization* 30.2 (2020), pp. 1327–1338.

[57] Damek Davis and Dmitriy Drusvyatskiy. "Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions". In: *arXiv preprint arXiv:1802.02988* (2018).

[58] Damek Davis et al. "Stochastic subgradient method converges on tame functions". In: *Foundations of computational mathematics* 20.1 (2020), pp. 119–154.

[59] Robin Delabays, Tommaso Coletta, and Philippe Jacquod. "Multistability of phase-locking in equal-frequency Kuramoto models on planar graphs". In: *Journal of Mathematical Physics* 58.3 (2017), p. 032703.

[60] Olivier Devolder, François Glineur, and Yurii Nesterov. "First-order methods of smooth convex optimization with inexact oracle". In: *Mathematical Programming* 146.1-2 (2014), pp. 37–75.

[61] Elisângela Silva Dias et al. "Efficient enumeration of chordless cycles". In: *arXiv preprint arXiv:1309.1051* (2013).

[62] Florian Dörfler, John W Simpson-Porco, and Francesco Bullo. "Electrical networks and algebraic graph theory: Models, properties, and applications". In: *Proceedings of the IEEE* 106.5 (2018), pp. 977–1005.

[63] Chao Duan et al. "Distributionally Robust Chance-Constrained Approximate AC-OPF With Wasserstein Metric". In: *IEEE Transactions on Power Systems* 33.5 (Sept. 2018), pp. 4924–4936. ISSN: 0885-8950, 1558-0679.

[64] K. Dvijotham, Enrique Mallada, and J.W. Simpson-Porco. "High-voltage solution in radial power networks: existence, properties, and equivalent algorithms". In: *IEEE Control Systems Letters* 1.2 (Oct. 2017), pp. 322–327.

[65] Krishnamurthy Dvijotham, Steven Low, and Michael Chertkov. "Solving the power flow equations: A monotone operator approach". In: *arXiv preprint arXiv:1506.08472* (2015).

[66] ME Dyer and LG Proll. "Note—On the validity of marginal analysis for allocating servers in M/M/c queues". In: *Management Science* 23.9 (1977), pp. 1019–1022.

[67] David J Eckman and Shane G Henderson. *Fixed-confidence, fixed-tolerance guarantees for selection-of-the-best procedures*. Tech. rep. Working paper, Cornell University, School of Operations Research and . . ., 2018.

[68] David J Eckman and Shane G Henderson. "Guarantees on the probability of good selection". In: *2018 Winter Simulation Conference (WSC)*. IEEE. 2018, pp. 351–365.

[69] David J Eckman, Matthew Plumlee, and Barry L Nelson. "Flat chance! using stochastic gradient estimators to assess plausible optimality for convex functions". In: *2021 Winter Simulation Conference (WSC)*. IEEE. 2021, pp. 1–12.

[70] David J Eckman, Matthew Plumlee, and Barry L Nelson. "Plausible screening using functional properties for simulations with large solution spaces". In: *Operations Research* 70.6 (2022), pp. 3473–3489.

[71] David J. Eckman and Shane G. Henderson. "Biased gradient estimators in simulation optimization". In: *Proceedings of the 2020 Winter Simulation Conference*. Ed. by K.-H. Bae et al. IEEE. Piscataway NJ, 2020, Submitted.

[72] D. Eppstein. "Parallel recognition of series-parallel graphs". In: *Information and Computation* 98.1 (1992), pp. 41–55.

[73] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. "PAC bounds for multi-armed bandit and Markov decision processes". In: *International Conference on Computational Learning Theory*. Springer. 2002, pp. 255–270.

[74] Weiwei Fan, L Jeff Hong, and Barry L Nelson. "Indifference-zone-free selection of the best". In: *Operations Research* 64.6 (2016), pp. 1499–1514.

[75] Salar Fattahi and Somayeh Sojoudi. "Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis". In: *Journal of machine learning research* (2020).

[76] Paola Favati. "Convexity in nonlinear integer programming". In: *Ricerca operativa* 53 (1990), pp. 3–44.

[77] Daniel Freund, Shane G Henderson, and David B Shmoys. "Minimizing multimodular functions and allocating capacity in bike-sharing systems". In: *International Conference on Integer Programming and Combinatorial Optimization*. Springer. 2017, pp. 186–198.

[78] Michael C Fu. "Optimization for simulation: Theory vs. practice". In: *INFORMS Journal on Computing* 14.3 (2002), pp. 192–215.

[79] Michael C Fu and Huashuai Qu. "Regression models augmented with direct stochastic gradient estimators". In: *INFORMS Journal on Computing* 26.3 (2014), pp. 484–499.

[80] Satoru Fujishige. *Submodular functions and optimization*. Elsevier, 2005.

[81] Satoru Fujishige. "Theory of submodular programs: A Fenchel-type min-max theorem and subgradients of submodular functions". In: *Mathematical programming* 29.2 (1984), pp. 142–155.

[82] A Futschik and Georg Pflug. "Confidence sets for discrete stochastic optimization". In: *Annals of Operations Research* 56.1 (1995), pp. 95–108.

[83] Andreas Futschik and G Ch Pflug. "Optimal allocation of simulation experiments in discrete stochastic optimization and approximative algorithms". In: *European Journal of Operational Research* 101.2 (1997), pp. 245–260.

[84] Michael R Garey and David S Johnson. *Computers and intractability*. Vol. 174. freeman San Francisco, 1979.

[85] Aurélien Garivier and Emilie Kaufmann. "Optimal best arm identification with fixed confidence". In: *Conference on Learning Theory*. 2016, pp. 998–1027.

[86] Rong Ge, Chi Jin, and Yi Zheng. "No spurious local minima in nonconvex low rank problems: A unified geometric analysis". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1233–1242.

[87] Rong Ge, Jason D Lee, and Tengyu Ma. "Matrix completion has no spurious local minimum". In: *Advances in Neural Information Processing Systems* (2016), pp. 2981–2989.

[88] Xiting Gong and Xiuli Chao. "Optimal control policy for capacitated inventory systems with remanufacturing". In: *Operations Research* 61.3 (2013), pp. 603–611.

[89] Michael Grant and Stephen Boyd. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1*. Mar. 2014. URL: `http://cvxr.com/cvx`.

[90] Andrei Graur et al. "New Query Lower Bounds for Submodular Function Minimization". In: *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2020.

[91] Oktay Günlük. "A branch-and-cut algorithm for capacitated network design problems". In: *Mathematical programming* 86.1 (1999), pp. 17–39.

[92] Yi Guo et al. "Data-based distributionally robust stochastic optimal power flow—Part I: Methodologies". In: *IEEE Transactions on Power Systems* 34.2 (2018), pp. 1483–1492.

[93] Yi Guo et al. "Data-based distributionally robust stochastic optimal power flow—Part II: Case studies". In: *IEEE Transactions on Power Systems* 34.2 (2018), pp. 1493–1503.

[94] Walter J Gutjahr and Georg Ch Pflug. "Simulated annealing for noisy cost functions". In: *Journal of global optimization* 8.1 (1996), pp. 1–13.

[95] Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. "An Equivalence between Critical Points for Rank Constraints Versus Low-Rank Factorizations". In: *SIAM Journal on Optimization* 30.4 (2020), pp. 2927–2955.

[96] Moritz Hardt. "Understanding alternating minimization for matrix completion". In: *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE. 2014, pp. 651–660.

[97] Moritz Hardt and Mary Wootters. "Fast matrix completion without the condition number". In: *Conference on learning theory*. PMLR. 2014, pp. 638–678.

[98] Elad Hazan and Satyen Kale. "Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization". In: *Proceedings of the 24th Annual Conference on Learning Theory*. 2011, pp. 421–436.

[99] Ian A. Hiskens and R. J. Davy. "Exploring the power flow solution space boundary". In: *IEEE Transactions on Power Systems* 16.3 (Aug. 2001), pp. 389–395.

[100] Bri-Mathias Hodge. *Final Report on the Creation of the Wind Integration National Dataset (WIND) Toolkit and API: October 1, 2013 - September 30, 2015*. 2016.

[101] L Jeff Hong, Weiwei Fan, and Jun Luo. "Review on ranking and selection: A new perspective". In: *Frontiers of Engineering Management* 8.3 (2021), pp. 321–343.

[102] L Jeff Hong and Barry L Nelson. "Discrete optimization via simulation using COMPASS". In: *Operations Research* 54.1 (2006), pp. 115–129.

[103] L Jeff Hong, Barry L Nelson, and Jie Xu. "Discrete optimization via simulation". In: *Handbook of simulation optimization*. Springer, 2015, pp. 9–44.

[104] L Jeff Hong, Barry L Nelson, and Jie Xu. "Speeding up COMPASS for high-dimensional discrete optimization via simulation". In: *Operations Research Letters* 38.6 (2010), pp. 550–555.

[105] L Jeff Hong and Xiaowei Zhang. "Surrogate-based simulation optimization". In: *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*. INFORMS, 2021, pp. 287–311.

[106] Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. "Fast global convergence for low-rank matrix recovery via Riemannian gradient descent with random initialization". In: *arXiv preprint arXiv:2012.15467* (2020).

[107] Jiaqiao Hu, Michael C Fu, and Steven I Marcus. "A model reference adaptive search method for global optimization". In: *Operations Research* 55.3 (2007), pp. 549–568.

[108] Jiaqiao Hu, Michael C Fu, Steven I Marcus, et al. "A model reference adaptive search method for stochastic global optimization". In: *Communications in Information & Systems* 8.3 (2008), pp. 245–276.

[109] Yifan Hu et al. "Biased Stochastic Gradient Descent for Conditional Stochastic Optimization". In: *arXiv preprint arXiv:2002.10790* (2020).

[110] Zhishen Huang and Stephen Becker. "Perturbed proximal descent to escape saddle points for non-convex and non-smooth objective functions". In: *INNS Big Data and Deep Learning conference*. Springer. 2019, pp. 58–77.

[111] Woonghee Tim Huh and Ganesh Janakiraman. "On the optimal policy structure in serial inventory systems with lost sales". In: *Operations Research* 58.2 (2010), pp. 486–491.

[112] Susan R Hunter and Barry L Nelson. "Parallel ranking and selection". In: *Advances in Modeling and Simulation*. Springer, 2017, pp. 249–275.

[113] Marija Ilic. "Network theoretic conditions for existence and uniqueness of steady state solutions to electric power circuits". In: *Proceedings of the IEEE International Symposium on Circuits and Systems*. 1992.

[114] Shinji Ito. "Submodular Function Minimization with Noisy Evaluation Oracle". In: *Advances in Neural Information Processing Systems*. 2019, pp. 12103–12113.

[115] Rabih A. Jabr, Sami Karaki, and Joe Akl Korbane. "Robust Multi-Period OPF With Storage and Renewables". In: *IEEE Transactions on Power Systems* 30.5 (Sept. 2015), pp. 2790–2799. ISSN: 0885-8950, 1558-0679.

[116] Saber Jafarpour et al. "Flow and Elastic Networks on the n-Torus: Geometry, Analysis, and Computation". In: *SIAM Review* 64.1 (2022), pp. 59–104.

[117] Prateek Jain, Raghu Meka, and Inderjit Dhillon. "Guaranteed rank minimization via singular value projection". In: *Advances in Neural Information Processing Systems* 23 (2010).

[118] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. "Low-rank matrix completion using alternating minimization". In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. 2013, pp. 665–674.

[119] Kevin Jamieson et al. "lil'ucb: An optimal exploration algorithm for multi-armed bandits". In: *Conference on Learning Theory*. PMLR. 2014, pp. 423–439.

[120] Nanjing Jian. "Exploring and Exploiting Structure in Large Scale Simulation Optimization". In: *Ph. D. thesis* (2017). Operations Research and Information Engineering, Cornell University, Ithaca NY.

[121] Nanjing Jian et al. "Simulation optimization for a large-scale bike-sharing system". In: *2016 Winter Simulation Conference (WSC)*. IEEE. 2016, pp. 602–613.

[122] Haotian Jiang. "Minimizing convex functions with integral minimizers". In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2021, pp. 976–985.

[123] Haotian Jiang et al. "An improved cutting plane method for convex optimization, convex-concave games, and its applications". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 944–953.

[124] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. "Accelerated gradient descent escapes saddle points faster than gradient descent". In: *Conference On Learning Theory*. PMLR. 2018, pp. 1042–1085.

[125] Chi Jin et al. "How to escape saddle points efficiently". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1724–1732.

[126] Chi Jin et al. "On the local minima of the empirical risk". In: *Advances in neural information processing systems*. 2018, pp. 4896–4905.

[127] Zohar Karnin, Tomer Koren, and Oren Somekh. "Almost optimal exploration in multi-armed bandits". In: *International Conference on Machine Learning*. 2013, pp. 1238–1246.

[128] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. "On the complexity of best-arm identification in multi-armed bandit models". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1–42.

[129] Emilie Kaufmann and Shivaram Kalyanakrishnan. "Information complexity in bandit subset selection". In: *Conference on Learning Theory*. 2013, pp. 228–251.

[130] Samir Khuller. "Ear decompositions". In: *SIGACT News* 20.1 (1989), p. 128.

[131] Seong-Hee Kim and Barry L Nelson. "Selecting the best system". In: *Handbooks in operations research and management science* 13 (2006), pp. 501–534.

[132] Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de-Mello. "The sample average approximation method for stochastic discrete optimization". In: *SIAM Journal on Optimization* 12.2 (2002), pp. 479–502.

[133] N.M. Korneyenko. "Combinatorial algorithms on a class of graphs". In: *Discrete Applied Mathematics* 54 (1994), pp. 215–217.

[134] Andrew J Korsak. "On the question of uniqueness of stable load-flow solutions". In: *IEEE Transactions on Power Apparatus and Systems* 3 (1972), pp. 1093–1100.

[135] Pierre L'Ecuyer. "A unified view of the IPA, SF, and LR gradient estimation techniques". In: *Management Science* 36.11 (1990), pp. 1364–1383.

[136] Tze Leung Lai and Herbert Robbins. "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22.

[137] Javad Lavaei and Steven H. Low. "Zero Duality Gap in Optimal Power Flow Problem". In: *IEEE Transactions on Power Systems* 27.1 (Feb. 2012), pp. 92–107. ISSN: 0885-8950, 1558-0679.

[138] Jason D Lee et al. "Gradient descent only converges to minimizers". In: *Conference on learning theory*. PMLR. 2016, pp. 1246–1257.

[139] Jasper CH Lee and Paul Valiant. "Optimal Sub-Gaussian Mean Estimation in $\mathbb{R}$". In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 672–683.

[140] Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. "A faster cutting plane method and its implications for combinatorial and convex optimization". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 1049–1065.

[141] Arjen K Lenstra, Hendrik Willem Lenstra, and László Lovász. "Factoring polynomials with rational coefficients". In: *Mathematische annalen* 261.ARTICLE (1982), pp. 515–534.

[142] Matthieu Lerasle. "Selected topics on robust statistical learning theory". In: *arXiv preprint arXiv:1908.10761* (2019).

[143] Eitan Levin, Joe Kileel, and Nicolas Boumal. "The effect of smooth parametrizations on nonconvex optimization landscapes". In: *Mathematical Programming* (2024), pp. 1–49.

[144] Bowen Li, Ruiwei Jiang, and Johanna L. Mathieu. "Distributionally Robust Chance-Constrained Optimal Power Flow Assuming Unimodal Distributions With Misspecified Modes". In: *IEEE Transactions on Control of Network Systems* 6.3 (Sept. 2019), pp. 1223–1234. ISSN: 2325-5870, 2372-2533.

[145] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. "The non-convex geometry of low-rank matrix optimization". In: *Information and Inference: A Journal of the IMA* 8.1 (2019), pp. 51–96.

[146] Xiao Li et al. "Nonconvex robust low-rank matrix recovery". In: *SIAM Journal on Optimization* 30.1 (2020), pp. 660–686.

[147] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations". In: *Conference On Learning Theory*. PMLR. 2018, pp. 2–47.

[148] Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. "On zeroth-order stochastic convex optimization via random walks". In: *arXiv preprint arXiv:1402.2667* (2014).

[149] Eunji Lim. "Stochastic approximation over multidimensional discrete sets with applications to inventory systems and admission control of queueing networks". In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 22.4 (2012), pp. 1–23.

[150] Johan Lofberg. "YALMIP: A toolbox for modeling and optimization in MATLAB". In: *2004 IEEE international conference on robotics and automation (IEEE Cat. No. 04CH37508)*. IEEE. 2004, pp. 284–289.

[151] Raphael Louca and Eilyan Bitar. "Robust AC Optimal Power Flow". In: *IEEE Transactions on Power Systems* 34.3 (May 2019), pp. 1669–1681. ISSN: 0885-8950, 1558-0679.

[152] László Lovász. "Submodular functions and convexity". In: *Mathematical programming the state of the art*. Springer, 1983, pp. 235–257.

[153] Steven H. Low. "Convex Relaxation of Optimal Power Flow—Part I: Formulations and Equivalence". In: *IEEE Transactions on Control of Network Systems* 1.1 (Mar. 2014), pp. 15–27. ISSN: 2325-5870.

[154] Miles Lubin, Yury Dvorkin, and Scott Backhaus. "A Robust Approach to Chance Constrained Optimal Power Flow With Renewable Generation". In: *IEEE Transactions on Power Systems* 31.5 (Sept. 2016), pp. 3840–3849. ISSN: 0885-8950, 1558-0679.

[155] Jun Luo et al. "Fully sequential procedures for large-scale ranking-and-selection problems in parallel computing environments". In: *Operations Research* 63.5 (2015), pp. 1177–1194.

[156] Yuetian Luo, Xudong Li, and Anru R Zhang. "Nonconvex Factorization and Manifold Formulations are Almost Equivalent in Low-rank Matrix Optimization". In: *arXiv preprint arXiv:2108.01772* (2021).

[157] Cong Ma et al. "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3345–3354.

[158] Jianhao Ma and Salar Fattahi. "Sign-RIP: A Robust Restricted Isometry Property for Low-rank Matrix Recovery". In: *arXiv preprint arXiv:2102.02969* (2021).

[159] Sijia Ma and Shane G Henderson. "An efficient fully sequential selection procedure guaranteeing probably approximately correct selection". In: *2017 Winter Simulation Conference (WSC)*. IEEE. 2017, pp. 2225–2236.

[160] Sijia Ma and Shane G. Henderson. "Predicting the simulation budget in ranking and selection procedures". In: *ACM Transactions on Modeling and Computer Simulation* 29.3 (2019), Article 14, 1–25.

[161] Ziye Ma et al. "Sharp restricted isometry property bounds for low-rank matrix recovery problems with corrupted measurements". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 7. 2022, pp. 7672–7681.

[162] Ramtin Madani, Javad Lavaei, and Ross Baldick. "Convexification of power flow problem over arbitrary networks". In: *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE. 2015, pp. 1–8.

[163] Vien Mai and Mikael Johansson. "Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization". In: *International conference on machine learning*. PMLR. 2020, pp. 6630–6639.

[164] Oren Mangoubi and Nisheeth K Vishnoi. "Convex optimization with unbounded nonconvex oracles using simulated annealing". In: *Conference On Learning Theory*. PMLR. 2018, pp. 1086–1124.

[165] Karen Nan Miu and Hsiao-Dong Chiang. "Existence, uniqueness, and monotonic properties of the feasible power flow solution for radial three-phase distribution networks". In: *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* 47.10 (2000), pp. 1502–1514.

[166] Peyman Mohajerin Esfahani and Daniel Kuhn. "Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations". en. In: *Mathematical Programming* 171.1-2 (Sept. 2018), pp. 115–166. ISSN: 0025-5610, 1436-4646.

[167] Daniel K. Molzahn et al. "Implementation of a Large-Scale Optimal Power Flow Solver Based on Semidefinite Programming". In: *IEEE Transactions on Power Systems* 28.4 (Nov. 2013), pp. 3987–3998. ISSN: 0885-8950, 1558-0679.

[168] Kazuo Murota. "Discrete Convex Analysis". In: *Society for Industrial and Applied Mathematics*. Citeseer. 2003.

[169] Barry L Nelson. "Optimization via simulation over discrete decision variables". In: *Risk and Optimization in an Uncertain World*. Informs, 2010, pp. 193–207.

[170] Arkadiĭ Semenovich Nemirovsky and David Borisovich Yudin. "Problem complexity and method efficiency in optimization." In: (1983).

[171] Yurii Nesterov. *Lectures on convex optimization*. Vol. 137. Springer, 2018.

[172] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. "Phase retrieval using alternating minimization". In: *Advances in Neural Information Processing Systems* 26 (2013).

[173] Praneeth Netrapalli et al. "Non-convex robust PCA". In: *Advances in Neural Information Processing Systems* 27 (2014).

[174] Tibor Neubrunn. "Quasi-continuity". In: *Real Analysis Exchange* 14.2 (1988), pp. 259–306.

[175] Hung D Nguyen and Konstantin S Turitsyn. "Appearance of multiple stable load flow solutions under power flow reversal conditions". In: *2014 IEEE PES General Meeting— Conference & Exposition*. IEEE. 2014, pp. 1–5.

[176] Eric C. Ni et al. "Efficient ranking and selection in high performance computing environments". In: *Operations Research* 65.3 (2017), pp. 821–836.

[177] Eugene Ostrovsky and Leonid Sirota. "Exact value for subgaussian norm of centered indicator random variable". In: *arXiv preprint arXiv:1405.6749* (2014).

[178] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro. "Sample Average Approximation Method for Chance Constrained Programming: Theory and Applications". en. In: *Journal of Optimization Theory and Applications* 142.2 (Aug. 2009), pp. 399–416. ISSN: 0022-3239, 1573-2878.

[179] Ioannis Panageas and Georgios Piliouras. "Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions". In: *arXiv preprint arXiv:1605.00405* (2016).

[180] Zhan Pang, Frank Y Chen, and Youyi Feng. "A note on the structure of joint inventory-pricing control with leadtimes". In: *Operations Research* 60.3 (2012), pp. 581–587.

[181] Chuljin Park and Seong-Hee Kim. "Penalty function with memory for discrete optimization via simulation with stochastic constraints". In: *Operations Research* 63.5 (2015), pp. 1195–1212.

[182] Chuljin Park et al. "Designing an optimal water quality monitoring network for river systems using constrained discrete optimization via simulation". In: *Engineering Optimization* 46.1 (2014), pp. 107–129.

[183] Dohyung Park et al. "Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably". In: *SIAM Journal on Imaging Sciences* 11.4 (2018), pp. 2165–2204.

[184] SangWoo Park et al. "Uniqueness of power flow solutions using monotonicity and network topology". In: *IEEE Transactions on Control of Network Systems* 8.1 (2020), pp. 319–330.

[185] Sriram Pemmaraju and Steven Skiena. *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica®*. Cambridge university press, 2003.

[186] Bala Kameshwar Poolla et al. "Wasserstein Distributionally Robust Look-Ahead Economic Dispatch". In: *IEEE Transactions on Power Systems* 36.3 (May 2021), pp. 2010–2022. ISSN: 0885-8950, 1558-0679.

[187] Yu. V. Prokhorov. "Convergence of Random Processes and Limit Theorems in Probability Theory". en. In: *Theory of Probability & Its Applications* 1.2 (Jan. 1956), pp. 157–214. ISSN: 0040-585X, 1095-7219.

[188] Huashuai Qu and Michael C Fu. "Gradient extrapolated stochastic kriging". In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 24.4 (2014), pp. 1–25.

[189] Prasanna K Ragavan et al. "Adaptive Sampling line search for local stochastic optimization with integer variables". In: *Mathematical Programming* 196.1-2 (2022), pp. 775–804.

[190] Hamed Rahimian and Sanjay Mehrotra. "Distributionally robust optimization: A review". In: *arXiv preprint arXiv:1908.05659* (2019).

[191] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3 (2010), pp. 471–501.

[192] James Renegar. "Condition numbers, the barrier method, and the conjugate-gradient method". In: *SIAM Journal on Optimization* 6.4 (1996), pp. 879–912.

[193] James Renegar. "Linear programming, complexity theory and elementary functional analysis". In: *Mathematical Programming* 70.1 (1995), pp. 279–351.

[194] Line Roald and Göran Andersson. "Chance-Constrained AC Optimal Power Flow: Reformulations and Efficient Algorithms". In: *IEEE Transactions on Power Systems* 33.3 (2018), pp. 2906–2918.

[195] S. Zampieri S. Bolognani. "On the existence and linear approximation of the power flow solution in power distribution networks". In: *IEEE Transactions on Power Systems* 31 (Jan. 2016), pp. 163–172.

[196] Mark Semelhago et al. "Rapid Discrete Optimization via Simulation with Gaussian Markov random fields". In: *INFORMS journal on Computing* Articles in Advance (2020).

[197] Suvrajeet Sen and Julia L. Higle. "Stabilization of cutting plane algorithms for stochastic linear programming problemsStabilization of Cutting Plane Algorithms for Stochastic Linear Programming Problems". In: *Encyclopedia of Optimization*. Ed. by Christodoulos A. Floudas and Panos M. Pardalos. Boston, MA: Springer US, 2001, pp. 2434–2440. ISBN: 978-0-306-48332-5.

[198] Moshe Shaked and J George Shanthikumar. "Stochastic convexity and its applications". In: *Advances in Applied Probability* 20.2 (1988), pp. 427–446.

[199] Ohad Shamir. "A variant of azuma's inequality for martingales with subgaussian tails". In: *arXiv preprint arXiv:1110.2392* (2011).

[200] Yoav Shechtman et al. "Phase retrieval with application to optical imaging: a contemporary overview". In: *IEEE signal processing magazine* 32.3 (2015), pp. 87–109.

[201] Leyuan Shi et al. "Nested partitions method for stochastic optimization". In: *Methodology and Computing in Applied probability* 2.3 (2000), pp. 271–291.

[202] John W. Simpson-Porco. "A Theory of Solvability for Lossless Power Flow Equations – Part I: Fixed-Point Power Flow". In: *IEEE Transactions on Control of Network Systems* 5.3 (Sept. 2018), pp. 1361–1372.

[203] John W. Simpson-Porco. "A Theory of Solvability for Lossless Power Flow Equations – Part II: Conditions for Radial Networks". In: *IEEE Transactions on Control of Network Systems* 5.3 (Sept. 2018), pp. 1373–1385.

[204] Amit Singer. "Angular synchronization by eigenvectors and semidefinite programming". In: *Applied and computational harmonic analysis* 30.1 (2011), pp. 20–36.

[205] Divya Singhvi et al. "Predicting Bike Usage for New York City's Bike Sharing System." In: *AAAI Workshop: Computational Sustainability*. Citeseer. 2015.

[206] Somayeh Sojoudi, Salar Fattahi, and Javad Lavaei. "Convexification of generalized network flow problem". In: *Mathematical Programming* 173.1 (2019), pp. 353–391.

[207] 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). Publisher: SIAM, pp. 1746–1778.

[208] Somayeh Sojoudi and Javad Lavaei. "Physics of power networks makes hard optimization problems easy to solve". In: *2012 IEEE Power and Energy Society General Meeting*. IEEE, July 2012.

[209] Dominik Stöger and Mahdi Soltanolkotabi. "Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction". In: *Advances in Neural Information Processing Systems* 34 (2021).

[210] Ju Sun, Qing Qu, and John Wright. "A geometric analysis of phase retrieval". In: *Foundations of Computational Mathematics* 18.5 (2018), pp. 1131–1198.

[211] 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 (2016), pp. 853–884.

[212] Lihua Sun, L Jeff Hong, and Zhaolin Hu. "Balancing exploitation and exploration in discrete optimization via simulation through a Gaussian process-based search". In: *Operations Research* 62.6 (2014), pp. 1416–1438.

[213] Ruoyu Sun and Zhi-Quan Luo. "Guaranteed matrix completion via non-convex factorization". In: *IEEE Transactions on Information Theory* 62.11 (2016), pp. 6535–6579.

[214] J. Thorp, D. Schulz, and M. Ilic-Spong. "Reactive power-voltage problem: conditions for the existence of solution and localized disturbance propagation". In: *International Journal of Electrical and Power Energy Systems* 9 (Jan. 1986).

[215] James C Tiernan. "An efficient search algorithm to find the elementary circuits of a graph". In: *Communications of the ACM* 13.12 (1970), pp. 722–726.

[216] Tian Tong, Cong Ma, and Yuejie Chi. "Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent". In: *Journal of Machine Learning Research* 22.150 (2021), pp. 1–63.

[217] Tian Tong, Cong Ma, and Yuejie Chi. "Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number". In: *IEEE Transactions on Signal Processing* 69 (2021), pp. 2396–2409.

[218] Tian Tong et al. "Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements". In: *arXiv preprint arXiv:2104.14526* (2021).

[219] Stephen Tu et al. "Low-rank solutions of linear matrix equations via procrustes flow". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 964–973.

[220] Takeaki Uno and Hiroko Satoh. "An efficient algorithm for enumerating chordless cycles and chordless paths". In: *International Conference on Discovery Science*. Springer. 2014, pp. 313–324.

[221] Pravin M Vaidya. "A new algorithm for minimizing convex functions over convex sets". In: *Mathematical programming* 73.3 (1996), pp. 291–341.

[222] Bart P. G. Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. "From Data to Decisions: Distributionally Robust Optimization Is Optimal". en. In: *Management Science* 67.6 (June 2021), pp. 3387–3402. ISSN: 0025-1909, 1526-5501.

[223] Andreas Venzke et al. "Convex Relaxations of Chance Constrained AC Optimal Power Flow". In: *IEEE Transactions on Power Systems* 33.3 (May 2018), pp. 2829–2841. ISSN: 0885-8950, 1558-0679.

[224] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[225] Cong Wang et al. "Explicit Conditions on Existence and Uniqueness of Load-Flow Solutions in Distribution Networks". In: *IEEE Transactions on Smart Grid* 9.2 (Mar. 2018), pp. 953–962.

[226] Honggang Wang, Raghu Pasupathy, and Bruce W Schmeiser. "Integer-ordered simulation optimization using R-SPLINE: Retrospective search with piecewise-linear interpolation and neighborhood enumeration". In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 23.3 (2013), pp. 1–24.

[227] Lingxiao Wang, Xiao Zhang, and Quanquan Gu. "A unified computational and statistical framework for nonconvex low-rank matrix estimation". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 981–990.

[228] Tianxiang Wang et al. "Optimal Computing Budget Allocation for regression with gradient information". In: *Automatica* 134 (2021), p. 109927.

[229] Ke Wei et al. "Guarantees of riemannian optimization for low rank matrix completion." In: *Inverse Problems & Imaging* 14.2 (2020).

[230] Ke Wei et al. "Guarantees of Riemannian optimization for low rank matrix recovery". In: *SIAM Journal on Matrix Analysis and Applications* 37.3 (2016), pp. 1198–1222.

[231] Wei Wei, Feng Liu, and Shengwei Mei. "Distributionally Robust Co-Optimization of Energy and Reserve Dispatch". In: *IEEE Transactions on Sustainable Energy* 7.1 (Jan. 2016), pp. 289–300. ISSN: 1949-3029, 1949-3037.

[232] Ronald W Wolff and Chia-Li Wang. "On the convexity of loss probabilities". In: *Journal of applied probability* (2002), pp. 402–406.

[233] Laurence A Wolsey and George L Nemhauser. *Integer and combinatorial optimization*. Vol. 55. John Wiley & Sons, 1999.

[234] Felix F Wu. "Theoretical study of the convergence of the fast decoupled load flow". In: *IEEE transactions on power apparatus and systems* 96.1 (1977), pp. 268–275.

[235] Weijun Xie and Shabbir Ahmed. "Distributionally Robust Chance Constrained Optimal Power Flow with Renewables: A Conic Reformulation". In: *IEEE Transactions on Power Systems* 33.2 (Mar. 2018), pp. 1860–1867. ISSN: 0885-8950, 1558-0679.

[236] Jie Xu, Barry L Nelson, and Jeff L Hong. "Industrial strength COMPASS: A comprehensive algorithm and software for optimization via simulation". In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 20.1 (2010), pp. 1–29.

[237] Jingxu Xu and Zeyu Zheng. "Gradient-Based Simulation Optimization Algorithms via Multi-Resolution System Approximations". In: *INFORMS Journal on Computing* 35.3 (2023), pp. 633–651.

[238] Wendy Lu Xu and Barry L Nelson. "Empirical stochastic branch-and-bound for optimization via simulation". In: *Iie Transactions* 45.7 (2013), pp. 685–698.

[239] Yi Xu, Qihang Lin, and Tianbao Yang. "Accelerated stochastic subgradient methods under local error bound condition". In: *arXiv preprint arXiv:1607.01027* (2016).

[240] Lun Yang et al. "Tractable Convex Approximations for Distributionally Robust Joint Chance-Constrained Optimal Power Flow Under Uncertainty". In: *IEEE Transactions on Power Systems* 37.3 (May 2022), pp. 1927–1941. ISSN: 0885-8950, 1558-0679.

[241] Tianbao Yang and Qihang Lin. "Rsg: Beating subgradient method without smoothness and strong convexity". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 236–268.

[242] Xinyang Yi et al. "Fast algorithms for robust PCA via gradient descent". In: *Advances in neural information processing systems* 29 (2016).

[243] Alp Yurtsever et al. "Scalable semidefinite programming". In: *SIAM Journal on Mathematics of Data Science* 3.1 (2021), pp. 171–200.

[244] Haixiang Zhang, Yingjie Bi, and Javad Lavaei. "General low-rank matrix optimization: Geometric analysis and sharper bounds". In: *Advances in Neural Information Processing Systems* 34 (2021).

[245] Jialun Zhang, Salar Fattahi, and Richard Zhang. "Preconditioned Gradient Descent for Over-Parameterized Nonconvex Matrix Factorization". In: *Advances in Neural Information Processing Systems* 34 (2021).

[246] Jingzhao Zhang et al. "Complexity of finding stationary points of nonconvex nonsmooth functions". In: (2020), pp. 11173–11182.

[247] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. "Sharp Restricted Isometry Bounds for the Inexistence of Spurious Local Minima in Nonconvex Matrix Recovery." In: *J. Mach. Learn. Res.* 20.114 (2019), pp. 1–34.

[248] Richard Y Zhang et al. "How Much Restricted Isometry is Needed In Nonconvex Matrix Recovery?" In: *NeurIPS*. 2018.

[249] Siqi Zhang and Niao He. "On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization". In: *arXiv preprint arXiv:1806.04781* (2018).

[250] Yiling Zhang, Siqian Shen, and Johanna L Mathieu. "Distributionally robust chance-constrained optimal power flow with uncertain renewables and uncertain reserves provided by loads". In: *IEEE Transactions on Power Systems* 32.2 (2016), pp. 1378–1388.

[251] Yu Zhang, Ramtin Madani, and Javad Lavaei. "Conic relaxations for power system state estimation with line measurements". In: *IEEE Transactions on Control of Network Systems* 5.3 (2017), pp. 1193–1205.

[252] Qinqing Zheng and John Lafferty. "A convergent Gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*. 2015, pp. 109–117.

[253] Ying Zhong and L Jeff Hong. "Fully sequential ranking and selection procedures with PAC guarantee". In: *2018 Winter Simulation Conference (WSC)*. IEEE. 2018, pp. 1898–1908.

[254] Ying Zhong and L Jeff Hong. "Knockout-tournament procedures for large-scale ranking and selection in parallel computing environments". In: *Operations Research* (2021).

[255] Anping Zhou et al. "A linear programming approximation of distributionally robust chance-constrained dispatch with Wasserstein distance". In: *IEEE Transactions on Power Systems* 35.5 (2020), pp. 3366–3377.

[256] Zhihui Zhu et al. "Global optimality in low-rank matrix optimization". In: *IEEE Transactions on Signal Processing* 66.13 (2018), pp. 3614–3628.

[257] Zhihui Zhu et al. "The global optimization geometry of low-rank matrix optimization". In: *IEEE Transactions on Information Theory* 67.2 (2021), pp. 1308–1331.

[258] Ray D. Zimmerman and Carlos E. Murillo-Sánchez. *MATPOWER*. Language: en. Oct. 2020.

[259] Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sanchez, 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 (Feb. 2011), pp. 12–19. ISSN: 0885-8950, 1558-0679.

[260] Martin Zinkevich. "Online convex programming and generalized infinitesimal gradient ascent". In: *Proceedings of the 20th international conference on machine learning (icml-03)*. 2003, pp. 928–936.

[261] Paul Zipkin. "On the structure of lost-sales inventory models". In: *Operations research* 56.4 (2008), pp. 937–944.